Ridge Regression Revisited: Debiasing, Thresholding and Bootstrap

Yunyi Zhang and Dimitris N. Politis

September 18, 2020

Abstract

Under high dimensional setting, the facts that classical ridge regression method cannot perform model selection on its own and it introduces large bias make this method an unsatisfactory tool for analyzing high dimensional linear models. In this paper, we propose a debiased and threshold ridge regression method which solves the aforementioned drawbacks. Besides, focus on performing statistical inference and prediction on linear combinations of parameters, we derive a normal approximation theorem for the estimator and introduce two bootstrap algorithms which provide simultaneous confidence region and prediction region for linear combinations of parameters. In statistical inference part, apart from the dimension of parameters, we allow the number of linear combinations to increase as sample size increases. From numerical experiments, we can see that the proposed regression method is robust with fluctuations in the ridge parameter and reduces estimation error compared to classical and threshold ridge regression methods. Apart from theoretical interests, the proposed methods can be applied to disciplines such as econometrics, finance, medical research and etc.

1 Introduction

Statistical inference on the linear model \( y = X\beta + \epsilon \) with \( \beta \) being \( p \) dimensional unknown parameters and \( \epsilon \) being residuals with mean 0 and marginal variance \( \sigma^2 \) is one of the fundamental topics in statistics. The dimension \( p \) is assumed to be fixed in the classical setting, but in the modern era data always have complex structures, correspondingly the dimension of data can be as large as, or even larger than the number of samples \( n \). The large dimension brings extra challenges to statisticians. Lasso and its modifications are among the most popular methods which solve this problem. For example, Meinshausen and Bühlmann \cite{1} applied Lasso for model selection and Meinshausen and Yu \cite{2} derived the sign consistency and \( L_2 \) consistency of Lasso for high dimensional data. Huang, Ma and Zhang \cite{3} applied adaptive Lasso for high dimensional regression problem. Fan and Li \cite{4} introduced SCAD penalty, Kim, Choi and Oh \cite{5} proved the model selection consistency of regression method based on SCAD penalty and Wang, Song and Tian \cite{6} proved the consistency of this method. In order to make statistical inference, Javanmard and Montanari \cite{7} proposed desparsifying Lasso. Focus on testing \( \beta_i = \beta_{0,i}, i \in G \subset \{1,2,\ldots,p\} \) with \( \beta_{0,i} \) being given, Zhang and Cheng \cite{8} applied desparsifying Lasso and multiplier bootstrap to create the simultaneous confidence region for parameters \( \beta \) and Dezeure, Bühlmann and Zhang \cite{9} solved the same test problem under heteroskedasticity. Chen and Zhou \cite{10} performed statistical inference based on Huber regression and the multiplier bootstrap. We also refer a two step method introduced by Liu and Yu \cite{11}. In this paper, the authors applied Lasso for model selection and then performed ordinary least square regression or ridge regression for estimating the parameters \( \beta \).

From the numerical experiments made by Zou and Hastie \cite{12}, we can see that the ridge regression also has good performance compared to Lasso. In addition, expression of the ridge regression estimator is simple, which means that we
can easily perform statistical inference on it and we do not need to use optimization tools to find the estimator. However, there are relatively few researches on ridge regression under high dimensional setting. Shao and Deng [13] proposed a threshold ridge regression method and proved its model selection consistency as well as $L_2$ consistency. Lopes [14] introduced a residual-based bootstrap algorithm to provide the confidence interval of linear combinations of parameters. Bühmann [15] introduced how to correct the bias in ridge regression through using Lasso. Based on our understanding, there are three drawbacks which stop classical ridge regression from being used to analyze high dimensional linear models:

1. The ridge regression cannot execute model selection on its own. A well-known fact of Lasso (see Tibshirani [16]) is that it produces parameters $\beta$ with lots of 0. This property is useful especially when the underlying model is sparse. Since ridge regression does not have this property, we need to apply extra measures, like threshold used by Shao and Deng [13], to facilitate model selection.

2. Under high dimensional setting, bias in the classical ridge regression brings critical troubles. This phenomenon can be seen in figure 1. Suppose we want to estimate the linear combination of parameters $a^T \beta$ with $a$ a constant vector and $p < n$, after performing tight singular value decomposition $X = P \Lambda Q^T$ (theorem 7.3.2 in [17]), classical ridge regression (with ridge parameter $\rho_n$) says

$$a^T(\beta - \beta) = -\rho_n a^T(X^TX + \rho_n I_p)^{-1} \beta + a^T(X^TX + \rho_n I_p)^{-1}X^T \epsilon$$

In the worst situation, according to Cauchy’s inequality, suppose $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$, we have

$$|\rho_n a^T(X^TX + \rho_n I_p)^{-1} \beta| \leq \frac{\rho_n}{\lambda_p} \|a\|_2 \times \|\beta\|_2$$

As a comparison, the standard deviation of $a^T Q (\Lambda^2 + \rho_n I_p)^{-1} \Lambda P^T \epsilon$ is less than $\frac{\rho_n}{\lambda_p} \|a\|_2$. If $p$ is fixed, then $\|\beta\|_2$ has order $O(1)$, correspondingly by choosing $\rho_n$ appropriately, the bias is smaller than the standard deviation. However, if $p$ is large and we do not assume that $\|\beta\|_2$ has order $O(1)$ (like Lopes [14] does), the introduced bias can be even larger than the standard deviation, which makes ridge regression meaningless. Worse still, it can be very hard to perform statistical inference if the undetectable bias has larger order than the stochastic error.

In order to solve this problem, one way is to estimate the bias by plugging in the ridge regression estimator of $\beta$ and use the estimator $a^T(\hat{\beta} + \rho_n a^T(X^TX + \rho_n I_p)^{-1} \hat{\beta})$ instead of $a^T \hat{\beta}$. According to (5), by choosing proper ridge parameter $\rho_n$, this modification does not enlarge the stochastic error significantly but helps reduce the order of bias.

3. The third problem comes when the dimension $p$ is greater than the sample size $n$. According to Shao and Deng [13] and Bühmann [15], when the dimension $p$ is greater than the sample size $n$, the underlying parameters $\beta$ are not identifiable.

In this situation, Lasso tends to select the parameters with lots of 0 but ridge regression prefers the projection of parameters on the space spanned by rows of the design matrix $X$, which seldom has lots of 0. Statisticians hope to find satisfactory parameters which are sparse, so they would prefer Lasso (or its modifications) in performing statistical inference or testing. However, as we can see in figure 1 if the underlying parameters $\beta$ is not sparse, then it is possible for the ridge regression to outperform the Lasso.

In this paper, our propose is to solve the first and the second drawbacks and to provide two statistical inference algorithms which generate simultaneous confidence regions and prediction regions for the modified ridge regression method. As a
generalization of Lopes \cite{14}, we decide to perform statistical inference on $\gamma = M \beta$ with $M$ being a $p_1 \times p$ known matrix and $\beta$ being the underlying parameters. Performing statistical inference on $\gamma$ is a common topic in econometrics (we refer Vogelsang \cite{18}, Ye and Sun \cite{19} and Gonçalves and Vogelsang \cite{20} as a background) but receives few attentions in high dimensional statistics. Besides, analyzing $\gamma$ directly leads to prediction (like generating prediction intervals for future observations), which is also a hot area for the time being. We refer Politis \cite{21} as an introduction of prediction and Stine \cite{22} as an example of how to perform prediction in a linear model.

Compared to the current linear regression methods and the statistical inference algorithms, the proposed methods have several advantages. The first one is that the modified ridge regression method has an explicit expression, so it is not difficult to perform statistical inference on the estimator and we do not need to use optimization algorithms to get the parameters $\beta$. From chapter \ref{section7} we can see that this method is robust to the fluctuation of the ridge parameter, which brings less pressure on model selection. The third advantage is that the associated bootstrap inference and prediction algorithms allow the design matrix $X$ to have relatively small singular values, which is frequently seen when the dimension $p$ is close to the sample size $n$. In addition, the proposed bootstrap inference algorithm allows the number of linear combinations $p_1$ to grow as the sample size $n$ increases.

We introduce the frequently used notations and assumptions in chapter \ref{section2} and several useful lemmas in chapter \ref{section3}. In chapter \ref{section4} we derive consistency of the proposed ridge regression method and prove that Gaussian approximation can be applied to the estimator of $\gamma = M \beta$, this lays the theoretical foundation for algorithm \ref{algorithm1}. In chapter \ref{section5} we introduce the bootstrap algorithm \ref{algorithm1} which provides simultaneous confidence region for $\gamma$. In chapter \ref{section6} we discuss prediction and provide bootstrap algorithm \ref{algorithm2} to create the simultaneous prediction region of $\gamma$. We demonstrate the finite sample performance in chapter \ref{section7} and make conclusions in chapter \ref{section8}. We postpone the proofs of mentioned theorems to the appendix.

## 2 preliminary

In this section we introduce the frequently used notations and assumptions.

Suppose $n \times p$ design matrix $X$ and random variables $y$ satisfy $y = X \beta + \epsilon$, we are interested in estimating the linear combinations of parameters $\gamma = M \beta$ with $M = (m_{ij})_{i=1,2,...,p_1 \times j=1,2,...,p}$ as a $p_1 \times p$ known matrix. By using thin singular value decomposition (theorem 7.3.2 in \cite{17}), we have $X = P \Lambda Q^T$ with $P, Q$ respectively being $n \times r$, $p \times r$ orthonormal matrix satisfying $P^T P = Q^T Q = I_r$ and $\Lambda = \text{diag}(\lambda_1, ..., \lambda_r)$ such that $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r > 0$ being non-zero singular values. Here $r$ is the rank of the design matrix $X$. We denote $Q_\perp$ as the $p \times (p - r)$ orthonormal complement of $Q$, so that we have $Q_\perp^T Q_\perp = I_{p-r}$, $Q_\perp^T Q = Q^T Q = 0$ and $QQ^T + Q_\perp Q_\perp^T = I_p$. We define $\zeta = Q^T \beta$ and $\theta = Q \zeta = QQ^T \beta$, which is the projection of parameters $\beta$ on the space spanned by rows of the design matrix $X$ (thus we have $X \beta = X \theta$ and $\theta^T \theta = \zeta^T Q_\perp Q \zeta = \zeta^T \zeta$). According to Shao and Deng \cite{13}, ridge regression estimates $\theta$ rather than $\beta$. We also define the unobservable part $\theta_\perp = Q_\perp Q^T \beta$ and correspondingly the equation $\beta = \theta + \theta_\perp$ happens. For a given positive number $b$, we define set $N_b = \{i \mid |\theta_i| > b\}$. Suppose $b$ is chosen, we define

$$c_{ik} = \sum_{j \in N_b} m_{ij} q_{jk}, \forall i = 1, 2, ..., p_1, k = 1, 2, ..., r, \text{ and } M = \{i \mid \sum_{k=1}^r c_{ik}^2 > 0\}$$

(3)

For a chosen ridge parameter $\rho_n > 0$ and a threshold level $b_n > 0$, we define the classical ridge regression statistics $\tilde{\theta}^*$ and...
the de-biased statistics $\tilde{\theta}$ as

$$\tilde{\theta}^* = (X^TX + \rho_n I_p)^{-1}X^Ty, \ \tilde{\theta} = \tilde{\theta}^* + \rho_n \times Q(\Lambda^2 + \rho_n I_r)^{-1}Q^T\tilde{\theta}^*$$  

(4)

From (4), we know that

$$\tilde{\theta} - \theta = -\rho_n^2 Q(\Lambda^2 + \rho_n I_r)^{-2}\zeta + Q \left((\Lambda^2 + \rho_n I_r)^{-1}\Lambda + \rho_n(\Lambda^2 + \rho_n I_r)^{-2}\Lambda\right)P^T\epsilon$$  

(5)

Similar as $N_{b_n}$, we define set $\tilde{N}_{b_n}$, statistics $\tilde{\theta} = (\tilde{\theta}_1, ..., \tilde{\theta}_p)^T$ and $\tilde{\gamma}$ as

$$\tilde{N}_{b_n} = \{i ||\tilde{\theta}_i| > b_n\}, \ \tilde{\theta}_i = \tilde{\theta}_i \times 1_{i \in \tilde{N}_{b_n}}, \ \tilde{\gamma} = M\tilde{\theta}$$  

(6)

We will need to estimate the marginal variance of residuals $\sigma^2 = \mathbb{E}[\epsilon_1]^2$, and the estimator we use is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij}\hat{\theta}_j)^2$$  

(7)

We define $\tau_i, \ \tilde{\tau}_i, \ i = 1, 2, ..., p$ and $H(x)$ for $x \geq 0$ as

$$\tau_i = \sqrt{\sum_{k=1}^{r} c_{ik} \left(\frac{\lambda_k}{\lambda_i + \rho_n} + \frac{\rho_n\lambda_k}{(\lambda_i^2 + \rho_n)^2}\right)^2 + \frac{1}{n}}, \ \tilde{\tau}_i = \sqrt{\sum_{k=1}^{r} \left(\sum_{j \in \tilde{N}_{b_n}} m_{ij}\eta_{jk}\right)^2 \left(\frac{\lambda_k}{\lambda_i^2 + \rho_n} + \frac{\rho_n\lambda_k}{(\lambda_i^2 + \rho_n)^2}\right)^2 + \frac{1}{n}}$$  

(8)

$$H(x) = \text{Prob} \left(\max_{i \in M} \frac{1}{\tau_i} \sum_{k=1}^{r} c_{ik} \left(\frac{\lambda_k}{\lambda_i^2 + \rho_n} + \frac{\rho_n\lambda_k}{(\lambda_i^2 + \rho_n)^2}\right) |\xi_k| \leq x\right)$$

Here $\xi_k, \ k = 1, 2, ..., r$ are independent normal random variables with mean 0 and variance $\sigma^2 = \mathbb{E}[\epsilon_1]^2$. If assumption 7) below happens, since matrix $\left(\frac{1}{\tau_i} c_{ik} \left(\frac{\lambda_k}{\lambda_i^2 + \rho_n} + \frac{\rho_n\lambda_k}{(\lambda_i^2 + \rho_n)^2}\right)\right)_{i \in M, j=1, 2, ..., r}$ has rank $|M|$ and we may apply lemma 2) to $H$. As we will show in theorem 2) $H(x)$ will be used to approximate the distribution of $\max_{i=1, 2, ..., p} \frac{|\tilde{\tau}_i - C_0|}{\tau_i}$. Similar as chapter 1.5.1 in [23], for a sequence $a_n \in \mathbb{R}$ and $b_n > 0$, we say $a_n = O(b_n)$ if $\exists C > 0$ such that $a_n \leq Cb_n$ for $n = 1, 2, ...$ and $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$. For random variables $X_n$ and $Y_n$, we say that $X_n = O_p(Y_n)$ if for any given $\delta > 0$, there exists $C_\delta > 0$ such that $\sup_{n=1, 2, ...} \text{Prob} \left(\left|X_n \right| \geq C_\delta |Y_n|\right) < \delta$ and $X_n = o_p(Y_n)$ if $\frac{X_n}{Y_n} \to_p 0$. For a finite set $A$, we use $|A|$ to denote the number of elements in $A$. In the following of this paper, we will use $\text{Prob}^\star(\cdot)$ to represent the conditional probability $\text{Prob}(|X, y)$ and $\mathbb{E}^\star$ to denote the conditional expectation $\mathbb{E}(|X, y)$. For we assume fixed design, $X$ is considered as a fixed numerical matrix and therefore $\text{Prob}^\star(\cdot) = \text{Prob}(|\epsilon)$ and $\mathbb{E}^\star = \mathbb{E}|\epsilon$. We adopt the definition of quantile in Politis et.al. [24]: Suppose $H(x)$ is a cumulative distribution function and $0 < \alpha < 1$, then $1 - \alpha$ quantile of $H$ is

$$c_{1-\alpha} = \inf \{x \in \mathbb{R} | H(x) \geq 1 - \alpha\}$$  

(10)
For a set of numbers $E_i, \ i = 1, 2, \ldots, B$ such that $E_1 \leq E_2 \leq \ldots \leq E_B$, we define its $1 - \alpha$ sample quantile $C_{1-\alpha}$ as

$$C_{1-\alpha} = E_{b_0} \text{ such that } b_0 = \min\{i| \sum_{j=1}^{B} 1_{E_j \leq E_i} \geq B \times (1 - \alpha)\} \quad (11)$$

Other symbols will be defined before being used.

Remark 1

We would like to explain why we decide to choose $\tilde{\tau}_i, i = 1, 2, \ldots, p_1$, which equal $\tau_i$ if $\hat{N}_{b_0} = N_{b_0}$, as normalizing parameters here. If we can make sure that all of the singular values of the design matrix $X$ have order $O(\sqrt{n})$, then normalizing parameters can be simply chosen as $1/\sqrt{n}$. However, from Table 1, we can see that some of the singular values of the design matrix $X$ can be significantly smaller than $\sqrt{n}$ if the dimension $p$ is as large as the sample size $n$. If we still adopt $1/\sqrt{n}$ as normalizing parameter, then the variance of random variable $\sqrt{n}(\tilde{\tau}_i - \tau_i)$ may tend to infinity if as sample size $n$ increases, which is not acceptable.

Another choice is to use the estimated marginal standard deviations as normalizing parameters. According to the aforementioned definition, if the threshold selects correct parameters, we have

$$\tilde{\tau}_i - \tau_i = \sum_{j \in N_{b_0}} m_{ij}(\hat{\theta}_j - \theta_j) - \sum_{j \notin N_{b_0}} m_{ij}\theta_j - \sum_{j=1}^{p} m_{ij}\theta_\perp,j \quad (12)$$

If assumption 1) and 5) happen and $i \in M$, for sufficiently large $n$, the marginal standard deviation of the first term is

$$\sigma \sqrt{\sum_{k=1}^{r} E_{\perp k} \left(\frac{\lambda_k}{\lambda_k + \rho_n} \right)^2} \geq \frac{\sigma \sqrt{C_M}}{2C_{1-\alpha} n^{1/2}} \quad (13)$$

while the second and the third term have order $o(1/\sqrt{n})$, which are significantly smaller than the standard deviation. However, if $i \notin M$, the standard deviation is 0 but the second and the third term are not guaranteed to be 0. If we want to provide the simultaneous confidence region of $\gamma_i, i = 1, 2, \ldots, p_1$ and unfortunately some of $i$ are not in $M$, then the bias introduced by the second and the third term will be expanded to infinity, which is not acceptable as well.

The advantages of using $\tilde{\tau}_i$ comes in two aspects. If $i \in M$, from (13) we know that the random variable $\frac{\tilde{\tau}_i - \tau_i}{\tau_i}$ does not degenerate. On the other hand, if $i \notin M$, according to assumption 5), since the normalizing parameter is larger than $1/\sqrt{n}$, the bias introduced by the second and the third term remains small after dividing $\tilde{\tau}_i$, which will not bring extra burdens for us to observe the behaviors of $\tilde{\tau}_i - \tau_i, i \in M$.

Now we introduce the main assumptions of this paper.

Assumptions

1). There exists constants $c_\lambda, C_\lambda > 0$ and $0 < \eta \leq 1/2$ such that singular values of design matrix $X$ satisfy

$$C_\lambda n^{1/2} \geq \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r \geq c_\lambda n^{\eta} \quad (14)$$

In addition, we assume that the Euclidean norm of $\theta$, $||\theta||_2 = \sqrt{\sum_{i=1}^{p} \theta_i^2}$ satisfies $||\theta||_2 = O(n^{\alpha_\theta})$ with $\alpha_\theta$ being a positive number such that $\alpha_\theta < 3\eta$.

2). Ridge parameter $\rho_n$ satisfies $\rho_n = O(n^{2\eta - \delta})$ with positive number $\frac{\eta + \delta}{2} < \delta < 2\eta$.
3. Residuals $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ are independently and identically distributed with $E\epsilon_1 = 0$ and there exists $m > 4$ such that $E|\epsilon_1|^m < \infty$.

4. Dimension of parameters $p$ satisfies $p = O(n^{\alpha_p})$ with $m \eta > \alpha_p > 0$. Threshold $b_n$ is chosen as $b_n = C_b \times n^{-\nu_b}$ with constants $C_b, \nu_b > 0$ such that $\nu_b + \frac{\alpha}{m} - \eta < 0$. Parameters $\theta$ satisfy $|\theta_i| \leq c_b \times b_n$ or $|\theta_i| \geq \frac{b_n}{c_b}$ for a constant $0 < c_b < 1$.

5. $\mathcal{M}$ is not empty and $|\mathcal{M}| = O(n^{\alpha_{\mathcal{M}}})$ with $\alpha_{\mathcal{M}} < m \eta$, in addition there exists constants $c_M, C_M$ such that $0 < c_M < \sum_{k=1}^r c_k^2 \leq C_M$ for all $i \in \mathcal{M}$, here $c_{ik}, i = 1, 2, \ldots, p, k = 1, 2, \ldots, r$ are defined. We also assume

$$
\max_{i=1,2,\ldots,p} \left| \sum_{j \in \mathcal{N}_n} m_{ij} \theta_{1,j} \right| = O\left( \frac{1}{\sqrt{n \log(n)}} \right), \quad \max_{i=1,2,\ldots,p} \left| \sum_{j=1}^p m_{ij} \theta_{1,j} \right| = O\left( \frac{1}{n \log(n)} \right)
$$

6. We assume that there exists a constant $\eta \geq \alpha_\sigma > 0$ such that

$$
\eta^{-\nu_b} \sum_{j \notin \mathcal{N}_n} |\theta_j| = O(n^{-\alpha_\sigma}), \quad \frac{|B_n|}{\eta^{\eta}} = O(n^{-\alpha_\sigma})
$$

7. We assume that $|\mathcal{M}| \leq r$, the rank of design matrix $X$, and the matrix $T = (c_{ik})_{i \in \mathcal{M}, k=1,2,\ldots,r}$ is of full rank($rank|\mathcal{M}|$). In addition, we assume one of the two following conditions happens:

7.1) 

$$
\max_{i \in \mathcal{M}, l=1,2,\ldots,n} \left| \frac{1}{\tau_i} \sum_{k=1}^r c_{ik}p_{lk} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right) \right| = o\left( \min\left( n^{(\alpha_\sigma - 1)/2} \times \log^{-3/2}(n), \ n^{-1/3} \times \log^{-3/2}(n) \right) \right)
$$

7.2) $\alpha_\sigma < 1/2$ and

$$
|\mathcal{M}| = o(n^{\alpha_\sigma} \times \log^{-3}(n)), \quad \max_{i \in \mathcal{M}, l=1,2,\ldots,n} \left| \frac{1}{\tau_i} \sum_{k=1}^r c_{ik}p_{lk} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right) \right| = O(n^{-\alpha_\sigma} \times \log^{-3/2}(n))
$$

**Remark 2**

The definition of $\tau_i$ requires that $\sqrt{n} \max_{i \in \mathcal{M}, l=1,2,\ldots,n} \left| \frac{1}{\tau_i} \sum_{k=1}^r c_{ik}p_{lk} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right) \right| > c > 0$ for some constant $c$, therefore if we need to apply assumption 7.2), then $\alpha_\sigma$ should be smaller than $1/2$.

Like the conditions used by Shao and Deng [13], assumptions 1) to 4) are applied for model selection consistency and consistency of estimator $\hat{\beta}$ and $\hat{\gamma}$ and assumption 6) is applied to make sure that the estimator of variance $\hat{\sigma}^2$ is consistent for real variance $\sigma^2$. Coincide with the illustration in remark 1 the key purpose for making assumption 5) is to make sure that the bias introduced by thresholding does not outweigh stochastic errors. The reason for making assumption 7) is to make sure that the residuals are sufficiently mixed so that individual residual does not make significant contribution on the stochastic error. Assumption 7) also shows a tradeoff between the number of linear combinations and how well the residuals are mixed. That is, if we want to provide the simultaneous confidence region for many linear combinations of parameters, then the residuals are required to be mixed well.

### 3 Some important lemmas

In this section, we introduce three lemmas which will be frequently used in the following sections. The first one comes from Whittle [25], which directly contributes to model selection consistency. The second one and the third one are similar with
Proof. According to theorem 2 in [25], for any \( i = 1, 2, \ldots, k \),

\[
\Pr\left( \max_{i=1,2,\ldots,k} \left| \sum_{j=1}^{n} \gamma_{ij} \epsilon_j \right| > \delta \right) \leq \frac{kED^{m/2}}{\delta^m}.
\]

(20)

Therefore, choose \( E = 2^m C(m) \mathbb{E}|\epsilon|^m \), we have

\[
\Pr\left( \max_{i=1,2,\ldots,k} \left| \sum_{j=1}^{n} \gamma_{ij} \epsilon_j \right| > \delta \right) \leq \frac{kED^{m/2}}{\delta^m}.
\]

(22)

**Lemma 2**

Suppose \( \epsilon = (\epsilon_1, \ldots, \epsilon_n)^T \) are joint normal random variables (not necessarily independent) with mean \( \mathbb{E}\epsilon_i = 0 \), full rank covariance matrix \( \mathbb{E}\epsilon \epsilon^T \) and marginal variance \( \sigma_i^2 = \mathbb{E}\epsilon_i^2 > 0 \) for \( i = 1, 2, \ldots, n \). In addition, suppose there exists two constants \( 0 < c_0 \leq C_0 < \infty \) such that \( c_0 \leq \sigma_i \leq C_0 \) for \( i = 1, 2, \ldots, n \), then for any given \( \delta > 0 \), we have

\[
\sup_{x \in \mathbb{R}} \left( \Pr( \max_{i=1,2,\ldots,n} |\epsilon_i| \leq x + \delta ) - \Pr( \max_{i=1,2,\ldots,n} |\epsilon_i| \leq x ) \right) \leq C\delta(\sqrt{\log(n)} + \sqrt{\log(\delta)} + 1)
\]

(23)

Here \( C \) is a constant which only depends on constants \( c_0, C_0 \).

**Lemma 3**

Suppose \( \epsilon = (\epsilon_1, \ldots, \epsilon_n)^T \) are independent and identically distributed random variables with \( \mathbb{E}\epsilon_1 = 0 \), \( \mathbb{E}\epsilon_i^2 = \sigma^2 \) and \( \mathbb{E}|\epsilon|^3 < \infty \), \( \Gamma = (\gamma_{ij})_{i=1,2,\ldots,n,j=1,2,\ldots,k} \) is an \( n \times k \) (\( 1 \leq k \leq n \)) rank \( k \) matrix such that there exists constants \( 0 < c_\Gamma \leq C_\Gamma < \infty \) and \( \sigma_i^2 \leq \sum_{j=1}^{n} \gamma_{ij}^2 \leq C_i^2 \) for \( i = 1, 2, \ldots, k \), \( \hat{\sigma}^2 = \hat{\sigma}^2(\epsilon) \) is an estimator of variance \( \sigma^2 \) and random variables \( \epsilon^*|\epsilon = (\epsilon_1^*, \ldots, \epsilon_n^*)^T \) are independent and identically distributed random variables with normal distribution \( \mathcal{N}(0, \hat{\sigma}^2) \) such that \( \epsilon^*_i/\hat{\sigma} \) is independent of \( \epsilon \) for \( i = 1, 2, \ldots, n \), in addition suppose one of the following conditions happens,
C1) There exists a constant $0 < \alpha \leq 1/2$ such that

$$|\sigma^2 - \hat{\sigma}^2| = O_p(n^{-\alpha}) \quad \text{and} \quad \max_{j=1,2,\ldots,n, i=1,2,\ldots,k} |\gamma_{ji}| = o(\min(n^{(\alpha-1)/2} \times \log^{-3/2}(n), \ n^{-1/3} \times \log^{-3/2}(n)))$$ (24)

C2) There exists a constant $0 < \alpha < 1/2$ such that

$$|\sigma^2 - \hat{\sigma}^2| = O_p(n^{-\alpha}), \quad k = o(n^{\alpha} \times \log^{-3}(n)), \quad \max_{j=1,\ldots,n, i=1,\ldots,k} |\gamma_{ji}| = O(n^{-\alpha} \times \log^{-3/2}(n))$$ (25)

Then we have

$$\sup_{x \in [0,\infty)} |\text{Prob}( \max_{i=1,2,\ldots,k} |\sum_{j=1}^{n} \gamma_{ji} \epsilon_j| \leq x) - \text{Prob}^*( \max_{i=1,2,\ldots,k} |\sum_{j=1}^{n} \gamma_{ji} \epsilon_j^*| \leq x)| = o_P(1)$$ (26)

In particular, if we choose $\hat{\sigma} = \sigma$, by assuming one of the following two conditions,

C1

$$\max_{j=1,2,\ldots,n, i=1,2,\ldots,k} |\gamma_{ji}| = o(n^{-1/3} \times \log^{-3/2}(n))$$ (27)

C2

$$k \times \max_{j=1,2,\ldots,n, i=1,2,\ldots,k} |\gamma_{ji}| = o(\log^{-9/2}(n))$$ (28)

We have

$$\lim_{n \to \infty} \sup_{x \in [0,\infty)} |\text{Prob}( \max_{i=1,2,\ldots,k} |\sum_{j=1}^{n} \gamma_{ji} \epsilon_j| \leq x) - \text{Prob}( \max_{i=1,2,\ldots,k} |\sum_{j=1}^{n} \gamma_{ji} \epsilon_j^*| \leq x)| = 0$$ (29)

A simple observation is that condition C1) implies C1’) and condition C2) implies C2’). If we need to estimate residuals’ variance $\sigma^2$, then we need stronger conditions to ensure normal approximation. Condition C1) is designed for the situation when the number of linear combinations $k$ is as large as the sample size $n$ and condition C2) is used when the number of linear combinations is significantly smaller than the sample size $n$.

The difference between lemma 3 and the classical central limit theorem is that we allow the number of linear combinations $k$ to go to infinity as the sample size $n$ increases. Asymptotically, since $k$ can be infinity, the random variable $\max_{i=1,2,\ldots,k} |\sum_{j=1}^{n} \gamma_{ji} \epsilon_j|$ may not have asymptotic distribution and central limit theorem fails. However, if the residuals are mixed well, according to lemma 3, using normal random variables to approximate the behavior of $\max_{i=1,2,\ldots,k} |\sum_{j=1}^{n} \gamma_{ji} \epsilon_j|$ is still a good idea.

With the help of lemma 3, we can establish the normal approximation theorem and construct the simultaneous confidence region for the estimator $\hat{\gamma}$.

4 Consistency and Gaussian approximation theorem for the debiased and threshold ridge regression method

In this section, we concentrate on showing that the debiased and threshold ridge regression statistics $\hat{\gamma}$ is consistent and its distribution can be approximated by the distribution of several joint normal random variables. In addition, we will show
that $\hat{\sigma}^2$ defined in (7) is consistent with the residuals' variance $\sigma^2$.

**Theorem 1**

1. Suppose assumptions 1) to 5) happen, then we have

$$\text{Prob}(\hat{N}_{b_n} \neq N_{b_n}) = O(n^{\alpha_p+m_p-m_0}) \quad \text{and} \quad \max_{i=1,2,\ldots,p_1} |\hat{\gamma}_i - \gamma_i| = O_p(|M|^{1/m} \times n^{-\eta}) \quad (30)$$

2. Suppose assumptions 1) to 6) happen, then we have

$$|\hat{\sigma}^2 - \sigma^2| = O_p(n^{-\alpha_p}) \quad (31)$$

Here $\hat{\sigma}^2$ is defined in (7).

When $|M|$ and $p$ are not very large, the first result in theorem 1 shows that the threshold ridge regression estimator is consistent under model selection and under infinity norm. In the proof of theorem 1, we see that $\max_{i\in M} \sqrt{\sum_{k=1}^{r} c_{ik}^2}$ can be of order larger than $O(1)$ and (15) can be relaxed, but we need these conditions to prove the normal approximation theorem.

In this paper, we allow the number of linear combinations $|M|$ to grow as the sample size $n$ increases, but an obvious problem is that the maximum $\max_{i=1,2,\ldots,p_1} |\hat{\gamma}_i - \gamma_i|$ may not have asymptotic distribution. We adopt the idea in Chernozhukov et.al. [26] and show that the distribution of maximum $\max_{i=1,2,\ldots,p_1} |\hat{\gamma}_i - \gamma_i|$ can be approximated by the maximum of joint normal random variables when sample size becomes large.

**Theorem 2**

Suppose assumptions 1) to 7) and define $H(x)$, $c_{1-\alpha}$ as in (8) and (10), then we have

1. $$\lim_{n \to \infty} \sup_{x \geq 0} |\text{Prob} \left( \max_{i=1,2,\ldots,p_1} \frac{|\hat{\gamma}_i - \gamma_i|}{\hat{\tau}_i} \leq x \right) - H(x)| = 0 \quad (32)$$

2. $$\lim_{n \to \infty} \sup_{0 < \alpha_0 \leq \alpha \leq \alpha_1} |\text{Prob} \left( \max_{i=1,2,\ldots,p_1} \frac{|\hat{\gamma}_i - \gamma_i|}{\hat{\tau}_i} \leq c_{1-\alpha} \right) - (1 - \alpha)| = 0 \quad (33)$$

Here $0 < \alpha_0 \leq \alpha_1 < 1$ are two given constants.

## 5 Bootstrap inference algorithm for linear combination of parameters

One of the key problems in theorem 2 is that the maximum of joint normal random variables has complex distribution and we are not able to directly calculate the $1 - \alpha$ quantile of $H(x)$. In order to solve this problem, we introduce a bootstrap algorithm(Algorithm 1). This algorithm helps approximate the $1 - \alpha$ quantile of $H(x)$ through Monte Carlo simulation.

**Algorithm 1** (Bootstrap algorithm for threshold ridge regression model)

**Input:** Design matrix $X$ and dependent variable $y = X\beta + \epsilon$, linear combination matrix $M$, ridge parameter $\rho_n$, threshold level $b_n$, confidence level $0 < 1 - \alpha < 1$ and number of bootstrap replicates $B$

1. Calculate $\hat{\theta}$, $\hat{\gamma}$ defined in (6) and $\hat{\tau}_i$, $i = 1,2,\ldots,p_1$, $\hat{\sigma}$ respectively defined in (8), (7)
2. Generate independent and identically distributed residuals \( e^* = (e_1^*, \ldots, e_n^*)^T \) with normal marginal distribution which has mean 0 and variance \( \hat{\sigma}^2 \), then calculate \( y^* = X\hat{\theta} + e^* \) and \( \hat{\theta}_1 = Q_1 Q_1^T \hat{\theta} \)

3. Calculate \( \hat{\theta}^{**} = (X^TX + \rho_n I_p)^{-1}X^Ty^* \) and \( \hat{\theta}^* = \hat{\theta}^{**} + \rho_n \times Q(\lambda^2 + \rho_n I_r)^{-1}Q^T \hat{\theta}^{**} + \hat{\theta}_1 \), then recalculate \( \hat{\gamma}_{i,n}^* = \{i | \hat{\theta}_i^* > b_n\} \), \( \hat{\theta}^* = (\hat{\theta}_1^*, \ldots, \hat{\theta}_p^*)^T \) such that \( \hat{\theta}_i^* = \hat{\theta}_i^* \times 1_{i \in \hat{\gamma}_{i,n}^*} \) for \( i = 1, 2, \ldots, p \)

4. Calculate \( \hat{\gamma}^* = M\hat{\theta}^* \) and \( E_b^* \) such that

\[
\hat{\gamma}_{i}^* = \left[ \sum_{i=1}^{r} \left( \sum_{j \in \hat{\gamma}_{i,n}^*} m_{ij} q_{jk} \right) \right]^{2} \times \left( \frac{\lambda_k}{\lambda_k + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k + \rho_n)^2} \right)^2 + \frac{1}{n}, \quad E_b^* = \max_{i=1,2,\ldots,p_1} \left( \frac{\hat{\gamma}_{i}^* - \gamma_i^*}{\gamma_i^*} \right)
\]

5. Repeat step 2. to 4. for \( B \) times and generate \( E_b^* \), \( b = 1, 2, \ldots, B \), then calculate the \( 1 - \alpha \) sample quantile \( C_{1-\alpha}^* \) of \( E_b^* \), the \( 1 - \alpha \) confidence region of \( \hat{\gamma}^* \) is given by

\[
\max_{i=1,2,\ldots,p_1} \left( \frac{\hat{\gamma}_{i}^* - \gamma_i^*}{\gamma_i^*} \right) \leq C_{1-\alpha}^*
\]

**Remark 3**

According to Gilvenko-Cantelli lemma and theorem 1.2.1. in [24], we have

\[
\lim_{B \to \infty} \sup_{x \in \mathbb{R}} \frac{1}{B} \sum_{i=1}^{B} \mathbf{1}_{E_i^* \leq x} - \text{Prob}^* \left( \max_{i=1,2,\ldots,p_1} \left( \frac{\hat{\gamma}_{i}^* - \gamma_i^*}{\gamma_i^*} \right) \leq x \right) = 0
\]

almost surely and \( C_{1-\alpha}^* \) converges to \( 1 - \alpha \) quantile \( c_{1-\alpha}^* \) of the conditional distribution \( \text{Prob}^* \left( \max_{i=1,2,\ldots,p_1} \left( \frac{\hat{\gamma}_{i}^* - \gamma_i^*}{\gamma_i^*} \right) \leq x \right) \) as \( B \to \infty \) if this distribution is continuous and strictly increasing at \( c_{1-\alpha}^* \). Thus, it is sufficient to show that

\[
\text{Prob} \left( \max_{i=1,2,\ldots,p_1} \left( \frac{\hat{\gamma}_{i}^* - \gamma_i^*}{\gamma_i^*} \right) \leq c_{1-\alpha}^* \right) \to 1 - \alpha
\]

with \( c_{1-\alpha}^* \) being the \( 1 - \alpha \) quantile of the conditional distribution \( \text{Prob}^* \left( \max_{i=1,2,\ldots,p_1} \left( \frac{\hat{\gamma}_{i}^* - \gamma_i^*}{\gamma_i^*} \right) \leq x \right) \).

If the dimension \( p \) is fixed, traditionally statisticians prefer the quadratic form of parameters to construct the simultaneous confidence region (like chapter 5 in Seber and Lee [27]). However, in order to avoid the accumulation of bias, in this paper we will use the weighted infinity norm \( \max_{i=1,2,\ldots,p_1} \frac{\hat{\gamma}_{i}^* - \gamma_i^*}{\gamma_i^*} \) to construct the simultaneous confidence region. Based on different considerations, infinite norm is frequently used in high dimensional statistics, like Zhang and Cheng [8], Chernozhukov et. al. [26] and Zhang and Wu [28]. We provide the theoretical justification of algorithm [1] in theorem 3.

**Theorem 3**

Suppose conditions 1) to 7), then we have

\[
\sup_{x \geq 0} \left| \text{Prob}^* \left( \max_{i=1,2,\ldots,p_1} \left( \frac{\hat{\gamma}_{i}^* - \gamma_i^*}{\gamma_i^*} \right) \leq x \right) - H(x) \right| = o_p(1)
\]

In addition, for any given \( 0 < \alpha < 1 \), suppose \( c_{1-\alpha}^* \) is the \( 1 - \alpha \) quantile of the conditional distribution
\[
\begin{align*}
\Pr\left(\max_{i=1,2,\ldots,p_1} \frac{|\hat{\gamma}_i - \gamma_i|}{\hat{\tau}_i} \leq x \right), \text{ we have} \\
\Pr\left(\max_{i=1,2,\ldots,p_1} \frac{|\hat{\gamma}_i - \gamma_i|}{\hat{\tau}_i} \leq c_{1-\alpha} \right) \to 1 - \alpha
\end{align*}
\]

as \( n \to \infty. \)

### 6 Bootstrap prediction algorithm for the regression method

Apart from classical statistical inference, in this chapter we also provide a bootstrap prediction algorithm which generates the simultaneous prediction region for future observations \( y_f \). Unlike statistical inference, prediction tries to analyze the behavior of one or several future observations \cite{21}. Since we are trying to analyze one specific instance rather than the underlying population, normal approximation does not work and the width of the prediction region does not shrink to 0 as sample size \( n \) increases. Suppose the future observation is \( y_f = x_f^T \beta + \epsilon_f \) and the predictor is \( \hat{y}_f = x_f^T \hat{\beta} \), according to chapter 3.6.2 in \cite{21}, the prediction root \( y_f - \hat{y}_f = x_f^T (\beta - \hat{\beta}) + \epsilon_f \) consists of asymptotically negligible error \( x_f^T (\beta - \hat{\beta}) \) and the non-negligible error \( \epsilon_f \). Distribution of the first term can be approximated by normal distribution but distribution of the second term needs to be estimated from data. This observation helps us create bootstrap algorithm \cite{2}. We adopt definition 2.4.1 in \cite{21} and define the asymptotically valid prediction region in definition \cite{1}.

#### Definition 1

Suppose \( n \times p \) design matrix \( X \) and dependent variable \( y \) satisfy \( y = X\beta + \epsilon \) with \( \epsilon = (\epsilon_1, \ldots, \epsilon_n)^T \) being independent and identically distributed random variables, and in addition suppose there are new observations \( X_f \) and \( y_f = X_f \beta + \epsilon_f \) with \( \epsilon_f = (\epsilon_{f1}, \ldots, \epsilon_{fk})^T \) being independent and identically distributed random variables which are independent of \( \epsilon \) and has the same marginal distribution of \( \epsilon_i \), the set \( \Gamma = \Gamma(X, y) \) is an asymptotically valid \( 1 - \alpha \) prediction region if

\[
\Pr(y_f \in \Gamma) \to 1 - \alpha
\]

as \( n \to \infty. \)

We show that the residuals’ distribution can be consistently estimated in lemma \cite{4}. In order to prove consistency of the bootstrap algorithm \cite{2} in addition to assumptions 1) to 7), we need the residuals’ cumulative distribution function to be continuous and the number of linear combinations to be finite.

#### Additional assumptions

8) Cumulative distribution function of residuals \( F(x) = \Pr(\epsilon_1 \leq x) \) is continuous

9) number of linear combinations \( p_1 = O(1) \)

If \( F(x) \) is continuous, for any \( a > 0 \), there exists a number \( Z > 0 \) such that \( F(x) > 1 - a \) for any \( x \geq Z \) and \( F(x) < a \) for any \( x \leq -Z \). Notice that continuous function is uniformly continuous in a compact set, we can choose \( 1/4 > \delta > 0 \) being sufficiently small so that \( \sup_{x,y \in [-Z-1,Z+1],|x-y| \leq \delta} |F(x) - F(y)| < a \) and correspondingly for any \( x,y \in \mathbb{R}, |x-y| < \delta, \) if \( |x| \leq Z + 1/2, \) then \( |y| \leq Z + 1 \) and \( |F(x) - F(y)| \leq a \) and if \( |x| > Z + 1/2, \) then \( |y| \geq Z + 1/4, \) which implies that \( |F(x) - F(y)| \leq 2a, \) thus \( F \) is uniformly continuous on \( \mathbb{R}. \) This property will be used in the proof of lemma \cite{4}.

#### Lemma 4
Supposes conditions 1) to 6) and 8), if we define the estimated un-centered residuals $\hat{e} = (\hat{e}_1, ..., \hat{e}_n)^T = y - X\hat{\theta}$ and the centered residuals $\tilde{e} = (\tilde{e}_1, ..., \tilde{e}_n)^T$ such that $\tilde{e}_i = \hat{e}_i - \frac{1}{n} \sum_{i=1}^{n} \hat{e}_i$, then we have

$$\sup_{x \in \mathbb{R}} |\hat{F}(x) - F(x)| \to \rho 0 \quad \text{Here } \hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{\hat{e}_i \leq x}$$

as $n \to \infty$.

Algorithm 2

**Input:** Design matrix $X$ and dependent variable $y = X\beta + \epsilon$, new $p_1 \times p$ design matrix $X_f$, ridge parameter $\rho_n$, threshold level $b_n$, confidence level $0 < 1 - \alpha < 1$ and the number of bootstrap replicates $B$

1. Calculate $\hat{\theta}$ defined in (6), $\tilde{y}_f = X_f\hat{\theta}$, $\tilde{\sigma}$ defined in (7) and $\tilde{e}$ defined in lemma 4
2. Generate independent and identically distributed residuals $e^* = (e^*_1, ..., e^*_n)^T$ with normal marginal distribution which has mean 0 and variance $\sigma^2$, and independent and identically distributed residuals $e_f^* = (e^*_f,1, ..., e^*_f,p_1)^T$ whose marginal distribution is $\tilde{F}$ defined in lemma 4, then calculate $y^* = X\hat{\theta} + e^*$ and $\tilde{\theta} = Q \tilde{\sigma}^2 \tilde{\theta}$, then recalculate $\tilde{N}_{b_n} = \{i||\tilde{\theta}^*_i| > b_n\}$, $\bar{\theta}^* = (\tilde{\theta}^*_1, ..., \tilde{\theta}^*_p)^T$ such that $\tilde{\theta}^*_i = \tilde{\theta}^*_i \times 1_{i \in \tilde{N}_{b_n}}$ for $i = 1, 2, ..., p$
3. Calculate $\hat{\theta}^{**} = (X^T X + \rho_n I_p)^{-1} X^T y^*$ and $\bar{\tilde{\theta}}^{**} = \tilde{\theta}^{**} + \rho_n \times Q(\Lambda_2^2 + \rho_n I_r)^{-1} Q^T \bar{\tilde{\theta}}^{**} + \tilde{\theta}_\perp$, then calculate $\tilde{N}_{b_n}^{**} = \{i||\tilde{\theta}^{**}_i| > b_n\}$, $\hat{\theta}^{**} = (\tilde{\theta}^{**}_1, ..., \tilde{\theta}^{**}_p)^T$ such that $\tilde{\theta}^{**}_i = \tilde{\theta}^{**}_i \times 1_{i \in \tilde{N}_{b_n}^{**}}$ for $i = 1, 2, ..., p$
4. Calculate $y_f^* = (y^*_f,1, ..., y^*_f,p_1)^T = X_f \hat{\theta} + e^*_f$ and $\tilde{y}_f^* = (\tilde{y}^*_f,1, ..., \tilde{y}^*_f,p_1)^T = X_f \tilde{\theta}^*$, define $E_b^* = \max_{i=1,2,...,p_1} |y_f^* - \tilde{y}_f^*|$ for $B$ times and generate $E_b^*$, $b = 1, 2, ..., B$, then calculate the $1 - \alpha$ sample quantile $C_{1-\alpha}^*$ of $E_b^*$, the $1 - \alpha$ prediction region of new observations $y_f = X_f\beta + \epsilon_f$ is given by

$$\max_{i=1,2,...,p_1} |y_f,i - \hat{y}_f,i| \leq C_{1-\alpha}^*$$

In theorem 4 we provide a theoretical justification of algorithm 2, and show that the prediction region generated by algorithm 2 satisfies definition 1. Similar as remark 3, we suppose $B \to \infty$ and show that $\text{Prob} \left( \max_{i=1,2,...,p_1} |y_f,i - \hat{y}_f,i| \leq c_{1-\alpha} \right) \to 1 - \alpha$ as $n \to \infty$, here $c_{1-\alpha}$ is the $1 - \alpha$ quantile of the conditional distribution $\text{Prob}^* \left( \max_{i=1,2,...,p_1} |y_f,i - \hat{y}_f,i| \leq x \right)$

Theorem 4

Suppose assumptions 1) to 6) and 8) to 9), then we have

$$\sup_{x \geq 0} |\text{Prob}^* \left( \max_{i=1,2,...,p_1} |y_f^*,i - \tilde{y}_f^*,i| \leq x \right) - \text{Prob} \left( \max_{i=1,2,...,p_1} |y_f,i - \hat{y}_f,i| \leq x \right) | = o_p(1)$$

In addition, suppose $c_{1-\alpha}^*$ is the $1 - \alpha$ quantile of conditional distribution $\text{Prob}^* \left( \max_{i=1,2,...,p_1} |y_f^*,i - \tilde{y}_f^*,i| \leq x \right)$ and $0 < \alpha < 1$ is given, then we have

$$\text{Prob} \left( \max_{i=1,2,...,p_1} |y_f,i - \hat{y}_f,i| \leq c_{1-\alpha}^* \right) \to 1 - \alpha$$

as $n \to \infty$. 
7 Numerical Simulation

In this section, we provide several numerical examples to illustrate the finite sample performance of the proposed ridge regression method and the bootstrap inference algorithms associated with this method. We define \( k_n = \sqrt{n \log(n)} \) and 4 terms \( K_i, i = 1, 2, 3, 4 \) as follow:

\[
K_1 = \max_{i=1,2,...,p_1} k_n \left| \sum_{j \in N_{n_i}} m_{ij} \theta_j \right|, \quad K_2 = \max_{i=1,2,...,p_1} k_n \left| \sum_{j=1}^r m_{ij} \theta_{i,j} \right|, \quad K_3 = b_n \sum_{j \notin N_{n_i}} |\theta_j|, \quad K_4 = \sqrt{|N_{n_i}|} \lambda_r
\]

(45)

Assumption 5) and 6) require that these four terms should be close to 0. In the numerical examples, we see that the proposed algorithms still have good performance even though some of the \( K_i \) are not very small. However if \( K_i, i = 1, 2, 3, 4 \) become very large (like case 5), then the proposed ridge regression method may have large error and the associated bootstrap algorithms fail to catch the correct confidence region.

We apply two types of strategies to generate the design matrix, linear combination matrix and the parameter \( \beta \). When \( p \leq n \), we choose \( \beta = (\beta_1, ..., \beta_n)^T \) such that \( \beta_i = 2, i = 1, 2, 3 \), \( \beta_i = -2, i = 4, 5, 6 \), \( \beta_i = 1, 0, i = 7, 8, 9 \), \( \beta_i = -1, i = 10, 11, 12 \), \( \beta_i = 0.004, i = 13, ..., 30 \) and 0 otherwise. We generate the design matrix \( X = (x_1^T, ..., x_n^T)^T \) by multivariate normal random variables with covariance matrix \( \Sigma \) which has diagonal 2.0 and off-diagonal 0.5 (this is similar with Shao [13]) and fix the design matrix after generating them. For the first \( |M| \) linear combinations, we generate them through independent normal random variables with mean 0.5 and variance 1.0, and for the remaining linear combinations, we let the first 50 elements to be 0.0 and generate the remaining elements by independent normal random variables with mean 1.0 and variance 4.0. We fix the linear combination matrix after generating it.

On the other hand, if \( p > n \), we choose the parameter \( \beta_0 \) such that \( \beta_{0,1} = \beta_{0,2} = \beta_{0,3} = 1.0, \beta_{0,4} = \beta_{0,5} = \beta_{0,6} = -1.0 \) and 0.0 otherwise. We generate the design matrix \( X = (x_1^T, ..., x_n^T)^T \) through multivariate normal random variables with mean 0 and covariance matrix \( \Sigma \) which has diagonal element 4.0 and off-diagonal element 0.2. We generate the linear combination matrix thought the following strategy: for the first \( |M| \) rows, we assign the first 6 columns values which are generated by normal random variables with mean 0.5 and variance 1.0 and for each row, we randomly choose 15 columns form the 7th column to the \( p \)th column and assign them values which are generated by normal random variables with mean 0 and variance 0.25. The other elements are assigned to be 0.0. For the remaining rows, for each row we randomly choose 15 columns form the 7th column to the \( p \)th column and assign them values which are generated by normal random variables with mean 0 and variance 0.25, then assign the other elements to be 0.0. We perform tight singular value decomposition \( X = P \Lambda Q^T \) and define \( \beta_1 = QQ^T \beta_0 \).

We define two types of methods to generate residuals:

R1) We independently generate \( z_1 = (z_{1,1}, ..., z_{n,1})^T, \; z_2 = (z_{1,2}, ..., z_{n,2})^T \) through exponential random variables with scale parameter \( \sqrt{2} \) (or variance 2), then we define \( \epsilon = z_1 - z_2 \), so that \( \epsilon, i = 1, 2, ..., n \) has variance 4.

R2) We independently generate \( \epsilon = (\epsilon_1, ..., \epsilon_n)^T \) through t-distribution with degrees of freedom 8/3, so that it still has marginal variance 4.

We list the information about how we generate simulation cases in table I.

When \( p \leq n \) and the design matrix has full rank, the proposed ridge regression method estimates \( \beta \) and there is no ambiguity. However, if \( p > n \), similar with Shao and Deng [13], the proposed ridge regression method estimates \( QQ^T \beta \) instead of \( \beta \). Unfortunately, the sparsity assumptions (like assumption 5) or 6)) may not be satisfied for \( QQ^T \beta \) and \( Q \bot Q^T \beta \).
may have large norm. If the underlying bias $Q_{\perp}Q_{\perp}^T\beta$ has large norm, then the performance of the proposed ridge regression method and the associated bootstrap inference algorithms will be affected.

We plot the error $\|\widehat{\gamma} - \gamma\|_2$ for different linear regression methods in figure 1. One advantage of the proposed ridge regression method is that it is robust with the fluctuation in ridge parameter $\rho_n$. As we can see, even though methods like Lasso or threshold ridge regression perform well with suitable $\rho_n$, as $\rho_n$ changes, the error enlarges drastically. On the contrary, the error of the proposed method does not increase significantly when $\rho_n$ deviates from its optimal value. This property ensures that the proposed method has good performance even when model selection procedures (like 10-fold cross validation) do not select the optimal ridge parameter. Case 4 and 5 illustrate how the underlying bias $Q_{\perp}Q_{\perp}^T\beta$ affects the performance of linear regression methods. Even though $X\beta_0 = X\beta_1$ (which means that we cannot tell the difference between $\beta_0$ and $\beta_1$ based on data $X$ and $y = X\beta_0 + \epsilon = X\beta_1 + \epsilon$), $M\beta_0$ is not necessarily equal to $M\beta_1$. Under this situation, Lasso methods tend to choose $\beta_0$ and Ridge regression methods tend to choose $\beta_1$. If the underlying parameters are not the ones favored by the linear regression method, then the underlying bias rather than the stochastic error will mainly contribute to the total estimation error.

Table 1: Characters about design matrix $X$, linear combination matrix $M$, parameters $\beta$ and residuals $\epsilon$ for simulations cases

| Case | # Samples | Dimension | Residual | # combinations / $|M|$ | $\lambda_p$ | Parameters |
|------|-----------|-----------|----------|-------------------------|------------|------------|
| 1    | 3000      | 1500      | R1)      | 800 / 300               | 22.103     | $\beta$   |
| 2    | 3000      | 1500      | R2)      | 800 / 300               | 22.103     | $\beta$   |
| 3    | 3000      | 2400      | R1)      | 800 / 300               | 8.244      | $\beta$   |
| 4    | 3000      | 4500      | R1)      | 800 / 300               | 24.774     | $\beta_1$ |
| 5    | 3000      | 4500      | R1)      | 800 / 300               | 24.774     | $\beta_0$ |

We list performance of bootstrap algorithm 1 on different simulation cases in table 2. From case 1 and 2, we can see that the Gaussian approximation theorem (theorem 2) works for the proposed ridge regression method. Residuals’ distribution in case 1 and 2 are different, but the 93% quantile of the statistics $\max_{i=1,2,\ldots,p_1}\frac{|\widehat{y}_{i}-\gamma_{i}|}{\tau}$ are approximately the same. Case 3 shows that algorithm 1 works even when the minimum singular value $\lambda_r$ is not very large. Case 4 and 5 provide similar 95% quantiles. However, the underlying bias $Q_{\perp}Q_{\perp}^T\beta$ makes $K_2$ large and therefore invalidates algorithm 1 in case 5.

Table 2: Performance of algorithm 1, the desired coverage probability is 95%, $\rho_n$ and $b_n$ are chosen by 10-fold cross validation.

| Case | $K_1$ | $K_2$ | $K_3$ | $K_4$ | $\rho_n$ | $b_n$ | Sos error | Probability | Average $C_{1-\alpha}^\perp$ |
|------|-------|-------|-------|-------|----------|-------|------------|-------------|-------------------|
| 1    | 11.728| 0.0   | 0.014 | 1.749 | 98.138   | 0.209 | 1.542      | 0.924       | 6.888            |
| 2    | 11.728| 0.0   | 0.011 | 1.749 | 169.583  | 0.166 | 2.420      | 0.928       | 6.824            |
| 3    | 13.534| 0.0   | 0.019 | 5.935 | 40.770   | 0.283 | 2.303      | 0.953       | 7.089            |
| 4    | 19.560| 5.831 × 10^{-13} | 6.604 | 2.707 | 5.557 | 0.166 | 1.447 | 0.960 | 7.422 |
| 5    | 19.560| 453.240 | 6.604 | 2.707 | 5.557 | 0.106 | 14.818 | 0.000 | 7.415 |

We list performance of the bootstrap prediction algorithm 2 in table 3. In order to satisfy assumption 9), we define the new prediction matrix $X_f$ as the first 200 rows of $M$, correspondingly we have $\#$ Combinations $= |M| = 200$. Unlike statistical inference, residuals’ distribution will make influence on the 95% quantile of $\max_{i=1,2,\ldots,p_1}|\widehat{y}_{f,i} - \widehat{y}_{f,i}|$, that is why case 1 and 2 have two different 95% quantiles. Compared to algorithm 1, the bootstrap prediction algorithm 2 can tolerate moderate bias in the parameters $\beta$. 

14
Figure 1: sum of square loss $\|\hat{\gamma} - \gamma\|_2$ for different regression methods, ridge(Lasso) parameter and threshold are chosen by 10-fold cross validation. 'Deb Thr' means the method proposed in this paper, 'Thr Lasso' and 'Thr Ridge' respectively means threshold Lasso and threshold ridge regression. Dots represent the threshold $b_n$ and the ridge parameter $\rho_n$ selected by 10-fold cross validation for different methods.
Table 3: Performance of bootstrap algorithm 2, the desired coverage probability is 95%, parameters are chosen the same as table 2.

| Case | Coverage Probability | Average $C_{1-\alpha}$ |
|------|----------------------|------------------------|
| 1    | 0.929                | 11.969                 |
| 2    | 0.934                | 33.940                 |
| 3    | 0.934                | 12.099                 |
| 4    | 0.962                | 13.141                 |
| 5    | 0.962                | 13.112                 |

8 Conclusion

In order to make ridge regression be suitable for high dimensional linear model, in this paper we propose a debiased and threshold ridge regression method which automatically performs model selection and avoids introducing large bias. Besides, focus on analyzing linear combinations of parameters $\gamma = M\beta$ with $M$ being a known matrix, we introduce two bootstrap algorithms (algorithm 1 and 2) which perform statistical inference and prediction for $\gamma$. Numerical performance shows that the proposed regression method is robust for the fluctuation in ridge parameter and achieves higher accuracy than classical ridge regression and threshold ridge regression method. The proposed bootstrap algorithms can provide accurate simultaneous confidence region for linear combinations $\gamma$ even when some of the assumptions are not perfectly satisfied. For statistical inference part, the number of linear combinations is allowed to increase as sample size $n$ increases. Apart from theoretical interests, the proposed methods can be applied to disciplines such as econometrics, finance, medical researches and etc.

References

[1] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 06 2006.

[2] Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.*, 37(1):246–270, 02 2009.

[3] Jian Huang, Shuangge Ma, and Cun-Hui Zhang. Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18(4):1603–1618, 2008.

[4] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[5] Yongdai Kim, Hosik Choi, and Hee-Seok Oh. Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, 103(484):1665–1673, 2008.

[6] Mingqiu Wang, Lixin Song, and Guo liang Tian. Scad-penalized least absolute deviation regression in high-dimensional models. *Communications in Statistics - Theory and Methods*, 44(12):2452–2472, 2015.

[7] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909, 2014.
[8] Xianyang Zhang and Guang Cheng. Simultaneous inference for high-dimensional linear models. *Journal of the American Statistical Association*, 112(518):757–768, 2017.

[9] Ruben Dezeure, Peter Bühlmann, and Cun-Hui Zhang. High-dimensional simultaneous inference with the bootstrap. *TEST*, 26:685–719, 2017.

[10] Xi Chen and Wen-Xin Zhou. Robust inference via multiplier bootstrap. *Ann. Statist.*, 48(3):1665–1691, 06 2020.

[11] Hanzhong Liu and Bin Yu. Asymptotic properties of lasso+mls and lasso+ridge in sparse high-dimensional linear regression. *Electron. J. Statist.*, 7:3124–3169, 2013.

[12] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[13] Jun Shao and Xinwei Deng. Estimation in high-dimensional linear models with deterministic design matrices. *Ann. Statist.*, 40(2):812–831, 04 2012.

[14] Miles Lopes. A residual bootstrap for high-dimensional regression with near low-rank designs. In *NIPS*, 2014.

[15] Peter Bühlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4):1212–1242, 09 2013.

[16] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[17] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2012.

[18] Timothy J. Vogelsang. Heteroskedasticity, autocorrelation, and spatial correlation robust inference in linear panel models with fixed-effects. *Journal of Econometrics*, 166(2):303 – 319, 2012.

[19] Xiaqing Ye and Yixiao Sun. Heteroskedasticity- and autocorrelation-robust f and t tests in stata. *The Stata Journal*, 18(4):951–980, 2018.

[20] Sílvia Gonçalves and Timothy J. Vogelsang. Block bootstrap hac robust tests: The sophistication of the naive bootstrap. *Econometric Theory*, 27(4):745791, 2011.

[21] Dimitris N. Politis. *Model-Free Prediction and Regression*. Springer International Publishing, 2015.

[22] Robert A. Stine. Bootstrap prediction intervals for regression. *Journal of the American Statistical Association*, 80(392):1026–1031, 1985.

[23] Jun Shao. *Mathematical Statistics*. Springer-Verlag New York, 2003.

[24] Dimitris N. Politis, Joseph P. Romano, and Michael Wolf. *Subsampling*. Springer-Verlag New York, 1999.

[25] Peter Whittle. Bounds for the moments of linear and quadratic forms in independent variables. *Theory of Probability & Its Applications*, 5(3):302–305, 1960.

[26] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.*, 41(6):2786–2819, 12 2013.
Since defining $x$ Thus, for any $\sigma$

First notice that for any $\sigma$

Thus, for any $x \in \mathbb{R}$, we have

Since $-\epsilon$ is also joint Gaussian with mean 0 and marginal variance $\mathbb{E}(\epsilon_j)^2 = \sigma_j^2$, from theorem 3 and (18), (19) in [29], by defining $\sigma = \min_{i=1,2,\ldots,n} \sigma_i \leq \max_{i=1,2,\ldots,n} \sigma_i = \overline{\sigma}$, we have

Define $C$ as the first term in (48), which only depends on $c_0$, $C_0$, we have

\[ \sup_{x \in \mathbb{R}} \left( \mathbb{P}(\max_{i=1,2,\ldots,n} \epsilon_i - x \leq \delta) \right) \leq \frac{\sqrt{2}\delta}{\overline{\sigma}} \left( \sqrt{\log(n)} + \sqrt{\max(1, \log(\sigma) - \log(\delta))} \right) \]

\[ + \frac{4\sqrt{2}\delta}{c_0} \left( \frac{\overline{\sigma}}{\sigma} \sqrt{\log(n)} + 2 + \frac{\overline{\sigma}}{\sigma} \sqrt{\max(0, \log(\sigma) - \log(\delta))} \right) \]

\[ \leq \frac{\sqrt{2}\delta}{c_0} \left( \sqrt{\log(n)} + 1 + \sqrt{\log(c_0) + \log(C_0)} + \sqrt{\log(\delta)} \right) \]

\[ + \frac{4\sqrt{2}\delta C_0}{c_0} \left( 2 + \sqrt{\log(c_0) + \log(C_0)} + \sqrt{\log(\delta)} \right) \]

\[ \leq \left( \sqrt{2} \times (1 + \log(c_0) + \log(C_0)) \right) + \frac{4\sqrt{2} C_0}{c_0} (2 + \sqrt{\log(c_0) + \log(C_0)} + \sqrt{\log(\delta)}) \times \delta \left( \sqrt{\log(n)} + 1 + \sqrt{\log(\delta)} \right) \]
For any given $x$,

In this proof, for convenience, we let $\Gamma = (\gamma_1, ..., \gamma_k)$, correspondingly for $i = 1, 2, ..., k$, $\gamma_i^T \epsilon = \sum_{j=1}^n \gamma_{ji} \epsilon_j$. From Lemma A.2 and (8) in [26] and (S1) to (S5) in the supplementary material of [30], for $x = (x_1, ..., x_n)$ and $y, z \in \mathbb{R}$, define

$$F_\beta(x) = \frac{1}{\beta} \log \left( \sum_{i=1}^n \exp(\beta x_i) \right), \quad g_0(y) = (1 - \min(1, \max(y, 0))^4), \quad g_\psi, z(y) = g_0(\psi(y - z)) \quad (50)$$

with $\beta, \psi > 0$, then we have $g_\psi, z \in C^3$ being nonincreasing function, $g_0 = 1$ with $y \leq 0$, $0$ with $y \geq 1$ and

$$g_* = \max_{y \in \mathbb{R}} |g_0'(y)| + |g_0''(y)| + |g_0'''(y)| < \infty, \quad 1_{y \leq z} \leq g_\psi, z(y) \leq 1_{y \geq z + \psi}^{-1}$$

$$\sup_{y, z \in \mathbb{R}} |g_\psi, z'(y)| \leq g_* \psi, \quad \sup_{y, z \in \mathbb{R}} |g_\psi, z''(y)| \leq g_* \psi^2, \quad \sup_{y, z \in \mathbb{R}} |g_\psi, z'''(y)| \leq g_* \psi^3$$

$$\partial F_\beta = \frac{\exp(\beta x_i)}{\sum_{j=1}^n \exp(\beta x_j)} \Rightarrow \frac{\partial F_\beta}{\partial x_i} \geq 0, \quad \sum_{i=1}^n \frac{\partial F_\beta}{\partial x_i} = 1,$$

$$\sum_{i=1}^n \sum_{j=1}^n |\frac{\partial^2 F_\beta}{\partial x_i \partial x_j}| \leq 2\beta, \quad \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n |\frac{\partial^3 F_\beta}{\partial x_i \partial x_j \partial x_k}| \leq 6\beta^2$$

$$F_\beta(x_1, ..., x_n) - \frac{\log(n)}{\beta} \leq \max_{i=1, ..., n} x_i \leq F_\beta(x_1, ..., x_n)$$

For any given $x = (x_1, ..., x_n) \in \mathbb{R}^n$, define function

$$G_\beta(x) = \frac{1}{\beta} \log \left( \sum_{i=1}^n \exp(\beta x_i) + \sum_{i=1}^n \exp(-\beta x_i) \right) = F_\beta(x_1, ..., x_n, -x_1, ..., -x_n) \quad (52)$$

Combine with (51) and (46), we have for $i, j, k = 1, ..., n$

$$G_\beta(x) - \frac{\log(2n)}{\beta} \leq \max_{i=1, ..., n} |x_i| \leq G_\beta(x), \quad \frac{\partial G_\beta}{\partial x_i} = \frac{\partial F_\beta}{\partial x_i} - \frac{\partial F_\beta}{\partial x_{i+n}} \Rightarrow \sum_{i=1}^n |\frac{\partial G_\beta}{\partial x_i}| \leq \sum_{i=1}^n |\frac{\partial F_\beta}{\partial x_i}| + \sum_{i=1}^n |\frac{\partial F_\beta}{\partial x_{i+n}}| = 1$$

$$\frac{\partial^2 G_\beta}{\partial x_i \partial x_j} = \frac{\partial^2 F_\beta}{\partial x_i \partial x_j} - \frac{\partial^2 F_\beta}{\partial x_{i+n} \partial x_j} + \frac{\partial^2 F_\beta}{\partial x_{i+n} \partial x_{j+n}} \Rightarrow \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n |\frac{\partial^2 F_\beta}{\partial x_i \partial x_j \partial x_k}| \leq 2\beta$$

$$\frac{\partial^3 G_\beta}{\partial x_i \partial x_j \partial x_k} = \frac{\partial^3 F_\beta}{\partial x_i \partial x_j \partial x_k} - \frac{\partial^3 F_\beta}{\partial x_{i+n} \partial x_j \partial x_k} - \frac{\partial^3 F_\beta}{\partial x_{i+n} \partial x_{j+n} \partial x_k} + \frac{\partial^3 F_\beta}{\partial x_{i+n} \partial x_{j+n} \partial x_{k+n}} \Rightarrow \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \sum_{l=1}^n |\frac{\partial^3 F_\beta}{\partial x_i \partial x_j \partial x_k \partial x_l}| \leq 6\beta^2$$

and we prove the result.
Consider the composition of $g_{\psi, \lambda}$ and $G_\beta$, direct calculation shows that

$$
\frac{\partial g_{\psi, \lambda}(G_\beta(x_1, \ldots, x_n))}{\partial x_i} = g_{\psi, \lambda}'(G_\beta(x_1, \ldots, x_n)) \frac{\partial G_\beta}{\partial x_i} \Rightarrow \sum_{i=1}^{n} \left| \frac{\partial g_{\psi, \lambda}(G_\beta(x_1, \ldots, x_n))}{\partial x_i} \right| \leq \sum_{i=1}^{n} \left| \frac{\partial g_{\psi, \lambda}(G_\beta(x_1, \ldots, x_n))}{\partial x_i} \right| \leq g_{\psi, \lambda} \psi
$$

$$
\frac{\partial^2 g_{\psi, \lambda}(G_\beta(x_1, \ldots, x_n))}{\partial x_i \partial x_j} = g_{\psi, \lambda}''(G_\beta(x_1, \ldots, x_n)) \frac{\partial^2 G_\beta}{\partial x_i \partial x_j} + g_{\psi, \lambda}(G_\beta(x_1, \ldots, x_n)) \frac{\partial^2 G_\beta}{\partial x_i \partial x_j}
$$

$$
\Rightarrow \sum_{i=1}^{n} \sum_{j=1}^{n} \left| \frac{\partial^2 g_{\psi, \lambda}(G_\beta(x_1, \ldots, x_n))}{\partial x_i \partial x_j} \right| \leq g_{\psi, \lambda} \psi \left( \sum_{i=1}^{n} \left| \frac{\partial G_\beta}{\partial x_i} \right| \right)^2 + \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \left| \frac{\partial^2 G_\beta}{\partial x_i \partial x_j} \right| \leq g_{\psi, \lambda} \psi^2 + 2g_{\psi, \lambda} \beta
$$

$$
\frac{\partial^3 g_{\psi, \lambda}(G_\beta(x_1, \ldots, x_n))}{\partial x_i \partial x_j \partial x_k} = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \left| \frac{\partial^3 g_{\psi, \lambda}(G_\beta(x_1, \ldots, x_n))}{\partial x_i \partial x_j \partial x_k} \right| \leq g_{\psi, \lambda} \psi^3 \left( \sum_{i=1}^{n} \left| \frac{\partial G_\beta}{\partial x_i} \right| \right)^3 + 3g_{\psi, \lambda} \psi^2 \left( \sum_{i=1}^{n} \sum_{j=1}^{n} \left| \frac{\partial^2 G_\beta}{\partial x_i \partial x_j} \right| \right) \times \left( \sum_{k=1}^{n} \left| \frac{\partial G_\beta}{\partial x_k} \right| \right)
$$

$$
+ g_{\psi, \lambda} \psi \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \left| \frac{\partial^3 G_\beta}{\partial x_i \partial x_j \partial x_k} \right| \leq g_{\psi, \lambda} \psi^3 + 6g_{\psi, \lambda} \psi^2 + 6g_{\psi, \lambda} \beta^2
$$

We define $\xi = (\xi_1, \ldots, \xi_n)$ as i.i.d. random variables with the same marginal distribution as $\epsilon_1$ and being independent of $\epsilon, \epsilon^*$, so that $Prob(\max_{i=1,2,\ldots,k} |\gamma_i^T \epsilon| \leq x) = Prob^*(\max_{i=1,2,\ldots,k} |\sum_{j=1}^{n} \gamma_i^T \xi_j| \leq x)$ for any $x$. For any given $x \geq 0$, according to (46), (53) and lemma 2 for any given $\psi, \lambda, \sigma > 0$, notice that

$$
c_i^2 \leq E^* \left( \sum_{i=1}^{n} \frac{\gamma_i \epsilon_i}{\sigma} \right)^2 = \sum_{i=1}^{n} \gamma_i^2 \leq C_i^2 \quad \text{for} \quad i = 1, 2, \ldots, k
$$

There exists a constant $C$ which only depends on $c_i$ and $C_i$ such that

$$
\sup_{x \in \mathbb{R}} \left( \text{Prob}^* \left( \max_{i=1,2,\ldots,k} |\gamma_i^T \epsilon^*| \leq x + \frac{1}{\psi} + \frac{\log(2k)}{\beta} \right) - \text{Prob}^* \left( \max_{i=1,2,\ldots,k} |\gamma_i^T \epsilon^*| \leq x \right) \right)
$$

$$
= \sup_{x \in \mathbb{R}} \left( \text{Prob}^* \left( \max_{i=1,2,\ldots,k} \frac{\gamma_i^T \epsilon^*}{\sigma} \right) \right) \leq C \times \left( \frac{1}{\psi \sigma} + \frac{\log(2k)}{\beta \sigma} \right) \times \left( 1 + \sqrt{\log(n)} + \sqrt{\log \left( \frac{1}{\psi \sigma} + \frac{\log(2k)}{\beta \sigma} \right)} \right)
$$

(56)
We define $z = C \times \left( \frac{1}{\psi \sigma} + \frac{\log(2k)}{\beta \sigma} \right) \times \left( 1 + \sqrt{\log(n)} + \sqrt{\log \left( \frac{1}{\psi \sigma} + \frac{\log(2k)}{\beta \sigma} \right)} \right)$, correspondingly, for any $x \geq 0$,

$$
\begin{align*}
&\text{Prob}(\max_{i=1,2,\ldots,k} |\gamma_i^T \epsilon| \leq x) - \text{Prob}^*(\max_{i=1,2,\ldots,k} |\gamma_i^T \epsilon^*| \leq x) \\
&\leq \text{Prob}^*(\max_{i=1,2,\ldots,k} |\gamma_i^T \xi| \leq x) - \text{Prob}^*(\max_{i=1,2,\ldots,k} |\gamma_i^T \xi^*| \leq x + \frac{1}{\psi} + \frac{\log(2k)}{\beta} + z) \\
&\leq \text{Prob}^*(G_\beta(\gamma_1^T \xi, \ldots, \gamma_k^T \xi) \leq x \frac{1}{\psi} + \frac{\log(2k)}{\beta}) - \text{Prob}^*(G_\beta(\gamma_1^T \xi^*, \ldots, \gamma_k^T \xi^*) \leq x + \frac{1}{\psi} + \frac{\log(2k)}{\beta} + z) \\
&\leq \mathbb{E}^* \psi,x+1 \frac{\log(2k)}{\beta} (G_\beta(\gamma_1^T \xi, \ldots, \gamma_k^T \xi)) - \mathbb{E}^* \psi,x \frac{\log(2k)}{\beta} (G_\beta(\gamma_1^T \xi^*, \ldots, \gamma_k^T \xi^*)) + z
\end{align*}
$$

(57)

Thus, we have

$$
\begin{align*}
&\sup_{x \in [0, \infty)} |\text{Prob}(\max_{i=1,2,\ldots,k} |\gamma_i^T \epsilon| \leq x) - \text{Prob}^*(\max_{i=1,2,\ldots,k} |\gamma_i^T \epsilon^*| \leq x)| \\
&\leq z + \sup_{x \in \mathbb{R}} |\mathbb{E}^* \psi,x(\gamma_1^T \xi, \ldots, \gamma_k^T \xi)) - \mathbb{E}^* \psi,x(\gamma_1^T \xi^*, \ldots, \gamma_k^T \xi^*))|
\end{align*}
$$

(58)

For any $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, n$, define $H_{ij} = \sum_{s=1}^{i-1} \gamma_s \xi_s + \sum_{s=j+1}^{n} \gamma_s \xi_s$, $m_{ij} = \gamma_i \xi_j$ and $m^*_{ij} = \gamma_i \xi_j^*$, we have $H_{ij} + m_{ij} = H_{ij+1} + m^*_{ij+1}$ and therefore

$$
\begin{align*}
&\sup_{x \in \mathbb{R}} |\mathbb{E}^* \psi,x(\gamma_1^T \xi, \ldots, \gamma_k^T \xi)) - \mathbb{E}^* \psi,x(\gamma_1^T \xi^*, \ldots, \gamma_k^T \xi^*))| \\
&= \sup_{x \in \mathbb{R}} |\mathbb{E}^* \psi,x(G_\beta(H_{1n} + m_{1n}, \ldots, H_{kn} + m_{kn})) - \mathbb{E}^* \psi,x(G_\beta(H_{11} + m^*_{11}, \ldots, H_{k1} + m^*_{k1}))| \\
&= \sup_{x \in \mathbb{R}} \sum_{s=1}^{n} |\mathbb{E}^* \psi,x(G_\beta(H_{1s} + m_{1s}, \ldots, H_{ks} + m_{ks})) - \mathbb{E}^* \psi,x(G_\beta(H_{1s} + m^*_{1s}, \ldots, H_{ks} + m^*_{ks}))| \\
&\leq \sum_{s=1}^{n} \sup_{x \in \mathbb{R}} |\mathbb{E}^* \psi,x(G_\beta(H_{1s} + m_{1s}, \ldots, H_{ks} + m_{ks})) - \mathbb{E}^* \psi,x(G_\beta(H_{1s} + m^*_{1s}, \ldots, H_{ks} + m^*_{ks}))|
\end{align*}
$$

(59)
For any \( x \in \mathbb{R} \) and \( s = 1, 2, ..., n \),

\[
E g_{\psi, x} (G_\beta (H_{1s} + m_1, ..., H_{ks} + m_{ks})) - g_{\psi, x} (G_\beta (H_{1s} + m^*_1, ..., H_{ks} + m^*_{ks})) | \epsilon, \xi, \epsilon^*_b, b \neq s
\]

\[
= \sum_{i=1}^{k} \frac{\partial g_{\psi, x} (G_\beta (H_{1s}, ..., H_{ks}))}{\partial x_i} \gamma_{s_i} E (\xi - \epsilon^*_s | \epsilon, \xi, \epsilon^*_b, b \neq s) + \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{\partial^2 g_{\psi, x} (G_\beta (H_{1s}, ..., H_{ks}))}{\partial x_i \partial x_j} \gamma_{s_i} \gamma_{s_j} E (\xi^2 - \epsilon^*_s^2 | \epsilon, \xi, \epsilon^*_b, b \neq s)
\]

\[
+ E (g_{\psi, x} (G_\beta (H_{1s} + m_1, ..., H_{ks} + m_{ks})) | \epsilon, \xi, \epsilon^*_b, b \neq s) - g_{\psi, x} (G_\beta (H_{1s}, ..., H_{ks})) - \frac{1}{2} \sum_{i=1}^{k} \frac{\partial^2 g_{\psi, x} (G_\beta (H_{1s}, ..., H_{ks}))}{\partial x_i} m_{is}
\]

\[
- \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{\partial^2 g_{\psi, x} (G_\beta (H_{1s}, ..., H_{ks}))}{\partial x_i \partial x_j} m_{is} m_{js}
\]

\[
- E (g_{\psi, x} (G_\beta (H_{1s} + m^*_1, ..., H_{ks} + m^*_{ks})) | \epsilon, \xi, \epsilon^*_b, b \neq s) + g_{\psi, x} (G_\beta (H_{1s}, ..., H_{ks})) + \frac{1}{2} \sum_{i=1}^{k} \frac{\partial^2 g_{\psi, x} (G_\beta (H_{1s}, ..., H_{ks}))}{\partial x_i} m^*_is
\]

\[
+ \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{\partial^2 g_{\psi, x} (G_\beta (H_{1s}, ..., H_{ks}))}{\partial x_i \partial x_j} m^*_is m^*_js
\]

\[(60)\]

Notice that \( E (\xi | \epsilon, \xi, \epsilon^*_b, b \neq s) = E (\epsilon^*_s | \epsilon, \xi, \epsilon^*_b, b \neq s) = 0 \), \( E (\xi^2 - \epsilon^*_s^2 | \epsilon, \xi, \epsilon^*_b, b \neq s) = \sigma^2 - \bar{\sigma}^2 \), from multivariate Taylor's theorem (see for example, theorem 5.2. in [31] and [54]), we have

\[
|E (g_{\psi, x} (G_\beta (H_{1s} + m_1, ..., H_{ks} + m_{ks})) - g_{\psi, x} (G_\beta (H_{1s}, ..., H_{ks})))| \leq \frac{\max_{i=1, ..., k} |\gamma_{s_i}|^3}{6} E (\sum_{i=1}^{k} \sum_{j=1}^{k} \sum_{l=1}^{k} \frac{\partial^3 g_{\psi, x} (G_\beta (z_1, ..., z_k))}{\partial x_i \partial x_j \partial x_l} | \xi^2 | \epsilon, \xi, \epsilon^*_b, b \neq s)
\]

\[
\leq \frac{\max_{i=1, ..., k} |\gamma_{s_i}|^3}{6} \times (g_\psi \beta^3 + 6g_\psi \beta^2 + 6g_\psi \beta^2) E (|\epsilon|)^3
\]

\[(61)\]

\[
|E (g_{\psi, x} (G_\beta (H_{1s} + m^*_1, ..., H_{ks} + m^*_{ks})) - g_{\psi, x} (G_\beta (H_{1s}, ..., H_{ks})))| \leq \frac{\max_{i=1, ..., k} |\gamma_{s_i}|^3}{6} E (\sum_{i=1}^{k} \sum_{j=1}^{k} \sum_{l=1}^{k} \frac{\partial^3 g_{\psi, x} (G_\beta (z_1, ..., z_k))}{\partial x_i \partial x_j \partial x_l} | \epsilon^*_s | \epsilon, \xi, \epsilon^*_b, b \neq s)
\]

\[
\leq \frac{\max_{i=1, ..., k} |\gamma_{s_i}|^3}{6} \times (g_\psi \beta^3 + 6g_\psi \beta^2 + 6g_\psi \beta^2) \bar{\sigma}^3 \times D
\]
Here \( D = E[Y|^3 \rightleftharpoons Y \sim N(0, 1) \) being a standard normal random variable. Combine with \[\{58\} \text{ to } \{61\}, \) we have

\[
\begin{align*}
&\left| E^*[g_{\psi,x}(G_\beta(H_{1s} + m_{1s}, ..., H_{ks} + m_{ks})) - g_{\psi,x}(G_\beta(H_{1s} + m_{1s}^*, ..., H_{ks} + m_{ks}^*))| \\
\leq & \frac{1}{2} \left| \sigma^2 - \hat{\sigma}^2 \right| \times \max_{i=1,2,...,k} \frac{\gamma_{sl}^2}{6} \times \sum_{i=1}^{k} \sum_{j=1}^{k} \left( \frac{\partial^2 g_{\psi,x}(G_\beta(H_{1s}, ..., H_{ks}))}{\partial x_i \partial x_j} \right) \\
\leq & \frac{1}{6} \left( g_\psi \sigma^3 + 6g_\psi \sigma^2 \beta + 6g_\psi \beta^2 \right) E[\epsilon_1^2] \left( \max_{i=1,2,...,k} \gamma_{sl}^2 \right) \times \left( 6 \left( g_\psi \sigma^3 + 6g_\psi \sigma^2 \beta + 6g_\psi \beta^2 \right) \right) \sigma^3 \times \delta^3
\end{align*}
\]

and correspondingly,

\[
\begin{align*}
&\sup_{x \in [0, \infty)} \left| Prob \left( \max_{i=1,2,...,k} \gamma_{sl}^T \epsilon \leq x \right) - Prob^* \left( \max_{i=1,2,...,k} \gamma_{sl}^T \epsilon^* \leq x \right) \right| \\
\leq & z + (g_\psi \sigma^2 + g_\psi \beta) \sigma^2 - \hat{\sigma}^2 \times \sum_{i=1}^{n} \max_{s=1} \gamma_{sl}^2 + \left( E[\epsilon_1^2] + D\hat{\sigma}^3 \right) \times (g_\psi \sigma^3 + \psi \beta + \beta^2) \times \sum_{s=1}^{n} \max_{s=1} \gamma_{sl}^3
\end{align*}
\]

In particular, for any given \( \delta > 0, \) if we choose \( \psi = \beta = \log^{3/2}(n)/\delta^{1/4} \) and \( \frac{3}{2} > \hat{\sigma} > \frac{1}{2}, \) then for sufficiently large \( n, \) we have

\[
\frac{1}{\psi^4} + \frac{\log(2k)}{\psi^2} \leq \frac{4 \log(n)}{\psi^4} \leq \frac{4 \delta^{1/4}}{\sigma \sqrt{\log(n)}} < 1 \quad \text{and correspondingly}
\]

\[
z \leq \frac{4 \log(n)}{\psi^4} \leq \frac{4 \delta^{1/4}}{\sigma} \times \left( 2 \sqrt{\log(n)} + \log(\psi^2) \right) \leq C' \delta^{1/4}
\]

with \( C' = \frac{12C}{\sigma}. \)

Suppose condition C1) happens, then for any \( 1 > \delta > 0, \) there exists a \( D_\delta > 0 \) such that for sufficiently large \( n,

\[
\begin{align*}
&\sup_{x \in [0, \infty)} \left| Prob \left( \max_{i=1,2,...,k} \gamma_{sl}^T \epsilon \leq x \right) - Prob^* \left( \max_{i=1,2,...,k} \gamma_{sl}^T \epsilon^* \leq x \right) \right| \\
\leq & C' \delta^{1/4} + 2g_\psi \sigma^2 \times D_\delta \times n^{-\alpha_\epsilon} \times \frac{\delta^2}{\log^2(n)} + \left( E[\epsilon_1^2] + \frac{27D}{8} \sigma^3 \right) \times 3g_\psi \sigma^3 \times \delta^3 \times n \times \frac{1}{\log^3/2(n)}
\end{align*}
\]

We choose \( \psi = \beta = \log^{3/2}(n)/\delta^{1/4}, \) then according to \[\{63\}, \) for sufficiently large \( n, \) we know that \[\{65\} \) happens and

\[
\frac{1}{2} \sigma < \hat{\sigma} < \frac{3}{2} \sigma \quad \text{with probability } 1 - \delta, \quad \text{if \[\{58\}\] happens,}
\]

\[
\sup_{x \in [0, \infty)} \left| Prob \left( \max_{i=1,2,...,k} \gamma_{sl}^T \epsilon \leq x \right) - Prob^* \left( \max_{i=1,2,...,k} \gamma_{sl}^T \epsilon^* \leq x \right) \right| = o_P(1)
\]
If condition C2) happens, we have for any \( \delta > 0 \), there exists \( D_\delta > 0 \) such that

\[
\Pr\left( |\sigma^2 - \hat{\sigma}^2| \leq D_\delta \times n^{-\alpha} \right) \geq 1 - \delta, \quad k \leq \frac{\delta n^{\alpha}}{\log^2(n)}.
\]

Then

\[
\max_{i=1,2,\ldots,k} \sum_{j=1}^{n} \gamma_{ji}^2 \leq D_\delta, \quad \max_{j=1,2,\ldots,n,i=1,2,\ldots,k} |\gamma_{ji}| \leq D_\delta \times n^{-\alpha} \log^{3/2}(n) \tag{68}
\]

Since

\[
\sum_{j=1}^{n} \max_{i=1,\ldots,k} \gamma_{ji}^2 \leq j=1,2,\ldots,n,i=1,2,\ldots,k |\gamma_{ji}| \times \sum_{j=1}^{n} \max_{i=1,\ldots,k} \gamma_{ji}^2 \leq kD_\delta
\]

we have

\[
\sup_{x \in [0,\infty)} |\Pr\left( \max_{i=1,2,\ldots,k} |\gamma_i^T \epsilon| \leq x \right) - \Pr\left( \max_{i=1,2,\ldots,k} |\gamma_i^* T \epsilon^*| \leq x \right)|
\]

\[
\leq C' \delta^{1/4} + 2g_2\sigma D_\delta \times \frac{\log^3(n)}{\delta^{1/2}} \times \frac{\delta n^{\alpha}}{\log^3(n)} \times n^{-\alpha} + 3(\mathbb{E}[\epsilon_1]^3 + \frac{27D}{8}\sigma^3)g_2\sigma D_\delta \times \frac{\log^9/2(n)}{\delta^{3/4}} \times \frac{\delta n^{\alpha}}{\log^3(n)} \times \frac{n^{-\alpha}}{\log^{3/2}(n)}
\]

\[
= C' \delta^{1/4} + 2g_2\sigma D_\delta \delta^{1/2} + 3(\mathbb{E}[\epsilon_1]^3 + \frac{27D}{8}\sigma^3)g_2\sigma D_\delta \times \delta^{1/4}
\]

Thus, we prove \( \text{(67)} \).

If we pick \( \delta = \sigma \) and choose \( \psi = \beta = \log^{3/2}(n)/\delta^{1/4} \), then \( \text{(63)} \) can be modified as

\[
\sup_{x \in [0,\infty)} |\Pr\left( \max_{i=1,2,\ldots,k} |\gamma_i^T \epsilon| \leq x \right) - \Pr\left( \max_{i=1,2,\ldots,k} |\gamma_i^T \epsilon^*| \leq x \right)|
\]

\[
\leq C' \delta^{1/4} + (\mathbb{E}[\epsilon_1]^3 + D\sigma^3) \times g_2(\psi^3 + \psi^2 \beta + \psi \beta^2) \times \sum_{i=1,\ldots,k} \max_{j=1}^{n} |\gamma_{ji}|^3 \tag{71}
\]

Suppose condition C1') happens, for any \( \delta > 0 \) and sufficiently large \( n \), \( \max_{j=1,2,\ldots,n,i=1,2,\ldots,k} |\gamma_{ji}| \leq \delta \times n^{-1/3} \log^{-3/2}(n) \), correspondingly we have

\[
\sup_{x \in [0,\infty)} |\Pr\left( \max_{i=1,2,\ldots,k} \sum_{j=1}^{n} \gamma_{ji} \epsilon_j \leq x \right) - \Pr\left( \max_{i=1,2,\ldots,k} \sum_{j=1}^{n} \gamma_{ji} \epsilon_j^* \leq x \right)| \leq C' \delta^{1/4} + 3(\mathbb{E}[\epsilon_1]^3 + D\sigma^3)g_2 \times \delta^{3/4}
\]

and we prove \( \text{(29)} \).

Suppose condition C2') happens, for any \( \delta > 0 \) and sufficiently large \( n \), \( k \times \max_{j=1,2,\ldots,n,i=1,2,\ldots,k} |\gamma_{ji}| \leq \delta \log^{-5/2}(n) \), thus according to \( \text{(69)} \), for sufficiently large \( n \) we have

\[
\sup_{x \in [0,\infty)} |\Pr\left( \max_{i=1,2,\ldots,k} \sum_{j=1}^{n} \gamma_{ji} \epsilon_j \leq x \right) - \Pr\left( \max_{i=1,2,\ldots,k} \sum_{j=1}^{n} \gamma_{ji} \epsilon_j^* \leq x \right)| \leq C' \delta^{1/4} + 3(\mathbb{E}[\epsilon_1]^3 + D\sigma^3)g_2 D_\delta \times \delta^{1/4}
\]

and we prove \( \text{(29)} \).
Proof of theorem 1. First from (6),

\[
\text{Prob} \left( \hat{N}_{b_n} \neq N_{b_n} \right) \leq \text{Prob} \left( \min_{i \in \mathcal{N}_{b_n}} |\theta_i| \leq b_n \right) + \text{Prob} \left( \max_{i \notin \mathcal{N}_{b_n}} |\theta_i| > b_n \right)
\]

\[
\leq \text{Prob} \left( \min_{i \in \mathcal{N}_{b_n}} |\theta_i| - \max_{i \notin \mathcal{N}_{b_n}} \rho_i^2 \sum_{j=1}^r q_{ij} \tilde{\zeta}_j \left( \frac{\lambda_j}{\lambda_j^2 + \rho_n} \right)^2 \right) - \max_{i \notin \mathcal{N}_{b_n}} \rho_n \lambda_j \sum_{j=1}^r \sum_{l=1}^r |p_{ij} \epsilon_l| \leq b_n
\]

\[+
\text{Prob} \left( \max_{i \in \mathcal{N}_{b_n}} |\theta_i| + \max_{i \notin \mathcal{N}_{b_n}} \rho_i^2 \sum_{j=1}^r q_{ij} \tilde{\zeta}_j \left( \frac{\lambda_j}{\lambda_j^2 + \rho_n} \right)^2 + \max_{i \notin \mathcal{N}_{b_n}} \rho_n \lambda_j \sum_{j=1}^r \sum_{l=1}^r |p_{ij} \epsilon_l| > b_n \right)
\]

(74)

From Cauchy inequality,

\[
\max_{i=1,2,\ldots,p} \rho_i^2 \sum_{j=1}^p q_{ij} \tilde{\zeta}_j \left( \frac{\lambda_j}{\lambda_j^2 + \rho_n} \right)^2 \leq \max_{i=1,2,\ldots,p} \rho_i^2 \sum_{j=1}^r q_{ij}^2 \left( \frac{\lambda_j}{\lambda_j^2 + \rho_n} \right)^2 \leq \max_{i=1,2,\ldots,p} \frac{4 \sum_{j=1}^r q_{ij}^2}{\lambda_j^2}
\]

Thus, for sufficiently large \( n \), from assumption 4) and lemma 1,

\[
\min_{i \in \mathcal{N}_{b_n}} |\theta_i| - \max_{i \notin \mathcal{N}_{b_n}} \rho_i^2 \sum_{j=1}^r q_{ij} \tilde{\zeta}_j \left( \frac{\lambda_j}{\lambda_j^2 + \rho_n} \right)^2 \]

\[
- b_n > \frac{1}{2} \left( \frac{1}{c_b} - 1 \right) b_n
\]

\[
\Rightarrow \text{Prob} \left( \hat{N}_{b_n} \neq N_{b_n} \right) = \frac{|N_{b_n}| \times E \times 2^m}{\lambda_r \times \left( \frac{2}{c_b} - 1 \right) b_n} + \frac{p - |N_{b_n}|}{\lambda_r \times \left( \frac{2}{c_b} - 1 \right) b_n}^m = O\left(p^{\alpha_p + m_\alpha b_m - m_\beta} \right)
\]

For \( \beta = \theta + \theta^\perp \), if \( \hat{N}_{b_n} = N_{b_n} \), suppose \( \hat{\gamma} = M \hat{\theta} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_p)^T \) and \( \gamma = M \beta = (\gamma_1, \ldots, \gamma_p)^T \), from (5) and (5),

\[
\max_{i=1,2,\ldots,p_1} |\tilde{\gamma}_i - \gamma_i| = \max_{i=1,2,\ldots,p_1} \left| \sum_{j \in \mathcal{N}_{b_n}} m_{ij} \tilde{\theta}_j - \sum_{j \notin \mathcal{N}_{b_n}} m_{ij} \theta_j - \sum_{j \notin \mathcal{N}_{b_n}} m_{ij} \theta_j - \sum_{j \notin \mathcal{N}_{b_n}} m_{ij} \theta^\perp_j \right|
\]

\[
\leq \max_{i=1,2,\ldots,p_1} \rho_i^2 \sum_{k=1}^r \frac{c_{ik} \tilde{\zeta}_k}{\left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} \right)^2} + \max_{i=1,2,\ldots,p_1} \left| \sum_{k=1}^r c_{ik} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{\left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} \right)^2} \right) \right|^n \sum_{l=1}^r |p_{ik} \epsilon_l|
\]

\[
+ \max_{i=1,2,\ldots,p_1} \left| \sum_{j \notin \mathcal{N}_{b_n}} m_{ij} \theta_j \right| + \max_{i=1,2,\ldots,p_1} \left| \sum_{j \notin \mathcal{N}_{b_n}} m_{ij} \theta^\perp_j \right|
\]

(77)

According to (3) and assumption 5), if \( i \notin \mathcal{M} \), then \( c_{ik} = 0 \) for \( k = 1, 2, \ldots, r \), thus from Cauchy inequality and lemma 1,

\[
\max_{i=1,2,\ldots,p_1} \rho_i^2 \sum_{k=1}^r \frac{c_{ik} \tilde{\zeta}_k}{\left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} \right)^2} \leq \max_{i \in \mathcal{M}} \rho_i^2 \sum_{k=1}^r \frac{c_{ik}^2}{\left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} \right)^4} \leq \sqrt{C_M} \rho_i^2 \times \left\| \frac{\theta}{\lambda_i^2} \right\| = O\left(p^{\alpha_s - 2 \delta} \right)
\]

\[
\max_{i \in \mathcal{M}} \sum_{k=1}^n \left( \sum_{k=1}^r c_{ik} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{\left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} \right)^2} \right) \right)^2 \leq \max_{i \in \mathcal{M}} \sum_{k=1}^r c_{ik} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{\left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} \right)^2} \right)^2 \leq 4C_M \lambda_i^2 \lambda_i
\]

\[
\Rightarrow \text{Prob} \left( \max_{i=1,2,\ldots,p_1} \left| \sum_{k=1}^r c_{ik} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{\left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} \right)^2} \right) \right| > \delta \right) \leq \frac{|\mathcal{M}| \times E \times 2^m C_M^{\alpha_s/2}}{\lambda_i^m \delta^m} \text{ for } \forall \delta > 0
\]

\[
\Rightarrow \max_{i=1,2,\ldots,p_1} \left| \sum_{k=1}^r c_{ik} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{\left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} \right)^2} \right) \right| \sum_{l=1}^n |p_{ik} \epsilon_l| = O_p\left(|\mathcal{M}|^{1/m} \times n^{-n} \right)
\]
According to (15), for any given $0 < \xi < 1$, choose $\delta_0 = C \times (n^{\alpha_x-2\delta} + |\mathcal{M}|^{1/m} \times n^{-\eta})$ with sufficiently large $C$, then for sufficiently large $n$,

$$
\text{Prob} \left( \max_{i=1,2,\ldots,p_1} |\tilde{\gamma}_i - \gamma_i| > \delta_0 \right) \leq \text{Prob} \left( \tilde{N}_{b_n} \neq N_{b_n} \right)
$$

$$
\leq C n^{\alpha_y + m\eta - m\eta + \xi}
$$

(79)

Combine with assumption 2), we prove the first result.

For the second result, if $\tilde{N}_{b_n} = N_{b_n}$, since $X\beta = X\theta$, we have

$$
\tilde{\sigma}^2 - \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} \left( \epsilon_i - \sum_{j \in N_{b_n}} x_{ij} (\tilde{\theta}_j - \theta_j) + \sum_{j \notin N_{b_n}} x_{ij} \theta_j \right)^2 - \sigma^2
\leq \frac{1}{n} \sum_{i=1}^{n} c_i^2 - \sigma^2 + \frac{1}{n} \sum_{i=1}^{n} \left( \sum_{j \in N_{b_n}} x_{ij} (\tilde{\theta}_j - \theta_j) \right)^2 + \frac{2}{n} \sum_{i=1}^{n} \sum_{j \notin N_{b_n}} \epsilon_i x_{ij} (\tilde{\theta}_j - \theta_j)
\leq \frac{2}{n} \sum_{i=1}^{n} \sum_{j \notin N_{b_n}} \epsilon_i x_{ij} (\tilde{\theta}_j - \theta_j)
$$

(80)

From assumption 3), we have $E \left( \frac{1}{n} \sum_{i=1}^{n} c_i^2 - \sigma^2 \right)^2 \leq \frac{2}{n} (E \epsilon_i^4 + \sigma^4) = O(1/n)$, this implies that $\frac{1}{n} \sum_{i=1}^{n} c_i^2 - \sigma^2 = O_p(1/\sqrt{n})$.

For the second term, define vector $Z = (Z_1, \ldots, Z_p) = (\tilde{\theta}_j - \theta_j)$ if $j \in N_{b_n}$ and 0 otherwise, then from assumption 1) and (75),

$$
\frac{1}{n} \sum_{i=1}^{n} \left( \sum_{j \in N_{b_n}} x_{ij} (\tilde{\theta}_j - \theta_j) \right)^2 = Z^T \left( \frac{X^T X}{n} \right) Z \leq C \lambda^2 \sum_{j \in N_{b_n}} (\tilde{\theta}_j - \theta_j)^2
\leq 2 C \lambda^2 \sum_{j \in N_{b_n}} \left( \rho_n^4 \left( \sum_{k=1}^{r} \frac{q_{jk} \tilde{\zeta}_k}{\lambda_k^2 + \rho_n} \right)^2 + \left( \sum_{k=1}^{r} \frac{q_{jk} \lambda_k^2}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right) \frac{\sum_{i=1}^{n} p_{ik} \epsilon_i}{n} \right)^2
$$

(81)

$$
= O(|N_{b_n}| \times n^{2\alpha_x - 4\delta}) + 2 C \lambda^2 \sum_{j \in N_{b_n}} \left( \sum_{k=1}^{r} \frac{q_{jk} \lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right) \left( \sum_{i=1}^{n} p_{ik} \epsilon_i \right)^2
$$

Since

$$
E \sum_{j \in N_{b_n}} \left( \sum_{k=1}^{r} \frac{q_{jk} \lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right) \left( \sum_{i=1}^{n} p_{ik} \epsilon_i \right)^2 = \sigma^2 \sum_{j \in N_{b_n}} \sum_{i=1}^{n} \left( \sum_{k=1}^{r} \frac{q_{jk} \lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right) p_{ik} \epsilon_i ^2
$$

(82)

$$
= \sigma^2 \sum_{j \in N_{b_n}} \left( \sum_{k=1}^{r} q_{jk} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right) \right)^2 \leq \frac{4 \sigma^2 |N_{b_n}|}{\lambda^2}
$$

We have $\frac{1}{n} \sum_{i=1}^{n} \left( \sum_{j \in N_{b_n}} x_{ij} (\tilde{\theta}_j - \theta_j) \right)^2 = O_p(|N_{b_n}| \times n^{2\alpha_x - 4\delta} + |N_{b_n}| \times n^{-2\eta})$.

For the third term, from assumption 6) we have

$$
\frac{1}{n} \sum_{i=1}^{n} \left( \sum_{j \notin N_{b_n}} x_{ij} \theta_j \right)^2 \leq C \lambda^2 \sum_{j \notin N_{b_n}} \theta_j^2 \leq C \lambda^2 \times b_n \sum_{j \notin N_{b_n}} |\theta_j| = O(n^{-\alpha_y})
$$

(83)
For the fourth term, from Cauchy inequality and (81),

\[
E \frac{1}{n} \sum_{i=1}^{n} \sum_{j \in \mathcal{N}_b} \epsilon_i x_{ij} (\tilde{\theta}_j - \theta_j) \leq \frac{1}{n} E \left( \sum_{i=1}^{n} \epsilon_i^2 \right)^{1/2} \left( \sum_{j \in \mathcal{N}_b} (\sum_{j \in \mathcal{N}_b} x_{ij} (\tilde{\theta}_j - \theta_j))^2 \right)^{1/2} \leq \sqrt{E} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2 \times \sqrt{\frac{1}{n} E \sum_{j \in \mathcal{N}_b} \left( \sum_{j \in \mathcal{N}_b} x_{ij} (\tilde{\theta}_j - \theta_j))^2 \right)} \]

\[
= \sigma \times O(\sqrt{|\mathcal{N}_b|} \times n^{2\alpha_s - 4\delta} + |\mathcal{N}_b| \times n^{-2\eta})
\]

\[
\Rightarrow \frac{1}{n} \sum_{i=1}^{n} \sum_{j \in \mathcal{N}_b} \epsilon_i x_{ij} (\tilde{\theta}_j - \theta_j) = O_p(\sqrt{|\mathcal{N}_b|} \times n^{\alpha_s - 2\delta} + |\mathcal{N}_b| \times n^{-\eta - \alpha_s/2})
\]

For the fifth term, notice that

\[
E \left[ \frac{1}{n} \sum_{i=1}^{n} \sum_{j \notin \mathcal{N}_b} \epsilon_i x_{ij} \theta_j \right] = \frac{\sigma^2}{n} \sum_{i=1}^{n} \left( \sum_{j \notin \mathcal{N}_b} x_{ij} \theta_j \right)^2 \leq \frac{\sigma^2 C^2}{n} \sum_{j \notin \mathcal{N}_b} \theta_j^2 \Rightarrow \frac{1}{n} \sum_{i=1}^{n} \sum_{j \notin \mathcal{N}_b} \epsilon_i x_{ij} \theta_j = O_p(n^{-1 + \alpha_s/2})
\]

For the last term, notice that

\[
\frac{1}{n} \sum_{i=1}^{n} \left( \sum_{j \in \mathcal{N}_b} x_{ij} (\tilde{\theta}_j - \theta_j) \right) \times \left( \sum_{j \notin \mathcal{N}_b} x_{ij} \theta_j \right) \leq C^2 \sum_{j \notin \mathcal{N}_b} (\tilde{\theta}_j - \theta_j)^2 \times \sum_{j \notin \mathcal{N}_b} \theta_j^2 \]

\[
= O_p(\sqrt{|\mathcal{N}_b|} \times n^{\alpha_s - 2\delta} + |\mathcal{N}_b| \times n^{-\eta - \alpha_s/2})
\]

For any given 1 > \xi > 0, from (30), for sufficiently large n, \(\text{Prob}(\mathcal{N}_b \neq \mathcal{N}_b) < \xi/2\) and thus we have

\[
\hat{\sigma}^2 - \sigma^2 = O_p \left( \frac{1}{\sqrt{n}} + \sqrt{|\mathcal{N}_b|} \times n^{\alpha_s - 2\delta} + \sqrt{|\mathcal{N}_b|} \times n^{-\eta} + n^{-\alpha_s} \right)
\]

(87)

From assumption 2) and 6), we prove the second result. \[\square\]

**Proof of theorem 4** First from Cauchy inequality and assumption 2), suppose \(\delta = \frac{\eta + \alpha_s + \delta_1}{2}\) with \(\delta_1 > 0\), for \(i \in \mathcal{M}\),

\[
\left| \sum_{k=1}^{r} \frac{c_{ik} \zeta_k}{(\lambda_k^2 + \rho_n^2)} \right|^2 \leq \sum_{k=1}^{r} \frac{c_{ik}^2 \lambda_k^2}{(\lambda_k^2 + \rho_n^2)^2} \times \sum_{k=1}^{r} \frac{c_{ik}^2 \lambda_k^2}{(\lambda_k^2 + \rho_n^2)^2} \leq \tau_i \times ||\theta||^2 / \lambda_i^2
\]

\[
\Rightarrow \max_{i \in \mathcal{M}} \frac{\rho_n^2}{\tau_i} \sum_{k=1}^{r} \frac{c_{ik} \zeta_k}{(\lambda_k^2 + \rho_n^2)} = O(n^{-\delta_1})
\]

By defining \(t_{il} = \frac{1}{\tau_i} \times \sum_{k=1}^{r} c_{ik} p_{ik} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right) \) for \(i \in \mathcal{M}\) and \(l = 1, 2, \ldots, n\), from (5), (6), (77) and assumption 5),
if $\tilde{N}_b = N_b$, we have $\tilde{\tau}_i = \tau_i \geq 1/\sqrt{n}$ and there exists a constant $C > 0$, for any $a > 0$ and sufficiently large $n,$

$$
\max_{i=1,2,\ldots,p_1} | - \rho_i^2 \sum_{k=1}^{r} \frac{c_{ik} \zeta_k}{(\lambda_k^2 + \rho_n)^2} \sum_{k=1}^{r} \sum_{l=1}^{n} c_{ik} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right) p_{ik} \ell_i \geq - \sum_{j \in N_b} m_{ij} \theta_j - \sum_{p=1}^{p_1} m_{ij} \theta_{p,j} | \frac{\tilde{\gamma}_i - \gamma_i}{\tilde{\tau}_i} \\
\leq \max_{i \in M} \frac{\rho_i^2}{\tau_i} | \sum_{k=1}^{r} \frac{c_{ik} \zeta_k}{(\lambda_k^2 + \rho_n)^2} \sum_{k=1}^{r} \sum_{l=1}^{n} t_{il} \ell_i | + \max_{i \in M} \sum_{l=1}^{n} t_{il} \ell_i | - \max_{i \in M} | \sum_{j \in N_b} m_{ij} \theta_j | \frac{\tilde{\gamma}_i - \gamma_i}{\tilde{\tau}_i} + \max_{i=1,2,\ldots,p_1} | \sum_{p=1}^{p_1} m_{ij} \theta_{p,j} | \frac{\tilde{\gamma}_i - \gamma_i}{\tilde{\tau}_i} \\
\geq \max_{i \in M} \sum_{l=1}^{n} t_{il} \ell_i | - Cn^{-\delta_1} - \frac{a}{\sqrt{\log(n)}} (89)
$$

According to theorem [1] and lemma [1] there exists a constant $C$ and for any given $a > 0,$ for sufficiently large $n,$ for any $x \geq 0,$

$$
\text{Prob} \left( \max_{i=1,2,\ldots,p_1} \frac{| \tilde{\gamma}_i - \gamma_i |}{\tilde{\tau}_i} \leq x \right) \leq \text{Prob} \left( \max_{i=1,2,\ldots,p_1} \frac{| \tilde{\gamma}_i - \gamma_i |}{\tilde{\tau}_i} \leq x \cap \tilde{N}_b = N_b \right) + \text{Prob} (\tilde{N}_b \neq N_b) \\
\leq \text{Prob} \left( \max_{i \in M} \sum_{l=1}^{n} t_{il} \ell_i | \leq x + Cn^{-\delta_1} + \frac{a}{\sqrt{\log(n)}} \right) + Cn^{\alpha + m_{h_k} - m_{\eta}} \\
\leq \text{Prob} \left( \max_{i \in M} \sum_{l=1}^{n} t_{il} \ell_i^* | \leq x \right) + Cn^{\alpha + m_{h_k} - m_{\eta}} + \sup_{x \geq 0} \text{Prob} \left( \max_{i \in M} \sum_{l=1}^{n} t_{il} \ell_i | \leq x \right) - \text{Prob} \left( \max_{i \in M} \sum_{l=1}^{n} t_{il} \ell_i | \leq x \right) \\
+ \sup_{x \in \mathbb{R}} \left( \text{Prob} \left( \max_{i \in M} \sum_{l=1}^{n} t_{il} \ell_i^* | \leq x + Cn^{-\delta_1} + \frac{a}{\sqrt{\log(n)}} \right) - \text{Prob} \left( \max_{i \in M} \sum_{l=1}^{n} t_{il} \ell_i^* | \leq x \right) \right) \\
\text{Prob} \left( \max_{i=1,2,\ldots,p_1} \frac{| \tilde{\gamma}_i - \gamma_i |}{\tilde{\tau}_i} \leq x \right) \geq \text{Prob} \left( \max_{i=1,2,\ldots,p_1} \frac{| \tilde{\gamma}_i - \gamma_i |}{\tilde{\tau}_i} \leq x \cap \tilde{N}_b = N_b \right) (90) \\
\geq \text{Prob} \left( \max_{i \in M} \sum_{l=1}^{n} t_{il} \ell_i | \leq x - Cn^{-\delta_1} - \frac{a}{\sqrt{\log(n)}} \right) - \text{Prob} (\tilde{N}_b \neq N_b) \\
\geq \text{Prob} \left( \max_{i \in M} \sum_{l=1}^{n} t_{il} \ell_i | \leq x - Cn^{-\delta_1} - \frac{a}{\sqrt{\log(n)}} \right) - Cn^{\alpha + m_{h_k} - m_{\eta}} \\
- \text{Prob} \left( \max_{i \in M} \sum_{l=1}^{n} t_{il} \ell_i | \leq x \right) - \text{Prob} \left( \max_{i \in M} \sum_{l=1}^{n} t_{il} \ell_i | \leq x - Cn^{-\delta_1} - \frac{a}{\sqrt{\log(n)}} \right) \\
- \left| \text{Prob} \left( \max_{i \in M} \sum_{l=1}^{n} t_{il} \ell_i | \leq x - Cn^{-\delta_1} - \frac{a}{\sqrt{\log(n)}} \right) - \text{Prob} \left( \max_{i \in M} \sum_{l=1}^{n} t_{il} \ell_i | \leq x - Cn^{-\delta_1} - \frac{a}{\sqrt{\log(n)}} \right) \right| \\
\text{From assumption 1), 2), 5) and 7), for sufficiently large } n \text{ we have}
$$

$$
\max_{i \in M} \mathbb{E} \left( \sum_{l=1}^{n} t_{il} \ell_i^* \right)^2 = \sigma^2 \max_{i \in M} \sum_{l=1}^{n} \ell_i^2 = \sigma^2 \max_{i \in M} \sum_{k=1}^{r} c_{ik}^2 \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right)^2 \leq \sigma^2 (91)
$$

$$
\min_{i \in M} \mathbb{E} \left( \sum_{l=1}^{n} t_{il} \ell_i^* \right)^2 = \sigma^2 \min_{i \in M} \frac{1}{1 + \frac{1}{n \sum_{k=1}^{r} c_{ik}^2 \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right)^2} \geq \sigma^2 \min_{i \in M} \frac{1}{1 + \frac{1}{n \sum_{k=1}^{r} c_{ik}^2 \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right)^2} } \geq \frac{\sigma^2}{1 + 4\mathbb{E}^T} > 0
$$

and $(t_{il})_{i \in M, l=1,2,\ldots,n} = D_1 T D_2 P^T,$ here $D_1, T, D_2$ coincides with [9], we know that $(t_{il})_{i \in M, l=1,2,\ldots,n}$ has full rank(rank

28
rank covariance matrix is positive for all \( x \) and we prove the first result. For the second result, notice that the density of a multivariate normal random variable with full rank covariance matrix is positive for all \( x \). If \( x < Cn^{-\delta_1} \), combine with (90) to (94), we have

\[
\left| \log(Cn^{-\delta_1} + \frac{a}{\sqrt{\log(n)}}) \right| \leq \log\left(\frac{\sqrt{\log(n)}}{a}\right) = \frac{\log(\log(n))}{2} - \log(a) \leq \log(\log(n))
\]

\[
\Rightarrow \sup_{x \in \mathbb{R}} \left( \begin{array}{c}
\left| \log(Cn^{-\delta_1} + \frac{a}{\sqrt{\log(n)}}) \right| \\
\sup_{x \in \mathbb{R}} \left( \begin{array}{c}
\left| \log(Cn^{-\delta_1} + \frac{a}{\sqrt{\log(n)}}) \right| \\
\sup_{x \in \mathbb{R}} \left( \begin{array}{c}
\left| \log(Cn^{-\delta_1} + \frac{a}{\sqrt{\log(n)}}) \right|
\end{array} \right) \end{array} \right) \end{array} \right)
\]

\[
\leq C' \left( Cn^{-\delta_1} + \frac{a}{\sqrt{\log(n)}} \right) \times \left( 1 + \sqrt{\log(|\mathcal{M}|)} + \sqrt{\log(Cn^{-\delta_1} + \frac{a}{\sqrt{\log(n)}})} \right)
\]

For sufficiently large \( n \), we have \( Cn^{-\delta_1} + \frac{a}{\sqrt{\log(n)}} \leq 1 \) and correspondingly

\[
\left| \log(Cn^{-\delta_1} + \frac{a}{\sqrt{\log(n)}}) \right| \leq \log\left(\frac{\sqrt{\log(n)}}{a}\right) = \frac{\log(\log(n))}{2} - \log(a) \leq \log(\log(n))
\]

From assumption 7), (91) and lemma 3, we have

\[
\sup_{x \geq 0} |\mathbb{P}(\max_{i \in \mathcal{M}} \sum_{t=1}^{n} t_{il} \epsilon_i^* | \leq x) - \mathbb{P}(\max_{i \in \mathcal{M}} \sum_{t=1}^{n} t_{il} \epsilon_i^* | \leq x) | < a
\]

for sufficiently large \( n \). If \( x < Cn^{-\delta_1} + \frac{a}{\sqrt{\log(n)}} \), we have \( \mathbb{P}(\max_{i \in \mathcal{M}} |\sum_{t=1}^{n} t_{il} \epsilon_i^* | \leq x - Cn^{-\delta_1} - \frac{a}{\sqrt{\log(n)}}) = 0 \) and \( \mathbb{P}(\max_{i \in \mathcal{M}} |\sum_{t=1}^{n} t_{il} \epsilon_i^* | \leq x - Cn^{-\delta_1} - \frac{a}{\sqrt{\log(n)}}) = 0 \), combine with (90) to (94), we have

\[
\sup_{x \geq 0} |\mathbb{P}(\max_{i=1,2,\ldots,p_1} \frac{\gamma_i - \gamma_1}{\tau_i} \leq x) - \mathbb{P}(\max_{i \in \mathcal{M}} \sum_{t=1}^{n} t_{il} \epsilon_i^* | \leq x) | \leq Cn^{\alpha_p + m_p - m} + 6C' a + a
\]

and we prove the first result. For the second result, notice that the density of a multivariate normal random variable with full rank covariance matrix is positive for all \( x \in \mathbb{R}^{\mathcal{M}} \) and for any \( x \geq 0, \delta > 0 \), set \( \{ t = (i, i \in \mathcal{M}) | x < \max_{i=1,2,\ldots,|\mathcal{M}|} |t_i| \leq x + \delta \} \) has positive Lebesgue measure, thus \( H(x) \) is strictly increasing and for any \( 0 < \alpha < 1 \), \( H(c_{1-\alpha}) = 1 - \alpha \). According to the first result,

\[
\sup_{\alpha_0 \leq \alpha \leq \alpha_1} |\mathbb{P}(\max_{i=1,2,\ldots,p_1} \frac{\gamma_i - \gamma_1}{\tau_i} \leq c_{1-\alpha}) - (1 - \alpha)| \leq \sup_{x \geq 0} |\mathbb{P}(\max_{i=1,2,\ldots,p_1} \frac{\gamma_i - \gamma_1}{\tau_i} \leq x) - H(x) | \to 0
\]

as \( n \to \infty \), and we prove the second result.

\[\square\]

**Proof of theorem 5** First according to theorem 1, we have \( \mathbb{P}(\hat{\mathcal{N}}_{b_n} \neq \mathcal{N}_{b_n}) = O(n^{\alpha_p + m_p - m}) \) and if \( \hat{\mathcal{N}}_{b_n} = \mathcal{N}_{b_n} \), from (5) we have

\[
\|\hat{\theta}\|_2^2 = \sum_{i \in \hat{\mathcal{N}}_{b_n}} \hat{\theta}_i^2 \leq 3 \sum_{i \in \hat{\mathcal{N}}_{b_n}} |\theta_i|^2 + 3p_n^4 \sum_{i \in \hat{\mathcal{N}}_{b_n}} \left( \sum_{j=1}^{r} q_{ij} \hat{\epsilon}_j \right)^2 + 3 \sum_{i \in \hat{\mathcal{N}}_{b_n}} \left( \sum_{j=1}^{r} \sum_{l=1}^{n} q_{ij} \left( \frac{\lambda_j}{\lambda_i^2 + \rho_n} + \frac{\rho_n \lambda_j}{(\lambda_j^2 + \rho_n)^2} p_{ij} \epsilon_i \right) \right)^2
\]

(97)
From assumption 2), we have $\sum_{i \in \mathcal{N}_b} |\theta_i|^2 \leq \|\theta\|_2^2 = O(n^{2\alpha_e})$. Similarly we have

$$
\rho_n^2 \sum_{i \in \mathcal{N}_b} \left( \sum_{j=1}^r \frac{q_{ij} \tilde{\sigma}_j}{(\lambda_j^2 + \rho_n)^2} \right)^2 
\leq \rho_n^2 \sum_{i \in \mathcal{N}_b} \sum_{j=1}^r \frac{q_{ij} \tilde{\sigma}_j^2}{(\lambda_j^2 + \rho_n)^2} = \rho_n^2 \times |\mathcal{N}_b| \times \|\theta\|_2^2 = o(n^{-2\alpha_e})
$$

(98)

From assumption 6), we have

$$
E \sum_{i \in \mathcal{N}_b} \left( \sum_{j=1}^r \sum_{l=1}^n q_{ij} \left( \frac{\lambda_j}{\lambda_j^2 + \rho_n} + \frac{\rho_n \lambda_j}{(\lambda_j^2 + \rho_n)^2} \right) p_{lj} \theta_l \right)^2 
= \sigma^2 \sum_{i \in \mathcal{N}_b} \sum_{j=1}^r \sum_{l=1}^n q_{ij} \left( \frac{\lambda_j}{\lambda_j^2 + \rho_n} + \frac{\rho_n \lambda_j}{(\lambda_j^2 + \rho_n)^2} \right)^2 p_{lj}^2

\leq \frac{4\sigma^2 |\mathcal{N}_b|}{\lambda^2}
$$

(99)

$$
\Rightarrow \sum_{i \in \mathcal{N}_b} \left( \sum_{j=1}^r \sum_{l=1}^n q_{ij} \left( \frac{\lambda_j}{\lambda_j^2 + \rho_n} + \frac{\rho_n \lambda_j}{(\lambda_j^2 + \rho_n)^2} \right) p_{lj} \theta_l \right)^2 = o_p(n^{-2\alpha_e})
$$

Since $\alpha_\theta, \alpha_\sigma \geq 0$, we have $\|\hat{\theta}\|_2 = O_p(n^{\alpha_e})$ according to [1] and [5], define $\tilde{\zeta} = Q^T \hat{\theta}$, we have

$$
\tilde{\theta}^* - \hat{\theta} = \left( I_p + \rho_n Q (\Lambda^2 + \rho_n I_r)^{-1} Q^T \right) Q (\Lambda^2 + \rho_n I_r)^{-1} \left( \Lambda^2 Q^T \hat{\theta} + \Lambda^T \hat{\theta} \right) + \tilde{\theta} - Q Q^T \hat{\theta} - Q \tilde{\theta}
$$

$$
\Rightarrow \tilde{\theta}^*_i - \hat{\theta}_i = -\rho_n^2 \sum_{j=1}^r \frac{q_{ij} \tilde{\sigma}_j}{(\lambda_j^2 + \rho_n)^2} + \sum_{j=1}^r \sum_{l=1}^n q_{ij} \left( \frac{\lambda_j}{\lambda_j^2 + \rho_n} + \frac{\rho_n \lambda_j}{(\lambda_j^2 + \rho_n)^2} \right) p_{lj} \theta_l \epsilon_l^*
$$

(100)

Similar as (78), suppose $\delta$ in assumption 2) as $\delta = \frac{\gamma + \alpha_\sigma + \delta_1}{2}$ with $\delta_1 > 0$, we have

$$
\max_{i=1,2,...,p} |\rho_n^2 \sum_{j=1}^r \frac{q_{ij} \tilde{\sigma}_j}{(\lambda_j^2 + \rho_n)^2}| \leq \max_{i=1,2,...,p} \rho_n^2 \sqrt{\sum_{j=1}^r q_{ij}^2 (\lambda_j^2 + \rho_n)^2} \leq \rho_n^2 \|\hat{\theta}\|_2 \leq \frac{\rho_n^2 \|\hat{\theta}\|_2}{\lambda^2} = o_p(n^{-\alpha_\sigma})
$$

(101)

For $\epsilon^*_i |\epsilon^*_i, i = 1, 2, ..., n$ are normal random variables with mean 0 and variance $\tilde{\sigma}^2$, we have $E[\epsilon^*_i |\epsilon^*_i]^m = D$, here constant $D = E(Y |\epsilon^*_i)^m$ with $Y$ being normal random variable with mean 0 and variance 1. From [75] and lemma [4] there exists a constant $E$ which depends on $m$ and $D$ such that for any $a > 0$, if $\tilde{\sigma} > 0$,

$$
\text{Prob}^* \left( \max_{i=1,2,...,p} \frac{\sum_{j=1}^r \sum_{l=1}^n q_{ij} \left( \frac{\lambda_j}{\lambda_j^2 + \rho_n} + \frac{\rho_n \lambda_j}{(\lambda_j^2 + \rho_n)^2} \right) p_{lj} \theta_l \epsilon_l^*}{a \tilde{\sigma}} \right) \leq \frac{pE \tilde{\sigma}^m}{\lambda^2 \rho_n^m}
$$

(102)

If $\hat{\mathcal{N}}_b = \mathcal{N}_b$, $\frac{\gamma}{2} < \tilde{\sigma} < \frac{\gamma}{2}$ and $\max_{i=1,2,...,p} |\rho_n^2 \sum_{j=1}^r \frac{q_{ij} \tilde{\sigma}_j}{(\lambda_j^2 + \rho_n)^2}| \leq C \times n^{-\alpha_\sigma - \delta_1}$ for some constant $C$, since $\tilde{\theta}_i = 0$ if $i \not\in \hat{\mathcal{N}}_b$, we have

$$
\text{Prob}^* \left( \hat{\mathcal{N}}_b = \mathcal{N}_b \right) \leq \text{Prob}^* \left( \min_{i \in \mathcal{N}_b} |\tilde{\theta}_i| \leq b_n \right) + \text{Prob}^* \left( \max_{i \not\in \mathcal{N}_b} |\tilde{\theta}_i| > b_n \right)
$$

$$
\leq \text{Prob}^* \left( \min_{i \in \mathcal{N}_b} |\tilde{\theta}_i| - \max_{i \not\in \mathcal{N}_b} |\rho_n^2 \sum_{j=1}^r \frac{q_{ij} \tilde{\sigma}_j}{(\lambda_j^2 + \rho_n)^2}| \leq b_n \right) + \text{Prob}^* \left( \max_{i \not\in \mathcal{N}_b} |\tilde{\theta}_i| > b_n \right)

+ \text{Prob}^* \left( \max_{i \not\in \mathcal{N}_b} \frac{\sum_{j=1}^r q_{ij} \left( \lambda_j \right)}{(\lambda_j^2 + \rho_n)^2} \right) \geq b_n - \rho_n^2 \max_{i \not\in \mathcal{N}_b} \left( \sum_{j=1}^r \frac{q_{ij} \tilde{\sigma}_j}{(\lambda_j^2 + \rho_n)^2} \right)
$$

(103)
From assumption 4, we have for sufficiently large $n$,

$$b_n - \rho_n^2 \max_{i \in \mathcal{N}_b} \left| \sum_{j=1}^{r} q_{ij} \hat{\sigma}_j \right| \geq C \beta n^{-\alpha} - C n^{-\eta - \delta_1} \geq \frac{b_n}{2} \quad (104)$$

From (79), lemma 1 and assumption 1) and 4), we have

$$\max_{i=1,2,\ldots,p} \left| \sum_{j=1}^{r} q_{ij} \left( \frac{\lambda_j}{\lambda_j^2 + \rho_n} + \frac{\rho_n \lambda_j}{\lambda_j^2 + \rho_n^2} \right) \right| \frac{n}{p} \epsilon_i \epsilon_i^* = O_p \left( n^{\alpha - \eta / m - \eta} \right) \quad (105)$$

If there exists a constant $C$ such that $\max_{i=1,2,\ldots,p} \left| \sum_{j=1}^{r} q_{ij} \left( \frac{\lambda_j}{\lambda_j^2 + \rho_n} + \frac{\rho_n \lambda_j}{\lambda_j^2 + \rho_n^2} \right) \right| \leq C n^{\alpha - \eta / m - \eta}$ and (since $\frac{\rho_n^2 \epsilon_i^*}{\lambda^2} = O(n^{-\eta - \delta_1}$)), from assumption 4) we have for sufficiently large $n$,

$$\min_{i \in \mathcal{N}_b} \hat{\sigma}_i \geq \min_{i \in \mathcal{N}_b} |\sigma_i| - \max_{i \in \mathcal{N}_b} \rho_n^2 \lambda_i \sum_{j=1}^{r} q_{ij} \left( \frac{\lambda_j}{\lambda_j^2 + \rho_n} + \frac{\rho_n \lambda_j}{\lambda_j^2 + \rho_n^2} \right) \sum_{l=1}^{n} p_{lj} \epsilon_i \epsilon_i^* \geq \frac{b_n}{c_{tb}} - C n^{\alpha - \eta / m - \eta} \geq \frac{1}{c_{tb} - 1} b_n - C n^{-\eta - \delta_1} - C n^{\alpha - \eta / m - \eta} - C n^{-\eta - \delta_1} \geq \frac{b_n}{2} \left( \frac{1}{c_{tb} - 1} \right) \quad (106)$$

Correspondingly we have

$$\text{Prob}^* \left( \hat{\sigma}_i \neq \sigma_i \right) \leq \text{Prob}^* \left( \max_{i \in \mathcal{N}_b} \sum_{j=1}^{r} q_{ij} \left( \frac{\lambda_j}{\lambda_j^2 + \rho_n} + \frac{\rho_n \lambda_j}{\lambda_j^2 + \rho_n^2} \right) p_{lj} \epsilon_i^* \epsilon_i^* \geq \frac{b_n}{2} (1/c_{tb} - 1) \right)$$

$$+ \text{Prob}^* \left( \max_{i \in \mathcal{N}_b} \sum_{j=1}^{r} q_{ij} \left( \frac{\lambda_j}{\lambda_j^2 + \rho_n} + \frac{\rho_n \lambda_j}{\lambda_j^2 + \rho_n^2} \right) p_{lj} \epsilon_i^* \epsilon_i^* > \frac{b_n}{2} \right) \quad (107)$$

If $\hat{\sigma}_i = \sigma_i$, then $\hat{\sigma}_i = \sigma_i$ for $i = 1, 2, ..., p_1$ and similar with (88), we have

$$\max_{i \in \mathcal{M}} \left| \hat{\sigma}_i - \sigma_i \right| \leq \max_{i \in \mathcal{M}} \rho_n^2 \lambda_i \sum_{k=1}^{r} q_{ki} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{\lambda_k^2 + \rho_n^2} \right) p_{lk} \epsilon_l \epsilon_l^* \quad (108)$$

From theorem 1 for any $a > 0$, there exists constant $D_a$ such that $|\hat{\sigma}^2 - \sigma^2| \leq D_a n^{-\alpha / a} \sigma + \frac{1}{2} a < \hat{\sigma} < \frac{3}{2} a \sigma$ with probability $1 - a$, and thus we have

$$|\sigma - \hat{\sigma}| = \frac{|\sigma^2 - \hat{\sigma}^2|}{\sigma + \hat{\sigma}} \leq \frac{D_a n^{-\alpha / a}}{\sigma} \quad (109)$$
If $0 < x \leq n^{\alpha_{\sigma}/2}$, according to lemma 2, assumption 7), (9) and (91), there exists a constant $C'$ which only depends on $\sigma, c_M, C_\lambda$ such that

$$|	ext{Prob}^*\left(\max_{i \in M} \left| \sum_{l=1}^{n} \sum_{k=1}^{r} c_{lk} \left( \frac{\lambda_k}{\lambda_k + \rho_k} + \frac{\rho_k \lambda_k}{(\lambda_k + \rho_k)^2} \right) \right| \leq x - H(x) = |H(x) - H(x)|$$

$$\leq C' \left( 1 + \frac{1}{\log(|M|)} \right) n^{-\alpha_{\sigma}/2} + C' \sqrt{\frac{|x| \sigma - \sigma|}{\sigma}} \frac{1}{\log(\frac{|x| \sigma - \sigma|}{\sigma})} \times \sqrt{\frac{2D_a}{\sigma^2} n^{-\alpha_{\sigma}/4}}$$

(110)

For function $x \log(x)$ is continuous when $x > 0$ and $x \log(x) \to 0$ as $x \to 0$ and $\frac{x|\sigma - \sigma|}{\sigma} \leq \frac{2D_a n^{-\alpha_{\sigma}/2}}{\sigma^2}$, we know that $\sqrt{\frac{|x| \sigma - \sigma|}{\sigma}} \log(\frac{|x| \sigma - \sigma|}{\sigma}) \leq \sup_{x \in (0, 1]} \sqrt{x \log(x)} < \infty$ for sufficiently large $n$.

On the other hand, if $x > n^{\alpha_{\sigma}/2}$, then $\frac{x \sigma - \sigma|}{\sigma} > \frac{2n^{\alpha_{\sigma}/2}}{3}$, from lemma 3 we may choose sufficiently large $m_1$ such that $m_1 \alpha_{\sigma}/2 > 2$, since $E[\xi_1] = \alpha_{\sigma}/2$ (where $\xi_1$ is a normal random variable with mean 0 and variance $\sigma^2$) is a constant for given $m_1$ and $\max_{i \in M} \sum_{k=1}^{r} \frac{1}{\tau_i} c_{lk}^2 \left( \frac{\lambda_k}{\lambda_k + \rho_k} + \frac{\rho_k \lambda_k}{(\lambda_k + \rho_k)^2} \right) \leq 1$, we have

$$\text{Prob}\left( \max_{i \in M} \left| \sum_{l=1}^{n} \sum_{k=1}^{r} c_{lk} \left( \frac{\lambda_k}{\lambda_k + \rho_k} + \frac{\rho_k \lambda_k}{(\lambda_k + \rho_k)^2} \right) \xi_k \right| > 2n^{\alpha_{\sigma}/2} \right) \leq \frac{2D_a n^{-\alpha_{\sigma}/2}}{m_1} \times \frac{E}{2m_1 n^{\alpha_{\sigma}/2}}$$

(111)

Since $H(0) = 0$, combine with (110) and (111), we have for any given $a > 0$, for sufficiently large $n$,

$$\sup_{x \geq 0} |\text{Prob}^*\left( \max_{i \in M} \left| \sum_{l=1}^{n} \sum_{k=1}^{r} c_{lk} \left( \frac{\lambda_k}{\lambda_k + \rho_k} + \frac{\rho_k \lambda_k}{(\lambda_k + \rho_k)^2} \right) \xi_k \right| \leq x \right) - H(x)| < a$$

(112)

As a summary, for any given $a > 0$, there exists a constant $D_a$ such that for sufficiently large $n$, event $|\hat{\Theta}^2 - \sigma^2| \leq D_a n^{-\alpha_{\sigma}}$, $\frac{1}{2} \sigma < \hat{\theta} < \frac{3}{2} \sigma, \tilde{N}_{b_n} = N_{b_n}, \|\tilde{\theta}\|_2 \leq D_a \times n^{\alpha_{\sigma}} \Rightarrow \frac{2}{\lambda_1} \|\tilde{\theta}\|_2 \leq D_a n^{-\delta_1}$ for constant $D_a$ and $\max_{i=1, 2, \ldots, p} |\rho_k^2 \sum_{j=1}^{r} \frac{a_j \xi_j}{(\lambda_j + \rho_k)^2}| \leq D_a \times n^{-\alpha_{\sigma}/2}$ happen with probability $1 - a$ and correspondingly from (108), assumption 5) and lemma 2, we have for any
$x \geq 0$, there exists a constant $C'$ such that

$$\text{Prob}^*\left(\max_{i \in \mathcal{M}} \frac{1}{n^2} \sum_{k=1}^n c_{ik} \left(\frac{\lambda_k + \rho_n \lambda_k}{\tau_k} \right) p_{ik} t_i \leq x \right) - H(x) \leq \text{Prob}^*\left(\hat{N}_{b_n} \neq N_{b_n}\right) - \text{Prob}^*\left(\hat{N}_{b_n} = N_{b_n}\right) - H(x)$$

$$+ \text{Prob}^*\left(\max_{i \in \mathcal{M}} \frac{1}{n^2} \sum_{k=1}^n c_{ik} \frac{\lambda_k + \rho_n \lambda_k}{\tau_k} p_{ik} t_i \leq x \right) - H(x) \geq \text{Prob}^*\left(\max_{i \in \mathcal{M}} \frac{1}{n^2} \sum_{k=1}^n c_{ik} \frac{\lambda_k + \rho_n \lambda_k}{\tau_k} p_{ik} t_i \leq x \right) - H(x)$$

$$- \text{Prob}^*\left(\hat{N}_{b_n} \neq N_{b_n}\right) - C' D'_a (1 + \sqrt{\log(n)}) n^{-\delta_1} - C' \sqrt{D'_a n^{-\delta_1/2}} \sqrt{\log\left(\frac{\rho_n^2 \|\hat{\theta}\|_2}{\lambda_r^2}\right)} \frac{\rho_n^2 \|\hat{\theta}\|_2}{\lambda_r^2}$$

If $0 \leq x \leq \frac{\sigma^2 \|\hat{\theta}\|_2}{\lambda_r^2}$, then $\text{Prob}^*\left(\max_{i \in \mathcal{M}} \frac{1}{n^2} \sum_{k=1}^n c_{ik} \frac{\lambda_k + \rho_n \lambda_k}{\tau_k} p_{ik} t_i \leq x \right) - H(x) \geq \text{Prob}^*\left(\max_{i \in \mathcal{M}} \frac{1}{n^2} \sum_{k=1}^n c_{ik} \frac{\lambda_k + \rho_n \lambda_k}{\tau_k} p_{ik} t_i \leq x \right) - H(x)$$

$$- \text{Prob}^*\left(\hat{N}_{b_n} \neq N_{b_n}\right) - C' D'_a (1 + \sqrt{\log(n)}) n^{-\delta_1} - C' \sqrt{D'_a n^{-\delta_1/2}} \sqrt{\log\left(\frac{\rho_n^2 \|\hat{\theta}\|_2}{\lambda_r^2}\right)} \frac{\rho_n^2 \|\hat{\theta}\|_2}{\lambda_r^2}$$

$$\leq C_n^{m(\nu_h + \alpha_m/m-n)} + 2a$$

and we prove the first result.

For the second result, for any $a > 0$, from the first result, for sufficiently large $n$, we have

$$\text{Prob} \left(\sup_{x \geq 0} |\text{Prob}^*\left(\max_{i \in \mathcal{M}} \frac{1}{n^2} \sum_{k=1}^n c_{ik} \frac{\lambda_k + \rho_n \lambda_k}{\tau_k} p_{ik} t_i \leq x \right) - H(x)\right) \leq a$$

$$- (1 - \alpha - 2a) \leq 1 - \alpha + 2a < 1,$$
Therefore, combine with theorem \[2\] we have for sufficiently large \(n\),

\[
\begin{align*}
\Prb\left(\max_{i=1,2,\ldots,p_1} \frac{\hat{\gamma}_i - \gamma_i}{\hat{\tau}_i} \leq c_{1-a}\right) \\
\leq \Prb\left(\sup_{x \geq 0} |\Prb^*\left(\max_{i=1,2,\ldots,p_1} \frac{\hat{\gamma}_i - \gamma_i}{\hat{\tau}_i} \leq x\right) - H(x)| > a\right) \\
+ \Prb\left(\max_{i=1,2,\ldots,p_1} \frac{\hat{\gamma}_i - \gamma_i}{\hat{\tau}_i} \leq c_{1-a+2a}\right)
\end{align*}
\]

\[
\leq a + (H(c_{1-a+2a}) + a) = 1 - a + 4a
\]

\[
\begin{align*}
\Prb\left(\max_{i=1,2,\ldots,p_1} \frac{\hat{\gamma}_i - \gamma_i}{\hat{\tau}_i} \leq c_{1-a}\right) &\geq \Prb\left(\max_{i=1,2,\ldots,p_1} \frac{\hat{\gamma}_i - \gamma_i}{\hat{\tau}_i} \leq c_{1-a} \cap \sup_{x \geq 0} |\Prb^*\left(\max_{i=1,2,\ldots,p_1} \frac{\hat{\gamma}_i - \gamma_i}{\hat{\tau}_i} \leq x\right) - H(x)| \leq a\right) \\
&\geq \Prb\left(\sup_{x \geq 0} |\Prb^*\left(\max_{i=1,2,\ldots,p_1} \frac{\hat{\gamma}_i - \gamma_i}{\hat{\tau}_i} \leq x\right) - H(x)| > a\right)
\end{align*}
\]

\[
\Rightarrow |\Prb\left(\max_{i=1,2,\ldots,p_1} \frac{\hat{\gamma}_i - \gamma_i}{\hat{\tau}_i} \leq c_{1-a}\right) - (1 - a) | \leq 4a
\]  

(117)

For \(a > 0\) can be arbitrarily small, we prove the second result.

\[\square\]

**Proof of lemma \[3\]** First if \(\hat{\gamma}_1 = \hat{\gamma}_2 = \ldots = \hat{\gamma}_n\) we have for \(i = 1, 2, \ldots, n\), by defining \(\pi_j = \frac{1}{n} \sum_{i=1}^n x_{ij}\) and \(x'_{ij} = x_{ij} - \pi_j\),

\[
\tilde{c}_i = \epsilon_i + \sum_{j \in N_{\hat{\gamma}_1}} x_{ij} \theta_j - \sum_{j \in N_{\hat{\gamma}_n}} x_{ij} \theta_j - \theta_j \Rightarrow \tilde{c}_i = \epsilon_i - \frac{1}{n} \sum_{i=1}^n \epsilon_i + \sum_{j \in N_{\hat{\gamma}_1}} x'_{ij} \theta_j - \sum_{j \in N_{\hat{\gamma}_n}} x'_{ij} \theta_j - \theta_j
\]

(118)

For any \(x \in \mathbb{R}\), define \(\tilde{F}(x) = \frac{1}{n} \sum_{i=1}^n 1_{\epsilon_i \leq x}\), first from \[51\], for any given \(\psi > 0\), we have

\[
\begin{align*}
\tilde{F}(x) - F(x) &= \left(\tilde{F}(x) - \tilde{F}(x + 1/\psi)\right) + \left(\tilde{F}(x + 1/\psi) - F(x + 1/\psi)\right) + \left(F(x + 1/\psi) - F(x)\right) \\
&\leq \frac{1}{n} \sum_{i=1}^n \left(g_{\psi,x}(\epsilon_i) - g_{\psi,x}(\epsilon_i)\right) + \sup_{x \in \mathbb{R}} |\tilde{F}(x) - F(x)| + (F(x + 1/\psi) - F(x)) \\
&\leq g_\psi \sqrt{\frac{1}{n} \sum_{i=1}^n (\epsilon_i - \epsilon_i)^2 + \sup_{x \in \mathbb{R}} |\tilde{F}(x) - F(x)| + (F(x + 1/\psi) - F(x))}
\end{align*}
\]

\[
\begin{align*}
\tilde{F}(x) - F(x) &= \left(\tilde{F}(x) - \tilde{F}(x - 1/\psi)\right) + \left(\tilde{F}(x - 1/\psi) - F(x - 1/\psi)\right) - (F(x) - F(x - 1/\psi)) \\
&\geq \frac{1}{n} \sum_{i=1}^n \left(g_{\psi,x-1/\psi}(\epsilon_i) - g_{\psi,x-1/\psi}(\epsilon_i)\right) - \sup_{x \in \mathbb{R}} |\tilde{F}(x) - F(x)| - (F(x) - F(x - 1/\psi)) \\
&\geq -g_\psi \sqrt{\frac{1}{n} \sum_{i=1}^n (\epsilon_i - \epsilon_i)^2 - \sup_{x \in \mathbb{R}} |\tilde{F}(x) - F(x)| - (F(x) - F(x - 1/\psi))}
\end{align*}
\]

\[
\begin{align*}
\Rightarrow \sup_{x \in \mathbb{R}} |\tilde{F}(x) - F(x)| &\leq g_\psi \sqrt{\frac{1}{n} \sum_{i=1}^n (\epsilon_i - \epsilon_i)^2 + \sup_{x \in \mathbb{R}} |\tilde{F}(x) - F(x)| + \sup_{x \in \mathbb{R}} |F(x + 1/\psi) - F(x)|}
\end{align*}
\]

Since assumptions 1) to 6) are satisfied, from \[81\], \[82\], \[83\], and \(\frac{1}{n} \sum_{i=1}^n \epsilon_i = O_p(1/\sqrt{n})\), for any \(0 < a < 1\), there exists a
constant $C_a$ such that with probability at least $1 - a$, for any $n = 1, 2, \ldots$,

$$
\frac{1}{n} \sum_{i=1}^{n} (\epsilon_i - \epsilon_i^2) = \frac{1}{n} \sum_{i=1}^{n} \left( \sum_{j \in \mathcal{N}_n} x_{ij} \theta_j - \sum_{j \in \mathcal{N}_n} x'_{ij} (\tilde{\theta}_j - \theta_j) \right) - \frac{1}{n} \sum_{j=1}^{n} \epsilon_j \right)^2 
\leq \frac{3}{n} \sum_{i=1}^{n} \left( \sum_{j \in \mathcal{N}_n} x'_{ij} (\tilde{\theta}_j - \theta_j) \right)^2 + \frac{3}{n} \sum_{j=1}^{n} \left( \sum_{j \in \mathcal{N}_n} x_{ij} (\tilde{\theta}_j - \theta_j) \right)^2 + 3 \left( \frac{1}{n} \sum_{j=1}^{n} \epsilon_j \right)^2 
\leq \frac{6}{n} \sum_{i=1}^{n} \left( \sum_{j \in \mathcal{N}_n} x_{ij} \theta_j \right)^2 + 6 \left( \sum_{j \in \mathcal{N}_n} \tau_j \theta_j \right)^2 + \frac{6}{n} \sum_{j=1}^{n} \left( \sum_{j \in \mathcal{N}_n} x_{ij} (\tilde{\theta}_j - \theta_j) \right)^2 + \frac{6}{n} \sum_{j=1}^{n} \left( \sum_{j \in \mathcal{N}_n} x_{ij} (\tilde{\theta}_j - \theta_j) \right)^2 + \frac{6}{n} \sum_{j=1}^{n} \left( \sum_{j \in \mathcal{N}_n} x_{ij} (\tilde{\theta}_j - \theta_j) \right)^2 
\leq C_a n^{-\alpha_\sigma} + \frac{6}{n} \left( \sum_{i=1}^{n} \sum_{j \in \mathcal{N}_n} x_{ij} \theta_j \right)^2 + C_a |\mathcal{N}_b| (n^{2\alpha_\sigma - 4\delta} + n^{-\alpha_\sigma}) + \frac{6}{n} \sum_{i=1}^{n} \left( \sum_{j \in \mathcal{N}_n} x_{ij} (\tilde{\theta}_j - \theta_j) \right)^2 + \frac{6}{n} \sum_{j=1}^{n} \left( \sum_{j \in \mathcal{N}_n} x_{ij} (\tilde{\theta}_j - \theta_j) \right)^2 
\Rightarrow \frac{1}{n} \sum_{i=1}^{n} (\epsilon_i - \epsilon_i^2) = O_p(n^{-\alpha_\sigma/2})
$$

(120)

According to Gilvenko-Cantelli lemma, we have $\sup_{x \in \mathbb{R}} |\tilde{F}(x) - F(x)| \rightarrow 0$ almost surely. Thus, for any $a > 0$ and sufficiently large $n$, $\text{Prob} \left( \sup_{x \in \mathbb{R}} |\tilde{F}(x) - F(x)| \leq a \right) > 1 - a$, by choosing sufficiently small $a$ and $\psi = 1/a$, from assumption 8) and \cite{120}, we show that $\sup_{x \in \mathbb{R}} |\tilde{F}(x) - F(x)| \rightarrow 0$, as $n \rightarrow \infty$.

\textbf{Proof of theorem 4} First from theorem 1 since $p_1 = O(1)$, define $X_f = (x_{f,i,j})_{i=1,\ldots,n, j=1,\ldots,p}$, we have

$$
\max_{i=1,2,\ldots,p_1} \left| \sum_{j=1}^{p} x_{f,i,j} \tilde{\theta}_j - \sum_{j=1}^{p} x_{f,i,j} \beta_j \right| = O_p(n^{-\eta})
$$

(121)

Thus for any given $0 < a < 1$, we can choose a constant $C_a$ such that

$$
\text{Prob} \left( \max_{i=1,2,\ldots,p_1} \left| \sum_{j=1}^{p} x_{f,i,j} \tilde{\theta}_j - \sum_{j=1}^{p} x_{f,i,j} \beta_j \right| \leq C_a n^{-\eta} \right) \geq 1 - a
$$

(122)

for any $n = 1, 2, \ldots$. We define $F^-(x) = \lim_{y \downarrow x} F(y)$ for any $x \in \mathbb{R}$ and $G(x) = \text{Prob} \left( \max_{i=1,2,\ldots,p_1} |f_{i,i}| \leq x \right) = (F(x) - F^-(x))^{p_1}$ for $x \geq 0$, which is continuous if assumption 8) is satisfied. By assuming assumption 8), we have for any $x \geq 0$,

$$
\text{Prob} \left( \max_{i=1,2,\ldots,p_1} \left| y_{f,i} - \sum_{j=1}^{p} x_{f,i,j} \tilde{\theta}_j \right| \leq x \right) - G(x) \leq \text{Prob} \left( \max_{i=1,2,\ldots,p_1} |f_{i,i}| \leq x + \max_{i=1,2,\ldots,p_1} \left| \sum_{j=1}^{p} x_{f,i,j} (\beta_j - \tilde{\theta}_j) \right| \right) - G(x)
\leq a + \text{Prob} \left( \max_{i=1,2,\ldots,p_1} |f_{i,i}| \leq x + C_a n^{-\eta} \right) - G(x)
$$

$$
\text{Prob} \left( \max_{i=1,2,\ldots,p_1} \left| y_{f,i} - \sum_{j=1}^{p} x_{f,i,j} \tilde{\theta}_j \right| \leq x \right) - G(x) \geq \text{Prob} \left( \max_{i=1,2,\ldots,p_1} |f_{i,i}| \leq x - \max_{i=1,2,\ldots,p_1} \left| \sum_{j=1}^{p} x_{f,i,j} (\beta_j - \tilde{\theta}_j) \right| \right) - G(x)
\geq \text{Prob} \left( \max_{i=1,2,\ldots,p_1} |f_{i,i}| \leq x - C_a n^{-\eta} \right) - a - G(x)
$$

(123)
Since for any \( \delta > 0 \) and any \( x \geq 0 \), from assumption 8,
\[
G(x + \delta) - G(x) = \sum_{i=1}^{p_1} (F(x + \delta) - F(-x - \delta))^{-1} \times (F(x) - F(-x))^{p_1-1} \times (F(x + \delta) - F(-x - \delta) - F(x) + F(-x)) \\
\leq 2p_1 \sup_{x \in \mathbb{R}} (F(x + \delta) - F(x)) \Rightarrow \sup_{x \geq 0} (G(x + \delta) - G(x)) \leq 2p_1 \sup_{x \in \mathbb{R}} (F(x + \delta) - F(x))
\]
(124)

If \( x < C_a n^{-\eta} \), since \( G(0) = 0 \), we have
\[
G(x) - G(x - C_a n^{-\eta}) = G(x) \leq G(C_a n^{-\eta}) - G(0) \leq \sup_{x \geq 0} (G(x + C_a n^{-\eta}) - G(x))
\]
(125)

Combine with (123), we have for sufficiently large \( n \),
\[
\sup_{x \geq 0} \left| \left( \max_{i=1,2,\ldots,p_1} \left| y_{f,i} - \sum_{j=1}^{p} x_{f,ij} \tilde{\theta}_j \right| \right) - G(x) \right| \leq a + \sup_{x \geq 0} (G(x + C_a n^{-\eta}) - G(x)) \\
\leq a + 2p_1 \sup_{x \in \mathbb{R}} (F(x + C_a n^{-\eta}) - F(x)) \leq 2a
\]
(126)

Now we concentrate on bootstrap world. If \( \hat{N}_{b_n} = N_{b_n}, \sigma \leq \tilde{\sigma} \leq \frac{\sigma}{2} \), \( \| \tilde{\theta} \|_2 \leq C \times n^{\alpha} \),
\[
\max_{i=1,2,\ldots,p} \sum_{j=1}^{r} \rho_n q_{ij} \langle \tilde{\gamma}_j \rangle \leq C \times n^{-\eta - \delta_i}, \text{ and } \max_{i=1,2,\ldots,p} \sum_{j=1}^{r} \rho_n \langle \tilde{\gamma}_j \rangle \leq C n^{\alpha/m - \eta}
\]
(127)

for some constant \( C \), from (107) there exists a constant \( E \) such that
\[
\text{Prob}^* \left( \hat{N}_{b_n}^* \neq N_{b_n} \right) \leq \frac{Ep}{n^{m\eta} b_n^m}
\]
(128)

If \( \hat{N}_{b_n} = N_{b_n} \), we have
\[
\left| \sum_{j=1}^{p} x_{f,ij} \hat{\theta}_j - \sum_{j=1}^{p} x_{f,ij} \tilde{\theta}_j \right| = \left| \sum_{j \in N_{b_n}} x_{f,ij} (\hat{\theta}_j - \tilde{\theta}_j) \right| \leq \rho_n^2 \left| \sum_{k=1}^{r} \sum_{l=1}^{l} c_{ik} \tilde{\gamma}_k \right| + \left| \sum_{k=1}^{r} \sum_{l=1}^{l} c_{ik} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{\lambda_k^2 + \rho_n} \right) p_k \epsilon_l^1 \right| \\
\leq \rho_n^2 \sqrt{CM} \| \tilde{\theta} \|_2 + \left| \sum_{k=1}^{r} \sum_{l=1}^{l} c_{ik} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{\lambda_k^2 + \rho_n} \right) p_k \epsilon_l^1 \right|
\]
(129)

Form (78) and lemma 1 there is a constant \( E \) which only depends on \( m \) such that for any \( 1 > a > 0 \), by choosing sufficiently large \( C_a > 0 \),
\[
\text{Prob}^* \left( \max_{i=1,2,\ldots,p} \left| \sum_{k=1}^{r} \sum_{l=1}^{l} c_{ik} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{\lambda_k^2 + \rho_n} \right) p_k \epsilon_l^1 \right| > \frac{C_a n^{-\eta}}{\sigma} \right) \leq \frac{p_1 E \tilde{\sigma}^m}{n^{m\eta} C_a n^{-m\eta}} < a
\]
(130)
Thus, combine with (128), there exists a constant $C_a$ such that we have with conditional probability at least $1 - a$

$$\max_{i=1,2,...,p_1} |\sum_{j=1}^{p} \tilde{y}_{f,i,j} - \sum_{j=1}^{p} \tilde{y}_{j,i}| \leq C_a n^{-\eta}$$

$$\Rightarrow \text{Prob}^* \left( \max_{i=1,2,...,p_1} |y_{f,i} - \tilde{y}_{f,i,j}| \leq x \right) - G(x) \leq a + \text{Prob}^* \left( \max_{i=1,2,...,p_1} |\epsilon_{f,i}^*| \leq x + C_a n^{-\eta} \right) - G(x)$$

$$\leq a + \sup_{x \geq 0} \text{Prob}^* \left( \max_{i=1,2,...,p_1} |\epsilon_{f,i}^*| \leq x \right) - G(x) + 2p_1 \sup_{x \in \mathbb{R}} (F(x + C_a n^{-\eta}) - F(x))$$

$$\leq a + \sup_{x \geq 0} \text{Prob}^* \left( \max_{i=1,2,...,p_1} |\epsilon_{f,i}^*| \leq x \right) - G(x) + 2p_1 \sup_{x \in \mathbb{R}} (F(x + C_a n^{-\eta}) - F(x))$$

(131)

Since $G(x) = 0$ and $\text{Prob}^* \left( \max_{i=1,2,...,p_1} |\epsilon_{f,i}^*| \leq x \right) = 0$ if $x < 0$, we have

$$\sup_{x \geq 0} \text{Prob}^* \left( \max_{i=1,2,...,p_1} |y_{f,i} - \tilde{y}_{f,i,j}| \leq x \right) - G(x) \leq a + \sup_{x \geq 0} \text{Prob}^* \left( \max_{i=1,2,...,p_1} |\epsilon_{f,i}^*| \leq x \right) - G(x) + 2p_1 \sup_{x \in \mathbb{R}} (F(x + C_a n^{-\eta}) - F(x))$$

(132)

From lemma 4, we have for any $x \geq 0$,

$$|\text{Prob}^* \left( \max_{i=1,2,...,p_1} |\epsilon_{f,i}^*| \leq x \right) - G(x)| = \left| \left( \hat{F}(x) - \hat{F}^-(x) \right)^{p_1} - (F(x) - F(-x))^{p_1} \right|$$

$$\leq \sum_{i=1}^{p_1} |\hat{F}(x) - \hat{F}^-(x)|^{-1} \times |F(x) - F(-x)|^{p_1-i} \times \left( |\hat{F}(x) - F(x)| + |\hat{F}^-(x) - F^-(x)| \right)$$

(133)

$$\leq 2p_1 \sup_{x \in \mathbb{R}} |\hat{F}(x) - F(x)| \to p 0$$

as $n \to \infty$. Combine with (126), for any $1 > a > 0$, with probability at least $1 - a$ there exists a constant $C_a > 0$ such that for sufficiently large $n$, (127) happens and $\sup_{x \geq 0} |\text{Prob}^* \left( \max_{i=1,2,...,p_1} |\epsilon_{f,i}^*| \leq x \right) - G(x)| < a$, correspondingly for sufficiently large $n$,

$$\sup_{x \geq 0} |\text{Prob}^* \left( \max_{i=1,2,...,p_1} |y_{f,i} - \tilde{y}_{f,i,j}| \leq x \right) - \text{Prob} \left( \max_{i=1,2,...,p_1} |y_{f,i} - \tilde{y}_{f,i}| \leq x \right)|$$

$$\leq \sup_{x \geq 0} |\text{Prob}^* \left( \max_{i=1,2,...,p_1} |y_{f,i} - \tilde{y}_{f,i,j}| \leq x \right) - G(x)| + \sup_{x \geq 0} |\text{Prob} \left( \max_{i=1,2,...,p_1} |y_{f,i} - \tilde{y}_{f,i}| \leq x \right) - G(x)|$$

(134)

$$\leq a + \sup_{x \geq 0} |\text{Prob}^* \left( \max_{i=1,2,...,p_1} |\epsilon_{f,i}^*| \leq x \right) - G(x)| + 2p_1 \sup_{x \in \mathbb{R}} (F(x + C_a n^{-\eta}) - F(x)) + 2a \leq 5a$$

and we prove the first result.

For the second result, for given $0 < \alpha < 1$ and sufficiently small $a > 0$ such that $0 < 1 - 1 - a < 1 - \alpha + a < 1$, define $c_{1-a}$ as $1 - \alpha$ quantile of $G(x)$, for $G(x)$ is continuous, we know that $G(c_{1-a}) = 1 - \alpha$, from (134), for sufficiently large $n$, $\sup_{x \geq 0} |\text{Prob} \left( \max_{i=1,2,...,p_1} |y_{f,i} - \tilde{y}_{f,i}| \leq x \right) - G(x)| < a/2$ and with probability at least $1 - a$ we have
\[
\sup_{x \geq 0} \Pr^* \left( \max_{i=1,2,\ldots,p_1} |y^*_f,i - \tilde{y}^*_f,i| \leq x \right) - G(x) < a/2, \text{ correspondingly}
\]

\[
\Pr^* \left( \max_{i=1,2,\ldots,p} |y^*_f,i - \tilde{y}^*_f,i| \leq c_1 - \alpha + a \right) \geq 1 - \alpha + a/2 \Rightarrow c_1^* \leq c_1 - \alpha + a
\]

\[
\Pr^* \left( \max_{i=1,2,\ldots,p} |y^*_f,i - \tilde{y}^*_f,i| \leq c_1 - \alpha - a \right) \leq 1 - \alpha - a/2 \Rightarrow c_1^* \geq c_1 - \alpha - a
\]

\[
\Rightarrow \Pr \left( \max_{i=1,2,\ldots,p} |y_f,i - \tilde{y}_f,i| \leq c_1 - \alpha \right)
\]

\[
\leq \Pr \left( \max_{i=1,2,\ldots,p} |y_f,i - \tilde{y}_f,i| \leq c_1 - \alpha + a \cap \sup_{x \geq 0} \Pr^* \left( \max_{i=1,2,\ldots,p_1} |y^*_f,i - \tilde{y}^*_f,i| \leq x \right) - G(x) < a/2 \right) + a
\]

\[
\leq \Pr \left( \max_{i=1,2,\ldots,p} |y_f,i - \tilde{y}_f,i| \leq c_1 - \alpha + a \right) - G(c_1 - \alpha + a) + a \leq 1 - \alpha + 3a
\]

\[
\geq \Pr \left( \max_{i=1,2,\ldots,p} |y_f,i - \tilde{y}_f,i| \leq c_1 - \alpha - a \cap \sup_{x \geq 0} \Pr^* \left( \max_{i=1,2,\ldots,p} |y^*_f,i - \tilde{y}^*_f,i| \leq x \right) - G(x) < a/2 \right)
\]

\[
\geq \Pr \left( \max_{i=1,2,\ldots,p} |y_f,i - \tilde{y}_f,i| \leq c_1 - \alpha - a \right) - a
\]

\[
\geq -\Pr \left( \max_{i=1,2,\ldots,p} |y_f,i - \tilde{y}_f,i| \leq c_1 - \alpha - a \right) - G(c_1 - \alpha - a) + G(c_1 - \alpha - a) - a \geq 1 - \alpha - 3a
\]

for \(a > 0\) can be arbitrarily small, we prove the second result.