On A Principal Varying Coefficient Model

Qian Jiang, Hansheng Wang, Yingcun Xia and Guohua Jiang

Abstract We propose a novel varying coefficient model, called principal varying coefficient model (PVCM), by characterizing the varying coefficients through linear combinations of a few principal functions. Compared with the conventional varying coefficient model (VCM; Chen and Tsay, 1993; Hastie and Tibshirani, 1993), PVCM reduces the actual number of nonparametric functions, and thus has better estimation efficiency. Compared with the semi-varying coefficient model (SVCM; Zhang et al, 2002; Fan and Huang, 2005), PVCM is more flexible but with the same estimation efficiency when the number of principal functions in PVCM and the number of varying coefficients in SVCM are the same. Model estimation and identification are investigated, and the better estimation efficiency is justified theoretically. Incorporating the estimation with the L_1-penalty, variables in the linear combinations can be selected automatically and hence the estimation efficiency can be further improved. Numerical experiments suggest that the model together with the estimation method are useful even when the number of covariates is large.

KEY WORDS: local linear estimator; L_1-penalty; principal function; profile least-squares estimation; semi-varying coefficient model; varying coefficient model.

Qian Jiang is a PhD student, the Department of Statistics, National University of Singapore. Hansheng Wang is a Professor of Statistics, Guanghua School of Management, Peking University, China. Yingcun Xia is a Professor of Statistics, Department of statistics, National University of Singapore, and a Professor of Finance, Nanjing University, China. Guohua Jiang is a Professor of Accounting, Guanghua School of Management, Peking University, China. Wang's work was supported in part by National Natural Science Foundation of China (11131002, 11271032), Fox Ying Tong Education Foundation, and the Center for Statistical Science at Peking University. Xia's work is partially supported by a grand from the National University of Singapore (R-155-000-121-112). Guohua Jiang's work was supported in part by the National Natural Science Foundation of China (71132004). The authors are very grateful for the helpful comments of 3 referees, the associate editor, and the editor.

1. INTRODUCTION

Let (Y, X, U) be a random triplet, where $Y \in \mathbb{R}^1$ is the response of interest, $X = (X_1, \cdots, X_p)^\top \in \mathbb{R}^p$ is the associated p-dimensional predictor, and $U \in \mathbb{R}^1$ is the so-called *index* variable. The conventional varying coefficient model (**??**, VCM) assumes that $Y = X^\top \boldsymbol{\beta}(U) + \varepsilon$, where ε is the random noise and $\boldsymbol{\beta}(u) = (\beta_1(u), ..., \beta_p(u))^\top \in \mathbb{R}^p$ is a vector of unknown smooth functions in u, called the varying coefficients. Ever since **?** and **?**, VCM has gained a lot of popularity in the literature attributing to the following three facts. Firstly, VCM is easy to interpret because conditioned on the index variable $U = u$, VCM reduces to a standard linear regression model which has been well understood in practice. Secondly, VCM allows the varying coefficient $\boldsymbol{\beta}(u)$ to be fully nonparametric. Thus, it has much stronger modeling capability than a standard linear regression model. Lastly, because the index variable U is typically univariate, VCM is free of the curse of dimensionality. VCM and its variants have been extensively studied in the literature during the past two decades (**?????????**).

It is remarkable that, although the estimation of VCM requires only univariate nonparametric smoothing, it is still very unstable when p is large or even moderately large, because there are p nonparametric functions to estimate. To improve the estimation efficiency, some estimation methods have been developed based on either kernel smoothing or splines smoothing, including Fan and Zhang (1999), Cheng and Hall (2003), Wu and Liang (2004), Huang et al (2002, 2004), Eubank et al (2004) and Kai et al (2011); their main idea is applying different smoothing parameter to different coefficients. However, the improvement based on their idea is limited especially when different coefficients need similar smoothing parameters. Another way to improve the efficiency is through further model specification without losing much information. The semi-varying coefficient model (SVCM) proposed by **?** and **?** is a good example for this purpose. SVCM confines some coefficients to be constant but allows the others to vary with the index variable U.

In this paper, we consider an extension of SVCM by allowing different varying coefficients to be linearly dependent and thus reduce the actual number of unknown functions in the model. To further illustrate the idea, let us revisit the Boston housing data. The response of interest is the median value of owner-occupied homes (MEDV, in \$1000) with 13 predictors, denoted by $X_1, ..., X_{13}$ respectively. More details will be stated in Section 4. As noticed by Fan and Huang (2006), the following varying coefficient model with the lower status of the population (U =LSTAT) being the index variable is appropriate for the data,

$$\text{MEDV} = \beta_1(U)X_1 + ... + \beta_{13}(U)X_{13} + \varepsilon. \tag{1.1}$$

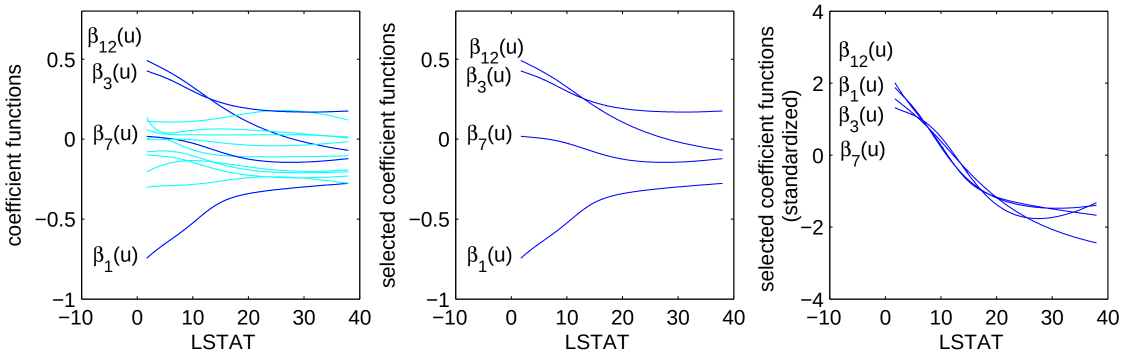

Figure 1: The estimated varying coefficients for the Boston housing dataset. The first panel shows all the coefficients, where coefficients with large variations (i.e., β_1, β_3, β_7, and β_{12}) are labeled and highlighted. For better visualization, those coefficients with large variations are redrawn in the second panel. To see similarity among those coefficients, their linearly transformed versions are shown in the third panel.

In (**??**), the varying coefficients can be estimated by the method based on the local linear smoothing; see for example **?** and Wu and Liang (2004). The estimated coefficients are shown in the first panel of Figure **??** where the coefficients with large variations are highlighted and labeled; those coefficients are redrawn in the second panel for better visualization. Remarkably similar shapes are discovered after linear transformations as shown in the third panel. The similarity implies that different varying coefficients are likely to be linearly dependent and that the index variable affects those coefficients in a similar manner.

Next, we quantify the above linear dependency amongst $\beta_1(U), ..., \beta_p(U)$ using the principal component analysis. Let $\theta = (\theta_1, \cdots, \theta_p)^\top = E\beta(U)$ with $p = 13$ in the above example and $\Sigma_\beta = \text{cov}\{\beta(U)\}$. Suppose the eigenvalue-eigenvector decomposition is

$$\Sigma_\beta = (b_1, ..., b_p) diag(\lambda_1, ..., \lambda_p)(b_1, ..., b_p)^\top$$

with $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$ and $b_k = (b_{k1}, ..., b_{kp})^\top$. Define the principal components for the varying coefficients as

$$\begin{pmatrix} g_1(U) \\ g_2(U) \\ \vdots \\ g_p(U) \end{pmatrix} = (b_1, ..., b_p)^\top \begin{pmatrix} \beta_1(U) - \theta_1 \\ \beta_2(U) - \theta_2 \\ \vdots \\ \beta_p(U) - \theta_p \end{pmatrix}.$$

Then it is easy to see that $E\{g_k(U)\} = 0$ and $Var(g_k(U)) = \lambda_k, k = 1, 2, ..., p$, and that

$$\beta_k(U) = \theta_k + b_{1k} g_1(U) + b_{2k} g_2(U) + ... + b_{pk} g_p(U), \quad k = 1, 2, ..., p. \qquad (1.2)$$

Because $Var(g_k(U))$ decreases with k, the contribution of $g_k(U)$ to the coefficient functions also decreases with k. In this example, the five largest eigenvalues are respectively 25.8584, 0.5668, 0.1445, 0.0370, and 0.0126. The rest $13 - 5 = 8$ eigenvalues are very close to 0. It is remarkable that the largest eigenvalue (i.e., 25.8584) by itself can explain about 97% of the total variation of $\beta_1(U), ..., \beta_p(U)$, which suggests that the first principal component contributes dominantly to the variation of $\beta_k(U)$'s; the others' contribution are very small. This fact motivates us to simplify model (??) into

$$\text{MEDV} = \Big(\theta_1 X_1 + \cdots + \theta_{13} X_{13}\Big) + \gamma_1(U)\Big(\phi_1 X_1 + \cdots + \phi_{13} X_{13}\Big) + \varepsilon. \tag{1.3}$$

Theoretically, the estimators produced by (??) are more efficient than those by (??) if the simplification does not lose much information, because only one nonparametric function $\gamma_1(.)$ needs to be estimated in (??) but a total of $p = 13$ functions need to be estimated in (??). Furthermore, model (??) identifies two important components given by $\theta_1 X_1 + \cdots + \theta_{13} X_{13}$ and $\phi_1 X_1 + \cdots + \phi_{13} X_{13}$ respectively. The first component is linearly related to the response, and the second nonlinearly in the sense that it has a nontrivial interaction with index variable U. Thus, model (??) is also more informative than model (??). See Section 4 for more discussions about this real example.

In this paper, we shall discuss a more general model of (??), called the principal varying coefficient model (PVCM). The rest of the article is organized as the follows. Next section introduces formally the model and discusses its identification. Model estimation and selection based on a profile approach is investigated in Section 3 theoretically. Incorporating the estimation with the adaptive L_1 penalty is studied in Section 4. Simulation studies are presented in Section 5 and the Boston housing data is further analyzed in Section 6. Lastly, the article is concluded with a brief discussion in Section 7. All technical details are left to the Appendix.

2. MODEL REPRESENTATION AND IDENTIFICATION

Let (Y_i, X_i, U_i) be the observation collected from the ith subject, $i = 1, 2, ..., n$, where $Y_i \in \mathbb{R}^1$ is the response, $X_i = (X_{i1}, \cdots, X_{ip})^\top \in \mathbb{R}^p$ is the p-dimensional predictor, and $U_i \in \mathbb{R}^1$ is the univariate index variable. The conventional VCM model assumes

$$Y_i = \beta_1(U_i)X_{i1} + \beta_2(U_i)X_{i2} + ... + \beta_p(U_i)X_{ip} + \varepsilon_i,$$

where $\beta_k(.), k = 1, ..., p$, are unknown coefficient functions and $E(\varepsilon_i|X_i, U_i) = 0$ almost surely. We can set $X_{i1} \equiv 1$ to allow the model to include an intercept function. Let $\boldsymbol{\beta}_0(u) = (\beta_1(u), \beta_2(u), ..., \beta_p(u))^\top$ be the true varying coefficients. Motivated by the example and (**??**) above, we consider a more general case where the variation in $\beta_1(U), ..., \beta_p(U)$ is dominantly explained by the first d_0 principal components while contribution of the other components is negligible, and thus assume the following principal component structure for the coefficient functions for modelling purpose,

$$\boldsymbol{\beta}_0(u) = \theta_0 + B_0\boldsymbol{\gamma}_0(u),$$

where $\theta_0 \in \mathbb{R}^p$ and $B_0 = (b_1, ..., b_{d_0}) \in \mathbb{R}^{p\times d_0}$, with $\text{rank}(B_0) = d_0 \leq p$, are parameters and $\boldsymbol{\gamma}_0(u) = (g_1(u), ..., g_{d_0}(u))^\top$ are unknown principal functions. As a consequence, we come up with the following principal varying coefficient model (PVCM)

$$Y_i = \theta_0^\top X_i + g_1(U_i)b_1^\top X_i + ... + g_{d_0}(U_i)b_{d_0}^\top X_i + \varepsilon_i. \tag{2.1}$$

For convenience, we refer to d_0 as the number of principal functions, $\theta_0^\top X_i$ as the linear part, and $X_i^\top B_0\boldsymbol{\gamma}_0(U_i)$ as the nonlinear part. We further assume that the principal functions $\boldsymbol{\gamma}_0(u) \in \mathbb{R}^{d_0}$ satisfy $\text{rank}\{\text{cov}(\boldsymbol{\gamma}_0(U_i))\} = d_0$. Otherwise, functional elements in $\boldsymbol{\gamma}_0(u)$ are still linearly dependent, and the rank of B_0 can be further reduced. Obviously, model (**??**) becomes a standard linear regression model if $d_0 = 0$, and a full VCM if $d_0 = p$. PVCM also includes SVCM of **?** as a special case if the last $p - q$ elements in θ_0 are zeros and the first q elements in all $b_k, k = 1, ..., d_0$, are zeros.

Compared with the conventional VCM (**????**), PVCM reduces the actual number of unknown nonparametric functions, and thus has better estimation efficiency. Compared with SVCM (**???**), PVCM is more flexible and informative by allowing a predictor to appear in both linear and nonlinear parts simultaneously. On the other hand, PVCM shares the same estimation efficiency with SVCM when the number of principal functions in PVCM and the number of varying coefficients in SVCM are the same.

Model (**??**) is not uniquely identifiable. For example, let C be an arbitrary $d_0 \times d_0$ orthonormal matrix. Then, by re-defining $B_0 := B_0 C$ and $\boldsymbol{\gamma}_0(u) := C^\top \boldsymbol{\gamma}_0(u)$, model (**??**) still holds. Parameter vector θ_0 is also not unique even if B_0 is fixed. For example, let $c \in \mathbb{R}^{d_0}$ be an arbitrary constant vector and re-define $\theta_0 := \theta_0 - B_0 c$ and $\boldsymbol{\gamma}_0(\cdot) := \boldsymbol{\gamma}_0(\cdot) + c$, then model (**??**) is still correct. To fix the identification problem, we can always appropriately select the vector c such that $E\boldsymbol{\gamma}_0(U) = 0$. If further $\mathrm{cov}\{\boldsymbol{\gamma}_0(U)\}$ is of full rank, we then have the following identification equations

$$\theta_0 = E\{\boldsymbol{\beta}_0(U)\}, \quad \mathcal{S}(B_0) = \mathcal{S}(\mathrm{cov}\{\boldsymbol{\beta}_0(U)\}), \tag{2.2}$$

where $\mathcal{S}(A)$ stands for the linear subspace spanned by the column vectors of an arbitrary matrix A. Because $\mathcal{S}(B_0) = \mathcal{S}(\Sigma_\beta)$ with $\Sigma_\beta = \mathrm{cov}\{\boldsymbol{\beta}_0(U)\}$, we can define $B_0 = (b_1, \cdots, b_{d_0}) \in \mathbb{R}^{p \times d_0}$, where b_j $(1 \leq j \leq d_0)$ are the eigenvectors associated with Σ_β's d_0 largest eigenvalues in descending order. In that case, $\boldsymbol{\gamma}_0(U) = B_0^\top\{\boldsymbol{\beta}_0(U) - E\boldsymbol{\beta}_0(U)\}$ with

$$E(\boldsymbol{\gamma}_0(U)) = 0, \qquad Cov(\boldsymbol{\gamma}_0(U)) = diag(\lambda_1, ..., \lambda_{d_0}), \qquad B_0^\top B_0 = I_{d_0}, \tag{2.3}$$

where $\lambda_1 \geq ... \geq \lambda_{d_0} > 0$. As long as the d_0 largest eigenvalues are mutually different,

B_0 is uniquely identifiable up to a sign difference. For convenience, we assume throughout the rest of this article that the non-zero eigenvalues of Σ_β are different from one another.

Based on (**??**) and (**??**), we can also give another way to identify the linear part. Write $\tilde{\theta}_0 = (I - B_0 B_0^\top)\theta_0 B_0^\top \theta_0$ and $\tilde{\boldsymbol{\gamma}}_0(U) = \boldsymbol{\gamma}_0(U) + B_0^\top \theta_0 \overset{def}{=} (\tilde{g}_1(U), ..., \tilde{g}_{d_0}(U))^\top$, then we have $\beta_0(U) = \theta_0 + B_0 \boldsymbol{\gamma}_0(U) = \tilde{\theta}_0 + B_0 \tilde{\boldsymbol{\gamma}}_0(U)$, such that

$$B_0^\top \tilde{\theta}_0 = 0, \qquad B_0^\top B_0 = I_{d_0}, \qquad Cov(\tilde{\boldsymbol{\gamma}}_0(U)) = diag(\lambda_1, ..., \lambda_{d_0}). \tag{2.4}$$

By (**??**), it is easy to see that $\tilde{\theta}_0$ and $\mathcal{S}(B_0)$ satisfying (**??**) are identifiable. This way of identifying the model is preferable because it has less parameters when $E\{\boldsymbol{\beta}_0(U)\} \in \mathcal{S}(\mathrm{cov}\{\boldsymbol{\beta}_0(U_i)\})$, in which case $\theta_0 = 0$. This fact will be used in our test for whether there exists a linear combination of X whose coefficient does not change with U. This fact can also be used to test whether there are constant coefficients in SVCM (**?**). It should be noted that the identification conditions $E(\gamma_0(U)) = 0$ in (**??**) and $B_0^\top \tilde{\theta}_0 = 0$ in (**??**) should not be used simultaneously. Otherwise, the PVCM model might be overspecified.

We end this section by mentioning some work in the literature that is related to principal functions. Factor models or principal component analysis that extract the main informative variables from a large number of variables are powerful approaches towards multivariate analysis. However, most of the existing models are under linear settings or under nonlinear framework; see for example Stock and Watson (2002) and Hastie and Stuetzle (1989). Our approach is under a functional framework. On the other hand, the principal components in our model are constructed for unobserved functions which is different from the fact that the usual factor models are proposed for observed data; see for example Forni et al (2000), Froni and Lippi (2001) and Bai (2003).

3. PROFILE LEAST-SQUARE ESTIMATION OF PVCM

In this section, we investigate the model estimation using the kernel smoothing approach. Estimation based on other nonparametric smoothing methods such as splines and penalized polynomial splines can be investigated similarly.

We firstly consider the estimation of θ_0 and B_0 under the assumption d_0 is known in advance. The estimation of d_0 will be addressed later. Equation (**??**) motivates a very convenient way to estimate θ_0 and B_0. Specifically, by the local linear estimation (**?**) we can estimate $\boldsymbol{\beta}_0(u)$ by $\hat{\boldsymbol{\beta}}(u)$, where $\hat{\boldsymbol{\beta}}(u)$ is the minimizer of a in

$$\min_{a\in\mathbb{R}^p, b\in\mathbb{R}^p} n^{-1}\sum_{i=1}^{n}\Big\{Y_i - a^\top X_i - b^\top X_i(U_i-u)\Big\}^2 K_h(U_i-u), \tag{3.1}$$

where $K_h(u) = K(u/h)/h$ and $K(\cdot)$ is a kernel function. Consequently, we estimate Σ_β by $\hat{\Sigma}_\beta = n^{-1}\sum\{\hat{\boldsymbol{\beta}}(U_i)-\bar{\beta}\}\{\hat{\boldsymbol{\beta}}(U_i)-\bar{\beta}\}^\top$, where $\bar{\beta} = n^{-1}\sum\hat{\boldsymbol{\beta}}(U_i)$. We then estimate θ_0 by $\theta^{(0)} \overset{def}{=} \bar{\beta}$ and B_0 by $B^{(0)} \overset{def}{=} (\hat{b}_1^{(0)}, \cdots, \hat{b}_{d_0}^{(0)})$, where $\hat{b}_j^{(0)}$ is the eigenvector associated with the jth largest eigenvalue of $\hat{\Sigma}_\beta$ for $1 \le j \le d_0$. Let A be an arbitrary matrix and vec(A) stand for a vector constructed by stacking A's columns. Denote by $\|A\|$ the operation norm, i.e., the maximal absolute singular value of A. The estimation error for $B^{(0)}$ can be then defined as $\|\hat{B}^{(0)}(\hat{B}^{(0)})^\top - B_0B_0^\top\|$. We have the following consistency for the estimators.

Theorem 1. *Under the conditions (C.1)–(C.4) in the Appendix, we have* $\|\theta^{(0)} - \theta_0\| = O_p\{h^2 + (nh/\log(n))^{-1/2}\}$ *and* $\|\hat{B}^{(0)}(\hat{B}^{(0)})^\top - B_0B_0^\top\| = O_p\{h^2 + (nh/\log(n))^{-1/2}\}$.

Fix B and θ in model (**??**), and consider the local linear smoother of model (**??**)

$$\min_{a(u)\in\mathbb{R}^p, b(u)\in\mathbb{R}^p} \sum_{i=1}^{n} \{Y_i - X_i^\top\theta - a(u)^\top B^\top X_i - b(u)^\top B^\top X_i(U_i-u)/h\}^2 K_h(U_i-u).$$

If B and θ are close to the true values, then the minimizer of $a(u)$ is a local linear estimator of the coefficient functions $\boldsymbol{\gamma}_0(u)$, denoted by $\hat{\boldsymbol{\gamma}}(u|B,\theta)$. By Fan and Zhang (1999), we have

$$\begin{aligned}\hat{\boldsymbol{\gamma}}(u|B,\theta) \;&=\; \{S_n(u,B)\}^{-1}B^\top[L_{n,0}(u) - S_{n,0}(u)\theta \\ &\qquad -S_{n,1}(u)B(B^\top S_{n,2}(u)B)^{-1}B^\top(L_{n,1}(u) - S_{n,1}(u)\theta)],\end{aligned} \tag{3.2}$$

where $S_{n,k}(u) = \sum_{i=1}^n K_h(U_i-u)\{(U_i-u)/h\}^k X_i X_i^\top$, $L_{n,k}(u) = \sum_{i=1}^n K_h(U_i-u)\{(U_i-u)/h\}^k X_i y_i$ for $k = 0, 1, 2$ and $S_n(u,B) = B^\top\{S_{n,0}(u) - S_{n,1}(u)B(B^\top S_{n,2}(u)B)^{-1}B^\top \times S_{n,1}(u)\}B$. Let $\bar{\boldsymbol{\gamma}}(B,\theta) = n^{-1}\sum_{i=1}^n \hat{\boldsymbol{\gamma}}(U_i|B,\theta)$ and $\tilde{\boldsymbol{\gamma}}(u|B,\theta) = \hat{\boldsymbol{\gamma}}(u|B,\theta) - \bar{\boldsymbol{\gamma}}(B,\theta)$. Substituting $\tilde{\boldsymbol{\gamma}}(U_i|\theta,B)$ into model (**??**), we have

$$Y_i \approx X_i^\top\theta + X_i^\top B\tilde{\boldsymbol{\gamma}}(U_i|\theta,B) + \varepsilon_i, \quad i = 1, 2, ..., n.$$

Thus, we consider

$$Q(\theta,B) = n^{-1}\sum_{i=1}^n \Big\{Y_i - X_i^\top\theta - X_i^\top B\tilde{\boldsymbol{\gamma}}(U_i|\theta,B)\Big\}^2,$$

and estimate θ_0 and B_0 by

$$(\hat{\theta},\hat{B}) = \arg\min_{\theta,B} Q(\theta,B). \tag{3.3}$$

Although the minimization is searched over the whole space, as in many model estimations, an initial estimator is sometimes essential. Estimator $\theta^{(0)}$ and $B^{(0)}$ can be used for this purpose. Other robust estimation method such as the back-fitting method of Wu and Liang (2004) is also helpful to find initial estimators. To facilitate the theoretical investigation, Theorem **??** allows us to restrict the parameter space in a small range of the true parameters, $\Theta_n = \{(\theta, B) : ||\theta - \theta_0|| + ||B - B_0|| \leq M(h^2 + \delta_n)\}$ for some constant $M > 0$. In what follows, $A \otimes B$ denotes the Kronecker product of two matrix A and B, and the notation $A^{\otimes 2}$ denotes $AA^\top$ for any matrix A.

Theorem 2. *Suppose conditions (C.1)–(C.4) in the Appendix hold. Let* $(\hat{\theta}, \hat{B}) = arg\min_{(\theta,B)\in\Theta_n} Q_n(\theta, B)$. *Then*

$$\sqrt{n}\begin{pmatrix} \hat{\theta} - \theta_0 \\ vec(\hat{B} - B_0) \end{pmatrix} \overset{D}{\to} N\{0, \Sigma_0^{-1}(\Sigma_1 + \Sigma_2)\Sigma_0^{-1}\}$$

in distribution, where

$$\Sigma_0 = E\Bigg\{ \begin{pmatrix} X \\ \boldsymbol{\gamma}_0(U) \otimes X \end{pmatrix} \begin{pmatrix} X \\ \boldsymbol{\gamma}_0(U) \otimes X \end{pmatrix}^\top \Bigg\},$$

$$\Sigma_1 = E\left\{ \Bigg[\Bigg\{ \begin{pmatrix} I_p \\ \boldsymbol{\gamma}_0(U) \otimes I \end{pmatrix} - \Bigg(\begin{pmatrix} W(U) \\ \boldsymbol{\gamma}_0(U) \otimes W(U) \end{pmatrix} - E \begin{pmatrix} W(U) \\ \boldsymbol{\gamma}_0(U) \otimes W(U) \end{pmatrix} \Bigg) V(U) \Bigg\} X\varepsilon \Bigg]^{\otimes 2} \right\},$$

$$\Sigma_2 = E \begin{pmatrix} W(U) \\ \boldsymbol{\gamma}_0(U) \otimes W(U) \end{pmatrix} B_0 E\{\boldsymbol{\gamma}_0(U)\boldsymbol{\gamma}_0^\top(U)\} B_0^\top E \begin{pmatrix} W(U) \\ \boldsymbol{\gamma}_0(U) \otimes W(U) \end{pmatrix}^\top,$$

with $W(U) = E(XX^\top|U)$ *and* $V(U) = B_0(B_0^\top W(U) B_0)^{-1} B_0^\top$.

After θ_0 and B_0 are estimated, we can estimate $\boldsymbol{\gamma}_0(u)$ immediately by $\hat{\boldsymbol{\gamma}}(u|\hat{\theta}, \hat{B})$ defined in (**??**) and have the following limiting distribution.

Theorem 3. *Under regularity conditions (C.1)-(C.4) in the Appendix, we have in distribution*

$$\sqrt{nh\hat{f}(u)}\Big\{\hat{\boldsymbol{\gamma}}(u|\hat{B},\hat{\theta}) - \boldsymbol{\gamma}_0(u) - \frac{1}{2}\mu_2\boldsymbol{\gamma}_0''(u)h^2\Big\} \overset{D}{\to}$$
$$N\Big(0, \{B_0^\top W(u)B_0\}^{-1}B_0^\top W_2(u)B_0\{B_0^\top W(u)B_0\}^{-1}\Big),$$

where $W_2(u) = \int K^2(v)dvE\{XX^\top\varepsilon^2|U=u\}$, $\mu_2 = \int v^2K(v)dv$ *and* $\hat{f}(u) = n^{-1}\sum_{i=1}^n$ $K_h(U_i - u)$.

Writing the model as a VCM, the estimated coefficient functions are $\hat{\beta}_{PVCM}(u) = \hat{\theta} + \hat{B}\hat{\boldsymbol{\gamma}}(u|\hat{B},\hat{\theta})$. It follows from Theorems **??** and **??** that

$$\sqrt{nh\hat{f}(u)}\{\hat{\beta}_{PVCM}(u) - \boldsymbol{\beta}_0(u) - \frac{1}{2}\mu_2\boldsymbol{\beta}_0''(u)h^2\} \overset{D}{\to} N\{0, \Sigma_{PVCM}(u)\},$$

where $\Sigma_{PVCM}(u) = B_0\{B_0^\top W(u)B_0\}^{-1}B_0^\top W_2(u)B_0\{B_0^\top W(u)B_0\}^{-1}B_0$. However, if we treat the model as a VCM in the estimation process and estimate it by the method in Fan and Zhang (1999), then the estimator $\hat{\beta}_{VCM}(u)$ has

$$\sqrt{nh\hat{f}(u)}\{\hat{\beta}_{VCM}(u) - \boldsymbol{\beta}_0(u) - \frac{1}{2}\mu_2\boldsymbol{\beta}_0''(u)h^2\} \overset{D}{\to} N\{0, \Sigma_{VCM}(u)\},$$

where $\Sigma_{VCM}(u) = \{W(u)\}^{-1}W_2(u)\{W(u)\}^{-1}$; see **?**. If $d_0 < p$, it is easy to see that

$$\Sigma_{PVCM}(u) < \Sigma_{VCM}(u),$$

indicating that the estimator based on a PVCM is indeed more efficient than that based on a VCM. The smaller d_0 is, the more efficient is PVCM compared with VCM.

To make statistical inference, we also need to estimate the variance-covariance matrices in the limiting distributions. These matrices can be estimated simply by their sample versions with the unknown functions and parameters being replaced by their estimators respectively. By the local linear kernel smoothing, $W(u)$ can be estimated consistently by

$$\hat{W}(u) = \sum_{i=1}^{n} w_{n,h}(U_i - u) X_i X_i^\top / \sum_{i=1}^{n} w_{n,h}(U_i - u),$$

where $w_{n,h}(U_i - u) = K_h(U_i - u) \sum_{i=1}^n K_h(U_i - u)\{(U_i - u)/h\}^2 - K_h(U_i - u)\{(U_i - u)/h\} \sum_{i=1}^n K_h(U_i - u)\{(U_i - u)/h\}$, and $E\{XX^\top \varepsilon^2 | U = u\}$ by

$$\sum_{i=1}^{n} w_{n,h}(U_i - u) X_i X_i^\top \{Y_i - X_i^\top \hat{\theta} - \hat{\gamma}(U_i)\hat{B}^\top X_i\}^2 / \sum_{i=1}^{n} w_{n,h}(U_i - u).$$

As an example of hypothesis testing, we consider whether there is a separate linear part in the model under identification (**??**), i.e. whether there exists a linear combination $\theta_0^\top X$ such that $\theta_0^\top B_0 = 0$ and $\theta_0 \neq 0$. The corresponding hypothesis is

$$H_0: \quad (I - B_0 B_0^\top)\theta_0 \neq 0.$$

With the identification of (**??**), we can construct a test statistic

$$ST = n(\hat{\theta} - \theta_0)^\top \hat{P}(\hat{P} S_{00} \hat{P})^+ \hat{P}(\hat{\theta} - \theta_0),$$

where $\hat{P} = (I - \hat{B}\hat{B}^\top)$ and S_{00} is the submatrix of estimated $\Sigma_0^{-1}(\Sigma_1 + \Sigma_2)\Sigma_0^{-1}$ in its first p rows and first p columns, and A^+ denotes the Moore-Penrose inverse of matrix A.

Corollary 1. *Under the model assumptions (C.1) and (C.4) and H_0, with identification (**??**) we have $ST \xrightarrow{D} \chi^2(p - d_0)$ as $n \to \infty$.*

By Corollary 1, we reject H_0 if $ST > \chi^2_{1-\alpha}(p - d_0)$ with significance level α.

Next, we consider the estimation of d_0. To this end, we propose here a BIC-type criterion,

$$\text{BIC}(d) = \log \hat{\sigma}_d^2 + d \times \frac{\log(nh)}{nh}, \tag{3.4}$$

where d is the working number of principal functions, nh is the effective sample size in nonparametric regression, and $\hat{\sigma}_d^2$ is given by

$$\hat{\sigma}_d^2 = n^{-1} \sum_{k=1}^{n} \left\{ Y_k - X_k^\top \hat{\theta} - \hat{\gamma}^\top(U_k) \hat{B}^\top X_k \right\}^2,$$

where estimators $\hat{\theta}$, $\hat{B}$ and $\hat{\gamma}(U_i)$ are all obtained under the working number, d, of principal functions. For the purpose of completeness, define $BIC(0) = n^{-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$ with $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i$. Then d_0 is estimated by $\hat{d} = \text{argmin}_{0 \le d \le p} \text{BIC}(d)$.

Theorem 4. *Assuming the technical conditions (C.1)–(C.4) in the Appendix hold, we have* $P(\hat{d} = d_0) \to 1$.

By Theorem **??**, it is also easy to see that Theorems 1-3 still hold if we replace d_0 by $\hat{d}$.

4. REFINEMENT OF ESTIMATION BASED ON L_1 PENALTY

In this section, we estimate the model by incorporating the kernel smoothing with the L_1 penalty. As well demonstrated in the literature, the L_1 penalty approach has several advantages. Specifically for PVCM, the L_1 penalty can achieve the following goals simultaneously. (1) To identify variables that have cross effect with the index variable on the response, and those that only have simple linear effect. (2) To identify unimportant variables and automatically remove them from the model. (3) To improve the estimation efficiency when there is sparsity and the number of covariates is large.

Let $\alpha = (\alpha_1, ..., \alpha_{p(d_0+1)})^\top = (\theta^\top, vec(B)^\top)^\top$, $S = \{1, 2, \ldots, p(d_0+1)\}$ and $\mathcal{A} = \{s \in S : \alpha_s \neq 0\}$. Then, $\mathcal{A}$ is the index set that contains only nonzero elements in α. Following Zou (2006) and Zhang and Lu (2007), consider the following adaptive LASSO estimation,

$$\begin{aligned} \tilde{\alpha}^{(n)} = \{\tilde{\theta}_n^\top, vec(\tilde{B}_n)^\top\}^\top &= \arg\min_{(\theta,B)} \Big\{ Q(\theta, B) + \lambda_n \sum_{i=1}^{p} (\hat{w}_i |\theta_i| + \sum_{j=1}^{d_0} \hat{w}_{ij} |B_{ij}|) \Big\} \\ &= \arg \min_{\alpha \in \mathcal{R}^{p(d_0+1)}} \Big\{ Q(\alpha) + \lambda_n \sum_{s=1}^{p(d_0+1)} \hat{w}_s |\alpha_s| \Big\}, \qquad (4.1) \end{aligned}$$

where $\hat{w}_s = 1/|\hat{\alpha}_s|^\tau$ with $\tau > 0$ and $\hat{\alpha}_s$ is the estimator of α_s defined in (**??**). Let $\mathcal{A}_n = \{s \in S : \tilde{\alpha}_s^{(n)} \neq 0\}$. Then $\mathcal{A}_n$ is the index set of variables that are selected in either the linear part or nonlinear part of PVCM or both. If a variable is selected neither in the linear nor in the nonlinear part, the variable is unimportant and will be removed automatically from the model.

Theorem 5. *Under the conditions of Theorem* **??** *and* $\lambda_n/\sqrt{n} \to 0$, $\lambda_n n^{\frac{\tau-1}{2}} \to \infty$, *we have the following asymptotic properties for estimators* $\tilde{\theta}_n$ *and* $\tilde{B}_n$.

(1) The coefficients with nonzero values in both θ_0 *and* B_0 *can be consistently identified, i.e.* $\lim_{n\to\infty} P(\mathcal{A}_n = \mathcal{A}) = 1$.

(2) The estimated parameters achieve the oracle efficiency where the zero coefficients are known and removed in advance, i.e.

$$\sqrt{n} \begin{pmatrix} \tilde{\theta} - \theta_0 \\ vec(\tilde{B} - B_0) \end{pmatrix}_{\mathcal{A}} \xrightarrow{D} N\Big\{0, \Big((\Sigma_0)_{\mathcal{A}}\Big)^{-1} \Big(\Sigma_1 + \Sigma_2\Big)_{\mathcal{A}} \Big((\Sigma_0)_{\mathcal{A}}\Big)^{-1}\Big\}, \qquad (4.2)$$

where notation $M_{\mathcal{A}}$ denotes the submatrix of M with jth row (and jth column if M is a matrix) being removed for all $j \in \mathcal{A}^c$, complement of set A.

The selection of the tuning parameter λ_n is essential in the estimation. We found the commonly used BIC criterion works well, which is stated below. To indicate the dependence of the estimators on λ_n, write the estimators in (**??**) as $\tilde{\theta}_{\lambda_n}$ and $\tilde{B}_{\lambda_n}$ respectively. Define

$$BIC(\lambda_n) = \log\{Q(\tilde{\theta}_{\lambda_n}, \tilde{B}_{\lambda_n})\} + \log(n)\frac{p_n}{n},$$

where p_n is the total number of nonzero values in $\tilde{\theta}_{\lambda_n}$ or $\tilde{B}_{\lambda_n}$. The asymptotic performance of $BIC(\lambda_n)$ in selecting λ_n can be similarly discussed as in Wang and Xia (2009). The details are omitted here.

5. SIMULATION STUDIES

Consider two varying coefficient models where the covariates $X_{i1} \equiv 1$, and X_{ij}s $(1 < j \le p)$ are simulated from a multivariate normal distribution with $\mathrm{cov}(X_{ij_1}, X_{ij_2}) = 0.5^{|j_1 - j_2|}$ for any $j_1, j_2 \ge 2$, and U_i is simulated from $U[0,1]$, and ε_i from $N(0,1)$. The parameters and principal functions are respectively

$$\begin{aligned} &\text{Model 1.} \qquad \theta_0 = b_0, \quad B_0 = b_1, \quad \gamma_0(u) = 10u(1-u) - 5/3, \\ &\text{Model 2.} \quad \theta_0 = b_0, \quad B_0 = (b_2, b_3), \quad \gamma_0(u) = \{\cos(2\pi u), \sin(2\pi u)\}^\top, \end{aligned}$$

where $b_0 = (\underbrace{1, 1, ..., 1}_{7}, 0, ..., 0)^\top$, $b_1 = (\underbrace{1, -1, ..., 1, -1}_{[(p-1)/3]}, 0, ..., 0)^\top$, $b_2 = (\underbrace{1, ..., 1}_{[(p-1)/3]}, 0, ..., 0)^\top$ and $b_3 = (\underbrace{0, ..., 0}_{[(p-1)/3]}, \underbrace{1, ..., 1}_{[(p-1)/3]}, 0..., 0)^\top$. It is easy to see that Model 1 has 1 principal function ($d_0 = 1$) and Model 2 has 2 ($d_0 = 2$).

In the following calculation, we use the Newton-Rahpson algorithm to solve the minimization problem in (**??**). For the minimization in (**??**), we use the quadratic norm to approximate the L_1 norm and then the Newton-Rahpson algorithm to solve the minimization numerically.

For each model setting, 500 simulation replications are conducted. For each simulation replication, we first estimate the varying coefficients $\hat{\boldsymbol{\beta}}(u)$ according to (**??**) by treating the model as a VCM. See Fan and Zhang (1999) for more details. In the estimation, bandwidth h is selected by the leave-one-out cross-validation. The same bandwidth is then used throughout the rest of the computational process, except for the estimation of B_0 and θ_0 where the bandwidth is multiplied by $n^{-0.1}$ for the purpose of undersmoothing; see **?**. We then apply the BIC criterion in (**??**) to estimate the number of principal functions, $\hat{d}$. The percentage of replications in which the number of principal functions is correctly estimated is summarized in the third column of Table 1. As we can see, the percentage converges to 100% quickly when sample size increases. This convergency supports the theory that $\hat{d}$ is a consistent estimator of d_0.

As shown in Theorem 5, the estimation method in (**??**) can also be used for variable selection. To check the performance, we count in each estimation the number of zero rows (i.e. the rows in which all elements are zeros) in the estimated θ and B respectively. The numbers are listed in the fourth and fifth columns of Table 1. Note that if a row of estimated θ is zero, it means the corresponding variable is not selected in the linear part; if all the elements in a row of B are zero, it means the corresponding variable is removed from the nonlinear part. By comparing the numbers with true numbers of zeros in θ_0 and B_0 respectively as given in the square brackets of the table, we see that as sample size increases, the estimation method in (**??**) is consistent in selecting the variables in the linear part and nonlinear part.

We evaluate the overall performance of model estimation by checking the estimation error of the coefficients $\boldsymbol{\beta}(u)$ after rewriting the estimated model as a VCM. With the estimated d_0, we compute $\hat{\theta}$ and $\hat{B}$ and thus $\hat{\boldsymbol{\beta}}(u_i) = \hat{\theta} + \hat{B}\hat{\boldsymbol{\gamma}}(u_i)$. The estimation error of the whole model is evaluated by

$$n^{-1}\sum_{i=1}^{n}\left|\hat{\boldsymbol{\beta}}(u_i) - \theta_0 - B_0\boldsymbol{\gamma}_0(u_i)\right|,$$

where $|\ell| = (|\ell_1| + |\ell_2| + ... + |\ell_p|)/p$ for any vector $\ell = (\ell_1, ..., \ell_p)^\top$. The average estimation errors across 500 simulation replications are summarized in columns 6, 7 and 8 of Table 1. In these columns, as the sample size increases, the estimation error steadily shrinks towards 0. These shrinkages support that all the estimators are consistent. However, treating a PVCM as a VCM, the estimation efficiency is very much adversely affected by noticing that column 6 is obviously bigger than column 7. By comparing the eighth column with the seventh column, we can see that imposing the adaptive L_1 penalty, the estimation efficiency can be substantially improved, especially when the number of covariates is large.

Next, we check the performance of the proposed statistic in Corollary 1 for testing hypothesis on the linear part. We allow the linear part θ_0 to change with c, i.e. $\theta_0 = c \times b_0$. The larger c is, the more influential the linear part is. We also vary the signal-to-noise ratios (SNR) by changing the variance of ε. With significance level $\alpha = 0.05$, we calculate the rejection frequencies for $H_0 : |\theta_0| = 0$ under model specification (**??**). In both models, when $c = 0$ there are no linear parts, and thus the rejection frequency should be around 0.05. As c increases, the rejection frequencies should also increase. For the two models with $p = 7$, our simulation results for $c = 0, 0.05, 0.1, 0.15$ and 0.2 reported in Figure **??** support our theory quite well, indicating that the hypothesis testing statistic has reasonable power with roughly correct significant level. It is also

Table 1: Estimation results based on 500 replications

			correct (and incorrect)		estimation errors		
Model	sample	correct	zeros in the rows of		(and their standard error)		
and (p)	Size	d_0	θ	B	VCM	PVCM	PVCM+L_1
I($p=7$)	100	98%	0.0(0.0)	4.5(0.0)	0.2287	0.1741	0.1409
					(0.0411)	(0.0457)	(0.0340)
	200	100%	0.0(0.0)	4.9(0)	0.1578	0.1121	0.0910
					(0.0270)	(0.0249)	(0.0243)
	500	100%	0.0(0.0)	5.0(0.0)	0.0972	0.0742	0.0576
					(0.0149)	(0.0140)	(0.0139)
			[0(0)]	[5(0)]			
II($p=7$)	100	90%	0(0)	2.9(0.1)	0.2584	0.2129	0.1887
					(0.0494)	(0.0399)	(0.0363)
	200	100%	0.0(0.0)	3.0(0.0)	0.1721	0.1407	0.1243
					(0.0275)	(0.0271)	(0.0273)
	500	100%	0(0)	3.0(0.0)	0.1117	0.0873	0.0861
					(0.0137)	(0.0132)	(0.0135)
			[0(0)]	[3(0)]			
I($p=13$)	100	93%	5.9(0.0)	8.8(0.1)	0.2796	0.2114	0.0998
					(0.0530)	(0.0449)	(0.0378)
	200	100%	6.0(0.0)	9.0(0.0)	0.1749	0.1327	0.0617
					(0.0216)	(0.0226)	(0.0151)
	500	100%	6.0(0.0)	9.0(0.0)	0.1030	0.0694	0.0365
					(0.0130)	(0.0110)	(0.0081)
			[6(0)]	[9(0)]			
II($p=13$)	100	86%	5.3(0.4)	4.9(1.8)	0.3701	0.2651	0.2273
					(0.0782)	(0.0476)	(0.0393)
	200	97%	5.6(0.0)	5(0.3)	0.2094	0.1478	0.1161
					(0.0298)	(0.0201)	(0.0211)
	500	100%	6.0(0.0)	5.0(0.0)	0.1241	0.0884	0.0759
					(0.0122)	(0.0101)	(0.0105)
			[6(0)]	[5(0)]			
I($p=21$)	100	72%	13.8(0.1)	14.4(2)	0.3409	0.2919	0.1180
					(0.0867)	(0.0931)	(0.0551)
	200	99%	14.0(0.0)	15.0(0.0)	0.1878	0.1400	0.0485
					(0.0231)	(0.0217)	(0.0191)
	500	100%	14.0(0.0)	15.0(0.0)	0.1124	0.0719	0.0298
					(0.0099)	(0.0102)	(0.0064)
			[14(0)]	[15(0)]			
II($p=21$)	100	84%	12.0(0.4)	9.0(5.7)	0.5395	0.4305	0.3197
					(0.1045)	(0.1025)	(0.0510)
	200	92%	13.5(0.0)	9.0(1.0)	0.2559	0.1704	0.1242
					(0.0300)	(0.0214)	(0.0173)
	500	100%	13.9(0.1)	9.0(0.2)	0.1334	0.0876	0.0584
					(0.0104)	(0.0082)	(0.0082)
			[14(0)]	[9(0)]			

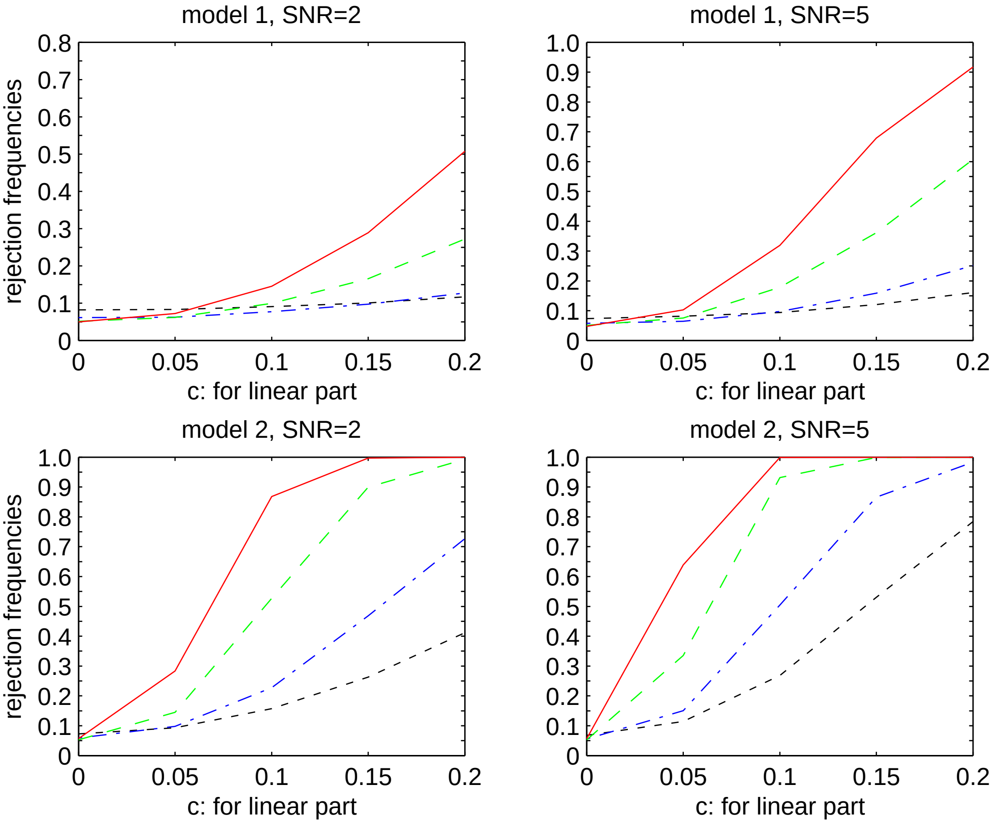

Figure 2: The simulation results for testing hypothesis H_0 with significance level 0.05 based on 5000 replications for each model setting. In each panel, from the bottom to the top the dotted, dash-dotted, dashed and solid lines correspond to sample sizes 100, 200, 500 and 1000 respectively.

reasonable to see that as the number of principal functions increases, the power of testing increases.

6. A REAL EXAMPLE

The Boston housing data of Harrison and Rubinfeld (1978) has attracted lots of attention in statistics. Various models have been applied to the data, including the linear regression model (Belsley et al, 1980), the additive model (Fan and Jiang, 2005) and the varying coefficient model (Fan and Huang, 2006). The response of interest is the median value of owner-occupied homes (MEDV, in $1000) with 13 predictors: lower status of the population (LSTAT), per capita crime rate (CRIM) by town, average number of rooms per dwelling (RM), full-value property-tax rate per \$10,000 (TAX), nitrogen dioxides concentration (NOX, parts per 10 million), pupil-teacher ratio by town (PTRATIO), proportion of owner-occupied units built prior to 1940 (AGE), proportion of residential land zoned for lots over 25,000 square feet (ZN), proportion of non-retail business acres per town (INDUS), Charles River dummy variable (1 if tract bounds river; 0 otherwise; CHAS), weighted distances to five Boston employment centres (DIS), index of accessibility to radial highways (RAD), $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town (B).

Fan and Huang (2005) fitted the data with a semi-varying coefficient model using $U = \sqrt{LSTAT}$ as the index variable. However, as the number of covariates $p = 13$ is too big for a VCM to be estimated well, Fan and Huang (2005) only included 6 variables in their model. With the superior estimation efficiency of PVCM over CVM, however, we are able to include all the variables into a PVCM which will be further identified as (**??**). We standardize all the variables before fitting the model.

As we mentioned in the first section, after linear transformation, remarkably similar shapes are shared among different coefficient functions. The eigenvalues of the estimated Σ_β suggest that the number of principal functions is $d_0 = 1$. The BIC defined in (**??**) for $d_0 = 0$(linear model), $d_0 = 1$, ..., and $d_0 = 10$ are respectively $-1.1593, -1.7199, -1.6950, -1.5482, -1.4933, -1.2018, -0.8020, -0.5044, -0.2011$ and -0.1034. Therefore, the number of principal functions is also selected as 1 by the BIC. The corresponding parameters in the model are estimated and listed in Table 2, where the standard errors of estimators are calculated based on Theorem 5 and are put in the parentheses. If a parameter is estimated as 0 and removed from the model, its standard error is not available. It is interesting to see that some of the covariates are eliminated from the model such as AGE, INDUS and CHAS because they do not appear in either the linear part or the nonlinear part. In a different model that only included the variables in the top panel of Table 2, AGE was also removed by Fan and Huang (2006) based on a statistical testing approach. Some other covariates have no cross effect with LSTAT on the response, such as TAX, PTRATIO, ZN, DIS and B, and are removed from the nonlinear part.

Table 2: Estimated parameters (and their standard errors in the parenthesis) in the model for the Boston housing data

coefficient	LSTAT	CRIM	RM	TAX	NOX	PTRATIO	AGE
θ_0	-0.5478 (0.0586)	0 (—)	0.1968 (0.0701)	-0.2335 (0.0482)	-0.1325 (0.0526)	-0.1756 (0.0235)	0 (—)
B_0	-0.2683 (0.1924)	0.3026 (0.1999)	0.6068 (0.1762)	0 (—)	0.2743 (0.1292)	0 (—)	0 (—)

	ZN	INDUS	CHAS	DIS	RAD	B	
θ_0	0.1003 (0.0390)	0 (—)	0 (—)	-0.2373 (0.0413)	0.3749 (0.0602)	0.1381 (0.0526)	
B_0	0 (—)	0 (—)	0 (—)	0 (—)	0.6245 (0.2303)	0 (—)	

The principal function $\boldsymbol{\gamma}(u)$ is estimated and shown in Figure **??**, where the centralized pointwise 95% confidence band based on Theorem **??** is also plotted.

Table 3: Average prediction errors based on 1000 partitions

size of		Linear	VCM	PVCM	PVCM
training set	prediction set	model			+ penality
200	306	0.3028	0.9312	0.2514	0.2434
300	206	0.2918	0.8210	0.2349	0.2262
400	106	0.2866	0.8661	0.2274	0.2215

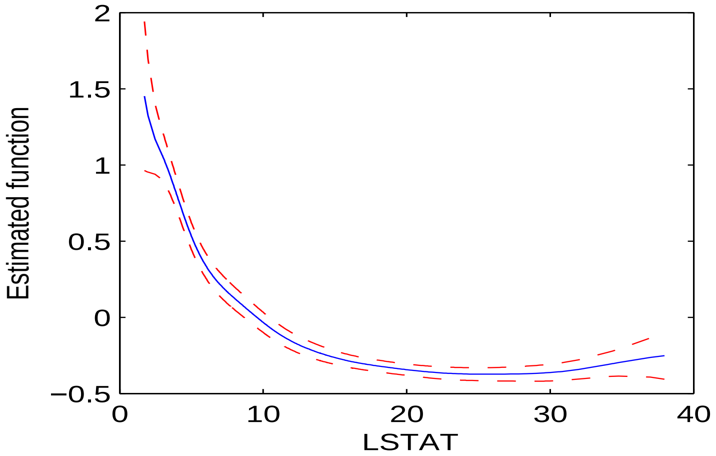

Figure 3: The estimated principal function (in the middle) for Boston housing dataset, and its 95% centralized pointwise confidence band denoted by the dash lines.

To further verify the appropriateness of different models for the data, we consider the prediction error of PVCM and compare it with linear regression model and the conventional VCM. We randomly partition all the 506 observations into a training set and a prediction set. We estimate the PVCM based on the training set, and use the estimated model to make prediction for observations in the prediction set. With different sizes of training set and prediction set, the average prediction errors based on 1000 random partitions are listed in Table 3. It is ease to see from Table 3 that VCM has very poor prediction capability, and is much worse than the simple linear regression model. However, PVCM with one principal function as identified by the proposed method (**??**) has much better prediction ability than VCM and even substantially better than the linear regression model. The prediction ability can be further improved when the L_1 penalty is imposed in the estimation, though the primary purpose of imposing the L_1 penalty is for variable selection.

7. CONCLUSION

Motivated by the compelling need to improve estimation efficiency of a VCM, especially when p is large, and by practical examples in which different coefficient functions are linearly dependent, this paper proposed a new varying coefficient model, PVCM, that incorporates the intrinsic patterns in the coefficients. The model possesses superior estimation efficiency over VCM.

The proposed model is a semiparametric model and thus can be estimated based on kernel smoothing and splines smoothing as well. The gain in estimation efficiency is due to further model identification that only a small number of principal functions need to be estimated nonparametrically, regardless of the smoothing method. Though only the estimation method based on kernel smoothing is discussed in this paper, the splines smoothing and the penalized splines enjoy many good properties (Wood, 2006; Ruppert et al, 2009). Thus estimation based on the splines smoothing needs further investigation.

The advantage of PVCM over VCM increases as p increases where the coefficient functions are more likely to be linearly dependent. Incorporating with the L_1 penalty, the estimation can automatically select variables in the linear part and the nonlinear part. The estimation efficiency only depends on the number of principal functions and the variables in the linear part and nonlinear part. Theoretical analysis and data study further confirm the advantages of PVCM. In conclusion, PVCM together with the estimation methods provide a powerful approach towards the analysis of complicated data.

APPENDIX: TECHNICAL DETAILS

To establish the asymptotic theory for the proposed estimation methods, we need the following technical assumptions.

(C.1) (*The Index Variable*). The *index* variable U has a bounded compact support $\mathcal{D}$ and a probability density function $f(u)$, which is Lipschitz continuous and bounded away from 0 on $\mathcal{D}$.

(C.2) (*Smoothness Assumptions*). Every component of $W(u) = E(XX^\top|U = u)$ and $L(u) = E(XY^\top|U = u)$ is Lipschitz continuous. In addition to that, we assume $\beta_0(u)$ has continuous second order derivatives in $u \in \mathcal{D}$. The matrix $W(u)$ is positive definite for all $u \in \mathcal{D}$.

(C.3) (*Moment Conditions*). There exist $s > 2$ and $\delta < 2 - s^{-1}$, such that $E\|X\|^s < \infty$ with $n^{2\delta-1}h \to \infty$, where $\|\cdot\|$ stands for a typical L_2 norm.

(C.4) (*The Kernel and Bandwidth*). We assume that the kernel function $K(\cdot)$ is a symmetric density function with a compact support. Moreover, we assume $h \propto n^{-c}$ with $c > 0$ such that $\sqrt{n}h^2 \to 0$ and $nh/\log n \to \infty$.

We remark that the above regularity conditions are rather standard. Similar assumptions have been used in, for example, **?** and ?. Let $\mu_k = \int t^k K(t)$. Then by (C.4) we have $\mu_0 = 1$ and $\mu_1 = 0$. For ease of exposition, we further standardize $K(.)$ such that $\mu_2 = 1$ in the following proofs. In addition, we denote $U_i - u$ by U_{iu} and $U_i - U_j$ by U_{ij} in the following proofs.

Lemma 1. *Under the regularity conditions (C.1)-(C.4), for the estimator defined in (??) we have the following expansion*

$$\begin{aligned}
\hat{\gamma}(u|B,\theta) &= \boldsymbol{\gamma}_0(u) + \frac{1}{2}\mu_2\boldsymbol{\gamma}_0''(u)h^2 + \{B^\top W(u)B\}^{-1}\{nf(u)\}^{-1}B^\top \sum_{i=1}^{n} K_h(U_{iu})X_i\varepsilon_i \\
&\quad +\{B^\top W(u)B\}^{-1}B^\top W(u)(B_0 - B)\boldsymbol{\gamma}_0(u) + \{B^\top W(u)B\}^{-1}B^\top W(u)(\theta_0 - \theta) \\
&\quad +O_p(h^3 + h\delta_n + \delta_n^2)
\end{aligned}$$

uniformly for any $u \in \mathcal{D}$ and $(\theta, B) \in \Theta_n$.

Proof. Write $Y_i - X_i^\top \theta = \varepsilon_i + X_i^\top B \boldsymbol{\gamma}_0(U_i) + X_i^\top (B_0 - B)\boldsymbol{\gamma}_0(U_i) + X_i^\top(\theta_0 - \theta)$. Thus

$$\begin{aligned} \sum_{i=1}^n K_h(U_{iu}) X_i \{Y_i - X_i^\top \theta\} &= \sum_{i=1}^n K_h(U_{iu}) X_i \varepsilon_i + \sum_{i=1}^n K_h(U_{iu}) X_i X_i^\top B \boldsymbol{\gamma}_0(U_i) \qquad \text{(A.1)} \\ &\quad + \sum_{i=1}^n K_h(U_{iu}) X_i X_i^\top (B_0 - B)\boldsymbol{\gamma}_0(U_i) + S_n(u)(\theta_0 - \theta). \end{aligned}$$

Let $s_n(u) = \sum_{i=1}^n K_h(U_{iu})$. By Mack and Silverman (1982), we have uniformly for $u \in \mathcal{D}$, $s_n^{-1}(u) = (nf(u))^{-1}(1 + O_p(h^2 + \delta_n))$, and

$$\frac{1}{n}\sum_{i=1}^n K_h(U_{iu}) X_i X_i^\top = f(u)W(u)(1 + O_p(h^2 + \delta_n)), \quad \frac{1}{n}\sum_{i=1}^n K_h(U_{iu}) X_i \varepsilon_i = O_p(\delta_n).$$

Thus,

$$s_n^{-1}(u) \sum_{i=1}^n K_h(U_{iu}) X_i X_i^\top = W(u) + O_p(h^2 + \delta_n),$$

$$s_n^{-1}(u) \sum_{i=1}^n K_h(U_{iu}) X_i X_i^\top \boldsymbol{\gamma}_0(U_i) = W(u)\boldsymbol{\gamma}_0(u) + O_p(h^2 + \delta_n),$$

$$s_n^{-1}(u) \sum_{i=1}^n K_h(U_{iu}) X_i \varepsilon_i = \{nf(u)\}^{-1} \sum_{i=1}^n K_h(U_{iu}) X_i \varepsilon_i + O_p(h^2\delta_n + \delta_n^2),$$

and

$$s_n^{-1}(u) \sum_{i=1}^n K_h(U_{iu}) X_i X_i^\top (B_0 - B)\boldsymbol{\gamma}_0(U_i) = W(u)(B_0 - B)\boldsymbol{\gamma}_0(u) + \|B_0 - B\| O_p(h^2 + \delta_n)$$

uniformly for $u \in \mathcal{D}$. Combining the above results yields that uniformly in $u \in \mathcal{D}$,

$$\begin{aligned}
& s_n^{-1}(u) \sum_{i=1}^n K_h(U_{iu}) X_i X_i^\top B \boldsymbol{\gamma}_0(U_i) \\
= \ & s_n^{-1}(u) \sum_{i=1}^n K_h(U_{iu}) X_i X_i^\top B \boldsymbol{\gamma}_0(u) + s_n^{-1}(u) \sum_{i=1}^n K_h(U_{iu}) X_i X_i^\top B \{\boldsymbol{\gamma}_0(U_i) - \boldsymbol{\gamma}_0(u)\} \\
= \ & s_n^{-1}(u) \sum_{i=1}^n K_h(U_{iu}) X_i X_i^\top B \boldsymbol{\gamma}_0(u) \\
& + s_n^{-1}(u) \sum_{i=1}^n K_h(U_{iu}) X_i X_i^\top B \{\boldsymbol{\gamma}_0'(u)(U_{iu}) + \frac{1}{2}\mu_2 \boldsymbol{\gamma}_0''(u)(U_{iu})^2 + O_p(U_{iu}^3)\} \\
= \ & s_n^{-1}(u) \sum_{i=1}^n K_h(U_{iu}) X_i X_i^\top B \boldsymbol{\gamma}_0(u) + \{f^{-1}(u) f'(u) W'(u) B \boldsymbol{\gamma}_0'(u) + \frac{1}{2}\mu_2 W(u) B \boldsymbol{\gamma}_0''(u)\} h^2 \\
& + O_p(h^3).
\end{aligned}$$

For $(\theta, B) \in \Theta_n$, we have

$$\begin{aligned}
\hat{\boldsymbol{\gamma}}(u|B,\theta) &= (B^\top S_n(u) B)^{-1} B^\top \sum_{i=1}^n K_h(U_{iu}) X_i \{Y_i - X_i^\top \theta\} \\
&= (B^\top s_n^{-1}(u) S_n(u) B)^{-1} B^\top \Big(s_n^{-1}(u) \sum_{i=1}^n K_h(U_{iu}) X_i \{Y_i - X_i^\top \theta\} \Big) \\
&= \boldsymbol{\gamma}_0(u) + \frac{1}{2}\mu_2 \boldsymbol{\gamma}_0''(u) h^2 + \{B^\top W(u) B\}^{-1} \{n f(u)\}^{-1} B^\top \sum_{i=1}^n K_h(U_{iu}) X_i \varepsilon_i \\
&\quad + \{B^\top W(u) B\}^{-1} B^\top W(u) (B_0 - B) \boldsymbol{\gamma}_0(u) + \{B^\top W(u) B\}^{-1} B^\top W(u) (\theta_0 - \theta) \\
&\quad + O_p(h^3 + h\delta_n + \delta_n^2).
\end{aligned}$$

As a special case,

$$\begin{aligned}
\hat{\boldsymbol{\gamma}}(u|B_0,\theta_0) &= \boldsymbol{\gamma}_0(u) + \frac{1}{2}\mu_2 \boldsymbol{\gamma}_0''(u) h^2 + \{B_0^\top W(u) B_0\}^{-1} \{n f(u)\}^{-1} B_0^\top \sum_{i=1}^n K_h(U_{iu}) X_i \varepsilon_i \\
&\quad + O_p(h^3 + h\delta_n + \delta_n^2).
\end{aligned}$$

We have completed the proof. □

Proof of Theorems 1. By Theorem 1 of **?** or Lemma **??**, we have

$$\sup_{u\in\mathcal{D}}|\hat{\boldsymbol{\beta}}(u)-\boldsymbol{\beta}_0(u)|=O_p(h^2+\delta_n), \tag{A.2}$$

where $\delta_n=\{nh/\log(n)\}^{-1/2}$. Theorems 1 follows immediately from (**??**). □

Proof of Theorem 2. Let $\alpha=(\theta^\top,vec(B)^\top)^\top$, $\alpha_0=(\alpha_{0,1},...,\alpha_{0,p(d_0+1)})^\top=(\theta_0^\top,vec(B_0)^\top)^\top$, $\hat{\alpha}=(\hat{\theta}^\top,vec(\hat{B})^\top)^\top$ and $Q(\alpha)=Q(\theta,B)$. By Taylor expansion about α_0, we have

$$0=\frac{\partial Q(\hat{\alpha})}{\partial\alpha}=\frac{\partial Q(\alpha_0)}{\partial\alpha}+\frac{\partial^2Q(\alpha^*)}{\partial\alpha\partial\alpha^\top}(\hat{\alpha}-\alpha_0),$$

where α^* lies on the line segment between α_0 and $\hat{\alpha}$. Let $\Delta_i(\alpha)=Y_i-X_i^\top\theta-X_i^\top B\tilde{\boldsymbol{\gamma}}(U_i)$, $\eta_i(\alpha)=Y_i-X_i^\top\theta-X_i^\top B\boldsymbol{\gamma}_0(U_i)$, then $\Delta_i(\alpha)=\eta_i(\alpha)-X_i^\top B(\tilde{\boldsymbol{\gamma}}(U_i)-\boldsymbol{\gamma}_0(U_i))$, $\eta_i(\alpha_0)=\varepsilon_i$, and

$$Q(\alpha)=\sum_{i=1}^n\Delta_i^2(\alpha).$$

Let $Q_0(\alpha)=\sum_{i=1}^n\eta_i^2(\alpha)$. From Lemma **??**, when $||\alpha-\alpha_0||=O_p(h^2+\delta_n)$ we have

$$\sup_{u\in\mathcal{D}}\|\tilde{\boldsymbol{\gamma}}(u)-\boldsymbol{\gamma}_0(u)\|=O_p(h^2+\delta_n)=o_p(1).$$

Thus $\Delta_i(\alpha)=\eta_i(\alpha)-X_i^\top B(\tilde{\boldsymbol{\gamma}}(U_i)-\boldsymbol{\gamma}_0(U_i))=\eta_i(\alpha)+o_p(1)$, $\partial\Delta_i(\alpha)/\partial\alpha=\partial\eta_i(\alpha)/\partial\alpha+$

$o_p(1)$. It follows that

$$\begin{aligned}
\frac{1}{2n}\frac{\partial^2 Q(\alpha)}{\partial\alpha\partial\alpha^\top} &= \frac{1}{n}\sum_{i=1}^n \frac{\partial\Delta_i(\alpha)}{\partial\alpha}\frac{\partial\Delta_i(\alpha)}{\partial\alpha^\top} + \frac{1}{n}\sum_{i=1}^n \Delta_i(\alpha)\frac{\partial^2\Delta_i(\alpha)}{\partial\alpha\partial\alpha^\top} \\
&= \frac{1}{n}\sum_{i=1}^n \frac{\partial\eta_i(\alpha)}{\partial\alpha}\frac{\partial\eta_i(\alpha)}{\partial\alpha^\top} + \frac{1}{n}\sum_{i=1}^n \eta_i(\alpha)\frac{\partial^2\eta_i(\alpha)}{\partial\alpha\partial\alpha^\top} + o_p(1) \\
&= \frac{1}{n}\sum_{i=1}^n \frac{\partial\eta_i(\alpha_0)}{\partial\alpha}\frac{\partial\eta_i(\alpha_0)}{\partial\alpha^\top} + \frac{1}{n}\sum_{i=1}^n \eta_i(\alpha_0)\frac{\partial^2\eta_i(\alpha_0)}{\partial\alpha\partial\alpha^\top} + o_p(1) \\
&\to E\Big\{\frac{\partial\eta_1(\alpha_0)}{\partial\alpha}\frac{\partial\eta_1(\alpha_0)}{\partial\alpha^\top}\Big\} \\
&= E\Big\{\begin{pmatrix} X \\ \boldsymbol{\gamma}_0(U)\otimes X\end{pmatrix}\begin{pmatrix} X \\ \boldsymbol{\gamma}_0(U)\otimes X\end{pmatrix}^\top\Big\} = \Sigma_0, \quad \text{in probability.}
\end{aligned}$$

In the last step, $\partial^2\eta_i(\alpha_0)/(\partial\alpha\partial\alpha^\top) = 0$ is used. Write

$$\frac{1}{2\sqrt{n}}\frac{\partial Q(\alpha_0)}{\partial\alpha} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \{\eta_i(\alpha_0) - X_i^\top B_0(\tilde{\boldsymbol{\gamma}}(U_i) - \boldsymbol{\gamma}_0(U_i))\}\{\frac{\partial\eta_i(\alpha_0)}{\partial\alpha} + \frac{\partial\Delta_i(\alpha_0)}{\partial\alpha} - \frac{\partial\eta_i(\alpha_0)}{\partial\alpha}\}. \tag{A.3}$$

Let $Z_{n0} = Z_{n1} + Z_{n2}$ with $Z_{n1} = n^{-1/2}\sum_{i=1}^n \eta_i(\alpha_0)\partial\eta_i(\alpha_0)/\partial\alpha$ and

$$Z_{n2} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \eta_i(\alpha_0)\Big(\frac{\partial\Delta_i(\alpha_0)}{\partial\alpha} - \frac{\partial\eta_i(\alpha_0)}{\partial\alpha}\Big) - \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i^\top B_0(\tilde{\boldsymbol{\gamma}}(U_i) - \boldsymbol{\gamma}_0(U_i))\frac{\partial\eta_i(\alpha_0)}{\partial\alpha}.$$

By Lemma **??**, we have

$$\begin{aligned}
\Big|\frac{1}{2\sqrt{n}}\frac{\partial Q(\alpha_0)}{\partial\alpha} - Z_{n0}\Big| &= \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i^\top B_0(\tilde{\boldsymbol{\gamma}}(U_i) - \boldsymbol{\gamma}_0(U_i))\Big(\frac{\partial\Delta_i(\alpha_0)}{\partial\alpha} - \frac{\partial\eta_i(\alpha_0)}{\partial\alpha}\Big)\Big| \\
&\le \sqrt{n}\max_{1\le i\le n}|X_i^\top B_0(\tilde{\boldsymbol{\gamma}}(U_i) - \boldsymbol{\gamma}_0(U_i))| \max_{1\le i\le n}\Big\|\frac{\partial\Delta_i(\alpha_0)}{\partial\alpha} - \frac{\partial\eta_i(\alpha_0)}{\partial\alpha}\Big\| \\
&= \sqrt{n}O_p(h^2+\delta_n)O_p(h^2+\delta_n) = o_p(1).
\end{aligned}$$

It is easy to check that

$$Z_{n1} \;=\; -n^{-1/2}\sum_{i=1}^{n}\begin{pmatrix} X_i \\ \boldsymbol{\gamma}_0(U_i)\otimes X_i \end{pmatrix}\varepsilon_i.$$

Let $\ell(U) = (1, \boldsymbol{\gamma}_0(U)^\top)^\top \otimes W(U)$ and $\bar{\ell} = E\ell(U)$. Write $Z_{n2} = E_{n1} - E_{n2}$, where

$$E_{n1} = n^{-1/2}\sum_{i=1}^{n}\eta_i(\alpha_0)\Big(\frac{\partial\Delta_i(\alpha_0)}{\partial\alpha} - \frac{\partial\eta_i(\alpha_0)}{\partial\alpha}\Big),$$

$$E_{n2} = n^{-1/2}\sum_{i=1}^{n}X_i^\top B_0(\tilde{\boldsymbol{\gamma}}(U_i) - \boldsymbol{\gamma}_0(U_i))\frac{\partial\eta_i(\alpha_0)}{\partial\alpha}.$$

Under assumptions (C.1)-(C.4), we can show that

$$E_{n1} = o_p(1) \tag{A.4}$$

and

$$\begin{aligned} E_{n2} \;&=\; \frac{1}{2}E\{(\ell(U)-\bar{\ell})B_0\boldsymbol{\gamma}_0''(U)\}n^{1/2}h^2 + \frac{1}{\sqrt{n}}\sum_{j=1}^{n}(\ell(U_j)-\bar{\ell})V(U_j)X_j\varepsilon_j \\ &\quad + \bar{\ell}B_0\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\gamma}_0(U_i) + o_p(1). \end{aligned} \tag{A.5}$$

Thus, we have

$$\begin{aligned} Z_{n2} \;&=\; \frac{1}{2}E\{(\ell(U)-\bar{\ell})B_0\boldsymbol{\gamma}_0''(U)\}n^{1/2}h^2 + \frac{1}{\sqrt{n}}\sum_{j=1}^{n}(\ell(U_j)-\bar{\ell})V(U_j)X_j\varepsilon_j \\ &\quad + \begin{pmatrix} EW(U) \\ E\{\boldsymbol{\gamma}_0(U)\otimes W(U)\} \end{pmatrix} B_0\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\gamma}_0(U_i) + o_p(1), \end{aligned}$$

where $W(u)$ and $V(u)$ are defined in Theorem **??**. By the central limit theorem (CLT), we have

$$Z_{n1} + \frac{1}{\sqrt{n}} \sum_{j=1}^{n} (\ell(U_j) - \bar{\ell}) V(U_j) X_j \varepsilon_j \to N(0, \Sigma_1),$$

where Σ_1 is given in Theorem **??**. On the other hand, since $E\boldsymbol{\gamma}_0(U) = 0$, we have

$$n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\gamma}_0(U_i) \to N\Big(0, E\{\boldsymbol{\gamma}_0(U)\boldsymbol{\gamma}_0^\top(U)\}\Big).$$

Theorem **??** follows from last three equations and (**??**).

Now, we turn to prove (**??**) and (**??**). We only give the details for the latter. Decompose E_{n2} into two terms.

$$E_{n2} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i^\top B_0 (\hat{\boldsymbol{\gamma}}(U_i) - \boldsymbol{\gamma}_0(U_i)) \frac{\partial \eta_i(\alpha_0)}{\partial \alpha} - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i^\top B_0 \bar{\boldsymbol{\gamma}} \frac{\partial \eta_i(\alpha_0)}{\partial \alpha} \triangleq E_{n2}^1 - E_{n2}^2, \tag{A.6}$$

where $\hat{\boldsymbol{\gamma}}(U_i) = \hat{\boldsymbol{\gamma}}(U_i|\theta_0, B_0)$ and $\bar{\boldsymbol{\gamma}} = n^{-1}\sum_{i=1}^{n} \hat{\boldsymbol{\gamma}}(U_i)$. From Lemma **??**, we have

$$E_{n2}^1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{pmatrix} X_i \\ \boldsymbol{\gamma}_0(U_i) \otimes X_i \end{pmatrix} X_i^\top B_0 \{\frac{1}{2}\boldsymbol{\gamma}_0''(U_i)h^2 + R_n(U_i) + O_p(h^3 + h\delta_n + \delta_n^2)\},$$

where $R_n(U_i) = \{nf(U_i)B_0^\top W(U_i) B_0\}^{-1} B_0^\top \sum_{j=1}^{n} K_h(U_{ij}) X_j \varepsilon_j$. It follows from the laws of large numbers

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{pmatrix} X_i \\ \boldsymbol{\gamma}_0(U_i) \otimes X_i \end{pmatrix} X_i^\top B_0 \boldsymbol{\gamma}_0''(U_i) h^2 = E\{\ell(U) B_0 \boldsymbol{\gamma}_0''(U)\} n^{1/2} h^2 + o_p(1). \tag{A.7}$$

As $f(u)$ is bounded away from 0, we then have

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix}X_i\\ \boldsymbol{\gamma}_0(U_i)\otimes X_i\end{pmatrix}X_i^{\top}B_0R_n(U_i) &= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\Big\{\sum_{i=1}^{n}\begin{pmatrix}X_i\\ \boldsymbol{\gamma}_0(U_i)\otimes X_i\end{pmatrix}X_i^{\top}V(U_i)\\ &\qquad\times\frac{1}{nf(U_i)}K_h(U_{ij})\Big\}X_j\varepsilon_j\\ &= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\ell(U_j)V(U_j)X_j\varepsilon_j+\Delta_n, \qquad (A.8)\end{aligned}$$

where

$$\Delta_n=\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\Big\{\sum_{i=1}^{n}\begin{pmatrix}X_i\\ \boldsymbol{\gamma}_0(U_i)\otimes X_i\end{pmatrix}X_i^{\top}V(U_i)\frac{1}{nf(U_i)}K_h(U_{ij})-\ell(U_j)V(U_j)\Big\}X_j\varepsilon_j.$$

By simple calculation, we have $Var(\Delta_n)=O\{(h^2+\delta_n)^2\}$ and thus

$$\Delta_n = O_p(h^2+\delta_n). \qquad (A.9)$$

For E_{n2}^2, by Lemma **??** we have $\bar{\boldsymbol{\gamma}}=O_p(h^2+\delta_n)$,

$$\bar{\boldsymbol{\gamma}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\gamma}_0(U_i)+\frac{1}{2}E\boldsymbol{\gamma}_0''(U)h^2+\frac{1}{n}\sum_{i=1}^{n}(B_0^{\top}W(U_i)B_0)^{-1}B_0^{\top}X_i\varepsilon_i+o_p(n^{-1/2}))$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\eta_i(\alpha_0)}{\partial\alpha}X_i^{\top}=\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}X_iX_i^{\top}\\ \boldsymbol{\gamma}_0(U_i)\otimes X_iX_i^{\top}\end{pmatrix}=\bar{\ell}+O_p(n^{-1/2}).$$

It follows from Lemma **??** that

$$E_{n2}^2=\bar{\ell}\Big\{B_0\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\gamma}_0(U_i)+\frac{1}{2}B_0E\boldsymbol{\gamma}_0''(U)\sqrt{n}h^2+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}V(U_i)X_i\varepsilon_i\Big\}+o_p(1). \quad (A.10)$$

Equation (**??**) follows from (**??**)-(**??**) and the following fact

$$\bar{\ell}B_0\frac{1}{2}E\boldsymbol{\gamma}_0''(U)h^2 - E\{\ell(U)B_0\frac{1}{2}\boldsymbol{\gamma}_0''(U)\}n^{-1/2}h^2 = -\frac{1}{2}E\{(\ell(U)-\bar{\ell})B_0\boldsymbol{\gamma}_0''(U)\}h^2.$$

This completes the proof. □

Proof of Theorem 4. For any fixed d, denote the estimators of θ_0, B_0 and $\boldsymbol{\gamma}_0(u)$ by $\hat{\theta}_d$, $\hat{B}_d$ and $\hat{\boldsymbol{\gamma}}_d(u)$ respectively.

Case 1. ($d < d_0$, underfitted model) By the proof of Theorem 1, $\hat{\theta}_d - \theta_d = O_p(h^2+\delta_n)$ and that there exist nonrandom matrix B_d and function $\boldsymbol{\gamma}_d(u)$ such that

$$\hat{B}_d - B_d = O_p(h^2+\delta_n), \quad \hat{\boldsymbol{\gamma}}_d(u) - \boldsymbol{\gamma}_d(u) = O_p(h^2+\delta_n)$$

uniformly for $u \in \mathcal{D}$. By the definition of d_0, if $d < d_0$ then $E||B_0\boldsymbol{\gamma}_0(U) - B_d\boldsymbol{\gamma}_d(U)|| > 0$. It is easy to see by the above facts and the CLT that

$$\begin{aligned}
\hat{\sigma}_d^2 &= n^{-1}\sum_{i=1}^n\{Y_i - (\hat{\theta}_d + \hat{B}_d\hat{\boldsymbol{\gamma}}_d(U_i))^\top X_i\}^2 \\
&= n^{-1}\sum_{i=1}^n\{Y_i - (\theta_d + B_d\boldsymbol{\gamma}_d(U_i))^\top X_i\}^2 + O_p(h^2+\delta_n) \\
&= n^{-1}\sum_{i=1}^n\{\varepsilon_i + (B_0\boldsymbol{\gamma}_0(U_i) - B_d\boldsymbol{\gamma}_d(U_i))^\top X_i\}^2 + O_p(h^2+\delta_n) \\
&= n^{-1}\sum_{i=1}^n\varepsilon_i^2 + 2n^{-1}\sum_{i=1}^n\varepsilon_i(B_0\boldsymbol{\gamma}_0(U_i) - B_d\boldsymbol{\gamma}_d(U_i))^\top X_i \\
&\quad + n^{-1}\sum_{i=1}^n\{(B_0\boldsymbol{\gamma}_0(U_i) - B_d\boldsymbol{\gamma}_d(U_i))^\top X_i\}^2 + O_p(h^2+\delta_n) \\
&= \sigma^2 + E\{(B_0\boldsymbol{\gamma}_0(U) - B_d\boldsymbol{\gamma}_d(U))^\top X\}^2 + O_p(h^2+\delta_n+n^{-1/2}). \qquad \text{(A.11)}
\end{aligned}$$

Therefore, as a special case we have $\hat{\sigma}^2_{d_0} = \sigma^2 + O_p(h^2 + \delta_n + n^{-1/2})$. Note that

$$\begin{aligned} E\{(B_0\boldsymbol{\gamma}_0(U) - B_d\boldsymbol{\gamma}_d(U))^\top X\}^2 &= E\{(B_0\boldsymbol{\gamma}_0(U) - B_d\boldsymbol{\gamma}_d(U))^\top W(U)(B_0\boldsymbol{\gamma}_0(U) - B_d\boldsymbol{\gamma}_d(U))\} \\ &\geq \lambda_1(W(u))E||B_0\boldsymbol{\gamma}_0(U) - B_d\boldsymbol{\gamma}_d(U)|| \overset{def}{=} c_0 > 0. \end{aligned}$$

Therefore, for $d < d_0$ we have $\hat{\sigma}^2_d \geq \sigma^2_{d_0} + c_0 + O_p(h^2 + \delta_n + n^{-1/2})$. Therefore

$$P\Big\{\mathrm{BIC}(d) > \mathrm{BIC}(d_0)\Big\} \to 1 \text{ for any } d < d_0. \tag{A.12}$$

Case 2. ($d \geq d_0$, overfitted model) By the definition of d_0, if $d \geq d_0$ then $B_d\boldsymbol{\gamma}_d(u) = B_0\boldsymbol{\gamma}_0(u)$. For ease of exposition, we only consider the case that ε_i is independent of (X_i, U_i). If $d > d_0$, following the same argument of Theorem **??** and Lemma **??** we have $\hat{\theta}_d - \theta_0 = O_p(n^{-1/2})$ and

$$\begin{aligned} B_d\boldsymbol{\gamma}_d(u) - B_0\boldsymbol{\gamma}_0(u) &= \frac{1}{2}\mu_2 B_d\boldsymbol{\gamma}''_d(u)h^2 + B_d\{nf(u)B_d^\top W(u)B_d\}^{-1}B_d^\top \sum_{i=1}^n K_h(U_{iu})X_i\varepsilon_i \\ &\quad + O_p(n^{-1/2} + h^3 + h\delta_n + \delta_n^2). \end{aligned}$$

where $O_p(n^{-1/2} + h^3 + h\delta_n + \delta_n^2)$ are independent of ε_i. Thus, by CLT we have

$$\begin{aligned} \hat{\sigma}^2_d &= n^{-1}\sum_{j=1}^n \Big(\varepsilon_j - \big(\frac{1}{2}\mu_2 B_d\boldsymbol{\gamma}''_d(U_j)h^2 + B_d\{nf(U_j)B_d^\top W(U_j)B_d\}^{-1}B_d^\top \sum_{i=1}^n K_h(U_{ij})X_i\varepsilon_i\big)^\top X_j\Big)^2 \\ &\quad + O_p\{n^{-1/2}(n^{-1/2} + h^3 + h\delta_n + \delta_n^2)\} \\ &= n^{-1}\sum_{i=1}^n \varepsilon_i^2 - 2n^{-1}\sum_{j=1}^n (B_d\{nf(U_j)B_d^\top W(U_j)B_d\}^{-1}B_d^\top \sum_{i=1}^n K_h(U_{ij})X_i\varepsilon_i)^\top X_j\varepsilon_j \\ &\quad + \frac{1}{4}\mu_2^2 E\{(B_d\boldsymbol{\gamma}''_d(U))^\top W(U)(B_d\boldsymbol{\gamma}''_d(U))\}h^4 + O_p((nh)^{-1} + n^{-1/2}h^2 + n^{-1}). \end{aligned}$$

It is easy to see that

$$Var(n^{-1}\sum_{j=1}^{n}(B_d\{nf(U_j)B_d^{\top}W(U_j)B_d\}^{-1}B_d^{\top}\sum_{i=1}^{n}K_h(U_{ij})X_i\varepsilon_i)^{\top}X_j\varepsilon_j)=O(\frac{1}{n^2h}).$$

Note that $B_d\boldsymbol{\gamma}_d''(U)$ are the same for different $d\geq d_0$. Thus, we have

$$\hat{\sigma}_d^2=\hat{\sigma}_{d_0}^2+O_p\{(nh)^{-1}+n^{-1/2}h^2\}.$$

It follows that $\log\hat{\sigma}_d^2-\log\hat{\sigma}_{d_0}^2=O_p\{(nh)^{-1}+n^{-1/2}h^2\}$. As a consequence, we have

$$\mathrm{BIC}(d)-\mathrm{BIC}(d_0)=(d-d_0)\frac{\log(nh)}{nh}+O_p\{(nh)^{-1}+n^{-1/2}h^2\},$$

where the first term on the right hand side dominates under the condition (C.4). Hence,

$$P\Big\{\mathrm{BIC}(d)>\mathrm{BIC}(d_0)\Big\}\to 1\text{ for any }d>d_0. \tag{A.13}$$

Equations (**??**) and (**??**) together imply that $P\{\mathrm{BIC}(d)>\mathrm{BIC}(d_0)\}\to 1$. This further implies that $P(\hat{d}=d_0)=1$. □

Proof of Theorem 5. The proof is an adaption to our case of Zou (2006). We first show (**??**).

Let $\tilde{\alpha}^{(n)}=\alpha_0+u/\sqrt{n}$ where $u=(u_1,\ldots,u_S)^{\top}\in\mathcal{R}^S$, the objective function (**??**) can be written as a function of u as

$$\tilde{Q}_n(u)=Q_n(\alpha_0+\frac{u}{\sqrt{n}})+\lambda_n\sum_{s=1}^{S}\hat{w}_s|\alpha_{0,s}+\frac{u}{\sqrt{n}}|.$$

Let $\tilde{u}=arg\min_{u\in\mathcal{R}^S}\tilde{Q}_n(u)$ and obviously $\tilde{Q}_n(u)$ is minimized at $\tilde{u}_n=\sqrt{n}(\tilde{\alpha}^{(n)}-\alpha_0)$.

Next, write

$$\begin{aligned} D_n(u) &= \tilde{Q}_n(u) - \tilde{Q}_n(0) \\ &= \big(Q_n(\alpha_0 + \frac{u}{\sqrt{n}}) - Q_n(\alpha_0)\big) + \lambda_n \sum_{s=1}^{S} \hat{w}_s\big(|\alpha_{0,s} + \frac{u_s}{\sqrt{n}}| - |\alpha_{0,s}|\big) \\ &\equiv I_{1,n}(u) + I_{2,n}(u), \end{aligned}$$

where $I_{1,n}(u) = Q_n(\alpha_0 + \frac{u}{\sqrt{n}}) - Q_n(\alpha_0)$ is due to the loss function and $I_{2,n}(u)$ is due to the penalty term. From the proof of theorem 2, we know that

$$\begin{aligned} \frac{1}{2n}\frac{\partial^2 Q(\alpha_0)}{\partial\alpha\partial\alpha^\top} &\to \Sigma_0 \text{ in probability,} \\ \frac{1}{2}n^{-\frac{1}{2}}\frac{\partial Q(\alpha_0)}{\partial\alpha} &\overset{D}{\to} Z = N(0, \Sigma_1 + \Sigma_2). \end{aligned}$$

Thus the loss function term

$$I_{1,n}(u) = \frac{1}{\sqrt{n}}u^\top\frac{\partial Q(\alpha_0)}{\partial\alpha} + \frac{1}{2n}u^\top\frac{\partial^2 Q(\alpha_0)}{\partial\alpha\partial\alpha^\top}u(1 + o_p(1)) \overset{D}{\to} 2u^\top Z + u^\top\Sigma_0 u.$$

Now, we consider the limiting behavior of the penalty term $I_{2,n}(u)$. If $s \in \mathcal{A}$, that is $\alpha_{0,s} \neq 0$, then $\hat{w}_s \to |\alpha_{0,s}|^{-\tau}$ in probability and $\sqrt{n}(|\alpha_{0,s} + u_s/\sqrt{n}| - |\alpha_{0,s}|) \to u_s\text{sgn}(\alpha_{0,s})$. Since $\lambda_n/\sqrt{n} \to 0$, we have

$$\frac{\lambda_n}{\sqrt{n}}\hat{w}_s\sqrt{n}(|\alpha_{0,s} + u_s/\sqrt{n}| - |\alpha_{0,s}|) \to 0.$$

If $s \notin \mathcal{A}$ then $\sqrt{n}(|\alpha_{0,s}+u_s/\sqrt{n}|-|\alpha_{0,s}|) = |u_s|$. Since $\sqrt{n}\hat{\alpha}_n = O_p(1)$ and $\lambda_n n^{\frac{\tau-1}{2}} \to \infty$, we have $\frac{\lambda_n}{\sqrt{n}}\hat{w}_s = \lambda_n n^{\frac{\tau-1}{2}}|\sqrt{n}\hat{\alpha}_s^{(n)}|^{-\tau} \to \infty$ in probability. It follows that

$$D_n(u) \Rightarrow D(u) = \begin{cases} 2(u_\mathcal{A})^\top Z_\mathcal{A} + (u_\mathcal{A})^\top(\Sigma_0)_\mathcal{A}(u_\mathcal{A}), & \text{if } u_s = 0, \forall s \notin \mathcal{A} \\ \infty, & \text{otherwise ,} \end{cases}$$

where $u_{\mathcal{A}}$ and $Z_{\mathcal{A}}$ are the j-th ($j \in \mathcal{A}^c$) elements deleted from u and Z respectively. Note that $D_n(u)$ is convex, and the unique minimum of $D(u)$ is

$$u_{min} = \begin{pmatrix} -\Big((\Sigma_0)_{\mathcal{A}}\Big)^{-1} Z_{\mathcal{A}} \\ 0 \end{pmatrix},$$

where 0 denotes a vector of zeros. Following the epi-convergence result of Geyer (1994), we have

$$\tilde{\alpha}_{\mathcal{A}}^{(n)} \overset{D}{\to} \Big((\Sigma_0)_{\mathcal{A}}\Big)^{-1} Z_{\mathcal{A}} = N\left(0, \Big((\Sigma_0)_{\mathcal{A}}\Big)^{-1}(\Sigma_1 + \Sigma_2)_{\mathcal{A}}\Big((\Sigma_0)_{\mathcal{A}}\Big)^{-1}\right) \qquad \text{(A.14)}$$

and $\tilde{\alpha}_{\mathcal{A}^c}^{(n)} \to 0$. Now we prove the consistency part. It suffices to show that $\forall s \in \mathcal{A}^c$, $P(s \in \mathcal{A}_n) \to 0$. By the KKT optimality conditions,

$$\frac{1}{\sqrt{n}} \frac{\partial Q_n(\tilde{\alpha}^{(n)})}{\partial \alpha_s} + \frac{\lambda_n}{\sqrt{n}} \hat{w}_s \mathrm{sgn}(\tilde{\alpha}_s^{(n)}) = 0.$$

If $s \in \mathcal{A}^c$, then

$$\frac{\lambda_n}{\sqrt{n}} \hat{w}_s = \lambda_n n^{\frac{\tau-1}{2}} |\sqrt{n}\hat{\alpha}_s^{(n)}|^{-\tau} \to \infty$$

in probability, whereas

$$\begin{aligned} \frac{1}{\sqrt{n}} \frac{\partial Q_n(\tilde{\alpha}^{(n)})}{\partial \alpha_s} &= \frac{1}{\sqrt{n}} \frac{\partial Q_n(\tilde{\alpha}^{(n)})}{\partial \alpha_s} + \frac{1}{n} \frac{\partial^2 Q_n(\tilde{\alpha}^{(n)})}{\partial \alpha_s^2} \sqrt{n}(\tilde{\alpha}_s^{(n)} - \alpha_{0,s})(1 + o_p(1)) \\ &\overset{D}{\to} \text{some normal distribution} \end{aligned}$$

by (**??**) and Slutsky's theorem. Thus, for $s \in \mathcal{A}^c$,

$$P(s \in \mathcal{A}^{(n)}) \leq P\Big(|\frac{1}{\sqrt{n}} \frac{\partial Q_n(\tilde{\alpha}^{(n)})}{\partial \alpha_s}| = \frac{\lambda_n}{\sqrt{n}} \hat{w}_s\Big) \to 0.$$

We have completed the proof. □

REFERENCES

Bai, J. (2003) Inferential theory for factor models of large dimensions, *Econometrica*, **71**, 135-171.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980) *Regression diagnostics: identifying influential data and sources of collinearity*, New York, John Wiley.

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.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997) Generalized partially linear single-index models, *Journal of the American Statistical Association*, **92**, 477-489.

Chen, R. and Tsay, R.S. (1993) Functional coefficient autoregressive models, *Journal of American Statistical Association*, **88**, 298-308.

Cheng, M.-Y., and Hall, P. (2003) Reducing variance in nonparametric surface estimation. *Journal of Multivariate Analysis*, **86**, 375-397.

Eubank, R. L., Huang, C. Maldonado, Y. M., Wang, N., and Wang, S. and and Buchanan, R. J. (2004) Smoothing spline estimation in varying-coefficient models, *J. R. Statist. Soc. B* **66**, 653-667.

Fan, J. and Gijbels, I. (1996) *Local Polynomial Modeling and Its Applications*, Chapman and Hall, 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 Jiang, J. (2005) Nonparametric inferences for additive models. *Journal of the American Statistical Association*, **100**, 890-907.

Fan, J. and Zhang, J. T. (2000) Two-step estimation of functional linear model with application to longitudinal data, *Journal of the Royal Statistical Society, Series B*, **62**, 303-322.

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

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

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

Forni, M. and Lippi, M. (2001) The Generalized Dynamic Factor Model: Representation Theory, *Econometric Theory*, **17(06)**, 1113-1141.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2000) The generalized dynamic factor model: identification and estimation, *Review of Economics and Statistics*,**82**, 540-554.

Geyer, C. (1994) On the asymptotics of constrained M-estimation, *Annals of Statistics*, **22**, 1993-2010.

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

Hastie, T. and Stuetzle, W. (1989) Principal curves, *Journal of the American Statistical Association*, **84** 502-516.

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 splines estimation and inference for varying coefficient models with longitudinal data, *Statistica Sinica* **14**, 763-788.

Kai, B., Li, R. and Zou, H. (2011) New Efficient Estimation and Variable Selection Methods for Semiparametric Varying-Coefficient Partially Linear Models. *The Annals of Statistics*, **39**, 305-332.

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

Ruppert, D., Wand, M.P. and Carroll, R.J. (2009) Semiparametric Regression During 2003-2007. *Electronic Journal of Statistics*, **3**, 1193-1256.

Wang, H. (2008) Rank reducible varying coefficient model, *Journal of Statistical Planning and Inference*, **138**, 236-245.

Wang, H. and Xia, Y. (2008) Sliced regression for dimension reduction, *Journal of the American Statistical Associate*, **103**, 811-821.

Wood S.N. (2006) *Generalized Additive Models: An Introduction with R.* Chapman and Hall/CRC Press.

Wu, H. and Liang, H. (2004) Backfitting random varying-coefficient models with time-dependent smoothing covariates. *Scandinavian Journal of Statistics*, **31**, 3-19.

Zhang, W., Lee, S. Y., and Song, X. (2002) Local polynomial fitting in semivarying coefficient model, *Journal of Multivariate Analysis*, **82**, 166-188.

Zhang, H. and Lu, W. (2007) Adaptive-LASSO for Cox's proportional hazards model. *Biometrika*, **94**, 1-13.

Zou, H. (2006) The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association,* **101**, 1418-1429.