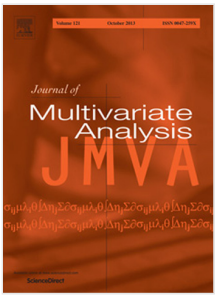

Profile forward regression screening for ultra-high dimensional semiparametric varying coefficient partially linear models

Yujie Li, Gaorong Li, Heng Lian, Tiejun Tong

PII: S0047-259X(16)30261-5
DOI: http://dx.doi.org/10.1016/j.jmva.2016.12.006
Reference: YJMVA 4200

To appear in: *Journal of Multivariate Analysis*

Received date: 23 March 2016

Please cite this article as: Y. Li, G. Li, H. Lian, T. Tong, Profile forward regression screening for ultra-high dimensional semiparametric varying coefficient partially linear models, *Journal of Multivariate Analysis* (2016), http://dx.doi.org/10.1016/j.jmva.2016.12.006

Profile forward regression screening for ultra-high dimensional semiparametric varying coefficient partially linear models

Yujie Li[a], Gaorong Li[*,b], Heng Lian[c], Tiejun Tong[d]

[a] *College of Applied Sciences, Beijing University of Technology, Beijing 100124, P. R. China*
[b] *Beijing Institute for Scientific and Engineering Computing, Beijing University of Technology, Beijing 100124, P. R. China*
[c] *Department of Mathematics, City University of Hong Kong, Hong Kong*
[d] *Department of Mathematics, Hong Kong Baptist University, Hong Kong*

Abstract

In this paper, we consider semiparametric varying coefficient partially linear models when the predictor variables of the linear part are ultra-high dimensional where the dimensionality grows exponentially with the sample size. We propose a profile forward regression (PFR) method to perform variable screening for ultra-high dimensional linear predictor variables. The proposed PFR algorithm can not only identify all relevant predictors consistently even for ultra-high semiparametric models including both nonparametric and parametric parts, but also possesses the screening consistency property. To determine whether or not to include the candidate predictor in the model of selected ones, we adopt an extended Bayesian information criterion (EBIC) to select the "best" candidate model. Simulation studies and a real data example are also carried out to assess the performance of the proposed method and to compare it with existing screening methods.

Key words: Varying coefficient partially linear model; profile forward regression; variable screening; screening consistency property; ultra-high dimension; EBIC

AMS 2010 subject classifications: primary 62G08; secondary 62J02

1. Introduction

In recent years, high-dimensional data analysis has become increasingly frequent and important in a large variety of areas such as health sciences, economics, finance, and epidemiology. The analysis of high-dimensional data poses many challenges for statisticians and thus calls for new statistical methodologies as well as theories; see Fan and Li [11].

To address these challenges, variable screening is an effective method of using a ranking criterion to select significant variables, particularly for statistical models with nonpolynomial dimensionality or "large p, small n" paradigms when p can be as large as an exponential of the sample size n; see Li et al. [24]. The main idea is to first apply a fast, reliable and efficient method to reduce the ultra-high dimensionality p from a large or huge

*Corresponding author. *E-mail address:* ligaorong@bjut.edu.cn (G. R. Li).
E-mail addresses: liyujie0207@emails.bjut.edu.cn (Yujie Li), henglian@cityu.edu.hk (Heng Lian), tongt@hkbu.edu.hk (Tiejun Tong)

scale to a relatively large scale s (say $s \leq n$), and then apply some well-developed variable selection techniques to perform the final variable selection and parameter estimation simultaneously. In the first screening step, the sure screening property introduced by Fan and Lv [12] needs to be satisfied such that all truly important predictors can be selected with probability tending to 1 as the sample size goes to infinity.

Since Fan and Lv [12] proposed the sure independence screening (SIS) procedure for ultra-high linear models, many authors had further developed the SIS method and applied it to various statistical models. For illustration, Fan et al. [14] and Fan and Song [15] extended SIS to generalized linear models. Wang [35] proposed a forward regression algorithm for ultra-high dimensional variable screening. Fan et al. [8] studied nonparametric independence screening (NIS) in sparse ultra-high dimensional additive models. Cui et al. [6] and Zhu et al. [23] proposed model-free variable screening methods, respectively. Li et al. [24, 25] proposed a robust rank correlation screening (RRCS) procedure based on Kendall's rank correlation coefficient. Li et al. [29] developed the sure independence screening procedure based on the distance correlation (DC-SIS) under general parametric models. Fan et al. [13] and Liu et al. [32] extended the NIS to sparse ultra-high dimensional varying coefficient models.

In this paper, we propose a new method for variable screening in the ultra-high dimensional semiparametric varying coefficient partially linear model (VCPLM). Suppose that Y is a response variable and $(U, \boldsymbol{X}^\top, \boldsymbol{Z}^\top)$ are the associated covariates, where $^\top$ denotes transposition. The VCPLM takes the form

$$Y = \boldsymbol{X}^\top \boldsymbol{\alpha}(U) + \boldsymbol{Z}^\top \boldsymbol{\beta} + \varepsilon, \tag{1.1}$$

where $\boldsymbol{\alpha}(\cdot) = (\alpha_1(\cdot), \ldots, \alpha_q(\cdot))^\top$ is a q-dimensional vector of unknown regression functions, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is a p-dimensional vector of unknown regression coefficients, ε is independent of $(U, \boldsymbol{X}^\top, \boldsymbol{Z}^\top)$ and follows a distribution with mean 0 and variance σ^2, and U is a univariate variable on the compact support Ω. From prior knowledge, we assume that some of the true predictors may have varying effects while the others have constant effects to the response variable. We further assume that the predictor variable $\boldsymbol{X}$ has fixed dimension q, while the predictor variable $\boldsymbol{Z}$ has ultra-high dimensionality or nonpolynomial dimensionality such that $\ln p = O(n^\kappa)$ for some $\kappa > 0$.

Model (1.1) retains the flexibility of the nonparametric regression model and has also the nice interpretability of the linear regression model. When the dimension p is fixed, model (1.1) has been extensively studied in the literature including, e.g., Ahmad et al. [1], Fan and Huang [10], Li et al. [21], Li et al. [22], Li et al. [27], Li and Liang [28], Wu et al. [37], Xia et al. [38], Xue and Zhu [39], You and Chen [40], Zhang et al. [42], Zhou and Liang [46]. When the dimension p grows with the sample size n, Lam and Fan [20] considered a generalized varying coefficient partially linear model, and studied the asymptotic properties of the profile likelihood estimator. Li et al. [23] proposed the bias-corrected empirical likelihood method to study the VCPLM with a diverging number of parameters.

Variable selection for Model (1.1) is challenging because it involves both nonparametric and parametric parts. We note that penalized variable selection methods have been successfully applied to Model (1.1) when $p < n$, such as Hong et al. [17], Kai et al. [19], Li et al. [26], Wang et al. [36], Zhao and Xue [43], and Zhao et al. [44]. Nevertheless, when $p > n$ or even grows exponentially with n, the aforementioned penalized variable selection methods may not work for the ultra-high dimensional VCPLM (1.1) due to the

simultaneous challenges of computational expediency, statistical accuracy and algorithm stability.

To cater for the demand, an alternative popular and classical variable screening method, namely the forward regression (FR) approach, has recently been proposed for ultra-high dimensional linear regression models in Wang [35]. Zhong et al. [45] proposed a stepwise procedure, correlation pursuit, for variable selection and screening under the sufficient dimension reduction framework. Cheng et al. [4] extended the FR algorithm of variable screening to sparse ultra-high dimensional varying coefficient models. Liang et al. [31] proposed the profile forward regression (PFR) algorithm of variable screening to ultra-high dimensional semiparametric partially linear models. Such methods enjoy desirable theoretical properties, including the screening consistency property, and have advantages from numerical aspects. Inspired by these advantages, we develop a new PFR algorithm of variable screening for the semiparametric varying coefficient partially linear models with ultra-high dimensional covariate for the linear part, where the dimension can be much larger than the sample size. The proposed PFR method can not only identify all relevant predictors consistently even for ultra-high semiparametric models including both nonparametric and parametric parts, but also possesses the screening consistency property.

For ease of notation, we use the boldface roman $\mathbf{B}$ to represent a matrix, boldface italics $\boldsymbol{B}$ to represent a vector, and B_{ik} to represent the (i,k)th entry of the matrix $\mathbf{B}$ throughout this paper. The remainder of this paper is organized as follows. In Section 2, the PFR procedure of variable screening is introduced. In Section 3, the asymptotic properties are derived under some regularity conditions. In Section 4, simulation studies are carried out to assess the performance of the proposed method and to compare it with existing methods. A real data example is used for illustration in Section 5. The technical proofs of the main results and some lemmas are given in the Appendix.

2. Profile forward regression method

We first present the main profile idea in the population form. Assuming that $\boldsymbol{\beta}$ is known, Model (1.1) becomes the following varying coefficient model:

$$Y - \boldsymbol{Z}^\top\boldsymbol{\beta} = \boldsymbol{X}^\top\boldsymbol{\alpha}(U) + \varepsilon.$$

For any given U, we can solve the following profile estimation equation:

$$\mathrm{E}[\boldsymbol{X}\{Y - \boldsymbol{Z}^\top\boldsymbol{\beta} - \boldsymbol{X}^\top\boldsymbol{\alpha}(U)\}|U] = 0.$$

For simplicity, let

$$\boldsymbol{\eta}(U) = \{\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^\top|U)\}^{-1}\mathrm{E}(\boldsymbol{X}Y|U), \quad \boldsymbol{\mu}(U) = \{\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^\top|U)\}^{-1}\mathrm{E}(\boldsymbol{X}\boldsymbol{Z}^\top|U).$$

Then

$$\begin{aligned}\boldsymbol{\alpha}(U,\boldsymbol{\beta}) &= \{\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^\top|U)\}^{-1}\mathrm{E}(\boldsymbol{X}Y|U) - \{\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^\top|U)\}^{-1}\mathrm{E}(\boldsymbol{X}\boldsymbol{Z}^\top|U)\boldsymbol{\beta} \\ &= \boldsymbol{\eta}(U) - \boldsymbol{\mu}(U)\boldsymbol{\beta}. \end{aligned} \tag{2.1}$$

Replacing $\boldsymbol{\alpha}(U)$ in Model (1.1) by (2.1), we get

$$Y - \boldsymbol{X}^\top\boldsymbol{\eta}(U) = \{\boldsymbol{Z} - \boldsymbol{\mu}^\top(U)\boldsymbol{X}\}^\top\boldsymbol{\beta} + \varepsilon. \tag{2.2}$$

Note that Model (2.2) reduces to the linear model with the unknown nonparametric functions $\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^\top|U)$, $\mathrm{E}(\boldsymbol{X}Y|U)$ and $\mathrm{E}(\boldsymbol{X}\boldsymbol{Z}^\top|U)$, which can be estimated consistently by using the kernel smoothing method, respectively. When the dimension p of the linear component $\boldsymbol{\beta}$ is fixed, many authors have shown that the profile least squares estimator of $\boldsymbol{\beta}$ is semiparametrically efficient in large samples; see, e.g., Theorem 4.1 in Fan and Huang [10]. For this reason, Li et al. [23] showed that $\boldsymbol{\mu}^\top(U)\boldsymbol{X}$ is the projection of $\boldsymbol{Z}$ onto the space spanned by $\boldsymbol{X}$, and $\boldsymbol{Z}-\boldsymbol{\mu}^\top(U)\boldsymbol{X}$ is orthogonal to $\boldsymbol{X}^\top$ for any given U. That is,

$$\mathrm{E}[\{\boldsymbol{Z}-\boldsymbol{\mu}^\top(U)\boldsymbol{X}\}\boldsymbol{X}^\top|U]=0.$$

In other words, this orthogonality will play a key role for the semiparametric efficiency (see Fan and Huang [10]) and for the asymptotic normality of the bias-corrected empirical log-likelihood ratio (see Li et al. [23]).

Let $\{(Y_i;\boldsymbol{X}_i^\top,\boldsymbol{Z}_i^\top,U_i):1\le i\le n\}$ be an independent and identically distributed random sample from Model (1.1) with predictor variable $\boldsymbol{X}_i$ having the fixed dimension q and predictor variable $\boldsymbol{Z}_i$ having the ultra-high dimension $p\gg n$ as $n\to\infty$. For ease of notation, let $\boldsymbol{X}_i=(X_{i1},\ldots,X_{iq})^\top\in\mathbb{R}^q$ and $\boldsymbol{Z}_i=(Z_{i1},\ldots,Z_{ip})^\top\in\mathbb{R}^p$ be the predictor variables. Define $\boldsymbol{Y}=(Y_1,\ldots,Y_n)^\top\in\mathbb{R}^n$ as the response vector, $\mathbf{X}=(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)^\top\in\mathbb{R}^{n\times q}$ and $\mathbf{Z}=(\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_n)^\top\in\mathbb{R}^{n\times p}$ as two matrices of explanatory variables, and $\boldsymbol{\varepsilon}=(\varepsilon_1,\ldots,\varepsilon_n)^\top$ as the vector of random error.

In order to perform variable screening for the linear part conveniently, we take Z_{ij} as a relevant predictor variable if $\beta_j\neq 0$; otherwise we refer to Z_{ij} as an irrelevant predictor variable if $\beta_j=0$. Let $\mathcal{M}=\{j_1,\ldots,j_{p*}\}$ denote an arbitrary model with $Z_{ij_1},\ldots,Z_{ij_{p*}}$ as relevant predictors, and let $\mathcal{M}_{\mathcal{F}}=\{1,\ldots,p\}$ and $\mathcal{M}_{\mathcal{T}}=\{j:\beta_j\neq 0\}$ represent the full model and the true model, respectively. In this paper, we use $|\mathcal{M}|$ to denote the size of model $\mathcal{M}$. Thus, $|\mathcal{M}_{\mathcal{F}}|=p$ and $|\mathcal{M}_{\mathcal{T}}|=p_0$, where p_0 is the size of the true model or the number of relevant predictors in the true model. For any candidate model $\mathcal{M}$, we use $\boldsymbol{Z}_{i(\mathcal{M})}=\{Z_{ij}:j\in\mathcal{M}\}$ to represent the subvector of $\boldsymbol{Z}_i$ corresponding to $\mathcal{M}$, and $\mathbf{Z}_{(\mathcal{M})}=\{Z_{ij}:i=1,\ldots,n,j\in\mathcal{M}\}$ to denote the matrix consisting of the column of $\mathbf{Z}$ with indices in $\mathcal{M}$. Similarly, let $\boldsymbol{\beta}_{\mathcal{M}}$ denote the vector consisting of the corresponding components of $\boldsymbol{\beta}$.

In the sample form, Model (2.2) can be written as

$$Y_i-\boldsymbol{X}_i^\top\boldsymbol{\eta}(U_i)=\{\boldsymbol{Z}_i-\boldsymbol{\mu}^\top(U_i)\boldsymbol{X}_i\}^\top\boldsymbol{\beta}+\varepsilon_i,\tag{2.3}$$

where

$$\boldsymbol{\eta}(U_i)=\{\mathrm{E}(\boldsymbol{X}_i\boldsymbol{X}_i^\top|U_i)\}^{-1}\mathrm{E}(\boldsymbol{X}_iY_i|U_i),\quad \boldsymbol{\mu}(U_i)=\{\mathrm{E}(\boldsymbol{X}_i\boldsymbol{X}_i^\top|U_i)\}^{-1}\mathrm{E}(\boldsymbol{X}_i\boldsymbol{Z}_i^\top|U_i).$$

The functions $\boldsymbol{\eta}(U_i)$ and $\boldsymbol{\mu}(U_i)$ contain the unknown nonparametric functions $\mathrm{E}(\boldsymbol{X}_i\boldsymbol{X}_i^\top|U_i)$, $\mathrm{E}(\boldsymbol{X}_iY_i|U)$ and $\mathrm{E}(\boldsymbol{X}_i\boldsymbol{Z}_i^\top|U_i)$, which need to be estimated by some nonparametric smoothing methods. For convenience, we define the following notations:

$$\mathbf{D}_u=\begin{pmatrix}\boldsymbol{X}_1^\top & \dfrac{U_1-u}{h}\boldsymbol{X}_1^\top\\ \vdots & \vdots\\ \boldsymbol{X}_n^\top & \dfrac{U_n-u}{h}\boldsymbol{X}_n^\top\end{pmatrix},\qquad \mathbf{W}_u=\mathrm{diag}\{K_h(U_1-u),\ldots,K_h(U_n-u)\},$$

where $K_h(\cdot) = K(\cdot/h)/h, K(\cdot)$ is a kernel function and h is the bandwidth. We further define the smoothing matrix by

$$\mathbf{S} = \begin{pmatrix} (\boldsymbol{X}_1^\top \ \ \mathbf{0}^\top)(\mathbf{D}_{u_1}^\top \mathbf{W}_{u_1} \mathbf{D}_{u_1})^{-1} \mathbf{D}_{u_1}^\top \mathbf{W}_{u_1} \\ \vdots \\ (\boldsymbol{X}_n^\top \ \ \mathbf{0}^\top)(\mathbf{D}_{u_n}^\top \mathbf{W}_{u_n} \mathbf{D}_{u_n})^{-1} \mathbf{D}_{u_n}^\top \mathbf{W}_{u_n} \end{pmatrix}, \tag{2.4}$$

where $\mathbf{0}$ is a q-dimensional vector of zeros. It is easy to see that the smoothing matrix $\mathbf{S}$ depends only on the observations $\{(U_i, \boldsymbol{X}_i) : 1 \le i \le n\}$. Similar to the results in Fan and Huang [10] and Li et al. [23], $\boldsymbol{X}_i^\top \boldsymbol{\eta}(U_i)$ and $\boldsymbol{\mu}^\top(U_i)\boldsymbol{X}_i$ can be directly estimated by, respectively,

$$\boldsymbol{X}_i^\top \widehat{\boldsymbol{\eta}}(U_i) = \sum_{k=1}^{n} S_{ik} Y_k, \qquad \widehat{\boldsymbol{\mu}}^\top(U_i)\boldsymbol{X}_i = \sum_{k=1}^{n} S_{ik} \boldsymbol{Z}_k, \tag{2.5}$$

where S_{ik} is the (i,k)th entry of the smoothing matrix $\mathbf{S}$. To facilitate the notation, we denote $\widehat{\boldsymbol{Y}} = (\mathbf{I}_n - \mathbf{S})\boldsymbol{Y} = (\widehat{Y}_1, \ldots, \widehat{Y}_n)^\top$, $\widehat{\mathbf{Z}} = (\mathbf{I}_n - \mathbf{S})\mathbf{Z} = (\widehat{\boldsymbol{Z}}_1, \ldots, \widehat{\boldsymbol{Z}}_n)^\top$, where $\mathbf{I}_n$ is an $n \times n$ identity matrix. This leads to the linear model as

$$\widehat{Y}_i \approx \widehat{\boldsymbol{Z}}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.6}$$

or, in matrix form, $\widehat{\boldsymbol{Y}} \approx \widehat{\mathbf{Z}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. For the ultra-high dimensional semiparametric varying coefficient partially linear Model (1.1), we have transformed the model into the ultra-high dimensional linear model (2.6) by using the profile technique. Thus, we can apply the forward regression method in Wang [35] to identify all relevant predictor variables in the linear part of model (1.1). The proposed algorithm is as follows.

(1) Initially we specify a null model $\mathcal{M}^{(0)}$, which can be taken as $\mathcal{M}^{(0)} = \emptyset$.

(2) (Profile forward regression screening)

a) (Evaluation). In Step k ($k \ge 1$), the model $\mathcal{M}^{(k-1)}$ is given based on a priori knowledge. Then, for every $j \in \mathcal{M}_{\mathcal{F}}/\mathcal{M}^{(k-1)}$, construct a candidate model as $\mathcal{M}_j^{(k-1)} = \mathcal{M}^{(k-1)} \bigcup \{j\}$, whose lack of fit can be quantified by

$$\mathrm{RSS}_j^{(k-1)} = \widehat{\boldsymbol{Y}}^\top \Big\{ \mathbf{I}_n - \mathbf{H}_{(\mathcal{M}_j^{(k-1)})} \Big\} \widehat{\boldsymbol{Y}},$$

where

$$\mathbf{H}_{(\mathcal{M}_j^{(k-1)})} = \widehat{\mathbf{Z}}_{(\mathcal{M}_j^{(k-1)})} \Big\{ \widehat{\mathbf{Z}}_{(\mathcal{M}_j^{(k-1)})}^\top \widehat{\mathbf{Z}}_{(\mathcal{M}_j^{(k-1)})} \Big\}^{-1} \widehat{\mathbf{Z}}_{(\mathcal{M}_j^{(k-1)})}^\top.$$

b) (Screening). Find

$$a_k = \arg\min_{j \in \mathcal{M}_{\mathcal{F}}/\mathcal{M}^{(k-1)}} \mathrm{RSS}_j^{(k-1)} \tag{2.7}$$

and update the candidate model as $\mathcal{M}^{(k)} = \mathcal{M}^{(k-1)} \bigcup \{a_k\}$.

(3) (Solution path). Iterate Step 2 for n times, then a total of n nested candidate models are obtained by the solution path $\mathbb{S} = \{\mathcal{M}^{(k)} : 1 \le k \le n\}$, where $\mathcal{M}^{(k)} = \{a_1, \ldots, a_k\}$.

3. Theoretical properties

Before we derive the theoretical properties, we present some regularity conditions. Throughout the paper, we denote $\gamma_{\min}(\mathbf{A})$ and $\gamma_{\max}(\mathbf{A})$ as the smallest and largest eigenvalues of an arbitrary positive definite matrix $\mathbf{A}$, respectively. We define the profile response and the profile predictor as $Y^* = Y - \boldsymbol{X}^\top \boldsymbol{\eta}(U)$ and $\boldsymbol{Z}^* = \boldsymbol{Z} - \boldsymbol{\mu}^\top(U)\boldsymbol{X} = (Z_1^*, \ldots, Z_p^*)^\top$, respectively.

(C1) The random variable U has a compact support Ω. The density function $f(u)$ of U has a continuous second derivative and is uniformly bounded away from zero and infinity.

(C2) The $q \times q$ matrix $\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^\top|U)$ is non-singular for each $U \in \Omega$. $\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^\top|U)$, $\mathrm{E}(\boldsymbol{X}\boldsymbol{Z}^\top|U)$ and $\{\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^\top|U)\}^{-1}$ are functions about U and all Lipschitz continuous. Further, assume that all the elements of $\{\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^\top|U)\}^{-1}$ and $\mathrm{E}(\boldsymbol{X}\boldsymbol{Z}^\top|U)$ are bounded.

(C3) $\alpha_1(\cdot), \ldots, \alpha_q(\cdot)$ have continuous second derivatives in $u \in \Omega$.

(C4) The kernel $K(\cdot)$ is a bounded symmetric density function with bounded support.

(C5) The bandwidth h satisfies that $nh^6 \to 0$ and $nh^3/(\ln n)^3 \to \infty$.

(C6) Assume that Σ is the covariance matrix of the profile predictor $\boldsymbol{Z}^*$, and is a positive definite matrix. There exist two positive constants $\tau_{\min}$ and $\tau_{\max}$ satisfying $0 < \tau_{\min} < \tau_{\max} < \infty$, such that $2\tau_{\min} < \gamma_{\min}(\Sigma) \le \gamma_{\max}(\Sigma) < 2^{-1}\tau_{\max}$.

(C7) Assume that $\|\boldsymbol{\beta}\| \le C_{\boldsymbol{\beta}}$ for some constant $C_{\boldsymbol{\beta}} > 0$ and $\beta_{\min} \ge \nu_{\boldsymbol{\beta}} n^{-\xi_{\min}}$ for some $\xi_{\min} > 0$ and $\nu_{\boldsymbol{\beta}} > 0$, where $\|\cdot\|$ denotes the standard L_2 norm and $\beta_{\min} = \min_{j \in \mathcal{M}_{\mathcal{T}}} |\beta_j|$.

(C8) (Divergence speed of p and p_0) There exist positive constants ξ, ξ_0 and ν, such that $\ln p \le \min(\nu n^{\xi}, n^{3/10})$, $p_0 \le \nu n^{\xi_0}$, and $\xi + 6\xi_0 + 12\xi_{\min} < 1$.

(C9) (Moment constraint) Assume that $\max_{0 \le j \le p} \mathrm{E}\{\exp(u|Z_j^*|)\} < \infty$ for all $0 \le u \le t_0/\sigma_v$, where t_0 and σ_v are positive constants, and the moment generating functions $M_j(u)$ of Z_j^* for $j = 0, \ldots, p$ satisfy

$$\max_{0 \le j \le p} \sup_{0 \le u \le t_0} \left| \frac{d^3}{du^3} \ln\{M_j(u)\} \right| < \infty.$$

Further, assume that $\max_{0 \le j \le p} \mathrm{E}|Z_j^*|^{2k} \le \sigma_v^2$ for some $k > 2$, and assume that ε follows a normal distribution.

Note that the above conditions are assumed to be held uniformly in $u \in \Omega$. Conditions (C1)–(C4) are common in semiparametric varying coefficient partially linear models. These conditions are mild and can be easily satisfied; see Fan and Huang [10], Li et al. [23], and You and Zhou [41]. Condition (C5) was used in Li et al. [23], and the range from $O(n^{-1/3} \ln n)$ to $O(n^{-1/6})$ includes the optimal bandwidth. Conditions (C6)–(C8) are technical requirements for the model selection or variable selection; see Liang et al. [31], and Wang [35]. Condition (C9) was used in Liang et al. [31] to obtain an exponential inequality for the sum of random variables; see the details in Chernoff [5].

THEOREM 3.1. *Under regularity conditions* (C1)–(C9)*, as* $n \to \infty$*, we have*

$$\Pr\left(\mathcal{M}_{\mathcal{T}} \subset \mathcal{M}^{([Kn^{\xi_0 + 4\xi_{\min}}])}\right) \longrightarrow 1,$$

where $\mathcal{M}_{\mathcal{T}} = \{j : \beta_j \neq 0\}$ *denotes the true model, and* $\mathcal{M}^{([Kn^{\xi_0+4\xi_{\min}}])}$ *denotes the selected* $[Kn^{\xi_0+4\xi_{\min}}]$*th model in the solution path* $\mathbb{S}$*. The constant* $K = 4\tau_{\max}\tau_{\min}^{-2}C_{\boldsymbol{\beta}}^2\nu_{\boldsymbol{\beta}}^{-4}\nu$ *is independent of the sample size* n*, the constants* $\tau_{\max}, \tau_{\min}, C_{\boldsymbol{\beta}}, \nu_{\boldsymbol{\beta}}$ *and* ν *are defined in Conditions* (C6)*,* (C7) *and* (C8)*, and* $[t]$ *denotes the smallest integer not less than* t*.*

REMARK 1. When $q = 1$ and $\boldsymbol{X} \equiv 1$, Model (1.1) is reduced to an ultra-high dimensional partially linear model. Thus, the result in Liang et al. [31] will be the special case of Theorem 3.1. Theorem 3.1 shows that the profile forward regression method can identify all relevant predictors within $O(n^{\xi_0+4\xi_{\min}})$ steps, which is smaller than the sample size n, with probability tending to 1.

Since the solution path $\mathbb{S}$ consists of n nested models, we need to determine which selected model includes the candidate predictor Z_{k^*} in the model of selected ones. To this end, we adopt an extended Bayesian information criterion (EBIC) as follows:

$$\mathrm{EBIC}(\mathcal{M}) = \ln(\widehat{\sigma}^2_{(\mathcal{M})}) + n^{-1}|\mathcal{M}|\,(\ln n + 2\zeta \ln p)\,, \tag{3.1}$$

where ζ is a fixed constant, $\mathcal{M}$ is an arbitrary candidate model with $|\mathcal{M}| \leq n$,

$$\widehat{\sigma}^2_{(\mathcal{M})} = n^{-1}\mathrm{RSS}(\mathcal{M}) = \widehat{\boldsymbol{Y}}^\top\{\mathbf{I}_n - \mathbf{H}_{(\mathcal{M})}\}\widehat{\boldsymbol{Y}}/n, \quad \mathbf{H}_{(\mathcal{M})} = \widehat{\mathbf{Z}}_{(\mathcal{M})}\{\widehat{\mathbf{Z}}^\top_{(\mathcal{M})}\widehat{\mathbf{Z}}_{(\mathcal{M})}\}^{-1}\widehat{\mathbf{Z}}^\top_{(\mathcal{M})}.$$

When $\zeta = 1$, the EBIC criterion has been used by Chen and Chen [2], Liang et al. [31], and Wang [35]. Let $\widehat{m} = \arg\min_{1\leq k\leq n}\mathrm{EBIC}(\mathcal{M}^{(k)})$, then the selected model is $\mathcal{M}^{(\widehat{m})}$. Then we may want to know whether the model chosen by the EBIC criterion can contain the true model with probability tending to 1. The following theorem answers the question: the EBIC criterion enjoys the screening consistency property.

THEOREM 3.2. *Under regularity conditions* (C1)–(C9)*, as* $n \to \infty$*, we have*

$$\Pr(\mathcal{M}_{\mathcal{T}} \subset \mathcal{M}^{(\widehat{m})}) \longrightarrow 1. \tag{3.2}$$

4. Numerical studies

In this section, we present the results of Monte Carlo simulations to evaluate the finite-sample performance of the proposed PFR algorithm of ultra-high dimensional variable screening. Throughout this section, we use the Epanechnikov kernel $K(u) = 0.75(1-u^2)_+$. For each setting, we repeat the experiment 200 times and compare the proposed PFR method with the FR method in Wang [35]. The FR method is treated as a standard method, and the results are obtained by assuming that the coefficient function vector $\boldsymbol{\alpha}(U)$ in (4.1) is known, then the response variable becomes $\tilde{Y} = Y - \boldsymbol{X}^\top\boldsymbol{\alpha}(U)$.

Consider the following varying coefficient partially linear model:

$$Y = \boldsymbol{X}^\top\boldsymbol{\alpha}(U) + \boldsymbol{Z}^\top\boldsymbol{\beta} + \varepsilon, \tag{4.1}$$

where the nonparametric component $\boldsymbol{\alpha}(U) = (\alpha_1(U), \alpha_2(U))^\top$ with $q = 2$. In Examples 1–3, we take $\alpha_1(U) = 4 + \sin(2\pi U)$ and $\alpha_2(U) = 2U(1-U)$, where the covariate U is uniformly distributed on [0,1], $X_1 = 1$, and X_2 follows the standard normal distribution except for Example 2. In all examples, the noise ε is generated from the normal distribution with mean 0 and variance σ^2. We consider different noise level to obtain different

signal-to-noise ratio $R^2 = \text{var}\{\boldsymbol{X}^\top\boldsymbol{\alpha}(U) + \boldsymbol{Z}^\top\boldsymbol{\beta}\}/\text{var}(Y)$. For the linear part, we consider the following three commonly adopted data structures.

Example 1 (Independent predictors). In this example, the linear predictor $\boldsymbol{Z}$ is an independent and standard normal random vector. The size of the true model is chosen to be $p_0 = 8$ with $\beta_j = (-1)^{V_j}(4\ln n/\sqrt{n} + |T_j|)$, where V_j is a binary random variable with $\Pr(V_j = 1) = 0.4$ and T_j is a standard normal random variable. The variance σ^2 of model error ε is selected so that the resulting theoretical R^2 is approximately 50%, 70% and 90%, respectively, to represent the different signal-to-noise ratios from weak to strong. For comparison, we consider three sample sizes ($n = 100, 150$ and 200) and three predictor dimensions ($p = 500, 1000$ and 2000) for the linear part.

We mainly demonstrate whether or not the PFR method can identify the relevant predictors as well as the FR method even if the model involves the nonparametric components. Let $\widehat{\boldsymbol{\beta}}_{(k)} = (\widehat{\beta}_{1(k)}, \ldots, \widehat{\beta}_{p(k)})^\top \in \mathbb{R}^p$ denote the estimator obtained in the kth simulation replication. Based on the EBIC, we use the proposed PFR algorithm to select the final candidate model. The selected model is taken as $\widehat{\mathcal{M}}_{(k)} = \{j \in \{1, \ldots, p\} : |\widehat{\beta}_{j(k)}| > 0\}$. Similar to Liang et al. [31] and Wang [35], we compute the following seven performance measures to evaluate the PFR method.

(1) The average model size (AMS) is computed as

$$\frac{1}{200}\sum_{k=1}^{200}|\widehat{\mathcal{M}}_{(k)}|.$$

(2) The coverage probability (CP) is computed as

$$\frac{1}{200}\sum_{k=1}^{200}\mathbf{1}(\mathcal{M}_{\mathcal{T}} \subset \widehat{\mathcal{M}}_{(k)}),$$

which is used to measure how likely all relevant predictors will be discovered by the PFR method.

(3) The percentage of correct zeros (CZ), which is used to characterize the PFR method's capability in producing sparse solutions, can be computed by

$$\frac{100}{200(p-p_0)}\sum_{k=1}^{200}\sum_{j=1}^{p}\mathbf{1}(\widehat{\beta}_{j(k)} = 0) \times \mathbf{1}(\beta_j = 0).$$

(4) The percentage of incorrect zeros (IZ), which is used to characterize the PFR method's under-fitting effects, can be computed by

$$\frac{100}{200p_0}\sum_{k=1}^{200}\sum_{j=1}^{p}\mathbf{1}(\widehat{\beta}_{j(k)} = 0) \times \mathbf{1}(\beta_j \neq 0).$$

(5) The percentage of correctly fitted (CF), which is used to measure the capability in identifying the true model correctly, can be computed by

$$\frac{1}{200}\sum_{k=1}^{200}\mathbf{1}(\widehat{\mathcal{M}}_{(k)} = \mathcal{M}_{\mathcal{T}}).$$

(6) The percentage of submodels $\widehat{\mathcal{M}}_{(k)}$ contains the jth covariate (P_j), which is used to measure the capability that the jth covariate will be discovered by the PFR method.

(7) The average estimation error (AEE) is computed by the L_2 error

$$\frac{1}{200}\sum_{k=1}^{200}\|\widehat{\boldsymbol{\beta}}_{(k)}-\boldsymbol{\beta}\|_2.$$

For getting the consistent estimators of the unknown nonparametric functions including $\mathrm{E}(\boldsymbol{X}_i\boldsymbol{X}_i^\top|U_i)$, $\mathrm{E}(\boldsymbol{X}_iY_i|U_i)$ and $\mathrm{E}(\boldsymbol{X}_i\boldsymbol{Z}_i^\top|U_i)$ by using the nonparametric smoothing methods, we need to choose an appropriate bandwidth. In the simulation studies, the rule of thumb is used to choose the bandwidth for convenience, that is, $h=c\widehat{\sigma}_U n^{-1/5}$, where $\widehat{\sigma}_U$ denotes the sample standard deviation of U. To check the effect of the bandwidth, we consider Example 1 and take $c=1,1.5,2$, three sample sizes $n=100,150$ and 200 and two predictor dimensions $p=500$ and 1000. The simulation results are presented in Table 1. It is evident that the proposed method is not sensitive to the choice of the bandwidth.

Table 1: Simulation results for Example 1 based on various bandwidths.

		$p=500$						$p=1000$					
c	n	CP	CZ	IZ	CF	AMS	AEE	CP	CZ	IZ	CF	AMS	AEE
	100	3.0	99.95	60.8	2.0	3.370	16.092	0.0	99.9	70.0	0.0	2.745	21.122
1	150	32.5	99.9	20.1	29.0	6.540	7.584	20.5	99.9	26.1	18.0	6.030	8.433
	200	71.5	99.9	5.4	69.5	7.615	3.129	59.5	99.9	8.0	55.5	7.425	4.160
	100	3.5	99.9	60.6	3.5	3.430	10.458	1.0	99.9	70.9	1.0	2.515	16.913
1.5	150	35.0	99.9	18.3	32.5	6.635	5.959	26.0	99.9	23.8	23.5	6.170	6.211
	200	63.5	99.9	6.7	60.5	7.495	3.054	57.5	99.9	7.7	55.0	7.420	2.780
	100	3.0	99.96	59.9	1.0	3.410	13.484	0.5	99.9	69.1	0.0	2.625	15.341
2	150	32.0	99.9	18.6	29.5	6.590	5.772	24.0	99.9	25.1	22.0	6.060	5.898
	200	71.0	99.9	5.4	70.0	7.585	2.215	61.0	99.9	7.6	58.0	7.445	3.459

Next, we provide the finite-sample performance of the proposed PFR method and the FR method under different signal-to-noise ratios R^2, different dimensions p and different sample sizes n. The corresponding simulation results are reported in Table 2.

From Table 2, we can find the following results according to the effects of the signal-to-noise ratios:

(1) For three different signal-to-noise ratios, the PFR and FR methods can almost truly identify the inactive variables, and have the higher percentage of correct zeros (CZ). This shows that two methods can produce sparse solutions with probability tending to one.

(2) The signal-to-noise ratios have certain effect for the PFR and FR methods' performance in terms of the results of the coverage probability (CP), the percentage of incorrect zeros (IZ), the percentage of correctly fitted (CF), the average model size (AMS) and the average estimation error (AEE). For the low signal-to-noise ratio ($R^2=50\%$), neither PFR nor FR performs well in terms of the coverage probability. Wang [35] also find the FR method performs worse for the low signal-to-noise ratios. For the high signal-to-noise ratio ($R^2=90\%$), the PFR and FR methods perform

Table 2: Simulation results for Example 1 based on the PFR and FR methods.

			PFR						FR					
R^2	p	n	CP	CZ	IZ	CF	AMS	AEE	CP	CZ	IZ	CF	AMS	AEE
		100	0.0	99.9	85.1	0.0	1.375	19.760	0.0	99.9	85.8	0.0	1.245	16.900
	500	150	0.5	99.9	70.4	0.5	2.415	12.553	0.5	99.9	71.8	0.5	2.275	10.742
		200	0.8	99.9	53.4	0.8	3.760	8.225	2.0	99.9	52.4	2.0	3.835	8.061
		100	0.0	99.9	87.7	0.0	1.230	25.468	0.0	99.9	87.4	0.0	1.135	19.119
0.5	1000	150	0.0	99.9	77.6	0.0	1.865	13.643	0.0	99.9	77.5	0.0	1.820	11.250
		200	0.5	99.9	56.6	0.5	3.530	10.110	0.5	99.9	57.8	0.5	3.410	9.229
		100	0.0	99.9	90.8	0.0	1.170	30.087	0.0	99.9	89.2	0.0	1.105	22.968
	2000	150	0.0	99.9	80.5	0.0	1.655	13.993	0.0	99.9	80.3	0.0	1.625	12.348
		200	0.0	99.9	63.3	0.0	2.995	10.112	0.0	99.9	64.6	0.0	2.870	10.067
		100	3.5	99.9	60.6	3.5	3.430	10.458	4.5	99.9	59.9	4.5	3.230	9.626
	500	150	35.0	99.9	18.3	32.5	6.635	5.959	40.0	99.9	16.0	38.0	6.765	4.686
		200	63.5	99.9	6.7	60.5	7.495	3.054	67.0	99.9	5.7	65.5	7.565	2.523
		100	1.0	99.9	70.9	1.0	2.515	16.913	1.5	99.9	69.9	1.5	2.480	12.789
0.7	1000	150	26.0	99.9	23.8	23.5	6.170	6.211	29.5	99.9	25.4	20.5	6.012	5.909
		200	57.5	99.9	7.7	55.0	7.420	2.780	60.5	99.9	8.0	60.0	7.375	2.958
		100	0.0	99.9	76.4	0.0	2.120	19.060	0.7	99.9	77.2	0.7	1.915	13.741
	2000	150	14.5	99.9	33.8	12.0	5.410	8.713	23.5	99.9	31.0	21.0	5.380	6.731
		200	50.0	99.9	10.5	48.0	7.205	3.995	53.5	99.9	11.4	52.5	7.100	2.646
		100	98.5	99.9	0.25	80.5	8.190	4.083	100.0	99.9	0.0	93.5	8.075	2.788
	500	150	100.0	99.9	0.0	93.5	8.070	1.965	100.0	99.9	0.0	97.5	8.025	1.365
		200	100.0	99.9	0.0	95.0	8.055	1.345	100.0	99.9	0.0	97.5	8.025	0.810
		100	97.5	99.9	0.9	73.5	8.255	5.043	98.0	99.9	0.9	88.0	8.035	3.018
0.9	1000	150	100.0	99.9	0.0	91.5	8.100	2.871	100.0	99.9	0.0	96.5	8.035	1.474
		200	100.0	99.9	0.0	95.5	8.045	1.813	100.0	99.9	0.0	97.5	8.025	1.208
		100	94.0	99.9	3.3	71.5	8.070	6.135	98.0	99.9	0.8	88.0	8.065	3.801
	2000	150	99.5	99.9	0.6	97.0	8.055	2.073	99.5	99.9	0.6	97.0	8.020	1.127
		200	100.0	99.9	0.0	95.5	8.045	1.194	100.0	99.9	0.0	99.0	8.010	0.773

better. The coverage probabilities approach 100% with the sample size n. We also note that the results are similar in other settings from Table 2.

To reduce the computational burden, we fix $c = 1.5$ in the bandwidth choice and the signal-to-signal ratio is 70% for comparing the proposed PFR method with the FR method proposed in Wang [35] in the following three examples.

Example 2 (Autoregressive correlation). The covariate $(\boldsymbol{Z}^\top, \boldsymbol{X}^\top)$ is a $(p+2)$-dimensional multivariate normal random vector with mean zero and covariance matrix $\Sigma = (\sigma_{ij})$ with element $\sigma_{ij} = 0.5^{|i-j|}$ for $1 \le i, j \le p+2$. The 1st, 4th and 7th components of $\boldsymbol{\beta}$ are 3, 1.5 and 2, respectively. Other elements of $\boldsymbol{\beta}$ are fixed to be zero. For comparison, we consider three sample sizes ($n = 100, 150$ and 200) and three predictor dimensions ($p = 500, 1000$ and 2000) for the linear part. The simulation results are reported in Table 3.

Example 3 (Compound symmetry). In this simulation, we consider the covariance structure of the linear part is compound symmetry. Specially, the covariate $\boldsymbol{Z}$ has a p-

Table 3: Simulation results for Example 2.

p	n	CP	CZ	IZ	CF	AMS	P_1	P_2	P_3	AEE
					Method:		PFR			
500	100	15.0	99.9	35.0	14.5	2.125	98.5	31.0	65.5	6.260
	150	53.0	99.9	16.0	52.5	2.615	99.5	61.0	91.5	3.202
	200	78.0	99.9	7.5	76.0	2.835	100.0	81.0	96.5	1.949
1000	100	9.5	99.9	38.0	9.5	2.020	96.5	26.5	63.0	6.251
	150	39.0	99.9	21.3	38.5	2.535	100.0	50.5	85.5	5.602
	200	77.0	99.9	8.0	76.0	2.810	100.0	80.5	95.5	3.026
2000	100	7.0	99.9	42.2	7.0	1.940	95.5	19.0	59.0	6.996
	150	37.0	99.9	22.2	36.5	2.475	99.0	42.5	92.0	4.671
	200	68.0	99.9	10.8	67.0	2.745	100.0	71.0	96.5	2.788
					Method:		FR			
500	100	22.0	99.9	31.3	21.5	2.130	99.0	34.0	73.0	4.383
	150	60.5	99.9	13.5	60.0	2.625	100.0	63.0	96.5	2.439
	200	82.0	99.9	6.2	80.0	2.840	100.0	82.0	100.0	1.406
1000	100	16.0	99.8	25.4	14.5	2.025	98.5	28.0	72.0	3.659
	150	52.5	99.9	17.0	52.0	2.545	100.0	57.0	92.0	3.535
	200	79.5	99.9	7.1	78.5	2.815	100.0	81.0	98.0	2.723
2000	100	15.5	99.9	36.5	15.5	1.945	98.5	23.5	68.5	3.621
	150	49.0	99.9	17.7	48.5	2.485	100.0	53.0	94.0	1.657
	200	74.5	99.9	8.5	73.5	2.755	100.0	75.0	99.5	1.027

dimensional multivariate normal distribution $\mathcal{N}(0, \Sigma)$. The covariance matrix has entries $\sigma_{ii} = 1$ for all $i \in \{1, \ldots, p\}$ and $\sigma_{ij} = \rho, i \neq j$. Furthermore, the nonzero coefficients $\beta_j = 5$ for $j = 1, 2, 3$. For comparison, we consider $n = 100$ and three predictor dimensions ($p = 500, 1000$ and 2000) for the linear part, and $\rho = 0.1, 0.5, 0.9$. The simulation results are reported in Table 4.

In the following example, we evaluate the performance of our PFR method in the setting that $\boldsymbol{Z}$ is highly correlated with $\boldsymbol{X}$, and meanwhile, is highly correlated with U.

Example 4. Let $\boldsymbol{W} = (W_1, \ldots, W_{p+2})^\top$ be an independent and standard normal random vector, and (T_1, T_2) be independent and standard uniformly distributed random variables. We construct $(U, \boldsymbol{X}^\top, \boldsymbol{Z}^\top)$ as follows:

$$X_i = \frac{W_i + t_1 T_1}{1 + t_1}, \quad Z_j = \frac{W_{j+2} + t_1 T_1}{1 + t_1}, \quad U = \frac{T_2 + t_2 T_1}{1 + t_2},$$

where $i = 1, 2$ and $j \in \{1, \ldots, p\}$. Let the nonparametric component $\boldsymbol{\alpha}(U) = (\alpha_1(U), \alpha_2(U))^\top$ in (4.1), where

$$\alpha_1(U) = (U + 1)^2, \quad \alpha_2(U) = \frac{4 \sin(2\pi U)}{2 - \sin(2\pi U)},$$

and the nonzero coefficients β_j be 3 for $j = 1, 2$. We take $(t_1, t_2) = (2, 1)$ and $(3, 1)$, corresponding to the correlation coefficient matrix of $(\boldsymbol{X}^\top, \boldsymbol{Z}^\top)$ with non-diagonal elements being 0.25 and 0.43, and the correlation coefficient matrix of $(\boldsymbol{X}^\top, U)$ and $(\boldsymbol{Z}^\top, U)$ with non-diagonal elements being 0.35 and 0.46, respectively. For comparison, we consider two

Table 4: Simulation results for Example 3.

p	ρ	CP	CZ	IZ	CF	AMS	P_1	P_2	P_3	AEE
					Method:	PFR				
	0.1	99.0	99.9	0.3	91.5	3.090	99.5	100.0	99.5	9.458
500	0.5	26.0	99.9	31.2	24.0	2.445	69.0	68.5	69.0	28.747
	0.9	1.0	99.8	92.2	0.0	1.050	8.0	8.0	7.5	116.017
	0.1	99.0	99.9	0.3	93.0	3.050	99.5	99.5	100.0	6.636
1000	0.5	24.0	99.9	37.0	23.0	2.430	62.5	64.5	62.0	40.541
	0.9	0.6	99.9	95.5	0.0	1.065	3.5	5.0	5.0	126.787
	0.1	97.0	99.9	1.2	88.5	3.070	98.0	99.5	99.0	9.001
2000	0.5	11.5	99.9	50.0	9.5	2.345	50.0	48.0	52.0	47.778
	0.9	0.4	99.9	97.3	0.0	1.015	3.0	2.0	3.0	130.374
					Method:	FR				
	0.1	99.5	99.9	0.2	97.5	3.025	99.5	100.0	100.0	4.846
500	0.5	34.5	99.94	26.8	34.5	2.465	76.0	71.0	72.5	24.039
	0.9	2.0	99.8	91.5	0.1	1.035	10.5	8.0	7.0	111.833
	0.1	99.5	99.9	0.5	95.5	3.015	99.5	99.0	100.0	5.671
1000	0.5	27.0	99.9	33.2	26.5	2.380	65.5	70.0	65.0	33.586
	0.9	1.0	99.9	93.8	0.0	1.010	7.0	6.5	5.0	124.342
	0.1	99.0	99.9	0.3	96.5	3.015	99.5	100.0	99.5	4.752
2000	0.5	16.0	99.9	42.5	14.5	2.275	54.5	58.0	60.0	40.715
	0.9	0.7	99.9	95.2	0.0	1.005	4.5	5.5	4.5	130.393

sample sizes ($n = 100, 150$) and three predictor dimensions ($p = 500, 1000$ and 2000) for the linear part. The simulation results are reported in Table 5.

From Tables 2–5, we have the comparison results as follows.

(1) The proposed PFR method is comparable with the FR method proposed by Wang [35], which is treated as a standard method. These simulation results numerically confirm that the proposed PFR method is screening consistent. As a byproduct of the screening consistency property, the percentage of incorrect zeros (IZ) approaches quickly toward 0 as the sample size increases.

(2) For fixed p, the proposed PFR performs better as the sample size increases. For example, the coverage probabilities become large with the sample size increasing. But the coverage probabilities are not enough large for the signal-to-noise ratio $R^2 = 70\%$, the reason may be that the signal-to-noise ratio is also not large and the sample size is not enough large to obtain a satisfactory coverage probability. We can clearly find that the coverage probabilities change substantially as the sample size increases. As long as the sample size is enough large, the coverage probabilities can approach approximately 100%. As is known to all, it is not sufficient that we only use the coverage probability to assess the screening consistency of the proposed PFR method, we can also use the other assessment criteria to illustrate the screening consistency, such as the capability of identifying the correct zero. It is easy to see

Table 5: Simulation results for Example 4.

p	n	(t_1, t_2)	CP	CZ	IZ	CF	AMS	AEE	P_1	P_2
					Method:	PFR				
500	100		96.5	99.9	2.0	91.5	2.035	4.213	98.5	97.5
	150		100.0	99.9	0.0	94.0	2.360	3.181	100.0	100.0
1000	100	(2,1)	96.0	99.9	2.0	93.5	2.020	4.441	100.0	96.0
	150		100.0	99.9	0.0	98.0	2.025	2.443	100.0	100.0
2000	100		93.5	99.9	3.3	87.0	2.015	4.622	98.0	95.5
	150		100.0	99.9	0.0	96.0	2.040	2.123	100.0	100.0
500	100		78.0	99.9	12.0	77.0	1.850	6.944	90.5	85.5
	150		97.5	99.9	1.3	96.0	2.010	3.514	98.5	99.0
1000	100	(3,1)	76.0	99.9	13.0	74.5	1.835	9.178	88.0	86.0
	150		96.5	99.9	1.8	93.5	2.005	3.784	97.0	99.5
2000	100		68.0	99.9	18.0	63.0	1.845	10.086	82.0	82.0
	150		94.5	99.9	2.8	91.0	2.000	4.161	97.5	97.0
					Method:	FR				
500	100		99.5	99.9	0.2	97.0	2.025	1.721	100.0	99.5
	150		100.0	99.9	0.0	97.5	2.025	2.064	100.0	100.0
1000	100	(2,1)	99.0	99.9	0.3	97.0	2.015	3.217	99.5	99.5
	150		100.0	99.9	0.0	97.5	2.025	2.227	100.0	100.0
2000	100		97.5	99.9	1.5	96.5	2.020	3.489	99.0	98.0
	150		100.0	99.9	0.0	99.0	2.010	1.309	100.0	100.0
500	100		92.0	99.9	4.0	90.0	2.015	7.077	97.5	94.5
	150		99.5	99.9	0.1	98.0	2.015	3.229	100.0	99.5
1000	100	(3,1)	89.0	99.9	6.3	87.5	1.995	7.453	93.0	94.5
	150		99.5	99.9	0.5	96.5	2.030	2.235	100.0	99.5
2000	100		84.0	99.9	9.5	82.5	1.985	10.238	90.0	91.0
	150		99.0	99.9	0.8	96.5	2.030	2.445	100.0	99.0

that the proposed PFR method identifies correct zeros almost 100%. In addition, the average model size is small, and is close to the true model size when the sample size increases. Consequently, the average estimation error decreases as the sample size increases.

(3) For the fixed sample size n, we consider the performance of the proposed PFR method for the different dimension p of the linear predictors. It is easy to see that the finite sample performance becomes worse as the dimension p of the covariates increases. Then we may focus on the variation rate, which does not deteriorate rapidly, when the dimension of the covariates increases. For example, Table 4 shows that the coverage probability drops from 78% to 68% as the dimension of the covariates increases from 500 to $4 \times 500 = 2000$ with $n = 200$. For comparison, we fix the dimension $p = 2000$, but increases the sample size from $n = 50$ to $n = 4 \times 50 = 200$, the coverage probability is computed. We find that the coverage probability (CP) increases from 4.5% to 68.0%. These results show that the sample size n is more important than the dimension of the covariates for ultra-high dimensional variable

screening.

(4) We also compare the finite sample performance of the proposed PFR method in Table 4 when the correlation between covariates drops from high to low. We find that the coverage probabilities deteriorate rapidly, for example, the coverage probability drops from 99.0% to 1%, as the correlation increases from 0.1 to 0.9 with the dimension of linear part $p = 500$.

(5) We evaluate the finite sample performance of the proposed PFR method when $\boldsymbol{Z}$ is highly correlated with the covariates $(\boldsymbol{X}, U)$ in Table 5. We find that the finite performance becomes worse as the correlation becomes higher. For example, the coverage probability drops from 96.5% to 78% as the value of (t_1, t_2) ranges from $(2, 1)$ to $(3, 1)$ for $p = 500$ and $n = 100$.

For other assessment criteria, we can find similar simulation results as in Tables 2–5. In conclusion, the numerical results demonstrate that the proposed PFR method performs better than existing methods.

5. Application to birth weight data

We demonstrate the effectiveness of the proposed PFR method by an application to the birth weight data. Votavová et al. [34] collected the samples of peripheral blood, placenta and cord blood from 91 women who gave birth to a baby in the Česke Budějovice Hospital (Czech Republic) from November 2008 to March 2009. Based on the smoking history, the women were divided into two groups, 20 smokers and 52 non-smokers, while 19 passive smokers were excluded from the study. Gene expression profiles were assayed using HumanRef-8 v3 Expresion BeadChips with 24,526 transcripts. The study was approved by the Local Institutional Review Board. All participants provided the written informed consent and completed an extensive questionnaire. Birth weight of baby (BW, in kilograms) was recorded along with mother's age (MOA), gestational age (GEA), mother's body mass index (BMI), and parity, measurement of the amount of cotinine, a chemical found in tobacco, in the blood. Of interest in this empirical analysis is to identify which genes are strongly associated with the birth weight of baby (BW).

The blood and placental samples include $n = 64$ subjects after dropping those with incomplete information. Dudoit et al. [7] proposed a three-step procedure to preprocess the gene expression data: remove genes having little variation in intensity, transform intensities to base 2 logarithms, and normalize each data vector to have sample mean 0 and standard deviation 1. This procedure results in $p = 5869$ genes. Based on the results in Votavová et al. [34], we consider the varying coefficient partially linear model to fit the birth weight data as follows:

$$\mathrm{BW} = \alpha_1(U) + \alpha_2(U)\mathrm{GEA} + \alpha_3(U)\mathrm{BMI} + \sum_{j=1}^{5869} \beta_j \mathrm{GE}_j + \varepsilon, \tag{5.1}$$

where the variable $U = \mathrm{MOA}$, GE_j is the jth gene, and GEA and BMI are normalized variables with sample mean 0 and standard deviation 1.

Epanechnikov kernel $K(u) = 0.75(1 - u^2)_+$ and the bandwidth $h = c\widehat{\sigma}_U n^{-1/5}$ are used to fit the coefficient functions, where $\widehat{\sigma}_U$ denotes the sample standard deviation of U. The

Table 6: The top ten genes and the corresponding ID numbers selected by the PFR method.

Genes	NTN3	SPDYE1	FRK	OR5P2	MRO
ID numbers	1656040	2250830	1727605	2059464	1800874
Genes	UTS2B	KCNC4	RGS9	LAGE1	CDC25A
ID numbers	2180232	1727850	2389984	1803412	1733396

proposed PFR algorithm is used to select the candidate model, and the selected top ten genes and their corresponding ID numbers are listed in Table 6.

We note that the gene OR5P2 is also identified in Sherwood and Wang [33], and the gene OR5P2 is lied on chromosome 11. Chromosome 11 contains the gene PHLDA2, which is reported in Ishida et al. [18]. Ishida et al. [18] found that the gene PHLDA2 is highly expressed in mothers that have children with low birth weight. Gilliam et al. [16] pointed out that the gene RGS9 plays a role in obesity and the parental obesity may have influence on the birth weight of baby (BW).

The EBIC criterion is used to select the top two genes (NTN3 and SPDYE1) in the solution path. We find that the gene NTN3 encodes a novel human netrin mapping to the autosomal dominant polycystic kidney disease region on chromosome 16p13.3, and the gene SPDYE1 is located at chromosome 7p13 which is close to the Williams Beuren syndrome chromosome region 7q11.23. It remains to be validated wether it is really related with the birth weight of baby (BW) by biologists. Please refer to `https://www.ncbi.nlm.nih.gov/gene` for the more details of other genes.

The estimated coefficient functions are reported in Figure 1. The latter shows the mother's age (MOA) has a positive impact on the birth weight of baby (BW) before age 30, and has a negative impact on the birth weight of baby (BW) after age 30. The fitted curve $\widehat{\alpha}_2(u)$ is always positive and has a rapid increasing after age 35. This implies that the variable GEA has a positive effect on the birth weight of baby (BW), and the value of effect increases rapidly after age 35. This also coincides with the intuition that the birth weight of baby (BW) increases with the variable GEA and premature birth is often strongly associated with low birth weight of baby. The fitted curve $\widehat{\alpha}_3(u)$ is decreasing with the mother's age (MOA), and turns to negative about age 35. This implies that BMI has a positive effect on the birth weight of baby (BW) before age 35, and has a negative effect after age 35. Our findings are consistent with the results in Gilliam er al. [16]. Hence, from a practical point of view, we have demonstrated that the proposed PFR method is an efficient method for analyzing the varying coefficient partially linear model.

Acknowledgements

Gaorong Li's research was supported by the National Natural Science Foundation of China (Grant No. 11471029), the Beijing Natural Science Foundation (Grant No. 1142002) and the Science and Technology Project of Beijing Municipal Education Commission (Grant No. KM201410005010). Tiejun Tong's research was supported by the Hong Kong Baptist University FRG grants FRG1/14-15/044, FRG2/15-16/019 and FRG2/15-16/038, and the National Natural Science Foundation of China (Grant No. 11671338). The authors would like to thank the Editor-in-Chief, Christian Genest, an Associate Editor, and two referees for their helpful comments that helped to improve an earlier version

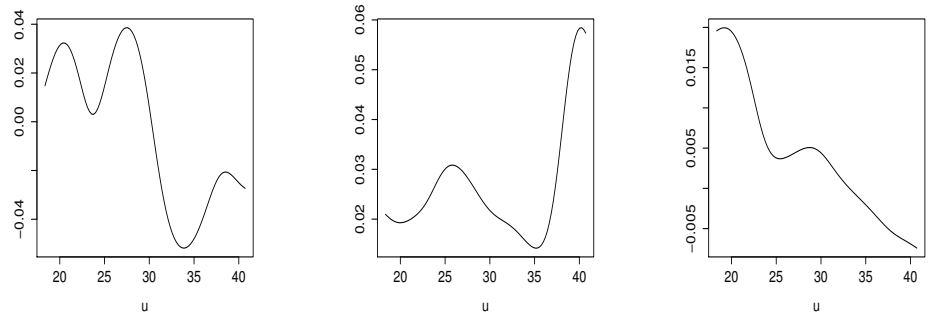

Figure 1: *The fitted coefficient functions* $\widehat{\alpha}_1(u), \widehat{\alpha}_2(u)$ *and* $\widehat{\alpha}_3(u)$ *from the left panel to the right panel, respectively.*

of this article.

6. Appendix

For the sake of convenience, let $C \in (0, \infty)$ denote a constant not depending on n and p, but may take difference values at each appearance. We introduce the following notations to simplify our presentation. Let $\Gamma(u) = \mathrm{E}(\boldsymbol{X}\boldsymbol{X}^\top | U = u)$, $\Psi(u) = \mathrm{E}(\boldsymbol{X}\boldsymbol{Y} | U = u)$ and $\Phi(u) = \mathrm{E}(\boldsymbol{X}\boldsymbol{Z}^\top | U = u)$. Let $Y_i^* = Y_i - \boldsymbol{X}_i^\top \boldsymbol{\eta}(U_i)$ be the profiled response variable and $\boldsymbol{Z}_i^* = \boldsymbol{Z}_i - \boldsymbol{\mu}(U_i)^\top \boldsymbol{X}_i = (Z_{i1}^*, \ldots, Z_{ip}^*)^\top$ be the profiled predictors. Define $\boldsymbol{Y}^* = (Y_1^*, \ldots, Y_n^*)^\top \in \mathbb{R}^n$ as the profiled response vector, and $\mathbf{Z}^* = (\boldsymbol{Z}_1^*, \ldots, \boldsymbol{Z}_n^*)^\top \in \mathbb{R}^{n \times p}$ as the matrix of the profiled predictors. Further define $\widehat{\Sigma} = \widehat{\mathbf{Z}}^\top \widehat{\mathbf{Z}}/n$ and $\Sigma^* = \mathbf{Z}^{*\top} \mathbf{Z}^*/n$. For any candidate model $\mathcal{M}$, define $\Sigma_{(\mathcal{M})} = \{\Sigma_{ij} \mid i, j \in \mathcal{M}\}$.

6.1. Some lemmas

LEMMA 6.1. *Let* $W_1, \ldots, W_n$ *be independent and identically distributed random variables with* $\mathrm{E}(W_i) = 0$ *and* $\mathrm{var}(W_i) = 1$. $M(t) = \mathrm{E}\{\exp(tW_i)\}$ *is the moment generating function of* W_i*, for each* $i \in \{1, \ldots, n\}$*, and assume that there exists a positive constant* t_0 *such that* $\mathrm{E}\{\exp(t|W_i|)\} < \infty$ *for all* $t \in [0, t_0]$. *Let* a_{nk}*, for any* $1 \le k \le n$*, be a sequence of constants and* $A, A_1, A_2, \ldots$ *be a sequence of constants satisfying*

$$A_n \ge \sum_{k=1}^{n} a_{nk}^2 \quad \textit{and} \quad A \ge \max_k |a_{nk}|/A_n.$$

If

$$M^* = \sup_{0 \le t \le t_0} \left| \frac{d^3}{dt^3} \ln M(t) \right| < \infty,$$

then, for $0 < \xi < t_0/A$*, we have*

$$\Pr\left(\left| \sum_{k=1}^{n} a_{nk} W_k \right| > \xi \right) \le 2 \exp\left\{ -\frac{\xi^2}{2A_n} \left(1 - \frac{1}{3} A M^* \xi \right) \right\}.$$

LEMMA 6.2. *Let $W_1, \ldots, W_n$ be independent random variables with mean $\mathrm{E}(W_i) = 0$ and variance $\mathrm{var}(W_i) = \sigma_i^2$. If $\mathrm{E}|W_i|^m \leq (m!/2)\sigma_i^2 t^{m-2}$ for $1 \leq i \leq n$, $0 < t < \infty$, and some $m > 2$, then for any $\delta > 0$,*

$$\Pr\left(\left|\sum_{i=1}^{n} W_i\right| > \delta\right) \leq 2\exp\left\{-\frac{\delta^2}{2(\sum_{i=1}^{n}\sigma_i^2 + t\delta)}\right\}.$$

LEMMA 6.3. *Let $(\boldsymbol{X}_i, U_i)$ be independent and identically distributed random variable vector, where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^\top \in \mathbb{R}^p$, and U_i is a univariate variable for each $i \in \{1, \ldots, n\}$. Define $G_j(U_i) = E(X_{ij}|U_i)$ for each $j \in \{1, \ldots, p\}$. The weight functions w_{nk}, $1 \leq k \leq n$, satisfy, with probability tending to 1,*

$$\max_{1\leq k\leq n}\sum_{i=1}^{n} w_{nk}(U_i) = O(1), \quad \max_{1\leq i,k\leq n} w_{nk}(U_i) = o(n^{-4/5})$$

and

$$\max_{1\leq k\leq n}\sum_{k=1}^{n} w_{nk}(U_i)\mathbf{1}(|U_i - U_k| > c_n) = o(c_n),$$

then we have

$$\max_{1\leq i\leq n}\max_{1\leq j\leq p}|G_j(U_i) - \sum_{k=1}^{n} w_{nk}(U_i)G_j(U_k)| = o_P(c_n),$$

with $c_n = n^{-1/4}\ln^{-1} n$.

The proofs of Lemma 6.1–6.3 can be found in Liang et al. [31]; hence we omit the details.

LEMMA 6.4. *Suppose Conditions* (C1)–(C5) *and* (C8)–(C9) *hold. We have*

$$\widehat{\boldsymbol{Y}} - \boldsymbol{Y}^* = o_P(1), \tag{6.1}$$

$$\widehat{\mathbf{Z}} - \mathbf{Z}^* = o_P(1). \tag{6.2}$$

Proof. The aim of the lemma is to show that the estimators of the profiled response vector $\boldsymbol{Y}^*$ and the profiled predictors matrix $\mathbf{Z}^*$ are consistent. Since the proof of (6.1) is similar to that of (6.2), we only present the proof of (6.2). If each element of $\widehat{\mathbf{Z}} - \mathbf{Z}^*$ is $o_P(1)$, then we can claim that the estimator of the profiled predictor is consistent. In other words, we only need to prove

$$\max_{1\leq i\leq n}\max_{1\leq j\leq p}|\widehat{Z}_{ij} - Z_{ij}^*| = o_P(1).$$

Note that $\widehat{Z}_{ij} - Z_{ij}^* = -\widehat{\mu}_j^\top(U_i)\boldsymbol{X}_i + \mu_j^\top(U_i)\boldsymbol{X}_i$, where $\mu_j(U_i)$ is the jth column of $\boldsymbol{\mu}(U_i)$, and its estimator is $\widehat{\mu}_j(U_i)$. By Condition (C2), it is easy to show that

$$\begin{aligned}
\max_{1\leq i\leq n}\max_{1\leq j\leq p}|\widehat{Z}_{ij} - Z_{ij}^*| &= \max_{1\leq i\leq n}\max_{1\leq j\leq p}\left|\{\widehat{\mu}_j(U_i) - \mu_j(U_i)\}^\top \boldsymbol{X}_i\right| \\
&\leq q\max_{1\leq i\leq n}\max_{1\leq j\leq p}\max_{1\leq \ell\leq q}|\{\widehat{\mu}_{\ell j}(U_i) - \mu_{\ell j}(U_i)\}X_{il}| \\
&= q\max_{1\leq i\leq n}\max_{1\leq j\leq p}\max_{1\leq \ell\leq q}|\widehat{\mu}_{\ell j}(U_i) - \mu_{\ell j}(U_i)|\{1 + O_P(1)\}. \qquad (6.3)
\end{aligned}$$

Since q is fixed, we only need to prove

$$\max_{1\leq i\leq n}\max_{1\leq j\leq p}|\widehat{\mu}_{\ell j}(U_i) - \mu_{\ell j}(U_i)| = o_P(1)$$

for each $1 \le \ell \le q$. For convenience, we first introduce some notations as follows:

$$\begin{aligned}
\widehat{\Gamma}(u) &= nf(u)[\mathbf{I}_q, \mathbf{0}_q](\mathbf{D}_u^\top \mathbf{W}_u \mathbf{D}_u)^{-1} = (\widehat{\Gamma}_1(u), \ldots, \widehat{\Gamma}_q(u))^\top, \\
\widehat{\Gamma}_\ell(u) &= \left(\widehat{\Gamma}_{\ell 1}(u), \ldots, \widehat{\Gamma}_{\ell 2q}(u)\right)^\top, \qquad 1 \le \ell \le q, \\
\tilde{\Gamma}(u) &= [\mathbf{I}_q, \mathbf{0}_q]\Gamma^{-1}(u) \otimes \begin{pmatrix} 1 & 0 \\ 0 & \mu_2^{-1} \end{pmatrix} = (\tilde{\Gamma}_1(u), \ldots, \tilde{\Gamma}_q(u))^\top, \\
\tilde{\Gamma}_\ell(u) &= (\tilde{\Gamma}_{\ell 1}, \ldots, \tilde{\Gamma}_{\ell 2q})^\top, \qquad 1 \le \ell \le q, \\
\widehat{\Phi}(u) &= \frac{1}{nf(u)} \mathbf{D}_u^\top \mathbf{W}_u \mathbf{Z} = (\widehat{\Phi}_1(u), \ldots, \widehat{\Phi}_{2q}(u))^\top, \\
\widehat{\Phi}_m(u) &= (\widehat{\Phi}_{m1}(u), \ldots, \widehat{\Phi}_{mp}(u))^\top, \qquad 1 \le m \le 2q, \\
\tilde{\Phi}(u) &= \Phi(u) \otimes (1, 0)^\top = (\tilde{\Phi}_1(u), \ldots, \tilde{\Phi}_{2q}(u))^\top, \\
\tilde{\Phi}_m(u) &= (\tilde{\Phi}_{m1}(u), \ldots, \tilde{\Phi}_{mp}(u))^\top, \qquad 1 \le m \le 2q,
\end{aligned}$$

where $\otimes$ is the Kronecker product, $\mu_2 = \int u^2 K(u) du$, $\mathbf{I}_q$ is a $q \times q$ identity matrix, and $\mathbf{0}_q$ is a $q \times q$ zero matrix. We then find that, for each $1 \le \ell \le q$,

$$\begin{aligned}
&\max_{1 \le i \le n} \max_{1 \le j \le p} |\widehat{\mu}_{\ell j}(U_i) - \mu_{\ell j}(U_i)| \\
&\qquad = \max_{1 \le i \le n} \max_{1 \le j \le p} \left| \sum_{m=1}^{2q} \{\widehat{\Gamma}_{\ell m}(U_i) \widehat{\Phi}_{mj}(U_i) - \tilde{\Gamma}_{\ell m}(U_i) \tilde{\Phi}_{mj}(U_i)\} \right|. \quad (6.4)
\end{aligned}$$

It is easy to see that (6.4) is bounded by the following three parts:

$$\begin{aligned}
I_1 &= \max_{1 \le i \le n} \max_{1 \le j \le p} \left| \sum_{m=1}^{2q} \tilde{\Gamma}_{\ell m}(U_i) \{\widehat{\Phi}_{mj}(U_i) - \tilde{\Phi}_{mj}(U_i)\} \right|, \\
I_2 &= \max_{1 \le i \le n} \max_{1 \le j \le p} \left| \sum_{m=1}^{2q} \{\widehat{\Gamma}_{\ell m}(U_i) - \tilde{\Gamma}_{\ell m}(U_i)\} \tilde{\Phi}_{mj}(U_i) \right|, \\
I_3 &= \max_{1 \le i \le n} \max_{1 \le j \le p} \left| \sum_{m=1}^{2q} \{\widehat{\Gamma}_{\ell m}(U_i) - \tilde{\Gamma}_{\ell m}(U_i)\} \{\widehat{\Phi}_{mj}(U_i) - \tilde{\Phi}_{mj}(U_i)\} \right|.
\end{aligned}$$

In the following, we mainly consider the convergence rate of I_1. The convergence rates of I_2 and I_3 can be obtained in a similar way. Note that

$$\widehat{\Phi}_{mj}(U_i) = \begin{cases} \sum_{t=1}^{n} w_{nt}(U_i) X_{tm} Z_{tj}, & 1 \le m \le q, \\ \sum_{t=1}^{n} w_{nt}(U_i) X_{t(m-q)} Z_{tj}, & q+1 \le m \le 2q, \end{cases}$$

with

$$w_{nt}(U_i) = \begin{cases} \{nf(U_i)\}^{-1} K_h(U_t - U_i), & 1 \le m \le q, \\ \{nhf(U_i)\}^{-1} (U_t - U_i) K_h(U_t - U_i), & q+1 \le m \le 2q. \end{cases}$$

We also note that $\tilde{\Phi}_{mj}(U_i) = \Phi_{mj}(U_i)$ for $1 \le m \le q$ and $\tilde{\Phi}_{mj}(U_i) = 0$ for $q+1 \le m \le 2q$. Then, we can show that I_1 is bounded above by

$$I_{11} = \max_{1 \le i \le n} \max_{1 \le j \le p} \left| \sum_{m=1}^{q} \tilde{\Gamma}_{\ell m}(U_i) \{\widehat{\Phi}_{mj}(U_i) - \Phi_{mj}(U_i)\} \right|$$

and

$$I_{12} = \max_{1\le i\le n}\max_{1\le j\le p}\left|\sum_{m=q+1}^{2q}\tilde{\Gamma}_{\ell m}(U_i)\widehat{\Phi}_{mj}(U_i)\right|.$$

From Condition (C2), we get

$$\begin{aligned} I_{11} \le\ & C\max_{1\le i\le n}\max_{1\le j\le p}\max_{1\le m\le q}\left|\sum_{t=1}^{n} w_{nt}(U_i)\{X_{tm}Z_{tj}-\Phi_{mj}(U_t)\}\right| \\ & + C\max_{1\le i\le n}\max_{1\le j\le p}\max_{1\le m\le q}\left|\sum_{t=1}^{n} w_{nt}(U_i)\Phi_{mj}(U_t)-\Phi_{mj}(U_i)\right|. \end{aligned} \tag{6.5}$$

As $w_{nt}(U_i)$, $1\le t\le n$, satisfy the conditions of Lemma 6.3, we obtain that the convergence rate of the second term of (6.5) is $o_P(c_n)$.

Next we consider the convergence rate of the first term of (6.5). To this end, let A be a constant such that $A \ge n^{4/5}\max_{1\le t\le n} w_{nt}(U_i)/C$ with $A_n = C\sigma_v^2 n^{-4/5}$. It is easy to show that A_n and A satisfy the conditions of Lemma 6.1. For each $1\le m\le q$, we have

$$\begin{aligned} &\Pr\left\{\max_{1\le i\le n}\max_{1\le j\le p}\left|\sum_{t=1}^{n} w_{nt}(U_i)\left\{X_{tm}Z_{tj}-\Phi_{mj}(U_t)\right\}\right| > c_n\right\} \\ &\quad\le pn\Pr\left\{\left|\sum_{t=1}^{n} w_{nt}(U_i)\left\{X_{tm}Z_{tj}-\Phi_{mj}(U_t)\right\}\right| > c_n\right\} \\ &\quad\le 2pn\exp\left\{-\frac{c_n^2}{2A_n}(1-AM_\nu c_n)\right\} \le 2pn\exp\left(-\frac{c_n^2}{2A_n}\right) \\ &\quad= 2\exp\left\{-\frac{c_n^2}{2A_n}+\ln(pn)\right\} \\ &\quad= 2\exp\left\{-n^{3/10}\ln^{-2}(n/C)+\ln(pn)\right\}. \end{aligned} \tag{6.6}$$

Then, based on Condition (C8), we see that the convergence rate of the first term of (6.5) is $o_P(c_n)$. By the above results, then I_{12} is bounded above by

$$\begin{aligned} I_{12} \le\ & C\max_{1\le i\le n}\max_{1\le j\le p}\max_{1\le m_*\le q}\left|\sum_{t=1}^{n} w_{nt}(U_i)(X_{tm_*}Z_{tj}-\mu_{m_*j})\right| \\ & + C\max_{1\le i\le n}\max_{1\le j\le p}\max_{1\le m_*\le q}\left|\sum_{t=1}^{n} w_{nt}(U_i)\mu_{m_*j}\right|, \end{aligned} \tag{6.7}$$

where $m_* = m-q$ with $q+1\le m\le 2q$, and $\mu_{m_*j} = \mathrm{E}(X_{tm_*}Z_{tj})$. Invoking a similar argument for I_{11}, we can obtain the convergence rate of $I_{12} = o_P(c_n)$. Combined with the convergence rate of I_{11}, this yields $I_1 = o_P(c_n)$ for each $1\le \ell\le q$. Again using the same argument, we can prove that $I_2 = o_P(c_n)$ and $I_3 = o_P(c_n)$ for each $1\le \ell\le q$. This completes the proof of (6.2). □

LEMMA 6.5. *Suppose Conditions (C1)–(C9) hold, and let* $\tilde{m} = O(n^{2\xi_0+4\xi_{\min}})$ *with probability tending to* 1*, then we have*

$$2\tau_{\min} < \min_{|\mathcal{M}|\le\tilde{m}}\gamma_{\min}\{\widehat{\Sigma}_{(\mathcal{M})}\} \le \max_{|\mathcal{M}|\le\tilde{m}}\gamma_{\max}\{\widehat{\Sigma}_{(\mathcal{M})}\} < 2^{-1}\tau_{\max}. \tag{6.8}$$

Proof. Let $a = (a_1, \ldots, a_p)^\top$ be an arbitrary p-dimensional vector and $a_{(\mathcal{M})}$ be the sub-vector corresponding to $\mathcal{M}$. From Condition (C6), we get

$$\begin{aligned} 2\tau_{\min} &< \min_{\mathcal{M}\in\mathcal{M}_{\mathcal{F}}} \inf_{\|a_{(\mathcal{M})}\|=1} a_{(\mathcal{M})}^\top \Sigma_{(\mathcal{M})} a_{(\mathcal{M})} \\ &\leq \max_{\mathcal{M}\in\mathcal{M}_{\mathcal{F}}} \sup_{\|a_{(\mathcal{M})}\|=1} a_{(\mathcal{M})}^\top \Sigma_{(\mathcal{M})} a_{(\mathcal{M})} < 2^{-1}\tau_{\max}. \end{aligned}$$

Here we first consider to prove the maximum eigenvalue of $\widehat{\Sigma}_{\mathcal{M}}$ for $|\mathcal{M}| \leq \tilde{m}$ satisfying (6.8), the result of the minimum eigenvalue of $\widehat{\Sigma}_{(\mathcal{M})}$ for $|\mathcal{M}| \leq \tilde{m}$ can similarly be proved. In other words, we need to show that

$$\Pr\left(\max_{|\mathcal{M}|\leq\tilde{m}} \sup_{||a_{(\mathcal{M})}\|=1} \left| a_{(\mathcal{M})}^\top \{\widehat{\Sigma}_{(\mathcal{M})} - \Sigma_{(\mathcal{M})}\} a_{(\mathcal{M})} \right| > \epsilon \right) \longrightarrow 0, \tag{6.9}$$

where ϵ is an arbitrary positive number. Note that

$$\begin{aligned} &\Pr\left(\max_{|\mathcal{M}|\leq\tilde{m}} \sup_{\|a_{(\mathcal{M})}\|=1} \left| a_{(\mathcal{M})}^\top \{\widehat{\Sigma}_{(\mathcal{M})} - \Sigma_{(\mathcal{M})}\} a_{(\mathcal{M})} \right| > \epsilon \right) \\ &\quad\leq \Pr\left(\max_{|\mathcal{M}|\leq\tilde{m}} \sup_{\|a_{(\mathcal{M})}\|=1} \left| a_{(\mathcal{M})}^\top \{\widehat{\Sigma}_{(\mathcal{M})} - \Sigma^*_{(\mathcal{M})}\} a_{(\mathcal{M})} \right| > \frac{\epsilon}{2} \right) \\ &\qquad + \Pr\left(\max_{|\mathcal{M}|\leq\tilde{m}} \sup_{\|a_{(\mathcal{M})}\|=1} \left| a_{(\mathcal{M})}^\top \{\Sigma^*_{(\mathcal{M})} - \Sigma_{(\mathcal{M})}\} a_{(\mathcal{M})} \right| > \frac{\epsilon}{2} \right). \end{aligned} \tag{6.10}$$

Now we consider the first term of the right-hand side of (6.10). Note that

$$\begin{aligned} \widehat{\Sigma}_{(\mathcal{M})} - \Sigma^*_{(\mathcal{M})} &= \frac{1}{n}\{\widehat{\mathbf{Z}}_{(\mathcal{M})}^\top \widehat{\mathbf{Z}}_{(\mathcal{M})} - \mathbf{Z}^{*\top}_{(\mathcal{M})} \mathbf{Z}^*_{(\mathcal{M})}\} \\ &= \frac{1}{n}\{\widehat{\mathbf{Z}}_{(\mathcal{M})} - \mathbf{Z}^*_{(\mathcal{M})}\}^\top \{\widehat{\mathbf{Z}}_{(\mathcal{M})} - \mathbf{Z}^*_{(\mathcal{M})}\} + \frac{1}{n}\{\widehat{\mathbf{Z}}_{(\mathcal{M})} - \mathbf{Z}^*_{(\mathcal{M})}\}^\top \mathbf{Z}^*_{(\mathcal{M})} \\ &\quad + \frac{1}{n}\mathbf{Z}^{*\top}_{(\mathcal{M})} \{\widehat{\mathbf{Z}}_{(\mathcal{M})} - \mathbf{Z}^*_{(\mathcal{M})}\}. \end{aligned} \tag{6.11}$$

For any $\mathcal{M}$ with $|\mathcal{M}| \leq \tilde{m}$, we get

$$\begin{aligned} &\frac{1}{n}\left| a_{(\mathcal{M})}^\top \left[\mathbf{Z}^{*\mathrm{T}}_{(\mathcal{M})} \{\widehat{\mathbf{Z}}_{(\mathcal{M})} - \mathbf{Z}^*_{(\mathcal{M})}\} \right] a_{(\mathcal{M})} \right| \\ &\qquad = \frac{1}{n}\left| \sum_{k,j\in\mathcal{M}} a_k \mathbf{Z}^{*\top}_{(k)} \{\widehat{\mathbf{Z}}_{(j)} - \mathbf{Z}^*_{(j)}\} a_j \right| \\ &\qquad \leq \frac{1}{n} \sum_{k,j\in\mathcal{M}} |a_k| \left| \mathbf{Z}^{*\top}_{(k)} \{\widehat{\mathbf{Z}}_{(j)} - \mathbf{Z}^*_{(j)}\} \right| |a_j| \\ &\qquad = \frac{|\mathcal{M}|}{n} \max_{\substack{1\leq i\leq n \\ j\in\mathcal{M}}} |\widehat{Z}_{ij} - Z^*_{ij}| \sum_{k\in\mathcal{M}} a_k^2 \left\{1 + O_P(1)\right\}. \end{aligned} \tag{6.12}$$

Then, by Lemma 6.4 and the Cauchy–Schwarz inequality, we have

$$\frac{1}{n} \max_{|\mathcal{M}|\leq\tilde{m}} \sup_{\|a_{(\mathcal{M})}\|=1} \left| a_{(\mathcal{M})}^\top [\mathbf{Z}^{*\top}_{(\mathcal{M})} \{\widehat{\mathbf{Z}}_{(\mathcal{M})} - \mathbf{Z}^*_{(\mathcal{M})}\}] a_{(\mathcal{M})} \right| = o_P(c_n).$$

Using the same argument, we can prove that

$$\frac{1}{n}\max_{|\mathcal{M}|\le\tilde{m}}\sup_{\|a_{(\mathcal{M})}\|=1}\left|a_{(\mathcal{M})}^{\top}[\{\widehat{\mathbf{Z}}_{(\mathcal{M})}-\mathbf{Z}^{*}_{(\mathcal{M})}\}^{\top}\mathbf{Z}^{*}_{(\mathcal{M})}]a_{(\mathcal{M})}\right| = o_P(c_n),$$

$$\frac{1}{n}\max_{|\mathcal{M}|\le\tilde{m}}\sup_{\|a_{(\mathcal{M})}\|=1}\left|a_{(\mathcal{M})}^{\top}[\{\widehat{\mathbf{Z}}_{(\mathcal{M})}-\mathbf{Z}^{*}_{(\mathcal{M})}\}^{\top}\{\widehat{\mathbf{Z}}_{(\mathcal{M})}-\mathbf{Z}^{*}_{(\mathcal{M})}\}]a_{(\mathcal{M})}\right| = o_P(c_n).$$

For an arbitrary positive number ϵ, invoking the above results, we have

$$\Pr\left(\max_{|\mathcal{M}|\le\tilde{m}}\sup_{\|a_{(\mathcal{M})}\|=1}\left|a_{(\mathcal{M})}^{\top}\{\widehat{\Sigma}_{(\mathcal{M})}-\Sigma^{*}_{(\mathcal{M})}\}a_{(\mathcal{M})}\right|>\frac{\epsilon}{2}\right)\longrightarrow 0.$$

Next we consider the second term of the right-hand side of (6.10). Lemma 6.2 relaxes the normality assumption on the covariates imposed in Wang [35] and provides an exponential inequality for the sum of random variables. Therefore, using a similar technique as in Lemma 1 of Wang [35] and Lemma 6.2, for an arbitrary positive number ϵ, we have

$$\Pr\left(\max_{|\mathcal{M}|\le\tilde{m}}\sup_{\|a_{(\mathcal{M})}\|=1}\left|a_{(\mathcal{M})}^{\top}\{\Sigma^{*}_{(\mathcal{M})}-\Sigma_{(\mathcal{M})}\}a_{(\mathcal{M})}\right|>\frac{\epsilon}{2}\right)\longrightarrow 0.$$

Using the same argument, we can obtain the result of the minimum eigenvalue in (6.8). This completes the proof of Lemma 6.5. □

6.2. Proof of Theorem 3.1

Let $m^{*}\triangleq[Kn^{\xi_0+4\xi_{min}}]$. For each $k\le m^{*}$, after some algebraic operations, we have

$$\Omega(k)\triangleq \mathrm{RSS}(\mathcal{M}^{(k)})-\mathrm{RSS}(\mathcal{M}^{(k+1)})=\|\mathbf{H}^{(k)}_{a_{k+1}}\mathbf{Q}_{(\mathcal{M}^{(k)})}\widehat{\boldsymbol{Y}}\|^2, \tag{6.13}$$

where

$$\mathbf{Q}_{(\mathcal{M}^{(k)})}=\mathbf{I}_n-\mathbf{H}_{(\mathcal{M}^{(k)})},\qquad \mathbf{H}_{(\mathcal{M}^{(k)})}=\widehat{\mathbf{Z}}_{(\mathcal{M}^{(k)})}(\widehat{\mathbf{Z}}^{\top}_{(\mathcal{M}^{(k)})}\widehat{\mathbf{Z}}_{(\mathcal{M}^{(k)})})^{-1}\widehat{\mathbf{Z}}^{\top}_{(\mathcal{M}^{(k)})},$$
$$\mathbf{H}^{(k)}_{j}=\widehat{\mathbf{Z}}^{(k)}_{(j)}\widehat{\mathbf{Z}}^{(k)\top}_{(j)}\|\widehat{\mathbf{Z}}^{(k)}_{(j)}\|^{-2},\quad \widehat{\mathbf{Z}}^{(k)}_{(j)}=\{\mathbf{I}_n-\mathbf{H}_{(\mathcal{M}^{(k)})}\}\widehat{\mathbf{Z}}_{(j)},$$

and $a_{k+1}=\arg\min_{j\in\mathcal{M}_{\mathcal{F}}/\mathcal{M}^{(k)}}\mathrm{RSS}^{(k)}_{j}$, here $\widehat{\mathbf{Z}}_{(j)}$ is the jth column of $\widehat{\mathbf{Z}}$. Supposing $\mathcal{M}_{\mathcal{T}}\not\subset\mathcal{M}^{(m^{*})}$, we get

$$\Omega(k) \ge \max_{j\in\mathcal{M}^{*}_{k}}\|\mathbf{H}^{(k)}_{j}\mathbf{Q}_{(\mathcal{M}^{(k)})}\hat{\boldsymbol{Y}}\|^2\ge\|\mathbf{H}^{(k)}_{\hat{j}}\mathbf{Q}_{(\mathcal{M}^{(k)})}\hat{\boldsymbol{Y}}\|^2,$$

where $\mathcal{M}^{*}_{k}=\mathcal{M}_{\mathcal{T}}/\mathcal{M}^{(k)}\neq\emptyset$, and

$$\hat{j}=\arg\max_{j\in\mathcal{M}^{*}_{k}}\|\mathbf{H}^{(k)}_{j}\mathbf{Q}_{(\mathcal{M}^{(k)})}\{\mathbf{Z}^{*}_{(\mathcal{M}_{\mathcal{T}})}\boldsymbol{\beta}_{\mathcal{M}_{\mathcal{T}}}\}\|^2.$$

Thus, we have

$$\begin{aligned}\|\mathbf{H}^{(k)}_{\hat{j}}\mathbf{Q}_{(\mathcal{M}^{(k)})}\hat{\boldsymbol{Y}}\|^2 &\ge \|\mathbf{H}^{(k)}_{\hat{j}}\mathbf{Q}_{(\mathcal{M}^{(k)})}\{\mathbf{Z}^{*}_{(\mathcal{M}_{\mathcal{T}})}\boldsymbol{\beta}_{\mathcal{M}_{\mathcal{T}}}\}\|^2\\ &\quad-\|\mathbf{H}^{(k)}_{\hat{j}}\mathbf{Q}_{(\mathcal{M}^{(k)})}\boldsymbol{\varepsilon}\|^2-\|\mathbf{H}^{(k)}_{\hat{j}}\mathbf{Q}_{(\mathcal{M}^{(k)})}(\hat{\boldsymbol{Y}}-\boldsymbol{Y}^{*})\|^2\\ &\ge \max_{j\in\mathcal{M}^{*}_{k}}\|\mathbf{H}^{(k)}_{j}\mathbf{Q}_{(\mathcal{M}^{(k)})}\{\mathbf{Z}^{*}_{(\mathcal{M}_{\mathcal{T}})}\boldsymbol{\beta}_{\mathcal{M}_{\mathcal{T}}}\}\|^2-\max_{j\in\mathcal{M}_{\mathcal{T}}}\|\mathbf{H}^{(k)}_{j}\mathbf{Q}_{(\mathcal{M}^{(k)})}\boldsymbol{\varepsilon}\|^2\\ &\quad-\max_{j\in\mathcal{M}_{\mathcal{T}}}\|\mathbf{H}^{(k)}_{j}\mathbf{Q}_{(\mathcal{M}^{(k)})}(\hat{\boldsymbol{Y}}-\boldsymbol{Y}^{*})\|^2.\end{aligned} \tag{6.14}$$

Since $\mathbf{H}_j^{(k)}$ and $\mathbf{Q}_{(\mathcal{M}^{(k)})}$ are projection matrices, it follows from Lemma 6.4 that the third term on the right-hand side of (6.14) is bounded above by $\|\hat{\boldsymbol{Y}} - \boldsymbol{Y}^*\|^2 = n\, o_P(c_n^2)$. Thus, we only need to consider the first two terms on the right-hand side of (6.14).

Now we deal with the first term on the right-hand side of (6.14). Given that $\hat{\mathbf{Z}}_{(j)}^{(k)\mathrm{T}}\mathbf{Q}_{(\mathcal{M}^{(k)})} = \hat{\mathbf{Z}}_{(j)}^{\top}\mathbf{Q}_{(\mathcal{M}^{(k)})}$, we get

$$
\begin{aligned}
&\max_{j\in\mathcal{M}_k^*} \|\mathbf{H}_j^{(k)}\mathbf{Q}_{(\mathcal{M}^{(k)})}\{\mathbf{Z}^*_{(\mathcal{M}_\mathcal{T})}\boldsymbol{\beta}_{(\mathcal{M}_\mathcal{T})}\}\|^2 \\
&\qquad = \max_{j\in\mathcal{M}_k^*} \|\mathbf{H}_j^{(k)}\mathbf{Q}_{(\mathcal{M}^{(k)})}\{\mathbf{Z}^*_{(\mathcal{M}_k^*)}\boldsymbol{\beta}_{(\mathcal{M}_k^*)}\}\|^2 \\
&\qquad = \max_{j\in\mathcal{M}_k^*} \left\{ \|\hat{\mathbf{Z}}_{(j)}^{(k)}\|^{-2} \left| \hat{\mathbf{Z}}_{(j)}^{(k)\top}\mathbf{Q}_{(\mathcal{M}^{(k)})}\{\mathbf{Z}^*_{(\mathcal{M}_k^*)}\boldsymbol{\beta}_{(\mathcal{M}_k^*)}\} \right|^2 \right\} \\
&\qquad \geq \|\hat{\mathbf{Z}}_{(j^*)}^{(k)}\|^{-2} \left| \hat{\mathbf{Z}}_{(j^*)}^{\top}\mathbf{Q}_{(\mathcal{M}^{(k)})}\{\mathbf{Z}^*_{(\mathcal{M}_k^*)}\boldsymbol{\beta}_{(\mathcal{M}_k^*)}\} \right|^2 \\
&\qquad \geq \min_{j\in\mathcal{M}_k^*} \left\{ \|\hat{\mathbf{Z}}_{(j)}^{(k)}\|^{-2} \right\} \left| \hat{\mathbf{Z}}_{(j^*)}^{\top}\mathbf{Q}_{(\mathcal{M}^{(k)})}\{\mathbf{Z}^*_{(\mathcal{M}_k^*)}\boldsymbol{\beta}_{(\mathcal{M}_k^*)}\} \right|^2 \\
&\qquad = \max_{j\in\mathcal{M}_k^*} \left\{ \|\hat{\mathbf{Z}}_{(j)}^{(k)}\|^{2} \right\}^{-1} \left| \hat{\mathbf{Z}}_{(j^*)}^{\top}\mathbf{Q}_{(\mathcal{M}^{(k)})}\{\mathbf{Z}^*_{(\mathcal{M}_k^*)}\boldsymbol{\beta}_{(\mathcal{M}_k^*)}\} \right|^2 \\
&\qquad \geq \left(\max_{j\in\mathcal{M}_k^*} \|\hat{\mathbf{Z}}_{(j)}\|^2 \right)^{-1} \max_{j\in\mathcal{M}_k^*} \left| \hat{\mathbf{Z}}_{(j)}^{\top}\mathbf{Q}_{(\mathcal{M}^{(k)})}\{\mathbf{Z}^*_{(\mathcal{M}_k^*)}\boldsymbol{\beta}_{(\mathcal{M}_k^*)}\} \right|^2, \qquad (6.15)
\end{aligned}
$$

where $j^* = \arg\max_{j\in\mathcal{M}_k^*} |\hat{\mathbf{Z}}_{(j)}^{\top}\mathbf{Q}_{(\mathcal{M}^{(k)})}\{\mathbf{Z}^*_{(\mathcal{M}_k^*)}\boldsymbol{\beta}_{(\mathcal{M}_k^*)}\}|^2$, and (6.15) is due to the fact that $\|\hat{\mathbf{Z}}_{(j)}\| \geq \|\hat{\mathbf{Z}}_{(j)}^{(k)}\|$. Note that

$$
\|\mathbf{Q}_{(\mathcal{M}^{(k)})}\{\mathbf{Z}^*_{(\mathcal{M}_k^*)}\boldsymbol{\beta}_{(\mathcal{M}_k^*)}\}\|^2 = \sum_{j\in\mathcal{M}_k^*} \beta_j [\mathbf{Z}_{(j)}^{*\top}\mathbf{Q}_{(\mathcal{M}^{(k)})}\{\mathbf{Z}^*_{(\mathcal{M}_k^*)}\boldsymbol{\beta}_{(\mathcal{M}_k^*)}\}].
$$

By Condition (C7), and invoking the Cauchy–Schwarz inequality, we can further get

$$
\begin{aligned}
&\|\mathbf{Q}_{(\mathcal{M}^{(k)})}\{\mathbf{Z}^*_{(\mathcal{M}_k^*)}\boldsymbol{\beta}_{(\mathcal{M}_k^*)}\}\|^2 \\
&\qquad \leq \Big(\sum_{j\in\mathcal{M}_k^*} \beta_j^2 \Big)^{1/2} \Big[\sum_{j\in\mathcal{M}_k^*} |\mathbf{Z}_{(j)}^{*\top} Q_{(\mathcal{M}^{(k)})}\{\mathbf{Z}^*_{(\mathcal{M}_k^*)}\boldsymbol{\beta}_{(\mathcal{M}_k^*)}\}|^2 \Big]^{1/2} \\
&\qquad \leq C_{\boldsymbol{\beta}} p_0^{1/2} \max_{j\in\mathcal{M}_k^*} |\mathbf{Z}_{(j)}^{*\top}\mathbf{Q}_{(\mathcal{M}^{(k)})}\{\mathbf{Z}^*_{(\mathcal{M}_k^*)}\boldsymbol{\beta}_{(\mathcal{M}_k^*)}\}|. \qquad (6.16)
\end{aligned}
$$

On the other hand, we can see that

$$
\begin{aligned}
&\max_{j\in\mathcal{M}_k^*} \left| \mathbf{Z}_{(j)}^{*\top}\mathbf{Q}_{(\mathcal{M}^{(k)})}\{\mathbf{Z}^*_{(\mathcal{M}_k^*)}\boldsymbol{\beta}_{(\mathcal{M}_k^*)}\} \right| \\
&\qquad = \max_{j\in\mathcal{M}_k^*} \left| \{\mathbf{Z}_{(j)}^* - \hat{\mathbf{Z}}_{(j)} + \hat{\mathbf{Z}}_{(j)}\}^{\top}\mathbf{Q}_{(\mathcal{M}^{(k)})}\{\mathbf{Z}^*_{(\mathcal{M}_k^*)}\boldsymbol{\beta}_{(\mathcal{M}_k^*)}\} \right| \\
&\qquad \leq \max_{j\in\mathcal{M}_k^*} \left| \hat{\mathbf{Z}}_{(j)}^{\top}\mathbf{Q}_{(\mathcal{M}^{(k)})}\{\mathbf{Z}^*_{(\mathcal{M}_k^*)}\boldsymbol{\beta}_{(\mathcal{M}_k^*)}\} \right| \\
&\qquad\quad + \max_{j\in\mathcal{M}_k^*} \left| \{\mathbf{Z}_{(j)}^* - \hat{\mathbf{Z}}_{(j)}\}^{\top}\mathbf{Q}_{(\mathcal{M}^{(k)})}\{\mathbf{Z}^*_{(\mathcal{M}_k^*)}\boldsymbol{\beta}_{(\mathcal{M}_k^*)}\} \right|. \qquad (6.17)
\end{aligned}
$$

By Lemma 6.4, it is easy to show that, with probability tending to 1,

$$\|\mathbf{Q}_{(\mathcal{M}^{(k)})}\{\mathbf{Z}^*_{(\mathcal{M}^*_k)}\boldsymbol{\beta}_{(\mathcal{M}^*_k)}\}\|^2 \leq C_{\boldsymbol{\beta}} p_0^{1/2} \max_{j\in\mathcal{M}^*_k} \left|\hat{\mathbf{Z}}_{(j)}^\top \mathbf{Q}_{(\mathcal{M}^{(k)})}\{\mathbf{Z}^*_{(\mathcal{M}^*_k)}\boldsymbol{\beta}_{(\mathcal{M}^*_k)}\}\right|. \tag{6.18}$$

Then, by (6.18) along with Lemma 6.5 and Conditions (C7)–(C8), it is easy to show that the right-hand side of (6.15) is bounded below by

$$\left\{\max_{j\in\mathcal{M}^*_k}\|\hat{\mathbf{Z}}_{(j)}\|^2\right\}^{-1}[\|\mathbf{Q}_{(\mathcal{M}^{(k)})}\{\mathbf{Z}^*_{(\mathcal{M}^*_k)}\boldsymbol{\beta}_{(\mathcal{M}^*_k)}\}\|^2 C_{\boldsymbol{\beta}}^{-1} p_0^{-1/2}]^2$$
$$\geq \frac{1}{2} n^{-1}\tau_{\max}^{-1} p_0^{-1} C_{\boldsymbol{\beta}}^{-2} \|Q_{(\mathcal{M}^{(k)})}\{\mathbf{Z}^*_{(\mathcal{M}^*_k)}\boldsymbol{\beta}_{(\mathcal{M}^*_k)}\}\|^4. \tag{6.19}$$

Together (6.15) and (6.19) with Conditions (C7)–(C8), this leads to the conclusion that

$$\max_{j\in\mathcal{M}^*_k}\|\mathbf{H}_j^{(k)}\mathbf{Q}_{(\mathcal{M}^{(k)})}\{\mathbf{Z}^*_{(\mathcal{M}_\mathcal{T})}\boldsymbol{\beta}_{(\mathcal{M}_\mathcal{T})}\}\|^2 \geq \frac{1}{2} C_{\boldsymbol{\beta}}^{-2}\tau_{\max}^{-1}\tau_{\min}^2\nu_{\boldsymbol{\beta}}^4\nu^{-1}n^{1-\xi_0-4\xi_{\min}}. \tag{6.20}$$

Next we consider the second term on the right-hand side of (6.14). A simple calculation shows that

$$\begin{aligned}\max_{j\in\mathcal{M}_\mathcal{T}}\|\mathbf{H}_{(j)}^{(k)}\mathbf{Q}_{(\mathcal{M}^{(k)})}\boldsymbol{\varepsilon}\|^2 &= \max_{j\in\mathcal{M}_\mathcal{T}}\|\hat{\mathbf{Z}}_{(j)}^{(k)}\|^{-4}\|\hat{\mathbf{Z}}_{(j)}^{(k)}\hat{\mathbf{Z}}_{(j)}^{(k)\top}\mathbf{Q}_{(\mathcal{M}^{(k)})}\boldsymbol{\varepsilon}\|^2 \\ &\leq \tau_{\min}^{-1}n^{-1}\max_{j\in\mathcal{M}_\mathcal{T}}\max_{|\mathcal{M}|\leq m^*}\{\mathbf{Z}_{(j)}^{*\top}\mathbf{Q}_{(\mathcal{M})}\boldsymbol{\varepsilon}\}^2. \end{aligned}\tag{6.21}$$

It is noteworthy that $\mathbf{Z}_{(j)}^{*\top}\mathbf{Q}_{(\mathcal{M})}\boldsymbol{\varepsilon}$ is a normal random variable with mean $\mathbf{0}$ and variance $\|\mathbf{Q}_{(\mathcal{M})}\mathbf{Z}^*_{(j)}\|^2 \leq \|\mathbf{Z}^*_{(j)}\|^2$. Thus, the right-hand side of (6.21) can be bounded above by

$$\tau_{\min}^{-1}n^{-1}\max_{j\in\mathcal{M}_\mathcal{T}}\|\mathbf{Z}^*_{(j)}\|^2\sigma^2\max_{j\in\mathcal{M}_\mathcal{T}}\max_{|\mathcal{M}|\leq m^*}\chi_1^2,$$

where χ_1^2 stands for a chi-squared random variable with one degree of freedom. As is shown in Wang [35], $\max_{j\in\mathcal{M}_\mathcal{T}}\max_{|\mathcal{M}|\leq m^*}\chi_1^2$ is less than $3K\nu n^{\xi+\xi_0+4\xi_{\min}}$ with probability tending to 1. From Lemma 6.5, we know that

$$\max_{j\in\mathcal{M}_\mathcal{T}}\|\mathbf{Z}^*_{(j)}\|^2 \leq 2^{-1}n\tau_{\max}.$$

Then the second term of (6.14) is bounded by $2^{-1}\tau_{\max}\tau_{\min}^{-1}3K\nu n^{\xi+\xi_0+4\xi_{\min}}\sigma^2$. Combining this result with (6.14) and (6.20), we have

$$\frac{1}{n}\Omega(k) \geq \frac{\tau_{\min}^2\nu_{\boldsymbol{\beta}}^4}{2\nu\tau_{\max}C_{\boldsymbol{\beta}}^2}n^{-\xi_0-4\xi_{\min}}\left(1-\frac{\tau_{\max}^2C_{\boldsymbol{\beta}}^2\nu^2}{\tau_{\min}^3\nu_{\boldsymbol{\beta}}^4}3\sigma^2Kn^{\xi+2\xi_0+8\xi_{\min}-1}\right) \tag{6.22}$$

uniformly for $k\leq m^*$. Under Condition (C8), and recalling that $K = 4\nu\tau_{\max}C_{\boldsymbol{\beta}}^2/(\tau_{\min}^2\nu_{\boldsymbol{\beta}}^4)$, we have

$$n^{-1}\|\hat{\boldsymbol{Y}}\|^2 \geq n^{-1}\sum_{k=1}^{[Kn^{\xi_0+4\xi_{\min}}]}\Omega(k) \geq 2\left(1-\frac{\tau_{\max}^2C_{\boldsymbol{\beta}}^2\nu^2}{\tau_{\min}^3\nu_{\boldsymbol{\beta}}^4}3\sigma^2Kn^{\xi+2\xi_0+8\xi_{\min}-1}\right)\xrightarrow{P}2. \tag{6.23}$$

Without loss of generality, we further assume that $\mathrm{var}(Y_i^*) = 1$. Then according to Lemma 6.4, we have $n^{-1}\|\hat{\boldsymbol{Y}}\|^2\xrightarrow{P}1$, which contradicts the result of (6.23). Based on the assumption that $\mathcal{M}_\mathcal{T}\not\subset\mathcal{M}^{(m^*)}$ with $m^*\triangleq[Kn^{\xi_0+4\xi_{min}}]$, we reach a contradiction, i.e., the assumption is false which means that $\mathcal{M}_\mathcal{T}\subset\mathcal{M}^{(m^*)}$ with probability tending to 1. Therefore, the proof of Theorem 3.1 is complete. □

6.3. Proof of Theorem 3.2

By Theorem 3.1, we know that $\mathcal{M}_{\mathcal{T}} \subset \mathcal{M}^{[Kn^{\xi_0+4\xi_{\min}}]}$ with probability tending to 1. Thus we only need to show that

$$\Pr\left(\min_{\mathcal{M}_k^* \neq \emptyset, k \leq m^*} \{\mathrm{BIC}(k) - \mathrm{BIC}(k+1)\} > 0\right) \to 1, \tag{6.24}$$

where $\mathcal{M}_k^* = \mathcal{M}_{\mathcal{T}}/\mathcal{M}^{(k)} \neq \emptyset$ and $m^* = [Kn^{\xi_0+4\xi_{min}}]$. Note that

$$\begin{aligned}
&\mathrm{BIC}(\mathcal{M}^{(k)}) - \mathrm{BIC}(\mathcal{M}^{(k+1)}) \\
&\quad = \ln\left(\frac{\widehat{\sigma}^2_{(\mathcal{M}^{(k)})}}{\widehat{\sigma}^2_{(\mathcal{M}^{(k+1)})}}\right) - n^{-1}(\ln n + 2\zeta \ln p) \\
&\quad \geq \ln\left(1 + \frac{\widehat{\sigma}^2_{(\mathcal{M}^{(k)})} - \widehat{\sigma}^2_{(\mathcal{M}^{(k+1)})}}{\widehat{\sigma}^2_{(\mathcal{M}^{(k+1)})}}\right) - n^{-1}(1+2\zeta)\ln p,
\end{aligned} \tag{6.25}$$

where, for the last inequality, we have used the assumption of $p > n$. It is easy to see that $\widehat{\sigma}^2_{(\mathcal{M}^{(k+1)})} \leq n^{-1}||\widehat{\boldsymbol{Y}}||^2$, and by Lemma 6.4, we have $n^{-1}\|\widehat{\boldsymbol{Y}}\|^2 \xrightarrow{P} 1$. Then with probability tending to 1, the right-hand side of (6.25) is bounded below by

$$\ln\{1 + 2^{-1}n^{-1}\Omega(k)\} - n^{-1}(1+2\zeta)\ln p,$$

where the definition of $\Omega(k)$ is given in (6.13). According to the element inequality $\ln(1+x) \geq \min(\ln 2, x/2)$ and the inequality (6.22), the right-hand side of (6.25) is no less than, with probability tending to 1,

$$\begin{aligned}
&\min\{\ln 2, 4^{-1}n^{-1}\Omega(k)\} - n^{-1}(1+2\zeta)\ln p \\
&\qquad \geq \min\{\ln 2, 5^{-1}K^{-1}n^{-\xi_0-4\xi_{\min}}\} - n^{-1}(1+2\zeta)\ln p.
\end{aligned} \tag{6.26}$$

Note that, under Condition (C8), the right-hand side of (6.26) is positive with probability tending to 1 uniformly for $k \leq m^*, \mathcal{M}_k^* \neq \emptyset$. Hence the proof is complete. □

References

[1] Ahmad, I., Leelahanon, S. and Li, Q. (2005), Efficient estimation of a semiparametric partially linear varying coefficient model. *The Annals of Statistics*, **33**, 258–283.

[2] Chen, J.H. and Chen, Z.H. (2008), Extended bayesian information criterion for model selection with large model spaces. *Biometrika*, **95**, 759–771.

[3] Cheng, M.Y., Feng, S.Y., Li, G.R., and Lian, H. (2015), Greedy forward regression for variable screening. *Preprint*, arXiv:1511.01124.

[4] Cheng, M.Y., Honda, T. and Zhang, J.T. (2016), Forward variable selection for sparse ultra-high dimensional varying coefficient models. *Journal of the American Statistical Association*, **111**, 1209–1221.

[5] Chernoff, H. (1952), A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, **23**, 493–507.

[6] Cui, H.J., Li, R.Z. and Zhong, W. (2015), Model-free feature screening for ultra-high dimensional discriminant analysis. *Journal of the American Statistical Association*, **110**, 630–641.

[7] Dudoit, S., Fridlyand, J. and Speed, T.P. (2002), Comparsion of discriminantion methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, **97**,77–87.

[8] Fan, J.Q., Feng, Y. and Song, R. (2011), Nonparametric independence screening in sparse ultra-high dimensional additive models. *Journal of the American Statistical Association*, **106**, 544–557.

[9] Fan, J.Q. and Gijbels, I. (1996), *Local Polynomial Modeling and Its Applications*. London: Chapman and Hall.

[10] Fan, J.Q. and Huang, T. (2005), Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, **11**, 1031–1057.

[11] Fan, J.Q. and Li, R.Z. (2006), Statistical challenges with high-dimensionality: feature selection in knowledge discovery, Proceedings of International Congress of Mathematicians (M. Sanz-Solé, J. Soria, J.L. Varona, J. Verdera, eds.), Vol. III, 595–622.

[12] Fan, J.Q. and Lv, J.C. (2008), Sure independence screening for ultra-high dimensional feature space (with discussion). *Journal of the Royal Statistical Society, Series B*, **70**, 849–911.

[13] Fan, J.Q., Ma, Y.B. and Dai, W. (2014), Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *Journal of the American Statistical Association*, **109**, 1270–1284.

[14] Fan, J.Q., Samworth, R. and Wu, Y.C. (2009), Ultra-high dimensional feature selection: beyond the linear model. *Journal of Machine Learning Research*, **10**, 2013–2038.

[15] Fan, J.Q. and Song, R. (2010), Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*, **38**, 3567–3604.

[16] Gilliam, M., Rifas-Shiman, S., Berkey, C., Field, A. and Colditz, G. (2003), Maternal gestational diabetes, birth weight and adolescent obesity. *Pediatrics*, **111**, 221–226.

[17] Hong, Z.P., Hu, Y. and Lian, H. (2013), Variable selection for high-dimensional varying coefficient partially linear models via nonconcave penalty. *Metrika*, **76**, 887–908.

[18] Ishida, M., Monk, D., Duncan, A. J., Abu-Amero, S., Chong, J., Ring, S.M., Pembrey, M.E., Hindmarsh, P.C., Stanier, P. and Moore, G.E. (2012), Maternal inheritance of a promoter variant in the imprinted PHLDA2 gene significantly increases birth weight. *The American Journal of Human Genetics*, **90**, 715–719.

[19] Kai, B., Li, R.Z. and Zou, H. (2011), New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. *The Annals of Statistics*, **39**, 305–332.

[20] Lam, C. and Fan, J.Q. (2008), Profile-kernel likelihood inference with diverging number of parameters. *The Annals of Statistics*, **36**, 2232–2260.

[21] Li, G.R., Feng, S.Y. and Peng, H. (2011), Profile-type smoothed score function for a varying coefficient partially linear model. *Journal of Multivariate Analysis*, **102**, 372–385.

[22] Li, G.R., Feng, S.Y. and Zhang, J. (2016), *Modern Measurement Error Models*. Beijing: Science Press.

[23] Li, G.R., Lin, L. and Zhu, L.X. (2012), Empirical likelihood for varying coefficient partially linear model with diverging number of parameters. *Journal of Multivariate Analysis*, **105**, 85–111.

[24] Li, G.R., Peng, H., Zhang, J. and Zhu, L.X. (2012), Robust rank correlation based screening. *The Annals of Statistics*, **40**, 1846–1877.

[25] Li, G.R., Peng, H. and Zhu, L.X. (2011), Nonconcave penalized M-estimation with a diverging number of parameters. *Statistica Sinica*, **21**, 391–419.

[26] Li, G.R., Xue, L.G. and Lian, H. (2011), Semi-varying coefficient models with a diverging number of components. *Journal of Multivariate Analysis*, **102**, 1166–1174.

[27] Li, Q., Huang, C.J., Li, D. and Fu, T.T. (2002), Semiparametric smooth coefficient models. *Journal of Business & Economic Statistics*, **20**, 412–422.

[28] Li, R.Z. and Liang, H. (2008), Variable selection in semiparametric regression modeling. *The Annals of Statistics*, **36**, 261–286.

[29] Li, R.Z., Zhong, W. and Zhu, L.P. (2012), Feature screening via distance correlation learning. *Journal of the American Statistical Association*, **107**, 1129–1139.

[30] Lian, H. (2012), Variable selection for high-dimensional generalized varying-coefficient models. *Statistica Sinica*, **22**, 1563–1588.

[31] Liang, H., Wang, H.S. and Tsai, C.L. (2012), Profile forward regression for ultrahigh dimensional variable screening in semiparametric partially linear models. *Statistica Sinica*, **22**, 531–554.

[32] Liu, J.Y., Li, R.Z. and Wu, R.L. (2014), Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *Journal of the American Statistical Association*, **109**, 266–274.

[33] Sherwood, B. and Wang, L. (2016), Partially linear additive quantile regression in ultrahigh dimension. *The Annals of Statistics*, **44**, 288–317.

[34] Votavová, H., Dostálová Merkerová, M., Fejglova, K., Vašíková, A., Krejčík, Z., Pastorková, A., Tabashidze, N., Topinka, J., Velemínský, M.Jr., Šrám R.J. and Brdička, R. (2011), Transcriptome alterations in maternal and fetal cells induced by tobacco smoke. *Placenta*, **32**, 763–770.

[35] Wang, H.S. (2009), Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, **104**, 1512–1524.

[36] Wang, L.F., Li, H.Z. and Huang, J. (2008), Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association*, **103**, 1556–1569.

[37] Wu, C., Cui, Y.H. and Ma, S.G. (2014), Integrative analysis of gene-environment interactions under a multi-response partially linear varying coefficient model. *Statistics in Medicine*, **33**, 4988–4998.

[38] Xia, Y.C., Zhang, W.Y. and Tong, H. (2004), Efficient estimation for semivarying-coefficient models. *Biometrika*, **91**, 661–681.

[39] Xue, L.G. and Zhu, L.X. (2007), Empirical likelihood for a varying coefficient model with longitudinal data. *Journal of the American Statistical Association*, **102**, 642–654.

[40] You, J.H. and Chen, G.M. (2006), Estimation of a semiparametric varying-coefficient partially linear errors-in-variables model. *Journal of Multivariate Analysis*, **97**, 324–341.

[41] You, J.H. and Zhou, Y. (2006), Empirical likelihood for semiparametric varying-coefficient partially linear regression models. *Statistics & Probability Letters*, **76**, 412–422.

[42] Zhang, W.W., Li, G.R. and Xue, L.G. (2011), Profile inference on partially linear varying-coefficient errors-in-variables models under restricted condition. *Computational Statistics & Data Analysis*, **55**, 3027–3040.

[43] Zhao, P.X. and Xue, L.G. (2009), Variable selection for semiparametric varying coefficient partially linear models. *Statistics & Probability Letters*, **79**, 2148–2157.

[44] Zhao, W.H., Zhang, R.Q., Liu, J. and Lv, Y.Z. (2014), Robust and efficient variable selection for semiparametric partially linear varying coefficient model based on modal regression. *Annals of the Institute of Statistical Mathematics*, **66**, 165–191.

[45] Zhong, W.X., Zhang, T.T., Zhu, Y. and Liu, J.S. (2012), Correlation pursuit: forward stepwise variable selection for index models. *Journal of the Royal Statistical Society, Series B*, **74**, 849–870.

[46] Zhou, Y. and Liang, H. (2009), Statistical inference for semiparametric varying-coefficient partially linear models with generated regressors. *The Annals of Statistics*, **37**, 427–458.

[47] Zhu, L.P., Li, L.X., Li, R.Z. and Zhu, L.X. (2011), Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, **106**, 1464–1475.