Electronic Journal of Statistics
Vol. 17 (2023) 823–857
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2117

Variable selection for single-index varying-coefficients models with applications to synergistic G × E interactions[*]

Shunjie Guan[†1] **and Mingtao Zhao**[†2]

[1] *Department of Statistics and Probability, Michigan State University, East Lansing, MI, 48824, USA*

[2] *School of Statistics and Applied Mathematics, Anhui University of Finance & Economics, Bengbu, Anhui, 233000, China*
e-mail: shunjie.guan@gmail.com; mingtao.zhao@outlook.com

Yuehua Cui[‡1]

[1] *Department of Statistics and Probability, Michigan State University, East Lansing, MI, 48824, USA*
e-mail: cuiy@msu.edu

Abstract: Epidemiological evidence suggests that simultaneous exposures to multiple environmental risk factors (Es) can increase disease risk larger than the additive effect of individual exposure acting alone. The interaction between a gene and multiple Es on a disease risk is termed as synergistic gene-environment interactions (synG × E). Single-index varying-coefficients models (SIVCM) have been a promising tool to model synergistic G × E effect and to understand how multiple Es jointly influence genetic risks on a disease outcome. In this work, we proposed a unified variable selection approach for SIVCM to estimate different effects of gene variables: varying, non-zero constant and zero effects which respectively correspond to nonlinear synG × E, no synG × E and no genetic effect. For multiple environmental exposure variables, we also estimated and selected important environmental variables that contribute to the synergistic interaction effect. We theoretically evaluated the oracle property of the proposed variable selection approach. Extensive simulation studies were conducted to evaluate the finite sample performance of the method, considering both continuous and discrete gene variables. Application to a real dataset further demonstrated the utility of the method. Our method has broad applications in areas where the purpose is to identify synergistic interaction effect.

MSC2020 subject classifications: Primary 62J99, 60K35; secondary 62P10.

[*] This work was supported in part by the National Institutes of Health [R21HG010073 to Y.C.], the University Social Science Research Project of Anhui Province (SK2020A0051 to M.Z.) and the Social Science Foundation of Ministry of Education of China [21YJAZH081 and 19YJCZH250 to M.Z.]. The funding agencies had no role in study design, data collection, analysis and interpretation.

[†] S. Guan is presently at Pfizer Inc. The first two authors contributed equally to the work.

[‡] Corresponding author.

Keywords and phrases: Variable selection, single-index varying-coefficients model, synergistic G × E interaction, oracle property.

Received April 2022.

Contents

1. Introduction

Genetic factors play fundamental roles in many complex diseases, and their effects are largely influenced by environmental factors. The same genetic factor can have different effects on disease risks under different environmental conditions, leading to the so called gene-environment (G × E) interaction [1]. The identification of G × E interactions has been one of the central foci in genetic studies.

Recently, Ma et al. [2] and Wu et al. [3] proposed a nonparametric method to capture nonlinear G × E interaction effects. Motivated by epidemiological evidence that simultaneously exposure to multiple environmental conditions would give rise to a higher risk than the simple addition of individual exposure acting alone, Liu et al. [4] proposed a partial linear varying multi-index coefficients model to capture the interaction effect between genetic factors and multiple exposures, termed as synergistic G × E (synG × E). The method can test the interaction between a gene and a mixture of environmental variables and further assess if the interaction effect is linear or nonlinear. While the method was proposed under a low dimensional framework, when the number of genetic variables is large, a variable selection method is needed.

Consider the following single-index varying coefficient model (SIVCM),

$$Y = f^T(X^T\beta)G + \epsilon, \tag{1.1}$$

where Y is a continuous response variable that measures certain phenotypic trait of interest; $X \in \mathbb{R}^q$ is a q dimensional environmental exposure variable and also called loading covariates; $G \in \mathbb{R}^{p+1}$ is a $p + 1$ dimensional genetic variable; $f(\cdot) = (f_0(\cdot), f_1(\cdot), \cdots, f_p(\cdot))^T$ is a $(p + 1) \times 1$ vector of unknown functions with $f_k(\cdot)$ $(k = 0, 1, \cdots, p)$ being the kth unknown non-parametric function; $\beta = (\beta_1, \beta_2, \cdots, \beta_q)^T$ is a vector of unknown loading parameters of dimension q. The model error ϵ has mean 0 and finite variance σ^2. Furthermore, according to Theorem 1 in Fan et al. (2003) [5], for the sake of identifiability, we assume $\|\beta\| = 1$, $\beta_1 > 0$, where $\| \cdot \|$ denotes the Euclidean norm operator; and $f_k(\cdot)$ $(k = 0, 1, \cdots, p)$ cannot be the form as $f(u) = \alpha^T u\beta^T u + \gamma^T u + c_0$, where $\alpha, \gamma \in \mathbb{R}^{p+1}, c_0 \in \mathbb{R}$ are constants, and α and β are not parallel to each other.

One of the main advantages of model (1.1) is that it models the effects of G on Y as functions of X without suffering the curse of dimensionality. One can interpret $f_k(X^T\beta)$ as the effect of G on Y, modified by multiple X variables through the index $X^T\beta$. In addition, model (1.1) is very flexible to cover a wide range of models. For instance, if $q = 1$ and $\beta = 1$, then it becomes a varying-coefficient model (VCM); and if $p = 0$ and $G = 1$, then it becomes a standard single-index model (SIM).

Variable selection has been a popular statistical strategy to solve large p small n problems in a regression setup. In the past, researchers often opted for forward/backward selection, as well as information based criteria such as AIC and BIC for variable selection. Recently, variable selection via penalized regression has been gaining more popularity since it features simultaneous estimation and selection of parameters. The idea is to add a penalty function to the loss function or log-likelihood function. Bridge regression [6], least absolute shrinkage and selection operator (LASSO) [7] and its extensions (adaptive-LASSO [8]), smoothly clipped absolute deviation (SCAD) [9] and minimax concave penalty (MCP) [10] are a few examples. To evaluate different penalized functions, Fan and Li [9] proposed three important criteria: sparsity, unbiasedness and continuity. They showed that SCAD penalty possess the oracle property, meaning that penalized regression featuring SCAD works as well as if the correct sub-model was known in advance. Adaptive LASSO [8], SCAD [9] and MCP [10] all possess the oracle property. However, for adaptive LASSO, determining weights for parameters might become problematic when the dimension of a model is higher than sample size. Zhang [10] proved that at a universal penalty level, MCP has high probability of matching the signs of the unknowns, and thus has nearly unbiased selection, without assuming the strong irrepresentable condition required by the LASSO. Therefore, in the current work, we applied MCP penalty function for its oracle property and fast algorithm [10].

As we all know, model (1.1) is a natural extension of SIM. Variable selection methods for SIM have been studied by many existing works. Naik [11] derived a unified model selection approach for SIM by minimizing the expected Kullback-Leibler distance between the true and candidate models. Naik and Tsai [12] developed a residual information criterion to select both the smoothing parameter and explanatory variables using a residual log-likelihood approach. Wang [13] proposed a fully Bayesian variable selection method for SIM. Peng and Huang [14] proposed a nonconcave penalized least squares method to estimate both the parameters and the link function of SIM. Zeng et al. [15] proposed a sim-lasso approach for estimation and variable selection under SIM. Li et al. [16] proposed a nonconcave penalized least squares variable selection method with the B-spline-based single index approximation. Luo and Ghosal [17] proposed a variable selection and estimation technique for high dimensional SIM with unknown monotone smooth link function. Cheng et al. [18] proposed an effective variable selection method for high-dimensional SIM. Zhang et al. [19] studied the estimation and variable selection for a partial linear SIM (PLSIM) when some linear covariates are not observed, which was extended by Li et al. [20] by implementing a bias correction step. Wang and Zhu [21] applied penalized spline to estimate the nonparametric function and SCAD penalty to achieve sparse estimates of regression parameters in both the linear and single-index parts of the model. Yu et al. [22] performed variable selection in linear term and index vector via binary indicators for PLSIVCM. In this work, we treat the single index functions as regression coefficients and do variable selections on the index functions as well as the index loading parameters. Treating the index functions as coefficients has natural interpretations in the current G × E interaction framework as explained later.

Considering the complicated structure of model (1.1), specifically, the nonlinear structure about the unknown non-parametric functions $f_k(\cdot)$ $(k = 0, 1, \cdots, p)$ and the unknown parameter β, we proposed a unified variable selection method for SIVCM and a three stage iterative variable selection strategy. Specifically, our goal is to: (1) classify the non-parametric functions $f_k(\cdot)$ $(k = 0, 1, \cdots, p)$ into three categories: varying, non-zero constant and zero; (2) select zero and non-zero component of loading parameters β; and (3) estimate $f_k(\cdot)$ $(k = 0, 1, \cdots, p)$ and β. Our approach was motivated by the practical need to classify three different mechanisms in G × E interaction. The zero function of $f_k(\cdot)$ $(k = 0, 1, \cdots, p)$ indicates no genetic effect at all; the constant function of $f_k(\cdot)$ $(k = 0, 1, \cdots, p)$ indicates the effect of G_k on Y does not change over $X^T\beta$, hence no G × E effect; while the varying function of $f_k(\cdot)$ $(k = 0, 1, \cdots, p)$ indicates the existence of G×E effect. In addition to the selection of the coefficient functions, we can also select important loading parameters inside each index coefficient function, to further quantify the relative importance of individual exposure variables. If more than one X variable is selected, we can conclude there is synG × E effect. As shown in Liu et al. [4], the model has the advantage to capture the joint interaction of a gene with multiple exposures as a whole. Novel insights about the underlying genetic mechanism can be revealed by the proposed model.

Feng and Xue [23] proposed a variable selection approach based on model (1.1) by applying a group SCAD penalty on B-spline coefficients and loading parameters β. They focused on either zero or non-zero coefficient functions $f_k(\cdot)$ $(k = 0, 1, \cdots, p)$. We are particularly interested in the constant coefficient since it corresponds to no G × E effect and has important practical implications. Tang et al. [24] and Wu et al. [25] proposed a 2-step unified variable selection approach based on an additive varying-coefficient model. Instead of assuming that the regression coefficients in the model must be variable or constant, they classified the non-parametric function into three categories: varying, constant or zero. Their model is a special case of our SIVCM model when the dimension of the X variable is reduced to one. No variable selection approach on SIVCM has been proposed to classify unknown non-parametric functions $f_k(\cdot)$ $(k = 0, 1, \cdots, p)$ into three categories (varying, constant or zero), while selecting non-zero loading parameter β simultaneously. Following the previous work, we used B-spline basis functions to approximate unknown non-parametric functions $f_k(\cdot)$ $(k = 0, 1, \cdots, p)$, then using penalized regression to classify $f_k(\cdot)$ $(k = 0, 1, \cdots, p)$ into varying, constant or zero. Further, we selected non-zero β via first order approximation and penalized regression. The proposed variable selection method does not assume apriori whether $f_k(\cdot)$ $(k = 0, 1, \cdots, p)$ is constant or a non-parametric function and can be regarded as a unified variable selection method for SIVCM. We showed that under some mild regularity conditions, our estimators possess the oracle property, indicating that our penalized estimators work as well as if the correct sub-model is known in advance.

The rest of the paper was organized as follows. Section 2 introduced our proposed variable selection approach, including the iterative estimation approach and how to select various tuning parameters for B-spline approximation and penalized regressions. Method on how to select initial values for β was discussed. In Section 3, we evaluated the theoretical properties of our approach. In Section 4, we performed simulations to evaluate the performance of our method in finite samples, followed by a real data application in Section 5 and a discussion in Section 6.

2. The variable selection method

2.1. Model setup

Consider model (1.1) with data $\{(Y_i, X_i, G_{ik}), i = 1, 2, \cdots n, k = 0, 1, 2, \cdots, p\}$ in the following form,

$$Y_i = f^T(X_i^T\beta)G_i + \epsilon_i, \quad i = 1, 2, \cdots, n, \tag{2.1}$$

where Y_i is a continuous response variable; $X_i = (X_{i1}, X_{i2}, \cdots, X_{iq})^T$ is q-dimensional continuous loading covariates; $X_i^T\beta$ is the so-called index; $G = (G_{ik})_{(p+1)\times n} = (G_1, G_2, \cdots, G_n)$, $G_i = (1, G_{i1}, \cdots, G_{ip})^T$; $G_{\cdot k} = (G_{1k}, \cdots, G_{nk})^T$ is a continuous or discrete vector of length n for $k = 0, 1, 2, \cdots p$. In model (1.1), $f_k(\cdot)$ is the effect of $G_{\cdot k}$ on Y for $k \neq 0$ and $f_0(\cdot)$ is the intercept function which models the marginal effect of X on Y; ϵ_i $(i = 1, 2, \cdots, n)$ are unknown random errors with mean 0 and finite variance σ^2. We further assume that ϵ_i and ϵ_j are independent of each other for $i \neq j$ ($\forall\ 1 \leq i, j \leq n$), $\{\epsilon_i, i = 1, 2, \cdots, n\}$ are independent of $\{(X_i, G_{ik}), i = 1, 2, \cdots, n, k = 0, 1, 2, \cdots, p\}$.

2.2. Estimation method

We approximate the unknown functions $\{f_k(u) : u \in \mathcal{U}\}$ ($k = 0, 1, 2, \cdots, p$) using B-spline basis functions. Here, we assume that $\mathcal{U}$ is a nondegenerate compact interval. Denote $\mathscr{F}$ to be a collection of functions $f(u)$ satisfying (A2) in Appendix. Let K be the number of interior knots and h be the order of the B-spline basis function. By Schumaker (1981, chapter 4) [26], we can normalize the B-spline basis function $\widetilde{B}(u) = (\tilde{B}_1(u), \tilde{B}_2(u), \cdots, \tilde{B}_L(u))^T$ for $\mathscr{F}$, and there exists a linear transformation matrix Π [24], such that

$$\Pi\widetilde{B}(u) = (\mathbf{1}, B_2(u), B_3(u), \cdots, B_L(u))^T = (\mathbf{1}, \bar{B}^T(u))^T \triangleq B(u) \tag{2.2}$$

where $\bar{B}(u) = (B_2(u), B_3(u), \cdots, B_L(u))^T$, $L = K + h$ and each component of $\bar{B}(u)$ and $\widetilde{B}(u)$ is a function of u. Clearly, $B(u)$ is also a basis function for $\mathscr{F}$. In our work, we assume that $f_k(u) \in \mathscr{F}$ for $k = 0, 1, 2, \cdots, p$. Therefore, we can approximate each $f_k(u)$ by

$$f_k(u) \approx B^T(u)\gamma_k = \gamma_{k1} + \bar{B}^T(u)\gamma_{k*}, \quad k = 0, 1, 2, \cdots, p, \tag{2.3}$$

where $\gamma_k = (\gamma_{k1}, \gamma_{k*}^T)^T$ and γ_{k1} corresponds to the constant part of the coefficient function and $\gamma_{k*} = (\gamma_{k2}, \gamma_{k3}, \cdots, \gamma_{kL})^T$ corresponds to the varying part. To fix notation, we take $\gamma = (\gamma_0^T, \gamma_1^T, \cdots, \gamma_p^T)^T$, $W_i(\beta) = I_{p+1} \otimes B(X_i^T\beta) \cdot G_i$, where I_{p+1} is the $(p+1) \times (p+1)$ identity matrix and "$\otimes$" is the Kronecker product operator. With the B-spline approximation same as (2.3), model (2.1) can be rewritten as

$$Y_i \approx W_i^T(\beta)\gamma + \epsilon_i, \qquad i = 1, 2, \cdots, n. \tag{2.4}$$

In matrix notation, we have

$$Y \approx W(\beta)\gamma + \epsilon \tag{2.5}$$

where $\epsilon = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)^T$ and $W(\beta) = (W_1(\beta), W_2(\beta), \cdots, W_n(\beta))^T \in \mathbb{R}^n \times \mathbb{R}^{L(p+1)}$. Thus, the original estimation problem can be transformed to estimate γ and β.

Remark 1. By some simple matrix calculation, we can see that

$$W_i^T(\beta)\gamma = G_i^T\gamma_{*1} + \bar{W}(\beta)_i^T\gamma_*, \qquad i = 1, 2, \cdots, n, \tag{2.6}$$

where $\bar{W}_i^T(\beta) = I_{p+1} \otimes \bar{B}(X_i^T\beta) \cdot G_i$, $\gamma_{*1} = (\gamma_{01}, \gamma_{11}, \cdots, \gamma_{p1})^T$ and $\gamma_* = (\gamma_{0*}^T, \gamma_{1*}^T, \cdots, \gamma_{p*}^T)^T$.

Remark 2. The transformation matrix Π can separate the main genetic and G × E effect from the total effect, which further enables us to assess if there exist genetic main and interaction effects, that is: (1) if $\|\gamma_{k*}\| = (\sum_{l=2}^{L} \gamma_{kl}^2)^{1/2} \neq 0$, then there exists interaction between $G_{\cdot k}$ and multiple X; (2) if $\|\gamma_{k*}\| = 0$ and $|\gamma_{k1}| \neq 0$, then $G_{\cdot k}$ has a constant effect on Y, i.e., no G × E interaction effect; and (3) if further $\|\gamma_{k*}\| = 0$ and $|\gamma_{k1}| = 0$ then $G_{\cdot k}$ has no effect on Y at all.

To select and estimate the parameters γ and β, we apply the penalized regression idea and minimize the following penalized least squares objective function

$$\begin{aligned} Q(\beta, \gamma) = & \sum_{i=1}^{n} \left(Y_i - W_i^T(\beta)\gamma\right)^2 + n \sum_{k=1}^{p} p_{\lambda_{1k}}(\|\gamma_{k*}\|) \\ & + n \sum_{k=1}^{p} p_{\lambda_{2k}}(|\gamma_{k1}|) I(\|\gamma_{k*}\| = 0) + n \sum_{d=2}^{q} p_{\lambda_{3d}}(|\beta_d|), \end{aligned} \tag{2.7}$$

where $p_{\lambda_{1k}}(\cdot), p_{\lambda_{2k}}(\cdot), p_{\lambda_{3d}}(\cdot)$ are penalty functions of the corresponding parameters, and $I(\cdot)$ is an indicator function. In our work, the penalty functions $p_{\lambda_{1k}}(\cdot), p_{\lambda_{2k}}(\cdot), p_{\lambda_{3d}}(\cdot)$ are MCP [10] penalty functions such that $p(x, \lambda) = \lambda \int_0^x (1 - \frac{s}{\tau\lambda})_+ ds$ with regularization parameters $\tau > 0$ and $\lambda > 0$.

Remark 3. (1) From the construction of the penalty function, we penalize γ_{k1} only if $\|\gamma_{k*}\| = 0$. If $\|\gamma_{k*}\| \neq 0$, it implies that the function is varying and no need to penalize the constant part; (2) No penalty is applied to the intercept function $f_0(\cdot)$. There is no practical motivation to penalize the marginal intercept function; and (3) No penalty is applied to the first loading parameter β_1 in β due to the constraint.

We now handle the constraints $\|\beta\| = 1$ and $\beta_1 > 0$ on the q-dimensional single-index parameter β with reparametrization. Denote $\phi = (\phi_2, \phi_3, \cdots, \phi_q)^T = (\beta_2, \beta_3, \cdots, \beta_q)^T$, and we can get

$$\beta = \left(\sqrt{1 - \|\phi\|^2}, \phi^T\right)^T, \quad \|\phi\| < 1.$$

Therefore, $\beta = \beta(\phi)$, and β is infinitely differentiable with respect to ϕ. The Jacobian matrix of β with respect to ϕ is

$$J_\phi = \begin{pmatrix} -(1 - \|\phi\|^2)^{-1/2} \phi^T \\ \mathrm{I}_{q-1} \end{pmatrix}. \tag{2.8}$$

Note that ϕ is one dimension lower than β, and $Q(\beta, \gamma)$ can be rewritten as

$$\begin{aligned} Q(\phi, \gamma) = & \sum_{i=1}^{n} \left(Y_i - W_i^T(\phi)\gamma\right)^2 + n \sum_{k=1}^{p} p_{\lambda_{1k}}(\|\gamma_{k*}\|) \\ & + n \sum_{k=1}^{p} p_{\lambda_{2k}}(|\gamma_{k1}|) I(\|\gamma_{k*}\| = 0) + n \sum_{d=2}^{q} p_{\lambda_{3d}}(|\phi_d|), \end{aligned} \tag{2.9}$$

where $W_i(\phi) = W_i(\beta)$. Then we can get the penalized least squares estimators $\hat{\phi}$, $\hat{\gamma}$ and $\hat{\beta}$ as

$$(\hat{\phi}, \hat{\gamma}) = \arg\min_{\phi,\gamma} Q(\phi, \gamma), \tag{2.10}$$

$$\hat{\beta} = \left(\sqrt{1 - \|\hat{\phi}\|^2}, \hat{\phi}^T \right)^T, \quad \|\hat{\phi}\| \le 1, \tag{2.11}$$

where $\hat{\gamma} = (\hat{\gamma}_0^T, \hat{\gamma}_1^T, \cdots, \hat{\gamma}_p^T)^T$. Therefore, the estimator of $f_k(u)$ can be obtained by

$$\hat{f}_k(u) = B^T(u)\hat{\gamma}_k, \qquad k = 0, 1, 2, \cdots, p. \tag{2.12}$$

2.3. Iterative algorithm

We can see that $\hat{\phi}$ and $\hat{\gamma}$ denoted by (2.10) do not have closed form. Thus, we propose a iterative approach to get the numerical solution of $\hat{\phi}$ and $\hat{\gamma}$. Our modeling purpose is to classify $f_k(\cdot)$ $(k = 0, 1, 2, \cdots, p)$ into three different categories: varying, non-zero constant or zero, denoted by $\mathcal{V}$, $\mathcal{C}$ and $\mathcal{Z}$ respectively. For $\forall k \in \{0, 1, 2, \cdots, p\}$, notations "$k \in \mathcal{V}$", "$k \in \mathcal{C}$" and "$k \in \mathcal{Z}$" mean that the function $f_k(\cdot)$ is varying, non-zero constant and zero respectively. Obviously, $\mathcal{V}, \mathcal{C}$ and $\mathcal{Z}$ are mutually disjoint, and $\mathcal{V} \cup \mathcal{C} \cup \mathcal{Z} = \{0, 1, 2, \cdots, p\}$. Furthermore, $k \notin \mathcal{V}$ means that $f_k(\cdot)$ is non-zero constant or zero, that is, $\{k \notin \mathcal{V}\} = \{k \in \mathcal{C}\} \cup \{k \in \mathcal{Z}\}$. Following Feng and Xue [23] and Tang et al. [24], we propose a stepwise iterative approach to obtain our penalized estimator.

Step 0: Set initial values $\hat{\beta}^{(0)}$ and $\hat{\gamma}^{(0)}$ to start the iteration. Setting $f_k(\cdot)$ $(k = 0, 1, 2, \cdots, p)$ as identity functions, we can get a simple linear additive model as

$$Y_i = X_i^T\beta + X_i^T\beta \cdot G_{i1} + X_i^T\beta \cdot G_{i2} + \cdots + X_i^T\beta \cdot G_{ip} + \epsilon_i, \ \ i = 1, 2, \cdots, n. \tag{2.13}$$

Therefore, we can set an initial estimator $\tilde{\beta} = (\tilde{\beta}_1, \tilde{\phi}^T)^T$ as

$$\tilde{\beta} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T Y, \tag{2.14}$$

where $\tilde{\phi} = (\tilde{\beta}_2, \tilde{\beta}_3, \cdots, \tilde{\beta}_q)^T$, $\tilde{X} = (X_1, \tilde{G}_2 X_2, \cdots, \tilde{G}_n X_n)^T$, $\tilde{G}_i = \sum_{k=1}^p G_{ik}$. Considering the constraints for β such that $\|\beta\| = 1$ and $\beta_1 > 0$, the initial estimator $\hat{\beta}^{(0)}$ can be chosen from (2.13) and (2.14) as

$$\hat{\beta}^{(0)} = \frac{\tilde{\beta}}{\|\tilde{\beta}\|} \cdot \mathrm{sgn}(\tilde{\beta}_1) \tag{2.15}$$

Then the initial estimator of $\hat{\gamma}^{(0)}$ can be obtained by

$$\hat{\gamma}^{(0)} = \left(\sum_{i=1}^n W_i(\hat{\beta}^{(0)}) W_i^T(\hat{\beta}^{(0)}) \right)^{-1} \sum_{i=1}^n W_i^T(\hat{\beta}^{(0)}) Y_i. \tag{2.16}$$

Step 1: In this step, we classify $f_k(\cdot)$ $(k = 0, 1, 2, \cdots, p)$ into varying $(k \in \mathcal{V})$ and non-varying $(k \in \mathcal{C} \cup \mathcal{Z})$. For a given initial value of β, denoted by $\hat{\beta}^{(0)}$ from

(2.15), we can obtain our 1st step estimation $\hat{\gamma}^{(1)} = ((\hat{\gamma}_0^{(1)})^T, (\hat{\gamma}_1^{(1)})^T, \cdots, (\hat{\gamma}_p^{(1)})^T)^T$ by following a group penalized regression

$$\hat{\gamma}^{(1)} = \min_{\gamma} Q_1(\gamma|\Lambda_1, \hat{\beta}^{(0)}), \tag{2.17}$$

where the kth coefficient $\hat{\gamma}_k^{(1)} = (\hat{\gamma}_{k1}^{(1)}, (\hat{\gamma}_{k*}^{(1)})^T)^T$ $(k = 0, 1, 2, \cdots, p)$, $\Lambda_1 = \{\lambda_{11}, \lambda_{12}, \cdots, \lambda_{1p}\}$ and

$$Q_1(\gamma|\Lambda_1, \hat{\beta}^{(0)}) = \sum_{i=1}^{n} \left(Y_i - W_i^T(\hat{\beta}^{(0)})\gamma\right)^2 + n\sum_{k=1}^{p} p_{\lambda_{1k}}(\|\gamma_{k*}\|). \tag{2.18}$$

Note that $\|\gamma_{k*}\| > 0$ and $\|\gamma_{k*}\| = 0$ respectively imply that $f_k(\cdot)$ is varying $(k \in \mathcal{V})$ and non-varying $(k \in \mathcal{C} \cup \mathcal{Z})$. Therefore, instead of penalizing each coordinate of $\gamma_{k*} = (\gamma_{k2}, \cdots, \gamma_{kL})^T$ $(k = 0, 1, 2, \cdots, p)$ separately, we penalized $\|\gamma_{k*}\|$ $(k = 0, 1, 2, \cdots, p)$ for the reason that we want to assess the presence of the joint varying effect of X and $G_{\cdot k}$ on Y. In particular, from (2.18), no penalty is applied to γ_{0*}, which means that the intercept function $f_0(\cdot)$ is treated as being varying in our work. Step 1 classifies $f_k(\cdot)$ $(k = 0, 1, 2, \cdots, p)$ into two categories, i.e., varying and non-varying. However, $\hat{\gamma}^{(1)}$ does not have a closed form. We can only get numerical solutions through an iterative algorithm. The detailed iterative algorithm for this step can be found in A.1 of the Appendix, with the initial iterative value of γ denoted by $\hat{\gamma}^{(0)}$ in (2.16).

Step 2: After Step 1, we would like to further select variables with constant effects and classify the non-varying functions $f_k(\cdot)$ $(k \in \mathcal{C} \cup \mathcal{Z})$ into non-zero constants $(k \in \mathcal{C})$ and zeros $(k \in \mathcal{Z})$ in this step, i.e., estimate and select γ_{k1} given $\hat{\gamma}_{k*}^{(1)} = 0$ for $k \in \mathcal{C} \cup \mathcal{Z}$. In order to do that, we penalize γ_{k1} only when $\|\hat{\gamma}_{k*}^{(1)}\| = 0$, i.e. $k \in \mathcal{C} \cup \mathcal{Z}$, and no penalty is applied to γ_{01}.

We obtain estimator $\hat{\gamma}^{(2)} = ((\hat{\gamma}_0^{(2)})^T, (\hat{\gamma}_1^{(2)})^T, \cdots, (\hat{\gamma}_p^{(2)})^T)^T$ via penalized regression

$$\hat{\gamma}^{(2)} = \min_{\gamma} Q_2(\gamma|\Lambda_2, \hat{\beta}^{(0)}, \hat{\gamma}^{(1)}), \tag{2.19}$$

where $(\hat{\gamma}_k^{(2)})_{k\in\mathcal{V}} = (\hat{\gamma}_{k1}^{(2)}, (\hat{\gamma}_{k*}^{(2)})^T)^T, (\hat{\gamma}_k^{(2)})_{k\in\mathcal{C}} = (\hat{\gamma}_{k1}^{(2)}, \mathbf{0}^T)^T, (\hat{\gamma}_k^{(2)})_{k\in\mathcal{Z}} = \mathbf{0}, k = 0, 1, 2, \cdots, p$, $\Lambda_2 = \{\lambda_{21}, \lambda_{22}, \cdots, \lambda_{2p}\}$ and

$$Q_2(\gamma|\Lambda_2, \hat{\beta}^{(0)}, \hat{\gamma}^{(1)}) = \sum_{i=1}^{n} \left(Y_i - W_i^T(\hat{\beta}^{(0)})\hat{\gamma}^{(1)}\right)^2 + n\sum_{k=1}^{p} p_{\lambda_{2k}}(|\gamma_{k1}^{(1)}|)I(\|\hat{\gamma}_{k*}^{(1)}\| = 0). \tag{2.20}$$

The detailed iterative algorithm for Step 2 can be found in A.1 of the Appendix. After Step 1 and 2, we can obtain the estimator of the B-spline coefficients γ denoted as $\hat{\gamma}^{(2)}$ and classify $f_k(\cdot)$ $(k = 0, 1, \cdots, p)$ into $\mathcal{V}$, $\mathcal{C}$ or $\mathcal{Z}$. Then the next step is to estimate and select loading parameter β given $\hat{\gamma}^{(2)}$.

Step 3: We obtain $\hat{\beta}$ via the penalized regression

$$\hat{\beta} = \min_{\|\beta\|=1} Q_3(\beta|\Lambda_3, \hat{\gamma}^{(2)}), \tag{2.21}$$

where $\Lambda_3 = \{\lambda_{32}, \cdots, \lambda_{3q}\}$ and

$$Q_3(\beta|\Lambda_3, \hat{\gamma}^{(2)}) = \sum_{i=1}^{n} \Big(Y_i - W_i^T(\beta)\hat{\gamma}^{(2)}\Big)^2 + n\sum_{d=2}^{q} p_{\lambda_{3d}}(|\phi_d|). \tag{2.22}$$

The detailed iterative algorithm for $\hat{\beta}$ can be found in A.1 of the Appendix. We then replace $\hat{\beta}^{(0)}$ by $\hat{\beta}$ and iterate between Step 1 and Step 3 until convergence.

2.4. Selection of tuning parameters

We use the Bayesian Information Criterion (BIC) to select the tuning parameters τ, λ_{1k}, λ_{2k} and λ_{3d} in the penalty functions [27]. Since there are too many tuning parameters in our penalty functions, and the minimization problem for the BIC method over a high-dimensional space is computationally intensive and difficult to track, similar to Feng and Xue [23], we take the adaptive tuning parameters λ_{1k}, λ_{2k} and λ_{3d} as

$$\lambda_{1k} = \frac{\lambda_1}{\|\hat{\gamma}_k^{un}\|},\ \lambda_{2k} = \frac{\lambda_2}{\|\hat{\gamma}_{k1}^{(1)}\|},\ \lambda_{3d} = \frac{\lambda_3}{|\hat{\beta}_d^{un}|},$$

where $\hat{\gamma}_k^{un}$ $(k = 0, 1, 2, \cdots, p)$ and $\hat{\beta}_d^{un}$ $(d = 2, \cdots, q)$ are the unpenalized estimators of γ_k $(k = 0, 1, 2, \cdots, p)$ and $\beta_d^{(0)}$ $(d = 2, \cdots, q)$. $\hat{\gamma}_k^{(1)} = (\gamma_{k1}^{(1)}, (\hat{\gamma}_{k*}^{(1)})^T)^T$ is denoted by (2.17) and satisfies $\|\hat{\gamma}_{k*}^{(1)}\| = 0$. Therefore, we transform the selection of tuning parameters λ_{1k}, λ_{2k} and λ_{3d} into a one-dimensional grid searching problem. We just need to chose optimal λ_1, λ_2 and λ_3 in the three step algorithm.

In Step 1, we take optimal λ_1 as the minimizer of

$$\text{BIC}_1(\lambda_1) = \log\left(\sum_{i=1}^{n}\Big(Y_i - W_i^T(\hat{\beta}^{(0)})\hat{\gamma}^{(\lambda_1)}\Big)^2\right) + \frac{\log(n)}{n}\cdot df_{\lambda_1}, \tag{2.23}$$

where $\hat{\gamma}^{(\lambda_1)} = \arg\min_\gamma Q_1(\gamma|\Lambda_1, \hat{\beta}^{(0)})$ is defined by (2.17) for a given λ_1, $\hat{\beta}^{(0)}$ is denoted as (2.15), df_{λ_1} is defined as the total number of non-zero coefficients of $\{\|\hat{\gamma}_k^{(\lambda_1)}\|, k = 0, 1, 2, \cdots, p\}$ for a given λ_1.

In Step 2, the optimal λ_2 is the minimizer of

$$\text{BIC}_2(\lambda_2) = \log\left(\sum_{i=1}^{n}\Big(Y_i - W_i^T(\hat{\beta}^{(0)})\hat{\gamma}^{(\lambda_2)}\Big)^2\right) + \frac{\log(n)}{n}\cdot df_{\lambda_2}, \tag{2.24}$$

where $\hat{\gamma}^{(\lambda_2)} = \arg\min_\gamma Q_2(\gamma|\Lambda_2, \hat{\beta}^{(0)}, \hat{\gamma}^{(1)})$ is defined by (2.19) for a given λ_2, df_{λ_2} is defined as the total number of non zero coefficients of $\{\|\hat{\gamma}_k^{(\lambda_2)}\|, k = 0, 1, 2, \cdots, p\}$ for a given λ_2.

In Step 3, we take optimal λ_3 as the minimizer of

$$\text{BIC}(\lambda_3) = \log\left(\sum_{i=1}^{n}\left(Y_i - W_i^T(\hat{\beta}^{(\lambda_3)})\hat{\gamma}^{(\lambda_2)}\right)^2\right) + \frac{\log n}{n}\cdot df_{\lambda_3}, \tag{2.25}$$

where $\hat{\gamma}^{(\lambda_2)} = \arg\min_\gamma Q_2(\gamma|\Lambda_2, \hat{\beta}^{(0)}, \hat{\gamma}^{(1)})$, and $\hat{\beta}^{(\lambda_3)} = \arg\min_\beta Q_3(\beta|\Lambda_3, \hat{\gamma}^{(2)})$ is defined by (2.21) for a given λ_3, and df_{λ_3} is defined as the total number of non-zero β_d $(d = 1, 2, \cdots, q)$ for a given λ_3. We search the optimal value of $\lambda_1, \lambda_2, \lambda_3$ over a grid of 100 exponentially decreasing values with the minimum being 1E-3, and the maximum of $\lambda_1, \lambda_2, \lambda_3$ is set to be the minimum value such that all of the penalized estimators are zeros.

2.5. Selection of the order h and the number of interior knots K

Since h is the order of the B-spline basis function, higher degree corresponds to more complicated interactions and is less interpretable in practice. In practice, there is no need to set the order too high, reasonable allocation of knots can make low-order splines achieve better fitting effect. Tang et al. [24] suggested using lower degree splines such as linear, quadratic or cubic splines corresponding to $h = 2, 3$ and 4, respectively. Hence, we search optimal order h_{opt} over the set $\mathcal{H} = \{2, 3, 4\}$. Futhermore, $K = O_p(n^{\frac{1}{2r+1}})$ is a necessary assumption for oracle properties of the proposed variable selection approach, where n is the sample size and r is defined in condition (A2) in Appendix. According to He et al. [28], in our work, the range of the interior knots is taken to be $\mathcal{K} = \left[\max(\lfloor 0.5\cdot n^{\frac{1}{2r+1}}\rfloor, 1), \lfloor 1.5\cdot n^{\frac{1}{2r+1}}\rfloor\right]$, where $\lfloor x\rfloor$ denotes the integer part of x.

In theory, we can select the optimal order h_{opt} and the number of interior knots K_{opt} for each nonparametric function $f_k(\cdot)$. However, this is practically infeasible due to the large searching space and the computational cost. We assume that all the nonparametric functions share common h and K. Thus, (K_{opt}, h_{opt}) can be achieved via a two-dimensional grid search for $(K_{opt}, h_{opt}) \in \mathcal{K}\times\mathcal{H}$ focusing only on the intercept function by the following criterion

$$(K_{opt}, h_{opt}) = \arg\min_{K,h}\left\{\log\left(\sum_{i=1}^{n}\left(Y_i - W_i^T\hat{\gamma}\right)^2\right) + \frac{\log(n)}{n}(K+h)\right\}, \tag{2.26}$$

where $\hat{\gamma} = (\hat{\gamma}_0^T, 0^T, \cdots, 0^T)^T$.

3. Theoretical properties

We first fix some notations. Let $f_0(\cdot) = (f_{00}(\cdot), f_{10}(\cdot), \cdots, f_{p0}(\cdot))^T$ and $\beta_0 = (\beta_{10}, \beta_{20}, \cdots, \beta_{q0})^T$ be the true value of $f(\cdot)$ and β respectively, and denote $\gamma_0 = (\gamma_{00}^T, \gamma_{10}^T, \cdots, \gamma_{p0}^T)^T$ be the true value of the B-spline coefficient γ, where $\gamma_{k0} = (\gamma_{k1}^0, \gamma_{k*}^{0T})^T$, $\gamma_{k*}^0 = (\gamma_{k2}^0, \gamma_{k3}^0, \cdots, \gamma_{kL}^0)^T$. Without loss of generality, we assume $\beta_{d0} \neq 0$ for $d = 1, \cdots s$, $\beta_{d0} = 0$ for $d = s+1, \cdots q$; $f_{k0}(\cdot)$ is varying for $k = 0, 1, \cdots, v$, $f_{k0}(\cdot)$ is non-zero constant for $k = v+1, \cdots, c$ and $f_{k0}(\cdot)$ is zero for $k = c+1, \cdots, p$. Clearly, we can see that $\mathcal{V} = \{0, 1, 2, \cdots, v\}$ and $\mathcal{C} = \{v+1, v+2, \ldots, c\}$, $\mathcal{Z} = \{c+1, c+2, \cdots, p\}$. The following theorem gives the consistency of the penalized least square estimators.

Theorem 3.1. *Suppose the regulatory conditions (A1) – (A8) in Appendix hold and the number of interior knots* $K = O_p(n^{1/(2r+1)})$. *Then*

(i) $\|\hat{\beta} - \beta_0\| = O_p(n^{-r/(2r+1)} + a_n)$*;*

(ii) $\|\hat{f}_k(\cdot) - f_{k0}(\cdot)\| = O_p(n^{-r/(2r+1)} + a_n)$, $k = 0, 1, 2, \cdots, p$*;*

where $a_n = \max_{k,l}\{p'_{\lambda_{1k}}(\|\gamma^0_{k*}\|), p'_{\lambda_{2k}}(|\gamma^0_{k1}|), p'_{\lambda_{3l}}(|\beta_{d0}|), \gamma^0_{k*} \neq 0, \gamma^0_{k1} \neq 0, \beta^0_d \neq 0, k = 0, 1, 2, \cdots, p, d = 1, 2, \cdots, q\}$.

Furthermore, under some regularity conditions, we can demonstrate that the above consistent estimators possess the following sparsity properties.

Theorem 3.2. *Suppose the regularity conditions (A1) – (A8) in Appendix hold and the number of interior knots* $K = O_p(n^{1/(2r+1)})$. *Let* $\lambda_{\max} = \max\{\lambda_{1k}, \lambda_{2k}, \lambda_{3d}, k = 0, 1, 2, \cdots, p; d = 2, \cdots, q\}$ *and* $\lambda_{\min} = \min\{\lambda_{1k}, \lambda_{2k}, \lambda_{3d}, k = 0, 1, 2, \cdots, p; d = 2, \cdots, q\}$. *Suppose* $\lambda_{\max} \to 0$ *and* $n^{r/(2r+1)}\lambda_{\min} \to \infty$ *as* $n \to \infty$. *Then with probability approaching to* 1, $\hat{\beta}$ *and* $\hat{f}_k(\cdot)$ *satisfy*

(i) $\hat{\beta}_d = 0$ *for* $d = s+1, \cdots, q$*;*

(ii) $\hat{f}_k(\cdot) = c_k$ *for* $k = v+1, \cdots, c$, *where* c_k *is some non-zero constant;*

(iii) $\hat{f}_k(\cdot) = 0$ *for* $k = c+1, \cdots, p$*;*

Next, we show that the asymptotic normality of the non-zero coefficients β and the spline coefficients γ. Obviously, if $\mathcal{C} \neq \varnothing$, model (2.1) degenerates into a partial linear single-index varying-coefficient model. However, the true model is unknown in advance. Without loss of generality, we treat all of functions $f_k(\cdot)$ $(k = 0, 1, 2, \cdots, p)$ as being varying in advance, then identify whether each $f_k(\cdot)$ is varying, non-zero constant or zero. Denote

$$\beta^* = (\beta_1, \beta_2, \cdots, \beta_s)^T, \quad f^*(\cdot) = (f^{*T}_{(\mathcal{V})}(\cdot), f^{*T}_{(\mathcal{C})}(\cdot))^T,$$

$$f^*_{(\mathcal{V})}(\cdot) = (f_0(\cdot), f_1(\cdot), \cdots, f_v(\cdot))^T, \quad f^*_{(\mathcal{C})}(\cdot) = (f_{v+1}(\cdot), f_{v+2}(\cdot), \cdots, f_c(\cdot))^T,$$

and the corresponding covariates are denoted by $X^*, G^*_i = (G^{*T}_{(\mathcal{V})i}, G^{*T}_{(\mathcal{C})i})^T$ $(i = 1, 2, \cdots, n)$. Let $\beta^*_0 = (\beta_{10}, \beta_{20}, \cdots, \beta_{s0})^T$ and $f^*_0(\cdot) = (f^{*T}_{(\mathcal{V})0}(\cdot), f^{*T}_{(\mathcal{C})0}(\cdot))^T$ to be the true values of β^* and $f^*(\cdot)$, where $f^*_{(\mathcal{V})0}(\cdot) = (f_{00}(\cdot), f_{10}(\cdot), \cdots, f_{v0}(\cdot))^T$, $f^*_{(\mathcal{C})0}(\cdot) = (f_{(v+1)0}(\cdot), f_{(v+2)0}(\cdot), \cdots, f_{c0}(\cdot))^T$. Obviously, $f_{k0}(u)$ $(k = v+1, v+2, \cdots, c)$ are non-zero constants for $\forall u \in \mathcal{U}$. Similarly, we have ϕ^*, $W^*_i = (W^T_{(\mathcal{V})i}, W^T_{(\mathcal{C})i})^T$ and $\gamma^* = (\gamma^{*T}_{(\mathcal{V})}, \gamma^{*T}_{(\mathcal{C})})^T$, where

$$W_{(\mathcal{V})i} = I_{v+1} \otimes B(X^{*T}_i\beta^*) \cdot G^*_{(\mathcal{V})i}, \quad W_{(\mathcal{C})i} = I_{c-v} \otimes B(X^{*T}_i\beta^*) \cdot G^*_{(\mathcal{C})i},$$

$$\gamma^*_{(\mathcal{V})} = (\gamma^T_0, \gamma^T_1, \cdots, \gamma^T_v)^T, \qquad \gamma^*_{(\mathcal{C})} = (\gamma^T_{v+1}, \gamma^T_{v+2}, \cdots, \gamma^T_c)^T.$$

Denote $\gamma^*_{(\mathcal{V})0} = (\gamma^T_{00}, \gamma^T_{10}, \cdots, \gamma^T_{v0})^T$, $\gamma^*_{(\mathcal{C})0} = (\gamma^T_{(v+1)0}, \gamma^T_{(v+2)0}, \cdots, \gamma^T_{c0})^T$ be the estimators of the B-spline approximation to $f^*_{(\mathcal{V})0}(\cdot)$ and $f^*_{(\mathcal{C})0}(\cdot)$, respectively.

We can see that $\gamma_{k0} = (\gamma_{k1}^0, 0, 0, \cdots, 0)^T$ for $k = v+1, v+2, \cdots, c$ and γ_{k1}^0 $(k = v+1, v+2, \cdots, c)$ are non-zero constants. Furthermore, we have $f_{(\mathcal{C})0}^* = \gamma_{(\mathcal{C})1}^{*0T} = (\gamma_{(v+1)1}^0, \gamma_{(v+2)1}^0, \cdots, \gamma_{c1}^0)^T$. Denote $\vartheta^* = (\gamma_{(\mathcal{C})1}^{*T}, \phi^{*T})^T$. The corresponding estimator and true value of ϑ^* are denoted by $\hat{\vartheta}^* = (\hat{\gamma}_{(\mathcal{C})1}^{*T}, \hat{\phi}^{*T})^T$ and $\vartheta_0^* = (\gamma_{(\mathcal{C})1}^{*0T}, \phi_0^{*T})^T$, respectively. In addition, let

$$\Sigma_1 = \mathrm{E}(G_{(\mathcal{C})}^* G_{(\mathcal{C})}^{*T}) - \mathrm{E}\{C_1(X_i^{*T}\beta^*)D^{-1}(X_i^{*T}\beta^*)C_1^T(X_i^{*T}\beta^*)\} \tag{3.1}$$

$$\Sigma_2 = \mathrm{E}(V^* V^{*T}) - \mathrm{E}\{C_2(X_i^{*T}\beta^*)D^{-1}(X_i^{*T}\beta^*)C_2^T(X_i^{*T}\beta^*)\} \tag{3.2}$$

where

$$V^* = \dot{f}^T(X_i^{*T}\beta^*)G^* X^*, \quad D(u) = \mathrm{E}\{G_{(\mathcal{V})}^* G_{(\mathcal{V})}^{*T} | X_i^{*T}\beta^* = u\}$$

$$C_1(u) = \mathrm{E}\{G_{(\mathcal{C})}^{*T} G_{(\mathcal{V})}^{*T} | X_i^{*T}\beta^* = u\}, \quad C_2(u) = \mathrm{E}\{V^* G_{(\mathcal{V})}^{*T} | X_i^{*T}\beta^* = u\}$$

Then, we can get the asymptotic normality of $\hat{\vartheta}^*$ in the following theorem.

Theorem 3.3. *Under the assumptions of Theorem 3.2, $\hat{\vartheta}^*$ is $\sqrt{n}$-consistent and*

$$\sqrt{n}(\hat{\vartheta}^* - \vartheta_0^*) \xrightarrow{\mathscr{D}} N(0, \Sigma) \tag{3.3}$$

where notation "$\xrightarrow{\mathscr{D}}$" represents "convergence in distribution" and

$$\Sigma = \begin{pmatrix} \Sigma_1^{-1} & 0 \\ 0 & J_{\phi^{*0}} \Sigma_2^{-1} J_{\phi^{*0}}^T \end{pmatrix}.$$

All the proofs can be found in Appendix.

4. Simulation

We conducted extensive simulations to evaluate the performance of the proposed approach. The performance is measured in several ways: (1) classification accuracy of the $f(\cdot)$ function denoted as the oracle percentage; (2) IMSE of the estimated f-function; (3) selection accuracy of β; and (4) estimation accuracy of β by MSE. Denote R as the total number of simulation runs.

Oracle percentage of $f(\cdot)$ is defined as the percentage of correct classification out of a total of R simulations, for example, if $k \in \mathcal{V}$, and out of R simulations, $f_k(\cdot)$ is classified as varying for g times, then the oracle percentage of $f_k(\cdot)$ is $\frac{g}{R} \times 100\%$. IMSE of $f_k(\cdot)$ is defined as

$$\mathrm{IMSE} = \frac{1}{R}\sum_{\ell=1}^{R}\left(\frac{1}{n_{grid}}\sum_{j=1}^{n_{grid}}\left(f_k(u_j) - B^T(u_j)\hat{\gamma}_k^{(\ell)}\right)^2\right) \tag{4.1}$$

where n_{grid} is the number of points used to estimate the IMSE of the predicted function; $\hat{\gamma}_k^{(\ell)}$ are the estimators of the B-spline coefficients for the ℓth simulation; $\hat{\beta}^{(\ell)}$ is the estimator of the loading parameter β for the ℓth simulation; u_j is taken at the $j/n_{grid} \times 100\%$ quantile among the range of $X^T\hat{\beta}^{(\ell)}$. For our simulations, n_{grid} was set to be 100.

Oracle percentage of β is defined as the percentage of correct selection of β out of R simulations. For example, if $\beta_d \neq 0$ and out of R simulations, β_d is selected to be non-zero for g times, then the oracle percentage of β_d is $\frac{g}{R} \times 100\%$. MSE of β_d is calculated as $\frac{1}{R}\sum_{\ell=1}^{R}(\hat{\beta}_d^{(\ell)} - \beta_d)^2$ where $\hat{\beta}_d^{(\ell)}$ is the estimator for β_d in the ℓth simulation.

The simulation data were generated according to model (2.1), where X were generated from a $Unif(0,1)$ distribution. For the loading parameter $\beta = (\beta_1, \beta_2, \cdots, \beta_q)^T$, $\beta_1 = \beta_2 = \frac{1}{\sqrt{2}}$ and the rest $\beta_j's$ were set as zeros. We evaluated the performance of the proposed approach with both continuous and discrete predictors $G_{\cdot k}$ $(k = 0, 1, 2, \cdots, p)$. For continuous variables $G_{\cdot k}$ $(k = 0, 1, 2, \cdots, p)$, they can be gene expressions. For discrete variables $G_{\cdot k}$ $(k = 0, 1, 2, \cdots, p)$, they can be single nucleotide polymorphism (SNP) variants. In either case, the dimension p can be large.

4.1. The continuous cases

In the continuous case, the nonparametric functions $f_k(u)$ $(k = 0, 1, 2, \cdots, p)$ were defined as follows: $f_0(u) = 2\sin(2\pi u)$, $f_1(u) = 2\cos(\pi u) + 2$ and $f_2(u) = \sin(2\pi u) + \cos(\pi u) + 1$ are varying functions; $f_3(u) = 2$ and $f_4(u) = 2.5$ are non-zero constants; $f_k(u) = 0$ are zeros for $k = 5, \cdots, p$. The number of loading parameters was set as $q = 5$ and $\beta_1 = \beta_2 = \frac{1}{\sqrt{2}}$, $\beta_3 = \beta_4 = \beta_5 = 0$. Both $G_{\cdot k}$ $(k = 0, 1, 2, \cdots, p)$ and ϵ were generated from independent $N(0,1)$. We run 1000 simulations (R = 1000) to evaluate the performance of the proposed variable selection approach under $p = 50, 100$.

Table 1 demonstrates the selection and estimation accuracy for continuous $G_{\cdot k}$. The left and right penal corresponds to the case where $p = 50$ and 100 respectively. For all the cases, the selection accuracy (oracle %) is very closed to 100% ($> 99\%$), IMSE for varying functions ($f_0(\cdot), f_1(\cdot)$ and $f_2(\cdot)$) are in the order of -2, and IMSE for non-zero constant functions ($f_3(\cdot)$ and $f_4(\cdot)$) are in the order of -3. All of the model IMSE and oracle IMSE are in the same order. These observations indicate that our proposed estimation and selection approach possesses reasonable selection and estimation accuracy for the non-parametric function $f_k(\cdot)$ $(k = 0, 1, 2, \cdots, p)$.

Table 2 presents the selection and estimation accuracy for the loading parameter β. The results shows that the selection accuracy for all β is reasonably good ($> 98\%$) in all cases. For most of the β, the MSE is in the order of -4 or lower, except for β_2, which is -3 for both $p = 50$ and $p = 100$ when $n = 500$. The order of the model estimation for β are at least the same as that of the

TABLE 1

Selection (%) and estimation accuracy (IMSE) of $f_k(\cdot)$ for continuous G.

Sample size	Function	$p = 50$			$p = 100$		
		Oracle %	Model	Oracle	Oracle %	Model	Oracle
$n = 500$	$f_0(\cdot)$	100.0%	3.87E-02	4.27E-02	100.0%	3.77E-02	4.51E-02
	$f_1(\cdot)$	99.6%	1.58E-02	2.42E-02	99.9%	1.57E-02	3.14E-02
	$f_2(\cdot)$	99.9%	2.33E-02	2.58E-02	99.9%	2.26E-02	2.96E-02
	$f_3(\cdot)$	100.0%	2.09E-03	2.11E-03	100.0%	1.90E-03	1.97E-03
	$f_4(\cdot)$	100.0%	2.04E-03	2.06E-03	100.0%	2.07E-03	2.12E-03
	Zero	99.7%	1.94E-05	0	99.9%	1.12E-05	0
$n = 1000$	$f_0(.)$	100.0%	3.23E-02	3.40E-02	100.0%	3.31E-02	3.47E-02
	$f_1(\cdot)$	100.0%	7.17E-03	1.21E-02	100.0%	7.07E-03	1.17E-02
	$f_2(\cdot)$	100.0%	1.46E-02	1.59E-02	100.0%	1.46E-02	1.64E-02
	$f_3(\cdot)$	100.0%	1.02E-03	1.02E-03	100.0%	9.60E-04	9.55E-04
	$f_4(\cdot)$	100.0%	1.09E-03	1.09E-03	100.0%	1.06E-03	1.07E-03
	Zero	99.8%	8.50E-06	0	99.9%	3.46E-06	0

TABLE 2

Selection (%) and estimation accuracy (MSE) of β for continuous G.

Sample size	β	$p = 50$			$p = 100$		
		Oracle %	Model	Oracle	Oracle %	Model	Oracle
$n = 500$	β_1	100.0%	1.15E-04	1.07E-04	100.0%	1.17E-04	1.30E-04
	β_2	100.0%	8.04E-03	4.12E-03	100.0%	2.26E-03	7.62E-03
	β_3	98.1%	9.98E-05	0	98.2%	3.64E-05	0
	β_4	98.8%	2.99E-05	0	99.1%	3.13E-05	0
	β_5	98.6%	1.00E-04	0	98.5%	7.73E-05	0
$n = 1000$	β_1	100.0%	5.30E-05	5.52E-05	100.0%	5.00E-05	5.49E-05
	β_2	100.0%	5.34E-05	1.86E-03	100.0%	5.04E-05	1.79E-03
	β_3	98.9%	9.36E-06	0	98.8%	1.16E-05	0
	β_4	99.4%	6.30E-06	0	99.5%	5.49E-06	0
	β_5	99.1%	7.17E-06	0	99.0%	6.93E-06	0

oracle model if not lower. These results indicate that our model possesses good selection and estimation accuracy for the loading parameters β.

4.2. The discrete case

We further evaluated how the proposed model performs with discrete $G_{\cdot k}$ $(k = 0, 1, 2, \cdots, p)$, i.e., SNP data. In this simulation, each $G_{\cdot k}$ $(k = 0, 1, 2, \cdots, p)$ variable was simulated from a multinomial distributions with minor allele frequency (MAF) P_a. The $G_{\cdot k}$ $(k = 0, 1, 2, \cdots, p)$ variable takes values $0, 1, 2$ corresponding to the genotype aa, Aa, and AA with corresponding genotype frequency P_a^2, $2P_a(1 - P_a)$ and $(1 - P_a)^2$, respectively. We set $P_a = 0.5$ for $k = 1, 2, 7$; $P_a = 0.3$ for $k = 3, 4, 8$; $P_a = 0.1$ for $k = 5, 6, 9$ and $P_a \sim Unif(0.05, 0.5)$ for $k = 10, 11, \cdots, p$. For the non-parametric functions, $f_0(u) = 2\sin(2\pi u)$, $f_1(u) = f_3(u) = f_5(u) = 2\cos(\pi u) + 2$, $f_2(u) = f_4(u) = f_6(u) = \sin(2\pi u) + \cos(\pi u) + 1$; $f_7(u) = f_8(u) = f_9(u) = 2$; and $f_k(u) = 0$ for $k = 10, 11, \cdots, p$. Under the setup, we had both varying and constant effect with different minor allele frequencies. X was generated from $Unif(0, 1)$ and ϵ was generated from $N(0, 1)$. Finally, Y was generated according to model (2.1). We evaluated the performance of the proposed model via $R = 1000$ simulations under $p = 50, 100$ and $n = 500, 1000$.

Table 3 presents the selection and estimation accuracy of the non-parametric

Table 3
Selection (%) and estimation accuracy (IMSE) of $f_k(\cdot)$ for discrete G.

Sample size	Function	$p = 50$			$p = 100$		
		Oracle %	Model	Oracle	Oracle %	Model	Oracle
$n = 500$	$f_0(.)$	100.0%	5.94E-02	5.66E-02	100.0%	7.42E-02	6.95E-02
	$f_1(\cdot)$	98.9%	3.71E-02	4.87E-02	98.4%	3.78E-02	5.44E-02
	$f_2(\cdot)$	99.0%	4.14E-02	3.79E-02	98.6%	4.30E-02	4.09E-02
	$f_3(\cdot)$	99.0%	3.50E-02	4.76E-02	98.5%	3.64E-02	5.81E-02
	$f_4(\cdot)$	98.9%	4.04E-02	3.63E-02	98.5%	4.48E-02	3.98E-02
	$f_5(\cdot)$	99.0%	4.02E-02	4.95E-02	98.6%	4.50E-02	7.29E-02
	$f_6(\cdot)$	98.8%	5.03E-02	4.52E-02	98.4%	4.98E-02	4.83E-02
	$f_7(\cdot)$	100.0%	2.37E-03	2.33E-03	99.9%	2.57E-03	2.51E-03
	$f_8(\cdot)$	100.0%	2.37E-03	2.37E-03	100.0%	2.55E-03	2.64E-03
	$f_9(\cdot)$	100.0%	2.66E-03	2.38E-03	100.0%	2.26E-03	2.24E-03
	Zero	99.6%	3.25E-05	0	99.7%	2.88E-05	0
$n = 1000$	$f_0(.)$	100.0%	3.12E-02	3.20E-02	100.0%	3.09E-02	3.44E-02
	$f_1(\cdot)$	99.9%	7.92E-03	1.22E-02	99.9%	7.96E-03	1.22E-02
	$f_2(\cdot)$	99.9%	1.50E-02	1.63E-02	99.9%	1.47E-02	1.59E-02
	$f_3(\cdot)$	99.9%	7.87E-03	1.21E-02	99.9%	8.19E-03	1.26E-02
	$f_4(\cdot)$	99.9%	1.44E-02	1.60E-02	99.9%	1.43E-02	1.58E-02
	$f_5(\cdot)$	99.9%	8.40E-03	1.17E-02	99.9%	8.54E-03	1.33E-02
	$f_6(\cdot)$	99.9%	1.48E-02	1.62E-02	99.9%	1.44E-02	1.64E-02
	$f_7(\cdot)$	100.0%	1.13E-03	1.14E-03	100.0%	9.55E-04	9.50E-04
	$f_8(\cdot)$	100.0%	1.14E-03	1.20E-03	100.0%	1.12E-03	1.16E-03
	$f_9(\cdot)$	100.0%	1.03E-03	1.04E-03	100.0%	1.13E-03	1.14E-03
	Zero	99.8%	9.21E-06	0	99.9%	4.38E-06	0

Table 4
Selection (%) and estimation accuracy (MSE) of β for discrete G.

Sample size	β	$p = 50$			$p = 100$		
		Oracle %	Model	Oracle	Oracle %	Model	Oracle
n = 500	β_1	100.0%	1.15E-04	1.07E-04	100.0%	1.17E-04	1.30E-04
	β_2	100.0%	8.04E-03	4.12E-03	100.0%	2.26E-03	7.62E-03
	β_3	98.1%	9.98E-05	0	98.2%	3.64E-05	0
	β_4	98.8%	2.99E-05	0	99.1%	3.13E-05	0
	β_5	98.6%	1.00E-04	0	98.5%	7.73E-05	0
n = 1000	β_1	100.0%	5.30E-05	5.52E-05	100.0%	5.00E-05	5.49E-05
	β_2	100.0%	5.34E-05	1.86E-03	100.0%	5.04E-05	1.79E-03
	β_3	98.9%	9.36E-06	0	98.8%	1.16E-05	0
	β_4	99.4%	6.30E-06	0	99.5%	5.49E-06	0
	β_5	99.1%	7.17E-06	0	99.0%	6.93E-06	0

function $f_k(\cdot)$. We observed that the oracle percentage are very high ($> 98.8\%$) for all cases, indicating our proposed model can correctly select the coefficient functions with high accuracy. Further, the IMSE for varying functions are of the order -2 or lower, while the IMSE for constant functions are of the order -3 or lower. Moreover, the IMSE of the proposed model are in the same order of the IMSE of the oracle model. These suggest that our model performs reasonably well in both selection and estimation for the non-parametric functions.

Table 4 presents the selection and estimation result of the loading parameters β. We observed that the oracle percentage in all the cases are above 98%, and the MSE for the estimation of β is in the order of -3 or lower in the proposed and oracle model. These suggests that our proposed model can correctly select and estimate the loading parameters with high accuracy.

In all the simulation studies, we observed improved performance when the sample size increases from 500 to 1000. For example, as shown in Table 4, the MSE for β_5 reduces from 1E-04 to 7.17E-06 when the sample size increases from

500 to 1000.

5. Real data application

We demonstrated the utility of the model with a human liver cohort (HLC) data set. The data set can be downloaded from www.synapse.org using synapse ID: syn4499 which contains gene expressions and phenotypes (activity of several liver enzymes). For more details regarding the data set, please refer to Schadt et al. [29] and Yang et al. [30]. In the HLC data set, the phenotypes are enzyme activity measurements of Cytochrom P450. There are a total of nine P450 enzymes (CYP1A2, 2A6, 2B6, 2C8, 2C9, 2C19, 2D6, 2E1, and 3A4). We chose CYP2E1 to demonstrate the utility of the method. For the environmental variable (X), we chose Age (=X_1), Aldehyde Oxydase (X_2), and Liver Triglyceride (X_3), then transformed each one of them to [0,1] with $\frac{X_i - \min(X_i)}{\max(X_i) - \min(X_i)}$. In this analysis, we focused on gene expressions which are treated as the G variable. After data cleaning, we had $n = 394$ (sample size) and $N = 19,172$ (number of gene expressions). Applying the proposed method, we would like to answer the following questions: (1) which gene is sensitive to the synergistic effect of the three X variables to affect the CYP2E1 activity? (2) what is the effect function of the three X variables as a whole, zero, constant or varying? and (3) which X variable contributes to the synergistic interaction effect?

We focused on the KEGG pathway “Metabolism of Xenobiotics by Cytochrome P450” (hsa00980) to select important genes associated with CYP2E1 activity. There are 76 genes in this pathway and 70 are mapped to our data set. After applying the proposed method, we identified one gene expression (SULT2A1) with varying effect and three gene expressions (FABP1, C15orf39, B3GNT5) with constant effect.

Figure 1 presents the plot of the intercept function (left panel) and the varying coefficient function for gene SULT2A1 (right panel) on CYP2E1 activity. After shrinkage, the coefficients for X_2 and X_3 were all zeros, leaving only Age as the effective environmental factor. The intercept function first increases before age 20, then it decreases dramatically for the rest of the life, showing the overall declining CYP2E1 enzyme activity over age. The effect of gene SULT2A1 on the CYP2E1 activity, however, behaves quite differently. The effect of this gene on CYP2E1 activity shows little change (around the zero line) before age 65. After that, it shows a positive effect on CYP2E1 activity as people become old. Gene SULT2A1 encodes sulfotransferase which aids in the metabolism of drugs and endogenous compounds. Study by Echchgadda et al. [31] showed that in senescent male rodents, SULT2A1 gene transcription in the liver is significantly enhanced due to the age-associated loss of the liver expression of androgen receptor. Although the study was conducted in rodents, it has implication on humans. Our result of enhanced function of SULT2A1 late in life agrees with the finding by Echchgadda et al. [31]. This result also demonstrates the unique strength of the proposed method to capture the non-linear interaction between environmental factors and genes. However, further biological investigation is

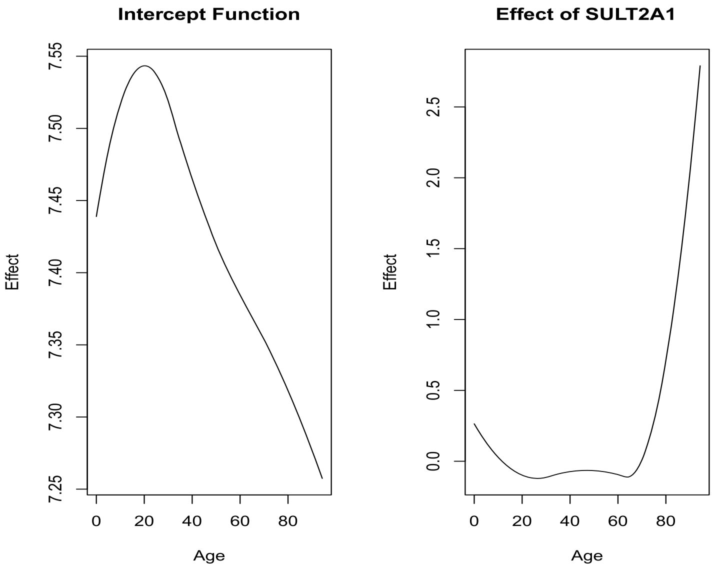

FIG 1. *Plot of the varying effect for gene SULT2A1.*

needed to confirm the real function of this gene modified by aging. In addition to this gene, genes with constant effect are FABP1 ($\hat{f} = 0.135$), C15orf39 ($\hat{f} = -0.112$) and B3GNT5 ($\hat{f} = -0.128$). The constant effects indicate that the effect of these genes on CYP2E1 does not change over age. In addition, the negative effect size tells that the CYP2E1 activity is negatively regulated by these genes. We did not find literature report to support that these genes show age-related expressions.

6. Discussion

SIVCM is a promising tool to model non-linear interactions between genes and multiple environments as a whole. It combines multiple exposure variables X into a single-index $X^T\beta$, hence can reduce model dimension and alleviate the curse of dimensionality. In this paper, we develop a three stage variable selection approach for SIVCM. Our goal is to identify varying, non-zero constant and zero effects which respectively correspond to nonlinear G × E effect, no G × E effect and no genetic effect. In the meantime, we also select important exposure variables. Rather than modeling the G × E effect for each X variable separately, our approach can model the joint effect of multiple environmental factors (X) as a whole, then identify how different genes interact with the environmental mixture to affect a disease trait, the so called synergistic G × E interaction. Our model is biologically motivated and attractive since it offers an alternative strategy to look for G×E interaction. In addition, our model is flexible to detect any potential non-linear interactions. We further studied the theoretical property of the proposed estimation and selection method. Both simulation and real data analysis demonstrate the utility of the proposed method.

In our model setup, the covariates X are assumed to be continuous. This is due to the fact that the index $u = X^T\beta$ has to be continuous in order to model the nonlinear function. In real applications, environmental variables can be discrete such as smoking, gender and ethnicity group. To accommodate the presence of discrete factors, the SIVCM can be generalized to a partial linear VMICM, i.e.,

$$Y = f(X^T\beta)G + Z\alpha + ZG\delta + \epsilon \tag{6.1}$$

where Z represent discrete covariates and α and δ represent the effects of Z and the interaction between Z and G, respectively. According to (2.3)–(2.5), we have

$$Y \approx W(\beta)\gamma + Z\alpha + ZG\delta + \epsilon \tag{6.2}$$

Our variable selection approach could be modified slightly to perform selection of non-parametric functions and the parametric components simultaneously. More specifically, the design matrix can be updated to $(W(\beta), Z, ZG)$ in Step 1 in the algorithm, then the rest follows.

So far we discussed the variable selection approach for SIVCM with a continuous response phenotype. In practice, many phenotype can be categorical such as a binary disease response in a case control study. It is natural to extend the current selection approach to a generalized SIVCM framework, which will be investigated in our future work.

In our model formulation, we assumed index coefficients share common loading parameters β. From a practical point of view, assuming different loading parameters makes perfect sense such as the model proposed by Ma and Song [32]. However, such a treatment imposes theoretical challenges when evaluating the theoretical properties such as the selection consistency. This is because that the loading coefficients for the kth index coefficient are not identifiable when $f_k(u) \notin \mathcal{V}$. When a coefficient function is not varying, β does not exists. Thus, the selection consistency for β does not exists. For this reason, we impose the same loading parameters for all the index coefficient functions. In addition to the application to G × E studies, our model has many applications in other fields where the purpose is to model the interaction between one variable and a mixture of a few other variables, the so called synergistic interaction.

Appendix A: Appendix

A.1. Computational algorithms

From (2.5), we have the design matrix $W(\beta)$ with the corresponding parameters $\gamma = (\gamma_0^T, \gamma_1^T, \cdots, \gamma_p^T)^T$ and $\gamma_k = (\gamma_{k1}, \gamma_{k*}^T)^T$. Then the detailed computational algorithms for Step 1, Step 2 and Step 3 are given as follows.

Computational algorithm for Step 1: In this step, we get the estimator $\hat{\gamma}^{(1)}$ denoted in (2.17) by minimizing the objective function $Q_1(\gamma|\Lambda_1, \hat{\beta}^{(0)})$ and using the group coordinate descent algorithm for iterative computation. We first assign a grouping index from 0 to M for each of the parameters. Furthermore, parameters with the same grouping index are in the same group and penalized as a group. Parameters with grouping index 0 are not penalized. Clearly, $\{\gamma_k, k = 0, 1, \ldots, p\} = \{\gamma_{(m)}, m = 0, 1, \ldots, M\}$, and $\{\hat{\gamma}_k, k = 0, 1, \ldots, p\} = \{\hat{\gamma}_{(m)}, m = 0, 1, \ldots, M\}$. Denote $W_{(m)}$ as the design matrix for group m, $m = 0, 1, \cdots M$. Given a tuning parameter λ and MCP tuning parameter τ^{MCP}, $\hat{\gamma}^{(1)}$ can be obtained through the following iteration.

(0) Run a Q-R decomposition on all $W_{(m)}$, i.e., $W_{(m)} = Q_{(m)} R_{(m)}$, $m = 0, 1, 2 \cdots M$, where $Q^T_{(m)} Q_{(m)} = I$ and $R_{(m)}$ is an upper triangular matrix, $Q_{(m)}$ is the normalized design matrix for group m.

(1) Assign the grouping index for the initial values $\hat{\gamma}^{(0)}$ from (2.16) such as $\{\hat{\gamma}^{(0)}_{(m)}, m = 0, 1, \cdots, M\}$, obtain the ordinary least squares (OLS) estimator $\hat{\gamma}^{OLS}_{(m)}$ via $\hat{\gamma}^{OLS}_{(m)} = Q^T_{(m)}(Y - Q_{-(m)}\hat{\gamma}_{-(m)}) = Q^T_{(m)} Y - Q^T_{(m)} Q_{-(m)} \hat{\gamma}_{-(m)}$, where subscript $Q_{-(m)}$ represents the normalized design matrix without group m and $\hat{\gamma}_{-(m)}$ represents the most updated values for γ without group m.

(2) For $m = 0$, set $\hat{\gamma}_{(0)} = \hat{\gamma}^{OLS}_{(0)}$.

(3) For $m = 1, \cdots, M$, obtain the MCP estimate $\hat{\gamma}_{(m)}$ via

$$\hat{\gamma}_{(m)} = \begin{cases} \hat{\gamma}^{OLS}_{(m)}, & \text{if } \|\hat{\gamma}^{OLS}_{(m)}\| > \lambda\tau^{MCP} \\ \frac{\tau}{1-\tau} S(\hat{\gamma}^{OLS}_{(m)}, \lambda), & \text{if } \|\hat{\gamma}^{OLS}_{(m)}\| \leq \lambda\tau^{MCP} \end{cases}, \tag{A.1}$$

where $S(\hat{\gamma}^{OLS}_{(m)}, \lambda) = \hat{\gamma}^{OLS}_{(m)} \left(1 - \frac{\lambda}{\|\hat{\gamma}^{OLS}_{(m)}\|}\right)_+$.

(4) Updated $\hat{\gamma}^{(0)}_{(m)}$ in step (1) by $\hat{\gamma}_{(m)}$.

Iterate step (1) through step (4) until convergence and get an unadjusted MCP estimator denoted as $\hat{\gamma}^{unadjusted}$. Then, we can get an adjusted MCP estimator as

$$\hat{\gamma}_{(m)} = R^{-1}_{(m)} \hat{\gamma}^{unadjusted}_{(m)}, \qquad m = 0, 1, \cdots, M \tag{A.2}$$

Accordingly, we have $\{\hat{\gamma}^{(1)}_k, k = 0, 1, \ldots, p\} = \{\hat{\gamma}_{(m)}, m = 0, 1, \ldots, M\}$. Finally, we can get our Step 1 estimator $\hat{\gamma}^{(1)} = ((\hat{\gamma}^{(1)}_0)^T, (\hat{\gamma}^{(1)}_1)^T, \cdots, (\hat{\gamma}^{(1)}_p)^T)^T$.

Computational algorithm for Step 2: In this step, given $\hat{\gamma}^{(1)}$ in Step 1, we get the estimator $\hat{\gamma}^{(2)}$ denoted in (2.19) and use the group coordinate descent algorithm for iterative computation, same as in Step 1. We first get different design matrix and grouping index according to $\hat{\gamma}^{(1)}$; then, repeat Step 1 until convergence to get $\hat{\gamma}^{(2)}$.

Computational algorithm for Step 3: In this step, given $\hat{\gamma}^{(2)}$ in Step 2, we get $\hat{\beta}$ denoted in (2.21). We adopt the idea of first order approximation and coordinate decent algorithm to estimate β by minimizing (2.22). Since $\bar{B}(X^T\beta)$ is not a linear function of β, there is no closed form solution of β. Hence, we apply a local linear approximation of $\bar{B}(X^T\beta)$ at $\tilde{\beta}$, and $\tilde{\beta}$ is the most updated value of β. We have

$$\bar{B}(X^T\beta)\hat{\gamma}_{k*} \approx \bar{B}(X^T\tilde{\beta})\hat{\gamma}_{k*} + \bar{B}'(X^T\tilde{\beta})\hat{\gamma}_{k*}X(\beta - \tilde{\beta}) \tag{A.3}$$

Working with β_d, $d = 1, 2, \cdots, q$, we have

$$\bar{B}(X^T\beta)\hat{\gamma}_k^* \approx \bar{B}(X^T\tilde{\beta})\hat{\gamma}_k^* + \bar{B}'(X^T\tilde{\beta})\hat{\gamma}_k^* X_d(\beta_d - \tilde{\beta}_d) \tag{A.4}$$

Then we can obtain $\hat{\beta}_d$ by minimizing the following penalized function,

$$Q_d = \|Y_d^* - X_d^*\beta_d\|^2 + np_{\lambda_3}(|\beta_d|) \tag{A.5}$$

where

$$Y_d^* = Y - \sum_{k=0}^{p}[\hat{\gamma}_{k1}G_k + \bar{B}(X^T\tilde{\beta})\hat{\gamma}_k^* G_k - \bar{B}^T(X\tilde{\beta})\hat{\gamma}_k^* G_k X_d \tilde{\beta}_d],$$

$$X_d^* = \sum_{k=0}^{p} \bar{B}^T(X\tilde{\beta})\hat{\gamma}_k^* G_k X_d.$$

Then, the MCP penalized estimator $\hat{\beta}^* = (\hat{\beta}_1^*, \cdots, \hat{\beta}_q^*)^T$ can be obtained via the coordinate descent algorithm. Since there are two constrains on β: (1) $\|\beta\|_2 = 1$ and (2) $\beta_1 > 0$. We do not penalize β_1 and normalize β after updating β, i.e., $\hat{\beta}_d = \frac{\hat{\beta}_d^*}{\|\hat{\beta}^*\|}\mathrm{sgn}(\hat{\beta}_1^*)$. The detailed algorithm for estimating $\beta_d, d = 1, 2, \cdots, q$, is given as follows:

(0) Get the initial estimator $\hat{\beta}^{(0)}$ from (2.15);

(1) Calculate Y_d^* and X_d^*;

(2) Normalized X_d^* by $\tilde{X}_d^* = X_d^*/\|X_d^*\|$;

(3) Calculate $\hat{\beta}_d^{OLS} = \tilde{X}_d^{*T} Y_d^*$

(4) Let $\hat{\beta}_1^* = \hat{\beta}_1^{OLS}$ and for $d \neq 1, \hat{\beta}_d^* = \frac{(\hat{\beta}_d^{OLS}-\lambda)_+}{1-1/\tau^{MCP}}$ if $|\hat{\beta}_d^{OLS}| \leq \lambda\tau^{MCP}$ and $\hat{\beta}_d^* = \hat{\beta}_d^{OLS}$ if $|\hat{\beta}_d^{OLS}| > \lambda\tau^{MCP}$;

(5) Normalized $\hat{\beta}^* = (\hat{\beta}_1^*, \cdots, \hat{\beta}_q^*)^T$, i.e., $\hat{\beta}_d = \frac{\hat{\beta}_d^*}{\|\hat{\beta}^*\|}\mathrm{sgn}(\hat{\beta}_1^*)$;

(6) Update $\hat{\beta}^{(0)}$ in step (0) with $\hat{\beta} = (\hat{\beta}_1, \cdots, \hat{\beta}_q)^T$, then iterate until convergence.

A.2. Proofs of theorems

The following regularity conditions are assumed.

(A1) The density function $f_U(u)$ of a random variable $U = X^T\beta$ is bounded away from 0 on $\mathcal{U} = \{u = X^T\beta : X \in \mathcal{X}\}$, where $\mathcal{X}$ is the compact support of X. Furthermore, we assume that $f_u(\cdot)$ satisfies the Lipschitz condition of order 1 on $\mathcal{U}$;

(A2) $f_k(\cdot)$ $(k = 0, 1, \cdots, p)$ have bounded and continuous derivatives up to order r on $\mathcal{U}$ and $r \geq 2$;

(A3) $\mathrm{E}(\|G\|^6) < \infty$ and $\mathrm{E}(|\epsilon|^6) < \infty$;

(A4) $\{(Y_i, X_i, G_i), i = 1, 2, \cdots, n\}$ is a strictly stationary and strongly mixing sequence with mixing coefficient $\alpha(n) = O(\rho^n)$ for some $0 < \rho < 1$;

(A5) Let $b_n = \max_{k,l}\{p''_{\lambda_1}(\|\gamma^0_{k*}\|), p''_{\lambda_2}(|\gamma^0_{k1}|), p''_{\lambda_3}(|\beta^0_d|), \gamma^0_{k*} \neq 0, \gamma^0_{k1} \neq 0, \beta^0_l \neq 0\}$ for $k = 1, \cdots, p, d = 2, \cdots, q$, then $b_n \to 0$ as $n \to 0$;

(A6) $\liminf_{n\to\infty} \liminf_{\|\gamma_{k*}\|\to 0^+} \frac{1}{\lambda_1}|p'_{\lambda_1}(\|\gamma_{k*}\|)| > 0$ for $k = v+1, \cdots, p$

$$\liminf_{n\to\infty} \liminf_{|\gamma_{k1}|\to 0^+} \frac{1}{\lambda_2}|p'_{\lambda_2}(|\gamma_{k1}|)| > 0 \text{ for } k = c+1, \cdots, p$$

$$\liminf_{n\to\infty} \liminf_{|\beta_d|\to 0^+} \frac{1}{\lambda_3}|p'_{\lambda_3}(|\beta_d|)| > 0 \text{ for } d = s+1, \cdots, q$$

(A7) Let $\kappa_1, \kappa_2, \cdots, \kappa_K$ be internal knots of $[a, b]$, where $a = \inf\{u : u \in \mathcal{U}\}$, $b = \sup\{u : u \in \mathcal{U}\}$. Furthermore, let $\kappa_1 = a$, $\kappa_{K+1} = b$, $h_i = \kappa_i - \kappa_{i-1}$, $h_{\max} = \max\{h_i\}$, $h_{\min} = \min\{h_i\}$. Then, there exist a constant C_0 such that $\frac{h_{\max}}{h_{\min}} < C_0$ and $\max\{h_{i+1} - h_i\} = o(K^{-1})$;

(A8) $D(u)$ is positive, and each element of $C_1(u)$ and $C_2(u)$ satisfy the Lipschitz condition of order 1 on $\mathcal{U}$.

Before the proof, we first define some notations as follows:

$$\Psi_{11} = \frac{1}{n}\sum_{i=1}^{n} W^*_{(\mathcal{V})i}(\phi^{*0})W^{*T}_{(\mathcal{V})i}(\phi^{*0}), \Psi_{12} = \frac{1}{n}\sum_{i=1}^{n} W^*_{(\mathcal{V})i}(\phi^{*0})G^{*T}_{(\mathcal{C})i},$$

$$\Psi_{13} = \frac{1}{n}\sum_{i=1}^{n} W^*_{(\mathcal{V})i}(\phi^{*0})V^{*T}_i, \Psi_{21} = \frac{1}{n}\sum_{i=1}^{n} G^*_{(\mathcal{C})i}W^{*T}_{(\mathcal{V})i}(\phi^{*0}),$$

$$\Psi_{22} = \frac{1}{n}\sum_{i=1}^{n} G^*_{(\mathcal{C})i}G^{*T}_{(\mathcal{C})i}, \Psi_{23} = \frac{1}{n}\sum_{i=1}^{n} G^*_{(\mathcal{C})i}V^{*T}_i,$$

$$\Psi_{31} = \frac{1}{n}\sum_{i=1}^{n} V_i W^{*T}_{(\mathcal{V})i}(\phi^{*0}), \Psi_{32} = \frac{1}{n}\sum_{i=1}^{n} V^*_i G^{*T}_{(\mathcal{C})i},$$

$$\Psi_{33} = \frac{1}{n}\sum_{i=1}^{n} V^*_i V^{*T}_i, \Lambda_{10} = \frac{1}{n}\sum_{i=1}^{n} W^*_{(\mathcal{V})i}(\phi^{*0})(\epsilon_i + R^T(\phi^{*0})G^*_i).$$

Lemma 1. *If $f_k(u)$ $(k = 0, 1, \cdots, p)$ satisfies condition (A2), then there exists a constat $C_0 > 0$ such that*

$$\sup_{u\in\mathcal{U}} |f_k(u) - B^T(u)\gamma_{k*}| \leq C_0 K^{-r} \tag{A.6}$$

Proof. This result follows directly from the standard B-spline theory. □

Lemma 2. *Suppose the regularity conditions (A1) – (A7) hold and the number of knots $K = O_p(n^{1/(2r+1)})$. Then we have*

$$\Psi_{22} - \Psi^T_{12}\Psi^{-1}_{11}\Psi_{12} \xrightarrow{P} \Sigma_1 \text{ and } \Psi_{33} - \Psi^T_{13}\Psi^{-1}_{11}\Psi_{13} \xrightarrow{P} \Sigma_2 \tag{A.7}$$

where notation "$\xrightarrow{P}$" represents convergence in probability.

Proof. The results of this lemma follow directly from [23] and [33]. □

Proof of Theorem 3.1. To show the consistency of $\hat{\beta}$ is equivalent to show the consistency of $\hat{\phi}$. Let $\alpha_n = n^{-r/(2r+1)} + a_n$, $\phi = \phi^0 + \delta\tau_1$, $\gamma = \gamma^0 + \delta\tau_2$ and $\tau = (\tau_1^T, \tau_2^T)^T$, where $\tau_2 = (\tau_{01}, \tau_{0*}, \cdots, \tau_{p1}, \tau_{p*})$ and $\{\tau_{k1}, \tau_{k*}\}$ corresponds to the B-spline coefficients $\{\gamma_{k1}, \gamma_{k*}\}$; $\tau_1 = (\tau_1^{\phi}, \cdots, \tau_{q-1}^{\phi})$; τ_l^{ϕ} corresponds to ϕ_l; and γ^0 and ϕ^0 are the true value of γ and ϕ, respectively.

To show the consistency of $\hat{\gamma}$ and $\hat{\phi}$, we need to show $\forall\epsilon > 0$, $\exists$ a large enough C such that

$$P\left\{\inf_{\|\tau\|=C}\{Q(\phi,\gamma)\} > Q(\phi^0,\gamma^0)\right\} \geq 1-\epsilon. \tag{A.8}$$

If (A.8) holds, we can say with probability at least $1-\epsilon$, there exists a local minimum in the ball $\{(\gamma^0,\phi^0) + \delta\tau : \|\tau\| \leq C\}$. Hence, there exists a local minimizer such that $\|(\hat{\gamma},\hat{\phi}) - (\gamma^0,\phi^0)\| = O_p(\delta)$.

Let $D_n(\tau) = K^{-1}\{Q(\gamma,\phi) - Q(\gamma^0,\phi^0)\}$, we can get

$$\begin{aligned}
D_n(\tau) = {}& \frac{1}{K}\sum_{i=1}^{n}\Big[\big(Y_i - W_i^T(\phi^0+\delta\tau_1)(\gamma^0+\delta\tau_2)\big)^2 - \big(Y_i - W_i^T(\phi^0)\gamma^0\big)^2\Big] \\
&+\frac{n}{K}\sum_{k=1}^{p}\left[p_{\lambda_{2k}}(|\gamma_{k1}^0+\delta\tau_{k1}|)I(\|\gamma_{k*}^0+\delta\tau_{k*}\|=0)-p_{\lambda_{2k}}(|\gamma_{k1}^0|)I(\|\gamma_{k*}^0\|=0)\right] \\
&+\frac{n}{K}\sum_{d=1}^{q-1}\Big[p_{\lambda_{3d}}(|\phi_d^0+\delta\tau_d^{\phi}|) - p_{\lambda_{3d}}(|\phi_d^0|)\Big]
\end{aligned}$$

Since $p_{\lambda_{1k}}(\|\gamma_{k*}^0\|)] = 0$ for $k = v+1,\cdots,p$ and $p_{\lambda_{3d}}(|\phi_d^0|) = 0$ for $d = s+1,\cdots,q-1$ and $I(\|\gamma_{k*}^0\| = 0) = 0$ for $k = 1,\cdots,v$, we have

$$\begin{aligned}
D_n(\tau) \geq {}& \frac{1}{K}\sum_{i=1}^{n}\Big[\big(Y_i - W_i^T(\phi^0+\delta\tau_1)(\gamma^0+\delta\tau_2)\big)^2 - \big(Y_i - W_i^T(\phi^0)\gamma^0\big)^2\Big] \\
&+\frac{n}{K}\sum_{k=1}^{v}\left[p_{\lambda_{1k}}(\|\gamma_{k*}^0+\alpha_n\tau_{k*}\|) - p_{\lambda_{1k}}(\|\gamma_{k*}^0\|)\right] \\
&+\frac{n}{K}\sum_{k=v+1}^{p}\left[p_{\lambda_{2k}}(|\gamma_{k1}^0+\delta\tau_{k1}|) - p_{\lambda_{2k}}(|\gamma_{k1}^0|)\right] \\
&+\frac{n}{K}\sum_{j=1}^{s-1}[p_{\lambda_{3d}}(|\phi_j^0+\alpha_n\tau_j^{\phi}|) - p_{\lambda_{3d}}(|\phi_j^0|)]
\end{aligned}$$

By Taylor Expansion at (γ^0,ϕ^0), following [23], we have

$$\begin{aligned}
D_n(\tau) \geq {}& \frac{-2\delta}{K}\sum_{i=1}^{n}\Big[(\epsilon_i + R^T(X_i^T\beta^0)G_i)(\dot{W}_i^T(\phi^0)\gamma^0 J_{\phi^0}^T X_i\tau_1 + W_i^T(\phi^0)\tau_2)\Big] \\
&+\frac{\delta^2}{K}\sum_{i=1}^{n}(\dot{W}_i^T(\phi^0)\gamma^0 J_{\phi^0}^T X_i\tau_1 + W_i^T(\phi^0)\tau_2)^2 + o_p(1)
\end{aligned}$$

$$
\begin{aligned}
&+\frac{n}{K}\sum_{k=1}^{v}\left[\delta p'_{\lambda_{1k}}(\|\gamma^0_{k*}\|)\frac{\gamma^0_{k*}}{\|\gamma^0_{k*}\|}\tau^T_{k*}+\delta^2 p''_{\lambda_{1k}}(\|\gamma^0_{k*}\|)\tau_{k*}\tau^T_{k*}(1+o_p(1))\right]\\
&+\frac{n}{K}\sum_{k=v+1}^{p}\left[\delta p'_{\lambda_{2k}}(|\gamma^0_{k1}|)\mathrm{sgn}(\gamma^0_{k1})\tau_{k1}+\delta^2 p''_{\lambda_{2k}}(|\gamma^0_{k1}|)(\tau_{k1})^2(1+o_p(1))\right]\\
&+\frac{n}{K}\sum_{d=1}^{s-1}\left[\delta p'_{\lambda_{3d}}(|\phi^0_{d}|)\mathrm{sgn}(\phi^0_{d})\tau^{\phi}_{d}+\delta^2 p''_{\lambda_{3d}}(|\phi^0_{d}|)(\tau^{\phi}_{d})^2(1+o_p(1))\right]\\
=:&\ S_1+S_2+o_p(1)+S_3+S_4+S_5
\end{aligned}
$$

where $\dot{W}_i(\phi^0) = I_{p+1}\otimes\dot{B}(X_i^T\beta^0)\cdot G_i,\ R(u) = (R_0(u), R_2(u),\cdots,R_p(u))^T$, $R_k(u) = f_k(u) - B^T(u)\gamma^0_k,\ k = 0,1,\cdots,p$. From Lemma 1, we have $|R_k(u)| = O(K^{-r})$ and

$$
|\dot{f}_k(X_i^T\beta^0)-\dot{B}^T(u)\gamma^0_k|\le C_0K^{-r+1} \tag{A.9}
$$

Note that ϵ_i is independent of (X_i, G_i), we have

$$
\frac{1}{\sqrt{n}}\sum_{1}^{n}\epsilon_i(\dot{W}_i^T(\phi^0)\gamma^0J^T_{\phi^0}X_i\tau_1+W_i^T(\phi^0)\tau_2)=O_p(\|\tau\|) \tag{A.10}
$$

In addition, from (A.4), we can get

$$
\begin{aligned}
&\sum_{i=1}^{n}R^T(X_i^T\beta^0)G_i(\dot{W}_i^T(\phi^0)\gamma^0J^T_{\phi^0}X_i\tau_1+W_i^T(\phi^0)\tau_2)\\
&\qquad=\sum_{i=1}^{n}R^T(X_i^T\beta^0)G_i\{\dot{f}^T(X_i^T\beta^0)G_iJ^T_{\phi^0}X_i\tau_1\\
&\qquad\quad+(\dot{W}_i^T(\phi^0)\gamma^0-\dot{f}^T(X_i^T\beta^0)G_iJ^T_{\phi^0}X_i\tau_1+W_i^T(\phi^0)\tau_2)\}\\
&\qquad=O_p(nK^{-r}\|\tau\|).
\end{aligned} \tag{A.11}
$$

Following [23], from (A.10), (A.11) and (A.4), it is easy to show that

$$
S_1=O_p(\sqrt{n}K^{-1}\delta)\|\tau\|+O_p(nK^{-1-r}\delta)\|\tau\|=O_p(1+n^{r/(2r+1)}a_n)\|\tau\|. \tag{A.12}
$$

Similarly, we can get

$$
S_2=O_p(\sqrt{n}K^{-1}\delta^2)\|\tau\|^2=O_p(1+2n^{r/(2r+1)}a_n)\|\tau\|^2. \tag{A.13}
$$

Hence, S_2 dominates S_1 uniformly in $\{\tau : \|\tau\| = C\}$ by choosing a sufficiently large C.

Further, by Taylor expansion at γ^0, we have

$$
\begin{aligned}
S_3&\le nK^{-1}\delta a_n\sum_{k=1}^{v}\frac{\gamma^0_{k*}}{\|\gamma^0_{k*}\|}\tau^T_{k*}+nK^{-1}\delta^2b_n\sum_{k=1}^{v}\tau_{k*}\tau^T_{k*}\\
&\le nK^{-1}\delta a_n\sqrt{v}\|\tau\|+nK^{-1}\delta^2b_n\|\tau\|^2
\end{aligned}
$$

Since $b_n \to 0$, then it is easy to show that S_3 is dominated by S_2 uniformly in $\|\tau\| = C$.

For S_4 and S_5, we have

$$S_4 \leq \delta a_n nK^{-1} \sum_{k=v+1}^{p} \tau_{k1} + nK^{-1}\delta^2 b_n \sum_{k=v+1}^{p} (\tau_{k1})^2 \leq nK^{-1}\delta^2 C + nK^{-1}\delta^2 C^2 b_n,$$

$$S_5 \leq \delta a_n nK^{-1} \sum_{j=1}^{s} \tau_j^{\phi} + nK^{-1}\delta^2 b_n \sum_{j=1}^{s} (\tau_j^{\phi})^2 \leq nK^{-1}\delta^2 C + nK^{-1}\delta^2 C^2 b_n.$$

With the same argument, we have S_4 and S_5 dominated by S_2 uniformly in $\|\tau\| = C$. Hence, by choosing a large enough C, (A.8) holds. Therefore, there exists local minimizers $\hat{\phi}$ and $\hat{\gamma}$ such that

$$\|\hat{\phi} - \phi^0\| = O_p(\delta), \;\; \|\hat{\gamma} - \gamma^0\| = O_p(\delta).$$

So we can get $\|\hat{\beta} - \beta^0\| = O_p(\delta)$, which completes the proof of (i). □

Note that

$$\begin{aligned}
\|\hat{f}_k(u) - f_k^0(u)\| &= \int_{\mathcal{U}} \{\hat{f}_k(u) - f_k^0(u)\}^2 du \\
&= \int_{\mathcal{U}} \{B^T(u)\hat{\gamma}_k - B^T(u)\gamma_k^0 + R_k(u)\}^2 du \\
&\leq 2\int_{\mathcal{U}} \{B^T(u)\hat{\gamma}_k - B^T(u)\gamma_k^0\}^2 du + 2\int_{\mathcal{U}} R_k^2(u) du \\
&= 2(\hat{\gamma}_k - \gamma_k^0)^T \left(\int_{\mathcal{U}} B^T(u)B(u)du\right)(\hat{\gamma}_k - \gamma_k^0) + 2\int_{\mathcal{U}} R_k^2(u)du.
\end{aligned}$$

It is obvious that $\int_{\mathcal{U}} B^T(u)B(u)du = O(1)$, so we can get

$$(\hat{\gamma}_k - \gamma_k^0)^T \left(\int_{\mathcal{U}} B^T(u)B(u)du\right)(\hat{\gamma}_k - \gamma_k^0) = O_p(n^{-2r/(2r+1)} + a_n^2). \tag{A.14}$$

In addition, from Lemma 1, it is easy to show that

$$\int_{\mathcal{U}} R_k^2(u)du = O_p(n^{-2r/(2r+1)}). \tag{A.15}$$

According to (A.14) and (A.15), we complete the proof of (ii).

Proof of Theorem 3.2. (i) Without loss of generality, we denote $\phi = (\phi^{nz}, \phi^z)$, where $\phi^{nz} = (\phi_1, \cdots, \phi_{s-1})$ and $\phi^z = (\phi_s, \cdots, \phi_{q-1})$. Since $\lambda_{\max} \to 0$, it can be seen $a_n = 0$ for large n. Then, by Theorem 3.1, it is sufficient to show

$$\|\phi_j - \phi_j^0\| = O_p(n^{-r/(2r+1)}), \quad d = 1, \cdots, s-1$$

for ϕ^{nz}. For ϕ^{z}, for some given small $\varepsilon = Cn^{-r/(2r+1)}$, with probability approaching 1 as $n \to \infty$, for $d = s, \cdots, q-1$, we have

$$\frac{\partial Q(\phi,\gamma)}{\partial \phi_d} > 0 \text{ when } 0 < \phi_d < \varepsilon \text{ and } \frac{\partial Q(\phi,\gamma)}{\partial \phi_d} < 0 \text{ when } -\varepsilon < \phi_d < 0.$$

We have

$$\frac{\partial Q(\phi,\gamma)}{\partial \phi_d} = \frac{\partial g(\gamma,\phi)}{\partial \phi_d} + np_{\lambda_{3d}}(|\phi_d|)\text{sgn}(\phi_d)$$

$$\begin{aligned}\frac{\partial Q(\phi,\gamma)}{\partial \phi_d} &= \sum_{i=1}^{n} \left(Y_i - W_i^T(\phi)\gamma\right) \dot{W}_i^T(\phi)\gamma e_{\phi_d}^T X_i + n\dot{p}_{\lambda_{3d}}(|\phi_d|)\text{sgn}(\phi_d) \\ &= \sum_{i=1}^{n} \{\epsilon_i + R^T(X_i^T\beta^0)G_i + (I_{p+1} \otimes B(X_i\beta^0) \cdot G_i)^T(\gamma^0 - \gamma) \\ &\quad + (I_p \otimes [B(X_i^T\beta^0) - B(X_i^T\beta)] \cdot G_i)^T\gamma\} W_i^T(\phi)\gamma e_{\phi_d}^T X_i \\ &\quad + np'_{3d}(|\phi_d|)\text{sgn}(\phi_d)\end{aligned}$$

where $e_{\phi_d} = (-(1-\|\phi\|^2)^{-1/2}\phi_d, 0, \cdots, 0, 1, 0, \cdots, 0)^T$ with $(d+1)$th component as 1. From conditions (A.1), (A.2), (A.4) and (A.9), similar to [23], we have

$$\frac{\partial Q(\phi,\gamma)}{\partial \phi_d} = n\lambda_{3d}\{\lambda_{3d}^{-1} p'_{\lambda_{3d}}(|\phi_d|)\text{sgn}(\phi_d) + O_p(n^{-r/(2r+1)}\lambda_{3d}^{-1})\} \tag{A.16}$$

Clearly we can see that $\lambda_{3d} n^{r/(2r+1)} \geq \lambda_{\min} n^{r/(2r+1)} \to \infty$, which implies $O_p(n^{-r/(2r+1)}\lambda_{3d}^{-1}) = o_p(1)$. From (A6), $\liminf_{n\to\infty} \liminf_{|\beta_d|\to 0^+} \frac{1}{\lambda_3}|p'_{\lambda_3}(|\beta_d|)| > 0$. So we can conclude that the sign of $\frac{\partial Q(\phi,\gamma)}{\partial \phi_j}$ is completely determined by sign of ϕ_j. Hence, we prove $\hat{\beta}_j = 0$ for $j = s+1, \cdots, q$. This completes the proof of (i).

(ii) & (iii) Applying similar arguments as in (i), we immediately have, with probability approaching 1, $\hat{\gamma}_{k*} = 0$ for $k = v+1, \cdots, p$ and $\hat{\gamma}_{k1} = 0$ for $k = c+1, \cdots, p$. Then by $\sup_u B(u) = O(1)$ and $\hat{f}_k(\cdot) = \hat{\gamma}_{k0} + \bar{B}(X\hat{\beta})\hat{\gamma}_{k*}$, we prove $\hat{f}_k(\cdot) = c_k$ for $k = v+1, \cdots, c$ where c_k is some constant and $\hat{f}_k(\cdot) = 0$ for $k = c+1, \cdots, p$. □

Proof of Theorem 3.3. By Theorems 3.1 and 3.2, we can see that, as $n \to \infty$, $Q(\phi,\gamma)$ attains the minimal value at $(\hat{\phi}^{*T}, 0)^T$ and $(\hat{\gamma}_{(\mathcal{V})}^{*T}, \hat{\gamma}_{(\mathcal{C})}^{*T}, 0)^T$. Obviously, according to (2.7), we can see that $\hat{\gamma}_{(\mathcal{C})}^{*} = (\hat{\gamma}_{v+1}^{*T}, \hat{\gamma}_{v+2}^{*T}, \cdots, \hat{\gamma}_{c+1}^{*T})^T$ and $\hat{\gamma}_k^* = (\hat{\gamma}_{k1}^*, 0, 0, \cdots, 0)^T$ for $k = v+1, v+2, \cdots, c$. Then, we have $\hat{f}_k(\cdot) = \hat{\gamma}_{k1}$ for $k = v+1, \cdots, c$. Denote $\theta^* = (\gamma_{(\mathcal{C})1}^{*T}, \phi^{*T})^T$, the real value of θ^* is $\theta^{*0} = (\gamma_{(\mathcal{C})1}^{*0T}, \phi^{*0T})^T$ and let

$$Q_{1n}(\phi,\gamma) = \frac{\partial Q(\phi,\gamma)}{\partial \gamma_{(\mathcal{V})}}, \quad Q_{2n}(\phi,\gamma) = \frac{\partial Q(\phi,\gamma)}{\partial \gamma_{(\mathcal{C})}}, \quad Q_{3n}(\phi,\gamma) = \frac{\partial Q(\phi,\gamma)}{\partial \phi}.$$

Then, $(\hat{\phi}^{*T}, 0)^T$ and $(\hat{\gamma}^{*T}_{(\mathcal{V})}, \hat{\gamma}^{*T}_{(\mathcal{C})}, 0)^T$ must satisfy

$$\begin{aligned}&\frac{1}{n} Q_{1n}((\hat{\phi}^{*T}, 0)^T, (\hat{\gamma}^{*T}_{(\mathcal{V})}, \hat{\gamma}^{*T}_{(\mathcal{C})}, 0)^T) \\ &\qquad = -\frac{2}{n} \sum_{i=1}^{n} W^{*}_{(\mathcal{V})i}(\hat{\phi}^{*}) \left(Y_i - W^{*T}_{(\mathcal{V})i}(\hat{\phi}^{*}) \hat{\gamma}^{*}_{(\mathcal{V})} - G^{*T}_{(\mathcal{C})i} \hat{\gamma}^{*}_{(\mathcal{C})1}\right) + V_1 = 0\end{aligned} \tag{A.17}$$

$$\begin{aligned}&\frac{1}{n} Q_{2n}((\hat{\phi}^{*T}, 0)^T, (\hat{\gamma}^{*T}_{(\mathcal{V})}, \hat{\gamma}^{*T}_{(\mathcal{C})}, 0)^T) \\ &\qquad = -\frac{2}{n} \sum_{i=1}^{n} W^{*}_{(\mathcal{C})i} \left(Y_i - W^{*T}_{(\mathcal{V})i}(\hat{\phi}^{*}) \hat{\gamma}^{*}_{(\mathcal{V})} - G^{*T}_{(\mathcal{C})i} \hat{\gamma}^{*}_{(\mathcal{C})1}\right) + V_2 = 0\end{aligned} \tag{A.18}$$

$$\begin{aligned}&\frac{1}{n} Q_{3n}((\hat{\phi}^{*T}, 0)^T, (\hat{\gamma}^{*T}_{(\mathcal{V})}, \hat{\gamma}^{*T}_{(\mathcal{C})}, 0)^T) \\ &\qquad = -\frac{2}{n} \sum_{i=1}^{n} \dot{W}^{*T}_{(\mathcal{V})i}(\hat{\phi}^{*}) \hat{\gamma}^{*}_{(\mathcal{V})} J^{T}_{\hat{\phi}^{*}} X^{*}_{(\mathcal{V})i} \left(Y_i - W^{*T}_{(\mathcal{V})i}(\hat{\phi}^{*}) \hat{\gamma}^{*}_{(\mathcal{V})} - G^{*T}_{(\mathcal{C})i} \hat{\gamma}^{*}_{(\mathcal{C})1}\right) \\ &\qquad\quad + V_3 = 0\end{aligned} \tag{A.19}$$

where

$$V_1 = \left(0, p'_{\lambda^{*}_{11}}(\|\hat{\gamma}^{*}_1\|) \frac{\hat{\gamma}^{*}_1}{\|\hat{\gamma}^{*}_1\|}, p'_{\lambda_{12}}(\|\hat{\gamma}^{*}_2\|) \frac{\hat{\gamma}^{*}_2}{\|\hat{\gamma}^{*}_2\|}, \cdots, p'_{\lambda_{1v}}(\|\hat{\gamma}_v\|) \frac{\hat{\gamma}_v}{\|\hat{\gamma}_v\|}\right)^T \in \mathbb{R}^{(K+h)(v+1)}$$

$$\begin{aligned} V_2 = \Big(&p'_{\lambda_{2(v+1)}}(|\hat{\gamma}^{*}_{(v+1)1}|) \mathrm{sgn}(|\hat{\gamma}^{*}_{(v+1)1}|), \\ &\ldots, p'_{\lambda_{2c}}(|\hat{\gamma}^{*}_{c}|) \mathrm{sgn}(|\hat{\gamma}^{*}_{c}|), 0, 0, \cdots, 0\Big)^T \in \mathbb{R}^{c-v+s-1}\end{aligned}$$

$$\begin{aligned} V_3 = \Big(&0, 0, \cdots, 0, p'_{\lambda_{31}}(|\hat{\phi}^{*}_1|) \mathrm{sgn}(|\hat{\phi}^{*}_1|), \\ &\ldots, p'_{\lambda_{3(s-1)}}(|\hat{\phi}^{*}_{s-1}|) \mathrm{sgn}(|\hat{\phi}^{*}_{s-1}|)\Big)^T \in \mathbb{R}^{c-v+s-1}.\end{aligned}$$

Applying Taylor expansion to $p'_{3d}(|\hat{\phi}^{*}_d|)$ $(d = 1, \cdots, s-1)$, we get

$$p'_{\lambda_{3d}}(|\hat{\phi}^{*}_d|) = p'_{\lambda_{3d}}(|\hat{\phi}^{0}_d|) + \{p''_{\lambda_{3d}}(|\hat{\phi}^{0}_d|) + o_p(1)\}(\hat{\phi}^{*0}_d - \phi^{*0}_d). \tag{A.20}$$

Furthermore, (A5) implies that $p''_{\lambda_{3d}}(|\hat{\phi}^{0}_d|) = o_p(1)$, and note that $p'_{\lambda_{3d}}(|\hat{\phi}^{0}_d|) = 0$ as $\lambda_{\max} \to 0$. Then, from Theorem 3.1 and 3.2, we have

$$p'_{\lambda_{3d}}(|\hat{\phi}_d|) \mathrm{sgn}(\hat{\phi}_d) = o_p(\hat{\phi}^{*} - \phi^{*0})$$

Similarly, we have

$$p'_{\lambda_{1k}}(\|\hat{\gamma}_k\|)\frac{\hat{\gamma}_k}{\|\hat{\gamma}_k\|} = o_p(\hat{\gamma}^*_{(\mathcal{V})} - \gamma^0_{(\mathcal{V})}), \quad k = 0, 1, 2, \cdots, v$$

$$p'_{\lambda_{2k}}(|\hat{\gamma}_k|)\mathrm{sgn}(\hat{\gamma}_k) = o_p(\hat{\gamma}^*_{(\mathcal{C})1} - \gamma^0_{(\mathcal{C})1}), \quad k = v+1, \cdots, c$$

Hence, by (A.17) and using Taylor expansion, a simple calculation yields

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} W^*_{(\mathcal{V})i}(\hat{\phi}^*)\left(Y_i - W^{*T}_{(\mathcal{V})i}(\hat{\phi}^*)\hat{\gamma}^*_{(\mathcal{V})} - G^{*T}_{(\mathcal{C})i}\hat{\gamma}^*_{(\mathcal{C})1}\right) \\
&\quad = \frac{1}{n}\sum_{i=1}^{n}[W^*_{(\mathcal{V})i}(\phi^{*0}) + W^*_{(\mathcal{V})i}(\hat{\phi}^*) - W^*_{(\mathcal{V})i}(\phi^{*0})]\Big(\epsilon_i + R^T(\phi^{*0})G^*_i \\
&\qquad - W^{*T}_{(\mathcal{V})i}(\phi^{*0})(\hat{\gamma}^*_{(\mathcal{V})} - \gamma^{*0}_{(\mathcal{V})}) - [W^{*T}_{(\mathcal{V})i}(\hat{\phi}^*) - W^{*T}_{(\mathcal{V})i}(\phi^{*0})]\hat{\gamma}^*_{(\mathcal{V})} \\
&\qquad - G^{*T}_{(\mathcal{C})i}(\hat{\gamma}^*_{(\mathcal{C})1} - \gamma^0_{(\mathcal{C})1})\Big) \\
&\quad = \frac{1}{n}\sum_{i=1}^{n} W^*_{(\mathcal{V})i}(\phi^{*0})\Big(\epsilon_i + R^T(\phi^{*0})G^*_i - W^{*T}_{(\mathcal{V})i}(\phi^{*0})(\hat{\gamma}^*_{(\mathcal{V})} - \gamma^{*0}_{(\mathcal{V})}) \\
&\qquad - [W^{*T}_{(\mathcal{V})i}(\hat{\phi}^*) - W^{*T}_{(\mathcal{V})i}(\phi^{*0})]\hat{\gamma}^*_{(\mathcal{V})} - G^{*T}_{(\mathcal{C})i}(\hat{\gamma}^*_{(\mathcal{C})1} - \hat{\gamma}^{*0}_{(\mathcal{C})1})\Big) + o_p(\hat{\phi}^* - \phi^{*0}) \\
&\quad = \frac{1}{n}\sum_{i=1}^{n} W^*_{(\mathcal{V})i}(\epsilon_i + R^T(\phi^{*0})G^*_i) - \frac{1}{n}\sum_{i=1}^{n} W^*_{(\mathcal{V})i}(\phi^{*0})W^{*T}_{(\mathcal{V})i}(\phi^{*0})(\hat{\gamma}^*_{(\mathcal{V})} - \gamma^{*0}_{(\mathcal{V})}) \\
&\qquad - \frac{1}{n}\sum_{i=1}^{n} W^*_{(\mathcal{V})i}[W^{*T}_{(\mathcal{V})i}(\hat{\phi}^*) - W^{*T}_{(\mathcal{V})i}(\phi^{*0})]\hat{\gamma}^*_{(\mathcal{V})} \\
&\qquad - \frac{1}{n}\sum_{i=1}^{n} W^*_{(\mathcal{V})i}G^{*T}_{(\mathcal{C})i}(\phi^{*0})(\hat{\gamma}^*_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1}) \\
&\qquad + o_p(\hat{\phi}^* - \phi^{*0}) \\
&\quad = \frac{1}{n}\sum_{i=1}^{n} W^*_{(\mathcal{V})i}(\epsilon_i + R^T(\phi^{*0})G^*_i) - \frac{1}{n}\sum_{i=1}^{n} W^*_{(\mathcal{V})i}W^{*T}_{(\mathcal{V})i}(\phi^*)(\hat{\gamma}^*_{(\mathcal{V})} - \gamma^{*0}_{(\mathcal{V})}) \\
&\qquad - \frac{1}{n}\sum_{i=1}^{n} W^*_{(\mathcal{V})i}V^{*T}_i(\hat{\phi}^* - \phi^{*0}) \\
&\qquad - \frac{1}{n}\sum_{i=1}^{n} W^*_{(\mathcal{V})i}G^{*T}_{(\mathcal{C})i}(\phi^*)(\hat{\gamma}^*_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1}) + o_p(\hat{\phi}^* - \phi^{*0})
\end{aligned}$$

Then, based on (A8), Theorem 3.1 and $\sup_u \|B(u)\| = O(1)$, we have

$$\hat{\gamma}^*_{(\mathcal{V})} - \gamma^{*0}_{(\mathcal{V})} = [\Psi_{11} + o_p(1)]^{-1}(\Lambda_{10} - \Psi_{12}(\hat{\gamma}^*_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1}) - \Psi_{13}(\hat{\phi}^* - \phi^{*0})) \quad \text{(A.21)}$$

Thus, according to (A.18), we can get

$$0 = \frac{1}{n}\sum_{i=1}^{n} G^*_{(\mathcal{C})i}\left(Y_i - W^{*T}_{(\mathcal{V})i}(\hat{\phi}^*)\hat{\gamma}^*_{(\mathcal{V})} - G^{*T}_{(\mathcal{C})i}\hat{\gamma}^*_{(\mathcal{C})1}\right)$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n} G^*_{(\mathcal{C})i}\Big(\epsilon_i + R^T(\phi^*)G^*_i - W^{*T}_{(\mathcal{V})i}(\phi^*)(\hat{\gamma}^*_{(\mathcal{V})} - \gamma^{*0}_{(\mathcal{V})}) \\
&\quad - [W^{*T}_{(\mathcal{V})i}(\hat{\phi}^*) - W^{*T}_{(\mathcal{V})i}(\phi^*)]\hat{\gamma}^*_{(\mathcal{V})} - G^{*T}_{(\mathcal{C})i}(\hat{\gamma}^*_{(\mathcal{C})1} - \hat{\gamma}^{*0}_{(\mathcal{C})1})\Big) + o_p(\hat{\gamma}^*_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1}) \\
&= \frac{1}{n}\sum_{i=1}^{n} G^*_{(\mathcal{C})i}\Big(\epsilon_i + R^T(\phi^*)G^*_i - W^{*T}_{(\mathcal{V})i}(\phi^*)[\Psi_{11} + o_p(1)]^{-1} \\
&\quad \times (\Lambda_{10} - \Psi_{12}(\hat{\gamma}^*_{(\mathcal{C})1} - \hat{\gamma}^{*0}_{(\mathcal{C})1}) \\
&\quad - \Psi_{13}(\hat{\phi}^* - \phi^*)) - [W^{*T}_{(\mathcal{V})i}(\hat{\phi}^*) - W^{*T}_{(\mathcal{V})i}(\phi^*)]\hat{\gamma}^*_{(\mathcal{V})} - G^{*T}_{(\mathcal{C})i}(\hat{\gamma}^*_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1})\Big) \\
&\quad + o_p(\hat{\gamma}^*_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1}) \\
&= \frac{1}{n}\sum_{i=1}^{n} G^*_{(\mathcal{C})i}\Big(\epsilon_i + R^T(\phi^*)G^*_i - W^{*T}_{(\mathcal{V})i}(\phi^*)[\Psi_{11} + o_p(1)]^{-1}\Lambda_{10}\Big) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} G^*_{(\mathcal{C})i}W^{*T}_{(\mathcal{V})i}(\phi^*)[\Psi_{11} + o_p(1)]\Psi_{12}(\hat{\gamma}^*_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1}) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} G^*_{(\mathcal{C})i}W^{*T}_{(\mathcal{V})i}(\phi^*)[\Psi_{11} + o_p(1)]\Psi_{13}(\hat{\phi}^* - \phi^{*0}) \\
&\quad - \frac{1}{n}\sum_{i=1}^{n} G^*_{(\mathcal{C})i}\Big([W^{*T}_{(\mathcal{V})i}(\hat{\phi}^*) - W^{*T}_{(\mathcal{V})i}(\phi^*)]\hat{\gamma}^*_{(\mathcal{V})}\Big) \\
&\quad - \frac{1}{n}\sum_{i=1}^{n} G^*_{(\mathcal{C})i}G^{*T}_{(\mathcal{C})i}(\hat{\gamma}^*_{(\mathcal{C})1} - \hat{\gamma}^{*0}_{(\mathcal{C})1}) + o_p(\hat{\gamma}^*_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1}) \\
&\triangleq J_1 + J_2 + J_3 - J_4 - J_5 + o_p(\hat{\gamma}^*_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1})
\end{aligned}
$$

Note that

$$
\frac{1}{n}\sum_{i=1}^{n} \Phi_{22}\Phi_{11}^{-1}W^*_{(\mathcal{V})i}(\phi^*)(\epsilon_i + R^T(\phi^*)G^*_i - W^{*T}_{(\mathcal{V})i}(\phi^*)\Psi_{11}^{-1}\Lambda_{10}) = 0
$$

$$
\frac{1}{n}\sum_{i=1}^{n}(G^*_{(\mathcal{C})i} - \Phi_{22}\Phi_{11}^{-1}W^*_{(\mathcal{V})i}(\phi^*))W^{*T}_{(\mathcal{V})i}(\phi^*) = 0
$$

Hence, we can get

$$
\begin{aligned}
J_1 = \frac{1}{n}\sum_{i=1}^{n}&(G^*_{(\mathcal{C})i} - \Psi_{22}\Psi_{11}^{-1}W^*_{(\mathcal{V})i}(\phi^*))\epsilon_i \\
&+ \frac{1}{n}\sum_{i=1}^{n}(G^*_{(\mathcal{C})i} - \Psi_{22}\Psi_{11}^{-1}W^*_{(\mathcal{V})i}(\phi^*))R(\phi^*)G^*_i \\
&+ \frac{1}{n}\sum_{i=1}^{n}(G^*_{(\mathcal{C})i} - \Psi_{22}\Psi_{11}^{-1}W^*_{(\mathcal{V})i}(\phi^*))W^{*T}_{(\mathcal{V})i}(\phi^*)[\Psi_{11} + o_p(1)]^{-1}
\end{aligned}
$$

$$
\begin{aligned}
&+ o_p(\hat{\gamma}^*_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1}) \\
&= \frac{1}{n}\sum_{i=1}^{n}(G^*_{(\mathcal{C})i} - \Psi_{22}\Psi_{11}^{-1}W^*_{(\mathcal{V})i}(\phi^*))\epsilon_i + o_p(\hat{\gamma}^*_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1})
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
J_2 &= \Phi_{22}\Psi_{11}^{-1}\Psi_{12}(\hat{\gamma}^*_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1}) + o_p(\hat{\gamma}^*_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1}) \\
J_3 &= \Psi_{22}\Phi_{11}^{-1}\Psi_{13}(\hat{\phi}^* - \phi^{*0}) + o_p(\hat{\phi}^* - \phi^{*0}) \\
J_4 &= \Psi_{23}(\hat{\phi}^* - \phi^{*0}) + o_p(\hat{\phi}^* - \phi^{*0}) \\
J_5 &= \Psi_{22}(\hat{\gamma}^*_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1}) + o_p(\hat{\gamma}^*_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1})
\end{aligned}
$$

So we can get

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(G^*_{(\mathcal{C})i} - \Psi_{22}\Psi_{11}^{-1}W^*_{(\mathcal{V})i}(\phi^*))\epsilon_i \\
&\quad= (\Psi_{22}\Psi_{11}^{-1}\Psi_{12} - \Psi_{22})(\hat{\gamma}^*_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1}) + (\Psi_{22}\Psi_{11}^{-1}\Psi_{13} - \Psi_{23})(\hat{\phi}^* - \phi^{*0}) \\
&\quad\quad + o_p(\hat{\gamma}^*_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1}) + o_p(\hat{\phi}^* - \phi^{*0}) \\
&\quad= (\Phi_{11}, \Phi_{12})(\hat{\theta}^* - \theta^{*0}) + o_p(\hat{\theta}^* - \theta^{*0}) \qquad \text{(A.22)}
\end{aligned}
$$

where $\Phi_{11} = \Psi_{22}\Psi_{11}^{-1}\Psi_{12} - \Psi_{22}, \Phi_{12} = \Psi_{22}\Psi_{11}^{-1}\Psi_{13} - \Psi_{23}$.

According to (A.19), we have

$$
\begin{aligned}
0 &= \frac{1}{n}\sum_{i=1}^{n}\hat{V}^*\left(Y_i - W^{*T}_{(\mathcal{V})i}(\hat{\phi}^*)\hat{\gamma}^*_{(\mathcal{V})} - G^{*T}_{(\mathcal{C})i}\hat{\gamma}^*_{(\mathcal{C})1}\right) + o_p(\hat{\gamma}^*_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1}) + o_p(\hat{\phi}^* - \phi^{*0}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\hat{V}^*\Big(\epsilon_i + R^T(\phi^*)G^*_i - W^{*T}_{(\mathcal{V})i}(\phi^*)(\hat{\gamma}^*_{(\mathcal{V})} - \gamma^{*0}_{(\mathcal{V})}) \\
&\quad - [W^{*T}_{(\mathcal{V})i}(\hat{\phi}^*) - W^{*T}_{(\mathcal{V})i}(\phi^*)]\hat{\gamma}^*_{(\mathcal{V})} \\
&\quad - G^{*T}_{(\mathcal{C})i}(\hat{\gamma}^*_{(\mathcal{C})1} - \hat{\gamma}^{*0}_{(\mathcal{C})1})\Big) + o_p(\hat{\gamma}^*_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1}) + o_p(\hat{\phi}^* - \phi^{*0}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\hat{V}^*\Big(\epsilon_i + R^T(\phi^*)G^*_i - W^{*T}_{(\mathcal{V})i}[\Psi_{11} + o_p(1)]^{-1}(\Lambda_{10} - \Psi_{12}(\hat{\gamma}^*_{(\mathcal{C})1} - \hat{\gamma}^{*0}_{(\mathcal{C})1}) \\
&\quad - \Psi_{13}(\hat{\phi}^* - \phi^*)) - [W^{*T}_{(\mathcal{V})i}(\hat{\phi}^*) - W^{*T}_{(\mathcal{V})i}(\phi^*)]\hat{\gamma}^*_{(\mathcal{V})} - G_i^{*T}(\hat{\gamma}^*_{(\mathcal{C})1} - \hat{\gamma}^{*0}_{(\mathcal{C})1})\Big) \\
&\quad + o_p(\hat{\gamma}^*_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1}) + o_p(\hat{\phi}^* - \phi^*) \\
&= \frac{1}{n}\sum_{i=1}^{n}\hat{V}^*\Big(\epsilon_i + R^T(\phi^*)G^*_i - W^{*T}_{(\mathcal{V})i}[\Psi_{11} + o_p(1)]^{-1}\Lambda_{10}\Big) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\hat{V}^*W^{*T}_{(\mathcal{V})i}[\Psi_{11} + o_p(1)]^{-1}\Psi_{12}(\hat{\gamma}^*_{(\mathcal{C})1} - \hat{\gamma}^{*0}_{(\mathcal{C})1})
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{1}{n}\sum_{i=1}^{n} \hat{V}^{*} W^{*T}_{(\mathcal{V})i}[\Psi_{11} + o_p(1)]^{-1}\Psi_{13}(\hat{\phi}^{*} - \phi^{*0}) \\
&- \frac{1}{n}\sum_{i=1}^{n} \hat{V}^{*} [W^{*T}_{(\mathcal{V})i}(\hat{\phi}^{*}) - W^{*T}_{(\mathcal{V})i}(\phi^{*})]\hat{\gamma}^{*}_{(\mathcal{V})} \\
&- \frac{1}{n}\sum_{i=1}^{n} \hat{V}^{*} G^{*T}_{(\mathcal{C})i}(\hat{\gamma}^{*}_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1}) + o_p(\hat{\gamma}^{*}_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1}) + o_p(\hat{\phi}^{*} - \phi^{*0}) \\
\triangleq\ & \Delta_1 + \Delta_2 + \Delta_3 - \Delta_4 - \Delta_5 + + o_p(\hat{\gamma}^{*}_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1}) + o_p(\hat{\phi}^{*} - \phi^{*0})
\end{aligned}
$$

where $\hat{V}^{*} = \dot{W}^{*T}_{(\mathcal{V})i}(\hat{\phi}^{*})\hat{\gamma}^{*}_{(\mathcal{V})} J^{T}_{\hat{\phi}^{*}} X^{*}_{(\mathcal{V})i}$. For Δ_1, we have

$$
\begin{aligned}
\Delta_1 &= \frac{1}{n}\sum_{i=1}^{n} \hat{V}^{*} M_1 = \frac{1}{n}\sum_{i=1}^{n} V^{*} M_1 + \frac{1}{n}\sum_{i=1}^{n}[\dot{f}(\phi^{*})G^{*}_{i} - \dot{W}(\phi^{*})\gamma^{*}]J^{T}_{\hat{\phi}^{*}} X^{*}_{i} M_1 \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \dot{W}(\phi^{*})(\hat{\gamma}^{*} - \gamma^{*})J^{T}_{\hat{\phi}^{*}} X^{*}_{i} M_1 + \frac{1}{n}\sum_{i=1}^{n}[\dot{W}(\phi^{*}) - \dot{W}(\hat{\phi}^{*})]^{T} J^{T}_{\hat{\phi}^{*}} X^{*}_{i} M_1 \\
&=: \Delta_{11} + \Delta_{12} + \Delta_{13} + \Delta_{14}
\end{aligned}
$$

where $M_1 = \epsilon_i + R^{T}(\phi^{*})G^{*}_{i} - W^{*T}_{(\mathcal{V})i}(\phi^{*})\Psi^{-1}_{11}\Lambda_{10}$.

Note that

$$
\frac{1}{n}\sum_{i=1}^{n} \Psi_{13}\Psi^{-1}_{11} W^{*}_{(\mathcal{V})i}(\phi^{*})(\epsilon_i + R^{T}(\phi^{*})G^{*}_{i} - W^{*T}_{(\mathcal{V})i}(\phi^{*})\Psi^{-1}_{11}\Lambda_{10}) = 0
$$

$$
\frac{1}{n}\sum_{i=1}^{n}(V^{*}_{i} - \Psi^{T}_{13}\Psi^{-1}_{11} W^{*}_{(\mathcal{V})i}(\phi^{*}))W^{*T}_{(\mathcal{V})i}(\phi^{*}) = 0
$$

Then, we can show that

$$
\begin{aligned}
\Delta_{11} &= \frac{1}{n}\sum_{i=1}^{n}(V^{*}_{i} - \Psi^{T}_{13}\Psi^{-1}_{11} W^{*}_{(\mathcal{V})i}(\phi^{*}))\epsilon_i \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}(V^{*}_{i} - \Psi^{T}_{13}\Psi^{-1}_{11} W^{*}_{(\mathcal{V})i}(\phi^{*}))R(\phi^{*})G^{*}_{i} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}(V^{*}_{i} - \Psi^{T}_{13}\Psi^{-1}_{11} W^{*}_{(\mathcal{V})i}(\phi^{*}))W^{*T}_{(\mathcal{V})i}(\phi^{*})[\Psi_{11} + o_p(1)]^{-1} \\
&\quad + o_p(\hat{\gamma}^{*}_{(\mathcal{C})} - \gamma^{*0}_{(\mathcal{C})1}) \\
&= \frac{1}{n}\sum_{i=1}^{n}(V^{*}_{i} - \Psi^{T}_{13}\Psi^{-1}_{11} W^{*}_{(\mathcal{V})i}(\phi^{*}))\epsilon_i + o_p(\hat{\gamma}^{*}_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1}) + o_p(\hat{\phi}^{*} - \phi^{*0})
\end{aligned}
$$

Similar to [23], we can get $\Delta_{12} = o_p(\hat{\gamma}^{*}_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1}) + o_p(\hat{\phi}^{*} - \phi^{*0})$,

$\Delta_{13} = o_p(\hat{\gamma}^{*}_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1}) + o_p(\hat{\phi}^{*} - \phi^{*0}), \Delta_{14} = o_p(\hat{\gamma}^{*}_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1}) + o_p(\hat{\phi}^{*} - \phi^{*0})$.

Hence, we have

$$\Delta_1 = \frac{1}{n}\sum_{i=1}^{n}(V_i^* - \Psi_{13}^T\Psi_{11}^{-1}W_{(\mathcal{V})i}^*(\phi^*))\epsilon_i + o_p(\hat{\gamma}_{(\mathcal{C})1}^* - \gamma_{(\mathcal{C})1}^{*0}) + o_p(\hat{\phi}^* - \phi^{*0}) \quad \text{(A.23)}$$

For Δ_2, we have

$$\begin{aligned}\Delta_2 &= \frac{1}{n}\sum_{i=1}^{n}\hat{V}^* M_2 = \frac{1}{n}\sum_{i=1}^{n}V^* M_2 + \frac{1}{n}\sum_{i=1}^{n}[\dot{f}^T(\phi^*)G_i^T - \dot{W}(\phi^*)\gamma^*]J_{\hat{\phi}^*}^T X_i^* \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\dot{W}_i^{*T}(\hat{\gamma}^* - \gamma^*)J_{\hat{\phi}^*}^T X_i^* M_2 + \frac{1}{n}\sum_{i=1}^{n}[\dot{W}_i^*(\phi^*) - \dot{W}_i^*(\hat{\phi}^*)]J_{\hat{\phi}^*}^T X_i^* M_2 \\ &\triangleq \Delta_{21} + \Delta_{22} + \Delta_{23} + \Delta_{24}\end{aligned}$$

where $M_2 = W_{(\mathcal{V})i}^{*T}[\Psi_{11} + o_p(1)]^{-1}\Psi_{12}(\hat{\gamma}_{(\mathcal{C})1}^* - \hat{\gamma}_{(\mathcal{C})1}^{*0})$. Hence, we have

$$\Delta_{21} = \Psi_{13}^T\Psi_{11}^{-1}\Psi_{12}(\hat{\gamma}_{(\mathcal{C})1}^* - \hat{\gamma}_{(\mathcal{C})1}^{*0}) + o_p(\hat{\gamma}_{(\mathcal{C})1}^* - \hat{\gamma}_{(\mathcal{C})1}^{*0})$$

Similar arguments to that of J_{12}, we have

$$\Delta_{22} = o_p(\hat{\gamma}_{(\mathcal{C})1}^* - \hat{\gamma}_{(\mathcal{C})1}^{*0}), \Delta_{23} = o_p(\hat{\gamma}_{(\mathcal{C})1}^* - \hat{\gamma}_{(\mathcal{C})1}^{*0}), \text{ and } \Delta_{24} = o_p(\hat{\gamma}_{(\mathcal{C})1}^* - \hat{\gamma}_{(\mathcal{C})1}^{*0}).$$

Therefore, we have

$$\Delta_2 = \Psi_{13}^T\Psi_{11}^{-1}\Psi_{12}(\hat{\gamma}_{(\mathcal{C})1}^* - \hat{\gamma}_{(\mathcal{C})1}^{*0}) + o_p(\hat{\gamma}_{(\mathcal{C})1}^* - \hat{\gamma}_{(\mathcal{C})1}^{*0}). \quad \text{(A.24)}$$

Similarly, we have

$$\Delta_3 = \Psi_{13}^T\Psi_{11}^{-1}\Psi_{13}(\hat{\phi}^* - \phi^{*0}) + o_p(\hat{\phi}^* - \phi^{*0}). \quad \text{(A.25)}$$

Now we consider Δ_4, applying Taylor expansion, we have

$$\begin{aligned}\Delta_4 &= \frac{1}{n}\sum_{i=1}^{n}\hat{V}^*[W_{(\mathcal{V})i}^{*T}(\hat{\phi}^*) - W_{(\mathcal{V})i}^{*T}(\phi^{*0})]\hat{\gamma}_{(\mathcal{V})}^* \\ &= \frac{1}{n}\sum_{i=1}^{n}\hat{V}^*[\dot{W}_i^{*T}(\phi^{*0})\hat{\gamma}_{(\mathcal{V})i}J_{\phi^*}^T X_i^{*T}(\hat{\phi}^* - \phi^{*0}) + o_p(\hat{\phi}^* - \phi^{*0})] \\ &= \frac{1}{n}\sum_{i=1}^{n}\hat{V}^*[V^{*T}(\hat{\phi}^* - \phi^{*0})) + o_p(\hat{\phi}^* - \phi^{*0})] \\ &= \frac{1}{n}\sum_{i=1}^{n}V^* V^{*T}(\hat{\phi}^* - \phi^{*0}) + \frac{1}{n}\sum_{i=1}^{n}V^*(\hat{\phi}^* - \phi^{*0})\dot{W}_i^{*T}(\phi^*)(\hat{\gamma}_{(\mathcal{V})}^* - \gamma_{(\mathcal{V})}^*)J_{\hat{\phi}^*}^T X_i^* \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}V^{*T}(\hat{\phi}^* - \phi^{*0})[\dot{f}(\phi^*)G_i^* - \dot{W}_i^{*T}(\phi^*)\gamma_{(\mathcal{V})}^*]J_{\hat{\phi}^*}^T X_i^* \\ &= \frac{1}{n}\sum_{i=1}^{n}V^* V^{*T}(\hat{\phi}^* - \phi^{*0}) + o_p(\hat{\phi}^* - \phi^{*0})\end{aligned}$$

$$= \Psi_{21}(\hat{\phi}^* - \phi^{*0}) + o_p(\hat{\phi}^* - \phi^{*0}).$$

Similarly, we have

$$\Delta_5 = \Psi_{23}^T(\hat{\gamma}^*_{(\mathcal{C})1} - \gamma^*_{(\mathcal{C})1}) + o_p(\hat{\gamma}^*_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1}) \tag{A.26}$$

So we can get

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(V_i^* - \Psi_{13}^T\Psi_{11}^{-1}W^*_{(\mathcal{V})i}(\hat{\phi}^*))\epsilon_i & \text{(A.27)}\\
&= \Psi_{23}^T(\hat{\gamma}^*_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1}) + \Psi_{21}^T(\hat{\phi}^* - \phi^{*0}) - \Psi_{13}^T\Psi_{11}^{-1}\Psi_{12}(\hat{\gamma}^*_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1})\\
&\quad - \Psi_{13}^T\Psi_{11}^{-1}\Psi_{13}(\hat{\phi}^* - \phi^{*0}) + o_p(\hat{\gamma}^*_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1}) + o_p(\hat{\phi}^* - \phi^{*0})\\
&= (\Psi_{21}^T - \Psi_{13}^T\Psi_{11}\Psi_{13})(\hat{\phi}^* - \phi^{*0}) + (\Psi_{23}^T - \Psi_{13}^T\Psi_{11}\Psi_{12})(\hat{\gamma}^*_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1})\\
&\quad + o_p(\hat{\gamma}^*_{(\mathcal{C})1} - \gamma^{*0}_{(\mathcal{C})1}) + o_p(\hat{\phi}^* - \phi^{*0})\\
&= (\Phi_{21}, \Phi_{22})(\hat{\theta}^* - \theta^{*0}) + o_p(\hat{\theta}^* - \theta^{*0}). & \text{(A.28)}
\end{aligned}$$

where $\Phi_{21} = \Psi_{23}^T - \Psi_{13}^T\Psi_{11}\Psi_{12}$ and $\Phi_{22} = \Psi_{21}^T - \Psi_{13}^T\Psi_{11}\Psi_{13}$.

According to (A.22) and (A.27), we have

$$\begin{aligned}
\sqrt{n}(\hat{\theta}^* - \theta^{*0}) = \begin{pmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{pmatrix}^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix} G^*_{(\mathcal{C})i} - \Psi_{22}\Psi_{11}^{-1}W^*_{(\mathcal{V})i}(\phi^*) \\ \mathrm{V}_i^* - \Psi_{13}^T\Psi_{11}^{-1}W^*_{(\mathcal{V})i}(\phi^*) \end{pmatrix}\epsilon_i&\\
+ o_p(1). & \quad \text{(A.29)}
\end{aligned}$$

By the central limit theorem and Slutsky's theorem, we can see that $\hat{\theta}^*$ is consistent and has asymptotic normality.

It follows from (2.8) that

$$\hat{\beta}^* - \beta_0^* = J_{\phi^{*0}}(\hat{\phi}^* - \phi^{*0}) + O_p(n^{-1}).$$

Hence, we can get

$$\sqrt{n}(\hat{\vartheta}^* - \vartheta_0^*) = \begin{pmatrix} 1 & 0 \\ 0 & J_{\phi^{*0}} \end{pmatrix}\sqrt{n}(\hat{\theta} - \theta^{*0}).$$

Therefore, we can get the asymptotic covariance matrix Σ as

$$\Sigma = \begin{pmatrix} \Sigma_1^{-1} & 0 \\ 0 & J_{\phi^{*0}}\Sigma_2^{-1}J_{\phi^{*0}}^T \end{pmatrix}$$

Then, the proof of Theorem 3.3 is completed. □

Acknowledgments

The authors thank the Editor and referees for their insightful and constructive comments that greatly improved the manuscript.

References

[1] Falconer, D. S. (1952). The Problem of Environment and Selection. *Am. Natural.* **86**: 293–299.

[2] Ma, S., Yang, L., Romero, R., and Cui, Y. (2011). Varying coefficient model for gene-environment interaction: a non-linear look. *Bioinformatics*, **27**: 2119–2126.

[3] Wu, C. and Cui, Y. (2013). A novel method for identifying nonlinear gene-environment interactions in case-control association studies. *Human Genetics*, **132**: 1413–1425.

[4] Liu, X., Cui, Y., and Li, R. (2016). Partial linear varying multi-index coefficient model for integrative gene-environment interactions. *Statistica Sinica*, **26**: 1037–1060. MR3559942

[5] Fan, J. Q., Yao, Q. W., and Cai, Z. W. (2003). Adaptive varying-coefficient linear models. *J. R. Stat. Soc. B*, **65**: 57–80. MR1959093

[6] Frank, L. E. and Friedman, J. H., (1993). A statistical view of some chemometrics regression tools. *Technometrics*, **35**(2): 109–135. MR1700749

[7] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, **58**(1): 267–288. MR2815776

[8] Zou, H.(2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**(476): 1418–1429. MR2279469

[9] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, **96**(456): 1348–1360. MR1946581

[10] Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, **38**(2): 894–942. MR2604701

[11] Naik, P.A. (2001). Single-index model selections. *Biometrika*, **88**(3): 821–832. MR1859412

[12] Naik, P.A. and Tsai C. L. (2004). Residual information criterion for single-index model selections. *Journal of Nonparametric Statistics*, **16**(1-2): 187–195. MR2053069

[13] Wang, H.B. (2009). Bayesian estimation and variable selection for single index models. *Computational Statistics & Data Analysis*, **53**(7): 2617–2627. MR2665912

[14] Peng, H. and Huang, T. (2011). Penalized least squares for single index models. *J. Statist. Plann. Inference*, **141** (4): 1362–1379. MR2747907

[15] Zeng, P., He, T. and Zhu, Y. (2012). A lasso-type approach for estimation and variable selection in single index models. *J. Comput. Graph. Statist.*, **21** (1): 92–109. MR2913358

[16] Li, J., Li, Y. and Zhang, R. (2017). B spline variable selection for the single index models. *Statist. Papers*, **58** (3): 691–706. MR3686846

[17] Luo, S., Ghosal, S(2016). Forward selection and estimation in high dimensional single index models. *Stat. Methodol*, 33: 172–179. MR3582782

[18] Cheng, L., Zeng, P. and Zhu, Y. (2017). BS-SIM: an effective variable selection method for high-dimensional single index model. *Electron. J.*

Stat., **11**(2): 3522–3548. MR3709862

[19] Zhang, J., Wang, X., Yu, Y. and Gai, Y. (2014). Estimation and variable selection in partial linear single index models with error-prone linear covariates. *Statistics*, **48**(5): 1048–1070. MR3259875

[20] Li, G., Lai, P. and Lian, H. (2015). Variable selection and estimation for partially linear single-index models with longitudinal data. *Stat. Comput.*, **(**3): 579–593. MR3334418

[21] Wang, W. and Zhu, Z.(2017). Variable selection for the partial linear single-index model. *Acta Math. Appl. Sin. Engl. Ser.*, **33**(2): 373–388. MR3646992

[22] Yu, Y., Zou, Z. and Wang, S. (2019). Bayesian quantile regression and variable selection for partial linear single-index model: using free knot spline. *Comm. Statist. Simulation Comput.*, **48**(5): 1429–1449. MR3945337

[23] Feng, S. and Xue, L. (2013). Variable selection for single-index varying-coefficient model. *Frontiers of Mathematics in China*, **8**(3): 541–565. MR3044669

[24] Tang, Y., Wang, H. J., Zhu, Z. and Song, X. (2012). A unified variable selection approach for varying coefficient models. *Statistica Sinica*, **7**: 601–628. MR2954354

[25] Wu, C., Zhong, P. S. and Cui, Y. (2018). Additive varying-coefficient model for nonlinear gene-environment interactions. *Statistical Applications in Genetics and Molecular Biology*, **17**(2): 1–18. MR3797928

[26] Schumaker, L. (2007). *Spline functions: basic theory.* Cambridge University Press. MR2348176

[27] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**(2): 461–464. MR0468014

[28] He, X., Wing K. F. and Zhu Z. Y. (2005). Robust estimation in generalized partial linear models for clustered data. *Journal of the American Statistical Association*, **100**(472): 1176–1184. MR2236433

[29] Schadt, E. E., Molony, C., ... and Zhu, J. (2008). Mapping the genetic architecture of gene expression in human liver. *PLoS Biology*, **6**(5): 107–118.

[30] Yang, X., Zhang, B., ...and Guengerich, F. P. (2010). Systematic genetic and genomic analysis of cytochrome P450 enzyme activities in human liver. *Genome Research*, **20**(8): 1020–1036.

[31] Echchgadda, I., Song, C.S.,... and Chatterjee, B. (2004). Gene regulation for the senescence marker protein DHEA-sulfotransferase by the xenobiotic-activated nuclear pregnane X receptor (PXR). *Mechanisms of Ageing and Development*, **125**(10-11): 733–745.

[32] Ma, S. and Song, P. X. K. (2015). Varying index coefficient models. *Journal of the American Statistical Association*, **110**(509): 341–356. MR3338507

[33] Zhao, P. and Xue, L. (2010). Variable selection for semiparametric varying coefficient partially linear errors in variables models. *Journal of Multivariate Analysis*, **101**(8): 1872–1883. MR2651962