International Journal of Statistical Distributions and Applications
2020; 6(1): 1-9
http://www.sciencepublishinggroup.com/j/ijsda
doi: 10.11648/j.ijsd.20200601.11
ISSN: 2472-3487 (Print); ISSN: 2472-3509(Online)

Variable Selection for Partially Linear Additive Model Based on Modal Regression Under High Dimensional Data

Yafeng Xia*, Lirong Zhang

School of Sciences, Lanzhou University of Technology, Lanzhou, P. R. China

Email address:
gsxyf01@163.com (Lirong Zhang)
*Corresponding author

To cite this article:
Yafeng Xia, Lirong Zhang. Variable Selection for Partially Linear Additive Model based on Modal Regression under High Dimensional Data. *International Journal of Statistical Distributions and Applications*. Vol. 6, No. 1, 2020, pp. 1-9. doi: 10.11648/j.ijsd.20200601.11

Received: December 19, 2019; **Accepted:** January 9, 2020; **Published:** April 17, 2020

Abstract: In this article, we focus on the variable selection for partially linear additive model under high dimensional data.Variable selection is proposed based on modal regression estimation with Adoptive Bridge Method.Using the B-spline basic function to approximate the additive function, a penalty estimation objective equation is constructed.It establishes and proves that the variable selection methods have oracle property.Numerical simulations tested the performance of the proposed methods in a finite sample and verified the significance of the proposed estimation and the variable selection methods. At the end of the article, we attach the detailed derivation of the theoretical results.Therefore, the correctness of the method used is verified theoretically and practically.

Keywords: High Dimensional Data, Partially Linear Additive Model, Modal Regression, Variable Selection, Adoptive Bridge, B-spline

1. Introduction

Variable selection of model is one of the hot topics in modern statistics.With the progress of science and technology continuously, statisticians need to process large scale data and select valuable information from these large scale data for statistical analysis.In modern scientific research and technological development, the emergence of high dimensional data has brought new challenges to statisticians.The emergence of high-dimensional data and new scientific problems have changed the ideas of traditional statistics and data analysis.Reducing dimension and screening characteristic variable have become the primary tasks for high dimensional statistical problems.The large scale calculations brought by the reducing dimension and screening characteristic variable process will also promote the continuous improvement of the algorithm.

The partially linear additive model is proposed by Hastie and Tibshirani [1] combining the basic additive model and the characteristics of the partially linear model.It can be defined as follows:

$$Y_i = {X_i}^T \beta_n + \sum_{l=1}^{d_n} g_l\left(Z_l\right) + \varepsilon_i \tag{1}$$

where Y is response variable, $X = \left({X_1}^T, {X_2}^T, \cdots, {X_{p_n}}^T\right)$ and $Z = \left({Z_1}^T, {Z_2}^T, \cdots {Z_{d_n}}^T\right)$ are two group of covariates, $g_1\left(Z_l\right), \cdots, g_{d_n}\left(Z_l\right)$ is d_n-dimensional vector of additive, $\beta = \left(\beta_1, \cdots, \beta_{d_n}\right)$ is p_n-dimensional vector of unknown regression coefficient, ε is model error, which is independent of X, Z, and $E(\varepsilon|X) = 0$.The dimensions p_n and d_n in the article increase with the increase of the sample size, that is, the sizes of p_n and d_n are related to that of n.

A large number of scholars have studied the partially linear model.Guo [2] proposed an estimation method based on compound quantile regression for semi-parametric partially linear additive model.Liu [3] studied the asymptotic normality of the parameter estimation, used the SCAD penalty function to identify important linear components, it proved that the estimation of non-zero components has Oracle properties.Xia

[4] focus on the variable selection for semiparametric model with response missing at random.Fan and Huang [5] used profile least squares to estimate the variables of the parameter part, discussed the estimation with asymptotic properties, and tested the model by the Profile generalized likelihood ratio method.Hoshino [6] studied the estimation problem for partially linear additive quantile regression model.

In the high-dimensional model, Meinshausen and Bhlmann [7] studies High-dimensional graphs and variable selection with the lasso.Zhang and Huang [8] discussed The sparsity and bias of the lasso selection in high-dimensional linear regression.Using the bridge penalty method, Wang et al. [9] studies variable selection and parameter estimation in the partially linear model and high-dimensional generalized linear models.Li et al.[10] discussed the variable selection of the generalized semi-varying coefficient model in the ultra-high-dimensional case.Inspired by the above literature, this paper considers the robust estimation and variable selection for partially linear additive model based on the method of modal regression and adoptive bridge estimation under high dimensional data.It is proved by theoretical properties that adoptive bridge estimation can accurately screen non-zero parameters with probability tending to 1 under high-dimensional data.

2. Variable Selection Method

Similar to the reference [11], the basis function approximation is used to replace the additive function in (1).

Let

$$B(U) = (B_1(U), \cdots, Bq_n(U))^T$$

be B-spline basic function with the order of $m+1$ where $q_n = K + m + 1$ and K is the number of interior knot.In order to obtain the consistent estimation of $\{g_l(Z_l)\}$, we use empirically centralized B-spline function subspace

$$S_l^0 = \left\{ s | \, s = \sum\nolimits_{j=1}^{q_n} B_j(z_l)^T \gamma_{lj}, \sum\nolimits_{i=1}^{n} s(Z_{il}) = 0 \right\},$$

its empirically centralized basis function is

$$\psi_{lj}(z_l) = B_j(z_l) - \sum\nolimits_{i=1}^{n} B_j(Z_{il})/n, j = 1, \cdots, q_n,$$

then

$$g_l(z_l) \approx \sum\nolimits_{l=1}^{q_n} B_j(z_l)^T \gamma_{lj}, l = 1, 2, \cdots, d_n . g_l(Z_l)$$

can be approximated by

$$g_l(z_l) \approx \sum_{j=1}^{q_n} \gamma_{lj} \psi_{lj}(z_l) = {\psi_l}^T \gamma_l,$$

where $\psi_l(z_l) = (\psi_{l1}(z_l), \cdots, \psi_{lq_n}(z_l))^T$ and $\gamma_l = (\gamma_{l1}, \gamma_{l2}, \cdots, \gamma_{lq_n})^T$ are B-spline coefficients.Denote $\psi_{il} = (\psi_{ll}(Z_{il}), \cdots, \psi_{lq}(Z_{il}))^T$ and $\Psi_i = \left(\psi_{i1}^T, \cdots, \psi_{id_n}^T\right)^T$.Using the modal regression method proposed by Yao [12], we obtain the estimation $\hat{\beta}_n$ and $\hat{\gamma_n}$ of β_n and γ_n by maximizing equation (1).

$$Q_h(\gamma_n, \beta_n) = \sum_{i=1}^{n} \phi_h\left(Y_i - X_i^T \beta_n - \Psi_i^T \gamma_n\right), \tag{2}$$

where $\phi_h(t)$ equals $h^{-1}\phi(t/h)$, h is bandwidth which plays the role of robust estimation, and $\phi(t)$ is a kern density function.In order to calculate and discuss properties conveniently, in this paper we use the normal kernel density function, that is, $\phi_h(t) = \frac{1}{\sqrt{2\pi}h} e^{-\frac{t^2}{2h^2}}$.Consider the following penalty function based on equation (1), we have

$$\begin{aligned} P_n(\gamma_n, \beta_n) = Q_h(\gamma_n, \beta_n) - \lambda_{1k} w_k \sum_{k=1}^{p_n} |\beta_{nk}|^{\zeta} \\ - \lambda_{2l} w_l \sum_{1=1}^{d_n} \|\gamma_{nl}\|_H{}^{\zeta}, \end{aligned} \tag{3}$$

where

$$\|\gamma_{nl}\|_H = \sqrt{\gamma_{nl}^T H \gamma_{nl}}, H = \int {}_0^1 \psi_{lk}(z) \psi_{lk'}(z) dz,$$

λ_{1k} and λ_{2l} are tuning parameters more than 0.w_k and w_l are the penalty weights for the k-th and l-th components, respectively.We default $w_k = \left|\tilde{\beta}_{nk}\right|^{-r}$ and $w_l = |\tilde{\gamma}_{nl}|^{-r}$ generally, $\tilde{\gamma}_{nl}$ and $\tilde{\beta}_{nk}$ are non-penalty estimation of γ_{nl} and β_{nl}, respectively.$0 < \zeta < 1$, the definition of the adoptive bridge estimation is as follows:

$$\hat{\theta}_n = \left(\hat{\gamma}_n, \hat{\beta}_n\right) = \arg\max_{\hat{\gamma}_n, \hat{\beta}_n} \{P_n(\gamma_n, \beta_n)\}. \tag{4}$$

It is more difficult to maximize the objective function (2) for a given tuning parameter λ_{1k} and λ_{2l} directly.Assuming that the initial value $\tilde{\theta}_{n0} = \left(\tilde{\beta}_{n0}, \tilde{\gamma}_{n0}\right)^T$ is very close to the maximum value of the objective function (2), the initial value is usually considered to be the non-penalty estimation of equation (2). We use the local quadratic algorithm (LQA) proposed by Fan [13] to approximate the penalty function. If $\tilde{\theta}_n^{(0)}$ is very close to 0, let $\tilde{\theta}_n^{(0)} = 0$, otherwise

$$\lambda_{1k} w_k |\beta_{nk}|^{\zeta} \approx \lambda_{1k} w_k \left|\beta_{nk}^{(0)}\right|^{\zeta} + \frac{1}{2} \left\{ \frac{\lambda_{1k} w_k \zeta \left|\beta_{nk}^{(0)}\right|^{\zeta-1}}{\left|\beta_{nk}^{(0)}\right|} \right\} \left(|\beta_{nk}|^2 - \left|\beta_{nk}^{(0)}\right|^2 \right), \tag{5}$$

$$\lambda_{2l} w_l \|\gamma_{nl}\|_H{}^{\zeta} \approx \lambda_{2l} w_l \left\|\gamma_{nl}^{(0)}\right\|_H{}^{\zeta} + \frac{1}{2} \left\{ \frac{\lambda_{2l} w_l \zeta \left\|\gamma_{nl}^{(0)}\right\|_H{}^{\zeta-1}}{\left\|\gamma_{nl}^{(0)}\right\|_H} \right\} \left(\gamma_{nl}^T H \gamma_{nl} - \gamma_{nl}^{(0)T} H \gamma_{nl}^{(0)} \right). \tag{6}$$

Denote

$$G(x,z,h) = E\left\{ {\phi'}_h(\varepsilon)^2 | X = x, Z = z \right\}, \quad F(x,z,h) = E\left\{ {\phi''}_h(\varepsilon) | X = x, Z = z \right\},$$

$$a_n = \max_{k,l} \{ \lambda_{1k}, \lambda_{2l} : k = 1,2,\cdots,s_1; l = 1,2,\cdots,s_2 \}, \quad b_n = \min_{k,l} \{ \lambda_{1k}, \lambda_{2l} : k = 1,2,\cdots,s_1; l = 1,2,\cdots,s_2 \}.$$

3. Asymptotic Properties of Variable Selection

Let H_r represent the whole of function $h(t)$ that satisfies certain conditions in interval $[0,1]$, $h^m(t)$ is m-th derivative of $h(t)$, it is continuous and satisfies v-th order condition of *Hölder*, $r = m + v$.That is, there is a constant value $M_0 \in (0,\infty)$ make $|h^m(s) - h^m(t)| \le M_0 |s-t|^v$ is true, where $s,t \in [0,1]$.In order to prove the conclusion of the theorem, the following regular conditions are needed in this paper.

(B1) $E(g_l(Z_l)) = 0$ and $g_l(z_l) \in H_r$, $l = 1,\cdots,d_n$, $r > 1/2$.

(B2) Covariates Z_l is a continuous density function $f_{z_l}(z_l)$, constants δ_1 and δ_2 enables $f_{z_l}(z_l)$ to satisfy $0 < \delta_1 \le f_{z_l}(z_l) \le \delta_2 < \infty$ on interval $[0,1]$.

(B3) Random variables X_{ik} and the eigenvalues of $E\left\{ X_i X_i^T \middle| Z_i \right\}$ are uniformly bounded, where $1 \le i \le n$, $1 \le k \le p_n$.

(B4) Let $t_1,\cdots,t_K$ be the interior knots of $[0,1]$. Moreover, let $t_0 = 0$, $t_{K+1} = 1, \xi_i = t_i - t_{i-1}$ and $\xi = max\{\xi_i\}$.Then, there exists a constant C_0 such that

$$\frac{\xi}{\min\{\xi_i\}} \le C_0, \quad \max\{|\xi_{i+1} - \xi_i|\} = o(K^{-1}).$$

(B5) Let $F(x,z,h)$ and $G(x,z,h)$ are continuous with respect to (x,z).In addition, $F(x,z,h) < 0, \forall h > 0$.

(B6) $E({\phi'}_h(\varepsilon)|x,z) = 0$, $E\left({\phi''}_h(\varepsilon)^2 \middle| x,z \right)$, $E\left({\phi'}_h(\varepsilon)^3 \middle| x,z \right)$ and $E({\phi'''}_h(\varepsilon)|x,z)$ are continuous with respect to (x,z).

(B7) $\lambda_n w_n {s_n}^{-1/2} \to 0$, where λ_n is λ_{1k} or λ_{2l} , ω_n is ω_{1k} or ω_{2l} and s_n is s_1 or s_2.

(B8) For a_n and b_n, there are $\sqrt{n}a_n \to 0$ and $\sqrt{n}b_n \to \infty$ where $n \to \infty$.

Remark 1 Conditions (B1)-(B3) are common conditions for partially linear additive model; Condition (B4) indicates that $c_0,\cdots,c_{K+1}$ is a C_0-quasi-uniform sequence of partitions of $[0,1]$; Conditions (B5)-(B6) are required for the modal regression.Reference Yao[12] thought $E({\phi'}_h(\varepsilon)|x,z) = 0$ guarantees the consistency of the estimation.If the error distribution density is symmetrical about 0, these conditions are automatically satisfied; Condition (B7) is an indispensable condition for the adoptive bridge estimation of convergence and oracle properties; Condition (B8) represents the assumption of adoptive bridge estimation.

Now, we discuss the asymptotic properties of penalty estimation $\hat{\beta}_n$ and $\hat{\gamma}_n$.Let $g_l{}^{(0)}(z_l)$ and $\beta_n{}^{(0)}$ be the real value of $g_l(z_l)$ and β_n.Default $\beta_{nk}{}^{(0)} = 0$ can represent an unrelated variable, where $k = s_1 + 1,\cdots,p_n$, $\beta_{nk}{}^{(0)} \ne 0$ can represent a function corresponding to an important variables where $k = 1,2,\cdots s_1$.Similarly, we have $g_l{}^{(0)}(z_l) = 0$ and $g_l{}^{(0)}(z_l) \ne 0$ where $l = s_2 + 1,\cdots,d_n$ and $l = 1,2,\cdots s_2$, respectively.

First, we give the consistency of the penalty estimations.

Theorem 1 Suppose that the regular conditions (B1)-(B6) hold, the number of knots $K = O(n^{1/(2r+1)})$, $0 < \zeta < 1$, then we have

(i) $\left\| \hat{\beta}_{nk} - \beta_{nk}{}^{(0)} \right\| = O_p\left(n^{-r/2r+1} + a_n \right)$,

(ii) $\left\| \hat{g}_l(z_l) - g_l{}^{(0)}(z_l) \right\| = O_p\left(n^{-r/2r+1} + a_n \right)$.

Next, we can obtain the sparsity of the penalty estimation under certain conditions.

Theorem 2 Suppose that the regular conditions (B1)-(B6) hold, $0 < \zeta < 1$, then we have

(i) $\hat{\beta}_{nk} = 0, k = s_1 + 1,\cdots,p_n$,

(ii) $\hat{g}_l = 0, l = s_2 + 1,\cdots,d_n$.

Denote

$$\beta_{n1} = \left(\beta_{n_1}^T,\cdots,\beta_{n s_1}^T \right)^T, \gamma_{n1} = \left(\gamma_{n_1}^T,\cdots,\gamma_{n s_2}^T \right)^T,$$

and let $\beta_{n1}{}^{(0)}$ and $\gamma_{n1}{}^{(0)}$ be the real true of β_{n1} and γ_{n1} respectively. Their respective corresponding covariates are labeled with X_1 and Ψ_1. In addition, define

$$\Upsilon = E\left({\phi''}_h(\varepsilon) \Psi_1 \Psi_1^T \right) = E\left(F(x,z,h) \Psi_1 \Psi_1^T \right),$$
$$\Omega = E\left({\phi''}_h(\varepsilon) \Psi_1 X_1^T \right) = E\left(F(x,z,h) \Psi_1 X_1^T \right).$$

Finally, we give the asymptotic normality of the non-zero components in parameters.

Theorem 3 Under the conditions of Theorem 2 and (B1)-(B8), we have

$$\sqrt{n}\left(\hat{\beta}_{n1} - \beta_{n1}{}^{(0)} \right) \xrightarrow{D} N\left(0, \sum{}^{-1} \Gamma \sum{}^{-1} \right),$$

where Γ is $E\left(G\left(X,Z,h\right)\breve{X}_1\breve{X}_1^T\right)$, Σ is $E\left(F\left(X,Z,h\right)\breve{X}_1\breve{X}_1^T\right)$ and $\breve{X}_1$ is $X_1-\Omega^T\Upsilon^{-1}\Psi_1$.

4. Estimation Algorithm and Parameter Selection

4.1. Estimation Algorithm

The following is the algorithm which has combined EM algorithm with LQA algorithm, it is suitable for this paper.Assume that the previous estimations are $\hat{\gamma}_{nl}^{(m)}$ and $\hat{\beta}_{nk}^{(m)}$ respectively, where $l=1,2,\cdots,p_n$ and $k=1,2,\cdots d_n$.If they are very close to 0, then let $\hat{\gamma}_{nj}^{(m+1)}=0$ and $\hat{\beta}_{nk}^{(m+1)}=0$; otherwise, iterative estimation is performed according to the following algorithm steps.

Denote

$$Z_i^T=\left(X_i^T,\Psi_i^T\right)^T,\theta_n=\left(\beta_n,\gamma_n\right)^T$$

and let

$$\Sigma_{\lambda_1,\lambda_2}\left(\theta_n^{(m)}\right)=diag\left\{\lambda_{11}w_1\zeta\left\|\gamma_{n1}^{(0)}\right\|_H^{\zeta-1}H,\cdots,\lambda_{1d_n}w_{d_n}\zeta\left\|\gamma_{nd_n}^{(0)}\right\|_H^{\zeta-1}H,\lambda_{21}w_1\zeta\left|\beta_{n1}{}^{(0)}\right|^{\zeta-1},\cdots,\lambda_{2p_n}w_{p_n}\zeta\left|\beta_{np_n}{}^{(0)}\right|^{\zeta-1}\right\}.$$

Based on the EM algorithm, we can obtain the penalty estimation of (3) as follows:

Step 1(E-step): First, the non-penalty estimation in (3) is the initial value $\tilde{\theta}_n^{(0)}=\left(\tilde{\beta}_n^{(0)},\tilde{\gamma}_n^{(0)}\right)^T$ and the termination error, $0<\varepsilon<10^{-6}$.Then let $m=0$ and update $\pi\left(i\left|\theta_n{}^{(m)}\right.\right)$ by

$$\pi\left(i\left|\theta_n{}^{(m)}\right.\right)\propto\sum_{i=1}^{n}\phi_h\left(Y_i-Z_i{}^T\theta_n{}^{(m)}\right),$$

where

$$\begin{aligned}\tilde{\theta}_n^{(0)}&=\arg\max_{\theta}{}_n\Sigma_{i=1}^n\{\pi(i|\theta_n^{(m)})log\phi_h(Y_i-Z_i^{*T}\theta_n^{(m)})\}\\&=\left(Z_i^{*T}DZ_i^*\right)^{-1}Z_i^{*T}DY_i,\end{aligned}$$

D is a diagonal matrix of $n\times n$ whose elements of the i-th diagonal are $\pi\left(i\left|\theta_n{}^{(m)}\right.\right)$.

Due to

$$\tilde{\theta}_n^{(0)}=\arg\max_{\theta}{}_n\Sigma_{i=1}^n\{\pi(i|\theta_n^{(m)})log\phi_h(Y_i-Z_i^{*T}\theta_n^{(m)})\},$$

and

$$\phi_h\left(t\right)=\frac{1}{\sqrt{2\pi}h}e^{-\frac{t^2}{2h^2}}.$$

By calculating the logarithmic function, we have the target function

$$Q=\Sigma_{i=1}^n\{\pi(i|\theta_n^{(m)})(Y_i-Z_i^{*T}\theta_n^{(m)})^2\},$$

Next, we maximize the target function by calculating partial derivatives, we have

$$\begin{aligned}\tilde{\theta}_n^{(0)}&=\arg\max_{\theta}{}_n\Sigma_{i=1}^n\{\pi(i|\theta_n^{(m)})log\phi_h(Y_i-Z_i^{*T}\theta_n^{(m)})\}\\&=\left(Z_i^{*T}DZ_i^*\right)^{-1}Z_i^{*T}DY_i.\end{aligned}$$

Step 2(M-step): Update $\hat{\theta}_n^{(m)}$ by

$$\begin{aligned}\hat{\theta}_n^{(m+1)}&=\arg\max\sum_{i=1}^{n}\left\{\pi\left(i\left|\theta_n^{(m)}\right.\right)\log\phi_h\left(Y_i-Z_i^T\theta_n^{(m)}\right)\right\}\\&\quad+\frac{n}{2}\theta_n^{(m)^T}\sum\nolimits_{\lambda_1,\lambda_2}\left(\theta_n^{(m)}\right)\theta_n^{(m)}\\&\approx\left(Z_i^TDZ_i+n\sum\nolimits_{\lambda_1,\lambda_2}\left(\theta_n^{(m)}\right)\right)^{-1}Z_i^TDY_i.\end{aligned}$$

Step 3: Repeat Step 1 and Step 2 continuously until the algorithms are convergent and the final estimation $\hat{\theta}_n=\left(\hat{\beta}_n,\hat{\gamma}_n\right)$, where $\hat{\beta}_n=\left(I_{p_n\times p_n},0_{p_n\times d_nq}\right)\hat{\theta}_n$ and $\hat{\gamma}_n=\left(0_{d_nq\times p_n},I_{d_nq\times d_nq}\right)\hat{\theta}_n$ is obtained.

4.2. The Selection of Bandwidth, Node and Tuning Parameter

4.2.1. Bandwidth Selection

In order to select the bandwidth handily, we assume that X and Z are independent of each other.On the basis of the asymptotic variance obtained by Theorem 3 and the variance of the Least square B-spline estimation proposed by Zhao [14], we can get the ratio of two asymptotic variances as

$$r(h)=\frac{G\left(h\right)F^{-2}\left(h\right)}{\sigma^2},$$

where

$$\sigma^2=E\left(\varepsilon^2\right),G\left(h\right)=E\{\phi'_h\left(\varepsilon\right)\}^2,F\left(h\right)=E\left\{\phi''_h\left(\varepsilon\right)\right\}.$$

Define the optimal bandwidth as follows:

$$h_{opt}=argmin_h\,r(h)=argmin\,G(H)F^{-2}(H).\qquad(7)$$

In practical application, since the distribution of errors are unknown, the bandwidth obtained by equation (5) cannot be used directly.A feasible method is to estimate $F(h)$ and $G(h)$

by

$$\hat{F}(h)=\frac{1}{n}\sum_{i=1}^{n}\phi_h''(\hat{\varepsilon}_i),\hat{G}(h)=\frac{1}{n}\sum_{i=1}^{n}\{\phi_h''(\hat{\varepsilon}_i)\}^2. \quad (8)$$

Then we can obtain the estimation of $r(h)$ by

$$\hat{r}(h)=\hat{G}(h)\hat{F}^{-2}(h)/\hat{\sigma}^2,$$

where

$$\hat{\varepsilon}_i=Y_i-X_i^T\hat{\beta}_n-\sum_{l=1}^{d_n}\hat{g}_l\left(z_l\right),$$

$\hat{g}_l\left(z_l\right)$ and $\hat{\beta}_n$ are initial estimations based on the robust method.

In practical, after getting the estimation of $r(h)$, we use the lattice point search method to obtain the h_{opt}.According to the suggestion in reference by Yao [12], some of the possible lattice shops can be taken as $h_{opt}=0.5\hat{\sigma}\times 1.02^j$, where $j=0,1,\cdots,k$, k is 50 or 100.

4.2.2. The Selection of Node K and Tuning Parameter

In order to implement the estimation algorithm handily, it is necessary to select the appropriate number of internal nodes K and the tuning parameters λ_{1k} and λ_{2l} in the adoptive bridge penalty estimation. Denote

$$\lambda_{1k}=\frac{\lambda}{\left|\tilde{\beta}_{nk}\right|},\lambda_{2l}=\frac{\lambda}{\|\tilde{\gamma}_{nl}\|} \quad (9)$$

where $\tilde{\gamma}_{nl}$ and $\tilde{\beta}_{nk}$ are non-penalty estimations of γ_{nl} and β_{nk} respectively. λ and K are selected by CV criterion:

$$CV\left(K,\lambda\right)=\sum_{i=1}^{n}\phi_h\left[Y_i-\Psi_i^T\hat{\gamma}_n^{(-i)}-X_i^T\hat{\beta}_n^{(-i)}\right], \quad (10)$$

where $\hat{\gamma}_n^{(-i)}$ and $\hat{\beta}_n^{(-i)}$ are penalty estimations obtained by excepting the i-th data from equation (4).The optimal node K_{opt} and tuning parameter λ_{opt} are obtained by

$$\left(K_{opt},\lambda_{opt}\right)=\max_{K,\lambda}CV\left(K,\lambda\right). \quad (11)$$

5. Numerical Simulations

We simulate data from model (1), where $g_l\left(z_l\right)$ are the d_n-dimensional vectors, $g_1\left(Z_{i1}\right)=2Z_{i1}-E\left(2Z_{i1}\right)$, $g_2\left(Z_{i2}\right)=\left(3Z_{i2}-1\right)^2-E\left(\left(3Z_{i2}-1\right)\right)$ and $g_3\left(Z_{i3}\right)=4\sin\left(4\pi Z_{i3}\right)-E\left(4\sin\left(4\pi Z_{i3}\right)\right)$, the rest of $g_l\left(z_l\right)$ are zero; β_n is p_n-dimensional vector, $\beta_1=1.8$, $\beta_2=1$, $\beta_3=0.5$, $\beta_4=-0.8$, $\beta_5=-0.3$, the rest of β_n are zero.In this paper, both p_n and d_n increase as the sample size n increases gradually, so we assume $p_n=d_n$.To perform this simulation, we take the covariates $X\sim N(0,1)$ and $Z\sim U(0,1)$.The model error ε_i is subject to the three cases of $N(0,1)$, $t(4)$ and $0.25N(0,6)+0.75N(0,1)$ respectively.The simulated sample size n is taken as 100, 200 and 400 respectively.And the corresponding p_n is taken as 10, 20 and 40.Repeat 400 times for each simulation and calculate its average.We use a third order B-spline basic function to approximate the function of the variable coefficient part in the simulation, and optimal node and tuning parameters are selected by CV method.In order to test the effect of variable selection under high-dimensional data, we also calculate the penalty estimations through LASSO and adoptive bridge estimation based on the modal regression in this paper.We consider the cases of $\zeta=0.6$ and $\zeta=0.85$ in the adoptive bridge estimation.

To illustrate the proposed method in this paper, the efficient estimations $\hat{g}_l\left(z_l\right)$ will be assessed by using the square root of average square errors $(RASE)$

$$RASE=\left\{n^{-1}\sum_{l=1}^{grid}\left\|\hat{g}_l\left(z_l\right)-g_l\left(z_l\right)\right\|^2\right\}^{1/2}.$$

The generalized mean square error $(GMSE)$ is used to evaluate the performance of $\hat{\beta}_n$ $GMSE=\left(\hat{\beta}_k-\beta_k\right)E\left(XX^T\right)\left(\hat{\beta}_k-\beta_k\right)$.CN and CP are used to represent the number of important variables for correctly estimating additive coefficients and parameters respectively.

It can be seen from Table (1) that when the error obeys standard normal distribution, t-distribution or mixed normal distribution, the values of MRSE and GMSE under adoptive bridge estimation method are generally smaller than those under LASSO.This conclusion is still true when the sample size and dimension are getting larger, and as the dimension and sample size increase, the estimation accuracy of the model also increases.If the dimension and sample size are fixed, the values of RASE and GMSE when the error obeys the normal distribution are smaller than those when the error obeys t-distribution, which indicates that the simulation effect is better under the normal distribution.

6. Proof of Theorems

Proof of Theorem 1 Let $\delta=n^{-r/2r+1}+a_n$ and $v=\left(v_1^T,v_2^T\right)^T$, where v_1^T is a p_n-dimensional vector, v_2^T is a d_n-dimensional vector.Denote $\beta_n=\beta_n{}^{(0)}+\delta v_1$ and $\gamma_n=\gamma_n{}^{(0)}+\delta v_2$.For part (i), we first prove for any given v_1^T and v_2^T, there exists

$$\Pr\left\{\sup P_n\left(\gamma_n,\beta_n\right)-P_n\left(\gamma_n{}^{(0)},\beta_n{}^{(0)}\right)>0\right\}\to 1,$$
$$n\to\infty. \quad (12)$$

Define

$$\begin{aligned}\Pi\left(\gamma_n,\beta_n\right)=&\frac{1}{K}\left\{P_n\left(\gamma_n,\beta_n\right)-P\left({\gamma_n}^{(0)},{\beta_n}^{(0)}\right)\right\}\\=&-\frac{-\delta}{K}\sum_{i=1}^{n}{\phi'}_h\left(\varepsilon_i+1_{d_n}^T R\left(Z_i\right)\right)\left(X_i^T v_1+\Psi_i^T v_2\right)+\frac{\delta^2}{K}\sum_{i=1}^{n}{\phi''}_h\left(\varepsilon_i+1_{d_n}^T R\left(Z_i\right)\right)\left(X_i^T v_1+\Psi_i^T v_2\right)^2\\&+\frac{\delta^3}{K}\sum_{i=1}^{n}{\phi''}_h\left(\varepsilon_i^*\right)\left(X_i^T v_1+\Psi_i^T v_2\right)^3\\&+\frac{n}{K}\lambda_{1l}w_l\sum_{l=1}^{d_n}\left\{\|\gamma_{nl}\|_H^{\zeta}-\left\|{\gamma_{nl}}^{(0)}\right\|_H^{\zeta}\right\}+\frac{n}{K}\lambda_{2k}w_k\sum_{k=1}^{p_n}\left\{|\beta_{nk}|^{\zeta}-\left|{\beta_{nk}}^{(0)}\right|^{\zeta}\right\}\\\triangleq&I_{n1}+I_{n2}+I_{n3}+I_{n4}+I_{n5},\end{aligned}$$

where ε_i^* is between $\varepsilon_i+1_{d_n}^T R\left(Z_i\right)$ and $\varepsilon_i+1_{d_n}^T R\left(Z_i\right)-\delta\left(X_i^T v_1+\Psi_i^T v_2\right)$, and we denote $R\left(U\right)=\left(R_1\left(U\right),\cdots,R_{d_n}\left(U\right)\right)^T$, $R_l\left(Z\right)=g_{0l}\left(Z\right)-{\psi_l}^T\gamma_{0l}$ and $l=1,2,\cdots,d_n$. According to the regular conditions (B1) and (B2) and the conclusions in Schumaker [15], there exists $\|R_l\left(Z_i\right)\|=O\left(K^{-r}\right)$.We get $E\left({\phi_h}'\left(\varepsilon_i\right)|x,z,u\right)=0$ by expanding I_{n1} by Taylor formula and regular conditions (B4) and (B6).By simple calculation, we get

$$\begin{aligned}I_{n1}&=-K^{-1}\delta\sum_{i=1}^{n}{\phi'}_h\left(\varepsilon_i+1_{d_n}^T R\left(Z_i\right)\right)\left(X_i^T v_1+\Psi_i^T v_2\right)\\&=O_p\left(nK^{-r}\right)\|v\|.\end{aligned}$$

Similarly, we get $I_{n1}=o_p\left(I_{n2}\right)$ and $I_{n3}=o_p\left(I_{n2}\right)$. For I_{n5}, by using the inequality Hölder, we get

$$|I_{n5}|\le C_2\lambda_2 w_k\zeta {p_n}^{-1}\left(\sum_{k=1}^{p_n}{p_n}^2\right)^{1/2}\left(\sum_{k=1}^{p_n}\left(\beta_{nk}-{\beta_{nk}}^{(0)}\right)^2\right)^{1/2},k=1,2,\cdots,s_2.$$

According to the regular condition (B7) we have $|I_{n5}|\le C_2\lambda_2 w_k\zeta {p_n}^{-1/2}\,\delta\|v\|\to 0$.That is, I_{n5} is consistently controlled by I_{n2}. Similarly, I_{n4} is consistently controlled by I_{n2}.

In summary, according to the regular condition (B5) $F\left(x,z,h\right)<0$, so if we select a sufficiently large C and a vector of $\|v\|$ that can satisfy the condition, $\Pi_n\left(\gamma_n,\beta_n\right)<0$, which means

$$\Pr\left\{\sup P_n\left(\gamma_n,\beta_n\right)-P_n\left({\gamma_n}^{(0)},{\beta_n}^{(0)}\right)>0\right\}\to 1.$$

Thus, there is a local maximum point that satisfies$\left\|\hat{\beta}_{nk}-{\beta_{nk}}^{(0)}\right\|=O_p\left(n^{-r/2r+1}+a_n\right)$. That is, part (i) is proven. We use the similar method in reference [12] to complete the second part of the proof.

Proof of Theorem 2 First, we prove part (i).According to the conclusion of Theorem 1, for any given sufficiently small $v=Cn^{-r/(2r+1)}$, when $n\to\infty$, for any γ_{nj} that satisfies $||\gamma_{nl}-\gamma_{nl}^{(0)}||=O_P(n^{\frac{-r}{2r+1}})$ and β_{nk} that satisfies $||\beta_{nk}-\beta_{nk}^{(0)}||=O_P(n^{\frac{-r}{2r+1}}),k=1,\cdots,s_1$, with probability tending to 1, we have

$$\frac{\partial P_n\left(\gamma_n,\beta_{nk}\right)}{\partial\beta_{nk}}<0,\ \text{if } 0<\beta_{\mathrm{nk}}<\mathrm{v},\text{where } \mathrm{k}=\mathrm{s}_1+1,\mathrm{s}_1+2,\cdots,\mathrm{p}_\mathrm{n},\tag{13}$$

$$\frac{\partial P_n\left(\gamma_n,\beta_{nk}\right)}{\partial\beta_{nk}}>0,\ \text{if } -\mathrm{v}<\beta_{\mathrm{nk}}<0,\text{where } \mathrm{k}=\mathrm{s}_1+1,\mathrm{s}_1+2,\cdots,\mathrm{p}_\mathrm{n}.\tag{14}$$

Similar to the proof of Theorem 1, we get

$$\frac{\partial P_n\left(\gamma_{nl},\beta_{nk}\right)}{\partial\beta_{nk}}=-n\lambda_{2k}\zeta\left[|\beta_{nk}|^{\zeta-1}sgn\left(\beta_{nk}\right)+O_p\left(n^{-1}{\lambda_{2k}}^{-1}\zeta^{-1}n^{\frac{-r}{2r+1}}\right)\right].$$

By the regular conditions (B3), (B7) and (B8), with probability tending to 1, we have $\lambda_{2k}n^{-r/2r+1}\ge b_n n^{-r/2r+1}\to\infty$ where $k=s_1+1,\cdots,p_n$. Therefore, the sign of $\partial P_n(\gamma_n,\beta_{nk})/\partial\beta_{nk}$ is completely determined by the sign of β_{nk}, that is, the equations of (15) and (16) are established.

For the proof of part (ii), using a method similar to the above, with probability to 1 we have $\hat{g}_{nl}(z_l) = 0$, where $l = s_2 + 1, \cdots, d_n$, that is, the conclusion of part (ii) is established.

Proof of Theorem 3 From the proof of Theorem 1 and Theorem 2, we know that $P_n(\gamma_n, \beta_n)$ attains the maximal value at $(\hat{\beta}_{n1}^T, 0)^T$ and $(\hat{\gamma}_{n1}^T, 0)^T$ with probability tending to 1 when $n \to \infty$. Denote

$$P_{n1}(\gamma_n, \beta_n) = \frac{\partial P_{n1}(\gamma_n, \beta_n)}{\partial \beta_{n1}} \text{ then } \left(\hat{\beta}_{n1}^{\mathrm{T}}, 0\right)^{\mathrm{T}}$$

must satisfy the following equation

$$\begin{aligned} &\frac{1}{n} P_{n1}\left(\left(\hat{\gamma}_{n1}^T, 0\right)^T, \left(\hat{\beta}_{n1}^T, 0\right)^T\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} X_{ia}\phi'_h\left(Y_i - \Psi_i^T\hat{\gamma}_{nl} - X_i^T\hat{\beta}_{nk}\right) - \lambda_{2k} w_k \zeta \sum_{k=1}^{p_n}\left|\hat{\beta}_{nk}\right|^{\zeta-1} sgn\left(\hat{\beta}_{nk}\right) = 0. \end{aligned} \tag{15}$$

Expanding the second term below the equation (17) by Taylor formula, we have

$$\begin{aligned} &n^{-1}\sum_{i=1}^{n} Z_{i1}\left\{\phi'_h(\epsilon_i) + \phi''_h(\epsilon_i)\left[X_i^T R^*(U_i) - Z_{i1}^T\left(\hat{\beta}_{n1} - \beta_{n1}{}^{(0)}\right) - W_{i1}^T\left(\hat{\gamma}_{n1} - \gamma_{n1}{}^{(0)}\right)\right]\right. \\ &\left. + \phi'''_h(\varsigma_i)\left[X_i^T R^*(U_i) - Z_{i1}^T\left(\hat{\beta}_{n1} - \beta_{n1}{}^{(0)}\right) - W_{i1}^T\left(\hat{\gamma}_{n1} - \gamma_{n1}{}^{(0)}\right)\right]^2\right\} + o_p\left(\hat{\beta}_{n1} - \beta_{n1}{}^{(0)}\right) = 0. \end{aligned} \tag{16}$$

where ς_i is between ε_i and $Y_i - \Psi_{i1}^T\hat{\gamma}_{n1} - X_{i1}^T\hat{\beta}_{n1}$, $R^*(u) = (R_1(u), \cdots, R_{s_1}(u))^T$. According to the regular conditions $\sqrt{n}b_n \to \infty$ and $\left(\hat{\beta}_{nk} - \beta_{nk}{}^{(0)}\right) = o_p(1)$, we obtain

$$\hat{\gamma}_{n1} - \gamma_{n1}{}^{(0)} = (\Upsilon_n + o_p(1))^{-1}\left\{-\Omega_n\left(\hat{\beta}_{n1} - \beta_{n1}{}^{(0)}\right) + \mathrm{T}_n\right\}. \tag{17}$$

where

$$\mathrm{T}_n = n^{-1}\sum_{i=1}^{n} W_{i1}\left[\phi'_h(\varepsilon_i) + \phi''_h(\varepsilon_i) X_{i1}^T R^*(U_i)\right].$$

Substituting into equation (16), we obtain

$$E\left(n^{-1}\sum_{i=1}^{n}\phi''_h(\varepsilon_i)\Omega^T\Upsilon^{-1}\Psi_{i1}\left[X_{i1}^T - \Psi_{i1}^T\Upsilon^{-1}\Omega^T\right]\right) = 0.$$

After calculations, we have

$$\begin{aligned} &\left\{n^{-1}\sum_{i=1}^{n}\phi''_h(\epsilon_i)\breve{Z}_{i1}\breve{Z}_{i1}^T + o_p(1)\right\}\sqrt{n}\left(\hat{\beta}_{n1} - \beta_{n10}\right) \\ &= n^{-1/2}\sum_{i=1}^{n}\breve{Z}_{i1}\phi'_h(\varepsilon_i) + n^{-1/2}\sum_{i=1}^{n}\breve{Z}_{i1}\phi''_h(\varepsilon_i) X_i^T R^*(U_i) - n^{-1/2}\sum_{i=1}^{n}\breve{Z}_{i1}\phi''_h(\varepsilon_i)\left(\Upsilon^{-1} + o_p(1)\right)\mathrm{T}_n \\ &= J_{n1} + J_{n2} + J_{n3}, \end{aligned}$$

where $\breve{X}_{i1} = X_{i1} - \Omega_n{}^T\Upsilon_n{}^{-1}\mathrm{T}_i$.According to the definition of $R^*(U_i)$, for J_{n2}, we have

$$\begin{aligned} J_{n2} &= n^{-1/2}\sum_{i=1}^{n}\phi''_h(\varepsilon_i)\left(Z_i^T - E(\Omega_n)^T E(\Upsilon_n)^{-1} W_{i1}\right) X_{i1}^T R^*(U_i) \\ &+ n^{-1/2}\sum_{i=1}^{n}\phi''_h(\varepsilon_i) E(\Omega_n)^T\left(E(\Upsilon_n)^{-1} - \Omega_n{}^T\Upsilon_n{}^{-1} W_{i1}\right) X_i^T R^*(U_i) \\ &= J_{n21} + J_{n22}. \end{aligned}$$

Since $E\left(\phi''_h(\varepsilon_i)\left\{X_i{}^T - E(\Omega_n)^T\left[(\Upsilon_n)^{-1}\Psi_{i1}{}^T\right]\Psi_{i1}\right\}\right) = 0$, we have

$$n^{-1/2}\sum_{i=1}^{n}\phi''_h(\varepsilon_i)\left(X_i{}^T - E(\Omega_n)^T E\ (\Upsilon_n)^{-1}\Psi_{i1}{}^T\right)\Psi_{i1} = o_p(1).$$

Combining $W_i = \left({W_{i1}}^T, {W_{i2}}^T, \cdots, {W_{ip_n}}^T\right)$ and $|R^*(U_i)| = o_p(1)$, we get $J_{n21} = o_p(1)$.Similarly, $J_{n22} = o_p(1)$.

Further, according to the regular conditions, we have $n^{-1}\sum_{i=1}^{n}{\phi''}_h(\varepsilon_i)\breve{X}_{i1}\breve{X}_{i1}^T \xrightarrow{P} \Sigma$.According to the Central limit theorem, there is

$$J_{n1} = n^{-1/2}\sum_{i=1}^{n}{\phi'}_h(\varepsilon_i)\breve{X}_{i1}^T \xrightarrow{d} N(0,\Gamma). \tag{18}$$

where $\Gamma = E\left(G(x,z,h)\breve{X}_{i1}\breve{X}_{i1}^T\right)$.

For any vectors ρ whose components are not all zero,

$$\rho^T P_{n1} = n^{-1/2}\sum_{i=1}^{n}\rho^T{\phi'}_h(\varepsilon_i)\breve{X}_{i1}^T \triangleq \sum_{i=1}^{n}o_i\xi_i.$$

where ${o_i}^2 = n^{-1}\sum_{i=1}^{n}G(X,Z,h)\rho^T\breve{X}_{i1}\breve{X}_{i1}^T\rho$.Under the given conditions of $\{X_i, Z_i\}$, ξ_i is independent with mean zero and variance one.

Then, we verify Linderberg's central limit theorem.If

$$\frac{\max {o_i}^2}{\sum_{i=1}^{n}{o_i}^2} \xrightarrow{P} 0, \tag{19}$$

holds, then

$$\sum_{i=1}^{n}o_i\xi_i/\sqrt{\sum_{i=1}^{n}{o_i}^2} \xrightarrow{D} N(0,1)$$

and thus the equation (18) holds.

Next, we prove (19). Owing to $\left(\rho^T\breve{Z}_{i1}^T\right)^2 \le \|\rho\|^2\left\|\breve{Z}_{i1}\right\|^2$, we have ${o_i}^2 \le n^{-1}G(X,Z,U,h)\|\rho\|^2\left\|\breve{Z}_{i1}\right\|^2$ and $\left\|\breve{Z}_{i1}\right\| = \left\|Z_{i1} - \Omega^T\Upsilon^{-1}W_{i1}\right\| \le \|Z_{i1}\| + \left\|\Omega^T\Upsilon^{-1}W_{i1}\right\|$. According to the regular conditions, we have $\max_i\|Z_i\|n^{-1/2} = o_p(1)$ and $\max_i\left\|\Omega^T\Upsilon^{-1}W_i\right\| = o_p(1)$. By applying Slutsky theorem again, then we can prove that equation (18) is true.

In conclusion, we obtain

$$\sqrt{n}\left(\hat{\beta}_{n1} - {\beta_{n1}}^{(0)}\right) \xrightarrow{D} N\left(0, \Sigma^{-1}\Gamma\Sigma^{-1}\right).$$

7. Conclusion

Combining the results of the above theoretical proof and numerical simulations, we can get the following conclusions.It is proved by theoretical properties that adoptive bridge estimation can accurately screen non-zero parameters with probability tending to 1 under high-dimensional data.Numerical simulations tested the performance of the proposed methods in a finite sample and verified the significance of modal regression estimation and the variable selection methods.

References

[1] Hastie, T. J, Tibshirani, R. J. (2017). Generalized Additive Models. New York, Routledge. doi: 10.1201/9780203753781.

[2] Guo, J, Tang, M. L, Tian, M. Z, Zhu, K. (2013).Variable selection in high-dimensional partially linear additive models for composite quantile regression. Computational Statistics and Data Analysis, 65 (9), 56. doi: 10.1016/j.csda.2013.03.017.

[3] Wang, L, Liu, X, Liang, H, Raymond, J. C. (2011). Estimation and variable selection for semiparmetric additive partial linear models. Statistica Sinica, 21 (3), 1225. doi: 10.2307/23033585.

[4] Xia, Y. F, Qu, Y. R, Sun, N. L. (2018). Variable selection for semiparametric varying coefficient partially linear model based on modal regression with missing data. Communications in Statistics Theory and Methods, 48 (20), 5121. doi: 10.1080/03610926.2018.1508712.

[5] Fan, J. Q, Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. Bernoulli, 11 (6), 1031. doi: 10.2307/25464778.

[6] Hoshino, T. (2014). Quantile regression estimation of partially linear additive models. Journal of Nonparametric Statistics, 26 (3), 509. doi: 10.1080/10485252.2014.929675.

[7] Meinshausen, N, Bhlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. The Annals of statistics, 34 (3), 1436. doi: 10.1214/009053606000000281.

[8] Zhang, C. H, Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. The Annals of Statistics, 36 (4), 1567. doi: 10.1214/07-AOS520.

[9] Wang, M. Q, Song, L. X, Wang, X. G. (2010). Bridge estimation for generalized linear models with a diverging number of parameters. Statistics Probability Letters, 80 (21), 1584. doi: 10.1016/j.spl.2010.06.012.

[10] Li, K. P, Li, D. G, Liang, Z. W. (2017). Estimation of semi-varying coefficient models with nonstationary regressors. Econometric Reviews, 36 (1), 354. doi: 10.1080/07474938.2015.1114563.

[11] Lam, C, Fan, J. Q. (2008). Profile-kernel likelihood inference with diverging number of parameters. The Annals of Statistics, 36(5), 2232. doi: 10.1214/07-AOS544.

[12] Yao, W. X, Lindsay, B. G, Li, R. Z. (2012). Local modal regression. Journal of Nonparametric Statistics, 24 (3), 647. doi: 10.1080/10485252.2012.678848.

[13] Fan, J. Q, Li, R. Z. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association, 96(456), 1348. doi: 10.1198/016214501753382273.

[14] Zhao, P. X, Xue, L. G. (2009). Variable selection for semiparametric varying coefficient partially linear models. Statistics Probability Letters, 79(20), 2148. doi: 10.1016/j.spl.2009.07.004.

[15] Schumaker, L. L. (2007).Splines Function: Basic Theory. Cambridge, Cambridge University Press. doi:10.1017/CBO9780511618994.