RIDGE REGRESSION REVISITED:
DEBIASING, THRESHOLDING AND BOOTSTRAP

BY YUNYI ZHANG 1, DIMITRIS N. POLITIS 2

1Department of Mathematics, Univ. of California–San Diego yuz334@ucsd.edu
2Department of Mathematics and Halicioglu Data Science Institute, Univ. of California–San Diego dpolitis@ucsd.edu

The success of the Lasso in the era of high-dimensional data can be attributed to its conducting an implicit model selection, i.e., zeroing out regression coefficients that are not significant. By contrast, classical ridge regression can not reveal a potential sparsity of parameters, and may also introduce a large bias under the high-dimensional setting. Nevertheless, recent work on the Lasso involves debiasing and thresholding, the latter in order to further enhance the model selection. As a consequence, ridge regression may be worth another look since –after debiasing and thresholding-- it may offer some advantages over the Lasso, e.g., it can be easily computed using a closed-form expression. In this paper, we define a debiased and thresholded ridge regression method, and prove a consistency result and a Gaussian approximation theorem. We further introduce a wild bootstrap algorithm to construct confidence regions and perform hypothesis testing for a linear combination of parameters. In addition to estimation, we consider the problem of prediction, and present a novel, hybrid bootstrap algorithm tailored for prediction intervals. Extensive numerical simulations further show that the debiased and thresholded ridge regression has favorable finite sample performance and may be preferable in some settings.

1. Introduction. Linear regression is a fundamental topic in statistical inference. The classical setting assumes the dimension of parameters in a linear model is constant. However, in the modern era, observations may have a comparable or even larger dimension than the number of samples. To perform a consistent estimation with high-dimensional data, statisticians often assume the underlying parameters are sparse (i.e., the parameter vector contains lots of zeros), and proceed with statistical inference based on this assumption.

The success of the Lasso in the setting of high-dimensional data can be attributed to its conducting an implicit model selection, i.e., zeroing out regression coefficients that are not significant; see Tibshirani (1996). More recent work includes: Meinshausen and Bühlmann (2006), Meinshausen and Yu (2009), and van de Geer (2008) for the Lasso estimator’s (model-selection) consistency and applications; Chatterjee and Lahiri (2010, 2011), Zhang and Cheng (2017), and Dezeure, Bühlmann and Zhang (2017) for confidence interval construction and hypothesis testing; and Javanmard and Montanari (2018), Fan and Li (2001), and Chen and Zhou (2020) for improvements of the Lasso estimator. Although the Lasso has the desirable property of zeroing out some regression coefficients, van de Geer, Bühlmann and Zhou (2011) proposed to further threshold the estimated coefficients, leading to a sparser fitted model. Furthermore, Bühlmann and van de Geer (2011), and Dezeure, Bühlmann and Zhang (2017), proposed to debias the Lasso in constructing confidence intervals; see van de Geer (2019) and Javanmard and Javadi (2019) for recent works on debiased Lasso.

Keywords and phrases: Gaussian approximation, high-dimensional data, Lasso, prediction, regression, resampling.
An alternative approach providing consistent estimators for a high dimensional linear model is the so-called post-selection inference. It first applies Lasso to select influential parameters, then fits an ordinary least squares regression on the selected parameters; see Lee et al. (2016), Liu and Yu (2013), and Tibshirani et al. (2018). We refer to Bühlmann and van de Geer (2011) for a comprehensive overview of the Lasso method for high dimensional data.

Ridge regression is a classical method, and its estimator has a closed-form expression, making statistical inference easier than Lasso. However, there is relatively little research on the ridge regression under the high-dimensional setting. Shao and Deng (2012) proposed a broken adaptive ridge estimator to approximate $L_0$ penalized regression. Dai et al. (2018) introduced the ridge regression under the high-dimensional setting. Shao and Deng (2012) proposed a broken adaptive ridge estimator to approximate $L_0$ penalized regression. Dai et al. (2018) introduced a threshold ridge regression method and proved its consistency.

Three issues have prevented ridge regression from being suitable for a high dimensional linear model:

1. **The ridge regression cannot preserve/recover sparsity.** Typically, a ridge regression estimator of the parameter vector will not contain any zeros, even though the parameters may be sparse.

2. **Bias in the ridge regression estimator can be large.** To illustrate this, suppose the parameter of interest is $a^T\beta$ in a linear model $y = X\beta + \epsilon$; here, the dimension $p \leq n$ (the sample size), $X$ has rank $p$, and $a$ is a known vector. The ridge estimator is $a^T\tilde{\theta}^*$ with $\tilde{\theta}^* = (X^TX + \rho_n I_p)^{-1}X^Ty$, for some $\rho_n > 0$, with $I_p$ denoting the $p$-dimensional identity matrix. Performing a thin singular value decomposition $X = P\Lambda Q^T$ (as in Theorem 7.3.2 in Horn and Johnson (2012)), and assuming the error vector $\epsilon$ consists of independent identically distributed (i.i.d.) components, the bias and the standard deviation can be calculated (and controlled) as follows:

   \[ Ea^T\theta^* - a^T\beta = -\rho_n a^TQ(\Lambda^2 + \rho_n I_p)^{-1}Q^T\beta \]

   \[ \text{and } \sqrt{\text{Var}(a^T\theta^*)} = \sqrt{\text{Var}(\epsilon_1) \times a^TQ(\Lambda^2 + \rho_n I_p)^{-2}\Lambda^2Q^T} \leq \frac{\sqrt{\text{Var}(\epsilon_1)} \times \|a\|_2}{\lambda_p}. \]

   In the above, $\lambda_p$ is the smallest singular value of $X$, and $\| \cdot \|_2$ is the Euclidean norm of a vector. If $\|\beta\|_2$ does not have a bounded order, the bias may tend to infinity. Another critical problem is that the absolute value of the bias can be significantly larger than the standard deviation, which makes constructing confidence intervals difficult.

3. **When the dimension of parameters is larger than the sample size, ridge regression estimates the projection of parameters on the linear space spanned by rows of $X$** (Shao and Deng (2012)). The projection (which can now be considered to be the ‘parameters’ of the linear model) is not sparse, bringing extra burdens for statistical inference.

The third issue comes from the nature of ridge regression, and it is not necessarily bad; our section 6 provides an example to illustrate this. The first two issues can be solved by thresholding and debiasing respectively, yielding an improved ridge regression that will be the focus of this paper. If the Lasso is in need of thresholding and debiasing –as van de Geer, Bühlmann and Zhou (2011), Dezeure, Bühlmann and Zhang (2017), and Bühlmann and van de Geer (2011) seem to suggest– then it loses some of its attractiveness, in which case (improved) ridge regression may be worth another look. If (improved) ridge regression turns out to have comparable performance to threshold Lasso, then the former would be preferable since it can be easily computed using a closed-form expression. Indeed, numerical simulations in section 6 indicate that improved ridge regression has favorable finite-sample performance.
performance, and has a further advantage over the Lasso: it is robust against a non-optimal choice of the hyperparameters.

Apart from point estimation using improved ridge regression, this paper presents a Gaussian approximation theorem for the improved ridge regression estimator. Applying this result, we propose a wild bootstrap algorithm to construct a confidence region for $\gamma = M\beta$ with $M$ a known matrix and/or test the null hypothesis $\gamma = \gamma_0$ with $\gamma_0$ a known vector, versus the alternative hypothesis $\gamma \neq \gamma_0$. The wild bootstrap was developed in the 1980s by Wu (1986) and Liu (1988); its applicability to high-dimensional problems was recognized early on by Mammen (1993). Here we will use the wild bootstrap in its Gaussian residuals version that has been found useful in high-dimensional regression; see Chernozhukov, Chetverikov and Kato (2013). Estimating and testing $\gamma$ are important problems in econometrics, e.g., Dolado and Lütkepohl (1996), Sun (2011, 2013), and Gonçalves and Vogelsang (2011). Besides, estimating $\gamma$ directly contributes to prediction, which is an important topic in modern age statistics.

Finally, we consider statistical prediction based on the improved ridge regression estimator for a high-dimensional linear model. For a regression problem, quantifying a predictor’s accuracy can be as important as predicting accurately. To do that, it is useful to be able to construct a prediction interval to accompany the point prediction; this is usually done by some form of bootstrap; see Stine (1985) for a classical result, and Politis (2015) for a comprehensive treatment of both model-based and model-free prediction intervals in regression. As an alternative to the bootstrap, conformal prediction may be a tool to yield prediction intervals; see e.g. Romano, Patterson and Candès (2019) and Romano, Sesia and Candès (2020). In our point of view, however, the bootstrap is preferable as it captures the underlying variability of estimated quantities; Section 5 in what follows gives the details.

The remainder of this paper is organized as follows: Section 2 introduces frequently used notations and assumptions. Section 3 presents the consistency result and the Gaussian approximation theorem for the improved ridge regression estimator. Section 4 constructs a confidence region for $\gamma = M\beta$, and tests the null hypothesis $\gamma = \gamma_0$ versus the alternative hypothesis $\gamma \neq \gamma_0$ via a bootstrap algorithm. Section 5 constructs bootstrap prediction intervals in our ridge regression setting using a novel, hybrid resampling procedure. Finally, Section 6 provides extensive simulations to illustrate the finite sample performance, while Section 7 gives some concluding remarks; technical proofs are deferred to the Appendix.

2. Preliminaries. Our work focuses on the fixed design linear model

$$y = X\beta + \epsilon$$

(2)

where the (unknown) parameter vector $\beta$ is $p$-dimensional, and the $n \times p$ fixed (nonrandom) design matrix $X$ is assumed to have rank $r$. The error vector $\epsilon$ has mean zero and satisfies assumptions to be specified later.

Define the known matrix of linear combination coefficients as $M = (m_{ij})_{i=1,2,\ldots,p_1;j=1,2,\ldots,p}$ so that $M$ has $p_1$ rows. The linear combination of interest are $\gamma = (\gamma_1, \ldots, \gamma_{p_1})^T = M\beta$.

Perform a thin singular value decomposition $X = P\Lambda Q^T$ as in Theorem 7.3.2 in Horn and Johnson (2012); here, $P$ and $Q$ respectively is $n \times r$ and $p \times r$ orthonormal matrices that satisfy $P^TP = Q^TQ = I_r$, where $I_r$ denotes the $r \times r$ identity matrix. Furthermore, $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_r)$, and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0$ are positive singular values of $X$.

Denote $Q_\perp$ as the $p \times (p - r)$ orthonormal complement of $Q$; then we have

$$Q^T\perp Q_\perp = I_{p-r}, \quad Q^TQ_\perp = 0, \quad \text{and } QQ^T + Q_\perp Q^T_\perp = I_p.$$  

(3)

in the above, $0$ is the $r \times (p - r)$ matrix having all elements $0$. Define $\zeta = Q^T\beta$ and $\theta = (\theta_1, \ldots, \theta_p)^T = Q\zeta = QQ^T\beta$, then $X\beta = X\theta$, $\beta^T\theta = \zeta^TQ^TQ\zeta = \zeta^T\zeta$. According to Shao and Deng (2012), the ridge regression estimates $\theta$ rather than $\beta$. 

Define $\theta_\perp = Q_\perp Q_\perp^T \beta$, so $\beta = \theta + \theta_\perp$. If the design matrix $X$ has rank $p \leq n$, then $Q_\perp$ does not exist. In this situation, we define $\theta_\perp = 0$, the $p$ dimensional vector with all elements 0. For a threshold $b_n$, define the set $\mathcal{N}_{b_n} = \{i \mid |\theta_i| > b_n\}$. After selecting a suitable $b_n$, define

$$c_{ik} = \sum_{j \in \mathcal{N}_{b_n}} m_{ij} q_{jk}, \forall i = 1, 2, \ldots, p, \ k = 1, 2, \ldots, r,$$ and $\mathcal{M} = \{i \mid \sum_{k=1}^{r} c_{ik}^2 > 0\}$

Define $\tau_i$, $i = 1, 2, \ldots, p_1$ as

$$\tau_i = \sqrt{\sum_{k=1}^{r} c_{ik}^2 \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right)^2 + \frac{1}{n}}$$

In section 3 and (88) to (90) in appendix B, we show that the estimation error $\widehat{\gamma} - \gamma$ (see (17) for the definition of $\widehat{\gamma}$) asymptotically can be approximated by the random vector

$$\left( \sum_{l=1}^{n} \sum_{k=1}^{r} c_{lk} \rho P \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right) \epsilon_l, \ldots, \sum_{l=1}^{n} \sum_{k=1}^{r} c_{lk} \rho P \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right) \epsilon_l \right)^T$$

where $P = (p_{lk})_{l=1, \ldots, n; k=1, \ldots, r}$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$. Moreover, if we assume that $\epsilon_i, i = 1, \ldots, n$ are i.i.d. with mean 0 and variance 1, then

$$V a r \left( \sum_{l=1}^{n} \sum_{k=1}^{r} c_{lk} \rho P \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right) \epsilon_l \right) = \sum_{l=1}^{n} \left( \sum_{k=1}^{r} c_{lk}^2 \rho P \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right)^2 \right)^2 = \sum_{k=1}^{r} c_{ik}^2 \rho P \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right)^2$$

In section 3, we will estimate $\tau_i$ by $\widehat{\tau}_i$ (defined in (25)) and use $\widehat{\tau}_i$ to normalize the estimation error. The extra $1/n$ in (5) is introduced to assure $\tau_i > 0$.

We will use the standard order notations $O(\cdot)$, $o(\cdot)$, $O_p(\cdot)$, and $o_p(\cdot)$. For two numerical sequences $a_n, b_n, n = 1, 2, \ldots$, we say $a_n = O(b_n)$ if $\exists$ a constant $C > 0$ such that $|a_n| \leq C b_n$ for all $n$, and $a_n = o(b_n)$ if $\lim_{n \to \infty} a_n/b_n = 0$. For two random variable sequences $X_n, Y_n$, we say $X_n = O_p(Y_n)$ if for any $0 < \epsilon < 1, \exists$ a constant $C_{\epsilon} > 0$ such that $\sup_n Prob(\{|X_n| \geq C_{\epsilon} \|Y_n\|) \leq \epsilon$; and $X_n = o_p(Y_n)$ if $X_n \to_p 0$; see e.g. Definition 1.9 and Chapter 1.5.1 of Shao (2003). All order notations and convergences in this paper will be understood as the sample size $n \to \infty$. For a vector $a = (a_1, \ldots, a_n)^T$ and a fixed number $q \geq 1$, define $||a||_q = (\sum_{i=1}^{n} |a_i|^q)^{1/q}$. For a finite set $A$, $|A|$ denotes the number of elements in $A$. Notations $\exists$ and $\forall$ denote “there exists” and “for all” respectively. $Prob(\cdot)$ and $E^*$ respectively represent probability and expectation in the “bootstrap world”, i.e., they are the conditional probability $Prob(\cdot | y)$ and the conditional expectation $E(\cdot | y)$.

Suppose $H(x)$ is a cumulative distribution function and $0 < \alpha < 1$; then the $1 - \alpha$ quantile of $H$ is defined as

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

In particular, given some order statistics $X_1 \leq X_2 \leq \ldots \leq X_B$, the $1 - \alpha$ sample quantile $C_{1-\alpha}$ is defined as

$$C_{1-\alpha} = X_{i_*} \text{ such that } i_* = \min \left\{i \mid \frac{1}{B} \sum_{j=1}^{B} 1_{X_j \leq X_i \geq 1 - \alpha} \right\}.$$
Other notations will be defined before being used. Without being explicitly specified, the convergence results in this paper assume the sample size \( n \to \infty \).

The high dimensionality in this paper comes from two aspects: the number of parameters \( p \) may increase with the sample size \( n \), and (for statistical inference/hypothesis testing) the number of simultaneous linear combinations \( p_1 \) and \( |\mathcal{M}| \) can also increase with \( n \).

Our work adopts the following assumptions:

**Assumptions**

1. Assume a fixed design, i.e., the design matrix \( X \) is deterministic. Also assume that there exists constants \( c_\lambda, C_\lambda > 0, 0 < \eta \leq 1/2 \), such that the positive singular values of \( X \) satisfy
   \[
   C_\lambda \sqrt{n} \geq \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r \geq c_\lambda n^\eta.
   \]

Furthermore, the Euclidean norm of \( \theta \) is assumed to satisfy \( \|\theta\|_2 = \sqrt{\sum_{i=1}^p \theta_i^2} = O(n^{\alpha_\theta}) \) with \( 0 < \alpha_\theta < 3\eta \).

2. The ridge parameter satisfies \( \rho_n = O(n^{2\eta-\delta}) \) with a positive constant \( \delta \) such that \( \frac{\eta+\alpha_\theta}{2} < \delta < 2\eta \).

3. The errors \( \epsilon = (\epsilon_1, \ldots, \epsilon_n)^T \) driving regression (2) are assumed to be i.i.d., with \( \mathbb{E}\epsilon_1 = 0 \), and \( \mathbb{E}|\epsilon_1|^{m} < \infty \) for some \( m > 4 \).

4. The dimension of the parameter vector \( \beta \) satisfies \( p = O(n^\alpha) \) for some constant \( \alpha_p \in [0, m\eta) \) where \( m, \eta \) are as defined in Assumptions 1–3. Furthermore, the threshold \( b_n \) is chosen as \( b_n = C_b \times n^{-\nu_b} \) with constants \( C_b, \nu_b > 0 \) and \( \nu_b + \frac{\alpha_\lambda}{m} - \eta < 0 \). We assume \( \exists \) a constant \( 0 < c_b < 1 \) such that \( \max_{i \notin \mathcal{N}_{b_n}} |\theta_i| \leq c_b \times b_n \), and \( \min_{i \in \mathcal{N}_{b_n}} |\theta_i| \geq \frac{b_n}{c_b} \).

The intuitive meaning of assumption 4 is that the \( \theta_i \)'s that are not being truncated should be significantly larger than the \( \theta_i \)'s that are being truncated.

5. \( \mathcal{M} \) (defined in (4)) is non-empty and \( |\mathcal{M}| = O(n^{\alpha_M}) \) with \( \alpha_M < m\eta \) where \( m, \eta \) are as defined in Assumptions 1–3. Besides, assume \( \exists \) constants \( c_M, C_M \) such that \( 0 < c_M < \sum_{k=1}^r c_{ik}^2 \leq C_M \) for all \( i \in \mathcal{M} \). Also assume
   \[
   \max_{i=1,2,\ldots,p_1} \sum_{j \in \mathcal{N}_{b_n}} m_{ij} \theta_j = o \left( \frac{1}{\sqrt{n \log(n)}} \right) \quad \text{and} \quad \max_{i=1,2,\ldots,p_1} \sum_{j=1}^p m_{ij} \theta_{i,j} = o \left( \frac{1}{\sqrt{n \log(n)}} \right)
   \]

We assume (11) to maintain the sparsity of \( \theta \) and assure that the projection bias \( \beta - \theta \) is negligible compared to the stochastic errors. It allows an inexact sparsity, i.e., some \( \theta_i \) may not equal 0 even if \( i \notin \mathcal{N}_{b_n} \). Theoretical results for other linear regression estimators (e.g., Lasso) need an exact sparse assumption (\( \theta_i = 0 \) for all \( i \notin \mathcal{N}_{b_n} \)), see Zhao and Yu (2006) and Basu and Michailidis (2015) for a further introduction.

6. \( \exists \) a constant \( \alpha_\sigma \) satisfying \( \eta \geq \alpha_\sigma > 0 \) such that
   \[
   n^{-\nu_b} \sum_{j \notin \mathcal{N}_{b_n}} |\theta_j| = O(n^{-\alpha_\sigma}), \quad \frac{\sqrt{|\mathcal{N}_{b_n}|}}{n^{\eta}} = O(n^{-\alpha_\sigma})
   \]

7. \( |\mathcal{M}| \leq r \), the matrix \( T = (c_{ik})_{i \in \mathcal{M}, k=1,2,\ldots,r} \) has rank \( |\mathcal{M}| \), and one of the two following conditions holds true:

   **7.1.**

   \[
   \max_{i \in \mathcal{M}, l=1,2,\ldots,n} \left| \frac{1}{\tau_k} \sum_{k=1}^r c_{ikl} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right) \right| = o(\min((n^{1/2} \log^{-3/2}(n), n^{-1/3} \log^{-3/2}(n)))
   \]
7.2. $\alpha_{\sigma} < 1/2$ and
\[ |\mathcal{M}| = o(n^{\alpha_{\sigma} \times \log^{-3}(n)}) \]

According to (6), the normalized estimation error $\frac{\tilde{\eta} - \eta}{\tilde{\tau}}$ asymptotically will be approximated by $\sum_{i=1}^{n} \left( \frac{1}{\tau_i} \sum_{k=1}^{r} c_{ik} p_{tk} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right) \right) \epsilon_i$. Therefore, the intuitive meaning of assumption 7 is that all terms $\frac{1}{\tau_i} \sum_{k=1}^{r} c_{ik} p_{tk} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right) \epsilon_i$ in the summation are negligible and the number of simultaneous linear combinations $|\mathcal{M}|$ cannot be too large.

Recall that the paper at hand focuses on fixed design regression, i.e., no randomness involves in the design matrix $X$. However, all results of this paper still hold true in the case of random design after conditioning on $X$, as long as $X$ can be assumed independent of the error vector $\epsilon$. In this case, to interpret the results we would need to replace $\text{Prob}(\cdot)$ by $\text{Prob}(\cdot|X)$, $\mathbb{E}$ by $\mathbb{E} \cdot |X$, $\text{Prob}^\ast(\cdot)$ by $\text{Prob}(\cdot|X,y)$ and $\mathbb{E}^\ast \cdot$ by $\mathbb{E} \cdot |X,y$.

**Remark 1.** We do not require that the design matrix has rank $\min(n, p)$ or that $p < n$. However, when these conditions are not satisfied, the sparsity of $\theta$, i.e., assumption 5 and 6, can be violated. Section 6 uses a numerical simulation to illustrate this problem.

Example 1 below provides an instance in which assumption 1 is satisfied.

**Example 1.** Suppose $n > p$ and $\lim_{n \to \infty} p/n = c \in (0, 1)$. Choose $X = (x_{ij})_{i=1,...,n,j=1,...,p}$ such that the $x_{ij}$ are a realization of i.i.d. random variables with mean 0, variance 1, and finite fourth order moment. According to Bai and Yin (1993), the smallest eigenvalue of $\frac{1}{n} X^T X$ would then converge to $(1 - \sqrt{\alpha})^2$ almost surely as $n \to \infty$. So the smallest singular value of $X$ (which is the smallest eigenvalue of the square root of $X^T X$) is greater than $\frac{1 - \sqrt{\alpha}}{2} \sqrt{n}$ for sufficiently large $n$, almost surely. On the other hand, the largest eigenvalue of $\frac{1}{n} X^T X$ converges to $(1 + \sqrt{\alpha})^2$ as $n \to \infty$. Hence, the largest singular value of $X$ also has order $O(\sqrt{n})$ almost surely.

### 3. Consistency and the Gaussian approximation theorem.
Throughout, we will use the notations developed in section 2. For a chosen ridge parameter $\rho_n > 0$, define the classical ridge regression estimator $\bar{\theta}^\ast$ and the de-biased estimator $\tilde{\theta}$ as
\[
\bar{\theta}^\ast = (X^T X + \rho_n I_p)^{-1} X^T y
\]
\[
\tilde{\theta} = (\bar{\theta}_1, ..., \bar{\theta}_p)^T = \bar{\theta}^\ast + \rho_n \times Q (\Lambda^2 + \rho_n I_r)^{-1} Q^T \bar{\theta}^\ast
\]

Then we have
\[
\tilde{\theta} - \theta = -\rho_n^2 Q (\Lambda^2 + \rho_n I_r)^{-2} \zeta + Q (\Lambda^2 + \rho_n I_r)^{-1} \Lambda + \rho_n (\Lambda^2 + \rho_n I_r)^{-2} \Lambda) P^T \epsilon
\]

Similar to $\mathcal{N}_{b_n}$, define the set $\tilde{\mathcal{N}}_{\tilde{b}_n}$, the estimator $\tilde{\theta} = (\tilde{\theta}_1, ..., \tilde{\theta}_p)^T$ and $\tilde{\gamma}$ as
\[
\tilde{\mathcal{N}}_{\tilde{b}_n} = \left\{ i \mid |\tilde{\theta}_i| > b_n \right\}, \quad \tilde{\theta}_i = \tilde{\theta}_i \times 1_{i \in \tilde{\mathcal{N}}_{\tilde{b}_n}}, \quad \tilde{\gamma} = M \tilde{\theta}
\]

Then, $\tilde{\theta}$ and $\tilde{\gamma}$ constitute the improved, i.e., debiased and thresholded, ridge regression estimator for the parameter vector $\theta$ and $\gamma = M \beta$ respectively. Apart from parameter estimation, we need to estimate the error variance $\sigma^2 = \mathbb{E} \epsilon^2_1$. The estimator for $\sigma^2$ is
Assume \( \rho \) decreases the estimation error. A decrease in bias is significant.

For \( \|\rho^2 Q (\Lambda^2 + \rho_n I_r)^{-1} - \Lambda \| \leq \rho_n^2 \|Q\|_2 \). In order to control the bias term, \( \|\beta\|_2 \) cannot be too large (which is achievable if \( \beta \) is sparse); in addition, \( \rho_n / \lambda^2 \) must be small.

We can now explain why debiasing helps decrease the estimation error; we will use the notation of section 2. According to (1), for a fixed vector \( a \in \mathbb{R}^p \),

\[
a^T \tilde{\theta} - a^T \theta = a^T \tilde{\theta} - a^T \theta - a^T Q \perp Q^T \beta
\]

Assume \( a^T Q \perp Q^T \beta = 0 \). Then the bias term of \( a^T \tilde{\theta} \) will be \( -\rho_n a^T Q (\Lambda^2 + \rho_n I_r)^{-1} - \Lambda \)\( \perp Q^T \beta \) and subtract the estimated bias from \( a^T \tilde{\theta} \), yielding the debiased estimator.

Compared to (16), the debiased estimator \( \tilde{\theta} \) changes the bias from \( -\rho_n a^T Q (\Lambda^2 + \rho_n I_r)^{-1} - \Lambda \) (having order \( \|a\|_2 \times \|\beta\|_2 / (\Lambda^2 + \rho_n) \)) to \( -\rho_n^2 Q (\Lambda^2 + \rho_n I_r)^{-1} - \Lambda \) (having order \( \rho_n^2 \|Q\|_2^2 / (\Lambda^2 + \rho_n)^2 \)).

At the same time, \( \tilde{\theta} \) will enlarge the variance from \( \text{Var}(\epsilon_1) \times a^T Q (\Lambda^2 + \rho_n I_r)^{-1} - \Lambda \) to

\[
\text{Var}(\epsilon_1) \times a^T Q (\Lambda^2 + \rho_n I_r)^{-1} - \Lambda + \rho_n (\Lambda^2 + \rho_n I_r)^{-2} \Lambda^2 Q^T a
\]

Assume \( \rho_n / \lambda^2 = o(1) \); then, \( \tilde{\theta} \)'s variance enlargement is asymptotically negligible but its decrease in bias is significant.

Even when \( \rho_n > \lambda^2 \), numerical simulations in figure 1 show that debiasing still may help decrease the estimation error.

![Figure 1](image-url)  
**Figure 1.** Estimation errors of the ridge regression estimator \( a^T \tilde{\theta}^* \), the debiased estimator \( a^T \tilde{\theta} \), the debiased and threshold ridge regression estimator \( a^T \tilde{\theta} \), and the threshold ridge regression estimator \( \tilde{\theta} \) with respect to different \( \rho_n \). The threshold \( b_n \) is chosen to be 0.1, \( a \) is a fixed linear combination vector with \( \|a\|_2 = 1 \), and \( \lambda_r = 12.684 \).
Remark 3 (Further discussion on the debiased estimator). Apart from our work, there are other procedures that help decrease the bias of an estimator. For example, B"uhlmann (2013) proposed a bias-corrected ridge regression estimator, and Zhang and Zhang (2014) considered correcting bias for a general linear regression estimator. However, the purpose of our work and those procedures are different. The bias-corrected ridge regression estimator focuses on eliminating $Q \perp Q^T \beta$ (i.e., the projection bias in B"uhlmann (2013)). Therefore, if $p<n$ and $X$ has rank $p$, then the bias-corrected ridge regression estimator equals the classical ridge regression estimator $\hat{\theta}^\star$. Our work does not focus on the projection bias but wants to diminish the estimation bias $\hat{\theta} - \rho_n Q (\lambda^2 + \rho_n I_p)^{-1} \zeta$. Thus, even if $p<n$ and $X$ has rank $p$, the debiased estimator $\hat{\theta}$ is still different from $\hat{\theta}^\star$ (which is demonstrated in figure 1).

Theorem 1. (i) Suppose assumptions 1 to 5 hold true. Then
\begin{equation}
\text{Prob} \left( \hat{N}_{b_n} \neq N_{b_n} \right) = O(n^{\alpha_p + mn_0 - mn})
\end{equation}
$N_{b_n}$ is defined in section 2. In other words, the variable selection consistency holds true asymptotically. Besides,
\begin{equation}
\max_{i=1,2,\ldots,p_1} |\hat{\gamma}_i - \gamma_i| = O_p(|M|^{1/m} \times n^{-\eta})
\end{equation}
where $\gamma_i, i = 1, \ldots, p_1$ are defined in section 2.

(ii). Suppose assumptions 1 to 6 hold true. Then
\begin{equation}
|\hat{\sigma}^2 - \sigma^2| = O_p(n^{-\alpha_\sigma}).
\end{equation}

An advantage of using $\hat{\beta}$ is that it can be computed by a closed-form formula, making it simpler to practically calculate as well as derive its theoretical guarantees. As an example, define $\hat{\beta}$’s prediction loss $\frac{1}{n} \| X \hat{\beta} - X \theta \|^2 = \frac{1}{n} \| X \hat{\beta} - X \beta \|^2$. If $N_{b_n} = \hat{N}_{b_n}$, then from (82) and (83) in appendix B we have
\begin{equation}
\frac{1}{n} \| X \hat{\beta} - X \theta \|^2 \leq 2 \frac{n}{2} \left( \sum_{i=1}^{n} x_{ij} (\hat{\theta}_j - \theta_j) \right)^2 + 2 \frac{n}{n} \left( \sum_{j \in N_{b_n}} x_{ij} \theta_j \right)^2
\end{equation}
\begin{equation*}
\Rightarrow \frac{1}{n} \| X \hat{\beta} - X \theta \|^2 = O_p \left( n^{-\alpha_\sigma} \right).
\end{equation*}

On the other hand, the prediction loss of other estimators (e.g., Lasso) can be hard to derive. Dalalyan, Hebiri and Lederer (2017), Bickel, Ritov and Tsybakov (2009) and Sun and Zhang (2012) provided oracle inequalities for the Lasso estimator. However, those inequalities depend on terms that are hard to bound. Numerical experiments in section 6 show that $\hat{\beta}$ has comparable performance with complex estimators like the threshold Lasso or post-selection estimators. In this case, it is beneficial to choose an estimator that has clear theoretical guarantees.

Define $\hat{\tau}_i, i = 1, 2, \ldots, p_1$ and $H(x), x \in \mathbb{R}$ as
\begin{equation}
\hat{\tau}_i = \frac{\sqrt{\sum_{k=1}^{r} \left( \sum_{j \in N_{b_n}} m_{ij} q_{jk} \right)^2 \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right)^2 + \frac{1}{n}}}{\sqrt{\lambda_k}}
\end{equation}
\begin{equation}
H(x) = \text{Prob} \left( \max_{i \in M} \frac{1}{\hat{\tau}_i} \sum_{k=1}^{r} c_{ik} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right) \xi_k \leq x \right)
\end{equation}
Here $\xi_k$, $k = 1, 2, ..., r$ are independent normal random variables with mean 0 and variance $\sigma^2 = E\xi^2_1$. $|M|$ (defined in (4)) and $p_1$ may grow as the sample size increases. In this case, the estimator $\max_{i=1,2,...,p_1} \frac{\hat{\gamma}_i - \gamma_i}{\hat{\tau}_i}$ does not have an asymptotic distribution. However, the cumulative distribution function of $\max_{i=1,2,...,p_1} \frac{\hat{\gamma}_i - \gamma_i}{\hat{\tau}_i}$ still can be approximated by $H(x)$ (whose expression changes as the sample size increases as well). Define $c_{1-\alpha}$ as the $1 - \alpha$ quantile of $H$; theorem 2 implies that the set

$$\left\{ \gamma = (\gamma_1, ..., \gamma_{p_1}) \mid \max_{i=1,...,p_1} \frac{\hat{\gamma}_i - \gamma_i}{\hat{\tau}_i} \leq c_{1-\alpha} \right\}$$

is an asymptotically valid $(1 - \alpha) \times 100\%$ confidence region for the parameter of interest $\gamma$.

**Theorem 2.** Suppose assumptions 1 to 7 hold true. Then

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

where $\gamma_i$, $i = 1, ..., p_1$ are defined in section 2.

Gaussian approximation theorems like theorem 2 are useful tools not only in linear models but also in other high dimensional statistics; e.g., Chernozhukov, Chetverikov and Kato (2013) and Zhang and Wu (2017).

4. **Bootstrap inference and hypothesis testing.** An obstacle for constructing a practical confidence region or testing a hypothesis via theorem 2 are the unknown $M$, $N_{b_n}$, and $\sigma$. Besides, $H$ is too complicated to have a closed-form formula. Fortunately, statisticians can simulate normal random variables on a computer, so they may use Monte-Carlo simulations to find the $1 - \alpha$ quantile of $H$. Based on this idea, this section develops a wild bootstrap algorithm similar to Mammen (1993) and Chernozhukov, Chetverikov and Kato (2013) for the following tasks: constructing the confidence region for the parameter of interest $\gamma = M \beta$; and testing the null hypothesis $\gamma = \gamma_0$ (for a known $\gamma_0$) versus the alternative hypothesis $\gamma \neq \gamma_0$. Similar to Zhang and Cheng (2017), Chernozhukov, Chetverikov and Kato (2013), and Zhang and Wu (2017), we use the maximum statistic $\max_{i=1,2,...,p_1} \frac{\hat{\gamma}_i - \gamma_i}{\hat{\tau}_i}$ to construct a simultaneous confidence region.

**Algorithm 1** (Wild bootstrap inference and hypothesis testing). **Input:** Design matrix $X$, dependent variables $y = X\beta + \epsilon$, linear combination matrix $M$, ridge parameter $\rho_n$, threshold $b_n$, nominal coverage probability $1 - \alpha$, number of bootstrap replicates $B$

**Additional input for testing:** $\gamma_0 = (\gamma_{0,1}, ..., \gamma_{0,p_1})^T$

1. Calculate $\hat{\theta}$, $\hat{\gamma} = (\hat{\gamma}_1, ..., \hat{\gamma}_{p_1})^T$ defined in (17), $\hat{\tau}_i$, $i = 1, 2, ..., p_1$ defined in (25), and $\hat{\sigma}$ defined in (18).
2. Generate i.i.d. errors $\epsilon^* = (\epsilon_1^*, ..., \epsilon_n^*)^T$ with $\epsilon_i^*, i = 1, ..., n$ having normal distribution with mean 0 and variance $\hat{\sigma}^2$, then calculate $y^* = X\hat{\theta} + \epsilon^*$ and $\hat{\theta}_1 = Q_1Q_1^T \hat{\theta}$ ($Q_1$ is defined in section 2).
3. Calculate $\hat{\theta}^{**} = (X^TX + \rho_nI_p)^{-1}X^Ty^*$ and $\hat{\theta}^{*} = (\hat{\theta}^{*}_1, ..., \hat{\theta}^{*}_p)^T = \hat{\theta}^{**} + \rho_n \times Q(\Lambda^2 + \rho_nI_r)^{-1}Q^T \hat{\theta}^{**} + \hat{\theta}_1$.
4. Calculate $N_{b_n} = \left\{ \hat{\theta}^{*}_i \right\}$ and $\hat{\theta}^{*} = (\hat{\theta}^{*}_1, ..., \hat{\theta}^{*}_p)^T$ with $\hat{\theta}^{*}_i = \hat{\theta}^{*}_i \times 1_{i \in N_{b_n}}$ for $i = 1, 2, ..., p$. 


5. Calculate \( \hat{\gamma}^* = M \hat{\theta}^* \), \( \hat{\tau}_i^* \), \( i = 1, 2, ..., p_1 \), and \( E_b^* \) such that
\[
\hat{\tau}_i^* = \sqrt{\frac{1}{n} \left( \sum_{j=1}^{r} \left( \sum_{k \in N_{s,n}} m_{ij} q_{jk} \right)^2 \times \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right)^2 \right) + \frac{1}{n} \max_{i=1,2,...,p_1} \frac{|\hat{\gamma}_i^* - \hat{\gamma}_i|}{\hat{\tau}_i^*}}
\]

6.a (For constructing a confidence region) Repeat steps 2 to 5 for \( B \) times to generate \( E_b^* \), \( b = 1, 2, ..., B \); then calculate the \( 1 - \alpha \) sample quantile \( C_{1-\alpha}^* \) of \( E_b^* \). The \( 1 - \alpha \) confidence region for the parameter of interest \( \gamma = M \beta \) is given by the set
\[
\left\{ \gamma = (\gamma_1, \ldots, \gamma_{p_1})^T \mid \max_{i=1,2,...,p_1} \frac{|\hat{\gamma}_i - \gamma_i|}{\hat{\tau}_i} \leq C_{1-\alpha}^* \right\}
\]

6.b (For hypothesis testing) Repeat steps 2 to 5 for \( B \) times to generate \( E_b^* \), \( b = 1, 2, ..., B \); then calculate the \( 1 - \alpha \) sample quantile \( C_{1-\alpha}^* \) of \( E_b^* \). Reject the null hypothesis \( \gamma = \gamma_0 \) when
\[
\max_{i=1,2,...,p_1} \frac{|\hat{\gamma}_i - \gamma_0|}{\hat{\tau}_i} > C_{1-\alpha}^*.
\]

As in section 2, if \( X \) has rank \( p \leq n \), we define \( \hat{\theta}_\perp = 0 \), the \( p \) dimensional vector with all elements 0.

According to theorem 1.2.1 in Politis, Romano and Wolf (1999), the consistency of algorithm 1—either for asymptotic validity of confidence regions or consistency of the hypothesis test—is ensured if
\[
Prob \left( \max_{i=1,2,...,p_1} \frac{|\hat{\gamma}_i - \gamma_i|}{\hat{\tau}_i} \leq c_{1-\alpha}^* \right) \rightarrow 1 - \alpha
\]

where \( c_{1-\alpha}^* \) is the \( 1 - \alpha \) quantile of the conditional distribution \( Prob^* \left( \max_{i=1,2,...,p_1} \frac{|\hat{\gamma}_i - \gamma_i|}{\hat{\tau}_i} \leq x \right) \); we prove this in theorem 3 below.

**THEOREM 3.** Suppose assumptions 1 to 7 hold true. Then
\[
\sup_{x \geq 0} |Prob^* \left( \max_{i=1,2,...,p_1} \frac{|\hat{\gamma}_i^* - \hat{\gamma}_i|}{\hat{\tau}_i^*} \leq x \right) - H(x)| = o_P(1).
\]

In addition, for any given \( 0 < \alpha < 1 \), (31) holds true.

Theorem 3 has two implications. On the one hand, the confidence region introduced in step 6.a of algorithm 1 is asymptotically valid, i.e., its coverage tends to \( 1 - \alpha \). On the other hand, consider the hypothesis test of step 6.b of algorithm 1; Theorem 3 implies that, if the null hypothesis is true, then the probability for incorrectly rejecting the null hypothesis is asymptotically \( \alpha \), i.e., the test is consistent.

5. **Bootstrap interval prediction.** Given our data from the linear model \( y = X \beta + \epsilon \), consider a new \( p_1 \times p \) regressor matrix \( X_f \), i.e., a collection of regressor (column) vectors that happen to be of interest; as with \( X \) itself, \( X_f \) is assumed given, i.e., deterministic. The prediction problem involves (a) finding a predictor for the future (still unobserved) vector \( y_f = X_f \beta + \epsilon_f \), and (b) finding a \( 1 - \alpha \) prediction region \( A \subset R^{p_1} \) so that \( Prob(y_f \in A) \rightarrow 1 - \alpha \) as the (original) sample size \( n \rightarrow \infty \). Here \( \epsilon_f = (\epsilon_{f,1}, \ldots, \epsilon_{f,p_1})^T \) are i.i.d. errors with the same marginal distribution as \( \epsilon_1 \), and \( \epsilon_f \) is independent with \( \epsilon \).
Finding a good predictor based on different criteria is a big topic. For example, Greenshtein and Ritov (2004) applied Lasso in constructing predictors and their predictor’s mean square error is minimal asymptotically. We construct an intuitive predictor based on the following idea: if $\beta$ were known, the predictor of $y_f$ that is optimal with respect to total mean squared error is $X_f \beta$; since $\beta$ is typically unknown, we can estimate it by $\hat{\theta}$ as in (17), yielding the practical predictor $\hat{y}_f = X_f \hat{\theta}$. In what follows, we would like to derive a $1 - \alpha$ prediction region for $y_f$ based on the intuitive predictor $\hat{y}_f$.

We adopt definition 2.4.1 of Politis (2015), and define a consistent prediction region in terms of conditional coverage as follows.

**Definition 1 (Consistent prediction region).** A set $\Gamma = \Gamma(X, y, X_f)$ is called a $1 - \alpha$ consistent prediction region for the future observation $y_f = X_f \beta + \epsilon_f$ if

$$\text{Prob}(y_f \in \Gamma | y) \to 1 - \alpha$$

as $n \to \infty$. Note that the convergence in (33) is “in probability” since $\text{Prob}(y_f \in \Gamma | y)$ is a function of $y$, and therefore random; see also Lei and Wasserman (2014) for more on the notion of conditional validity.

Other authors, including Stine (1985), Romano, Patterson and Candès (2019), and Chernozhukov, Wüthrich and Zhu (2019), considered another definition of prediction interval consistency focusing on unconditional coverage, i.e., insisting that

$$\text{Prob}(y_f \in \Gamma) \to 1 - \alpha.$$  

However, the conditional coverage of definition 1 is a stronger property. To see why, define the random variables $U_n = \text{Prob}(y_f \in \Gamma | y)$, noting that $y$ has dimension $n$. Then, the boundedness of $U_n$ can be invoked to show that if $U_n \to 1 - \alpha$, then $\mathbb{E}U_n \to 1 - \alpha$ as well. Hence, (33) implies (34); see Zhang and Politis (2021) for a further discussion on conditional vs. unconditional coverage.

Consider the prediction error $y_f - X_f \hat{\theta} = \epsilon_f - X_f (\hat{\theta} - \beta)$. If we can put bounds on the prediction error that are valid with conditional probability $1 - \alpha$ (asymptotically), then a consistent prediction region ensues. Note that the prediction error has two parts: $\epsilon_f$ and $-X_f (\hat{\theta} - \beta)$. Although the latter may be asymptotically negligible, it is important in practice to not approximate it by zero as it would yield finite-sample undercoverage; see e.g. Ch. 3 of Politis (2015) for an extensive discussion.

Theorem 2 indicates that the asymptotically negligible estimation error can be approximated by normal random variables. On the other hand, the non-negligible error $\epsilon_f$ may not have a normal distribution; so in order to approximate the distribution of $\epsilon_f - X_f (\hat{\theta} - \beta)$, we need to estimate the errors’ marginal distribution as well.

This section requires some additional assumptions.

**Additional assumptions**

8. The cumulative distribution function of errors $F(x) = \text{Prob}(\epsilon_1 \leq x)$ is continuous

9. The number of regressors of interest is bounded, i.e., $p_1 = O(1)$

Since $F$ is increasing and bounded, if $F(x)$ is continuous, then $F$ is uniformly continuous on $\mathbb{R}$, this property is useful in the proof of lemma 1.
Lemma 1. Suppose assumptions 1 to 6 and 8 hold true. Define the residuals \( \epsilon = (\epsilon_1, \ldots, \epsilon_n)^T = y - X\hat{\theta} \), as well as the centered residuals \( \bar{\epsilon} = (\bar{\epsilon}_1, \ldots, \bar{\epsilon}_n)^T \) with \( \bar{\epsilon}_i = \epsilon_i - \frac{1}{n} \sum_{i=1}^n \epsilon_i \). If we let \( \bar{F}(x) = \frac{1}{n} \sum_{i=1}^n 1_{\epsilon_i \leq x} \), then

\[
\sup_{x \in \mathbb{R}} |\bar{F}(x) - F(x)| \to_p 0 \text{ as } n \to \infty.
\]

We emphasize that the dimension of parameters \( p \) in lemma 1 can grow to infinity as long as assumption 4 is satisfied. Furthermore, the validity of lemma 1 – as well as that of theorem 4 that follows – does not require assumption 7.

We will resample the centered residuals \( \bar{\epsilon}_i, i = 1, 2, \ldots, n \) (in other words, generate random variables with distribution \( \bar{F} \)) in algorithm 2. Lemma 1 will ensure that the centered residuals can capture the distribution of the non-negligible errors.

For a high dimensional linear model, lemma 1 is not an obvious result; see Mammen (1996) for a detailed explanation. Lemma 1 is the foundation for a new resampling procedure as follows; this is a hybrid bootstrap as it combines the residual-based bootstrap to replicate the new error \( \epsilon_f \) with the normal approximation to the estimation error \( -X_f(\hat{\theta} - \beta) \).

Algorithm 2 (Hybrid bootstrap for prediction region). Input: Design matrix \( X \), dependent variables \( y = X\beta + \epsilon \), a new \( p_1 \times p \) linear combination matrix \( X_f \), ridge parameter \( \rho_n \), threshold \( b_n \), nominal coverage probability \( 0 < 1 - \alpha < 1 \), the number of bootstrap replicates \( B \)

1. Calculate \( \hat{\theta} \) defined in (17), \( \hat{\sigma} \) defined in (18), \( \hat{\epsilon} \) defined in lemma 1, \( \hat{y}_f = (\hat{y}_{f,1}, \ldots, \hat{y}_{f,p_1})^T = X_f\hat{\theta} \), and \( \hat{\theta}_\perp = Q_{\perp}Q_{\perp}^T \hat{\theta} \).

2. Generate i.i.d. errors \( \epsilon^* = (\epsilon^*_1, \ldots, \epsilon^*_n)^T \) with \( \epsilon^*_i, i = 1, \ldots, n \) having mean 0 and variance \( \hat{\sigma}^2 \). Then generate i.i.d. errors \( \epsilon^*_f = (\epsilon^*_{f,1}, \ldots, \epsilon^*_{f,p_1})^T \) with \( \epsilon^*_{f,i}, i = 1, \ldots, p_1 \) having cumulative distribution function \( \hat{F} \) defined in lemma 1. Calculate \( y^* = X\hat{\theta} + \epsilon^* \).

3. Calculate \( \theta^{**} = (X^T X + \rho_n I_{p_1})^{-1} X^T y^* \) and \( \theta^* = \hat{\theta}^* + \rho_n \times Q(\Lambda^2 + \rho_n I_{p})^{-1}Q^T \hat{\theta}^{**} + \hat{\theta}_\perp \). Then derive \( \hat{\theta}^{**} = \left\{ i \mid \theta^{**}_i > b_n \right\} \), \( \theta^* = (\theta^*_1, \ldots, \theta^*_p)^T \) with \( \theta^*_i = \theta^*_i \times 1_{i \in \hat{\theta}^{**}} \) for \( i = 1, 2, \ldots, p \).

4. Calculate \( y^{**}_f = (y^{**}_{f,1}, \ldots, y^{**}_{f,p_1})^T = X_f\hat{\theta} + \epsilon^* \) and \( \hat{y}^{**}_f = (\hat{y}^{**}_{f,1}, \ldots, \hat{y}^{**}_{f,p_1})^T = X_f\hat{\theta}^* \). Define \( E^*_b = \max_{i=1,2,\ldots,p_1} |y^*_f - \hat{y}^{**}_f| \).

5. Repeat steps 2 to 4 for \( B \) times, and generate \( E^*_b \), \( b = 1, 2, \ldots, B \). Calculate the \( 1 - \alpha \) sample quantile \( C_{1-\alpha}^* \) of \( E^*_b \). Then, the \( 1 - \alpha \) prediction region for \( y_f = X_f\beta + \epsilon_f \) is given by

\[
\left\{ y_f = (y_{f,1}, \ldots, y_{f,p_1})^T \mid \max_{i=1,2,\ldots,p_1} \left| y_{f,i} - \hat{y}_{f,i} \right| \leq C_{1-\alpha}^* \right\}.
\]

If the design matrix \( X \) has rank \( p \), then \( \hat{\theta}_\perp \) is defined to be 0.

Similar to section 4, here we define \( \epsilon_{1-\alpha}^* \) as the \( 1 - \alpha \) quantile of the conditional distribution \( \text{Prob}^* \left( \max_{i=1,2,\ldots,p_1} \left| y_{f,i} - \hat{y}_{f,i} \right| \leq x \right) \), which can be approximated by \( C_{1-\alpha}^* \) by letting \( B \to \infty \). Theorem 4 below proves \( \text{Prob} \left( \max_{i=1,2,\ldots,p_1} \left| y_{f,i} - \hat{y}_{f,i} \right| \leq \epsilon_{1-\alpha}^* \right) \to 1 - \alpha \) as the sample size \( n \to \infty \), which justifies the consistency of the prediction region (36).
THEOREM 4. Suppose assumptions 1 to 6 and 8 to 9 hold (here consider \( M = (m_{ij})_{i=1,\ldots,p, j=1,\ldots,p} \) in assumption 5 as \( X_P \)). Then

\[
\sup_{x \geq 0} \left| \text{Prob}^* \left( \max_{i=1,2,\ldots,p} |y_{f,i} - \hat{y}_{f,i}| \leq x \right) - \text{Prob}^* \left( \max_{i=1,2,\ldots,p} |y_{f,i} - \hat{y}_{f,i}| \leq x \right) \right| = o_p(1).
\]

For any fixed \( 0 < \alpha < 1 \), it follows that

\[
\text{Prob}^* \left( \max_{i=1,2,\ldots,p} |y_{f,i} - \hat{y}_{f,i}| \leq c_{1-\alpha} \right) \to_p 1 - \alpha \text{ as } n \to \infty.
\]

Note that the bootstrap probability \( \text{Prob}^*(\cdot) \) is probability conditional on the data \( y \), thus justifying the notion of conditional validity of our definition 1.

A version of the algorithm 2 can be constructed where the residual-based bootstrap part is conducted by resampling from the empirical distribution of the (centered) predictive, i.e., leave-one-out, residuals instead of the fitted residuals \( \hat{c}_i \); see Ch. 3 of Politis (2015) for a discussion.

6. Numerical Simulations. Define \( k_n = \sqrt{n \log(n)} \) and the following four terms

\[
\mathcal{K}_1 = \max_{i=1,2,\ldots,p} k_n \sum_{j \notin N_n} m_{ij} \theta_j, \quad \mathcal{K}_2 = \max_{i=1,2,\ldots,p} k_n \sum_{j=1}^p m_{ij} \theta_{i,j},
\]

\[
\mathcal{K}_3 = b_n \sum_{j \notin N_n} |\theta_j|, \quad \mathcal{K}_4 = \sqrt{\frac{|N_n|}{b_n}};
\]

see section 2 for the meaning of notations in the above. Assumptions 5 and 6 imply that these terms converge to 0 as the sample size \( n \to \infty \). Indeed, if one of the \( \mathcal{K}_i \) is large, the debiased and threshold ridge regression estimator may have a large bias, which affects the performance of the bootstrap algorithms.

In this section, we generate the design matrix \( X \), the linear combination matrix \( M \), and the parameters \( \beta \) through the following strategies:

**Design matrix** \( X \): define \( X = [x_1, \ldots, x_n]^T \) with \( x_i = (x_{i1}, \ldots, x_{ip})^T \in \mathbb{R}^p, i = 1, \ldots, n \). Generate \( x_1, x_2, \ldots \) as i.i.d. normal random vectors with mean 0 and covariance matrix \( \Sigma \in \mathbb{R}^{p \times p} \). We choose \( \Sigma \) with diagonal elements equal to 2.0 and off-diagonal elements equal to 0.5.

\[
M = \begin{bmatrix}
m_{11} & m_{12} & \cdots & m_{1\tau} & m_{1\tau+1} & \cdots & m_{1p} \\
m_{21} & m_{22} & \cdots & m_{2\tau} & m_{2\tau+1} & \cdots & m_{2p} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
m_{|\mathcal{M}|1} & m_{|\mathcal{M}|2} & \cdots & m_{|\mathcal{M}|\tau} & m_{|\mathcal{M}|\tau+1} & \cdots & m_{|\mathcal{M}|p} \\
0 & 0 & \cdots & 0 & m_{|\mathcal{M}|+1\tau} & \cdots & m_{|\mathcal{M}|+1p} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 0 & m_{p,\tau+1} & \cdots & m_{pp}
\end{bmatrix}
\]

\( M \) and \( \beta \) when \( p < n \): choose \( \tau = 50 \) in (40). Generate \( m_{ij}', i = 1, 2, \ldots, |\mathcal{M}|, j = 1, 2, \ldots, \tau \) as i.i.d. normal with mean 0.5 and variance 1.0, and generate \( m_{ij}', i = 1, 2, \ldots, p, j = \tau + 1, \ldots, p \) as i.i.d. normal with mean 1.0 and variance 4.0. Use \( m_{ij} = 2.0 \times m_{ij}' / \sqrt{\sum_{j=1}^\tau m_{ij}'^2} \) for \( i = 1, 2, \ldots, |\mathcal{M}|, j = 1, 2, \ldots, \tau; m_{ij} = 4.0 \times m_{ij}' / \sqrt{\sum_{j=\tau+1}^p m_{ij}'^2} \) for \( i = 1, 2, \ldots, |\mathcal{M}|, j = \tau + 1, \ldots, p \).
\[ \tau + 1, \ldots, p; \text{ and } m_{ij} = 6.0 \times m'_{ij}/\sqrt{\sum_{j=\tau+1}^{p} m_{ij}^2} \text{ for } i = |M| + 1, \ldots, p, j = \tau + 1, \ldots, p. \]

Choose \( \beta = (\beta_1, \ldots, \beta_p)^T \) with \( \beta_i = 2.0, i = 1, 2, 3, \beta_i = -2.0, i = 4, 5, 6, \beta_i = 1.0, i = 7, 8, 9, \beta_i = -1.0, i = 10, 11, 12, \beta_i = 0.01, i = 13, 14, 15, 16, \) and 0 otherwise.

**M and \( \beta \) when \( p > n \):** Choose \( \tau = 6 \) in (40). Generate \( m_{ij}, i = 1, 2, \ldots, |M|, j = 1, 2, \ldots, \tau \) as i.i.d. normal with mean 1.0 and variance 1.0, and generate \( m_{ij}, i = 1, 2, \ldots, p, j = \tau + 1, \ldots, p \) as i.i.d. normal with mean 1.0 and variance 4.0. Use \( m_{ij} = 2.0 \times m'_{ij}/\sqrt{\sum_{j=\tau+1}^{p} m_{ij}^2} \) for \( i = 1, 2, \ldots, |M|, j = 1, 2, \ldots, \tau; \) and \( m_{ij} = m'_{ij}/\sqrt{\sum_{j=\tau+1}^{p} m_{ij}^2} \) for \( i = 1, 2, \ldots, p, j = \tau + 1, \ldots, p. \) Choose \( \beta_i = 1.0, i = 1, 2, 3, \beta_i = -1.0, i = 4, 5, 6, \) and 0 otherwise. When \( p > n, \beta \) may not be identifiable (Shao and Deng (2012)), and \( \beta \) may not equal \( \theta \) (defined in section 2) despite \( X(\hat{\beta} - X\theta) \). We consider both situations and evaluate the performance of proposed methods on the linear model \( y = X(\beta + \epsilon) \) and \( y = X\theta + \epsilon \). We fix \( X \) and \( M \) in each simulation.

The different regression algorithms considered are the debiased and threshold ridge regression (Thr Thr), ridge regression, Lasso, threshold ridge regression (Thr Ridge), threshold Lasso (Thr Lasso), and the post-selection algorithms, i.e., Lasso + OLS (Post OLS), and Lasso + thresholding (Deb Thr), ridge regression, Lasso, threshold ridge regression (Thr Ridge), threshold parameter \( \rho \) estimation/hypothesis testing and interval-prediction in Table 3 and Figure 5. The optimal ridge method on the linear model \( y = X(\beta + \epsilon) \) and \( y = X\theta + \epsilon \). We fix \( X \) and \( M \) in each simulation.

**Table 1**

| Case | \( n \) | \( p \) | Residual | \( |M| \) | \( \lambda_{\tau} \) | \( \rho_n \) | \( b_n \) | \( K_1 \) | \( K_2 \) | \( K_3 \) | \( K_4 \) |
|------|------|------|--------|------|------|------|------|------|------|------|------|
| 1    | 1000 | 500  | Normal | 800  | 300  | 12.978 | 56.453 | 0.343 | 1.370 | 0.0  | 0.013 | 1.712 |
| 2    | 1000 | 500  | Laplace| 800  | 300  | 12.561 | 36.728 | 0.354 | 1.636 | 0.0  | 0.014 | 1.769 |
| 3    | 1000 | 650  | Laplace| 800  | 300  | 8.226  | 56.432 | 0.396 | 1.553 | 0.0  | 0.016 | 3.085 |
| 4    | 1000 | 500  | Laplace| 800  | 700  | 12.847 | 55.317 | 0.346 | 1.510 | 0.0  | 0.014 | 1.730 |
| 5    | 1000 | 1500 | Normal | 800  | 300  | 9.766  | 1.201  | 0.228 | 6.938 | 129/0.0 | 8.214 | 3.962 |
| 6    | 1000 | 1500 | Laplace| 800  | 300  | 9.766  | 1.201  | 0.228 | 6.938 | 129/0.0 | 8.214 | 3.962 |

Case 5 and 6 consider both the linear model \( y = X(\beta + \epsilon) \) and \( y = X\theta + \epsilon \), here \( \beta \neq \theta = QQ^T \beta. \) The difference in \( \beta \) and \( \theta \) affects the value of \( K_2 \) (but does not affect others), so we have two values in table 1.

Figure 2 plots the Euclidean norm \( \|\hat{\gamma} - \gamma\|_2 \), with \( \hat{\gamma} \) defined in (17), and \( \gamma \) defined in Section 2, for various linear regression methods. When the underlying linear model is sparse, thresholding decreases the ridge regression estimator’s error (from around 10 to around 2 in our experiment). However, the performance of the threshold ridge regression method is sensitive to the ridge parameter \( \rho_n \), i.e., \( \|\hat{\gamma} - \gamma\|_2 \) can be significantly larger than its minimum despite \( \rho_n \) is close to the minimizer of \( \|\hat{\gamma} - \gamma\|_2 \).

In reality, cross validation does not necessarily guarantee selection of the optimal \( \rho_n \) and \( b_n \), so it is risky to use the threshold ridge regression method. Debiasing helps decrease the ridge regression estimator’s error; more importantly, it is robust to changes in the choice of \( \rho_n \). Even if a cross validation selects a sub-optimal \( \rho_n \), the error of the debiased and threshold
RIDGE REGRESSION REVISITED: DEBIASING, THRESHOLDING AND BOOTSTRAP

Fig 2. Estimation performance of various linear regression methods over case 1 to 4. 'Deb' abbreviates 'Debiased', 'Thr' abbreviates 'Threshold', 'Post' abbreviates 'Post-selection', and 'OLS' abbreviates 'ordinary least square'. Red dots represent the ridge/Lasso parameters $\rho_n$ and the thresholds $b_n$ selected by 5-fold cross validation. The vertical axis represents the Euclidean norm of $\hat{\gamma} - \gamma$ where $\hat{\gamma}$ is defined in (17) and $\gamma$ is defined in Section 2. The little graphs in the middle of each of the four graphs show a zoomed-in part of the graph above it. The little graphs below each of the four graphs show the estimation performance of the debiased and threshold ridge regression method with respect to different thresholds.
Fig 3. Estimation performance of various linear regression methods over case 5 to 6. The meaning of symbols coincides with figure 2.
ridge regression estimator does not surge, and the estimator’s performance does not notably deteriorate. On the other hand, in Figure 2 and 3, $\|\hat{\gamma} - \gamma\|_2$ reaches its minimum and does not increase for a wide range of $b_n$. For example, in case 3 the cross validation chooses $b_n = 0.395$, but any value between 0.30 and 0.75 can be the optimal threshold (the threshold having the smallest $\|\hat{\gamma} - \gamma\|_2$). Since there is a wide interval of thresholds $b_n$ that have small $\|\hat{\gamma} - \gamma\|_2$, the regression algorithm has robustness regarding the choice of $b_n$. Because of these good properties, we consider the debiased and threshold ridge regression as a practical method to handle real-life data.

Thresholding also helps improve the performance of Lasso, especially when the Lasso parameter is small. However, when the Lasso parameter becomes large, Lasso method already recovers the underlying sparsity of the linear model, and thresholding becomes unnecessary (but large Lasso parameters tend to introduce large bias).

When the dimension of parameters $p$ is greater than the sample size $n$, both parameters $\beta$ and $\theta$(see section 2) could be considered as the ‘parameters’ for the linear model. Lasso methods estimate linear combinations of $\beta$, while ridge regression methods estimate linear combinations of $\theta$. Under this situation, the difference between $\beta$ and $\theta$ is the main factor for the estimators’ error. In reality, statisticians cannot distinguish between $\beta$ and $\theta$ based on data. So they need to design which parameters to estimate a priori and select a suitable regression method (e.g., Lasso, ridge regression, or their variations) reflecting their preferences.

The optimal prediction loss of the debiased and threshold ridge regression method is comparable to the threshold Lasso and the post-selection algorithms. Furthermore, the prediction loss is robust to changes in the choice of $\rho_n$, $b_n$. When $p > n$, the sparsity of $\theta$ is violated and
a large bias is introduced to the estimator (see Table 1). As a result the prediction loss will be enlarged.

As a summary of Figure 2 to 4, apart from having a closed-form formula, the debiased and threshold ridge regression has the smallest estimation error and prediction loss among all ridge regression variations, and has comparable performance to the threshold Lasso. Furthermore, it is not overly sensitive on changes in the ridge parameter $\rho_n$ as well as the threshold $b_n$. Therefore, even when a sub-optimal $\rho_n$ or $b_n$ are selected, the performance of the debiased and threshold ridge regression is not severely affected. When $p > n$, this method (and other ridge regression methods) considers $\theta$ rather than $\beta$ to be the parameter of the linear model. So, in this case, ridge regression methods are suitable if the underlying linear model is indeed $y = X\theta + \epsilon$ (in other words, the projection does not have effect on the parameters of the linear model).

Table 2 demonstrates the model selection performance of various linear regression algorithms. Following Fithian, Sun and Taylor (2017), we evaluate the algorithms through the frequency of model misspecification (i.e., $\widehat{N}_{b_n} \neq N_{b_n}$), $P(\widehat{N}_{b_n} \neq N_{b_n})$; the average size of model misspecification $|\widehat{N}_{b_n} \Delta N_{b_n}|$ (here $\Delta$ denotes the symmetric difference, i.e. $A \Delta B = (A - B) \cup (B - A)$); and the average false discovery rate $|\widehat{N}_{b_n} - N_{b_n}| / \max(|\widehat{N}_{b_n}|, 1)$. Notice that Lasso and the ridge regression do not have thresholds. For these algorithms we say $i \in \widehat{N}_{b_n}$ if the estimated parameter $|\widehat{\beta}_i| > 0.001$. When the sparsity assumption is not violated, the debiased and threshold ridge regression can perfectly recover the model sparsity, and thresholding is also an essential tool that improves Lasso’s model selection performance. On the other hand, if the sparsity assumption is violated, then $|\theta_i|$ can be close to $b_n$ even if $i \notin N_{b_n}$. Despite the stochastic errors are still small, the summation of $\theta_i$ and the stochastic error can exceed $b_n$, which results in selecting a false model.

| Case | Algorithm | $P(\widehat{N}_{b_n} \neq N_{b_n})$ | $|\widehat{N}_{b_n} \Delta N_{b_n}|$ | False discovery rate |
|------|-----------|-------------------------------|---------------------------------|---------------------|
| 1    | Deb Thr   | 0.0                           | 0.0                             | 0.0                 |
|      | Lasso     | 1.0                           | 9.674                           | 0.463               |
|      | Ridge     | 1.0                           | 240.53                          | 0.967               |
|      | Thr Lasso | 0.009                         | 0.009                           | 0.001               |
|      | Thr Ridge | 0.0                           | 0.0                             | 0.0                 |
| 5(use $\theta$) | Deb Thr   | 0.132                         | 0.140                           | 0.034               |
|      | Lasso     | 1.0                           | 761.59                          | 0.308               |
|      | Ridge     | 1.0                           | 452.69                          | 0.283               |
|      | Thr Lasso | 0.004                         | 0.004                           | 0.001               |
|      | Thr Ridge | 0.508                         | 0.729                           | 0.157               |

Table 3 records the average errors of the proposed statistics $\widehat{\gamma}$ (defined in (17)), $\widehat{\sigma}^2$ (defined in (18)), and the coverage probability of the confidence region (29) as well as the coverage probability of the prediction region (36), in 1000 numerical simulations. We also record the frequency of model misspecification $P(\widehat{N}_{b_n} \neq N_{b_n})$. When the sample size $n$ is greater than the dimension of parameters $p$, thresholding is likely to recover the sparsity of the parameters. In all these cases, i.e., Case 1–4, our confidence intervals achieve near-perfect coverage. The slight under-coverage in prediction intervals is a well-known phenomenon; see e.g. Ch. 3.7 of Politis (2015).
However, in cases 5 and 6 where \( p > n \), \( \theta \) is not necessarily sparse, and model misspecification may happen. Notably, \( \hat{\gamma} \)’s error in estimating linear combinations of \( \theta \) does not surge even when \( p > n \). However, the difference between \( \beta \) and \( \theta \) introduces a large bias to \( \hat{\gamma} \). Besides, when \( p > n \), assumption 6 can be violated. Correspondingly the variance estimator \( \hat{\sigma}^2 \) may have a large error. The difference between \( \beta \) and \( \theta \) invalidates the confidence region (29).

For prediction region (36), this problem still exists. However, the prediction region catches non-negligible errors apart from the asymptotically negligible errors and it is wider than the confidence region. Consequently, as long as the absolute values of difference are small, the prediction interval’s performance will not be severely affected.

**Table 3**

Frequency of model misspecification; average errors of \( \hat{\gamma} \) and \( \hat{\sigma}^2 \); and the coverage probability for the confidence region (29) and the prediction region (36). The nominal coverage probability is \( 1 - \alpha = 95\% \). The overscore represents calculating the sample mean among 1000 simulations. We choose the number of bootstrap replicates \( B = 500 \).

<table>
<thead>
<tr>
<th>Estimation and Confidence region construction</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case #</td>
<td>( P(\hat{N}<em>{b_n} \neq N</em>{b_n}) )</td>
</tr>
<tr>
<td>1</td>
<td>0.0</td>
</tr>
<tr>
<td>2</td>
<td>0.0</td>
</tr>
<tr>
<td>3</td>
<td>0.0</td>
</tr>
<tr>
<td>4</td>
<td>0.0</td>
</tr>
<tr>
<td>5(use ( \beta ))</td>
<td>0.129</td>
</tr>
<tr>
<td>5(use ( \theta ))</td>
<td>0.122</td>
</tr>
<tr>
<td>6(use ( \beta ))</td>
<td>0.126</td>
</tr>
<tr>
<td>6(use ( \theta ))</td>
<td>0.137</td>
</tr>
</tbody>
</table>

Figure 5 plots the power curve of the hypothesis test of \( \gamma = \gamma_0 \) versus \( \gamma \neq \gamma_0 \); here, we use \( \gamma_0 = \gamma + \delta \times (1, 1, \ldots, 1)^T \) and \( \delta > 0 \).

**Figure 5.** Power of the test for cases 1 and 2; the x-axis represents \( \max_{i=1,\ldots,p} |\gamma_{0,i} - \gamma_{i}| \). Nominal size for the test is \( 5\% \); see algorithm 1 for the meaning of notations.
7. Conclusion. The paper at hand proposes an improved, i.e., debiased and thresholded, ridge regression method that recovers the sparsity of parameters and avoids introducing a large bias. Besides, it derives a consistency result and the Gaussian approximation theorem for the improved ridge estimator. An asymptotically valid confidence region for $\gamma = M\beta$ and a hypothesis test of $\gamma = \gamma_0$ are also constructed based on a wild bootstrap algorithm. In addition, a novel, hybrid resampling procedure was proposed that can be used to perform interval prediction based on the improved ridge regression. When the dimension of parameters $p$ is larger than the sample size $n$, the proposed method estimates linear combinations of $\theta = QQ^T\beta$ instead of linear combinations of $\beta$. If the underlying parameter is indeed $\beta$ and the projection bias $\theta - \beta$ is not negligible, then the proposed methods may fail to provide a consistent result.

Numerical simulations indicate that improved ridge regression has comparable performance to the threshold Lasso while having at least two major advantages: (a) Ridge regression is easily computed using a closed-form expression, and (b) it appears to be quite robust against a non-optimal choice of the ridge parameter $\rho_0$ as well as the threshold $b_0$. Therefore, ridge regression may be found useful again in applied work using high-dimensional data as long as practitioners make sure to include debiasing and thresholding.

Acknowledgements. The authors are grateful to the anonymous referees for their valuable suggestions that significantly improve the content of this article. This research was partially supported by NSF Grant DMS 19-14556.

REFERENCES


APPENDIX A: SOME IMPORTANT LEMMAS

This section introduces three useful lemmas. Lemma 2 comes from Whittle (1960), which directly contributes to the model selection consistency. Lemma 3 and 4 are similar to Chernozhukov, Chetverikov and Kato (2013), they used a joint normal distribution to approximate the distribution of linear combinations of independent random variables.

**LEMMA 2.** Suppose random variables $\epsilon_1, \ldots, \epsilon_n$ are i.i.d., $\mathbf{E}\epsilon_1 = 0$, and $\exists$ a constant $m > 0$ such that $\mathbf{E}|\epsilon_1|^m < \infty$. In addition suppose the matrix $\Gamma = (\gamma_{ij})_{i=1,2,\ldots,k,j=1,2,\ldots,n}$ satisfies

$$\max_{i=1,2,\ldots,k} \sum_{j=1}^n \gamma_{ij}^2 \leq D, \ D > 0$$

Then $\exists$ a constant $E$ which only depends on $m$ and $\mathbf{E}|\epsilon_1|^m$ such that for $\forall \delta > 0$,

$$\text{Prob}\left( \max_{i=1,2,\ldots,k} \sum_{j=1}^n \gamma_{ij}\epsilon_j > \delta \right) \leq \frac{kED^{m/2}}{\delta^m}$$

**PROOF.** From theorem 2 in Whittle (1960), for any $i = 1, 2, \ldots, k$,

$$\text{Prob}\left( \sum_{j=1}^n \gamma_{ij}\epsilon_j > \delta \right) \leq \frac{\mathbf{E}\sum_{j=1}^n \gamma_{ij}\epsilon_j^m}{\delta^m} \leq \frac{2^mC(m)\mathbf{E}|\epsilon_1|^m(\sum_{j=1}^n \gamma_{ij}^2)^{m/2}}{\delta^m} \leq \frac{2^mC(m)\mathbf{E}|\epsilon_1|^mD^{m/2}}{\delta^m}$$

Here $C(m)$ is a constant depending on $m$. Choose $E = 2^mC(m)\mathbf{E}|\epsilon_1|^m$.

$$\text{Prob}\left( \max_{i=1,2,\ldots,k} \sum_{j=1}^n \gamma_{ij}\epsilon_j > \delta \right) \leq \sum_{i=1}^k \text{Prob}\left( \sum_{j=1}^n \gamma_{ij}\epsilon_j > \delta \right) \leq \frac{kED^{m/2}}{\delta^m}$$

$\square$

**LEMMA 3.** Suppose $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ are joint normal random variables with mean $\mathbf{E}\epsilon = 0$, non-singular covariance matrix $\mathbf{E}\epsilon\epsilon^T$, and positive marginal variance $\sigma_i^2 = \mathbf{E}\epsilon_i^2 > 0$, $i = 1, 2, \ldots, n$. In addition, suppose $\exists$ 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 $\delta > 0$,

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

$C$ only depends on $c_0$ and $C_0$.

**PROOF OF LEMMA 3.** First for any $i = 1, 2, \ldots, n$,

$$|\epsilon_i| = \max(\epsilon_i, -\epsilon_i) \Rightarrow \max_{i=1,\ldots,n} |\epsilon_i| = \max(\max_{i=1,\ldots,n} \epsilon_i, \max_{i=1,\ldots,n} -\epsilon_i)$$

Therefore, for any $x \in \mathbb{R}$,

$$\text{Prob}(\max_{i=1,2,\ldots,n} |\epsilon_i| \leq x + \delta) - \text{Prob}(\max_{i=1,2,\ldots,n} |\epsilon_i| \leq x)$$

$$= \text{Prob}(0 < \max_{i=1,2,\ldots,n} \epsilon_i, \max_{i=1,2,\ldots,n} -\epsilon_i) - x \leq \delta)$$

$$\leq \text{Prob}(0 < \max_{i=1,\ldots,n} \epsilon_i - x \leq \delta) + \text{Prob}(0 < \max_{i=1,\ldots,n} -\epsilon_i - x \leq \delta)$$

$$\leq \text{Prob}(\max_{i=1,\ldots,n} |\epsilon_i - x| \leq \delta) + \text{Prob}(\max_{i=1,\ldots,n} |\epsilon_i - x| \leq \delta)$$
Choose $C = \frac{\sqrt{2} \times (1 + |\log(c_0)| + |\log(C_0)|)}{c_0} + \frac{4\sqrt{2}C_0}{c_0^2} \bigg(2 + \sqrt{|\log(c_0)| + |\log(C_0)|}\bigg)$, which only depends on $c_0, C_0$. Then

$$\sup_{x \in \mathbb{R}} \left| \text{Prob} \left( \max_{i=1,2,\ldots,n} \epsilon_i \leq x + \delta \right) - \text{Prob} \left( \max_{i=1,2,\ldots,n} \epsilon_i \leq x \right) \right| \leq 2C\delta \big(1 + \sqrt{|\log(n)| + |\log(\epsilon)|}\big)$$

\[\square\]

**Lemma 4.** Suppose $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ are i.i.d. random variables with $E\epsilon_1 = 0, E\epsilon_1^2 = \sigma^2$ and $E|\epsilon_1|^3 < \infty$. $\Gamma = (\gamma_{ij})_{i=1,2,\ldots,n, j=1,2,\ldots,k}$ is an $n \times k$ ($1 \leq k \leq n$) rank $k$ matrix. And \(\exists\) constants $0 < c_\Gamma \leq C_\Gamma < \infty$ such that $c_\Gamma^2 \leq \sum_{i=1}^{n} \gamma_{ij}^2 \leq C_\Gamma^2$ for $i = 1, 2, \ldots, k$. $\hat{\sigma}^2 = \hat{\sigma}^2(\epsilon)$ is an estimator of $\sigma^2$ and random variables $\epsilon^*|\epsilon = (\epsilon_1^*, \ldots, \epsilon_n^*)^T$ are i.i.d. with $\epsilon_1^*$ having normal distribution $\mathcal{N}(0, \hat{\sigma}^2)$. $\epsilon_1^*$ is independent of $\epsilon$ for $i = 1, 2, \ldots, n$. In addition, suppose one of the following conditions:

**C1.** \(\exists\) a constant $0 < \alpha_\sigma \leq 1/2$ such that

$$|\sigma^2 - \hat{\sigma}^2| = O_p(n^{-\alpha_\sigma}) \text{ and } \max_{j=1,2,\ldots,k} \gamma_{ji} = o \left( n^{(\alpha_\sigma - 1)/2} \times \log^{-3/2}(n), n^{-1/3} \times \log^{-3/2}(n) \right)$$

**C2.** \(\exists\) a constant $0 < \alpha_\sigma < 1/2$ such that

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

Then we have

$$\sup_{x \in \mathbb{R}} \left| \text{Prob} \left( \max_{i=1,2,\ldots,k} \left| \sum_{j=1}^{n} \gamma_{ji} \epsilon_j \right| \leq x \right) - \text{Prob}^* \left( \max_{i=1,2,\ldots,k} \left| \sum_{j=1}^{n} \gamma_{ji} \epsilon_j^* \right| \leq x \right) \right| = o_p(1)$$

In particular, if $\hat{\sigma} = \sigma$, by assuming one of the following conditions, $C_1'$.

$$\max_{j=1,2,\ldots,n, i=1,2,\ldots,k} |\gamma_{ji}| = o(n^{-1/3} \times \log^{-3/2}(n))$$
\[ C_2'. \]

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

Then we have

\[ \sup_{x \in [0,\infty)} |\text{Prob}(\max_{i=1,2,\ldots,k} \left| \sum_{j=1}^{n} \gamma_{ji} \epsilon_{j} \right| \leq x) - \text{Prob}(\max_{i=1,2,\ldots,k} \left| \sum_{j=1}^{n} \gamma_{ji}^* \epsilon_{j}^* \right| \leq x)| = o(1) \]

**Proof of Lemma 4.** In this proof we define \( \Gamma = (\gamma_1, \ldots, \gamma_k) \) with \( \gamma_i = (\gamma_{1i}, \gamma_{2i}, \ldots, \gamma_{ni})^T \in \mathbb{R}^n \). For \( i = 1, 2, \ldots, k, \gamma_i^T \epsilon = \sum_{j=1}^{n} \gamma_{ji} \epsilon_j \). From lemma A.2 and (8) in Chernozhukov, Chetverikov and Kato (2013), and (S1) to (S5) in Xu, Zhang and Wu (2019), for \( x = (x_1, \ldots, x_n) \) and \( y, z \in \mathbb{R} \), define

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

Here \( \beta, \psi > 0 \). Then \( g_{\psi,z} \in C^3 \) is nonincreasing function. \( g_0 = 1 \) with \( y \leq 0 \), \( 0 \) with \( y \geq 1 \), and

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

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

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

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

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

For any given \( x = (x_1, \ldots, x_n) \in \mathbb{R}^n \), define function
\( G_\beta(x) = \frac{1}{\beta} \log(\sum_{i=1}^{n} \exp(\beta x_i) + \sum_{i=1}^{n} \exp(-\beta x_i)) = F_\beta(x_1, \ldots, x_n, -x_1, \ldots, -x_n). \) From (57) and (46), for \( i, j, k = 1, \ldots, n \)

(58)

\[
G_\beta(x) - \frac{\log(2n)}{\beta} \leq \max_{i=1, \ldots, n} |x_i| \leq G_\beta(x)
\]

\[
\frac{\partial G_\beta}{\partial x_i} = \frac{\partial F_\beta}{\partial x_i} - \frac{\partial F_\beta}{\partial x_{i+n}} \Rightarrow \sum_{i=1}^{n} \left| \frac{\partial G_\beta}{\partial x_i} \right| \leq \sum_{i=1}^{n} \frac{\partial F_\beta}{\partial x_i} + \frac{\partial F_\beta}{\partial x_{i+n}} = 1
\]

\[
\frac{\partial^2 G_\beta}{\partial x_i \partial x_j} = \frac{\partial^2 F_\beta}{\partial x_i \partial x_j} - \frac{\partial^2 F_\beta}{\partial x_{i+n} \partial x_j} - \frac{\partial^2 F_\beta}{\partial x_i \partial x_{j+n}} + \frac{\partial^2 F_\beta}{\partial x_{i+n} \partial x_j} + \frac{\partial^2 F_\beta}{\partial x_i \partial x_{j+n}}
\]

\[
\Rightarrow \sum_{i=1}^{n} \sum_{j=1}^{n} \left| \frac{\partial^2 G_\beta}{\partial x_i \partial x_j} \right| \leq \sum_{i=1}^{n} \sum_{j=1}^{n} \left| \frac{\partial^2 F_\beta}{\partial x_i \partial x_j} \right| \leq 2\beta
\]

Define \( h_{\beta, \psi, x}(x_1, \ldots, x_n) = g_{\psi, x}(G_\beta(x_1, \ldots, x_n)) \). Direct calculation shows \( \frac{\partial h_{\beta, \psi, x}(x_1, \ldots, x_n)}{\partial x_i} = g_{\psi, x}(G_\beta(x_1, \ldots, x_n)) \frac{\partial G_\beta}{\partial x_i} \frac{\partial G_\beta}{\partial x_j} \)

(59)

\[
\frac{\partial^2 h_{\beta, \psi, x}(x_1, \ldots, x_n)}{\partial x_i \partial x_j} = g''_{\psi, x}(G_\beta(x_1, \ldots, x_n)) \frac{\partial G_\beta}{\partial x_i} \frac{\partial G_\beta}{\partial x_j} + g'_{\psi, x}(G_\beta(x_1, \ldots, x_n)) \frac{\partial^2 G_\beta}{\partial x_i \partial x_j} \]

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

and

\[
\frac{\partial^3 h_{\beta, \psi, x}(x_1, \ldots, x_n)}{\partial x_i \partial x_j \partial x_k} = g'''_{\psi, x}(G_\beta(x_1, \ldots, x_n)) \frac{\partial G_\beta}{\partial x_i} \frac{\partial G_\beta}{\partial x_j} \frac{\partial G_\beta}{\partial x_k} + g''_{\psi, x}(G_\beta(x_1, \ldots, x_n)) \frac{\partial^2 G_\beta}{\partial x_i \partial x_j} \frac{\partial G_\beta}{\partial x_k} + g''_{\psi, x}(G_\beta(x_1, \ldots, x_n)) \frac{\partial^2 G_\beta}{\partial x_i \partial x_k} \frac{\partial G_\beta}{\partial x_j} + g''_{\psi, x}(G_\beta(x_1, \ldots, x_n)) \frac{\partial^2 G_\beta}{\partial x_j \partial x_k} \frac{\partial G_\beta}{\partial x_i} + g''_{\psi, x}(G_\beta(x_1, \ldots, x_n)) \frac{\partial^2 G_\beta}{\partial x_k \partial x_i} \frac{\partial G_\beta}{\partial x_j} + g''_{\psi, x}(G_\beta(x_1, \ldots, x_n)) \frac{\partial^2 G_\beta}{\partial x_k \partial x_j} \frac{\partial G_\beta}{\partial x_i} + g''_{\psi, x}(G_\beta(x_1, \ldots, x_n)) \frac{\partial^2 G_\beta}{\partial x_i \partial x_k} \frac{\partial G_\beta}{\partial x_j}
\]

\[
\Rightarrow \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \left| \frac{\partial^3 h_{\beta, \psi, x}(x_1, \ldots, x_n)}{\partial x_i \partial x_j \partial x_k} \right| \leq g_{\psi, x} \left( \sum_{i=1}^{n} \left| \frac{\partial G_\beta}{\partial x_i} \right| \right)^3 + 3g_{\psi, x}^2 \left( \sum_{i=1}^{n} \sum_{j=1}^{n} \left| \frac{\partial^2 G_\beta}{\partial x_i \partial x_j} \right| \right) \times \left( \sum_{k=1}^{n} \left| \frac{\partial G_\beta}{\partial x_k} \right| \right) + g_{\psi, x}^3 \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \left| \frac{\partial^3 G_\beta}{\partial x_i \partial x_j \partial x_k} \right| \leq g_{\psi, x}^3 + 6g_{\psi, x}^2 \beta + 6g_{\psi, x} \beta^2
\]

Define \( \xi = (\xi_1, \ldots, \xi_n) \) as i.i.d. random variables with the same marginal distribution as \( \epsilon_1, \epsilon^* \), and is independent of \( \epsilon, \epsilon^* \). Therefore, \( \text{Prob}(\max_{i=1, \ldots, k} |\gamma_1^T \xi| \leq x) = \text{Prob}^* \left( \max_{i=1, \ldots, k} |\gamma_i^T \xi| \leq x \right) \) for any \( x \). Since \( \gamma_i^T \xi \leq \mathbf{E}^* \left( \sum_{i=1}^{n} \frac{\gamma_i^T \xi}{\sigma} \right)^2 = \sum_{i=1}^{n} \gamma_i^2 \leq C^2 \) for \( i = 1, 2, \ldots, k \). According to (46), (58) and lemma 3, \( \exists \) a constant \( C \) which only depends on \( c_T \) and \( C_T \) such that for any
given $\psi, \beta, \sigma > 0$,

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

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

\[
\leq C \times \left( \frac{1}{\psi\sigma} + \frac{\log(2k)}{\beta\sigma} \right) \times \left( 1 + \sqrt{\log(k) + \sqrt{\log \left( \frac{1}{\psi\sigma} + \frac{\log(2k)}{\beta\sigma} \right)}} \right).
\]

Define $z = C \times \left( \frac{1}{\psi\sigma} + \frac{\log(2k)}{\beta\sigma} \right) \times \left( 1 + \sqrt{\log(k) + \sqrt{\log \left( \frac{1}{\psi\sigma} + \frac{\log(2k)}{\beta\sigma} \right)}} \right)$. For any $x \geq 0$,

\[
\text{Prob}^*(\max_{i=1,2,\ldots,k} |\gamma_i^T e| \leq x) - \text{Prob}^*(\max_{i=1,2,\ldots,k} |\gamma_i^T e^*| \leq x)
\]

\[
\leq \text{Prob}^*(\max_{i=1,2,\ldots,k} |\gamma_i^T \xi| \leq x) - \text{Prob}^*(\max_{i=1,2,\ldots,k} |\gamma_i^T e^*| \leq x + \frac{1}{\psi} + \frac{\log(2k)}{\beta}) + z
\]

\[
\leq \text{Prob}(G_{\beta}(\gamma_1^T \xi, \ldots, \gamma_k^T \xi) \leq x + \frac{\log(2k)}{\beta}) - \text{Prob}(G_{\beta}(\gamma_1^T e^*, \ldots, \gamma_k^T e^*) \leq x + \frac{1}{\psi} + \frac{\log(2k)}{\beta}) + z
\]

\[
\leq \mathbb{E}^* h_{\beta,\psi,x} + \frac{\log(2k)}{\beta} (\gamma_1^T \xi, \ldots, \gamma_k^T \xi) - \mathbb{E}^* h_{\beta,\psi,x} + \frac{\log(2k)}{\beta} (\gamma_1^T e^*, \ldots, \gamma_k^T e^*) + z
\]

\[
\text{Prob}^*(\max_{i=1,2,\ldots,k} |\gamma_i^T e| \leq x) - \text{Prob}^*(\max_{i=1,2,\ldots,k} |\gamma_i^T e^*| \leq x)
\]

\[
\geq \text{Prob}^*(\max_{i=1,2,\ldots,k} |\gamma_i^T \xi| \leq x) - \text{Prob}^*(\max_{i=1,2,\ldots,k} |\gamma_i^T e^*| \leq x - \frac{1}{\psi} - \frac{\log(2k)}{\beta}) - z
\]

\[
\geq \text{Prob}(G_{\beta}(\gamma_1^T \xi, \ldots, \gamma_k^T \xi) \leq x) - \text{Prob}(G_{\beta}(\gamma_1^T e^*, \ldots, \gamma_k^T e^*) \leq x - \frac{1}{\psi}) - z
\]

\[
\geq \mathbb{E}^* h_{\beta,\psi,x - \frac{1}{\psi}} (\gamma_1^T \xi, \ldots, \gamma_k^T \xi) - \mathbb{E}^* h_{\beta,\psi,x - \frac{1}{\psi}} (\gamma_1^T e^*, \ldots, \gamma_k^T e^*) - z
\]

Therefore, we have

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

\[
\leq z + \sup_{x \in \mathbb{R}} \left| \mathbb{E}^* h_{\beta,\psi,x} (\gamma_1^T \xi, \ldots, \gamma_k^T \xi) - \mathbb{E}^* h_{\beta,\psi,x} (\gamma_1^T e^*, \ldots, \gamma_k^T e^*) \right|
\]

For any $i = 1, 2, \ldots, k, j = 1, 2, \ldots, n$, define $H_{ij} = \sum_{s=j}^{j-1} \gamma_{si} e_s + \sum_{s=j+1}^{n} \gamma_{se_s}$, $m_{ij} = \gamma_{ji} e_j$, and $m_{ij}^* = \gamma_{ji} e_j^*$, we have $H_{ij} + m_{ij} = H_{ij+1} + m_{ij+1}$, and

\[
\sup_{x \in \mathbb{R}} \left| \mathbb{E}^* h_{\beta,\psi,x} (\gamma_1^T \xi, \ldots, \gamma_k^T \xi) - \mathbb{E}^* h_{\beta,\psi,x} (\gamma_1^T e^*, \ldots, \gamma_k^T e^*) \right|
\]

\[
= \sup_{x \in \mathbb{R}} \left| \sum_{s=1}^{n} \mathbb{E}^* h_{\beta,\psi,x} (H_{1s} + m_{1s}, \ldots, H_{ks} + m_{ks}) - \mathbb{E}^* h_{\beta,\psi,x} (H_{1s} + m_{1s}^*, \ldots, H_{ks} + m_{ks}^*) \right|
\]

\[
\leq \sum_{s=1}^{n} \sup_{x \in \mathbb{R}} \left| \mathbb{E}^* h_{\beta,\psi,x} (H_{1s} + m_{1s}, \ldots, H_{ks} + m_{ks}) - \mathbb{E}^* h_{\beta,\psi,x} (H_{1s} + m_{1s}^*, \ldots, H_{ks} + m_{ks}^*) \right|
\]
Since $E(\xi_s|\epsilon, \xi_b, \epsilon^*_b, b \neq s) = E(\epsilon^*_s|\epsilon, \xi_b, \epsilon^*_b, b \neq s) = 0$, $E(\xi^2_s - \epsilon^2_s|\epsilon, \xi_b, \epsilon^*_b, b \neq s) = \sigma^2 - \bar{\sigma}^2$, from multivariate Taylor's theorem and (59), for any $s = 1, 2, \ldots, n$ and $x \in \mathbb{R}$,

\[ |E(h_{j, \psi, x}(H_{1s} + m_{1s}, \ldots, H_{ks} + m_{ks}) - h_{j, \psi, x}(H_{1s} + m_{1s}^*, \ldots, H_{ks} + m_{ks}^*))| \leq \sum_{i=1}^{k} \frac{\partial h_{j, \psi, x}(H_{1s}, \ldots, H_{ks})}{\partial x_i} |\gamma_{si}| \left( E(\epsilon^*_s - \epsilon^*_s|\epsilon, \xi_b, \epsilon^*_b, b \neq s) \right) \]

\[ + \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{\partial^2 h_{j, \psi, x}(H_{1s}, \ldots, H_{ks})}{\partial x_i \partial x_j} |\gamma_{si} \gamma_{sj}| E(\epsilon^*_s - \epsilon^*_s|\epsilon, \xi_b, \epsilon^*_b, b \neq s) \]

\[ + (g_\ast \psi^2 + g_\ast \psi^2 \beta + g_\ast \psi^2 \beta^2) \max_{i=1,2,\ldots,k} |\gamma_{si}|^3 \times (E|\epsilon_1|^3 + \bar{\sigma}^3 D) \]

\[ \Rightarrow \sup_{x \in \mathbb{R}} |E(h_{j, \psi, x}(H_{1s} + m_{1s}, \ldots, H_{ks} + m_{ks}) - h_{j, \psi, x}(H_{1s} + m_{1s}^*, \ldots, H_{ks} + m_{ks}^*))| \leq g_\ast (\psi^2 + \psi \beta) \sigma^2 - \bar{\sigma}^2 \times \max_{i=1,2,\ldots,k} \gamma_{si}^2 \times (E|\epsilon_1|^3 + \bar{\sigma}^3 D) \times g_\ast (\psi^3 + \psi^2 \beta + \psi \beta^2) \times \max_{i=1,2,\ldots,k} |\gamma_{si}|^3 \]

Here $D = E|Y|^3$ with $Y$ having normal distribution with mean 0 and variance 1. Then

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

\[ \leq z + (g_\ast \psi^2 + g_\ast \psi \beta) \sigma^2 - \bar{\sigma}^2 \times \sum_{s=1}^{n} \max_{i=1,2,\ldots,k} \gamma_{si}^2 \]

\[ + (E|\epsilon_1|^3 + \bar{\sigma}^3 D) \times g_\ast (\psi^3 + \psi^2 \beta + \psi \beta^2) \times \sum_{s=1}^{n} \max_{i=1,2,\ldots,k} |\gamma_{si}|^3 \]

In particular, for any given $\delta > 0$, choose $\psi = \beta = \log^{3/2}(n)/\delta^{1/4}$ and suppose $\frac{3\sigma}{2} > \bar{\sigma} > \frac{\sigma}{2}$. For sufficiently large $n$ we have $\frac{1}{\psi^\sigma} + \frac{\log(2k)}{\psi \sigma} \leq \frac{\log(n)}{\psi \sigma} \leq \frac{4\delta^{1/4}}{\sigma \log(n)} < 1$ and

\[ z \leq \frac{4C \log(n)}{\psi \sigma} \times \left( 2\sqrt{\log(n)} + \sqrt{\log(\psi \bar{\sigma})} \right) \leq \frac{4C \delta^{1/4}}{\sigma} \left( 2 + \frac{\log(\log(n)) + \log(3\sigma/2\delta^{1/4})}{\log(n)} \right) \leq C' \delta^{1/4} \]

Here $C' = \frac{12C}{\sigma}$. Suppose condition C1. For any $1 > \delta > 0$, $\exists D_\delta > 0$ such that for sufficiently large $n$,

\[ \text{Prob}(|\sigma^2 - \bar{\sigma}^2| \leq D_\delta \times n^{-a_\sigma}) > 1 - \delta, \max_{j=1,2,\ldots,n} \gamma_{ji} < \delta \times n^{(a_\sigma - 1)/2} \times \log^{-3/2}(n), \]

and $\max_{j=1,2,\ldots,n} \gamma_{ji} < \delta \times n^{-1/3} \times \log^{-3/2}(n)$.
Choose \( \psi = \beta = \log^{3/2}(n)/\delta^{1/4} \). According to (65), for sufficiently large \( n \), (67) happens and \( \frac{1}{2} \sigma < \hat{\sigma} < \frac{3}{2} \sigma \) with probability \( 1 - \delta \). If (67) happens, (68)

\[
\sup_{x \in [0, \infty)} |\text{Prob}(\max_{i=1,2,\ldots,k} |\gamma_i^T e| \leq x) - \text{Prob}^*(\max_{i=1,2,\ldots,k} |\gamma_i^T e^*| \leq x)| \\
\leq C' \delta^{1/4} + 2g_* \psi^2 \times D_\delta \times n^{-\alpha_*} \times \frac{\delta^2 \times n^{-\alpha_*}}{\log^3(n)} + (E|\epsilon_1|^3 + \frac{27D}{8} \sigma^3) \times 3g_* \psi^3 \times \delta^3 \times n \times \frac{1}{n \log^{9/2}(n)} \\
= C' \delta^{1/4} + 2g_* D_\delta \delta^{3/2} + 3g_* (E|\epsilon_1|^3 + \frac{27D}{8} \sigma^3) \times \delta^{9/4}
\]

For \( \delta > 0 \) can be arbitrarily small, we prove (52).

Suppose condition C2. For any \( \delta > 0 \), there exists \( D_\delta > 0 \) such that for sufficiently large \( n \)

\[
\text{Prob}( |\sigma^2 - \hat{\sigma}^2| \leq D_\delta \times n^{-\alpha_*}) \geq 1 - \delta, \quad k \leq \frac{\delta n^{\alpha_*}}{\log^3(n)}, \quad \max_{i=1,2,\ldots,k} \sum_{j=1}^n \gamma_{ji}^2 \leq D_\delta \\
\text{and} \quad \max_{j=1,2,\ldots,n, i=1,2,\ldots,k} |\gamma_{ji}| \leq \frac{D_\delta \times n^{-\alpha_*}}{\log^3(n)}
\]

Since (70)

\[
\sum_{j=1}^n \max_{i=1,\ldots,k} \gamma_{ji}^3 \leq \max_{j=1,2,\ldots,n, i=1,2,\ldots,k} |\gamma_{ji}| \times \sum_{j=1}^n \max_{i=1,\ldots,k} \gamma_{ji}^2 \leq k D_\delta \\
\sum_{j=1}^n \gamma_{ji}^3 \leq \sum_{j=1}^n \gamma_{ji}^2 \leq k D_\delta \times \max_{j=1,2,\ldots,n, i=1,2,\ldots,k} |\gamma_{ji}|
\]

If (69) happens, by choosing \( \psi = \beta = \log^{3/2}(n)/\delta^{1/4} \)

\[
\sup_{x \in [0, \infty)} |\text{Prob}(\max_{i=1,2,\ldots,k} |\gamma_i^T e| \leq x) - \text{Prob}^*(\max_{i=1,2,\ldots,k} |\gamma_i^T e^*| \leq x)| \\
\leq C' \delta^{1/4} + 2g_* \psi^2 D_\delta n^{-\alpha_*} \times k D_\delta + (E|\epsilon_1|^3 + \frac{27D}{8} \sigma^3) \times 3g_* \psi^3 \times k D_\delta \max_{j=1,2,\ldots,n, i=1,2,\ldots,k} |\gamma_{ji}| \\
\leq C' \delta^{1/4} + 2g_* D_\delta^2 \times \frac{\log^3(n)}{\delta^{1/2}} \times \frac{\delta n^{\alpha_*}}{\log^3(n)} \times n^{-\alpha_*} \\
+ 3(E|\epsilon_1|^3 + \frac{27D}{8} \sigma^3) g_* D_\delta^2 \times \frac{\log^{9/2}(n)}{\delta^{3/4}} \times \frac{\delta n^{\alpha_*}}{\log^3(n)} \times \frac{n^{-\alpha_*}}{\log^{3/2}(n)} \\
= C' \delta^{1/4} + 2g_* D_\delta^2 \delta^{1/2} + 3(E|\epsilon_1|^3 + \frac{27D}{8} \sigma^3) g_* D_\delta^2 \times \delta^{1/4}
\]

and we prove (52).

If \( \hat{\sigma} = \sigma \). We choose \( \psi = \beta = \log^{3/2}(n)/\delta^{1/4} \), (65) can be modified to

\[
\sup_{x \in [0, \infty)} |\text{Prob}(\max_{i=1,2,\ldots,k} |\gamma_i^T e| \leq x) - \text{Prob}(\max_{i=1,2,\ldots,k} |\gamma_i^T e^*| \leq x)| \\
\leq C' \delta^{1/4} + (E|\epsilon_1|^3 + D \sigma^3) g_* \psi^2 + \psi \beta + \beta^2 \sum_{s=1}^n \max_{i=1,2,\ldots,k} |\gamma_{si}^3|
\]

(72)
Suppose condition $C1'$. For any $\delta > 0$ and sufficiently large $n$, $\max_{j=1,2,...,n,i=1,2,...,k} |\gamma_{ji}| \leq \delta \times n^{-1/3} \log^{-3/2}(n)$,

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

(73) $$\leq C' \delta^{1/4} + 3(\mathbb{E}|\epsilon_1|^3 + D\sigma^3)g_* \times \delta^{9/4}$$

and we prove (55).

Suppose condition $C2'$. For any $\delta > 0$ and sufficiently large $n$, $k \times \max_{j=1,2,...,n,i=1,2,...,k} |\gamma_{ji}| \leq \delta \log^{-9/2}(n)$. According to (70), for sufficiently large $n$ we have

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

(74) $$\leq C' \delta^{1/4} + 3(\mathbb{E}|\epsilon_1|^3 + D\sigma^3)g_* D_\delta \times \delta^{1/4}$$

and we prove (55).

Condition $C1'$ implies $C1$, and condition $C2'$ implies $C2'$. The additional proportions in $C1$ and $C2$ accommodate the error introduced in estimating errors’ variance $\sigma^2$. Condition $C1$ is designed for the situation when the number of linear combinations $k$ is as large as the sample size $n$; and condition $C2$ can be used when $k$ is significantly smaller than $n$.

The difference between lemma 4 and the classical central limit theorem is that $k$ can grow as $n$ increases. The maximum $\max_{i=1,2,...,k} | \sum_{j=1}^{n} \gamma_{ji} \epsilon_j |$ does not have an asymptotic distribution if $k \rightarrow \infty$. However, if the random variables are mixed well, approximating the distribution of $\max_{i=1,2,...,k} | \sum_{j=1}^{n} \gamma_{ji} \epsilon_j |$ by the distribution of the maximum of normal random variables is still applicable. With the help of lemma 4, we can establish the normal approximation theorem and construct the simultaneous confidence region for $\hat{\gamma}$(defined in (17)).

APPENDIX B: PROOFS OF THEOREMS IN SECTION 3

This section applies notations in section 2.

PROOF OF THEOREM 1. From (16),

$$\text{Prob}(N_{b_n} \neq N_{b_n}) \leq \text{Prob}(\min_{i \in \mathcal{N}_{b_n}} |\tilde{\theta}_i| \leq b_n) + \text{Prob}(\max_{i \notin \mathcal{N}_{b_n}} |\tilde{\theta}_i| > b_n)$$

(75) $$\leq \text{Prob}\left(\min_{i \in \mathcal{N}_{b_n}} |\tilde{\theta}_i| - \max_{i \notin \mathcal{N}_{b_n}} \rho_n^{2} \sum_{j=1}^{r} q_{ij} \frac{\lambda_j}{(\lambda_j^2 + \rho_n^2)^2} - \max_{i \notin \mathcal{N}_{b_n}} \sum_{j=1}^{r} q_{ij} \frac{\lambda_j}{\lambda_j^2 + \rho_n^2} \left(\frac{\rho_n \lambda_j}{(\lambda_j^2 + \rho_n^2)^2}\right) \sum_{l=1}^{n} p_{lj} \epsilon_l \leq b_n\right)$$

+ $$\text{Prob}\left(\max_{i \notin \mathcal{N}_{b_n}} |\tilde{\theta}_i| + \max_{i \notin \mathcal{N}_{b_n}} \rho_n^{2} \sum_{j=1}^{r} q_{ij} \frac{\lambda_j}{(\lambda_j^2 + \rho_n^2)^2} + \max_{i \notin \mathcal{N}_{b_n}} \sum_{j=1}^{r} q_{ij} \frac{\lambda_j}{\lambda_j^2 + \rho_n^2} \left(\frac{\rho_n \lambda_j}{(\lambda_j^2 + \rho_n^2)^2}\right) \sum_{l=1}^{n} p_{lj} \epsilon_l > b_n\right)$$
From Cauchy inequality,
(76)
\[
\max_{i=1,2,\ldots,p} \rho_n^2 \left| \sum_{j=1}^{r} q_{ij} \xi_j \right|^2 \leq \max_{i=1,2,\ldots,p} \rho_n^2 \left[ \sum_{j=1}^{r} q_{ij}^2 \right] \times \left( \sum_{j=1}^{r} \xi_j^2 \right) = O(n^{\alpha_\theta - 2\delta})
\]
\[
\max_{i=1,2,\ldots,p} \sum_{l=1}^{n} \left( \sum_{j=1}^{r} q_{ij} \left( \frac{\lambda_j}{\lambda_j^2 + \rho_n} + \frac{\rho_n \lambda_j}{(\lambda_j^2 + \rho_n)^2} \right) p_{lj} \right)^2 = \max_{i=1,2,\ldots,p} \sum_{j=1}^{r} q_{ij}^2 \left( \frac{\lambda_j}{\lambda_j^2 + \rho_n} + \frac{\rho_n \lambda_j}{(\lambda_j^2 + \rho_n)^2} \right)^2 \leq \max_{i=1,2,\ldots,p} \frac{4 \sum_{j=1}^{r} q_{ij}^2}{\lambda_r^2}
\]

Therefore, for sufficiently large \( n \), from assumption 4 and lemma 2
(77)
\[
\min_{i \in N_{b_n}} |\theta_i| - \max_{i \notin N_{b_n}} \rho_n^2 \left| \sum_{j=1}^{r} q_{ij} \xi_j \right|^2 - b_n > \frac{1}{2} \left( \frac{1}{c_b} - 1 \right) b_n
\]
\[
b_n - \max_{i \notin N_{b_n}} |\theta_i| - \max_{i \in N_{b_n}} \rho_n^2 \left| \sum_{j=1}^{r} q_{ij} \xi_j \right|^2 > \frac{1}{2} \left( 1 - c_b \right) b_n
\]
\[
\Rightarrow \text{Prob} \left( \tilde{N}_{b_n} \neq N_{b_n} \right) \leq \left| N_{b_n} \right| \times E \times 2^m \lambda_m n \times \left( \frac{1}{c_b} - 1 \right) b_n + \left( p - \left| N_{b_n} \right| \right) \times E \times 2^m \lambda_m \times \left( 1 - c_b \right) b_n = O(n^{\alpha_\theta + m\alpha_\theta - mn})
\]

and we prove (21).

Define \( \hat{\gamma} = M \hat{\theta} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_p)^T \) and \( \gamma = M \beta = (\gamma_1, \ldots, \gamma_p)^T \). For \( \beta = \theta + \theta_\perp \), if \( \tilde{N}_{b_n} = N_{b_n} \), (16) and (4) imply
(78)
\[
\max_{i=1,2,\ldots,p} \left| \hat{\gamma}_i - \gamma_i \right| = \max_{i=1,2,\ldots,p} \left| \sum_{j \in N_{b_n}} m_{ij} \hat{\theta}_j - \sum_{j \notin N_{b_n}} m_{ij} \theta_j - \sum_{j \in N_{b_n}} m_{ij} \theta_j - \sum_{j=1}^{p} m_{ij} \theta_{\perp j} \right|
\]
\[
\leq \max_{i=1,2,\ldots,p} \rho_n^2 \left| \sum_{k=1}^{r} c_{ik} \xi_k \right|^2 \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right) \sum_{l=1}^{p} \eta_{ik} \epsilon_l \left| \sum_{j \notin N_{b_n}} m_{ij} \theta_j + \sum_{j=1}^{p} m_{ij} \theta_{\perp j} \right|
\]
From (4) and assumption 5, if \( i \notin M \), then \( c_{ik} = 0 \) for \( k = 1, 2, \ldots, r \), so from Cauchy inequality and lemma 2, (79)

\[
\max_{i=1,2,\ldots,p_i} \rho_n^2 \left| \sum_{k=1}^{r} \frac{c_{ik}\zeta_k}{(\lambda_k^2 + \rho_n)^2} \right| \leq \max_{i \in M} \rho_n^2 \left( \sum_{k=1}^{r} c_{ik}^2 \right)^{1/2} \leq \sqrt{CM\rho_n^2} \times \frac{||\theta||_2}{\lambda^2} = O(n^{\alpha_\theta-2\delta})
\]

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

\[
\Rightarrow \text{Prob} \left( \max_{i=1,2,\ldots,p_i} \left| \sum_{k=1}^{r} c_{ik} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right) \sum_{l=1}^{n} p_l \epsilon_l \right| > \delta \right) \leq \frac{|M| \times E \times 2^n C_M^{n/2}}{\lambda^n \delta^m}
\]

for \( \forall \delta > 0 \)

\[
\Rightarrow \max_{i=1,2,\ldots,p_i} \left| \sum_{k=1}^{r} c_{ik} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right) \sum_{l=1}^{n} p_l \epsilon_l \right| = O_p(|M|^{1/m} \times n^{-\eta})
\]

Here \( E \) is the constant defined in lemma 2. Combine with assumption 2, assumption 5, and (77), we prove (22).

If \( \hat{N}_b = N_b^\prime \), since \( X \beta = X \theta \), we have (80)

\[
\hat{\sigma}^2 - \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} \left( \epsilon_i - \sum_{j \in N_b} x_{ij} (\widetilde{\theta}_j - \theta_j) + \sum_{j \notin N_b} x_{ij} \theta_j \right)^2 - \sigma^2
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2 - \sigma^2 + \frac{1}{n} \sum_{i=1}^{n} \left( \sum_{j \in N_b} x_{ij} (\widetilde{\theta}_j - \theta_j) \right)^2 - \frac{1}{n} \sum_{i=1}^{n} \sum_{j \in N_b} x_{ij} (\widetilde{\theta}_j - \theta_j) \sum_{j \notin N_b} x_{ij} \theta_j
\]

\[
+ \frac{2}{n} \sum_{i=1}^{n} \sum_{j \notin N_b} \epsilon_i x_{ij} \theta_j - \frac{2}{n} \sum_{i=1}^{n} \left( \sum_{j \in N_b} x_{ij} (\widetilde{\theta}_j - \theta_j) \right) \times \left( \sum_{j \notin N_b} x_{ij} \theta_j \right)
\]

From assumption 3, \( E (\frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2 - \sigma^2)^2 \leq \frac{2}{n} (E \epsilon^4 + \sigma^4) = O(1/n) \Rightarrow \frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2 - \sigma^2 = O_p(1/\sqrt{n}) \). For the second term, from assumption 1 and (76), (81)

\[
\frac{1}{n} \sum_{i=1}^{n} \left( \sum_{j \in N_b} x_{ij} (\widetilde{\theta}_j - \theta_j) \right)^2 \leq C_\lambda^2 \sum_{j \in N_b} (\widetilde{\theta}_j - \theta_j)^2
\]

\[
\leq 2C_\lambda^2 \sum_{j \in N_b} \left( \rho_n^4 \left( \sum_{k=1}^{r} q_{jk} \zeta_k \right)^2 \right) + \left( \sum_{k=1}^{r} q_{jk} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right) \right) \sum_{l=1}^{n} p_l \epsilon_l \right)^2
\]

\[
= O(|N_b| \times n^{2\alpha_\theta-4\delta}) + 2C_\lambda^2 \sum_{j \in N_b} \left( \rho_n^4 \left( \sum_{k=1}^{r} q_{jk} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right) \right) \right) \sum_{l=1}^{n} p_l \epsilon_l \right)^2
\]
Since
\[
E \sum_{j \in \mathcal{N}_b} \left( \sum_{k=1}^r q_{jk} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right) \sum_{l=1}^n p_{lk} \epsilon_l \right)^2
\]
\[
= \sigma^2 \sum_{j \in \mathcal{N}_b} \sum_{k=1}^n \left( \sum_{l=1}^r q_{jk} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right) p_{lk} \right)^2
\]
\[
= \sigma^2 \sum_{j \in \mathcal{N}_b} \sum_{k=1}^n q_{jk}^2 \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right)^2 \leq \frac{4\sigma^2 |\mathcal{N}_b|}{\lambda^2}
\]

We have \( \frac{1}{n} \sum_{i=1}^n \left( \sum_{j \in \mathcal{N}_b} x_{ij} (\bar{\theta}_j - \theta_j) \right)^2 \) = \( O_p(\sqrt{|\mathcal{N}_b|} \times n^{2\alpha - 4\delta} + |\mathcal{N}_b| \times n^{-2\eta}) \). For the third term, from assumption 6 we have
\[
\frac{1}{n} \sum_{i=1}^n \left( \sum_{j \notin \mathcal{N}_b} x_{ij} \theta_j \right)^2 \leq C^2_X \sum_{j \notin \mathcal{N}_b} \theta_j^2 \leq C^2_X \times b_n \sum_{j \notin \mathcal{N}_b} \theta_j = O(n^{-\alpha})
\]

For the fourth term, from Cauchy inequality and (81),
\[
\frac{1}{n} \sum_{i=1}^n \sum_{j \in \mathcal{N}_b} \epsilon_i x_{ij} (\bar{\theta}_j - \theta_j) \leq \frac{1}{n} \sqrt{\sum_{i=1}^n \epsilon_i^2} \times \sqrt{\sum_{i=1}^n \left( \sum_{j \in \mathcal{N}_b} x_{ij} (\bar{\theta}_j - \theta_j) \right)^2}
\]
\[
\leq \sqrt{\frac{\sigma^2}{n}} \sum_{i=1}^n \sum_{j \in \mathcal{N}_b} \epsilon_i x_{ij} (\bar{\theta}_j - \theta_j) \leq \sigma \times O(\sqrt{|\mathcal{N}_b|} \times n^{2\alpha - 4\delta} + |\mathcal{N}_b| \times n^{-2\eta})
\]
\[
\Rightarrow \frac{1}{n} \sum_{i=1}^n \sum_{j \in \mathcal{N}_b} \epsilon_i x_{ij} (\bar{\theta}_j - \theta_j) = O_p(\sqrt{|\mathcal{N}_b|} \times n^{\alpha - 2\delta} + |\mathcal{N}_b| \times n^{-\eta})
\]

For the fifth term,
\[
\frac{1}{n} \sum_{i=1}^n \sum_{j \notin \mathcal{N}_b} \epsilon_i x_{ij} \theta_j \leq \frac{\sigma^2}{n^2} \sum_{i=1}^n \left( \sum_{j \notin \mathcal{N}_b} x_{ij} \theta_j \right)^2 \leq \frac{\sigma^2 C^2_X}{n} \sum_{j \notin \mathcal{N}_b} \theta_j^2
\]
\[
\Rightarrow \frac{1}{n} \sum_{i=1}^n \sum_{j \notin \mathcal{N}_b} \epsilon_i x_{ij} \theta_j = O_p(n^{-\left(1+\alpha\right)/2})
\]

For the last term,
\[
\frac{1}{n} \sum_{i=1}^n \left( \sum_{j \in \mathcal{N}_b} x_{ij} (\bar{\theta}_j - \theta_j) \right) \times \left( \sum_{j \notin \mathcal{N}_b} x_{ij} \theta_j \right) \leq C^2_X \sqrt{\sum_{j \in \mathcal{N}_b} (\bar{\theta}_j - \theta_j)^2} \times \sqrt{\sum_{j \notin \mathcal{N}_b} \theta_j^2}
\]
\[
= O_p(\sqrt{|\mathcal{N}_b|} \times n^{\alpha - 2\delta - \alpha\eta/2} + |\mathcal{N}_b| \times n^{-\eta - \alpha\eta/2})
\]

From (21), \( \text{Prob}(\hat{\mathcal{N}}_b \neq \mathcal{N}_b) \to 0 \). So we have
\[
|\hat{\sigma}^2 - \sigma^2| = O_p\left( \frac{1}{\sqrt{n}} + \sqrt{|\mathcal{N}_b|} \times n^{\alpha - 2\delta} + \sqrt{|\mathcal{N}_b|} \times n^{-\eta} + n^{-\alpha\eta/2} \right)
\]

From assumption 2 and 6, we prove the second result.
Define $T = (c_{ik})_{i \in \mathcal{M}, k = 1, 2, \ldots, r}$. From assumption 7, since the matrix 
\[
\left( \frac{1}{\tau_i} c_{ik} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{\lambda_k^2 + \rho_n} \right) \right)_{i \in \mathcal{M}, j = 1, 2, \ldots, r} = D_1 T D_2 \] with $D_1 = \text{diag}(1/\tau_i, i \in \mathcal{M})$ and 
\[
D_2 = \text{diag} \left( \frac{\lambda_1}{\lambda_1^2 + \rho_n}, \ldots, \frac{\lambda_r}{\lambda_r^2 + \rho_n} \right), \] the matrix 
\[
\left( \frac{1}{\tau_i} c_{ik} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{\lambda_k^2 + \rho_n} \right) \right)_{i \in \mathcal{M}, j = 1, 2, \ldots, r}
\]
also has rank $|\mathcal{M}|$. The proof of theorem 2 uses this result.

**Proof of Theorem 2.** From Cauchy inequality and assumption 2, suppose $\delta = \frac{\eta + \alpha_n + \delta_1}{2}$ with $\delta_1 > 0$. For $i \in \mathcal{M}$,

\[
\left| \sum_{k=1}^{r} \frac{c_{ik} \zeta_k}{(\lambda_k^2 + \rho_n)^2} \right| \leq \sqrt{\sum_{k=1}^{r} \frac{c_{ik}^2 \lambda_k^2}{(\lambda_k^2 + \rho_n)^2}} \times \sqrt{\sum_{k=1}^{r} \frac{\zeta_k^2}{(\lambda_k^2 + \rho_n)^2}} \leq \tau_i \times \frac{|\theta|_2}{\lambda_1^3} \times \max_{i \in \mathcal{M}} \frac{\rho_n^2}{\tau_i} \sum_{k=1}^{r} \frac{c_{ik} \zeta_k}{(\lambda_k^2 + \rho_n)^2} = O(n^{-\delta_1})
\]

Define $t_{il} = \frac{1}{\tau_i} \times \sum_{k=1}^{r} c_{ik} p_{lk} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{\lambda_k^2 + \rho_n} \right)$ for $i \in \mathcal{M}$ and $l = 1, 2, \ldots, n$. From (16), (5), (78) and assumption 5, if $\hat{N}_b_z = N_{b_z}$, we have $\hat{\tau}_i = \tau_i \geq 1/\sqrt{n}$ and $\exists$ a constant $C > 0$, for any $a > 0$ and sufficiently large $n$,

\[
\max_{i=1, 2, \ldots, p_l} \frac{\hat{\gamma}_i - \gamma_i}{\hat{\tau}_i} \leq \max_{i \in \mathcal{M}} \frac{\rho_n^2}{\tau_i} \left| \sum_{k=1}^{r} \frac{c_{ik} \zeta_k}{(\lambda_k^2 + \rho_n)^2} \right| + \max_{i \in \mathcal{M}} \left| \sum_{l=1}^{n} t_{il} \epsilon_i \right| + \max_{i=1, 2, \ldots, p_l} \frac{|\sum_{j \notin N_{b_z}^i} m_{ij} \theta_j|}{\tau_i} + \max_{i=1, 2, \ldots, p_l} \frac{|\sum_{j=1}^{p} m_{ij} \theta_{-j}|}{\tau_i}
\]

\[
\leq \max_{i \in \mathcal{M}} \left| \sum_{l=1}^{n} t_{il} \epsilon_i \right| + Cn^{-\delta_1} + \frac{a}{\sqrt{\log(n)}}
\]

\[
\max_{i=1, 2, \ldots, p_l} \frac{\hat{\gamma}_i - \gamma_i}{\hat{\tau}_i} \geq \max_{i \in \mathcal{M}} \left| \sum_{l=1}^{n} t_{il} \epsilon_i \right| - \max_{i \in \mathcal{M}} \frac{\rho_n^2}{\tau_i} \left| \sum_{k=1}^{r} \frac{c_{ik} \zeta_k}{(\lambda_k^2 + \rho_n)^2} \right| - \max_{i=1, 2, \ldots, p_l} \frac{|\sum_{j \notin N_{b_z}^i} m_{ij} \theta_j|}{\tau_i} - \max_{i=1, 2, \ldots, p_l} \frac{|\sum_{j=1}^{p} m_{ij} \theta_{-j}|}{\tau_i}
\]

\[
\geq \max_{i \in \mathcal{M}} \left| \sum_{l=1}^{n} t_{il} \epsilon_i \right| - Cn^{-\delta_1} - \frac{a}{\sqrt{\log(n)}}
\]
According to theorem 1 and lemma 1, \( \exists \) a constant \( C \) and for any given \( a > 0 \), for sufficiently large \( n \) and any \( x \geq 0 \),

\[
\text{(90)}
\]

\[
\begin{align*}
\text{Prob} \left( \max_{i=1,2,\ldots,n} \frac{\gamma_i - \gamma_i}{\tilde{\sigma}_i} \leq x \right) & \leq \text{Prob} \left( \max_{i=1,2,\ldots,n} \frac{\gamma_i - \gamma_i}{\tilde{\sigma}_i} \leq x \cap \tilde{N}_n = N_n \right) + \text{Prob} \left( \tilde{N}_n \neq N_n \right) \\
& \leq \text{Prob} \left( \max_{i \in M} \sum_{l=1}^{n} t_{il} \epsilon_l \leq x + Cn^{-\delta_1} + \frac{a}{\sqrt{\log(n)}} \right) + Cn^{\alpha_p + \mu_n - mn} \\
& \leq \text{Prob} \left( \max_{i \in M} \sum_{l=1}^{n} t_{il} \epsilon_l^* \leq x \right) + Cn^{\alpha_p + \mu_n - mn} \\
& \quad + \sup_{x \geq 0} |\text{Prob} \left( \max_{i \in M} \sum_{l=1}^{n} t_{il} \epsilon_l \leq x \right) - \text{Prob} \left( \max_{i \in M} \sum_{l=1}^{n} t_{il} \epsilon_l \leq x \right)| \\
& \quad + \sup_{x \in \mathbb{R}} \left( \text{Prob} \left( \max_{i \in M} \sum_{l=1}^{n} t_{il} \epsilon_l \leq x + Cn^{-\delta_1} + \frac{a}{\sqrt{\log(n)}} \right) - \text{Prob} \left( \max_{i \in M} \sum_{l=1}^{n} t_{il} \epsilon_l \leq x \right) \right) \\
& \quad \geq \text{Prob} \left( \max_{i \in M} \sum_{l=1}^{n} t_{il} \epsilon_l \leq x - Cn^{-\delta_1} - \frac{a}{\sqrt{\log(n)}} \right) - \text{Prob} \left( \tilde{N}_n \neq N_n \right) \\
& \quad \geq \text{Prob} \left( \max_{i \in M} \sum_{l=1}^{n} t_{il} \epsilon_l^* \leq x \right) - Cn^{\alpha_p + \mu_n - mn} \\
\end{align*}
\]

\[
\begin{align*}
& \quad - \sup_{x \in \mathbb{R}} \left( \text{Prob} \left( \max_{i \in M} \sum_{l=1}^{n} t_{il} \epsilon_l \leq x \right) - \text{Prob} \left( \max_{i \in M} \sum_{l=1}^{n} t_{il} \epsilon_l \leq x - Cn^{-\delta_1} - \frac{a}{\sqrt{\log(n)}} \right) \right) \\
& \quad - |\text{Prob} \left( \max_{i \in M} \sum_{l=1}^{n} t_{il} \epsilon_l \leq x - Cn^{-\delta_1} - \frac{a}{\sqrt{\log(n)}} \right) - \text{Prob} \left( \max_{i \in M} \sum_{l=1}^{n} t_{il} \epsilon_l \leq x - Cn^{-\delta_1} - \frac{a}{\sqrt{\log(n)}} \right)| \\
\end{align*}
\]

From assumption 1, 2, 5 and 7, for sufficiently large \( n \) we have

\[
\text{(91)}
\]

\[
\begin{align*}
\text{max}_{i \in M} \left( \sum_{l=1}^{n} t_{il} \epsilon_l^* \right)^2 &= \sigma^2 \text{max}_{i \in M} \sum_{l=1}^{n} t_{il}^2 = \sigma^2 \text{max}_{i \in M} \sum_{k=1}^{r} \left( \frac{\lambda_k}{\lambda_k + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k + \rho_n)^2} \right)^2 \\
\text{min}_{i \in M} \left( \sum_{l=1}^{n} t_{il} \epsilon_l^* \right)^2 &= \sigma^2 \text{min}_{i \in M} \frac{1}{1 + \frac{1}{n \sum_{k=1}^{c_k} \left( \frac{\lambda_k}{\lambda_k + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k + \rho_n)^2} \right)^2}} \\
& \geq \sigma^2 \text{min}_{i \in M} \frac{1}{1 + \frac{1}{n \sum_{k=1}^{c_k} \left( \frac{\lambda_k^2}{\lambda_k^2 + \rho_n^2} \right)}} \geq \frac{\sigma^2}{4C \epsilon_{c_M}} > 0 \\
\end{align*}
\]

and \((t_{il})_{i \in M, l=1,2,\ldots,n} = D_1 T D_2 P^T\), here \(T = (c_{ik})_{i \in M, k=1,2,\ldots,r}\), \(D_1 = \text{diag}(1/\tau_i, i \in M)\), and
\[ D_2 = \text{diag} \left( \frac{\lambda_1}{\lambda_1 + \rho_n}, \frac{\rho_n \lambda_1}{(\lambda_1 + \rho_n)^2}, \ldots, \frac{\lambda_r}{\lambda_r + \rho_n}, \frac{\rho_n \lambda_r}{(\lambda_r + \rho_n)^2} \right). \]

So \((t_{il})_{i \in \mathcal{M}, l=1,2,\ldots,n}\) has full rank \(|\mathcal{M}|\). From lemma 3, \(\exists\) a constant \(C'\) which only depends on \(\sigma, c, C_\lambda\) such that

\[
\sup_{x \in \mathbb{R}} \left( \text{Prob} \left( \max_{i \in \mathcal{M}} \left| \sum_{l=1}^{n} t_{il} \epsilon_l^* \right| \leq x + Cn^{-\delta_1} + \frac{a}{\sqrt{\log(n)}} \right) - \text{Prob} \left( \max_{i \in \mathcal{M}} \left| \sum_{l=1}^{n} t_{il} \epsilon_l^* \right| \leq x \right) \right) \\
\leq C' \left( Cn^{-\delta_1} + \frac{a}{\sqrt{\log(n)}} \right) \times \left( 1 + \sqrt{\log(|\mathcal{M}|)} + \sqrt{\log(Cn^{-\delta_1} + \frac{a}{\sqrt{\log(n)}})} \right)
\]

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

\[
\Rightarrow \sup_{x \in \mathbb{R}} \left( \text{Prob} \left( \max_{i \in \mathcal{M}} \left| \sum_{l=1}^{n} t_{il} \epsilon_l^* \right| \leq x + Cn^{-\delta_1} + \frac{a}{\sqrt{\log(n)}} \right) - \text{Prob} \left( \max_{i \in \mathcal{M}} \left| \sum_{l=1}^{n} t_{il} \epsilon_l^* \right| \leq x \right) \right) \\
\leq C' \left( Cn^{-\delta_1} + \frac{a}{\sqrt{\log(n)}} \right) \times \left( 1 + \sqrt{\log(n)} + \sqrt{\log(|\mathcal{M}|)} \right) \leq 6C'a
\]

From assumption 7, (91) and lemma 4, we have

\[
\sup_{x \geq 0} |\text{Prob} \left( \max_{i \in \mathcal{M}} \left| \sum_{l=1}^{n} t_{il} \epsilon_l^* \right| \leq x \right) - \text{Prob} \left( \max_{i \in \mathcal{M}} \left| \sum_{l=1}^{n} t_{il} \epsilon_l^* \right| \leq x \right) | < a \text{ for sufficiently large } n
\]

If \(x < Cn^{-\delta_1} + \frac{a}{\sqrt{\log(n)}}\), then \(\text{Prob} \left( \max_{i \in \mathcal{M}} \left| \sum_{l=1}^{n} t_{il} \epsilon_l^* \right| \leq x - Cn^{-\delta_1} - \frac{a}{\sqrt{\log(n)}} \right) = 0\) and

\[
\text{Prob} \left( \max_{i \in \mathcal{M}} \left| \sum_{l=1}^{n} t_{il} \epsilon_l^* \right| \leq x - Cn^{-\delta_1} - \frac{a}{\sqrt{\log(n)}} \right) = 0. \text{ Combine with (90) to (94), we have}
\]

\[
\sup_{x \geq 0} |\text{Prob} \left( \max_{i=1,2,\ldots,p} \frac{\hat{G}_i - \gamma_i}{\tau_i} \leq x \right) - \text{Prob} \left( \max_{i \in \mathcal{M}} \left| \sum_{l=1}^{n} t_{il} \epsilon_l^* \right| \leq x \right) | \leq Cn^{\alpha p + mn - mn} + 6C'a + a
\]

and we prove (27). \(\Box\)

Define \(c_{1-\alpha}\) as the \(1 - \alpha\) quantile of \(H\). The density of a multivariate normal random variable with a full rank covariance matrix is positive for \(\forall x \in \mathbb{R}^{\mathcal{M}}\). And \(\forall x \geq 0, \delta > 0\), the set \(\begin{Bmatrix} t = (t_{il}, i \in \mathcal{M}) | x < \max_{i \in \mathcal{M}} |t_i| \leq x + \delta \end{Bmatrix}\) has positive Lebesgue measure. Therefore, \(H(x)\) is strictly increasing, and for any \(0 < \alpha < 1\), \(H(c_{1-\alpha}) = 1 - \alpha\). From theorem 2, for any given \(0 < \alpha_0 < \alpha_1 < 1\),

\[
\sup_{\alpha_0 \leq \alpha \leq \alpha_1} |\text{Prob} \left( \max_{i=1,2,\ldots,p} \frac{\hat{G}_i - \gamma_i}{\tau_i} \leq c_{1-\alpha} \right) - (1 - \alpha) | \\
\leq \sup_{x \geq 0} |\text{Prob} \left( \max_{i=1,2,\ldots,p} \frac{\hat{G}_i - \gamma_i}{\tau_i} \leq x \right) - H(x) | \rightarrow 0
\]
as \(n \rightarrow \infty\).
APPENDIX C: PROOFS OF THEOREMS IN SECTION 4

PROOF OF THEOREM 3. According to theorem 1, \( P_{\text{prob}} \left( \hat{\mathcal{N}}_{b} \neq \mathcal{N}_{b} \right) = O(n^{\alpha_{p}+m_{n}-m_{n}}) \).

If \( \hat{\mathcal{N}}_{b_n} = \mathcal{N}_{b_n} \), from (16)

\[
\| \hat{\theta} \|_{2}^{2} = \sum_{i \in \mathcal{N}_{b_n}} \theta_{i}^{2} \leq 3 \sum_{i \in \mathcal{N}_{b_n}} | \theta_{i} |^{2} + 3 \rho_{n}^{4} \sum_{i \in \mathcal{N}_{b_n}} \left( \sum_{j=1}^{r} \frac{q_{ij}\zeta_{j}}{\lambda_{j}^{2} + \rho_{n}} \right)^{2}
\]

(97)

\[
+ 3 \sum_{i \in \mathcal{N}_{b_n}} \left( \sum_{j=1}^{r} \sum_{l=1}^{n} q_{ij} \left( \frac{\lambda_{j}}{\lambda_{j}^{2} + \rho_{n}} + \frac{\rho_{n}\lambda_{j}}{(\lambda_{j}^{2} + \rho_{n})^{2}} \right) \right) p_{ij} \epsilon_{l}
\]

From assumption 2, \( \sum_{i \in \mathcal{N}_{b_n}} | \theta_{i} |^{2} \leq \| \theta \|_{2}^{2} = O(n^{2\alpha_{\sigma}}) \). Similarly

\[
p_{n}^{4} \sum_{i \in \mathcal{N}_{b_n}} \left( \sum_{j=1}^{r} \frac{q_{ij}\zeta_{j}}{\lambda_{j}^{2} + \rho_{n}} \right)^{2} \leq \frac{\rho_{n}^{2}}{\lambda_{p}^{2}} \sum_{i \in \mathcal{N}_{b_n}} \sum_{j=1}^{r} q_{ij}^{2} \sum_{j=1}^{r} \zeta_{j}^{2} = \frac{\rho_{n}^{4} \times | \mathcal{N}_{b_n} | \times \| \theta \|_{2}^{2}}{\lambda_{p}^{8}} = o(n^{-2\alpha_{\sigma}})
\]

From assumption 6,

\[
\mathbb{E} \sum_{i \in \mathcal{N}_{b_n}} \left( \sum_{j=1}^{r} \sum_{l=1}^{n} q_{ij} \left( \frac{\lambda_{j}}{\lambda_{j}^{2} + \rho_{n}} + \frac{\rho_{n}\lambda_{j}}{(\lambda_{j}^{2} + \rho_{n})^{2}} \right) \right) p_{ij} \epsilon_{l}
\]

(98)

\[
= \sigma_{p}^{2} \sum_{i \in \mathcal{N}_{b_n}} \sum_{i=1}^{n} \left( \sum_{j=1}^{r} q_{ij} \left( \frac{\lambda_{j}}{\lambda_{j}^{2} + \rho_{n}} + \frac{\rho_{n}\lambda_{j}}{(\lambda_{j}^{2} + \rho_{n})^{2}} \right) \right) p_{ij}
\]

\[
= \sigma_{p}^{2} \sum_{i \in \mathcal{N}_{b_n}} \sum_{j=1}^{r} q_{ij}^{2} \left( \frac{\lambda_{j}}{\lambda_{j}^{2} + \rho_{n}} + \frac{\rho_{n}\lambda_{j}}{(\lambda_{j}^{2} + \rho_{n})^{2}} \right) \leq 4 \sigma_{p}^{2} | \mathcal{N}_{b_n} | \frac{\lambda_{p}^{2}}{\lambda_{p}^{8}}
\]

\[
\Rightarrow \sum_{i \in \mathcal{N}_{b_n}} \left( \sum_{j=1}^{r} \sum_{l=1}^{n} q_{ij} \left( \frac{\lambda_{j}}{\lambda_{j}^{2} + \rho_{n}} + \frac{\rho_{n}\lambda_{j}}{(\lambda_{j}^{2} + \rho_{n})^{2}} \right) \right) p_{ij} \epsilon_{l}
\]

(99)

\[
\Rightarrow \cap_{i \in \mathcal{N}_{b_n}} \left( \sum_{j=1}^{r} \sum_{l=1}^{n} q_{ij} \left( \frac{\lambda_{j}}{\lambda_{j}^{2} + \rho_{n}} + \frac{\rho_{n}\lambda_{j}}{(\lambda_{j}^{2} + \rho_{n})^{2}} \right) \right) p_{ij} \epsilon_{l} = O_{p} \left( n^{-2\alpha_{\sigma}} \right)
\]

Since \( \alpha_{\theta}, \alpha_{\sigma} \geq 0 \), \( \| \hat{\theta} \|_{2} = O_{p}(n^{\alpha_{\sigma}}) \). According to (15) and (16), define \( \zeta = Q^{T} \tilde{\theta} \).

\[
\tilde{\theta} - \hat{\theta} = (I_{p} + \rho_{n}Q(\Lambda^{2} + \rho_{n}I_{r})^{-1}Q^{T})Q(\Lambda^{2} + \rho_{n}I_{r})^{-1} \left( \Lambda^{2}Q^{T} \hat{\theta} + \Lambda P^{T} \epsilon^{*} \right) + \hat{\theta} - QQ^{T} \tilde{\theta} - Q_{\perp}Q^{T} \hat{\theta}
\]

(100)

\[
\Rightarrow \tilde{\theta}^{*} - \hat{\theta} = -\rho_{n}^{2} \sum_{j=1}^{r} q_{ij} \zeta_{j} \sum_{j=1}^{r} \sum_{l=1}^{n} q_{ij} \left( \frac{\lambda_{j}}{\lambda_{j}^{2} + \rho_{n}} + \frac{\rho_{n}\lambda_{j}}{(\lambda_{j}^{2} + \rho_{n})^{2}} \right) p_{ij} \epsilon_{l}^{*}
\]

Similar to (79), rewrite \( \delta \) in assumption 2 as \( \delta = \frac{\eta + \alpha_{\sigma} + \delta_{1}}{2} \) with \( \delta_{1} > 0 \), we have

\[
\max_{i=1,2,\ldots,p} \rho_{n}^{2} \sum_{j=1}^{r} q_{ij} \zeta_{j} \leq \max_{i=1,2,\ldots,p} \frac{\rho_{n}^{2}}{\lambda_{p}^{4}} \sum_{j=1}^{r} q_{ij}^{2} \times \sum_{j=1}^{r} \zeta_{j}^{2} = \frac{\rho_{n}^{2} \| \hat{\theta} \|_{2}}{\epsilon_{n}^{m} n^{m}} = O_{p} \left( n^{-\eta - \delta} \right)
\]

(101)

\[
\epsilon_{1}^{*} \epsilon_{i}, i = 1, 2, \ldots, n \text{ are normal random variables with mean } 0 \text{ and variance } \hat{\sigma}^{2}. \text{ Therefore } \mathbb{E}^{*} \epsilon_{1}^{*} = \sigma m D, D = \mathbb{E}[Y]^{m}, Y \text{ is a normal random variable with mean } 0 \text{ and variance } 1.
\]
If \( \hat{\sigma} > 0 \), from (76) and lemma 2, \( \exists \) a constant \( E \) which depends on \( m \) and \( D \) such that for any \( a > 0 \),

\[
Prob^* \left( \max_{i=1,2,\ldots,p} \left| \sum_{j=1}^{r} \sum_{l=1}^{n} q_{ij} \left( \frac{\lambda_j}{\lambda_j^2 + \rho_n} + \frac{\rho_n \lambda_j}{(\lambda_j^2 + \rho_n)^2} \right) p_{ij} \epsilon^*_i \right| > \frac{a}{\sigma} \right) \leq \frac{pE\hat{\sigma}^m}{\lambda_\rho \gamma_m a^m}
\]

Suppose \( \tilde{N}_{b_n} = N_{b_n}, \frac{a}{2} < \sigma < \frac{3a}{2}, \) and \( \max_{i=1,2,\ldots,p} \rho_n^2 \sum_{j=1}^{r} q_{ij} \left( \frac{\lambda_j}{\lambda_j^2 + \rho_n} + \frac{\rho_n \lambda_j}{(\lambda_j^2 + \rho_n)^2} \right) \leq C \) \( n^{-\eta-\delta_1} \) for a constant \( C \). Since \( \hat{\theta}_i = 0 \) if \( i \notin \tilde{N}_{b_n} \),

\[
Prob^* \left( \tilde{N}_{b_n}^* \neq N_{b_n} \right) \leq Prob^* \left( \min_{i \in \tilde{N}_{b_n}} |\hat{\theta}_i| \leq b_n \right) + Prob^* \left( \max_{i \notin \tilde{N}_{b_n}} |\hat{\theta}_i| > b_n \right)
\]

\[
\leq Prob^* \left( \min_{i \in \tilde{N}_{b_n}} |\hat{\theta}_i| - \max_{i \notin \tilde{N}_{b_n}} \frac{\rho_n^2}{(\lambda_j^2 + \rho_n)^2} \sum_{j=1}^{r} q_{ij} \left( \frac{\lambda_j}{\lambda_j^2 + \rho_n} + \frac{\rho_n \lambda_j}{(\lambda_j^2 + \rho_n)^2} \right) p_{ij} \epsilon^*_i | \leq b_n \right) + Prob^* \left( \max_{i \notin \tilde{N}_{b_n}} \sum_{j=1}^{r} \sum_{l=1}^{n} q_{ij} \left( \frac{\lambda_j}{\lambda_j^2 + \rho_n} + \frac{\rho_n \lambda_j}{(\lambda_j^2 + \rho_n)^2} \right) p_{ij} \epsilon^*_i | > b_n \right)
\]

From assumption 4, for sufficiently large \( n \),

\[
b_n - \rho_n^2 \max_{i \notin \tilde{N}_{b_n}} \sum_{j=1}^{r} q_{ij} \left( \frac{\lambda_j}{\lambda_j^2 + \rho_n} + \frac{\rho_n \lambda_j}{(\lambda_j^2 + \rho_n)^2} \right) \geq C b_n \eta_b - C n^{-\eta-\delta_1} \geq \frac{b_n}{2}
\]

From (76), lemma 2, assumption 1 and 4, we have

\[
\max_{i=1,2,\ldots,p} \left| \sum_{j=1}^{r} q_{ij} \left( \frac{\lambda_j}{\lambda_j^2 + \rho_n} + \frac{\rho_n \lambda_j}{(\lambda_j^2 + \rho_n)^2} \right) \right| \leq O_p \left( n^{\alpha_p/m-\eta} \right)
\]

Suppose a constant \( C \) such that \( \max_{i=1,2,\ldots,p} \sum_{j=1}^{r} q_{ij} \left( \frac{\lambda_j}{\lambda_j^2 + \rho_n} + \frac{\rho_n \lambda_j}{(\lambda_j^2 + \rho_n)^2} \right) \mid_{\hat{\theta}_j} \leq C n^{-\eta-\delta_1} \). From assumption 4, for sufficiently large \( n \),

\[
\min_{i \in \tilde{N}_{b_n}} |\hat{\theta}_i| \geq \min_{i \notin \tilde{N}_{b_n}} |\theta_i| - \max_{i \notin \tilde{N}_{b_n}} \rho_n^2 \sum_{j=1}^{r} q_{ij} \left( \frac{\lambda_j}{\lambda_j^2 + \rho_n} + \frac{\rho_n \lambda_j}{(\lambda_j^2 + \rho_n)^2} \right) \sum_{l=1}^{n} p_{ij} \epsilon^*_i \]

\[
\geq b_n - \rho_n^2 \sum_{j=1}^{r} \sum_{l=1}^{n} \frac{\lambda_j}{\lambda_j^2 + \rho_n} - C n^{\alpha_p/m-\eta} \Rightarrow \min_{i \in \tilde{N}_{b_n}} |\hat{\theta}_i| - \max_{i \notin \tilde{N}_{b_n}} \rho_n^2 \sum_{j=1}^{r} q_{ij} \left( \frac{\lambda_j}{\lambda_j^2 + \rho_n} + \frac{\rho_n \lambda_j}{(\lambda_j^2 + \rho_n)^2} \right) \sum_{l=1}^{n} p_{ij} \epsilon^*_i \]

\[
\geq \left( \frac{1}{c_b} - 1 \right) b_n - C n^{-\eta-\delta_1} - C n^{\alpha_p/m-\eta} - b_n > \frac{b_n}{2} \left( \frac{1}{c_b} - 1 \right)
\]

Correspondingly

\[
Prob^* \left( \tilde{N}_{b_n}^* \neq N_{b_n} \right) \leq Prob^* \left( \max_{i \in \tilde{N}_{b_n}} \left| \sum_{j=1}^{r} \sum_{l=1}^{n} q_{ij} \left( \frac{\lambda_j}{\lambda_j^2 + \rho_n} + \frac{\rho_n \lambda_j}{(\lambda_j^2 + \rho_n)^2} \right) p_{ij} \epsilon^*_i \right| > \frac{b_n}{2} \left( \frac{1}{c_b} - 1 \right) \right)
\]

\[
+ Prob^* \left( \max_{i \notin \tilde{N}_{b_n}} \sum_{j=1}^{r} \sum_{l=1}^{n} q_{ij} \left( \frac{\lambda_j}{\lambda_j^2 + \rho_n} + \frac{\rho_n \lambda_j}{(\lambda_j^2 + \rho_n)^2} \right) p_{ij} \epsilon^*_i \right) > \frac{b_n}{2} \left( \frac{1}{c_b} - 1 \right)
\]

\[
= \frac{pE\hat{\sigma}^m}{\lambda_\rho \gamma_m a^m} \times \left( \frac{2^m}{(1/c_b - 1)^m + 2^m} \right)
\]
which has order $O_p(n^{m_2m_0-m_0})$. If $\tilde{N}_{b_n}^* = N_{b_n}$, then $\tilde{\tau}_i = \tau_i$ for $i = 1, 2, \ldots, p_1$. Similar to (88),
\begin{equation}
\max_{i=1,2,\ldots,p_1} \left| \frac{\tilde{\gamma}_i^* - \gamma_i}{\tilde{\tau}_i^*} \right| = \max_{i=1,2,\ldots,p_1} \left| -\rho_n^2 \sum_{k=1}^{r} \frac{c_{ik} \tilde{\gamma}_i}{(\lambda_k^2 + \rho_n)^2} + \sum_{l=1}^{n} \sum_{k=1}^{r} c_{ik} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right) \rho_{lk} \epsilon_l^* \right| \frac{\tau_i}{\tilde{\tau}_i} \leq \max_{i \in M} \rho_n^2 \left| \sum_{k=1}^{r} c_{ik} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right) \rho_{lk} \epsilon_l^* \right| \frac{\tau_i}{\tilde{\tau}_i} \leq \rho_n^2 \frac{||\tilde{\theta}||_2}{\lambda^3 \rho_n^3} + \max_{i \in M} \left| \sum_{l=1}^{n} \sum_{k=1}^{r} c_{ik} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right) \rho_{lk} \epsilon_l^* \right| \frac{\tau_i}{\tilde{\tau}_i} \leq \rho_n^2 \frac{||\tilde{\theta}||_2}{\lambda^3 \rho_n^3}
\end{equation}

From theorem 1, for any $a > 0$ and sufficiently large $n$, $\exists$ constant $D_a$ such that $|\tilde{\sigma}^2 - \sigma^2| \leq D_a n^{-a}$ and $\frac{1}{2} \sigma < \tilde{\sigma} < \frac{3}{2} \sigma$ with probability $1 - a$.
\begin{equation}
|\sigma - \tilde{\sigma}| = \frac{|\sigma^2 - \tilde{\sigma}^2|}{\sigma + \tilde{\sigma}} \leq \frac{D_a n^{-a}}{\sigma}
\end{equation}

If $0 < x \leq n^{m_2/2}$, according to lemma 3, assumption 7 and (91), $\exists$ a constant $C'$ which only depends on $\sigma, c_M, C_\lambda$ such that
\begin{equation}
|\text{Prob} \left( \max_{i \in M} \left| \sum_{l=1}^{n} \sum_{k=1}^{r} c_{ik} \left( \frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right) \rho_{lk} \epsilon_l^* \right| \leq x \right) - H(x)| = |H \left( \frac{x \sigma}{\sigma} \right) - H(x)| \leq C' \left( 1 + \sqrt{\log(|M|)} + \sqrt{\log \left( \frac{x \sigma^2}{\sigma^2} \right)} \right) \frac{x |\sigma - \tilde{\sigma}|}{\sigma} \leq \frac{2D_a C'}{\sigma^2} \left( 1 + \sqrt{\log(n)} \right) n^{-a/2} + C' \sqrt{\frac{x |\sigma - \tilde{\sigma}|}{\sigma} \log \left( \frac{x |\sigma - \tilde{\sigma}|}{\sigma} \right)} \frac{x |\sigma - \tilde{\sigma}|}{\sigma} \times \sqrt{\frac{2D_a n^{-a/4}}{\sigma^2}}
\end{equation}

Function $x \log(x)$ is continuous when $x > 0$, $x \log(x) \to 0$ as $x \to 0$, and $\frac{x |\sigma - \tilde{\sigma}|}{\sigma} \leq \frac{2D_a n^{-a/2}}{\sigma^2} \to 0$ as $n \to \infty$. So $\sqrt{\frac{x |\sigma - \tilde{\sigma}|}{\sigma} \log \left( \frac{x |\sigma - \tilde{\sigma}|}{\sigma} \right)} \leq \sup_{x \in (0,1]} \sqrt{x \log(x)} < \infty$ for sufficiently large $n$.

On the other hand, if $x > n^{a/2}$, then $\frac{x \sigma}{\sigma} > \frac{2n^{a/2}}{3}$. From lemma 2, we may choose sufficiently large $m_1$ such that $m_1 \alpha \sigma / 2 > 2$, since $E|\xi_1|^{m_1} < \infty$ (Here $\xi_1$ is a normal random variable with mean 0 and variance $\sigma^2$) is a constant for given $m_1$ and
\[
\max_{i \in M} \sum_{k=1}^{r} \frac{1}{\tau_i} c_{ik} \left( \frac{\lambda_k}{\lambda_i^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_i^2 + \rho_n)^2} \right)^2 \leq 1, \text{ we have}
\]
\[
\Prob \left( \max_{i \in M} \left| \sum_{k=1}^{r} c_{ik} \left( \frac{\lambda_k}{\lambda_i^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_i^2 + \rho_n)^2} \right) \xi_k \right| > \frac{2n^{\alpha_\sigma}/2}{3} \right) \leq \frac{3m_1 |M| \times E}{2m_1 n^{m_1 \alpha_\sigma}/2}
\]
\[
\Rightarrow \left| \Prob^* \left( \max_{i \in M} \left| \sum_{l=1}^{n} \sum_{k=1}^{r} c_{ik} \left( \frac{\lambda_k}{\lambda_i^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_i^2 + \rho_n)^2} \right) p_l k \xi_{i_l} \right| > x \right) - H(x) \right| \leq 2 \times \frac{3m_1 |M| \times E}{2m_1 n^{m_1 \alpha_\sigma}/2}
\]

Since \( H(0) = 0 \), from (110) and (111), for any given \( a > 0 \) and sufficiently large \( n \),
\[
\sup_{x \geq 0} |\Prob^* \left( \max_{i \in M} \left| \sum_{l=1}^{n} \sum_{k=1}^{r} c_{ik} \left( \frac{\lambda_k}{\lambda_i^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_i^2 + \rho_n)^2} \right) p_l k \xi_{i_l} \right| \leq x \right) - H(x) | < a
\]

As a summary, for any given \( a > 0 \), \( \exists \) a constant \( D_a \) such that for sufficiently large \( n \),
the event \( |\hat{\sigma}^2 - \sigma^2| \leq D_a n^{-\alpha_\sigma} \), \( \frac{1}{2} \sigma < \hat{\sigma} < \frac{3}{2} \sigma \), \( \hat{N}_{b_n} = N_{b_n} \), \( \|\hat{\theta}\|_2 \leq D_a \times n^{\alpha_\sigma} \Rightarrow \frac{\sigma^2}{\lambda_i^2} \leq D_a' n^{-\delta_1} \) for constant \( D_a' \) and \( \max_{i=1,2,\ldots,p} |p_n^2 \sum_{j=1}^{r} \frac{\hat{q}_j \xi_j}{(\lambda_i^2 + \rho_n)^2}| \leq D_a \times n^{-\eta-\delta_1} \) happen with probability \( 1 - a \). From (108), assumption 5 and lemma 3, we have for any \( x \geq 0 \), there exists
a constant $C'$ such that

\[(113) \quad \text{Prob}^*\left(\max_{i=1,2,\ldots,p_1} \left| \frac{\hat{\gamma}_i^* - \gamma_i}{\tau_i^*} \right| \leq x \right) - H(x) \leq \text{Prob}^*\left(\hat{\mathcal{N}}_{b_n}^* \neq \mathcal{N}_{b_n}\right)
\]

\[+ C' \frac{\rho_n^2 \|\hat{\theta}\|_2}{\lambda^2} \times \left(1 + \sqrt{\log(|\mathcal{M}|)} + \sqrt{\log\left(\frac{\rho_n^2 \|\hat{\theta}\|_2}{\lambda^2}\right)}\right)\]

\[\leq a + C' D'_a(1 + \sqrt{\log(n)})n^{-\delta_1} + C' \sqrt{D'_a n^{-\delta_1/2}} \sqrt{\log\left(\frac{\rho_n^2 \|\hat{\theta}\|_2}{\lambda^2}\right)} + \text{Prob}^*\left(\hat{\mathcal{N}}_{b_n}^* \neq \mathcal{N}_{b_n}\right)\]

\[\geq \text{Prob}^*\left(\max_{i=1,2,\ldots,p_1} \left| \frac{\hat{\gamma}_i^* - \gamma_i}{\tau_i^*} \right| \leq x \right) - H(x) \geq \text{Prob}^*\left(\max_{i=1,2,\ldots,p_1} \left| \frac{\hat{\gamma}_i^* - \gamma_i}{\tau_i^*} \right| \leq x \cap \hat{\mathcal{N}}_{b_n}^* = \mathcal{N}_{b_n}\right) - H(x)\]

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

If $0 \leq x \leq \frac{\rho_n^2 \|\hat{\theta}\|_2}{\lambda^2}$, then $\text{Prob}^*\left(\max_{i=1,2,\ldots,p_1} \left| \frac{\hat{\gamma}_i^* - \gamma_i}{\tau_i^*} \right| \leq x \right) - H(x) = \text{Prob}^*\left(\max_{i=1,2,\ldots,p_1} \left| \frac{\hat{\gamma}_i^* - \gamma_i}{\tau_i^*} \right| \leq x \right) - H(x) = 0$. Therefore, for sufficiently large $n$, from (112) and (107), $\exists$ a constant $C$ such that

\[(114) \quad \sup_{x \geq 0} |\text{Prob}^*\left(\max_{i=1,2,\ldots,p_1} \left| \frac{\hat{\gamma}_i^* - \gamma_i}{\tau_i^*} \right| \leq x \right) - H(x)| \leq \frac{p E \sigma^m}{c_\alpha n^{m \eta / m - \eta}} \times \left(2^m + \frac{2^m}{(1/c_\beta - 1)^m}\right) + a + C' D'_a(1 + \sqrt{\log(n)})n^{-\delta_1} + C' \sqrt{D'_a n^{-\delta_1/2}} \sqrt{\sup_{x \in (0,1]} |x \log(x)|} \leq C n^{m(\nu_0 + \alpha \eta / m - \eta)} + 2a\]

and we prove (32).

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

\[(115) \quad \text{Prob}\left(\sup_{x \geq 0} |\text{Prob}^*\left(\max_{i=1,2,\ldots,p_1} \left| \frac{\hat{\gamma}_i^* - \gamma_i}{\tau_i^*} \right| \leq x \right) - H(x)| \leq a\right) > 1 - a\]

Choose sufficiently small $a$ such that $0 < 1 - \alpha - 2a < 1 - \alpha + 2a < 1$. If (115) happens, for any $1 > \alpha > 0$, define $c_{1-\alpha}$ as the $1 - \alpha$ quantile of $H(x)$,

\[\text{Prob}^*\left(\max_{i=1,2,\ldots,p_1} \left| \frac{\hat{\gamma}_i^* - \gamma_i}{\tau_i^*} \right| \leq c_{1-\alpha} + 2a\right) - (1 - \alpha + 2a) \geq -a \Rightarrow c_{1-\alpha} \leq c_{1-\alpha} + 2a\]

\[(116) \quad \text{Prob}^*\left(\max_{i=1,2,\ldots,p_1} \left| \frac{\hat{\gamma}_i^* - \gamma_i}{\tau_i^*} \right| \leq c_{1-\alpha} - 2a\right) - (1 - \alpha - 2a) \leq a \Rightarrow c_{1-\alpha} > c_{1-\alpha} - 2a\]
From theorem 2, we have for sufficiently large $n$, (117)

$$\text{Prob}\left(\max_{i=1,2,...,p_1} \frac{|\hat{\gamma}_i - \gamma_i|}{\hat{\tau}_i} \leq c_{1-\alpha}\right)$$

$$\leq \text{Prob}\left(\sup_{x \geq 0} \text{Prob}^*\left(\max_{i=1,2,...,p_1} \frac{|\hat{\gamma}_i - \gamma_i|}{\hat{\tau}_i} \leq x\right) - H(x) > a\right) + \text{Prob}\left(\max_{i=1,2,...,p_1} \frac{|\hat{\gamma}_i - \gamma_i|}{\hat{\tau}_i} \leq c_{1-\alpha+2a}\right)$$

$$\leq a + (H(c_{1-\alpha+2a}) + a) = 1 - \alpha + 4a$$

$$\text{Prob}\left(\max_{i=1,2,...,p_1} \frac{|\hat{\gamma}_i - \gamma_i|}{\hat{\tau}_i} \leq c_{1-\alpha}\right) \geq \text{Prob}\left(\sup_{x \geq 0} \text{Prob}^*\left(\max_{i=1,2,...,p_1} \frac{|\hat{\gamma}_i - \gamma_i|}{\hat{\tau}_i} \leq x\right) - H(x) \leq a\right)$$

$$\geq \text{Prob}\left(\max_{i=1,2,...,p_1} \frac{|\hat{\gamma}_i - \gamma_i|}{\hat{\tau}_i} \leq c_{1-\alpha-2a}\right) - \text{Prob}\left(\sup_{x \geq 0} \text{Prob}^*\left(\max_{i=1,2,...,p_1} \frac{|\hat{\gamma}_i^* - \hat{\gamma}_i|}{\hat{\tau}_i^*} \leq x\right) - H(x) > a\right)$$

$$\geq (H(c_{1-\alpha-2a}) - a) - a = 1 - \alpha - 4a \Rightarrow \text{Prob}\left(\max_{i=1,2,...,p_1} \frac{|\hat{\gamma}_i - \gamma_i|}{\hat{\tau}_i} \leq c_{1-\alpha}\right) - (1 - \alpha) \leq 4a$$

For $a > 0$ can be arbitrarily small, we prove (31).

**APPENDIX D: PROOFS OF THEOREMS IN SECTION 5**

**PROOF OF LEMMA 1.** Define the design matrix $X = (x_{ij})_{i=1,...,n,j=1,...,p}$, $\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$, and $x'_{ij} = x_{ij} - \bar{x}_j$. If $\hat{\mathcal{N}}_{b_n} = \mathcal{N}_b$, for $i = 1, 2, ..., n$, (118)

$$\hat{\epsilon}'_i = \epsilon_i + \sum_{j \notin \mathcal{N}_b} x_{ij} \theta_j - \sum_{j \in \mathcal{N}_b} x_{ij} (\tilde{\theta}_j - \theta_j) \Rightarrow \hat{\epsilon}_i = \epsilon_i - \frac{1}{n} \sum_{i=1}^n \epsilon_i + \sum_{j \notin \mathcal{N}_b} x'_{ij} \theta_j - \sum_{j \in \mathcal{N}_b} x'_{ij} (\tilde{\theta}_j - \theta_j)$$
Define $\tilde{F}(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{\epsilon_i \leq x}, \ x \in \mathbb{R}$. From (57), for any given $\psi > 0$,
\begin{equation}
\hat{F}(x) - F(x) = \left( \hat{F}(x) - \tilde{F}(x + 1/\psi) \right) + \left( \tilde{F}(x + 1/\psi) - F(x + 1/\psi) \right) + (F(x + 1/\psi) - F(x))
\end{equation}
\begin{equation}
\leq \frac{1}{n} \sum_{i=1}^{n} (g_{\psi,x}(\epsilon_i) - g_{\psi,x}(\epsilon_i)) + \sup_{x \in \mathbb{R}} |\tilde{F}(x) - F(x)| + (F(x + 1/\psi) - F(x))
\end{equation}
\begin{equation}
\leq g_* \psi \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\tilde{\epsilon}_i - \epsilon_i)^2 + \sup_{x \in \mathbb{R}} |\tilde{F}(x) - F(x)| + (F(x + 1/\psi) - F(x))}
\end{equation}
\begin{equation}
\hat{F}(x) - F(x) = \left( \hat{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))
\end{equation}
\begin{equation}
\geq \frac{1}{n} \sum_{i=1}^{n} (g_{\psi,x-1/\psi}(\epsilon_i) - g_{\psi,x-1/\psi}(\epsilon_i)) - \sup_{x \in \mathbb{R}} |\tilde{F}(x) - F(x)| - (F(x) - F(x - 1/\psi))
\end{equation}
\begin{equation}
\geq -g_* \psi \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\tilde{\epsilon}_i - \epsilon_i)^2 - \sup_{x \in \mathbb{R}} |\tilde{F}(x) - F(x)| - (F(x) - F(x - 1/\psi))}
\end{equation}
\begin{equation}
\Rightarrow \sup_{x \in \mathbb{R}} |\hat{F}(x) - F(x)| \leq g_* \psi \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\tilde{\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{equation}
Suppose assumption 1 to 6. 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 $1 - a$

\[(120)\]

\[
\frac{1}{n} \sum_{i=1}^{n} (\epsilon_i - \epsilon_i)^2 = \frac{1}{n} \sum_{i=1}^{n} \left( \sum_{j \not\in N_{n_i}} x_{ij} \bar{\theta}_j - \sum_{j \in N_{n_i}} x_{ij} (\bar{\theta}_j - \theta_j) - \frac{1}{n} \sum_{j=1}^{n} \epsilon_j \right)^2
\]

\[
\leq \frac{3}{n} \sum_{i=1}^{n} \left( \sum_{j \not\in N_{n_i}} x_{ij} \bar{\theta}_j \right)^2 + \frac{3}{n} \sum_{i=1}^{n} \left( \sum_{j \in N_{n_i}} x_{ij} (\bar{\theta}_j - \theta_j) \right)^2 + \frac{3}{n} \sum_{j=1}^{n} \epsilon_j^2
\]

\[
\leq \frac{6}{n} \sum_{i=1}^{n} \left( \sum_{j \not\in N_{n_i}} x_{ij} \bar{\theta}_j \right)^2 + 6 \left( \sum_{j \not\in N_{n_i}} \bar{\theta}_j \right)^2 + \frac{6}{n} \sum_{i=1}^{n} \left( \sum_{j \in N_{n_i}} x_{ij} (\bar{\theta}_j - \theta_j) \right)^2 + \frac{3}{n} \sum_{j=1}^{n} \epsilon_j^2
\]

\[
\leq C_a n^{-\alpha} + \frac{6}{n^2} \left( \sum_{i=1}^{n} \sum_{j \not\in N_{n_i}} x_{ij} \bar{\theta}_j \right)^2 + C_a |N_{n_i}| (n^{2\alpha_0 - 4\delta} + n^{-2\eta}) + \frac{6}{n^2} \left( \sum_{i=1}^{n} \sum_{j \in N_{n_i}} x_{ij} (\bar{\theta}_j - \theta_j) \right)^2 + \frac{C_a}{n}
\]

\[
\leq C_a n^{-\alpha} + \frac{6}{n} \sum_{i=1}^{n} \left( \sum_{j \not\in N_{n_i}} x_{ij} \bar{\theta}_j \right)^2 + C_a |N_{n_i}| (n^{2\alpha_0 - 4\delta} + n^{-2\eta}) + \frac{6}{n} \sum_{i=1}^{n} \left( \sum_{j \in N_{n_i}} x_{ij} (\bar{\theta}_j - \theta_j) \right)^2 + \frac{C_a}{n}
\]

\[
\Rightarrow \left[ \frac{1}{n} \sum_{i=1}^{n} (\epsilon_i - \epsilon_i)^2 = O_p(n^{-\alpha/2}) \right]
\]

According to Gilvenko-Cantelli lemma, $\sup_{x \in \mathbb{R}} |\bar{F}(x) - F(x)| \to 0$ almost surely. Therefore, for any $a > 0$ and sufficiently large $n$, $\text{Prob} \left( \sup_{x \in \mathbb{R}} |\bar{F}(x) - F(x)| \leq a \right) > 1 - a$. Choose sufficiently small $a$ and $\psi = 1/a$, from assumption 8 and (120), we prove (35). ☐

**Proof of Theorem 4.** Define $X_f = (x_{f,ij})_{i=1,\ldots,p_1; j=1,\ldots,p}$. From theorem 1, since $p_1 = O(1)$.

\[(121)\]

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

For any given $0 < a < 1$, choose a constant $C_a$ such that

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

for any $n = 1, 2, \ldots$.

Define $F^-(x) = \lim_{y \rightarrow x, y \to x} F(y)$ for any $x \in \mathbb{R}$, and $G(x) = \text{Prob} \left( \max_{i=1,2,\ldots,p_1} |\epsilon_i| \leq x \right) = (F(x) - F^-(x))^{p_1}$ for $x \geq 0$. $G$ is continuous according to assumption 8. With probability at least $1 - a$
From (122) and assumption 8

\[
\sup_{x \geq 0} |\text{Prob}^*\left(\max_{i=1,2,\ldots,p_1} |\epsilon_{f,i} - x + \sum_{j=1}^p x_{f,ij}(j - \hat{\theta}_j)| - G(x)\right) | \leq \sup_{x \geq 0} |\text{Prob}^*(\max_{i=1,2,\ldots,p_1} |\epsilon_{f,i} - x + \sum_{j=1}^p x_{f,ij}(j - \hat{\theta}_j)| - G(x)\right) |
\]

(122)

\[
= \sup_{x \geq 0} |\text{Prob}^*\left(\max_{i=1,2,\ldots,p_1} |\epsilon_{f,i} - x + \sum_{j=1}^p x_{f,ij}(j - \hat{\theta}_j)| - G(x)\right) | \leq \sup_{x \geq 0} |G(x + C_a n^{-\eta}) - G(x)| + \sup_{x \geq 0} |G(x) - G(\max(0, x - C_a n^{-\eta}))|
\]

For any \(\delta > 0\) and any \(x \geq 0\)

\[
G(x + \delta) - G(x) = \sum_{i=1}^{p_1} (F(x + \delta) - F(-x - \delta))^{i-1} \times (F(x) - F(-x))^{p_1-i} \times (F(x + \delta) - F(-x - \delta) - F(x) + F(-x)) \leq 2p_1 \times \sup_{x \in \mathbb{R}} (F(x + \delta) - F(x)) \Rightarrow \sup_{x \geq 0} G(x + \delta) - G(x) \leq 2p_1 \times \sup_{x \in \mathbb{R}} (F(x + \delta) - F(x))
\]

From (122) and assumption 8

(123)

If \(\tilde{\mathcal{N}}_{b_n} = \mathcal{N}_{b_n}, \frac{\sigma}{\sqrt{T}} < \hat{\sigma} < \frac{3\sigma}{\sqrt{T}}, \) and \(||\tilde{\theta}\|_2 \leq C \times n^{\alpha_\theta} \) (see (97) to (99)), then

(124)

\[
\sup_{x \geq 0} |\text{Prob}^*\left(\max_{i=1,2,\ldots,p_1} |\epsilon_{f,i} - x + \sum_{j=1}^p x_{f,ij}(j - \hat{\theta}_j)| - G(x)\right) | = o_p(1)
\]

\[
\max_{i=1,2,\ldots,p_1} \sum_{j=1}^r q_{ij} \left(\frac{\lambda_j}{\lambda_j^2 + \rho_n} + \frac{\rho_n \lambda_j}{(\lambda_j^2 + \rho_n)^2}\right) \sum_{l=1}^n p_{ij} \epsilon_l \leq C n^{\alpha_\epsilon/m - \eta}
\]

(125)

for some constant \(C\) with probability at least 1 \(- a\). Here \(2\delta = \eta + \alpha_\theta + \delta_1\). From (107), \(\exists\) a constant \(E\) such that

(126)

\[
\text{Prob}^*\left(\tilde{\mathcal{N}}_{b_n}^* \neq \mathcal{N}_{b_n}\right) \leq \frac{Ep}{n^{\alpha_\eta L_n}}
\]

If \(\tilde{\mathcal{N}}_{b_n} = \mathcal{N}_{b_n},\)

\[
\left|\sum_{j=1}^p x_{f,ij} \hat{\theta}_{ij}^* - \sum_{j=1}^p x_{f,ij} \hat{\theta}_j\right| = \left|\sum_{j \in \mathcal{N}_{b_n}} x_{f,ij}(\hat{\theta}_{ij}^* - \hat{\theta}_j)\right|
\]

(127)

\[
\leq \rho_n^2 \sum_{k=1}^r \left|\frac{c_{ik} \tilde{\lambda}_k}{\lambda_k^2 + \rho_n}\right|^2 + \left|\sum_{k=1}^r \sum_{l=1}^n c_{ik} \left(\frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2}\right) p_{ik} \epsilon_l^*\right|
\]

\[
\leq \rho_n^2 \sqrt{C_M} \|\tilde{\theta}\|_2 \lambda_k + \sum_{k=1}^r \sum_{l=1}^n c_{ik} \left(\frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2}\right) p_{ik} \epsilon_l^*\]
Form (79) and lemma 2, $\exists$ a constant $E$ which only depends on $m$, and for any $1 > a > 0$ with a sufficiently large $C_a > 0$

$$\begin{align*}
\text{(128)} & \quad Prob^\ast \left( \max_{i=1,2,...,p_1} \left| y_{f,i}^* - \hat{y}_{f,i}^* \right| \leq x \right) - G(x) \leq a + Prob^\ast \left( \max_{i=1,2,...,p_1} \left| \epsilon_{f,i}^* \right| \leq x + C_a n^{-\eta} \right) - G(x) \\
& \leq a + \sup_{x \geq 0} \left[ \left| Prob^\ast \left( \max_{i=1,2,...,p_1} \left| \epsilon_{f,i}^* \right| \leq x \right) - G(x) \right| + 2p_1 \sup_{x \in \mathbb{R}} (F(x + C_a n^{-\eta}) - F(x)) \right] \\
& \qquad \text{prob}\ast \left( \max_{i=1,2,...,p_1} \left| \epsilon_{f,i}^* \right| \leq x - C_a n^{-\eta} \right) - G(x) \\
& \geq -a + Prob^\ast \left( \max_{i=1,2,...,p_1} \left| \epsilon_{f,i}^* \right| \leq x - C_a n^{-\eta} \right) - G(x) - C_a n^{-\eta} - 2p_1 \sup_{x \in \mathbb{R}} (F(x + C_a n^{-\eta}) - F(x)) \\
& \quad \text{Since } G(x) = 0 \text{ and } Prob^\ast \left( \max_{i=1,2,...,p_1} \left| \epsilon_{f,i}^* \right| \leq x \right) = 0 \text{ if } x < 0, \text{ we have} \\
\text{(130)} & \quad \sup_{x \geq 0} \left| Prob^\ast \left( \max_{i=1,2,...,p_1} \left| y_{f,i}^* - \hat{y}_{f,i}^* \right| \leq x \right) - G(x) \right| \\
& \leq a + \sup_{x \geq 0} \left[ \left| Prob^\ast \left( \max_{i=1,2,...,p_1} \left| \epsilon_{f,i}^* \right| \leq x \right) - G(x) \right| + 2p_1 \sup_{x \in \mathbb{R}} (F(x + C_a n^{-\eta}) - F(x)) \right] \\
& \quad \text{From lemma 1, for any } x \geq 0, \\
\text{(131)} & \quad \left| Prob^\ast \left( \max_{i=1,2,...,p_1} \left| \epsilon_{f,i}^* \right| \leq x \right) - G(x) \right| = \left| \left( \hat{F}(x) - \hat{F}^\ast (-x) \right) \right|^{p_1} \left| (F(x) - F(-x))^{p_1} \right| \\
& \leq \sum_{i=1}^{p_1} \left| \hat{F}(x) - \hat{F}^\ast (-x) \right|^i \times \left| (F(x) - F(-x))^{p_1 - i} \times \left| \hat{F}(x) - F(x) \right| + \hat{F}^\ast (-x) - F(-x) \right| \\
& \leq 2p_1 \sup_{x \in \mathbb{R}} |\hat{F}(x) - F(x)| \to_0 0
\end{align*}$$

as $n \to \infty$. From theorem 1 and (97) to (99), for any $1 > a > 0$, with probability at least $1 - a$ $\exists$ a constant $C_a > 0$ such that for sufficiently large $n$, (125) happens with $C = C_a$ and $\sup_{x \geq 0} \left| Prob^\ast \left( \max_{i=1,2,...,p_1} \left| \epsilon_{f,i}^* \right| \leq x \right) - G(x) \right| < a$. Correspondingly for sufficiently
large $n$, with probability at least $1 - a$,

\begin{equation}
\sup_{x \geq 0} |\text{Prob}^* \left( \max_{i=1,2,\ldots,p_1} |y_{f,i} - \hat{y}_{f,i}| \leq x \right) - \text{Prob}^* \left( \max_{i=1,2,\ldots,p_1} |y_{f,i} - \hat{y}_{f,i}| \leq x \right) | \\
\leq \sup_{x \geq 0} |\text{Prob}^* \left( \max_{i=1,2,\ldots,p_1} |y_{f,i} - \hat{y}_{f,i}| \leq x \right) - G(x) | + \sup_{x \geq 0} |\text{Prob}^* \left( \max_{i=1,2,\ldots,p_1} |y_{f,i} - \hat{y}_{f,i}| \leq x \right) - \text{Prob}^* \left( \max_{i=1,2,\ldots,p_1} |y_{f,i} - \hat{y}_{f,i}| \leq x \right) - G(x) | \\
\leq a + \sup_{x \geq 0} |\text{Prob}^* \left( \max_{i=1,2,\ldots,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)) + a \leq 4a
\end{equation}

and we prove (37).

For given $0 < \alpha < 1$ and sufficiently small $a > 0$ such that $0 < 1 - \alpha - a < 1 - \alpha + a < 1$, define $c_{1-\alpha}$ as the $1 - \alpha$ quantile of $G(x)$. For $G(x)$ is continuous, $G(c_{1-\alpha}) = 1 - \alpha$. With probability at least $1 - a$, $\sup_{x \geq 0} |\text{Prob}^* \left( \max_{i=1,2,\ldots,p_1} |y_{f,i} - \hat{y}_{f,i}| \leq x \right) - G(x) | < a/2$.

Correspondingly with probability at least $1 - a$

\begin{equation}
\text{Prob}^* \left( \max_{i=1,2,\ldots,p_1} |y_{f,i}^* - \hat{y}_{f,i} | \leq c_{1-\alpha} + a \right) \geq 1 - a + a/2 \Rightarrow c_{1-\alpha} - a \leq c_{1-\alpha} + a
\end{equation}

\begin{equation}
\text{Prob}^* \left( \max_{i=1,2,\ldots,p_1} |y_{f,i}^* - \hat{y}_{f,i} | \leq c_{1-\alpha} - a \right) \leq 1 - a - a/2 \Rightarrow c_{1-\alpha} \geq c_{1-\alpha} - a
\end{equation}

From (124), for sufficiently large $n$, with probability at least $1 - a$

\begin{equation}
\text{Prob}^* \left( \max_{i=1,2,\ldots,p_1} |y_{f,i} - \hat{y}_{f,i}| \leq c_{1-\alpha} \right) \leq \text{Prob}^* \left( \max_{i=1,2,\ldots,p_1} |y_{f,i} - \hat{y}_{f,i}| \leq c_{1-\alpha} + a \right) \leq 1 - \alpha + 2a
\end{equation}

\begin{equation}
\text{Prob}^* \left( \max_{i=1,2,\ldots,p_1} |y_{f,i} - \hat{y}_{f,i}| \leq c_{1-\alpha} \right) \geq \text{Prob}^* \left( \max_{i=1,2,\ldots,p_1} |y_{f,i} - \hat{y}_{f,i}| \leq c_{1-\alpha} - a \right) \geq 1 - \alpha - 2a
\end{equation}

For $a > 0$ can be arbitrarily small, we prove (38). \qed