mathematics

MDPI

Article

Estimation for Varying Coefficient Models with Hierarchical Structure

Feng Li [1], Yajie Li [1] and Sanying Feng [1,2,*]

[1] School of Mathematics and Statistics, Zhengzhou University, Zhengzhou 450001, China; lifengstat@zzu.edu.cn (F.L.); liyajiehandy@gs.zzu.edu.cn (Y.L.)
[2] Henan Key Laboratory of Financial Engineering, Zhengzhou University, Zhengzhou 450001, China
* Correspondence: fsy5801@zzu.edu.cn

Abstract: The varying coefficient (VC) model is a generalization of ordinary linear model, which can not only retain strong interpretability but also has the flexibility of the nonparametric model. In this paper, we investigate a VC model with hierarchical structure. A unified variable selection method for VC model is proposed, which can simultaneously select the nonzero effects and estimate the unknown coefficient functions. Meanwhile, the selected model enforces the hierarchical structure, that is, interaction terms can be selected into the model only if the corresponding main effects are in the model. The kernel method is employed to estimate the varying coefficient functions, and a combined overlapped group Lasso regularization is introduced to implement variable selection to keep the hierarchical structure. It is proved that the proposed penalty estimators have oracle properties, that is, the coefficients are estimated as well as if the true model were known in advance. Simulation studies and a real data analysis are carried out to examine the performance of the proposed method in finite sample case.

Keywords: varying coefficient model; variable selection; interaction term; hierarchical structure; group lasso

Citation: Li, F.; Li, Y.; Feng, S. Estimation for Varying Coefficient Models with Hierarchical Structure. *Mathematics* **2021**, *9*, 132. https://doi.org/10.3390/math9020132

Received: 30 November 2020
Accepted: 7 January 2021
Published: 9 January 2021

1. Introduction

The varying coefficient model [1] is defined as

$$Y = \sum_{j=1}^{d} X_j \beta_j(U) + \varepsilon, \tag{1}$$

where Y is the response variable and $(X_1, X_2, \cdots, X_d, U)$ are its associated covariates, ε is the error term, and $\beta_j(U)$ are unknown smooth coefficient functions of observable continuous covariate U. Compared with the linear model, the predictors X_j $(j = 1, 2, \cdots, d)$ affect the response variable Y linearly, but their coefficients are allowed to change smoothly with other covariate U. Thus, each value of U is associated with a different linear model, and it also allows one to examine the extent to which the association between response Y and covariates X_j varies over covariate U (see [1,2]). In environmental data analysis [2], one objective of the study is to investigate the association between the level of the pollutants and the number of daily total hospital admissions for circulatory and respiratory problems, as well as to examine how the association varies over time, where $U = t = time$. Besides, without the model specification of parametric structure and multivariate nonparametric structure, the VC model can significantly reduce the modelling bias and avoid the "curse of dimensionality".

As the merit of good interpretability and flexibility of the varying coefficient model, a considerable amount of literature has been published on estimation and hypothesis test of the VC model since it was initiated (see, e.g., [2–6]). Recently, variable selection and model detection for varying coefficient model have gained much attention, for instance Wang and Xia [7]

combined the ideas of the local polynomial smoothing and Lasso (least absolute shrinkage and selection operator [8]) to estimate the coefficients and select variables simultaneously; Zhao and Xue [9] employed basis function approximations and SCAD (smoothly clipped absolute deviation [10]) penalty for the semiparametric varying coefficient partially linear model; Tang et al. [11] developed a unified variable selection approach for varying coefficient models; Li et al. [12] studied the model selection and structure specification for the generalized semi-varying coefficient models; and He et al. [13] introduced a dimensionality reduction and variable selection method for multivariate varying-coefficient models with a large number of covariates, and so on.

In many complex situations, main effects combined with interaction effects may be sufficient to characterize the relationship between the response and predictors. In social, political, and economic problems and genome-wide association studies, it is useful to identify nontrivial interactions between covariates in modeling selection results, product sales, social networks, stock market changes, and disease risk. Recent years have seen a surge of interests in interaction identification in the high dimensional setting by many researchers. For instance, Hall and Xue [14] proposed a recursive approach to identify important interactions among covariates; Niu et al. [15] proposed a forward selection based screening method for identifying interactions for ultra-high dimensional data; Kong et al. [16] suggested a two-stage interaction identification method, called the interaction pursuit via distance correlation in high dimensional multi-response regression; and Radchenko and James [17] investigated variable selection for nonlinear additive regression models with interaction structures by group regularization method. Specifically, for the high dimensional linear model with interaction terms,

$$Y = \sum_{j=1}^{d} X_j \beta_j + \sum_{1 \le j < k \le d} X_j X_k \phi_{jk} + \varepsilon, \tag{2}$$

some important works include but are not limited to the following: Choi et al. [18] reparameterized the coefficients for the interaction terms; Bien et al. [19] added a set of convex constraints to the Lasso to produce sparse interaction terms; Zhao et al. [20] introduced the composite absolute penalties family by defining groups with particular overlapping patterns to express the relationships between the predictors; and Lim and Hastie [21] developed a method for learning pairwise interactions via hierarchical group-Lasso regularization. A key feature of the models is its hierarchical structure, as the interaction effects are derived from the main effects, which means that the interaction terms exist only if the main terms are significant in the model. This is also referred to the marginality principal in generalized linear models [22] or the strong heredity in the analysis of designed experiments [23].

Compared to parametric model (2), the VC model with interaction terms is the direct extension to the nonparametric case, where the coefficients are unknown smoothing functions of some covariates. However, the estimation methods for model (2) cannot be directly extended to the VC model with interaction terms. Variable selection for the VC model including interaction effects is also important, since ignoring important predictors can lead to biased results, while including irrelevant predictors may lead to efficiency loss. Moreover, variable selection for the VC model with interaction terms is even more complex, since nonzero functional coefficients rather than nonzero parameters need to be identified. In this paper, we aim to develop a unified variable selection method for VC model with hierarchical structure, which not merely can identify the significant variables with nonzero functional coefficients. Moreover, the selected model keeps the hierarchical structure, that is, interaction terms can be selected into the model only if the corresponding main effects are in the model. Firstly, kernel smoothing method is employed to obtain the initial estimates of the varying coefficient functions. Secondly, the local penalized least squares estimates with overlapped group Lasso penalty are proposed to simultaneously achieve variable selection and coefficients estimation, and the estimators enforce the hierarchical

structure. Thirdly, it is proved that the proposed estimators have the oracle properties, that is, the functional coefficients are estimated as well as if the true model were known in advance.

The rest of the paper is organized as follows. In Section 2, we propose the local penalized least squares estimator with group Lasso penalty, which enforces the hierarchical structure. The asymptotic properties of the estimators are investigated in Section 3. In Section 4, we conduct some simulations and the Boston housing data analysis to assess the finite sample performance of the new estimators. The conclusion and future works are discussed in Section 5. Proofs of the theorems are postponed to Appendix A.

2. Modeling and Estimation

The varying coefficient model with hierarchical structure is defined as follows,

$$Y_i = \sum_{j=1}^{d} X_{ij}\beta_j(U_i) + \sum_{1\le j<k\le d} X_{ij}X_{ik}\phi_{jk}(U_i) + \varepsilon_i, \quad i = 1, 2, \cdots, n, \tag{3}$$

where Y_i is the response variable, $X_{ij}, j = 1, 2, \cdots, d$ are the predictive variables, U_i is the index covariate, ε_i is the random error and satisfies $\mathrm{E}(\varepsilon_i|X_i, U_i) = 0$, $\mathrm{Var}(\varepsilon_i|X_i, U_i) = \sigma^2(U_i)$, $\beta_j(U_i)$ is the unknown smooth coefficient function, and $\phi_{jk}(U_i)$ is the coefficient function associated with the interaction term $X_{ij}X_{ik}$. The hierarchical structure means that the interaction term $X_{ij}X_{ik}$ exists if and only if both X_{ij} and X_{ik} exist in the model, namely $\int \phi_{jk}^2(u)du \neq 0$ holds if and only if both $\int \beta_j^2(u)du \neq 0$ and $\int \beta_k^2(u)du \neq 0$ hold for any $i = 1, 2, \cdots, n$, $1 \le j < k \le d$. The model (3) is a useful extension of the linear model with hierarchical structure (2) (see, e.g., [18,19,24–27]), which can maintain the good interpretability of parameter models and also has the flexibility of the nonparametric models.

To simplify the representation of model (3), we list the following definitions,

$$X_i = (X_{i1}, X_{i2}, \cdots, X_{id})^\tau, \quad Z_i = (X_{i1}X_{i2}, \cdots, X_{i1}X_{id}, X_{i2}X_{i3}, \cdots, X_{i2}X_{id}, \cdots, X_{id-1}X_{id})^\tau,$$
$$\boldsymbol{\phi}(U_i) = (\phi_{12}(U_i), \phi_{13}(U_i), \cdots, \phi_{1d}(U_i), \phi_{23}(U_i), \phi_{24}(U_i), \cdots, \phi_{2d}(U_i), \cdots, \phi_{(d-1)d}(U_i))^\tau,$$
$$\boldsymbol{\beta}(U_i) = (\beta_1(U_i), \beta_2(U_i), \cdots, \beta_d(U_i))^\tau, \quad W_i = (X_i^\tau, Z_i^\tau)^\tau, \quad \boldsymbol{\alpha}(U_i) = (\boldsymbol{\beta}(U_i)^\tau, \boldsymbol{\phi}(U_i)^\tau)^\tau,$$
$$\boldsymbol{\beta}_j = (\beta_j(U_1), \beta_j(U_2), \cdots, \beta_j(U_n))^\tau, \quad \boldsymbol{\phi}_{jk} = (\phi_{jk}(U_1), \phi_{jk}(U_2), \cdots, \phi_{jk}(U_n))^\tau,$$

where the superscript "τ" means transposition operation. Then, model (3) can be reformulated as

$$Y_i = W_i^\tau \boldsymbol{\alpha}(U_i) + \varepsilon_i, i = 1, 2, \cdots, n.$$

Let $\boldsymbol{\Lambda} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_d, \boldsymbol{\phi}_{12}, \boldsymbol{\phi}_{13}, \cdots, \boldsymbol{\phi}_{(d-1)d})$, the support set of the important main effects be $\mathcal{M} = \{j : \|\boldsymbol{\beta}_j\| > 0\}$, and the support set of the important interaction effects be $\mathcal{I} = \{(j,k) : \|\boldsymbol{\phi}_{jk}\| > 0, j, k \in \mathcal{M}\}$, where$\|\cdot\|$ is the L_2 norm.

To obtain the initial estimate of coefficient matrix $\boldsymbol{\Lambda}$, we minimize the following objective function,

$$Q(\boldsymbol{\Lambda}) = \sum_{t=1}^{n}\sum_{i=1}^{n}[Y_i - W_i^\tau \boldsymbol{\alpha}(U_t)]^2 K_h(U_t - U_i), \tag{4}$$

where $K_h(\cdot) = \frac{1}{h}K(\cdot/h)$, and $K(\cdot)$ is a kernel function which satisfies the Condition C5 and $h \to 0$ is the bandwidth. Denote the solution to the objective function (4) by $\widetilde{\boldsymbol{\Lambda}}$, the tth row of $\widetilde{\boldsymbol{\Lambda}}$ for $t = 1, 2, \cdots, n$ has the closed form

$$\widetilde{\boldsymbol{\alpha}}(U_t) = \left[\frac{1}{n}\sum_{i=1}^{n} W_i W_i^\tau K_h(U_t - U_i)\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n} W_i Y_i K_h(U_t - U_i)\right].$$

Corresponding to the assumption, only the columns of $\mathbf{\Lambda}$ indexed by the support sets $\mathcal{M}$ and $\mathcal{I}$ are nonzero, so the main task of variable selection is to identify the sparse columns of $\mathbf{\Lambda}$ efficiently. Meanwhile, to maintain the hierarchical structure of the model, we apply the idea of group Lasso proposed by Yuan and Lin [28] and give the following local penalized least squares estimation,

$$\begin{aligned}\widehat{\mathbf{\Lambda}}_\lambda = \arg\min_{\mathbf{\Lambda}} Q_\lambda(\mathbf{\Lambda}) &= \arg\min_{\boldsymbol{\beta}_j, \boldsymbol{\phi}_{jk}, 1\le j<k\le d} \left\{ \sum_{t=1}^{n}\sum_{i=1}^{n} [Y_i - W_i^\tau \boldsymbol{\alpha}(U_t)]^2 K_h(U_t - U_i) \right. \\ &\quad \left. + \sum_{j=1}^{d} \left(\lambda_j^1 \sqrt{\|\boldsymbol{\beta}_j\|^2 + \sum_{k:k\neq j} \|\boldsymbol{\phi}_{jk}\|^2} + \sum_{k:j<k} \lambda_{jk}^2 \|\boldsymbol{\phi}_{jk}\| \right) \right\}, \end{aligned} \tag{5}$$

where we assume $\boldsymbol{\phi}_{jk} = \boldsymbol{\phi}_{kj}$, which is commonly used in the model with hierarchical structure, and the assumption means the interaction effects are independent of the order of the both covariates, which also significantly reduces the computation burden by lessening the number of the functional coefficients from the order of $O(d^2)$ to $O(d^2/2)$ (see [18,19,25,27]). λ_j^1 and λ_{jk}^2 are tuning parameters. For simplicity of calculations, we use the local quadratic approximation (see [10,29,30]) in each step of the iteration. Take $\widetilde{\mathbf{\Lambda}}$ as the initial estimator, which is $\widehat{\mathbf{\Lambda}}_\lambda^{(0)} = \widetilde{\mathbf{\Lambda}}$, and, for the $(m+1)$th step, the objective function can be approximately represented as follows,

$$\begin{aligned} Q_\lambda^{(m+1)}(\mathbf{\Lambda}) &\approx \sum_{t=1}^{n}\sum_{i=1}^{n} \{Y_i - W_i^\tau \boldsymbol{\alpha}(U_t)\}^2 K_h(U_t - U_i) \\ &\quad + \sum_{j=1}^{d} \left(\lambda_j^1 \frac{\|\boldsymbol{\beta}_j\|^2 + \sum_{k:k\neq j} \|\boldsymbol{\phi}_{jk}\|^2}{\sqrt{\|\widehat{\boldsymbol{\beta}}_j^{(m)}\|^2 + \sum_{k:k\neq j} \|\widehat{\boldsymbol{\phi}}_{jk}^{(m)}\|^2}} + \sum_{k:j<k} \lambda_{jk}^2 \frac{\|\boldsymbol{\phi}_{jk}\|^2}{\|\widehat{\boldsymbol{\phi}}_{jk}^{(m)}\|} \right) \\ &= \sum_{t=1}^{n} \left\{ \sum_{i=1}^{n} \{Y_i - W_i^\tau \boldsymbol{\alpha}(U_t)\}^2 K_h(U_t - U_i) \right. \\ &\quad \left. + \sum_{j=1}^{d} \left(\lambda_j^1 \frac{\beta_j^2(U_t) + \sum_{k:k\neq j} \phi_{jk}^2(U_t)}{\sqrt{\|\widehat{\boldsymbol{\beta}}_j^{(m)}\|^2 + \sum_{k:k\neq j} \|\widehat{\boldsymbol{\phi}}_{jk}^{(m)}\|^2}} + \sum_{k:j<k} \lambda_{jk}^2 \frac{\phi_{jk}^2(U_t)}{\|\widehat{\boldsymbol{\phi}}_{jk}^{(m)}\|} \right) \right\}. \end{aligned} \tag{6}$$

By minimizing $Q_\lambda^{(m+1)}(\mathbf{\Lambda})$, we have

$$\widehat{\boldsymbol{\alpha}}_\lambda(U_t)^{(m+1)} = \left[\sum_{i=1}^{n} W_i W_i^\tau K_h(U_t - U_i) + G^{(m)} + H^{(m)} \right]^{-1} \left[\sum_{i=1}^{n} W_i Y_i K_h(U_t - U_i) \right], \tag{7}$$

where

$$G^{(m)} = \begin{pmatrix} M_1^{(m)} & 0 & 0 & \cdots & 0 \\ 0 & I_{d-1}\lambda_1^1\gamma_1^{(m)} + M_2^{(m)} & 0 & \cdots & 0 \\ 0 & 0 & I_{d-2}\lambda_2^1\gamma_2^{(m)} + M_3^{(m)} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_{d-1}^1\gamma_{d-1}^{(m)} + M_d^{(m)} \end{pmatrix},$$

$$H^{(m)} = \begin{pmatrix} 0_{d\times d} & 0 & 0 & \cdots & 0 \\ 0 & \lambda_{12}^2\zeta_{12}^{(m)} & 0 & \cdots & 0 \\ 0 & 0 & \lambda_{13}^2\zeta_{13}^{(m)} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_{(d-1)d}^2\zeta_{(d-1)d}^{(m)} \end{pmatrix},$$

and $M_j^{(m)} = \mathrm{diag}(\lambda_j^1\gamma_j^{(m)}, \lambda_{j+1}^1\gamma_{j+1}^{(m)}, \cdots, \lambda_d^1\gamma_d^{(m)})$, $\gamma_j^{(m)} = \frac{1}{\sqrt{\|\widehat{\boldsymbol{\beta}}_j^{(m)}\|^2 + \sum_{k:k\neq j}\|\widehat{\boldsymbol{\phi}}_{jk}^{(m)}\|^2}}$, $\zeta_{jk}^{(m)} = \frac{1}{\|\widehat{\boldsymbol{\phi}}_{jk}^{(m)}\|}$, $1 \le j < k \le d$.

Regard $\widehat{\boldsymbol{\Lambda}}_\lambda$ as the Kernel Lasso (KLasso) estimator, and the specific implementation procedures can be arranged as follows:

(1) Take the non-penalized estimator $\widetilde{\boldsymbol{\Lambda}}$ as the initial estimator $\widehat{\boldsymbol{\Lambda}}_\lambda^{(0)} = \widetilde{\boldsymbol{\Lambda}}$.
(2) According to the method discussed above, iterate (7) until convergence; specifically, the iteration stops when $\max\left(|\|\widehat{\boldsymbol{\beta}}_j^{(m+1)}\| - \|\widehat{\boldsymbol{\beta}}_j^{(m)}\||, |\|\widehat{\boldsymbol{\phi}}_{jk}^{(m+1)}\| - \|\widehat{\boldsymbol{\phi}}_{jk}^{(m)}\||, 1 \le j < k \le d\right)$ is less than 10^{-3}. For model sparsity, $\|\widehat{\boldsymbol{\beta}}_j\|$ and $\|\widehat{\boldsymbol{\phi}}_{jk}\|$ should be set 0 when they are less than a small real value c_r (in our simulation, we choose $c_r = 0.5$).
(3) Get the KLasso estimator $\widehat{\boldsymbol{\Lambda}}_\lambda$.

Selection of bandwidth h and tuning parameters λ_j^1 and λ_{jk}^2 ($1 \le j < k \le d$) is another important issue for VC model. There are two types of parameters to be considered; one common method to solve this problem is grid searching, such as cross validation (CV) or generalized cross validation (GCV) [31]. However, it would be too expensive in computation, thus here we choose h in the kernel estimation of coefficient functions by CV criterion. For the tuning parameters, we apply the idea of Zou and Li [32], where large coefficients should be given small penalties, while small coefficients should be given large penalties, and the tuning parameters can be chosen as

$$\lambda_j^1 = \frac{\lambda_0}{n^{-\frac{1}{2}}(\sqrt{\|\widetilde{\boldsymbol{\beta}}_j\|^2 + \sum_{k:k\neq j}\|\widetilde{\boldsymbol{\phi}}_{jk}\|^2})}, \quad \lambda_{jk}^2 = \frac{\lambda_0}{n^{-\frac{1}{2}}\|\widetilde{\boldsymbol{\phi}}_{jk}\|}.$$

Then, only one parameter, λ_0, needs to be considered, which is selected according to BIC criterion

$$\mathrm{BIC}_\lambda = \log(\mathrm{RSS}_\lambda) + d_\lambda\frac{\log(nh)}{nh},$$

where d_λ is the number of nonzero coefficient functions determined by $\widehat{\boldsymbol{\Lambda}}_\lambda$ and

$$\mathrm{RSS}_\lambda = \frac{1}{n^2}\sum_{t=1}^n\sum_{i=1}^n\{Y_i - W_i^\tau\widehat{\boldsymbol{\alpha}}_\lambda(U_t)\}^2K_h(U_t - U_i).$$

Then, λ_0 can be obtained by minimizing BIC_λ.

3. Theoretical Properties

In this section, we establish asymptotic properties of the proposed estimator, including the model selection consistency and the oracle properties. First, we make some notations and list some regular conditions. Define $a_n = \max\{\lambda_j^1, \lambda_{kl}^2 : j \in \mathcal{M}, (k,l) \in \mathcal{I}\}$, $b_n = \min\{\lambda_j^1, \lambda_{kl}^2 : j \in \mathcal{M}^c, (k,l) \in \mathcal{I}^c\}$. Let $\boldsymbol{X}_\mathcal{M}$ denote a matrix which is generated from $\boldsymbol{X}$ with columns indexed by $\mathcal{M}$, and $\boldsymbol{Z}_\mathcal{I}$ denotes a matrix which is generated from $\boldsymbol{Z}$ with columns indexed by $\mathcal{I}$, $\boldsymbol{W}_\mathcal{S} = (\boldsymbol{X}_\mathcal{M}, \boldsymbol{Z}_\mathcal{I})$, $\boldsymbol{W}_{\mathcal{S}^c} = (\boldsymbol{X}_{\mathcal{M}^c}, \boldsymbol{Z}_{\mathcal{I}^c})$, $\widehat{\mathcal{M}} = \{j : \|\widehat{\boldsymbol{\beta}}_j\| > 0\}$, $\widehat{\mathcal{I}} = \{(j,k) : \|\widehat{\boldsymbol{\phi}}_{j,k}\| > 0\}$, $\widehat{\boldsymbol{\alpha}}_\mathcal{S}(U_t) = (\widehat{\boldsymbol{\beta}}_\mathcal{M}^\tau(U_t), \widehat{\boldsymbol{\phi}}_\mathcal{I}^\tau(U_t))^\tau$, $\widehat{\boldsymbol{\alpha}}_{\mathcal{S}^c}(U_t) = (\widehat{\boldsymbol{\beta}}_{\mathcal{M}^c}^\tau(U_t), \widehat{\boldsymbol{\phi}}_{\mathcal{I}^c}^\tau(U_t))^\tau$. The convergence of $o_p(\cdot)$ and $O_p(\cdot)$ are defined, respectively, as follows, for random variables ξ and η: $\xi = o_p(\eta)$ means that, for all $\epsilon > 0$, $P(|\xi/\eta| > \epsilon) \to 0$ as $n \to \infty$; and

$\xi = O_p(\eta)$ means that, for all $\epsilon > 0$, there exists $c > 0$ such that $P(|\xi/\eta| > c) < \epsilon$ as n is sufficiently large. The following traditional conditions (see [3]) are also needed.

C1. For $1 \leq i \leq n$, the covariate X_i is independent of the error ε_i.
C2. The covariate W_i has finite p-order moment, i.e., $E\|W_i\|^p < \infty$, where $p \geq 2$.
C3. The density function of U, $f(U)$, is continuous and has second-order derivative.
C4. $\Omega(U) = E(W_i W_i^\tau \mid U_i = U)$ has second-order derivative, while $E(\|W_i\|^4 \mid U_i = U)$ and $E(\|\varepsilon_i\|^2 \mid U_i = U)$ are both bounded.
C5. $K(\cdot)$ is a symmetric kernel function, which satisfies $\int K(s)ds = 1$, $\int s^2 K(s)ds = \mu_2 < \infty$, $\int K^2(s)ds = \nu < \infty$, $\int s^2 K^2(s)ds = \nu_2 < \infty$.
C6. $\beta_j(U)$ and $\phi_{jk}(U)$ are all bounded and have second-order continuous derivatives for $j \in \mathcal{M}$ and $(j,k) \in \mathcal{I}$.

Theorem 1. *Suppose C1-C6 hold; if* $h \propto n^{-\frac{1}{5}}, (nh)^{-\frac{1}{2}} a_n \to 0$, $(nh)^{-\frac{1}{2}} b_n \to \infty$, *then the following results hold:*

(1) $P(\widehat{\mathcal{M}} = \mathcal{M}) \to 1$ *as* $n \to \infty$
(2) $P(\widehat{\mathcal{I}} = \mathcal{I}) \to 1$ *as* $n \to \infty$

Theorem 1 shows that the KLasso estimator can select the true model consistently. Then, we discuss the oracle properties of the KLasso estimator in Theorem 2.

Theorem 2. *Suppose C1-C6 hold; if* $h \propto n^{-\frac{1}{5}}, (nh)^{-\frac{1}{2}} a_n \to 0$, $(nh)^{-\frac{1}{2}} b_n \to \infty$, *then we have*

$$\sup_{U_t \in [0,1]} \|\widehat{\boldsymbol{\alpha}}_{\lambda,\mathcal{S}}(U_t) - \widetilde{\boldsymbol{\alpha}}_{\mathcal{S}}(U_t)\| = o_p(n^{-\frac{2}{5}}),$$

for $1 \leq t \leq n$.

Note that the optimal convergence rate of oracle estimator is $O_p(n^{-\frac{2}{5}})$. We also observe that the difference of convergence rate between the KLasso estimator and the oracle estimator can be ignored over the univariate indicator set. Thus, we can conclude that the KLasso estimator shares the same asymptotic properties with the oracle estimator. Proofs of these two theorems are given in the Appendix.

4. Simulation Study and Real Data Analysis

4.1. Simulation Study

In this section, three examples are applied to assess the proposed procedures in terms of varying coefficient functions estimations and variable selection. The data with sample size $n = 100, 200, 500$ are independently generated from the following models:

$$\begin{aligned}
\text{Model 1:}\quad & Y_i = \sin(\pi U_i)X_{i1} + 0.5\exp(U_i)X_{i2} + U_i^3 X_{i3} + \cos(\pi U_i)X_{i2}X_{i3} + \varepsilon_i,\\
\text{Model 2:}\quad & Y_i = [(U_i - 0.5)^2 + 1]X_{i1} + 0.5\exp(U_i)X_{i2} + U_i^3 X_{i3} + \cos(\pi U_i)X_{i4}\\
& \quad + [U_i^3 + \log(2U_i + 1)]X_{i1}X_{i2} + \sin(2\pi U_i)X_{i2}X_{i3} + \varepsilon_i,\\
\text{Model 3:}\quad & Y_i = 2U_i^2 X_{i1} + 0.5\exp(U_i)X_{i2} + U^{1/3}X_{i3} + [\cos(2\pi U_i) + \cos(\pi U_i)]X_{i4}\\
& \quad + \sin(2\pi U_i)X_{i2}X_{i3} + [\sin(2\pi U_i) + \sin(\pi U_i)]X_{i2}X_{i4} + \varepsilon_i,
\end{aligned}$$

where the functional coefficients mainly include trigonometric function with different periods, exponential function, and power function with different locations and power. The random vector $(X_{i1}, \cdots, X_{i10})^\tau$ follows the multivariate normal distribution with zero means and $\mathrm{Cov}(X_{ij}, X_{ik}) = 0.5^{|k-j|}$, for $1 \leq j < k \leq 10$. The index variable $U_i \sim$Uniform[0,1] and the random error ε_i follow a standard normal distribution. In the estimation procedures, Gaussian kernel function $K(u) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{u^2}{2})$ is employed. The initial estimated coeffi-

cient matrix $\widetilde{\mathbf{\Lambda}}$ is obtained by minimizing (4), and the optimal bandwidth is selected via CV criterion, the selected bandwidth is also used for KLasso estimate procedure.

To assess the performance of the KLasso estimator in estimation accuracy, the empirical integrated squared error defined as follows is computed,

$$\mathrm{ISE}(\hat{\beta}_j) = \frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_j(U_i) - \beta_j(U_i))^2, \quad \mathrm{ISE}(\hat{\phi}_{jk}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{\phi}_{jk}(U_i) - \phi_{jk}(U_i))^2, \text{for } j \in \mathcal{M}, (j,k) \in \mathcal{I}.$$

As a benchmark of comparison, the mean empirical integrated squared error (MISE) and its standard error (in parenthesis) of the oracle estimators and the proposed estimators for the three models, are respectively, reported in Tables 1–3. All the empirical results were computed by R software [33] based on 1000 replications. In Tables 1–3, we can see that: (1) the MISEs decrease quickly as the sample size increases for both oracle estimator and KLasso estimator; (2) the proposed estimator performs comparably well with the oracle estimator in terms of coefficients estimations for moderate sample sizes; and (3) the estimators for Model 1 have smaller MISEs compared to those of the other two models, so the number and complexity of the nonzero coefficient functions to be estimated may affect the performance of the proposed procedure in finite samples.

Table 1. MISEs of the functional coefficients estimators for Model 1.

Estimator	$n = 100$	$n = 200$	$n = 500$
$\hat{\beta}_1(u)$	0.0684 (0.0856)	0.0285 (0.0189)	0.0174 (0.0087)
$\hat{\beta}_{ora,1}(u)$	0.0553 (0.0385)	0.0277 (0.0181)	0.0170 (0.0087)
$\hat{\beta}_2(u)$	0.0531 (0.0508)	0.0309 (0.0213)	0.0213 (0.0110)
$\hat{\beta}_{ora,2}(u)$	0.0476 (0.0406)	0.0305 (0.0211)	0.0211 (0.0107)
$\hat{\beta}_3(u)$	0.0483 (0.0403)	0.0277 (0.0199)	0.0183 (0.0096)
$\hat{\beta}_{ora,3}(u)$	0.0410 (0.0325)	0.0256 (0.0167)	0.0169 (0.0086)
$\hat{\phi}_{23}(u)$	0.0811 (0.0715)	0.0231 (0.0256)	0.0124 (0.0082)
$\hat{\phi}_{ora,23}(u)$	0.0435 (0.0343)	0.0198 (0.0134)	0.0111 (0.0056)

Table 2. MISEs of the functional coefficients estimators for Model 2.

Estimator	$n = 100$	$n = 200$	$n = 500$
$\hat{\beta}_1(u)$	0.0605 (0.0433)	0.0332 (0.0215)	0.0173 (0.0084)
$\hat{\beta}_{ora,1}(u)$	0.0576 (0.0407)	0.0329 (0.0212)	0.0172 (0.0083)
$\hat{\beta}_2(u)$	0.0832 (0.0608)	0.0428 (0.0278)	0.0221 (0.0107)
$\hat{\beta}_{ora,2}(u)$	0.0789 (0.0551)	0.0426 (0.0270)	0.0221 (0.0106)
$\hat{\beta}_3(u)$	0.0820 (0.0584)	0.0411 (0.0253)	0.0221 (0.0116)
$\hat{\beta}_{ora,3}(u)$	0.0753 (0.0498)	0.0419 (0.0254)	0.0225 (0.0117)
$\hat{\beta}_4(u)$	0.0866 (0.0882)	0.0358 (0.0323)	0.0182 (0.0095)
$\hat{\beta}_{ora,4}(u)$	0.0605 (0.0429)	0.0325 (0.0212)	0.0173 (0.0088)
$\hat{\phi}_{12}(u)$	0.0815 (0.0685)	0.0329 (0.0213)	0.0147 (0.0081)
$\hat{\phi}_{ora,12}(u)$	0.0658 (0.0486)	0.0317 (0.0202)	0.0143 (0.0077)
$\hat{\phi}_{23}(u)$	0.1695 (0.1023)	0.0486 (0.0328)	0.0185 (0.0093)
$\hat{\phi}_{ora,23}(u)$	0.1123 (0.0652)	0.0428 (0.0242)	0.0168 (0.0082)

Let CM denote the frequency of the nonzero coefficients being correctly estimated as nonzero, CZ denote the frequency of the zero coefficients being correctly estimated as zero, and CS denote the frequency of the correctly selected model, which means that only nonzero coefficients are estimated as nonzero. CM, CZ, and CS are summarized in Table 4. For the three models, as the sample size increases, CM, CZ, and CS all can be as large as

100%, which implies that the proposed method can identify the model well. In addition, Model 1 has the largest possibility of being correctly selected among the three models.

Besides, we also depict the quantiles curves of the estimated coefficient functions at fixed series $U_1, U_2, \cdots, U_{46}$, where $U_i = 0.05 + (i-1) \times 0.02, i = 1, 2, \cdots, 46$. Figures 1–3 are quantile curves for Models 1–3 with sample size 200, respectively. In these figures, we can see that the main effects and interaction effects can be correctly selected and consistently estimated. Meanwhile, the estimated curves usually underestimate at the peaks while overestimate at the valley of the curves. In summary, the proposed method for VC models with hierarchical structure works well.

Table 3. MISEs of the functional coefficients estimators for Model 3.

Estimator	$n = 100$	$n = 200$	$n = 500$
$\hat{\beta}_1(u)$	0.0784 (0.0751)	0.0357 (0.0232)	0.0188 (0.0096)
$\hat{\beta}_{ora,1}(u)$	0.0702 (0.0536)	0.0345 (0.0224)	0.0189 (0.0096)
$\hat{\beta}_2(u)$	0.0938 (0.0753)	0.0483 (0.0338)	0.0253 (0.0127)
$\hat{\beta}_{ora,2}(u)$	0.0918 (0.0717)	0.0484 (0.0334)	0.0253 (0.0125)
$\hat{\beta}_3(u)$	0.0819 (0.0654)	0.0435 (0.0305)	0.0242 (0.0119)
$\hat{\beta}_{ora,3}(u)$	0.0795 (0.0595)	0.0435 (0.0297)	0.0244 (0.0119)
$\hat{\beta}_4(u)$	0.2343 (0.1158)	0.0658 (0.0385)	0.0227 (0.0113)
$\hat{\beta}_{ora,4}(u)$	0.2127 (0.1052)	0.0628 (0.0367)	0.0216 (0.0104)
$\hat{\phi}_{23}(u)$	0.1462 (0.1213)	0.0734 (0.0660)	0.0247 (0.0126)
$\hat{\phi}_{ora,23}(u)$	0.1265 (0.0897)	0.0668 (0.0483)	0.0239 (0.0118)
$\hat{\phi}_{24}(u)$	0.2088 (0.1638)	0.0933 (0.0616)	0.0330 (0.0166)
$\hat{\phi}_{ora,24}(u)$	0.1842 (0.1159)	0.0909 (0.0534)	0.0305 (0.0152)

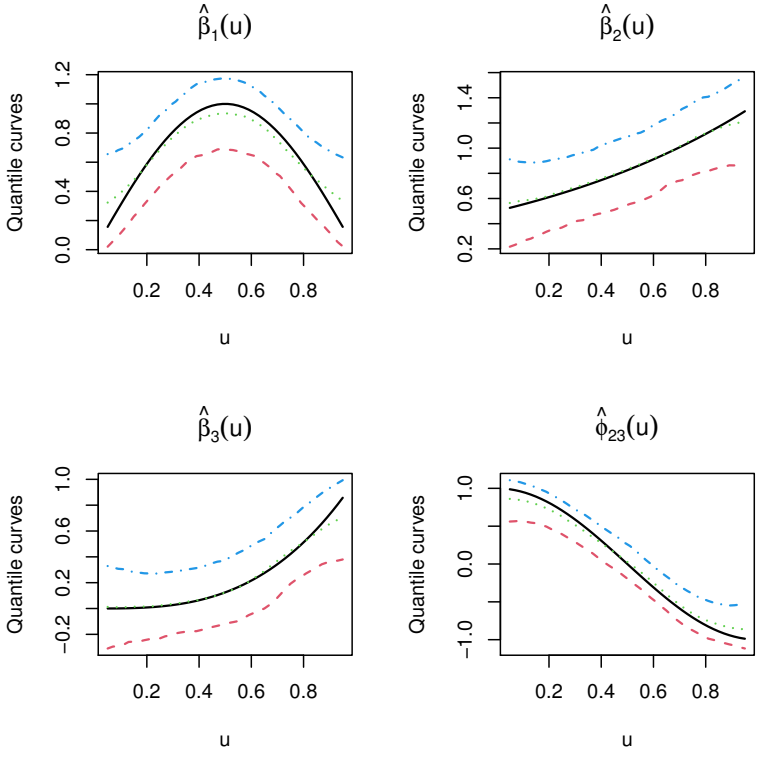

Figure 1. Estimated quantile curves of Model 1. Dash-dot lines are 0.95 quantile curves, dotted lines are 0.50 quantile curves, dashed lines are 0.05 quantile curves, and solid lines are the real function curves. The dash-dot and dashed lines indicate 90% confident bands.

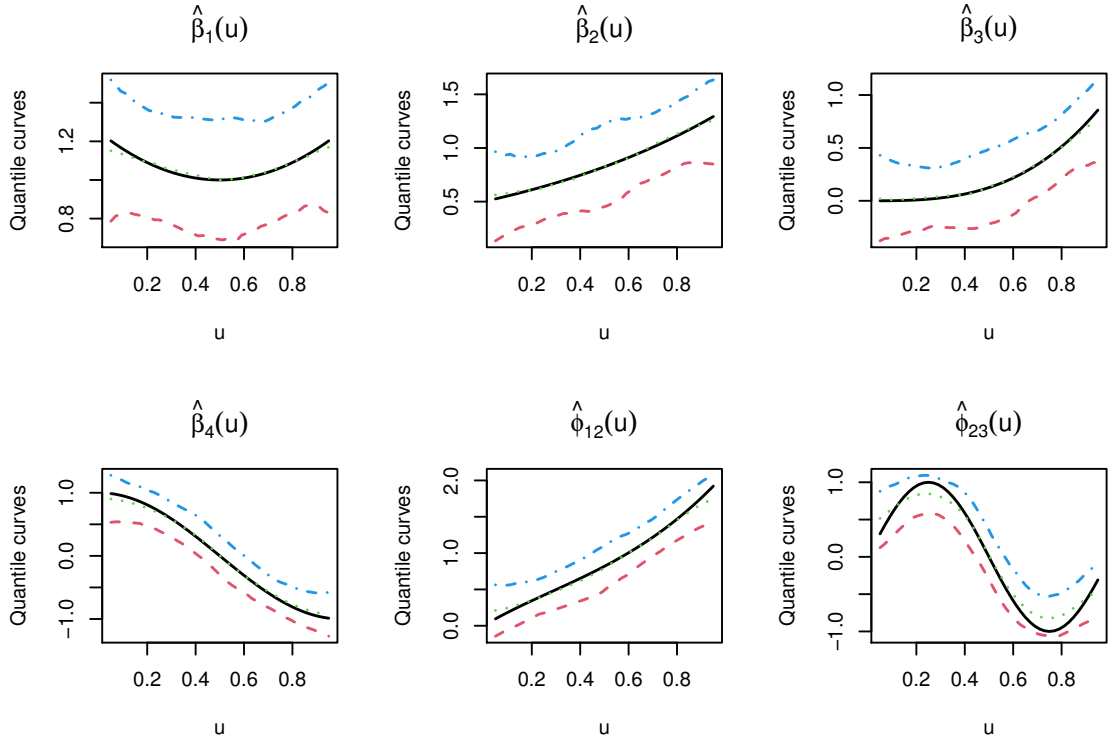

Figure 2. Estimated quantile curves of Model 2. Dash-dot lines are 0.95 quantile curves, dotted lines are 0.50 quantile curves, dashed lines are 0.05 quantile curves, and solid lines are the real function curves. The dash-dot and dashed lines indicate 90% confident bands.

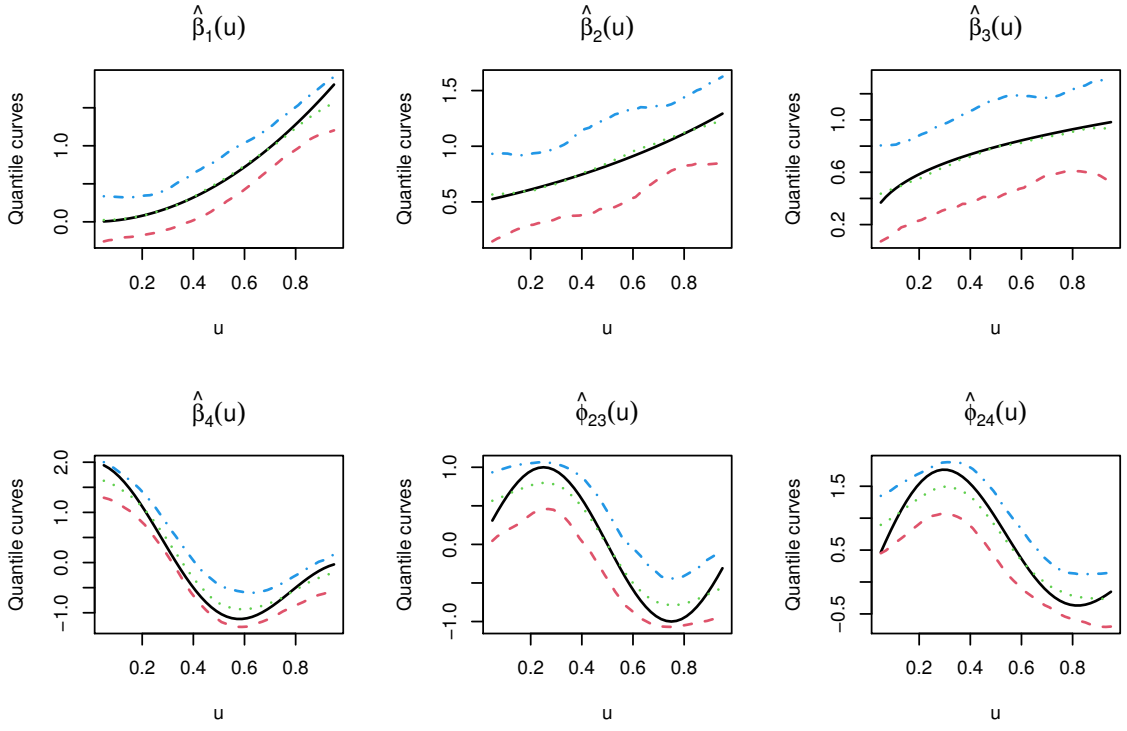

Figure 3. Estimated quantile curves of Model 3. Dash-dot lines are 0.95 quantile curves, dotted lines are 0.50 quantile curves, dashed lines are 0.05 quantile curves, and solid lines are the real function curves. The dash-dot and dashed lines indicate 90% confident bands.

Table 4. CM, CZ, and CS for Models 1–3.

	n	CM	CZ	CS
Model 1	100	0.922	0.711	0.658
	200	0.996	0.985	0.982
	500	1.000	1.000	1.000
Model 2	100	0.884	0.575	0.524
	200	0.987	0.972	0.960
	500	1.000	1.000	1.000
Model 3	100	0.932	0.533	0.517
	200	0.973	0.952	0.937
	500	1.000	1.000	1.000

4.2. The Boston Housing Data Analysis

To further investigate the performance of our method, we apply the proposed method to the Boston housing data which concerns the median value of owner-occupied homes (MV) for 506 census tracts in 1970. The dataset "Boston" is available in the R package "MASS" ([34]). Following the basic housing value equation of Harrison and Rubinfeld [35] and the study of Fan and Huang [3], Wang and Xia [7], we consider seven covariates here: CRIM (per capita crime rate by town), RM (average number of rooms per dwelling), NOX (nitric oxides concentration), PTRATIO (pupil-teacher ratio by town), TAX (full-value property-tax rate per $10,000), AGE (proportion of owner-occupied units built prior to 1940), and LSTAT (percentage of lower status of the population). The log transformation of MV and power transformation of RM, NOX, and LSTAT are employed to fit the Boston data well. For simplicity of representation, CRIM, RM^2, TAX, NOX^2, PTRATIO, and AGE are, respectively, denoted by $X_1, \cdots, X_6$, and they are all scaled with mean zero and standard deviation 1. Meanwhile, we take $X_0 \equiv 1$ as the intercept term (INT), log(MV) as the response Y, and $\text{LSTAT}^{1/2}$ as the index variable U. Consequently, the varying coefficient model with hierarchical structure

$$Y_i = \beta_0(U_i) + \sum_{j=1}^{6} X_{ij}\beta_j(U_i) + \sum_{1\le j<k\le 6} X_{ij}X_{ik}\phi_{jk}(U_i) + \varepsilon_i$$

is fitted to the Boston housing data.

The data are divided into training data and testing data with sample size 405 and 101, respectively, by sampling. Estimates are based on the training data, and performance of the proposed procedures are evaluated on the testing data. The CV criterion suggests a bandwidth $h = 0.31$, and the optimal tuning parameter selected by BIC criterion is $\hat{\lambda}_0 = 1.8$. During the implementation procedures, the variables are regarded insignificant for the mode of whose estimated functional coefficients are less than 0.1. It shows that INT, CRIM, RM^2, TAX, PTRATIO, AGE, $\text{RM}^2 \times$TAX are significant but the others are not. The estimated coefficient function curves for these relevant variables are depicted in Figure 4. For the testing data, compared with the VC model without considering the interaction terms, the multiple R^2 for our proposed procedure is 0.8314, and it is 0.8021 without interaction terms. Thus, the proposed VC model with hierarchical structure can fit the testing data slightly better.

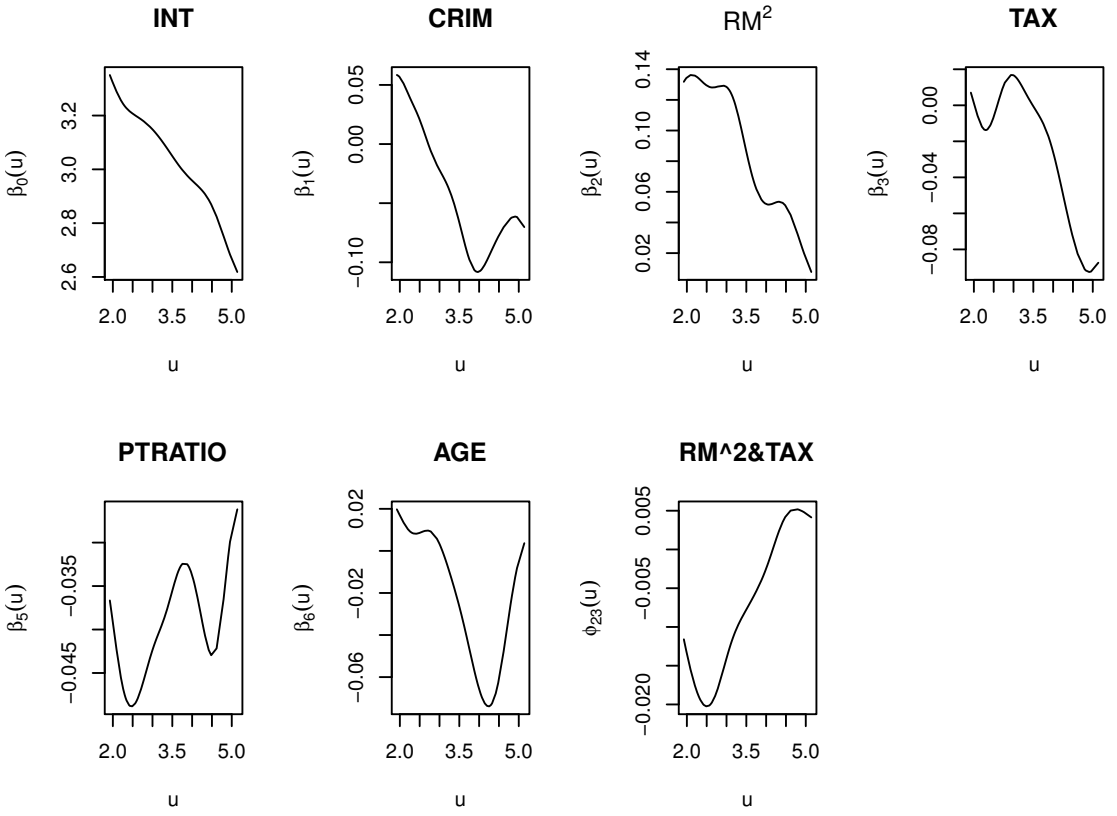

Figure 4. Estimated functional coefficients curves of the relevant variables.

5. Conclusions and Future Works

In this paper, the VC models with hierarchical structure are investigated, and a unified variable selection procedure is proposed, which can simultaneously select the nonzero effects and estimate the unknown coefficient functions, while the selected model enforces the hierarchical structure. It is proved that the proposed penalty estimators have the oracle properties, that is the coefficients are estimated as well as if the true model were known in advance. Simulation studies and Boston housing data analysis are carried out to examine the performance of the proposed method in finite sample case.

However, we mainly focus on the fixed dimensionality of the predictive covariates in the paper. We will investigate the VC model with hierarchical structure in the case of diverging dimensionality of the predictors in the future. Estimation and variable selection for the generalized varying coefficient model with hierarchical structure, as well as estimation for the semiparametric varying coefficient partially linear model with hierarchical structure, are also interesting topics that deserve to be studied.

Appendix A

Before proving Theorems 1 and 2, we first investigate the following lemmas, which aim for the asymptotic properties of $\widehat{\mathbf{\Lambda}}_\lambda$. Without loss of generality, we suppose $\mathcal{M} = \{1, 2, \cdots, d_0\}$, $\mathcal{I} = \{(j,k) : 1 \le j < k \le d_1\}$, where $d_1 < d_0$. Let $\boldsymbol{\alpha}_\mathcal{S}(U_i) = (\beta_1(U_i), \cdots, \beta_{d_0}(U_i)$, $\phi_{12}(U_i), \cdots, \phi_{(d_1-1)d_1}(U_i))^\tau$, $\boldsymbol{\beta}_\mathcal{M} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_{d_0})$, $\boldsymbol{\phi}_\mathcal{I} = (\boldsymbol{\phi}_{12}, \boldsymbol{\phi}_{13}, \cdots, \boldsymbol{\phi}_{(d_1-1)d_1})$.

Lemma A1. *Suppose Conditions C1–C6 hold,* $h \propto n^{-\frac{1}{5}}$, $(nh)^{-\frac{1}{2}} a_n \to 0, (nh)^{-\frac{1}{2}} b_n \to \infty$; *we have*

$$\frac{1}{n} \sum_{t=1}^{n} \|\widehat{\boldsymbol{\alpha}}_\lambda(U_t) - \boldsymbol{\alpha}(U_t)\|^2 = O_p(n^{-\frac{4}{5}}). \tag{A1}$$

Proof. We use $A = (a_{ij}) \in R^{n\times\frac{d(d+1)}{2}}$ to denote an arbitrary $n \times \frac{d(d+1)}{2}$ matrix, the rows of A are defined as $\boldsymbol{\eta}_1^\tau, \cdots, \boldsymbol{\eta}_n^\tau$, and the columns of which are defined as $\boldsymbol{\rho}_1, \cdots, \boldsymbol{\rho}_d, \boldsymbol{\rho}_{12}, \cdots, \boldsymbol{\rho}_{1d}$, $\boldsymbol{\rho}_{23}, \cdots, \boldsymbol{\rho}_{2d}, \cdots, \boldsymbol{\rho}_{(d-1)d}$. Use L_2 norm to define $\|A\|^2 = \Sigma a_{ij}^2$. According Fan and Li [10], it suffices to show that, for any small $\epsilon > 0$, there exists a constant $C > 0$; we have

$$\liminf_{n\to\infty} P\left(\inf_{n^{-1}\|A\|^2=C} Q_\lambda(\boldsymbol{\Lambda} + (nh)^{-\frac{1}{2}}A) > Q_\lambda(\boldsymbol{\Lambda})\right) = 1-\epsilon. \tag{A2}$$

Define $D_n(A) = Q_\lambda(\boldsymbol{\Lambda} + (nh)^{-1/2}A) - Q_\lambda(\boldsymbol{\Lambda})$. Then, we have

$$\begin{aligned}
D_n(A) =& \sum_{t=1}^n\sum_{i=1}^n \Big[(nh)^{-1}\boldsymbol{\eta}_t^\tau W_i W_i^\tau \boldsymbol{\eta}_t + 2(nh)^{-\frac{1}{2}}(Y_i - W_i^\tau \boldsymbol{\alpha}(U_t))W_i^\tau\boldsymbol{\eta}_t\Big]K_h(U_t-U_i)\\
&+\sum_{j=1}^d \lambda_j^1\left(\sqrt{\|\boldsymbol{\beta}_j + (nh)^{-\frac{1}{2}}\boldsymbol{\rho}_j\|^2 + \sum_{k:k\neq j}\|\boldsymbol{\phi}_{jk} + (nh)^{-\frac{1}{2}}\boldsymbol{\rho}_{jk}\|^2} - \sqrt{\|\boldsymbol{\beta}_j\|^2 + \sum_{k=j+1}^d\|\boldsymbol{\phi}_{jk}\|^2}\right)\\
&+\sum_{j=1}^d\sum_{k=j+1}^d \lambda_{jk}^2\Big(\|\boldsymbol{\phi}_{jk} + (nh)^{-\frac{1}{2}}\boldsymbol{\rho}_{jk}\| - \|\boldsymbol{\phi}_{jk}\|\Big)\\
=& R_0 + R_1 + R_2.
\end{aligned}$$

Next, we discuss $R_i, i = 0, 1, 2$, respectively. For R_0,

$$\begin{aligned}
R_0 =& \sum_{t=1}^n\sum_{i=1}^n \Big[(nh)^{-1}\boldsymbol{\eta}_t^\tau W_i W_i^\tau \boldsymbol{\eta}_t + 2(nh)^{-\frac{1}{2}}(Y_i - W_i^\tau \boldsymbol{\alpha}(U_t))W_i^\tau\boldsymbol{\eta}_t\Big]K_h(U_t-U_i)\\
=& \sum_{t=1}^n \frac{1}{h}\boldsymbol{\eta}_t^\tau\left(\frac{1}{n}\sum_{i=1}^n W_i W_i^\tau K_h(U_t-U_i)\right)\boldsymbol{\eta}_t\\
&+2(nh)^{-\frac{1}{2}}\sum_{t=1}^n\sum_{i=1}^n \boldsymbol{\eta}_t^\tau W_i[(\boldsymbol{\alpha}(U_i) - \boldsymbol{\alpha}(U_t))^\tau W_i + \varepsilon_i]K_h(U_t-U_i)\\
=& \frac{1}{h}\sum_{t=1}^n \boldsymbol{\eta}_t^\tau\Sigma(U_t)\boldsymbol{\eta}_t + 2\frac{1}{h}\sum_{t=1}^n\boldsymbol{\eta}_t^\tau \boldsymbol{e}_t\\
\geq& nh^{-1}\lambda_{min}(n^{-1}\sum_{t=1}^n\|\boldsymbol{\eta}_t\|^2) - 2nh^{-1}(\sum_{t=1}^n n^{-\frac{1}{2}}\|\boldsymbol{\eta}_t\|)(\sum_{t=1}^n n^{-\frac{1}{2}}\|\boldsymbol{e}_t\|)\\
\geq& nh^{-1}\lambda_{min}C - 2nh^{-1}C^{\frac{1}{2}},
\end{aligned}$$

where $\Sigma(U_t) = n^{-1}\sum_{i=1}^n W_i W_i^\tau K_h(U_t-U_i)$, $\boldsymbol{e}_t = n^{-\frac{1}{2}}h^{\frac{1}{2}}\sum_{i=1}^n W_i[(\boldsymbol{\alpha}(U_i) - \boldsymbol{\alpha}(U_t))^\tau W_i + \varepsilon_i]K_h(U_t-U_i)$, $\lambda_{min}(U_t)$ is the minimum eigenvalue of $\Sigma(U_t)$, λ_{min} is the minimizer of $\{\lambda_{min}(U_t), t = 1, 2, \cdots, n\}$, and $n^{-1}\sum_{t=1}^n\|\boldsymbol{e}_t\|^2 = O_p(1)$, which we prove in Lemma 2. According to $n^{-1}\|A\|^2 = C$ and $a_n = \max\{\lambda_j^1, \lambda_{jk}^2 : j \in \mathcal{M}, (j,k) \in \mathcal{I}\}$, it is easy to show that

$$\begin{aligned}
R_1 =& \sum_{j=1}^d \lambda_j^1\left(\sqrt{\|\boldsymbol{\beta}_j + (nh)^{-\frac{1}{2}}\boldsymbol{\rho}_j\|^2 + \sum_{k:k\neq j}\|\boldsymbol{\phi}_{jk} + (nh)^{-\frac{1}{2}}\boldsymbol{\rho}_{jk}\|^2} - \sqrt{\|\boldsymbol{\beta}_j\|^2 + \sum_{k=j+1}^d\|\boldsymbol{\phi}_{jk}\|^2}\right)\\
\geq& \sum_{j=1}^{d_0}\lambda_j^1\left(\sqrt{\|\boldsymbol{\beta}_j\|^2 + \sum_{k=j+1}^{d_0}\|\boldsymbol{\phi}_{jk}\|^2} - \sqrt{(nh)^{-1}\|\boldsymbol{\rho}_j\|^2 + (nh)^{-1}\sum_{k=j+1}^{d_0}\|\boldsymbol{\rho}_{jk}\|^2}\right.\\
&\left.-\sqrt{\|\boldsymbol{\beta}_j\|^2 + \sum_{k=j+1}^{d_0}\|\boldsymbol{\phi}_{jk}\|^2}\right)\\
\geq& -d_0 C^{\frac{1}{2}}h^{-\frac{1}{2}}a_n.
\end{aligned}$$

$$\begin{aligned}
R_2 &= \sum_{j=1}^{d} \sum_{k:j<k\le d} \lambda_{jk}^2 \Big(\|\boldsymbol{\phi}_{jk} + (nh)^{-\frac{1}{2}} \boldsymbol{\rho}_{jk}\| - \|\boldsymbol{\phi}_{jk}\| \Big) \\
&\ge \sum_{j=1}^{d_0} \sum_{k:j<k\le d_0} \lambda_{jk}^2 \Big(\|\boldsymbol{\phi}_{jk} + (nh)^{-\frac{1}{2}} \boldsymbol{\rho}_{jk}\| - \|\boldsymbol{\phi}_{jk}\| \Big) \\
&\ge - \sum_{j=1}^{d_0} \sum_{k:j<k\le d_0} \lambda_{jk}^2 (nh)^{-\frac{1}{2}} \|\boldsymbol{\rho}_{jk}\| \\
&\ge - (nh)^{-\frac{1}{2}} n^{\frac{1}{2}} C^{\frac{1}{2}} a_n \\
&= - C^{\frac{1}{2}} h^{-\frac{1}{2}} a_n.
\end{aligned}$$

Consequently, we have

$$\begin{aligned}
hn^{-1} D_n(A) =& hn^{-1}(R_0 + R_1 + R_2) \\
\ge& hn^{-1}(nh^{-1}\lambda_{min} C - 2nh^{-1} C^{\frac{1}{2}} - d_0 a_n C^{\frac{1}{2}} h^{-\frac{1}{2}} - C^{\frac{1}{2}} h^{-\frac{1}{2}} a_n) \\
=& \lambda_{min} C - 2C^{\frac{1}{2}} - n^{-\frac{1}{2}} h (d_0 a_n + a_n) C^{\frac{1}{2}} \\
=& \lambda_{min} C - 2C^{\frac{1}{2}} - n^{-\frac{1}{5}} ((d_0 + 1)(nh)^{-\frac{1}{2}} a_n) C^{\frac{1}{2}} \\
\ge& 0,
\end{aligned}$$

when C is large enough, the result of Lemma 1 holds. □

Lemma A2. *Suppose C1 − C6 hold; then,* $n^{-1} \sum_{t=1}^{n} \|\boldsymbol{e}_t\|^2 = O_p(1)$.

Proof. By straightforward algebra, we have

$$\begin{aligned}
& \mathrm{E}(n^{-1} \sum_{t=1}^{n} \|\boldsymbol{e}_t\|^2) \\
=& n^{-2} h \sum_{t=1}^{n} \mathrm{E}[\sum_{i\ne j}^{n} (\boldsymbol{\alpha}(U_i) - \boldsymbol{\alpha}(U_t))^\tau W_i W_i^\tau W_j W_j^\tau (\boldsymbol{\alpha}(U_j) - \boldsymbol{\alpha}(U_t)) K_h(U_t - U_i) \\
& K_h(U_t - U_j) + \sum_{i=j}^{n} (\boldsymbol{\alpha}(U_i) - \boldsymbol{\alpha}(U_t))^\tau W_i W_i^\tau W_i W_i^\tau (\boldsymbol{\alpha}(U_i) - \boldsymbol{\alpha}(U_t)) K_h^2(U_t - U_i) \\
& + \sum_{i=1}^{n} \varepsilon_i^2 W_i^\tau W_i K_h^2(U_t - U_i)] \\
=& n^{-2} h \sum_{t=1}^{n} \mathrm{E}(\sum_{i\ne j, i=1}^{n} \boldsymbol{e}_{t1} + \sum_{i=j, i=1}^{n} \boldsymbol{e}_{t2} + \sum_{i=1}^{n} \boldsymbol{e}_{t3}).
\end{aligned}$$

By Tylor expansion (7), we have

$$\begin{aligned}
& \mathrm{E}(e_{t1}) \\
=& \mathrm{E}(\mathrm{E}(e_{t1} \mid U_i, U_j)) \\
=& \mathrm{E}(\mathrm{E}((\boldsymbol{\alpha}(U_i) - \boldsymbol{\alpha}(U_t))^\tau W_i W_i^\tau W_j W_j^\tau (\boldsymbol{\alpha}(U_j) - \boldsymbol{\alpha}(U_t)) K_h(U_t - U_i) K_h(U_t - U_j) \mid U_i, U_j))) \\
=& \mathrm{E}(\mathrm{E}(\boldsymbol{\alpha}'(U_t)^\tau W_i W_i^\tau W_j W_j^\tau \boldsymbol{\alpha}'(U_t)(U_t - U_i)(U_t - U_j) K_h(U_t - U_i) K_h(U_t - U_j) \\
& + \|W_i\|^2 \|W_j\|^2 (U_t - U_i)^2 (U_t - U_j)^2 K_h(U_t - U_i) K_h(U_t - U_j) \mid U_i, U_j) \\
=& E_1 + E_2.
\end{aligned}$$

When $t = i$ or $t = j$, $E_1 = E_2 = 0$; $t \ne i$ and $t \ne j$, according to C3,

$$
\begin{aligned}
E_1 =&\mathrm{E}(\mathrm{E}(\boldsymbol{\alpha}'(U_t)^\tau W_i W_i^\tau W_j W_j^\tau \boldsymbol{\alpha}'(U_t)(U_t-U_i)(U_t-U_j)K_h(U_t-U_i)K_h(U_t-U_j) \mid U_t,U_i,U_j))\\
=&\mathrm{E}(\boldsymbol{\alpha}'(U_t)^\tau E(W_i W_i^\tau W_j W_j^\tau \mid U_i,U_j)\boldsymbol{\alpha}'(U_t)(U_t-U_i)(U_t-U_j)K_h(U_t-U_i)K_h(U_t-U_j))\\
=&\int \boldsymbol{\alpha}'(U_t)^\tau \Omega(U_i)\Omega(U_j)\boldsymbol{\alpha}'(U_t)(U_t-U_i)(U_t-U_j)K_h(U_t-U_i)K_h(U_t-U_j)\\
&f(U_t)f(U_i)f(U_j)dU_t dU_i dU_j\\
=&h^2\int \boldsymbol{\alpha}'(U_t)^\tau \Omega(U_t+hs_1)\Omega(U_t+hs_1)f(U_t+hs_1)f(U_t+hs_2)\boldsymbol{\alpha}'(U_t)s_1s_2K(s_1)K(s_2)dU_t ds_1 ds_2\\
=&h^2\int \boldsymbol{\alpha}'(U_t)^\tau(\tilde{\omega}(U_t)+\tilde{\omega}_1'(U_t)hs_1+\tilde{\omega}_2'(U_t)hs_2+C_3(s_1^2+s_2^2))\boldsymbol{\alpha}'(U_t)h^2s_1s_2K(s_1)K(s_2)dU_t ds_1 ds_2\\
=&O(h^4),
\end{aligned}
$$

for $\int(\tilde{\omega}(U_t)+\tilde{\omega}_1'(U_t)hs_1+\tilde{\omega}_2'(U_t)hs_2)s_1s_2K(s_1)K(s_2)ds_1ds_2 = 0$. Similarly, we can get $E_2 = O(h)$, so $\mathrm{E}(\boldsymbol{e}_{t1}) = O(h^4)$, $\mathrm{E}(\boldsymbol{e}_{t2}) = O(h)$.

Next, we consider $\mathrm{E}(\sum_{i=1}^n \boldsymbol{e}_{t3})$, define $g(U_t) = f^2(U_t)$, and suppose that $\widetilde{\mathrm{E}}(\|W_t\varepsilon_t\|^2|U_t = u) = \int \|W_t\varepsilon_t\|^2 g(U_t)dU_t$ is bounded.

$$
\begin{aligned}
\mathrm{E}(\sum_{i=1}^n \boldsymbol{e}_{t3}) =&(n-1)\mathrm{E}(\|W_i\varepsilon_i\|^2K_h^2(U_t-U_i)) + K^2(0)\mathrm{E}(\|W_i\varepsilon_i\|^2)\\
=&(n-1)\mathrm{E}\{\mathrm{E}(\|W_i\varepsilon_i\|^2K_h^2(U_t-U_i)|U_t,U_i) + K^2(0)\mathrm{E}(\|W_i\varepsilon_i\|^2|U_i=U)\}\\
=&(n-1)\int \varepsilon_i^2W_i^\tau W_iK_h^2(U_t-U_i)f(U_t)f(U_i)dU_tdU_i + K^2(0)\mathrm{E}(\|W_i\varepsilon_i\|^2|U_i=U)\}\\
=&(n-1)h^{-1}\int \varepsilon_i^2W_i^\tau W_iK^2(v)f(U_i+hv)f(U_i)dU_idv + K^2(0)\mathrm{E}(\|W_i\varepsilon_i\|^2|U_i=U)\}\\
=&(n-1)h^{-1}\int \varepsilon_i^2W_i^\tau W_iK^2(v)f(U_i)\{f(U_i)+f'(U_i)hv+C_1h^2v^2\}dU_idv\\
&+K^2(0)\mathrm{E}(\|W_i\varepsilon_i\|^2|U_i=U)\\
=&(n-1)h^{-1}\int \varepsilon_i^2W_i^\tau W_iK^2(v)f(U_i)\{f(U_i)+C_1h^2v^2\}dU_idv + K^2(0)\mathrm{E}(\|W_i\varepsilon_i\|^2|U_i=U)\\
=&(n-1)h^{-1}\{\int \varepsilon_i^2W_i^\tau W_ig(U_i)dU_i\int K^2(v)dv + C_1h^3\int \varepsilon_i^2W_i^\tau W_if(U_i)dU_i\int K_h^2(v)v^2dv\}\\
&+K^2(0)\mathrm{E}(\|W_i\varepsilon_i\|^2|U_i=U)\\
\le&(n-1)h^{-1}\{\nu\widetilde{\mathrm{E}}(\|W_i\varepsilon_i\|^2|U_i=U) + C_1h^3\nu_2\mathrm{E}(\|W_i\varepsilon_i\|^2|U_i=U)\}\\
&+K^2(0)\mathrm{E}(\|W_i\varepsilon_i\|^2|U_i=U),
\end{aligned}
$$

where $\nu = \int K^2(v)dv$ and $\nu_2 = \int K^2(v)v^2dv$, we can get

$$
\begin{aligned}
n^{-1}h\mathrm{E}(\sum_{i=1}^n \boldsymbol{e}_{t3}) =&n^{-1}h\{(n-1)h^{-1}\nu\widetilde{\mathrm{E}}(\|W_i\varepsilon_i\|^2|U_i=U) + C_1h^3\nu_2\mathrm{E}(\|W_i\varepsilon_i\|^2|U_i=U)\\
&+K^2(0)\mathrm{E}(\|W_i\varepsilon_i\|^2|U_i=U)\}\\
\le&\widetilde{\mathrm{E}}(\|W_i\varepsilon_i|^2|U_i=U)\nu + n^{-\frac{9}{5}}C_1\mathrm{E}(\|W_i\varepsilon_t\|^2|U_i=U)\nu_2 + n^{-\frac{6}{5}}K^2(0)\mathrm{E}(\|W_i\varepsilon_i\|^2|U_i=U)\\
<&\infty
\end{aligned}
$$

for $n^{-\frac{9}{5}}, n^{-\frac{6}{5}} \to 0$. Thus, $n^{-1}\sum_{t=1}^n \|\boldsymbol{e}_t\|^2 = n^{-1}h(O_p(n^2h^4)+nO_p(h)) + O_p(1) = O_p(1)$ can be proved. □

Proof of Theorem 1. According to the definition above, we have $\mathcal{M}^c = \{d_0+1, d_0+2, \cdots, d\} = \{j : \|\boldsymbol{\beta}_j\| = 0\}$, $\mathcal{I}^c = \{(j,k) : d_1 < j < k \le d_0 \le d$, or $d_1 \le j < d_0 < k \le d$, or $d_0 < j < k \le d\} = \{(j,k) : \|\boldsymbol{\phi}_{jk}\| = 0\}$. Meanwhile, $\widehat{\mathcal{M}^c} = \{j : \|\widehat{\boldsymbol{\beta}}_j\| = 0\}$, $\widehat{\mathcal{I}^c} = \{(j,k) : \|\widehat{\boldsymbol{\phi}}_{j,k}\| = 0\}$.

We first prove that $P(\widehat{\mathcal{M}^c} = \mathcal{M}^c) \to 1$ as $n \to \infty$. That is for any $j \in \mathcal{M}^c$, $P(\widehat{\beta}_j(U_t) = 0) \to 1$ for $1 \le t \le n$. If it is not true, $\widehat{\beta}_j(U_t) \neq 0$ must be the solution of the following normal equation,

$$\begin{aligned}
0 = \frac{\partial Q_\lambda(\widehat{\Lambda})}{\partial \beta_j(U_t)} &= -2\sum_{t=1}^{n}\sum_{i=1}^{n} X_{ij}(Y_i - W_i^\tau \widehat{\boldsymbol{\alpha}}_\lambda(U_t))K_h(U_t - U_i) + 2\lambda_j^1 \left(\frac{\widehat{\beta}_j(U_t)}{\sqrt{\|\widehat{\boldsymbol{\beta}}_j\|^2 + \sum_{k:k\neq j}\|\widehat{\boldsymbol{\phi}}_{jk}\|^2}} \right) \\
&= -2\sum_{t=1}^{n}\sum_{i=1}^{n} X_{ij}(Y_i - W_i^\tau \boldsymbol{\alpha}(U_t) - W_i^\tau(\widehat{\boldsymbol{\alpha}}_\lambda(U_t) - \boldsymbol{\alpha}(U_t)))K_h(U_t - U_i) \\
&\quad + 2\lambda_j^1 \left(\frac{\widehat{\beta}_j(U_t)}{\sqrt{\|\widehat{\boldsymbol{\beta}}_j\|^2 + \sum_{k:k\neq j}\|\widehat{\boldsymbol{\phi}}_{jk}\|^2}} \right) \\
&= -2\sum_{t=1}^{n}\sum_{i=1}^{n} X_{ij}\varepsilon_i K_h(U_t - U_i) + 2\sum_{t=1}^{n}\sum_{i=1}^{n} X_{ij}W_i^\tau(\widehat{\boldsymbol{\alpha}}_\lambda(U_t) - \boldsymbol{\alpha}(U_t))K_h(U_t - U_i) \\
&\quad + 2(nh)^{\frac{1}{2}} \frac{\lambda_j^1}{(nh)^{\frac{1}{2}}} \left(\frac{\widehat{\beta}_j(U_t)}{\sqrt{\|\widehat{\boldsymbol{\beta}}_j\|^2 + \sum_{k:k\neq j}\|\widehat{\boldsymbol{\phi}}_{jk}\|^2}} \right),
\end{aligned}$$

by Conditions C1-C6 and Lemma 2, we know that $\sum_{t=1}^{n}\sum_{i=1}^{n} X_{ij}\varepsilon_i K_h(U_t - U_i) = O_p(n^2)$. Meanwhile, according to Lemma 1, $\|\widehat{\boldsymbol{\alpha}}_\lambda(U_t) - \boldsymbol{\alpha}(U_t)\| = O_p(n^{-\frac{2}{5}})$, $\|\sum_{t=1}^{n}\sum_{i=1}^{n} X_{ij}W_i^\tau(\widehat{\boldsymbol{\alpha}}_\lambda(U_t) - \boldsymbol{\alpha}(U_t))K_h(U_t - U_i)\| \le \sum_{t=1}^{n}\sum_{i=1}^{n}\|X_{ij}W_i^\tau K_h(U_t - U_i)\|\|\widehat{\boldsymbol{\alpha}}_\lambda(U_t) - \boldsymbol{\alpha}(U_t)\| = O_p(n^2h^2)$, and then we have that $\frac{\widehat{\beta}_j(U_t)}{\sqrt{\|\widehat{\boldsymbol{\beta}}_j\|^2 + \sum_{k:k\neq j}\|\widehat{\boldsymbol{\phi}}_{jk}\|^2}} = O_p(1)$. In addition, for $b_n = \min\{\lambda_j^1, \lambda_{kl}^2 : j \in \mathcal{M}^c, (k,l) \in \mathcal{I}^c\}$, we get $\frac{\partial Q_\lambda(\widehat{\Lambda})}{\partial \beta_j(U_t)} = O_p(n^2) + O_p(n^2h^2) + (nh)^{\frac{1}{2}}O_p((nh)^{-\frac{1}{2}}\lambda_j^1) \ge O_p(n^2) + O_p(n^2h^2) + (nh)^{\frac{1}{2}}O_p((nh)^{-\frac{1}{2}}b_n) \to \infty$ for $(nh)^{-\frac{1}{2}}b_n \to \infty$. It clearly shows that for $j \in \mathcal{M}^c$, there are no solutions for $\frac{\partial Q_\lambda(\widehat{\Lambda})}{\partial \beta_j(U_t)} = 0$ with $\hat{\beta}_j(U_t) \neq 0$, which is to say for any $j \in \mathcal{M}^c$, then $j \in \widehat{M^c}$, so $P(\widehat{\mathcal{M}^c} = \mathcal{M}^c) \to 1$ is proved, natually we get that $P(\widehat{\mathcal{M}} = \mathcal{M}) \to 1$.

Then, we pay attention to $P(\widehat{\mathcal{I}^c} = \mathcal{I}^c) \to 1$ as $n \to \infty$, i.e., for any $(j,k) \in \mathcal{I}^c$, $P(\widehat{\phi}_{jk}(U_t) = 0) \to 1$ for $1 \le t \le n$. In addition, if the claim is not true, $\hat{\phi}_{jk}(U_t) \neq 0$ must be the solution of the following normal equation,

$$\begin{aligned}
0 = \frac{\partial Q_\lambda(\widehat{\Lambda})}{\partial \phi_{jk}(U_t)} &= -2\sum_{t=1}^{n}\sum_{i=1}^{n} Z_{ij}(Y_i - W_i^\tau \widehat{\boldsymbol{\alpha}}_\lambda(U_t))K_h(U_t - U_i) + 2\lambda_j^1 \frac{\widehat{\phi}_{jk}(U_t)}{\sqrt{\|\widehat{\boldsymbol{\beta}}_j\|^2 + \sum_{k:k\neq j}\|\widehat{\boldsymbol{\phi}}_{jk}\|^2}} \\
&\quad + 2\lambda_k^1 \frac{\widehat{\phi}_{jk}(U_t)}{\sqrt{\|\widehat{\boldsymbol{\beta}}_k\|^2 + \sum_{j:k\neq j}\|\widehat{\boldsymbol{\phi}}_{jk}\|^2}} + 2\lambda_{jk}^2 \frac{\widehat{\phi}_{jk}(U_t)}{\|\widehat{\boldsymbol{\phi}}_{jk}\|} \\
&= -2\sum_{t=1}^{n}\sum_{i=1}^{n} Z_{ij}\varepsilon_i K_h(U_t - U_i) + 2\sum_{t=1}^{n}\sum_{i=1}^{n} Z_{ij}W_i^\tau(\widehat{\boldsymbol{\alpha}}_\lambda(U_t) - \boldsymbol{\alpha}(U_t))K_h(U_t - U_i) \\
&\quad + 2(nh)^{\frac{1}{2}} \frac{\lambda_j^1}{(nh)^{\frac{1}{2}}} \frac{\widehat{\phi}_{jk}(U_t)}{\sqrt{\|\widehat{\boldsymbol{\beta}}_j\|^2 + \sum_{k:k\neq j}\|\widehat{\boldsymbol{\phi}}_{jk}\|^2}} \\
&\quad + 2(nh)^{\frac{1}{2}} \frac{\lambda_k^1}{(nh)^{\frac{1}{2}}} \frac{\widehat{\phi}_{jk}(U_t)}{\sqrt{\|\widehat{\boldsymbol{\beta}}_k\|^2 + \sum_{j:k\neq j}\|\widehat{\boldsymbol{\phi}}_{jk}\|^2}} + 2(nh)^{\frac{1}{2}} \frac{\lambda_{jk}^2}{(nh)^{\frac{1}{2}}} \frac{\widehat{\phi}_{jk}(U_t)}{\|\widehat{\boldsymbol{\phi}}_{jk}\|}.
\end{aligned}$$

Referring to the definition, we consider the following three situations for $(j,k) \in \mathcal{I}^c$,

(1) $(j,k) \in \mathcal{I}^c$, and $j,k \in \mathcal{M}$;
(2) $(j,k) \in \mathcal{I}^c$, and $j \in \mathcal{M}$ and $k \in \mathcal{M}^c$; and
(3) $(j,k) \in \mathcal{I}^c$, and $j,k \in \mathcal{M}^c$.

By Conditions C1-C6 and Lemma 2, we know that $\sum_{t=1}^{n}\sum_{i=1}^{n} Z_{ij}\varepsilon_i K_h(U_t - U_i) = O_p(n^2)$. Meanwhile, according to Lemma 1, $\|\widehat{\boldsymbol{\alpha}}_\lambda(U_t) - \boldsymbol{\alpha}(U_t)\| = O_p(n^{-\frac{2}{5}})$, $\|\sum_{t=1}^{n}\sum_{i=1}^{n} Z_{ij}W_i^\tau (\widehat{\boldsymbol{\alpha}}_\lambda(U_t) - \boldsymbol{\alpha}(U_t))K_h(U_t - U_i)\| \le \sum_{t=1}^{n}\sum_{i=1}^{n} \|Z_{ij}W_i^\tau K_h(U_t - U_i)\|\|\widehat{\boldsymbol{\alpha}}_\lambda(U_t) - \boldsymbol{\alpha}(U_t)\| = O_p(n^2h^2)$, and have that $\frac{\widehat{\phi}_{jk}(U_t)}{\sqrt{\|\widehat{\boldsymbol{\beta}}_j\|^2 + \sum_{k:k\neq j}\|\widehat{\boldsymbol{\phi}}_{jk}\|^2}} = O_p(1)$, $\frac{\widehat{\phi}_{jk}(U_t)}{\|\widehat{\boldsymbol{\phi}}_{jk}\|} = O_p(1)$. We get $\frac{\partial Q_\lambda(\widehat{\Lambda})}{\partial \phi_{jk}(U_t)} = O_p(n^2) + O_p(n^2h^2) + (nh)^{\frac{1}{2}}O_p((nh)^{-\frac{1}{2}}(\lambda_j^1 + \lambda_k^1 + \lambda_{jk}^2))$. In addition, for $a_n = \max\{\lambda_j^1, \lambda_k^1, \lambda_{kl}^2 : j \in \mathcal{M}, (k,l) \in \mathcal{I}\}$, we can get that $\nu_n = \min\{\lambda_j^1, \lambda_k^1, \lambda_{kl}^2 : j \in \mathcal{M}, (k,l) \in \mathcal{I}\}$ and $(nh)^{-\frac{1}{2}}\nu_n \le (nh)^{-\frac{1}{2}}a_n \to 0$, while $b_n = \min\{\lambda_j^1, \lambda_k^1, \lambda_{kl}^2 : j \in \mathcal{M}^c, (k,l) \in \mathcal{I}^c\}$.

Then, for Situations (1) and (2), $\frac{\partial Q_\lambda(\widehat{\Lambda})}{\partial \beta_j(U_t)} \ge O_p(n^2) + O_p(n^2h^2) + (nh)^{\frac{1}{2}}O_p((nh)^{-\frac{1}{2}}(\nu_n + b_n)) \to \infty$ for $(nh)^{-\frac{1}{2}}\nu_n \to 0$, $(nh)^{-\frac{1}{2}}b_n \to \infty$.

For Situation (3), $\frac{\partial Q_\lambda(\widehat{\Lambda})}{\partial \beta_j(U_t)} \ge O_p(n^2) + O_p(n^2h^2) + (nh)^{\frac{1}{2}}O_p((nh)^{-\frac{1}{2}}(b_n + b_n)) \to \infty$ for $(nh)^{-\frac{1}{2}}b_n \to \infty$.

It clearly shows that, for $(j,k) \in \mathcal{I}^c$, there are no solutions for $\frac{\partial Q_\lambda(\widehat{\Lambda})}{\partial \phi_{jk}(U_t)} = 0$ with $\hat{\phi}_{jk}(U_t) \neq 0$, which is to say that, for any $(j,k) \in \mathcal{I}^c$, $(j,k) \in \widehat{\mathcal{I}^c}$, thus $P(\widehat{\mathcal{I}^c} = \mathcal{I}^c) \to 1$ is proved; naturally, we get that $P(\widehat{\mathcal{I}} = \mathcal{I}) \to 1$. □

Proof of Theorem 2. By Theorem 1, we know immediately that $\widehat{\boldsymbol{\alpha}}_{\lambda,\mathcal{S}^c}(U_t) = 0$ with probability tending to 1, thus we know $\widehat{\boldsymbol{\alpha}}_{\lambda,\mathcal{S}}(U_t)$ must be the solution of

$$-\frac{1}{n}\sum_{t=1}^{n}\sum_{i=1}^{n} W_{i\mathcal{S}}(Y_i - W_{i\mathcal{S}}^\tau\widehat{\boldsymbol{\alpha}}_{\lambda,\mathcal{S}}(U_t))K_h(U_t - U_i) + \frac{1}{n}\sum_{j=1}^{d_0}\lambda_j^1\gamma_j(\mathbf{1}_j\mathbf{1}_j^\tau + \sum_{k:j<k}\mathbf{1}_{jk}\mathbf{1}_{jk}^\tau)\widehat{\boldsymbol{\alpha}}_{\lambda,\mathcal{S}}(u_t)$$

$$+\frac{1}{n}\sum_{j=1}^{d_0}\sum_{k:j<k}\lambda_k^1\gamma_k\mathbf{1}_{jk}\mathbf{1}_{jk}^\tau\widehat{\alpha}_{\lambda,\mathcal{S}}(u_t) + \frac{1}{n}\sum_{j=1}^{d_1}\sum_{k:j<k}\lambda_{jk}^2\zeta_{jk}\mathbf{1}_{jk}\mathbf{1}_{jk}^\tau\widehat{\boldsymbol{\alpha}}_{\lambda,\mathcal{S}}(u_t) = 0,$$

which means $\widehat{\boldsymbol{\alpha}}_{\lambda,\mathcal{S}}$ has the closed form

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}_{\lambda,\mathcal{S}}(U_t) = &\Big[\frac{1}{n}\sum_{i=1}^{n} W_{i\mathcal{S}}W_{i\mathcal{S}}^\tau K_h(U_t - U_i)\Big]^{-1}\frac{1}{n}\Big[\sum_{i=1}^{n} W_{i\mathcal{S}}Y_iK_h(U_t - U_i)\\ &+\sum_{j=1}^{d_0}\lambda_j^1\gamma_j(\mathbf{1}_j\mathbf{1}_j^\tau + \sum_{k:j<k}\mathbf{1}_{jk}\mathbf{1}_{jk}^\tau)\widehat{\boldsymbol{\alpha}}_{\lambda,\mathcal{S}}(u_t)\\ &+\sum_{j=1}^{d_0}\sum_{k:j<k}\lambda_k^1\gamma_k\mathbf{1}_{jk}\mathbf{1}_{jk}^\tau\widehat{\alpha}_{\lambda,\mathcal{S}}(u_t) + \sum_{j=1}^{d_1}\sum_{k:j<k}\lambda_{jk}^2\zeta_{jk}\mathbf{1}_{jk}\mathbf{1}_{jk}^\tau\widehat{\boldsymbol{\alpha}}_{\lambda,\mathcal{S}}(u_t)\Big],\end{aligned}$$

where both $\mathbf{1}_j$ and $\mathbf{1}_{jk}$ are $d_0 + \frac{d_1(d_1+1)}{2}$-dimensional unit vectors, with the components either 1 or 0, and it satisfies $\widehat{\boldsymbol{\alpha}}_{\lambda,\mathcal{S}}(U_t)^\tau\mathbf{1}_j = \widehat{\beta}_j(U_t)$, $\widehat{\boldsymbol{\alpha}}_{\lambda,\mathcal{S}}(U_t)^\tau\mathbf{1}_{jk} = \widehat{\phi}_{jk}(U_t)$. The oracle estimator is defined as follows,

$$\widetilde{\boldsymbol{\alpha}}_{\mathcal{S}}(U_t) = \Big[\frac{1}{n}\sum_{i=1}^{n} W_{i\mathcal{S}}W_{i\mathcal{S}}^\tau K_h(U_t - U_i)\Big]^{-1}\Big[\frac{1}{n}\sum_{i=1}^{n} W_{i\mathcal{S}}Y_iK_h(U_t - U_i)\Big].$$

Then, we have

$$\begin{aligned}\|\widehat{\boldsymbol{\alpha}}_{\lambda,\mathcal{S}}(U_t)-\widetilde{\boldsymbol{\alpha}}_{\mathcal{S}}(U_t)\| &= \|\Sigma_{\mathcal{S}}^{-1}(U_t)\times\frac{1}{n}[\sum_{j=1}^{d_0}\lambda_j^1\gamma_j(\mathbf{1}_j\mathbf{1}_j^\tau+\sum_{k:j<k}\mathbf{1}_{jk}\mathbf{1}_{jk}^\tau)\widehat{\boldsymbol{\alpha}}_\lambda(u_t)\\ &+\sum_{j=1}^{d_0}\sum_{k:j<k}\lambda_k^1\gamma_k\mathbf{1}_{jk}\mathbf{1}_{jk}^\tau\widehat{\alpha}_{\lambda,\mathcal{S}}(u_t)+\sum_{j=1}^{d_1}\sum_{k:j<k\le d_1}\lambda_{jk}^2\zeta_{jk}\mathbf{1}_{jk}\mathbf{1}_{jk}^\tau\widehat{\boldsymbol{\alpha}}_\lambda(u_t)]\|\\ &\le \lambda_{\mathcal{S},max}\cdot a_n\times\frac{1}{n}\Big[2\sum_{j=1}^{d_0}\frac{\sqrt{\widehat{\boldsymbol{\beta}}_j^2(U_t)+\sum_{k:j<k}\widehat{\boldsymbol{\phi}}_{jk}^2(U_t)}}{\sqrt{\|\widehat{\boldsymbol{\beta}}_j\|^2+\sum_{k:j<k}\|\widehat{\boldsymbol{\phi}}_{jk}\|^2}}+\sum_{j=1}^{d_1}\sum_{k:j<k\le d_1}\frac{|\widehat{\boldsymbol{\phi}}_{jk}(U_t)|}{\|\widehat{\boldsymbol{\phi}}_{jk}\|}\Big]\\ &\le \lambda_{\mathcal{S},max}\cdot a_n\times\frac{1}{n}(2d_0+\frac{d_1(d_1+1)}{2})\\ &= (d_0+\frac{d_1(d_1+1)}{2})\lambda_{\mathcal{S},max}n^{-\frac{3}{5}}(nh)^{-\frac{1}{2}}a_n,\end{aligned}$$

where $\lambda_{\mathcal{S},max}$ is the maximum eigenvalue of $\Sigma_{\mathcal{S}}^{-1}(U_t)$, and, referring to $(nh)^{-\frac{1}{2}}a_n\to 0$, we know that $max\|\widehat{\boldsymbol{\alpha}}_{\lambda,\mathcal{S}}(U_t)-\widetilde{\boldsymbol{\alpha}}_{\mathcal{S}}(U_t)\|\le(2d_0+\frac{d_1(d_1+1)}{2})\lambda_{\mathcal{S},max}n^{-\frac{3}{5}}(nh)^{-\frac{1}{2}}a_n\to o_p(n^{-\frac{3}{5}})$, and Theorem 2 is proved. □

Author Contributions: Methodology, F.L. and S.F.; Software, F.L.; Writing—original draft, Y.L.; Writing—review—editing, S.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China (Grant Nos.U1404104 and 11501522), the National Statistical Science Research Project of China (Grant No. 2019LY18), the Foundation of Henan Educational Committee (Grant No. 21A910004), and the Training Fund for Basic Research Program of Zhengzhou University (Grant No. 32211591).

Institutional Review Board Statement: not applicable.

Informed Consent Statement: not applicable.

Data Availability Statement: not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hastie, T.; Tibshirani, R. Varying-coefficient models. *J. R. Stat. Soc. Ser. B* **1993**, *55*, 757–779. [CrossRef]
2. Fan, J.; Zhang, W. Statistical estimation in varying coefficient models. *Ann. Stat.* **1999**, *27*, 1491–1518.
3. Fan, J.; Huang, T. Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli* **2005**, *11*, 1031–1057. [CrossRef]
4. Fan, J.; Zhang, W. Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scand. J. Stat.* **2000**, *27*, 715–731. [CrossRef]
5. Zhu, N.H. Two-stage local Walsh average estimation of generalized varying coefficient models. *Acta Math. Appl. Sin. Engl. Ser.* **2015**,*31*, 623–642. [CrossRef]
6. Li, Z.H.; Liu, J.S.; Wu, X.L. Variable bandwidth and one step local M-estimation of varying coefficient models. *Appl. Math. A J. Chin. Univ.* **2009**, *4*, 379–390.
7. Wang, H.; Xia, Y. Shrinkage estimation of the varying coefficient model. *J. Am. Stat. Assoc.* **2009**, *104*, 747–757. [CrossRef]
8. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288. [CrossRef]
9. Zhao, P.; Xue, L. Variable selection for semiparametric varying coefficient partially linear models. *Stat. Probab. Lett.* **2009**, *79*, 2148–2157. [CrossRef]
10. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [CrossRef]
11. Tang, Y.; Wang, H.J.; Zhu, Z.; Song, X. A unified variable selection approach for varying coefficient models. *Stat. Sin.* **2012**, *22*, 601–628. [CrossRef]
12. Li, D.; Ke, Y.; Zhang, W. Model selection and structure specification in ultra-high dimensional generalised semi-varying coefficient models. *Ann. Stat.* **2015**, *43*, 2676–2705. [CrossRef]
13. He, K.; Lian, H.; Ma, S.; Huang, J. Dimensionality reduction and variable selection in multivariate varying-coefficient models with a large number covariates. *J. Am. Stat. Assoc.* **2018**, *113*, 746–754. [CrossRef]

14. Hall, P.; Xue, J.H. On selecting interacting features from high-dimensional data. *Comput. Stat. Data Anal.* **2014**, *71*, 694–708. [CrossRef]
15. Niu, Y.S.; Hao, N.; Zhang, H.H. Interaction screening by partial correlation. *Stat. Interface* **2018**, *11*, 317–325. [CrossRef]
16. Kong, Y.; Li, D.; Fan, Y.; Lv, J. Interaction pursuit in high-dimensional multi-response regression via distance correlation. *Ann. Stat.* **2017**, *45*, 897–922. [CrossRef]
17. Radchenko, P.; James, G. Variable selection using adaptive nonlinear interaction structure in high dimensions. *J. Am. Stat. Assoc.* **2010**, *105*, 1541–1553. [CrossRef]
18. Choi, N.; Li, W.; Zhu, J. Variable selection with strong heredity constraint and its oracle property. *J. Am. Stat. Assoc.* **2010**, *105*, 354–364. [CrossRef]
19. Bien, J.; Taylor, J.; Tibshirani, R. A lasso for hierarchical interactions. *Ann. Stat.* **2013**, *41*, 1111–1141. [CrossRef]
20. Zhao, P.; Rocha, G.; Yu, B. The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Stat.* **2009**, *37*, 3468–3497. [CrossRef]
21. Lim, M.; Hastie, T. Learning interactions via hierarchical group lasso regularization. *J. Comput. Graph. Stat.* **2015**, *24*, 627–654. [CrossRef] [PubMed]
22. Nelder, J.A. The statistics of linear models: Back to basics. *Stat. Comput.* **1994**, *4*, 221–234. [CrossRef]
23. Hamada, M.; Wu, C.J. Analysis of designed experiments with complex aliasing. *J. Qual. Technol.* **1992**,*24*, 130–137. [CrossRef]
24. Haris, A.; Witten, D.; Simon, N. Convex modeling of interactions with strong heredity. *J. Comput. Graph. Stat.* **2016**, *25*, 981–1004. [CrossRef] [PubMed]
25. Ning, H.; Zhang, H.H. A note on high-dimensional linear regression with interactions. *Am. Stat.* **2017**, *71*, 291–297.
26. Ning, H.; Yang, F.; Zhang, H.H. Model selection for high dimensional quadratic regression via regularization. *J. Am. Stat. Assoc.* **2018**, *113*, 615–625.
27. She, Y.; Wang, Z.; Jiang, H. Group regularized estimation under structural hierarchy. *J. Am. Stat. Assoc.* **2018**, *113*, 445–454. [CrossRef]
28. Yuan, M.; Lin, Y. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B* **2006**, *68*, 49–67. [CrossRef]
29. Hu, T.; Xia, Y. Adaptive semi-varying coefficient model selection. *Stat. Sin.* **2012**, *22*, 575–599. [CrossRef]
30. Hunter, D.; Li, R. Variable selection using MM algorithms. *Ann. Stat.* **2005**, *33*, 1617–1642. [CrossRef]
31. Craven, P.; Wahba, G. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **1979**, *31*, 377–403. [CrossRef]
32. Zou, H.; Li, R. One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Stat.* **2008**, *36*, 1509–1533. [CrossRef] [PubMed]
33. R Core Team. *R: A Language and Environment for Statistical Computating*; R Foundation for Statistical Computing: Vienna, Austria, 2020. Available online: https://www.R-project.org/ (accessed on 16 March 2020).
34. Venables, W.; Ripley, B. *Modern Applied Statistics with S*, 4th ed.; Springer: New York, NY, USA, 2020; ISBN 0-387-95457-0.
35. Harrison D. and Rubinfeld D. Hedonic housing prices and the demand for clean air. *J. Environ. Econ. Manag.* **1978**, *5*, 81–102. [CrossRef]