International Journal of Statistical Distributions and Applications
2023; 9(2): 49-61
http://www.sciencepublishinggroup.com/j/ijsda
doi: 10.11648/j.ijsd.20230902.11
ISSN: 2472-3487 (Print); ISSN: 2472-3509 (Online)

Variable Selection for Semi-Parametric Models with Interaction Under High Dimensional Data

Yafeng Xia, Na Kui*

School of Science, Lanzhou University of Technology, Lanzhou, China

Email address:

gsxyf01@163.com (Yafeng Xia), 2624714229@qq.com (Na Kui)

*Corresponding author

To cite this article:

Yafeng Xia, Na Kui. Variable Selection for Semi-Parametric Models with Interaction Under High Dimensional Data. *International Journal of Statistical Distributions and Applications*. Vol. 9, No. 2, 2023, pp. 49-61. doi: 10.11648/j.ijsd.20230902.11

Received: March 9, 2023; **Accepted**: April 2, 2023; **Published**: April 15, 2023

Abstract: With the continuous development of modern science and technology and the continuous improvement of data collection technology, researchers can collect a lot of high-dimensional data from various fields. At present, there has been some development in the selection of variables under high-dimensional data, but most of these studies only consider the selection of variables for main effects. However, when modeling many important practical problems, the main effects alone may not be enough to describe the relationship between the response variable and the predictor variable. Therefore, the variable selection problem with interaction terms under high-dimensional data is more meaningful. Based on this, this article focus on the robust estimation for semi-parametric models with interactions in high-dimensional data under the framework of mode regression. And the two-stage regularization method is applied to implement variable selection with high-dimensional data. At Stage 1, using the B-spline basic function to approximate the non-parametric function. Both parametric and non-parametric components were selected simultaneously based on mode regression and the adaptive least absolute shrinkage and selection operator (LASSO) estimation. At Stage 2, the model variables are composed of the selected variables at Stage 1 and interaction terms are derived from the main effects. To maintain the heredity structure between main effects of linear part and interaction effects, we only selected the interaction terms to obtain important interaction effects. Then, under proper regularization conditions, oracle properties of variable selection and the consistency of the hierarchical structure are proved. Numerical results are also shown to demonstrate performance of the methods.

Keywords: Semi-Parametric Models with Interaction, Variable Selection, Modal Regression, Adaptive LASSO

1. Introduction

In many practical problems, the main terms $X_1,\cdots,X_p$ alone may not be sufficient to depict the relationship between response and predictor variables. Therefore, the problem of variable selection with interaction terms under high-dimensional data is of more practical significance. In this paper, we consider the semi-parametric model with interaction, that is

$$Y_i=\beta_0+\beta_1X_1+\cdots+\beta_pX_p+\beta_{1,1}X_1^2+\beta_{1,2}X_1X_2+\cdots+\beta_{p,p}X_p^2+\sum_{l=1}^{d}g_l\left(Z_{il}\right)+\varepsilon_i, \tag{1}$$

where Y_i is the response, $\beta=\left(\beta_1,\cdots,\beta_p\right)^T$ is the p-dimension unknown regression parameter vector, $g_1(\cdot),\cdots,g_d(\cdot)$ are respectively one-dimensional unknown smooth function, $X_1,\cdots,X_p$ are main effects, and order-2 terms $X_jX_k\left(1\le j\le k\le p\right)$ include quadratic main effects $\left(j=k\right)$ and two-way interaction effect $\left(j\ne k\right)$. $Z_i\in R^d$ is d-dimensional covariable, and ε is the noise with mean zero. A key feature of model (1) is that order-2 terms are derived from the main effects. We call X_jX_k the child of X_j and X_k, and X_j and X_k the parents of X_jX_k.

In recent years, the semi-parametric variable selection method has attracted the attention of statisticians. Liang [1] proposed a partial linear model-based profiled forward regression (PFR) screening method, which transformed the variable coefficient model into a traditional linear model with a profile least square method and obtained the PFR screening method through the forward regression algorithm. For robust variable selection, Zhao [2] approximated the non-parametric function with B-spline and realized simultaneously variable selection of parameters and variable coefficients based on the double penalty mode regression objective function. Based on mode regression, Zhang and Zhao [3] studied the two-step estimation of the partial variable coefficient model by using local polynomials, and obtained the estimation of the parameters and non-parametric functions respectively and reached the optimal convergence rate respectively.

Since Efron [4] proposed the two-stage method of linear models with interaction effects, a large number of literatures have extended the two-stage method to various models. Hao and Zhang [5] proposed a forward-selection-based algorithms for interaction selection (iFOR) in the high dimensional interactive selection, which can effectively identify interaction effects and maintain the hierarchy. Subsequently, Hao and Feng [6] proposed regularization methods for high-dimensional quadratic regression models with second-order interaction terms and established theoretical properties for two-stage LASSO. To deal with ultra-high-dimensional problems, Dong and Jiang [7] studied a two-stage method that requires sparsity in the high-order interaction model and used the square root hard ridge (SRHR) method to discover relevant variables in each stage.

Inspired by the above methods, this paper proposes and studies the robust variable selection of a semi-parametric model with interaction in high dimensional data, and adopts a two-stage regularization method for variable selection.

2. Variable Selection Method

We define some notations used in the paper. Let $X=\left(X_1,\cdots,X_n\right)^T$ be the $n\times p$ design matrix of the main effect and $\left(y_1,\cdots,y_n\right)^T$ be the n -dimensional response vector, and let the index set of the linear term be $M=\{1,2,\cdots,p\}$ and the index set of the second-order term be $\mathrm{I}=\{(j,k):1\le j\le k\le p\}$. The regression coefficient vector $\beta=\left(\beta_0,\beta_M^T,\beta_{\mathrm{I}}^T\right)^T$, where $\beta_{\mathrm{M}}=\left(\beta_1,\cdots,\beta_p\right)^{\mathrm{T}}$ and $\beta_{\mathrm{I}}=\left(\beta_{1,1},\beta_{1,2},\cdots,\beta_{p,p}\right)^{\mathrm{T}}$. For a subset $A\subset M$, β_A is a subvector of β_M , X_A is a submatrix of X , and X_j represents the jth column of X . In addition, the subscript (j,k) and (k,j) are the same, that is, $\beta_{j,k}=\beta_{k,j}$.

To estimate the non-parametric part of the semi-parametric model (1), we use the B-spline basis function to approximate every non-parametric function $g_l(z)$, and model (1) is expressed as

$$Y_i\approx\beta_0+\beta_1X_1+\cdots+\beta_pX_p+\beta_{1,1}X_1^2+\beta_{1,2}X_1X_2+\cdots+\beta_{p,p}X_p^2+\Psi_i^T\gamma+\varepsilon_i, \tag{2}$$

where $\gamma=\left(\gamma_1^T,\cdots,\gamma_d^T\right)$, and the definition of Ψ_i is described in lv [8].

2.1. Variable Selection at Stage 1

The objective function of the first stage is given below

$$L(\beta_0,\beta,\gamma)=\sum_{i=1}^{n}\phi_h\left\{Y_i-\beta_0-X_i^T\beta_M-\Psi_i^T\gamma\right\}, \tag{3}$$

where $\phi_h(t)=h^{-1}\phi(t/h)$, and $\phi(t)$ is a kernel density function and h is window width. $\tilde{\beta}$ and $\tilde{\gamma}$ are estimated by maximizing the objective function (3). In order to realize variable selection, according to the mode regression estimation method of Yao [9], the adaptive LASSO penalty function is introduced, then the objective function of the first stage becomes

$$G_1(\beta_0,\beta_M,\gamma)=\sum_{i=1}^{n}\phi_h\left\{Y_i-\beta_0-X_i^T\beta_{\mathrm{M}}-\Psi_i^T\gamma\right\}-n\sum_{k=1}^{p}\lambda_{1k}\omega_k\left|\beta_k\right|-n\sum_{l=1}^{d}\lambda_{2l}\omega_l\left\|\gamma_l\right\|_{H_l}, \tag{4}$$

where $\phi_h(t)=h^{-1}\phi(t/h)$, $\phi(t)$ is a kernel density function, h is window width, λ_1 and λ_2 are regularization parameters, and ω_k and ω_l are the penalty weights. Generally, $\omega_k=\left|\tilde{\beta}_k\right|^{-v}$ and $\omega_l=\left|\tilde{\gamma}_l\right|^{-v}$ are taken, where v is a positive constant, $\left\|\gamma_l\right\|_{H_l}=\left(\gamma_l^TH\gamma_l\right)^{1/2}$ and the definition of H_l are shown in literature Li [10]. Sparse solutions $\hat{\beta}$ and $\hat{\gamma}$ of λ_1 and λ_2 can be obtained by selecting appropriate regularization parameters λ_1 and λ_2 and maximizing objective function (4), and the estimation of $\hat{g}_l$ can be expressed by $\hat{\gamma}$.

Since it is difficult to maximize equation (4) directly, local quadratic estimation algorithms (LQA) and expectation

maximization (EM) proposed by Fan et al. (2001) [11] are adopted to obtain penalty estimation. Given initial values $\left|\beta_k^0\right|>0, k=1,\cdots,p$ and $\left\|\gamma_l^0\right\|>0, l=1,\cdots,d$, the penalty function can be expressed as

$$\lambda_{1k}\omega_k\left|\beta_k\right| \approx \lambda_{1k}\omega_k\left|\beta_k^{(0)}\right| + \frac{1}{2}\frac{\omega_k\lambda_{1k}}{\left(\left|\beta_k^{(0)}\right|\right)}\left(\left|\beta_k\right|^2 - \left|\beta_k^{(0)}\right|^2\right), \tag{5}$$

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

So for ease of calculation, let $\theta=\left(\beta^T,\gamma^T\right)^T$, $Z_i=\left(X_i^T,\Psi_i^T\right)$ and $m=0$, we have

$$\Sigma_{\lambda_1,\lambda_2}\left(\theta^{(m)}\right) = diag\left\{\lambda_{11}\omega_1\left|\beta_1^0\right|^{-1},\cdots,\lambda_{1p}\omega_p\left|\beta_p^0\right|^{-1},\lambda_{21}\omega_1\left\|\gamma_l^0\right\|_H^{-1}H,\cdots,\lambda_{2d}\omega_d\left\|\gamma_d^0\right\|_H^{-1}H\right\},$$

The EM algorithm to obtain the penalty estimation in (4) is as follows:

Step 1 (E-Step): Update $\pi\left(i\middle|\theta^{(m)}\right)$ by the following formula

$$\pi\left(i\middle|\theta^{(m)}\right)=\frac{\phi_h\left\{Y_i-Z_i^T\theta^{(m)}\right\}}{\sum_{i=1}^n\phi_h\left\{Y_i-Z_i^T\theta^{(m)}\right\}},$$

Step 2 (M-Step): Update $\theta^{(m+1)}$

$$\theta^{(m+1)}=\arg\max_\theta\sum_{i=1}^n\left\{\pi\left(i\middle|\theta^{(m)}\right)\log\phi_h\left\{Y_i-Z_i^T\theta^{(m)}\right\}-\frac{n}{2}\theta^T\Sigma_{\lambda_1,\lambda_2}\left(\theta^{(m)}\right)\theta\right\}$$

$$=\left(Z^TWZ+n\Sigma_{\lambda_1,\lambda_2}\left(\theta^{(m)}\right)\right)^{-1}Z^TWZ,$$

where $Z=\left(Z_1,\cdots,Z_n\right)^T$, W is a $n\times n$ diagonal matrix and its i th diagonal element is $\pi\left(i\middle|\theta^{(m)}\right)$.

Step 3: Iterate E-Step and M-Step repeatedly until the algorithm converges. Let the estimate of θ be $\hat{\theta}$, then $\hat{\beta}=\left(I_{p\times p},0_{p\times dK_n}\right)\hat{\theta}$, $\hat{\gamma}=\left(0_{dK_n\times p},I_{dK_n\times dK_n}\right)\hat{\theta}$, and $\hat{g}_l=\psi_l\left(z_l\right)^T\hat{\gamma}_l$, $l=1,\cdots,d$.

2.2. Variable Selection at Stage 2

Define $X^{\circ 2}=X\circ X$ as an $n\times\frac{p(p+1)}{2}$ matrix consisting of all pairwise column products, that is, for $X=\left(X_1,\cdots,X_p\right)$, $X^{\circ 2}=X\circ X=\left(X_1*X_1,X_1*X_2,\cdots,X_p*X_p\right)$, where $*$ represents the entry-wise product of two column vectors. For an index set $A\subset M$, define $A^{\circ 2}=A\circ A=\left\{(j,k):1\le j\le k\le p;j,k\in A\right\}\subset I$, and use $X_A^{\circ 2}$ as a short notation for $\left(X_A\right)^{\circ 2}$, and the columns of $\left(X_A\right)^{\circ 2}$ are indexed by $A^{\circ 2}$. $\hat{A}=\left\{j:\hat{\beta}_j\neq 0,j=1,\cdots,p\right\}\subset M$ is used to represent the model selected in the first stage. In the second stage, $\hat{A}$ is extended by all the two-way interactions of those main effects within $\hat{A}$, and the nonparametric part of the first stage is denoted by $\hat{L}=\left\{l:\hat{g}_l\neq 0,l=1,\cdots,s\right\}$.

To maintain the hierarchy structure, only interaction terms are penalized in the second stage, and the coefficient of the interaction term x_jx_k should satisfy $\beta_{j,k}\neq 0$, if and only if $\beta_j\neq 0$ and $\beta_k\neq 0$. The objective function of the second stage is

$$G_2\left(\beta_0,\beta_{\hat{A}},\beta_{\hat{A}^{\circ 2}},\gamma_{\hat{L}}\right)=\sum_{i=1}^n\phi_h\left\{Y_i-\beta_0-X_{\hat{A}}\beta_{\hat{A}}-X_{\hat{A}}^{\circ 2}\beta_{\hat{A}^{\circ 2}}-\Psi_{\hat{L}}\gamma_{\hat{L}}\right\}-n\sum_{\alpha\in\hat{A}^{\circ 2}}\lambda\omega_\alpha\left|\beta_{k,j}\right|. \tag{7}$$

For the variable selection in the second stage, the following formula is solved

$$\hat{\mu}^{AL} = \arg\max_{\mu} \left\{ \sum_{i=1}^{n} \phi_h \left\{ Y_i - \beta_0 - X_{\hat{A}} \beta_{\hat{A}} - \Psi_{\hat{L}} \gamma_{\hat{L}} - X_{\hat{A}}^{\circ 2} \beta_{\hat{A}^{\circ 2}} \right\} - n \sum_{\alpha \in \hat{A}^{\circ 2}} \lambda \omega_{\alpha} \left| \beta_{j,k} \right| \right\}, \tag{8}$$

where λ is the regularization parameter, and ω_{α} is the penalty weight of the α th component of $\beta_{k,j}$. Generally, $\omega_{\alpha} = \left| \beta_{\alpha} \right|^{-v}$ is taken, where v is a positive constant.

By using appropriate penalty parameters, coefficient estimators can be obtained and variable selection can be achieved. In this stage, only interaction terms are penalized. The iterative calculation process is the same as that in the first stage.

3. Theoretical Properties

If we ignore the interaction terms, let β_M^0 and $g_{0l}(z_l)$ be the true values of β_M and $g_l(z_l)$, respectively. Let $a_n = \max_{k,l} \left\{ \lambda_{1k}, \lambda_{2l} \left| \beta_{0k} \neq 0, \gamma_{0l} \neq 0 \right. \right\}$ and $\sqrt{n} a_n \to 0$, where $n \to \infty$. Let $b_n = \min_{k,l} \left\{ \lambda_{1k}, \lambda_{2l} \left| \beta_{0k} = 0, \gamma_{0l} = 0 \right. \right\}$ and $\sqrt{n} b_n \to \infty$, where $n \to \infty$. At Stage 2, to keep the hierarchy, the coefficient of the interaction term $x_j x_k$ should satisfy $\beta_{j,k} \neq 0$, if and only if $\beta_j \neq 0$ and $\beta_j \neq 0$, $1 \le j \le k \le p; j,k \in \hat{A}$.

Theorem 1. Suppose that (C-1) - (C-6) hold, and the number of nodes is $K = O\left(n^{1/2r+1}\right)$, then

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

(2) $\left\| \hat{g}_l(z_l) - g_{0l}(z_l) \right\| = O_p\left(n^{-r/2r+1} + a_n\right), l = 1, \cdots, s$.

The definition of r is shown in the regular condition (C-2).

Let $\beta_{0k} \neq 0, k = 1, \cdots, t$, and $\beta_{0k} = 0, k = t+1, \cdots, p$. Similarly, let $g_{0l} \neq 0, l = 1, \cdots, s$ and $g_{0l} = 0, l = s+1, \cdots, d$. Define $\beta = \left(\beta_a^T, \beta_b^T\right)^T$, and $\gamma = \left(\gamma_a^T, \gamma_b^T\right)^T$, where $\beta_a = (\beta_1, \cdots, \beta_t)^T$ and $\gamma_a = (\gamma_1, \cdots, \gamma_s)^T$ correspond to covariates X_a and Ψ_a. Under the regular condition, the first-stage penalty estimation has sparsity.

Theorem 2. Suppose that (C-1) - (C-6) hold, and the number of nodes is $K = O\left(n^{1/2r+1}\right)$, then

(1) $\hat{\beta}_k = 0, k = t+1, \cdots, p$,

(2) $\hat{g}_l = 0, l = s+1, \cdots, d$.

The asymptotic normal distribution of the nonzero component of the parameter is given. Let

$$F(x,z,h) = E\left\{ \phi_h''(\varepsilon) \middle| X = x, Z = z \right\},$$

$$G(x,z,h) = E\left\{ \phi_h'(\varepsilon)^2 \middle| X = x, Z = z \right\},$$

$$\Phi = E\left\{ \phi_h''(\varepsilon) \Psi \Psi^T \right\} = E\left\{ F(x,z,h) \Psi \Psi^T \right\},$$

$$\Gamma = E\left\{ \phi_h''(\varepsilon) \Psi X^T \right\} = E\left\{ F(x,z,h) \Psi X^T \right\},$$

$X^* = X - \Gamma^T \Phi^{-1} \Psi$, $\Sigma = E\left\{ F(X,Z,h) X^* X^{*T} \right\}$,

$$\Omega = E\left\{ G(X,Z,h) X^* X^{*T} \right\}.$$

Theorem 3. Under the condition of Theorem 3.2, then

$$\sqrt{n}\left(\hat{\beta}_a - \beta_{0a} \right) \xrightarrow{d} N\left(0, \Sigma_a^{-1} \Omega_a \Sigma_a^{-1}\right),$$

where Σ_a and Ω_a are the pre-$t \times t$ submatrices of Σ and Ω, respectively.

Theorem 4. Suppose the regularization condition (C-1) - (C-6) hold, and the number of nodes is $K = O\left(n^{1/2r+1}\right)$, then

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

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

(3) $\left\| \hat{g}_l(z_l) - g_{0l}(z_l) \right\| = O_p\left(n^{-r/2r+1} + a_n\right), l \in \hat{L}$.

Theorem 5. Suppose the regularization condition (C-1) - (C-6) hold, and the number of nodes $K = O\left(n^{1/2r+1}\right)$, let $\tilde{c} = \left\{ (k,j) : \hat{\beta}_{k,j} = 0, t+1 \le k \le j \le p \right\}$, then

$$\hat{\beta}_{k,j} = 0, (k,j) \in \tilde{c}.$$

Let Σ_{AB} be the submatrix of Σ, the row index is A and the column index is B. Refer to Hao and Zhang [12].

and define the index set of important main effects as $S = \left\{ j : \beta_j^2 + \sum_{k=1}^{p} \beta_{j,k}^2 > 0 \right\}$, and let $s = |S|$, $T = \left\{ (k,l) : \beta_{k,l} \neq 0 \right\}$, and $T \subset S^{\circ 2}$. In addition, we follow Wainwright [13] and define $\Sigma_{S^c|S} = \Sigma_{S^c S^c} - \Sigma_{S^c S} \left(\Sigma_{SS}\right)^{-1} \Sigma_{SS^c}$, where $S^c = M - S$. Let $\Lambda_{\min}(A)$ be the smallest eigenvalue of A, and $\rho_u(A) = \max_i A_{ii}$. Assume the following technical conditions:

(B1) $\left\| \Sigma_{S^c S} \left(\Sigma_{SS}\right)^{-1} \right\|_{\infty} \le 1 - \gamma, \gamma \in (0,1]$,

(B2) $\Lambda_{\min}\left(\Sigma_{SS}\right) \ge C_{\min} > 0$.

Theorem 6. Suppose that the regular condition (B-1)-(B-2) hold, consider the family of regularized parameters

$$\lambda_n\left(\phi_p\right)=\sqrt{\frac{\phi_p\rho_u\left(\Sigma_{S^c|S}\right)}{\gamma^2}\frac{4\left(\sigma^2+\tau^2\right)\log\left(p\right)}{n}},$$

where $\phi_p \geq 2$. For some fixed $\delta>0$, the sequence (n,p,s) and regularization sequence $\{\lambda_n\}$ satisfy

$$\frac{n}{2s\log\left(p-s\right)}>\left(1+\delta\right)\frac{\rho_u\left(\Sigma_{S^c|S}\right)}{C_{\min}\gamma^2}\left(1+\frac{2\left(\sigma^2+\tau^2\right)C_{\min}}{\lambda_n^2 s}\right).$$

then the following holds with probability greater than

$$1-c_1\exp\left(-c_2\min\left\{s,\log\left(p-s\right),n^{\frac{1}{2}}\right\}\right).$$

(i) The adaptive LASSO has a unique solution;
(ii) Define the gap

$$g\left(\lambda_n\right)=c_3\lambda_n\left\|\Sigma_{SS}^{-1/2}\right\|_\infty^2+20\sqrt{\frac{\sigma^2 s}{C_{\min}n}}+\frac{9\left\|\beta_I\right\|_2\sqrt{s}}{C_{\min}n^{1/3}}.$$

Then if $\beta_{\min}=\min_{j\in S}\left|\beta_j\right|>g\left(\lambda_n\right)$, then $sign\left(\hat{\beta}_L\right)=sign\left(\beta_M\right)$.

4. Selection of Window Width and Adjustment Parameters

4.1. Selection of Window Width

Referring to theorem 1 of Liu [14], we can obtain that the ratio of the spline mode estimate to the asymptotic variance of the least squares estimate is

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

where $\sigma^2=E\left(\varepsilon^2\right)$, $F\left(h\right)=E\left\{\phi_h''\left(\varepsilon\right)\right\}$, and $G\left(h\right)=E\left\{\phi_h'\left(\varepsilon\right)\right\}^2$. The optimal window width h can be obtained by minimizing $R\left(h\right)$, that is,

$$h_{opt}=\arg\min_h R\left(h\right)=\arg\min_h G\left(h\right)F^{-2}\left(h\right).$$

Therefore, h_{opt} does not depend on n but only on the distribution of ε. In practical problems, the distribution of model error ε is usually unknown, so the window width cannot be solved directly by equation (3). Considering the method proposed by Yao [9], the estimation of $F\left(h\right)$ and $G\left(h\right)$ is used in equation (3), that is,

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

then the estimate of $r\left(h\right)$ is $\hat{r}\left(h\right)=\frac{\hat{G}\left(h\right)\hat{F}^{-2}\left(h\right)}{\hat{\sigma}^2}$, where $\hat{\varepsilon}_i=Y_i-X_i^T\hat{\beta}-\Psi_i^T\hat{\gamma}$ represents the estimation of the residuals, $\hat{\gamma}$, $\hat{\beta}$, and $\hat{\sigma}$ are the initial estimates, and the lattice point method can be used to obtain the optimal window width.

4.2. Selection of Adjustment Parameters

According to Zhao [15], let

$$\lambda_{1k}=\frac{\lambda_1}{\left|\tilde{\beta}_k\right|},\lambda_{2l}=\frac{\lambda_2}{\left\|\tilde{\gamma}_l\right\|_{H_l}},$$

where $\tilde{\beta}_k$ and $\tilde{\gamma}_l$ are the no-penalty estimates of β_k and γ_l, respectively. To choose the optimal regularization parameters λ_1 and λ_2, Bayesian criterion (BIC) is applied in this paper. Particularly, BIC is defined as

$$\mathrm{BIC}\left(\lambda_1,\lambda_2\right)=-\frac{1}{n}\sum_{i=1}^n\phi_h\left\{Y_i-\beta_0-X_i^T\hat{\beta}_{\lambda_1}-\Psi_i^T\hat{\gamma}_{\lambda_2}\right\}+df_n\frac{\log\left(n/K_n\right)}{n/K_n}+df_c\frac{\log\left(n\right)}{n},$$

where $\hat{\beta}_{\lambda_1}$ and $\hat{\gamma}_{\lambda_2}$ represent the penalty estimates given λ_1 and λ_2, df_n is the number of nonzero additive functions and df_c is the number of nonzero parameter.

5. Simulation Results

In this section, an example is used as a simulation study to test the performance of the two-stage regularization method. The two-stage adaptive LASSO method studied in this paper is compared with other penalty estimates, including LASSO and SCAD. In the simulation study, we generate data from the semi-parametric model. The sample size is $n=200$ and $n=400$, the number of covariables is set as $p=10$, and $d=5$, X_i is subject to a multivariate normal distribution, Z_i is the independent uniform distribution on the interval $[0,1]$, and the error distribution is considered as a standard normal distribution. In the real model, let the coefficient of the linear part be $\beta=(3,2,1,0,0,0,0,0,0,0)$, the coefficient of the linear interaction item be $\beta_{3,2}=2$, and the non-parametric part be $g_1\left(Z_{i1}\right)=5*\cos\left(Z_{i1}\right)+3\sin\left(Z_{i1}\right)$, $g_l\left(Z_{i1}\right)=0, l=2,\ldots,d$. I represents the number of interaction items non-zero coefficient estimated to be 0 in the simulation, and C represents the number of interaction item zero coefficient estimated to be 0 in the simulation. The generalized mean square error (GMSE) to evaluate the

accuracy of the estimated parameters is defined as

$$GMSE = \left(\hat{\beta}-\beta\right)^T E\left(XX^T\right)\left(\hat{\beta}-\beta\right),$$

The square root of the mean square error (RASE) to evaluate the estimation accuracy of the nonparametric component $g_1\left(Z_{i1}\right)$ is defined as

$$RASE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{d}\left(\hat{g}_l\left(Z_{i1}\right)-g_l\left(Z_{i1}\right)\right)^2}.$$

As shown in Table 1, in the first stage of simulation, adaptive LASSO is selected for variable selection, and the interaction term is not considered in the process of variable selection. The main effect β_1,β_2,β_3 is selected, while the non-parametric part $g_1\left(Z_{i1}\right)$ is selected. In the second stage, to maintain the hierarchical structure of the model, only the interaction terms are punished, and the punishment methods include adaptive LASSO, smoothly clipped absolute deviation (SCAD), and LASSO estimation. Table 2 shows the results of the second stage simulation, comparing the performance of the three methods. It is not difficult to see that there is little difference in the performance of the three methods in selecting covariables, but adaptive LASSO is a little smaller than the GMSE and RASE of the other two methods. On the other hand, with the increase of GMSE and RASE, the three different methods decreased, and the simulation results were closer to the real model. It can be seen that adaptive LASSO is a relatively effective variable selection method.

Table 1. *Variable Selection at Stage 1.*

(*n*, *p*, *d*)			**GMSE**	**RASE**
(200,10,5)	β_1,β_2,β_3	$g_1\left(Z_{i1}\right)$	0.049	0.182
(400,10,5)	β_1,β_2,β_3	$g_1\left(Z_{i1}\right)$	0.032	0.290

Table 2. *Variable Selection at Stage 2.*

(*n*, p_1, p_2, *d*)	**Method**	β_1	β_2	β_3	$\beta_{3,2}$	**I**	**C**	**GMSE**	**RASE**
(200,3,9,1)	adaptive LASSO	1.0053	1.0152	1.9639	2.2475	0	8	0.0481	0.5901
	SCAD	1.1309	1.1147	1.7983	3.1058	0	8	0.0563	0.6301
(400,3,9,1)	adaptive LASSO	1.0037	1.0078	1.9491	2.2191	0	8	0.0392	0.5710
	SCAD	1.3530	1.6407	1.1449	2.1472	0	8	0.0404	0.5863

6. Technical Proofs

To prove the conclusion of the theorem, some regular conditions are given:

Let H_r denote the totality of the function $h(t)$ on $[0,1]$ that satisfies the following conditions, that the m -order derivative $h^m(t)$ of $h(t)$ is continuous and satisfies the v -order *Hölder* condition, and that $r = m+v$, that is, that there is a constant $M_0 \in (0,\infty)$ such that the absolute value inequality $\left|h^m(s)-h^m(t)\right| \le M_0|s-t|^v$ holds for any $s,t \in [0,1]$.;

(C-1) $E\left(g_l\left(z_l\right)\right)=0$ and $g_l\left(z_l\right) \in H_r$, $l=1,\cdots,d, r>1/2$;

(C-2) The covariate Z_l has a continuous density function $f_{z_l}\left(z_l\right)$ and the presence of constants δ_1 and δ_2 such that $f_{z_l}\left(z_l\right)$ satisfies $0<\delta_1 \le f_{z_l}\left(z_l\right) \le \delta_2 < \infty, l=1,\cdots,d$ on the interval $[0,1]$;

(C-3) For $1 \le i \le n$, the random variable is uniformly bounded and the eigenvalues of $E\left\{X_i X_i^T \middle| Z_i\right\}$ are bounded away from 0 and infinity is consistent for $1 \le i \le n$;

(C-4) Let $t_1,\cdots,t_{k_n}$ be an inner node on the interval $[0,1]$, and with $t_0=0, t_{k_n+1}=1, \xi_i = t_i - t_{i-1}$, and $\xi = \max\{\xi_i\}$, there is a constant C_0 such that

$$\frac{\xi}{\min\{\xi_i\}} \le C_0, \max\left\{\left|\xi_{i+1}-\xi_i\right|\right\} = o\left(k_n^{-1}\right)$$

holds;

(C-5) $F(x,z,h)$ and $G(x,z,u,h)$ are continuous concerning (x,z), and $F(x,z,h)<0, \forall h>0$;

(C-6) The random error ε satisfies $E\left(\phi_h'(\varepsilon)\middle|X=x,Z=z\right)=0$, $E\left(\phi_h''(\varepsilon)^2\middle|X=x,Z=z\right)$, $E\left(\phi_h'(\varepsilon)^3\middle|X=x,Z=z\right)$, and $E\left(\phi_h'''(\varepsilon)\middle|X=x,Z=z\right)$ concerning the variables x and z are continuous.

Proof of Theorem 1. Let $\delta = n^{\frac{-r}{2r+1}} + a_n$, $\beta_M = \beta_M^0 + \delta\alpha_1$, $\gamma = \gamma_0 + \delta\alpha_2$, $\alpha = \left(\alpha_1^T, \alpha_2^T\right)$. Now we work on the first part of Theorem 1. For $\forall \eta > 0$, there is a large constant C, such that

$$\mathrm{P}\left\{\sup_{\|\alpha\|=C} G(\theta_0+\delta\alpha) < G(\theta_0)\right\} \le 1-\eta . \tag{9}$$

According to the arbitrariness of η, there is a local maximum point $\hat{\theta}$ in the interior of ball $\{\theta_0+\delta\alpha : \|\alpha\| \le C\}$, that is, $\|\hat{\theta}-\theta_0\| = Op(\delta)$. Let $\Delta(\theta_0) = K^{-1}\{G(\theta_0+\delta\alpha)-G(\theta_0)\}$,

$$\Delta(\theta_0) = K^{-1}\{G(\theta_0+\delta\alpha)-G(\theta_0)\}$$

$$= K^{-1}\sum_{i=1}^{n}\phi_h\left\{Y_i - X_i^T\left(\beta_M^0+\delta\alpha_1\right) - \Psi_i^T\left(\gamma_0+\delta\alpha_2\right)\right\} - K^{-1}\sum_{i=1}^{n}\phi_h\left\{Y_i - X_i^T\beta_M^0 - \Psi_i^T\gamma_0\right\}$$

$$-\frac{n}{K}\sum_{k=1}^{p}\omega_k\left(\lambda_{1k}|\beta_k| - \lambda_{1k}|\beta_{0k}|\right) - \frac{n}{K}\sum_{l=1}^{d}\omega_l\left(\lambda_{2l}\|\gamma_l\|_{H_l} - \lambda_{2l}\|\gamma_{0l}\|_{H_l}\right).$$

Taylor expansion is performed on the above formula. The above equality becomes

$$\Delta(\theta_0) \le -\frac{\delta}{K}\sum_{i=1}^{n}\phi_h'\{\varepsilon_i + 1_d R(Z_i)\}\left(X_i^T\alpha_1 + \Psi_i^T\alpha_2\right) + \frac{1}{2}\frac{\delta^2}{K}\sum_{i=1}^{n}\phi_h''\{\varepsilon_i + 1_d R(Z_i)\}\left(X_i^T\alpha_1 + \Psi_i^T\alpha_2\right)^2$$

$$-\frac{1}{6}\frac{\delta^3}{K}\sum_{i=1}^{n}\phi_h'''(\xi_i)\left(X_i^T\alpha_1 + \Psi_i^T\alpha_2\right)^3$$

$$-\frac{n}{K}\sum_{k=1}^{p}\omega_k\left(\lambda_{1k}|\beta_k| - \lambda_{1k}|\beta_{0k}|\right) - \frac{n}{K}\sum_{l=1}^{d}\omega_l\left(\lambda_{2l}\|\gamma_l\|_{H_l} - \lambda_{2l}\|\gamma_{0l}\|_{H_l}\right)$$

$$\triangleq I_1 + I_2 + I_3 + I_4 + I_5 .$$

Where ξ_i is between $\varepsilon_i + 1_d R(Z_i)$ and $\varepsilon_i + 1_d R(Z_i) - \delta\left(X_i^T\alpha_1 + \Psi_i^T\alpha_2\right)$, 1_d represents a d-dimensional column vector with all 1 elements. $R(Z_i) = \left(R_1(Z_{i1}), \cdots, R_d(Z_{id})\right)^T$, $R_l(Z_{i1}) = g_{0l}(z_{il}) - \psi_{il}^T\gamma_{0l}$, $l = 1, \cdots, d$. According to de Boor [16], we get $\|R(Z_i)\| = Op(\delta)$. Using Taylor expansion again for I_1, we have

$$I_1 = -\frac{\delta}{K}\sum_{i=1}^{n}\left\{\phi_h'(\varepsilon_i) + \phi_h''(\varepsilon_i)1_d R(Z_i) + \frac{1}{2}\phi_h'''\left(\varepsilon_i^*\right)\left[1_d R(Z_i)\right]^2\right\}\left(X_i^T\alpha_1 + \Psi_i^T\alpha_2\right)$$

Where ε_i^* is between ε_i and $\varepsilon_i + 1_d R(Z_i)$. By the condition (C-4) and a_n, we can get

$$I_1 = Op\left(nK^{-1}\delta K^{-r}\|\alpha\|\right) = Op\left(n\delta^2 K^{-1}\|\alpha\|\right)$$

$$I_2 = E\left(F(X,Z,h)\right)Op\left(n\delta^2 K^{-1}\|\alpha\|^2\right)$$

Thus, for sufficiently large positive integers C, I_2 is able to consistently control I_1 when $\|\alpha\| = C$.

In a similar way, we have $I_3 = Op\left(n\delta^2 K^{-1}\|\alpha\|^3\right)$. If $\delta \to 0$, then $\delta\|T\| \to 0\left(\|\alpha\| = C\right)$. We can get $I_3 = op(I_2)$, that is, I_2 is able to consistently control I_3 when $\|\alpha\| = C$. On the other hand,

$$I_4 \le tnK^{-1}\delta a_n\|\alpha\| = sn^{r/(2r+1)}a_n\|\alpha\|$$

Conditioned on a_n, $I_4 = op\left(\|\alpha\|\right)$ holds for $\|\alpha\| = C$. Therefore, I_4 is consistently controlled by I_2, and I_5 is consistently controlled by I_2. Moreover, conditioned on (C-6) and $F(x,z,h) < 0$, then $\Delta(\theta_0) < 0$. Therefore, inequality (9) holds with probability $1-\eta$. So there are local maximum points $\hat{\beta}_M$ and $\hat{\gamma}$ that satisfies

$$\left\|\hat{\beta}_M-\beta_0\right\|=Op(\delta),\left\|\hat{\gamma}-\gamma_0\right\|=Op(\delta). \quad (10)$$

Now we work on the second part of Theorem 1, let $\psi_l(z)=\left(\psi_{l1}(z),\cdots,\psi_{lK_n}(z)\right)^T, l=1,\cdots,d$,

$$\left\|\hat{g}_l(z)-g_{0l}(z)\right\|^2=\int_0^1\left|\hat{g}_l(z)-g_{0l}(z)\right|^2dz$$

$$=\int_0^1\left\{\psi_l^T(z)\hat{\gamma}_l-\psi_l^T(z)\gamma_{0l}+R_l(z)\right\}^2dz$$

$$\le 2\int_0^1\left\{\psi_l^T(z)\hat{\gamma}_l-\psi_l^T(z)\gamma_{0l}\right\}^2dz+2\int_0^1R_l(z)^2dz$$

$$=2\left(\hat{\gamma}_l-\gamma_{0l}\right)^T H_l\left(\hat{\gamma}_l-\gamma_{0l}\right)dz+2\int_0^1R_l(z)^2dz.$$

Where H_l is a $K_n\times K_n$ matrix, and the (k,k')th element is $\int_0^1\psi_{1k}(z)\psi_{1k'}(z)dz$. If $\|H_l\|=O(1)$, then

$$\left(\hat{\gamma}_l-\gamma_{0l}\right)^T H_l\left(\hat{\gamma}_l-\gamma_{0l}\right)=Op\left(n^{\frac{-2r}{2r+1}}+a_n^2\right), \quad (11)$$

$$\int_0^1R_l(z)^2dz=Op\left(n^{\frac{-2r}{2r+1}}\right), \quad (12)$$

Combining (11) and (12), the second part of Theorem 1 is easily obtained.

Proof of Theorem 2. No we will prove the first part of the Theorem 2. The goal is to show that, with overwhelming probability, under Theorem 1, inequality

$$\frac{\partial G(\beta_M,\gamma)}{\partial\beta_k}<0, \qquad 0<\beta_k<v,k=t+1,\cdots,p, \quad (13)$$

$$\frac{\partial G(\beta_M,\gamma)}{\partial\beta_k}>0, \qquad -v<\beta_k<0,k=t+1,\cdots,p, \quad (14)$$

hold for each $v=Cn^{-r/(2r+1)}$, γ_l and β_k, where $\left\|\gamma_l-\gamma_{0l}\right\|=Op\left(n^{-r/(2r+1)}\right)$ and $\left\|\beta_k-\beta_{0k}\right\|=Op\left(n^{-r/(2r+1)}\right)$. Now the goal is to prove that (13) and (14) are true. It is straightforward to get

Moreover,

$$\frac{\partial G(\beta_M,\gamma)}{\partial\beta_k}=-\sum_{i=1}^n X_i\phi_h'\left(Y_i-X_i^T\beta_M-\Psi_i^T\gamma\right)-n\lambda_{1k}\omega_k\operatorname{sgn}(\beta_k)$$

$$=-\sum_{i=1}^n X_i\left\{\phi_h'\left(\varepsilon_i+1_dR(Z_i)\right)-\phi_h''\left(\varepsilon_i+1_dR(Z_i)\right)\left[\Psi_i^T(\gamma-\gamma_0)+X_i^T\left(\beta_M-\beta_M^0\right)\right]\right.$$

$$\left.+\frac{1}{2}\phi_h'''(\zeta_i)\left[\Psi_i^T(\gamma-\gamma_0)+X_i^T\left(\beta_M-\beta_M^0\right)\right]^2\right\}-n\lambda_{1k}\omega_k\operatorname{sgn}(\beta_k)$$

$$=-n\lambda_{1k}\omega_k\left[\operatorname{sgn}(\beta_k)+Op\left(\lambda_{1k}^{-1}\omega_k^{-1}n^{-r/(2r+1)}\right)\right]$$

where ζ_i is between $\varepsilon_i+1_dR(Z_i)$ and $Y_i-X_i^T\beta_M-\Psi_i^T\gamma$. By b_n, we have $\lambda_{1k}n^{-r/(2r+1)}\ge b_nn^{-r/(2r+1)}\to\infty$, $k=t+1,\cdots,p$. Therefore, we have the sign of $\partial G(\beta_M,\gamma)/\partial\beta_k$ determined by β_k, that is, (13) and (14) are true.

Based on Theorem 1, for $\|\psi_l\|=O(1)$ and $\hat{g}_l(z_l)=\psi(l)(z_l)^T\hat{\gamma}_l$, equality $\hat{g}_l(\cdot)=0, l=s+1,\cdots,d$ holds with overwhelming probability.

Proof of Theorem 3. By Theorem 1 and 2, when $n\to\infty$, $G(\gamma,\beta)$ reaches its maximum at points $\left(\hat{\beta}_a^T,0\right)^T$ and $\left(\hat{\gamma}_a^T,0\right)^T$ with probability 1. By definition of $G(\gamma,\beta)$, we have

$$-\frac{1}{n}\frac{\partial G(\beta,\gamma)}{\partial\beta}\bigg|_{\beta=\left(\hat{\beta}_a^T,0^T\right)^T}=\frac{1}{n}\sum_{i=1}^n X_i\phi_h'\left(Y_i-X_i^T\hat{\beta}_a-\Psi_a^T\hat{\gamma}_a\right)+\sum_{k=1}^p\lambda_{1k}\omega_k\operatorname{sgn}\left(\hat{\beta}_k\right)=0,$$

$$-\frac{1}{n}\frac{\partial G(\beta,\gamma)}{\partial\gamma}\bigg|_{\beta=\left(\hat{\gamma}_a^T,0^T\right)^T}=\frac{1}{n}\sum_{i=1}^n\Psi_i\phi_h'\left(Y_i-X_i^T\hat{\beta}_a-\Psi_a^T\hat{\gamma}_a\right)+\sum_{l=1}^d\lambda_{2l}\omega_l\frac{H\hat{\gamma}_a}{\left\|\hat{\gamma}_a\right\|_{H_l}}=0,$$

By Theorem 1 and a_n, we have

$$\sum_{k=1}^{p}\lambda_{1k}\omega_k \operatorname{sgn}\left(\hat{\beta}_k\right)=o_p\left(\hat{\beta}_a-\beta_{0a}\right),\ \left\|\hat{\beta}-\beta_0\right\|=Op\left(n^{-r/(2r+1)}\right)=o_p(1),\ \left\|\hat{\gamma}-\gamma_0\right\|=o_p(1),$$

Combining Theorem 1 and Theorem 2, we can get

$$\frac{1}{n}\sum_{i=1}^{n}X_i\left\{\phi_h'(\varepsilon_i)+\phi_h''(\varepsilon_i)\left[1_d^TR(Z_i)-\Psi_i^T\left(\hat{\gamma}_a-\gamma_{0a}\right)-X_i^T\left(\hat{\beta}_a-\beta_{0a}\right)\right]\right.$$

$$\left.+\frac{1}{2}\phi_h'''(\xi_i)\left[1_d^TR(Z_i)-\Psi_i^T\left(\hat{\gamma}_a-\gamma_{0a}\right)-X_i^T\left(\hat{\beta}_a-\beta_{0a}\right)\right]^2\right\}+o_p\left(\hat{\beta}_a-\beta_{0a}\right)=0, \tag{15}$$

$$\frac{1}{n}\sum_{i=1}^{n}\Psi_i\left\{\phi_h'(\varepsilon_i)+\phi_h''(\varepsilon_i)\left[1_d^TR(Z_i)-\Psi_i^T\left(\hat{\gamma}_a-\gamma_{0a}\right)-X_i^T\left(\hat{\beta}_a-\beta_{0a}\right)\right]\right.$$

$$\left.+\frac{1}{2}\phi_h'''(\xi_i)\left[1_d^TR(Z_i)-\Psi_i^T\left(\hat{\gamma}_a-\gamma_{0a}\right)-X_i^T\left(\hat{\beta}_a-\beta_{0a}\right)\right]^2\right\}+o_p\left(\hat{\gamma}_a-\gamma_{0a}\right)=0, \tag{16}$$

where ξ_i is between ε_i and $Y_i-X_i^T\hat{\beta}_a-\Psi_a^T\hat{\gamma}_a$. Let $\Lambda_n=\frac{1}{n}\sum_{i=1}^{n}\Psi_i\left[\phi_h'(\varepsilon_i)+\phi_h''(\varepsilon_i)1_d^TR(Z_i)\right]$, $\Phi_n=\frac{1}{n}\sum_{i=1}^{n}\phi_h''(\varepsilon_i)\Psi_i\Psi_i^T$, $\Gamma_n=\frac{1}{n}\sum_{i=1}^{n}\phi_h''(\varepsilon_i)\Psi_iX_i^T$. By Theorem 2, (C-3) and a_n, equation (16) becomes

$$\hat{\gamma}_a-\gamma_{0a}=\left(\Phi_n+o_p(1)\right)^{-1}\left\{-\Gamma_n\left(\hat{\beta}_a-\beta_{0a}\right)+\Lambda_n\right\}, \tag{17}$$

Moreover, we have

$$\frac{1}{n}\sum_{i=1}^{n}\phi_h''(\varepsilon_i)\left[X_i-\Gamma_n^T\left(\Phi_n^{-1}+o_p(1)\right)\Psi_i\right]^T\left(\hat{\beta}_a-\beta_{0a}\right)+o_p\left(\hat{\beta}_a-\beta_{0a}\right)$$

$$=\frac{1}{n}\sum_{i=1}^{n}X_i\left[\phi_h'(\varepsilon_i)+\phi_h''(\varepsilon_i)1_d^TR(Z_i)-\phi_h''(\varepsilon_i)\Psi_i^T\left(\Phi^{-1}+o_p(1)\right)\Lambda_n\right], \tag{18}$$

On the other hand,

$$E\left[\frac{1}{n}\sum_{i=1}^{n}\phi_h''(\varepsilon_i)\Gamma_n^T\Phi^{-1}\Psi_i\left(X_i^T-\Psi_i^T\Phi_n^{-1}\Gamma_n\right)\right]=0,$$

$$Var\left[\frac{1}{n}\sum_{i=1}^{n}\phi_h''(\varepsilon_i)\Gamma_n^T\Phi^{-1}\Psi_i\left(X_i^T-\Psi_i^T\Phi_n^{-1}\Gamma_n\right)\right]=o_p(1),$$

then

$$\left\{\frac{1}{n}\sum_{i=1}^{n}\phi_h''(\varepsilon_i)X_i^*X_i^{*T}+o_p(1)\right\}\sqrt{n}\left(\hat{\beta}_a-\beta_{0a}\right)$$

$$=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_h'(\varepsilon_i)X_i^*+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_h''(\varepsilon_i)X_i^*1_d^TR(Z_i)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_h''(\varepsilon_i)X_i^*\left(\Phi_n^{-1}+o_p(1)\right)\Lambda_n. \tag{19}$$

Let $J_1=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_h'(\varepsilon_i)X_i^*$, $J_2=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_h''(\varepsilon_i)X_i^*1_d^TR(Z_i)$, and $J_3=-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_h''(\varepsilon_i)X_i^*\left(\Phi_n^{-1}+o_p(1)\right)\Lambda_n$, then equation (19) can be

expressed as $J_1+J_2+J_3$, where $X_i^*=X_i-\Gamma_n^T\Phi_n^{-1}\Psi_i$, $J_3=0$. For J_2, we can get

$$J_2=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_h''(\varepsilon_i)\left[X_i-E(\Gamma_n)^T E(\Phi_n)^{-1}\Psi_i\right]1_d^T R(Z_i)$$

$$+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_h''(\varepsilon_i)E(\Gamma_n)^T\left(E(\Phi_n)^{-1}-\Gamma_n^T\Phi_n^{-1}\Psi_i\right)1_d^T R(Z_i)$$

Let $J_{21}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_h''(\varepsilon_i)\left[X_i-E(\Gamma_n)^T E(\Phi_n)^{-1}\Psi_i\right]1_d^T R(Z_i)$ and $J_{22}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_h''(\varepsilon_i)E(\Gamma_n)^T\left(E(\Phi_n)^{-1}-\Gamma_n^T\Phi_n^{-1}\Psi_i\right)1_d^T R(Z_i)$, then the above equation can be expressed as $J_{21}+J_{22}$.
when

$$E\left\{\phi_h''(\varepsilon_i)\left[X_i-E(\Gamma_n)^T E(\Phi_n)^{-1}\Psi_i\right]\Psi_i^T\right\}=0,$$

implies

$$\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n}\phi_h''(\varepsilon_i)\left(X_i-E(\Gamma_n)^T E(\Phi_n)^{-1}\Psi_i\right)\Psi_i^T=O_p(1).$$

Combining $\Psi_i=\left(\psi_{il}^T,\cdots,\psi_{id}^T\right)$, $\|\psi_{il}\|=O(1)$ and $\|R(Z_i)\|=O\left(n^{-r/(2r+1)}\right)$, we have $J_{21}=o_p(1)$. Apply the same technique and combine the results, we have $J_{22}=o_p(1)$. Equation (19) can be expressed as

$$n^{-1}\sum_{i=1}^{n}\phi_h''(\varepsilon_i)X_i^*X_i^{*T}\sqrt{n}\left(\hat{\beta}_a-\beta_{a0}\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_h'(\varepsilon_i)X_i^*+o_p(1),\tag{20}$$

Combining regular condition and Law of Large Number, we can get

$$\frac{1}{n}\sum_{i=1}^{n}\phi_h''(\varepsilon_i)X_i^*X_i^{*T}\xrightarrow{p}\Sigma,\tag{21}$$

On the other hand, by the Central Limit Theorem, we have

$$\sqrt{n}\sum_{i=1}^{n}\phi_h'(\varepsilon_i)X_i^*\xrightarrow{d}N(0,\Omega),\tag{22}$$

Therefore, combining (20)-(22) and Slutsky Theorem, we can get

$$\sqrt{n}\left(\hat{\beta}_a-\beta_{0a}\right)\xrightarrow{d}N\left(0,\Sigma^{-1}\Omega\Sigma^{-1}\right).\tag{23}$$

Now we work on (22). For any vector ς that is not all 0, such that

$$\varsigma^T G_1=\sum_{i=1}^{n}\frac{1}{\sqrt{n}}\varsigma^T\phi_h'(\varepsilon_i)X_i^*\ \square\ \sum_{i=1}^{n}a_i\xi_i,$$

where $a_i^2=\frac{1}{n}G(X_i,Z_i,h)\varsigma^T\overset{\vee}{Z}_{ia}\overset{\vee}{Z}_{ia}^T\varsigma$, Conditioned on X_i and Z_i, we have that ξ_i is an independent random variables with zero mean. Now verify the condition of Lindeberg Central Limit Theorem, if

$$\frac{\max_i a_i^2}{\sum_{i=1}^{n}a_i^2}\xrightarrow{p}0,\tag{24}$$

then $\sum_{i=1}^{n}a_i\xi_i\Big/\sqrt{\sum_{i=1}^{n}a_i^2}\xrightarrow{d}N(0,1)$. (22) is easily obtained. For (24), if $\left(\varsigma^T X_i^*\right)^2\le\|\varsigma\|^2\left\|X_i^*\right\|^2$, then $a_i^2\le n^{-1}G(X_i,Z_i,h)\|\varsigma\|^2\left\|X_i^*\right\|^2$ and $\left\|X_i^*\right\|=\left\|X_i^*-\Gamma_n^T\Phi_n^{-1}\Psi_i\right\|\le\left\|X_i^*\right\|+\left\|\Gamma_n^T\Phi_n^{-1}\Psi_i\right\|$. Combining (C-3), we have $\max_i\|X_i\|/\sqrt{n}=o_p(1)$ and $\max_i\left\|\Gamma_n^T\Phi_n^{-1}\Psi_i\right\|/\sqrt{n}=o_p(1)$. By Slutsky Theorem, (24) is easily obtained. Hence, it is straightforward to get Theorem 3.

Proof of Theorem 4. The proof is similar to Theorem 1. Conditioned on (C-9), let $\beta_{\hat{A}}^0$ be the true value of $\beta_{\hat{A}}$, let $\beta_{\hat{A}^{\circ 2}}^0$ be the true value of $\beta_{\hat{A}^{\circ 2}}$, and let $\tilde{\gamma}_0$ be the true value of $\tilde{\gamma}$. If $\delta=n^{-r/(2r+1)}+a_n$, then $\beta_{\hat{A}}=\beta_{\hat{A}}^0+\delta\alpha_1'$, $\beta_{\hat{A}^{\circ 2}}=\beta_{\hat{A}^{\circ 2}}^0+\delta\alpha_2'$ and $\tilde{\gamma}=\tilde{\gamma}_0+\delta\alpha_3'$. Let

$\alpha' = \left(\alpha_1'^T, \alpha_2'^T, \alpha_3'^T\right)$. Now we work on the first part of Theorem 4. For $\eta_2 > 0$, there is a larger constant C_2 satisfies

$$\mathrm{P}\left\{\sup_{\|\alpha'\|=C'} G_2\left(\beta_{\hat{A}}, \beta_{\hat{A}^{\circ 2}}, \tilde{\gamma}\right) < G_2\left(\beta_{\hat{A}}^0, \beta_{\hat{A}^{\circ 2}}^0, \tilde{\gamma}_0\right)\right\} \le 1-\eta_2. \quad (25)$$

Combining the results of Theorem 1, we can get (25). So there is a local maximum point $\left(\hat{\beta}_{\hat{A}}^{\mathrm{T}}, \hat{\beta}_{\hat{A}^{\circ 2}}^{\mathrm{T}}, \hat{\tilde{\gamma}}^{\mathrm{T}}\right)$ that satisfies $\left\|\hat{\beta}_{\hat{A}} - \beta_{\hat{A}}^0\right\| = Op(\delta)$, $\left\|\hat{\beta}_{\hat{A}^{\circ 2}} - \beta_{\hat{A}^{\circ 2}}^0\right\| = Op(\delta)$ and $\left\|\hat{\tilde{\gamma}} - \tilde{\gamma}_0\right\| = Op(\delta)$. Moreover, we have $\left\|\hat{g}_l(Z_l) - g_{0l}(Z_l)\right\| = Op(\delta)$. It is straightforward to get Theorem 4.

Proof of Theorem 5. The proof is similar to Theorem 2. By

$$\frac{\partial G_2\left(\beta_{\hat{A}}, \beta_{\hat{A}^{\circ 2}}, \gamma_{\hat{L}}\right)}{\partial \beta_{j,k}} = -\sum_{i=1}^{n} X_{\hat{A}}^{\circ 2} \phi_h\left\{Y_i - X_{\hat{A}}\beta_{\hat{A}} - X_{\hat{A}}^{\circ 2}\beta_{\hat{A}^{\circ 2}} - \Psi_{\hat{L}}\gamma_{\hat{L}}\right\} - n\lambda\omega_\alpha \operatorname{sgn}\left(\beta_{j,k}\right)$$

$$= -n\lambda\omega_\alpha\left[\operatorname{sgn}\left(\beta_{j,k}\right) + Op\left(\lambda^{-1}\omega_\alpha^{-1} n^{-r/(2r+1)}\right)\right],$$

we know that the sign of $\frac{\partial G_2\left(\beta_{\hat{A}}, \beta_{\hat{A}^{\circ 2}}, \gamma_{\hat{L}}\right)}{\partial \beta_{j,k}}$ is determined by $\beta_{j,k}$. It is straightforward to get Theorem 5.

Proof of Theorem 6. We will adopt the primal-dual witness (PDW) mothod and refer to the proof of Theorem 3 in Wainwright (2009) [13]. Meanwhile, using (W1), (W2), …, to denote the formula (1), (2), …, in Wainwright (2009). Therefore, the proof of this theorem only needs to verify the strict dual feasibility and sign consistency.

(i) Verifying strict dual feasibility

The proof of the strict dual feasibility only needs to prove that under condition (10), for any $j \in S^c$, the inequality $|Z_j| < 1$ holds, where Z_j is defined in (W10).

For each $j \in S^c$, conditioned on X_S, then $X_j^T = \Sigma_{jS}\left(\Sigma_{SS}\right)^{-1} X_S^T + E_j^T$, where X_j is a normal distribution with mean zero. E_j is independent and has the same distribution, and $E_{ij} \sim N\left(0, \left[\Sigma_{S^c|S}\right]_{jj}\right)$. In addition, under condition X_S, (W37) gives a decomposition $Z_j = A_j + B_j$, where

$$A_j = E_j^T\left\{X_S\left(X_S^T X_S\right)^{-1}\tilde{z}_S + \Pi_{X_S^\perp}\left(\frac{\omega}{\lambda_n n}\right)\right\},$$

$$B_j = \Sigma_{jS}\left(\Sigma_{SS}\right)^{-1}\tilde{z}_S.$$

Theorem 4, we just have to prove that for any $v' = Cn^{-r/(2r+1)}$ and $\left|\beta_{k,j} - \beta_{k,j}^0\right| = O_p\left(n^{-r/(2r+1)} + a_n\right)$, when $n \to \infty$, with probability 1, we have

$$\frac{\partial G_2\left(\beta_{\hat{A}}, \beta_{\hat{A}^{\circ 2}}, \gamma_{\hat{L}}\right)}{\partial \beta_{j,k}} > 0, 0 < \beta_{j,k} < v', (j,k) \in \hat{A}^{\circ 2}, \quad (26)$$

$$\frac{\partial G_2\left(\beta_{\hat{A}}, \beta_{\hat{A}^{\circ 2}}, \gamma_{\hat{L}}\right)}{\partial \beta_{j,k}} < 0, -v' < \beta_{j,k} < 0, (j,k) \in \hat{A}^{\circ 2}. \quad (27)$$

(26) and (27) imply that $G_2\left(\beta_{\hat{A}}, \beta_{\hat{A}^{\circ 2}}, \gamma_{\hat{L}}\right)$ is greatest at $\beta_{k,j} = 0, (k,j) \in \hat{A}^{\circ 2}$. So we just have to prove that (26) and (27) are true. By equality

Hence, according to condition (B1), then $\max_{j \in S^c}|B_j| \le 1-\gamma$. Conditioned on X_S and ω, for $\operatorname{var}(E_{ij}) = \left[\Sigma_{S^c|S}\right]_{jj} \le \rho_u\left(\Sigma_{S^c|S}\right)$, then A_j is Gaussian with mean zero and variance $\operatorname{var}(A_j) \le \rho_u\left(\Sigma_{S^c|S}\right) M_n$, where $M_n = \frac{1}{n}\tilde{z}_S^{\mathrm{T}}\left(\frac{X_S^T X_S}{n}\right)^{-1}\tilde{z}_S + \left\|\Pi_{X_S^\perp}\left(\frac{\omega}{\lambda_n n}\right)\right\|_2^2$.

From Lemma 1 in Hao and Zhang (2018) [6],

$$\mathrm{P}\left(\max_{j \in S^c}|Z_j| \ge 1\right) \le \mathrm{P}\left(\max_{j \in S^c}|A_j| \ge \gamma\right)$$

$$\le \mathrm{P}\left(\max_{j \in S^c}|A_j| \ge \gamma \middle| \bar{\mathrm{T}}^c(\varepsilon)\right) + C_1 \exp\left(-C_2 \min\left\{\sqrt{n}\varepsilon^2, s\right\}\right). \quad (28)$$

Given $\bar{\mathrm{T}}(\varepsilon) = \left\{M_n > \bar{M}_n(\varepsilon)\right\}$, conditioned on $\bar{T}^c(\varepsilon)$, $\operatorname{var}(A_j) \le \rho_u\left(\Sigma_{S^c|S}\right)\bar{M}_n(\varepsilon)$, then

$$\mathrm{P}\left(\max_{j \in S^c}|A_j| \ge \gamma \middle| \bar{\mathrm{T}}^c(\varepsilon)\right) \le 2(p-s)\exp\left(-\frac{\gamma^2}{2\rho_u\left(\Sigma_{S^c|S}\right)\bar{M}_n(\varepsilon)}\right),$$

From $\frac{a}{n} = o(1)$, $\frac{1}{\lambda_n^2 n} = o(1)$, we have $\bar{M}_n(\varepsilon) = o(1)$.

Therefore, it is easy to check that (10) can guarantee that $\max_{j\in S^c}\left|Z_j\right|<1$ holds with probability at least $1-c_1\exp\left(-c_2\min\left\{s,\log\left(p-s\right),n^{1/2}\right\}\right)$.

(ii) Sign consistency

To prove sign consistency, we just have to prove that (W13) holds, that is,

$$sign\left(\beta_j+\Delta_j\right)=sign\left(\beta_j\right),j\in S \tag{29}$$

where $\Delta_j=e_j^T\left(\frac{X_S^TX_S}{n}\right)^{-1}\left[\frac{1}{n}X_S^T\omega-\lambda_n sign\left(\beta_S\right)\right]$. From the definition of Δ_j, applying the triangle inequality, we have

$$\max_{j\in S}\left|\Delta_j\right|\le\lambda_n\left\|\left(\frac{1}{n}X_S^TX_S\right)^{-1}\operatorname{sgn}\left(\beta_S\right)\right\|_\infty+\left\|\left(\frac{1}{n}X_S^TX_S\right)^{-1}\frac{1}{n}X_S^T\omega\right\|_\infty$$

$$\because F_1+F_2\le F_1+\left(F_{2,1}+F_{2,2}\right),$$

where $F_{2,1}=\left\|\left(\frac{1}{n}X_S^TX_S\right)^{-1}\frac{1}{n}X_S^T\varepsilon\right\|_\infty$,

$F_{2,2}=\left\|\left(\frac{1}{n}X_S^TX_S\right)^{-1}\frac{1}{n}X_S^T\gamma_I\right\|_\infty$.

From (W41) and (W42), we have

$$\mathrm{P}\left(F_1>c_3\lambda_n\left\|\Sigma_{SS}^{-1/2}\right\|_\infty^2\right)\le 4\exp\left(-c_2\min\left\{s,\log\left(p-s\right)\right\}\right), \tag{30}$$

$$\mathrm{P}\left(F_{2,1}\ge 20\sqrt{\frac{\sigma^2s}{C_{\min}n}}\right)\le 4\exp\left(-c_1s\right). \tag{31}$$

Now we work on $F_{2,2}$, by (W60),

$$\mathrm{P}\left(\left\|\left(\frac{1}{n}X_S^TX_S\right)^{-1}\right\|_2\ge\frac{9}{C_{\min}}\right)\le 2\exp\left(-n/2\right),$$

$$\left\|\frac{1}{n}X_S^TX_S\right\|_2\le\left\|\beta_I\right\|_2\max_{j\in S;(k,l)\in T}\left\{\left|\frac{1}{n}X_j^T\left(X_k*X_l\right)\right|\right\},$$

where $\frac{1}{n}X_j^T\left(X_k*X_l\right)$ is a sample third moment. By Lemma B.5 in Hao and Zhang [12],

$$\mathrm{P}\left(\left|\frac{1}{n}X_j^T\left(X_k*X_l\right)\right|>\varepsilon\right)\le c_4\exp\left(-c_5n^{2/3}\varepsilon^2\right).$$

Therefore, we have

$$\mathrm{P}\left(\left\|\frac{1}{n}X_S^T\gamma_I\right\|_2\ge\left\|\beta_I\right\|_2\varepsilon\right)\le s^3c_4\exp\left(-c_5n^{\frac{2}{3}}\varepsilon^2\right),$$

$$\mathrm{P}\left(F_{2,2}\ge\frac{9}{C_{\min}}\left\|\beta_I\right\|_2\varepsilon\right)\le s^3c_6\exp\left(-c_7n^{\frac{2}{3}}\varepsilon^2\right).$$

Let $\varepsilon=\frac{s^{1/2}}{n^{1/3}}$,

$$\mathrm{P}\left(F_{2,2}\ge\frac{9\left\|\beta_I\right\|_2\sqrt{s}}{C_{\min}n^{\frac{1}{3}}}\right)\le c_8\exp\left(-c_9s\right), \tag{32}$$

Combining (30), (31) and (32), we have that with probability greater than $1-c_1'\exp\left(-c_2'\min\left\{s,\log\left(p-s\right)\right\}\right)$,

$$\max_{j\in S}\left|\Delta_j\right|\le c_3\lambda_n\left\|\Sigma_{SS}^{-1/2}\right\|_\infty^2+20\sqrt{\frac{\sigma^2s}{C_{\min}n}}+\frac{9\left\|\beta_I\right\|_2\sqrt{s}}{C_{\min}n^{1/3}}=g\left(\lambda_n\right).$$

Therefore, (29) holds when $\beta_{\min}>g\left(\lambda_n\right)$.

7. Conclusions and Future Studies

In this paper, we have used the two-stage regularization method for simultaneously fitting a regression model and identifying interaction terms. The proposed method automatically enforces the heredity constraint. In addition, it enjoys the "oracle" property under regularity conditions. Due to the property of the two-stage regularization method, the whole solution path is highly dependent on the result of variable selection in the first stage, so there will be some errors in the selected interaction terms. Based on this, future research should consider the regularization method under the marginal principle, so that main terms and interaction terms can be selected simultaneously.

References

[1] Liang, H., H. Wang a, and C.-L. Tsai, Profiled forward regression for ultrahigh dimensional variable screening in semiparametric partially linear models. Statistica Sinica, 2012. 22 (2): p. 531-554.

[2] Zhao, W., et al., Robust and efficient variable selection for semiparametric partially linear varying coefficient model based on modal regression. Annals of the Institute of Statistical Mathematics, 2013. 66 (1): p. 165-191.

[3] Zhang, R., W. Zhao, and J. Liu, Robust estimation and variable selection for semiparametric partially linear varying coefficient model based on modal regression. Journal of Nonparametric Statistics, 2013. 25 (2): p. 523-544.

[4] Bradley, E., et al., Least angle regression. The Annals of Statistics, 2004. 32 (2): p. 407-499.

[5] Hao, N. and H. H. Zhang, Interaction Screening for Ultrahigh-Dimensional Data. Journal of the American Statistical Association, 2014. 109 (507): p. 1285-1301.

[6] Hao, N., Y. Feng, and H. H. Zhang, Model Selection for High-Dimensional Quadratic Regression via Regularization. Journal of the American Statistical Association, 2018. 113 (522): p. 615-625.

[7] Dong, Y. and H. Jiang, A Two-Stage Regularization Method for Variable Selection and Forecasting in High-Order Interaction Model. Complexity, 2018. 2018: p. 1-12.

[8] Lv, J., H. Yang, and C. Guo, Variable selection in partially linear additive models for modal regression. Communications in Statistics - Simulation and Computation, 2017. 46 (7): p. 5646-5665.

[9] Yao, W., B. G. Lindsay, and R. Li, Local Modal Regression. J Nonparametr Stat, 2012. 24 (3): p. 647-663.

[10] Li, J., S. Ray, and B. G. Lindsay, A nonparametric statistical approach to clustering via mode identification. Journal of Machine Learning Research, 2007. 8: p. 1687-1723.

[11] Fan, J. and R. Li, Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. Journal of the American Statistical Association, 2001. 96 (456): p. 1348-1360.

[12] Hao, N. and H. H. Zhang, A Note on High-Dimensional Linear Regression With Interactions. The American Statistician, 2018. 71 (4): p. 291-297.

[13] Wainwright, M. J., Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using-Constrained Quadratic Programming (Lasso). IEEE Transactions on Information Theory, 2009. 55 (5): p. 2183-2202.

[14] Liu, X., L. Wang, and H. Liang, Estimation and Variable Selection for Semiparametric Additive Partial Linear Models STATISTICA SINICA, 2011. 21 (3): p. 1225-1248.

[15] Zou, H., The Adaptive Lasso and Its Oracle Properties. Journal of the American Statistical Association, 2012. 101 (476): p. 1418-1429.

[16] R, J. and C. de Boor, A Practical Guide to Splines. Mathematics of Computation, 1980. 34 (149).