Journal of Multivariate Analysis 167 (2018) 418–434

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Variable selection for partially linear models via partial correlation

Jingyuan Liu [a], Lejia Lou [b], Runze Li [c,*]

[a] Department of Statistics in School of Economics, Wang Yanan Institute for Studies in Economics and Fujian Key Laboratory of Statistical Science, Xiamen University, Xiamen, Fujian, 361005, China
[b] Ernst & Young, 5 Times Square, New York, NY 10036, USA
[c] Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, PA 16802, USA

ARTICLE INFO

Article history:
Received 13 September 2017
Available online 20 June 2018

AMS 2010 subject classifications:
62J05

Keywords:
Model selection consistency
Partial faithfulness
Semiparametric regression modeling

ABSTRACT

The partially linear model (PLM) is a useful semiparametric extension of the linear model that has been well studied in the statistical literature. This paper proposes a variable selection procedure for the PLM with ultrahigh dimensional predictors. The proposed method is different from the existing penalized least squares procedure in that it relies on partial correlation between the partial residuals of the response and the predictors. We systematically study the theoretical properties of the proposed procedure and prove its model consistency property. We further establish the root-n convergence of the estimator of the regression coefficients and the asymptotic normality of the estimate of the baseline function. We conduct Monte Carlo simulations to examine the finite-sample performance of the proposed procedure and illustrate the proposed method with a real data example.

1. Introduction

Let y be a response variable, u be a univariate continuous covariate and $\mathbf{x} = (x_1, \ldots, x_p)^\top$ be a p-dimensional covariate vector. The partially linear model (PLM) assumes that

$$y = g(u) + \mathbf{x}^\top \boldsymbol{\beta} + \epsilon, \tag{1}$$

where g is an unspecified baseline function, and $\boldsymbol{\beta}$ is a vector of unknown regression coefficients. The PLM thus assumes that the regression function linearly depends on the covariates $\mathbf{x}$ while depending nonparametrically on u. This model increases the flexibility of linear models by allowing the intercept to be a nonparametric function of the covariate u. It is one of the most popular semiparametric regression models in the literature [12].

This work aims to develop a variable selection procedure for the PLM with ultrahigh dimensional $\mathbf{x}$, i.e, $p = \mathcal{O}\{\exp(n^a)\}$ for some positive constant a, where n is the sample size. PLM estimation has been well studied in the case where p is finite and fixed; see, e.g., [7,13]. Variable selection procedures have also been developed in this case, e.g., by Fan and Li [5] via penalized least squares, and by Liang and Li [9] who employed the penalized least squares method for variable selection in the PLM in the presence of error in variables. Xie and Huang [15] studied the penalized least squares method with the SCAD penalty [4] for variable selection in the PLM with $p = o(\sqrt{n})$.

In this paper, we propose a new variable selection procedure for PLM. This procedure differs from the aforementioned penalized least squares methods in that it is a partial correlation learning procedure based on the notion of partial faithfulness

* Corresponding author.
E-mail address: rzli@psu.edu (R. Li).

https://doi.org/10.1016/j.jmva.2018.06.005
0047-259X/

that was first advocated by Bühlmann et al. [1] for normal linear models and further used for elliptical linear models in [8]. We first utilize partial residual techniques to eliminate the nonparametric baseline function, and then conduct variable selection by recursively testing the partial correlations between the partial residual of the response and that of the linear covariates. That is, we recursively compare the partial correlations with some threshold values, and therefore refer to this method as the thresholded partial correlation on partial residuals (TPC-PR). Thus, the TPC-PR can be carried out by using the algorithm proposed in [8].

This paper's main purpose is to study the theoretical properties of the TPC-PR, and to ensure that partial correlation learning works properly for the PLM. Developing the asymptotic theory of the TPC-PR is challenging since we have to deal with the approximation errors due to nonparametric estimation involved in the partial residuals. Furthermore, we need to study the partial faithfulness under the PLM setting without assuming normality. We first establish the concentration inequality of the partial correlations of the partial residuals. We then prove the model selection consistency of the TPC-PR under the PLM with ultrahigh dimensional $\mathbf{x}$. We further establish the $\sqrt{n}$-consistency of the regression coefficient estimate and the asymptotic normality of the nonparametric baseline estimation.

The rest of the paper is organized as follows. In Section 2, we discuss partial faithfulness under the PLM setting, and systematically study the asymptotic theory of partial correlations between partial residuals. We propose the TPC-PR in Section 3, and carefully study its theoretical properties. Section 4 provides the results of Monte Carlo studies and a real data example. Technical proofs are given in Section 5, along with the corresponding regularity conditions to facilitate the proofs. A conclusion is provided in Section 6.

2. Partial faithfulness and partial correlations for the PLM

In model (1), assume that the random error ϵ satisfies $\mathrm{E}(\epsilon|u) = \mathrm{E}(\epsilon|x_j) = 0$ for all $j \in \{1, \ldots, p\}$ and $\mathrm{E}(\epsilon^2) < \infty$. The objective is to recover the truly active model $\mathcal{A} = \{j \in \{1, \ldots, p\} : \beta_j \neq 0\}$ with cardinality $|\mathcal{A}|$, as well as to estimate $g(u)$ and the nonzero coefficients in $\boldsymbol{\beta}$. As most variable selection procedures do, we impose here a sparsity assumption, namely that $|\mathcal{A}| = \mathcal{O}(n^b)$, where b is defined in Theorems 1 and 2.

2.1. Partial faithfulness in partially linear models

We first discuss a partial faithfulness assumption for the PLM in (1) as a theoretical basis for our proposed variable selection procedure. This assumption, initially formulated by Bühlmann et al. [1], states that if the partial correlation between a given predictor and the response is zero given some other predictors, then the correlation between this predictor and the response is also zero given all other predictors. However, when the nonparametric baseline function is taken into consideration in model (1), this assumption is not directly applicable. Thus we need to apply first a partial residual technique to deal with the nonparametric part. Specifically, note that $\mathrm{E}(y|u) = g(u) + \mathrm{E}(\mathbf{x}^\top|u)\boldsymbol{\beta} + \mathrm{E}(\epsilon|u)$.

Model (1) can be written in the form

$$y^* = \mathbf{x}^{*\top}\boldsymbol{\beta} + \epsilon^*, \tag{2}$$

where $y^* = y - \mathrm{E}(y|u)$ and $\mathbf{x}^* = (x_1^*, \ldots, x_p^*)^\top = \mathbf{x} - \mathrm{E}(\mathbf{x}|u)$ are called partial residuals of y and $\mathbf{x}$ on u, and $\epsilon^* = \epsilon - \mathrm{E}(\epsilon|u)$. It is easy to show that when the covariance matrix of $\mathbf{x}^*$, denoted by $\boldsymbol{\Sigma}_{\mathbf{x}^*}$, is positive definite, we have

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\mathbf{x}^*}^{-1}\mathrm{cov}(y^*, \mathbf{x}^{*\top}). \tag{3}$$

Therefore, whatever $j \in \{1, \ldots, p\}$,

$$\beta_j = 0 \quad \Leftrightarrow \quad \rho(y^*, x_j^*|x_k^*, k \in \{j\}^{\mathrm{C}}) = 0, \tag{4}$$

where $\rho(z_1, z_2|z_3)$ represents the partial correlation between z_1 and z_2 after regressing on z_3, and $\{j\}^{\mathrm{C}} = \{1, \ldots, j-1, j+1, \ldots, p\}$. This provides the rationale for recovering the nonzero coefficients in $\mathcal{A}$ by evaluating partial correlations.

However, the computation of $\rho(y^*, x_j^*|x_k^*, k \in \{j\}^{\mathrm{C}})$ is infeasible under the high-dimensional setting when p is large. To address this issue, we adapt the concept of partial faithfulness [1] specifically for PLM, based on which we can convert the problem of evaluating $\rho(y^*, x_j^*|x_k^*, k \in \{j\}^{\mathrm{C}})$ to recursively assess the partial correlations with lower dimensions in a backward direction. The partial faithfulness of PLM is defined as follows.

Definition 1. The partially linear model (1) is said to be partially faithful if for every $j \in \{1, \ldots, p\}$, $\rho(y^*, x_j^*|x_{\mathcal{S}}^*) = 0$ for some $\mathcal{S} \subset \{j\}^{\mathrm{C}}$ implies that $\rho(y^*, x_j^*|x_k^*, k \in \{j\}^{\mathrm{C}}) = 0$, where x_j^* and y^* are defined in model (2), and $x_{\mathcal{S}}^* = \{x_j^* : j \in \mathcal{S}\}$ for some index set $\mathcal{S}$.

To fully understand the implications of the partial faithfulness, recall that (4) indicates that for every $j \in \{1, \ldots, p\}$ in model (2), $\beta_j = 0$ is equivalent to $\{\rho(y^*, x_j^*|x_k^*, k \in \{j\}^{\mathrm{C}}) = 0\}$. Therefore,

$$\beta_j = 0 \quad \Leftrightarrow \quad \rho(y^*, x_j^*|x_{\mathcal{S}}^*) = 0 \text{ for some } \mathcal{S} \subseteq \{j\}^{\mathrm{C}},$$

or equivalently,

$$\beta_j \neq 0 \quad \Leftrightarrow \quad \rho(y^*, x_j^* | x_{\mathcal{S}}^*) \neq 0 \text{ for all } \mathcal{S} \subseteq \{j\}^{\complement}.$$

That is, under the partial faithfulness assumption, the predictor x_j may be removed if there exists one subset $x_{\mathcal{S}}^*$ such that x_j^* is no longer needed when $x_{\mathcal{S}}^*$ is in the model. When the subset is taken to be empty, the marginal correlation should also be non-trivial. Thus the partial faithfulness rules out the situation where some predictors are marginally uncorrelated with the response, but possess joint effects with other covariates. This coincides with the assumption of any sure screening procedure, initiated by [6].

Lemma 1 provides sufficient conditions for partial faithfulness of PLM.

Lemma 1. *Assume that*

(A1) $\mathbf{\Sigma}_{\mathbf{x}^*}$ *is positive definite for all* u;
(A2) $\{\beta_j; j \in \mathcal{A}\} \sim f(b)db$, *where* f *denotes the density on a subset of* $\mathbb{R}^{|\mathcal{A}|}$ *of an absolutely continuous distribution with respect to Lebesgue measure.*

Then $(\mathbf{x}^*, y^*)$ *satisfies partial faithfulness almost surely with respect to the distribution generating non-zero regression coefficients.*

Conditions (A1) and (A2) are inspired by [1]. Condition (A1) guarantees the identifiability of $\boldsymbol{\beta}$ due to (3). Condition (A2) may be interpreted from a Bayesian point of view. We can treat nonzero β_js as independently and identically distributed random variables from a population with a non-trivial density. This condition is mild in the sense that from a Bayesian perspective, the zero coefficients can arise in an arbitrary fashion. We remark here that though already mild, (A1) and (A2) may not be the weakest conditions to guarantee partial faithfulness.

Based on Lemma 1, in order to identify nonzero β_js, it is sufficient to test recursively the above partial correlations with index set $\mathcal{S}$, with sequentially increasing cardinality $|\mathcal{S}|$. Lemma 1 is a direct corollary from Theorem 1 in [1].

2.2. Asymptotics of sample partial correlations for PLM

The problem of comparing $\rho(y^*, x_j^* | x_{\mathcal{S}}^*)$ with 0 becomes testing the null hypothesis $\mathcal{H}_0 : \rho(y^*, x_j^* | x_{\mathcal{S}}^*) = 0$ in practice. This requires to study the asymptotic performance of the estimated partial correlations $\hat{\rho}(y^*, x_j^* | x_{\mathcal{S}}^*)$, which is computed through several estimated conditional means. We first apply local linear regression [3] to estimate $\mathrm{E}(y|u)$ and $\mathrm{E}(\mathbf{x}|u)$ in y^* and $\mathbf{x}^*$ based on the random sample $(u_1, \mathbf{x}_1^\top, y_1), \ldots, (u_n, \mathbf{x}_n^\top, y_n)$. The smoothing matrix $\mathbf{S}(h)$ is computed as

$$\mathbf{S}(h) = \begin{pmatrix} (1,0)\{Z^\top(u_1)W(u_1,h)Z(u_1)\}^{-1}Z^\top(u_1)W(u_1,h) \\ \vdots \\ (1,0)\{Z^\top(u_n)W(u_n,h)Z(u_n)\}^{-1}Z^\top(u_n)W(u_n,h) \end{pmatrix},$$

where

$$Z(u) = \begin{pmatrix} 1 & u_1 - u \\ \vdots & \vdots \\ 1 & u_n - u \end{pmatrix} \quad \text{and} \quad W(u,h) = \mathrm{diag}\{K_h(u_1 - u), \ldots, K_h(u_n - u)\},$$

with $K_h(\cdot) = K(\cdot/h)/h$, and K being a kernel function with bandwidth h. Then the sample of the partial residuals y^* and $\mathbf{x}^*$ can be obtained by

$$\hat{\mathbf{y}}^* = \{\mathbf{1} - \mathbf{S}(h_y)\}\mathbf{y}, \quad \hat{\mathbf{X}}^* = [\{\mathbf{1} - \mathbf{S}(h_1)\}\mathbf{X}_1, \ldots, \{\mathbf{1} - \mathbf{S}(h_p)\}\mathbf{X}_p], \tag{5}$$

where $(\mathbf{X}_1, \ldots, \mathbf{X}_p) = \mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$, and $\mathbf{y} = (y_1, \ldots, y_n)^\top$, h_y and h_j are bandwidths for estimating $\mathrm{E}(y|u)$ and $\mathrm{E}(x_1|u), \ldots, \mathrm{E}(x_p|u)$. The marginal correlations between the partial residuals y^* and $x_1^*, \ldots, x_p^*$ are then estimated by the Pearson correlation between $\hat{\mathbf{y}}^*$ and each column of $\hat{\mathbf{X}}^*$, viz.

$$\hat{\rho}(y^*, x_j^*) = \frac{\langle \{\mathbf{1} - \mathbf{S}(h_y)\}\mathbf{y}, \{\mathbf{1} - \mathbf{S}(h_j)\}\mathbf{X}_j \rangle}{\|\{\mathbf{1} - \mathbf{S}(h_y)\}\mathbf{y}\| \|\{\mathbf{1} - \mathbf{S}(h_j)\}\mathbf{X}_j\|}.$$

Following [1], the partial correlations can be computed recursively, for any $k \in \mathcal{S}$, by

$$\hat{\rho}(y^*, x_j^* | x_{\mathcal{S}}^*) = \frac{\hat{\rho}(y^*, x_j^* | x_{\mathcal{S}\setminus\{k\}}^*) - \hat{\rho}(y^*, x_k^* | x_{\mathcal{S}\setminus\{k\}}^*)\hat{\rho}(x_j^*, x_k^* | x_{\mathcal{S}\setminus\{k\}}^*)}{[\{1 - \hat{\rho}^2(y^*, x_k^* | x_{\mathcal{S}\setminus\{k\}}^*)\}\{1 - \hat{\rho}^2(x_j^*, x_k^* | x_{\mathcal{S}\setminus\{k\}}^*)\}]^{1/2}}. \tag{6}$$

Next, we discuss the asymptotic normality of the partial correlations $\hat{\rho}(y^*, x_j^* | x_{\mathcal{S}}^*)$ under a partially linear model setting.

Lemma 2. *For any $j \in \{1, \ldots, p\}$ and $\mathcal{S} \subset \{j\}^{\complement}$, under regularity conditions (B1)–(B8) in Section 5, we have*

$$\sqrt{n}\,\{\hat{\rho}(y^*, x_j^*|x_{\mathcal{S}}^*) - \rho(y^*, x_j^*|x_{\mathcal{S}}^*)\} \rightsquigarrow \mathcal{N}[0, (1+\kappa)\{1-\rho^2(y^*, x_j^*|x_{\mathcal{S}}^*)\}^2],$$

where κ is the marginal kurtosis for distribution of $(u, \mathbf{x}, y)$, and $\rightsquigarrow$ denotes convergence in distribution.

3. Variable selection via partial correlations of partial residuals

Lemma 1 states that identifying nonzero coefficients is equivalent to recursively testing $\mathcal{H}_0 : \rho(y^*, x_j^*|x_{\mathcal{S}}^*) = 0$ for different $\mathcal{S}$. Lemma 2 further implies conducting Z test based on the asymptotically normal distribution of $\hat{\rho}(y^*, x_j^*|x_{\mathcal{S}}^*)$. We claim $\beta_j = 0$ and delete x_j if only one $\mathcal{S} \in \{j\}^{\complement}$ can be found such that $\mathcal{H}_0$ cannot be rejected for this $\mathcal{S}$.

Specifically, first set $\mathcal{S} = \emptyset$, and by Lemma 1, we can delete x_j if $\mathcal{H}_0$ is not rejected for x_j^* and $x_{\mathcal{S}}^* = \emptyset$, and obtain the first-step active set $\hat{\mathcal{A}}^{[1]}$. Note that only marginal utilities are involved in this step, hence the procedure for obtaining $\hat{\mathcal{A}}^{[1]}$ can be viewed as a feature screening technique for PLM. Among the candidate covariates x_j in $\hat{\mathcal{A}}^{[1]}$, we continue to assess the partial correlations given each individual x_k, $k \in \hat{\mathcal{A}}^{[1]}$. The insignificant x_js are further deleted, and the second-step active set $\hat{\mathcal{A}}^{[2]} \subseteq \hat{\mathcal{A}}^{[1]}$ is obtained. Then the partial correlations given two and more covariates in the current active set are evaluated in a sequential fashion. The procedure naturally stops when the cardinality of the given covariate set exceeds that of the current active set. Then any model fitting techniques for PLM in literature can be applied for estimating the nonzero coefficients of the linear term, as well as the nonparametric baseline function. For the sake of simplicity, the least squared estimates $\hat{\beta}_j$s are computed for the nonzero coefficients, and the nonparametric function is estimated by $\hat{g}(\mathbf{u}) = S(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ with the plug-in $\hat{\boldsymbol{\beta}}$. We summarize the whole procedure in Algorithm 1, in which we follow [8] to set

$$T(\alpha, n, \kappa, |\mathcal{S}|) = \frac{\exp\left\{2\sqrt{1+\kappa}\,\Phi^{-1}(1-\alpha/2)/\sqrt{n-|\mathcal{S}|-1}\right\} - 1}{\exp\left\{2\sqrt{1+\kappa}\,\Phi^{-1}(1-\alpha/2)/\sqrt{n-|\mathcal{S}|-1}\right\} + 1} \quad \text{and}$$

$$\hat{\kappa} = \frac{1}{p}\sum_{j=1}^{p}\left[\frac{n^{-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^4}{3\{n^{-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2\}^2} - 1\right],$$

where Φ^{-1} is the inverse function of the cumulative distribution function of $\mathcal{N}(0, 1)$, $\bar{x}_j$ is the sample mean of the $j$the element of $\mathbf{x}$, and x_{ij} is the jth element of $\mathbf{x}_i$.

Algorithm 1 TPC-PR Procedure for PLM

1. Compute the sample partial residuals $\hat{\mathbf{y}}^*$ and $\hat{\mathbf{X}}^*$ by (5).
2. Set $m = 1$, construct the first-step estimated active set by evaluating the marginal correlations between partial residuals:

   $\hat{\mathcal{A}}^{[1]} = \{j \in \{1, \ldots, p\} : |\hat{\rho}(y^*, x_j^*)| > T(\alpha, n, \hat{\kappa}, 0)\}$.
3. Let $m = m + 1$. Establish the mth-step estimated active set as

   $$\hat{\mathcal{A}}^{[m]} = \{j \in \hat{\mathcal{A}}^{[m-1]} : |\hat{\rho}(y^*, x_j^*|x_{\mathcal{S}}^*)| > T(\alpha, n, \hat{\kappa}, m-1),\ \forall_{\mathcal{S}\subseteq\hat{\mathcal{A}}^{[m-1]}\setminus\{j\}}\ |\mathcal{S}| = m-1\}.$$
4. Repeat Step 3 until $|\hat{\mathcal{A}}^{[m]}| \leq m$.
5. The estimated coefficient vector $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^{\top}$ is defined as follows: $\hat{\beta}_j = 0$ if $j \notin \hat{\mathcal{A}}^{[m]}$; $\hat{\beta}_j$ is the least squares estimate by regressing the partial residuals for $j \in \hat{\mathcal{A}}^{[m]}$.
6. Obtain the estimated nonparametric baseline function by $\hat{g}(\mathbf{u}) = \mathbf{S}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$.

We remark here that bandwidths h used in the smoothing matrix $\mathbf{S}(h)$s are chosen differently as h_y and $h_1, \ldots, h_p$ for estimating the conditional means of y and every x_j. Several bandwidth selection techniques can be adopted, and we use the plug-in method by [11] rather than the cross-validation method for saving computational cost. In addition, one needs to select the bandwidth again after the active predictors are detected and the nonparametric baseline function is refitted.

Next we discuss the theoretical properties of the proposed TPC-PR procedure. We first advocate the model selection consistency in the following theorem under the ultrahigh dimensional PLM setting.

Theorem 1. *Assume regularity conditions (B1)–(B8) in* Section 5 *for the partially linear model (1). Further assume the partial faithfulness from Definition 1. Then there exist a sequence $\alpha \to 0$ and constants a and b, where $0 < a + b < (1-2d)/5$ with $d \in (0, 1/2)$ such that for $p = \mathcal{O}\{\exp(n^a)\}$ and $|\mathcal{A}| = \mathcal{O}(n^b)$, we have $\Pr(\hat{\mathcal{A}}^{[m]} \neq \mathcal{A}) \to 0$, as $n \to \infty$.*

Theorem 1 implies that the selected model successfully captures the true one with probability tending to 1. In the proof of this theorem, we indeed show the probability of selected model being the true one tends to 1 at an exponential rate. The result is more challenging than [8] since the approximation error of partial residuals has to be taken into consideration. Theorem 2 states the $\sqrt{n}$-consistency of the estimated linear coefficients, as well as the asymptotical normality of the estimated nonparametric baseline function.

Theorem 2. *Under the same conditions as in Theorem 1, for $p = \mathcal{O}\{\exp(n^a)\}$ and $|\mathcal{A}| = \mathcal{O}(n^b)$, where a and b are defined as Theorem 1, we have $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| = \mathcal{O}_p(n^{-1/2})$, where $\|\cdot\|$ refers to the L_2 norm. Furthermore,*

$$\sqrt{nh}\,\{\hat{g}(u) - g(u) - g''(u)\mu_2 h^2/2\} \rightsquigarrow \mathcal{N}[0, \sigma^2\nu_0/f(u)],$$

as $n \to \infty$, $\mu_j = \int u^j K(u)du$, $\nu_0 = \int K^2(u)du$, and h is the bandwidth for computing the smoothing matrix for estimating the nonparametric function in the last step of the algorithm.

From Theorem 2, we further derive the asymptotic bias and variance of $\hat{g}(u)$ at any given value of u, viz.

$$\text{bias}(\hat{g}(u)) = g''(u)\mu_2 h^2/2 + o(h^2) + \mathcal{O}_p(n^{-1/2}), \quad \text{var}\{\hat{g}(u)\} = \frac{\sigma^2\nu_0}{nhf(u)}\{1 + o(1)\}.$$

Theorems 1–2 ensure the theoretical validity of using the selected TPC-PR model for subsequent inference.

4. Numerical studies

In this section, we conduct simulation studies to assess the finite-sample performance of TPC-PR and to empirically verify the theoretical properties stated in the last section. We then illustrate the proposed methodology on a real data example.

4.1. Simulations studies

We evaluate the performance of TPC-PR by comparing it to the penalized regression on partial residuals, with the SCAD penalty [4] and the LASSO penalty [14], respectively. That is, we transform the PLM to linear models via partial residual technique, followed by the penalized least squared estimation procedure. The nonparametric baseline is estimated in the same fashion as TPC-PR. Furthermore, the PC-simple algorithm proposed by [1] is also studied based on the partial residuals, and is denoted as PC-PR. The distinction between TPC-PR and PC-PR is that we take the kurtosis into consideration in TPC-PR, and hence the normality assumption is not necessary for conducting the algorithm. The PC-PR, however, relies heavily on the normal distribution of the error term when sequentially testing the partial correlations.

To further enhance the finite-sample performance of TPC-PR, we in practice may consider a fine tuning on the critical value. Specifically, we use $cT(\alpha, n, \hat{\kappa}, m)$ as the threshold, where c is the tuning parameter chosen by minimizing the extended Bayesian information criterion [2],

$$\ln(\hat{\sigma}^2) + df \times \ln(p) \times \ln(n)/n,$$

where $\hat{\sigma}^2$ is the estimated error variance of the PLM and df is the number of nonzero estimated coefficients. The modified TPC-PR is denoted by TPC-PR-EBIC.

We conduct the simulation study under three dimension settings: low dimension ($p = 20$), medium dimension ($p = 200$), and high dimension ($p = 500$), with sample size $n = 200$. For each setting, we consider two distributions: normal distribution and mixture normal distribution, to the covariate vector $\mathbf{x}$. The normal samples with autoregressive correlation are generated in the following fashion. We first draw $(x_1, \ldots, x_{p+1})^\top$ from a multivariate normal distribution $\mathcal{N}(0, \Sigma)$, where Σ is the covariance matrix with correlation $\rho^{|i-j|}$ between x_i and x_j, $\rho = 0.5$ and 0.8, and variance is taken to be 0.25 and 1 to represent strong and weak signal, respectively. Then let $u = \Phi(x_{p+1})$, where Φ is the cumulative distribution function of standard normal distribution, $\mathcal{N}(0, 1)$. By doing this, u and $\mathbf{x}$ are generated correlated, and u follows a uniform distribution. In the same routine, we draw random samples from a mixture of two normal distributions, viz. $0.9\mathcal{N}(0, \Sigma) + 0.1\mathcal{N}(0, 9\Sigma)$. The true coefficient vector $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, \ldots, 0)^\top$, and hence $\mathcal{A} = \{1, 2, 5\}$. Finally, we define the nonparametric baseline function $g(u) = u^2$ and $\sin(2\pi u)$ in two scenarios. The experiment is repeated 1000 times, and the following criteria are adopted to assess the performance of all procedures.

(a) For evaluating model selection consistency:

- False positive number (FP): the number of zero coefficients erroneously detected to be nonzero, i.e.,

$$\text{FP} = \sum_{j=1}^{p} \mathbf{1}(\hat{\beta}_j \neq 0, \beta_j = 0).$$

- True positive number (TP): the number of nonzero coefficients correctly detected to be nonzero, i.e.,

$$\text{TP} = \sum_{j=1}^{p} \mathbf{1}(\hat{\beta}_j \neq 0, \beta_j \neq 0).$$

Table 1
Simulation results for mixture of normals when $p = 500$ and $\rho = 0.5$.

Method	ME(Devi)	TP	FP	Under-fit	Cor-fit	Over-fit	RASE(Devi)
$\sigma^2 = 0.25, g(u) = u^2$							
SCAD	0.0074 (0.0042)	3.000	1.245	0.000	0.640	0.360	0.0917 (0.0247)
LASSO	0.0446 (0.0120)	3.000	32.170	0.000	0.000	1.000	0.1027 (0.0267)
PC-PR	0.0096 (0.0051)	3.000	0.260	0.000	0.760	0.240	0.0893 (0.0255)
TPC-PR	0.0066 (0.0040)	3.000	0.005	0.000	0.995	0.005	0.0895 (0.0250)
TPC-PR-EBIC	0.0066 (0.0040)	3.000	0.000	0.000	1.000	0.000	0.0895 (0.0251)
$\sigma^2 = 0.25, g(u) = \sin(2\pi u)$							
SCAD	0.0072 (0.0040)	3.000	1.215	0.000	0.650	0.350	0.1117 (0.0204)
LASSO	0.0649 (0.0213)	3.000	28.750	0.000	0.005	0.995	0.1541 (0.0322)
PC-PR	0.0142 (0.0100)	3.000	0.220	0.000	0.785	0.215	0.1351 (0.0279)
TPC-PR	0.0117 (0.0076)	2.990	0.015	0.010	0.990	0.000	0.1340 (0.0281)
TPC-PR-EBIC	0.0114 (0.0072)	3.000	0.000	0.000	1.000	0.000	0.1340 (0.0278)
$\sigma^2 = 1, g(u) = u^2$							
SCAD	0.0342 (0.0182)	3.000	4.965	0.000	0.535	0.465	0.1803 (0.0525)
LASSO	0.1736 (0.0430)	3.000	42.945	0.000	0.000	1.000	0.1913 (0.0488)
PC-PR	0.0605 (0.0274)	3.000	0.935	0.000	0.290	0.710	0.1763 (0.0513)
TPC-PR	0.0248 (0.0134)	3.000	0.040	0.000	0.960	0.040	0.1737 (0.0513)
TPC-PR-EBIC	0.0241 (0.0134)	3.000	0.005	0.000	0.995	0.005	0.1737 (0.0508)
$\sigma^2 = 1, g(u) = \sin(2\pi u)$							
SCAD	0.0329 (0.0170)	3.000	4.680	0.000	0.555	0.445	0.2010 (0.0395)
LASSO	0.1748 (0.0445)	3.000	41.800	0.000	0.000	1.000	0.2172 (0.0408)
PC-PR	0.0614 (0.0277)	3.000	0.920	0.000	0.295	0.705	0.1991 (0.0402)
TPC-PR	0.0257 (0.0140)	3.000	0.040	0.000	0.960	0.040	0.1968 (0.0402)
TPC-PR-EBIC	0.0250 (0.0136)	3.000	0.005	0.000	0.995	0.005	0.1947 (0.0419)

  - Under-fit percentage (Under-fit): the proportion of missing at least one of truly active covariates in the linear part.
  - Correctly-fit percentage (Cor-fit): the proportion of identifying exactly the truly active set.
  - Over-fit percentage (Over-fit): the proportion of identifying all the truly active covariates, but including at least one inactive covariate erroneously.

(b) For evaluating the $\sqrt{n}$-consistency of linear coefficients:

  - Model error (ME) due to the linear part: $\mathrm{ME} = \mathrm{E}[\{\mathbf{x}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\}^2] = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \mathrm{cov}(\mathbf{x})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$.

(c) For evaluating the performance of the estimated nonparametric baseline:

  - Square root of average squared errors (RASE) defined by

$$\mathrm{RASE} = \left\{ \frac{1}{N_g} \sum_{k=1}^{N_g} \{\hat{g}(v_k) - g(v_k)\}^2 \right\}^{1/2},$$

where $v_1, \ldots, v_{N_g}$ are the grid points at which the functions are evaluated, and N_g is the number of grid points.

The medians of ME and RASE, along with the respective medians of their absolute deviations (Devi) among the 1000 simulations are recorded. For other criteria, we report the average over the 1000 simulations.

We present the simulation results in Table 1 for the high-dimensional case ($p = 500$) with mixture normal distribution imposed on the error term and the correlation $\rho = 0.5$, and in Table 2 for that with $\rho = 0.8$. The rest of results are attached in the Online Supplement.

From Table 1–2, the methods developed in this paper (TPC-PR and TPC-PR-EBIC) can successfully identify the three truly active covariates (TP $\approx$ 3), with fairly low average numbers of falsely including inactive covariates (FP $\approx$ 0). Similarly, the correct-fit rates approach 1, illustrating the model selection consistency. The under-fit and over-fit rate are both close to 0. The model error (ME) and the square root of average squared errors (RASE) are small enough to show the $\sqrt{n}$-consistency of the coefficient estimations and the validity of the nonparametric baseline estimation.

In terms of comparison, the selection methods based on partial correlations favors sparser models than the penalized regression approaches in general. According to Tables 1 and 2, the last three methods consistently outperform the first two penalized regression approaches, especially LASSO, which, as expected, overfits data and yields conservative models. Although the correct-fit probability for SCAD (around 0.6) is much larger than LASSO, but is still left behind by the other three methods. The ME and RASE illustrate the same phenomenon, no matter which functional form of the baseline $g(u)$ is assumed.

Table 2
Simulation results for mixture of normals when $p = 500$ and $\rho = 0.8$.

Method	ME(Devi)	TP	FP	Under-fit	Cor-fit	Over-fit	RASE(Devi)
$\sigma^2 = 0.25, g(u) = u^2$							
SCAD	0.0100 (0.0057)	3.000	1.185	0.000	0.640	0.360	0.1121 (0.0316)
LASSO	0.0658 (0.0204)	3.000	29.150	0.000	0.000	1.000	0.1915 (0.0571)
PC-PR	0.0121 (0.0067)	3.000	0.195	0.000	0.815	0.185	0.1089 (0.0345)
TPC-PR	0.0107 (0.0064)	2.995	0.015	0.005	0.985	0.010	0.1089 (0.0332)
TPC-PR-EBIC	0.0101 (0.0060)	3.000	0.000	0.000	1.000	0.000	0.1084 (0.0328)
$\sigma^2 = 0.25, g(u) = \sin(2\pi u)$							
SCAD	0.0102 (0.0060)	3.000	1.120	0.000	0.660	0.340	0.1272 (0.0293)
LASSO	0.0902 (0.0300)	3.000	32.095	0.000	0.000	1.000	0.2326 (0.0555)
PC-PR	0.0150 (0.0099)	2.975	0.215	0.025	0.800	0.175	0.1527 (0.0313)
TPC-PR	0.0122 (0.0084)	2.950	0.065	0.050	0.935	0.015	0.1529 (0.0318)
TPC-PR-EBIC	0.0117 (0.0078)	2.995	0.020	0.005	0.980	0.015	0.1512 (0.0304)
$\sigma^2 = 1, g(u) = u^2$							
SCAD	0.0459 (0.0232)	3.000	4.160	0.000	0.515	0.485	0.2121 (0.0557)
LASSO	0.1736 (0.0430)	3.000	42.945	0.000	0.000	1.000	0.1913 (0.0488)
PC-PR	0.0605 (0.0274)	3.000	0.935	0.000	0.290	0.710	0.1763 (0.0513)
TPC-PR	0.0248 (0.0134)	3.000	0.040	0.000	0.960	0.040	0.1737 (0.0513)
TPC-PR-EBIC	0.0241 (0.0134)	3.000	0.005	0.000	0.995	0.005	0.1737 (0.0508)
$\sigma^2 = 1, g(u) = \sin(2\pi u)$							
SCAD	0.0456 (0.0235)	3.000	4.120	0.000	0.540	0.460	0.2277 (0.0504)
LASSO	0.2464 (0.0676)	3.000	37.635	0.000	0.000	1.000	0.3521 (0.0853)
PC-PR	0.0595 (0.0307)	3.000	0.730	0.000	0.430	0.570	0.2317 (0.0504)
TPC-PR	0.0343 (0.0193)	2.970	0.055	0.030	0.945	0.020	0.2296 (0.0550)
TPC-PR-EBIC	0.0327 (0.0176)	3.000	0.000	0.000	1.000	0.000	0.2290 (0.0544)

Among the three partial-correlation-based methods, PC-PR yields the worst results, especially when the noise-to-signal ratio is high ($\sigma^2 = 1$): PC-PR yields only about 29% among all experiments that can identify the true model, while the TPC-PR and TPC-PR-EBIC both exceed 95%. This is due to the fact that PC-PR uses the wrong variance estimation for the testing statistics when normality assumptions are not satisfied. Meanwhile, TPC-PR involves the kurtosis into the limiting distribution of the partial correlations, and hence corrects the variance estimations. Under the normal distribution setting (the corresponding results are provided in the Online Supplement), PC-PR and TPC-PR perform similarly, since the asymptotic distributions are identical for the two testing statistics under normality assumption. TPC-PR with EBIC fine tuning indeed enhances the finite-sample performances, although TPC-PR already behaves satisfactory under most circumstances and is sufficient for practical application. The same pattern can be observed for both signal-to-noise ratios and both values that correlation ρ takes.

Comparing Tables 1 and 2, we observe that as ρ increases from 0.5 to 0.8, it becomes slightly more challenging to identify the true covariates. The model tends to be overfitted and the covariates are highly correlated with each other. And in the high-correlation scenario, the improvement by TPC-PR-EBIC is more significant. Finally, when the signal-to-noise ratio increases from 0.25 to 1, PC-PR works dramatically worse, but the other methods behave relatively robust. We have also compared computing times of these methods. In general, the LASSO method uses the least computing time. For a simulated data set with $p = 500$, the TPC-PR takes about 20 s, while the LASSO takes 0.1 s, and the one-step SCAD takes about 3 s. It is clear that the TPC-PR is slower than the LASSO and the one-step SCAD due to the recursive nature of the proposed Algorithm 1. However, TPC-PR can be carried out with a reasonable amount of time.

4.2. Supermarket data analysis

In this section, we apply the TPC-PR method to analyze a high dimensional data set from a supermarket. The data set consists of 464 daily records of the number of customers entering the supermarket, as well as the sales volume of 6398 products in the market. We suspect nonlinear relation between the dates and popularity of the store, thus a partially linear model is a plausible choice for fitting the data.

We apply SCAD, LASSO, PC-PR, TPC-PR, and TPC-PR-EBIC as the stock example. The model sizes and the prediction errors are reported in Table 3. SCAD and LASSO still yields much more conservative models with 28 and 39 selected variables than the partial correlation based methods, while the corresponding prediction errors are higher. Compared with PC-PR and TPC-PR, the TPC-PR-EBIC model is even sparser, with a slight sacrifice of prediction error. The time effect on the number of customers are depicted in Fig. 1. Some periodic pattern is observed.

Table 3
Model sizes and prediction errors for the market data analysis.

Approach	Model size	Prediction error
SCAD	28	16.87
LASSO	39	17.53
PC-PR	10	16.36
TPC-PR	10	16.36
TPC-PR-EBIC	7	16.97

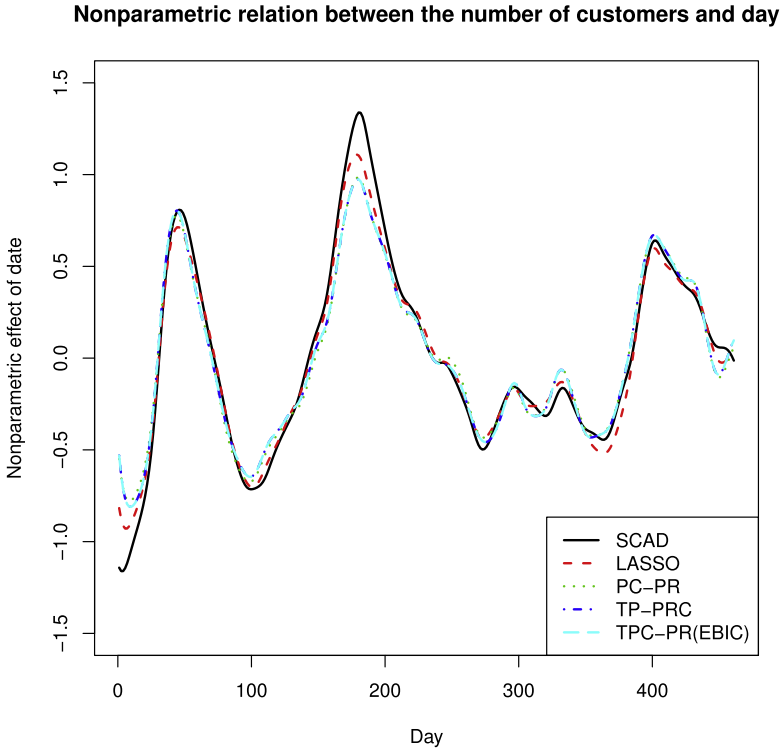

Fig. 1. The estimated curve of the number of customers against dates.

5. Conditions and proofs

5.1. Regularity conditions and lemmas

The following regularity conditions are imposed to facilitate the proofs.

(B1) For $j \in \{1, \ldots, p\}$, the conditional expectations $\mathrm{E}(y|u)$, $\mathrm{E}(x_j|u)$, $\mathrm{E}(yx_j|u)$, $\mathrm{E}(y^2|u)$, and $\mathrm{E}(x_j^2|u)$ are all uniformly bounded in $\mathbb{U}$, where $\mathbb{U}$ is the bounded support of u. Furthermore, we assume there exists $\delta_1 > 0$ such that (i) $\mathrm{E}\{\mathrm{var}(y|u)\} \geq \delta_1$ and (ii) $\mathrm{E}\{\mathrm{var}(x_j|u)\} \geq \delta_1$.

(B2) x_j and y satisfy the sub-exponential tail probability uniformly in u. That is, there exists $s_0 > 0$ such that for $s \in (0, s_0)$,

$$\sup_{u\in\mathbb{U}} \max_{j\in\{1,\ldots,p\}} \mathrm{E}\{\exp(sx_j^2)|u\} < \infty, \quad \sup_{u\in\mathbb{U}} \max_{j\in\{1,\ldots,p\}} \mathrm{E}\{\exp(sx_jy)|u\} < \infty, \quad \sup_{u\in\mathbb{U}} \mathrm{E}\{\exp(sy^2)|u\} < \infty.$$

(B3) The partial correlations $\rho(y^*, x_j^* | x_{\mathcal{S}}^*)$ satisfy

$$\inf\{|\rho(y^*, x_j^*|x_{\mathcal{S}}^*)| : j \in \{1, \ldots, p\}, \mathcal{S} \subseteq \{j\}^{\mathtt{C}}, |\mathcal{S}| \leq |\mathcal{A}|, \rho(y^*, x_j^*|x_{\mathcal{S}}^*) \neq 0\} \geq c_n,$$

where $c_n = \mathcal{O}(n^{-d})$, and $d \in (0, 1/2)$.

(B4) The partial correlations $\rho(y^*, x_j^*|x_{\mathcal{S}}^*)$ and $\rho(x_j^*, x_k^*|x_{\mathcal{S}}^*)$ satisfy

(i) $\sup\{|\rho(y^*, x_j^*|x_{\mathcal{S}}^*)| : j \in \{1, \ldots, p\}, \mathcal{S} \subseteq \{j\}^{\mathtt{C}}, |\mathcal{S}| \leq |\mathcal{A}|\} \leq \tau < 1$

(ii) $\sup\{|\rho(x_i^*, x_j^*|x_{\mathcal{S}}^*)| : i, j \in \{1, \ldots, p\}, i \neq j, \mathcal{S} \subseteq \{i, j\}^{\mathtt{C}}, |\mathcal{S}| \leq |\mathcal{A}|\} \leq \tau < 1.$

(B5) Let $f(u)$ be the density function of $u \in \mathbb{U}$. Assume $f(u)$ is bounded from 0; and $f(u)$, its first derivative $f'(u)$ and second derivative $f''(u)$ are bounded uniformly in u. That is, there exist $\delta_3 > 0$ and $M_1 > 0$ such that

$$\inf_{u\in\mathbb{U}}|f(u)| \geq \delta_3, \quad \sup_{u\in\mathbb{U}}|f(u)| \leq M_1, \quad \sup_{u\in\mathbb{U}}|f'(u)| \leq M_1, \quad \sup_{u\in\mathbb{U}}|f''(u)| \leq M_1.$$

(B6) The kernel function K is a symmetric density function, and it is bounded and has finite second moment, i.e.,

$$\sup_{u\in\mathbb{U}}|K(u)| < \infty \quad \text{and} \quad \int_{u\in\mathbb{U}} u^2K(u)du < \infty.$$

(B7) $(v, \mathbf{x}, y)$ has an elliptical distribution $\mathrm{EC}_{p+2}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi)$, and there exists a function ψ such that the σ-field generated by $u = \psi(v)$ is the same as that generated by v.

(B8) The bandwidths satisfy $h_y \to 0$, $h_j \to 0$, $nh_y^3 \to \infty$, and $nh_j^3 \to \infty$, for all $j \in \{1, \ldots, p\}$.

We further state some lemmas to facilitate the technical proofs of theorem. The proofs of Lemmas 3 to 5 follow the same techniques as in [10].

Lemma 3. *We adopt the following notation for simplicity:*

$$Z_1(u,h) = \frac{1}{n}\sum_{i=1}^{n} K_h(u_i - u), \quad Z_2(u,h) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{u_i - u}{h}\right)K_h(u_i - u),$$

$$Z_3(u,h) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{u_i - u}{h}\right)^2 K_h(u_i - u), \quad Z_4(u,h) = \frac{1}{n}\sum_{i=1}^{n} x_{ij}K_h(u_i - u),$$

$$Z_5(u,h) = \frac{1}{n}\sum_{i=1}^{n} x_{ij}\left(\frac{u_i - u}{h}\right)K_h(u_i - u), \quad Z_6(u,h) = \frac{1}{n}\sum_{i=1}^{n} y_iK_h(u_i - u),$$

$$Z_7(u,h) = \frac{1}{n}\sum_{i=1}^{n} y_i\left(\frac{u_i - u}{h}\right)K_h(u_i - u).$$

Then under conditions (B1), (B3)–(B6) and (B8), for some small $s > 0$, and any $\epsilon > 0$, we have the following results:

(1) $\sup_{u\in\mathbb{U}} \Pr\{|Z_1(u,h) - f(u)| > \epsilon\} \leq 4(1 - s\epsilon/4)^n$,

(2) $\sup_{u\in\mathbb{U}} \Pr\left\{|Z_2(u,h) - f(u)\int_{-\infty}^{+\infty} tK(t)dt| > \epsilon\right\} \leq 4(1 - s\epsilon/4)^n$,

(3) $\sup_{u\in\mathbb{U}} \Pr\left\{|Z_3(u,h) - f(u)\int_{-\infty}^{+\infty} t^2K(t)dt| > \epsilon\right\} \leq 4(1 - s\epsilon/4)^n$,

(4) $\sup_{u\in\mathbb{U}} \Pr\{|Z_4(u,h) - f(u)E(x_j|u)| > \epsilon\} \leq 4(1 - s\epsilon/4)^n$,

(5) $\sup_{u\in\mathbb{U}} \Pr\left\{|Z_5(u,h) - f(u)E(x_j|u)\int_{-\infty}^{+\infty} tK(t)dt| > \epsilon\right\} \leq 4(1 - s\epsilon/4)^n$,

(6) $\sup_{u\in\mathbb{U}} \Pr\{|Z_6(u,h) - f(u)E(y|u)| > \epsilon\} \leq 4(1 - s\epsilon/4)^n$,

(7) $\sup_{u\in\mathbb{U}} \Pr\left\{|Z_7(u,h) - f(u)E(y|u)\int_{-\infty}^{+\infty} tK(t)dt| > \epsilon\right\} \leq 4(1 - s\epsilon/4)^n$.

Lemma 4. *Assume $A(u)$ and $B(u)$ are two uniformly bounded functions of u. That is, there exist M_4 and M_5 such that*

$$\sup_{u\in\mathbb{U}}|A(u)| \leq M_4, \quad \sup_{u\in\mathbb{U}}|B(u)| \leq M_5.$$

For any given u, $\hat{A}(u)$ and $\hat{B}(u)$ are estimates of $A(u)$ and $B(u)$ based on n samples. Suppose there exist $C_1, \ldots, C_4$ and $q > 0$ such that

$$\sup_{u\in\mathbb{U}} \Pr\{|\hat{A}(u) - A(u)| > \epsilon\} \leq C_1n\{(1 - C_2\epsilon^2/n^{4q})^n + \exp(-C_3n^q)\},$$

$$\sup_{u\in\mathbb{U}} \Pr\{|\hat{B}(u) - B(u)| > \epsilon\} \leq C_1n\{(1 - C_2\epsilon^2/n^{4q})^n + \exp(-C_3n^q)\}.$$

Then

$$\sup_{u\in\mathbb{U}} \Pr\{|\hat{A}(u)\hat{B}(u) - A(u)B(u)| > \epsilon\} \leq C_1n\{(1 - C_2\epsilon^2/n^{4q})^n + \exp(-C_3n^q)\}.$$

Furthermore, if $\inf_{u\in\mathbb{U}}|B(u)| \geq \delta_3 > 0$, then

$$\sup_{u\in\mathbb{U}} \Pr\{|\hat{A}(u)/\hat{B}(u) - A(u)/B(u)| > \epsilon\} \leq C_1n\{(1 - C_2\epsilon^2/n^{4q})^n + \exp(-C_3n^q)\}.$$

$$\sup_{u\in\mathbb{U}} \Pr[|\{\hat{B}(u)\}^{1/2} - \{B(u)\}^{1/2}| > \epsilon] \leq C_1n\{(1 - C_2\epsilon^2/n^{4q})^n + \exp(-C_3n^q)\}.$$

Lemma 5. *Define*

$$W_j(u,h)=\frac{Z_3(u,h)Z_4(u,h)-Z_2(u,h)Z_5(u,h)}{Z_1(u,h)Z_3(u,h)-Z_2(u,h)Z_2(u,h)},\quad V(u,h)=\frac{Z_3(u,h)Z_6(u,h)-Z_2(u,h)Z_7(u,h)}{Z_1(u,h)Z_3(u,h)-Z_2(u,h)Z_2(u,h)}.$$

Then under the same conditions as Lemma 3, we have

(1) $\sup_{u\in\mathbb{U}}\Pr\{|W_j(u,h)-\mathrm{E}(x_j|u)|>\epsilon\}\le 4(1-s\epsilon/4)^n$.
(2) $\sup_{u\in\mathbb{U}}\Pr\{|V(u,h)-\mathrm{E}(y|u)|>\epsilon\}\le 4(1-s\epsilon/4)^n$.

5.2. Proof of Theorem 1

We divide the proof into six steps.

Step 1: First note that

$$\hat\rho(y^*,x_j^*|x_k^*)=\frac{\hat\rho(y^*,x_j^*)-\hat\rho(y^*,x_k^*)\hat\rho(x_j^*,x_k^*)}{[\{1-\hat\rho^2(y^*,x_k^*)\}\{1-\hat\rho^2(x_j^*,x_k^*)\}]^{1/2}}$$

is a function of $\hat\rho(y^*,x_j^*)$, $\hat\rho(y^*,x_k^*)$, and $\hat\rho(x_j^*,x_k^*)$. Let $g(x,y,z)=(x-yz)/\sqrt{(1-y^2)(1-z^2)}$, and $x,y,z\in(-1,1)$, then all the first and second derivatives are bounded away from 1, given y and z are bounded away from 1.

By Theorem 1 in [8],

$$\begin{aligned}\Pr\{|\hat\rho(y^*,x_j^*|x_k^*)-\rho(y^*,x_j^*|x_k^*)|>\epsilon\}&=\Pr\{|g\{\hat\rho(y^*,x_j^*),\hat\rho(y^*,x_k^*),\hat\rho(x_j^*,x_k^*)\}-g\{\rho(y^*,x_j^*),\rho(y^*,x_k^*),\rho(x_j^*,x_k^*)\}|>\epsilon\}\\&\le\Pr\left\{\left\|\begin{pmatrix}\hat\rho(y^*,x_j^*)\\\hat\rho(y^*,x_k^*)\\\hat\rho(x_j^*,x_k^*)\end{pmatrix}-\begin{pmatrix}\rho(y^*,x_j^*)\\\rho(y^*,x_k^*)\\\rho(x_j^*,x_k^*)\end{pmatrix}\right\|_2>C\epsilon\right\}\\&\le\Pr\{|\hat\rho(y^*,x_j^*)-\rho(y^*,x_j^*)|>C\epsilon/\sqrt{3}\}+\Pr\{|\hat\rho(y^*,x_k^*)-\rho(y^*,x_k^*)|>C\epsilon/\sqrt{3}\}\\&\qquad+\Pr\{|\hat\rho(x_j^*,x_k^*)-\rho(x_j^*,x_k^*)|>C\epsilon/\sqrt{3}\}\\&\le 3C_1n\{(1-C_2\epsilon^2/n^{4q})^n+\exp(-C_3n^q)\}.\end{aligned}$$

Similarly, for any $\mathcal{S}\subseteq\{j\}^{\mathfrak{C}}$, and $|\mathcal{S}|\le|\mathcal{A}|$, we can have

$$\Pr\{|\hat\rho(y^*,x_j^*|x_{\mathcal{S}}^*)-\rho(y^*,x_j^*|x_{\mathcal{S}}^*)|>\epsilon\}\le 3^{|\mathcal{S}|}C_1n\{(1-C_2\epsilon^2/n^{4q})^n+\exp(-C_3n^q)\}.$$

Step 2: If the distribution is assumed to be elliptical, we can use the sample version of the marginal kurtosis, viz.

$$\hat\kappa=\frac{1}{n}\sum_{j=1}^{p}\left[-1+\frac{1}{n}\sum_{i=1}^{n}(\tilde x_{ij}-\bar{\tilde x}_j)^4\Big/\left\{\frac{3}{n}\sum_{i=1}^{n}(\tilde x_{ij}-\bar{\tilde x}_j)^2\right\}^2\right],$$

to estimate the kurtosis. Similar to the proof in Step 1, we can obtain the following inequality:

$$\Pr(|\hat\kappa-\kappa|>\epsilon)\le C_4\exp(-C_3n^{1-4q}\epsilon^2)+C_2n\exp(-C_1n^q).$$

Step 3: To study $\Pr(|\hat Z_n(y^*,x_j^*|x_{\mathcal{S}}^*)/\sqrt{1+\hat\kappa}-Z_n(y^*,x_j^*|x_{\mathcal{S}}^*)/\sqrt{1+\kappa}|>\epsilon)$, define $g_2(u,v)=\ln\{(1+u)/(1-u)\}/(2\sqrt{1+v})$ for all $u\in(-1,1)$ and $v\in(-1,\infty)$. Then

$$\hat Z_n(y^*,x_j^*|x_{\mathcal{S}}^*)/\sqrt{1+\hat\kappa}=g_2\{\hat\rho(y^*,x_j^*|x_{\mathcal{S}}^*),\hat\kappa\}\quad\text{and}\quad Z_n(y^*,x_j^*|x_{\mathcal{S}}^*)/\sqrt{1+\kappa}=g_2\{\rho(y^*,x_j^*|x_{\mathcal{S}}^*),\kappa\}.$$

All the first and second derivatives are continuous and bounded for $u\in(-\tau,\tau)$, $v\in(-\delta,+\infty)$. By Theorem 1 in [8], under (B3) and (B5),

$$\begin{aligned}\Pr\left\{\left|\frac{\hat Z_n(y^*,x_j^*|x_{\mathcal{S}}^*)}{\sqrt{1+\hat\kappa}}-\frac{Z_n(y^*,x_j^*|x_{\mathcal{S}}^*)}{\sqrt{1+\kappa}}\right|>\epsilon\right\}&\le\Pr\left\{\left\|\begin{pmatrix}\hat\rho(y^*,x_j^*|x_{\mathcal{S}}^*)\\\hat\kappa\end{pmatrix}-\begin{pmatrix}\rho(y^*,x_j^*|x_{\mathcal{S}}^*)\\\kappa\end{pmatrix}\right\|>C\epsilon\right\}\\&\le\Pr(|\hat\rho(y^*,x_j^*|x_{\mathcal{S}}^*)-\rho(y^*,x_j^*|x_{\mathcal{S}}^*)|>C\epsilon/\sqrt{2})+\Pr(|\hat\kappa-\kappa|>C\epsilon/\sqrt{2})\\&\le(3^{|\mathcal{S}|}+1)C_1n\{(1-C_2\epsilon^2/n^{4q})^n+\exp(-C_3n^q)\}\\&\le 3^{|\mathcal{S}|}C_1n\{(1-C_2\epsilon^2/n^{4q})^n+\exp(-C_3n^q)\}.\end{aligned}$$

Step 4: Next we compute $\Pr(E_{j|\mathcal{S}})$. When testing the jth predictor given $\mathcal{S}\subseteq\{j\}^{\mathfrak{C}}$, denote the event

$$E_{j|\mathcal{S}}=\{\text{an error occurs when testing }\rho(y^*,x_j^*|x_{\mathcal{S}}^*)=0\}=E_{j|\mathcal{S}}^{I}\cup E_{j|\mathcal{S}}^{II},$$

where $E^I_{j|\mathcal{S}}$ denotes Type I error while $E^{II}_{j|\mathcal{S}}$ represents Type II error. We have

$$E^I_{j|\mathcal{S}} = \{(n-|\mathcal{S}|-1)^{1/2}|\hat{Z}_n(y^*, x_j^*|x^*_{\mathcal{S}})/\sqrt{1+\hat{\kappa}}| > \Phi^{-1}(1-\alpha/2) \text{ when } Z_n(y^*, x_j^*|x_{\mathcal{S}}) = 0\}.$$

Then

$$\begin{aligned}
\Pr(E^I_{j|\mathcal{S}}) &= \Pr\left\{(n-\mathcal{S}|-1)^{1/2}\left|\frac{\hat{Z}_n(y^*, x_j^*|x^*_{\mathcal{S}})}{\sqrt{1+\hat{\kappa}}}\right| > \Phi^{-1}(1-\alpha/2) \text{ when } Z_n(y^*, x_j^*|x^*_{\mathcal{S}}) = 0\right\} \\
&\le \Pr\left\{(n-|\mathcal{S}|-1)^{1/2}\left|\frac{\hat{Z}_n(y^*, x_j^*|x^*_{\mathcal{S}})}{\sqrt{1+\hat{\kappa}}} - \frac{Z_n(y^*, x_j^*|x^*_{\mathcal{S}})}{\sqrt{1+\kappa}}\right| > \Phi^{-1}(1-\alpha/2)\right\} \\
&= \Pr\left\{\left|\frac{\hat{Z}_n(y^*, x_j^*|x^*_{\mathcal{S}})}{\sqrt{1+\hat{\kappa}}} - \frac{Z_n(y^*, x_j^*|x^*_{\mathcal{S}})}{\sqrt{1+\kappa}}\right| > \sqrt{\frac{n}{n-|\mathcal{S}|-1}}\frac{c_n}{2\sqrt{1+\kappa}}\right\} \\
&\le \Pr\left\{\left|\frac{\hat{Z}_n(y^*, x_j^*|x^*_{\mathcal{S}})}{\sqrt{1+\hat{\kappa}}} - \frac{Z_n(y^*, x_j^*|x^*_{\mathcal{S}})}{\sqrt{1+\kappa}}\right| > \frac{c_n}{2\sqrt{1+\kappa}}\right\} \\
&\le 3^{|\mathcal{S}|}C_1 n\{(1-C_2c_n^2/n^{4q})^n + \exp(-C_3n^q)\},
\end{aligned}$$

by choosing $\alpha = 2\{1-\Phi(c_n\sqrt{n/(1+\kappa)}/2)\}$. Furthermore,

$$E^{II}_{j|\mathcal{S}} = \{(n-|\mathcal{S}|-1)^{1/2}|\hat{Z}_n(y^*, x_j^*|x^*_{\mathcal{S}})/\sqrt{1+\hat{\kappa}}| \le \Phi^{-1}(1-\alpha/2) \text{ when } Z_n(y^*, x_j^*|x^*_{\mathcal{S}}) \ne 0\}.$$

By choosing $\alpha = 2[1-\Phi\{\sqrt{n/(1+\kappa)}\,c_n/2\}]$, we can get the following inequality:

$$\begin{aligned}
\Pr(E^{II}_{j|\mathcal{S}}) &= \Pr\{(n-|\mathcal{S}|-1)^{1/2}|\hat{Z}_n(y^*, x_j^*|x^*_{\mathcal{S}})/\sqrt{1+\hat{\kappa}}| \le \Phi^{-1}(1-\alpha/2) \text{ when } Z_n(y^*, x_j^*|x^*_{\mathcal{S}}) \ne 0\} \\
&= \Pr\left\{\left|\frac{\hat{Z}_n(y^*, x_j^*|x^*_{\mathcal{S}})}{\sqrt{1+\hat{\kappa}}}\right| \le \sqrt{\frac{n}{n-|\mathcal{S}|-1}}\frac{c_n}{2\sqrt{1+\kappa}} \text{ when } Z_n(y^*, x_j^*|x^*_{\mathcal{S}}) \ne 0\right\} \\
&\le \Pr\left\{\left|\frac{Z_n(y^*, x_j^*|x^*_{\mathcal{S}})}{\sqrt{1+\kappa}}\right| - \left|\frac{\hat{Z}_n(y^*, x_j^*|x^*_{\mathcal{S}})}{\sqrt{1+\hat{\kappa}}} - \frac{Z_n(y^*, x_j^*|x^*_{\mathcal{S}})}{\sqrt{1+\kappa}}\right| \le \sqrt{\frac{n}{n-|\mathcal{S}|-1}}\frac{c_n}{2\sqrt{1+\kappa}}\right\} \\
&= \Pr\left\{\left|\frac{\hat{Z}_n(y^*, x_j^*|x^*_{\mathcal{S}})}{\sqrt{1+\hat{\kappa}}} - \frac{Z_n(y^*, x_j^*|x^*_{\mathcal{S}})}{\sqrt{1+\kappa}}\right| \ge \left|\frac{Z_n(y^*, x_j^*|x^*_{\mathcal{S}})}{\sqrt{1+\kappa}}\right| - \sqrt{\frac{n}{n-|\mathcal{S}|-1}}\frac{c_n}{2\sqrt{1+\kappa}}\right\}.
\end{aligned}$$

Let $g_3(u) = (1/2)\times\ln\{(1+u)/(1-u)\}$, then $|g_3(u)| = |1/2\times\ln\{(1+u)/(1-u)\}| \ge |u|$, for all $u\in(-1,1)$, and according to (B4), $|\rho(y^*, x_j^*|x^*_{\mathcal{S}})| \ge c_n$, then

$$\begin{aligned}
\Pr(E^{II}_{j|\mathcal{S}}) &\le \Pr\left\{\left|\frac{\hat{Z}_n(y^*, x_j^*|x^*_{\mathcal{S}})}{\sqrt{1+\hat{\kappa}}} - \frac{Z_n(y^*, x_j^*|x^*_{\mathcal{S}})}{\sqrt{1+\kappa}}\right| \ge \frac{c_n}{\sqrt{1+\kappa}} - \sqrt{\frac{n}{n-|\mathcal{S}|-1}}\frac{c_n}{2\sqrt{1+\kappa}}\right\} \\
&\le \Pr\left\{\left|\frac{\hat{Z}_n(y^*, x_j^*|x^*_{\mathcal{S}})}{\sqrt{1+\hat{\kappa}}} - \frac{Z_n(y^*, x_j^*|x^*_{\mathcal{S}})}{\sqrt{1+\kappa}}\right| \ge \frac{c_n}{\sqrt{1+\kappa}}\left(1-\sqrt{\frac{n}{n-|\mathcal{S}|-1}}\frac{1}{2}\right)\right\} \\
&\le \Pr\left\{\left|\frac{\hat{Z}_n(y^*, x_j^*|x^*_{\mathcal{S}})}{\sqrt{1+\hat{\kappa}}} - \frac{Z_n(y^*, x_j^*|x^*_{\mathcal{S}})}{\sqrt{1+\kappa}}\right| \ge \frac{3c_n}{8\sqrt{1+\kappa}}\right\} \\
&\le 3^{|\mathcal{S}|}C_1 n\{(1-C_2c_n^2/n^{4q})^n + \exp(-C_3n^q)\},
\end{aligned}$$

as for large n, $\sqrt{n/(n-|\mathcal{S}|-1)} \le 5/4$. Combining the above results, we have

$$\Pr(E_{j|\mathcal{S}}) = \Pr(E^I_{j|\mathcal{S}}) + \Pr(E^{II}_{j|\mathcal{S}}) \le 3^{|\mathcal{S}|}C_1 n\{(1-C_2c_n^2/n^{4q})^n + \exp(-C_3n^q)\}.$$

Step 5: To study $\Pr(\hat{\mathcal{A}}^{[m]} \ne \mathcal{A})$, we consider all $j\in\{1,\ldots,p\}$ and all $\mathcal{S}\subseteq\{j\}^{\complement}$ subject to $|\mathcal{S}|\le|\mathcal{A}|$, for any $b>0$, under the partial faithfulness assumption and (B2), for $j\in\{1,\ldots,p\}$, define $K_j = \{\mathcal{S}\subseteq\{j\}^{\complement} : |\mathcal{S}|\le|\mathcal{A}|\}$. Now

$$\begin{aligned}
\Pr\{\hat{\mathcal{A}}^{[m]} \ne \mathcal{A}\} &= \Pr\{\text{an error occurs for some } j \text{ and some } \mathcal{S}\} \\
&= \Pr\left\{\bigcup_{j\in\{1,\ldots,p\},\mathcal{S}\in K_j} E_{j|\mathcal{S}}\right\} \le \sum_{j\in\{1,\ldots,p\},\mathcal{S}\in K_j}\Pr(E_{j|\mathcal{S}}) \\
&\le p\times p^{|\mathcal{A}|}\sup_{j\in\{1,\ldots,p\},\mathcal{S}\in K_j}\Pr(E_{j|\mathcal{S}}) \le 3^{|\mathcal{A}|}p^{|\mathcal{A}|+1}C_1 n\{(1-C_2c_n^2/n^{4q})^n + \exp(-C_3n^q)\}.
\end{aligned}$$

The second last inequality holds since the number of possible choices of j is p, and there are $p^{|\mathcal{A}|}$ possible choices for $\mathcal{S}$. Furthermore, similar to lemma 3 in [1], we can show that $\Pr(m = |\mathcal{A}|) \to 1$.

Step 6: Under (B2), for $p = \mathcal{O}\{\exp(n^a)\} = C_4 \exp(n^a)$, and $|\mathcal{A}| = \mathcal{O}(n^b) = C_5 n^b$, we have

$$\Pr(\hat{\mathcal{A}}^{[m]} \neq \mathcal{A}) \leq 3^{|\mathcal{A}|} C_1 n (p)^{|\mathcal{A}|+1} C_1 n \{(1 - C_2 c_n^2 / n^{4q})^n + \exp(-C_3 n^q)\}.$$

Therefore,

$$\begin{aligned}\ln\{\Pr(\hat{\mathcal{A}}^{[m]} \neq \mathcal{A})\} &\leq |\mathcal{A}| \ln 3 + (|\mathcal{A}| + 1)\ln(p) + \ln(C_1 n) + \ln\{(1 - C_2 c_n^2 / n^{4q})^n + \exp(-C_3 n^q)\}. \\ &\leq C_1(|\mathcal{A}| + 1)\ln(p) + \ln\{\{(1 - C_2 c_n^2 / n^{4q})^n + \exp(-C_3 n^q)\}\} \\ &\leq C_1(C_5 n^b + 1)(\ln C_4 + n^a) + \ln\{(1 - C_2 c_n^2 / n^{4q})^n + \exp(-C_3 n^q)\}.\end{aligned}$$

Note that $e^{-x} \leq 1 - x/2$, for $x \in [0, \ln 2]$. Since $q \in (0, 1)$, then $C_3 / n^{1-q} \in [0, \ln 2]$.

$$\exp(-C_3 n^q) = \{\exp(-C_3 n^q / n)\}^n \leq (1 - C_3 / 2n^{1-q})^n = (1 - C_3 / n^{1-q})^n.$$

Therefore,

$$\ln\{\Pr(\hat{\mathcal{A}}^{[m]} \neq \mathcal{A})\} \leq C_1(C_5 n^b + 1)(\ln C_4 + n^a) + \ln\{(1 - C_2 / n^{2d+4q})^n + (1 - C_3 / n^{1-q})^n\}.$$

Let $(1 - 2d)/5 < q < (1 - a - b - 2d)/4$, then $1/n^{2d+4q} < 1/n^{1-q}$. Thus $1 - C_2/n^{2d+4q} \geq 1 - C_3/n^{1-q}$. Then

$$\ln\{\Pr(\hat{\mathcal{A}}^{[m]} \neq \mathcal{A})\} \leq C_1 n^{a+b} + \ln\{2(1 - C_2/n^{2d+4q})^n\} \leq C_1 n^{a+b} + n \ln\{(1 - C_2/n^{2d+4q})\}.$$

Since $2d + 5(a + b) < 1$, then $(1 - 2d)/5 < (1 - a - b - 2d)/4$, then $n^{4q} < n^{(1-a-b-2d)/4}$, then

$$\frac{n^{a+b}}{n \ln\left\{(1 - C_2/n^{2d+4q})\right\}} \approx \frac{n^{a+b}}{n C_2 / n^{2d+4q}\{-1 + o(1)\}} \approx -\frac{C_2 n^{4q}}{n^{1-a-b-2d}} \to 0$$

and

$$C_1 n \ln\{(1 - C_2/n^{2d+4q})\} \approx C_1 n C_2 / n^{2d+4q}\{-1 + o(1)\} \to -\infty.$$

Thus,

$$\begin{aligned}\ln\{\Pr(\hat{\mathcal{A}}^{[m]} \neq \mathcal{A})\} &\leq C_1 n^{a+b} + n \ln\{(1 - C_2/n^{2d+4q})\} \\ &\leq C_1 n \ln\{(1 - C_2/n^{2d+4q})\} \left\{\frac{n^{a+b}}{n \ln\left\{(1 - C_2/n^{2d+4q})\right\}} + 1\right\} \to -\infty.\end{aligned}$$

As a result, $\Pr(\hat{\mathcal{A}}^{[m]} \neq \mathcal{A}) \to 0$. This completes the proof of Theorem 1. □

5.3. Proof of Theorem 2

Since we are applying the least squares on the estimated active set, then as $n \to \infty$, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| = \mathcal{O}_p(-n^{1/2})$ when $\hat{\mathcal{A}}^{[m]} = \mathcal{A}$. And according to Theorem 1, $\Pr(\hat{\mathcal{A}}^{[m]} = \mathcal{A}) \to 1$. Thus $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| = \mathcal{O}_p(-n^{1/2})$.

We next focus on the asymptotic normality of the nonparametric estimation of $g(u)$. We complete the proof in three steps.

Step 1: First we derive the bias of $\hat{g}(u)$. After obtaining the estimated active set, for any fixed u, we have

$$\hat{\mathbf{b}}(u) = \begin{pmatrix}\hat{b}_0(u) \\ \hat{b}_1(u)\end{pmatrix} = \arg\min_{b_0, b_1} \sum_{i=1}^{n} \{y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}} - b_0 - b_1(u_i - u_0)\}^2 K_h(u_i - u_0) = (Z_u^\top W_u Z_u)^{-1} Z_u^\top W_u (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}),$$

where

$$Z_u = \begin{pmatrix}1 & u_1 - u \\ \vdots & \vdots \\ 1 & u_n - u\end{pmatrix} \quad \text{and} \quad W_u = \operatorname{diag}\{K_h(u_1 - u), \ldots, K_h(u_n - u)\}.$$

We note the following facts:

(1) $\mathrm{E}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} | \mathbf{X}, \mathbf{u}) = g(\mathbf{u})$.
(2) $(Z_u^\top W_u Z_u)^{-1} Z_u^\top W_u \mathbf{X} = \mathcal{O}_p(1)$, and $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| = \mathcal{O}_p(n^{-1/2})$.
(3) $g(u_i) = b_0 + b_1(u_i - u) + b_2(u_i - u)^2 + \cdots$, and $g(u) = b_0$.

Now

$$\begin{aligned}
\mathrm{E}(\hat{\mathbf{b}}|\mathbf{X},\mathbf{u}) &= (Z_u^\top W_u Z_u)^{-1} Z_u^\top W_u \mathrm{E}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}|\mathbf{X},\mathbf{u}) = (Z_u^\top W_u Z_u)^{-1} Z_u^\top W_u \mathrm{E}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}+\mathbf{X}\boldsymbol{\beta}-\mathbf{X}\hat{\boldsymbol{\beta}}|\mathbf{X},\mathbf{u}) \\
&= (Z_u^\top W_u Z_u)^{-1} Z_u^\top W_u \{g(\mathbf{u}) + \mathbf{X}\mathrm{E}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}|\mathbf{X},\mathbf{u})\} \\
&= (Z_u^\top W_u Z_u)^{-1} Z_u^\top W_u \begin{pmatrix} b_0 + b_1(u_1-u) + g(u) - b_0 - b_1(u_1-u) \\ \vdots \\ b_0 + b_1(u_n-u) + g(u) - b_0 - b_1(u_n-u) \end{pmatrix} + \mathcal{O}_p(n^{-1/2}) \\
&= \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} + (Z_u^\top W_u Z_u)^{-1} Z_u^\top W_u \begin{pmatrix} b_2(u_1-u)^2 + \cdots \\ \vdots \\ b_2(u_n-u)^2 + \cdots \end{pmatrix} + \mathcal{O}_p(n^{-1/2}) \\
&= \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} + J_1 + \mathcal{O}_p(n^{-1/2}).
\end{aligned}$$

Set

$$S_n = Z_u^\top W_u Z_u = \begin{pmatrix} \sum_{i=1}^n K_h(u_i-u) & \sum_{i=1}^n (u_i-u)K_h(u_i-u) \\ \sum_{i=1}^n (u_i-u)K_h(u_i-u) & \sum_{i=1}^n (u_i-u)^2 K_h(u_i-u) \end{pmatrix}, \quad H = \begin{pmatrix} 1 & 0 \\ 0 & h \end{pmatrix}.$$

Then

$$\begin{aligned}
\frac{1}{n} H^{-1} S_n H^{-1} &= \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n K_h(u_i-u) & \frac{1}{nh}\sum_{i=1}^n (u_i-u)K_h(u_i-u) \\ \frac{1}{nh}\sum_{i=1}^n (u_i-u)K_h(u_i-u) & \frac{1}{nh^2}\sum_{i=1}^n (u_i-u)^2 K_h(u_i-u) \end{pmatrix} \\
&= f(u)\begin{pmatrix} 1 & 0 \\ 0 & \mu_2 \end{pmatrix} + \mathcal{O}\{h + \sqrt{1/(nh)}\}.
\end{aligned}$$

Next,

$$Z_u^\top W_u \begin{pmatrix} (u_1-u)^\ell \\ \vdots \\ (u_n-u)^\ell \end{pmatrix} = \begin{pmatrix} \frac{nh^\ell}{n}\sum_{i=1}^n \left(\frac{u_i-u}{h}\right)^\ell K_h(u_i-u) \\ \frac{nh^{\ell+1}}{n}\sum_{i=1}^n \left(\frac{u_i-u}{h}\right)^{\ell+1} K_h(u_i-u) \end{pmatrix}.$$

Then J_1 can be expressed as follows:

$$\begin{aligned}
J_1 &= (Z_u^\top W_u Z_u)^{-1} Z_u^\top W_u \begin{pmatrix} b_2(u_1-u)^2 + \cdots \\ \vdots \\ b_2(u_n-u)^2 + \cdots \end{pmatrix} \\
&= b_2 S_n^{-1} \begin{pmatrix} nh^2 \frac{1}{n}\sum_{i=1}^n \left(\frac{u_i-u}{h}\right)^2 K_h(u_i-u) \\ nh^3 \frac{1}{n}\sum_{i=1}^n \left(\frac{u_i-u}{h}\right)^3 K_h(u_i-u) \end{pmatrix} + o(h^2) \\
&= b_2 H^{-1}(n^{-1}H^{-1}S_n H^{-1})^{-1} H^{-1} \begin{pmatrix} \frac{h^2}{n}\sum_{i=1}^n \left(\frac{u_i-u}{h}\right)^2 K_h(u_i-u) \\ \frac{h^3}{n}\sum_{i=1}^n \left(\frac{u_i-u}{h}\right)^3 K_h(u_i-u) \end{pmatrix} + o(h^2)
\end{aligned}$$

$$= b_2h^2H^{-1}(n^{-1}H^{-1}S_nH^{-1})^{-1}\begin{pmatrix}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{u_i-u}{h}\right)^2K_h(u_i-u)\\ \frac{1}{n}\sum_{i=1}^{n}\left(\frac{u_i-u}{h}\right)^3K_h(u_i-u)\end{pmatrix}+o(h^2)$$
$$= b_2h^2H^{-1}\left[f^{-1}(u)\begin{pmatrix}1&0\\0&1/\mu_2\end{pmatrix}+\mathcal{O}_p\{h+\sqrt{1/(nh)}\}\right]\left[f(u)\begin{pmatrix}\mu_2\\ \mu_3\end{pmatrix}+\mathcal{O}_p\{h+\sqrt{1/(nh)}\}\right]+o(h^2)$$
$$= b_2h^2H^{-1}\left[\begin{pmatrix}\mu_2\\ \mu_3/\mu_2\end{pmatrix}+\mathcal{O}_p\{h+\sqrt{1/(nh)}\}\right]+o(h^2).$$

Thus,

$$\begin{aligned}\mathrm{E}(\hat{g}(u)|\mathbf{X},\mathbf{u}) &= (1\,0)\mathrm{E}(\hat{\mathbf{b}}|\mathbf{X},\mathbf{u}) = b_0+b_2h^2[\mu_2+\mathcal{O}_p\{h+\sqrt{1/(nh)}\}]+o(h^2)+\mathcal{O}_p(n^{-1/2})\\ &= b_0+b_2\mu_2h^2+h^3\mathcal{O}_p\{1+\sqrt{1/(nh^3)}\}+o(h^2)+\mathcal{O}_p(n^{-1/2})\\ &= b_0+b_2\mu_2h^2+o(h^2)+\mathcal{O}_p(n^{-1/2}),\end{aligned}$$

and the last equality holds because $nh^3\to\infty$, as $n\to 0$. Note that the right-hand side does not depend on $\mathbf{X}$ and $\mathbf{u}$. Therefore,

$$\mathrm{E}\left\{\hat{g}(u)\right\} = b_0+b_2\mu_2h^2+o(h^2)+\mathcal{O}_p(n^{-1/2}) = g(u)+g''(u)\mu_2h^2/2+o(h^2)+\mathcal{O}_p(n^{-1/2}).$$

Step 2: Derive the variance of $\hat{g}(u)$. Note that $\mathrm{var}(\hat{\mathbf{b}}|\mathbf{X},\mathbf{u}) = (Z_u^\top W_uZ_u)^{-1}Z_u^\top W_u\mathrm{var}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}|\mathbf{X},\mathbf{u})W_uZ_u(Z_u^\top W_uZ_u)^{-1}$, and

$$\begin{aligned}\mathrm{var}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}|\mathbf{X},\mathbf{u}) &= \mathrm{var}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}+\mathbf{X}\boldsymbol{\beta}-\mathbf{X}\hat{\boldsymbol{\beta}}|\mathbf{X},\mathbf{u})\\ &= \mathrm{var}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}|\mathbf{X},\mathbf{u})+\mathrm{var}(\mathbf{X}\boldsymbol{\beta}-\mathbf{X}\hat{\boldsymbol{\beta}}|\mathbf{X},\mathbf{u})+2\mathrm{cov}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta},\mathbf{X}\boldsymbol{\beta}-\mathbf{X}\hat{\boldsymbol{\beta}}|\mathbf{X},\mathbf{u})\\ &= \{\sigma^2+\mathcal{O}_p(n^{-1/2})\}I_n.\end{aligned}$$

Then

$$\begin{aligned}\mathrm{var}(\hat{\mathbf{b}}|\mathbf{X},\mathbf{u}) &= (Z_u^\top W_uZ_u)^{-1}Z_u^\top W_u\{\sigma^2I_n+\mathcal{O}_p(n^{-1/2})\}W_uZ_u(Z_u^\top W_uZ_u)^{-1}\\ &= \{\sigma^2+\mathcal{O}_p(n^{-1/2})\}S_n^{-1}\left(Z_u^\top W_u^2Z_u\right)S_n^{-1}\\ &= \{\sigma^2+\mathcal{O}_p(n^{-1/2})\}H^{-1}(H^{-1}S_nH^{-1})^{-1}H^{-1}\left(Z_u^\top W_u^2Z_u\right)H^{-1}(H^{-1}S_nH^{-1})^{-1}\\ &\equiv \{\sigma^2+\mathcal{O}_p(n^{-1/2})\}J_2.\end{aligned}$$

Furthermore,

$$\begin{aligned}H^{-1}\left(Z_u^\top W_u^2Z_u\right)H^{-1} &= \begin{pmatrix}1&0\\0&1/h\end{pmatrix}\begin{pmatrix}\sum_{i=1}^{n}K_h^2(u_i-u) & \sum_{i=1}^{n}(u_i-u)K_h^2(u_i-u)\\ \sum_{i=1}^{n}(u_i-u)K_h^2(u_i-u) & \sum_{i=1}^{n}(u_i-u)^2K_h^2(u_i-u)\end{pmatrix}\begin{pmatrix}1&0\\0&1/h\end{pmatrix}\\ &= \begin{pmatrix}\sum_{i=1}^{n}K_h^2(u_i-u) & \sum_{i=1}^{n}\left(\frac{u_i-u}{h}\right)K_h^2(u_i-u)\\ \sum_{i=1}^{n}\left(\frac{u_i-u}{h}\right)K_h^2(u_i-u) & \sum_{i=1}^{n}\left(\frac{u_i-u}{h}\right)^2K_h^2(u_i-u)\end{pmatrix}\\ &= \frac{nf(u)}{h}\begin{pmatrix}\nu_0&\nu_1\\ \nu_1&\nu_2\end{pmatrix}+n\mathcal{O}_p\{1+\sqrt{1/(nh^3)}\}\end{aligned}$$

and

$$\begin{aligned}J_2 &= \frac{1}{n}H^{-1}(n^{-1}H^{-1}S_nH^{-1})^{-1}\{n^{-1}H^{-1}\left(Z_u^\top W_u^2Z_u\right)H^{-1}\}(n^{-1}H^{-1}S_nH^{-1})^{-1}H^{-1}\\ &= \frac{1}{n}H^{-1}\left[f^{-1}(u)\begin{pmatrix}1&0\\0&1/\mu_2\end{pmatrix}+\mathcal{O}_p\{h+\sqrt{1/(nh)}\}\right]\left[\frac{f(u)}{h}\begin{pmatrix}\nu_0&\nu_1\\ \nu_1&\nu_2\end{pmatrix}+\mathcal{O}_p\{1+\sqrt{1/(nh^3)}\}\right]\\ &\quad\times\left[f^{-1}(u)\begin{pmatrix}1&0\\0&1/\mu_2\end{pmatrix}+\mathcal{O}_p\{h+\sqrt{1/(nh)}\}\right]H^{-1}\\ &= \frac{1}{nh}H^{-1}\left[\frac{1}{f(u)}\begin{pmatrix}\nu_0&\nu_1/\mu_2\\ \nu_1/\mu_2&\nu_2/\mu_2\end{pmatrix}+\mathcal{O}_p\{h+\sqrt{1/(nh)}\}\right]H^{-1}.\end{aligned}$$

Therefore,

$$\begin{aligned}\operatorname{var}\left\{\hat{g}(u)|\mathbf{X},\mathbf{u}\right\} &= (1\,0)\operatorname{var}(\hat{\mathbf{b}}|\mathbf{X},\mathbf{u})(1\,0)^{\top} = \frac{\sigma^2+\mathcal{O}_p(n^{-1/2})}{nh}\left[\frac{1}{f(u)}\nu_0+\mathcal{O}_p\{h+\sqrt{1/(nh)}\}\right]\\ &= \frac{1}{nhf(u)}[\sigma^2\nu_0+\mathcal{O}_p(n^{-1/2})+\mathcal{O}_p\{h+\sqrt{1/(nh)}\}].\end{aligned}$$

Notice that the right-hand side does not depend on $\mathbf{X}$ and $\mathbf{u}$. Therefore,

$$\operatorname{var}\left\{\hat{g}(u)\right\} = \frac{1}{nhf(u)}[\sigma^2\nu_0+\mathcal{O}_p(n^{-1/2})+\mathcal{O}_p\{h+\sqrt{1/(nh)}\}].$$

Step 3: In order to derive the asymptotic distribution of $\hat{g}(u)$, we can use the following facts:

(1) $\hat{g}(u) = (1\,0)\hat{\mathbf{b}} = (1\,0)(Z_u^{\top}W_uZ_u)^{-1}Z_u^{\top}W_u(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})$.
(2) $\mathrm{E}\{\hat{g}(u)|\mathbf{X},\mathbf{u}\} = (1\,0)\mathrm{E}(\hat{\mathbf{b}}|\mathbf{X},\mathbf{u}) = g(u)+g''(u)\mu_2h^2/2+\mathcal{O}_p(n^{-1/2})+o(h^2)$.
(3) $\mathrm{E}\{\hat{g}(u)|\mathbf{X},\mathbf{u}\} = (1\,0)(Z_u^{\top}W_uZ_u)^{-1}Z_u^{\top}W_u\mathrm{E}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}|\mathbf{X},\mathbf{u})$.
(4) $\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathrm{E}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}|\mathbf{X},\mathbf{u}) = (\epsilon_1,\ldots,\epsilon_n)^{\top}$.

We first study $\hat{g}(u)-g(u)-g''(u)\mu_2h^2/2$:

$$\begin{aligned}\hat{g}(u)-g(u)-g''(u)\mu_2h^2/2 &= \hat{g}(u)-\mathrm{E}(\hat{g}(u)|\mathbf{X},\mathbf{u})+\mathcal{O}_p(n^{-1/2})+o(h^2)\\ &= (1\,0)(Z_u^{\top}W_uZ_u)^{-1}Z_u^{\top}W_u\{\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}-\mathrm{E}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}|\mathbf{X},\mathbf{u})\}+\mathcal{O}_p(n^{-1/2})+o(h^2)\\ &= (1\,0)S_n^{-1}Z_u^{\top}W_u\{\mathbf{y}-\mathbf{X}\boldsymbol{\beta}+\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})-\mathrm{E}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}+\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})|\mathbf{X},\mathbf{u})\}+\mathcal{O}_p(n^{-1/2})+o(h^2)\\ &= (1\,0)S_n^{-1}Z_u^{\top}W_u\{\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathrm{E}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}|\mathbf{X},\mathbf{u})\\ &\quad+\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})-E(\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})|\mathbf{X},\mathbf{u})\}+\mathcal{O}_p(n^{-1/2})+o(h^2).\end{aligned}$$

Because $S_n^{-1}Z_u^{\top}W_u\mathbf{X} = \mathcal{O}_p(1)$, $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\| = \mathcal{O}_p(n^{-1/2})$, then

$$\begin{aligned}\hat{g}(u)-g(u)-g''(u)\mu_2h_n^2/2 &= [1\,0]S_n^{-1}Z_u^{\top}W_u\{\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathrm{E}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}|\mathbf{X},\mathbf{u})\}+\mathcal{O}_p(n^{-1/2})+o(h^2)\\ &= [1\,0]S_n^{-1}Z_u^{\top}W_u(\epsilon_1,\ldots,\epsilon_n)^{\top}+\mathcal{O}_p(n^{-1/2})+o(h^2)\\ &= [1\,0]H^{-1}(nHS_n^{-1}H)H^{-1}\{Z_u^{\top}W_u(\epsilon_1,\ldots,\epsilon_n)^{\top}/n\}+\mathcal{O}_p(n^{-1/2})+o(h^2).\end{aligned}$$

Furthermore, given that

$$n^{-1}H^{-1}S_nH^{-1}\xrightarrow{\mathcal{P}} f(u)\begin{pmatrix}1 & 0\\ 0 & \mu_2\end{pmatrix}$$

we have

$$nHS_n^{-1}H\xrightarrow{\mathcal{P}}\frac{1}{f(u)}\begin{pmatrix}1 & 0\\ 0 & 1/\mu_2\end{pmatrix}.$$

Moreover,

$$H^{-1}\{b^{-1}Z_u^{\top}W_u(\epsilon_1,\ldots,\epsilon_n)^{\top}\} = \begin{pmatrix}1 & 0\\ 0 & 1/h\end{pmatrix}\begin{pmatrix}\frac{1}{n}\sum_{i=1}^{n}K_h(u_i-u)\epsilon_i\\ \frac{1}{n}\sum_{i=1}^{n}(u_i-u)K_h(u_i-u)\epsilon_i\end{pmatrix} = \begin{pmatrix}\frac{1}{n}\sum_{i=1}^{n}K_h(u_i-u)\epsilon_i\\ \frac{1}{n}\sum_{i=1}^{n}\left(\frac{u_i-u}{h}\right)K_h(u_i-u)\epsilon_i\end{pmatrix}.$$

We now turn to $n^{-1}\sum_{i=1}^{n}K_h(u_i-u)\epsilon_i$ and find

$$\xi_n^2 = \operatorname{var}\left\{\sum_{i=1}^{n}\frac{1}{n}K_h(u_i-u)\epsilon_i\right\} = \frac{1}{n}\mathrm{E}\{K_h^2(u_i-u)\epsilon_i^2\} = \frac{\sigma^2}{n}\mathrm{E}\{K_h^2(u_i-u)\}.$$

Note that

$$\begin{aligned}\mathrm{E}\{K_h^2(u_i-u)\} &= \int\left\{h^{-1}K\left(\frac{u_i-u}{h}\right)\right\}^2 f(u_i)du_i = h^{-1}\int K^2(t)f(u+th)dt\\ &= h^{-1}\int K^2(t)\{f(u)+thf'(u)+o(h)\}dt = h^{-1}\{f(u)\nu_0+hf'(u)\nu_1+o(h)\}.\end{aligned}$$

Thus $\xi_n^2 = \sigma^2\{f(u)\nu_0 + hf'(u)\nu_1 + o(h)\}/(nh) = \sigma^2\{C_1 + o(h)\}/(nh)$. Similarly, we have

$$\sum_{i=1}^{n} \mathrm{E}|K_h(u_i - u)\epsilon_i/n|^3 = n^{-2}\mathrm{E}|K_h^3(u_i - u)| \times \mathrm{E}|\epsilon_i^3| = n^{-2}h^{-2}\{C_2 + o(h)\}.$$

Therefore, as $n \to \infty$, $nh^3 \to \infty$, we have $nh \to \infty$, then

$$\frac{1}{\xi_n^2}\sum_{i=1}^{n} \mathrm{E}|K_h(u_i - u)\epsilon_i/n|^3 = \frac{\{C_2 + o(h)\}/(n^2h^2)}{\sigma^2\{C_1 + o(h)\}/(nh)} = \{C_3 + o(h)\}/(nh) \to 0,$$

By the Lyapunov Central Limit Theorem, we get

$$\frac{n^{-1}\sum_{i=1}^{n} K_h(u_i - u)\epsilon_i}{\sqrt{\sigma^2\{f(u)\nu_0 + hf'(u)\nu_1 + o(h)\}/(nh)}} \rightsquigarrow \mathcal{N}(0, 1).$$

That is,

$$\sqrt{h/n}\sum_{i=1}^{n} K_h(u_i - u)\epsilon_i \rightsquigarrow \mathcal{N}[0, \sigma^2 f(u)\nu_0].$$

Similarly,

$$\sqrt{h/n}\sum_{i=1}^{n} h^{-1}(u_i - u)K_h(u_i - u)\epsilon_i \rightsquigarrow \mathcal{N}[0, \sigma^2 f(u)\nu_2].$$

Applying Slutsky's Lemma, we find

$$\begin{aligned}
&\sqrt{nh}\{\hat{g}(u) - g(u) - g''(u)\mu_2 h_n^2/2\} \\
&\quad = \sqrt{nh}\,(1\,0)H^{-1}(nHS_n^{-1}H)H^{-1}\{Z_u^\top W_u(\epsilon_1, \ldots, \epsilon_n)^\top/n\} + \mathcal{O}_p(\sqrt{h}) + o(\sqrt{nh^5}) \\
&\quad = (1\,0)H^{-1}(nHS_n^{-1}H)\begin{pmatrix} \sqrt{h/n}\sum_{i=1}^{n} K_h(u_i - u)\epsilon_i \\ \sqrt{h/n}\sum_{i=1}^{n} h^{-1}(u_i - u)K_h(u_i - u)\epsilon_i \end{pmatrix} + \mathcal{O}_p(\sqrt{h}) + o(\sqrt{nh^5}) \\
&\quad \rightsquigarrow \frac{1}{f(u)}\mathcal{N}[0, \sigma^2 f(u)\nu_0] = \mathcal{N}[0, \sigma^2\nu_0/f(u)],
\end{aligned}$$

which completes the proof of Theorem 2. □

6. Conclusion

In this paper, we advocated a new approach to select significant variables in the partially linear models via partial correlation learning. Under the partial faithfulness framework, the nonparametric smoothing techniques are adopted to obtain the partial residuals, and then the recursive hypothesis tests of partial correlations between partial residuals are conducted to select linear covariates in a backward direction. Model selection consistency is proved and empirically verified through simulations. Furthermore, the $\sqrt{n}$-consistency of the estimated linear coefficients and the asymptotic normality of the nonparametric baseline estimations are provided. The performance of the method is further illustrated by supermarket data analysis.

Acknowledgments

The authors are grateful to the Editor-in-Chief, the Associate Editor and the referees for comments and suggestions that led to significant improvements. Liu's research was supported by National Natural Science Foundation of China (NNSFC) grants 11771361 and 11671334, JAS14007 and the Fundamental Research Funds for the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry. Li's research was supported by National Institute on Drug Abuse (NIDA) grants P50 DA039838 and P50 DA036107, and National Science Foundation grants DMS 1512422 and DMS 1820702. This work was also partially supported by NNSFC grant 11690015. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF, the NIDA, the NIH, or the NNSFC.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmva.2018.06.005. The Online Supplement includes two parts. The first half provides the proof of Lemma 2 and the second half reports the additional simulation results mentioned in Section 4.

References

[1] P. Bühlmann, M. Kalisch, M. Maathuis, Variable selection in high-dimensional linear models: Partially faithful distributions and the PC-simple algorithm, Biometrika 97 (2010) 261–278.
[2] J. Chen, Z. Chen, Extended Bayesian information criterion for model selection with large model space, Biometrika 94 (2008) 759–771.
[3] J. Fan, I. Gijbels, Local Polynomial Modelling and its Applications, Chapman & Hall, London, 1996.
[4] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and it oracle properties, J. Amer. Statist. Assoc. 96 (2001) 1348–1360.
[5] J. Fan, R. Li, New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis, J. Amer. Statist. Assoc. 99 (2004) 710–723.
[6] J. Fan, J. Lv, Sure independence screening for ultrahigh dimensional feature space (with discussion), J. R. Stat. Soc. Ser. B 70 (2008) 849–911.
[7] N. Heckman, Spline smoothing in a partly linear model, J. R. Stat. Soc. Ser. B 48 (1986) 244–248.
[8] R. Li, J. Liu, L. Lou, Variable selection via partial correlation, Statist. Sinica 27 (2017) 983–996.
[9] H. Liang, R. Li, Variable selection for partially linear models with measurement errors, J. Amer. Statist. Assoc. 104 (2009) 234–248.
[10] J. Liu, R. Li, R. Wu, Feature selection for varying coefficient models with ultrahigh dimensional covariates, J. Amer. Statist. Assoc. 109 (2014) 266–274.
[11] D. Ruppert, S.J. Sheather, M.P. Wand, An effective bandwidth selector for local least squares regression, J. Amer. Statist. Assoc. 90 (1995) 1257–1270.
[12] D. Ruppert, M.P. Wand, R.J. Carroll, Semiparametric Regression, Cambridge Press, New York, 2003.
[13] P. Speckman, Kernel smoothing in partial linear models, J. R. Stat. Soc. Ser. B 50 (1988) 413–436.
[14] R.J. Tibshirani, Regression shrinkage and selection via LASSO, J. R. Stat. Soc. Ser. B 58 (1996) 267–288.
[15] H. Xie, J. Huang, SCAD-penalized regression in high-dimensional partially linear models, Ann. Statist. 37 (2009) 673–696.