International Journal of Statistical Distributions and Applications
2022; 8(1): 14-23
http://www.sciencepublishinggroup.com/j/ijsda
doi: 10.11648/j.ijsd.20220801.12
ISSN: 2472-3487 (Print); ISSN: 2472-3509 (Online)

Variable Selection Based on Profile Forward Selection of Partial Linear Models with Interactive Terms

Yafeng Xia*, Lin Yan

School of Sciences, Lanzhou University of Technology, Department of University, Lanzhou, China

Email address:
gsxyf@163.com (Yafeng Xia), 2633735739@qq.com (Lin Yan)
*Corresponding author

To cite this article:
Yafeng Xia, Lin Yan. Variable Selection Based on Profile Forward Selection of Partial Linear Models with Interactive Terms. *International Journal of Statistical Distributions and Applications*. Vol. 8, No. 1, 2022, pp. 14-23. doi: 10.11648/j.ijsd.20220801.12

Received: March 17, 2022; **Accepted:** April 6, 2022; **Published:** April 23, 2022

Abstract: Due to the rapid development of information technology and data acquisition technology, the model which only considers the linear main effect can not provide accurate prediction results, and the interaction between the predictor and response variables can not be ignored, so the variable selection problem of the model with interaction terms has become an important research topic in the statistical analysis today. In this paper, we discuss the problem of variable selection for a partially linear model with interaction terms using the profile forward selection method under high dimensional data. We propose the two-stage interactive selection algorithm (iPFST) under strong genetic condition and the profile forward selection algorithm (iPFSM) under marginality principle respectively. Theoretically, we use the consistency of profile estimators to prove that profile estimators have uniform convergence rate, and use the screening consistency to prove that iPFST algorithm and iPFSM algorithm can uniformly identify all important linear main effect terms and important interaction effect terms with probability 1. Seven regularization conditions for the theorem are given. Numerical simulation shows the superiority of iPFST and iPFSM in variable selection, and the two algorithms are compared, then iPFST algorithm is better than iPFSM algorithm. Finally, we give detailed technical proof.

Keywords: Profile Forward Selection, Strong Genetic Condition, Marginality Principle, Screening Consistency, Variable Selection

1. Introduction

With the increasing complexity of data, the interaction between predictors and response variables is difficult to ignore. Schwender and Ickstadt found that SNPs interactions play an important role in cancer diagnosis [1], Assary et al. argue that interactions with multiple genes are crucial in molecular analysis [2]. For the partial linear model proposed by Engle et al [3], consider the following partially linear model with interaction terms

$$Y_i = \sum_{d=1}^{p_n} \beta_d X_{id} + \sum_{k,l=1}^{p_n} \gamma_{kl} X_{ik} X_{il} + m(U) + \varepsilon_i, \qquad (1)$$

where $Y = (Y_1, Y_2, \cdots, Y_n)^T$ is the n-dimension response variable, $X_i = (X_{i1}, X_{i2}, \cdots, X_{ip_n})^T$ is the $n \times p_n$ order main effects, $\beta = (\beta_1, \beta_2, \cdots, \beta_{p_n})$ is the $p_n \times 1$ dimension unknown parameter vector, $U = (U_1, U_2, \cdots, U_L)$ is the indicator variable, $m(U)$ is the unknown smooth function, $X_{jk}^2 (i = 1, 2, \cdots, n, 1 \leq k = l \leq p_n)$ and $X_{ik} X_{il} (i = 1, 2, \cdots, n, k < l = 1, 2, \cdots, p_n)$ are the quadratic terms and the second-order interaction terms respectively, and $\gamma_{kl} (k \leq l = 1, 2, \cdots, p_n)$ is the regression parameter vector of the interaction terms. ε is a random error independent of X_i, U_i, and its conditional mean is 0, its variance is σ^2, and σ^2 is finite.

There have been many studies on the variable selection of models with interaction terms. Hao, Feng and Zhang proposed a two-stage regularization method for linear regression models with interaction terms, and introduced sign consistency to prove that the main effect and interaction terms satisfy the hierarchical structure [4]. Yao and He proposed the two-stage square root hard ridge method for high-order interaction models, and provided the prediction and estimation error bounds of the algorithm by using the overfitting hypothesis and the weak factorability hypothesis [5]. Radchenko and James proposed a "Variable selection using Adaptive Nonlinear Interaction Structures in High dimensions" method (VANISH) based on the penalty least squares criterion for nonlinear models with interaction terms, and demonstrated that VANISH can select the correct model with probability 1 as n and p approach infinity [6].

However, there is no research on variable selection for a partially linear model with interaction terms at present. Wang proved that the forward selection method can select all main effects with probability 1 under high-dimensional data [7]. Fan and Huang used the profile-least-square method to transform a partially linear model into a linear regression model, and obtained a consistent bound on the absolute difference between the profile predictors and their estimators [8]. Based on this, Liang, Wang and Tsai proposed a profile forward regression algorithm for partially linear models, which can select relevant predictors within a limited number of steps, even though the dimension of the predictors is much larger than the sample size [9]. In this paper, we use the profile forward selection method to study the variable selection problem of partially linear models with interaction terms. We propose two algorithms, that is, iPFST and iPFSM, and prove that both algorithms can identify all important interaction terms with probability 1 by screening consistency.

The rest of the organization is as follows: The second section introduces the profile forward selection method, iPFST algorithm and iPFSM algorithm. Regularization conditions and screening consistency are given in section 3. The fourth section gives the numerical simulation results. All technical proofs are given in section 5.

2. Variable Selection Method

2.1. Profile Forward Selection (PFS)

Liang, Wang and Tsai proposed a profile forward selection (PFS) method for variable selection in a partially linear model with ultra-high-dimensional data [9], which can select all relevant variables in a limited number of steps. Next, consider a partially linear model (1) with interaction terms. Write $Z = (X_1^2, X_1X_2, \cdots, X_1X_{p_n}, X_2^2, X_2X_3, \cdots, X_{p_n}^2)$ $= (Z_{11}, Z_{12}, \cdots, Z_{1p_n}, Z_{22}, Z_{23}, \cdots, Z_{p_np_n}), \gamma = (\gamma_{11}, \gamma_{12}, \cdots, \gamma_{1p_n}, \gamma_{22}, \gamma_{23}, \cdots, \gamma_{p_np_n})$, then equation (1) can be written

$$Y = X^T\beta + Z^T\gamma + m(U) + \varepsilon, \tag{2}$$

Assume $E(X_d) = 0$,$Var(X_d) = 1$,$E(Y) = 0$,$Var(Y) = 1$,$d = 1, 2, \cdots, p_n$,$\varepsilon \sim N(0, \sigma^2)$,$\sigma^2$ is finite. Let matrix $\mathbb{X} = (X_1, X_2, \cdots, X_n)^T$ and $\mathbb{Y} = (Y_1, Y_2, \cdots, Y_n)^T$. The index set of linear main effects is defined as $\mathcal{P}_1 = \{1, 2, \cdots, p_n\}$, and the index set of second-order terms is defined as $\mathcal{P}_2 = \{(k, l), 1 \le k \le l \le p_n\}$, the non-zero linear main effect terms is defined as $\mathcal{T}_1 = \{d : \beta_d \neq 0, d \in \mathcal{P}_1\}$, and the non-zero second-order interaction effect terms is defined as $\mathcal{T}_2 = \{(k, l), \gamma_{kl} \neq 0, (k, l) \in \mathcal{P}_2\}$.

According to Fan and Huang [8], the profile-least-squares method was used to transform the semi-parametric models into the least-squares models, combined with local linear regression technique [10], then Equation (2) can be written

$$Y_i - E(Y_i \mid U_i) = \sum_{d=1}^{p_n} \beta_d[X_{id} - E(X_{id} \mid U_i)] + \sum_{k,l=1}^{p_n} \gamma_{i,kl}[Z_{i,kl} - E(Z_{i,kl} \mid U_i)] + \varepsilon_i, \tag{3}$$

Let $X_i^* = X_i - E(X_i|U_i) = (X_{i1}^*, X_{i2}^*, \cdots, X_{ip_n}^*)$, $Y_i^* = Y_i - E(Y_i|U_i)$, $Z_i^* = Z_i - E(Z_i|U_i) = (Z_{i,11}^*, Z_{i,12}^*, \cdots, Z_{i,p_np_n}^*)$, then we can get

$$Y_i^* = X_i^{*T}\beta + Z_i^{*T}\gamma + \varepsilon_i. \tag{4}$$

According to the solution path of PFS algorithm proposed by Liang, Wang and Tsai [9], $W_k (1 \le k \le n)$ represents the selected variable at the end of step k. The PFS algorithm is briefly described below:

Step 1(Initialize): $W_0 = \varnothing$,

Step 2(Forward regression): In the kth$(k \ge 1)$ step, given W_{k-1}, for every $d \in \mathcal{P}_1 \setminus W_{k-1}$, construct a candidate model $\mathcal{D}_{d,k-1} = W_{k-1} \cup d$, calculate the sum of residual squares $RSS_{d,k-1}$ of each d, $a_k = argmin_{d \in \mathcal{P}_1 \setminus W_{k-1}} RSS_{d,k-1}$, update $W_k = W_{k-1} \cup a_k$, and repeat this step until it stops.

2.2. Two-stage PFS Algorithm for Partially Linear Models with Interaction Terms (iPFST)

In this section, based on the PFS algorithm, a two-stage PFS algorithm is proposed to select variables for partial linear models with interactive terms. In the first stage, PFS only selects the main effect, and all the second-order terms are not included in the model. In the second stage, the selected the main effect set is represented by $\breve{\mathcal{M}}$, and the strong genetic condition is automatically satisfied by expanding $\breve{\mathcal{M}}$ by adding all bidirectional interactions within $\breve{\mathcal{M}}$, and then performing PFS on the expanded set while forcing $\breve{\mathcal{M}}$ to remain in the final model. The iPFST algorithm is briefly described below:

Step 1: define $\mathcal{D} = \mathcal{P}_1$, implement PFS on $\mathcal{D}$, the final solution path is $\{\varphi_t^{(1)}, t = 1, 2, \cdots\}$, and the main effect chosen is $\breve{\mathcal{M}} = \{j_1, j_2, \cdots, j_{t_1}\}$.

Step 2: update $\mathcal{D} = \breve{\mathcal{M}} \cup \{(d, j) : d, j \in \breve{\mathcal{M}}\}$, implement PFS on $\mathcal{D}$ by forcing $\breve{\mathcal{M}}$.

2.3. PFS Algorithm Based on Marginality Principle for Partially Linear Models with Interaction Terms (iPFSM)

iPFST algorithm regresses the response variable of each step to the most relevant covariable, calculates the residuals as the new response variable of the next step, after selecting the most relevant covariable X_1, regresses all other covariables to X_1, and then replaces these covariables with the corresponding normalized residuals as the new covariable of the next step.

The iPFSM algorithm is proposed by combining marginality principle [11] and PFS algorithm. The main idea of iPFSM algorithm is to apply PFS algorithm to a dynamic candidate index set $\mathcal{C}$ in the model (4), and in step t, use $\mathcal{S}_t$, $\mathcal{M}_t$ and $\mathcal{C}_t$ to represent all selected items, selected linear main effect and current candidate index set respectively. Let $\mathcal{C} = \mathcal{P}_1$ represent all linear main effects, and update candidate index set $\mathcal{C}$ by adding interaction effect terms among existing linear main effects in the model, that is, by defining $\mathcal{C}_t = \mathcal{P}_1 \cup \{(k,l) : k,l \in \mathcal{M}_t\}$. iPFSM algorithm allows interaction effect terms to enter the model as soon as possible, so as to select weakly correlation linear main effects. The iPFSM algorithm is briefly described below:

Step 1: let $\mathcal{S}_0 = \varnothing, \mathcal{M}_0 = \varnothing, \mathcal{C}_0 = \mathcal{P}_1$,

Step 2: at step t, given $\mathcal{S}_{t-1}, \mathcal{M}_{t-1}, \mathcal{C}_{t-1}$, PFS algorithm is used to select a predictor from $\mathcal{S}_{t-1}, \mathcal{C}_{t-1}$ (set as X_1) into the model, add X_1 to $\mathcal{S}_{t-1}$ to get $\mathcal{S}_t$, if the newly selected predictor is an important linear main effect, update $\mathcal{M}_t, \mathcal{C}_t$, otherwise $\mathcal{M}_t = \mathcal{M}_{t-1}, \mathcal{C}_t = \mathcal{C}_{t-1}$.

Step 3: Step 2 is iterated until the reasonable upper bound D of the total number of important linear main effect terms is reached, and a solution path $\{\mathcal{S}_t, t = 1, 2, \cdots, D\}$ is obtained.

3. Asymptotic Properties of Variable Selection

For convenience, some notations are given. $\lambda_{min}(A)$ and $\lambda_{max}(A)$ is used to represent the minimum and maximum eigenvalues of arbitrary square matrix A, $\Sigma^{(1)}$ and $\Sigma^{(2)}$ are used to represent the covariance matrices of linear principal effect term and bidirectional interaction effect term respectively, and Σ is used to represent the total covariant matrix. Relabel Profile response variable Y^*, linear main effect X_d^* and interaction effect term Z_{dj}^* are P, L, respectively. In addition, let $G_0(t) = E(Y|t)$,$G_d(t) = E(X_d|t)$,$G_{dj}(t) = E(Z_{dj}|t)$, $\hat{G_d}(t)$ be the estimator of $G_d(t)$,$M_d(u)$ be the generating function of V_d. Here is the regularization conditions required for the proof:

(B1) The error ε is normal.

(B2) Suppose there are two constants $0 < \tau_{min} < \tau_{max} < \infty$, then $2\tau_{min} < \lambda_{min}(\Sigma^{(1)}) \leq \lambda_{max}(\Sigma^{(1)}) < \frac{1}{2}\tau_{max}$.

(B3) Assume $\| \beta \| \leq C_\beta$, where $C_\beta > 0$, such that $\beta_{min} \geq v_\beta n^{-\xi_{min}}$, where $\beta_{min} = \min_{d\in\mathcal{T}} | \beta_d |, v_\beta > 0$.

(B4) There are constants ξ, ξ_0 and v, such that $\log p_n \leq vn^{\xi}, d_0 \leq vn^{\xi_0}$ and $\xi + 6\xi_0 + 12\xi_{min} < 1$.

(B5) $G_d(\cdot), d = 0, 1, 2, \cdots, p_n$ is first order Lipschitz uniformly continuous.

(B6) The weight function $\omega_{nk}(\cdot)$ satisfies with probability 1:

(i) $\max_{1\leq k\leq n} \sum_{i=1}^{n} \omega_{nk}(U_i) = O(1)$;

(ii) $\max_{1\leq i,k\leq n} \omega_{nk}(U_i) = O(b_n)$, where $b_n = n^{-\frac{4}{5}}$;

(iii)$\max_{1\leq i\leq n} \sum_{k=1}^{n} \omega_{nk}(U_i) I(| U_i - U_k |> c_n) = O(c_n)$, where $c_n = n^{-\frac{2}{5}} \log n$.

(B7) $\max_{0\leq d\leq p_n} E\{\exp(u \mid V_d)\} < \infty$, where $0 \leq u \leq \frac{t_0}{\sigma_v}$, constants $t_0 > 0, \sigma_v^2 > 0$, generating function $M_d(u), d = 0, 1, 2, \cdots, p_n$ satisfies $\max_{0\leq d\leq p_n} \sup_{0\leq u\leq t_0} | \frac{d^3 \log\{M_d(u)\}}{du^3} |< \infty, \max_{0\leq d\leq p_n} E \mid V_d |^{2k} \leq \sigma_v^2, k > 2$.

First, the consistency of profile estimators is given, and it is shown that profile estimators have uniform convergence speed $n^{-\frac{1}{4}} \log^{-1} n$.

Theorem 1 Suppose that the regular conditions (B4)-(B7) hold, then we have

$$\max_{0\leq d\leq p_n} \max_{1\leq i\leq n} | \hat{G}_d(U_i) - \sum_{k=1}^{n} \omega_{nk}(U_i) G_d(U_k) |= o_p\left(n^{-\frac{1}{4}} \log^{-1} n\right). \tag{5}$$

In the process of variable selection, Fan and Lv proposed that the solution path has screening consistency [12]. Next, the screening consistency of the iPFST algorithm is established, as shown in Theorem 2 and Corollary 1, indicating that iPFST can detect all relevant predictors within the step size Q with probability 1.

Theorem 2 Suppose that the regular conditions (B1)-(B4) hold, define $K = 2\tau_{max} v C_\beta^2 \tau_{min}^{-2} v_\beta^{-4}$, the first stage of iPFST is a consistent screening of linear main effects, for $t_1 \geq Kn^{\xi_0+4\xi_{min}}$, then we have

$$P\left(\mathcal{T}_1 \subset \varphi_{t_1}^{(1)}\right) \to 1, n \to \infty. \tag{6}$$

In the first stage of iPFST, iPFST reduces the dimension of linear main effect from p to $o(n^{\frac{1}{3}})$. Next, the asymptotic condition of iPFST under strong genetic condition [13] is studied:

$$\beta_{kl} \neq 0 \Leftrightarrow \beta_k\beta_l \neq 0. \tag{7}$$

Since the number of linear main effects is $o(n^{\frac{1}{3}})$, under strong genetic conditions, the interactive selection of the second stage of iPFST no longer needs to deal with high-dimensional predictors. The consistency of the determined screening of the interactive selection of the second stage of iPFST is given below.

Corollary 1 Under the conditions of (6) and (7), $t_2 \geq Kn^{\xi_0+4\xi_{min}}$, then

$$P\left(\mathcal{T} \subset \varphi_{t_1+t_2}^{(2)}\right) \to 1, n \to \infty. \tag{8}$$

Next, the consistency of iPFSM algorithm is established under the condition that only $\Sigma^{(1)}$ is related.

(H1): Assume there are two constants $0 < \tau_{min} < \frac{1}{4} < 1 < \tau_{max} < \infty$, then we have $\sqrt{\tau_{min}} < \lambda_{min}(\Sigma^{(1)}) \leq \lambda_{max}(\Sigma^{(1)}) < \frac{1}{2}\tau_{max}$.

(H2): There are constants ξ, ξ_0 and v, then $\log p_n \leq vn^{\xi}, d_0 \leq vn^{\xi_0}$ and $\xi + 6\xi_0 + 12\xi_{min} < \frac{1}{2}$.

Finally, we give the screening consistency of iPFSM algorithm and show that iPFSM algorithm can identify all important predictors with probability 1.

Theorem 3 Under the conditions of (B1), (B3), (H1), (H2) and strong genetic condition, iPFSM is screening consistent, for $t \geq Kvn^{2\xi_0+4\xi_{min}}$, then

$$P(\mathcal{T} \subset \varphi_t) \to 1, n \to \infty. \quad (9)$$

4. Numerical Simulations

This section uses numerical simulation to verify the validity of iPFST algorithm and iPFSM algorithm for variable selection. In the whole simulation process, we use Chen and Chen [14] extended Bayesian information criterion(EBIC), data were generated by the model (1), each covariable followed the standard normal distribution $N(0,1)$, and the model error followed the normal distribution $N(0,1)$ and the normal distribution $N(0,5)$. The correlation between covariable X_i and X_j was $Cov(X_i, X_j) = 0.5^{|i-j|}$, and the indicator variable U followed the uniform distribution $[0,1]$. The sample size n was 100,200,400 respectively, let $m(u) = 5\sin(4\pi u)$, the coefficient vector of the linear main effect be $\beta = (1,3,2,-1,0,0,0,0,0,0)$, the coefficient of the interaction term be $\gamma_{12} = \gamma_{13} = \gamma_{24} = 2$, and the remaining γ be zero vectors.

The variable selection results and estimation accuracy of the proposed method were evaluated by the following indexes.

The square root of the mean square error of nonparametric components($RMSE$):

$$RMSE = \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\hat{m}_l(u_{il}) - m_{0l}(u_{il}) \right)^2 \right\}^{\frac{1}{2}}$$

the square root of generalized mean square error of parameter part($GMSE$):

$$GMSE = \left(\hat{\beta}_d - \beta_{0d}\right) E\left(ZZ^T\right) \left(\hat{\beta}_k - \beta_k\right).$$

The simulation results are shown in Table 1 and Table 2, where CL represents the average number of zero coefficients correctly identified as zero in the linear main effect part. The average number of zero coefficients in the CI interaction section that are correctly identified as zero.

Table 1. *The numerical result at* $\varepsilon \sim N(0,1)$.

n	Method	RMSE	GMSE	CL	CI
$n=100$	iPFST	0.1135	0.0947	3.98	2.97
	iPFSM	0.1088	0.0819	3.99	2.99
$n=200$	iPFST	0.0976	0.0608	4.01	2.99
	iPFSM	0.0764	0.0515	4.00	2.99
$n=400$	iPFST	0.0616	0.0573	4.01	3.00
	iPFSM	0.0549	0.0416	4.02	3.01

Table 2. *The numerical result at* $\varepsilon \sim N(0,5)$.

n	Method	RMSE	GMSE	CL	CI
$n=100$	iPFST	0.1896	0.1016	3.95	2.95
	iPFSM	0.1651	0.0756	3.97	2.95
$n=200$	iPFST	0.1746	0.0994	3.96	2.96
	iPFSM	0.1512	0.0445	3.99	2.97
$n=400$	iPFST	0.1573	0.0766	3.98	2.94
	iPFSM	0.1533	0.0374	3.99	2.96

5. Proof of Theorems

Let's first give three lemmas to prove the theorem. Since iPFST algorithm only focuses on the screening consistency of linear main effect, the proof of Theorem 2 is similar to theorem 3, and theorem 2 will not be proved in this paper.

Lemma 1 Under the (B5) and (B6) of (i) and (iii), we have,

$$\max_{0 \leq d \leq p_n} \max_{1 \leq i \leq n} \left| G_d(U_i) - \sum_{k=1}^{n} \omega_{nk}(U_i) G_d(U_k) \right| = O(c_n)$$

Proof: Given $G_d(U_i), d = 1,2,\cdots,p_n, i = 1,2,\cdots,n$, we have

$$\begin{aligned} & G_d(U_i) - \sum_{k=1}^{n} \omega_{nk}(U_i) G_d(U_k) \\ = & \sum_{k=1}^{n} \omega_{nk}(U_i) \left[G_d(U_i) - G_d(U_k) \right] \\ = & \sum_{k=1}^{n} \omega_{nk}(U_i) \left[G_d(U_i) - G_d(U_k) \right] I\left(\mid U_i - U_k \mid > c_n\right) \\ & + \sum_{k=1}^{n} \omega_{nk}(U_i) \left[G_d(U_i) - G_d(U_k) \right] I\left(\mid U_i - U_k \mid \leq c_n\right) \end{aligned}$$

by (B6) of (iii), we can get

$$\max_{0\le d\le p_n}\max_{1\le i\le n}\left|\sum_{k=1}^{n}\omega_{nk}(U_i)\left[G_d(U_i)-G_d(U_k)\right]I\left(|\ U_i-U_k\ |>c_n\right)\right|$$
$$\le C\max_{0\le d\le p_n}\max_{1\le i\le n}\left|\sum_{k=1}^{n}\omega_{nk}(U_i)I\left(|\ U_i-U_k\ |>c_n\right)\right|=O(c_n)$$

by (B5) and (B6) of (i), we can get

$$\max_{0\le d\le p_n}\max_{1\le i\le n}\left|\sum_{k=1}^{n}\omega_{nk}(U_i)\left[G_d(U_i)-G_d(U_k)\right]I\left(|\ U_i-U_k\ |\le c_n\right)\right|$$
$$\le\max_{0\le d\le p_n}\max_{1\le i\le n}\left|\sum_{k=1}^{n}\omega_{nk}(U_i)c_n\right|=O(c_n).$$

This completes the proof of Lemma 1.

Lemma 2 : Let $W_1,W_2,\cdots,W_n$ be iid with variance σ^2,$R_k=\frac{W_k-EW_k}{\sigma}$, and $M(u)=E\{\exp(uR_k)\}$ is generating function of $R_k,k=1,2,\cdots,n$. Assume there is a constant $t_0>0$, we have $E\{\exp(t|W_k|)\}<\infty$, where $0\le t\le\frac{t_0}{\sigma}$. Let the constant sequence $A_1,A_2,\cdots$ satisfy $A_n\ge\sum_{k=1}^{n}a_{nk}^2\sigma^2$, and constant sequence $a_{nk},1\le k\le n$ satisfy $A\ge\frac{\max_k|a_{nk}\sigma|}{A_n}$. If

$$M^*\doteq\sup_{0\le u\le t_0}\left|\frac{d^3\log M(u)}{dt^3}\right|<\infty,\tag{10}$$

for $0<\zeta<\frac{t_0}{A}$, we have

$$P\left\{\left|\sum_{k=1}^{n}a_{nk}\left(W_k-EW_k\right)\right|>\zeta\right\}\le\exp\left\{-\frac{\zeta^2}{2A_n}\left(1-\frac{1}{3}AM^*\zeta\right)\right\}.\tag{11}$$

Proof Let $t_\zeta=\frac{\zeta}{A_n}$, then $|\ \frac{a_{nk}\sigma\zeta}{A_n}\ |\le A\zeta\le t_0$. Taylor expansion for $\log M(u)$ at $u=0$, we can get

$$\log M(u)=\log M(0)+t\ \frac{d\log M(u)}{du}\bigg|_{u=0}+\frac{u^2}{2}\cdot\frac{d^2\log M(u)}{du^2}\bigg|_{u=0}+\frac{u^3}{6}\ \frac{d^3\log M(u)}{du^3}\bigg|_{u=u_*},$$

where $0<u_*<u$. Note that $\log M(0)=0$, $\frac{d\log M(u)}{du}\Big|_{u=0}=E(R_1)=0$, $\frac{d^2\log M(u)}{du^2}\Big|_{u=0}=1$ and $\left|\frac{d^3\log M(u)}{du^3}\Big|_{u=u_*}\right|\le M^*$, then

$$\log\left\{M(a_{nk}\sigma\zeta)\right\}\le\frac{1}{2}\left(\frac{a_{nk}\sigma\zeta}{A_n}\right)^2+\frac{1}{6}\left|\frac{a_{nk}\sigma\zeta}{A_n}\right|^3M^*\le\frac{a_{nk}^2\sigma^2\zeta^2}{2A_n^2}\left(1+\frac{1}{3}AM^*\zeta\right).$$

After a simple calculation, we can get

$$\log P\left\{\sum_{k=1}^{n}a_{nk}\left(W_k-EW_k\right)>\zeta\right\}=\log P\left\{\sum_{k=1}^{n}a_{nk}\sigma R_k\right\}$$
$$\le\log E\left\{\exp\left(t_\zeta\left(\sum_{k=1}^{n}a_{nk}\sigma R_k-\zeta\right)\right)\right\}=\zeta t_\zeta+\sum_{k=1}^{n}\log M(a_{nk}\sigma\zeta)\le-\frac{\zeta^2}{A_n}+\sum_{k=1}^{n}\frac{a_{nk}^2\sigma^2\zeta^2}{2A_n^2}\left(1+\frac{1}{3}AM^*\zeta\right)$$
$$\le--\frac{\zeta^2}{A_n}+-\frac{\zeta^2}{2A_n^2}\left(1+\frac{1}{3}AM^*\zeta\right)=-\frac{\zeta^2}{2A_n}\left(1-\frac{1}{3}AM^*\zeta\right)$$

then,

$$\log P\left\{\sum_{k=1}^{n} a_{nk}\left(W_k - EW_k\right) > \zeta\right\} \leq \exp\left\{-\frac{\zeta^2}{2A_n}\left(1-\frac{1}{3}AM^*\zeta\right)\right\}. \tag{12}$$

Similarly, we have

$$\log P\left\{\sum_{k=1}^{n} a_{nk}\left(W_k - EW_k\right) < -\zeta\right\} \leq \exp\left\{-\frac{\zeta^2}{2A_n}\left(1-\frac{1}{3}AM^*\zeta\right)\right\}. \tag{13}$$

By the equation (12) and (13), this completes the proof of Lemma 2.

Lemma 3 : Under the conditions (B2) and (B4)-(B7), let $\hat{\Sigma} = \frac{1}{n}\hat{\mathbb{X}}^T\hat{\mathbb{X}}$, $\Sigma^* = \frac{1}{n}\mathbb{X}^{*T}\mathbb{X}^*$. For any submodel $\mathcal{F}$, set $\hat{\Sigma}_{(\mathcal{F})}$ and $\Sigma^*_{(\mathcal{F})}$ is submatrix of $\hat{\Sigma}$ and Σ^* respectively. If $\tilde{m} = O\left(n^{2\xi_0+4\xi_{min}}\right)$, then

$$P\left\{\tau_{min} \leq \min_{|\mathcal{F}|\leq\tilde{m}} \lambda_{min}\{\hat{\Sigma}_{(\mathcal{F})}\} \leq \max_{|\mathcal{F}|\leq\tilde{m}} \lambda_{max}\{\hat{\Sigma}_{(\mathcal{F})}\} \leq \tau_{max}\right\} \to 1. \tag{14}$$

Proof: Let $r = (r_1, r_2, \cdots, r_{p_n})^T \in R^{p_n}$ is any p_n dimensional vector, $r_{(\mathcal{F})}$ is subvector of $\mathcal{F}$. by the condition (B2), we have

$$2\tau_{min} \leq \min_{\mathcal{F}\subset\mathcal{P}_1} \inf_{\|r_{(\mathcal{F})}\|=1} r^T_{(\mathcal{F})}\Sigma_{(\mathcal{F})}r_{(\mathcal{F})} \leq \max_{\mathcal{F}\subset\mathcal{P}_1} \sup_{\|r_{(\mathcal{F})}\|=1} r^T_{(\mathcal{F})}\Sigma_{(\mathcal{F})}r_{(\mathcal{F})} \leq \frac{1}{2}\tau_{max}.$$

So to prove that equation (14) is true, we just have to prove that

$$P\left\{\max_{|\mathcal{F}|\leq\tilde{m}} \sup_{\|r_{(\mathcal{F})}\|=1} \left|r^T_{(\mathcal{F})}\left(\hat{\Sigma}_{(\mathcal{F})} - \Sigma_{(\mathcal{F})}\right)r_{(\mathcal{F})}\right| > \tilde{\varepsilon}\right\} \to 0 \tag{15}$$

where $\tilde{\varepsilon} > 0$ is any constant. Note that for any $\mathcal{F}$, $|\mathcal{F}| \leq \tilde{m}$, we have

$$\begin{aligned}\hat{\Sigma}_{(\mathcal{F})} - \Sigma_{(\mathcal{F})} =& \frac{1}{n}\left\{\hat{\mathbb{X}}^T_{(\mathcal{F})}\hat{\mathbb{X}}_{(\mathcal{F})} - \mathbb{X}^{*T}_{(\mathcal{F})}\mathbb{X}^*_{(\mathcal{F})}\right\}\\ =& \frac{1}{n}\left\{\hat{\mathbb{X}}_{(\mathcal{F})} - \mathbb{X}^*_{(\mathcal{F})}\right\}^T\left\{\hat{\mathbb{X}}_{(\mathcal{F})} - \mathbb{X}^*_{(\mathcal{F})}\right\} + \frac{1}{n}\left\{\hat{\mathbb{X}}_{(\mathcal{F})} - \mathbb{X}^*_{(\mathcal{F})}\right\}^T\mathbb{X}^*_{(\mathcal{F})} + \frac{1}{n}\mathbb{X}^{*T}_{(\mathcal{F})}\left\{\hat{\mathbb{X}}_{(\mathcal{F})} - \mathbb{X}^*_{(\mathcal{F})}\right\}.\end{aligned}$$

According to conditions (B4)-(B7), Theorem 1 and Cauchy inequality, we can get

$$\frac{1}{n}\max_{|\mathcal{F}|\leq\tilde{m}} \sup_{\|r_{(\mathcal{F})}\|=1} \left|r^T_{(\mathcal{F})}\left[\mathbb{X}^{*T}_{(\mathcal{F})}\left(\hat{\mathbb{X}}_{(\mathcal{F})} - \mathbb{X}^*_{(\mathcal{F})}\right)\right]r_{(\mathcal{F})}\right| = o_p\left(n^{\frac{1}{4}\log^{-1}n}\right) \tag{16}$$

Similarly, we can prove that equations (17) and (18) are hold.

$$\frac{1}{n}\max_{|\mathcal{F}|\leq\tilde{m}} \sup_{\|r_{(\mathcal{F})}\|=1} \left|r^T_{(\mathcal{F})}\left[\left(\hat{\mathbb{X}}_{(\mathcal{F})} - \mathbb{X}^*_{(\mathcal{F})}\right)^T\mathbb{X}^*_{(\mathcal{F})}\right]r_{(\mathcal{F})}\right| = o_p\left(n^{\frac{1}{4}\log^{-1}n}\right) \tag{17}$$

$$\frac{1}{n}\max_{|\mathcal{F}|\leq\tilde{m}} \sup_{\|r_{(\mathcal{F})}\|=1} \left|r^T_{(\mathcal{F})}\left[\left(\hat{\mathbb{X}}_{(\mathcal{F})} - \mathbb{X}^*_{(\mathcal{F})}\right)^T\left(\hat{\mathbb{X}}_{(\mathcal{F})} - \mathbb{X}^*_{(\mathcal{F})}\right)\right]r_{(\mathcal{F})}\right| = o_p\left(n^{\frac{1}{4}\log^{-1}n}\right) \tag{18}$$

then by (16), (17) and (18), we have

$$P\left\{\max_{|\mathcal{F}|\leq\tilde{m}} \sup_{\|r_{(\mathcal{F})}\|=1} \left|r^T_{(\mathcal{F})}\left(\hat{\Sigma}_{(\mathcal{F})} - \Sigma^*_{(\mathcal{F})}\right)r_{(\mathcal{F})}\right| > \tilde{\varepsilon}\right\} \to 0 \tag{19}$$

So, next we only need to prove equation (20) is hold.

$$P\left\{\max_{|\mathcal{F}|\leq\tilde{m}}\sup_{\|r_{(\mathcal{F})}\|=1}\left|r_{(\mathcal{F})}^{T}\left(\Sigma_{(\mathcal{F})}^{*}-\Sigma_{(\mathcal{F})}\right)r_{(\mathcal{F})}\right|>\tilde{\varepsilon}\right\}\rightarrow 0 \tag{20}$$

Because for any $\mathcal{F}$, we have

$$\begin{aligned}
&\left|r_{(\mathcal{F})}^{T}\left(\Sigma_{(\mathcal{F})}^{*}-\Sigma_{(\mathcal{F})}\right)r_{(\mathcal{F})}\right| \\
&\leq \sum_{f_1,f_2\in\mathcal{F}}|r_{f_1}|\times|r_{f_2}|\times\left|\sigma_{f_1f_2}^{*}-\sigma_{f_1f_2}\right| \\
&\leq \max_{1\leq f_1,f_2\leq p_n}\left|\sigma_{f_1f_2}^{*}-\sigma_{f_1f_2}\right|\sum_{f_1,f_2\in\mathcal{F}}|r_{f_1}|\times|r_{f_2}| \\
&= \max_{1\leq f_1,f_2\leq p_n}\left|\sigma_{f_1f_2}^{*}-\sigma_{f_1f_2}\right|\left(\sum_{f\in\mathcal{F}}|r_f|\right)^2 \\
&\leq |\mathcal{F}|\max_{1\leq f_1,f_2\leq p_n}\left|\sigma_{f_1f_2}^{*}-\sigma_{f_1f_2}\right| \\
&\leq \tilde{m}\max_{1\leq f_1,f_2\leq p_n}\left|\sigma_{f_1f_2}^{*}-\sigma_{f_1f_2}\right|
\end{aligned}$$

So we can get

$$\begin{aligned}
&\sum_{|\mathcal{F}|\leq\tilde{m}}P\left\{\sup_{\|r_{(\mathcal{F})}\|=1}\left|r_{(\mathcal{F})}^{T}\left(\Sigma_{(\mathcal{F})}^{*}-\Sigma_{(\mathcal{F})}\right)r_{(\mathcal{F})}\right|>\tilde{\varepsilon}\right\} \\
&\leq \sum_{|\mathcal{F}|\leq\tilde{m}}P\left\{\max_{1\leq f_1,f_2\leq p_n}\left|\sigma_{f_1f_2}^{*}-\sigma_{f_1f_2}\right|>\frac{\tilde{\varepsilon}}{\tilde{m}}\right\} \\
&\leq \sum_{|\mathcal{F}|\leq\tilde{m}}\sum_{1\leq f_1,f_2\leq p_n}P\left\{\left|\sigma_{f_1f_2}^{*}-\sigma_{f_1f_2}\right|>\frac{\tilde{\varepsilon}}{\tilde{m}}\right\}
\end{aligned}$$

By the regularization conditions (B1)-(B2) and Lemma A.3 of Bickel and Levina [15], there are constants $C_1>0, C_2>0$, we have $P\left\{\left|\sigma_{f_1f_2}^{*}-\sigma_{f_1f_2}\right|>\tilde{\varepsilon}\right\}\leq C_1e^{-C_2n\tilde{\varepsilon}^2}$. Since the number of models satisfying $|\mathcal{F}|\leq\tilde{m}$ does not exceed $p_n^{\tilde{m}+1}$, then we can get

$$\begin{aligned}
&\sum_{|\mathcal{F}|\leq\tilde{m}}\sum_{1\leq f_1,f_2\leq p_n}P\left\{\left|\sigma_{f_1f_2}^{*}-\sigma_{f_1f_2}\right|>\frac{\tilde{\varepsilon}}{\tilde{m}}\right\} \\
&\leq p_n^{\tilde{m}+1}p_n^2C_1e^{-C_2n\tilde{\varepsilon}^2\tilde{m}^{-2}} \\
&= C_1e^{(\tilde{m}+3)\log p_n-C_2n\tilde{\varepsilon}^2\tilde{m}^{-2}}.
\end{aligned}$$

Since $\tilde{m}=O\left(n^{2\xi_0+4\xi_{min}}\right)$, then for a constant $C>0$, there is $\tilde{m}+3\leq Cn^{2\xi_0+4\xi_{min}}$, by the (B4), we can get

$$\begin{aligned}
&C_1e^{(\tilde{m}+3)\log p_n-C_2n\tilde{\varepsilon}^2\tilde{m}^{-2}} \\
&\leq C_1e^{Cvn^{\xi+2\xi_0+4\xi_{min}}}-C_2\tilde{\varepsilon}^2C^{-2}n^{1-4\xi_0-8\xi_{min}} \\
&= C_1e^{Cvn^{\xi+2\xi_0+4\xi_{min}}\left(1-C_2\tilde{\varepsilon}^2C^{-2}v^{-1}n^{1-\xi-6\xi_0-12\xi_{min}}\right)},
\end{aligned}$$

Because of $\xi+6\xi_0+12\xi_{min}<1$, when $n\rightarrow\infty$, $C_1e^{Cvn^{\xi+2\xi_0+4\xi_{min}}\left(1-C_2\tilde{\varepsilon}^2C^{-2}v^{-1}n^{1-\xi-6\xi_0-12\xi_{min}}\right)}\rightarrow 0$, so equation (20) is hold. This completes the proof of Lemma 3.

Proof of Theorem 1

It's easy to prove in the case of $j = 0$, next let's prove the case of $j \geq 0$. Since $\hat{G}_j(U_i) = \sum_{k=1}^{n} \omega_{nk}(U_i)G_j(U_k) + \sum_{i=1}^{n} \omega_{nk}(U_i)V_{jk}$, where V_{jk} is the kth column of V_j, that is $V_{jk} = X_{jk}^*$, then

$$\begin{aligned}&\max_{1\leq j\leq d}\max_{1\leq i\leq n}\left|\hat{G}_j(U_i) - G_j(U_i)\right| \\ \leq& \max_{1\leq j\leq d}\max_{1\leq i\leq n}\left|\sum_{k=1}^{n}\omega_{nk}(U_i)G_j(U_k) - G_j(U_i)\right| + \max_{1\leq j\leq d}\max_{1\leq i\leq n}\left|\omega_{nk}(U_i)V_{jk}\right| \\ \triangleq& I_1 + I_2\end{aligned}$$

By the (B5)-(B6) and Lemma 1, we have $I_1 = O(c_n)$. Let $A_n = C\sigma^2 b_n, A \geq \max_{1\leq k\leq n} \frac{\omega_{nk}(U_i)}{Cb_n}$, where $C > 0$ is a constant,then we can prove that $A_n \geq \sum_{k=1}^{n} a_{nk}^2\sigma^2, A \geq \max_k \frac{|a_{nk}\sigma|}{A_n}$, where $a_{nk} = \omega_{nk}(U_i)$. According to equation (10), (B4), (B7) and set $\zeta = n^{-\frac{1}{4}}\log^{-1} n$, we can get

$$\begin{aligned}&P\left\{\max_{1\leq j\leq d}\max_{1\leq i\leq n}\left|\sum_{i=1}^{n}\omega_{nk}(U_i)V_{jk}\right| > \zeta\right\} \\ \leq& dn\max_{1\leq j\leq d}\max_{1\leq i\leq n} P\left\{\left|\sum_{i=1}^{n}\omega_{nk}(U_i)V_{jk}\right| > \zeta\right\} \\ \leq& 2dn\exp\left\{-\frac{\zeta^2}{2A_n}(1 + AM_v\zeta)\right\} \\ \leq& 2dn\exp\left\{-\frac{\zeta^2}{4b_nC}\right\} \\ =& 2\exp\left\{-n^{\frac{3}{10}}\log^{-2}\frac{n}{C} + \log(dn)\right\} \to 0\end{aligned}$$

Therefore

$$\max_{1\leq j\leq d}\max_{1\leq i\leq n}\left|\hat{G}_j(U_i) - G_j(U_i)\right| = o_p\left(n^{-\frac{1}{4}}\log^{-1} n\right).$$

Proof of Theorem 3

Let $K = 6\tau_{max}\tau_{min}^{-2}C_\beta^2 v_\beta^{-4} v$, $L = Kn^{\xi_0+4\xi_{min}}$, $d_0 \leq vn^{\xi_0}$, note that $|\mathcal{S}_t| < d_0 L \leq Kn^{\xi_0+4\xi_{min}}$, then the eigenvalues of $\Sigma_{(\mathcal{F})}$ are controlled by lemma 3. Next we only need to prove that $n^{-1}\Omega(t) \geq 2L^{-1}(1 + o(1)), 1 \leq t \leq L$ is hold. By the (B.2) and (B.3) of Wang(2009), we can get

$$\Omega(t)^{\frac{1}{2}} \geq \frac{\sqrt{3}}{3}\max_{j\in\varphi_t^*}\left\|H_j^{(t)}Q_{(\varphi_t)}\mathbb{X}_{(\mathcal{T})}^*\beta_{(\mathcal{T})}\right\| - \max_{j\in\mathcal{T}}\left\|H_j^{(t)}Q_{(\varphi_t)}\varepsilon\right\| - \max_{j\in\mathcal{T}}\left\|H_j^{(t)}Q_{(\varphi_t)}\left(\hat{\mathbb{Y}} - \mathbb{Y}^*\right)\right\|, \tag{21}$$

where $Q_{(\varphi_t)} = I_n - H_{(\varphi_t)}$, $H_{(\varphi_t)} = \hat{\mathbb{X}}_{(\varphi_t)}\left(\hat{\mathbb{X}}_{(\varphi_t)}^T\hat{\mathbb{X}}_{(\varphi_t)}\right)^{-1}\hat{\mathbb{X}}_{(\varphi_t)}^T$, $\varphi_t^* = \frac{\mathcal{T}}{\varphi_t}$, $\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)^T$. Since $\max_i\left|\hat{Y}_i - Y_i^*\right| = o_p\left(n^{-\frac{1}{4}}\log^{-1} n\right)$, then $\left\|\hat{\mathbb{Y}} - \mathbb{Y}^*\right\|^2 = n \cdot o_p\left(n^{-\frac{1}{2}}\right)$. According to Theorem 1, we can get

$$\max_{i,j}\left\|\hat{X}_{ij} - X_{ij}^*\right\| = o_p\left(n^{-\frac{1}{4}}\log^{-1} n\right). \tag{22}$$

Since

$$\begin{aligned}&\max_{j\in\varphi_t^*}\left\|H_j^{(t)}Q_{(\varphi_t)}\left\{\mathbb{X}_{(\mathcal{T})}^*\beta_{(\mathcal{T})}\right\}\right\|^2 = \max_{j\in\varphi_t^*}\left\|H_j^{(t)}Q_{(\varphi_t)}\left\{\mathbb{X}_{(\varphi_t^*)}^*\beta_{(\varphi_t^*)}\right\}\right\|^2 \\ \geq& \left\{\max_{j\in\mathcal{T}}\left\|\hat{\mathbb{X}}_j\right\|^2\right\}^{-1}\left[\max_{j\in\varphi_t^*}\left|\hat{\mathbb{X}}_j^T Q_{(\varphi_t)}\left\{\mathbb{X}_{(\varphi_t^*)}^*\beta_{(\varphi_t^*)}\right\}\right|^2\right],\end{aligned}$$

by the Cauchy inequality and (B3), we have

$$
\begin{aligned}
&\left\|Q_{(\varphi_t)}\left\{\mathbb{X}^*_{(\varphi_t^*)}\beta_{(\varphi_t^*)}\right\}\right\|^2\\
=&\sum_{j\in\varphi_t^*}\beta_j\left(\hat{\mathbb{X}}_j^T Q_{(\varphi_t)}\left\{\mathbb{X}^*_{(\varphi_t^*)}\beta_{(\varphi_t^*)}\right\}\right)\\
\le&\left(\sum_{j\in\varphi_t^*}\beta_j^2\right)^{\frac{1}{2}}\left(\sum_{j\in\varphi_t^*}\left(\hat{\mathbb{X}}_j^T Q_{(\varphi_t)}\left\{\mathbb{X}^*_{(\varphi_t^*)}\beta_{(\varphi_t^*)}\right\}\right)^2\right)^{\frac{1}{2}}\\
\le& C_\beta\cdot|\mathcal{T}|^{\frac{1}{2}}\cdot\max_{j\in\varphi_t^*}\left|\hat{\mathbb{X}}_j^T Q_{(\varphi_t)}\left\{\mathbb{X}^*_{(\varphi_t^*)}\beta_{(\varphi_t^*)}\right\}\right|,
\end{aligned}
$$

By the Bernstein's inequality, we can get

$$
\begin{aligned}
&\left\{\max_{j\in\mathcal{T}}\left\|\hat{\mathbb{X}}_j\right\|^2\right\}^{-1}\left[\left\|Q_{(\varphi_t)}\left\{\mathbb{X}^*_{(\varphi_t^*)}\beta_{(\varphi_t^*)}\right\}\right\|^2\cdot C_\beta^{-1}\cdot|\mathcal{T}|^{-\frac{1}{2}}\right]^2\\
&\ge n\tau_{max}^{-1}\tau_{min}^2\beta_{min}^4|\mathcal{T}|^{-1}C_\beta^{-2}\\
&\ge \tau_{max}^{-1}\tau_{min}^2 C_\beta^{-2}v_\beta^4 v^{-1}n^{1-\xi_0-4\xi_{min}},
\end{aligned}
$$

and $\left\|\hat{\mathbb{X}}_j^{(t)}\right\|^2\ge n\tau_{min}$, so we can get

$$
\left\|H_j^{(t)}Q_{(\varphi_t)}\varepsilon\right\|^2\le\tau_{min}^{-1}n^{-1}\max_{j\in\mathcal{T}}\max_{|\mu|\le m^*}\left(\mathbb{X}_j^{*T}Q_{(\mu)}\varepsilon\right)^2, \tag{23}
$$

where $m^*=Kn^{\xi_0+4\xi_{min}}$,$\mathbb{X}_j^{*T}Q_{(\mu)}\varepsilon$ is a normal random variable with mean 0 and variance $\left\|Q_{(\mu)}\mathbb{X}_j^*\right\|^2\le\left\|\mathbb{X}_j^*\right\|^2$. So,we have

$$
\tau_{min}^{-1}n^{-1}\max_{j\in\mathcal{T}}\max_{|\mu|\le m^*}\left(\mathbb{X}_j^{*T}Q_{(\mu)}\varepsilon\right)^2\le\tau_{min}^{-1}n^{-1}\max_{j\in\mathcal{T}}\left|\mathbb{X}_j^*\right|^2\max_{j\in\mathcal{T}}\max_{|\mu|\le m^*}\chi_1^2, \tag{24}
$$

where χ_1^2 represents a Chi-squared random variable with 1 degree of freedom. Then, $\left\|H_j^{(t)}Q_{(\varphi_t)}\varepsilon\right\|^2\le 3Kvn^{\xi+\xi_0+4\xi_{min}}$ is hold with probability 1, therefore,

$$
\begin{aligned}
&n^{-1}\Omega(t)\\
\ge&3^{-1}\tau_{max}^{-1}\tau_{min}^2C_\beta^{-2}v_\beta^4v^{-1}n^{-\xi_0-4\xi_{min}}\cdot\{1+o_p(1)\}\cdot\left\{1-9Kv^2\tau_{max}\tau_{min}^{-2}C_\beta^2v_\beta^{-4}n^{\xi+2\xi_0+8\xi_{min}-1}\right\}\\
=&2L^{-1}\{1+o_p(1)\}.
\end{aligned}
$$

References

[1] Schwender, H., Ickstadt, K. (2008). Identification of SNP interactions using logic regression. Biostatistics, 9 (1): 187-198.

[2] Assary, E., Vincent J. P., and Keers, P., et al. (2018). Gene-environment interaction and psychiatric disorders: review and future directions. Seminars in Cell & Development Biology, 77: 133-143.

[3] Engle, R., Granger, C., and Rice, J., et al. (1986). Semiparametric estimates of the relation between weather and electricity sales. Journal of the American Statistical Association, 81 (394): 310-320.

[4] Hao, N., Feng, Y., and Zhang. H. (2018). Model selection for high dimensional quadratic regression via regularization. Journal of the American Statistical Association, 113 (522): 615-625.

[5] Yao, D., He, J. (2018). A two-stage regularization method for variable selection and forecasing in high-order interaction model. Complexity, 4: 1-12.

[6] Radchenko, P., James, G. M. (2010). Variable selection using adaptive nonlinear interaction structures in high dimensions. Journal of the American Statistical Association, 105 (492): 1541-1553.

[7] Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. Journal of the American Statistical Association, 104 (488): 1512-1524.

[8] Fan, J., Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. Bernoulli, 11 (6): 1031-1057.

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

[10] Fan, J., Gijbels, I. (1996). Local polynomial modelling and its applications. London: Monographs on Statistics and Applied Probability.

[11] McCullagh, P. (2002). "What is Statistical Model?". The Annals of Statistics, 30: 1225-1267.

[12] Fan, J., Lv, J. (2008). Sure indenpence screening for ultra-high dimensional feature space. Journal of the Royal Statistical Society, 70 (5): 849-911.

[13] Chipman, H., Hamada, M., and Wu, C. F. J. (1997). A bayesian variable selection approach for analyzing designed experiments with complex aliasing. Technometrics, 39: 372-381.

[14] Chen, J., Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. Biometrika, 95 (3): 759-771.

[15] Bickel, P., Levina, E. (2008). Regularized estimation of large covariance matrices. The Annals of Statistics, 36 (1): 199-227.