UNIVERSITY OF CALIFORNIA SAN DIEGO

Regression with complex data: regularization, prediction and bootstrap

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Mathematics with a Specialization in Statistics

by

Yunyi Zhang

Committee in charge:

Professor Dimitris N. Politis, Chair
Professor Ery Arias-Castro
Professor Jason Schweinsberg
Professor Yixiao Sun
Professor Danna Zhang

2022

The Dissertation of Yunyi Zhang is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

To

My advisor Dimitris N. Politis

For his invaluable guidance on being a good researcher

My parents Jianping Zhang & Yunmei Huang

For their unconditional love and support

My fiancée Tingting Wang and her mom Qinghan Liang

For their love, encouragement, and confidence in me

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First of all, I want to thank my advisor, Professor Dimitris N. Politis. I started my research career by attending his reading class in 2018. Since then, we have collaborated on several research projects with topics range from the classical linear model to high dimensional data and non-stationary time series. His wisdom and profound understanding of statistics have provided me with lots of inspiration. Besides, we have gone through many discussions and editing works. Those experiences taught me how to write a clear and informative paper. His enthusiasm for statistics research and consistent confidence in me strongly encouraged me to follow my dream and pursue an academic career.

Then I want to thank Prof. Ery Arias-Castro, Prof. Yixiao Sun, Prof. Jason Schweinsberg and Prof. Danna Zhang for serving on my doctoral committee. I am grateful to Prof. Ronghui Xu. She was my advisor when I first came to UCSD, and her advice helped me get used to the new environment quickly. I want to thank Yiren Wang and Xiaoou Pan for their useful suggestions on parts of this dissertation. I also want to express my thankfulness to the math department and UCSD. The five years I have spent in La Jolla will be an invaluable treasure of my lifetime.

Finally, this work is dedicated to my dad Jianping Zhang, my mom Yunmei Huang, my fiancèe Tingting Wang and her mom Qinghan Liang. There have been many hard times during my Ph.D. study, but their support and encouragement helped me overcome those hardships.

Parts of the thesis are based on papers I have co-authored with my advisor Prof. Dimitris N. Politis.

Chapter 2 is based on the paper "Ridge Regression Revisited: Debiasing, Thresholding and Bootstrap" by Y.Zhang and D.N. Politis and has been accepted for publication in *Annals of Statistics*. The dissertation author was the primary investigator and author of this paper.

Chapter 3 is based on the paper "Debiased and thresholded ridge regression for linear models with heteroskedastic and correlated errors" by Y.Zhang and D.N.Politis and has been submitted for publication. The dissertation author was the primary investigator and author of this paper.

Chapter 4 is based on the paper "Bootstrap prediction intervals with asymptotic conditional validity and unconditional guarantees" by Y.Zhang and D.N. Politis and has been accepted for publication in *Information and Inference: A journal of the IMA*. The dissertation author was the primary investigator and author of this paper.

The following grants/award partially supported this dissertation which I would like to acknowledge: National Science Foundation Grant DMS 19-14556.

VITA

2017	B. S. in Flight Vehicle Power Engineering, Beihang University
2017	B. S. in Applied Mathematics, Beihang University
2022	Ph. D. in Mathematics with a Specialization in Statistics, University of California San Diego

PUBLICATIONS

Dandan Zhang, Yunyi Zhang, Ke Dong and Shaoping Wang, "Control Allocation Strategy for Optimal Digital Microthruster Arrays in Orbit Control Applications", *Journal of Spacecraft and Rockets*, 56(3), pp. 836 - 843, 2019

Yunyi Zhang, Jiazheng Liu, Zexin Pan, Dimitris N. Politis, "Estimating transformation function", *Electron. J. Stat.*, 13(2), pp. 3095 - 3119, 2019

Yunyi Zhang, Dimitris N. Politis, "Ridge Regression Revisited: Debiasing, Thresholding and Bootstrap", accepted by *Annals of Statistics*, 2022

Yunyi Zhang, Dimitris N. Politis, "Bootstrap prediction intervals with asymptotic conditional validity and unconditional guarantees", accepted by *Information and Inference: A Journal of the IMA*, 2022

Yunyi Zhang, Dimitris N. Politis, " Debiased and thresholded ridge regression for linear models with heteroskedastic and correlated errors", submitted

ABSTRACT OF THE DISSERTATION

Regression with complex data: regularization, prediction and bootstrap

by

Yunyi Zhang

Doctor of Philosophy in Mathematics with a Specialization in Statistics

University of California San Diego, 2022

Professor Dimitris N. Politis, Chair

Analyzing a linear model is a fundamental topic in statistical inference and has been well-studied. However, the complex nature of modern data brings new challenges to statisticians, i.e., the existing theories and methods may fail to provide consistent results. Focusing on a high dimensional linear model with i.i.d. errors or heteroskedastic and dependent errors, this dissertation introduces a new ridge regression method called 'the debiased and thresholded ridge regression'; then adopts this method to fit the linear model. After that, it introduces new bootstrap algorithms and applies them to generate consistent simultaneous confidence intervals/performs hypothesis testing for linear combinations of parameters in the linear model. In addition, this paper applies bootstrap algorithm to construct the simultaneous prediction intervals for future

observations. Numerical algorithms show that the new ridge regression method has a good performance compared to other complex methods like Lasso or the threshold Lasso.

This thesis also studies the properties of a residual-based bootstrap prediction interval. It derives the asymptotic distribution of the difference between the conditional coverage probability of a nominal prediction interval and the conditional coverage probability of a prediction interval obtained via a residual-based bootstrap. This result shows that the residual-based bootstrap prediction interval has about 50% possibility of yielding conditional under-coverage. Moreover, it introduces a new bootstrap prediction interval that has the desired asymptotic conditional coverage probability and the possibility of conditional under-coverage.

Chapter 1

Linear regression with complex data

Analyzing a linear model

$$\begin{aligned} y &= X\beta + \varepsilon \\ X \text{ is an } n \times p &\text{ design matrix} \\ \text{and } \varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T &\text{ are the errors} \end{aligned} \tag{1.1}$$

is a fundamental topic in statistical inference, and there has been extensive research on this topic; see Wu [1986] for constructing confidence intervals, Stine [1985] for constructing prediction intervals and Seber and Lee [2003] for a complete introduction. However, in the modern era, data always exhibit complex nature, which brings new challenges to the existing methods and theories; see Tibshirani [2011] and the reference therein. Our research considers the following properties:

- **High-dimensionality** The dimension p can have a comparable size to or even larger than the number of observations n. Classical linear regression theories and methods rely on the assumption that p is significantly smaller than n. If this assumption is violated, then those methods may fail to provide a consistent result, see, e.g., Mammen [1993, 1996].

- **Heteroskedasticity and dependency** In the linear regression literature, this property means that the variance $\mathbf{E}\varepsilon_i^2 \neq \mathbf{E}\varepsilon_j^2$ and the covariance $\mathbf{E}\varepsilon_i\varepsilon_j \neq 0$ for $i \neq j$. In practice, the

dependent variables y (e.g., daily stock price of a company or PM 2.5 concentration in Los Angeles) may come from a stochastic process. If this happens, we cannot simply assume the errors ε are i.i.d. (i.e., independent and identically distributed).

Our research focuses on providing consistent confidence intervals/making hypothesis testing for a high dimensional linear model with i.i.d. or heteroskedastic and dependent errors. Specifically, chapter 2 adopts a high dimensional linear model with i.i.d. errors, introduces a new ridge regression method called 'the debiased and thresholded ridge regression', and provides a (bootstrapped) simultaneous consistent confidence interval for linear combinations of parameters β. In addition, it constructs the simultaneous prediction interval for future observations; see Politis [2015] for a detailed introduction to prediction intervals. Chapter 3 considers a high dimensional linear model with heteroskedastic and dependent(correlated) errors and applies the debiased and thresholded ridge regression method to fit the linear model. After that, it constructs a consistent simultaneous confidence interval/performs hypothesis testing for linear combinations of parameters β. This chapter also provides some theoretical results for a new class of heteroskedastic, dependent(non-stationary) random variables. These results should be useful not only in linear regression but also in other statistics aspects.

In chapter 4, we study the properties of a bootstrapped prediction interval. Suppose we have a linear model (1.1) and a new regressor x_f, then we may apply the residual-based bootstrap(e.g., Stine [1985]) and make a prediction interval for the new dependent variable y_f. However, the bootstrapped prediction interval always manifests under-coverage(i.e., the conditional coverage probability of the prediction interval is smaller than the nominal coverage probability) in practice, see Politis [2013]. Our work derives the asymptotic distribution of the difference between the conditional coverage probability of a nominal prediction interval and the conditional coverage probability of a prediction interval obtained via a residual-based bootstrap. A corollary of this result is that the residual-based bootstrapped prediction interval has 50% possibility of yielding under-coverage. We also develop a new bootstrap prediction interval

that has the desired asymptotic conditional coverage probability and the desired possibility of yielding under-coverage.

We postpone the detailed proofs of the theorems in chapter 2 to 4 to appendices A to C.

Chapter 2

Ridge Regression Revisited: Debiasing, Thresholding and Bootstrap

2.1 Abstract

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.

2.2 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 et al. [2017] for confidence interval construction and hypothesis testing; and Javanmard and Montanari, 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 et al. [2011] proposed to further *threshold* the estimated coefficients, leading to a sparser fitted model. Furthermore, Bühlmann and van de Geer [2011], and Dezeure et al. [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.

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 threshold ridge regression method and proved its consistency. Dai et al. [2018] introduced a broken adaptive ridge estimator to approximate L_0 penalized regression. Dobriban and Wager [2018] derived the limit of high dimensional ridge regression's expected predictive risk. Bühlmann [2013] used Lasso to correct the bias in a ridge regression estimator, while Lopes [2014] applied a residual-based bootstrap to construct confidence intervals.

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 $y = X\beta + \varepsilon$; here, the dimension $p < n$ (the sample size), X has rank p, and a is a known vector. The ridge estimator is $a^T\widetilde{\theta}^\star$ with $\widetilde{\theta}^\star = (X^TX + \rho_n I_p)^{-1}X^Ty$, 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 [2013]), and assuming the error vector ε consists of independent identically distributed (i.i.d.) components, the bias and the standard deviation can be calculated (and controlled) as follows:

$$\begin{aligned} \mathbf{E}a^T\widetilde{\theta}^\star - a^T\beta &= -\rho_n a^T Q(\Lambda^2 + \rho_n I_p)^{-1}Q^T\beta \\ \text{which implies } \ |\mathbf{E}a^T\widetilde{\theta}^\star - a^T\beta| &\le \frac{\rho_n\|a\|_2 \times \|\beta\|_2}{\lambda_p^2 + \rho_n} \\ \text{and } \ \sqrt{Var(a^T\widetilde{\theta}^\star)} = \sqrt{Var(\varepsilon_1) \times a^T Q(\Lambda^2 + \rho_n I_p)^{-2}\Lambda^2 Q^T a} & \\ \le \frac{\sqrt{Var(\varepsilon_1)} \times \|a\|_2}{\lambda_p}. & \end{aligned} \tag{2.1}$$

In the above, λ_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 2.7 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 et al. [2011], Dezeure et al. [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 2.7 indicate that improved ridge regression has favorable finite-sample 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 γ_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 et al. [2013]. Estimating and testing γ are important problems in econometrics, e.g., Dolado and Lütkepohl [1996], Sun [2011], Sun, and Gonçalves and Vogelsang [2011]. Besides, estimating γ 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 et al. and Romano et al. [2020]. In our point of view, however, the bootstrap is preferable as it captures the underlying variability of estimated quantities; Section 2.6 in what follows gives the details.

The remainder of this paper is organized as follows: Section 2.3 introduces frequently used notations and assumptions. Section 2.4 presents the consistency result and the Gaussian approximation theorem for the improved ridge regression estimator. Section 2.5 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 2.6 constructs bootstrap prediction intervals in our ridge regression setting using a novel, hybrid resampling procedure. Finally, Section 2.7 provides extensive simulations to illustrate the finite sample performance, while Section 2.8 gives some concluding remarks; technical proofs are deferred to chapter A.

2.3 Preliminaries

Our work focuses on the fixed design linear model

$$y = X\beta + \varepsilon \tag{2.2}$$

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

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

Perform a thin singular value decomposition $X = P\Lambda Q^T$ as in Theorem 7.3.2 in Horn and Johnson [2013]; here, P and Q respectively is $n \times r$ and $p \times r$ orthonormal matrices that satisfy $P^T P = Q^T Q = I_r$, where I_r denotes the $r \times r$ identity matrix. Furthermore, $\Lambda = diag(\lambda_1, ..., \lambda_r)$, and $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r > 0$ are positive singular values of X.

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

$$Q_\perp^T Q_\perp = I_{p-r},\ Q^T Q_\perp = 0,\ \text{and}\ QQ^T + Q_\perp Q_\perp^T = I_p; \tag{2.3}$$

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

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 $\mathcal{N}_{b_n} = \{i \,||\theta_i| > b_n\}$. After selecting a suitable b_n, define

$$c_{ik} = \sum_{j \in \mathcal{N}_{b_n}} m_{ij} q_{jk},\ \forall\ i = 1, 2, ..., p_1,\ k = 1, 2, ..., r,\ \text{and}\ \mathcal{M} = \{i \mid \sum_{k=1}^{r} c_{ik}^2 > 0\} \tag{2.4}$$

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

$$\tau_i = \sqrt{\sum_{k=1}^{r} c_{ik}^2 \left(\frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right)^2 + \frac{1}{n}} \tag{2.5}$$

In section 2.4 and (A.2.14) to (A.2.16), we show that the estimation error $\widehat{\gamma} - \gamma$ (see (2.17) for

the definition of $\widehat{\gamma}$) asymptotically can be approximated by the random vector

$$\begin{aligned}&(\sum_{l=1}^{n}\sum_{k=1}^{r} c_{1k}p_{lk}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)\varepsilon_l,\\ &...,\sum_{l=1}^{n}\sum_{k=1}^{r} c_{p_1k}p_{lk}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)\varepsilon_l)^T\end{aligned} \tag{2.6}$$

here $P=(p_{lk})_{l=1,...,n,k=1,...,r}$ and $\varepsilon=(\varepsilon_1,...,\varepsilon_n)^T$. Moreover, if we assume that $\varepsilon_i, i=1,...,n$ are i.i.d. with mean 0 and variance 1, then

$$\begin{aligned}&Var\left(\sum_{l=1}^{n}\sum_{k=1}^{r} c_{ik}p_{lk}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)\varepsilon_l\right)\\ =&\sum_{l=1}^{n}\left(\sum_{k=1}^{r} c_{ik}p_{lk}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)\right)^2=\sum_{k=1}^{r} c_{ik}^2\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)^2\end{aligned} \tag{2.7}$$

In section 2.4, we will estimate τ_i by $\widehat{\tau}_i$(defined in (2.25)) and use $\widehat{\tau}_i$ to normalize the estimation error. The extra $1/n$ in (2.5) 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 $a_n, b_n, n=1,2,...$, we say $a_n=O(b_n)$ if $\exists$ a constant $C>0$ such that $|a_n|\le C|b_n|$ for all n, and $a_n=o(b_n)$ if $\lim_{n\to\infty}\frac{a_n}{b_n}=0$. For two random variable sequences X_n, Y_n, we say $X_n=O_p(Y_n)$ if for any $0<\varepsilon<1$, $\exists$ a constant $C_\varepsilon>0$ such that $\sup_n Prob(|X_n|\ge C_\varepsilon|Y_n|)\le\varepsilon$; and $X_n=o_p(Y_n)$ if $\frac{X_n}{Y_n}\to_p 0$; 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 to hold as the sample size $n\to\infty$.* For a vector $a=(a_1,...,a_n)^T$ and a fixed number $q\ge 1$, 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 $\mathbf{E}^*\cdot$ respectively represent probability and expectation in the *"bootstrap world"*, i.e., they are the conditional probability $Prob(\cdot|y)$ and the conditional expectation $\mathbf{E}(\cdot|y)$.

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

of H is defined as

$$c_{1-\alpha} = \inf\{x \in \mathbf{R} | H(x) \geq 1-\alpha\}. \tag{2.8}$$

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

$$C_{1-\alpha} = X_{i_*} \text{ such that } i_* = \min\left\{ i \;\middle|\; \frac{1}{B}\sum_{j=1}^{B} \mathbf{1}_{X_j \leq X_i} \geq 1-\alpha \right\}. \tag{2.9}$$

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 $|\mathscr{M}|$ 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

$$C_\lambda n^{1/2} \geq \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r \geq c_\lambda n^{\eta}. \tag{2.10}$$

Furthermore, the Euclidean norm of θ 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 δ such that $\frac{\eta+\alpha_\theta}{2} < \delta < 2\eta$

3. The errors $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T$ driving regression (2.2) are assumed to be i.i.d., with $\mathbf{E}\varepsilon_1 = 0$, and $\mathbf{E}|\varepsilon_1|^m < \infty$ for some $m > 4$.

4. The dimension of the parameter vector β satisfies $p = O(n^{\alpha_p})$ for some constant $\alpha_p \in [0, m\eta)$ where m, η 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 $\nu_b + \frac{\alpha_p}{m} - \eta < 0$. We assume $\exists$ a constant $0 < c_b < 1$ such that $\max_{i \notin \mathscr{N}_{b_n}} |\theta_i| \leq c_b \times b_n$, and $\min_{i \in \mathscr{N}_{b_n}} |\theta_i| \geq \frac{b_n}{c_b}$.

The intuitive meaning of assumption 4 is that the θ_is that are not being truncated should be significantly larger than the θ_i being truncated.

5. $\mathscr{M}$ (defined in (2.4)) is not empty and $|\mathscr{M}| = O(n^{\alpha_{\mathscr{M}}})$ with $\alpha_{\mathscr{M}} < m\eta$ where m, η are as defined in Assumptions 1–3. Besides, assume $\exists$ constants $c_{\mathscr{M}}, C_{\mathscr{M}}$ such that $0 < c_{\mathscr{M}} < \sum_{k=1}^{r} c_{ik}^2 \leq C_{\mathscr{M}}$ for all $i \in \mathscr{M}$. Also assume

$$\begin{aligned} \max_{i=1,2,\ldots,p_1} |\sum_{j \notin \mathscr{N}_{b_n}} m_{ij}\theta_j| &= o\left(\frac{1}{\sqrt{n\log(n)}}\right) \\ \text{and } \max_{i=1,2,\ldots,p_1} |\sum_{j=1}^{p} m_{ij}\theta_{\perp,j}| &= o\left(\frac{1}{\sqrt{n\log(n)}}\right) \end{aligned} \tag{2.11}$$

We assume (2.11) to maintain the sparsity of θ and assure that the projection bias $\beta - \theta$ is negligible compared to the stochastic errors. It allows an inexact sparsity, i.e., some θ_i may not equal 0 even if $i \notin \mathscr{N}_{b_n}$. Theoretical results for other linear regression estimators (e.g., Lasso) need an exact sparse assumption ($\theta_i = 0$ for all $i \notin \mathscr{N}_{b_n}$), see Zhao and Yu [2006] and Basu and Michailidis [2015] for a further introduction.

6. $\exists$ a constant α_σ satisfying $\eta \geq \alpha_\sigma > 0$ such that

$$n^{-\nu_b} \sum_{j \notin \mathscr{N}_{b_n}} |\theta_j| = O(n^{-\alpha_\sigma}), \quad \frac{\sqrt{|\mathscr{N}_{b_n}|}}{n^\eta} = O(n^{-\alpha_\sigma}) \tag{2.12}$$

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

7.1.

$$\begin{aligned} &\max_{i \in \mathscr{M}, l=1,2,\ldots,n} |\frac{1}{\tau_i} \times \sum_{k=1}^{r} c_{ik} p_{lk} \left(\frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2}\right)| \\ &= o(\min(n^{(\alpha_\sigma - 1)/2} \times \log^{-3/2}(n),\ n^{-1/3} \times \log^{-3/2}(n))) \end{aligned} \tag{2.13}$$

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

$$\begin{aligned} |\mathscr{M}| &= o(n^{\alpha_\sigma} \times \log^{-3}(n)) \\ \max_{i\in\mathscr{M}, l=1,2,\ldots,n} |\frac{1}{\tau_i} \times \sum_{k=1}^{r} c_{ik} p_{lk} \left(\frac{\lambda_k}{\lambda_k^2+\rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2+\rho_n)^2} \right)| &= O(n^{-\alpha_\sigma} \times \log^{-3/2}(n)) \end{aligned} \tag{2.14}$$

According to (2.6), the normalized estimation error $\frac{\widehat{\gamma}_i - \gamma_i}{\widehat{\tau}_i}$ asymptotically will be approximated by $\sum_{l=1}^{n} \left(\frac{1}{\tau_i} \sum_{k=1}^{r} c_{ik} p_{lk} \left(\frac{\lambda_k}{\lambda_k^2+\rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2+\rho_n)^2} \right) \right) \varepsilon_l$. Therefore, the intuitive meaning of assumption 7 is that all terms $\frac{1}{\tau_i} \sum_{k=1}^{r} c_{ik} p_{lk} \left(\frac{\lambda_k}{\lambda_k^2+\rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2+\rho_n)^2} \right) \varepsilon_l$ in the summation are negligible and the number of simultaneous linear combinations $|\mathscr{M}|$ 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 ε. In this case, to intrepret the results we would need to replace $Prob(\cdot)$ by $Prob(\cdot|X)$, $\mathbf{E}\cdot$ by $\mathbf{E}\cdot|X$, $Prob^*(\cdot)$ by $Prob(\cdot|X,y)$ and $\mathbf{E}^*\cdot$ by $\mathbf{E}\cdot|X,y$.

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* θ*, i.e., assumption 5 and 6, can be violated. Section 2.7 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* $X = (x_{ij})_{i=1,\ldots,n, j=1,\ldots,p}$ *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* $\frac{1}{n}X^TX$ *would then converge to* $(1-\sqrt{c})^2$ *almost surely as* $n \to \infty$*. So the smallest singular value of X (which is the smallest eigenvalue of the square root of* X^TX*) is greater than* $\frac{1-\sqrt{c}}{2}\sqrt{n}$ *for sufficiently large n, almost surely. On the other hand, the largest eigenvalue of* $\frac{1}{n}X^TX$ *converges to* $(1+\sqrt{c})^2$ *as* $n \to \infty$*. Hence, the largest singular value of X also has order* $O(\sqrt{n})$ *almost surely.*

2.4 Consistency and the Gaussian approximation theorem

Throughout, we will use the notations developed in section 2.3. For a chosen ridge parameter $\rho_n > 0$, define the classical ridge regression estimator $\widetilde{\theta}^\star$ and the de-biased estimator $\widetilde{\theta}$ as

$$\begin{aligned} \widetilde{\theta}^\star &= (X^T X + \rho_n I_p)^{-1} X^T y \\ \widetilde{\theta} = (\widetilde{\theta}_1, ..., \widetilde{\theta}_p)^T &= \widetilde{\theta}^\star + \rho_n \times Q(\Lambda^2 + \rho_n I_r)^{-1} Q^T \widetilde{\theta}^\star \end{aligned} \tag{2.15}$$

Then we have

$$\widetilde{\theta} - \theta = -\rho_n^2 Q(\Lambda^2 + \rho_n I_r)^{-2} \zeta + Q\left((\Lambda^2 + \rho_n I_r)^{-1}\Lambda + \rho_n(\Lambda^2 + \rho_n I_r)^{-2}\Lambda\right) P^T \varepsilon \tag{2.16}$$

Similar to $\mathscr{N}_{b_n}$, define the set $\widehat{\mathscr{N}_{b_n}}$, the estimator $\widehat{\theta} = (\widehat{\theta}_1, ..., \widehat{\theta}_p)^T$ and $\widehat{\gamma}$ as

$$\widehat{\mathscr{N}_{b_n}} = \left\{ i \,\middle|\, |\widetilde{\theta}_i| > b_n \right\}, \;\; \widehat{\theta}_i = \widetilde{\theta}_i \times \mathbf{1}_{i \in \widehat{\mathscr{N}_{b_n}}}, \;\; \widehat{\gamma} = M\widehat{\theta} \tag{2.17}$$

Then, $\widehat{\theta}$ and $\widehat{\gamma}$ constitute the improved, i.e., debiased and thresholded, ridge regression estimator for the parameter vector θ and $\gamma = M\beta$ respectively. Apart from parameter estimation, we need to estimate the error variance $\sigma^2 = \mathbf{E}\varepsilon_1^2$. The estimator for σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\widehat{\theta}_j)^2 \tag{2.18}$$

Here $X = (x_{ij})_{i=1,...,n, j=1,...,p}$.

Remark 2. *According to* (2.16)*, the estimation error $\widetilde{\theta} - \theta$ is decomposed into a bias term $-\rho_n^2 Q(\Lambda^2 + \rho_n I_r)^{-2}\zeta$ and a variance term $Q\left((\Lambda^2 + \rho_n I_r)^{-1}\Lambda + \rho_n(\Lambda^2 + \rho_n I_r)^{-2}\Lambda\right) P^T\varepsilon$. For $\|\rho_n^2 Q(\Lambda^2 + \rho_n I_r)^{-2}\zeta\|_2 \leq \frac{\rho_n^2\|\beta\|_2}{(\lambda_r^2 + \rho_n)^2}$. In order to control the bias term, $\|\beta\|_2$ cannot be too large (which is achievable if β is sparse); in addition, ρ_n/λ_r^2 must be small.*

We can now explain why debiasing helps decrease the estimation error; we will use the

notation of section 2.3. According to (2.1)*, for a fixed vector* $a \in \mathbf{R}^p$*,*

$$\begin{aligned} a^T\widetilde{\theta}^\star - a^T\beta &= a^T\widetilde{\theta}^\star - a^T\theta - a^T Q_\perp Q_\perp^T\beta \\ &= a^T Q(\Lambda^2+\rho_n I_r)^{-1}\Lambda P^T\varepsilon - \rho_n a^T Q(\Lambda^2+\rho_n I_r)^{-1}\zeta - a^T Q_\perp Q_\perp^T\beta \end{aligned} \tag{2.19}$$

Assume $a^T Q_\perp Q_\perp^T\beta = 0$*. Then the bias term of* $a^T\widetilde{\theta}^\star$ *will be* $-\rho_n a^T Q(\Lambda^2+\rho_n I_r)^{-1}\zeta$*. We can estimate this by* $-\rho_n a^T Q(\Lambda^2+\rho_n I_r)^{-1}Q^T\widetilde{\theta}^\star$ *and subtract the estimated bias from* $a^T\widetilde{\theta}^\star$*, yielding the debiased estimator.*

Compared to (2.16)*, the debiased estimator* $\widetilde{\theta}$ *changes the bias term from* $-\rho_n a^T Q(\Lambda^2+\rho_n I_r)^{-1}\zeta$ *(having order* $O\left(\frac{\rho_n\|a\|_2\times\|\beta\|_2}{\lambda_r^2+\rho_n}\right)$*) to* $-\rho_n^2 Q(\Lambda^2+\rho_n I_r)^{-2}\zeta$ *(having order* $O\left(\frac{\rho_n^2\|a\|_2\times\|\beta\|_2}{(\lambda_r^2+\rho_n)^2}\right)$*). At the same time,* $\widetilde{\theta}$ *will enlarge the variance from* $Var(\varepsilon_1)\times a^T Q(\Lambda^2+\rho_n I_r)^{-2}\Lambda^2 Q^T a$ *to*

$$Var(\varepsilon_1)\times a^T Q\left((\Lambda^2+\rho_n I_r)^{-1}\Lambda+\rho_n(\Lambda^2+\rho_n I_r)^{-2}\Lambda\right)^2 Q^T a. \tag{2.20}$$

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

Even when $\rho_n > \lambda_r^2$*, numerical simulations in figure 2.1 show that debiasing still may help decrease the estimation error.*

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, Bühlmann [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* $Q_\perp Q_\perp^T\beta$*(i.e., the projection bias in Bühlmann [2013]). Therefore, if* $p < n$ *and X has rank p, then the bias-corrected ridge regression estimator equals the classical ridge regression estimator* $\widetilde{\theta}^\star$*. Our work does not focus on the projection bias but wants to diminish the estimation bias*

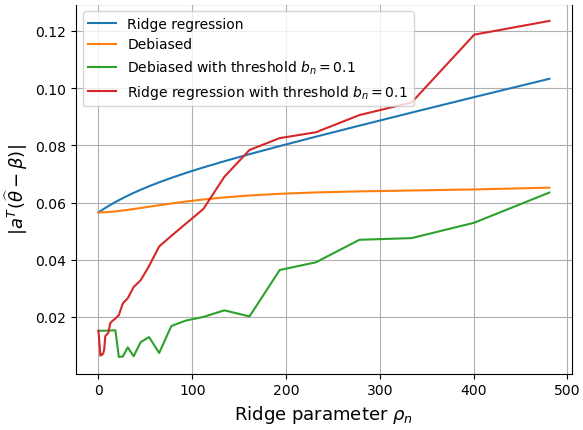


Figure 2.1. Estimation errors of the ridge regression estimator $a^T\widetilde{\theta}^\star$, the debiased estimator(Debiased) $a^T\widetilde{\theta}$, the debiased and threshold ridge regression estimator(Debiased with threshold $b_n = 0.1$) $a^T\widehat{\theta}$ and the threshold ridge regression estimator(Ridge regression with threshold $b_n = 0.1$ as in section 4 in Shao and Deng [2012]) with respect to different ρ_n. The threshold b_n is chosen to be 0.1, a is a fixed linear combination vector with $\|a\|_2 = 1$, and $\lambda_r = 12.684$.

$-\rho_n Q(\Lambda^2 + \rho_n I_r)^{-1}\zeta$. *Thus, even if* $p < n$ *and* X *has rank* p*, the debiased estimator* $\widetilde{\theta}$ *is still different from* $\widetilde{\theta}^\star$*(which is demonstrated in figure 2.1).*

Theorem 1. *(i). Suppose assumptions 1 to 5 hold true. Then*

$$Prob\left(\widehat{\mathscr{N}_{b_n}} \neq \mathscr{N}_{b_n}\right) = O(n^{\alpha_p + m\nu_b - m\eta}) \tag{2.21}$$

$\mathscr{N}_{b_n}$ *is defined in section 2.3. In other words, the variable selection consistency holds true asymptotically. Besides,*

$$\max_{i=1,2,\ldots,p_1} |\widehat{\gamma}_i - \gamma_i| = O_p(|\mathscr{M}|^{1/m} \times n^{-\eta}) \tag{2.22}$$

where $\gamma_i, i = 1, \ldots, p_1$ *are defined in section 2.3.*

(ii). Suppose assumptions 1 to 6 hold true. Then

$$|\widehat{\sigma}^2 - \sigma^2| = O_p(n^{-\alpha_\sigma}). \tag{2.23}$$

An advantage of using $\widehat{\theta}$ 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 $\widehat{\theta}$'s prediction loss $\frac{1}{n}\|X\widehat{\theta}-X\theta\|_2^2 = \frac{1}{n}\|X\widehat{\theta}-X\beta\|_2^2$. If $\widehat{\mathscr{N}_{b_n}} = \mathscr{N}_{b_n}$, then from (A.2.8) and (A.2.9) in section A.2,

$$\begin{aligned}\frac{1}{n}\|X\widehat{\theta}-X\theta\|_2^2 \leq \frac{2}{n}\sum_{i=1}^{n}\left(\sum_{j\in\mathscr{N}_{b_n}} x_{ij}(\widetilde{\theta}_j-\theta_j)\right)^2 + \frac{2}{n}\sum_{i=1}^{n}\left(\sum_{j\notin\mathscr{N}_{b_n}} x_{ij}\theta_j\right)^2 \\ \Rightarrow \frac{1}{n}\|X\widehat{\theta}-X\theta\|_2^2 = O_p\left(n^{-\alpha_\sigma}\right).\end{aligned} \tag{2.24}$$

On the other hand, the prediction loss of other estimators (e.g., Lasso) can be hard to derive. Dalalyan et al. [2017], Bickel et al. [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 2.7 show that $\widehat{\theta}$ 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 $\widehat{\tau}_i,\ i=1,2,...,p_1$ and $H(x), x\in\mathbf{R}$ as

$$\begin{aligned}\widehat{\tau}_i &= \sqrt{\sum_{k=1}^{r}\left(\sum_{j\in\widehat{\mathscr{N}_{b_n}}} m_{ij}q_{jk}\right)^2 \times \left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)^2 + \frac{1}{n}} \\ H(x) &= Prob\left(\max_{i\in\mathscr{M}}\frac{1}{\tau_i}\left|\sum_{k=1}^{r} c_{ik}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)\xi_k\right| \leq x\right)\end{aligned} \tag{2.25}$$

Here $\xi_k,\ k=1,2,...,r$ are independent normal random variables with mean 0 and variance $\sigma^2 = \mathbf{E}\varepsilon_1^2$. $|\mathscr{M}|$ (defined in (2.4)) and p_1 may grow as the sample size increases. In this case, the estimator $\max_{i=1,2,...,p_1}\frac{|\widehat{\gamma}_i-\gamma_i|}{\widehat{\tau}_i}$ does not have an asymptotic distribution. However, the cumulative distribution function of $\max_{i=1,2,...,p_1}\frac{|\widehat{\gamma}_i-\gamma_i|}{\widehat{\tau}_i}$ 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

$$\left\{\gamma=(\gamma_1,...,\gamma_{p_1})\Big|\max_{i=1,...,p_1}\frac{|\widehat{\gamma}_i-\gamma_i|}{\widehat{\tau}_i}\leq c_{1-\alpha}\right\} \tag{2.26}$$

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

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

$$\limsup_{n\to\infty}\sup_{x\geq 0}|Prob\left(\max_{i=1,2,...,p_1}\frac{|\widehat{\gamma}_i-\gamma_i|}{\widehat{\tau}_i}\leq x\right)-H(x)|=0 \tag{2.27}$$

where $\gamma_i, i=1,...,p_1$ are defined in section 2.3.

Gaussian approximation theorems like theorem 2 are useful tools not only in linear models but also in other high dimensional statistics; e.g., Chernozhukov et al. [2013] and Zhang and Wu [2017].

2.5 Bootstrap inference and hypothesis testing

An obstacle for constructing a practical confidence region or testing a hypothesis via theorem 2 are the unknown $\mathscr{M}$, $\mathscr{N}_{b_n}$, and σ. 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 et al. [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 γ_0) versus the alternative hypothesis $\gamma\neq\gamma_0$. Similar to Zhang and Cheng [2017], Chernozhukov et al. [2013], and Zhang and Wu [2017], we use the maximum statistic $\max_{i=1,2,...,p_1}\frac{|\widehat{\gamma}_i-\gamma_i|}{\widehat{\tau}_i}$ to construct a simultaneous confidence region.

Algorithm 1 (Wild bootstrap inference and hypothesis testing)**.** ***Input:*** *Design matrix X, dependent variables $y=X\beta+\varepsilon$, linear combination matrix M, ridge parameter ρ_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 $\widehat{\theta}$, $\widehat{\gamma} = (\widehat{\gamma}_1, ..., \widehat{\gamma}_{p_1})^T$ *defined in* (2.17), $\widehat{\tau}_i$, $i = 1, 2, ..., p_1$ *defined in* (2.25), *and* $\widehat{\sigma}$ *defined in* (2.18).

2. Generate i.i.d. errors $\varepsilon^* = (\varepsilon_1^*, ..., \varepsilon_n^*)^T$ *with* $\varepsilon_i^*, i = 1, ..., n$ *having normal distribution with mean* 0 *and variance* $\widehat{\sigma}^2$, *then calculate* $y^* = X\widehat{\theta} + \varepsilon^*$ *and* $\widehat{\theta}_\perp = Q_\perp Q_\perp^T \widehat{\theta}$ *(*$Q_\perp$ *is defined in section 2.3).*

3. Calculate $\widetilde{\theta}^{\star *} = (X^T X + \rho_n I_p)^{-1} X^T y^*$ *and*

$$\widetilde{\theta}^* = (\widetilde{\theta}_1^*, ..., \widetilde{\theta}_p^*)^T = \widetilde{\theta}^{\star *} + \rho_n \times Q(\Lambda^2 + \rho_n I_r)^{-1} Q^T \widetilde{\theta}^{\star *} + \widehat{\theta}_\perp$$

4. Calculate $\widehat{\mathcal{N}_{b_n}^*} = \left\{ i \middle| |\widetilde{\theta}_i^*| > b_n \right\}$ *and* $\widehat{\theta}^* = (\widehat{\theta}_1^*, ..., \widehat{\theta}_p^*)^T$ *with* $\widehat{\theta}_i^* = \widetilde{\theta}_i^* \times \mathbf{1}_{i \in \widehat{\mathcal{N}_{b_n}^*}}$ *for* $i = 1, 2, ..., p$.

5. Calculate $\widehat{\gamma}^* = M\widehat{\theta}^*$, $\widehat{\tau}_i^*, i = 1, 2, ..., p_1$, *and* E_b^* *such that*

$$\widehat{\tau}_i^* = \sqrt{\sum_{k=1}^{r} \left(\sum_{j \in \widehat{\mathcal{N}_{b_n}^*}} m_{ij} q_{jk} \right)^2 \times \left(\frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right)^2 + \frac{1}{n}}, \qquad E_b^* = \max_{i=1,2,...,p_1} \frac{|\widehat{\gamma}_i^* - \widehat{\gamma}_i|}{\widehat{\tau}_i^*} \tag{2.28}$$

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*

$$\left\{ \gamma = (\gamma_1, ..., \gamma_{p_1})^T \middle| \max_{i=1,2,...,p_1} \frac{|\widehat{\gamma}_i - \gamma_i|}{\widehat{\tau}_i} \leq C_{1-\alpha}^* \right\} \tag{2.29}$$

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*

$$\max_{i=1,2,\ldots,p_1} \frac{|\widehat{\gamma}_i - \gamma_{0,i}|}{\widehat{\tau}_i} > C^*_{1-\alpha}. \tag{2.30}$$

As in section 2.3, if X has rank $p \leq n$*, we define* $\widehat{\theta}_\perp = 0$*, the p dimensional vector with all elements* 0*.*

According to theorem 1.2.1 in Politis et al. [1999], the consistency of algorithm 1 –either for asymptotic validity of confidence regions or consistency of the hypothesis test– is ensured if

$$Prob\left(\max_{i=1,2,\ldots,p_1} \frac{|\widehat{\gamma}_i - \gamma_i|}{\widehat{\tau}_i} \leq c^*_{1-\alpha}\right) \to 1-\alpha \tag{2.31}$$

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

Theorem 3. *Suppose assumptions 1 to 7 hold true. Then*

$$\sup_{x\geq 0} |Prob^*\left(\max_{i=1,2,\ldots,p_1} \frac{|\widehat{\gamma}^*_i - \widehat{\gamma}_i|}{\widehat{\tau}^*_i} \leq x\right) - H(x)| = o_P(1). \tag{2.32}$$

In addition, for any given $0 < \alpha < 1$*, (2.31) 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 α, i.e., the test is consistent.

2.6 Bootstrap interval prediction

Given our data from the linear model $y = X\beta + \varepsilon$, 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 $y_f = X_f\beta + \varepsilon_f$, and (b) finding a $1-\alpha$ prediction region $A \subset R^{p_1}$ so that $Prob(y_f \in A) \to 1-\alpha$ as the (original) sample size $n \to \infty$. Here $\varepsilon_f = (\varepsilon_{f,1}, ..., \varepsilon_{f,p_1})^T$ are i.i.d. errors with the same marginal distribution as ε_1, and ε_f is independent with ε.

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 β were known, the predictor of y_f that is optimal with respect to total mean squared error is $X_f\beta$; since β is typically unknown, we can estimate it by $\widehat{\theta}$ as in (2.17), yielding the practical predictor $\widehat{y}_f = X_f\widehat{\theta}$. In what follows, we would like to derive a $1-\alpha$ prediction region for y_f based on the intuitive predictor $\widehat{y}_f$.

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 $y_f = X_f\beta + \varepsilon_f$ if*

$$Prob\left(y_f \in \Gamma | y\right) \to_p 1-\alpha \ \ as\ n \to \infty. \tag{2.33}$$

Note that the convergence in (2.33) is "in probability" since $Prob\left(y_f \in \Gamma | y\right)$ 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 et al., and Chernozhukov et al. [2021], considered another definition of prediction interval consistency focusing on unconditional coverage, i.e., insisting that

$$Prob(y_f \in \Gamma) \to 1-\alpha. \tag{2.34}$$

However, the conditional coverage of definition 1 is a stronger property. To see why, define the random variables $U_n = Prob\left(y_f \in \Gamma | y\right)$, noting that y has dimension n. Then, the boundedness of U_n can be invoked to show that if $U_n \to_p 1-\alpha$, then $\mathbf{E}U_n \to 1-\alpha$ as well. Hence, (2.33) implies (2.34); see Zhang and Politis [2021a] for a further discussion on conditional vs. unconditional coverage.

Consider the prediction error $y_f - X_f\widehat{\theta} = \varepsilon_f - X_f(\widehat{\theta}-\beta)$. 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: ε_f and $-X_f(\widehat{\theta}-\beta)$. 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 ε_f may not have a normal distribution; so in order to approximate the distribution of $\varepsilon_f - X_f(\widehat{\theta}-\beta)$, 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 $F(x) = Prob\left(\varepsilon_1 \leq x\right)$ 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 $\mathbf{R}$.

this property is useful in the proof of lemma 1.

Lemma 1. *Suppose assumption 1 to 6 and 8 hold true. Define the residuals* $\widehat{\varepsilon}' = (\widehat{\varepsilon}_1', ..., \widehat{\varepsilon}_n')^T = y - X\widehat{\theta}$*, as well as the centered residuals* $\widehat{\varepsilon} = (\widehat{\varepsilon}_1, ..., \widehat{\varepsilon}_n)^T$ *with* $\widehat{\varepsilon}_i = \widehat{\varepsilon}_i' - \frac{1}{n}\sum_{i=1}^n \widehat{\varepsilon}_i'$*. If we let* $\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\widehat{\varepsilon}_i \leq x}$*, then*

$$\sup_{x \in \mathbf{R}} |\widehat{F}(x) - F(x)| \to_p 0 \text{ as } n \to \infty. \tag{2.35}$$

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 $\widehat{\varepsilon}_i, i = 1, 2, ..., n$ (in other words, generate random variables with distribution $\widehat{F}$) 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 ε_f with the normal approximation to the estimation error $-X_f(\widehat{\theta} - \beta)$.

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

1. Calculate $\widehat{\theta}$ *defined in* (2.17)*,* $\widehat{\sigma}$ *defined in* (2.18)*,* $\widehat{\varepsilon}$ *defined in lemma 1,* $\widehat{y}_f = (\widehat{y}_{f,1}, ..., \widehat{y}_{f,p_1})^T = X_f\widehat{\theta}$*, and* $\widehat{\theta}_\perp = Q_\perp Q_\perp^T \widehat{\theta}$*.*

2. Generate i.i.d. errors $\varepsilon^* = (\varepsilon_1^*, ..., \varepsilon_n^*)^T$ *with* $\varepsilon_i^*, i = 1, ..., n$ *having normal distribution with mean* 0 *and variance* $\widehat{\sigma}^2$*. Then generate i.i.d. errors* $\varepsilon_f^* = (\varepsilon_{f,1}^*, ..., \varepsilon_{f,p_1}^*)^T$ *with* $\varepsilon_{f,i}^*, i = 1, ..., p_1$ *having cumulative distribution function* $\widehat{F}$ *defined in lemma 1. Calculate* $y^* = X\widehat{\theta} + \varepsilon^*$*.*

3. Calculate $\widetilde{\theta}^{**} = (X^T X + \rho_n I_p)^{-1} X^T y^*$ *and* $\widetilde{\theta}^* = \widetilde{\theta}^{**} + \rho_n \times Q(\Lambda^2 + \rho_n I_r)^{-1} Q^T \widetilde{\theta}^{**} + \widehat{\theta}_\perp$. *Then derive* $\widehat{\mathscr{N}_{b_n}^*} = \left\{ i \middle| |\widetilde{\theta}_i^*| > b_n \right\}$, $\widehat{\theta}^* = (\widehat{\theta}_1^*, ..., \widehat{\theta}_p^*)^T$ *with* $\widehat{\theta}_i^* = \widetilde{\theta}_i^* \times \mathbf{1}_{i \in \widehat{\mathscr{N}_{b_n}^*}}$ *for* $i = 1, 2, ..., p$.

4. Calculate $y_f^* = (y_{f,1}^*, ..., y_{f,p_1}^*)^T = X_f \widehat{\theta} + \varepsilon_f^*$ *and* $\widehat{y}_f^* = (\widehat{y}_{f,1}^*, ..., \widehat{y}_{f,p_1}^*)^T = X_f \widehat{\theta}^*$. *Define* $E_b^* = \max_{i=1,2,...,p_1} |y_{f,i}^* - \widehat{y}_{f,i}^*|$.

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

$$\left\{ y_f = (y_{f,1}, ..., y_{f,p_1})^T \middle| \max_{i=1,2,...,p_1} |y_{f,i} - \widehat{y}_{f,i}| \leq C_{1-\alpha}^* \right\}. \tag{2.36}$$

If the design matrix X has rank p, then $\widehat{\theta}_\perp$ *is defined to be* 0.

Similar to section 2.5, here we define $c_{1-\alpha}^*$ as the $1-\alpha$ quantile of the conditional distribution $Prob^* \left(\max_{i=1,...,p_1} |y_{f,i}^* - \widehat{y}_{f,i}^*| \leq x \right)$, which can be approximated by $C_{1-\alpha}^*$ by letting $B \to \infty$. Theorem 4 below proves $Prob \left(\max_{i=1,2,...,p_1} |y_{f,i} - \widehat{y}_{f,i}| \leq c_{1-\alpha}^* \right) \to 1-\alpha$ as the sample size $n \to \infty$, which justifies the consistency of the prediction region (2.36).

Theorem 4. *Suppose assumptions 1 to 6 and 8 to 9 hold true (here consider* $M = (m_{ij})_{i=1,...,p_1, j=1,...,p}$ *in assumption 5 as* X_f*). Then*

$$\sup_{x \geq 0} |Prob^* \left(\max_{i=1,2,...,p_1} |y_{f,i}^* - \widehat{y}_{f,i}^*| \leq x \right) - Prob^* \left(\max_{i=1,2,...,p_1} |y_{f,i} - \widehat{y}_{f,i}| \leq x \right)| = o_p(1). \tag{2.37}$$

For any fixed $0 < \alpha < 1$, *it follows that*

$$Prob^* \left(\max_{i=1,2,...,p_1} |y_{f,i} - \widehat{y}_{f,i}| \leq c_{1-\alpha}^* \right) \to_p 1-\alpha \text{ as } n \to \infty. \tag{2.38}$$

Note that the bootstrap probability $Prob^*(\cdot)$ 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 $\widehat{\varepsilon}_i$; see Ch. 3 of Politis [2015] for a discussion.

2.7 Numerical Simulations

Define $k_n = \sqrt{n\log(n)}$ and the following four terms

$$\begin{aligned} \mathscr{K}_1 = \max_{i=1,2,\ldots,p_1} k_n \Big| \sum_{j\notin\mathscr{N}_{b_n}} m_{ij}\theta_j \Big|, \quad \mathscr{K}_2 = \max_{i=1,2,\ldots,p_1} k_n \Big| \sum_{j=1}^{r} m_{ij}\theta_{\perp,j} \Big|, \\ \mathscr{K}_3 = b_n \sum_{j\notin\mathscr{N}_{b_n}} |\theta_j|, \quad \mathscr{K}_4 = \frac{\sqrt{|\mathscr{N}_{b_n}|}}{\lambda_r}; \end{aligned} \tag{2.39}$$

see section 2.3 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 $\mathscr{K}_i$ 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 β through the following strategies:

Design matrix X: define $X = [x_1,\ldots,x_n]^T$ with $x_i = (x_{i1},\ldots,x_{ip})^T \in \mathbf{R}^p, i = 1,\ldots,n$. Generate $x_1, x_2, \ldots$ as i.i.d. normal random vectors with mean 0 and covariance matrix $\Sigma \in \mathbf{R}^{p\times p}$. We

choose Σ with diagonal elements equal to 2.0 and off-diagonal elements equal to 0.5.

$$M = \begin{bmatrix} m_{11} & m_{12} & ... & m_{1\tau} & m_{1\tau+1} & ... & m_{1p} \\ m_{21} & m_{22} & ... & m_{2\tau} & m_{2\tau+1} & ... & m_{2p} \\ \vdots & \vdots & ... & \vdots & \vdots & ... & \vdots \\ m_{|\mathscr{M}|1} & m_{|\mathscr{M}|2} & ... & m_{|\mathscr{M}|\tau} & m_{|\mathscr{M}|\tau+1} & ... & m_{|\mathscr{M}|p} \\ 0 & 0 & ... & 0 & m_{|\mathscr{M}|+1\tau+1} & ... & m_{|\mathscr{M}|+1p} \\ \vdots & \vdots & ... & \vdots & \vdots & ... & \vdots \\ 0 & 0 & ... & 0 & m_{p_1\tau+1} & ... & m_{p_1 p} \end{bmatrix} \tag{2.40}$$

***M* and β when $p < n$:** choose $\tau = 50$ in (2.40). Generate $m_{ij}^{'}, i = 1,2,...,|\mathscr{M}|,\ j = 1,2,...,\tau$ as i.i.d. normal with mean 0.5 and variance 1.0, and generate $m_{ij}^{'}, i = 1,2,...,p_1, j = \tau + 1,...,p$ 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 $i = 1,2,...,|\mathscr{M}|,\ j = 1,2,...,\tau$; $m_{ij} = 4.0 \times m_{ij}^{'} / \sqrt{\sum_{j=\tau+1}^{p} m_{ij}^{'2}}$ for $i = 1,2,...,|\mathscr{M}|,\ j = \tau+1,...,p$; and $m_{ij} = 6.0 \times m_{ij}^{'} / \sqrt{\sum_{j=\tau+1}^{p} m_{ij}^{'2}}$ for $i = |\mathscr{M}|+1,...,p_1, j = \tau+1,...,p$. Choose $\beta = (\beta_1,...,\beta_p)^T$ 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 β when $p > n$:** choose $\tau = 6$ in (2.40). Generate $m_{ij}^{'}, i = 1,2,...,|\mathscr{M}|,\ j = 1,2,...,\tau$ as i.i.d. normal with mean 0.5 and variance 1.0, and generate $m_{ij}^{'}, i = 1,2,...,p_1,\ j = \tau+1,...,p$ 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 $i = 1,2,...,|\mathscr{M}|,\ j = 1,2,...,\tau$; and $m_{ij} = m_{ij}^{'} / \sqrt{\sum_{j=\tau+1}^{p} m_{ij}^{'2}}$ for $i = 1,2,...,p_1,\ j = \tau+1,...,p$. Choose $\beta_i = 1.0, i = 1,2,3$, $\beta_i = -1.0, i = 4,5,6$, and 0 otherwise. When $p > n$, β may not be identifiable (Shao and Deng [2012]), and β may not equal θ(defined in section 2.3) despite $X\beta = X\theta$. We consider both situations and evaluate the performance of proposed methods on the linear model $y = X\beta + \varepsilon$ and $y = X\theta + \varepsilon$. We fix X and M in each simulation.

The different regression algorithms considered are the debiased and threshold ridge regression(Deb 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 + Ridge (Post Ridge). We consider 6 cases for simulation involving a different p/n ratio, and Normal vs. Laplace (2-sided exponential) errors; we present detailed information about each simulation case in table 2.1, compare the performance of different regression algorithms in Figure 2.2 to 2.4 and Table 2.2, and record the performance of bootstrap algorithms on estimation/hypothesis testing and interval-prediction in Table 2.3 and Figure 2.5. The optimal ridge parameter ρ_n and threshold b_n are chosen by 5-fold cross validation. To adapt to assumption 9, we choose X_f as the first 100 lines of M for prediction.

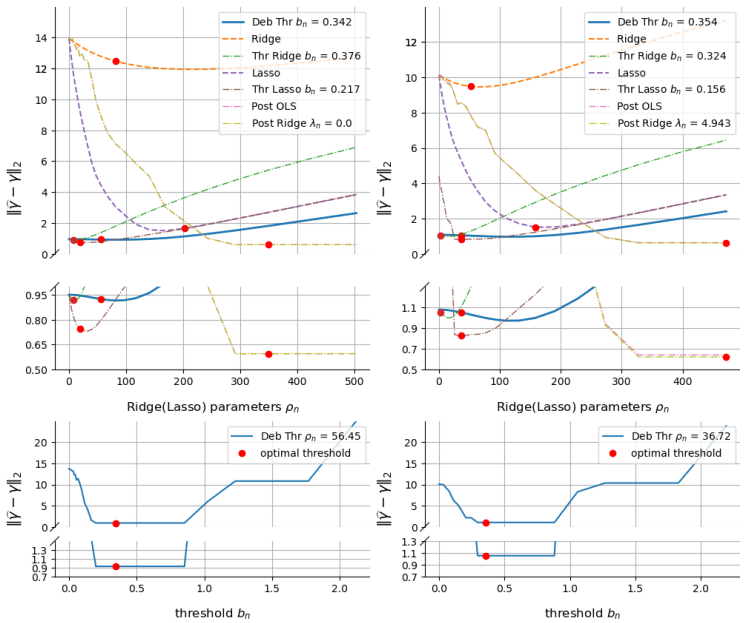
Table 2.1. Information about X, M and ε in each simulation case. For the normal distribution we choose variance 4, for the Laplace distribution we choose the scale $\sqrt{2}$. By doing this, the variance of residuals is 4. When $p > n$, $\beta \neq \theta$. **The left(right) side of the slashes represent $\mathscr{K}_2$ calculated by the linear model $y = X\beta + \varepsilon$ ($y = X\theta + \varepsilon$).** The difference between β and θ does not change other terms in case 5 and 6.

Case	p	p_1	$\lvert\mathscr{M}\rvert$	λ_r	ρ_n	b_n	$\mathscr{K}_1$	$\mathscr{K}_2$	$\mathscr{K}_3$	$\mathscr{K}_4$
1	500	800	300	12.978	56.453	0.343	1.370	0.0	0.013	1.712
2	500	800	300	12.561	36.728	0.354	1.636	0.0	0.014	1.769
3	650	800	300	8.226	56.432	0.396	1.553	0.0	0.016	3.085
4	500	800	700	12.847	55.317	0.346	1.510	0.0	0.014	1.730
5	1500	800	300	9.766	1.201	0.228	6.938	129 / 0.0	8.214	3.962
6	1500	800	300	9.766	1.201	0.228	6.938	129 / 0.0	8.214	3.962

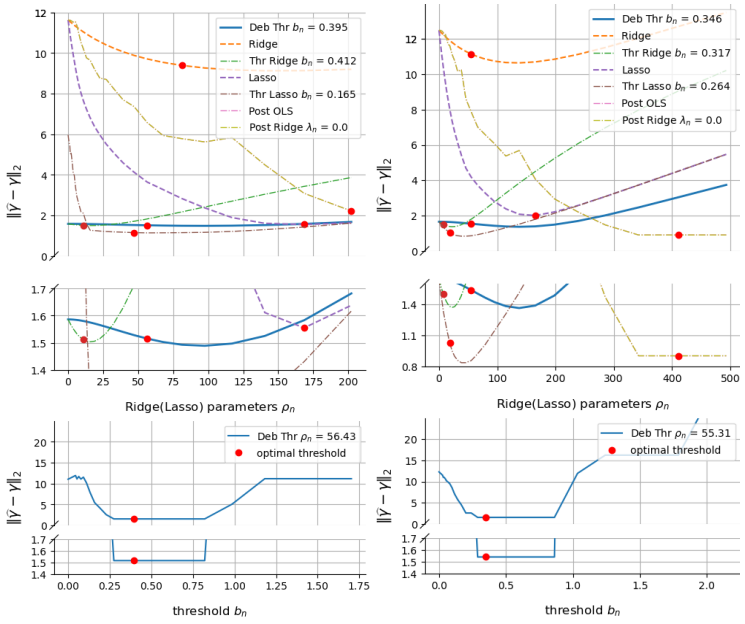
Case 5 and 6 consider both the linear model $y = X\beta + \varepsilon$ and $y = X\theta + \varepsilon$, here $\beta \neq \theta = QQ^T\beta$. The difference in β and θ affects the value of $\mathscr{K}_2$(but does not affect others), so we have two values in table 2.1.

Figure 2.2 plots the Euclidean norm $\|\widehat{\gamma} - \gamma\|_2$, with $\widehat{\gamma}$ defined in (2.17), and γ defined in Section 2.3, 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 ρ_n, i.e., $\|\widehat{\gamma} - \gamma\|_2$ can be significantly larger than its minimum despite ρ_n is close to the minimizer of $\|\widehat{\gamma} - \gamma\|_2$.

In reality, cross validation does not necessarily guarantee selection of the optimal ρ_n and b_n, so it is risky to use the threshold ridge regression method. Debiasing helps decrease the

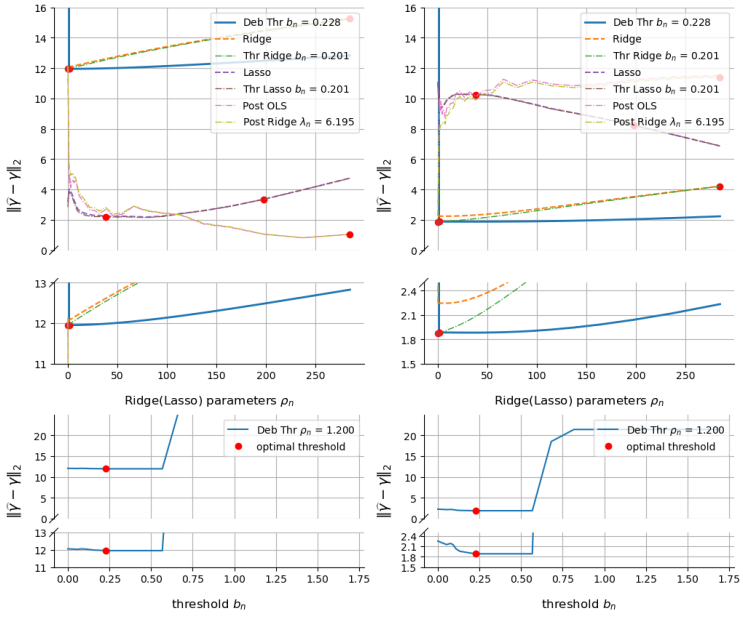


(a) Case 1 **(b)** Case 2



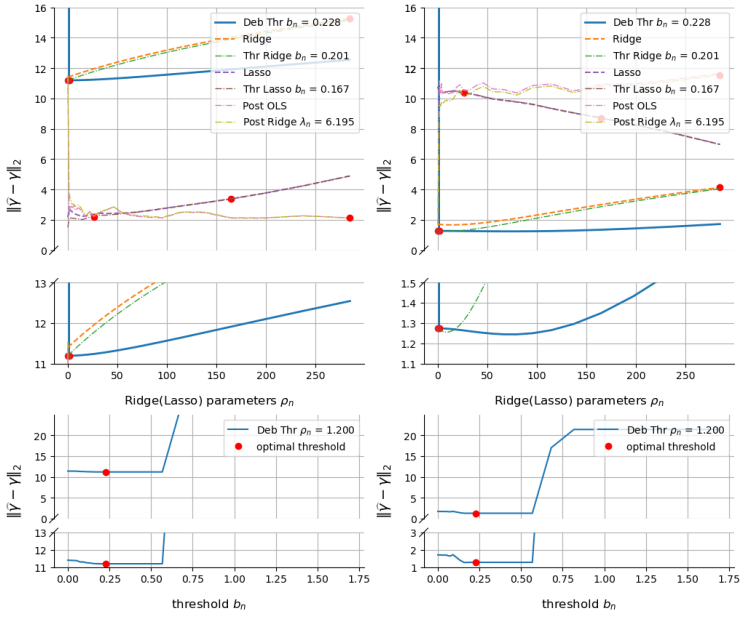
(c) Case 3 **(d)** Case 3

Figure 2.2. 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 parameters selected by 5-fold cross validation. The vertical axis represents the Euclidean norm of $\widehat{\gamma} - \gamma$(see (2.17) and section 2.3). 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.



(a) Case 5(use β)

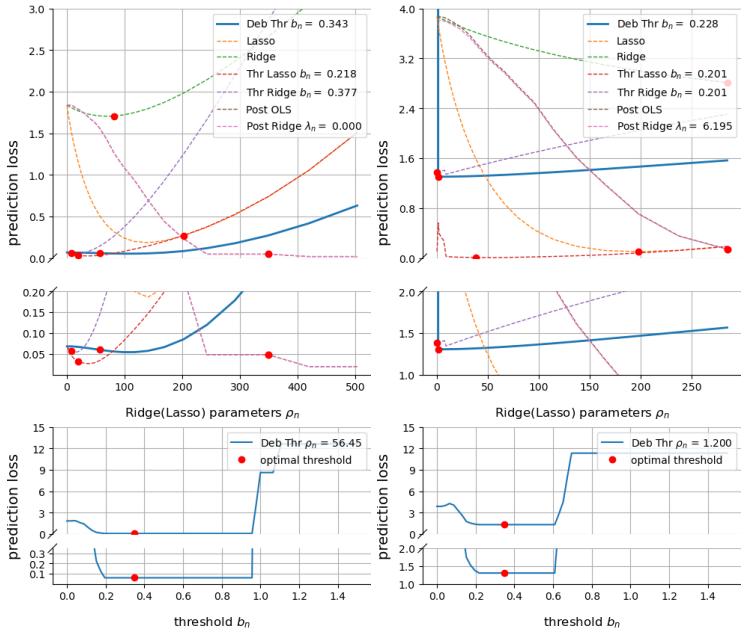
(b) Case 5(use θ)



(c) Case 6(use β)

(d) Case 6(use θ)

Figure 2.3. Estimation performance of various linear regression methods over case 5 to 6. The meaning of symbols coincides with figure 2.2.



(a) Case 1 **(b)** Case 5

Figure 2.4. Prediction loss(see section 2.4) of various linear regression methods. The meaning of symbols coincides with figure 2.2.

ridge regression estimator's error; more importantly, it is robust to changes in the choice of ρ_n. Even if a cross validation selects a sub-optimal ρ_n, the error of the debiased and threshold ridge regression estimator does not surge, and the estimator's performance does not notably deteriorate. On the other hand, in Figure 2.2 and 2.3, $\|\widehat{\gamma}-\gamma\|_2$ 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 $\|\widehat{\gamma}-\gamma\|_2$). Since there is a wide interval of thresholds b_n that have small $\|\widehat{\gamma}-\gamma\|_2$, 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 β and θ(see section 2.3) could be considered as the 'parameters' for the linear model. Lasso methods estimate linear combinations of β, while ridge regression methods estimate linear combinations of θ. Under this situation, the difference between β and θ is the main factor for the estimators' error. In reality, statisticians cannot distinguish between β and θ 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 ρ_n, b_n. When $p > n$, the sparsity of θ is violated and a large bias is introduced to the estimator(see table 2.1). As a result the prediction loss will be enlarged.

As a summary of Figure 2.2 to 2.4, 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 ρ_n as well as the threshold b_n. Therefore, even when a sub-optimal ρ_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 θ rather than β to be the parameter of the linear model. So, in this case, ridge regression methods are suitable if the underlying linear model is indeed $y = X\theta + \varepsilon$(in other words, the projection does not have effect on the parameters of the linear model).

Table 2.2 demonstrates the model selection performance of various linear regression algorithms. Following Fithian et al. [2017], we evaluate the algorithms through the frequency of model misspecification(i.e., $\widehat{\mathscr{N}_{b_n}} \neq \mathscr{N}_{b_n}$), $P(\widehat{\mathscr{N}_{b_n}} \neq \mathscr{N}_{b_n})$; the average size of model misspecification $|\widehat{\mathscr{N}_{b_n}} \Delta \mathscr{N}_{b_n}|$ (here Δ denotes the symmetric difference, i.e. $A \Delta B = (A - B) \cup (B - A)$); and the average false discovery rate $|\widehat{\mathscr{N}_{b_n}} - \mathscr{N}_{b_n}| / \max(|\widehat{\mathscr{N}_{b_n}}|, 1)$. Notice that Lasso and the ridge

regression do not have thresholds. For these algorithms we say $i \in \widehat{\mathscr{N}_{b_n}}$ 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 $i \notin \mathscr{N}_{b_n}$. Despite the stochastic errors are still small, the summation of θ_i and the stochastic error can exceed b_n, which results in selecting a false model.

Table 2.2. Model selection performance of various linear regression methods over case 1 and 5. The hyper-parameters are chosen by 5-fold cross validation. The overscore represents calculating the sample mean among 1000 simulations.

Case	Algorithm	$P(\widehat{\mathscr{N}_{b_n}} \neq \mathscr{N}_{b_n})$	$\overline{\lvert\widehat{\mathscr{N}_{b_n}} \Delta \mathscr{N}_{b_n}\rvert}$	$\overline{\text{False discovery rate}}$
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 θ)	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 2.3 records the average errors of the proposed statistics $\widehat{\gamma}$(defined in (2.17)), $\widehat{\sigma}^2$(defined in (2.18)), and the coverage probability of the confidence region (2.29) as well as the coverage probability of the prediction region (2.36), in 1000 numerical simulations. We also record the frequency of model misspecification $P(\widehat{\mathscr{N}_{b_n}} \neq \mathscr{N}_{b_n})$. 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 $p > n$, θ is not necessarily sparse, and model misspecification may happen. Notably, $\widehat{\gamma}$'s error in estimating linear combinations of θ does not surge even when $p > n$. However, the difference between β and θ introduces a large bias to $\widehat{\gamma}$.

Besides, when $p > n$, assumption 6 can be violated. Correspondingly the variance estimator $\widehat{\sigma}^2$ may have a large error. The difference between β and θ invalidates the confidence region (2.29). For prediction region (2.36), 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 2.3. Frequency of model misspecification; average errors of $\widehat{\gamma}$ and $\widehat{\sigma}^2$; and the coverage probability for the confidence region (2.29) and the prediction region (2.36). 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 $B = 500$.

Estimation and Confidence region construction					Prediction
Case #	$P(\widehat{\mathcal{N}_{b_n}} \neq \mathcal{N}_{b_n})$	$\overline{\max_{i=1,2,\ldots,p_1} \lvert\widehat{\gamma}_i - \gamma_i\rvert}$	$\overline{\lvert\widehat{\sigma}^2 - \sigma^2\rvert}$	coverage	coverage
1	0.0	0.185	0.144	95.4%	91.5%
2	0.0	0.183	0.228	93.6%	90.4%
3	0.0	0.209	0.232	95.9%	92.6%
4	0.0	0.191	0.224	95.3%	90.6%
5(use β)	0.129	1.578	1.341	0.0%	97.2%
5(use θ)	0.122	0.258	1.354	97.6%	98.2%
6(use β)	0.126	1.579	1.342	0.0%	94.6%
6(use θ)	0.137	0.258	1.364	97.3%	92.8%

Figure 2.5 plots the power curve of the hypothesis test of $\gamma = \gamma_0$ versus $\gamma \neq \gamma_0$; here, we use $\gamma_0 = \gamma + \delta \times (1,1,...,1)^T$ and $\delta > 0$.

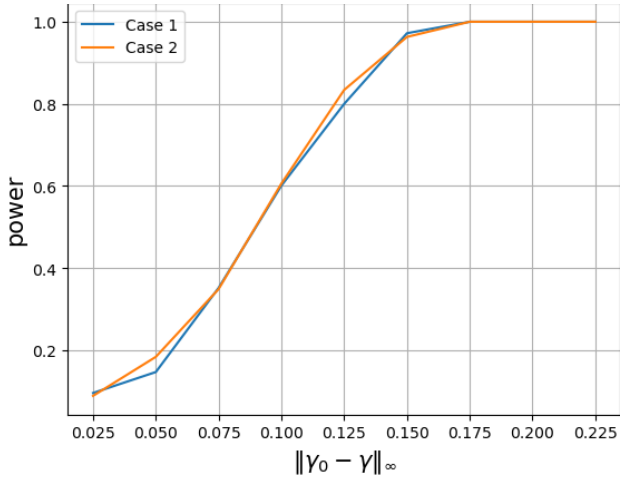


Figure 2.5. Power of the test for cases 1 and 2; the x-axis represents $\max_{i=1,\ldots,p_1} |\gamma_{0,i} - \gamma_i|$. Nominal size for the test is 5%; see algorithm 1 for the meaning of notations.

2.8 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 β. If the underlying parameter is indeed β 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 ρ_n as well as the threshold b_n. 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.

2.9 Acknowledgement

Chapter 2 is based on the paper "Ridge Regression Revisited: Debiasing, Thresholding and Bootstrap" by Y.Zhang and D.N. Politis and has been accepted for publication in *Annals of Statistics*. The dissertation author was the primary investigator and author of this paper.

Chapter 3

Debiased and thresholded ridge regression for linear models with heteroskedastic and correlated errors

3.1 Abstract

High-dimensional linear models with independent errors have been well-studied. However, statistical inference on a high-dimensional linear model with heteroskedastic, dependent (and possibly non-stationary) errors is still a novel topic. Under such complex assumptions, the paper at hand introduces a debiased and thresholded ridge regression estimator that is consistent, and is able to recover the model sparsity. Moreover, we derive a Gaussian approximation theorem for the estimator, and apply a dependent wild bootstrap algorithm to construct simultaneous confidence interval and hypothesis tests for linear combinations of parameters. Numerical experiments with both real and simulated data show that the proposed estimator has good finite sample performance. Of independent interest is the development of a new class of heteroscedastic, (weakly) dependent, and non-stationary random variables that can be used as a general model for regression errors.

Keywords: Linear regression, High dimensional data, Regularization, Dependent errors, Bootstrap.

3.2 Introduction

Linear regression is a fundamental topic in statistical inference. The classical setting assumes the dimension of parameters in a linear model is constant, and the errors are independent and identically distributed (i.i.d.). Recently, researchers have been working with high-dimensional linear models, i.e., the case where the dimension is allowed to diverge, but typically with i.i.d. errors. Under such a setting, research has been carried our on parameter estimation, e.g., Zou and Hastie [2005] and Zou [2006]; confidence interval construction/hypothesis testing, e.g., Chatterjee and Lahiri [2010, 2011]; and prediction, e.g., Stine [1985] and Zhang and Politis [2021a]. We also refer to Seber and Lee [2003], Hastie et al. [2009] and Fan et al. [2020] for a comprehensive introduction.

In practice, however, errors in a linear model can be dependent, and may have different distributions. As suggested by Vogelsang [2012] and Petersen [2008], heteroscedasticity, autocorrelation and spatial correlation can be present in panel data. If the errors are not i.i.d., then confidence intervals developed under the i.i.d. assumption may fail to capture the correct coverage probability. Several tools are developed to adapt to non-i.i.d. errors. 10. and Kim and Sun [2011] considered estimating the ordinary least square estimator's covariance matrix; Kelejian and Prucha [2007] and Vogelsang [2012] proposed the consistent test statistics for parameters; Sun and Wang [2021] and Conley et al. [2019] worked on statistical inference and hypothesis testing, etc. Resampling methods can also be used with dependent errors; see e.g. Politis et al. [1999] and Shao [2010]. Despite such accomplishments, the aforementioned works assumed that the dimension of parameters is fixed.

In the Big Data era, a practical situation to be handled via linear model may require many parameters, sometimes even more than the sample size. If this happens, statisticians cannot assume that the number of parameters is fixed, and the theoretical results, including the consistency and central limit theorem (CLT) of the estimators, are no longer obvious. In order to perform statistical inference, statisticians need to impose restrictions on the parameters. A

typical restriction is that the underlying parameter vector is sparse, i.e., containing many zeros. For a sparse linear model, Lasso is a suitable algorithm since it conducts an implicit model selection, i.e., zeroing out parameters that are not significant, see Tibshirani [2011]. More recent work includes Zhao and Yu [2006], Meinshausen and Bühlmann [2006] and Meinshausen and Yu [2009] for model selection; Zhang and Zhang [2014], Zhang and Cheng [2017] and Chatterjee and Lahiri [2010, 2011] for statistical inference and hypothesis testing; Greenshtein and Ritov [2004] for prediction and Zou [2006] for algorithm improvement. We refer to Bühlmann and van de Geer [2011] for a comprehensive overview of the Lasso method on high dimensional data sets.

Lasso is not the unique choice for regularizing a high-dimensional linear model. Fan and Li [2001] introduced a new penalty function, called SCAD, that is continuously differentiable and maintains the sparsity of the underlying model. Lee et al. [2016], Liu and Yu [2013] and Tibshirani et al. [2018] introduced *Post-selection inference*, i.e., performing model selection with Lasso, then fitting ordinary least square regression on the selected model. Shao and Deng [2012] applied threshold on the ridge regression estimator to recover the sparsity of a linear model. Zhang and Politis [2020] recently showed that, after debiasing and thresholding, the ridge regression estimator had a comparable performance to threshold Lasso and post-selection inference.

All the above works operate under the assumption of i.i.d. errors. The paper at hand focuses on statistical inference, i.e., point estimation, construction of confidence intervals and hypothesis tests, in a high dimensional linear model with the presence of dependent and heteroskedastic errors. Non-i.i.d. errors have been studied in fixed dimensional linear models. However, performing statistical inference for a high dimensional linear model with dependent errors is challenging. Wu and Wu [2016] proposed an oracle inequality; Han and Tsay [2020] proved the consistency of the Lasso estimator; Yuan and Guo derived the central limit theorem for the desparsified Lasso estimator. All these works rely on the assumption that the errors are stationary—see e.g., definition 1.3.3 in Brockwell and Davis [1991]. In order to address

the general problem with possibly non-stationary errors, we chose to work with the (debiased and thresholded) ridge estimator as it admits a closed-form formula, making it easier to derive theoretical guarantees. Moreover, it has good performance in the i.i.d. error case as Zhang and Politis [2020] showed.

To define our setup, consider a high dimensional, sparse linear model $y = X\beta + \varepsilon$ where $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T$. X is the fixed design matrix, $\beta = (\beta_1, ..., \beta_p)^T$ are the unknown parameters, and $y = (y_1, ..., y_n)^T$ are the response variables. The vector ε is a finite stretch of the stochastic process $\varepsilon_t, t = 1, 2, \ldots$ which is not assumed to be stationary or linear. We focus on estimating linear combinations of parameters, say $\gamma = M\beta$ with M a given matrix. Our work includes proving the consistency of the (debiased and thresholded) ridge estimator for γ. We will also derive a Gaussian approximation theorem and construct simultaneous confidence intervals for the coordinates of γ. We are also interested in testing the statistical hypothesis

$$\text{null: } M\beta = \zeta \text{ versus the alternative: } M\beta \neq \zeta \tag{3.1}$$

with ζ a given vector. To achieve this goal, we adapt the dependent wild bootstrap of Shao [2010], and provide its theoretical guarantee in our context .

The novelty of the paper at hand comes from the following aspects:

- It constructs consistent simultaneous confidence interval for γ under the situation that the errors ε have a complex covariance matrix. Notably, the high dimensional linear regression literature —see e.g., Zhang and Zhang [2014]— has invariably assumed that the errors were independent.

- The existing literature has focused on hypothesis testing for elements of β, while our work allows statisticians to test the linear combinations $M\beta$.

- A new class of heteroscedastic, (weakly) dependent, and non-stationary random variables is developed that can be used as a general model for regression errors in many contexts.

- In comparison to the paper by Zhang and Politis [2020], the current work not only allows dependent and heteroskedastic errors; it further identifies realistic assumptions on the parameters that allow for consistent estimation when the parameter dimension is larger than the sample size.

Since there is little research on statistical inference for a high dimensional linear model with non-i.i.d. errors, this paper at hand should shed some light on this field.

The remainder of this paper is organized as follows: section 3.3 introduces a new class of non-stationary random variables, called $(m,\alpha)-$short range dependent random variables. Section 3.4 introduces the debiased and threshold ridge regression estimator. Moreover, it presents the consistency results and the Gaussian approximation theorem for the proposed estimator. Section 3.5 constructs simultaneous confidence intervals for $\gamma = M\beta$, and tests the null hypothesis $\gamma = \zeta$ versus the alternative hypothesis $\gamma \neq \zeta$ via the dependent wild bootstrap. Section 3.6 presents numerical experiments with both real and simulated data to demonstrate the finite sample performance of the proposed estimator and the bootstrap algorithm. Section 3.7 contains our conclusions; technical proofs are deferred to the online appendix .

Notations: This paper applies the standard order notation $O(\cdot), o(\cdot), O_p(\cdot), o_p(\cdot)$: for two numerical sequences a_n, b_n, we say $a_n = O(b_n)$ if there exists a constant $C > 0$ such that $|a_n| \leq C|b_n|$ for $n = 1,2,...$; 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 $0 < \varepsilon < 1$, there exists a constant $C_\varepsilon > 0$ such that $Prob(|X_n| \leq C_\varepsilon |Y_n|) \geq 1 - \varepsilon$ for any n; and $X_n = o_p(Y_n)$ if $X_n/Y_n \to_p 0$ where the latter denotes convergence in probability; see definition 1.9 and chapter 1.5.1 in Shao [2003] for further details. *All order notations and convergence results are understood to hold true as the sample size* $n \to \infty$.

The symbol $\exists$ and $\forall$ respectively means 'there exists' and 'for all'. For a vector $a = (a_1,...,a_p)^T \in \mathbf{R}^p$, define its norm $|a|_q = (\sum_{i=1}^p |a_i|^q)^{1/q}$, here $q \geq 1$. Moreover, define $|a|_\infty = \max_{i=1,...,p} |a_i|$ and $|a|_0 = \sum_{i=1}^p \mathbf{1}_{a_i \neq 0}$, i.e., the number of non-zero elements presenting in a. For

a matrix T, define the operator norm $|T|_2 = \max_{|a|_2=1} |Ta|_2$. For a random variable X, define its m norm $\|X\|_m = (\mathbf{E}|X|^m)^{1/m}$. Define $a \vee b = \max(a,b)$ and $a \wedge b = \min(a,b)$. We use the notation C to represent a generic constant, i.e., the value of C may change in different locations.

3.3 $(m,\alpha)-$short range dependent random variables

In order to derive theoretical results, statisticians need to assume that the random variables satisfy some conditions. In the time series literature, random variables are often assumed to be stationary; see definition 1.3.3 in Brockwell and Davis [1991]. If this assumption holds true, then the covariance matrix of a finite stretch of these random variables will be a Toeplitz matrix, see section 2 in McElroy and Politis [2020] and section 0.9.7 in Horn and Johnson [2013]. However, the Toeplitz structure is often too restrictive to model the covariance matrix of regression errors. In other words, we may need to assume that the errors in the linear model are not stationary.

To perform time series analysis, going beyond stationarity often entails some form of local stationarity, a concept that was pioneered by Priestley [1988] and Dahlhaus [1997]. Examples include the time-varying coefficient models, e.g. Giraitis et al. [2014], and Dahlhaus and Subba Rao [2006], as well as nonparametric locally stationary setups, e.g. Das and Politis [2021], Dahlhaus et al. [2019], Zhou [2014], and Dette and Wu [2022]. Zhang and Wu [2021] introduced a special form of nonparametric locally stationary process and Dahlhaus et al. [2019] derived the law of large number and the central limit theorem for that process. Besides, Wu and Zhou [2011] introduced a Gaussian approximation theorem for the partial sum process of another type of non-stationary process.

In this section, we would like to introduce a new class of non-stationary random variables, called $(m,\alpha)-$short range dependent random variables that are not (necessarily) locally stationary. Specifically, this section provides a Gaussian approximation theorem for linear combinations of those random variables. Analyzing linear combinations of random variables is important in many statistical application, such as linear regression, non-linear regression, time series, etc.

Therefore, this section should be helpful to readers with different backgrounds.

Suppose $e_i, i \in \mathbf{Z}$ are independent (non necessarily identically distributed) random variables. Using these, we may define a new set of random variables $\varepsilon_i, i = 1, 2, ..., n$ by the relation

$$\varepsilon_i = g_i(..., e_{i-2}, e_{i-1}, e_i) \tag{3.2}$$

where g_i is a measurable function for each i; note that g_i can vary with respect to different i. Define $\mathscr{F}_i$ as the $\sigma-$field generated by $..., e_{i-2}, e_{i-1}, e_i$; hence, ε_i is $\mathscr{F}_i$ measurable. Consider independent random variables $e_i^\dagger, i \in \mathbf{Z}$ such that $e_i^\dagger$ has the same distribution as e_i for any i. Also assume that the sequence $e_i^\dagger, i \in \mathbf{Z}$ is independent to the sequence $e_i, i \in \mathbf{Z}$.

Define

$$\varepsilon_{i,j} = \begin{cases} g_i(..., e_{i-j-2}, e_{i-j-1}, e_{i-j}^\dagger, e_{i-j+1}, ..., e_{i-1}, e_i) \text{ for } j \geq 0 \\ \varepsilon_i \text{ for } j < 0 \end{cases} \tag{3.3}$$

and note that $\varepsilon_{i,j}$ has the same marginal distribution as ε_i. For any $j \geq 0$, define $\mathscr{F}_{i,j}$ as the $\sigma-$field generated by $e_{i-j}, e_{i-j+1}, ..., e_i$. Let $\delta_{i,j,m} = \|\varepsilon_i - \varepsilon_{i,j}\|_m$; clearly, $\delta_{i,j,m} = 0$ for $j < 0$.

We are now ready to define a new class of non-stationary random variables that are not (necessarily) locally stationary and may be helpful in general regression settings.

Definition 2 ($(m, \alpha)-$short range dependent random variables)**.** *Consider two constants $m \geq 2$ and $\alpha > 1$. We say the sequence $\{\varepsilon_i\}_{i=1,...,n}$ is $(m, \alpha)-$short range dependent if ε_i satisfies* (3.2)*, $\mathbf{E}\varepsilon_i = 0$ for any $i = 1, 2, ..., n$, and $\forall n \in \mathbf{N}$*

$$\begin{gathered} \sup_{k=0,1,...} (k+1)^\alpha \sum_{j=k}^{\infty} \max_{i=1,2,...,n} \delta_{i,j,m} = O(1) \\ \text{and} \quad \max_{i=1,2,...,n} \|\varepsilon_i\|_m = O(1) \end{gathered} \tag{3.4}$$

In definition 2, the sequence $\{\varepsilon_i\}_{i=1,...,n}$ should be recognized as the nth row of a triangular array

of random variables, i.e., ε_i may depend on n; for conciseness, the dependence of ε_i on n will not be explicitly denoted.

Example 2 shows that definition 2 is not a special case of the locally stationary process of Zhang and Wu [2021].

Example 2. *Suppose $e_i, i \in \mathbf{Z}$ are i.i.d. standard normal random variables. Set $\varepsilon_{2i} = e_{2i}$ and $\varepsilon_{2i-1} = e_{2i-2}e_{2i-1}$ for $i \in \mathbf{Z}$. Then $\delta_{i,j,m} = 0$ for $j > 2$, so ε_i is $(m,\alpha)-$short range dependent for any m, α. On the other hand, $\|e_{2i} - e_{2i}e_{2i-1}\|_m = \|e_{2i}\|_m \times \|1 - e_{2i-1}\|_m$, which does not shrink to 0 as the sample size $n \to \infty$. In other words, eq. (1.2) in Zhang and Wu [2021] is not satisfied.*

Remark 4. *In time series literature (e.g., section 1.2 in Reinsel [1993]), ε_i is considered to be causal, i.e., ε_i does not depend on the future innovations e_j with $j > i$. Wu [2005] introduced the form* (3.2)*, but assumed that the e_i in* (3.2) *were i.i.d. and g_i did not depend on the index i, which made ε_i stationary. The assumptions used in Wu and Zhou [2011] was similar to definition 2. However, they assumed that g_i in* (3.2) *was fixed, while we allow g_i to change with respect to the sample size n. On the other hand, the locally stationary condition(e.g., eq.(1.2) in Zhang and Wu [2021]) assumed that*

$$\varepsilon_i = g\left(\frac{i}{n}, ..., e_{i-1}, e_i\right) \quad \text{where } g(t,\cdot) \text{ is a continuous function in } t \tag{3.5}$$

but our work does not require this continuity.

From corollary C.9 in Øksendal [2003], it follows that

$$\varepsilon_i = \lim_{j\to\infty} \mathbf{E}\varepsilon_i|\mathscr{F}_{i,j} = \mathbf{E}\varepsilon_i|\mathscr{F}_{i,0} + \sum_{j=1}^{\infty}\left(\mathbf{E}\varepsilon_i|\mathscr{F}_{i,j} - \mathbf{E}\varepsilon_i|\mathscr{F}_{i,j-1}\right) \tag{3.6}$$

almost surely. Moreover, $\lim_{j\to\infty}\|\varepsilon_i - \mathbf{E}\varepsilon_i|\mathscr{F}_{i,j}\|_m = 0$. Besides,

$$\|\mathbf{E}\varepsilon_i|\mathscr{F}_{i,j} - \mathbf{E}\varepsilon_i|\mathscr{F}_{i,j-1}\|_m = \|\mathbf{E}(\varepsilon_i - \varepsilon_{i,j})|\mathscr{F}_{i,j}\|_m \le \|\varepsilon_i - \varepsilon_{i,j}\|_m \le \max_{i=1,...,n}\delta_{i,j,m} \tag{3.7}$$

so definition 2 implies

$$\sup_{i=1,...,n,k\geq 1}(k+1)^{\alpha}\sum_{j=k}^{\infty}\|\mathbf{E}\varepsilon_i|\mathscr{F}_{i,j}-\mathbf{E}\varepsilon_i|\mathscr{F}_{i,j-1}\|_m = O(1) \tag{3.8}$$

The first lemma is similar to Whittle [1960], which bounds the moments of linear forms of ε_i.

Lemma 2. *Suppose $\varepsilon_i, i = 1,...,n$ are $(m,\alpha)-$short range dependent random variables with constants $m \geq 2$ and $\alpha > 1$.*

(i). There exists a constant C only depending on m such that

$$\|\sum_{i=1}^{n} a_i\varepsilon_i\|_m \leq C\sqrt{\sum_{i=1}^{n} a_i^2} \text{ for } \forall a_i \in \mathbf{R} \tag{3.9}$$

(ii). In particular, we have

$$\|\max_{i=1,...,p}|\sum_{j=1}^{p} a_{ij}\varepsilon_j|\,\|_m \leq Cp^{1/m}\max_{i=1,...,p}\sqrt{\sum_{j=1}^{n} a_{ij}^2} \text{ for } \forall a_{ij} \in \mathbf{R} \tag{3.10}$$

Define $\Sigma = \{\sigma_{ij}\}_{i,j=1,...,n}$ such that $\sigma_{ij} = \mathbf{E}\varepsilon_i\varepsilon_j$. If $m \geq 2$, then lemma 2 implies

$$|\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j\sigma_{ij}| = \|\sum_{i=1}^{n} a_i\varepsilon_i\|_2^2 \leq \|\sum_{i=1}^{n} a_i\varepsilon_i\|_m^2 \leq C\sum_{i=1}^{n} a_i^2 \tag{3.11}$$

with a constant C; recall that the value of the constant C may be different in different places. So the largest eigenvalue of Σ has order $O(1)$. Besides, for $i > j$

$$\begin{aligned}|\mathbf{E}\varepsilon_i\varepsilon_j| &= |\mathbf{E}\varepsilon_j\left(\varepsilon_i - \mathbf{E}\varepsilon_i|\mathscr{F}_{i,i-j-1}\right)| \leq \|\varepsilon_j\|_2 \times \|\varepsilon_i - \mathbf{E}\varepsilon_i|\mathscr{F}_{i,i-j-1}\|_2 \\ &\leq \|\varepsilon_j\|_m \times \sum_{s=i-j}^{\infty}\|\mathbf{E}\varepsilon_i|\mathscr{F}_{i,s} - \mathbf{E}\varepsilon_i|\mathscr{F}_{i,s-1}\|_m \leq \frac{C}{(1+i-j)^{\alpha}}\end{aligned} \tag{3.12}$$

for another constant C. Therefore, definition 2 implies that the covariance of ε_i exhibits a *polynomial decay* with respect to the lag $|i-j|$.

We will now derive a Gaussian approximation theorem for linear combinations of ε_i, i.e., $\sum_{j=1}^n a_{ij}\varepsilon_j$ with $a_{ij} \in \mathbf{R}, i = 1,...,p_1, j = 1,...,n$. In a classical central limit theorem (e.g., theorem 1.15 in Shao [2003]), each term $a_{ij}\varepsilon_j$ in the summation is assumed to be negligible. Our work also requires this condition.

Lemma 3 (Gaussian approximation theorem)**.** *Suppose $\varepsilon_1,...,\varepsilon_n$ are $(m,\alpha)-$short range dependent random variables with $m > 6$ and $\alpha > 1$. Define $\Sigma = \{\mathbf{E}\varepsilon_i\varepsilon_j\}_{i,j=1,...,n}$, suppose $\exists$ a constant $c_\Sigma > 0$ such that Σ's smallest eigenvalue is greater than c_Σ for any n. Let $a_{ij}, i = 1,...,p_1, j = 1,...,n$ be real numbers with $p_1 = O(1)$. Suppose $\exists$ two constants $0 < c \leq C < \infty$ such that $c \leq \sum_{j=1}^n a_{ij}^2 \leq C$ for $i = 1,...,p_1$; and $a^* = \max_{i=1,...,p_1, j=1,...,n} |a_{ij}| = o(n^{-1/4}\log^{-z}(n))$ with $z = \max\left(\frac{9}{2}, \frac{3\alpha}{2\alpha-2}\right)$. Then*

$$\sup_{x\in\mathbf{R}} |Prob\left(\max_{i=1,...,p_1} |\sum_{j=1}^n a_{ij}\varepsilon_j| \leq x\right) - Prob\left(\max_{i=1,...,p_1} |\sum_{j=1}^n a_{ij}\xi_j| \leq x\right)| = o(1) \tag{3.13}$$

where $\xi_1,...,\xi_n$ have joint normal distribution with $\mathbf{E}\xi_i = 0$ and $\mathbf{E}\xi_i\xi_j = \mathbf{E}\varepsilon_i\varepsilon_j$ for $i, j = 1,...,n$.

Notably, $1/\sqrt{n} = O(a^*)$. Otherwise the condition $\sum_{j=1}^n a_{ij}^2 \geq c$ cannot be satisfied.

Finally, we would like to present a method that estimates the variances and covariances of linear combinations $\sum_{j=1}^n a_{ij}\varepsilon_j$. If ε_i are i.i.d., this problem is quite simple. For example, we can estimate the variance of ε_i through $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \varepsilon_i^2$ and the estimator for $Var(\sum_{j=1}^n a_{ij}\varepsilon_j)$ is given by $\widehat{\sigma}^2\sum_{j=1}^n a_{ij}^2$. However, ε_i may have different variances and covariances. In this case $Var(\sum_{j=1}^n a_{ij}\varepsilon_j)$ has a complex expression and we need a new method to estimate $Var(\sum_{j=1}^n a_{ij}\varepsilon_j)$. We will use a kernel-based method; a kernel function $K(\cdot)$ is defined as follows:

Definition 3 (kernel function)**.** *Let a function $K(\cdot) : \mathbf{R} \to [0,\infty)$ be symmetric, continuously differentiable, $K(0) = 1$, $\int_\mathbf{R} K(x)dx < \infty$ and $K(x)$ is decreasing on $[0,\infty)$. Define the Fourier transformation of K as $\mathscr{F}K(x) = \int_\mathbf{R} K(t)\exp(-2\pi\mathrm{i}\times tx)dt$; here, $\mathrm{i} = \sqrt{-1}$. Assume $\mathscr{F}K(x) \geq 0$ for all $x \in \mathbf{R}$ and $\int_\mathbf{R} \mathscr{F}K(x)dx < \infty$.*

Following Shao [2010], we will call $K(\cdot)$ the kernel function. In the time series literature (e.g.,

Politis [2003] or Politis and White [2004]), $K(\cdot)$ might alternatively be called a 'covariance taper' or a 'lag-window'.

Remark 5. *Definition 3 is a bit different than the usual definition of a kernel function (see e.g., Hall and Huang [2001]) but yields some desirable properties. According to Shao [2010] and the Fourier inversion theorem (e.g. theorem 8.26 in Folland [1999]),* $\forall x = (x_1, ..., x_n)^T \in \mathbf{R}^n$ *and any positive number k,*

$$\begin{aligned} \sum_{s=1}^{n}\sum_{j=1}^{n} x_s x_j K\left(\frac{s-j}{k}\right) &= \int_{\mathbf{R}} \sum_{s=1}^{n}\sum_{j=1}^{n} x_s x_j \mathscr{F}K(z) \exp\left(2\pi \mathrm{i} z \frac{s-j}{k}\right) dz \\ &= \int_{\mathbf{R}} \mathscr{F}K(z) |\sum_{s=1}^{n} x_s \exp\left(\frac{2\pi \mathrm{i} z s}{k}\right)|^2 dz \geq 0 \end{aligned} \tag{3.14}$$

so the matrix $\left\{K\left(\frac{s-j}{k}\right)\right\}_{s,j=1,2,...,n}$ *is positive semi-definite. One possible kernel function is* $K(x) = \exp(-x^2/2)$*, whose Fourier transform is* $\mathscr{F}K(x) = \sqrt{2\pi}\exp(-2\pi^2 x^2)$.

Lemma 4 (estimated covariance matrix)**.** *Suppose random variables* $\varepsilon_i, i = 1, ..., n$ *are* (m, α) *- short range dependent random variables with* $m > 6$ *and* $\alpha > 1$*. Suppose* $\{a_{ij}\}_{i=1,...,p_1, j=1,...,n}$ *satisfy assumptions in lemma 3. Suppose* $K(\cdot)$ *is a kernel function (i.e., satisfies definition 3) and* k_n *is a positive number such that* $k_n \to \infty$ *as* $n \to \infty$*. Then*

$$\begin{aligned} \max_{i_1, i_2 = 1, ..., p_1} |\sum_{j_1=1}^{n}\sum_{j_2=1}^{n} a_{i_1 j_1} a_{i_2 j_2} K\left(\frac{j_1 - j_2}{k_n}\right) \varepsilon_{j_1}\varepsilon_{j_2} - \sum_{j_1=1}^{n}\sum_{j_2=1}^{n} a_{i_1 j_1} a_{i_2 j_2} \sigma_{j_1 j_2}| \\ = o_p(k_n \times n^{-1/4} \log^{-z}(n)) + O_p(v_n) \end{aligned} \tag{3.15}$$

where $\sigma_{j_1 j_2} = \mathbf{E}\varepsilon_{j_1}\varepsilon_{j_2}$ *and* $z = \max\left(\frac{9}{2}, \frac{3\alpha}{2\alpha - 2}\right)$.

$$v_n = \begin{cases} k_n^{1-\alpha} \text{ if } 1 < \alpha < 2 \\ \log(k_n)/k_n \text{ if } \alpha = 2 \\ 1/k_n \text{ if } \alpha > 2 \end{cases} \tag{3.16}$$

Notice that

$$Cov(\sum_{j=1}^{n} a_{i_1 j}\varepsilon_j, \sum_{j=1}^{n} a_{i_2 j}\varepsilon_j) = \sum_{j_1=1}^{n}\sum_{j_2=1}^{n} a_{i_1 j_1}a_{i_2 j_2}\sigma_{j_1 j_2}; \tag{3.17}$$

so lemma 4 gives us a way to consistently estimate the covariances of the linear combinations $\sum_{j=1}^{n} a_{ij}\varepsilon_j, i = 1,2,...,p_1$ using the estimator

$$\sum_{j_1=1}^{n}\sum_{j_2=1}^{n} a_{i_1 j_1}a_{i_2 j_2}K\left(\frac{j_1 - j_2}{k_n}\right)\varepsilon_{j_1}\varepsilon_{j_2}.$$

According to lemma 4, the above estimator will be consistent as long as we choose k_n in such a way that $k_n \times n^{-1/4}\log^{-z}(n) = O(1)$.

We also want to stress that the factor k_n is related to the notion of the bandwidth in kernel methods like Fan and Gijbels [1995] and Paparoditis and Politis [2000]. However, the usual notion is that the bandwidth converges to 0 whereas lemma 4 requires $k_n \to \infty$ so that the estimator will not ignore the long-term covariances (i.e., $\sigma_{j_1 j_2}$ with large $|j_1 - j_2|$). In this sense, the factor k_n can be understood as the inverse of the usual bandwidth; see also Politis [2003] for an analogous construction.

3.4 Consistency and Gaussian approximation

This section goes back to the original problem: suppose the $n \times p$ **fixed** design matrix $X = \{X_{ij}\}_{i=1,...,n, j=1,...,p}$ and the dependent variables $y = (y_1,...,y_n)^T$ satisfy a linear model

$$y = X\beta + \varepsilon, \text{ or equivalently } y_i = \sum_{j=1}^{p} X_{ij}\beta_j + \varepsilon_i \tag{3.18}$$

here $\beta = (\beta_1,...,\beta_p)^T$ is the parameter vector and $\varepsilon = (\varepsilon_1,...,\varepsilon_n)^T$ are the errors. Classical linear regression theory assumes that p is significantly smaller than the sample size n. However, in

many situations p may have a comparable size to n, or even $p > n$. Meanwhile, we suppose the errors are not i.i.d., i.e., $\mathbf{E}\varepsilon_i^2$ may vary with i, $\mathbf{E}\varepsilon_i\varepsilon_j$ may not equal 0 for $i \neq j$ and $\mathbf{E}\varepsilon_i\varepsilon_j$ may not be a function of just the lag $|i-j|$ indicating nonstationarity.

Suppose X has full rank, i.e., the rank $r = \min(p,n)$. Apply the thin singular value decomposition(theorem 7.3.2 in Horn and Johnson [2013]) on X, i.e., $X = P\Lambda Q^T$, P,Q is respectively $n \times r, p \times r$ orthonormal matrix, i.e., $P^T P = Q^T Q = I_r$, the $r-$th identity matrix. If $p > r$, define $Q_\perp \in \mathbf{R}^{p\times(p-r)}$ as the orthonormal counterpart of Q, so we have $QQ^T + Q_\perp Q_\perp^T = I_p$, the $p-$th identity matrix; $Q_\perp^T Q = 0$ and $Q_\perp^T Q_\perp = I_{p-r}$. $\Lambda = diag(\lambda_1, ..., \lambda_r)^T$ is the $r \times r$ diagonal matrix, $\lambda_i > 0$ for any i. Define $\theta = QQ^T\beta$ and $\theta_\perp = (\theta_{\perp,1}, ..., \theta_{\perp,p})^T = Q_\perp Q_\perp^T \beta$. If $p \leq n$, $Q_\perp$ does not exist and we define $\theta_\perp = 0$, the $p-$dimensional vector with elements 0. With this definition we have $\beta = \theta + \theta_\perp$.

In order to estimate β, first consider a ridge regression method

$$\begin{aligned} \widetilde{\beta}^\dagger &= (X^T X + \rho_{n,r} I_p)^{-1} X^T y \\ \text{implying } \widetilde{\beta}^\dagger - \beta &= -\rho_{n,r} Q(\Lambda^2 + \rho_{n,r} I_r)^{-1} Q^T \theta - \theta_\perp + Q(\Lambda^2 + \rho_{n,r} I_r)^{-1} \Lambda P^T \varepsilon. \end{aligned} \tag{3.19}$$

Following Bühlmann [2013] and Zhang and Politis [2020], we call the term $-\rho_{n,r} Q(\Lambda^2 + \rho_n I_r)^{-1} Q^T \theta$ the 'estimation bias' and $-\theta_\perp$ the 'projection bias'. Notably, when $p \leq n$, the projection bias vanishes. If p is large, then both estimation bias and projection bias will affect the performance of the ridge regression estimator; see remark 2 and 3 in Zhang and Politis [2020] and section 2.3 in Bühlmann [2013]. Worse still, the biases can have a larger order than the stochastic error, making it hard to construct a confidence interval.

To avoid these problems, this section proposes the debiased estimator $\widetilde{\beta}$ that diminishes the biases; first define the Lasso estimator

$$\widetilde{\beta}^{lasso} = (\widetilde{\beta}_1^{lasso}, ..., \widetilde{\beta}_p^{lasso})^T = \underset{z\in\mathbf{R}^p}{\operatorname{argmin}} \frac{1}{2n}|y - Xz|_2^2 + \rho_{n,l}|z|_1. \tag{3.20}$$

Define $\widetilde{\theta}_{\perp}^{\dagger} = (\widetilde{\theta}_{\perp,1}^{\dagger}, ..., \widetilde{\theta}_{\perp,p}^{\dagger})^T = Q_{\perp} Q_{\perp}^T \widetilde{\beta}^{lasso}$ if $p > n$ and $\widetilde{\theta}_{\perp}^{\dagger} = 0$ if $p \leq n$, and $s = (s_1, ..., s_r)^T = Q^T \beta (= Q^T \theta)$. Then, define the debiased estimator $\widetilde{\beta} = (\widetilde{\beta}_1, ..., \widetilde{\beta}_p)^T$ as

$$\begin{aligned} \widetilde{\beta} &= \widetilde{\beta}^{\dagger} + \rho_{n,r} Q(\Lambda^2 + \rho_{n,r} I_r)^{-1} Q^T \widetilde{\beta}^{\dagger} + \widetilde{\theta}_{\perp}^{\dagger} \\ \text{implying } \widetilde{\beta} - \beta &= -\rho_{n,r}^2 Q(\Lambda^2 + \rho_{n,r} I_r)^{-2} Q^T \theta \\ &+ Q\left((\Lambda^2 + \rho_{n,r} I_r)^{-1} + \rho_{n,r}(\Lambda^2 + \rho_{n,r} I_r)^{-2}\right) \Lambda P^T \varepsilon + \widetilde{\theta}_{\perp}^{\dagger} - \theta_{\perp} \\ \text{or equivalently } \widetilde{\beta}_i - \beta_i &= -\rho_{n,r}^2 \sum_{j=1}^{r} \frac{q_{ij} s_j}{(\lambda_j^2 + \rho_{n,r})^2} \\ &+ \sum_{j=1}^{r} \sum_{l=1}^{n} q_{ij} p_{lj} \left(\frac{\lambda_j}{\lambda_j^2 + \rho_{n,r}} + \frac{\rho_{n,r} \lambda_j}{(\lambda_j^2 + \rho_{n,r})^2} \right) \varepsilon_l + (\widetilde{\theta}_{\perp,i}^{\dagger} - \theta_{\perp,i}); \end{aligned} \tag{3.21}$$

here $P = \{p_{ij}\}_{i=1,...,n, j=1,...,r}$ and $Q = \{q_{ij}\}_{i=1,...,p, j=1,...,r}$. For a threshold $b_n > 0$, define

$$\widehat{\beta} = (\widehat{\beta}_1, ..., \widehat{\beta}_p)^T \text{ such that } \widehat{\beta}_i = \widetilde{\beta}_i \times \mathbf{1}_{|\widetilde{\beta}_i| > b_n} \tag{3.22}$$

In this paper, we call $\widehat{\beta}$ 'the debiased and thresholded ridge regression estimator'. The assumptions for this section are presented below.

Assumptions

1. The **fixed** design matrix X has rank $r = n \wedge p$. There exist constants $c_\lambda, C_\lambda > 0$ and $\eta \in (0, 1/2]$ such that

$$C_\lambda n^{1/2} \geq \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r \geq c_\lambda n^{\eta} \text{ for sufficiently large } n; \tag{3.23}$$

here $\lambda_1, ..., \lambda_r$ are the singular values of X. In addition, $\max_{i=1,...,n, j=1,...,p} |X_{ij}| = O(1)$.

2. $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T$ are $(m, \alpha_\varepsilon)-$ short range dependent random variables, i.e., ε satisfies definition 2. Here m and α_ε are constants such that $m > 3/\eta$ and $\alpha_\varepsilon > 1$. Define $\Sigma = \{\sigma_{ij}\}_{i,j=1,...,n} = \mathbf{E}\varepsilon\varepsilon^T$, and assume $\exists$ a constant $c_\Sigma > 0$ such that the smallest eigenvalue of Σ is greater than or equal to c_Σ.

3. $p = O(n^{\alpha_p})$ where α_p is a constant such that $\alpha_p > 0$ and $\alpha_p + 3 < m\eta$.

4. Define $\mathscr{N}_{b_n} = \{i = 1,2,...,p : |\beta_i| > b_n\}$, assume $|\mathscr{N}_{b_n}| = O(n^{\alpha_{\mathscr{N}}})$. $\alpha_{\mathscr{N}}$ is a constant such that $\alpha_{\mathscr{N}} < \frac{1}{3} - \frac{2\alpha_p}{3m}$, $\alpha_{\mathscr{N}} < \eta - \frac{\alpha_p+1}{m}$ and $\alpha_{\mathscr{N}} \geq 0$. $\rho_{n,r} = C_{\rho,r} \times n^{\alpha_r}$; here α_r is a constant satisfying $0 < \alpha_r < \frac{3\eta}{2} - \frac{\alpha_{\mathscr{N}}}{4}$. Meanwhile, $\rho_{n,l} = C_{\rho,l} \times n^{-\alpha_l}$; here α_l is a constant satisfying $\frac{3\alpha_{\mathscr{N}}}{2} \vee \frac{3}{m} < \alpha_l < \frac{1}{2} - \frac{\alpha_p}{m}$. $C_{\rho,r}$ and $C_{\rho,l}$ are two constants such that $0 < C_{\rho,r}, C_{\rho,l} < \infty$.

5. (Restricted eigenvalue condition) The restricted eigenvalue condition holds true, i.e.,

$$|Xz|_2 \geq c_\lambda n^{\frac{1}{2}} |z|_2 \text{ for all } z \in \mathscr{A} = \left\{ z = (z_1,...,z_p)^T \in \mathbf{R}^p : \sum_{i \notin \mathscr{N}_{b_n}} |z_i| \leq 3 \sum_{i \in \mathscr{N}_{b_n}} |z_i| \right\} \tag{3.24}$$

Moreover, assume $|\beta|_\infty = O(1)$ and $b_n = C_b \times n^{-\alpha_b}$. Here C_b is a constant such that $0 < C_b < \infty$ and $0 < \alpha_b < (\eta - \frac{\alpha_p+1}{m}) \wedge (\alpha_l - \frac{\alpha_{\mathscr{N}}}{2})$. Assume that there exists a constant $0 < c_b < 1$ such that $\min_{i \in \mathscr{N}_{b_n}} |\beta_i| > b_n / c_b$ and $|\beta_i| = 0$ for $i \notin \mathscr{N}_{b_n}$.

We will call $\rho_{n,r}, \rho_{n,l}$ and b_n the 'hyperparameters' for $\widehat{\beta}$. We will introduce a method to fine-tune these hyperparameters in section 3.6.1.

Intuitively, we require that the design matrix X is well-behaved, the parameter vector β is sparse and the errors ε have a finite high order moment. Our work assumes fixed design, i.e., no randomness involved in the design matrix X. However, in the case of random design, we would need to assume that the design matrix is independent of the errors ε; in that case, the results in our paper would still hold true after conditioning on X.

Remark 6. *To show that Assumption 1 is achievable, suppose X_{ij} are generated by i.i.d. random variables with mean* 0*, variance* 1 *and finite* 4*th moment. Assume* $\lim_{n\to\infty} p/n = z \in (0,\infty)$ *exists. Bai and Yin [1993] proved that the smallest non-zero eigenvalue of $\frac{1}{n}X^TX$ tends to $(1-\sqrt{z})^2$ and the largest eigenvalue of $\frac{1}{n}X^TX$ tends to $(1+\sqrt{z})^2$ almost surely. For the singular values of X are the square roots of the eigenvalues of X^TX. Then, Assumption 1 is achieved with $\eta = 1/2$ for sufficiently large n.*

Remark 7. *If $|\beta|_\infty = O(1)$(see section 3.2 for the definition of $|\cdot|_\infty$), we have*

$$|\beta|_2 = \sqrt{\sum_{i \in \mathcal{N}_{b_n}} \beta_i^2} = O\left(n^{\frac{1}{2}\alpha_{\mathcal{N}}}\right) \tag{3.25}$$

which allows us to discuss assumption 5, i.e., the restricted eigenvalue condition. This condition was presented in section 6.8 of Bühlmann and van de Geer [2011] and also in Raskutti et al. [2010] as a sufficient condition to prove the consistency of the Lasso estimator. As we use the Lasso to diminish the projection bias, we also need this condition to maintain the consistency of Lasso. Notably, eq.(3.24) *is not the only form of the restricted eigenvalue condition; see e.g., theorem 1 in Raskutti et al. [2010].*

Note that Bühlmann [2013], Shao and Deng [2012], and Zhang and Politis [2020], all applied ridge regression to fit a high dimensional linear model. However, Bühlmann [2013] only fixed the projection bias while Zhang and Politis [2020] mainly handled the estimation bias. Besides, those works supposed the errors in the linear model were independent. Our paper takes both projection and estimation bias into consideration, and also allows the errors to be correlated.

We present the (model-selection) consistency result in theorem 5.

Theorem 5. *Suppose assumptions 1 to 5 hold true. Define $\widehat{\mathcal{N}_{b_n}} = \{i = 1, 2, ..., p : |\widetilde{\beta}_i| > b_n\}$, then we have*

$$|\widetilde{\beta} - \beta|_\infty = O_p\left(n^{\frac{\alpha_p}{m} - \eta} + n^{\frac{\alpha_{\mathcal{N}}}{2} - \alpha_l}\right) \text{ and } Prob\left(\widehat{\mathcal{N}_{b_n}} \neq \mathcal{N}_{b_n}\right) \to 1 \tag{3.26}$$

as the sample size $n \to \infty$.

Remark 8. *In the online supplement (lemma B.3.1), we prove the consistency of the Lasso estimator, i.e.,*

$$|\widetilde{\beta}^{lasso} - \beta|_2 = O_p(n^{\frac{\alpha_{\mathcal{N}}}{2} - \alpha_l}) \text{ and } |\widetilde{\beta}^{lasso} - \beta|_1 = O_p(n^{\alpha_{\mathcal{N}} - \alpha_l}) \tag{3.27}$$

If $p \leq n$ and X has rank p, then $|\widetilde{\theta}_{\perp}^{\dagger} - \theta_{\perp}|_{\infty} = 0$ and eq.(3.26) *can be improved to*

$$|\widetilde{\beta} - \beta|_{\infty} = O_p\left(n^{\frac{\alpha_p}{m}-\eta}\right) \tag{3.28}$$

We now turn to estimating the linear combinations of β, i.e.,

$$\zeta = M\beta, \text{ where } M \text{ is a given } p_1 \times p \text{ linear combination matrix.} \tag{3.29}$$

An intuitive estimator for ζ is $\widetilde{\zeta}^{\dagger} = (\widetilde{\zeta}_1^{\dagger}, ..., \widetilde{\zeta}_{p_1}^{\dagger})^T = M\widehat{\beta}$. Define $c_{ij}^{\dagger} = \sum_{k \in \mathscr{N}_{b_n}} m_{ik} q_{kj}$. Assume $\widehat{\mathscr{N}_{b_n}} = \mathscr{N}_{b_n}$; then the difference between $\widetilde{\zeta}^{\dagger}$ and ζ becomes

$$\begin{aligned} \widetilde{\zeta}_i^{\dagger} - \zeta_i = \sum_{j \in \mathscr{N}_{b_n}} m_{ij}(\widetilde{\beta}_j - \beta_j) = -\rho_{n,r}^2 \sum_{k=1}^{r} \frac{c_{ik}^{\dagger} s_k}{(\lambda_k^2 + \rho_{n,r})^2} \\ + \sum_{k=1}^{r}\sum_{l=1}^{n} c_{ik}^{\dagger} p_{lk} \left(\frac{\lambda_k}{\lambda_k^2 + \rho_{n,r}} + \frac{\rho_{n,r}\lambda_k}{(\lambda_k^2 + \rho_{n,r})^2}\right) \varepsilon_l + \sum_{j \in \mathscr{N}_{b_n}} m_{ij}(\widetilde{\theta}_{\perp,j}^{\dagger} - \theta_{\perp,j}) \end{aligned} \tag{3.30}$$

If assumption 2 holds true and $\sum_{k=1}^{r} c_{ik}^{\dagger 2} \neq 0$, then

$$C_{\Sigma} \geq \frac{Var\left(\sum_{k=1}^{r}\sum_{l=1}^{n} c_{ik}^{\dagger} p_{lk} \left(\frac{\lambda_k}{\lambda_k^2+\rho_{n,r}} + \frac{\rho_{n,r}\lambda_k}{(\lambda_k^2+\rho_{n,r})^2}\right) \varepsilon_l\right)}{\sum_{k=1}^{r} c_{ik}^{\dagger 2} \left(\frac{\lambda_k}{\lambda_k^2+\rho_{n,r}} + \frac{\rho_{n,r}\lambda_k}{\lambda_k^2+\rho_{n,r}}\right)^2} \geq c_{\Sigma} \tag{3.31}$$

while from Cauchy inequality

$$|\sum_{k=1}^{r} \frac{c_{ik}^{\dagger} s_k}{(\lambda_k^2 + \rho_{n,r})^2}| \leq \sqrt{\sum_{k=1}^{r} \frac{c_{ik}^{\dagger 2} \lambda_k^2}{(\lambda_k^2 + \rho_{n,r})^2}} \times \sqrt{\sum_{k=1}^{r} \frac{s_k^2}{\lambda_k^2 \times (\lambda_k^2 + \rho_{n,r})^2}} \tag{3.32}$$

Eq.(3.31) and (3.32) ensure that the estimation bias $\rho_{n,r}^2 \sum_{k=1}^{r} \frac{c_{ik}^{\dagger} s_k}{(\lambda_k^2+\rho_{n,r})^2}$ can have a smaller order than the stochastic error $\sum_{k=1}^{r}\sum_{l=1}^{n} c_{ik}^{\dagger} p_{lk} \left(\frac{\lambda_k}{\lambda_k^2+\rho_{n,r}} + \frac{\rho_{n,r}\lambda_k}{(\lambda_k^2+\rho_{n,r})^2}\right) \varepsilon_l$ under some conditions. But the order of the projection error $\sum_{j \in \mathscr{N}_{b_n}} m_{ij}(\widetilde{\theta}_{\perp,j}^{\dagger} - \theta_{\perp,j})$ is difficult to control, and can be larger than the order of the stochastic error.

Constructing a consistent confidence interval for the Lasso estimator is difficult; see e.g., Zhang and Zhang [2014] and Celentano et al. [2020]. So we aim to find an estimator of ζ whose projection error has smaller order than the stochastic error. Define $V = \{V_{ij}\}_{i,j=1,...,p} = Q_\perp Q_\perp^T$ if $p > n$ and 0 if $p \le n$. Define the estimator $\widehat{\zeta}_i = \sum_{j=1}^p m_{ij}\nu_j \times \widehat{\beta}_j + \sum_{j\in\widehat{\mathscr{N}_{b_n}}} m_{ij} \times (1-\nu_j)\widetilde{\beta}_j^{lasso}$ and the term $c_{ij} = \sum_{k\in\mathscr{N}_{b_n}} m_{ik}\nu_k q_{kj}$. Here $\nu_j \in \mathbf{R}, j = 1, ..., p$ are parameters to derive. If $\widehat{\mathscr{N}_{b_n}} = \mathscr{N}_{b_n}$,

$$\begin{aligned}
\widehat{\zeta}_i - \zeta_i &= \sum_{j\in\mathscr{N}_{b_n}} m_{ij}\nu_j \times (\widetilde{\beta}_j - \beta_j) + \sum_{j\in\mathscr{N}_{b_n}} m_{ij}(1-\nu_j) \times (\widetilde{\beta}_j^{lasso} - \beta_j) \\
&= -\rho_{n,r}^2 \sum_{j=1}^r \frac{c_{ij}s_j}{(\lambda_j^2+\rho_{n,r})^2} + \sum_{j=1}^r \sum_{l=1}^n c_{ij}p_{lj}\left(\frac{\lambda_j}{\lambda_j^2+\rho_{n,r}} + \frac{\rho_{n,r}\lambda_j}{(\lambda_j^2+\rho_{n,r})^2}\right)\varepsilon_l \qquad (3.33) \\
&+ \sum_{j\in\mathscr{N}_{b_n}} m_{ij}(\nu_j V_{jj} + 1 - \nu_j) \times (\widetilde{\beta}_j^{lasso} - \beta_j) + \sum_{j\in\mathscr{N}_{b_n}} \sum_{k\neq j} m_{ij}\nu_j V_{jk}(\widetilde{\beta}_k^{lasso} - \beta_k)
\end{aligned}$$

In practice, $|V_{jk}|, j \neq k$ are always small but V_{jj} can be significantly larger than 0. We choose ν_j such that $\nu_j \times V_{jj} + 1 - \nu_j = 0$, which implies $\nu_j = \frac{1}{1-V_{jj}}$. Correspondingly

$$c_{ij} = \sum_{k\in\mathscr{N}_{b_n}} \frac{m_{ik}}{1-V_{kk}} q_{kj} \text{ and } \widehat{\zeta}_i = \sum_{j=1}^p \frac{m_{ij}}{1-V_{jj}} \times \widehat{\beta}_j - \sum_{j\in\widehat{\mathscr{N}_{b_n}}} \frac{m_{ij}V_{jj}}{1-V_{jj}} \widetilde{\beta}_j^{lasso} \qquad (3.34)$$

Notably, if $p \le n$ and the design matrix X has rank p, then $\nu_j = 1$ and $\widehat{\zeta}_i = \widetilde{\zeta}_i^{\dagger}$. We introduce extra assumptions needed to derive the asymptotic distribution of $\widehat{\zeta}_i$.

Additional assumptions I:

6. Define

$$\begin{aligned}
c_{ij} = \sum_{k\in\mathscr{N}_{b_n}} \frac{m_{ik}q_{kj}}{1-V_{kk}},\ \mathscr{M} &= \{i = 1, 2, ..., p_1 : \sum_{j=1}^r c_{ij}^2 > 0\} \\
w_{il} &= \sum_{j=1}^r c_{ij}p_{lj}\left(\frac{\lambda_j}{\lambda_j^2+\rho_{n,r}} + \frac{\rho_{n,r}\lambda_j}{(\lambda_j^2+\rho_{n,r})^2}\right) \qquad (3.35) \\
\text{and } \tau_i &= \sqrt{\sum_{k=1}^r c_{ij}^2\left(\frac{\lambda_k}{\lambda_k^2+\rho_{n,r}} + \frac{\rho_{n,k}\lambda_k}{(\lambda_k^2+\rho_{n,r})^2}\right)^2 + \frac{1}{n}}
\end{aligned}$$

Assume $\mathscr{M}$ is not empty, $p_1 = O(1)$ and

$$\max_{i\in\mathscr{M},l=1,\ldots,n} |\frac{w_{il}}{\tau_i}| = o(n^{-1/4}\log^{-z}(n)) \text{ with } z = \frac{9}{2} \vee \frac{3\alpha_\varepsilon}{2\alpha_\varepsilon - 2} \tag{3.36}$$

Besides, assume $\exists$ constants $0 < c_{\mathscr{M}} < C_{\mathscr{M}} < \infty$ such that

$$c_{\mathscr{M}} \leq \sum_{k=1}^{r} c_{ik}^2 \leq C_{\mathscr{M}} \text{ for all } i \in \mathscr{M} \tag{3.37}$$

7. Recall $V = \{V_{ij}\}_{i,j=1,\ldots,p} = Q_\perp Q_\perp^T$. Assume $V_{jj} < 1$ for all j and

$$\max_{k=1,\ldots,p,i=1,\ldots,p_1} |\frac{1}{\tau_i} \sum_{j\in\mathscr{N}_{b_n}, j\neq k} \frac{m_{ij}V_{jk}}{1-V_{jj}}| = o(n^{\alpha_l - \alpha_{\mathscr{N}}}) \tag{3.38}$$

Remark 9. *Notably, if $p \leq n$ and X has full rank, then $V_{jk} = 0$ and assumption 7 is automatically satisfied. To explain assumption 6, adopt notations as in assumption 5; we then have*

$$\sum_{l=1}^{n} w_{il}^2 = \sum_{k=1}^{r} c_{ij}^2 \left(\frac{\lambda_k}{\lambda_k^2 + \rho_{n,r}} + \frac{\rho_{n,k}\lambda_k}{(\lambda_k^2 + \rho_{n,r})^2} \right)^2 < \tau_i^2 \tag{3.39}$$

Intuitively, eq. (3.36) *requires all terms in the summation $\sum_{l=1}^{n} \frac{w_{il}}{\tau_i}\varepsilon_l$ to be negligible.*

Theorem 6. *Suppose assumptions 1 to 7 hold true. Define*

$$\widehat{c}_{ij} = \sum_{k\in\widehat{\mathscr{N}_{b_n}}} \frac{m_{ik}q_{kj}}{1-V_{kk}} \text{ and } \widehat{\tau}_i = \sqrt{\sum_{k=1}^{r} \widehat{c}_{ij}^2 \left(\frac{\lambda_k}{\lambda_k^2 + \rho_{n,r}} + \frac{\rho_{n,r}\lambda_k}{(\lambda_k^2 + \rho_{n,r})^2} \right)^2 + \frac{1}{n}} \tag{3.40}$$

Then

$$\sup_{x\in\mathbf{R}} |Prob\left(\max_{i=1,\ldots,p_1} \frac{|\widehat{\zeta}_i - \zeta_i|}{\widehat{\tau}_i} \leq x \right) - Prob\left(\max_{i\in\mathscr{M}} |\xi_i| \leq x \right)| = o(1); \tag{3.41}$$

here $\xi_i, i \in \mathscr{M}$ are joint normal random variables such that $\mathbf{E}\xi_i = 0$ and

$$\mathbf{E}\xi_{i_1}\xi_{i_2} = \frac{1}{\tau_{i_1}\tau_{i_2}} \sum_{l_1=1}^{n}\sum_{l_2=1}^{n} \sigma_{l_1 l_2} \times w_{i_1 l_1} w_{i_2 l_2},\ i_1, i_2 \in \mathscr{M}. \tag{3.42}$$

The added term $1/n$ in $\widehat{\tau}_i$ is introduced just to ensure $\widehat{\tau}_i > 0$. If we can calculate $C_{1-\alpha}$, the $1-\alpha$ quantile of $\max_{i\in\mathscr{M}} |\xi_i|$, then the set

$$\left\{ x = (x_1, ..., x_{p_1})^T \in \mathbf{R} : \max_{i=1,...,p_1} \frac{|x_i - \widehat{\zeta}_i|}{\widehat{\tau}_i} \leq C_{1-\alpha} \right\} \tag{3.43}$$

will be a consistent confidence region for $\zeta = M\beta$. Despite the fact that the distribution of $\max_{i\in\mathscr{M}} |\xi_i|$ does not have a closed-form formula, the $\xi_i, i \in \mathscr{M}$ are joint normal random variables. Therefore, we can use a computer to generate pseudo-random numbers, simulate ξ_i many times, and calculate $C_{1-\alpha}$ through Monte Carlo simulation. The remaining problem is that the covariance matrix of ξ_i is unknown.

Since $\varepsilon_i, i = 1, 2, ..., n$ do not have identical distribution, estimating a specific $\sigma_{ij} = \mathbf{E}\varepsilon_i\varepsilon_j$, i.e., for some given values of i, j, is hopeless. Fortunately, lemma 4 tells us that estimating the covariances of the linear combinations $\sum_{l=1}^{n} \frac{w_{il}\varepsilon_l}{\tau_i}$ is still possible; this idea is what drives the well-known heteroscedastic standard errors of White [1980].

In the next section, we adopt the idea of lemma 4 and present a consistent estimator for the covariance matrix of $\xi_i, i \in \mathscr{M}$. Moreover, we will provide a bootstrap algorithm that automatically generates the desired confidence intervals without analytical calculations.

3.5 Bootstrap confidence intervals

This section focuses on constructing a simultaneous bootstrap confidence interval for the entries of $\zeta = M\beta$. Before presenting our work, we introduce an additional assumption.

Additional assumptions II:

8. Suppose $K(\cdot) : \mathbf{R} \to [0, \infty)$ is a given kernel function, i.e., $K(\cdot)$ satisfies definition 3,

and $k_n > 0$ is a given bandwidth sequence satisfying

$$\begin{gathered} k_n \to \infty,\ \frac{k_n}{n^{1/4}} = O(1),\ n^{\alpha_b + \frac{\alpha_{p+1}}{m} - \eta}\sqrt{k_n} \to 0 \\ k_n \times n^{\frac{2}{m} + \frac{\alpha_{\mathscr{N}}}{2} - \eta} \to 0,\ k_n \times n^{\frac{1}{m} + \alpha_{\mathscr{N}} - \alpha_l} \to 0 \text{ as } n \to \infty \end{gathered} \tag{3.44}$$

where α_b is defined in assumption 5 and $\alpha_{\mathscr{N}}, \alpha_l$ are defined in assumption 4.

Eq. (3.44) omits the $\log^z(n)$ in eq. (3.15) for convenience. However, $\log^z(n)$ is negligible compared to n^a(here $a > 0$), so assumption 8 is not very restrictive. We define the conditional probability and expectation

$$Prob^*(\cdot) = Prob(\cdot|y) \text{ and } \mathbf{E}^*\cdot = \mathbf{E}(\cdot|y). \tag{3.45}$$

In the bootstrap literature, these are recognized as 'the probability and the expectation in the bootstrap world'; see e.g., Cheng and Huang [2010]. Theorem 7 provides a consistent estimator for the covariance matrix $\{\mathbf{E}\xi_i\xi_j\}_{i,j\in\mathscr{M}}$.

Theorem 7. *Suppose assumptions 1 to 8 hold true. Define* $\widehat{\varepsilon} = (\widehat{\varepsilon}_1, ..., \widehat{\varepsilon}_n)^T$ *such that* $\widehat{\varepsilon}_i = y_i - \sum_{j=1}^p X_{ij}\widehat{\beta}_j$. *Define the matrix* $\widehat{\Gamma} = \{\widehat{\Gamma}_{ij}\}_{i,j=1,...,p_1}$ *by*

$$\widehat{\Gamma}_{ij} = \frac{1}{\widehat{\tau}_i\widehat{\tau}_j}\sum_{l_1=1}^n\sum_{l_2=1}^n K\left(\frac{l_1 - l_2}{k_n}\right)\widehat{\varepsilon}_{l_1}\widehat{\varepsilon}_{l_2} \times \widehat{w}_{il_1}\widehat{w}_{jl_2} \tag{3.46}$$

where

$$\widehat{w}_{il} = \sum_{j=1}^r \widehat{c}_{ij}p_{lj}\left(\frac{\lambda_j}{\lambda_j^2 + \rho_{n,r}} + \frac{\rho_{n,r}\lambda_j}{(\lambda_j^2 + \rho_{n,r})^2}\right) \tag{3.47}$$

and $\widehat{c}_{ij}$, $\widehat{\tau}_i$ *are defined in* (3.40). *Then we have*

$$\max_{i,j=1,...,p_1} |\widehat{\Gamma}_{ij} - \frac{1}{\tau_i\tau_j}\sum_{l_1=1}^n\sum_{l_2=1}^n \sigma_{l_1l_2} \times w_{il_1}w_{jl_2}| = o_p(1). \tag{3.48}$$

Theorem 7 adopts the idea of lemma 4, i.e., estimates the covariances of the linear combinations $\sum_{l=1}^{n} \frac{w_{il}\varepsilon_l}{\tau_i}$ with a kernel estimator. However, it is more complex than lemma 4 since w_{il} needs to be estimated, and the estimated errors $\widehat{\varepsilon}_i$ does not equal the real errors ε_i.

Although we have the consistent estimator $\widehat{\Gamma}$ for $\{\mathbf{E}\xi_i\xi_j\}_{i,j\in\mathscr{M}}$, computing $\widehat{\Gamma}$ is still a tedious undertaking in practice. Therefore, we hope to find an algorithm that is easy to implement and can automatically generate the desired simultaneous confidence intervals and/or perform hypothesis testing for ζ. Some form of resampling may be the way out; see e.g., Politis et al. [1999] and Zhang and Zhang [2014]. For example, the dependent wild bootstrap algorithm introduced by Shao [2010] has wide applicability in dependent data settings. Conley et al. [2019] applied the dependent wild bootstrap on linear regression while Zhang and Politis [2021b] used this algorithm for statistical inference on autoregressive models. We will focus on testing

$$\begin{gathered}\text{null: } M\beta = \psi \text{ versus alternative: } M\beta \neq \psi \\ \text{where } M \text{ is a given } p_1 \times p \text{ matrix and } \psi \text{ is a given vector.}\end{gathered} \tag{3.49}$$

Algorithm 3 (Dependent wild bootstrap)**.** ***Input:*** *Design matrix X, dependent variable vector y, the new linear combination matrix M, the ridge regression parameter $\rho_{n,r}$, the Lasso parameter $\rho_{n,l}$, threshold $b_n > 0$, the nominal coverage probability $1-\alpha, 0<\alpha<1$, the kernel function $K(\cdot)$, the bandwidth $k_n > 0$ and the number of bootstrap replicates B.*

Additional input for testing: *$\psi \in \mathbf{R}^{p_1}$, the expected value of $\zeta = M\beta$ under the null.*

Algorithm steps:

1. Calculate $\widehat{\beta}$ in (3.22)*, $\widehat{\zeta}$ in* (3.34) *and $\widehat{\tau}_i$ in* (3.40)*. Then derive $\widehat{\varepsilon} = (\widehat{\varepsilon}_1, ..., \widehat{\varepsilon}_n)^T$ such that $\widehat{\varepsilon}_i = y_i - \sum_{j=1}^{p} X_{ij}\widehat{\beta}_j$. Define $V = \{V_{ij}\}_{i,j=1,...,p} = Q_\perp Q_\perp^T$ if $p > n$ and 0 if $p \leq n$. Calculate $\widehat{\beta}_\perp = V\widehat{\beta}$.*

2. Generate jointly normal random variables $\varepsilon_1, ..., \varepsilon_n$ such that $\mathbf{E}\varepsilon_i = 0$ and $\mathbf{E}\varepsilon_i\varepsilon_j =$

$K\left(\frac{i-j}{k_n}\right)$. *Define* $\varepsilon^* = (\varepsilon_1^*, ..., \varepsilon_n^*)^T$ *such that* $\varepsilon_i^* = \widehat{\varepsilon}_i \times \varepsilon_i$. *Set* $y^* = X\widehat{\beta} + \varepsilon^*$. *Then calculate*

$$\begin{aligned}
\widetilde{\beta}^{\dagger *} &= (X^T X + \rho_{n,r} I_p)^{-1} X^T y^* \\
\widetilde{\beta}^* = (\widetilde{\beta}_1^*, ..., \widetilde{\beta}_p^*)^T &= \widetilde{\beta}^{\dagger *} + \rho_{n,r} Q(\Lambda^2 + \rho_{n,r} I_r)^{-1} Q^T \widetilde{\beta}^{\dagger *} + \widehat{\beta}_\perp \\
\widehat{\mathscr{N}_{b_n}^*} &= \{i = 1, 2, ..., p : |\widetilde{\beta}_i^*| > b_n\} \\
\widehat{\beta}^* = (\widehat{\beta}_1^*, ..., \widehat{\beta}_p^*)^T &\text{ such that } \widehat{\beta}_i^* = \widetilde{\beta}_i^* \times \mathbf{1}_{i \in \widehat{\mathscr{N}_{b_n}^*}} \\
\widehat{\zeta}^* = (\widehat{\zeta}_1^*, ..., \widehat{\zeta}_{p_1}^*)^T &\text{ such that } \widehat{\zeta}_i^* = \sum_{j=1}^{p} \frac{m_{ij}}{1 - V_{jj}} \times \widehat{\beta}_j^* - \sum_{j \in \widehat{\mathscr{N}_{b_n}^*}} \frac{m_{ij} V_{jj}}{1 - V_{jj}} \widehat{\beta}_j
\end{aligned} \tag{3.50}$$

3. Define

$$\widehat{c}_{ij}^* = \sum_{k \in \widehat{\mathscr{N}_{b_n}^*}} \frac{m_{ik} q_{kj}}{1 - V_{kk}} \text{ and } \widehat{\tau}_i^* = \sqrt{\sum_{k=1}^{r} \widehat{c}_{ij}^{*2} \left(\frac{\lambda_k}{\lambda_k^2 + \rho_{n,r}} + \frac{\rho_{n,k} \lambda_k}{(\lambda_k^2 + \rho_{n,r})^2} \right)^2 + \frac{1}{n}} \tag{3.51}$$

then calculate

$$\delta_b^* = \max_{i=1,...,p_1} \frac{|\widehat{\zeta}_i^* - \sum_{j=1}^{p} m_{ij} \widehat{\beta}_j|}{\widehat{\tau}_i^*} \tag{3.52}$$

4. Repeat step 2 to 3 for $b = 1, 2, ..., B$. *Compute* $\delta_1^*, ...\delta_B^*$ *from (3.52), and let* $C_{1-\alpha}^*$ *denote the empirical* $1 - \alpha$ *quantile of the values* $\delta_1^*, ...\delta_B^*$.

5.a (constructing confidence interval) The $1 - \alpha$ *simultaneous confidence intervals for* $\zeta = (\zeta_1, ..., \zeta_{p_1})^T$ *are given by*

$$\left\{ x = (x_1, ..., x_{p_1})^T \in \mathbf{R}^{p_1} : \max_{i=1,2,...,p_1} \frac{|x_i - \widehat{\zeta}_i|}{\widehat{\tau}_i} \leq C_{1-\alpha}^* \right\} \tag{3.53}$$

5.b (hypothesis testing) Reject the null hypothesis at level α *if*

$$\max_{i=1,2,...,p_1} \frac{|\psi_i - \widehat{\zeta}_i|}{\widehat{\tau}_i} > C_{1-\alpha}^*. \tag{3.54}$$

Remark 10. *Although in practice the bootstrap is carried out with B being a large —but finite— number, the theoretical analysis assumes $B \to \infty$. If $B \to \infty$, theorem 1.2.1 in Politis et al. [1999] shows that $C^*_{1-\alpha}$ converges to $c^*_{1-\alpha}$, the $1-\alpha$ quantile of the conditional distribution $Prob^*\left(\max_{i=1,\dots,p_1} \frac{|\widehat{\zeta}^*_i - \sum_{j=1}^p m_{ij}\widehat{\beta}_j|}{\widehat{\tau}^*_i} \leq x\right)$, i.e.,*

$$c^*_{1-\alpha} = \inf\left\{ x \in \mathbf{R} : Prob^*\left(\max_{i=1,\dots,p_1} \frac{|\widehat{\zeta}^*_i - \sum_{j=1}^p m_{ij}\widehat{\beta}_j|}{\widehat{\tau}^*_i} \leq x \right) \geq 1-\alpha \right\} \tag{3.55}$$

Therefore, in order to prove the consistency of algorithm 3, it suffices to show

$$\sup_{x\in\mathbf{R}} |Prob^*\left(\max_{i=1,\dots,p_1} \frac{|\widehat{\zeta}^*_i - \sum_{j=1}^p m_{ij}\widehat{\beta}_j|}{\widehat{\tau}^*_i} \leq x \right) - Prob\left(\max_{i\in\mathscr{M}} |\xi_i| \leq x \right)| = o_p(1); \tag{3.56}$$

see theorem 6 for the meaning of ξ_i.

Theorem 8 proves the consistency of algorithm 3.

Theorem 8. *Suppose assumptions 1 to 8 hold true. Then eq. (3.56) holds, where $\mathscr{M}$ and ξ_i are defined in theorem 6. Consequently, as both n and B tend to ∞, the confidence intervals and hypothesis test of algorithm 3 have asymptotically correct coverage and significance level respectively.*

3.6 Numerical experiment

This section presents several numerical experiments to illustrate the finite sample performance of the debiased and thresholded ridge regression estimator as well as the wild bootstrap algorithm 3. In addition, we apply the estimator and the algorithm to a real-life data set.

3.6.1 Selection of hyper-parameters

In order to use the debiased and thresholded ridge regression method (3.22) and the bootstrap algorithm 3, statisticians need to fine-tune the ridge regression parameter $\rho_{n,r}$, the

Lasso parameter $\rho_{n,l}$, the threshold b_n and the bandwidth k_n. E.g., $\rho_{n,r}, \rho_{n,l}$ and b_n can be chosen by cross-validation, i.e., separate the design matrix X and the dependent variables y into disjoint training set (X_{train}, y_{train}) and validation set (X_{valid}, y_{valid}); for each choice of parameters, use (X_{train}, y_{train}) to fit $\widehat{\beta}$; then calculate $\left| y_{valid} - X_{valid}\widehat{\beta} \right|_2$. The optimal parameters should minimize $\left| y_{valid} - X_{valid}\widehat{\beta} \right|_2$. See Arlot and Celisse [2010] for a further introduction on the cross validation methods.

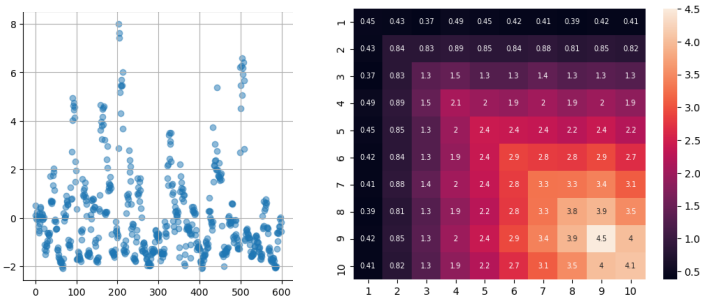
Grid search on the parameter tuple $(\rho_{n,r}, \rho_{n,l}, b_n)$ is time-consuming, so we adopt a two-stage search, i.e., first fit the Lasso regression (3.20) and choose $\rho^*_{n,l}$ that minimizes $|y_{valid} - X_{valid}\widetilde{\beta}^{lasso}|_2$. Fixing $\rho_{n,l}$ at the value $\rho^*_{n,l}$, perform grid search on the parameter tuple $(\rho_{n,r}, b_n)$ to find $\rho^*_{n,r}$ and b^*_n that minimize $|y_{valid} - X_{valid}\widehat{\beta}|_2$. According to the simulations, this method selects suitable parameters.

Fine-tuning k_n is more challenging. Politis and White [2004] introduced an automatic bandwidth selection algorithm; Shao [2010] applied this algorithm for the dependent wild bootstrap. 10. and Kim and Sun [2011] considered selecting bandwidth in the HAC estimation setting. Following Shao [2010], this paper applies Politis and White's algorithm Politis and White [2004] on the fitted residuals $\widehat{\varepsilon} = (\widehat{\varepsilon}_1, ..., \widehat{\varepsilon}_n)^T = y - X\widehat{\beta}$ to select k_n; to do this, we use R-Package 'np' Hayfield and Racine [2008] that implements the selection algorithm. However, ε is not assumed to be stationary in our setting; so this algorithm may result in a suboptimal bandwidth.

3.6.2 Simulated Data

The numerical experiment fits the linear model $y = X\beta + \varepsilon$. Here, X is generated by i.i.d. standard normal random variables, and is fixed in each experiment. Parameter $\beta = (\beta_1, ..., \beta_p)^T$ is generated by the following scheme

$$\beta_i = 0.1 \times (i+5) \text{ for } i = 1, 2, ..., 15 \text{ and 0 otherwise.} \tag{3.57}$$



(a) **(b)**

Figure 3.1. Figure 3.1a plots an observation of the errors ε and figure 3.1b plots the heatmap for the first 10×10 elements of ε's covariance matrix. Values in each grid represent the corresponding covariance. The covariance matrix is calculated by simulating 40000 samples of the random vector $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T$.

Define $e_i, i \in \mathbf{Z}$ as i.i.d. standard normal random variables. Choose $a = (a_1, a_2, ..., a_n)^T$ such that $a_i = e_i^2 e_{i-1}^2 - 1$. Then define $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T = (1/4) \times Ha$(the factor $1/4$ avoids ε's variances from being too large). Here $H = (h_{ij})_{i,j=1,...,n}$, $h_{ij} = 0$ for $j > i$, $h_{ii} = 1$, h_{ij} is generated by uniform distribution in $[0.6, 0.9]$ for $i - 10 \leq j < i$ and $h_{ij} = s_{ij}/(i-j)^3$ for $j < i - 10$. s_{ij} is generated by uniform distribution on $[-1, 1]$. H is fixed in each experiment. For $\mathbf{E}a_i = \mathbf{E}e_i^2 \times \mathbf{E}e_{i-1}^2 - 1 = 0$, we have $\mathbf{E}\varepsilon = 0$.

Note that the a_is are not white noise since $\mathbf{E}a_i a_{i-1} = 2$. Moreover, the ε_i are not a stretch of a linear process with independent increments because of the nonlinearity of the a_is, and are not stationary because of H. Figure 3.1 plots an observation of the errors ε, and the first 10×10 elements of ε's covariance matrix. In figure 3.1a, the errors demonstrate strong dependence, i.e., ε_{i+1} is likely to be large if ε_i is large. Figure 3.1b shows that ε's covariance matrix is not a Toeplitz matrix, so the distribution of ε is not stationary.

The linear combination matrix M is generated by i.i.d. normal random variables with mean 0.5 and variance 0.25, and is fixed in each experiment. The hyper-parameters $\rho_{n,r}, \rho_{n,l}$, b_n, k_n are tuned by the methods described in section 3.6.1. The sample size n and the dimension p vary in each experiment. We present the information on our experiments in table 3.1.

Simulation result The performance of the debiased and thresholded ridge regression

Table 3.1. Experiment parameters. 'No.' abbreviates 'the experiment number'. Denote $\rho_{n,r}, \rho_{n,l}, b_n, k_n$ the selected ridge parameter, the selected Lasso parameter, the selected threshold and the selected bandwidth respectively as defined in section 3.4. Denote λ_r the smallest singular value of the design matrix X. The number of linear combinations p_1 is 15.

No.	sample size	dimension	$\rho_{n,r}$	$\rho_{n,l}$	b_n	k_n	λ_r
1	300	400	26.908	0.104	0.353	13.720	2.694
2	600	800	18.288	0.118	0.422	16.994	3.614
3	1200	1600	19.859	0.184	0.470	28.704	5.208
4	300	200	5.305	0.202	0.288	14.943	3.299
5	600	400	11.544	0.095	0.408	12.182	4.667
6	1200	800	69.568	0.109	0.360	21.602	6.317

estimator (thsDeb, defined in (3.22)) and the bootstrap method is presented in figure 3.2, 3.3 and table 3.2 .The alternative methods are Lasso, Ridge regression, the ElasticNet, the threshold Lasso (thsLas) and the threshold ridge regression (thsRid); see Tibshirani [2011] and Zou and Hastie [2005]. We refer to Chatterjee and Lahiri [2011] and Meinshausen and Yu [2009] for the threshold Lasso, and Shao and Deng [2012] for the threshold ridge regression.

Apart from $\widehat{\zeta}_i$ (defined in (3.34)), we are also interested in $\widetilde{\zeta}_i^{\dagger} = \sum_{j=1}^{p} m_{ij}\widehat{\beta}_j$, named 'thsRaw' in this section. Compared to $\widehat{\zeta}_i$, $\widetilde{\zeta}_i^{\dagger}$ is a natural predictor for the linear combination $\zeta_i = \sum_{j=1}^{p} m_{ij}\beta_j$. However, figure 3.2 shows that $\widetilde{\zeta}_i^{\dagger}$ has a larger error than $\widehat{\zeta}_i$.

This section uses the following indices to evaluate the performance of linear regression methods: the probability of model misspecification $Prob(\widehat{\mathscr{N}_{b_n}} \neq \mathscr{N}_{b_n})$; the average size of model misspecification $|\widehat{\mathscr{N}_{b_n}} \Delta \mathscr{N}_{b_n}|$($\Delta$ represents the symmetric difference, i.e., $A\Delta B = (A-B)\cup(B-A)$); the average false discovery rate $\frac{|\widehat{\mathscr{N}_{b_n}} - \mathscr{N}_{b_n}|}{\max(1, |\widehat{\mathscr{N}_{b_n}}|)}$ and the average prediction loss $|X\widehat{\beta} - X\beta|_2$; see Shao and Deng [2012]. These terms are calculated over 1000 simulations.

The numerical experiments focus on two situations, i.e, $p > n$ (experiment 1 - 3) and $p < n$ (experiment 4 - 6). When $p > n$, the threshold Lasso has good performance. However, statistical inference (i.e., constructing confidence intervals or performing hypothesis testing) based on the threshold Lasso estimator is quite difficult. Chatterjee and Lahiri [2011] proposed a bootstrap algorithm to generate a consistent confidence interval when the dimension of the linear model is fixed. However, to the best of our knowledge, there has not been a discussion of this problem under the high dimensional setting. On the other hand, statisticians can generate

consistent confidence intervals for the ridge regression (e.g., Bühlmann [2013]) and the Lasso estimator(e.g., Zhang and Zhang [2014], Zhang and Cheng [2017]). However, the performance of these methods is significantly worse compared to the threshold Lasso and the debiased and thresholded ridge regression that is the subject of this paper.

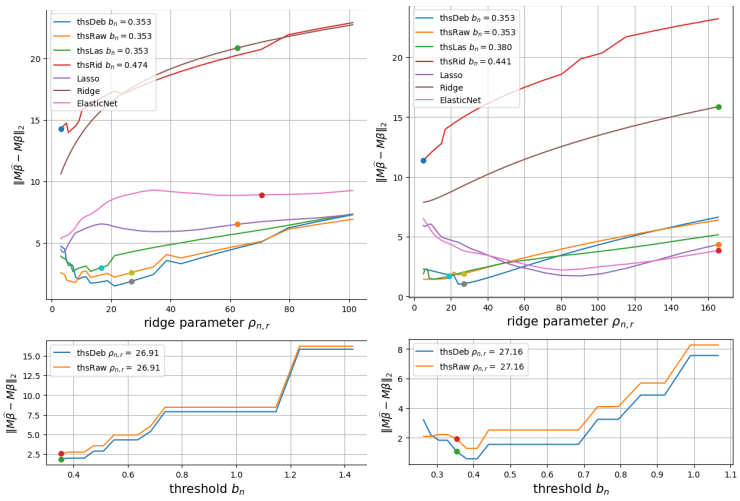
The debiased and thresholded ridge regression estimator has moderate probability of model misspecification. Even if the estimator selects a wrong model, table 3.2 shows that the deviation between $\widehat{\mathcal{N}_{b_n}}$ and $\mathcal{N}_{b_n}$ is small on average. Moreover, in figure 3.2 we see that the proposed method has a smaller estimation error than the threshold Lasso, which makes this method competitive. Besides, the modification (3.34) can further decrease the estimation error. Similar things happen when $p < n$, i.e., the threshold Lasso and the debiased and threshold ridge regression have small estimation errors. Notably, when $p < n$, the modified estimator (3.34) equals $\widetilde{\zeta}^{\dagger} = M\widehat{\beta}$. This observation coincides with section 3.4.

We also test algorithm 3 and present the results in table 3.2 . Simulations show that the bootstrap algorithm 3 can generate a simultaneous confidence interval with desired coverage probability.

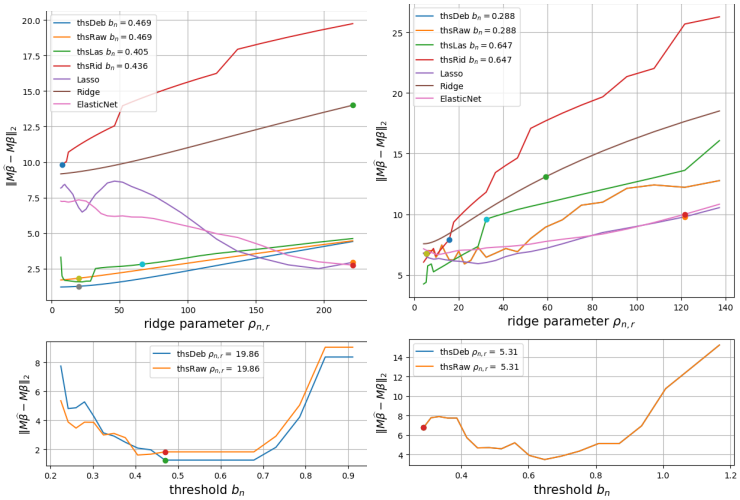
3.6.3 Real-life data

We analyze the 'Market data' that can be downloaded from Fan [Accessed: 2022]; see section 8.8 in Fan et al. [2020] and Chen et al. [2018] for a description. The response y represents the daily number of customers, and each column of the design matrix X records the sale amount of a product. The dataset has 464 observations and 6398 features (products). The response y and the design matrix X are standardized so that y and each column of X have (sample) mean 0 and variance 1.

For computational reasons, we explore the relation between y and just the first 2000 features. The first 450 observations will be used to train the model and the last 14 observations will serve as the test set. We select hyperparameters as described in Section 3.6.1, and present the selected values in table 3.3 . We evaluate the estimator's performance through the prediction



(a) Experiment 1 **(b)** Experiment 2

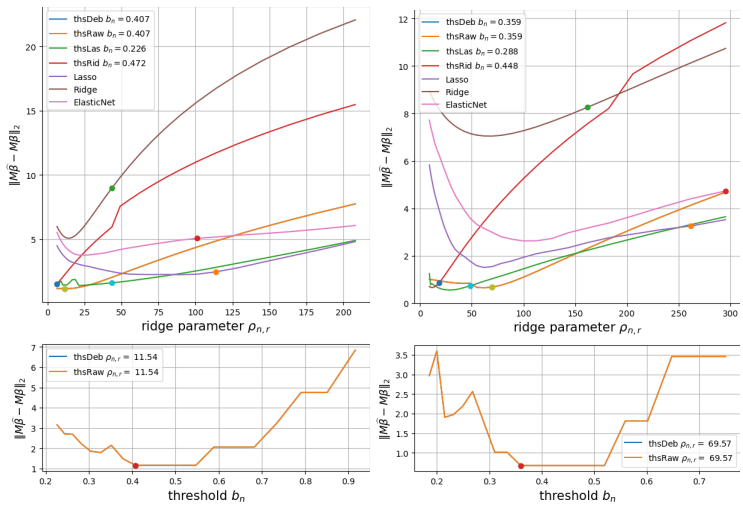


(c) Experiment 3 **(d)** Experiment 4

Figure 3.2. The estimation errors of various linear regression algorithms. The upper figure plots the estimation error $|M\widehat{\beta} - M\beta|_2$ with respect to different ridge(Lasso) parameter $\rho_{n,r}$(for the Lasso (3.20) and the threshold Lasso, we recognize the x-axis as $2n \times \rho_{n,l}$, n is the sample size. Otherwise, $\rho_{n,l}$ will be too small to be plotted on the figure.). The lower figure plots the estimation error of the proposed linear regression method (3.22) with respect to different threshold b_n. The dots represent the optimal parameters selected by ten-fold cross validation.

Table 3.2. Performance of various linear regression algorithms. The terms are calculated through 1000 simulations.'Prob' represents $Prob(\widehat{\mathcal{N}_{b_n}} \neq \mathcal{N}_{b_n})$, 'FDR' abbreviates 'the false discovery rate' and 'coverage' abbreviates 'the coverage probability of the confidence interval (3.53)'. 'length' represents the average length of the confidence interval (3.53), which equals $2C^*_{1-\alpha}$. We select the nominal coverage probability $1-\alpha = 90\%$.

No	Algorithm	*Prob*	$\lvert\widehat{\mathcal{N}_{b_n}}\Delta\mathcal{N}_{b_n}\rvert$	FDR	Prd-loss	coverage	length
1	thsDeb	71.4%	6.80	0.21	23.70	91.2%	9.678
	Lasso	100%	39.16	0.70	19.39		
	thsLas	54.6%	4.22	0.13	19.17		
	Ridge	100%	174.86	0.92	31.10		
	thsRid	92.1%	6.67	0.20	38.46		
	ElasticNet	100%	69.70	0.82	23.96		
2	thsDeb	53%	3.20	0.11	22.98	92.8%	20.843
	Lasso	100%	27.95	0.59	22.28		
	thsLas	36.9%	1.35	0.04	17.70		
	Ridge	100%	354.15	0.96	44.77		
	thsRid	98.2%	3.69	0.08	44.64		
	ElasticNet	100%	93.39	0.85	28.76		
3	thsDeb	21.7%	0.29	0.0037	16.38	85.7%	8.850
	Lasso	100%	56.71	0.75	23.16		
	thsLas	5.2%	0.08	0.0010	13.16		
	Ridge	100%	690.44	0.98	63.83		
	thsRid	78.5%	1.04	0.0100	56.43		
	ElasticNet	100%	179.17	0.92	38.20		
4	thsDeb	98.0%	13.51	0.40	27.91	90.6%	7.725
	Lasso	95.8%	7.57	0.29	17.67		
	thsLas	81.7%	1.32	0.009	19.16		
	Ridge	100%	89.33	0.86	27.43		
	thsRid	81.8%	1.49	0.02	21.93		
	ElasticNet	100%	27.36	0.62	20.61		
5	thsDeb	53.2%	2.54	0.10	23.08	96.9%	9.399
	Lasso	100%	32.71	0.65	18.65		
	thsLas	47.4%	2.40	0.10	15.58		
	Ridge	100%	170.31	0.92	38.39		
	thsRid	42.0%	1.10	0.04	18.76		
	ElasticNet	100%	81.71	0.84	26.74		
6	thsDeb	20.2%	0.46	0.024	16.62	96.9%	9.486
	Lasso	99.9%	21.03	0.52	19.50		
	thsLas	7.5%	0.15	0.008	11.51		
	Ridge	100%	356.76	0.96	54.89		
	thsRid	11.6%	0.15	0.003	15.72		
	ElasticNet	100%	79.58	0.83	29.18		



(a) Experiment 5 **(b)** Experiment 6

Figure 3.3. The estimation errors of various linear regression algorithms for experiment 5 and 6. The meaning of figures coincides with that of figure 3.2.

error (i.e., the test set 2-norm $|y_{test} - X_{test}\widehat{\beta}|_2$, where the subscript 'test' represents the test set) and the number of non-zero elements $|\widehat{\mathscr{N}_{b_n}}|$(see theorem 5). Ideally, a good estimator should have small prediction error and small $|\widehat{\mathscr{N}_{b_n}}|$. Figure 3.4 plots the predicted responses on the test set and the estimated autocorrelation function (ACF) of the fitted errors (i.e., $y - X\widehat{\beta}$). Since the errors have non-trivial ACF, they should be considered dependent.

According to table 3.3 , the debiased and thresholded ridge regression estimator has the smallest prediction error. Meanwhile, it maintains the model sparsity, i.e., it has small $|\widehat{\mathscr{N}_{b_n}}|$. On the other hand, despite the Lasso and the elasticnet having small prediction errors as well, they tend to select a more complex model than the debiased and threshold ridge regression estimator.

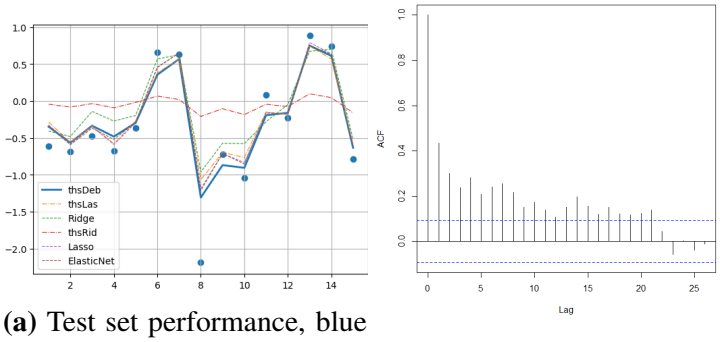
Furthermore, as an illustration, we construct a simultaneous confidence interval for the first 8 parameters, i.e., choose $M = [I_8, 0]$. Here I_8 is the 8×8 identity matrix and 0 represents the $8 \times (2000 - 8)$ matrix with elements 0. The result is demonstrated in table 3.4 . In algorithm 3, we estimate the parameters through (3.34), i.e., $\frac{1}{1-V_{ii}}\widehat{\beta}_i - \frac{V_{ii}}{1-V_{ii}}\widetilde{\beta}_i^{lasso} \times \mathbf{1}_{i \in \widehat{\mathscr{N}_{b_n}}}$ instead of $\widehat{\beta}_i$. So $\widehat{\beta}$ does not lie in the center of the confidence interval in table 3.4.

Table 3.3. Selected hyper-parameters(HP) of thsDeb and the test set performance of various linear regression methods. The meaning of symbols and abbreviations coincide with table 3.1 and 3.2. Lasso, ridge regression and elastic net do not have well-defined $\widehat{\mathcal{N}_{b_n}}$. For those methods, we consider $i \in \widehat{\mathcal{N}_{b_n}}$ if $|\widehat{\beta}_i| > 0.0001$. 'samp' represents the number of samples; while 'dim' represents the dimension of parameters.

The selected hyper-parameters(HP) of thsDeb							
	samp	dim	$\rho_{n,r}$	$\rho_{n,l}$	b_n	k_n	λ_r
HP of thsDeb	450	2000	39.934	0.044	0.013	28.488	15.497
Performance of various linear regression methods							
	thsDeb	thsLas	thsRid	Lasso	Ridge	ElasticNet	
$\lvert y_{test} - X_{test}\widehat{\beta}\rvert_2$	1.09	1.31	2.94	1.15	1.56	1.13	
$\lvert\widehat{\mathcal{N}_{b_n}}\rvert$	46	36	15	88	1938	141	

Table 3.4. The 95% simultaneous confidence interval for thsDeb's first 8 parameters. The 95% simultaneous confidence interval will be given by $\{x = (x_1, ..., x_8) : \text{Conf.low} \leq x_i \leq \text{Conf.high}\}$

$i =$	1	2	3	4	5	6	7	8
$\widehat{\beta}_i$	0.0	0.0	0.019	0.0	0.0	0.072	0.098	0.037
Conf.low	-0.046	-0.046	-0.057	-0.046	-0.046	-0.059	0.025	-0.064
Conf.high	0.046	0.046	0.122	0.046	0.046	0.213	0.228	0.146



(a) Test set performance, blue dots represent real values

(b) ACF of Residuals

Figure 3.4. The predicted responses on the test set for various linear regression methods and the ACF of residuals $(y - X\widehat{\beta})$.

3.7 Conclusion

Focusing on a high dimensional, sparse linear regression model $y = X\beta + \varepsilon$ with heteroscedastic, dependent (correlated) and non-stationary errors ε, the paper at hand proposes a debiased and thresholded ridge regression estimator that is consistent and able to recover the model sparsity. We also develop a Gaussian approximation theorem for the estimator. Moreover, we construct a dependent wild bootstrap algorithm that automatically yields consistent simultaneous confidence intervals and/or hypothesis tests for $\zeta = M\beta$, where M is a given linear combination matrix. Numerical simulations and real-life data analysis show that the proposed estimator has good finite sample performance, complementing our theoretical (asymptotic) results.

To the best of our knowledge, there is little research on the high dimensional linear model with non-stationary errors. Compared to the state-of-the-art of the literature, the paper at hand:

- introduces a set of theoretical tools to analyze non-stationary random variables as potential regression errors;
- enables consistent statistical inference even when the errors in the linear model are heteroscedastic, with an arbitrarily complex covariance matrix;
- allows practitioners to construct simultaneous confidence intervals for linear combinations of β;
- and allows for consistent debiased and thresholded ridge regression with parameter dimension larger than sample size — a result that is novel even in the context of i.i.d. errors.

3.8 Acknowledgement

Chapter 3 is based on the paper “Debiased and thresholded ridge regression for linear models with heteroskedastic and correlated errors” by Y.Zhang and D.N.Politis and has been

submitted for publication. The dissertation author was the primary investigator and author of this paper.

Chapter 4

Bootstrap prediction intervals with asymptotic conditional validity and unconditional guarantees

4.1 Abstract

It can be argued that optimal prediction should take into account all available data. Therefore, to evaluate a prediction interval's performance one should employ conditional coverage probability, conditioning on all available observations. Focusing on a linear model, we derive the asymptotic distribution of the difference between the conditional coverage probability of a nominal prediction interval and the conditional coverage probability of a prediction interval obtained via a residual-based bootstrap. Applying this result, we show that a prediction interval generated by the residual-based bootstrap has approximately 50% probability to yield conditional under-coverage. We then develop a new bootstrap algorithm that generates a prediction interval that asymptotically controls both the conditional coverage probability as well as the possibility of conditional under-coverage. We complement the asymptotic results with several finite-sample simulations.

Keywords: Prediction, regression, bootstrap, conditional validity

4.2 Introduction

Statistical inference comes in two flavors: explaining the world and predicting the future state of the world. To explain the world based on data, statisticians create models like linear regression and use data to fit the models. After doing that, they will gauge the goodness-of-fit, and assess the accuracy of estimation, e.g., via confidence intervals of the fitted model. Focusing on regression, the literature is huge; to pick 3-4 papers, see Shao [1996] on model selection, Xie and Huang [2009] or Liu and Yu [2013] on model fitting, and Freedman [1981] on statistical analysis.

Prediction is not a new topic in statistical inference; we refer to Geisser [1993] for a comprehensive introduction, or Politis [2015] for a more recent exposition. Notably, prediction has seen a resurgence in the 21st century with the advent of statistical learning; see Hastie et al. [2009] for an introduction. Similarly to the aforementioned linear model procedure, statisticians use data to fit a model that can yield a predictor for future observations, and use prediction intervals to quantify uncertainty in the prediction; see e.g. Romano et al. and Wang and Politis [2021]. Under a regression setting, there are several ways to construct a prediction interval. The classical prediction interval was typically obtained under a Gaussian assumption on the errors; see Section 4.3 in that follows. One of the earliest methods foregoing the restrictive normality assumption employed the residual-based bootstrap; see Stine [1985] and the references therein. More recent methods include the *Model-free* (MF) bootstrap and the hybrid *Model-free/Model-based* (MF/MB) bootstrap of Politis [2015].

For all bootstrap methods, the aim is to provide an asymptotically valid prediction interval. Suppose Γ is a prediction interval for the future observation $\mathscr{Y}_f$. If $Prob(\mathscr{Y}_f \in \Gamma) \approx 1-\alpha$ (where $\approx$ indicates an asymptotic approximation), then Γ is an asymptotically valid $1-\alpha$ prediction interval for $\mathscr{Y}_f$. On the other hand, if we wish to ensure that $Prob(\mathscr{Y}_f \in \Gamma) \geq 1-\alpha$, i.e., an *unconditional lower-bound guarantee*, then we may apply the conformal prediction idea of Shafer and Vovk [2008] and Vovk et al. [2005], which has been applied to several complex

models, including non-parametric regression; see Lei and Wasserman [2014], Lei et al. [2018], Romano et al., and Sesia and Candès.

In the paper at hand, we assume a linear model and discuss how to construct an asymptotically valid prediction interval in the context of conditional coverage that also possesses some unconditional guarantees as discussed above. To be more concrete, suppose we have an $n \times p$ design matrix $\mathscr{X}$, independent and identically distributed residuals $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T$, dependent variables $\mathscr{Y} = (\mathscr{Y}_1, ..., \mathscr{Y}_n)^T$ where $\mathscr{Y} = \mathscr{X}\beta + \varepsilon$ and a fixed new regressor(a vector) $\mathscr{X}_f$ that is of interest. We would like to provide a $1-\alpha$ prediction interval $\Gamma = \Gamma(\mathscr{X}, \mathscr{Y}, \mathscr{X}_f)$ for the future observation $\mathscr{Y}_f = \mathscr{X}_f^T\beta + \xi$; here ξ is independent of $\mathscr{X}, \mathscr{Y}$ and has the same distribution as ε_1. The aforementioned bootstrap methods will ensure that $Prob\left(\mathscr{Y}_f \in \Gamma\right) \approx 1-\alpha$, but without a lower-bound guarantee. On the other hand, the conformal prediction(e.g., Chernozhukov et al. [2021]) method yields an interval Γ such that $Prob\left(\mathscr{Y}_f \in \Gamma\right) \geq 1-\alpha$, i.e., an *unconditional* lower-bound guarantee. However, we are more interested in quantifying the performance of a prediction interval in terms of its conditional coverage probability $Prob\left(\mathscr{Y}_f \in \Gamma | \mathscr{Y}\right)$(or $Prob\left(\mathscr{Y}_f \in \Gamma | \mathscr{Y}, \mathscr{X}_f, \mathscr{X}\right)$ under random design).

The reason for our interest comes from two aspects. On one hand, the conditional probability precisely describes how statisticians make prediction in practice. By using the unconditional probability

$$Prob\left(\mathscr{Y}_f \in \Gamma\right) = E\left(Prob\left(\mathscr{Y}_f \in \Gamma | \mathscr{Y}\right)\right) \tag{4.1}$$

it is as if we assume that the statistician has not observed $\mathscr{Y}$ before making the prediction.

Realistically, however, statisticians have observed $\mathscr{Y}$ and have fitted the model before they make predictions. Therefore, it is informative to understand what happens to $\mathscr{Y}_f$ given our knowledge of all data (including $\mathscr{Y}$) rather than "on average" among all possible $\mathscr{Y}$.

On the other hand, according to eq. (4.1), analysis of the conditional probability is a more fundamental topic than the unconditional one. For example, if for any given $\delta > 0$,

$Prob\left(\{|Prob\left(\mathscr{Y}_f \in \Gamma|\mathscr{Y}\right) - (1-\alpha)| > \delta\}\right) \to 0$ as $n \to \infty$, then we can take the conditional expectation and have

$$\begin{aligned} |Prob\left(\mathscr{Y}_f \in \Gamma\right) - (1-\alpha)| &\leq E\left(|Prob\left(\mathscr{Y}_f \in \Gamma|\mathscr{Y}\right) - (1-\alpha)|\right) \\ &\leq \delta + Prob\left(\{|Prob\left(\mathscr{Y}_f \in \Gamma|\mathscr{Y}\right) - (1-\alpha)| > \delta\}\right) \end{aligned} \tag{4.2}$$

which implies $Prob\left(\mathscr{Y}_f \in \Gamma\right) \to 1-\alpha$.

Consequently, the aforementioned performance goals of asymptotic validity and lower bound guarantee should be recast in terms of conditional coverage. Note, however, that $Prob\left(\mathscr{Y}_f \in \Gamma|\mathscr{Y}\right)$ is a random variable itself – see e.g. definition 1.3 in Çinlar [2011]. Hence, the performance goals are now stochastic, i.e., $Prob\left(\mathscr{Y}_f \in \Gamma|\mathscr{Y}\right) \to_p 1-\alpha$ and $Prob\left(\mathscr{Y}_f \in \Gamma|\mathscr{Y}\right) \geq 1-\alpha$ with a specific probability. Surprisingly, we can achieve these goals simultaneously through a careful re-design of our prediction intervals. Definition 5 in what follows describes our new performance aim. Before stating it, however, we need to clarify our notation since our results hold true for both fixed and random design. In the latter case, however, all probabilities and expectations will be understood as being conditional on $\mathscr{X}$; see Definition 4 below.

Definition 4. *Consider the two cases:*

(a) **Fixed design***, i.e., there is no randomness involved in the design matrix $\mathscr{X}$ and the new regressor $\mathscr{X}_f$. In this case, we define $\mathbf{P}(\cdot) = Prob(\cdot)$, $\mathbf{P}^*(\cdot) = Prob(\cdot|\mathscr{Y})$, $\mathbf{E}\cdot = E\cdot$, and $\mathbf{E}^*\cdot = E(\cdot|\mathscr{Y})$.*

(b) **Random design***, i.e., there is randomness involved in the design matrix $\mathscr{X}$ (and possibly in the new regressor $\mathscr{X}_f$ as well). In this case, we define $\mathbf{P}(\cdot) = Prob(\cdot|\mathscr{X}, \mathscr{X}_f)$, $\mathbf{P}^*(\cdot) = Prob(\cdot|\mathscr{Y}, \mathscr{X}, \mathscr{X}_f)$, $\mathbf{E}\cdot = E(\cdot|\mathscr{X}, \mathscr{X}_f)$, and $\mathbf{E}^*\cdot = E(\cdot|\mathscr{Y}, \mathscr{X}, \mathscr{X}_f)$. Furthermore, convergences and probability statements will be understood to hold almost surely in $\mathscr{X}$ and $\mathscr{X}_f$.*

We can now state our new performance aims in general.

Definition 5 (Prediction interval with unconditional guarantee)**.** *Assume an $n \times p$ design matrix $\mathscr{X}$, independent and identically distributed (i.i.d.) residuals $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T \in \mathbf{R}^n$, and that the*

dependent variables $\mathscr{Y}$ satisfy a linear model $\mathscr{Y} = \mathscr{X}\beta + \varepsilon$. For a new regressor $\mathscr{X}_f \in \mathbf{R}^p$ and a potential future observation $\mathscr{Y}_f$, we say that $\Gamma = \Gamma(\mathscr{X}, \mathscr{Y}, \mathscr{X}_f)$ is the $1-\alpha$ prediction interval with $1-\gamma$ unconditional guarantee if the following conditions hold true:

1. For any given $\delta > 0$,

$$\mathbf{P}\left(\{|\mathbf{P}^*\left(\mathscr{Y}_f \in \Gamma\right) - (1-\alpha)| > \delta\}\right) \to 0 \tag{4.3}$$

2.

$$\mathbf{P}\left(\{\mathbf{P}^*(\mathscr{Y}_f \in \Gamma) \geq 1-\alpha\}\right) \to 1-\gamma \tag{4.4}$$

as $n \to \infty$; here, α, γ are constants in $(0,1)$. We call $1-\alpha$ the nominal (conditional) coverage probability and $1-\gamma$ the guarantee level.

Intuitively, Definition 5 requires the prediction interval Γ to have an asymptotically correct conditional coverage probability $1-\alpha$. Meanwhile, the hope is that Γ's conditional coverage probability is greater than $1-\alpha$ with a specific (unconditional) probability.

Remark 11. *In Definition 5, the validity condition (eq.*(4.3)*) is ubiquitous and easily understood, but the second condition (eq.*(4.4)*) needs some clarifications. This remark aims to stress that the second condition is not redundant.*

Suppose a prediction interval Γ satisfies (4.3) *with $1-\alpha = 95\%$. If the sample size n is very large, then Γ's conditional coverage probability is close to* 95%*. In this situation, whether or not the conditional coverage probability is greater than* 95% *is not important. However, if the sample size is merely moderate, then Γ's conditional coverage probability can be significantly smaller than* 95%*. Indeed, in table 4.2 and 4.3(in section 4.7), a nominal* 95% *prediction interval may have a conditional coverage probability less than* 91%*.*

In addition suppose Γ satisfies (4.4) *with $1-\gamma = 85\%$. When the sample size is moderate, the guarantee level may also be smaller than* 85%*. However, this condition still gives us an extra assurance that Γ is 'not likely' to have an under-coverage issue. Moreover, it is even unlikely for*

Γ*'s conditional coverage probability to be far less than* 95%.

Notably, statisticians have already noticed a gap between theoretical validity and finite sample performance. That is, an asymptotic valid prediction interval(e.g., Stine [1985]) will often manifest under-coverage in practice; see Politis [2013] for a discussion. In order to fill this gap, Politis [2015] proposed the definition of a 'pertinent prediction interval', which is a notion stronger than (4.3)*. Definition 5 provides a new perspective on this problem.*

Remark 12 (Further discussion on eq. (4.4))**.** *A drawback of eq.* (4.3) *is that it takes place asymptotically (as the sample size $n \to \infty$). Hence, a prediction interval may satisfy* (4.3) *with a given $\delta > 0$, but for a given sample size n, the probability of the event $\{|\mathbf{P}^*\left(\mathscr{Y}_f \in \Gamma\right) - (1-\alpha)| > \delta\}$ may not be negligible. If the event $\{|\mathbf{P}^*\left(\mathscr{Y}_f \in \Gamma\right) - (1-\alpha)| > \delta\}$ is to happen, we may prefer $\mathbf{P}^*\left(\mathscr{Y}_f \in \Gamma\right) > 1-\alpha+\delta$ (i.e., overcoverage) to $\mathbf{P}^*\left(\mathscr{Y}_f \in \Gamma\right) < 1-\alpha-\delta$ (i.e., undercoverage). Eq.*(4.4) *reflects the intensity of this preference, i.e., overcoverage is more likely to happen if we choose large $1-\gamma$. Notably, we require* (4.3) *and* (4.4) *to happen simultaneously. Therefore,* (4.4) *calibrates the usual prediction interval—e.g., the prediction interval generated by the residual-based bootstrap (Stine [1985])—instead of creating a new one.*

Remark 13. *This remark compares definition 5 with classical bootstrap methods and conformal predictions. Recall bootstrap methods always require $\mathbf{P}(\mathscr{Y}_f \in \Gamma) \to 1-\alpha$ like Stine [1985], or $\mathbf{P}^*(\mathscr{Y}_f \in \Gamma) \to_p 1-\alpha$ like Politis [2015]. On the other hand, conformal prediction is considered a model-free, non-asymptotic method to generate a prediction interval. But its guarantee is* ***on average over the observations and over the future random regressor*** *$\mathscr{X}_f$. In table 4.1, it appears that the guarantee level of a conformal prediction is only* 10.2% *even when the sample size is* 1600*, implying that in* 89.8% *of the samples we have conditional coverage probability less than $1-\alpha$. The new regressor $\mathscr{X}_f$ is fixed (or conditioned upon) in our paper, so a complete model-free procedure (i.e., a procedure that constructs a consistent prediction interval for any models) is impossible; see Barber et al. [2021].*

In order to increase the guarantee level, Vovk [2012] introduced the idea of a tolerance

region; Vovk's tolerance region is constructed as follows. First, perform the split-conformal prediction introduced in Lei et al. [2018] to make the $1-\alpha$ *prediction interval* $C_{1-\alpha}(\mathscr{X}_f)$ *for* $\mathscr{Y}_f$*. Denote* n_{calib} *the size of the calibration set (i.e.,* I_2 *in algorithm 2 of Lei et al. [2018]). Then choose* α' *such that*

$$\gamma \geq binom_{n_{calib},\alpha}(\lfloor \alpha'(n_{calib}+1)-1 \rfloor) \tag{4.5}$$

where $binom_{n,\alpha}$ *denotes the cumulative distribution function of a binominal distribution with size n and probability* α*, and* $\lfloor x \rfloor$ *denotes the largest integer that is smaller than or equal to x. Then Vovk's tolerance region is defined as* $C_{1-\alpha'}(\mathscr{X}_f)$*. According to proposition 2b in Vovk [2012], this prediction interval satisfies*

$$\mathbf{P}(\mathbf{P}^*(\mathscr{Y}_f \in C_{1-\alpha'}(\mathscr{X}_f)) \geq 1-\alpha) \geq 1-\gamma \tag{4.6}$$

which is similar to condition (4.4)*. However, Vovk's tolerance region might not satisfy* (4.3)*; that is why* (4.5) *is an inequality rather than an equality. In section 4.7, we compare several prediction methods via finite-sample simulations; it looks like Vovk's tolerance region is typically wider than other prediction intervals.*

Table 4.1 shows that this tolerance region has high guarantee levels among various linear models.

Definition 5 still follows a bootstrap framework but additionally requires $\mathbf{P}^*(\mathscr{Y}_f \in \Gamma) \geq 1-\alpha$ *for a specific proportion of observations. This definition is useful for understanding an existing bootstrap algorithm, like corollary 1. It also maintains the balance between* Γ*'s length and its possibility of under-coverage.*

Definition 5 is not easy to achieve; to see why, we present a simulation in table 4.1. The guarantee level(i.e., proportion of observations having conditional coverage probability $\geq 1-\alpha$) of the aforementioned methods are not very high.

Our paper has two main contributions. On the one hand, it derives the Gaussian approxi-

Table 4.1. Quantiles of conditional coverage probabilities and guarantee levels of prediction intervals on the **Experiment model** (see section 4.7). The errors are generated by i.i.d. normal random variables with mean 0 and variance 1. The nominal coverage probability is 95%. We use the R-package maintained by Tibshirani et al. [2021] to perform conformal predictions. For Vovk's tolerance region, we chose $\gamma = 15\%$ in (4.5). The notation 'Quantiles' denotes the quantiles of conditional coverage probabilities and 'Guarantee' denotes the guarantee level(see definition 5).

Sample size	Algorithm	Quantiles			Guarantee
		15%	30%	50%	
100	Residual bootstrap	91.0%	92.5%	93.9%	31.3%
	MF/MB bootstrap	93.5%	94.7%	95.8%	66.7%
	Conformal prediction	90.0%	91.8%	93.5%	27.9%
	Split conformal prediction	95.2%	96.7%	97.8%	87.0%
	Jackknife conformal prediction	92.7%	95.3%	97.2%	56.5%
	Vovk's tolerance region	97.3%	98.4%	99.2%	95.5%
400	Residual bootstrap	93.3%	94.0%	94.7%	40.8%
	MF/MB bootstrap	93.8%	94.6%	95.2%	56.3%
	Conformal prediction	91.9%	92.7%	93.6%	15.5%
	Split conformal prediction	93.9%	94.8%	95.6%	66.4%
	Jackknife conformal prediction	93.8%	95.0%	96.2%	52.6%
	Vovk's tolerance region	96.1%	96.8%	97.5%	95.2%
1600	Residual bootstrap	94.0%	94.5%	95.0%	48.0%
	MF/MB bootstrap	94.2%	94.6%	95.0%	52.8%
	Conformal prediction	92.0%	92.7%	93.4%	10.2%
	Split conformal prediction	94.0%	94.6%	95.2%	57.7%
	Jackknife conformal prediction	93.0%	93.6%	94.3%	25.5%
	Vovk's tolerance region	94.8%	95.3%	95.9%	81.3%

mation for the difference between the conditional probability of a nominal prediction interval and the conditional probability of a prediction interval based on residual-based bootstrap. In practice, bootstrap approximates the former by the latter, and the non-zero difference will make the former deviate from $1-\alpha$. This leads to the fact that the *residual-based bootstrap algorithm asymptotically has guarantee level of* 50%. On the other hand, we develop a new method to construct a prediction interval satisfying definition 5 with arbitrarily chosen α, γ.

We employ a simple example to illustrate why a classical prediction interval becomes problematic under the conditional coverage context in section 4.3. After that, we introduce the frequently used notations and assumptions in section 4.4. In section 4.5, we derive the Gaussian approximation result. In section 4.6, we develop the algorithm to construct the newly proposed prediction interval. We perform some simulations to illustrate the proposed algorithm's finite sample performance in section 4.7, and provide some conclusions in section 4.8. The proofs of the theoretical results will be deferred to chapter C.

4.3 An intuitive illustration in the Gaussian case

For the sake of illustration, in this section only we suppose the residual ε_1 has a normal distribution with mean 0 and known variance σ^2. Assume $\mathscr{X}^T\mathscr{X}$ is invertible. Denote $\Phi(x)$ as the cumulative distribution function of the standard normal distribution and $\Phi^{-1}(\alpha), \alpha \in (0,1)$ as its $\alpha-$quantile, i.e., $\Phi(\Phi^{-1}(\alpha)) = \alpha$. Adopt the notations $\mathbf{P}, \mathbf{P}^*$ in definition 4. If we do not care about the conditional coverage, we can define $\widehat{\beta} = (\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\mathscr{Y}$ and use the normal distribution $1-\alpha$ prediction interval $\mathscr{P}_1 = [\mathscr{X}_f^T\widehat{\beta} + \sigma\Phi^{-1}(\frac{\alpha}{2})\sqrt{1+\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}_f},\ \mathscr{X}_f^T\widehat{\beta} + \sigma\Phi^{-1}(1-\frac{\alpha}{2})\sqrt{1+\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}_f}]$ for the future response $\mathscr{Y}_f$. Since the random variable $\mathscr{Y}_f - \mathscr{X}_f^T\widehat{\beta}$ has normal distribution with mean 0 and variance $\sigma^2(1+\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}_f)$, it follows that

$$\mathbf{P}\left(\mathscr{Y}_f \in \mathscr{P}_1\right) = \mathbf{P}\left(\Phi^{-1}(\frac{\alpha}{2}) \le \frac{\mathscr{Y}_f - \mathscr{X}_f^T\widehat{\beta}}{\sigma\sqrt{1+\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}_f}} \le \Phi^{-1}(1-\frac{\alpha}{2})\right) = 1-\alpha. \quad (4.7)$$

In other words, $\mathscr{P}_1$ has precise unconditional coverage probability. However, if we take the conditional coverage into consideration, the random variable $\mathscr{Y}_f - \mathscr{X}_f^T\widehat{\beta}|\mathscr{Y}$ (or $\mathscr{Y}_f - \mathscr{X}_f^T\widehat{\beta}|\mathscr{Y}, \mathscr{X}_f, \mathscr{X}$ under random design) has normal distribution with mean $\mathscr{X}_f^T\beta - \mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\mathscr{Y}$ and variance σ^2. According to Taylor's theorem,

$$\begin{aligned}
\mathbf{P}^*\left(\mathscr{Y}_f \in \mathscr{P}_1\right) &= \mathbf{P}^*\left(\Phi^{-1}(\frac{\alpha}{2}) \le \frac{\mathscr{Y}_f - \mathscr{X}_f^T\widehat{\beta}}{\sigma\times\sqrt{1+\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}_f}} \le \Phi^{-1}(1-\frac{\alpha}{2})\right) \\
&= \Phi\left(\sqrt{1+\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}_f}\times\Phi^{-1}(1-\frac{\alpha}{2}) + \frac{\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon}{\sigma}\right) \\
&\quad -\Phi\left(\sqrt{1+\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}_f}\times\Phi^{-1}(\frac{\alpha}{2}) + \frac{\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon}{\sigma}\right) \\
&\approx 1-\alpha+\Phi'(\Phi^{-1}(1-\frac{\alpha}{2}))\times\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}_f\times\Phi^{-1}(1-\frac{\alpha}{2}) \\
&\quad +\Phi''(\Phi^{-1}(1-\frac{\alpha}{2}))\times(\frac{\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon}{\sigma})^2
\end{aligned} \quad (4.8)$$

The last line of (4.8) is derived by expanding the second line on $\Phi^{-1}(1-\frac{\alpha}{2})$, and expanding the third line on $\Phi^{-1}(\frac{\alpha}{2})$. Since $\Phi^{''}(\Phi^{-1}(1-\frac{\alpha}{2}))<0$, $\frac{(\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon)^2}{\sigma^2(\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}_f)}$ has χ_1^2 distribution and $\Phi^{''}(x)=-x\Phi^{'}(x)$ for any x,

$$\begin{aligned}
&\mathbf{P}\left(\{\mathbf{P}^*\left(\mathscr{Y}_f\in\mathscr{P}_1\right)\geq 1-\alpha\}\right)\\
\approx\mathbf{P}&\left(\frac{(\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon)^2}{\sigma^2\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}_f}\leq\frac{\Phi^{'}(\Phi^{-1}(1-\frac{\alpha}{2}))\times\Phi^{-1}(1-\frac{\alpha}{2})}{-\Phi^{''}(\Phi^{-1}(1-\frac{\alpha}{2}))}\right)
\end{aligned}\tag{4.9}$$

which approximately equals 0.683. Therefore, the prediction interval $\mathscr{P}_1$ has about 68% guarantee level.

However, it is possible to find a prediction interval with a desired guarantee level, say $1-\gamma$. Wallis [1951], Lieberman and Miller [1963] and De Gryze et al. [2007] considered this problem and defined the 'tolerance interval' that controlled the guarantee level. However, their work assumed that the residuals ε_1 had normal distribution. Moreover, an $1-\gamma$ tolerance interval does not ensure having asymptotic coverage probability $1-\alpha$. We define $C_{1-\gamma}$ as the $1-\gamma$ quantile of a χ_1^2 distribution, and let $c_{1-\gamma}=-\Phi^{''}(\Phi^{-1}(1-\frac{\alpha}{2}))\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}_f\times C_{1-\gamma}\ /\ (2\Phi^{'}(\Phi^{-1}(1-\frac{\alpha}{2})))>0$. We construct the prediction interval $\mathscr{P}_2=[\mathscr{X}_f^T\widehat{\beta}+\sigma\times$

$(\Phi^{-1}(\frac{\alpha}{2})-c_{1-\gamma}),\ \mathscr{X}_f^T\widehat{\beta}+\sigma\times(\Phi^{-1}(1-\frac{\alpha}{2})+c_{1-\gamma})]$. We can now compute

$$\begin{aligned}
\mathbf{P}^*\left(\mathscr{Y}_f\in\mathscr{P}_2\right)&=\mathbf{P}^*\left(\Phi^{-1}\left(\frac{\alpha}{2}\right)-c_{1-\gamma}\leq\frac{\mathscr{Y}_f-\mathscr{X}_f^T\widehat{\beta}}{\sigma}\leq\Phi^{-1}\left(1-\frac{\alpha}{2}\right)+c_{1-\gamma}\right)\\
&=\Phi\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right)+c_{1-\gamma}+\frac{\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon}{\sigma}\right)\\
&\quad-\Phi\left(\Phi^{-1}\left(\frac{\alpha}{2}\right)-c_{1-\gamma}+\frac{\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon}{\sigma}\right)\\
&\approx 1-\alpha+2\Phi'\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right)\times c_{1-\gamma}\\
&\quad+\Phi''\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right)\times(\frac{\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon}{\sigma})^2\\
&\text{which implies that}\quad\mathbf{P}\left(\{\mathbf{P}^*\left(\mathscr{Y}_f\in\mathscr{P}_2\right)\geq 1-\alpha\}\right)\\
&\approx\mathbf{P}(-\Phi''\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right)\frac{(\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon)^2}{\sigma^2\times\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}_f}\\
&\leq-\Phi''(\Phi^{-1}(1-\frac{\alpha}{2}))\times C_{1-\gamma})\\
&=1-\gamma
\end{aligned}\tag{4.10}$$

Hence, prediction interval $\mathscr{P}_2$ has guarantee level about $1-\gamma$. Note that since $c_{1-\gamma}$ has order $O(1/n)$, this correction does not significantly enlarge the width of the prediction interval. In other words, if the dimension of the parameter vector is fixed, then **the uncorrected and the corrected prediction intervals coincide with each other asymptotically.**

In the end of this section, we would like to briefly discuss the prediction problem under the high dimensional setting, i.e., $p/n\to s\in(0,1)$. Bates et al. [2021] and Dobriban and Wager [2018] also considered this problem but they focused on estimating the prediction error. Steinberger and Leeb [2016] and Zhang and Politis [2020] constructed asymptotically valid prediction intervals for a (sparse) high dimensional linear model. Suppose $\exists\ 0<c\leq C<\infty$ such that all eigenvalues of $\frac{1}{n}\mathscr{X}^T\mathscr{X}$ is greater than c and smaller than C. This assumption is achievable according to Bai and Yin [1993]. If p is large and the new regressor $\mathscr{X}_f$ is not sparse, then the term $\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}_f\geq\frac{\mathscr{X}_f^T\mathscr{X}_f}{Cn}$, which does not tend to 0 as $n\to\infty$. Therefore,

despite that $\frac{\mathscr{Y}_f - \mathscr{X}_f^T \widehat{\beta}}{\sigma\sqrt{1+\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}_f}}$ has normal distribution(so (4.7) is satisfied), we cannot use Taylor expansion in (4.8) and (4.10). So we need a new method to construct a prediction interval in order to satisfy Definition 5. Moreover, $c_{1-\gamma}$ will not converge to 0 as $n \to \infty$, and

$$\begin{aligned} &\Phi^{-1}(1-\frac{\alpha}{2}) + c_{1-\gamma} - \Phi^{-1}(1-\frac{\alpha}{2})\sqrt{1+\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}_f} \\ = \Phi^{-1}(1-\frac{\alpha}{2})&\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}_f \times \left(\frac{C_{1-\gamma}}{2} - \frac{1}{1+\sqrt{1+\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}_f}}\right) \end{aligned} \tag{4.11}$$

which does not converge to 0 as the sample size $n \to \infty$. So modification (4.10) will not be negligible asymptotically, and the prediction intervals (4.8) and (4.10) will not be close to each other even when n is large. In other words, constructing a 'good' prediction interval(e.g., a prediction interval satisfying definition 5) can be a challenging problem if the dimension of parameters is large. This paper will focus on the finite dimensional situation. However, our work should lay a good foundation for the high dimensional prediction problem.

Another limitation in this section is that the marginal distribution of the errors is assumed to be normal with known variance σ^2, which is always not true. In the general situation, the marginal distribution of the errors is not normal and is unknown. As a consequence, the correction can be significantly larger than $1/n$. Besides, we need to use resampling to find a satisfactory correction; this will be the subject of the following sections.

4.4 Preliminary notions

For the remainder of the paper, we revert to the general setup: an $n \times p$ design matrix $\mathscr{X}$ (assumed to have full-rank), the dependent variable $\mathscr{Y}$ satisfying the linear model $\mathscr{Y} = (\mathscr{Y}_1, ..., \mathscr{Y}_n)^T = \mathscr{X}\beta + \varepsilon$ with respect to the i.i.d. errors $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T$; here, ε_1 has mean zero, unknown variance σ^2, and cumulative distribution function denoted by F. We denote $\mathscr{X}^T = (\mathscr{X}_1, ..., \mathscr{X}_n)$, $\mathscr{X}_i = (\mathscr{X}_{i1}, ..., \mathscr{X}_{ip})^T \in \mathbf{R}^p, i = 1, 2, ..., n$, the new regressor $\mathscr{X}_f \in \mathbf{R}^p$ and the new dependent variable $\mathscr{Y}_f$(the subscript 'f' will only be used for future observations).

Define

$$\widehat{\beta} = (\widehat{\beta}_1, ..., \widehat{\beta}_p)^T = (\mathscr{X}^T \mathscr{X})^{-1} \mathscr{X}^T \mathscr{Y} \tag{4.12}$$

as the least squares estimator of the parameter vector β. Then, define the centered estimated residual $\widehat{\varepsilon} = (\widehat{\varepsilon}_1, ..., \widehat{\varepsilon}_n)^T$ and the residual empirical process $\widehat{F}(x)$ for any $x \in \mathbf{R}$ respectively as

$$\begin{aligned} \widetilde{\varepsilon}_i = \mathscr{Y}_i - \mathscr{X}_i^T \widehat{\beta} = \varepsilon_i - \mathscr{X}_i^T (\widehat{\beta} - \beta) \\ \widehat{\varepsilon}_i = \widetilde{\varepsilon}_i - \frac{1}{n} \sum_{j=1}^{n} \widetilde{\varepsilon}_j \\ \widehat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\widehat{\varepsilon}_i \le x}. \end{aligned} \tag{4.13}$$

We also define $\overline{\mathscr{X}}_n = \frac{1}{n} \sum_{i=1}^{n} \mathscr{X}_i \in \mathbf{R}^p$. From (4.13),

$$\int x d\widehat{F} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\varepsilon}_i = 0, \;\; \widehat{\sigma}^2 = \int x^2 d\widehat{F} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\varepsilon}_i^2. \tag{4.14}$$

Here and in the rest of this paper, the lower case letters x, y, z will be used to represent a scalar. For a function $g : \mathbf{R} \to \mathbf{R}$, define g' as its derivative. Denote $\mathbf{D} = \mathbf{D}[0,1]$ the space of *càdlàg* functions on $[0,1]$ with Skorohod topology–see chapter 3 of Billingsley [1999].

To derive our results, we need the following assumptions.

Assumptions:

1. ε_1's distribution is absolutely continuous with respect to Lebesgue measure. F is second order continuous differentiable and $\sup_{x \in \mathbf{R}} |F''(x)| < \infty$, $\mathbf{E}\varepsilon_1 = 0$, $\mathbf{E}|\varepsilon_1|^4 < \infty$. The new regressor $\mathscr{X}_f \in \mathbf{R}^p$ and the new dependent variable $\mathscr{Y}_f$ satisfy $\mathscr{Y}_f = \mathscr{X}_f^T \beta + \xi$. ξ is independent of ε and has the same distribution as ε_1.

2. One of the two following conditions holds true:

   2.1. **Fixed design:** $\mathscr{X}$ and $\mathscr{X}_f$ are fixed, i.e., non-random.

   2.2. **Random design:** $\mathscr{X}$ and $\mathscr{X}_f$ are random. However, $\mathscr{X}_f$ is independent of ε, ξ;

and $\mathscr{X}$ is independent of $\varepsilon, \xi, \mathscr{X}_f$.

3. $\mathscr{X}^T\mathscr{X}$ is invertible for $\forall n \geq p$ and $\lim_{n\to\infty} \frac{\mathscr{X}^T\mathscr{X}}{n} = A$, $\lim_{n\to\infty} \overline{\mathscr{X}}_n = b$; here A is an invertible matrix and $b \in \mathbf{R}^p$. Besides, there exists a constant $M > 0$ such that $\|\mathscr{X}_i\|_2 \leq M$ for $i = 1, 2, ..., n$ and $\|\mathscr{X}_f\|_2 \leq M$. $\|.\|_2$ denotes the Euclidean norm in $\mathbf{R}^p$.

4. Define $H(x) = \mathbf{E}\varepsilon_1 \mathbf{1}_{\varepsilon_1 \leq x}$ and for $\forall x, z \in \mathbf{R}$.

$$\begin{aligned} \mathscr{V}(x,z) &= \sigma^2 F'(x)F'(z)\left(\mathscr{X}_f^T A^{-1}\mathscr{X}_f + 1 - 2\mathscr{X}_f^T A^{-1} b\right) \\ -(F'(x)H(z) + F'(z)H(x))(\mathscr{X}_f^T A^{-1} b - 1) &+ F(\min(x,z)) - F(x)F(z) \\ \text{We also define } \mathscr{U}(x) &= \mathscr{V}(x,x) + \mathscr{V}(-x,-x) - 2\mathscr{V}(x,-x) \end{aligned} \tag{4.15}$$

Assume $F'(x) > 0, \forall x \in \mathbf{R}$, and $\mathscr{U}(x) > 0$ for $\forall\ 0 < x < \infty$.

For a function $g : \mathbf{R} \to \mathbf{R}$ and a point $x \in \mathbf{R}$, we define the limit from the left as

$$g^-(x) = \lim_{y\to x, y<x} g(y) \tag{4.16}$$

if this limit exists. Note that $g \in \mathbf{D}$ implies that $g^-(x)$ exists for $\forall x \in (0,1)$. As in section 1.1.4 of Politis et al. [1999], for any $0 < \alpha < 1$, we define the α quantile of a cumulative distribution function g as

$$c_\alpha = \inf\{x \in \mathbf{R} : g(x) \geq \alpha\}. \tag{4.17}$$

The meaning of notations $\mathbf{P}, \mathbf{P}^*, \mathbf{E}, \mathbf{E}^*$ is presented in definition 4. The symbol $\to$ represents convergence in $\mathbf{R}$, and $\to_{\mathscr{L}}$ represents convergence in distribution. Without being specified, the convergence assumes the sample size $n \to \infty$. $\Phi(\cdot)$ and $\Phi^{-1}(\cdot)$ respectively represents the cumulative distribution function and the quantile of the standard normal distribution. In the case of random design, the convergence results hold true for almost sure $\mathscr{X}$ and $\mathscr{X}_f$.

Remark 14. *(a) We centered $\widetilde{\varepsilon}_i$ in eq.* (4.13)*, but if the design matrix $\mathscr{X}$ has a column of ones, then summation of the estimated residuals will be* 0 *exactly, and re-centering is superfluous.*

(b) In the case of random design, we assume assumption 3 and 4 happen for almost sure $\mathscr{X}$ and $\mathscr{X}_f$.

(c) There are various linear model settings, e.g., presence of outliers, errors being dependent, errors being being heteroskedastic, etc. This paper cannot discuss all situations simultaneously. So we focus on the classical setting, i.e., without outliers and errors are i.i.d., to present our work.

4.5 Gaussian approximation in bootstrap prediction

Residual-based bootstrap has been widely used in interval prediction for various models, such as Thombs and Schucany [1990], and Pan and Politis [2016b]. Stine [1985] introduced a residual-based bootstrap algorithm for prediction, but this algorithm is typically characterized by finite sample undercoverage; see Pan and Politis [2016a]. To alleviate the finite-sample undercoverage, Politis [2015] proposed the *Model-free/Model-Based (MF/MB) bootstrap*, that resamples the *predictive* residuals $\widehat{r} = (\widehat{r}_1, ..., \widehat{r}_n)^T$ instead of the usual fitted residuals. The predictive residuals are sometimes called the 'leave-one-out' residuals, and are defined as:

$$\widetilde{r}_i = \mathscr{Y}_i - \mathscr{X}_i^T (\mathscr{X}_{-i}^T \mathscr{X}_{-i})^{-1} \mathscr{X}_{-i}^T \mathscr{Y}_{-i}, \;\; \widehat{r}_i = \widetilde{r}_i - \frac{1}{n} \sum_{j=1}^{n} \widetilde{r}_j, \; i = 1, 2, ..., n \tag{4.18}$$

here $\mathscr{X}_{-i}$ and $\mathscr{Y}_{-i}$ are the design matrix $\mathscr{X}$ and the dependent variable vector $\mathscr{Y}$ respectively, having left out the *i*th row. For a least squares estimator, the predictive residuals can be efficiently computed using the hat matrix; see theorem 10.1 in Seber and Lee Seber and Lee [2003].

For concreteness, the algorithms are as follows:

Algorithm 4 (Residual-based bootstrap)**.** ***Input:*** *Design matrix $\mathscr{X}$ and dependent variable data vector $\mathscr{Y}$ satisfying $\mathscr{Y} = \mathscr{X}\beta + \varepsilon$, the new regression vector $\mathscr{X}_f$ of interest, number of bootstrap replicates B, nominal coverage probability $1 - \alpha$*

1. Calculate statistics $\widehat{\beta} = (\mathscr{X}^T \mathscr{X})^{-1} \mathscr{X}^T \mathscr{Y}$ and $\widehat{\varepsilon} = (\widehat{\varepsilon}_1, ..., \widehat{\varepsilon}_n)^T$ as in eq. (4.13)*.*

2. Generate i.i.d. residuals $\varepsilon^* = (\varepsilon_1^*, ..., \varepsilon_n^*)^T$ *and* ξ^* *by drawing from* $\widehat{\varepsilon}_1, ..., \widehat{\varepsilon}_n$ *with replacement. Then calculate* $\mathscr{Y}^* = \mathscr{X}\widehat{\beta} + \varepsilon^*$ *and* $\mathscr{Y}_f^* = \mathscr{X}_f^T\widehat{\beta} + \xi^*$. *Re-estimate* $\widehat{\beta}^* = (\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\mathscr{Y}^*$ *and calculate the prediction root* $\delta_b^* = \mathscr{Y}_f^* - \mathscr{X}_f^T\widehat{\beta}^*$

3. Repeat step 2 for $b = 1, 2, ..., B$, *and calculate the* $1-\alpha$ *(unadjusted) sample quantile* $\widehat{c}_{1-\alpha}^*$ *of* $|\delta_b^*|$, $b = 1, 2, ..., B$.

4. The prediction interval of $\mathscr{Y}_f$ *is given by* $\left\{\mathscr{Y}_f : |\mathscr{Y}_f - \mathscr{X}_f^T\widehat{\beta}| \leq \widehat{c}_{1-\alpha}^*\right\}$

Remark 15. *If we replace* $\widehat{\varepsilon}$ *by* $\widehat{r}$ *in algorithm 4, we then obtain the MF/MB bootstrap algorithm.*

The Glivenko - Cantelli theorem ensures that the empirical process of the bootstrapped prediction root $\mathscr{Y}_f^* - \mathscr{X}_f^T\widehat{\beta}^*$ converges to $\mathbf{P}^*\left(\mathscr{Y}_f^* - \mathscr{X}_f^T\widehat{\beta}^* \leq x\right)$ for any $x \in \mathbf{R}$ $\mathbf{P}^*$ almost surely as $B \to \infty$. Therefore, the residual-based bootstrap approximates the unobservable conditional cumulative distribution function $\mathbf{P}^*(|\mathscr{Y}_f - \mathscr{X}_f^T\widehat{\beta}| \leq x)$ by $\mathbf{P}^*\left(|\mathscr{Y}_f^* - \mathscr{X}_f^T\widehat{\beta}^*| \leq x\right)$, and estimates the latter distribution by the bootstrapped prediction root's empirical process; see Politis et al. [1999].

Notably, the notation $\mathbf{P}^*$ and $\mathbf{E}^*$ are used for the conditional probability and expectation conditioning on all observed data in this paper. Note that this definition coincides with 'the probability and expectation in the bootstrap world' which is typical in the bootstrap literature; see e.g., Cheng and Huang [2010]. The bootstrap approximation inevitably introduces errors. This section focuses on understanding the asymptotic behavior of the error process.

$$\mathscr{S}(x) = \sqrt{n}\left(\mathbf{P}^*(|\mathscr{Y}_f - \mathscr{X}_f^T\widehat{\beta}| \leq x) - \mathbf{P}^*(|\mathscr{Y}_f^* - \mathscr{X}_f^T\widehat{\beta}^*| \leq x)\right) \tag{4.19}$$

here $\mathscr{Y}_f^*$ and $\widehat{\beta}^*$ are defined in algorithm 4. We refer to Bickel and Freedman [1981] and Politis et al. [1999] for the related researches.

The asymptotic behavior of $\mathscr{S}$ is summarized in theorem 9.

Theorem 9. *Suppose assumption 1 to 4 hold true. Then for any given real numbers* $0 < r < s < \infty$,

$$\sup_{x\in[r,s]}\sup_{y\in\mathbf{R}}|\mathbf{P}\left(\mathscr{S}(x)\leq y\right)-\Phi\left(\frac{y}{\sqrt{\mathscr{U}(x)}}\right)|\to 0 \tag{4.20}$$

here $\mathscr{U}$ is defined in (4.15).

Hence, if a prediction interval Γ has the form $\left\{y\in\mathbf{R}:|y-\mathscr{X}_f^T\widehat{\beta}|\leq x\right\}$ (where x is a given positive number), then the conditional probability $\mathbf{P}^*\left(\mathscr{Y}_f\in\Gamma\right)$ and Γ's coverage probability estimated by the residual-based bootstrap algorithm (i.e., $\mathbf{P}^*(|\mathscr{Y}_f^*-\mathscr{X}_f^T\widehat{\beta}^*|\leq x)$, where $\mathscr{Y}_f^*$ and $\widehat{\beta}^*$ are defined in algorithm 4) has an error. Moreover, $\sqrt{n}\times$ this error has an asymptotic normal distribution with mean 0 and a specific variance $\mathscr{U}(x)$(depending on x).

In the conditional coverage context, an application of theorem 9 is to calculate a prediction interval's guarantee level. For example, by choosing $y=0$, and $x=c^*_{1-\alpha}$ which denotes the $1-\alpha$ quantile of the distribution $\mathbf{P}^*(|\mathscr{Y}_f^*-\mathscr{X}_f^T\widehat{\beta}^*|\leq x)$, we have the following corollary

Corollary 1. *Under assumptions 1 to 4, the prediction interval generated by residual-based bootstrap has an asymptotically* 50% *guarantee level.*

Alternatively, for a given $\gamma\in(0,1)$, we could choose $y=\Phi^{-1}(\gamma)$, the γ quantile of the standard normal distribution, and $x=c^*_{1-\alpha-\Phi^{-1}(\gamma)\times\sqrt{\mathscr{U}(c^*_{1-\alpha})}/\sqrt{n}}$. Since $\mathscr{U}$ is continuous, theorem 9 implies the event $\{\mathbf{P}^*(|\mathscr{Y}_f-\mathscr{X}_f^T\widehat{\beta}|\leq c^*_{1-\alpha-\Phi^{-1}(\gamma)\times\sqrt{\mathscr{U}(c^*_{1-\alpha})}/\sqrt{n}})-(1-\alpha)\geq 0\}$, which is equivalent to the event

$$\begin{aligned}\sqrt{n}(\mathbf{P}^*(|\mathscr{Y}_f-\mathscr{X}_f^T\widehat{\beta}|\leq c^*_{1-\alpha-\Phi^{-1}(\gamma)\times\sqrt{\mathscr{U}(c^*_{1-\alpha})}/\sqrt{n}})\\ -(1-\alpha-\frac{\Phi^{-1}(\gamma)\times\sqrt{\mathscr{U}(c^*_{1-\alpha})}}{\sqrt{n}}))\\ \geq\Phi^{-1}(\gamma)\times\sqrt{\mathscr{U}(c^*_{1-\alpha})}\end{aligned} \tag{4.21}$$

asymptotically has unconditional probability $1-\gamma$. In other words, the prediction interval $\{y\in\mathbf{R}:|y-\mathscr{X}_f^T\widehat{\beta}|\leq c^*_{1-\alpha-\Phi^{-1}(\gamma)\times\sqrt{\mathscr{U}(c^*_{1-\alpha})}/\sqrt{n}}\}$ has an asymptotic guarantee level $1-\gamma$. Section 4.6 adopts this idea. However, in order to estimate $\mathscr{U}$, statisticians need to estimate

$F(x) = Prob(\varepsilon_1 \leq x)$, the derivative $F^{'}(x)$ and $H(x) = \mathbf{E}\varepsilon_1 \mathbf{1}_{\varepsilon_1 \leq x}$, which is complex. To make our work practical, section 4.6 presents a resampling algorithm that automatically generates the desired prediction interval without estimating $\mathscr{U}$.

4.6 Bootstrap prediction interval with unconditional guarantee

For a fixed dimensional linear model, bootstrap algorithms like the residual-based bootstrap and the *MF/MB bootstrap* generate asymptotically valid prediction intervals. Besides, Steinberger and Leeb [2016] and Zhang and Politis [2020] constructed asymptotically valid prediction intervals for high dimensional linear models. However, the statistician cannot adjust those prediction intervals' guarantee level; for example, corollary 1 says that the residual-based bootstrap has asymptotic guarantee level 50%. Therefore, in practice, the statistician cannot expect the possibility for a prediction interval to have a conditional coverage probability less than the nominal coverage probability. Ideally, we would wish for an algorithm that can generate an asymptotic valid prediction interval with a suitable guarantee level which is useful for both fixed and high dimensional regression. However, if the dimension is large, eq.(4.11) shows that the prediction intervals satisfying different purposes may not coincide with each other asymptotically. Therefore, finding a 'good' prediction interval can be a subtle problem for a high dimensional regression.

Focus on the fixed dimensional linear regression, this section proposes two new variations on these bootstrap methods, namely the *Residual bootstrap with unconditional guarantee (RBUG)* and the *Predictive residual bootstrap with unconditional guarantee (PRBUG)*, that maintain the asymptotic validity but also allows us to choose the prediction interval's guarantee level. These algorithms involve two steps: generating a valid prediction interval by residual-based bootstrap or *MF/MB bootstrap*; then calibrating the length of the prediction interval. Calibration of a confidence/prediction interval is not a new idea; see Loh [1991, 1987], Politis et al. [1999] and

Beran [1990]. Their work calibrated a confidence interval based on the Edgeworth expansion. Our method does not use Edgeworth expansion. Instead, our method calibrates the prediction interval based on theorem 9 and the idea of eq. (4.21).

In order to use eq.(4.21), we need to estimate $\mathscr{U}$. In section C.1.2, we show that the error process $\mathscr{S}(x)$(defined in (4.19)) can be approximated by a special stochastic process $\widetilde{M}_m\left(\frac{x+m}{2m}\right)-\widetilde{M}_m^-\left(\frac{-x+m}{2m}\right)$, here

$$\widetilde{M}_m(x)=\sqrt{n}F^{'}(x_m)\left(\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon-\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j\right)-\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left(\mathbf{1}_{\varepsilon_j\leq x_m}-F(x_m)\right) \quad (4.22)$$

m is a sufficiently large positive integer and $x_m=2mx-m$. As long as m is large, changing m does not affect the value of $\widetilde{M}_m\left(\frac{x+m}{2m}\right)-\widetilde{M}_m^-\left(\frac{-x+m}{2m}\right)$. Fortunately, simulating $\widetilde{M}_m$ in the bootstrap world is not difficult. So we can implicitly estimate $\mathscr{U}$ by simulating $\widetilde{M}_m$. Algorithm 5 adopts this idea, i.e., first estimate $c^*_{1-\alpha}$, the $1-\alpha$ (unadjusted) quantile of the conditional distribution $\mathbf{P}^*(|\mathscr{Y}_f^*-\mathscr{X}_f^T\widehat{\beta}^*|\leq x)$. Then estimate the coverage probability adjustment $-\frac{\Phi^{-1}(\gamma)\times\sqrt{\mathscr{U}(c^*_{1-\alpha})}}{\sqrt{n}}$ in eq.(4.21) by simulating $\widetilde{M}_m$. Finally, calibrate the prediction interval based on the adjustment.

Algorithm 5 (*RBUG/PRBUG*)**.** ***Input:*** *Design matrix $\mathscr{X}$ and dependent variable data vector $\mathscr{Y}$ satisfying $\mathscr{Y}=\mathscr{X}\beta+\varepsilon$, the new regression vector $\mathscr{X}_f$ of interest, and number of bootstrap replicates B, number of replicates to find quantile's adjustment $\mathscr{B}_1$, nominal coverage probability $1-\alpha$, and nominal guarantee level $1-\gamma$*

Note: *For RBUG, we define $\widehat{\tau}=(\widehat{\tau}_1,...,\widehat{\tau}_n)^T=\widehat{\varepsilon}$ as in* (4.13)*, while for PRBUG, we define $\widehat{\tau}=\widehat{r}$ as in* (4.18)*.*

Calculate an unadjusted sample quantile

1. Calculate the statistics $\widehat{\beta}=(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\mathscr{Y}$ and $\widehat{\tau}$.

2. Generate i.i.d. residuals $\varepsilon^*=(\varepsilon_1^*,...,\varepsilon_n^*)^T$ and ξ^* by drawing from $\widehat{\tau}_1,...,\widehat{\tau}_n$ with replacement; calculate $\mathscr{Y}^*=\mathscr{X}\widehat{\beta}+\varepsilon^*$, $\mathscr{Y}_f^*=\mathscr{X}_f^T\widehat{\beta}+\xi^*$ and $\widehat{\beta}^*=(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\mathscr{Y}^*$; derive the prediction root $\delta_b^*=\mathscr{Y}_f^*-\mathscr{X}_f^T\widehat{\beta}^*$.

3. Repeat 2 for $b=1,2,...,B$, and calculate the $1-\alpha$ unadjusted sample quantile

(denoted as $\widehat{c}^*_{1-\alpha}$) of $|\delta^*_b|$, $b = 1,2,...,B$.

Find the quantile adjustment

4. Generate i.i.d. $e^* = (e^*_1,...,e^*_n)^T$ by drawing from $\widehat{\tau}_1,...,\widehat{\tau}_n$ with replacement, then derive $\mathscr{Y}^\dagger = \mathscr{X}\widehat{\beta} + e^*$, $\widehat{\beta}^\dagger = (\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\mathscr{Y}^\dagger$. Then define $\widehat{\zeta}^*_i = \mathscr{X}_f^T\widehat{\beta} + \widehat{\tau}_i - \mathscr{X}_f^T\widehat{\beta}^\dagger + \frac{1}{n}\sum_{j=1}^n e^*_j$ for $i = 1,2,...,n$. Calculate

$$p^*_{b_1} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{1}_{|\widehat{\zeta}^*_i| \le \widehat{c}^*_{1-\alpha}} - \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{1}_{|e^*_i| \le \widehat{c}^*_{1-\alpha}} \tag{4.23}$$

5. Repeat step 4 for $b_1 = 1,2,...,\mathscr{B}_1$, then calculate the $1-\gamma$ sample quantile (denoted as $\widehat{d}^*_{1-\gamma}$) of $p^*_{b_1}, b_1 = 1,2,...,\mathscr{B}_1$.

Calibrate the prediction interval

6. Calculate $\widehat{c}^*_{1-\alpha+\widehat{d}^*_{1-\gamma}/\sqrt{n}}$, the $1-\alpha+\widehat{d}^*_{1-\gamma}/\sqrt{n}$ sample quantile of $|\delta^*_b|$, $b = 1,2,...,B$

7. The prediction interval with $1-\alpha$ coverage probability and $1-\gamma$ guarantee level is given by the set

$$\left\{x \in \mathbf{R} : |x - \mathscr{X}_f^T\widehat{\beta}| \le \widehat{c}^*_{1-\alpha+\widehat{d}^*_{1-\gamma}/\sqrt{n}}\right\}. \tag{4.24}$$

Remark 16. *This remark explains why step 4 and 5 in RBUG/PRBUG simulates $\widetilde{M}_m$. Suppose we use RBUG. Then*

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{1}_{|\widehat{\zeta}^*_i| \le x} &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{1}_{-x+\mathscr{X}_f^T(\widehat{\beta}^\dagger-\widehat{\beta})-\frac{1}{n}\sum_{j=1}^n e^*_j \le \widehat{\tau}_i \le x+\mathscr{X}_f^T(\widehat{\beta}^\dagger-\widehat{\beta})-\frac{1}{n}\sum_{j=1}^n e^*_j} \\ &= \sqrt{n}\left(\widehat{F}\left(x+\mathscr{X}_f^T(\widehat{\beta}^\dagger-\widehat{\beta})-\frac{1}{n}\sum_{j=1}^n e^*_j\right) - \widehat{F}^-\left(-x+\mathscr{X}_f^T(\widehat{\beta}^\dagger-\widehat{\beta})-\frac{1}{n}\sum_{j=1}^n e^*_j\right)\right)\end{aligned} \tag{4.25}$$

so $p^*_{b_1}$ *equals*

$$\begin{aligned}
&\sqrt{n}\left(\widehat{F}\left(\widehat{c}^*_{1-\alpha}+\mathscr{X}_f^T(\widehat{\beta}^\dagger-\widehat{\beta})-\frac{1}{n}\sum_{j=1}^n e^*_j\right)-\widehat{F}(\widehat{c}^*_{1-\alpha})\right)\\
&\qquad -\frac{1}{\sqrt{n}}\sum_{j=1}^n\left(\mathbf{1}_{e^*_j\le\widehat{c}^*_{1-\alpha}}-\widehat{F}(\widehat{c}^*_{1-\alpha})\right)\\
&-\sqrt{n}\left(\widehat{F}^-\left(-\widehat{c}^*_{1-\alpha}+\mathscr{X}_f^T(\widehat{\beta}^\dagger-\widehat{\beta})-\frac{1}{n}\sum_{j=1}^n e^*_j\right)-\widehat{F}^-(-\widehat{c}^*_{1-\alpha})\right)\\
&\qquad +\frac{1}{\sqrt{n}}\sum_{j=1}^n\left(\mathbf{1}_{e^*_j<-\widehat{c}^*_{1-\alpha}}-\widehat{F}^-(-\widehat{c}^*_{1-\alpha})\right)
\end{aligned}\tag{4.26}$$

which simulates $\widetilde{M}_m\left(\frac{x+m}{2m}\right)-\widetilde{M}^-_m\left(\frac{-x+m}{2m}\right)$ *in the bootstrap world. The same discussion applies to PRBUG as well.*

We focus on proving *RBUG*'s validity, i.e., that prediction interval (4.24) satisfies definition 5. Define the simulated stochastic process

$$\begin{aligned}
\widehat{\mathscr{M}}(x)=\sqrt{n}\widehat{F}\left(x+\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^Te^*-\frac{1}{n}\sum_{j=1}^n e^*_j\right)-\frac{1}{\sqrt{n}}\sum_{j=1}^n\mathbf{1}_{e^*_j\le x}\\
\text{and } \widehat{\mathscr{S}}(x)=\widehat{\mathscr{M}}(x)-\widehat{\mathscr{M}^-}(-x)
\end{aligned}\tag{4.27}$$

and the quantiles

$$\begin{aligned}
c^*_{1-\alpha}&=\inf\left\{x\in\mathbf{R}:\mathbf{P}^*\left(|\mathscr{Y}^*_f-\mathscr{X}_f^T\widehat{\beta}^*|\le x\right)\ge 1-\alpha\right\}\\
\text{and } d^*_{1-\gamma}(x)&=\inf\left\{z\in\mathbf{R}:\mathbf{P}^*\left(\widehat{\mathscr{S}}(x)\le z\right)\ge 1-\gamma\right\}
\end{aligned}\tag{4.28}$$

See algorithm 5 for the meaning of the notations. Denote

$$c^*(1-\alpha,1-\gamma)=c^*_{1-\alpha+d^*_{1-\gamma}(c^*_{1-\alpha})/\sqrt{n}}\tag{4.29}$$

From theorem 1.2.1 of Politis et al. [1999], $\widehat{c}^*_{1-\alpha+\widehat{d}^*_{1-\gamma}/\sqrt{n}}$ converges to $c^*(1-\alpha,1-\gamma)$ almost surely as $B,\mathscr{B}_1\to\infty$. Therefore, the theoretical justification only focuses on $c^*(1-\alpha,1-\gamma)$.

Theorem 10. *Consider the RBUG algorithm, i.e, algorithm 5 with* $\widehat{\tau} = \widehat{\varepsilon}$ *as in* (4.13)*. Suppose assumption 1 to 4 hold true. Then, for any given* $0 < \alpha, \gamma < 1, \delta > 0$*,*

$$\begin{aligned} \mathbf{P}\left(|\mathbf{P}^*\left(|\mathscr{Y}_f - \mathscr{X}_f^T\widehat{\beta}| \le c^*(1-\alpha, 1-\gamma)\right) - (1-\alpha)| \le \delta\right) &\to 1 \\ \mathbf{P}\left(\{\mathbf{P}^*\left(|\mathscr{Y}_f - \mathscr{X}_f^T\widehat{\beta}| \le c^*(1-\alpha, 1-\gamma)\right) \ge 1-\alpha\}\right) &\to 1-\gamma \end{aligned} \tag{4.30}$$

In other words, *RBUG* is able to generate a prediction interval with desired asymptotic coverage probability and guarantee level.

Corollary 2 proves the validity of *PRBUG*. In corollary 2, we choose $\widehat{\tau} = \widehat{r}$ in algorithm 5 and define $C^*_{1-\alpha} = \inf\left\{x \in \mathbf{R} : \mathbf{P}^*(|\mathscr{Y}_f^* - \mathscr{X}_f^T\widehat{\beta}^*| \le x) \ge 1-\alpha\right\}$;
$D^*_{1-\gamma}(x) = \inf\left\{z \in \mathbf{R} : \mathbf{P}^*\left(\widehat{\mathscr{S}}(x) \le z\right) \ge 1-\gamma\right\}$. We define
$C^*(1-\alpha, 1-\gamma) = C^*_{1-\alpha+D^*_{1-\gamma}(C^*_{1-\alpha})/\sqrt{n}}$. That is, $C^*_{1-\alpha}$, $D^*_{1-\gamma}(x)$ and $C^*(1-\alpha, 1-\gamma)$ play the same roles as $c^*_{1-\alpha}$, $d^*_{1-\gamma}$ and $c^*(1-\alpha, 1-\gamma)$. The only reason for using another set of notations is that we change the sampling mechanism (i.e., replace $\widehat{\varepsilon}$ in algorithm 5 by $\widehat{r}$).

Corollary 2. *Consider the PRBUG algorithm, i.e, algorithm 5 with* $\widehat{\tau} = \widehat{r}$*. Suppose assumptions 1 to 4 hold true. Then, for any given* $0 < \alpha, \gamma < 1, \delta > 0$*,*

$$\begin{aligned} \mathbf{P}\left(|\mathbf{P}^*\left(|\mathscr{Y}_f - \mathscr{X}_f^T\widehat{\beta}| \le C^*(1-\alpha, 1-\gamma)\right) - (1-\alpha)| \le \delta\right) &\to 1 \\ \mathbf{P}\left(\{\mathbf{P}^*\left(|\mathscr{Y}_f - \mathscr{X}_f^T\widehat{\beta}| \le C^*(1-\alpha, 1-\gamma)\right) \ge 1-\alpha\}\right) &\to 1-\gamma. \end{aligned} \tag{4.31}$$

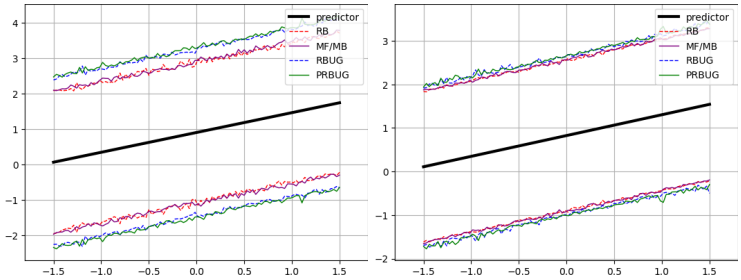
Remark 17. *Similar to residual-based bootstrap and MF / MB bootstrap, section 4.7 shows that PRBUG tends to generate a wider, and of higher guarantee level, prediction interval than RBUG.*

4.7 Numerical justification

This section applies numerical simulations to demonstrate the finite sample performance of *RBUG/PRBUG*. The alternatives are the residual-based bootstrap(RB), the *MF/MB* boot-

strap(*MF/MB*), the split conformal prediction defined in Lei et al. [2018] and Vovk's tolerance region (Vovk [2012]). The classical conformal prediction algorithm (e.g., Vovk et al. [2005]) assumed $\mathscr{X}_f$ is random, which is unsuitable for our setting. Vovk's tolerance region yields a prediction interval satisfying eq. (4.4) but not condition (4.3). Lei et al. [2018] showed that the split conformal prediction could generate an asymptotic valid prediction interval when $\mathscr{X}_f$ is fixed, which coincides with our setting.

Figure 4.1 plots point-wise prediction intervals for the linear model $\mathscr{Y}_i = 0.8 + 0.5\mathscr{X}_i + \varepsilon_i$. I.i.d. residuals are generated by normal distribution with mean 0 and variance 1. When the sample size is small, the prediction intervals generated by RBUG / PRBUG is significantly wider than the prediction intervals generated by classical bootstrap methods. On the other hand, when the sample size is large, the prediction intervals generated by different algorithms coincide with each other.



(a) Sample size = 80 **(b)** Sample size = 400

Figure 4.1. Predictors and point-wise prediction intervals for the linear model $\mathscr{Y}_i = 0.8 + 0.5\mathscr{X}_i + \varepsilon_i,\ i = 1,2,....$ The prediction intervals are generated by the following methods: RB for the residual-based bootstrap(Stine [1985]); MF/MB for the model-free / model-based bootstrap(Politis [2015]); RBUG and PRBUG for algorithm 5. We choose the nominal coverage probability $1-\alpha = 95\%$ and the nominal guarantee level(in RBUG / PRBUG) $1-\gamma = 90\%$.

Our linear model of choice is denoted as the **Experiment model** and defined as follows: $\mathscr{Y} = \mathscr{X}\beta + \varepsilon$, and β's dimension is 8. $\beta = (\beta_0, \beta_1, ..., \beta_7)^T$ with $\beta_0 = 1.0,\ \beta_1 = 0.5,\ \beta_2 = -1.0,\ \beta_3 = -0.5,\ \beta_4 = 1.5,\ \beta_5 = -1.5$ and $\beta_i = 0$ for $i \geq 6$. The design matrix $\mathscr{X}$ is generated by i.i.d. standard normal random variables, and is fixed in each experiment. The new regressor $\mathscr{X}_f = (\mathscr{X}_{f,0}, ..., \mathscr{X}_{f,7})^T$ is given by $\mathscr{X}_{f,i} = 0.1 \times i, i = 0, 1, ..., 7$. The i.i.d. error vector ε is

generated by various distributions. We choose the sample size $n = 50, 100, 200, 400, 1200$. The result is demonstrated in table 4.2, table 4.3, figure 4.2 and figure 4.3.

When the sample size is small (e.g. 50 or 100 in the example), the *MF/MB bootstrap* alleviates the residual-based bootstrap's under-coverage nature. Therefore, it has a higher guarantee level than the residual-based bootstrap. Yet this modification **does not** change the asymptotic guarantee level (in other words, the *MF/MB* bootstrap still has 50% asymptotic guarantee level). The split conformal prediction also has a high guarantee level when the sample size is small and a low guarantee level when the sample size is moderate or large. Vovk's tolerance region has the desired guarantee level when the sample size is large. However, when the sample size is small (e.g., 50 or 100), the tolerance region is always too wide. On the other hand, the *RBUG* and the *PRBUG* algorithms improve the residual-based bootstrap's performance by controlling the asymptotic guarantee level. *PRBUG* reaches the desired guarantee levels when the sample size is moderate, while *RBUG* needs a large number of data in order to achieve the desired guarantee level. So we recommend using *PRBUG* in practice. When the sample size is large, the bootstrap algorithms' conditional coverage probabilities are close to 95%, and the adjustments made by *RBUG / PRBUG* are not significant.

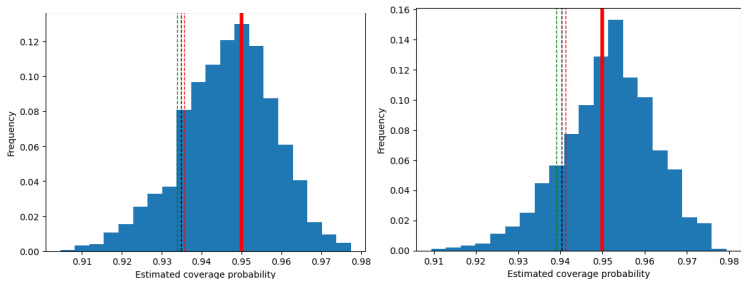
In practice, our work can be particularly useful when the sample size n is not very large. Suppose we use the residual-based bootstrap. In table 4.2 we see that the 15% quantile of conditional coverage probabilities is 91.0% when the sample size is 100, which means 15% of the nominal 95% prediction intervals' conditional coverage probabilities are less than 91%. On the other hand, the *RBUG*'s 15% quantile is 93.5% and the *PRBUG*'s 15% quantile is 95.5%, which is significantly larger than the residual-based bootstrap's quantile.

4.8 Conclusion

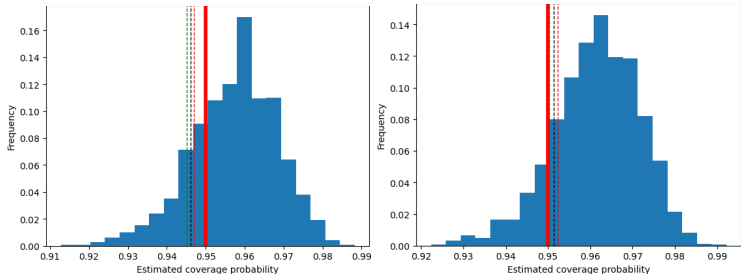
Focusing on the fixed dimensional linear model, in this paper we derive the asymptotic distribution of the difference between the conditional coverage probability of a nominal prediction

Table 4.2. Performance of different algorithms on the **Experiment model**. The nominal coverage probability is 95% and the nominal guarantee level is 85% (we also choose $\gamma = 15\%$ in (4.5)). The residuals are generated by normal random variables with mean 0 and variance 1. In the 'Algorithm' column, 'RB' means residual-based bootstrap; '*MF/MB*' means *MF/MB* bootstrap; 'split-conformal' means the split conformal prediction(defined in Lei et al. [2018]), Vovk's tolerance region was defined in remark 13, and *RBUG / PRBUG* mean algorithm 5. We use the R package maintained by Tibshirani et al. [2021] to perform the split conformal prediction. 'Length' represents the average length of the prediction interval. The number of bootstrap replicates is $B = 3000$, the number of replicates to find quantile's adjustment is $\mathscr{B}_1 = 3000$. The result is generated by 1500 simulations. In table 4.2, 'Quantiles' represents the quantiles of conditional coverage probabilities, and 'Guarantee' represents the guarantee level.

Sample size	Algorithm	Quantiles			Guarantee	Length
		15%	30%	50%		
50	RB	87.8%	90.2%	92.4%	21.1%	3.63
	MF / MB	93.8%	95.5%	96.9%	75.6%	4.40
	split-conformal	95.9%	97.8%	99.0%	89.2%	5.65
	Vovk's region	95.9%	98.0%	99.1%	89.2%	5.69
	RBUG	91.2%	93.7%	95.7%	57.9%	4.19
	PRBUG	95.8%	97.3%	98.5%	90.3%	5.02
100	RB	91.0%	92.6%	93.9%	29.0%	3.78
	MF / MB	93.7%	94.9%	95.9%	69.3%	4.14
	split-conformal	95.1%	96.7%	98.0%	86.0%	4.78
	Vovk's region	97.3%	98.5%	99.2%	96.5%	5.55
	RBUG	93.5%	94.9%	96.1%	68.1%	4.22
	PRBUG	95.5%	96.6%	97.6%	89.1%	4.58
200	RB	92.5%	93.4%	94.3%	34.7%	3.83
	MF / MB	93.7%	94.5%	95.3%	58.9%	4.00
	split-conformal	93.6%	94.9%	96.0%	69.8%	4.19
	Vovk's region	95.9%	97.0%	97.9%	92.7%	4.69
	RBUG	94.2%	95.0%	95.8%	71.0%	4.12
	PRBUG	95.1%	95.9%	96.7%	87.5%	4.29
400	RB	93.5%	94.1%	94.7%	41.3%	3.88
	MF / MB	94.0%	94.6%	95.2%	58.3%	3.96
	split-conformal	93.7%	94.7%	95.5%	65.2%	4.05
	Vovk's region	96.1%	96.8%	97.5%	95.5%	4.50
	RBUG	94.6%	95.3%	95.9%	75.5%	4.08
	PRBUG	95.1%	95.7%	96.2%	87.5%	4.16
1200	RB	94.0%	94.5%	94.9%	47.9%	3.91
	MF / MB	94.2%	94.7%	95.1%	55.8%	3.94
	split-conformal	94.1%	94.6%	95.2%	60.5%	3.97
	Vovk's region	95.1%	95.6%	96.2%	88.1%	4.15
	RBUG	94.7%	95.1%	95.6%	76.7%	4.03
	PRBUG	94.9%	95.3%	95.7%	81.0%	4.05



(a) Residual-based bootstrap **(b)** *MF / MB* bootstrap

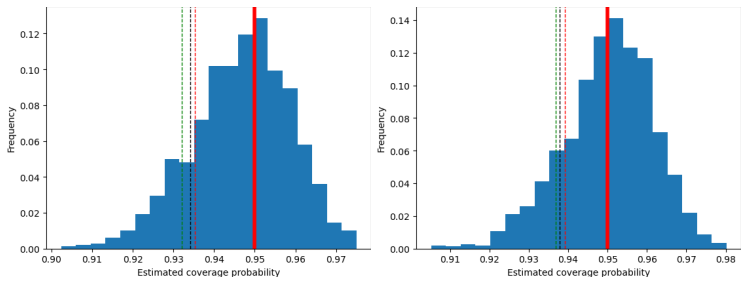


(c) *RBUG* **(d)** *PRBUG*

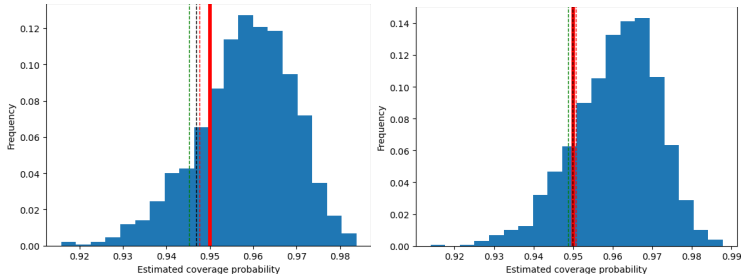
Figure 4.2. Histograms for the conditional coverage probabilities. Here we use the **Experiment model** with sample size 400. The residuals are generated by i.i.d. normal random variables with mean 0 and variance 1. The solid red line is the nominal coverage probability(95%); the green, black and red dashed lines respectively represents the 13%, 15%, 17% quantile of conditional coverage probabilities. In order to have a $1-\gamma = 85\%$ guarantee level, the solid red line should be close to the black dashed line.

Table 4.3. Performance of different algorithms on the **Experiment model**. The nominal coverage probability is 95%, and the nominal guarantee level is 85%. The residuals are generated by the Laplace distribution with mean 0 and scale $1/\sqrt{2}$; which makes the residuals' variance 1.

Sample size	Algorithm	Quantiles			Guarantee	Length
		15%	30%	50%		
50	RB	87.7%	90.4%	92.4%	23.3%	3.83
	MF / MB	92.2%	94.1%	95.5%	58.7%	4.60
	split-conformal	94.4%	96.6%	98.1%	82.1%	6.14
	Vovk's region	94.6%	96.6%	98.2%	82.7%	6.15
	RBUG	91.1%	93.6%	95.7%	57.3%	4.75
	PRBUG	94.5%	96.1%	97.6%	81.4%	5.67
100	RB	91.0%	92.6%	93.9%	33.9%	4.03
	MF / MB	92.9%	94.2%	95.3%	57.1%	4.41
	split-conformal	94.5%	96.1%	97.3%	81.5%	5.30
	Vovk's region	96.8%	98.0%	98.9%	94.6%	6.68
	RBUG	93.6%	95.1%	96.4%	72.1%	4.80
	PRBUG	94.9%	96.2%	97.3%	84.7%	5.21
200	RB	92.6%	93.6%	94.5%	37.7%	4.13
	MF / MB	93.4%	94.4%	95.2%	54.7%	4.32
	split-conformal	93.3%	94.6%	95.6%	63.9%	4.54
	Vovk's region	95.6%	96.7%	97.6%	91.0%	5.37
	RBUG	94.5%	95.3%	96.1%	76.7%	4.64
	PRBUG	95.1%	95.8%	96.6%	86.0%	4.83
400	RB	93.4%	94.1%	94.8%	42.6%	4.18
	MF / MB	93.8%	94.5%	95.1%	53.7%	4.28
	split-conformal	93.5%	94.5%	95.4%	60.4%	4.40
	Vovk's region	95.9%	96.6%	97.3%	94.5%	5.16
	RBUG	94.7%	95.3%	95.9%	78.6%	4.53
	PRBUG	95.0%	95.6%	96.2%	84.5%	4.63
1200	RB	94.1%	94.5%	94.9%	47.9%	4.23
	MF / MB	94.2%	94.6%	95.0%	52.5%	4.26
	split-conformal	93.9%	94.5%	95.2%	57.1%	4.28
	Vovk's region	95.1%	95.6%	96.1%	86.4%	4.61
	RBUG	94.8%	95.2%	95.6%	77.6%	4.42
	PRBUG	94.9%	95.3%	95.7%	81.8%	4.45



(a) Residual-based bootstrap **(b)** *MF / MB* bootstrap



(c) *RBUG* **(d)** *PRBUG*

Figure 4.3. Histograms for the conditional coverage probabilities of the **Experiment model**. The sample size is 400 and the residuals are generated by i.i.d. Laplace random variables with mean 0 and the scale parameter $1/\sqrt{2}$, which makes the variance 1. The meaning of lines coincide with figure 4.2.

interval $\mathbf{P}^*\left(|\mathscr{Y}_f - \mathscr{X}_f^T\widehat{\beta}| \leq x\right)$ and the conditional coverage probability of a prediction interval for residual-based bootstrapped observations $\mathbf{P}^*\left(|\mathscr{Y}_f^* - \mathscr{X}_f^T\widehat{\beta}^*| \leq x\right)$. According to this result, the prediction interval generated by residual-based bootstrap has approximately 50% probability to yield conditional under-coverage.

We then develop a new bootstrap algorithm that generates prediction intervals with arbitrarily assigned conditional coverage probability and guarantee level, and prove its asymptotic validity. Our theoretical results are corroborated by several finite-sample simulations.

Residual-based and the MF/MB bootstrap are widely used for prediction in numerous settings like nonparametric/nonlinear regression, quantile regression, time series analysis (regression with dependent errors, autoregression, etc.), and others. We expect our ideas to be applicable in those settings as well; future work will address the details. Furthermore, the case of high-dimensional linear regression is of current interest, i.e., where p is allowed to diverge as $n \to \infty$; this can also be the subject of future work.

4.9 Acknowledgement

Chapter 4 is based on the paper “Bootstrap prediction intervals with asymptotic conditional validity and unconditional guarantees” by Y.Zhang and D.N. Politis and has been accepted for publication in *Information and Inference: A journal of the IMA*. The dissertation author was the primary investigator and author of this paper.

Appendix A

Proofs of theorems in chapter 2

In this chapter, 'assumption 1' to 'assumption 9' represent assumption 1 to 9 in section 2.3 and section 2.6.

A.1 Some important lemmas

This section introduces three useful lemmas. Lemma A.1.1 comes from Whittle [1960], which directly contributes to the model selection consistency. Lemma A.1.2 and A.1.3 are similar to Chernozhukov et al. [2013], they used a joint normal distribution to approximate the distribution of linear combinations of independent random variables.

Lemma A.1.1. *Suppose random variables* $\varepsilon_1, ..., \varepsilon_n$ *are i.i.d.,* $\mathbf{E}\varepsilon_1 = 0$*, and* $\exists$ *a constant* $m > 0$ *such that* $\mathbf{E}|\varepsilon_1|^m < \infty$*. In addition suppose the matrix* $\Gamma = (\gamma_{ij})_{i=1,2,...,k, j=1,2,...,n}$ *satisfies*

$$\max_{i=1,2,...,k} \sum_{j=1}^{n} \gamma_{ij}^2 \leq D,\ D > 0 \tag{A.1.1}$$

Then $\exists$ *a constant E which only depends on m and* $\mathbf{E}|\varepsilon_1|^m$ *such that for* $\forall \delta > 0$*,*

$$Prob\left(\max_{i=1,2,...,k} |\sum_{j=1}^{n} \gamma_{ij}\varepsilon_j| > \delta \right) \leq \frac{kED^{m/2}}{\delta^m} \tag{A.1.2}$$

Proof. From theorem 2 in Whittle [1960], for any $i = 1,2,...,k$,

$$\begin{aligned} Prob\left(|\sum_{j=1}^{n}\gamma_{ij}\varepsilon_j| > \delta\right) &\leq \frac{\mathbf{E}|\sum_{j=1}^{n}\gamma_{ij}\varepsilon_j|^m}{\delta^m} \\ \leq \frac{2^m C(m)\mathbf{E}|\varepsilon_1|^m(\sum_{j=1}^{n}\gamma_{ij}^2)^{m/2}}{\delta^m} &\leq \frac{2^m C(m)\mathbf{E}|\varepsilon_1|^m D^{m/2}}{\delta^m} \end{aligned} \tag{A.1.3}$$

Here $C(m)$ is a constant depending on m. Choose $E = 2^m C(m)\mathbf{E}|\varepsilon_1|^m$,

$$Prob\left(\max_{i=1,2,...,k}|\sum_{j=1}^{n}\gamma_{ij}\varepsilon_j| > \delta\right) \leq \sum_{i=1}^{k} Prob\left(|\sum_{j=1}^{n}\gamma_{ij}\varepsilon_j| > \delta\right) \leq \frac{kED^{m/2}}{\delta^m} \tag{A.1.4}$$

□

Lemma A.1.2. *Suppose $\varepsilon = (\varepsilon_1,...,\varepsilon_n)^T$ are joint normal random variables with mean $\mathbf{E}\varepsilon = 0$, non-singular covariance matrix $\mathbf{E}\varepsilon\varepsilon^T$, and positive marginal variance $\sigma_i^2 = \mathbf{E}\varepsilon_i^2 > 0$, $i = 1,2,...,n$. In addition, suppose $\exists$ two constants $0 < c_0 \leq C_0 < \infty$ such that $c_0 \leq \sigma_i \leq C_0$ for $i = 1,2,...,n$. Then for any given $\delta > 0$,*

$$\begin{aligned} \sup_{x\in\mathbf{R}}\left(Prob(\max_{i=1,2,...,n}|\varepsilon_i| \leq x+\delta) - Prob(\max_{i=1,2,...,n}|\varepsilon_i| \leq x)\right) \\ \leq C\delta(\sqrt{\log(n)} + \sqrt{|\log(\delta)|} + 1) \end{aligned} \tag{A.1.5}$$

C only depends on c_0 and C_0.

Proof of lemma A.1.2. First for any $i = 1,2,...,n$,

$$|\varepsilon_i| = \max(\varepsilon_i, -\varepsilon_i) \Rightarrow \max_{i=1,...,n}|\varepsilon_i| = \max(\max_{i=1,...,n}\varepsilon_i, \max_{i=1,...,n}-\varepsilon_i) \tag{A.1.6}$$

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

$$\begin{aligned}
&Prob(\max_{i=1,2,\ldots,n}|\varepsilon_i| \le x+\delta) - Prob(\max_{i=1,2,\ldots,n}|\varepsilon_i| \le x) \\
&= Prob(0 < \max(\max_{i=1,\ldots,n}\varepsilon_i, \max_{i=1,\ldots,n}-\varepsilon_i) - x \le \delta) \\
&\le Prob(0 < \max_{i=1,\ldots,n}\varepsilon_i - x \le \delta) + Prob(0 < \max_{i=1,\ldots,n}-\varepsilon_i - x \le \delta) \\
&\le Prob(|\max_{i=1,\ldots,n}\varepsilon_i - x| \le \delta) + Prob(|\max_{i=1,\ldots,n}-\varepsilon_i - x| \le \delta)
\end{aligned} \tag{A.1.7}$$

$-\varepsilon$ is also joint normal with mean 0 and marginal variance $\mathbf{E}(-\varepsilon_j)^2 = \sigma_j^2$. From theorem 3 and (18), (19) in Chernozhukov et al. [2015], by defining $\underline{\sigma} = \min_{i=1,2,\ldots,n}\sigma_i \le \max_{i=1,2,\ldots,n}\sigma_i = \overline{\sigma}$, we have

$$\begin{aligned}
\sup_{x\in\mathbf{R}} Prob\left(|\max_{i=1,2,\ldots,n}\varepsilon_i - x| \le \delta\right) &\le \frac{\sqrt{2}\delta}{\underline{\sigma}}\left(\sqrt{\log(n)} + \sqrt{\max(1, \log(\underline{\sigma}) - \log(\delta))}\right) \\
&\quad + \frac{4\sqrt{2}\delta}{\underline{\sigma}} \times \left(\frac{\overline{\sigma}}{\underline{\sigma}}\sqrt{\log(n)} + 2 + \frac{\overline{\sigma}}{\underline{\sigma}}\sqrt{\max(0, \log(\underline{\sigma}) - \log(\delta))}\right) \\
&\le \frac{\sqrt{2}\delta}{c_0}\left(\sqrt{\log(n)} + \sqrt{1 + |\log(c_0)| + |\log(C_0)|} + \sqrt{|\log(\delta)|}\right) \\
&\quad + \frac{4\sqrt{2}\delta C_0}{c_0^2}\left(\sqrt{\log(n)} + 2 + \sqrt{|\log(c_0)| + |\log(C_0)|} + \sqrt{|\log(\delta)|}\right) \\
&\le \left(\frac{\sqrt{2\times(1 + |\log(c_0)| + |\log(C_0)|)}}{c_0} + \frac{4\sqrt{2}C_0}{c_0^2}(2 + \sqrt{|\log(c_0)| + |\log(C_0)|})\right) \\
&\quad \times \delta\left(\sqrt{\log(n)} + 1 + \sqrt{|\log(\delta)|}\right)
\end{aligned} \tag{A.1.8}$$

Choose $C = \frac{\sqrt{2\times(1+|\log(c_0)|+|\log(C_0)|)}}{c_0} + \frac{4\sqrt{2}C_0}{c_0^2}(2 + \sqrt{|\log(c_0)| + |\log(C_0)|})$, which only depends on c_0, C_0. Then

$$\begin{aligned}
&\sup_{x\in\mathbf{R}}(Prob(\max_{i=1,2,\ldots,n}|\varepsilon_i| \le x+\delta) - Prob(\max_{i=1,2,\ldots,n}|\varepsilon_i| \le x)) \\
&\le 2C\delta(1 + \sqrt{\log(n)} + \sqrt{|\log(\delta)|})
\end{aligned} \tag{A.1.9}$$

□

Lemma A.1.3. *Suppose* $\varepsilon = (\varepsilon_1, ...\varepsilon_n)^T$ *are i.i.d. random variables with* $\mathbf{E}\varepsilon_1 = 0$, $\mathbf{E}\varepsilon_1^2 = \sigma^2$ *and* $\mathbf{E}|\varepsilon_1|^3 < \infty$. $\Gamma = (\gamma_{ij})_{i=1,2,...,n, j=1,2,...,k}$ *is an* $n \times k$ $(1 \le k \le n)$ *rank k matrix. And* $\exists$ *constants* $0 < c_\Gamma \le C_\Gamma < \infty$ *such that* $c_\Gamma^2 \le \sum_{j=1}^n \gamma_{ji}^2 \le C_\Gamma^2$ *for* $i = 1, 2, ..., k$. $\widehat{\sigma}^2 = \widehat{\sigma}^2(\varepsilon)$ *is an estimator of* σ^2 *and random variables* $\varepsilon^*|\varepsilon = (\varepsilon_1^*, ..., \varepsilon_n^*)^T|\varepsilon$ *are i.i.d. with* ε_1^* *having normal distribution* $\mathcal{N}(0, \widehat{\sigma}^2)$. $\frac{\varepsilon_i^*}{\widehat{\sigma}}$ *is independent of* ε *for* $i = 1, 2, ..., n$. *In addition, suppose one of the following conditions:*

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

$$\begin{aligned} |\sigma^2 - \widehat{\sigma}^2| &= O_p(n^{-\alpha_\sigma}) \text{ and} \\ \max_{j=1,2,...,n,\ i=1,2,...,k} |\gamma_{ji}| &= o(\min(n^{(\alpha_\sigma - 1)/2} \times \log^{-3/2}(n),\ n^{-1/3} \times \log^{-3/2}(n)) \end{aligned} \tag{A.1.10}$$

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

$$\begin{aligned} |\sigma^2 - \widehat{\sigma}^2| &= O_p(n^{-\alpha_\sigma}),\ k = o(n^{\alpha_\sigma} \times \log^{-3}(n)), \\ \max_{j=1,...,n, i=1,...,k} |\gamma_{ji}| &= O(n^{-\alpha_\sigma} \times \log^{-3/2}(n)) \end{aligned} \tag{A.1.11}$$

Then we have

$$\sup_{x \in [0,\infty)} |Prob(\max_{i=1,2,...,k} |\sum_{j=1}^n \gamma_{ji}\varepsilon_j| \le x) - Prob^*(\max_{i=1,2,...,k} |\sum_{j=1}^n \gamma_{ji}\varepsilon_j^*| \le x)| = o_P(1) \tag{A.1.12}$$

In particular, if $\widehat{\sigma} = \sigma$, *by assuming one of the following conditions,*

C_1'.

$$\max_{j=1,2,...,n, i=1,2,...,k} |\gamma_{ji}| = o(n^{-1/3} \times \log^{-3/2}(n)) \tag{A.1.13}$$

C_2'.

$$k \times \max_{j=1,2,...,n, i=1,2,...,k} |\gamma_{ji}| = o(\log^{-9/2}(n)) \tag{A.1.14}$$

Then we have

$$\sup_{x\in[0,\infty)} |Prob(\max_{i=1,2,...,k} |\sum_{j=1}^{n} \gamma_{ji}\varepsilon_j| \le x) - Prob(\max_{i=1,2,...,k} |\sum_{j=1}^{n} \gamma_{ji}\varepsilon_j^*| \le x)| = o(1) \tag{A.1.15}$$

Proof of lemma A.1.3. In this proof we define $\Gamma = (\gamma_1, ..., \gamma_k)$ with $\gamma_i = (\gamma_{1i}, \gamma_{2i}, ..., \gamma_{ni})^T \in \mathbf{R}^n$. For $i = 1, 2, ..., k$, $\gamma_i^T \varepsilon = \sum_{j=1}^n \gamma_{ji}\varepsilon_j$. From lemma A.2 and (8) in Chernozhukov et al. [2013], and (S1) to (S5) in Xu et al. [2019], for $x = (x_1, ..., x_n)$ and $y, z \in \mathbf{R}$, define

$$\begin{gathered} F_\beta(x) = \frac{1}{\beta}\log\left(\sum_{i=1}^{n}\exp(\beta x_i)\right),\ g_0(y) = (1-\min(1,\max(y,0))^4)^4, \\ g_{\psi,z}(y) = g_0(\psi(y-z)) \end{gathered} \tag{A.1.16}$$

Here $\beta, \psi > 0$. Then $g_{\psi,z} \in \mathbf{C}^3$ is nonincreasing function. $g_0 = 1$ with $y \le 0$, 0 with $y \ge 1$, and

$$\begin{gathered} g_* = \max_{y\in\mathbf{R}}(|g_0^{'}(y)| + |g_0^{''}(y)| + |g_0^{'''}(y)|) < \infty,\ \mathbf{1}_{y\le z} \le g_{\psi,z}(y) \le \mathbf{1}_{y\le z+\psi^{-1}} \\ \sup_{y,z\in\mathbf{R}} |g_{\psi,z}^{'}(y)| \le g_*\psi,\ \sup_{y,z\in\mathbf{R}} |g_{\psi,z}^{''}(y)| \le g_*\psi^2,\ \sup_{y,z\in\mathbf{R}} |g_{\psi,z}^{'''}(y)| \le g_*\psi^3 \\ \frac{\partial F_\beta}{\partial x_i} = \frac{\exp(\beta x_i)}{\sum_{j=1}^n \exp(\beta x_j)} \Rightarrow \frac{\partial F_\beta}{\partial x_i} \ge 0,\ \sum_{i=1}^{n}\frac{\partial F_\beta}{\partial x_i} = 1 \\ \sum_{i=1}^{n}\sum_{j=1}^{n} |\frac{\partial^2 F_\beta}{\partial x_i \partial x_j}| \le 2\beta,\ \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} |\frac{\partial^3 F_\beta}{\partial x_i \partial x_j \partial x_k}| \le 6\beta^2 \\ F_\beta(x_1, ..., x_n) - \frac{\log(n)}{\beta} \le \max_{i=1,...,n} x_i \le F_\beta(x_1, ..., x_n) \end{gathered} \tag{A.1.17}$$

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

$G_\beta(x) = \frac{1}{\beta}\log(\sum_{i=1}^n \exp(\beta x_i) + \sum_{i=1}^n \exp(-\beta x_i)) = F_\beta(x_1, ..., x_n, -x_1, ..., -x_n)$. From (A.1.17)

and (A.1.6), for $i,j,k = 1,...,n$

$$
\begin{aligned}
G_\beta(x) - \frac{\log(2n)}{\beta} \leq \max_{i=1,...,n} |x_i| \leq G_\beta(x) \\
\frac{\partial G_\beta}{\partial x_i} = \frac{\partial F_\beta}{\partial x_i} - \frac{\partial F_\beta}{\partial x_{i+n}} \Rightarrow \sum_{i=1}^{n} |\frac{\partial G_\beta}{\partial x_i}| \leq \sum_{i=1}^{n} \frac{\partial F_\beta}{\partial x_i} + \frac{\partial F_\beta}{\partial x_{i+n}} = 1 \\
\frac{\partial^2 G_\beta}{\partial x_i \partial x_j} = \frac{\partial^2 F_\beta}{\partial x_i \partial x_j} - \frac{\partial^2 F_\beta}{\partial x_i \partial x_{j+n}} - \frac{\partial^2 F_\beta}{\partial x_{i+n} \partial x_j} + \frac{\partial^2 F_\beta}{\partial x_{i+n} \partial x_{j+n}} \\
\Rightarrow \sum_{i=1}^{n} \sum_{j=1}^{n} |\frac{\partial^2 G_\beta}{\partial x_i \partial x_j}| \leq \sum_{i=1}^{2n} \sum_{j=1}^{2n} |\frac{\partial^2 F_\beta}{\partial x_i \partial x_j}| \leq 2\beta \\
\frac{\partial^3 G_\beta}{\partial x_i \partial x_j \partial x_k} = \frac{\partial^3 F_\beta}{\partial x_i \partial x_j \partial x_k} - \frac{\partial^3 F_\beta}{\partial x_i \partial x_j \partial x_{k+n}} - \frac{\partial^3 F_\beta}{\partial x_i \partial x_{j+n} \partial x_k} \\
+ \frac{\partial^3 F_\beta}{\partial x_i \partial x_{j+n} \partial x_{k+n}} - \frac{\partial^3 F_\beta}{\partial x_{i+n} \partial x_j \partial x_k} + \frac{\partial^3 F_\beta}{\partial x_{i+n} \partial x_j \partial x_{k+n}} \\
+ \frac{\partial^3 F_\beta}{\partial x_{i+n} \partial x_{j+n} \partial x_k} - \frac{\partial^3 F_\beta}{\partial x_{i+n} \partial x_{j+n} \partial x_{k+n}} \\
\Rightarrow \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} |\frac{\partial^3 G_\beta}{\partial x_i \partial x_j \partial x_k}| \leq \sum_{i=1}^{2n} \sum_{j=1}^{2n} \sum_{k=1}^{2n} |\frac{\partial^3 F_\beta}{\partial x_i \partial x_j \partial x_k}| \leq 6\beta^2
\end{aligned}
\tag{A.1.18}
$$

Define $h_{\beta,\psi,x}(x_1,...,x_n) = g_{\psi,x}(G_\beta(x_1,...,x_n))$. Direct calculation shows

$\frac{\partial h_{\beta,\psi,x}(x_1,...,x_n)}{\partial x_i} = g'_{\psi,x}(G_\beta(x_1,...,x_n))\frac{\partial G_\beta}{\partial x_i} \Rightarrow \sum_{i=1}^n |\frac{\partial h_{\beta,\psi,x}(x_1,...,x_n)}{\partial x_i}| \leq g_*\psi;$

$$\begin{aligned}
\frac{\partial^2 h_{\beta,\psi,x}(x_1,...,x_n)}{\partial x_i \partial x_j} &= g''_{\psi,x}(G_\beta(x_1,...,x_n))\frac{\partial G_\beta}{\partial x_i}\frac{\partial G_\beta}{\partial x_j} \\
&\quad + g'_{\psi,x}(G_\beta(x_1,...,x_n))\frac{\partial^2 G_\beta}{\partial x_i \partial x_j} \\
\Rightarrow \sum_{i=1}^n\sum_{j=1}^n |\frac{\partial^2 h_{\beta,\psi,x}(x_1,...,x_n)}{\partial x_i \partial x_j}| &\leq g_*\psi^2\left(\sum_{i=1}^n |\frac{\partial G_\beta}{\partial x_i}|\right)^2 \\
&\quad + g_*\psi\sum_{i=1}^n\sum_{j=1}^n |\frac{\partial^2 G_\beta}{\partial x_i \partial x_j}| \leq g_*\psi^2 + 2g_*\psi\beta \\
\text{and } \frac{\partial^3 h_{\beta,\psi,x}(x_1,...,x_n)}{\partial x_i \partial x_j \partial x_k} &= g'''_{\psi,x}(G_\beta(x_1,...,x_n))\frac{\partial G_\beta}{\partial x_i}\frac{\partial G_\beta}{\partial x_j}\frac{\partial G_\beta}{\partial x_k} \\
&\quad + g''_{\psi,x}(G_\beta(x_1,...,x_n))\frac{\partial^2 G_\beta}{\partial x_i \partial x_k}\frac{\partial G_\beta}{\partial x_j} \\
&\quad + g''_{\psi,x}(G_\beta(x_1,...,x_n))\frac{\partial G_\beta}{\partial x_i}\frac{\partial^2 G_\beta}{\partial x_j \partial x_k} + g''_{\psi,x}(G_\beta(x_1,...,x_n))\frac{\partial^2 G_\beta}{\partial x_i \partial x_j}\frac{\partial G_\beta}{\partial x_k} \\
&\quad + g'_{\psi,x}(G_\beta(x_1,...,x_n))\frac{\partial^3 G_\beta}{\partial x_i \partial x_j \partial x_k} \\
\Rightarrow \sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n |\frac{\partial^3 h_{\beta,\psi,x}(x_1,...,x_n)}{\partial x_i \partial x_j \partial x_k}| &\leq g_*\psi^3\left(\sum_{i=1}^n |\frac{\partial G_\beta}{\partial x_i}|\right)^3 \\
&\quad + 3g_*\psi^2\left(\sum_{i=1}^n\sum_{j=1}^n |\frac{\partial^2 G_\beta}{\partial x_i \partial x_j}|\right)\times\left(\sum_{k=1}^n |\frac{\partial G_\beta}{\partial x_k}|\right) \\
+ g_*\psi\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n |\frac{\partial^3 G_\beta}{\partial x_i \partial x_j \partial x_k}| &\leq g_*\psi^3 + 6g_*\psi^2\beta + 6g_*\psi\beta^2
\end{aligned} \tag{A.1.19}$$

Define $\xi = (\xi_1,...,\xi_n)$ as i.i.d. random variables with the same marginal distribution as ε_1, and is independent of $\varepsilon, \varepsilon^*$. Therefore, $Prob(\max_{i=1,2,...,k}|\gamma_i^T\varepsilon| \leq x) = Prob^*(\max_{i=1,2,...,k}|\gamma_i^T\xi| \leq x)$ for any x. Since $c_\Gamma^2 \leq \mathbf{E}^*\left(\sum_{l=1}^n \frac{\gamma_{il}\varepsilon_l^*}{\hat{\sigma}}\right)^2 = \sum_{l=1}^n \gamma_{il}^2 \leq C_\Gamma^2$ for $i = 1,2,...,k$. According to (A.1.6), (A.1.18) and lemma A.1.2, $\exists$ a constant C which only depends on c_Γ and C_Γ such that for any

given $\psi, \beta, \widehat{\sigma} > 0$,

$$\begin{aligned}
&\sup_{x\in\mathbf{R}}\left(Prob^*\left(\max_{i=1,2,...,k}|\gamma_i^T\varepsilon^*|\le x+\frac{1}{\psi}+\frac{\log(2k)}{\beta}\right)-Prob^*\left(\max_{i=1,2,...,k}|\gamma_i^T\varepsilon^*|\le x\right)\right)\\
&\qquad=\sup_{x\in\mathbf{R}}(Prob^*\left(\max_{i=1,2,...,k}|\frac{\gamma_i^T\varepsilon^*}{\widehat{\sigma}}|\le\frac{x}{\widehat{\sigma}}+\frac{1}{\psi\widehat{\sigma}}+\frac{\log(2k)}{\beta\widehat{\sigma}}\right)\\
&\qquad\qquad -Prob^*\left(\max_{i=1,2,...,k}|\frac{\gamma_i^T\varepsilon^*}{\widehat{\sigma}}|\le\frac{x}{\widehat{\sigma}}\right))\\
&\qquad\le C\times\left(\frac{1}{\psi\widehat{\sigma}}+\frac{\log(2k)}{\beta\widehat{\sigma}}\right)\times\left(1+\sqrt{\log(k)}+\sqrt{|\log\left(\frac{1}{\psi\widehat{\sigma}}+\frac{\log(2k)}{\beta\widehat{\sigma}}\right)|}\right)
\end{aligned}\tag{A.1.20}$$

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

$$\begin{aligned}
&Prob(\max_{i=1,2,...,k}|\gamma_i^T\varepsilon|\le x)-Prob^*(\max_{i=1,2,...,k}|\gamma_i^T\varepsilon^*|\le x)\\
&\le Prob^*(\max_{i=1,2,...,k}|\gamma_i^T\xi|\le x)-Prob^*(\max_{i=1,2,...,k}|\gamma_i^T\varepsilon^*|\le x+\frac{1}{\psi}+\frac{\log(2k)}{\beta})+z\\
&\le Prob^*(G_\beta(\gamma_1^T\xi,...,\gamma_k^T\xi)\le x+\frac{\log(2k)}{\beta})\\
&\quad -Prob^*(G_\beta(\gamma_1^T\varepsilon^*,...,\gamma_k^T\varepsilon^*)\le x+\frac{1}{\psi}+\frac{\log(2k)}{\beta})+z\\
&\le \mathbf{E}^*h_{\beta,\psi,x+\frac{\log(2k)}{\beta}}(\gamma_1^T\xi,...,\gamma_k^T\xi)-\mathbf{E}^*h_{\beta,\psi,x+\frac{\log(2k)}{\beta}}(\gamma_1^T\varepsilon^*,...,\gamma_k^T\varepsilon^*)+z\\
&Prob(\max_{i=1,2,...,k}|\gamma_i^T\varepsilon|\le x)-Prob^*(\max_{i=1,2,...,k}|\gamma_i^T\varepsilon^*|\le x)\\
&\ge Prob^*(\max_{i=1,2,...,k}|\gamma_i^T\xi|\le x)-Prob^*(\max_{i=1,2,...,k}|\gamma_i^T\varepsilon^*|\le x-\frac{1}{\psi}-\frac{\log(2k)}{\beta})-z\\
&\ge Prob^*(G_\beta(\gamma_1^T\xi,...,\gamma_k^T\xi)\le x)-Prob^*(G_\beta(\gamma_1^T\varepsilon^*,...,\gamma_k^T\varepsilon^*)\le x-\frac{1}{\psi})-z\\
&\ge \mathbf{E}^*h_{\beta,\psi,x-\frac{1}{\psi}}(\gamma_1^T\xi,...,\gamma_k^T\xi)-\mathbf{E}^*h_{\beta,\psi,x-\frac{1}{\psi}}(\gamma_1^T\varepsilon^*,...,\gamma_k^T\varepsilon^*)-z
\end{aligned}\tag{A.1.21}$$

Therefore, we have

$$\begin{aligned}
&\sup_{x\in[0,\infty)} |Prob(\max_{i=1,2,...,k} |\gamma_i^T \varepsilon| \le x) - Prob^*(\max_{i=1,2,...,k} |\gamma_i^T \varepsilon^*| \le x)| \\
&\le z + \sup_{x\in\mathbf{R}} |\mathbf{E}^* h_{\beta,\psi,x}(\gamma_1^T \xi, ..., \gamma_k^T \xi) - \mathbf{E}^* h_{\beta,\psi,x}(\gamma_1^T \varepsilon^*, ..., \gamma_k^T \varepsilon^*)|
\end{aligned} \tag{A.1.22}$$

For any $i = 1,2,...,k, j = 1,2,...,n,$ define $H_{ij} = \sum_{s=1}^{j-1} \gamma_{si}\xi_s + \sum_{s=j+1}^{n} \gamma_{si}\varepsilon_s^*$, $m_{ij} = \gamma_{ji}\xi_j$ and $m_{ij}^* = \gamma_{ji}\varepsilon_j^*$, we have $H_{ij} + m_{ij} = H_{ij+1} + m_{ij+1}^*$, and

$$\begin{aligned}
&\sup_{x\in\mathbf{R}} |\mathbf{E}^* h_{\beta,\psi,x}(\gamma_1^T \xi, ..., \gamma_k^T \xi) - \mathbf{E}^* h_{\beta,\psi,x}(\gamma_1^T \varepsilon^*, ..., \gamma_k^T \varepsilon^*)| \\
&= \sup_{x\in\mathbf{R}} |\sum_{s=1}^{n} \mathbf{E}^* h_{\beta,\psi,x}(H_{1s} + m_{1s}, ..., H_{ks} + m_{ks}) - \mathbf{E}^* h_{\beta,\psi,x}(H_{1s} + m_{1s}^*, ..., H_{ks} + m_{ks}^*)| \\
&\le \sum_{s=1}^{n} \sup_{x\in\mathbf{R}} |\mathbf{E}^* h_{\beta,\psi,x}(H_{1s} + m_{1s}, ..., H_{ks} + m_{ks}) - \mathbf{E}^* h_{\beta,\psi,x}(H_{1s} + m_{1s}^*, ..., H_{ks} + m_{ks}^*)|
\end{aligned} \tag{A.1.23}$$

Since $\mathbf{E}(\xi_s|\varepsilon, \xi_b, \varepsilon_b^*, b \ne s) = \mathbf{E}(\varepsilon_s^*|\varepsilon, \xi_b, \varepsilon_b^*, b \ne s) = 0$, $\mathbf{E}(\xi_s^2 - \varepsilon_s^{*2}|\varepsilon, \xi_b, \varepsilon_b^*, b \ne s) = \sigma^2 - \widehat{\sigma}^2$,

from multivariate Taylor's theorem and (A.1.19), for any $s = 1, 2, ..., n$ and $x \in \mathbf{R}$,

$$
\begin{aligned}
&|\mathbf{E}\left(h_{\beta,\psi,x}(H_{1s}+m_{1s},...,H_{ks}+m_{ks}) - h_{\beta,\psi,x}(H_{1s}+m^*_{1s},...,H_{ks}+m^*_{ks})\right) \\
&\qquad\qquad \Big| \varepsilon, \xi_b, \varepsilon^*_b, b \neq s| \\
&\leq |\sum_{i=1}^{k} \frac{\partial h_{\beta,\psi,x}(H_{1s},...,H_{ks})}{\partial x_i} \gamma_{si} \mathbf{E}(\xi_s - \varepsilon^*_s | \varepsilon, \xi_b, \varepsilon^*_b, b \neq s)| \\
&+\frac{1}{2}|\sum_{i=1}^{k}\sum_{j=1}^{k} \frac{\partial^2 h_{\beta,\psi,x}(H_{1s},...,H_{ks})}{\partial x_i \partial x_j} \gamma_{si}\gamma_{sj} \mathbf{E}(\xi_s^2 - \varepsilon^{*2}_s | \varepsilon, \xi_b, \varepsilon^*_b, b \neq s)| \\
&+(g_*\psi^3 + g_*\psi^2\beta + g_*\psi\beta^2) \max_{i=1,2,...,k} |\gamma_{si}|^3 \times (\mathbf{E}|\varepsilon_1|^3 + \widehat{\sigma}^3 D) \qquad \text{(A.1.24)} \\
&\Rightarrow \sup_{x\in\mathbf{R}} |\mathbf{E} h_{\beta,\psi,x}(H_{1s}+m_{1s},...,H_{ks}+m_{ks}) \\
&-h_{\beta,\psi,x}(H_{1s}+m^*_{1s},...,H_{ks}+m^*_{ks})|\varepsilon, \xi_b, \varepsilon^*_b, b \neq s| \\
&\leq g_*(\psi^2 + \psi\beta)|\sigma^2 - \widehat{\sigma}^2| \times \max_{i=1,...,k} \gamma^2_{si} + (\mathbf{E}|\varepsilon_1|^3 + \widehat{\sigma}^3 D) \\
&\qquad \times g_*(\psi^3 + \psi^2\beta + \psi\beta^2) \times \max_{i=1,...,k} |\gamma_{si}|^3
\end{aligned}
$$

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

$$
\begin{aligned}
&\sup_{x\in[0,\infty)} |Prob(\max_{i=1,2,...,k} |\gamma_i^T \varepsilon| \leq x) - Prob^*(\max_{i=1,2,...,k} |\gamma_i^T \varepsilon^*| \leq x)| \\
&\qquad \leq z + (g_*\psi^2 + g_*\psi\beta)|\sigma^2 - \widehat{\sigma}^2| \times \sum_{s=1}^{n} \max_{i=1,...,k} \gamma^2_{si} \qquad \text{(A.1.25)} \\
&+(\mathbf{E}|\varepsilon_1|^3 + \widehat{\sigma}^3 D) \times g_*(\psi^3 + \psi^2\beta + \psi\beta^2) \times \sum_{s=1}^{n} \max_{i=1,...,k} |\gamma_{si}|^3
\end{aligned}
$$

In particular, for any given $\delta > 0$, choose $\psi = \beta = \log^{3/2}(n)/\delta^{1/4}$ and suppose $\frac{3\sigma}{2} > \widehat{\sigma} > \frac{\sigma}{2}$.

For sufficiently large n we have $\frac{1}{\psi\widehat{\sigma}}+\frac{\log(2k)}{\beta\widehat{\sigma}}\leq\frac{4\log(n)}{\psi\sigma}\leq\frac{4\delta^{1/4}}{\sigma\sqrt{\log(n)}}<1$ and

$$\begin{aligned} z &\leq \frac{4C\log(n)}{\psi\sigma}\times\left(2\sqrt{\log(n)}+\sqrt{\log(\psi\widehat{\sigma})}\right) \\ &\leq \frac{4C\delta^{1/4}}{\sigma}\left(2+\sqrt{\frac{\frac{3}{2}\log(\log(n))+\log(3\sigma/2\delta^{1/4})}{\log(n)}}\right)\leq C^{'}\delta^{1/4} \end{aligned} \tag{A.1.26}$$

Here $C^{'}=\frac{12C}{\sigma}$.

Suppose condition C1. For any $1>\delta>0$, $\exists D_\delta>0$ such that for sufficiently large n,

$$\begin{aligned} &Prob\left(|\sigma^2-\widehat{\sigma}^2|\leq D_\delta\times n^{-\alpha_\sigma}\right)>1-\delta \\ &\max_{j=1,2,\ldots,n,i=1,2,\ldots,k}|\gamma_{ji}|<\delta\times n^{(\alpha_\sigma-1)/2}\times\log^{-3/2}(n), \\ &\text{and }\max_{j=1,2,\ldots,n,i=1,2,\ldots,k}|\gamma_{ji}|<\delta\times n^{-1/3}\times\log^{-3/2}(n) \end{aligned} \tag{A.1.27}$$

Choose $\psi=\beta=\log^{3/2}(n)/\delta^{1/4}$. According to (A.1.25), for sufficiently large n, (A.1.27) happens and $\frac{1}{2}\sigma<\widehat{\sigma}<\frac{3}{2}\sigma$ with probability $1-\delta$. If (A.1.27) happens,

$$\begin{aligned} &\sup_{x\in[0,\infty)}|Prob(\max_{i=1,2,\ldots,k}|\gamma_i^T\varepsilon|\leq x)-Prob^*(\max_{i=1,2,\ldots,k}|\gamma_i^T\varepsilon^*|\leq x)| \\ &\leq C^{'}\delta^{1/4}+2g_*\psi^2\times D_\delta\times n^{-\alpha_\sigma}\times\frac{\delta^2\times n^{\alpha_\sigma}}{\log^3(n)} \\ &+(\mathbf{E}|\varepsilon_1|^3+\frac{27D}{8}\sigma^3)\times 3g_*\psi^3\times\delta^3\times n\times\frac{1}{n\log^{9/2}(n)} \\ &=C^{'}\delta^{1/4}+2g_*D_\delta\delta^{3/2}+3g_*(\mathbf{E}|\varepsilon_1|^3+\frac{27D}{8}\sigma^3)\times\delta^{9/4} \end{aligned} \tag{A.1.28}$$

For $\delta>0$ can be arbitrarily small, we prove (A.1.12).

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

n

$$
\begin{aligned}
Prob\left(|\sigma^2-\widehat{\sigma}^2|\le D_\delta\times n^{-\alpha_\sigma}\right)\ge 1-\delta,\quad k\le\frac{\delta n^{\alpha_\sigma}}{\log^3(n)},\\
\max_{i=1,2,\ldots,k}\sum_{j=1}^{n}\gamma_{ji}^2\le D_\delta\\
\text{and }\max_{j=1,2,\ldots,n,i=1,2,\ldots,k}|\gamma_{ji}|\le\frac{D_\delta\times n^{-\alpha_\sigma}}{\log^{3/2}(n)}
\end{aligned}
\tag{A.1.29}
$$

Since

$$
\begin{aligned}
\sum_{j=1}^{n}\max_{i=1,\ldots,k}\gamma_{ji}^2\le\sum_{j=1}^{n}\sum_{i=1}^{k}\gamma_{ji}^2\le kD_\delta\\
\sum_{j=1}^{n}\max_{i=1,\ldots,k}\gamma_{ji}^3\le\max_{j=1,2,\ldots,n,i=1,2,\ldots,k}|\gamma_{ji}|\times\sum_{j=1}^{n}\max_{i=1,\ldots,k}\gamma_{ji}^2\\
\le kD_\delta\times\max_{j=1,2,\ldots,n,i=1,2,\ldots,k}|\gamma_{ji}|
\end{aligned}
\tag{A.1.30}
$$

If (A.1.29) happens, by choosing $\psi=\beta=\log^{3/2}(n)/\delta^{1/4}$

$$
\begin{aligned}
\sup_{x\in[0,\infty)}|Prob(\max_{i=1,2,\ldots,k}|\gamma_i^T\varepsilon|\le x)-Prob^*(\max_{i=1,2,\ldots,k}|\gamma_i^T\varepsilon^*|\le x)|\\
\le C^{'}\delta^{1/4}+2g_*\psi^2D_\delta n^{-\alpha_\sigma}\times kD_\delta\\
+(\mathbf{E}|\varepsilon_1|^3+\frac{27D}{8}\sigma^3)\times 3g_*\psi^3\times kD_\delta\max_{j=1,2,\ldots,n,i=1,2,\ldots,k}|\gamma_{ji}|\\
\le C^{'}\delta^{1/4}+2g_*D_\delta^2\times\frac{\log^3(n)}{\delta^{1/2}}\times\frac{\delta n^{\alpha_\sigma}}{\log^3(n)}\times n^{-\alpha_\sigma}\\
+3(\mathbf{E}|\varepsilon_1|^3+\frac{27D}{8}\sigma^3)g_*D_\delta^2\times\frac{\log^{9/2}(n)}{\delta^{3/4}}\times\frac{\delta n^{\alpha_\sigma}}{\log^3(n)}\times\frac{n^{-\alpha_\sigma}}{\log^{3/2}(n)}\\
=C^{'}\delta^{1/4}+2g_*D_\delta^2\delta^{1/2}+3(\mathbf{E}|\varepsilon_1|^3+\frac{27D}{8}\sigma^3)g_*D_\delta^2\times\delta^{1/4}
\end{aligned}
\tag{A.1.31}
$$

and we prove (A.1.12).

If $\widehat{\sigma}=\sigma$. We choose $\psi=\beta=\log^{3/2}(n)/\delta^{1/4}$, (A.1.25) can be modified to

$$
\begin{aligned}
\sup_{x\in[0,\infty)}|Prob(\max_{i=1,2,\ldots,k}|\gamma_i^T\varepsilon|\le x)-Prob(\max_{i=1,2,\ldots,k}|\gamma_i^T\varepsilon^*|\le x)|\\
\le C^{'}\delta^{1/4}+(\mathbf{E}|\varepsilon_1|^3+D\sigma^3)g_*\psi(\psi^2+\psi\beta+\beta^2)\sum_{s=1}^{n}\max_{i=1,\ldots,k}|\gamma_{si}|^3
\end{aligned}
\tag{A.1.32}
$$

Suppose condition $C1^{'}$. For any $\delta > 0$ and sufficiently large n,

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

$$\begin{aligned} \sup_{x\in[0,\infty)} |Prob(\max_{i=1,2,\ldots,k} |\sum_{j=1}^{n} \gamma_{ji}\varepsilon_j| \leq x) - Prob(\max_{i=1,2,\ldots,k} |\sum_{j=1}^{n} \gamma_{ji}\varepsilon_j^*| \leq x)| \\ \leq C^{'}\delta^{1/4} + 3(\mathbf{E}|\varepsilon_1|^3 + D\sigma^3)g_* \times \delta^{9/4} \end{aligned} \tag{A.1.33}$$

and we prove (A.1.15).

Suppose condition $C2^{'}$. For any $\delta > 0$ and sufficiently large n,

$k \times \max_{j=1,2,\ldots,n,i=1,2,\ldots,k} |\gamma_{ji}| \leq \delta \log^{-9/2}(n)$. According to (A.1.30), for sufficiently large n we have

$$\begin{aligned} \sup_{x\in[0,\infty)} |Prob(\max_{i=1,2,\ldots,k} |\sum_{j=1}^{n} \gamma_{ji}\varepsilon_j| \leq x) - Prob(\max_{i=1,2,\ldots,k} |\sum_{j=1}^{n} \gamma_{ji}\varepsilon_j^*| \leq x)| \\ \leq C^{'}\delta^{1/4} + 3(\mathbf{E}|\varepsilon_1|^3 + D\sigma^3)g_* D_\delta \times \delta^{1/4} \end{aligned} \tag{A.1.34}$$

and we prove (A.1.15). □

Condition C1 implies $C1^{'}$, and condition C2 implies $C2^{'}$. The additional proportions in C1 and C2 accommodate the error introduced in estimating errors' variance σ^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 A.1.3 and the classical central limit theorem is that *k can grow as n increases*. The maximum $\max_{i=1,2,\ldots,k} |\sum_{j=1}^{n} \gamma_{ji}\varepsilon_j|$ does not have an asymptotic distribution if $k \to \infty$. However, if the random variables are mixed well, approximating the distribution of $\max_{i=1,2,\ldots,k} |\sum_{j=1}^{n} \gamma_{ji}\varepsilon_j|$ by the distribution of the maximum of normal random variables is still applicable. With the help of lemma A.1.3, we can establish the normal approximation theorem and construct the simultaneous confidence region for $\widehat{\gamma}$(defined in (2.17)).

A.2 Proofs of theorems in section 2.4

This section applies notations in section 2.3.

Proof of theorem 1. From (2.16),

$$
\begin{aligned}
Prob\left(\widehat{\mathscr{N}_{b_n}} \neq \mathscr{N}_{b_n}\right) &\leq Prob\left(\min_{i\in\mathscr{N}_{b_n}} |\widetilde{\theta}_i| \leq b_n\right) + Prob\left(\max_{i\notin\mathscr{N}_{b_n}} |\widetilde{\theta}_i| > b_n\right) \\
&\leq Prob(\min_{i\in\mathscr{N}_{b_n}} |\theta_i| - \max_{i\in\mathscr{N}_{b_n}} \rho_n^2 |\sum_{j=1}^{r} \frac{q_{ij}\zeta_j}{(\lambda_j^2+\rho_n)^2}| \\
&- \max_{i\in\mathscr{N}_{b_n}} |\sum_{j=1}^{r} q_{ij}\left(\frac{\lambda_j}{\lambda_j^2+\rho_n} + \frac{\rho_n\lambda_j}{(\lambda_j^2+\rho_n)^2}\right)\sum_{l=1}^{n} p_{lj}\varepsilon_l| \leq b_n) \\
&+ Prob(\max_{i\notin\mathscr{N}_{b_n}} |\theta_i| + \max_{i\notin\mathscr{N}_{b_n}} \rho_n^2 |\sum_{j=1}^{r} \frac{q_{ij}\zeta_j}{(\lambda_j^2+\rho_n)^2}| \\
&+ \max_{i\notin\mathscr{N}_{b_n}} |\sum_{j=1}^{r} q_{ij}\left(\frac{\lambda_j}{\lambda_j^2+\rho_n} + \frac{\rho_n\lambda_j}{(\lambda_j^2+\rho_n)^2}\right)\sum_{l=1}^{n} p_{lj}\varepsilon_l| > b_n)
\end{aligned}
\tag{A.2.1}
$$

From Cauchy inequality,

$$
\begin{aligned}
\max_{i=1,2,\ldots,p} \rho_n^2 |\sum_{j=1}^{r} \frac{q_{ij}\zeta_j}{(\lambda_j^2+\rho_n)^2}| &\leq \max_{i=1,2,\ldots,p} \rho_n^2 \sqrt{\sum_{j=1}^{r} q_{ij}^2} \times \sqrt{\sum_{j=1}^{r} \frac{\zeta_j^2}{(\lambda_j^2+\rho_n)^4}} \\
&= O(n^{\alpha_\theta - 2\delta}) \\
\max_{i=1,2,\ldots,p} \sum_{l=1}^{n} \left(\sum_{j=1}^{r} q_{ij}\left(\frac{\lambda_j}{\lambda_j^2+\rho_n} + \frac{\rho_n\lambda_j}{(\lambda_j^2+\rho_n)^2}\right) p_{lj}\right)^2 \\
= \max_{i=1,2,\ldots,p} \sum_{j=1}^{r} q_{ij}^2 \left(\frac{\lambda_j}{\lambda_j^2+\rho_n} + \frac{\rho_n\lambda_j}{(\lambda_j^2+\rho_n)^2}\right)^2 \\
\leq \max_{i=1,2,\ldots,p} \frac{4\sum_{j=1}^{r} q_{ij}^2}{\lambda_r^2}
\end{aligned}
\tag{A.2.2}
$$

Therefore, for sufficiently large n, from assumption 4 and lemma A.1.1

$$\begin{aligned}
&\min_{i\in\mathscr{N}_{b_n}}|\theta_i| - \max_{i\in\mathscr{N}_{b_n}}\rho_n^2|\sum_{j=1}^{r}\frac{q_{ij}\zeta_j}{(\lambda_j^2+\rho_n)^2}| - b_n > \frac{1}{2}(\frac{1}{c_b}-1)b_n \\
&b_n - \max_{i\notin\mathscr{N}_{b_n}}|\theta_i| - \max_{i\in\mathscr{N}_{b_n}}\rho_n^2|\sum_{j=1}^{r}\frac{q_{ij}\zeta_j}{(\lambda_j^2+\rho_n)^2}| > \frac{1}{2}(1-c_b)b_n \\
\Rightarrow Prob\left(\widehat{\mathscr{N}_{b_n}}\neq\mathscr{N}_{b_n}\right) &\leq \frac{|\mathscr{N}_{b_n}|\times E\times 2^m}{\lambda_r^m\times(\frac{1}{2}(\frac{1}{c_b}-1)b_n)^m} + \frac{(p-|\mathscr{N}_{b_n}|)\times E\times 2^m}{\lambda_r^m\times(\frac{1}{2}(1-c_b)b_n)^m} \\
&= O(n^{\alpha_p+m\nu_b-m\eta})
\end{aligned} \tag{A.2.3}$$

and we prove (2.21).

Define $\widehat{\gamma}=M\widehat{\theta}=(\widehat{\gamma}_1,...,\widehat{\gamma}_{p_1})^T$ and $\gamma=M\beta=(\gamma_1,...,\gamma_{p_1})^T$. For $\beta=\theta+\theta_\perp$, if $\widehat{\mathscr{N}_{b_n}}=\mathscr{N}_{b_n}$, (2.16) and (2.4) imply

$$\begin{aligned}
&\max_{i=1,2,...,p_1}|\widehat{\gamma}_i-\gamma_i| = \max_{i=1,2,...,p_1}|\sum_{j\in\mathscr{N}_{b_n}}m_{ij}\widetilde{\theta}_j - \sum_{j\in\mathscr{N}_{b_n}}m_{ij}\theta_j - \sum_{j\notin\mathscr{N}_{b_n}}m_{ij}\theta_j - \sum_{j=1}^{p}m_{ij}\theta_{\perp,j}| \\
&\leq \max_{i=1,2,...,p_1}\rho_n^2|\sum_{k=1}^{r}\frac{c_{ik}\zeta_k}{(\lambda_k^2+\rho_n)^2}| + \max_{i=1,2,...,p_1}|\sum_{k=1}^{r}c_{ik}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda^2+\rho_n)^2}\right)\sum_{l=1}^{n}p_{lk}\varepsilon_l| \\
&\qquad + \max_{i=1,2,...,p_1}|\sum_{j\notin\mathscr{N}_{b_n}}m_{ij}\theta_j| + \max_{i=1,2,...,p_1}|\sum_{j=1}^{p}m_{ij}\theta_{\perp,j}|
\end{aligned} \tag{A.2.4}$$

From (2.4) and assumption 5, if $i\notin\mathscr{M}$, then $c_{ik}=0$ for $k=1,2,...,r$, so from Cauchy inequality

and lemma A.1.1,

$$\begin{aligned}
\max_{i=1,2,\ldots,p_1} \rho_n^2 |\sum_{k=1}^{r} \frac{c_{ik}\zeta_k}{(\lambda_k^2+\rho_n)^2}| &\le \max_{i\in\mathscr{M}} \rho_n^2 \sqrt{\sum_{k=1}^{r} c_{ik}^2} \times \sqrt{\sum_{k=1}^{r} \frac{\zeta_k^2}{(\lambda_k^2+\rho_n)^4}} \\
&\le \sqrt{C_{\mathscr{M}}}\rho_n^2 \times \frac{\|\theta\|_2}{\lambda_r^4} = O(n^{\alpha_\theta - 2\delta}) \\
\max_{i\in\mathscr{M}} \sum_{l=1}^{n} \left(\sum_{k=1}^{r} c_{ik} p_{lk} \left(\frac{\lambda_k}{\lambda_k^2+\rho_n} + \frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2} \right) \right)^2 & \\
= \max_{i\in\mathscr{M}} \sum_{k=1}^{r} c_{ik}^2 \left(\frac{\lambda_k}{\lambda_k^2+\rho_n} + \frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2} \right)^2 &\le \frac{4C_{\mathscr{M}}}{\lambda_r^2} \qquad \text{(A.2.5)} \\
\Rightarrow Prob\left(\max_{i=1,2,\ldots,p_1} |\sum_{k=1}^{r} c_{ik} \left(\frac{\lambda_k}{\lambda_k^2+\rho_n} + \frac{\rho_n\lambda_k}{(\lambda^2+\rho_n)^2} \right) \sum_{l=1}^{n} p_{lk}\varepsilon_l| > \delta \right) & \\
\le \frac{|\mathscr{M}| \times E \times 2^m C_{\mathscr{M}}^{m/2}}{\lambda_r^m \delta^m} \quad \text{for } \forall \delta > 0 & \\
\Rightarrow \max_{i=1,2,\ldots,p_1} |\sum_{k=1}^{r} c_{ik} \left(\frac{\lambda_k}{\lambda_k^2+\rho_n} + \frac{\rho_n\lambda_k}{(\lambda^2+\rho_n)^2} \right) \sum_{l=1}^{n} p_{lk}\varepsilon_l| = O_p(|\mathscr{M}|^{1/m} \times n^{-\eta}) &
\end{aligned}$$

Here E is the constant defined in lemma A.1.1. Combine with assumption 2, assumption 5, and (A.2.3), we prove (2.22).

If $\widehat{\mathscr{N}_{b_n}} = \mathscr{N}_{b_n}$, since $X\beta = X\theta$, we have

$$\begin{aligned}
\widehat{\sigma}^2 - \sigma^2 &= \frac{1}{n}\sum_{i=1}^{n} \left(\varepsilon_i - \sum_{j\in\mathscr{N}_{b_n}} x_{ij}(\widetilde{\theta}_j - \theta_j) + \sum_{j\notin\mathscr{N}_{b_n}} x_{ij}\theta_j \right)^2 - \sigma^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 - \sigma^2 + \frac{1}{n}\sum_{i=1}^{n} \left(\sum_{j\in\mathscr{N}_{b_n}} x_{ij}(\widetilde{\theta}_j - \theta_j) \right)^2 + \frac{1}{n}\sum_{i=1}^{n} \left(\sum_{j\notin\mathscr{N}_{b_n}} x_{ij}\theta_j \right)^2 \\
&\qquad - \frac{2}{n}\sum_{i=1}^{n} \sum_{j\in\mathscr{N}_{b_n}} \varepsilon_i x_{ij}(\widetilde{\theta}_j - \theta_j) \qquad \text{(A.2.6)} \\
&\quad + \frac{2}{n}\sum_{i=1}^{n} \sum_{j\notin\mathscr{N}_{b_n}} \varepsilon_i x_{ij}\theta_j - \frac{2}{n}\sum_{i=1}^{n} \left(\sum_{j\in\mathscr{N}_{b_n}} x_{ij}(\widetilde{\theta}_j - \theta_j) \right) \times \left(\sum_{j\notin\mathscr{N}_{b_n}} x_{ij}\theta_j \right)
\end{aligned}$$

From assumption 3,

$\mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 - \sigma^2\right)^2 \le \frac{2}{n}(\mathbf{E}\varepsilon_1^4 + \sigma^4) = O(1/n) \Rightarrow \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 - \sigma^2 = O_p(1/\sqrt{n})$. For the second

term, from assumption 1 and (A.2.2),

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j\in\mathscr{N}_{b_n}} x_{ij}(\widetilde{\theta}_j-\theta_j)\right)^2 &\le C_\lambda^2 \sum_{j\in\mathscr{N}_{b_n}}(\widetilde{\theta}_j-\theta_j)^2 \\
&\le 2C_\lambda^2 \sum_{j\in\mathscr{N}_{b_n}}(\rho_n^4\left(\sum_{k=1}^{r}\frac{q_{jk}\zeta_k}{(\lambda_k^2+\rho_n)^2}\right)^2 \\
&+\left(\sum_{k=1}^{r} q_{jk}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)\sum_{l=1}^{n}p_{lk}\varepsilon_l\right)^2) \\
= O(|\mathscr{N}_{b_n}|\times n^{2\alpha_\theta-4\delta}) &+ 2C_\lambda^2\sum_{j\in\mathscr{N}_{b_n}}\left(\sum_{k=1}^{r} q_{jk}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)\sum_{l=1}^{n}p_{lk}\varepsilon_l\right)^2
\end{aligned} \tag{A.2.7}$$

Since

$$\begin{aligned}
&\mathbf{E}\sum_{j\in\mathscr{N}_{b_n}}\left(\sum_{k=1}^{r} q_{jk}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)\sum_{l=1}^{n}p_{lk}\varepsilon_l\right)^2 \\
&= \sigma^2\sum_{j\in\mathscr{N}_{b_n}}\sum_{l=1}^{n}\left(\sum_{k=1}^{r} q_{jk}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)p_{lk}\right)^2 \\
&= \sigma^2\sum_{j\in\mathscr{N}_{b_n}}\sum_{k=1}^{r} q_{jk}^2\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)^2 \le \frac{4\sigma^2|\mathscr{N}_{b_n}|}{\lambda_r^2}
\end{aligned} \tag{A.2.8}$$

We have $\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j\in\mathscr{N}_{b_n}} x_{ij}(\widetilde{\theta}_j-\theta_j)\right)^2 = O_p(|\mathscr{N}_{b_n}|\times n^{2\alpha_\theta-4\delta}+|\mathscr{N}_{b_n}|\times n^{-2\eta})$. For the third term, from assumption 6 we have

$$\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j\notin\mathscr{N}_{b_n}} x_{ij}\theta_j\right)^2 \le C_\lambda^2\sum_{j\notin\mathscr{N}_{b_n}}\theta_j^2 \le C_\lambda^2\times b_n\sum_{j\notin\mathscr{N}_{b_n}}|\theta_j| = O(n^{-\alpha_\sigma}) \tag{A.2.9}$$

For the fourth term, from Cauchy inequality and (A.2.7),

$$\begin{aligned}
\mathbf{E}\frac{1}{n}|\sum_{i=1}^{n}\sum_{j\in\mathscr{N}_{b_n}}\varepsilon_i x_{ij}(\widetilde{\theta}_j-\theta_j)| &\le \frac{1}{n}\mathbf{E}\sqrt{\sum_{i=1}^{n}\varepsilon_i^2}\times\sqrt{\sum_{i=1}^{n}(\sum_{j\in\mathscr{N}_{b_n}}x_{ij}(\widetilde{\theta}_j-\theta_j))^2} \\
&\le \sqrt{\frac{\mathbf{E}\sum_{i=1}^{n}\varepsilon_i^2}{n}}\times\sqrt{\frac{1}{n}\mathbf{E}\sum_{i=1}^{n}(\sum_{j\in\mathscr{N}_{b_n}}x_{ij}(\widetilde{\theta}_j-\theta_j))^2} \\
&= \sigma\times O(\sqrt{|\mathscr{N}_{b_n}|\times n^{2\alpha_\theta-4\delta}+|\mathscr{N}_{b_n}|\times n^{-2\eta}}) \\
\Rightarrow \frac{1}{n}|\sum_{i=1}^{n}\sum_{j\in\mathscr{N}_{b_n}}\varepsilon_i x_{ij}(\widetilde{\theta}_j-\theta_j)| &= O_p(\sqrt{|\mathscr{N}_{b_n}|}\times n^{\alpha_\theta-2\delta}+\sqrt{|\mathscr{N}_{b_n}|}\times n^{-\eta})
\end{aligned} \tag{A.2.10}$$

For the fifth term,

$$\begin{aligned}
\mathbf{E}|\frac{1}{n}\sum_{i=1}^{n}\sum_{j\notin\mathscr{N}_{b_n}}\varepsilon_i x_{ij}\theta_j|^2 = \frac{\sigma^2}{n^2}\sum_{i=1}^{n}\left(\sum_{j\notin\mathscr{N}_{b_n}}x_{ij}\theta_j\right)^2 &\le \frac{\sigma^2 C_\lambda^2}{n}\sum_{j\notin\mathscr{N}_{b_n}}\theta_j^2 \\
\Rightarrow \frac{1}{n}\sum_{i=1}^{n}\sum_{j\notin\mathscr{N}_{b_n}}\varepsilon_i x_{ij}\theta_j &= O_p(n^{-(1+\alpha_\sigma)/2})
\end{aligned} \tag{A.2.11}$$

For the last term,

$$\begin{aligned}
&\frac{1}{n}|\sum_{i=1}^{n}\left(\sum_{j\in\mathscr{N}_{b_n}}x_{ij}(\widetilde{\theta}_j-\theta_j)\right)\times\left(\sum_{j\notin\mathscr{N}_{b_n}}x_{ij}\theta_j\right)| \\
&\le C_\lambda^2\sqrt{\sum_{j\in\mathscr{N}_{b_n}}(\widetilde{\theta}_j-\theta_j)^2}\times\sqrt{\sum_{j\notin\mathscr{N}_{b_n}}\theta_j^2} \\
&= O_p(\sqrt{|\mathscr{N}_{b_n}|}\times n^{\alpha_\theta-2\delta-\alpha_\sigma/2}+\sqrt{|\mathscr{N}_{b_n}|}\times n^{-\eta-\alpha_\sigma/2})
\end{aligned} \tag{A.2.12}$$

From (2.21), $Prob(\widehat{\mathscr{N}_{b_n}}\neq\mathscr{N}_{b_n})\to 0$. So we have

$$|\widehat{\sigma}^2-\sigma^2| = O_p\left(\frac{1}{\sqrt{n}}+\sqrt{|\mathscr{N}_{b_n}|}\times n^{\alpha_\theta-2\delta}+\sqrt{|\mathscr{N}_{b_n}|}\times n^{-\eta}+n^{-\alpha_\sigma}\right) \tag{A.2.13}$$

From assumption 2 and 6, we prove the second result. □

Define $T=(c_{ik})_{i\in\mathscr{M},k=1,2,\ldots,r}$. From assumption 7, since the matrix

$\left(\frac{1}{\tau_i}c_{ik}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)\right)_{i\in\mathscr{M},j=1,2,...,r} = D_1TD_2$ with $D_1 = diag(1/\tau_i, i\in\mathscr{M})$ and $D_2 = diag\left(\frac{\lambda_1}{\lambda_1^2+\rho_n}+\frac{\rho_n\lambda_1}{(\lambda_1^2+\rho_n)^2},...,\frac{\lambda_r}{\lambda_r^2+\rho_n}+\frac{\rho_n\lambda_r}{(\lambda_r^2+\rho_n)^2}\right)$, the matrix $\left(\frac{1}{\tau_i}c_{ik}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)\right)_{i\in\mathscr{M},j=1,2,...,r}$ also has rank $|\mathscr{M}|$. The proof of theorem 2 uses this result.

Proof of theorem 2. From Cauchy inequality and assumption 2, suppose $\delta = \frac{\eta+\alpha_\theta+\delta_1}{2}$ with $\delta_1 > 0$. For $i\in\mathscr{M}$,

$$\begin{aligned} |\sum_{k=1}^{r}\frac{c_{ik}\zeta_k}{(\lambda_k^2+\rho_n)^2}| \le \sqrt{\sum_{k=1}^{r}\frac{c_{ik}^2\lambda_k^2}{(\lambda_k^2+\rho_n)^2}}\times\sqrt{\sum_{k=1}^{r}\frac{\zeta_k^2}{\lambda_k^2(\lambda_k^2+\rho_n)^2}} \le \tau_i\times\frac{\|\theta\|_2}{\lambda_r^3} \\ \Rightarrow \max_{i\in\mathscr{M}}\frac{\rho_n^2}{\tau_i}|\sum_{k=1}^{r}\frac{c_{ik}\zeta_k}{(\lambda_k^2+\rho_n)^2}| = O(n^{-\delta_1}) \end{aligned} \tag{A.2.14}$$

Define $t_{il} = \frac{1}{\tau_i}\times\sum_{k=1}^{r}c_{ik}p_{lk}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)$ for $i\in\mathscr{M}$ and $l = 1,2,...,n$. From (2.16), (2.5), (A.2.4) and assumption 5, if $\widehat{\mathscr{N}_{b_n}} = \mathscr{N}_{b_n}$, we have $\widehat{\tau}_i = \tau_i \ge 1/\sqrt{n}$ and $\exists$ a constant $C > 0$, for any $a > 0$ and sufficiently large n,

$$\begin{aligned} \max_{i=1,2,...,p_1}\frac{|\widehat{\gamma}_i-\gamma_i|}{\widehat{\tau}_i} &\le \max_{i\in\mathscr{M}}\frac{\rho_n^2}{\tau_i}|\sum_{k=1}^{r}\frac{c_{ik}\zeta_k}{(\lambda_k^2+\rho_n)^2}| \\ +\max_{i\in\mathscr{M}}|\sum_{l=1}^{n}t_{il}\varepsilon_l| + \max_{i=1,2,...,p_1}\frac{|\sum_{j\notin\mathscr{N}_{b_n}}m_{ij}\theta_j|}{\tau_i} + \max_{i=1,2,...,p_1}\frac{|\sum_{j=1}^{p}m_{ij}\theta_{\perp,j}|}{\tau_i} & \\ \le \max_{i\in\mathscr{M}}|\sum_{l=1}^{n}t_{il}\varepsilon_l| + Cn^{-\delta_1} + \frac{a}{\sqrt{\log(n)}} & \\ \max_{i=1,2,...,p_1}\frac{|\widehat{\gamma}_i-\gamma_i|}{\widehat{\tau}_i} \ge \max_{i\in\mathscr{M}}|\sum_{l=1}^{n}t_{il}\varepsilon_l| - \max_{i\in\mathscr{M}}\frac{\rho_n^2}{\tau_i}|\sum_{k=1}^{r}\frac{c_{ik}\zeta_k}{(\lambda_k^2+\rho_n)^2}| & \\ -\max_{i=1,2,...,p_1}\frac{|\sum_{j\notin\mathscr{N}_{b_n}}m_{ij}\theta_j|}{\tau_i} - \max_{i=1,2,...,p_1}\frac{|\sum_{j=1}^{p}m_{ij}\theta_{\perp,j}|}{\tau_i} & \\ \ge \max_{i\in\mathscr{M}}|\sum_{l=1}^{n}t_{il}\varepsilon_l| - Cn^{-\delta_1} - \frac{a}{\sqrt{\log(n)}} & \end{aligned} \tag{A.2.15}$$

According to theorem 1, $\exists$ a constant C and for any given $a > 0$, for sufficiently large n

and any $x \geq 0$, define $\mathscr{V} = Cn^{-\delta_1} + \frac{a}{\sqrt{\log(n)}}$,

$$
\begin{aligned}
Prob\left(\max_{i=1,2,\ldots,p_1} \frac{|\widehat{\gamma}_i - \gamma_i|}{\widehat{\tau}_i} \leq x\right) &\leq Prob\left(\max_{i=1,2,\ldots,p_1} \frac{|\widehat{\gamma}_i - \gamma_i|}{\widehat{\tau}_i} \leq x \cap \widehat{\mathscr{N}_{b_n}} = \mathscr{N}_{b_n}\right) \\
&\quad + Prob\left(\widehat{\mathscr{N}_{b_n}} \neq \mathscr{N}_{b_n}\right) \\
&\leq Prob\left(\max_{i\in\mathscr{M}} |\sum_{l=1}^{n} t_{il}\varepsilon_l| \leq x + \mathscr{V}\right) + Cn^{\alpha_p + m\nu_b - m\eta} \\
&\leq Prob\left(\max_{i\in\mathscr{M}} |\sum_{l=1}^{n} t_{il}\varepsilon_l^*| \leq x\right) + Cn^{\alpha_p + m\nu_b - m\eta} \\
&\quad + \sup_{x\geq 0} |Prob\left(\max_{i\in\mathscr{M}} |\sum_{l=1}^{n} t_{il}\varepsilon_l| \leq x\right) - Prob\left(\max_{i\in\mathscr{M}} |\sum_{l=1}^{n} t_{il}\varepsilon_l^*| \leq x\right)| \\
&\quad + \sup_{x\in\mathbf{R}} \left(Prob\left(\max_{i\in\mathscr{M}} |\sum_{l=1}^{n} t_{il}\varepsilon_l^*| \leq x + \mathscr{V}\right) - Prob\left(\max_{i\in\mathscr{M}} |\sum_{l=1}^{n} t_{il}\varepsilon_l^*| \leq x\right)\right) \qquad \text{(A.2.16)} \\
Prob\left(\max_{i=1,2,\ldots,p_1} \frac{|\widehat{\gamma}_i - \gamma_i|}{\widehat{\tau}_i} \leq x\right) &\geq Prob\left(\max_{i=1,2,\ldots,p_1} \frac{|\widehat{\gamma}_i - \gamma_i|}{\widehat{\tau}_i} \leq x \cap \widehat{\mathscr{N}_{b_n}} = \mathscr{N}_{b_n}\right) \\
&\geq Prob\left(\max_{i\in\mathscr{M}} |\sum_{l=1}^{n} t_{il}\varepsilon_l| \leq x - \mathscr{V}\right) - Prob\left(\widehat{\mathscr{N}_{b_n}} \neq \mathscr{N}_{b_n}\right) \\
&\geq Prob\left(\max_{i\in\mathscr{M}} |\sum_{l=1}^{n} t_{il}\varepsilon_l^*| \leq x\right) - Cn^{\alpha_p + m\nu_b - m\eta} \\
&\quad - \sup_{x\in\mathbf{R}} \left(Prob\left(\max_{i\in\mathscr{M}} |\sum_{l=1}^{n} t_{il}\varepsilon_l^*| \leq x\right) - Prob\left(\max_{i\in\mathscr{M}} |\sum_{l=1}^{n} t_{il}\varepsilon_l^*| \leq x - \mathscr{V}\right)\right) \\
&\quad - |Prob\left(\max_{i\in\mathscr{M}} |\sum_{l=1}^{n} t_{il}\varepsilon_l| \leq x - \mathscr{V}\right) - Prob\left(\max_{i\in\mathscr{M}} |\sum_{l=1}^{n} t_{il}\varepsilon_l^*| \leq x - \mathscr{V}\right)|
\end{aligned}
$$

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

$$\begin{aligned}
\max_{i\in\mathscr{M}}\mathbf{E}\left(\sum_{l=1}^{n}t_{il}\varepsilon_l^*\right)^2 &= \sigma^2\max_{i\in\mathscr{M}}\sum_{l=1}^{n}t_{il}^2 = \sigma^2\max_{i\in\mathscr{M}}\frac{\sum_{k=1}^{r}c_{ik}^2\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)^2}{\tau_i^2} \leq \sigma^2 \\
\min_{i\in\mathscr{M}}\mathbf{E}\left(\sum_{l=1}^{n}t_{il}\varepsilon_l^*\right)^2 &= \sigma^2\min_{i\in\mathscr{M}}\frac{1}{1+\frac{1}{n\sum_{k=1}^{r}c_{ik}^2\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)^2}} \\
&\geq \sigma^2\min_{i\in\mathscr{M}}\frac{1}{1+\frac{1}{n\sum_{k=1}^{r}c_{ik}^2\frac{\lambda_k^2}{(\lambda_k^2+\rho_n)^2}}} \geq \frac{\sigma^2}{1+\frac{4C_\lambda^2}{c_{\mathscr{M}}}} > 0
\end{aligned} \tag{A.2.17}$$

and $(t_{il})_{i\in\mathscr{M},l=1,2,\ldots,n} = D_1TD_2P^T$, here $T=(c_{ik})_{i\in\mathscr{M},k=1,2,\ldots,r}$, $D_1 = diag(1/\tau_i, i\in\mathscr{M})$, and $D_2 = diag\left(\frac{\lambda_1}{\lambda_1^2+\rho_n}+\frac{\rho_n\lambda_1}{(\lambda_1^2+\rho_n)^2},\ldots,\frac{\lambda_r}{\lambda_r^2+\rho_n}+\frac{\rho_n\lambda_r}{(\lambda_r^2+\rho_n)^2}\right)$. So $(t_{il})_{i\in\mathscr{M},l=1,2,\ldots,n}$ has full rank(rank $|\mathscr{M}|$). From lemma A.1.2, $\exists$ a constant $C^{'}$ which only depends on $\sigma, c_{\mathscr{M}}, C_\lambda$ such that

$$\begin{aligned}
&\sup_{x\in\mathbf{R}}(Prob\left(\max_{i\in\mathscr{M}}|\sum_{l=1}^{n}t_{il}\varepsilon_l^*| \leq x + Cn^{-\delta_1} + \frac{a}{\sqrt{\log(n)}}\right) \\
&\qquad - Prob\left(\max_{i\in\mathscr{M}}|\sum_{l=1}^{n}t_{il}\varepsilon_l^*| \leq x\right)) \\
&\leq C^{'}\left(Cn^{-\delta_1}+\frac{a}{\sqrt{\log(n)}}\right)\times\left(1+\sqrt{\log(|\mathscr{M}|)}+\sqrt{|\log(Cn^{-\delta_1}+\frac{a}{\sqrt{\log(n)}})|}\right)
\end{aligned} \tag{A.2.18}$$

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

$$\begin{aligned}
&|\log(Cn^{-\delta_1}+\frac{a}{\sqrt{\log(n)}})| \leq \log(\frac{\sqrt{\log(n)}}{a}) = \frac{\log(\log(n))}{2} - \log(a) \leq \log(\log(n)) \\
&\Rightarrow \sup_{x\in\mathbf{R}}(Prob\left(\max_{i\in\mathscr{M}}|\sum_{l=1}^{n}t_{il}\varepsilon_l^*| \leq x + Cn^{-\delta_1} + \frac{a}{\sqrt{\log(n)}}\right) \\
&\qquad - Prob\left(\max_{i\in\mathscr{M}}|\sum_{l=1}^{n}t_{il}\varepsilon_l^*| \leq x\right)) \\
&\leq C^{'}\left(Cn^{-\delta_1}+\frac{a}{\sqrt{\log(n)}}\right)\times\left(1+\sqrt{\log(n)}+\sqrt{\log(\log(n))}\right) \leq 6C^{'}a
\end{aligned} \tag{A.2.19}$$

From assumption 7, (A.2.17) and lemma A.1.3, we have

$$\sup_{x\geq 0}|Prob\left(\max_{i\in\mathscr{M}}|\sum_{l=1}^{n}t_{il}\varepsilon_l|\leq x\right)-Prob\left(\max_{i\in\mathscr{M}}|\sum_{l=1}^{n}t_{il}\varepsilon_l^*|\leq x\right)|<a \tag{A.2.20}$$

for sufficiently large n. If $x<Cn^{-\delta_1}+\frac{a}{\sqrt{\log(n)}}$, then
$Prob\left(\max_{i\in\mathscr{M}}|\sum_{l=1}^{n}t_{il}\varepsilon_l|\leq x-Cn^{-\delta_1}-\frac{a}{\sqrt{\log(n)}}\right)=0$ and
$Prob\left(\max_{i\in\mathscr{M}}|\sum_{l=1}^{n}t_{il}\varepsilon_l^*|\leq x-Cn^{-\delta_1}-\frac{a}{\sqrt{\log(n)}}\right)=0$. Combine with (A.2.16) to (A.2.20), we have

$$\begin{aligned}\sup_{x\geq 0}|Prob\left(\max_{i=1,2,\ldots,p_1}\frac{|\widehat{\gamma}_i-\gamma_i|}{\widehat{\tau}_i}\leq x\right)-Prob\left(\max_{i\in\mathscr{M}}|\sum_{l=1}^{n}t_{il}\varepsilon_l^*|\leq x\right)| \\ \leq Cn^{\alpha_p+m\nu_b-m\eta}+6C^{'}a+a\end{aligned} \tag{A.2.21}$$

and we prove (2.27). □

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 $\forall x\in\mathscr{R}^{|\mathscr{M}|}$. And $\forall x\geq 0,\ \delta>0$, the set $\{t=(t_i,i\in\mathscr{M})|\ x<\max_{i\in\mathscr{M}}|t_i|\leq x+\delta\}$ 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 $0<\alpha_0<\alpha_1<1$,

$$\begin{aligned}\sup_{\alpha_0\leq\alpha\leq\alpha_1}|Prob\left(\max_{i=1,2,\ldots,p_1}\frac{|\widehat{\gamma}_i-\gamma_i|}{\widehat{\tau}_i}\leq c_{1-\alpha}\right)-(1-\alpha)| \\ \leq\sup_{x\geq 0}|Prob\left(\max_{i=1,2,\ldots,p_1}\frac{|\widehat{\gamma}_i-\gamma_i|}{\widehat{\tau}_i}\leq x\right)-H(x)|\to 0\end{aligned} \tag{A.2.22}$$

as $n\to\infty$.

A.3 Proofs of theorems in section 2.5

Proof of theorem 3. According to theorem 1, $Prob\left(\widehat{\mathscr{N}_{b_n}} \neq \mathscr{N}_{b_n}\right) = O(n^{\alpha_p + m\nu_b - m\eta})$. If $\widehat{\mathscr{N}_{b_n}} = \mathscr{N}_{b_n}$, from (2.16)

$$\begin{aligned} \|\widehat{\theta}\|_2^2 = \sum_{i\in\mathscr{N}_{b_n}} \widetilde{\theta}_i^2 \leq 3 \sum_{i\in\mathscr{N}_{b_n}} |\theta_i|^2 + 3\rho_n^4 \sum_{i\in\mathscr{N}_{b_n}} \left(\sum_{j=1}^{r} \frac{q_{ij}\zeta_j}{(\lambda_j^2+\rho_n)^2} \right)^2 \\ +3 \sum_{i\in\mathscr{N}_{b_n}} \left(\sum_{j=1}^{r}\sum_{l=1}^{n} q_{ij} \left(\frac{\lambda_j}{\lambda_j^2+\rho_n} + \frac{\rho_n\lambda_j}{(\lambda_j^2+\rho_n)^2} \right) p_{lj}\varepsilon_l \right)^2 \end{aligned} \tag{A.3.1}$$

From assumption 2, $\sum_{i\in\mathscr{N}_{b_n}} |\theta_i|^2 \leq \|\theta\|_2^2 = O(n^{2\alpha_\theta})$. Similarly

$$\rho_n^4 \sum_{i\in\mathscr{N}_{b_n}} \left(\sum_{j=1}^{r} \frac{q_{ij}\zeta_j}{(\lambda_j^2+\rho_n)^2} \right)^2 \leq \frac{\rho_n^4}{\lambda_r^8} \sum_{i\in\mathscr{N}_{b_n}} \sum_{j=1}^{r} q_{ij}^2 \sum_{j=1}^{r} \zeta_j^2 = \frac{\rho_n^4 \times |\mathscr{N}_{b_n}| \times \|\theta\|_2^2}{\lambda_r^8} = o(n^{-2\alpha_\sigma}) \tag{A.3.2}$$

From assumption 6,

$$\begin{aligned} &\mathbf{E} \sum_{i\in\mathscr{N}_{b_n}} \left(\sum_{j=1}^{r}\sum_{l=1}^{n} q_{ij} \left(\frac{\lambda_j}{\lambda_j^2+\rho_n} + \frac{\rho_n\lambda_j}{(\lambda_j^2+\rho_n)^2} \right) p_{lj}\varepsilon_l \right)^2 \\ &= \sigma^2 \sum_{i\in\mathscr{N}_{b_n}} \sum_{l=1}^{n} \left(\sum_{j=1}^{r} q_{ij} \left(\frac{\lambda_j}{\lambda_j^2+\rho_n} + \frac{\rho_n\lambda_j}{(\lambda_j^2+\rho_n)^2} \right) p_{lj} \right)^2 \\ &= \sigma^2 \sum_{i\in\mathscr{N}_{b_n}} \sum_{j=1}^{r} q_{ij}^2 \left(\frac{\lambda_j}{\lambda_j^2+\rho_n} + \frac{\rho_n\lambda_j}{(\lambda_j^2+\rho_n)^2} \right)^2 \leq \frac{4\sigma^2|\mathscr{N}_{b_n}|}{\lambda_r^2} \\ \Rightarrow &\sum_{i\in\mathscr{N}_{b_n}} \left(\sum_{j=1}^{r}\sum_{l=1}^{n} q_{ij} \left(\frac{\lambda_j}{\lambda_j^2+\rho_n} + \frac{\rho_n\lambda_j}{(\lambda_j^2+\rho_n)^2} \right) p_{lj}\varepsilon_l \right)^2 = O_p\left(n^{-2\alpha_\sigma}\right) \end{aligned} \tag{A.3.3}$$

Since $\alpha_\theta, \alpha_\sigma \geq 0$, $\|\widehat{\theta}\|_2 = O_p(n^{\alpha_\theta})$. According to (2.15) and (2.16), define $\widehat{\zeta} = Q^T\widehat{\theta}$,

$$\begin{aligned}\widetilde{\theta}^* - \widehat{\theta} = \left(I_p + \rho_n Q(\Lambda^2+\rho_n I_r)^{-1}Q^T\right) Q(\Lambda^2+\rho_n I_r)^{-1}\left(\Lambda^2 Q^T\widehat{\theta} + \Lambda P^T\varepsilon^*\right) \\ + \widehat{\theta}_\perp - QQ^T\widehat{\theta} - Q_\perp Q_\perp^T\widehat{\theta} \\ \Rightarrow \widetilde{\theta}_i^* - \widehat{\theta}_i = -\rho_n^2\sum_{j=1}^r \frac{q_{ij}\widehat{\zeta}_j}{(\lambda_j^2+\rho_n)^2} + \sum_{j=1}^r\sum_{l=1}^n q_{ij}\left(\frac{\lambda_j}{\lambda_j^2+\rho_n} + \frac{\rho_n\lambda_j}{(\lambda_j^2+\rho_n)^2}\right)p_{lj}\varepsilon_l^*\end{aligned} \tag{A.3.4}$$

Similar to (A.2.5), rewrite δ in assumption 2 as $\delta = \frac{\eta+\alpha_\theta+\delta_1}{2}$ with $\delta_1 > 0$, we have

$$\begin{aligned}&\max_{i=1,2,...,p}|\rho_n^2\sum_{j=1}^r\frac{q_{ij}\widehat{\zeta}_j}{(\lambda_j^2+\rho_n)^2}| \\ \leq &\max_{i=1,2,...,p}\frac{\rho_n^2}{\lambda_r^4}\sqrt{\sum_{j=1}^r q_{ij}^2}\times\sqrt{\sum_{j=1}^r\widehat{\zeta}_j^2} \leq \frac{\rho_n^2\|\widehat{\theta}\|_2}{c_\lambda^4 n^{4\eta}} = O_p\left(n^{-\eta-\delta_1}\right)\end{aligned} \tag{A.3.5}$$

$\varepsilon_i^*|\varepsilon, i = 1,2,...,n$ are normal random variables with mean 0 and variance $\widehat{\sigma}^2$. Therefore $\mathbf{E}^*|\varepsilon_1^*|^m = \widehat{\sigma}^m D$, $D = \mathbf{E}|Y|^m$, Y is a normal random variable with mean 0 and variance 1. If $\widehat{\sigma} > 0$, from (A.2.2) and lemma A.1.1, $\exists$ a constant E which depends on m and D such that for any $a > 0$,

$$Prob^*\left(\max_{i=1,2,...,p}|\sum_{j=1}^r\sum_{l=1}^n q_{ij}\left(\frac{\lambda_j}{\lambda_j^2+\rho_n} + \frac{\rho_n\lambda_j}{(\lambda_j^2+\rho_n)^2}\right)p_{lj}\frac{\varepsilon_l^*}{\widehat{\sigma}}| > \frac{a}{\widehat{\sigma}}\right) \leq \frac{pE\widehat{\sigma}^m}{\lambda_r^m a^m} \tag{A.3.6}$$

Suppose $\widehat{\mathscr{N}_{b_n}} = \mathscr{N}_{b_n}$, $\frac{\sigma}{2} < \widehat{\sigma} < \frac{3\sigma}{2}$, and $\max_{i=1,2,...,p}|\rho_n^2\sum_{j=1}^r\frac{q_{ij}\widehat{\zeta}_j}{(\lambda_j^2+\rho_n)^2}| \leq C\times n^{-\eta-\delta_1}$ for a

constant C. Since $\widehat{\theta}_i = 0$ if $i \notin \widehat{\mathscr{N}_{b_n}}$,

$$\begin{aligned}
Prob^*\left(\widehat{\mathscr{N}_{b_n}^*} \neq \mathscr{N}_{b_n}\right) &\leq Prob^*\left(\min_{i\in\mathscr{N}_{b_n}} |\widetilde{\theta}_i^*| \leq b_n\right) + Prob^*\left(\max_{i\notin\mathscr{N}_{b_n}} |\widetilde{\theta}_i^*| > b_n\right) \\
&\leq Prob^*(\min_{i\in\mathscr{N}_{b_n}} |\widehat{\theta}_i| - \max_{i\in\mathscr{N}_{b_n}} |\rho_n^2 \sum_{j=1}^{r} \frac{q_{ij}\widehat{\zeta}_j}{(\lambda_j^2+\rho_n)^2}| - b_n \\
&\leq \max_{i\in\mathscr{N}_{b_n}} |\sum_{j=1}^{r}\sum_{l=1}^{n} q_{ij}\left(\frac{\lambda_j}{\lambda_j^2+\rho_n} + \frac{\rho_n\lambda_j}{(\lambda_j^2+\rho_n)^2}\right) p_{lj}\varepsilon_l^*|) \qquad \text{(A.3.7)} \\
&+Prob^*(\max_{i\notin\mathscr{N}_{b_n}} |\sum_{j=1}^{r}\sum_{l=1}^{n} q_{ij}\left(\frac{\lambda_j}{\lambda_j^2+\rho_n} + \frac{\rho_n\lambda_j}{(\lambda_j^2+\rho_n)^2}\right) p_{lj}\varepsilon_l^*| \\
&> b_n - \rho_n^2 \max_{i\notin\mathscr{N}_{b_n}} |\sum_{j=1}^{r} \frac{q_{ij}\widehat{\zeta}_j}{(\lambda_j^2+\rho_n)^2}|)
\end{aligned}$$

From assumption 4, for sufficiently large n,

$$b_n - \rho_n^2 \max_{i\notin\mathscr{N}_{b_n}} |\sum_{j=1}^{r} \frac{q_{ij}\widehat{\zeta}_j}{(\lambda_j^2+\rho_n)^2}| \geq C_b n^{-\nu_b} - Cn^{-\eta-\delta_1} \geq \frac{b_n}{2} \qquad \text{(A.3.8)}$$

From (A.2.2), lemma A.1.1, assumption 1 and 4, we have

$$\max_{i=1,2,\ldots,p} |\sum_{j=1}^{r} q_{ij}\left(\frac{\lambda_j}{\lambda_j^2+\rho_n} + \frac{\rho_n\lambda_j}{(\lambda_j^2+\rho_n)^2}\right)\sum_{l=1}^{n} p_{lj}\varepsilon_l| = O_p\left(n^{\alpha_p/m-\eta}\right) \qquad \text{(A.3.9)}$$

Suppose a constant C such that
$\max_{i=1,2,\ldots,p} |\sum_{j=1}^{r} q_{ij}\left(\frac{\lambda_j}{\lambda_j^2+\rho_n} + \frac{\rho_n\lambda_j}{(\lambda_j^2+\rho_n)^2}\right)\sum_{l=1}^{n} p_{lj}\varepsilon_l| \leq Cn^{\alpha_p/m-\eta}$ (from lemma A.1.1), and (since $\frac{\rho_n^2\|\theta\|_2}{\lambda_r^4} = O(n^{-\eta-\delta_1})$) $\frac{\rho_n^2\|\theta\|_2}{\lambda_r^4} \leq Cn^{-\eta-\delta_1}$. From assumption 4, for sufficiently large n,

$$
\begin{aligned}
\min_{i\in\mathscr{N}_{b_n}}|\widehat{\theta}_i| &\geq \min_{i\in\mathscr{N}_{b_n}}|\theta_i| - \max_{i\in\mathscr{N}_{b_n}}\rho_n^2|\sum_{j=1}^{r}\frac{q_{ij}\zeta_j}{(\lambda_j^2+\rho_n)^2}| \\
&\quad - \max_{i\in\mathscr{N}_{b_n}}|\sum_{j=1}^{r}q_{ij}\left(\frac{\lambda_j}{\lambda_j^2+\rho_n}+\frac{\rho_n\lambda_j}{(\lambda_j^2+\rho_n)^2}\right)\sum_{l=1}^{n}p_{lj}\varepsilon_l| \\
\geq \frac{b_n}{c_b} - \frac{\rho_n^2\|\theta\|_2}{\lambda_r^4} - Cn^{\alpha_p/m-\eta} &\Rightarrow \min_{i\in\mathscr{N}_{b_n}}|\widehat{\theta}_i| - \max_{i\in\mathscr{N}_{b_n}}|\rho_n^2\sum_{j=1}^{r}\frac{q_{ij}\widehat{\zeta}_j}{(\lambda_j^2+\rho_n)^2}| - b_n \\
&\geq \left(\frac{1}{c_b}-1\right)b_n - Cn^{-\eta-\delta_1} - Cn^{\alpha_p/m-\eta} - Cn^{-\eta-\delta_1} > \frac{b_n}{2}\left(\frac{1}{c_b}-1\right)
\end{aligned}
\tag{A.3.10}
$$

Correspondingly

$$
\begin{aligned}
Prob^*\left(\widehat{\mathscr{N}_{b_n}^*}\neq\mathscr{N}_{b_n}\right) &\leq Prob^*(\max_{i\in\mathscr{N}_{b_n}}|\sum_{j=1}^{r}\sum_{l=1}^{n}q_{ij}\left(\frac{\lambda_j}{\lambda_j^2+\rho_n}+\frac{\rho_n\lambda_j}{(\lambda_j^2+\rho_n)^2}\right)p_{lj}\varepsilon_l^*| \\
&\qquad > \frac{b_n}{2}(\frac{1}{c_b}-1)) \\
&\quad + Prob^*\left(\max_{i\notin\mathscr{N}_{b_n}}|\sum_{j=1}^{r}\sum_{l=1}^{n}q_{ij}\left(\frac{\lambda_j}{\lambda_j^2+\rho_n}+\frac{\rho_n\lambda_j}{(\lambda_j^2+\rho_n)^2}\right)p_{lj}\varepsilon_l^*| > b_n/2\right) \\
&\leq \frac{pE\widehat{\sigma}^m}{c_\lambda^m n^{m\eta}b_n^m}\times\left(\frac{2^m}{(1/c_b-1)^m}+2^m\right)
\end{aligned}
\tag{A.3.11}
$$

which has order $O_p(n^{\alpha_p+m\nu_b-m\eta})$. If $\widehat{\mathscr{N}_{b_n}^*}=\mathscr{N}_{b_n}$, then $\widehat{\tau}_i^*=\tau_i$ for $i=1,2,...,p_1$. Similar to (A.2.14),

$$
\begin{aligned}
\max_{i=1,2,...,p_1}\frac{|\widehat{\gamma}_i^*-\widehat{\gamma}_i|}{\widehat{\tau}_i^*} &= \max_{i=1,2,...,p_1}\frac{|-\rho_n^2\sum_{k=1}^{r}\frac{c_{ik}\widehat{\zeta}_k}{(\lambda_k^2+\rho_n)^2}+\sum_{l=1}^{n}\sum_{k=1}^{r}c_{ik}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)p_{lk}\varepsilon_l^*|}{\tau_i} \\
&\leq \max_{i\in\mathscr{M}}\rho_n^2\frac{|\sum_{k=1}^{r}\frac{c_{ik}\widehat{\zeta}_k}{(\lambda_k^2+\rho_n)^2}|}{\tau_i}+\max_{i\in\mathscr{M}}\frac{|\sum_{l=1}^{n}\sum_{k=1}^{r}c_{ik}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)p_{lk}\varepsilon_l^*|}{\tau_i} \\
&\leq \frac{\rho_n^2\|\widehat{\theta}\|_2}{\lambda_r^3}+\max_{i\in\mathscr{M}}\frac{|\sum_{l=1}^{n}\sum_{k=1}^{r}c_{ik}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)p_{lk}\varepsilon_l^*|}{\tau_i} \\
\max_{i=1,2,...,p_1}\frac{|\widehat{\gamma}_i^*-\widehat{\gamma}_i|}{\widehat{\tau}_i^*} &\geq \max_{i\in\mathscr{M}}\frac{|\sum_{l=1}^{n}\sum_{k=1}^{r}c_{ik}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)p_{lk}\varepsilon_l^*|}{\tau_i}-\frac{\rho_n^2\|\widehat{\theta}\|_2}{\lambda_r^3}
\end{aligned}
\tag{A.3.12}
$$

From theorem 1, for any $a>0$ and sufficiently large n, $\exists$ constant D_a such that $|\widehat{\sigma}^2-\sigma^2|\leq D_a n^{-\alpha_\sigma}$ and $\frac{1}{2}\sigma<\widehat{\sigma}<\frac{3}{2}\sigma$ with probability $1-a$,

$$|\sigma-\widehat{\sigma}|=\frac{|\sigma^2-\widehat{\sigma}^2|}{\sigma+\widehat{\sigma}}\leq\frac{D_a n^{-\alpha_\sigma}}{\sigma} \tag{A.3.13}$$

If $0<x\leq n^{\alpha_\sigma/2}$, according to lemma A.1.2, assumption 7 and (A.2.17), $\exists$ a constant $C^{'}$ which only depends on $\sigma, c_{\mathscr{M}}, C_\lambda$ such that

$$\begin{aligned}
&|Prob^*\left(\max_{i\in\mathscr{M}}\frac{|\sum_{l=1}^n\sum_{k=1}^r c_{ik}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)p_{lk}\varepsilon_l^*|}{\tau_i}\leq x\right)-H(x)|\\
&\qquad=|H(\frac{x\sigma}{\widehat{\sigma}})-H(x)|\\
&\qquad\leq C^{'}\left(1+\sqrt{\log(|\mathscr{M}|)}+\sqrt{|\log(\frac{x|\sigma-\widehat{\sigma}|}{\widehat{\sigma}})|}\right)\frac{x|\sigma-\widehat{\sigma}|}{\widehat{\sigma}}\\
&\qquad\leq\frac{2D_aC^{'}}{\sigma^2}\left(1+\sqrt{\log(n)}\right)n^{-\alpha_\sigma/2}\\
&\qquad+C^{'}\sqrt{\frac{x|\sigma-\widehat{\sigma}|}{\widehat{\sigma}}|\log(\frac{x|\sigma-\widehat{\sigma}|}{\widehat{\sigma}})|}\times\sqrt{\frac{2D_a}{\sigma^2}}n^{-\alpha_\sigma/4}
\end{aligned} \tag{A.3.14}$$

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

On the other hand, if $x>n^{\alpha_\sigma/2}$, then $\frac{x\sigma}{\widehat{\sigma}}>\frac{2n^{\alpha_\sigma/2}}{3}$. From lemma A.1.1, we may choose sufficiently large m_1 such that $m_1\alpha_\sigma/2>2$, since $\mathbf{E}|\xi_1|^{m_1}<\infty$(Here ξ_1 is a normal random variable with mean 0 and variance σ^2) is a constant for given m_1 and

$\max_{i\in\mathscr{M}}\sum_{k=1}^{r}\frac{1}{\tau_i^2}c_{ik}^2\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)^2\le 1$, we have

$$
\begin{aligned}
&Prob\left(\max_{i\in\mathscr{M}}\frac{|\sum_{k=1}^{r}c_{ik}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)\xi_k|}{\tau_i}>\frac{2n^{\alpha_\sigma/2}}{3}\right)\le\frac{3^{m_1}|\mathscr{M}|\times E}{2^{m_1}n^{m_1\alpha_\sigma/2}}\\
&\Rightarrow|Prob^*\left(\max_{i\in\mathscr{M}}\frac{|\sum_{l=1}^{n}\sum_{k=1}^{r}c_{ik}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)p_{lk}\varepsilon_l^*|}{\tau_i}\le x\right)-H(x)|\\
&\qquad\le Prob\left(\max_{i\in\mathscr{M}}\frac{|\sum_{k=1}^{r}c_{ik}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)\xi_k|}{\tau_i}>\frac{2n^{\alpha_\sigma/2}}{3}\right)\\
&\qquad\quad+Prob\left(\max_{i\in\mathscr{M}}\frac{|\sum_{k=1}^{r}c_{ik}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)\xi_k|}{\tau_i}>n^{\alpha_\sigma/2}\right)\\
&\qquad\qquad\le 2\times\frac{3^{m_1}|\mathscr{M}|\times E}{2^{m_1}n^{m_1\alpha_\sigma/2}}
\end{aligned}
\tag{A.3.15}
$$

Since $H(0)=0$, from (A.3.14) and (A.3.15), for any given $a>0$ and sufficiently large n,

$$
\sup_{x\ge 0}|Prob^*\left(\max_{i\in\mathscr{M}}\frac{|\sum_{l=1}^{n}\sum_{k=1}^{r}c_{ik}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)p_{lk}\varepsilon_l^*|}{\tau_i}\le x\right)-H(x)|<a \tag{A.3.16}
$$

As a summary, for any given $a>0$, $\exists$ a constant D_a such that for sufficiently large n, the event $|\widehat{\sigma}^2-\sigma^2|\le D_a n^{-\alpha_\sigma}$, $\frac{1}{2}\sigma<\widehat{\sigma}<\frac{3}{2}\sigma$, $\widehat{\mathscr{N}_{b_n}}=\mathscr{N}_{b_n}$, $\|\widehat{\theta}\|_2\le D_a\times n^{\alpha_\theta}\Rightarrow\frac{\rho_n^2\|\widehat{\theta}\|_2}{\lambda_r^3}\le D_a' n^{-\delta_1}$ for constant D_a' and $\max_{i=1,2,\ldots,p}|\rho_n^2\sum_{j=1}^{r}\frac{q_{ij}\widehat{\zeta}_j}{(\lambda_j^2+\rho_n)^2}|\le D_a\times n^{-\eta-\delta_1}$ happen with probability $1-a$. From (A.3.12), assumption 5 and lemma A.1.2, we have for any $x\ge 0$, there exists a constant C'

such that

$$
\begin{aligned}
Prob^*\left(\max_{i=1,2,\ldots,p_1}\frac{|\widehat{\gamma}_i^*-\widehat{\gamma}_i|}{\widehat{\tau}_i^*}\le x\right)-H(x) &\le Prob^*\left(\widehat{\mathscr{N}_{b_n}^*}\neq\mathscr{N}_{b_n}\right)\\
&+C'\frac{\rho_n^2\|\widehat{\theta}\|_2}{\lambda_r^3}\times\left(1+\sqrt{\log(|\mathscr{M}|)}+\sqrt{|\log(\frac{\rho_n^2\|\widehat{\theta}\|_2}{\lambda_r^3})|}\right)\\
+Prob^*\left(\max_{i\in\mathscr{M}}\frac{|\sum_{l=1}^n\sum_{k=1}^r c_{ik}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)p_{lk}\varepsilon_l^*|}{\tau_i}\le x+\frac{\rho_n^2\|\widehat{\theta}\|_2}{\lambda_r^3}\right)&\\
-H(x+\frac{\rho_n^2\|\widehat{\theta}\|_2}{\lambda_r^3})&\\
\le a+C'D_a'(1+\sqrt{\log(n)})n^{-\delta_1}+C'\sqrt{D_a'}n^{-\delta_1/2}\sqrt{|\log(\frac{\rho_n^2\|\widehat{\theta}\|_2}{\lambda_r^3})\times\frac{\rho_n^2\|\widehat{\theta}\|_2}{\lambda_r^3}|}&\\
+Prob^*\left(\widehat{\mathscr{N}_{b_n}^*}\neq\mathscr{N}_{b_n}\right)&\qquad\text{(A.3.17)}\\
Prob^*\left(\max_{i=1,2,\ldots,p_1}\frac{|\widehat{\gamma}_i^*-\widehat{\gamma}_i|}{\widehat{\tau}_i^*}\le x\right)-H(x)&\\
\ge Prob^*\left(\max_{i=1,2,\ldots,p_1}\frac{|\widehat{\gamma}_i^*-\widehat{\gamma}_i|}{\widehat{\tau}_i^*}\le x\cap\widehat{\mathscr{N}_{b_n}^*}=\mathscr{N}_{b_n}\right)-H(x)&\\
\ge Prob^*\left(\max_{i\in\mathscr{M}}\frac{|\sum_{l=1}^n\sum_{k=1}^r c_{ik}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)p_{lk}\varepsilon_l^*|}{\tau_i}\le x-\frac{\rho_n^2\|\widehat{\theta}\|_2}{\lambda_r^3}\right)&\\
-H(x-\frac{\rho_n^2\|\widehat{\theta}\|_2}{\lambda_r^3})&\\
-Prob^*\left(\widehat{\mathscr{N}_{b_n}^*}\neq\mathscr{N}_{b_n}\right)&\\
-C'D_a'(1+\sqrt{\log(n)})n^{-\delta_1}-C'\sqrt{D_a'}n^{-\delta_1/2}\sqrt{|\log(\frac{\rho_n^2\|\widehat{\theta}\|_2}{\lambda_r^3})\times\frac{\rho_n^2\|\widehat{\theta}\|_2}{\lambda_r^3}|}&
\end{aligned}
$$

If $0\le x\le\frac{\rho_n^2\|\widehat{\theta}\|_2}{\lambda_r^3}$, then
$Prob^*\left(\max_{i\in\mathscr{M}}\frac{|\sum_{l=1}^n\sum_{k=1}^r c_{ik}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)p_{lk}\varepsilon_l^*|}{\tau_i}\le x-\frac{\rho_n^2\|\widehat{\theta}\|_2}{\lambda_r^3}\right)=H(x-\frac{\rho_n^2\|\widehat{\theta}\|_2}{\lambda_r^3})=0$. There-

fore, for sufficiently large n, from (A.3.16) and (A.3.11), $\exists$ a constant C such that

$$\begin{aligned}
&\sup_{x\geq 0}|Prob^*\left(\max_{i=1,2,\ldots,p_1}\frac{|\widehat{\gamma}_i^*-\widehat{\gamma}_i|}{\widehat{\tau}_i^*}\leq x\right)-H(x)| \\
&\leq \frac{pE\widehat{\sigma}^m}{c_\lambda^m n^{m\eta}b_n^m}\times\left(2^m+\frac{2^m}{(1/c_b-1)^m}\right)+a \\
&+C^{'}D_a^{'}(1+\sqrt{\log(n)})n^{-\delta_1}+C^{'}\sqrt{D_a^{'}}n^{-\delta_1/2}\sqrt{\sup_{x\in(0,1]}|x\log(x)|} \\
&\leq Cn^{m(\nu_b+\alpha_p/m-\eta)}+2a
\end{aligned} \tag{A.3.18}$$

and we prove (2.32).

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

$$Prob\left(\sup_{x\geq 0}|Prob^*\left(\max_{i=1,2,\ldots,p_1}\frac{|\widehat{\gamma}_i^*-\widehat{\gamma}_i|}{\widehat{\tau}_i^*}\leq x\right)-H(x)|\leq a\right)>1-a \tag{A.3.19}$$

Choose sufficiently small a such that $0<1-\alpha-2a<1-\alpha+2a<1$. If (A.3.19) happens, for any $1>\alpha>0$, define $c_{1-\alpha}$ as the $1-\alpha$ quantile of $H(x)$,

$$\begin{aligned}
&Prob^*\left(\max_{i=1,2,\ldots,p_1}\frac{|\widehat{\gamma}_i^*-\widehat{\gamma}_i|}{\widehat{\tau}_i^*}\leq c_{1-\alpha+2a}\right)-(1-\alpha+2a)\geq -a \\
&\Rightarrow c_{1-\alpha}^*\leq c_{1-\alpha+2a} \\
&Prob^*\left(\max_{i=1,2,\ldots,p_1}\frac{|\widehat{\gamma}_i^*-\widehat{\gamma}_i|}{\widehat{\tau}_i^*}\leq c_{1-\alpha-2a}\right)-(1-\alpha-2a)\leq a \\
&\Rightarrow c_{1-\alpha}^*>c_{1-\alpha-2a}
\end{aligned} \tag{A.3.20}$$

From theorem 2, we have for sufficiently large n,

$$
\begin{aligned}
& Prob\left(\max_{i=1,2,\ldots,p_1}\frac{|\widehat{\gamma}_i-\gamma_i|}{\widehat{\tau}_i}\le c^*_{1-\alpha}\right)\\
\le\ & Prob\left(\sup_{x\ge 0}|Prob^*\left(\max_{i=1,2,\ldots,p_1}\frac{|\widehat{\gamma}^*_i-\widehat{\gamma}_i|}{\widehat{\tau}^*_i}\le x\right)-H(x)|>a\right)\\
& +Prob\left(\max_{i=1,2,\ldots,p_1}\frac{|\widehat{\gamma}_i-\gamma_i|}{\widehat{\tau}_i}\le c_{1-\alpha+2a}\right)\\
\le\ & a+(H(c_{1-\alpha+2a})+a)=1-\alpha+4a\\
& Prob\left(\max_{i=1,2,\ldots,p_1}\frac{|\widehat{\gamma}_i-\gamma_i|}{\widehat{\tau}_i}\le c^*_{1-\alpha}\right)\\
\ge\ & Prob(\max_{i=1,2,\ldots,p_1}\frac{|\widehat{\gamma}_i-\gamma_i|}{\widehat{\tau}_i}\le c^*_{1-\alpha}\\
& \cap\sup_{x\ge 0}|Prob^*\left(\max_{i=1,2,\ldots,p_1}\frac{|\widehat{\gamma}^*_i-\widehat{\gamma}_i|}{\widehat{\tau}^*_i}\le x\right)-H(x)|\le a)\\
\ge\ & Prob\left(\max_{i=1,2,\ldots,p_1}\frac{|\widehat{\gamma}_i-\gamma_i|}{\widehat{\tau}_i}\le c_{1-\alpha-2a}\right)\\
& -Prob\left(\sup_{x\ge 0}|Prob^*\left(\max_{i=1,2,\ldots,p_1}\frac{|\widehat{\gamma}^*_i-\widehat{\gamma}_i|}{\widehat{\tau}^*_i}\le x\right)-H(x)|>a\right)\\
\ge\ & (H(c_{1-\alpha-2a})-a)-a=1-\alpha-4a\\
\Rightarrow\ & |Prob\left(\max_{i=1,2,\ldots,p_1}\frac{|\widehat{\gamma}_i-\gamma_i|}{\widehat{\tau}_i}\le c^*_{1-\alpha}\right)-(1-\alpha)|\le 4a
\end{aligned}
\tag{A.3.21}
$$

For $a>0$ can be arbitrarily small, we prove (2.31). □

A.4 Proofs of theorems in section 5

Proof of lemma 1. Define the design matrix $X=(x_{ij})_{i=1,\ldots,n,j=1,\ldots,p}$, $\bar{x}_j=\frac{1}{n}\sum_{i=1}^n x_{ij}$, and $x^{'}_{ij}=x_{ij}-\bar{x}_j$. If $\widehat{\mathscr{N}_{b_n}}=\mathscr{N}_{b_n}$, for $i=1,2,\ldots,n$,

$$
\begin{aligned}
\widehat{\varepsilon}^{'}_i &= \varepsilon_i+\sum_{j\notin\mathscr{N}_{b_n}}x_{ij}\theta_j-\sum_{j\in\mathscr{N}_{b_n}}x_{ij}(\widetilde{\theta}_j-\theta_j)\\
\Rightarrow\ \widehat{\varepsilon}_i &= \varepsilon_i-\frac{1}{n}\sum_{i=1}^n\varepsilon_i+\sum_{j\notin\mathscr{N}_{b_n}}x^{'}_{ij}\theta_j-\sum_{j\in\mathscr{N}_{b_n}}x^{'}_{ij}(\widetilde{\theta}_j-\theta_j)
\end{aligned}
\tag{A.4.1}
$$

Define $\widetilde{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\varepsilon_i \le x},\ x \in \mathbf{R}$. From (A.1.17), for any given $\psi > 0$,

$$
\begin{aligned}
\widehat{F}(x) - F(x) &= \left(\widehat{F}(x) - \widetilde{F}(x + 1/\psi)\right) + \left(\widetilde{F}(x + 1/\psi) - F(x + 1/\psi)\right) \\
&\quad + (F(x + 1/\psi) - F(x)) \\
&\le \frac{1}{n}\sum_{i=1}^{n} (g_{\psi,x}(\widehat{\varepsilon}_i) - g_{\psi,x}(\varepsilon_i)) + \sup_{x \in \mathbf{R}} |\widetilde{F}(x) - F(x)| + (F(x + 1/\psi) - F(x)) \\
&\le g_* \psi \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\widehat{\varepsilon}_i - \varepsilon_i)^2} + \sup_{x \in \mathbf{R}} |\widetilde{F}(x) - F(x)| + (F(x + 1/\psi) - F(x)) \\
\widehat{F}(x) - F(x) &= \left(\widehat{F}(x) - \widetilde{F}(x - 1/\psi)\right) + \left(\widetilde{F}(x - 1/\psi) - F(x - 1/\psi)\right) \\
&\quad - (F(x) - F(x - 1/\psi)) \qquad \text{(A.4.2)} \\
&\ge \frac{1}{n}\sum_{i=1}^{n} (g_{\psi,x-1/\psi}(\widehat{\varepsilon}_i) - g_{\psi,x-1/\psi}(\varepsilon_i)) - \sup_{x \in \mathbf{R}} |\widetilde{F}(x) - F(x)| - (F(x) - F(x - 1/\psi)) \\
&\ge -g_* \psi \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\widehat{\varepsilon}_i - \varepsilon_i)^2} - \sup_{x \in \mathbf{R}} |\widetilde{F}(x) - F(x)| - (F(x) - F(x - 1/\psi)) \\
&\Rightarrow \sup_{x \in \mathbf{R}} |\widehat{F}(x) - F(x)| \le g_* \psi \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\widehat{\varepsilon}_i - \varepsilon_i)^2} + \sup_{x \in \mathbf{R}} |\widetilde{F}(x) - F(x)| \\
&\quad + \sup_{x \in \mathbf{R}} |F(x + 1/\psi) - F(x)|
\end{aligned}
$$

Suppose assumption 1 to 6. From (A.2.7), (A.2.8), (A.2.9) and $\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i = O_p(1/\sqrt{n})$, for any $0<a<1$, $\exists$ a constant C_a such that with probability at least $1-a$

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(\widehat{\varepsilon}_i-\varepsilon_i)^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j\notin\mathscr{N}_{b_n}} x_{ij}^{'}\theta_j - \sum_{j\in\mathscr{N}_{b_n}} x_{ij}^{'}(\widetilde{\theta}_j-\theta_j) - \frac{1}{n}\sum_{j=1}^{n}\varepsilon_j\right)^2 \\
&\le \frac{3}{n}\sum_{i=1}^{n}\left(\sum_{j\notin\mathscr{N}_{b_n}} x_{ij}^{'}\theta_j\right)^2 + \frac{3}{n}\sum_{i=1}^{n}\left(\sum_{j\in\mathscr{N}_{b_n}} x_{ij}^{'}(\widetilde{\theta}_j-\theta_j)\right)^2 + 3\left(\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j\right)^2 \\
&\le \frac{6}{n}\sum_{i=1}^{n}\left(\sum_{j\notin\mathscr{N}_{b_n}} x_{ij}\theta_j\right)^2 + 6\left(\sum_{j\notin\mathscr{N}_{b_n}} \overline{x}_j\theta_j\right)^2 + \frac{6}{n}\sum_{i=1}^{n}\left(\sum_{j\in\mathscr{N}_{b_n}} x_{ij}(\widetilde{\theta}_j-\theta_j)\right)^2 \\
&\qquad +6\left(\sum_{j\in\mathscr{N}_{b_n}} \overline{x}_j(\widetilde{\theta}_j-\theta_j)\right)^2 + 3\left(\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j\right)^2 \\
&\le C_a n^{-\alpha_\sigma} + \frac{6}{n^2}\left(\sum_{i=1}^{n}\sum_{j\notin\mathscr{N}_{b_n}} x_{ij}\theta_j\right)^2 + C_a|\mathscr{N}_{b_n}|(n^{2\alpha_\theta-4\delta}+n^{-2\eta}) \\
&\qquad +\frac{6}{n^2}\left(\sum_{i=1}^{n}\sum_{j\in\mathscr{N}_{b_n}} x_{ij}(\widetilde{\theta}_j-\theta_j)\right)^2 + \frac{C_a}{n} \\
&\le C_a n^{-\alpha_\sigma} + \frac{6}{n}\sum_{i=1}^{n}\left(\sum_{j\notin\mathscr{N}_{b_n}} x_{ij}\theta_j\right)^2 + C_a|\mathscr{N}_{b_n}|(n^{2\alpha_\theta-4\delta}+n^{-2\eta}) \\
&\qquad +\frac{6}{n}\sum_{i=1}^{n}\left(\sum_{j\in\mathscr{N}_{b_n}} x_{ij}(\widetilde{\theta}_j-\theta_j)\right)^2 + \frac{C_a}{n} \\
&\Rightarrow \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{\varepsilon}_i-\varepsilon_i)^2} = O_p(n^{-\alpha_\sigma/2})
\end{aligned}
\tag{A.4.3}
$$

According to Gilvenko-Cantelli lemma, $\sup_{x\in\mathbf{R}}|\widetilde{F}(x)-F(x)|\to 0$ almost surely. Therefore, for any $a>0$ and sufficiently large n, $Prob\left(\sup_{x\in\mathbf{R}}|\widetilde{F}(x)-F(x)|\le a\right)>1-a$. Choose sufficiently small a and $\psi = 1/a$, from assumption 8 and (C.2.40), we prove (2.35). □

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

$$
\max_{i=1,2,\ldots,p_1}\left|\sum_{j=1}^{p} x_{f,ij}\widehat{\theta}_j - \sum_{j=1}^{p} x_{f,ij}\beta_j\right| = O_p(n^{-\eta}) \tag{A.4.4}
$$

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

$Prob\left(\max_{i=1,2,...,p_1} |\sum_{j=1}^{p} x_{f,ij}\widehat{\theta}_j - \sum_{j=1}^{p} x_{f,ij}\beta_j| \le C_a n^{-\eta}\right) \ge 1-a$ for any $n = 1,2,....$ Define

$F^{-}(x) = \lim_{y<x, y\to x} F(y)$ for any $x \in \mathbf{R}$, and

$G(x) = Prob\left(\max_{i=1,2,...,p_1} |\varepsilon_{f,i}| \le x\right) = (F(x) - F^{-}(-x))^{p_1}$ for $x \ge 0$. G is continuous according to assumption 8. With probability at least $1-a$

$$
\begin{aligned}
&\sup_{x\ge 0} |Prob^*\left(\max_{i=1,2,...,p_1} |y_{f,i} - \sum_{j=1}^{p} x_{f,ij}\widehat{\theta}_j| \le x\right) - G(x)| \\
\le &\sup_{x\ge 0} |Prob^*\left(\max_{i=1,2,...,p_1} |\varepsilon_{f,i}| \le x + \max_{i=1,2,...,p_1} |\sum_{j=1}^{p} x_{f,ij}(\beta_j - \widehat{\theta}_j)|\right) - G(x)| \\
+ &\sup_{x\ge 0} |Prob^*\left(\max_{i=1,2,...,p_1} |\varepsilon_{f,i}| \le x - \max_{i=1,2,...,p_1} |\sum_{j=1}^{p} x_{f,ij}(\beta_j - \widehat{\theta}_j)|\right) - G(x)| \\
\le &\sup_{x\ge 0} |G(x + C_a n^{-\eta}) - G(x)| + \sup_{x\ge 0} |G(x) - G(\max(0, x - C_a n^{-\eta}))|
\end{aligned}
\tag{A.4.5}
$$

For any $\delta > 0$ and any $x \ge 0$

$$
\begin{aligned}
&G(x+\delta) - G(x) \\
= &\sum_{i=1}^{p_1} (F(x+\delta) - F(-x-\delta))^{i-1} \times (F(x) - F(-x))^{p_1 - i} \\
&\times (F(x+\delta) - F(-x-\delta) - F(x) + F(-x)) \\
\le &2p_1 \times \sup_{x\in\mathbf{R}} (F(x+\delta) - F(x)) \Rightarrow \sup_{x\ge 0} (G(x+\delta) - G(x)) \\
&\le 2p_1 \times \sup_{x\in\mathbf{R}} (F(x+\delta) - F(x))
\end{aligned}
\tag{A.4.6}
$$

From (A.4.5) and assumption 8

$$
\sup_{x\ge 0} |Prob^*\left(\max_{i=1,2,...,p_1} |y_{f,i} - \sum_{j=1}^{p} x_{f,ij}\widehat{\theta}_j| \le x\right) - G(x)| = o_p(1) \tag{A.4.7}
$$

If $\widehat{\mathscr{N}_{b_n}} = \mathscr{N}_{b_n}$, $\frac{\sigma}{2} < \widehat{\sigma} < \frac{3\sigma}{2}$, and $\|\widehat{\theta}\|_2 \leq C \times n^{\alpha_\theta}$ (see (A.3.1) to (A.3.3)), then

$$\begin{aligned} \max_{i=1,2,\ldots,p} |\rho_n^2 \sum_{j=1}^{r} \frac{q_{ij}\widehat{\zeta}_j}{(\lambda_j^2+\rho_n)^2}| &\leq Cn^{-\eta-\delta_1} \\ \max_{i=1,2,\ldots,p} |\sum_{j=1}^{r} q_{ij}\left(\frac{\lambda_j}{\lambda_j^2+\rho_n} + \frac{\rho_n\lambda_j}{(\lambda_j^2+\rho_n)^2}\right)\sum_{l=1}^{n} p_{lj}\varepsilon_l| &\leq Cn^{\alpha_p/m-\eta} \end{aligned} \tag{A.4.8}$$

for some constant C with probability at least $1-a$. Here $2\delta = \eta + \alpha_\theta + \delta_1$. From (A.3.11), $\exists$ a constant E such that

$$Prob^*\left(\widehat{\mathscr{N}_{b_n}^*} \neq \mathscr{N}_{b_n}\right) \leq \frac{Ep}{n^{m\eta}b_n^m} \tag{A.4.9}$$

If $\widehat{\mathscr{N}_{b_n}^*} = \mathscr{N}_{b_n}$,

$$\begin{aligned} |\sum_{j=1}^{p} x_{f,ij}\widehat{\theta}_j^* - \sum_{j=1}^{p} x_{f,ij}\widehat{\theta}_j| &= |\sum_{j\in\mathscr{N}_{b_n}} x_{f,ij}(\widetilde{\theta}_j^* - \widehat{\theta}_j)| \\ \leq \rho_n^2|\sum_{k=1}^{r}\frac{c_{ik}\widetilde{\zeta}_k}{(\lambda_k^2+\rho_n)^2}| + |\sum_{k=1}^{r}\sum_{l=1}^{n} c_{ik}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n} + \frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)p_{lk}\varepsilon_l^*| \\ \leq \rho_n^2\frac{\sqrt{C_{\mathscr{M}}}\|\widehat{\theta}\|_2}{\lambda_r^4} + |\sum_{k=1}^{r}\sum_{l=1}^{n} c_{ik}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n} + \frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)p_{lk}\varepsilon_l^*| \end{aligned} \tag{A.4.10}$$

Form (A.2.5) and lemma A.1.1, $\exists$ a constant E which only depends on m, and for any $1 > a > 0$ with a sufficiently large $C_a > 0$

$$\begin{aligned} Prob^*\left(\max_{i=1,2,\ldots,p_1} |\sum_{k=1}^{r}\sum_{l=1}^{n} c_{ik}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n} + \frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)p_{lk}\frac{\varepsilon_l^*}{\widehat{\sigma}}| > \frac{C_a n^{-\eta}}{\widehat{\sigma}}\right) \\ \leq \frac{p_1 E\widehat{\sigma}^m}{n^{m\eta}C_a^m n^{-m\eta}} < a \end{aligned} \tag{A.4.11}$$

Combine with (A.4.9), there exists a constant C_a, with conditional probability at least $1-a$

$$\begin{aligned}
&\max_{i=1,2,\ldots,p_1} |\sum_{j=1}^{p} x_{f,ij}\widehat{\theta}_j^* - \sum_{j=1}^{p} x_{f,ij}\widehat{\theta}_j| \le C_a n^{-\eta} \\
&\Rightarrow Prob^*\left(\max_{i=1,2,\ldots,p_1} |y_{f,i}^* - \widehat{y}_{f,i}^*| \le x\right) - G(x) \\
&\le a + Prob^*\left(\max_{i=1,2,\ldots,p_1} |\varepsilon_{f,i}^*| \le x + C_a n^{-\eta}\right) - G(x) \\
&\le a + \sup_{x\ge 0} |Prob^*\left(\max_{i=1,2,\ldots,p_1} |\varepsilon_{f,i}^*| \le x\right) - G(x)| \\
&\qquad + 2p_1 \sup_{x\in\mathbf{R}} (F(x+C_a n^{-\eta}) - F(x)) \qquad \text{(A.4.12)} \\
&Prob^*\left(\max_{i=1,2,\ldots,p_1} |y_{f,i}^* - \widehat{y}_{f,i}^*| \le x\right) - G(x) \\
&\ge -a + Prob^*\left(\max_{i=1,2,\ldots,p_1} |\varepsilon_{f,i}^*| \le x - C_a n^{-\eta}\right) - G(x) \\
&\ge -a + Prob^*\left(\max_{i=1,2,\ldots,p_1} |\varepsilon_{f,i}^*| \le x - C_a n^{-\eta}\right) \\
&\qquad - G(x - C_a n^{-\eta}) - 2p_1 \sup_{x\in\mathbf{R}} (F(x+C_a n^{-\eta}) - F(x))
\end{aligned}$$

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

$$\begin{aligned}
&\sup_{x\ge 0} |Prob^*\left(\max_{i=1,2,\ldots,p_1} |y_{f,i}^* - \widehat{y}_{f,i}^*| \le x\right) - G(x)| \\
&\le a + \sup_{x\ge 0} |Prob^*\left(\max_{i=1,2,\ldots,p_1} |\varepsilon_{f,i}^*| \le x\right) - G(x)| \qquad \text{(A.4.13)} \\
&\qquad + 2p_1 \sup_{x\in\mathbf{R}} (F(x+C_a n^{-\eta}) - F(x))
\end{aligned}$$

From lemma 1, for any $x \geq 0$,

$$
\begin{aligned}
& |Prob^*\left(\max_{i=1,2,\ldots,p_1} |\varepsilon^*_{f,i}| \leq x\right) - G(x)| \\
= & |\left(\widehat{F}(x) - \widehat{F}^-(-x)\right)^{p_1} - (F(x) - F(-x))^{p_1}| \\
\leq & \sum_{i=1}^{p_1} |\widehat{F}(x) - \widehat{F}^-(-x)|^{i-1} \times |F(x) - F(-x)|^{p_1 - i} \\
& \times \left(|\widehat{F}(x) - F(x)| + |\widehat{F}^-(-x) - F^-(-x)|\right) \\
& \leq 2p_1 \sup_{x \in \mathbf{R}} |\widehat{F}(x) - F(x)| \to_p 0
\end{aligned}
\tag{A.4.14}
$$

as $n \to \infty$. From theorem 1 and (A.3.1) to (A.3.3), for any $1 > a > 0$, with probability at least $1 - a$ $\exists$ a constant $C_a > 0$ such that for sufficiently large n, (A.4.8) happens with $C = C_a$ and $\sup_{x \geq 0} |Prob^*\left(\max_{i=1,2,\ldots,p_1} |\varepsilon^*_{f,i}| \leq x\right) - G(x)| < a$. Correspondingly for sufficiently large n, with probability at least $1 - a$,

$$
\begin{aligned}
& \sup_{x \geq 0} |Prob^*\left(\max_{i=1,2,\ldots,p_1} |y^*_{f,i} - \widehat{y}^*_{f,i}| \leq x\right) - Prob^*\left(\max_{i=1,2,\ldots,p_1} |y_{f,i} - \widehat{y}_{f,i}| \leq x\right)| \\
& \leq \sup_{x \geq 0} |Prob^*\left(\max_{i=1,2,\ldots,p_1} |y^*_{f,i} - \widehat{y}^*_{f,i}| \leq x\right) - G(x)| \\
& + \sup_{x \geq 0} |Prob^*\left(\max_{i=1,2,\ldots,p_1} |y_{f,i} - \widehat{y}_{f,i}| \leq x\right) - G(x)| \\
& \leq a + \sup_{x \geq 0} |Prob^*\left(\max_{i=1,2,\ldots,p_1} |\varepsilon^*_{f,i}| \leq x\right) - G(x)| \\
& + 2p_1 \sup_{x \in \mathbf{R}} (F(x + C_a n^{-\eta}) - F(x)) + a \leq 4a
\end{aligned}
\tag{A.4.15}
$$

and we prove (2.37).

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 $1 - a$, $\sup_{x \geq 0} |Prob^*\left(\max_{i=1,2,\ldots,p_1} |y^*_{f,i} - \widehat{y}^*_{f,i}| \leq x\right) - G(x)| < a/2$.

Correspondingly with probability at least $1-a$

$$\begin{aligned} &Prob^*\left(\max_{i=1,2,\ldots,p_1}|y^*_{f,i}-\widehat{y}^*_{f,i}|\le c_{1-\alpha+a}\right)\ge 1-\alpha+a/2\Rightarrow c^*_{1-\alpha}\le c_{1-\alpha+a}\\ &Prob^*\left(\max_{i=1,2,\ldots,p_1}|y^*_{f,i}-\widehat{y}^*_{f,i}|\le c_{1-\alpha-a}\right)\le 1-\alpha-a/2\Rightarrow c^*_{1-\alpha}\ge c_{1-\alpha-a} \end{aligned} \tag{A.4.16}$$

From (A.4.7), for sufficiently large n, with probability at least $1-a$

$$\begin{aligned} Prob^*\left(\max_{i=1,2,\ldots,p_1}|y_{f,i}-\widehat{y}_{f,i}|\le c^*_{1-\alpha}\right)&\le Prob^*\left(\max_{i=1,2,\ldots,p_1}|y_{f,i}-\widehat{y}_{f,i}|\le c_{1-\alpha+a}\right)\\ &\le 1-\alpha+2a\\ Prob^*\left(\max_{i=1,2,\ldots,p_1}|y_{f,i}-\widehat{y}_{f,i}|\le c^*_{1-\alpha}\right)&\ge Prob^*\left(\max_{i=1,2,\ldots,p_1}|y_{f,i}-\widehat{y}_{f,i}|\le c_{1-\alpha-a}\right)\\ &\ge 1-\alpha-2a \end{aligned} \tag{A.4.17}$$

For $a>0$ can be arbitrarily small, we prove (2.38). □

Appendix B

Proof of theorems in chapter 3

This chapter adopts assumption 1 to 7 in section 3.4.

B.1 Appendix: Preliminary Results

This section introduces some special functions that are helpful in the following proofs. For any $\tau, \psi > 0$, $z \in \mathbf{R}$, define $F_\tau(x_1,...,x_s) = \frac{1}{\tau}\log(\sum_{i=1}^s \exp(\tau x_s))$;

$$\begin{aligned} G_\tau(x_1,...,x_s) &= \frac{1}{\tau}\log\left(\sum_{i=1}^s \exp(\tau x_i) + \sum_{i=1}^s \exp(-\tau x_i)\right) \\ &= F_\tau(x_1,...,x_s,-x_1,...,-x_s) \end{aligned} \tag{B.1.1}$$

Define $g_0(x) = (1-\min(1,\max(x,0)))^4)^4$ and $g_{\psi,z}(x) = g_0(\psi(x-z))$. Then define $h_{\tau,\psi,z}(x_1,...,x_n) = g_{\psi,z}(G_\tau(x_1,...,x_n))$. From lemma A.2 and (8) in Chernozhukov et al. [2013] and (S1) to (S5) in Xu et al. [2019], $g_* = \sup_{x\in\mathbf{R}}(|g_0'(x)| + |g_0''(x)| + |g_0'''(x)|) < \infty$; $\mathbf{1}_{x\le z} \le g_{\psi,z}(x) \le \mathbf{1}_{x\le z+1/\psi}$; $\sup_{x,z\in\mathbf{R}}|g_{\psi,z}'(x)| \le g_*\psi$, $\sup_{x,z\in\mathbf{R}}|g_{\psi,z}''(x)| \le g_*\psi^2$ and $\sup_{x,z\in\mathbf{R}}|g_{\psi,z}'''(x)| \le g_*\psi^3$. Define the operator $\partial_i f = \frac{\partial f}{\partial x_i}$. Then $\partial_i F_\tau \ge 0$; $\sum_{i=1}^s \partial_i F_\tau = 1$; $\sum_{i=1}^s\sum_{j=1}^s |\partial_i\partial_j F_\tau| \le 2\tau$; $\sum_{i=1}^s\sum_{j=1}^s\sum_{k=1}^s |\partial_i\partial_j\partial_k F_\tau| \le 6\tau^2$. Moreover,

$$\begin{aligned} F_\tau(x_1,...,x_s) - \frac{\log(s)}{\tau} &\le \max_{i=1,...,s} x_i \le F_\tau(x_1,...,x_s) \\ \Rightarrow G_\tau(x_1,...,x_s) - \frac{\log(2s)}{\tau} &\le \max_{i=1,...,s}|x_i| \le G_\tau(x_1,...,x_s) \end{aligned} \tag{B.1.2}$$

Since $\partial_i G_\tau = \partial_i F_\tau - \partial_{s+i} F_\tau$, we get $\sum_{i=1}^s |\partial_i G_\tau| \le 1$. For $\partial_i\partial_j G_\tau = \partial_i\partial_j F_\tau - \partial_i\partial_{j+s} F_\tau - \partial_{i+s}\partial_j F_\tau +$ $\partial_{i+s}\partial_{j+s} F_\tau$, we have $\sum_{i=1}^s \sum_{j=1}^s |\partial_i\partial_j G_\tau| \le 2\tau$. Since
$\partial_i\partial_j\partial_k G_\tau = \partial_i\partial_j\partial_k F_\tau - \partial_{i+s}\partial_j\partial_k F_\tau - \partial_i\partial_{j+s}\partial_k F_\tau - \partial_i\partial_j\partial_{k+s} F_\tau + \partial_{i+s}\partial_{j+s}\partial_k F_\tau + \partial_i\partial_{j+s}\partial_{k+s} F_\tau +$ $\partial_{i+s}\partial_j\partial_{k+s} F_\tau - \partial_{i+s}\partial_{j+s}\partial_{k+s} F_\tau$, $\sum_{i=1}^s \sum_{j=1}^s \sum_{k=1}^s |\partial_i\partial_j\partial_k G_\tau| \le 6\tau^2$. For
$\partial_i h_{\tau,\psi,z} = g'_{\psi,z}(G_\tau(x_1,...,x_s)) \times \partial_i G_\tau$, we get $\sum_{i=1}^s |\partial_i h_{\tau,\psi,z}| \le g_*\psi$. Moreover,

$$\begin{aligned}
\partial_i\partial_j h_{\tau,\psi,z} &= g''_{\psi,z}(G_\tau(x_1,...,x_s)) \times \partial_i G_\tau \partial_j G_\tau + g'_{\psi,z}(G_\tau(x_1,...,x_s)) \times \partial_i\partial_j G_\tau \\
&\Rightarrow \sum_{i=1}^s \sum_{j=1}^s |\partial_i\partial_j h_{\tau,\psi,z}| \le g_*\psi^2 + 2g_*\psi\tau \\
\partial_i\partial_j\partial_k h_{\tau,\psi,z} &= g'''_{\psi,z}(G_\tau(x_1,...,x_s)) \times \partial_i G_\tau \partial_j G_\tau \partial_k G_\tau \\
+g''_{\psi,z}(G_\tau(x_1,...,x_s)) &\times \left(\partial_i\partial_j G_\tau \times \partial_k G_\tau + \partial_i\partial_k G_\tau \times \partial_j G_\tau + \partial_j\partial_k G_\tau \times \partial_i G_\tau\right) \\
&+g'(G_\tau(x_1,...,x_s)) \times \partial_i\partial_j\partial_k G_\tau \\
\Rightarrow \sum_{i=1}^s \sum_{j=1}^s \sum_{k=1}^s |\partial_i\partial_j\partial_k h_{\tau,\psi,z}| &\le g_*\psi^3 + 6g_*\tau\psi^2 + 6g_*\psi\tau^2
\end{aligned} \tag{B.1.3}$$

B.2 Proof of theorems in section 3.3

proof of lemma 2. For any $i = 1,2,...,n$ and $s = 1,2,...$, define $M_{i,s} = \sum_{j=n+1-i}^n a_j(\mathbf{E}\varepsilon_j|\mathscr{F}_{j,s} -$ $\mathbf{E}\varepsilon_j|\mathscr{F}_{j,s-1})$. Then $M_{i,s}$ is $\mathscr{F}_{n,i+s-1}$ measurable. Besides,

$$M_{i+1,s} - M_{i,s} = a_{n-i}(\mathbf{E}\varepsilon_{n-i}|\mathscr{F}_{n-i,s} - \mathbf{E}\varepsilon_{n-i}|\mathscr{F}_{n-i,s-1}) \tag{B.2.1}$$

Apply $\pi - \lambda$ theorem to the λ(Dynkin) system

$$\left\{A \in \mathscr{F}_{n,i+s-1} : \mathbf{E}\left(\mathbf{E}\varepsilon_{n-i}|\mathscr{F}_{n-i,s}\right) \times \mathbf{1}_A = \mathbf{E}\left(\mathbf{E}\varepsilon_{n-i}|\mathscr{F}_{n-i,s-1}\right) \times \mathbf{1}_A\right\} \tag{B.2.2}$$

and the π system $\{A_n \times A_{n-1} \times ... \times A_{n-i-s+1}\}$, A_i is generated by e_i. Then
$\mathbf{E}\left(\mathbf{E}\varepsilon_{n-i}|\mathscr{F}_{n-i,s} - \mathbf{E}\varepsilon_{n-i}|\mathscr{F}_{n-i,s-1}\right)|\mathscr{F}_{n,i+s-1} = 0$ almost surely, which implies that

$\{M_{i,s}\}_{i=1,2,...,n}$ form a martingale. From Burkholder's inequality(theorem 1.1 in Burkholder et al. [1972]), there exists a constant C depending only on m such that

$$\begin{aligned}\|\sum_{j=1}^{n} a_j(\mathbf{E}\varepsilon_j|\mathscr{F}_{j,s}-\mathbf{E}\varepsilon_j|\mathscr{F}_{j,s-1})\|_m &\le C\sqrt{\|\sum_{j=1}^{n} a_j^2(\mathbf{E}\varepsilon_j|\mathscr{F}_{j,s}-\mathbf{E}\varepsilon_j|\mathscr{F}_{j,s-1})^2\|_{m/2}} \\ &\le C\sqrt{\sum_{j=1}^{n} a_j^2\|\mathbf{E}\varepsilon_j|\mathscr{F}_{j,s}-\mathbf{E}\varepsilon_j|\mathscr{F}_{j,s-1}\|_m^2} \\ &\le C\sqrt{\sum_{j=1}^{n} a_j^2}\times\max_{i=1,...,n}\delta_{i,s,m}\end{aligned} \tag{B.2.3}$$

Therefore, from theorem 2 in Whittle [1960], there exists a constant C depending only on m such that

$$\begin{aligned}\|\sum_{j=1}^{n} a_j\varepsilon_j\| &\le \|\sum_{j=1}^{n} a_j(\mathbf{E}\varepsilon_j|\mathscr{F}_{j,0})\|_m+\sum_{s=1}^{\infty}\|\sum_{j=1}^{n} a_j(\mathbf{E}\varepsilon_j|\mathscr{F}_{j,s}-\mathbf{E}\varepsilon_j|\mathscr{F}_{j,s-1})\|_m \\ &\le C\sqrt{\sum_{i=1}^{n} a_i^2\|\varepsilon_i\|_m^2}+C\sqrt{\sum_{i=1}^{n} a_i^2}\times\sum_{s=1}^{\infty}\max_{i=1,...,n}\delta_{i,s,m}\end{aligned} \tag{B.2.4}$$

and we prove (3.9).

For

$$\|\max_{i=1,...,p}|\sum_{j=1}^{p} a_{ij}\varepsilon_j|\ \|_m \le p^{1/m}\max_{i=1,...,p}\|\sum_{j=1}^{p} a_{ij}\varepsilon_j\|_m \tag{B.2.5}$$

we prove (3.10). □

Before proving lemma 3, we derive a lemma which is a corollary of Chernozhukov et al. [2015]. It introduces some properties of joint Gaussian random variables.

Lemma B.2.1. *(i). Suppose $\xi_1,...,\xi_s$ are joint normal random variables with $\mathbf{E}\xi_i=0$. Besides, $\exists$ two constants $0<c_0\le C_0<\infty$ such that $c_0\le\mathbf{E}\xi_i^2\le C_0$ for $i=1,...,s$. Then*

$$\begin{aligned}\sup_{x\in\mathbf{R}}|Prob\left(\max_{i=1,...,s}|\xi_i|\le x+\delta\right)-Prob\left(\max_{i=1,...,s}|\xi_i|\le x\right)| \\ \le C\delta(1+\sqrt{\log(s)}+\sqrt{|\log(\delta)|})\end{aligned} \tag{B.2.6}$$

for any $\delta > 0$. Here C is a constant only depending on c_0, C_0.

(ii). Define $\Sigma = \{\sigma_{ij}\}_{i,j=1,...,s}$ such that $\sigma_{ij} = \mathbf{E}\xi_i\xi_j$. Suppose $\xi_1^\dagger, ..., \xi_s^\dagger$ are joint normal random variables with $\mathbf{E}\xi_i^\dagger = 0$. Define $\Sigma^\dagger = \{\sigma_{ij}^\dagger\}_{i,j=1,...,s}$ such that $\sigma_{ij}^\dagger = \mathbf{E}\xi_i^\dagger\xi_j^\dagger$. Define $\Delta = \max_{i,j=1,...,s}|\sigma_{ij} - \sigma_{ij}^\dagger|$ and suppose $\Delta < 1$. Then

$$\begin{aligned} &\sup_{x\in\mathbf{R}}|Prob\left(\max_{i=1,...,s}|\xi_i| \le x\right) - Prob\left(\max_{i=1,...,s}|\xi_i^\dagger| \le x\right)| \\ &\qquad \le C^*\left(\Delta^{1/3}(1+\log^3(s)) + \frac{\Delta^{1/6}}{1+\log^{1/4}(s)}\right) \end{aligned} \tag{B.2.7}$$

here C^* only depends on c_0, C_0.

Notably, we do not impose any assumptions on $\Sigma^\dagger$, the covariance matrix of $\varepsilon_i^\dagger$.

proof of lemma B.2.1. For $|\xi_i| = \max(\xi_i, -\xi_i)$,

$\max_{i=1,...,s}|\xi_i| = \max(\max_{i=1,...,s}\xi_i, \max_{i=1,...,s}(-\xi_i))$, so

$$\begin{aligned} &\sup_{x\in\mathbf{R}}\left(Prob\left(\max_{i=1,...,s}|\xi_i| \le x+\delta\right) - Prob\left(\max_{i=1,...,s}|\xi_i| \le x\right)\right) \\ \le &\sup_{x\in\mathbf{R}}Prob\left(x < \max_{i=1,...,s}\xi_i \le x+\delta\right) + \sup_{x\in\mathbf{R}}Prob\left(x < \max_{i=1,...,s}(-\xi_i) \le x+\delta\right) \\ &\qquad \le 2\sup_{x\in\mathbf{R}}Prob\left(|\max_{i=1,...,s}\xi_i - x| \le \delta\right) \end{aligned} \tag{B.2.8}$$

here $-\xi_1, ..., -\xi_s$ has the same joint distribution as $\xi_1, ..., \xi_s$. From theorem 3 and (18), (19) in

Chernozhukov et al. [2015], define $\underline{\sigma} = \min_{i=1,...,s} \sigma_{ii}$ and $\overline{\sigma} = \max_{i=1,...,s} \sigma_{ii}$,

$$\begin{aligned}
\sup_{x\in\mathbf{R}} Prob\left(|\max_{i=1,...,s}\xi_i - x| \le \delta\right) &\le \frac{\sqrt{2}\delta}{\underline{\sigma}}\left(\sqrt{\log(s)} + \sqrt{\max(1, \log(\underline{\sigma}) - \log(\delta))}\right) \\
&+ \frac{4\sqrt{2}\delta}{\underline{\sigma}} \times \left(\frac{\overline{\sigma}}{\underline{\sigma}}\sqrt{\log(s)} + 2 + \frac{\overline{\sigma}}{\underline{\sigma}}\sqrt{\max(0, \log(\underline{\sigma}) - \log(\delta))}\right) \\
&\le \frac{\sqrt{2}\delta}{c_0}\left(\sqrt{\log(s)} + \sqrt{1 + |\log(c_0)| + |\log(C_0)|} + \sqrt{|\log(\delta)|}\right) \\
&+ \frac{4\sqrt{2}C_0\delta}{c_0^2}\left(\sqrt{\log(s)} + 2 + \sqrt{|\log(c_0)| + |\log(C_0)|} + \sqrt{|\log(\delta)|}\right) \\
&\le C\delta(1 + \sqrt{\log(s)} + \sqrt{|\log(\delta)|})
\end{aligned} \tag{B.2.9}$$

here $C = \frac{\sqrt{2\times(1+|\log(c_0)|+|\log(C_0)|)}}{c_0} + \frac{4\sqrt{2}C_0}{c_0^2} \times (2 + \sqrt{|\log(c_0)| + |\log(C_0)|})$, and we prove (B.2.6).

Without loss of generality, assume ξ_i is independent of $\xi_j^\dagger$ for any i, j. Similar to Chernozhukov et al. [2015], for any $0 \le t \le 1$, define random variables $Z_i(t) = \sqrt{t}\xi_i + \sqrt{1-t}\xi_i^\dagger$. According to theorem 2.27 in Folland [1999] and lemma 2 in Chernozhukov et al. [2015]

$$\begin{aligned}
&\mathbf{E}h_{\tau,\psi,x}(\xi_1, ..., \xi_s) - \mathbf{E}h_{\tau,\psi,x}(\xi_1^\dagger, ..., \xi_s^\dagger) \\
&= \frac{1}{2}\sum_{i=1}^{s}\int_0^1 t^{-1/2}\mathbf{E}\left(\xi_i \partial_i h_{\tau,\psi,x}(Z_1(t), ..., Z_s(t))\right) dt \\
&- \frac{1}{2}\sum_{i=1}^{s}\int_0^1 (1-t)^{-1/2}\mathbf{E}\left(\xi_i^\dagger \partial_i h_{\tau,\psi,x}(Z_1(t), ..., Z_s(t))\right) dt \\
&= \frac{1}{2}\sum_{i=1}^{s}\sum_{j=1}^{s}\left(\sigma_{ij} - \sigma_{ij}^\dagger\right) \times \int_0^1 \mathbf{E}\left(\partial_i\partial_j h_{\tau,\psi,x}(Z_1(t), ..., Z_s(t))\right) dt \\
&\Rightarrow \sup_{x\in\mathbf{R}} |\mathbf{E}h_{\tau,\psi,x}(\xi_1, ..., \xi_s) - \mathbf{E}h_{\tau,\psi,x}(\xi_1^\dagger, ..., \xi_s^\dagger)| \\
&\le \frac{\Delta}{2} \times \int_0^1 \sum_{i=1}^{s}\sum_{j=1}^{s}\mathbf{E}|\partial_i\partial_j h_{\tau,\psi,x}(Z_1(t), ..., Z_s(t))| dt \\
&\le g_* \Delta \times (\psi^2 + \psi\tau)
\end{aligned} \tag{B.2.10}$$

Define $t = \frac{1}{\psi} + \frac{\log(2s)}{\tau}$, then

$$\begin{aligned}
& Prob\left(\max_{i=1,...,s}|\xi_i| \le x\right) - Prob\left(\max_{i=1,...,s}|\xi_i^\dagger| \le x\right) \\
\le\ & Prob\left(\max_{i=1,...,s}|\xi_i| \le x-t\right) + Ct(1+\sqrt{\log(s)}+\sqrt{|\log(t)|}) \\
& - Prob\left(\max_{i=1,...,s}|\xi_i^\dagger| \le x\right) \\
\le\ & \mathbf{E}h_{\tau,\psi,x-\frac{1}{\psi}}(\xi_1,...,\xi_s) - \mathbf{E}h_{\tau,\psi,x-\frac{1}{\psi}}(\xi_1^\dagger,...,\xi_s^\dagger) + Ct(1+\sqrt{\log(s)}+\sqrt{|\log(t)|}) \\
& Prob\left(\max_{i=1,...,s}|\xi_i| \le x\right) - Prob\left(\max_{i=1,...,s}|\xi_i^\dagger| \le x\right) \\
\ge\ & Prob\left(\max_{i=1,...,s}|\xi_i| \le x+t\right) - Prob\left(\max_{i=1,...,s}|\xi_i^\dagger| \le x\right) \\
& - Ct(1+\sqrt{\log(s)}+\sqrt{|\log(t)|}) \\
\ge\ & \mathbf{E}h_{\tau,\psi,x+\frac{\log(2s)}{\tau}}(\xi_1,...,\xi_s) - \mathbf{E}h_{\tau,\psi,x+\frac{\log(2s)}{\tau}}(\xi_1^\dagger,...,\xi_s^\dagger) \\
& - Ct(1+\sqrt{\log(s)}+\sqrt{|\log(t)|}) \\
\Rightarrow & \sup_{x\in\mathbf{R}}|Prob\left(\max_{i=1,...,s}|\xi_i| \le x\right) - Prob\left(\max_{i=1,...,s}|\xi_i^\dagger| \le x\right)| \\
\le & \sup_{x\in\mathbf{R}}|\mathbf{E}h_{\tau,\psi,x}(\xi_1,...,\xi_s) - \mathbf{E}h_{\tau,\psi,x}(\xi_1^\dagger,...,\xi_s^\dagger)| + Ct(1+\sqrt{\log(s)}+\sqrt{|\log(t)|})
\end{aligned} \tag{B.2.11}$$

Choose $\tau = \psi = \left(1+\log^{3/2}(s)\right)/\Delta^{1/3}$, then $\exists$ a constant $C_1 > 0$ such that $\frac{1}{C_1}\frac{\Delta^{1/3}}{1+\log^{1/2}(s)} \le t = \Delta^{1/3}\left(\frac{1+\log(2)}{1+\log^{3/2}(s)}\right) + \frac{\Delta^{1/3}\log(s)}{1+\log^{3/2}(s)} \le \frac{C_1\Delta^{1/3}}{1+\log^{1/2}(s)}$ for $s = 1,2,...$; and we prove (B.2.7). □

proof of lemma 3. For any given $\psi > 0$, define $t = \frac{1}{\psi} + \frac{\log(2p_1)}{\psi}$. Notice that $c \times c_\Sigma \le \mathbf{E}(\sum_{j=1}^n a_{ij}\xi_j)^2 = O(1)$. From lemma B.2.1 and (B.2.11)

$$\begin{aligned}
& \sup_{x\in\mathbf{R}}|Prob\left(\max_{i=1,...,p_1}|\sum_{j=1}^n a_{ij}\varepsilon_j| \le x\right) - Prob\left(\max_{i=1,...,p_1}|\sum_{j=1}^n a_{ij}\xi_j| \le x\right)| \\
\le & \sup_{x\in\mathbf{R}}|\mathbf{E}h_{\psi,\psi,x}\left(\sum_{j=1}^n a_{1j}\varepsilon_j,...,\sum_{j=1}^n a_{p_1j}\varepsilon_j\right) - \mathbf{E}h_{\psi,\psi,x}\left(\sum_{j=1}^n a_{1j}\xi_j,...,\sum_{j=1}^n a_{p_1j}\xi_j\right)| \\
& + Ct(1+\sqrt{\log(p_1)}+\sqrt{|\log(t)|})
\end{aligned} \tag{B.2.12}$$

here C is a constant. For any integer $s > 0$, (B.2.3) implies

$$\begin{aligned} &\max_{i=1,...,p_1} \| \sum_{j=1}^{n} a_{ij}\varepsilon_j - \left(\sum_{j=1}^{n} a_{ij}\mathbf{E}\varepsilon_j|\mathscr{F}_{j,s} \right) \|_m \\ \leq &\max_{i=1,...,p_1} \sum_{k=s+1}^{\infty} \| \sum_{j=1}^{n} a_{ij} \left(\mathbf{E}\varepsilon_j|\mathscr{F}_{j,k} - \mathbf{E}\varepsilon_j|\mathscr{F}_{j,k-1} \right) \|_m \\ \leq &C \max_{i=1,...,p_1} \sqrt{\sum_{j=1}^{n} a_{ij}^2} \sum_{k=s+1}^{\infty} \max_{i=1,...,n} \delta_{i,k,m} \leq \frac{C_1}{(1+s)^{\alpha}} \end{aligned} \tag{B.2.13}$$

here C_1 is a constant. Therefore

$$\begin{aligned} &\sup_{x\in\mathbf{R}} |\mathbf{E}h_{\psi,\psi,x} \left(\sum_{j=1}^{n} a_{1j}\varepsilon_j, ..., \sum_{j=1}^{n} a_{p_1 j}\varepsilon_j \right) \\ &-\mathbf{E}h_{\psi,\psi,x} \left(\sum_{j=1}^{n} a_{1j}\mathbf{E}\varepsilon_j|\mathscr{F}_{j,s}, ..., \sum_{j=1}^{n} a_{p_1 j}\mathbf{E}\varepsilon_j|\mathscr{F}_{j,s} \right) | \\ &\leq g_* \psi \mathbf{E} \max_{i=1,...,p_1} | \sum_{j=1}^{n} a_{ij}(\varepsilon_j - \mathbf{E}\varepsilon_j|\mathscr{F}_{j,s})| \\ \leq g_* \psi p_1^{1/m} \times &\max_{i=1,...,p_1} \| \sum_{j=1}^{n} a_{ij}(\varepsilon_j - \mathbf{E}\varepsilon_j|\mathscr{F}_{j,s}) \|_m = O\left(\frac{\psi}{(1+s)^{\alpha}} \right) \end{aligned} \tag{B.2.14}$$

For any integer $k > s$, define the big block S_{il} and the small block s_{il} as

$$S_{il} = \sum_{j=(l-1)\times(k+s)+1}^{((l-1)\times(k+s)+k)\wedge n} a_{ij}\mathbf{E}\varepsilon_j|\mathscr{F}_{j,s} \text{ and } s_{il} = \sum_{j=(l-1)\times(k+s)+k+1}^{(l\times(k+s))\wedge n} a_{ij}\mathbf{E}\varepsilon_j|\mathscr{F}_{j,s} \tag{B.2.15}$$

here $l = 1, 2, ..., l_0$, $l_0 = \lceil \frac{n}{k+s} \rceil$, i.e., the smallest integer that is larger than or equal to $\frac{n}{k+s}$. Then the vector $(S_{1l}, ..., S_{p_1 l}), l = 1, 2, ..., l_0$ are mutually independent, and the vector $(s_{1l}, ..., s_{p_1 l}), l = 1, 2, ..., l_0$ are mutually independent. Moreover,

$$\sum_{j=1}^{n} a_{ij}\mathbf{E}\varepsilon_j|\mathscr{F}_{j,s} = \sum_{l=1}^{l_0} S_{il} + \sum_{l=1}^{l_0} s_{il} \tag{B.2.16}$$

Besides,

$$\begin{aligned}
&\sup_{x\in\mathbf{R}}|\mathbf{E}h_{\psi,\psi,x}\left(\sum_{j=1}^{l_0}S_{1j}+\sum_{j=1}^{l_0}s_{1j},...,\sum_{j=1}^{l_0}S_{p_1j}+\sum_{j=1}^{l_0}s_{p_1j}\right)\\
&\qquad -\mathbf{E}h_{\psi,\psi,x}\left(\sum_{j=1}^{l_0}S_{1j},...,\sum_{j=1}^{l_0}S_{p_1j}\right)|\\
&\qquad \le g_*\psi p_1^{1/m}\times\max_{i=1,...,p_1}\|\sum_{j=1}^{l_0}s_{ij}\|_m\\
&\le g_*\psi p_1^{1/m}\times\max_{i=1,...,p_1}\|\sum_{l=1}^{l_0}\sum_{j=(l-1)\times(k+s)+k+1}^{(l\times(k+s))\wedge n}a_{ij}\mathbf{E}\varepsilon_j|\mathscr{F}_{j,0}\|_m\\
&+g_*\psi p_1^{1/m}\times\max_{i=1,...,p_1}\sum_{k=1}^{s}\|\sum_{l=1}^{l_0}\sum_{j=(l-1)\times(k+s)+k+1}^{(l\times(k+s))\wedge n}a_{ij}(\mathbf{E}\varepsilon_j|\mathscr{F}_{j,k}-\mathbf{E}\varepsilon_j|\mathscr{F}_{j,k-1})\|_m\\
&\le C\psi\max_{i=1,...,p_1}\sqrt{\sum_{l=1}^{l_0}\sum_{j=(l-1)\times(k+s)+k+1}^{(l\times(k+s))\wedge n}a_{ij}^2}=O\left(\psi a^*\sqrt{\frac{ns}{k}}\right)
\end{aligned}\tag{B.2.17}$$

Define $\xi_{i,s}, i=1,2,...,n$ as joint normal random variables such that $\mathbf{E}\xi_{i,s}=0$ and $\mathbf{E}\xi_{i,s}\xi_{j,s}=\mathbf{E}\left((\mathbf{E}\varepsilon_i|\mathscr{F}_{i,s})\times\left(\mathbf{E}\varepsilon_j|\mathscr{F}_{j,s}\right)\right)$ for any $i,j=1,2,...,n$, and is independent with e_k for any $k\in\mathbf{Z}$. Define the corresponding big block S_{il}^* and the small block s_{il}^* as

$$S_{il}^*=\sum_{j=(l-1)\times(k+s)+1}^{((l-1)\times(k+s)+k)\wedge n}a_{ij}\xi_{j,s}\text{ and }\ s_{il}^*=\sum_{j=(l-1)\times(k+s)+k+1}^{(l\times(k+s))\wedge n}a_{ij}\xi_{j,s}\tag{B.2.18}$$

For any i_1,i_2 and any $l_1\neq l_2$,

$$\begin{aligned}
&\mathbf{E}S_{i_1l_1}^*S_{i_2l_2}^*\\
&=\sum_{j_1=(l_1-1)\times(k+s)+1}^{((l_1-1)\times(k+s)+k)\wedge n}\ \sum_{j_2=(l_2-1)\times(k+s)+1}^{((l_2-1)\times(k+s)+k)\wedge n}a_{i_1j_1}a_{i_2j_2}\mathbf{E}\left(\left(\mathbf{E}\varepsilon_{j_1}|\mathscr{F}_{j_1,s}\right)\times\left(\mathbf{E}\varepsilon_{j_2}|\mathscr{F}_{j_2,s}\right)\right)=0
\end{aligned}\tag{B.2.19}$$

So the random vectors $(S_{1l}^*,S_{2l}^*,...,S_{p_1l}^*)^T$ are mutually independent for $l=1,2,...,l_0$. Define

$H_{ij} = \sum_{l=1}^{j-1} S_{il} + \sum_{l=j+1}^{l_0} S^*_{il}$, then $H_{ij} + S_{ij} = H_{ij+1} + S^*_{ij+1}$. From Taylor's theorem

$$
\begin{aligned}
&|\mathbf{E}[h_{\psi,\psi,x}(H_{1j}+S_{1j},...,H_{p_1j}+S_{p_1j}) \\
&-h_{\psi,\psi,x}(H_{1j}+S^*_{1j},...,H_{p_1j}+S^*_{p_1j})]|H_{1j},...,H_{p_1j}| \\
&\leq |\sum_{i=1}^{p_1} \partial_i h_{\psi,\psi,x}(H_{1j},...,H_{p_1j})\mathbf{E}(S_{ij}-S^*_{ij})| \\
&+|\frac{1}{2}\sum_{i_1=1}^{p_1}\sum_{i_2=1}^{p_1} \partial_{i_1}\partial_{i_2} h_{\psi,\psi,x}(H_{1j},...,H_{p_1j})\mathbf{E}(S_{i_1j}S_{i_2j}-S^*_{i_1j}S^*_{i_2j})| \\
&+3g_*\psi^3\mathbf{E}\max_{i=1,...,p_1}|S_{ij}|^3 + 3g_*\psi^3\mathbf{E}\max_{i=1,...,p_1}|S^*_{ij}|^3 \\
&\leq 3g_*\psi^3 p_1^{3/m}\max_{i=1,...,p_1}\|S_{ij}\|_m^3 + 3g_*\psi^3 p_1^{3/m}\max_{i=1,...,p_1}\|S^*_{ij}\|_m^3 \\
&\leq C\psi^3\max_{i=1,...,p_1}\|S_{ij}\|_m^3
\end{aligned}
\tag{B.2.20}
$$

here C is a constant. The last inequality holds since S^*_{kj} has normal distribution(therefore $\|S^*_{ij}\|_m \leq C\|S_{ij}\|_2 \leq C\|S_{ij}\|_m$ for a constant C). For

$$
\begin{aligned}
\|S_{ij}\|_m &= \|\sum_{l=(j-1)\times(k+s)+1}^{((j-1)\times(k+s)+k)\wedge n} a_{il}\mathbf{E}\varepsilon_l|\mathscr{F}_{l,s}\|_m \\
&\leq \|\sum_{l=(j-1)\times(k+s)+1}^{((j-1)\times(k+s)+k)\wedge n} a_{il}\mathbf{E}\varepsilon_l|\mathscr{F}_{l,0}\|_m \\
&+\sum_{v=1}^{s}\|\sum_{l=(j-1)\times(k+s)+1}^{((j-1)\times(k+s)+k)\wedge n} a_{il}(\mathbf{E}\varepsilon_l|\mathscr{F}_{l,v}-\mathbf{E}\varepsilon_l|\mathscr{F}_{l,v-1})\|_m \\
&\leq C\sqrt{\sum_{l=(j-1)\times(k+s)+1}^{((j-1)\times(k+s)+k)\wedge n} a_{il}^2} \leq Ca^*\sqrt{k}
\end{aligned}
\tag{B.2.21}
$$

with a constant C. We have

$$\sup_{x\in\mathbf{R}}|\mathbf{E}h_{\psi,\psi,x}\left(\sum_{j=1}^{l_0}S_{1j},...,\sum_{j=1}^{l_0}S_{p_1j}\right)-\mathbf{E}h_{\psi,\psi,x}\left(\sum_{j=1}^{l_0}S^*_{1j},...,\sum_{j=1}^{l_0}S^*_{p_1j}\right)|$$

$$\leq\sum_{j=1}^{l_0}\sup_{x\in\mathbf{R}}|\mathbf{E}h_{\psi,\psi,x}(H_{1j}+S_{1j},...,H_{p_1j}+S_{p_1j})-\mathbf{E}h_{\psi,\psi,x}(H_{1j}+S^*_{1j},...,H_{p_1j}+S^*_{p_1j})| \quad \text{(B.2.22)}$$

$$\leq C\psi^3\sum_{l=1}^{l_0}\sum_{i=1}^{p_1}\|S_{ij}\|_m^3\leq C'\psi^3\sum_{i=1}^{p_1}\sum_{j=1}^{l_0}\left(\sum_{l=(j-1)\times(k+s)+1}^{((j-1)\times(k+s)+k)\wedge n}a_{il}^2\right)^{3/2}\leq C''\psi^3a^*\sqrt{k}$$

here C,C',C'' are constants. The last inequality comes from $\sum_{j=1}^{l_0}\sum_{l=(j-1)\times(k+s)+1}^{((j-1)\times(k+s)+k)\wedge n}a_{il}^2\leq\sum_{j=1}^{n}a_{ij}^2=O(1)$. Since

$$|\mathbf{E}h_{\psi,\psi,x}(\sum_{j=1}^{l_0}S^*_{1j},...,\sum_{j=1}^{l_0}S^*_{p_1j})-\mathbf{E}h_{\psi,\psi,x}(\sum_{j=1}^{n}a_{1j}\xi_{j,s},...,\sum_{j=1}^{n}a_{p_1j}\xi_{j,s})|$$

$$\leq g_*\psi\mathbf{E}\max_{i=1,...,p_1}|\sum_{l=1}^{l_0}s^*_{il}| \quad \text{(B.2.23)}$$

$$\leq g_*\psi p_1^{1/m}\times\max_{i=1,2,...,p_1}\|\sum_{l=1}^{l_0}s^*_{il}\|_m\leq C\psi\max_{i=1,...,p_1}\|\sum_{l=1}^{l_0}s_{il}\|_2$$

with a constant C. From (B.2.17), this has order $O\left(\psi a^*\sqrt{\frac{ns}{k}}\right)$. From section 0.9.7 in Horn and

Johnson [2013] and (B.2.10),

$$
\begin{aligned}
&|\mathbf{E}h_{\psi,\psi,x}(\sum_{j=1}^{n} a_{1j}\xi_{j,s},\ldots,\sum_{j=1}^{n} a_{p_1 j}\xi_{j,s}) - \mathbf{E}h_{\psi,\psi,x}(\sum_{j=1}^{n} a_{1j}\xi_{j},\ldots,\sum_{j=1}^{n} a_{p_1 j}\xi_{j})| \\
&\leq 2g_*\psi^2 \\
&\times \max_{i_1,i_2=1,\ldots,p_1} |\sum_{j_1=1}^{n}\sum_{j_2=1}^{n} a_{i_1 j_1} a_{i_2 j_2} \left(\mathbf{E}\left(\left(\mathbf{E}\varepsilon_{j_1}|\mathscr{F}_{j_1,s}\right)\times\left(\mathbf{E}\varepsilon_{j_2}|\mathscr{F}_{j_2,s}\right)\right) - \mathbf{E}\varepsilon_{j_1}\varepsilon_{j_2}\right)| \\
&\leq 2g_*\psi^2 \\
&\times \max_{i_1,i_2=1,\ldots,p_1} \sum_{|j_1-j_2|\leq s, j_1,j_2=1,\ldots,n} |a_{i_1 j_1} a_{i_2 j_2}| \times \|\mathbf{E}\varepsilon_{j_1}|\mathscr{F}_{j_1,s}\|_2 \times \|\mathbf{E}\varepsilon_{j_2}|\mathscr{F}_{j_2,s} - \varepsilon_{j_2}\|_2 \\
&+2g_*\psi^2 \max_{i_1,i_2=1,\ldots,p_1} \sum_{|j_1-j_2|\leq s, j_1,j_2=1,\ldots,n} |a_{i_1 j_1} a_{i_2 j_2}| \times \|\varepsilon_2\|_2 \times \|\mathbf{E}\varepsilon_{j_1}|\mathscr{F}_{j_1,s} - \varepsilon_{j_1}\|_2 \\
&+2g_*\psi^2 \max_{i_1,i_2=1,\ldots,p_1} \sum_{|j_1-j_2|> s, j_1,j_2=1,\ldots,n} |a_{i_1 j_1} a_{i_2 j_2}| \times |\mathbf{E}\varepsilon_{j_1}\varepsilon_{j_2}| \\
&\leq \frac{C\psi^2 s}{(s+1)^\alpha}\sqrt{\sum_{j=1}^{n} a_{i_1 j}^2}\sqrt{\sum_{j=1}^{n} a_{i_2 j}^2} + C\psi^2\sqrt{\sum_{j=1}^{n} a_{i_1 j}^2}\sqrt{\sum_{j=1}^{n} a_{i_2 j}^2}\times\sum_{l=s}^{\infty}\frac{1}{(l+1)^\alpha} \\
&= O\left(\frac{\psi^2}{s^{\alpha-1}}\right)
\end{aligned}
\tag{B.2.24}
$$

Define V such that $1/V = a^* \times n^{1/4} \times \log^z(n) \to 0$ as $n \to \infty$. Then choose $k = \lfloor\sqrt{n}\rfloor$, $\psi = V^{(\alpha-1)/(3\alpha+3)}$, $s = \lfloor V^{2/(\alpha+1)} \log^{3/(\alpha-1)}(n)\rfloor \to \infty$. Here $\lfloor x\rfloor$ denotes the largest integer that is smaller than or equal to x. From (B.2.12), (B.2.14), (B.2.17), (B.2.22) and (B.2.24), we prove (3.13). □

proof of lemma 4. From section 0.9.7 in Horn and Johnson [2013]

$$
\begin{aligned}
&|\sum_{j_1=1}^{n}\sum_{j_2=1}^{n}\left(1-K\left(\frac{j_1-j_2}{k_n}\right)\right) a_{i_1 j_1} a_{i_2 j_2}\sigma_{j_1 j_2}| \\
&\leq C\sum_{j_1=1}^{n}\sum_{j_2=1}^{n} |a_{i_1 j_1} a_{i_2 j_2}| \times \frac{1-K\left(\frac{j_1-j_2}{k_n}\right)}{(1+|j_1-j_2|)^\alpha} \\
&\leq 2C\sqrt{\sum_{j=1}^{n} a_{i_1 j}^2}\times\sqrt{\sum_{j=1}^{n} a_{i_2 j}^2}\times\sum_{s=0}^{\infty}\frac{1-K(s/k_n)}{1+s^\alpha}
\end{aligned}
\tag{B.2.25}
$$

K is continuous differentiable, so

$$\begin{aligned}\sum_{s=0}^{\infty}\frac{1-K(s/k_n)}{1+s^{\alpha}} &\leq \frac{\max_{x\in[0,1]}|K'(x)|}{k_n}\sum_{s=0}^{k_n}\frac{s}{1+s^{\alpha}}+\sum_{s=k_n+1}^{\infty}\frac{1}{1+s^{\alpha}}\\ &= O\left(\frac{1}{k_n}+\frac{1}{k_n}\int_{[1,k_n]}x^{1-\alpha}dx+\int_{[k_n,\infty)}x^{-\alpha}dx\right)\end{aligned} \tag{B.2.26}$$

which implies

$$|\sum_{j_1=1}^{n}\sum_{j_2=1}^{n}\left(1-K\left(\frac{j_1-j_2}{k_n}\right)\right)a_{i_1j_1}a_{i_2j_2}\sigma_{j_1j_2}| = O(v_n) \tag{B.2.27}$$

On the other hand, define $\xi_{i,k}=\varepsilon_i\varepsilon_{i+k}-\sigma_{ii+k}=h_{i,i+k}(...,e_{i+k-1},e_{i+k})$ for $i=1,2,...,n-k$ and $k\geq 0$(in other words, $\xi_{i,k}$ is $\mathscr{F}_{i+k}$ measurable), then $\mathbf{E}\xi_{i,k}=0$. Define $\xi_{i,k,t}=h_{i,i+k}(...,e_{i+k-t-2},e_{i+k-t-1},e^{\dagger}_{i+k-t},e_{i+k-t+1},...,e_{i+k})$. Here $t\geq 0$. We have

$$\begin{aligned}\psi_{i,k,t,m/2} &= \|\xi_{i,k}-\xi_{i,k,t}\|_{m/2}=\|\varepsilon_i\varepsilon_{i+k}-\varepsilon_{i+k,t}\varepsilon_{i,t-k}\|_{m/2}\\ &\leq \|\varepsilon_i\|_m\times\|\varepsilon_{i+k}-\varepsilon_{i+k,t}\|_m+\|\varepsilon_{i+k,t}\|_m\times\|\varepsilon_i-\varepsilon_{i,t-k}\|_m\\ &\leq C\max_{i=1,...,n}\delta_{i,t,m}+C\max_{i=1,...,n}\delta_{i,t-k,m}\end{aligned} \tag{B.2.28}$$

Here C is a constant and $\delta_{i,j,m}=0$ if $j<0$. For a given $i_1,i_2=1,...,p_1$, define the term $N_{s,k,t}=\sum_{j=n-k+1-s}^{n-k}a_{i_1j}a_{i_2j+k}\left(\mathbf{E}\zeta_{j,k}|\mathscr{F}_{j+k,t}-\mathbf{E}\zeta_{j,k}|\mathscr{F}_{j+k,t-1}\right)$. $N_{s,k,t}$ is $\mathscr{F}_{n,s+t-1}$ measurable. Moreover, $N_{s+1,k,t}-N_{s,k,t}=a_{i_1n-k-s}a_{i_2n-s}(\mathbf{E}\zeta_{n-k-s,k}|\mathscr{F}_{n-s,t}-\mathbf{E}\zeta_{n-k-s,k}|\mathscr{F}_{n-s,t-1})$. Apply $\pi-\lambda$ theorem to the $\lambda-$system

$$\left\{A\in\mathscr{F}_{n,s+t-1}:\mathbf{E}(\mathbf{E}\zeta_{n-k-s,k}|\mathscr{F}_{n-s,t})\times\mathbf{1}_A=\mathbf{E}(\mathbf{E}\zeta_{n-k-s,k}|\mathscr{F}_{n-s,t-1})\times\mathbf{1}_A\right\} \tag{B.2.29}$$

and the $\pi-$system $\{A_n\times A_{n-1}\times...\times A_{n-s-t+1}\}$, A_i is generated by e_i. We know that $\{N_{s,k,t}:$

$s = 1, ..., n-k\}$ form a martingale. From (B.2.3) and theorem 2 in Whittle [1960]

$$\begin{aligned} \| \sum_{j=1}^{n-k} a_{i_1 j} a_{i_2 j+k} \zeta_{j,k} \|_{m/2} &\leq \| \sum_{j=1}^{n-k} a_{i_1 j} a_{i_2 j+k} (\mathbf{E}\zeta_{j,k} | \mathscr{F}_{j+k,0}) \|_{m/2} + \sum_{t=1}^{\infty} \| N_{n-k,k,t} \|_{m/2} \\ &\leq C \sqrt{\sum_{j=1}^{n-k} a_{i_1 j}^2 a_{i_2 j+k}^2} + C \sqrt{\sum_{j=1}^{n-k} a_{i_1 j}^2 a_{i_2 j+k}^2} \sum_{t=1}^{\infty} \max_{i=1,...,n-k} \psi_{i,k,t,m/2} \\ &\leq C' \times \max_{i=1,...,p_1, j=1,...,n} |a_{ij}| \end{aligned} \tag{B.2.30}$$

here C, C' are constants. Therefore from (B.2.28)

$$\begin{aligned} & \| \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} a_{i_1 j_1} a_{i_2 j_2} K\left(\frac{j_1 - j_2}{k_n}\right) (\varepsilon_{j_1} \varepsilon_{j_2} - \sigma_{j_1 j_2}) \|_{m/2} \\ \leq & \sum_{l=0}^{n-1} K\left(\frac{l}{k_n}\right) \| \sum_{j=1}^{n-l} a_{i_1 j} a_{i_2 j+l} \zeta_{j,l} \|_{m/2} + \sum_{l=1}^{n-1} K\left(\frac{l}{k_n}\right) \| \sum_{j=1}^{n-l} a_{i_1 j+l} a_{i_2 j} \zeta_{j,l} \|_{m/2} \\ \leq & C \times \max_{i=1,...,p_1, j=1,...,n} |a_{ij}| \sum_{l=0}^{\infty} K\left(\frac{l}{k_n}\right) \end{aligned} \tag{B.2.31}$$

with a constant C. Since

$$\sum_{l=0}^{\infty} K\left(\frac{l}{k_n}\right) \leq 1 + \int_{[0,\infty)} K(x/k_n) dx = O(k_n) \tag{B.2.32}$$

(B.2.27), (B.2.31) and $p_1 = O(1)$ imply (3.15). □

B.3 Proof of theorems in section 3.4

This section starts with the consistency of the Lasso estimator.

Lemma B.3.1. *Suppose assumption 1 to 5 hold true. Then*

$$|\widetilde{\beta}^{lasso} - \beta|_2 = O_p(n^{\frac{\alpha_{\mathscr{N}}}{2} - \alpha_l}) \text{ and } |\widetilde{\beta}^{lasso} - \beta|_1 = O_p(n^{\alpha_{\mathscr{N}} - \alpha_l}) \tag{B.3.1}$$

the Lasso estimator is defined in (3.20)*.*

In particular, if $p > n$, define $V = \{V_{ij}\}_{i,j=1,\ldots,p} = Q_\perp Q_\perp^T$ (see section 3.4 for the definition of $Q_\perp$),

$$\begin{aligned} |\widetilde{\theta}_\perp^\dagger - \theta_\perp|_\infty = |V(\widetilde{\beta}^{lasso} - \beta)|_\infty &\leq \max_{i=1,\ldots,p} \sqrt{\sum_{j=1}^p V_{ij}^2} \times |\widetilde{\beta}^{lasso} - \beta|_2 \\ &\leq |\widetilde{\beta}^{lasso} - \beta|_2 = O_p(n^{\frac{\alpha_{\mathscr{N}}}{2} - \alpha_l}) \\ \text{and } |\widetilde{\theta}_\perp^\dagger - \theta_\perp|_2 &\leq |\widetilde{\beta}^{lasso} - \beta|_2 = O_p(n^{\frac{\alpha_{\mathscr{N}}}{2} - \alpha_l}) \end{aligned} \tag{B.3.2}$$

We follow the proofs from Bühlmann and van de Geer [2011] to show the consistency of the Lasso estimator.

proof of lemma B.3.1. Define $\omega_i = \widetilde{\beta}_i^{lasso} - \beta_i$ and $\omega = (\omega_1, \ldots, \omega_p)^T$, then

$$\begin{aligned} &\frac{1}{2n}\sum_{i=1}^n \left(\varepsilon_i - \sum_{j=1}^p X_{ij}(\widetilde{\beta}_j^{lasso} - \beta_j)\right)^2 + \rho_{n,l}\sum_{j=1}^p |\widetilde{\beta}_j^{lasso}| \leq \frac{1}{2n}\sum_{i=1}^n \varepsilon_i^2 + \rho_{n,l}\sum_{j=1}^p |\beta_j| \\ &\Rightarrow \frac{1}{2n}|X\omega|_2^2 \leq \frac{1}{n}\varepsilon^T X\omega + \rho_{n,l}\sum_{j\in\mathscr{N}_{b_n}}(|\beta_j| - |\widetilde{\beta}_j^{lasso}|) - \rho_{n,l}\sum_{j\notin\mathscr{N}_{b_n}}|\omega_j| \end{aligned} \tag{B.3.3}$$

Since $\max_{j=1,\ldots,p}\sum_{i=1}^n X_{ij}^2 \leq Cn$ for a constant C, form lemma 2

$$\max_{j=1,\ldots,p} \|\sum_{i=1}^n X_{ij}\varepsilon_i\|_m = O(\sqrt{n}) \Rightarrow |\varepsilon^T X\omega| \leq |X^T\varepsilon|_\infty \times |\omega|_1 = O_p(n^{(\alpha_p/m)+1/2} \times |\omega|_1) \tag{B.3.4}$$

Therefore, for sufficiently large n, with probability tending to 1

$$0 \leq -\frac{\rho_{n,l}}{2}\sum_{j\notin\mathscr{N}_{b_n}}|\omega_j| + \frac{3\rho_{n,l}}{2}\sum_{j\in\mathscr{N}_{b_n}}|\omega_j| \Rightarrow \omega \in \mathscr{A} \text{ (defined in (3.24))} \tag{B.3.5}$$

Therefore, with probability tending to 1

$$\begin{aligned} \frac{c_\lambda^2|\omega|_2^2}{2} \leq \frac{3\rho_{n,l}}{2}\sum_{j\in\mathscr{N}_{b_n}}|\omega_j| &\leq \frac{3\rho_{n,l}}{2}\sqrt{|\mathscr{N}_{b_n}|} \times |\omega|_2 \\ &\Rightarrow |\widetilde{\beta}^{lasso} - \beta|_2 = O_p(n^{\frac{\alpha_{\mathscr{N}}}{2} - \alpha_l}) \end{aligned} \tag{B.3.6}$$

Meanwhile, from (B.3.5)

$$|\widetilde{\beta}^{lasso}-\beta|_1 \le 4\sum_{j\in\mathscr{N}_{b_n}}|\omega_j| \le 4\sqrt{|\mathscr{N}_{b_n}|}\times|\omega|_2 = O_p(n^{\alpha_{\mathscr{N}}-\alpha_l}) \tag{B.3.7}$$

□

proof of theorem 5. From (3.21), lemma 2 and lemma B.3.1, since

$$\begin{aligned}
\rho_{n,r}^2|\sum_{j=1}^{r}\frac{q_{ij}s_j}{(\lambda_j^2+\rho_{n,r})^2}| &\le \frac{\rho_{n,r}^2}{\lambda_r^4}\sqrt{\sum_{j=1}^{r}q_{ij}^2}\times\sqrt{\sum_{j=1}^{r}s_j^2} \le \frac{\rho_{n,r}^2\times|\beta|_2}{\lambda_r^4}\\
&\|\sum_{j=1}^{r}\sum_{l=1}^{n}q_{ij}p_{lj}\left(\frac{\lambda_j}{\lambda_j^2+\rho_{n,r}}+\frac{\rho_{n,r}\lambda_j}{(\lambda_j^2+\rho_{n,r})^2}\right)\varepsilon_l\|_m\\
&\le C\sqrt{\sum_{l=1}^{n}\left(\sum_{j=1}^{r}q_{ij}p_{lj}\left(\frac{\lambda_j}{\lambda_j^2+\rho_{n,r}}+\frac{\rho_{n,r}\lambda_j}{(\lambda_j^2+\rho_{n,r})^2}\right)\right)^2}\\
&= C\sqrt{\sum_{j=1}^{r}q_{ij}^2\left(\frac{\lambda_j}{\lambda_j^2+\rho_{n,r}}+\frac{\rho_{n,r}\lambda_j}{(\lambda_j^2+\rho_{n,r})^2}\right)^2} \le \frac{2C}{\lambda_r}\\
\Rightarrow \max_{i=1,\ldots,p}|\sum_{j=1}^{r}\sum_{l=1}^{n}q_{ij}p_{lj}&\left(\frac{\lambda_j}{\lambda_j^2+\rho_{n,r}}+\frac{\rho_{n,r}\lambda_j}{(\lambda_j^2+\rho_{n,r})^2}\right)\varepsilon_l| = O_p(n^{\alpha_p/m-\eta})\\
&|\widetilde{\theta}_{\perp}^{\dagger}-\theta_{\perp}|_{\infty} = O_p(n^{\frac{\alpha_{\mathscr{N}}}{2}-\alpha_l})
\end{aligned} \tag{B.3.8}$$

we have

$$|\widetilde{\beta}-\beta|_{\infty} = O_p\left(n^{2\alpha_r+\frac{\alpha_{\mathscr{N}}}{2}-4\eta}+n^{\alpha_p/m-\eta}+n^{\frac{\alpha_{\mathscr{N}}}{2}-\alpha_l}\right) \tag{B.3.9}$$

From assumption 4, $2\alpha_r+\frac{\alpha_{\mathscr{N}}}{2}-4\eta<-\eta$, so $|\widetilde{\beta}-\beta|_{\infty}=O_p(n^{\alpha_p/m-\eta}+n^{\frac{\alpha_{\mathscr{N}}}{2}-\alpha_l})$

In particular

$$\begin{aligned}
Prob\left(\widehat{\mathscr{N}_{b_n}}\neq\mathscr{N}_{b_n}\right) &\le Prob\left(\min_{i\in\mathscr{N}_{b_n}}|\widetilde{\beta}_i|\le b_n\right)+Prob\left(\max_{i\notin\mathscr{N}_{b_n}}|\widetilde{\beta}_i|>b_n\right)\\
&\le Prob\left(|\widetilde{\beta}-\beta|_{\infty}\ge(\frac{1}{c_b}-1)b_n\right)+Prob\left(|\widetilde{\beta}-\beta|_{\infty}>b_n\right)
\end{aligned} \tag{B.3.10}$$

and we prove (3.26). □

proof of theorem 6. Suppose $\widehat{\mathscr{N}_{b_n}} = \mathscr{N}_{b_n}$, then $\widehat{c}_{ij} = c_{ij}$ and $\widehat{\tau}_i = \tau_i$. If $i \in \mathscr{M}$, from Cauchy inequality

$$\frac{\rho_{n,r}^2}{\tau_i}|\sum_{j=1}^{r}\frac{c_{ij}s_j}{(\lambda_j^2+\rho_{n,r})^2}| \leq \frac{\rho_{n,r}^2}{\tau_i} \times \sqrt{\sum_{j=1}^{r}\frac{c_{ij}^2\lambda_j^2}{(\lambda^2+\rho_{n,r})^2}}\sqrt{\sum_{j=1}^{r}\frac{s_j^2}{\lambda_j^2(\lambda_j^2+\rho_{n,r})^2}} \leq \frac{\rho_{n,r}^2|\beta|_2}{\lambda_r^3} \quad \text{(B.3.11)}$$

which has order $o(1)$. Besides,

$$\begin{aligned}
\frac{1}{\tau_i^2}\sum_{l=1}^{n}w_{il}^2 = \frac{1}{\tau_i^2}\sum_{j=1}^{r}c_{ij}^2\left(\frac{\lambda_j}{\lambda_j^2+\rho_{n,r}}+\frac{\rho_{n,r}\lambda_j}{(\lambda_j^2+\rho_{n,r})^2}\right)^2 \leq 1 \\
\text{and for sufficiently large } n \quad \sum_{j=1}^{n}c_{ij}^2\left(\frac{\lambda_j}{\lambda_j^2+\rho_{n,r}}+\frac{\rho_{n,r}\lambda_j}{(\lambda_j^2+\rho_{n,r})^2}\right)^2 \geq \frac{c_{\mathscr{M}}}{4C_\lambda^2 n} \quad \text{(B.3.12)} \\
\Rightarrow \frac{1}{\tau_i^2}\sum_{l=1}^{n}w_{il}^2 = \frac{\sum_{j=1}^{n}c_{ij}^2\left(\frac{\lambda_j}{\lambda_j^2+\rho_{n,r}}+\frac{\rho_{n,r}\lambda_j}{(\lambda_j^2+\rho_{n,r})^2}\right)^2}{\sum_{j=1}^{n}c_{ij}^2\left(\frac{\lambda_j}{\lambda_j^2+\rho_{n,r}}+\frac{\rho_{n,r}\lambda_j}{(\lambda_j^2+\rho_{n,r})^2}\right)^2+\frac{1}{n}} \geq \frac{1}{1+\frac{4C_\lambda^2}{c_{\mathscr{M}}}}
\end{aligned}$$

From lemma 3

$$\sup_{x\in\mathbf{R}}|Prob\left(\max_{i=1,\ldots,p_1}|\frac{1}{\tau_i}\sum_{l=1}^{n}w_{il}\varepsilon_l| \leq x\right) - Prob\left(\max_{i\in\mathscr{M}}|\xi_i| \leq x\right)| = o(1) \quad \text{(B.3.13)}$$

From assumption 7 and lemma B.3.1

$$\begin{aligned}
\frac{1}{\tau_i}|\sum_{j\in\mathscr{N}_{b_n}}\sum_{k\neq j}\frac{m_{ij}V_{jk}}{1-V_{jj}}(\widetilde{\beta}_k^{lasso}-\beta_k)| &= \frac{1}{\tau_i}|\sum_{k=1}^{p}(\widetilde{\beta}_k^{lasso}-\beta_k) \times \sum_{j\in\mathscr{N}_{b_n},j\neq k}\frac{m_{ij}V_{jk}}{1-V_{jj}}| \\
&\leq |\widetilde{\beta}^{lasso}-\beta|_1 \times \max_{k=1,\ldots,p}|\frac{1}{\tau_i}\sum_{j\in\mathscr{N}_{b_n},j\neq k}\frac{m_{ij}V_{jk}}{1-V_{jj}}| \quad \text{(B.3.14)} \\
\Rightarrow \max_{i=1,\ldots,p_1}\frac{1}{\tau_i}|\sum_{j\in\mathscr{N}_{b_n}}\sum_{k\neq j}\frac{m_{ij}V_{jk}}{1-V_{jj}}(\widetilde{\beta}_k^{lasso}-\beta_k)| &= o_p(1)
\end{aligned}$$

Define

$$L = \frac{\rho_{n,r}^2 |\beta|_2}{\lambda_r^3} + \max_{i=1,\ldots,p_1} \frac{1}{\tau_i} | \sum_{j \in \mathscr{N}_{b_n}} \sum_{k \neq j} \frac{m_{ij} V_{jk}}{1 - V_{jj}} (\widetilde{\beta}_k^{lasso} - \beta_k) | = o_p(1) \tag{B.3.15}$$

then for sufficiently large n

$$\begin{aligned}
\max_{i \in \mathscr{M}} | \frac{1}{\tau_i} \sum_{l=1}^{m} w_{il} \varepsilon_l | - L \leq \max_{i=1,\ldots,p_1} \frac{|\widehat{\zeta}_i - \zeta_i|}{\tau_i} \leq L + \max_{i \in \mathscr{M}} | \frac{1}{\tau_i} \sum_{l=1}^{m} w_{il} \varepsilon_l | \\
\Rightarrow Prob \left(\max_{i=1,\ldots,p_1} \frac{|\widehat{\zeta}_i - \zeta_i|}{\widehat{\tau}_i} \leq x \right) - Prob \left(\max_{i \in \mathscr{M}} |\xi_i| \leq x \right) \\
\leq Prob(\widehat{\mathscr{N}_{b_n}} \neq \mathscr{N}_{b_n}) + Prob \left(\max_{i \in \mathscr{M}} | \frac{1}{\tau_i} \sum_{l=1}^{m} w_{il} \varepsilon_l | \leq x + L \right) \\
- Prob \left(\max_{i \in \mathscr{M}} |\xi_i| \leq x \right) \\
\text{and } Prob \left(\max_{i=1,\ldots,p_1} \frac{|\widehat{\zeta}_i - \zeta_i|}{\widehat{\tau}_i} \leq x \right) - Prob \left(\max_{i \in \mathscr{M}} |\xi_i| \leq x \right) \\
\geq - Prob \left(\widehat{\mathscr{N}_{b_n}} \neq \mathscr{N}_{b_n} \right) + Prob \left(\max_{i \in \mathscr{M}} | \frac{1}{\tau_i} \sum_{l=1}^{m} w_{il} \varepsilon_l | \leq x - L \right) \\
- Prob \left(\max_{i \in \mathscr{M}} |\xi_i| \leq x \right)
\end{aligned} \tag{B.3.16}$$

Apply lemma B.2.1, we prove (3.41). □

B.4 Proof of theorems in section 3.5

proof of theorem 7. From theorem 5, we have $Prob \left(\widehat{\mathscr{N}_{b_n}} \neq \mathscr{N}_{b_n} \right) \to 0$ as the sample size $n \to \infty$. If $\widehat{\mathscr{N}_{b_n}} = \mathscr{N}_{b_n}$, then $\widehat{c}_{ij} = c_{ij}$, $\widehat{\tau}_i = \tau_i$, $\widehat{w}_{il} = w_{il}$ and $\widehat{\Gamma}_{ij} = \frac{1}{\tau_i \tau_j} \sum_{l_1=1}^{n} \sum_{l_2=1}^{n} \sigma_{l_1 l_2} \times w_{il_1} w_{jl_2} = 0$ for all pairs i, j such that $i \notin \mathscr{M}$ or $j \notin \mathscr{M}$. From lemma 4 and (B.3.12),

$$\max_{i,j \in \mathscr{M}} | \sum_{l_1=1}^{n} \sum_{l_2=1}^{n} \frac{w_{il_1} w_{jl_2}}{\tau_i \tau_j} \times \left(\varepsilon_{l_1} \varepsilon_{l_2} K \left(\frac{l_1 - l_2}{k_n} \right) - \sigma_{l_1 l_2} \right) | = o_p(1) \tag{B.4.1}$$

Besides, if $\widehat{\mathscr{N}_{b_n}} = \mathscr{N}_{b_n}$, define $\delta_i = \sum_{j=1}^{p} X_{ij}(\widehat{\beta}_j - \beta_j)$, from section 0.9.7 in Horn and Johnson [2013]

$$\begin{aligned}
&\max_{i,j\in\mathscr{M}} |\widehat{\Gamma}_{ij} - \sum_{l_1=1}^{n}\sum_{l_2=1}^{n} \frac{w_{il_1}w_{jl_2}}{\tau_i\tau_j} K\left(\frac{l_1-l_2}{k_n}\right)\varepsilon_{l_1}\varepsilon_{l_2}| \\
&\leq \max_{i,j\in\mathscr{M}} |\sum_{l_1=1}^{n}\sum_{l_2=1}^{n} \frac{w_{il_1}w_{jl_2}}{\tau_i\tau_j} K\left(\frac{l_1-l_2}{k_n}\right)\varepsilon_{l_1}\delta_{l_2}| \\
&+ \max_{i,j\in\mathscr{M}} |\sum_{l_1=1}^{n}\sum_{l_2=1}^{n} \frac{w_{il_1}w_{jl_2}}{\tau_i\tau_j} K\left(\frac{l_1-l_2}{k_n}\right)\delta_{l_1}\varepsilon_{l_2}| \\
&+ \max_{i,j\in\mathscr{M}} |\sum_{l_1=1}^{n}\sum_{l_2=1}^{n} \frac{w_{il_1}w_{jl_2}}{\tau_i\tau_j} K\left(\frac{l_1-l_2}{k_n}\right)\delta_{l_1}\delta_{l_2}| \\
&\leq 2\sum_{l=0}^{\infty} K\left(\frac{l}{k_n}\right) \times \sqrt{\sum_{l=1}^{n}\frac{w_{il}^2\varepsilon_l^2}{\tau_i^2}} \times \sqrt{\sum_{l=1}^{n}\frac{w_{jl}^2\delta_l^2}{\tau_j^2}} \\
&+2\sum_{l=0}^{\infty} K\left(\frac{l}{k_n}\right) \times \sqrt{\sum_{l=1}^{n}\frac{w_{jl}^2\varepsilon_l^2}{\tau_j^2}} \times \sqrt{\sum_{l=1}^{n}\frac{w_{il}^2\delta_l^2}{\tau_i^2}} \\
&+2\sum_{l=0}^{\infty} K\left(\frac{l}{k_n}\right) \times \sqrt{\sum_{l=1}^{n}\frac{w_{il}^2\delta_l^2}{\tau_i^2}} \times \sqrt{\sum_{l=1}^{n}\frac{w_{jl}^2\delta_l^2}{\tau_j^2}} \\
&\leq Ck_n \times \left(2\max_{l=1,\ldots,n}|\varepsilon_l| \times \max_{l=1,\ldots,n}|\delta_l| + \max_{l=1,\ldots,n}|\delta_l|^2\right)
\end{aligned} \tag{B.4.2}$$

here C is a constant. Define $U_{ij} = \sum_{k\in\mathscr{N}_{b_n}} X_{ik}q_{kj}$. Then

$$\max_{i=1,\ldots,n}\sum_{j=1}^{r} U_{ij}^2 \leq \max_{i=1,\ldots,n}\sum_{k\in\mathscr{N}_{b_n}} X_{ik}^2 = O(|\mathscr{N}_{b_n}|) \tag{B.4.3}$$

If $\widehat{\mathscr{N}_{b_n}} = \mathscr{N}_{b_n}$, then from (3.21)

$$\begin{aligned}
\delta_i &= \sum_{j\in\mathscr{N}_{b_n}} X_{ij}(\widetilde{\beta}_j - \beta_j) \\
&= -\rho_{n,r}^2 \sum_{k=1}^{r} \frac{U_{ik}s_k}{(\lambda_k^2+\rho_{n,r})^2} + \sum_{k=1}^{r}\sum_{l=1}^{n} U_{ik}p_{lk}\left(\frac{\lambda_k}{\lambda_k^2+\rho_{n,r}} + \frac{\rho_{n,r}\lambda_k}{(\lambda_k^2+\rho_{n,r})^2}\right)\varepsilon_l \\
&\quad + \sum_{j\in\mathscr{N}_{b_n}} X_{ij}(\widetilde{\theta}_{\perp,j}^{\dagger} - \theta_{\perp,j})
\end{aligned} \tag{B.4.4}$$

For

$$\begin{aligned}
\max_{i=1,\ldots,p} |\rho_{n,r}^2 \sum_{k=1}^{r} \frac{U_{ik}s_k}{(\lambda_k^2+\rho_{n,r})^2}| &\leq \rho_{n,r}^2 \times \max_{i=1,\ldots,p} \frac{\sqrt{\sum_{k=1}^r U_{ik}^2}\sqrt{\sum_{k=1}^r s_k^2}}{\lambda_r^4} \\
&= O\left(n^{2\alpha_r-4\eta+\alpha_{\mathscr{N}}}\right), \\
\| \sum_{k=1}^{r}\sum_{l=1}^{n} U_{ik}p_{lk}\left(\frac{\lambda_k}{\lambda_k^2+\rho_{n,r}} + \frac{\rho_{n,r}\lambda_k}{(\lambda_k^2+\rho_{n,r})^2}\right)\varepsilon_l\|_m & \\
\leq C\left(\sqrt{\sum_{l=1}^{n}\left(\sum_{k=1}^{r} U_{ik}p_{lk}\left(\frac{\lambda_k}{\lambda_k^2+\rho_{n,r}} + \frac{\rho_{n,r}\lambda_k}{(\lambda_k^2+\rho_{n,r})^2}\right)\right)^2}\right) & \quad \text{(B.4.5)} \\
= C\sqrt{\sum_{k=1}^{r} U_{ik}^2\left(\frac{\lambda_k}{\lambda_k^2+\rho_{n,r}} + \frac{\rho_{n,r}\lambda_k}{(\lambda_k^2+\rho_{n,r})^2}\right)^2} \leq \frac{C'\sqrt{|\mathscr{N}_{b_n}|}}{\lambda_r} &
\end{aligned}$$

$$\Rightarrow \max_{i=1,\ldots,n} |\sum_{k=1}^{r}\sum_{l=1}^{n} U_{ik}p_{lk}\left(\frac{\lambda_k}{\lambda_k^2+\rho_{n,r}} + \frac{\rho_{n,r}\lambda_k}{(\lambda_k^2+\rho_{n,r})^2}\right)\varepsilon_l| = O_p\left(n^{1/m+\alpha_{\mathscr{N}}/2-\eta}\right)$$

$$\text{and } \max_{i=1,\ldots,p} |\sum_{j\in\mathscr{N}_{b_n}} X_{ij}(\widetilde{\theta}_{\perp,j}^{\dagger} - \theta_{\perp,j})| \leq \max_{i=1,\ldots,p}\sqrt{\sum_{j\in\mathscr{N}_{b_n}} X_{ij}^2}|\widetilde{\theta}_{\perp}^{\dagger} - \theta_{\perp}|_2 = O_p\left(n^{\alpha_{\mathscr{N}}-\alpha_l}\right)$$

Form assumption 4, $2\alpha_r - 4\eta + \alpha_{\mathscr{N}} < -\eta + \frac{1}{2}\alpha_{\mathscr{N}}$, so

$$\max_{i=1,\ldots,n} |\delta_i| = O_p\left(n^{1/m+\alpha_{\mathscr{N}}/2-\eta} + n^{\alpha_{\mathscr{N}}-\alpha_l}\right) \quad \text{(B.4.6)}$$

For $\max_{i=1,\ldots,n}|\varepsilon_i| = O_p(n^{1/m})$, from assumption 8

$$k_n \times \max_{l=1,\ldots,n} |\varepsilon_l| \times \max_{l=1,\ldots,n} |\delta_l| = O_p(k_n n^{\frac{2}{m}+\frac{\alpha_{\mathscr{N}}}{2}-\eta} + k_n n^{\frac{1}{m}+\alpha_{\mathscr{N}}-\alpha_l}) = o_p(1) \quad \text{(B.4.7)}$$

and we prove (3.48). □

proof of theorem 8. Suppose $\widehat{\mathscr{N}_{b_n}} = \mathscr{N}_{b_n}$. Define $\widehat{s} = (\widehat{s}_1, ..., \widehat{s}_r)^T = Q^T \widehat{\beta}$.

$$\begin{aligned} \widetilde{\beta}^* - \widehat{\beta} &= -\rho_{n,r}^2 Q \left(\Lambda^2 + \rho_{n,r}\right)^{-2} Q^T \widehat{\beta} + Q\left((\Lambda^2 + \rho_{n,r})^{-1} + \rho_{n,r}(\Lambda^2 + \rho_{n,r})^{-2}\right) \Lambda P^T \varepsilon^* \\ \Rightarrow \widetilde{\beta}_i^* - \widehat{\beta}_i &= -\rho_{n,r}^2 \sum_{j=1}^{r} \frac{q_{ij}\widehat{s}_j}{(\lambda_j^2 + \rho_{n,r})^2} + \sum_{j=1}^{r}\sum_{l=1}^{n} q_{ij} p_{lj} \left(\frac{\lambda_j}{\lambda_j^2 + \rho_{n,r}} + \frac{\rho_{n,r}\lambda_j}{(\lambda_j^2 + \rho_{n,r})^2}\right) \varepsilon_l^* \end{aligned} \tag{B.4.8}$$

Define the conditional norm $\|\cdot\|_m^* = (\mathbf{E}^*|\cdot|^m)^{1/m}, m \geq 1$ and $t_{il} = \sum_{j=1}^{r} q_{ij} p_{lj} \left(\frac{\lambda_j}{\lambda_j^2 + \rho_{n,r}} + \frac{\rho_{n,r}\lambda_j}{(\lambda_j^2 + \rho_{n,r})^2}\right)$. For ε_l^* has joint normal distribution,

$$\begin{aligned} \| \sum_{j=1}^{r}\sum_{l=1}^{n} q_{ij} p_{lj} \left(\frac{\lambda_j}{\lambda_j^2 + \rho_{n,r}} + \frac{\rho_{n,r}\lambda_j}{(\lambda_j^2 + \rho_{n,r})^2}\right) \varepsilon_l^* \|_m^* &= C \| \sum_{l=1}^{n} t_{il} \varepsilon_l^* \|_2^* \\ = C \sqrt{\sum_{l_1=1}^{n}\sum_{l_2=1}^{n} t_{il_1} t_{il_2} \widehat{\varepsilon}_{l_1} \widehat{\varepsilon}_{l_2} K\left(\frac{l_1 - l_2}{k_n}\right)} &\leq C \sqrt{\left(2\sum_{i=0}^{\infty} K(i/k_n)\right) \times \left(\sum_{l=1}^{n} t_{il}^2 \widehat{\varepsilon}_l^2\right)} \\ &\leq \frac{C'\sqrt{k_n}}{\lambda_r} \times \max_{i=1,...,n} |\widehat{\varepsilon}_i| \\ \Rightarrow \| \max_{i=1,2,...,p} | \sum_{j=1}^{r}\sum_{l=1}^{n} q_{ij} p_{lj} \left(\frac{\lambda_j}{\lambda_j^2 + \rho_{n,r}} + \frac{\rho_{n,r}\lambda_j}{(\lambda_j^2 + \rho_{n,r})^2}\right) \varepsilon_l^* | \|_m^* \\ &\leq \frac{C' p^{1/m}\sqrt{k_n}}{\lambda_r} \times \max_{i=1,...,n} |\widehat{\varepsilon}_i| \end{aligned} \tag{B.4.9}$$

here C and C' are constants. From (B.4.6), $\max_{i=1,...,n} |\widehat{\varepsilon}_i| \leq \max_{i=1,...,n} |\varepsilon_i| + \max_{i=1,...,n} |\widehat{\varepsilon}_i - \varepsilon_i| = O_p(n^{1/m})$.

On the other hand,

$$\begin{aligned} \max_{i=1,...,p} |\rho_{n,r}^2 \sum_{j=1}^{r} \frac{q_{ij}\widehat{s}_j}{(\lambda_j^2 + \rho_{n,r})^2}| &\leq \max_{i=1,2,...,p} \frac{\rho_{n,r}^2}{\lambda_r^4} \times \sqrt{\sum_{j=1}^{r} q_{ij}^2} \times |\widehat{\beta}|_2 \\ &\leq \frac{\rho_{n,r}^2}{\lambda_r^4} \times (|\beta|_2 + |\widehat{\beta} - \beta|_2) \end{aligned} \tag{B.4.10}$$

For

$$|\widehat{\beta} - \beta|_2 = \sqrt{\sum_{j \in \mathscr{N}_{b_n}} |\widetilde{\beta}_j - \beta_j|^2} \leq \sqrt{|\mathscr{N}_{b_n}|} \times |\widetilde{\beta} - \beta|_\infty = o_p(\sqrt{|\mathscr{N}_{b_n}|}) \tag{B.4.11}$$

we have for sufficiently large n,

$$\begin{aligned} Prob^*\left(\widehat{\mathscr{N}_{b_n}^*} \neq \mathscr{N}_{b_n}\right) &\leq Prob^*\left(\min_{i\in\mathscr{N}_{b_n}} |\widetilde{\beta}_i^*| \leq b_n\right) + Prob^*\left(\max_{i\notin\mathscr{N}_{b_n}} |\widetilde{\beta}_i^*| > b_n\right) \\ &\leq Prob^*\left(|\widetilde{\beta}^* - \widehat{\beta}|_\infty \geq (1/c_b - 1)b_n - |\widehat{\beta} - \beta|_\infty\right) \\ &\quad + Prob^*\left(|\widetilde{\beta}^* - \widehat{\beta}|_\infty \geq b_n - |\widehat{\beta} - \beta|_\infty\right) \\ &\leq \frac{\| \, |\widetilde{\beta}^* - \widehat{\beta}|_\infty \, \|_m^{*m}}{((1/c_b - 1)b_n - |\widehat{\beta} - \beta|_\infty)^m} + \frac{\| \, |\widetilde{\beta}^* - \widehat{\beta}|_\infty \, \|_m^{*m}}{(b_n - |\widehat{\beta} - \beta|_\infty)^m} \end{aligned} \tag{B.4.12}$$

which tends to 0 with probability tending to 1. If $\widehat{\mathscr{N}_{b_n}^*} = \widehat{\mathscr{N}_{b_n}} = \mathscr{N}_{b_n}$, then $\widehat{c}_{ij}^* = c_{ij}$ and $\widehat{\tau}_i^* = \tau_i$. Moreover,

$$\begin{aligned} \frac{\widehat{\zeta}_i^* - \sum_{j=1}^p m_{ij}\widehat{\beta}_j}{\widehat{\tau}_i^*} &= -\frac{\rho_{n,r}^2}{\tau_i}\sum_{k=1}^r \frac{c_{ik}\widehat{s}_k}{(\lambda_k^2 + \rho_{n,r})^2} \\ &+ \frac{1}{\tau_i}\sum_{k=1}^r\sum_{l=1}^n c_{ik}p_{lk}\left(\frac{\lambda_k}{\lambda_k^2 + \rho_{n,r}} + \frac{\rho_{n,r}\lambda_k}{(\lambda_k^2 + \rho_{n,r})^2}\right)\varepsilon_l^* \end{aligned} \tag{B.4.13}$$

From Cauchy inequality,

$$\begin{aligned} |\frac{\rho_{n,r}^2}{\tau_i}\sum_{k=1}^r \frac{c_{ik}\widehat{s}_k}{(\lambda_k^2 + \rho_{n,r})^2}| &\leq \frac{\rho_{n,r}^2}{\tau_i}\sqrt{\sum_{k=1}^r \frac{c_{ik}^2\lambda_k^2}{(\lambda_k^2 + \rho_{n,r})^2}} \times \sqrt{\sum_{k=1}^r \frac{\widehat{s}_k^2}{\lambda_k^2 \times (\lambda_k^2 + \rho_{n,r})^2}} \\ &\leq \frac{\rho_{n,r}^2 \times |\widehat{\beta}|_2}{\lambda_r^3} \end{aligned} \tag{B.4.14}$$

which has order $O_p(n^{2\alpha_r + \alpha_{\mathscr{N}}/2 - 3\eta})$. Define w_{il} as in (3.35), we have

$$\mathbf{E}^*\frac{1}{\tau_i}\frac{1}{\tau_j}\sum_{l_1=1}^n\sum_{l_2=1}^n w_{il_1}w_{jl_2}\varepsilon_{l_1}^*\varepsilon_{l_2}^* = \frac{1}{\tau_i}\frac{1}{\tau_j}\sum_{l_1=1}^n\sum_{l_2=1}^n w_{il_1}w_{jl_2}\widehat{\varepsilon}_{l_1}\widehat{\varepsilon}_{l_2}K\left(\frac{l_1 - l_2}{k_n}\right) \tag{B.4.15}$$

From lemma B.2.1 and theorem 7, we have

$$\sup_{x\in\mathbf{R}} |Prob^*\left(\max_{i\in\mathscr{M}} |\frac{1}{\tau_i}\sum_{l=1}^n w_{il}\varepsilon_l^*| \leq x\right) - Prob\left(\max_{i\in\mathscr{M}} |\xi_i| \leq x\right)| = o_p(1) \tag{B.4.16}$$

For

$$
\begin{aligned}
Prob^*\left(\max_{i=1,\ldots,p_1}\frac{|\widehat{\zeta}_i^*-\sum_{j=1}^p m_{ij}\widehat{\beta}_j|}{\widehat{\tau}_i^*}\le x\right) &\le Prob^*\left(\widehat{\mathscr{N}_{b_n}^*}\neq\mathscr{N}_{b_n}\right)\\
&+Prob^*\left(\max_{i\in\mathscr{M}}|\frac{1}{\tau_i}\sum_{l=1}^n w_{il}\varepsilon_l^*|\le x+\frac{\rho_{n,r}^2\times|\widehat{\beta}|_2}{\lambda_r^3}\right)\\
Prob^*\left(\max_{i=1,\ldots,p_1}\frac{|\widehat{\zeta}_i^*-\sum_{j=1}^p m_{ij}\widehat{\beta}_j|}{\widehat{\tau}_i^*}\le x\right) &\ge -Prob^*\left(\widehat{\mathscr{N}_{b_n}^*}\neq\mathscr{N}_{b_n}\right)\\
&+Prob^*\left(\max_{i\in\mathscr{M}}|\frac{1}{\tau_i}\sum_{l=1}^n w_{il}\varepsilon_l^*|\le x-\frac{\rho_{n,r}^2\times|\widehat{\beta}|_2}{\lambda_r^3}\right)
\end{aligned}
\tag{B.4.17}
$$

and (3.56) is proven. □

Appendix C

Proofs of theorems in chapter 4

C.1 Proof of theorem 9 in section 4.5

Adopt the notations in section 4.4 and 4.5 of the paper. Suppose assumption 1 to 4 in section 4.4 hold true. For any given positive integer $0 < m < \infty$, define $x_m = 2mx - m$ and the stochastic process

$$\widetilde{M}_m(x) = \sqrt{n}F^{'}(x_m)\left(\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon - \frac{1}{n}\sum_{j=1}^{n}\varepsilon_j\right) - \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left(\mathbf{1}_{\varepsilon_j \le x_m} - F(x_m)\right) \quad \text{(C.1.1)}$$

Then $\mathbf{E}\widetilde{M}_m(x) = 0$ for any given x.

Define the Gaussian process $\mathscr{M}_m(x) \in \mathbf{D}, x \in [0,1]$ such that

$$\mathbf{E}\mathscr{M}_m(x) = 0,\ \mathbf{E}\mathscr{M}_m(x)\mathscr{M}_m(z) = \mathscr{V}(x_m, z_m) \text{ for } \forall x, z \in [0,1] \quad \text{(C.1.2)}$$

here $z_m = 2mz - m$ and $\mathscr{M}_m$ has continuous sample paths almost surely. $\mathscr{V}$ is defined in (4.15) of the paper. The proof of theorem 9 has 4 steps:

1. Show the existence of $\mathscr{M}_m$ for any m
2. Prove that $\widetilde{M}_m \to_{\mathscr{L}} \mathscr{M}_m$ (i.e., converges in distribution) under the Skohord topology. Then lemma C.1.1 below implies that $\widetilde{M}_m(x)$'s sample paths will be similar to a continuous function, i.e., $|\widetilde{M}_m(x) - \widetilde{M}_m(z)|$ can be arbitrarily small as $|x - z| \to 0$ with probability

tending to 1.

3. Prove that the random variable $\widetilde{M}_m\left(\frac{x+m}{2m}\right) - \widetilde{M}_m^-(\frac{-x+m}{2m})$'s asymptotic distribution will be a normal distribution with mean 0 and variance $\mathscr{U}(x)$ for any $x \in (0,m]$. See (4.16) in the paper for the definition of the superscript $^-$.

4. Approximate $\mathscr{S}(x)$(see (4.19)) by $\widetilde{M}_m\left(\frac{x+m}{2m}\right) - \widetilde{M}_m^-(\frac{-x+m}{2m})$.

But before presenting the proof, we would like to introduce some useful lemmas.

C.1.1 Useful lemmas

Suppose random variables A, B satisfy $|A-B| \leq \delta, \delta > 0$. Then $\forall x \in \mathbf{R}$, $-\mathbf{1}_{x-\delta<B\leq x} \leq \mathbf{1}_{A\leq x} - \mathbf{1}_{B\leq x} \leq \mathbf{1}_{x<B\leq x+\delta}$, which implies

$$\begin{aligned} \mathbf{E}|\mathbf{1}_{A\leq x} - \mathbf{1}_{B\leq x}| &\leq \mathbf{E}|\mathbf{1}_{A\leq x} - \mathbf{1}_{B\leq x}| \times \mathbf{1}_{|A-B|\leq\delta} + Prob(|A-B|>\delta) \\ &\leq Prob(|A-B|>\delta) + Prob(x-\delta < B \leq x+\delta) \end{aligned} \tag{C.1.3}$$

For any given positive integer r and $\forall t_i \in [0,1], s_i \in \mathbf{R}, i = 1,2,...,r$, define $t_{i,m} = 2mt_i - m$, then

$$\begin{aligned} 0 &\leq \lim_{n\to\infty} \mathbf{E}\left(\sum_{i=1}^r s_i \widetilde{M}_m(t_i)\right)^2 \\ &= \lim_{n\to\infty} \sigma^2 \sum_{i=1}^r \sum_{j=1}^r s_i s_j F^{'}(t_{i,m}) F^{'}(t_{j,m}) \\ &\quad \times \left(\mathscr{X}_f^T \left(\frac{\mathscr{X}^T\mathscr{X}}{n}\right)^{-1} \mathscr{X}_f + 1 - 2\mathscr{X}_f^T \left(\frac{\mathscr{X}^T\mathscr{X}}{n}\right)^{-1} \overline{\mathscr{X}}_n \right) \\ &\quad - \lim_{n\to\infty} \sum_{i=1}^r \sum_{j=1}^r s_i s_j \left(\mathscr{X}_f^T \left(\frac{\mathscr{X}^T\mathscr{X}}{n}\right)^{-1} \overline{\mathscr{X}}_n - 1 \right) \times \left(F^{'}(t_{i,m})H(t_{j,m}) + F^{'}(t_{j,m})H(t_{i,m})\right) \\ &\quad + \lim_{n\to\infty} \sum_{i=1}^r \sum_{j=1}^r s_i s_j \left(F(\min(t_{i,m},t_{j,m})) - F(t_{i,m})F(t_{j,m})\right) \\ &= \sum_{i=1}^r \sum_{j=1}^r s_i s_j \mathscr{V}(t_{i,m}, t_{j,m}) \end{aligned} \tag{C.1.4}$$

Eq.(C.1.4) implies that $\mathscr{V}(2m\cdot -m, 2m\cdot -m)$ will be the asymptotic covariance function of the stochastic process $\widetilde{M}_m(\cdot)$. Moreover, for any real number sequence $\{z_i\}_{i=1,...,r}$, the matrix $\{\mathscr{V}(z_i,z_j)\}_{i,j=1,...,r}$ is positive semi-definite. From assumption 3, define $\widehat{\sigma}^2$ as in section (4.14) of the paper

$$\begin{aligned}\mathbf{E}\widehat{\sigma}^2 \leq \frac{2}{n}\sum_{i=1}^{n}\mathbf{E}\left(\varepsilon_i - \frac{\sum_{j=1}^{n}\varepsilon_j}{n}\right)^2 + \frac{2}{n}\sum_{i=1}^{n}\mathbf{E}((\mathscr{X}_i - \overline{\mathscr{X}}_n)^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon)^2 \\ \leq 2\sigma^2 + \frac{8M^2\sigma^2}{n}\|\left(\frac{\mathscr{X}^T\mathscr{X}}{n}\right)^{-1}\|_2\end{aligned} \tag{C.1.5}$$

so $\mathbf{E}\widehat{\sigma}^2 = O(1)$. Here $\|\left(\frac{\mathscr{X}^T\mathscr{X}}{n}\right)^{-1}\|_2$ is the matrix 2-norm of $\left(\frac{\mathscr{X}^T\mathscr{X}}{n}\right)^{-1}$. (C.1.3), (C.1.4) and (C.1.5) will be frequently used in the following sections. Then we introduce some lemmas. Lemma C.1.1 focuses on showing the existence of $\mathscr{M}_m$ and deriving its properties.

Lemma C.1.1. *Suppose assumption 1 to 4 hold true.*

1. For $\forall 0 < m \in \mathbf{N}$, $\exists$ *a Gaussian process* $\mathscr{M}_m$ *in* $\mathbf{D}$ *satisfying* (C.1.2) *and having continuous sample paths almost surely.*

2. For any given $\xi > 0$,

$$\lim_{\delta\to 0,\delta>0}\mathbf{P}\left(\sup_{y,z\in[0,1],|y-z|<\delta}|\mathscr{M}_m(y)-\mathscr{M}_m(z)| > \xi\right) = 0 \tag{C.1.6}$$

In addition, suppose a sequence of stochastic processes $\widetilde{\mathscr{N}_{m,n}} \in \mathbf{D}, n = 1,2,...$ *satisfy* $\widetilde{\mathscr{N}_{m,n}} \to_{\mathscr{L}} \mathscr{M}_m$ *under Skohord topology as* $n \to \infty$. *Then* $\exists\ \delta > 0$ *such that*

$$\limsup_{n\to\infty}\mathbf{P}\left(\sup_{y,z\in[0,1],|y-z|<\delta}|\widetilde{\mathscr{N}_{m,n}}(y)-\widetilde{\mathscr{N}_{m,n}}(z)| \geq \xi\right) \leq \xi \tag{C.1.7}$$

proof of lemma C.1.1. From (C.1.4), for any $t_i \in [0,1], i = 1,2,...,r$, the random vector $(\mathscr{M}_m(t_1),...,\mathscr{M}_m(t_r))^T$ has joint normal distribution with mean 0 and covariance matrix

$\{\mathscr{V}(2mt_i - m, 2mt_j - m)\}_{i,j=1,2,\ldots,r}$, so the consistency conditions in Kolmogorov extension theorem are satisfied. $\forall 0 \le t_1 \le t \le t_2 \le 1$,

$$\begin{aligned} &\mathbf{E}|\mathscr{M}_m(t) - \mathscr{M}_m(t_1)|^2 |\mathscr{M}_m(t) - \mathscr{M}_m(t_2)|^2 \\ \le &\frac{1}{2}\left(\mathbf{E}|\mathscr{M}_m(t) - \mathscr{M}_m(t_1)|^4 + \mathbf{E}|\mathscr{M}_m(t) - \mathscr{M}_m(t_2)|^4\right) \\ \le &\frac{3}{2}\left(\mathbf{E}(\mathscr{M}_m(t) - \mathscr{M}_m(t_1))^2 + \mathbf{E}(\mathscr{M}_m(t) - \mathscr{M}_m(t_2))^2\right)^2 \end{aligned} \tag{C.1.8}$$

The last inequality comes from the fact that $\mathscr{M}_m(t) - \mathscr{M}_m(t_1)$ and $\mathscr{M}_m(t) - \mathscr{M}_m(t_2)$ have normal distribution. Define $t_{i,m} = 2mt_i - m$ for $i = 1, 2$. Form assumption 1, $\exists$ a constant $C > 0$ with

$$\begin{aligned} &\mathbf{E}(\mathscr{M}_m(t) - \mathscr{M}_m(t_1))^2 \\ = &\sigma^2(\mathscr{X}_f^T A^{-1} \mathscr{X}_f + 1 - 2\mathscr{X}_f^T A^{-1} b)(F^{'}(2mt - m) - F^{'}(t_{1,m}))^2 \\ &+F(2mt - m) - F(t_{1,m}) - (F(2mt - m) - F(t_{1,m}))^2 \\ -2(\mathscr{X}_f^T A^{-1} b - 1)&(F^{'}(2mt - m) - F^{'}(t_{1,m}))(H(2mt - m) - H(t_{1,m})) \\ &\le C(t - t_1) \end{aligned} \tag{C.1.9}$$

Similarly, $\mathbf{E}(\mathscr{M}_m(t_2) - \mathscr{M}_m(t))^2 \le C(t_2 - t)$. Then (C.1.8) implies $\mathbf{E}|\mathscr{M}_m(t) - \mathscr{M}_m(t_1)|^2 |\mathscr{M}_m(t) - \mathscr{M}_m(t_2)|^2 \le \frac{3}{2}C^2(t_2 - t_1)^2$. Set $\alpha = \beta = 1$ and choose the non-decreasing, continuous function $F(x) = \frac{\sqrt{3}}{\sqrt{2}}Cx$ in eq.(13.15) of Billingsley [1999]. (C.1.9) also implies (13.16) in Billingsley [1999]. From theorem 13.6 in Billingsley [1999], $\exists\ \mathscr{M}_m \in \mathbf{D}$ satisfying (C.1.2). According to (C.1.9),

$$\mathbf{E}(\mathscr{M}_m(t) - \mathscr{M}_m(t_1))^4 \le 3C^2(t - t_1)^2 \tag{C.1.10}$$

so theorem 2.3 in Hahn [1977] is satisfied by choosing $r = 4$ and the function

$$f(x) = 3C^2x^2 \Rightarrow \int_{[0,1]} x^{-(r+1)/r} f^{1/r}(x)dx = 4(3C^2)^{1/4} < \infty \tag{C.1.11}$$

In particular, we can choose $\mathscr{M}_m \in \mathbf{D}$ such that $|\mathscr{M}_m(t) - \mathscr{M}_m(t_1)| \leq AH(|t-t_1|)$ almost surely, A is a random variable with $\mathbf{E}A^4 < \infty$, H is a continuous nondecreasing function on $[0,1]$ such that $H(0) = 0$. This implies $\mathscr{M}_m$ has continuous sample paths almost surely.

We prove (C.1.6) by

$$\mathbf{P}\left(\sup_{y,z\in[0,1],|y-z|<\delta} |\mathscr{M}_m(y) - \mathscr{M}_m(z)| > \xi\right) \leq \frac{\mathbf{E}A^4}{\xi^4} \times H^4(\delta) \tag{C.1.12}$$

.

For any given $\delta > 0$, define a function

$$h_\delta(f) = \sup_{x,y\in[0,1],|x-y|<\delta} |f(x) - f(y)|, \text{ here } f \in \mathbf{D} \tag{C.1.13}$$

From section 12, Billingsley [1999], if $f_n, n = 1, \ldots$ converges to f in $\mathbf{D}$, then $\exists$ strictly increasing mappings $\lambda_n : [0,1] \to [0,1], n = 1,2,\ldots$ such that
$\lim_{n\to\infty} \sup_{x\in[0,1]} |\lambda_n(x) - x| = 0$ and $\lim_{n\to\infty} \sup_{x\in[0,1]} |f_n(\lambda_n(x)) - f(x)| = 0$; so

$$\begin{aligned} |h_\delta(f_n) - h_\delta(f)| &\leq \sup_{x,y\in[0,1],|x-y|<\delta} |f_n(x) - f_n(y) - f(x) + f(y)| \\ &\leq \sup_{x\in[0,1]} |f_n(x) - f(x)| + \sup_{y\in[0,1]} |f_n(y) - f(y)| \\ &\leq 2(\sup_{x\in[0,1]} |f_n(x) - f(\lambda_n^{-1}(x))| + \sup_{x\in[0,1]} |f(\lambda_n^{-1}(x)) - f(x)|) \end{aligned} \tag{C.1.14}$$

If f is continuous on $[0,1]$, then $\lim_{n\to\infty} |h_\delta(f_n) - h_\delta(f)| = 0$. For $\mathscr{M}_m$ is continuous almost surely, and $\mathbf{R}, \mathbf{D}$ are Polish spaces(theorem 12.2 in Billingsley [1999]), 3.8, page 348 in Jacod and Shiryaev [2003] implies $h_\delta(\widetilde{\mathscr{N}_{m,n}}) \to_{\mathscr{L}} h_\delta(\mathscr{M}_m)$, and theorem 1.9 in Shao [2003] implies

$$\begin{aligned} &\limsup_{n\to\infty} \mathbf{P}\left(\sup_{x,y\in[0,1],|x-y|<\delta} |\widetilde{\mathscr{N}_{m,n}}(x) - \widetilde{\mathscr{N}_{m,n}}(y)| \geq \xi\right) \\ &\leq \mathbf{P}\left(\sup_{x,y\in[0,1],|x-y|<\delta} |\mathscr{M}_m(x) - \mathscr{M}_m(y)| \geq \xi\right) < \xi \end{aligned} \tag{C.1.15}$$

for sufficiently small $\delta > 0$. □

Notably, $\widetilde{\mathscr{N}_{m,n}}$ may not be continuous for finite n. However, if $\widetilde{\mathscr{N}_{m,n}} \to_{\mathscr{L}} \mathscr{M}_m$, lemma C.1.1 implies that the discontinuity in $\widetilde{\mathscr{N}_{m,n}}$ should vanish asymptotically. Combine lemma C.1.1 with (C.1.3), we derive the following corollary:

Corollary 3. *Suppose assumption 1 to 4 hold true. Then for any given* $0 < c < 1/4$,

$$\lim_{\delta \to 0} \sup_{|x-y|+|z-w|<\delta} |\mathbf{P}(\mathscr{M}_m(x) - \mathscr{M}_m(1-x) \le z) - \mathbf{P}(\mathscr{M}_m(y) - \mathscr{M}_m(1-y) \le w)| = 0 \quad \text{(C.1.16)}$$

and if $\widetilde{\mathscr{N}_{m,n}} \to_{\mathscr{L}} \mathscr{M}_m$, *then*

$$\lim_{n \to \infty} \sup_{x \in [\frac{1}{2}+c, 1-c], z \in \mathbf{R}} |\mathbf{P}(\widetilde{\mathscr{N}_{m,n}}(x) - \widetilde{\mathscr{N}_{m,n}^{-}}(1-x) \le z) - \mathbf{P}(\mathscr{M}_m(x) - \mathscr{M}_m(1-x) \le z)| = 0 \quad \text{(C.1.17)}$$

Here $(x,z),(y,w) \in [\frac{1}{2}+c, 1-c] \times \mathbf{R}$. *See* (4.16) *in the paper for the definition of the superscript* $^{-}$.

proof of corollary 3. Without loss of generality, assume $z \le w$. From (C.1.3), for $\forall\ \xi > 0$,

$$\begin{aligned} &|\mathbf{P}(\mathscr{M}_m(x) - \mathscr{M}_m(1-x) \le z) - \mathbf{P}(\mathscr{M}_m(y) - \mathscr{M}_m(1-y) \le w)| \\ \le\ &\mathbf{P}(|\mathscr{M}_m(x) - \mathscr{M}_m(y)| > \xi/2) + \mathbf{P}(|\mathscr{M}_m(1-x) - \mathscr{M}_m(1-y)| > \xi/2) \\ +&\mathbf{P}(z - \xi < \mathscr{M}_m(y) - \mathscr{M}_m(1-y) \le z + \xi) + \mathbf{P}(z < \mathscr{M}_m(y) - \mathscr{M}_m(1-y) \le w) \end{aligned} \quad \text{(C.1.18)}$$

Define $y_m = 2my - m$. From assumption 4, $\min_{y \in [\frac{1}{2}+c, 1-c]} \mathscr{U}(y_m) > 0$ so

$$\begin{aligned} \mathbf{P}(z < \mathscr{M}_m(y) - \mathscr{M}_m(1-y) \le w) &= \Phi\left(\frac{w}{\sqrt{\mathscr{U}(y_m)}}\right) - \Phi\left(\frac{z}{\sqrt{\mathscr{U}(y_m)}}\right) \\ &\le \frac{\delta}{\min_{y \in [\frac{1}{2}+c, 1-c]} \sqrt{\mathscr{U}(y_m)}} \end{aligned} \quad \text{(C.1.19)}$$

Similarly $\mathbf{P}(z - \xi < \mathscr{M}_m(y) - \mathscr{M}_m(1-y) \le z + \xi) \le \frac{2\xi}{\min_{y \in [\frac{1}{2}+c, 1-c]} \sqrt{\mathscr{U}(y_m)}}$. (C.1.16) is proved by

applying lemma C.1.1 to (C.1.18).

For $\forall x \in [\frac{1}{2}+c, 1-c]$, define $g_x : \mathbf{D} \to \mathbf{R} : g_x(f) = f(x) - f^-(1-x)$. We use the same notation as (C.1.14). If f_n converges to f in $\mathbf{D}$ and f is continuous,

$$\begin{aligned} |g_x(f_n) - g_x(f)| &\leq |f_n(x) - f(\lambda_n^{-1}(x))| + |f(\lambda_n^{-1}(x)) - f(x)| \\ + \limsup_{t \to 1-x, t<1-x} |f_n(t) - f(\lambda_n^{-1}(t))| &+ \limsup_{t \to 1-x, t<1-x} |f(\lambda_n^{-1}(t)) - f(t)| \end{aligned} \tag{C.1.20}$$

which tends to 0 as $n \to \infty$. Therefore, 3.8, page 348 in Jacod and Shiryaev [2003] implies $g_x(\widetilde{\mathscr{N}_{m,n}}) \to_{\mathscr{L}} g_x(\mathscr{M}_m)$. $\forall \psi > 0, t \in \mathbf{R}$, define $G_0(x) = (1 - \min(1, \max(x,0))^4)^4$, and $G_{\psi,t}(x) = G_0(\psi x - \psi t)$. From Xu et al. [2019], $\exists$ a constant $C > 0$ with

$$\begin{aligned} \mathbf{1}_{x \leq t} \leq G_{\psi,t}(x) \leq \mathbf{1}_{x \leq t+1/\psi}, \ \sup_{x,t} |G'_{\psi,t}(x)| \leq C\psi \\ \sup_{x,t} |G''_{\psi,t}(x)| \leq C\psi^2, \ \sup_{x,t} |G'''_{\psi,t}(x)| \leq C\psi^3 \end{aligned} \tag{C.1.21}$$

For $\forall \psi > 0$, define the set $\mathscr{A}_\psi = \{G_{\psi,t} : t \in \mathbf{R}\}$. $\forall \delta > 0$, choose $\gamma = \delta/(C\psi)$, then $\forall G_{\psi,t} \in \mathscr{A}_\psi, x, y \in \mathbf{R}$ with $|x-y| < \gamma$, $|G_{\psi,t}(x) - G_{\psi,t}(y)| \leq C\psi|x-y| < \delta \Rightarrow \mathscr{A}_\psi$ is equi-continuous and uniformly bounded by 1. From theorem 3.1 in Ranga Rao [1962],

$$\lim_{n \to \infty} \sup_{G_{\psi,t} \in \mathscr{A}_\psi} |\mathbf{E} G_{\psi,t}\left(\widetilde{\mathscr{N}_{m,n}}(x) - \widetilde{\mathscr{N}_{m,n}^-}(1-x)\right) - \mathbf{E} G_{\psi,t}\left(\mathscr{M}_m(x) - \mathscr{M}_m(1-x)\right)| = 0 \tag{C.1.22}$$

for any fixed $x \in [\frac{1}{2}+c, 1-c]$. From (C.1.21),

$$\begin{aligned}
&\mathbf{P}(\widetilde{\mathscr{N}_{m,n}}(x) - \widetilde{\mathscr{N}_{m,n}^{-}}(1-x) \le z) - \mathbf{P}(\mathscr{M}_m(x) - \mathscr{M}_m(1-x) \le z) \\
&\le \mathbf{E}G_{\psi,z}\left(\widetilde{\mathscr{N}_{m,n}}(x) - \widetilde{\mathscr{N}_{m,n}^{-}}(1-x)\right) - \mathbf{E}G_{\psi,z-1/\psi}(\mathscr{M}_m(x) - \mathscr{M}_m(1-x)) \\
&\le \sup_{G_{\psi,t}\in\mathscr{A}_\psi} |\mathbf{E}G_{\psi,t}\left(\widetilde{\mathscr{N}_{m,n}}(x) - \widetilde{\mathscr{N}_{m,n}^{-}}(1-x)\right) - \mathbf{E}G_{\psi,t}\left(\mathscr{M}_m(x) - \mathscr{M}_m(1-x)\right)| \\
&\qquad + \mathbf{P}\left(z - \frac{1}{\psi} < \mathscr{M}_m(x) - \mathscr{M}_m(1-x) \le z + \frac{1}{\psi}\right) \\
&\mathbf{P}(\widetilde{\mathscr{N}_{m,n}}(x) - \widetilde{\mathscr{N}_{m,n}^{-}}(1-x) \le z) - \mathbf{P}(\mathscr{M}_m(x) - \mathscr{M}_m(1-x) \le z) \\
&\ge \mathbf{E}G_{\psi,z-1/\psi}\left(\widetilde{\mathscr{N}_{m,n}}(x) - \widetilde{\mathscr{N}_{m,n}^{-}}(1-x)\right) - \mathbf{E}G_{\psi,z}\left(\mathscr{M}_m(x) - \mathscr{M}_m(1-x)\right) \\
&\ge -\sup_{G_{\psi,t}\in\mathscr{A}_\psi} |\mathbf{E}G_{\psi,t}\left(\widetilde{\mathscr{N}_{m,n}}(x) - \widetilde{\mathscr{N}_{m,n}^{-}}(1-x)\right) - \mathbf{E}G_{\psi,t}\left(\mathscr{M}_m(x) - \mathscr{M}_m(1-x)\right)| \\
&\qquad - \mathbf{P}\left(z - \frac{1}{\psi} < \mathscr{M}_m(x) - \mathscr{M}_m(1-x) \le z + \frac{1}{\psi}\right)
\end{aligned} \tag{C.1.23}$$

Choose $y = x, z = z + \frac{1}{\psi}, w = z - \frac{1}{\psi}$ in (C.1.16) and let $\psi \to \infty$,

$$\lim_{n\to\infty} \sup_{z\in\mathbf{R}} |\mathbf{P}(\widetilde{\mathscr{N}_{m,n}}(x) - \widetilde{\mathscr{N}_{m,n}^{-}}(1-x) \le z) - \mathbf{P}(\mathscr{M}_m(x) - \mathscr{M}_m(1-x) \le z)| = 0 \tag{C.1.24}$$

Finally, for any given $\xi > 0$, we choose $\frac{1}{2}+c = x_0 < x_1 < ... < x_M = 1-c$ and $x_i - x_{i-1} < \delta, i = 1,2,...,M$ with sufficiently small $\delta > 0$. For $\forall x \in [\frac{1}{2}+c, 1-c]$, $\exists\, I \in \{0,1,...,M\}$ such that $|x - x_I| < \delta$, and

$$\begin{aligned}
&\sup_{z\in\mathbf{R}} |\mathbf{P}(\widetilde{\mathscr{N}_{m,n}}(x) - \widetilde{\mathscr{N}_{m,n}^{-}}(1-x) \le z) - \mathbf{P}(\mathscr{M}_m(x) - \mathscr{M}_m(1-x) \le z)| \\
&\le \sup_{z\in\mathbf{R}} |\mathbf{P}(\widetilde{\mathscr{N}_{m,n}}(x) - \widetilde{\mathscr{N}_{m,n}^{-}}(1-x) \le z) - \mathbf{P}(\widetilde{\mathscr{N}_{m,n}}(x_I) - \widetilde{\mathscr{N}_{m,n}^{-}}(1-x_I) \le z)| \\
&+ \max_{I=1,2,...,M} \sup_{z\in\mathbf{R}} |\mathbf{P}(\widetilde{\mathscr{N}_{m,n}}(x_I) - \widetilde{\mathscr{N}_{m,n}^{-}}(1-x_I) \le z) - \mathbf{P}(\mathscr{M}_m(x_I) - \mathscr{M}_m(1-x_I) \le z)| \\
&\qquad + \sup_{z\in\mathbf{R}} |\mathbf{P}(\mathscr{M}_m(x_I) - \mathscr{M}_m(1-x_I) \le z) - \mathbf{P}(\mathscr{M}_m(x) - \mathscr{M}_m(1-x) \le z)|
\end{aligned} \tag{C.1.25}$$

From (C.1.3), $\forall\ \xi > 0$,

$$\begin{aligned}
&\sup_{z\in\mathbf{R}}|\mathbf{P}(\widetilde{\mathscr{N}_{m,n}}(x)-\widetilde{\mathscr{N}_{m,n}^{-}}(1-x)\le z)-\mathbf{P}(\widetilde{\mathscr{N}_{m,n}}(x_I)-\widetilde{\mathscr{N}_{m,n}^{-}}(1-x_I)\le z)| \\
\le\ &\mathbf{P}\left(|\widetilde{\mathscr{N}_{m,n}^{-}}(1-x)-\widetilde{\mathscr{N}_{m,n}^{-}}(1-x_I)|>\frac{\xi}{2}\right)+\mathbf{P}\left(|\widetilde{\mathscr{N}_{m,n}}(x)-\widetilde{\mathscr{N}_{m,n}}(x_I)|>\frac{\xi}{2}\right) \\
+2\max_{I=1,2,\ldots,M}&\sup_{z\in\mathbf{R}}|\mathbf{P}(\widetilde{\mathscr{N}_{m,n}}(x_I)-\widetilde{\mathscr{N}_{m,n}^{-}}(1-x_I)\le z)-\mathbf{P}(\mathscr{M}_m(x_I)-\mathscr{M}_m(1-x_I)\le z)| \\
&+\sup_{z\in\mathbf{R}}\mathbf{P}(z-\xi<\mathscr{M}_m(x_I)-\mathscr{M}_m(1-x_I)\le z+\xi)
\end{aligned} \tag{C.1.26}$$

Since $\sup_{x\in[\frac{1}{2}+c,1-c]}\mathbf{P}\left(|\widetilde{\mathscr{N}_{m,n}}(x)-\widetilde{\mathscr{N}_{m,n}}(x_I)|>\frac{\xi}{2}\right)$ and
$\sup_{x\in[\frac{1}{2}+c,1-c]}\mathbf{P}\left(|\widetilde{\mathscr{N}_{m,n}^{-}}(1-x)-\widetilde{\mathscr{N}_{m,n}^{-}}(1-x_I)|>\frac{\xi}{2}\right)$ are less or equal to
$\mathbf{P}\left(\sup_{y,z\in[0,1],|y-z|<\delta}|\widetilde{\mathscr{N}_{m,n}}(y)-\widetilde{\mathscr{N}_{m,n}}(z)|>\frac{\xi}{2}\right)$, (C.1.7), (C.1.16) and (C.1.24) imply (C.1.17).

□

The second lemma focuses on showing the asymptotic continuity of the residuals' empirical process in the real world and in the bootstrap world. Define the stochastic processes

$$\widehat{\alpha}(x)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathbf{1}_{\widehat{\varepsilon}_i\le x}-F(x)) \text{ and } \widetilde{\alpha}^*(x)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathbf{1}_{\varepsilon_i^*\le x}-\widehat{F}(x)) \tag{C.1.27}$$

Here $\widehat{\varepsilon}_i$ and $\widehat{F}$ are defined in (4.13). $\varepsilon_i^*, i=1,2,...,n$ are i.i.d. random variables generated from $\widehat{F}$. In algorithm 4, ε_i^* serves as the bootstrapped residuals. Define two assistant processes

$$\widetilde{F}(x)=\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\varepsilon_i\le x} \text{ and } \widetilde{\alpha}(x)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathbf{1}_{\varepsilon_i\le x}-F(x)), \text{ here } \forall x\in\mathbf{R} \tag{C.1.28}$$

The notation O_p and o_p have the same meaning as definition 1.9 in Shao [2003], i.e., two random variable sequences $X_n, Y_n, n=1,2,...$ satisfy $X_n=O_p(Y_n)$ if for $\forall t>0$, $\exists$ a constant C_t such that $Prob(|X_n|\ge C_t|Y_n|)\le t$ for $n=1,2,....$ $X_n=o_p(Y_n)$ if $X_n/Y_n\to_p 0$ as $n\to\infty$.

Lemma C.1.2. *Suppose assumption 1 to 4 hold true. Then for any given $\xi>0$ and $-\infty<r\le$*

$s < \infty$, $\exists\ \delta > 0$ *such that*

$$\limsup_{n\to\infty} \mathbf{P}\left(\sup_{x,y\in[r,s],|x-y|<\delta} |\widehat{\alpha}(x) - \widehat{\alpha}(y)| > \xi \right) < \xi \tag{C.1.29}$$

Besides, $\exists \delta > 0$ *and* $N > 0$ *such that* $\forall n \geq N$,

$$\mathbf{P}\left(\left\{ \mathbf{P}^* \left(\sup_{x,y\in[r,s],|x-y|<\delta} |\widetilde{\alpha}^*(x) - \widetilde{\alpha}^*(y)| > \xi \right) > \xi \right\} \right) < \xi \tag{C.1.30}$$

proof of lemma C.1.2. From assumption 4, F is strictly increasing in $\mathbf{R}$. From lemma 4.1 and 4.2, Bickel and Freedman [1981], $\exists$ independent random variables $U_i, i = 1,2,...$ with uniform distribution on $[0,1]$, a Brownian bridge B and a constant C such that

$$\mathbf{P}\left(\sup_{x\in[0,1]} |\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathbf{1}_{U_i \leq x} - x) - B(x)| \geq C\log(n)/\sqrt{n} \right) \leq C\log(n)/\sqrt{n} \tag{C.1.31}$$

and $\forall 0 < \delta < 1/2$, $\xi > 0$,

$$\begin{aligned} \mathbf{E} \sup_{x,y\in[0,1],|x-y|<\delta} |B(x) - B(y)| &\leq C(-\delta\log(\delta))^{1/2} \\ \Rightarrow \mathbf{P}\left(\sup_{x,y\in[0,1],|x-y|<\delta} |B(x) - B(y)| > \xi \right) &\leq \frac{C(-\delta\log(\delta))^{1/2}}{\xi} \end{aligned} \tag{C.1.32}$$

We choose $\varepsilon_i = F^{-1}(U_i), i = 1,2,...,n$ (ε_i has distribution F according to page 150, Billingsley [1999]),

$$\begin{aligned} \widetilde{\alpha}(x) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathbf{1}_{U_i \leq F(x)} - F(x)) \\ \Rightarrow \mathbf{P}\left(\sup_{x\in\mathbf{R}} |\widetilde{\alpha}(x) - B(F(x))| \geq C\log(n)/\sqrt{n} \right) &\leq C\log(n)/\sqrt{n} \end{aligned} \tag{C.1.33}$$

From assumption 3,

$$\max_{i=1,\ldots,n} \frac{1}{n} \mathscr{X}_i^T (\frac{\mathscr{X}^T \mathscr{X}}{n})^{-1} \mathscr{X}_i \leq \frac{1}{n} M^2 \|(\frac{\mathscr{X}^T \mathscr{X}}{n})^{-1}\|_2 = O(1/n) \tag{C.1.34}$$

and $\mathbf{E}\|(\mathscr{X}^T \mathscr{X})^{1/2}(\widehat{\beta} - \beta)\|_2^2 = p\sigma^2$ implies $\|(\mathscr{X}^T \mathscr{X})^{1/2}(\widehat{\beta} - \beta)\|_2 = O_p(1)$. Define

$$\widehat{\lambda} = \frac{1}{n} \sum_{j=1}^{n} \widetilde{\varepsilon}_j = \frac{1}{n} \sum_{j=1}^{n} \varepsilon_j - \overline{\mathscr{X}}_n^T (\widehat{\beta} - \beta) \tag{C.1.35}$$

here $\widetilde{\varepsilon}_i$ and $\overline{\mathscr{X}}_n$ are defined in (4.13). With this definition we have $\widehat{\varepsilon}_i = \widetilde{\varepsilon}_i - \widehat{\lambda}$. Besides,

$$\begin{aligned} \mathbf{E}\widehat{\lambda}^2 \leq 2\mathbf{E} \left(\frac{1}{n} \sum_{j=1}^{n} \varepsilon_j \right)^2 + 2\mathbf{E} \left(\overline{\mathscr{X}}_n^T (\widehat{\beta} - \beta) \right)^2 = \frac{2\sigma^2}{n} + \frac{2\sigma^2 \overline{\mathscr{X}}_n (\mathscr{X}^T \mathscr{X}/n)^{-1} \overline{\mathscr{X}}_n}{n} \\ \Rightarrow \widehat{\lambda} = O_p(1/\sqrt{n}) \end{aligned} \tag{C.1.36}$$

Define

$$\widetilde{\alpha}^{\dagger}(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\mathbf{1}_{\widetilde{\varepsilon}_i \leq x} - F(x) \right) \Rightarrow \widehat{\alpha}(x) = \widetilde{\alpha}^{\dagger}(x + \widehat{\lambda}) + \sqrt{n}(F(x + \widehat{\lambda}) - F(x)) \tag{C.1.37}$$

From theorem 6.2.1 in Koul [2002],

$$\sup_{x \in \mathbf{R}} |\widetilde{\alpha}^{\dagger}(x) - \widetilde{\alpha}(x) - \sqrt{n} F'(x) \overline{\mathscr{X}}_n^T (\widehat{\beta} - \beta)| = o_p(1) \tag{C.1.38}$$

Therefore,

$$\begin{aligned} &\sup_{x \in \mathbf{R}} |\widehat{\alpha}(x) - \widetilde{\alpha}(x) - \sqrt{n} F'(x) \overline{\mathscr{X}}_n^T (\widehat{\beta} - \beta) - \sqrt{n} F'(x) \widehat{\lambda}| \\ \leq &\sup_{x \in \mathbf{R}} |\widetilde{\alpha}^{\dagger}(x) - \widetilde{\alpha}(x) - \sqrt{n} F'(x) \overline{\mathscr{X}}_n^T (\widehat{\beta} - \beta)| + \sup_{x \in \mathbf{R}} |\widetilde{\alpha}(x + \widehat{\lambda}) - \widetilde{\alpha}(x)| \\ + &\sup_{x \in \mathbf{R}} \sqrt{n} |(F'(x + \widehat{\lambda}) - F'(x)) \overline{\mathscr{X}}_n^T (\widehat{\beta} - \beta)| + \sup_{x \in \mathbf{R}} \sqrt{n} |F(x + \widehat{\lambda}) - F(x) - F'(x) \widehat{\lambda}| \end{aligned} \tag{C.1.39}$$

From assumption 1 and 3 and Taylor's theorem, $\sup_{x\in\mathbf{R}}\sqrt{n}|(F^{'}(x+\widehat{\lambda})-F^{'}(x))\overline{\mathscr{X}}_n^T(\widehat{\beta}-\beta)|$ and $\sup_{x\in\mathbf{R}}\sqrt{n}|F(x+\widehat{\lambda})-F(x)-F^{'}(x)\widehat{\lambda}|$ have order $O_p(1/\sqrt{n})$. From (C.1.33), with probability tending to 1,

$$\sup_{x\in\mathbf{R}}|\widetilde{\alpha}(x+\widehat{\lambda})-\widetilde{\alpha}(x)|\leq\frac{2C\log(n)}{\sqrt{n}}+\sup_{x\in\mathbf{R}}|B(F(x+\widehat{\lambda}))-B(F(x))| \tag{C.1.40}$$

F is uniform continuous according to assumption 1, so

$$\begin{aligned}&\sup_{x\in\mathbf{R}}|\widetilde{\alpha}(x+\widehat{\lambda})-\widetilde{\alpha}(x)|=o_p(1)\\ \Rightarrow&\sup_{x\in\mathbf{R}}|\widehat{\alpha}(x)-\widetilde{\alpha}(x)-\sqrt{n}F^{'}(x)\overline{\mathscr{X}}_n^T(\widehat{\beta}-\beta)-\sqrt{n}F^{'}(x)\widehat{\lambda}|=o_p(1)\end{aligned} \tag{C.1.41}$$

For any given $-\infty<r\leq s<\infty$ and sufficiently small $\delta>0$,

$$\begin{aligned}&\sup_{x,y\in[r,s],|x-y|<\delta}|\widehat{\alpha}(x)-\widehat{\alpha}(y)|\leq\sup_{x,y\in[r,s],|x-y|<\delta}|\widetilde{\alpha}(x)-\widetilde{\alpha}(y)|\\ +&\sup_{x,y\in[r,s],|x-y|<\delta}\sqrt{n}|(F^{'}(x)-F^{'}(y))\times(\overline{\mathscr{X}}_n^T(\widehat{\beta}-\beta)+\widehat{\lambda})|+o_p(1)\end{aligned} \tag{C.1.42}$$

From assumption 1, (C.1.33) and (C.1.32), we prove (C.1.29).

Define the function $\widehat{\phi}(x)=\inf\{t|x\leq\widehat{F}(t)\},x\in[0,1]$. Page 150, Billingsley [1999] implies $\widehat{\phi}(x)\leq t\Leftrightarrow x\leq\widehat{F}(t)$. If U has uniform distribution on $[0,1]$, then $\widehat{\phi}(U)$ has distribution $\widehat{F}$. Without loss of generality, we choose $\varepsilon_i^*=\widehat{\phi}(U_i),i=1,2,...,n$; (C.1.31) implies

$$\begin{aligned}&\widetilde{\alpha}^*(x)=\frac{1}{\sqrt{n}}\sum_{i=1}^n(\mathbf{1}_{U_i\leq\widehat{F}(x)}-\widehat{F}(x))\\ \Rightarrow&\mathbf{P}^*\left(\sup_{x\in\mathbf{R}}|\widetilde{\alpha}^*(x)-B(\widehat{F}(x))|\geq C\log(n)/\sqrt{n}\right)\leq C\log(n)/\sqrt{n}\end{aligned} \tag{C.1.43}$$

From assumption 3

$$\begin{aligned}\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\varepsilon_i \le x+\mathscr{X}_i^T(\widehat{\beta}-\beta)+\widehat{\lambda}} &\le \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\varepsilon_i \le x+M\|\widehat{\beta}-\beta\|_2+|\widehat{\lambda}|} \\ &= \widetilde{F}(x+M\|\widehat{\beta}-\beta\|_2+|\widehat{\lambda}|)\end{aligned} \tag{C.1.44}$$

Similarly, $\widehat{F}(x) \ge \widetilde{F}(x-M\|\widehat{\beta}-\beta\|_2-|\widehat{\lambda}|)$. For any given $\omega > 0$, we can find $C_\omega > 0$ with $\mathbf{P}(\|\widehat{\beta}-\beta\|_2 > \frac{C_\omega}{2\sqrt{n}}) < \omega$ and $\mathbf{P}(|\widehat{\lambda}| > \frac{MC_\omega}{2\sqrt{n}}) < \omega$ for any n. From Glivenko - Cantelli theorem and dominated convergence theorem, $\lim_{n\to\infty}\mathbf{P}(\sup_{x\in\mathbf{R}}|\widetilde{F}(x)-F(x)| > \omega) = 0$. If $\|\widehat{\beta}-\beta\|_2 \le \frac{C_\omega}{2\sqrt{n}}$, $|\widehat{\lambda}| \le \frac{MC_\omega}{2\sqrt{n}}$ and $\sup_{x\in\mathbf{R}}|\widetilde{F}(x)-F(x)| \le \omega$, then for any given $-\infty < r \le s < \infty$, $\delta > 0$, $-\omega + F(x-\frac{MC_\omega}{\sqrt{n}}) \le \widehat{F}(x) \le \omega + F(x+\frac{MC_\omega}{\sqrt{n}})$, and

$$\sup_{r\le x\le y\le s, y-x<\delta}\widehat{F}(y)-\widehat{F}(x) \le 2\omega + \sup_{r\le x\le y\le s, y-x<\delta}F(y+\frac{MC_\omega}{\sqrt{n}})-F(x-\frac{MC_\omega}{\sqrt{n}}) \tag{C.1.45}$$

For any given $-\infty < r \le s < \infty$ and $\xi > 0$, we choose sufficiently small $\omega, \delta > 0$ and define $\zeta = 2\omega + \sup_{r\le x\le y\le s, y-x<\delta}F(y+\frac{MC_\omega}{\sqrt{n}})-F(x-\frac{MC_\omega}{\sqrt{n}})$,

$$\begin{aligned}&\sup_{x,y\in[r,s],|x-y|<\delta}|\widetilde{\alpha}^*(x)-\widetilde{\alpha}^*(y)| \le 2\sup_{x\in\mathbf{R}}|\widetilde{\alpha}^*(x)-B(\widehat{F}(x))| \\ &\qquad + \sup_{x,y\in[r,s],|x-y|<\delta}|B(\widehat{F}(x))-B(\widehat{F}(y))| \\ \le\ & 2\sup_{x\in\mathbf{R}}|\widetilde{\alpha}^*(x)-B(\widehat{F}(x))| + \sup_{x,y\in[-\omega+F(r-\frac{MC_\omega}{\sqrt{n}}),\omega+F(s+\frac{MC_\omega}{\sqrt{n}})],|x-y|\le\zeta}|B(x)-B(y)| \\ &\qquad \Rightarrow \mathbf{P}^*\left(\sup_{x,y\in[r,s],|x-y|<\delta}|\widetilde{\alpha}^*(x)-\widetilde{\alpha}^*(y)| > \xi\right) \\ \le\ & \mathbf{P}^*\left(\sup_{x\in\mathbf{R}}|\widetilde{\alpha}^*(x)-B(\widehat{F}(x))| > \frac{\xi}{4}\right) + \mathbf{P}^*\left(\sup_{x,y\in[0,1],|x-y|\le\zeta}|B(x)-B(y)| > \frac{\xi}{2}\right)\end{aligned} \tag{C.1.46}$$

For F is uniform continuous, (C.1.32) and (C.1.43) imply (C.1.30). □

C.1.2 Proof of theorem 9

The existence of $\mathscr{M}_m$ has been shown in lemma C.1.1, and this section will complete the remaining steps.

proof of theorem 9. **Prove $\widetilde{M}_m \to_{\mathscr{L}} \mathscr{M}_m$. Here $\widetilde{M}_m$ is defined in** (C.1.1)**.**

According to theorem 13.5 in Billingsley [1999], it suffices to verify the following conditions:

1. $\forall z_1,...,z_k \in [0,1]$, $(\widetilde{M}_m(z_1),...,\widetilde{M}_m(z_k)) \to_{\mathscr{L}} (\mathscr{M}_m(z_1),...,\mathscr{M}_m(z_k))$ in $\mathbf{R}^k$. According to Cramér-Wold device(theorem 1.9 in Shao [2003]), this condition can be proved by showing

$$\sum_{j=1}^{k} s_j \widetilde{M}_m(z_j) \to_{\mathscr{L}} \sum_{j=1}^{k} s_j \mathscr{M}_m(z_j) \tag{C.1.47}$$

here $s_1,...,s_k \in \mathbf{R}$ are any given real numbers.

2. $\mathscr{M}_m(1) - \mathscr{M}_m(1-\delta) \to_{\mathscr{L}} 0$ in $\mathbf{R}$ as $\delta \to 0, \delta > 0$

3. $\exists b \geq 0, a > 1/2$, and a non-decreasing, continuous function G on $[0,1]$ such that

$$\mathbf{E}|\widetilde{M}_m(t) - \widetilde{M}_m(s)|^{2b} |\widetilde{M}_m(s) - \widetilde{M}_m(r)|^{2b} \leq (G(t) - G(r))^{2a} \text{ for } \forall 1 \geq t > s > r \geq 0 \tag{C.1.48}$$

For the first condition: define
$c^T = (c_1,...,c_n) = \mathscr{X}_f^T (\mathscr{X}^T \mathscr{X})^{-1} \mathscr{X}^T - \frac{1}{n} e^T \Rightarrow c_i = \mathscr{X}_f^T (\mathscr{X}^T \mathscr{X})^{-1} \mathscr{X}_i - \frac{1}{n}$. Here
$e = (1,1,...,1)^T$. Define $z_{j,m} = 2mz_j - m, j = 1,2,...,k$. For any given $s_1,...,s_k \in \mathbf{R}$

$$\sum_{j=1}^{k} s_j \widetilde{M}_m(z_j) = \sum_{i=1}^{n} \left((\sum_{j=1}^{k} \sqrt{n} s_j F^{'}(z_{j,m})) c_i \varepsilon_i - \frac{1}{\sqrt{n}} (\sum_{j=1}^{k} s_j (\mathbf{1}_{\varepsilon_i \leq z_{j,m}} - F(z_{j,m}))) \right)$$
$$\Rightarrow \mathbf{E} \sum_{j=1}^{k} s_j \widetilde{M}_m(z_j) = 0 \tag{C.1.49}$$

Form assumption 3 and (5.8.4) in Horn and Johnson [2013], we define

$Y_i = (\sum_{j=1}^{k} \sqrt{n} s_j F^{'}(z_{j,m})) c_i \varepsilon_i - \frac{1}{\sqrt{n}} (\sum_{j=1}^{k} s_j (\mathbf{1}_{\varepsilon_i \leq z_{j,m}} - F(z_{j,m})))$,

$$
\begin{aligned}
\mathbf{E}Y_i^2 &= n\sigma^2 c_i^2 \sum_{j=1}^{k}\sum_{l=1}^{k} s_j s_l F^{'}(z_{j,m}) F^{'}(z_{l,m}) \\
&+ \frac{1}{n}\sum_{j=1}^{k}\sum_{l=1}^{k} s_j s_l (F(\min(z_{j,m}, z_{l,m})) - F(z_{j,m})F(z_{l,m})) \\
&- 2c_i \sum_{j=1}^{k}\sum_{l=1}^{k} s_j s_l F^{'}(z_{j,m}) H(z_{l,m}) \\
&\Rightarrow \lim_{n\to\infty} \sum_{i=1}^{n} \mathbf{E}Y_i^2 \\
&= \lim_{n\to\infty} \sigma^2 \sum_{j=1}^{k}\sum_{l=1}^{k} s_j s_l F^{'}(z_{j,m}) F^{'}(z_{l,m}) \\
&\times \left(\mathscr{X}_f^T (\frac{\mathscr{X}^T \mathscr{X}}{n})^{-1} \mathscr{X}_f + 1 - 2\mathscr{X}_f^T (\frac{\mathscr{X}^T \mathscr{X}}{n})^{-1} \overline{\mathscr{X}}_n \right) \\
&+ \sum_{j=1}^{k}\sum_{l=1}^{k} s_j s_l (F(\min(z_{j,m}, z_{l,m})) - F(z_{j,m})F(z_{l,m})) \\
&- \lim_{n\to\infty} 2\sum_{j=1}^{k}\sum_{l=1}^{k} s_j s_l F^{'}(z_{j,m}) H(z_{l,m}) \times (\mathscr{X}_f^T (\frac{\mathscr{X}^T \mathscr{X}}{n})^{-1} \overline{\mathscr{X}}_n - 1) \\
&= \sigma^2 K \times (\mathscr{X}_f^T A^{-1} \mathscr{X}_f + 1 - 2\mathscr{X}_f^T A^{-1} b) + N - 2R \times (\mathscr{X}_f^T A^{-1} b - 1)
\end{aligned}
\tag{C.1.50}
$$

here we define

$$
\begin{aligned}
K = \sum_{j=1}^{k}\sum_{l=1}^{k} s_j s_l F^{'}(z_{j,m}) F^{'}(z_{l,m}),\ N = \sum_{l=1}^{k} s_j s_l (F(\min(z_{j,m}, z_{l,m})) - F(z_{j,m})F(z_{l,m})) \\
\text{and } R = \sum_{j=1}^{k}\sum_{l=1}^{k} s_j s_l F^{'}(z_{j,m}) H(z_{l,m})
\end{aligned}
\tag{C.1.51}
$$

From mean value inequality,

$$\begin{aligned}\sum_{i=1}^{n}\mathbf{E}|Y_i|^3 \leq 4k^2\mathbf{E}|\varepsilon_1|^3\sum_{j=1}^{k}|s_jF^{'}(z_{j,m})|^3\times n\sqrt{n}\sum_{i=1}^{n}|c_i|^3 \\ +4k^2\sum_{j=1}^{k}|s_j|^3\times\frac{1}{n\sqrt{n}}\sum_{i=1}^{n}\mathbf{E}|\mathbf{1}_{\varepsilon_i\leq z_{j,m}}-F(z_{j,m})|^3\end{aligned} \tag{C.1.52}$$

From assumption 3,

$$\begin{aligned}n\sqrt{n}\sum_{i=1}^{n}|c_i|^3 \leq n\sqrt{n}\max_{i=1,2,\ldots,n}|c_i|\times\sum_{i=1}^{n}c_i^2 \\ \leq\frac{1+M^2\|(\mathscr{X}^T\mathscr{X}/n)^{-1}\|_2}{\sqrt{n}}\times\left(\mathscr{X}_f^T(\frac{\mathscr{X}^T\mathscr{X}}{n})^{-1}\mathscr{X}_f+1-2\mathscr{X}_f^T(\frac{\mathscr{X}^T\mathscr{X}}{n})^{-1}\overline{\mathscr{X}}_n\right)\end{aligned} \tag{C.1.53}$$

which has order $O(1/\sqrt{n})$. $\|(\mathscr{X}^T\mathscr{X}/n)^{-1}\|_2$ is the matrix 2 norm of the matrix $(\mathscr{X}^T\mathscr{X}/n)^{-1}$. If $\sigma^2K\times(\mathscr{X}_f^TA^{-1}\mathscr{X}_f+1-2\mathscr{X}_f^TA^{-1}b)+N-2R\times(\mathscr{X}_f^TA^{-1}b-1)\neq 0$, from Theorem 1.15, Theorem 1.11, and (1.97) in Shao [2003],

$$\begin{aligned}\sum_{j=1}^{k}s_j\widetilde{M}_m(z_j)=\frac{\sum_{j=1}^{k}s_j\widetilde{M}_m(z_j)}{\sqrt{\sum_{i=1}^{n}\mathbf{E}Y_i^2}}\times\sqrt{\sum_{i=1}^{n}\mathbf{E}Y_i^2} \\ \rightarrow_{\mathscr{L}}N(0,\sigma^2K\times(\mathscr{X}_f^TA^{-1}\mathscr{X}_f+1-2\mathscr{X}_f^TA^{-1}b)+N-2R\times(\mathscr{X}_f^TA^{-1}b-1))\end{aligned} \tag{C.1.54}$$

On the other hand, if $\sigma^2K\times(\mathscr{X}_f^TA^{-1}\mathscr{X}_f+1-2\mathscr{X}_f^TA^{-1}b)+N-2R\times(\mathscr{X}_f^TA^{-1}b-1)=0$, then $\forall\delta>0$, from (C.1.50), $\lim_{n\rightarrow\infty}\mathbf{P}(|\sum_{j=1}^{k}s_j\widetilde{M}_m(z_j)|\geq\delta)\leq\lim_{n\rightarrow\infty}\frac{\mathbf{E}|\sum_{j=1}^{k}s_j\widetilde{M}_m(z_j)|^2}{\delta^2}=0\Rightarrow$ $\sum_{j=1}^{k}s_j\widetilde{M}_m(z_j)\rightarrow_{\mathscr{L}}0$. From theorem 1.9, Shao [2003], we prove (C.1.47) and the first condition.

The second condition: $\forall \xi > 0$,

$$\begin{aligned}
\mathbf{P}\left(|\mathscr{M}_m(1)-\mathscr{M}_m(1-\delta)|\geq\xi\right) &\leq \frac{\mathbf{E}|\mathscr{M}_m(1)-\mathscr{M}_m(1-\delta)|^2}{\xi^2} \\
&\leq \frac{\sigma^2(\mathscr{X}_f^T A^{-1}\mathscr{X}_f - 2\mathscr{X}_f^T A^{-1}b+1)(F^{'}(m)-F^{'}(m-2m\delta))^2}{\xi^2} \\
+\frac{2|\mathscr{X}_f^T A^{-1}b-1|\times|F^{'}(m)-F^{'}(m-2m\delta)|\times|H(m)-H(m-2m\delta)|}{\xi^2} \\
&+\frac{|F(m)-F(m-2m\delta)|+|F(m)-F(m-2m\delta)|^2}{\xi^2}
\end{aligned} \tag{C.1.55}$$

From assumption 1, $\lim_{\delta\to 0,\delta>0} Prob\left(|\mathscr{M}_m(1)-\mathscr{M}_m(1-\delta)|\geq\xi\right)=0$, and we prove the second condition.

The third condition: we choose $b=a=1$ in (C.1.48) and define $t_m = 2mt-m, \forall t$. For $\forall t,s\in[0,1]$, we define

$$\begin{aligned}
\mathscr{A}(t,s) &= \sqrt{n}\left(F^{'}(t_m)-F^{'}(s_m)\right)\times\left(\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon-\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j\right) \\
\text{and } \mathscr{B}(t,s) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathbf{1}_{s_m<\varepsilon_i\leq t_m}-F(t_m)+F(s_m))
\end{aligned} \tag{C.1.56}$$

From mean value inequality,

$$\begin{aligned}
&\mathbf{E}|\widetilde{M}_m(t)-\widetilde{M}_m(s)|^2|\widetilde{M}_m(s)-\widetilde{M}_m(r)|^2 \\
&\leq 4\mathbf{E}\left(\mathscr{A}(t,s)^2\mathscr{A}(s,r)^2+\mathscr{B}(t,s)^2\mathscr{A}(s,r)^2+\mathscr{A}(t,s)^2\mathscr{B}(s,r)^2+\mathscr{B}(t,s)^2\mathscr{B}(s,r)^2\right)
\end{aligned} \tag{C.1.57}$$

From assumption 3, $\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X}/n)^{-1}\mathscr{X}_f \to \mathscr{X}_f^T A^{-1}\mathscr{X}_f$ and $\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X}/n)^{-1}\overline{\mathscr{X}}_n \to \mathscr{X}_f^T A^{-1}b$. Therefore, $\exists C>0$ such that $|\mathscr{X}_f^T(\frac{\mathscr{X}^T\mathscr{X}}{n})^{-1}\mathscr{X}_f+1-2\mathscr{X}_f^T(\frac{\mathscr{X}^T\mathscr{X}}{n})^{-1}\overline{\mathscr{X}}_n|\leq C$ for $\forall n$. Define $c=(c_1,...,c_n)^T$, $c_i=\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}_i-1/n, \forall 1\geq$

$t > s > r \geq 0$

$$\begin{aligned}
\mathbf{E}\mathscr{A}(t,s)^2\mathscr{A}(s,r)^2 &= n^2\left(F^{'}(t_m)-F^{'}(s_m)\right)^2\left(F^{'}(s_m)-F^{'}(r_m)\right)^2\mathbf{E}(c^T\varepsilon)^4 \\
&= n^2\left(F^{'}(t_m)-F^{'}(s_m)\right)^2\left(F^{'}(s_m)-F^{'}(r_m)\right)^2 \\
&\quad\times\left(\mathbf{E}\varepsilon_1^4\times\sum_{i=1}^n c_i^4+3\sigma^4\sum_{i=1}^n\sum_{j=1,j\neq i}^n c_i^2c_j^2\right) \\
&\leq 16m^4(\mathbf{E}\varepsilon_1^4+3\sigma^4)(t-s)^2(s-r)^2\times\sup_{x\in\mathbf{R}}|F^{''}(x)|^4 \\
&\quad\times\left(\mathscr{X}_f^T(\frac{\mathscr{X}^T\mathscr{X}}{n})^{-1}\mathscr{X}_f+1-2\mathscr{X}_f^T(\frac{\mathscr{X}^T\mathscr{X}}{n})^{-1}\overline{\mathscr{X}}_n\right)^2 \\
&\leq 16C^2m^4(\mathbf{E}\varepsilon_1^4+3\sigma^4)(t-r)^2\times\sup_{x\in\mathbf{R}}|F^{''}(x)|^4
\end{aligned} \tag{C.1.58}$$

$$\begin{aligned}
&\mathbf{E}\mathscr{B}(t,s)^2\mathscr{B}(s,r)^2 \\
&\leq\frac{1}{n}\mathbf{E}(\mathbf{1}_{s_m<\varepsilon_1\leq t_m}-F(t_m)+F(s_m))^2(\mathbf{1}_{r_m<\varepsilon_1\leq s_m}-F(s_m)+F(r_m))^2 \\
&+\mathbf{E}(\mathbf{1}_{s_m<\varepsilon_1\leq t_m}-F(t_m)+F(s_m))^2\times\mathbf{E}(\mathbf{1}_{r_m<\varepsilon_1\leq s_m}-F(s_m)+F(r_m))^2 \\
&+2\left(\mathbf{E}(\mathbf{1}_{s_m<\varepsilon_1\leq t_m}-F(t_m)+F(s_m))(\mathbf{1}_{r_m<\varepsilon_1\leq s_m}-F(s_m)+F(r_m))\right)^2 \\
&\leq\frac{3}{n}(F(t_m)-F(s_m))(F(s_m)-F(r_m))+(F(t_m)-F(s_m))(F(s_m)-F(r_m)) \\
&+2(F(t_m)-F(s_m))^2(F(s_m)-F(r_m))^2\leq 6(F(t_m)-F(r_m))^2
\end{aligned} \tag{C.1.59}$$

$$\begin{aligned}
\mathbf{E}\mathscr{A}(t,s)^2\mathscr{B}(s,r)^2 &= \left(F^{'}(t_m)-F^{'}(s_m)\right)^2 \\
&\times(\sum_{i=1}^n c_i^2\mathbf{E}\varepsilon_i^2(\mathbf{1}_{r_m<\varepsilon_i\leq s_m}-F(s_m)+F(r_m))^2 \\
&+\sigma^2(n-1)\sum_{i=1}^n c_i^2\mathbf{E}(\mathbf{1}_{r_m<\varepsilon_1\leq s_m}-F(s_m)+F(r_m))^2 \\
&+2\sum_{i=1}^n\sum_{j=1,j\neq i}^n c_ic_j\times(\mathbf{E}\varepsilon_1\mathbf{1}_{r_m<\varepsilon_1\leq s_m})^2) \\
&\leq\sigma^2\left(F^{'}(t_m)-F^{'}(s_m)\right)^2\times(n\sum_{i=1}^n c_i^2+2((\sum_{i=1}^n c_i)^2+\sum_{i=1}^n c_i^2))
\end{aligned} \tag{C.1.60}$$

Notice that $n\sum_{i=1}^{n} c_i^2 = \mathscr{X}_f^T(\frac{\mathscr{X}^T\mathscr{X}}{n})^{-1}\mathscr{X}_f + 1 - 2\mathscr{X}_f^T(\frac{\mathscr{X}^T\mathscr{X}}{n})^{-1}\overline{\mathscr{X}}_n$ and $\sum_{i=1}^{n} c_i = \mathscr{X}_f^T(\mathscr{X}^T\mathscr{X}/n)^{-1}\overline{\mathscr{X}}_n - 1$, (C.1.48) is satisfied by choosing $G(x) = C'x$ with a sufficiently large constant C'. Then we prove $\widetilde{M}_m \to_{\mathscr{L}} \mathscr{M}_m$. In particular, for any given $0 < r < s < \infty$, choose sufficiently large integer $m > s+1$, from corollary 3

$$\begin{aligned}\sup_{x\in[r,s],z\in\mathbf{R}} &|\mathbf{P}\left(\widetilde{M}_m\left(\frac{x+m}{2m}\right) - \widetilde{M}_m^-\left(\frac{-x+m}{2m}\right) \le z\right) \\ &-\mathbf{P}\left(\mathscr{M}_m\left(\frac{x+m}{2m}\right) - \mathscr{M}_m\left(\frac{-x+m}{2m}\right) \le z\right)| = o(1)\end{aligned} \tag{C.1.61}$$

see(4.16) for the definition of the superscript $^-$. Since the random variable $\mathscr{M}_m\left(\frac{x+m}{2m}\right) - \mathscr{M}_m\left(\frac{-x+m}{2m}\right)$ has normal distribution with mean 0 and variance $\mathscr{U}(x)$ (see (4.15)), the proof remains showing that $\mathscr{S}(x)$ approximately equals $\widetilde{M}_m\left(\frac{x+m}{2m}\right) - \widetilde{M}_m^-\left(\frac{-x+m}{2m}\right)$.

Prove $\mathscr{S}(x)$ approximately equals $\widetilde{M}_m\left(\frac{x+m}{2m}\right) - \widetilde{M}_m^-\left(\frac{-x+m}{2m}\right)$

Recall the definition

$$\mathscr{S}(x) = \sqrt{n}\left(\mathbf{P}^*(|\mathscr{Y}_f - \mathscr{X}_f^T\widehat{\beta}| \le x) - \mathbf{P}^*(|\mathscr{Y}_f^* - \mathscr{X}_f^T\widehat{\beta}^*| \le x)\right) \tag{C.1.62}$$

here the condtional probability $\mathbf{P}^*$ is defined in definition 4. Since $\mathscr{Y}_f - \mathscr{X}_f^T\widehat{\beta} = \xi - \mathscr{X}_f^T\left(\mathscr{X}^T\mathscr{X}\right)^{-1}\mathscr{X}^T\varepsilon$, here ξ is a random variable being independent of ε and having the same distribution as ε_1. We have

$$\begin{aligned}&\mathbf{P}^*\left(|\mathscr{Y}_f - \mathscr{X}_f^T\widehat{\beta}| \le x\right) \\ =\ &\mathbf{P}^*\left(-x + \mathscr{X}_f^T\left(\mathscr{X}^T\mathscr{X}\right)^{-1}\mathscr{X}^T\varepsilon \le \xi \le x + \mathscr{X}_f^T\left(\mathscr{X}^T\mathscr{X}\right)^{-1}\mathscr{X}^T\varepsilon\right) \\ =\ &F\left(x + \mathscr{X}_f^T\left(\mathscr{X}^T\mathscr{X}\right)^{-1}\mathscr{X}^T\varepsilon\right) - F(-x + \mathscr{X}_f^T\left(\mathscr{X}^T\mathscr{X}\right)^{-1}\mathscr{X}^T\varepsilon)\end{aligned} \tag{C.1.63}$$

On the other hand, we have $\mathscr{Y}_f^* - \mathscr{X}_f^T\widehat{\beta}^* = \xi^* - \mathscr{X}_f^T\left(\mathscr{X}^T\mathscr{X}\right)^{-1}\mathscr{X}^T\varepsilon^*$, here $\varepsilon^* = (\varepsilon_1^*, ..., \varepsilon_n^*)^T$ and ξ^*, ε^* are independent with distribution $\widehat{F}$(see algorithm 4). Take the conditional distribution,

we have

$$\begin{aligned}
&\mathbf{P}^*(|\mathscr{Y}_f^* - \mathscr{X}_f^T\widehat{\beta}^*| \le x) \\
&= \begin{cases} \mathbf{E}^* Prob\left(|\mathscr{Y}_f^* - \mathscr{X}_f^T\widehat{\beta}^*| \le x \,|\mathscr{Y}, \varepsilon^*\right) \text{ for fixed design} \\ \mathbf{E}^* Prob\left(|\mathscr{Y}_f^* - \mathscr{X}_f^T\widehat{\beta}^*| \le x \,|\mathscr{Y}, \mathscr{X}, \mathscr{X}_f, \varepsilon^*\right) \text{ for random design} \end{cases} \\
&= \mathbf{E}^*\widehat{F}\left(x + \mathscr{X}_f^T\left(\mathscr{X}^T\mathscr{X}\right)^{-1}\mathscr{X}^T\varepsilon^*\right) - \mathbf{E}^*\widehat{F}^-\left(-x + \mathscr{X}_f^T\left(\mathscr{X}^T\mathscr{X}\right)^{-1}\mathscr{X}^T\varepsilon^*\right)
\end{aligned} \tag{C.1.64}$$

Choose $m > s+1$ and define $\widehat{\alpha}(x) = \sqrt{n}(\widehat{F}(x) - F(x))$ (the same as in (C.1.27)). $\forall x \in [r,s]$, from Taylor's theorem

$$\begin{aligned}
\mathscr{S}(x) &= \sqrt{n}(F(x + \mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon) - F(-x + \mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon)) \\
&-\sqrt{n}\left(\mathbf{E}^*\left(\widehat{F}(x + \mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon^*) - \widehat{F}^-(-x + \mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon^*)\right)\right) \\
&= \left(F'(x) - F'(-x)\right) \times \sqrt{n}\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon \\
&+\frac{F''(\eta_1) - F''(\eta_2)}{2} \times \sqrt{n}(\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon)^2 \\
&-\mathbf{E}^*\left(\widehat{\alpha}(x + \mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon^*) - \widehat{\alpha}(x)\right) \\
&+\mathbf{E}^*\left(\widehat{\alpha}^-(-x + \mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon^*) - \widehat{\alpha}^-(-x)\right) \\
&-\widehat{\alpha}(x) + \widehat{\alpha}^-(-x) - \sqrt{n}\mathbf{E}^*\left(F(x + \mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon^*) - F(x)\right) \\
&+\sqrt{n}\mathbf{E}^*(F(-x + \mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon^*) - F(-x))
\end{aligned} \tag{C.1.65}$$

which implies

$$\begin{aligned}
&\sup_{x\in[r,s]}|\mathscr{S}(x)-\left(\widetilde{M}_m(\frac{x+m}{2m})-\widetilde{M}_m^-(\frac{-x+m}{2m})\right)| \\
&\leq \sqrt{n}(\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon)^2\times\sup_{x\in\mathbf{R}}|F^{''}(x)| \\
&+\sup_{x\in[r,s]}|\mathbf{E}^*(\widehat{\alpha}(x+\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon^*)-\widehat{\alpha}(x))| \\
&+\sup_{x\in[r,s]}|\mathbf{E}^*(\widehat{\alpha}^-(-x+\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon^*)-\widehat{\alpha}^-(-x))| \\
&+\sup_{x\in[r,s]}|\widehat{\alpha}(x)-\widetilde{\alpha}(x)-\frac{F^{'}(x)}{\sqrt{n}}\sum_{i=1}^n\varepsilon_i|+\sup_{x\in[r,s]}|\widehat{\alpha}^-(-x)-\widetilde{\alpha}^-(-x)-\frac{F^{'}(-x)}{\sqrt{n}}\sum_{i=1}^n\varepsilon_i| \\
&+\sqrt{n}\mathbf{E}^*(\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon^*)^2\times\sup_{x\in\mathbf{R}}|F^{''}(x)|
\end{aligned} \tag{C.1.66}$$

here η_1,η_2 are two arbitrary real numbers. From lemma C.1.2, for any given $\xi>0$, $\exists 1/2>\delta>0$ such that for sufficiently large n, $\mathbf{P}\left(\sup_{x,y\in[-m,m],|x-y|<\delta}|\widehat{\alpha}(x)-\widehat{\alpha}(y)|\leq\xi\right)>1-\xi$. If $\sup_{x,y\in[-m,m],|x-y|<\delta}|\widehat{\alpha}(x)-\widehat{\alpha}(y)|\leq\xi$, then $\forall x\in[r,s]$,

$$\begin{aligned}
&|\mathbf{E}^*(\widehat{\alpha}^-(-x+\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon^*)-\widehat{\alpha}^-(-x))| \\
&\text{and }|\mathbf{E}^*(\widehat{\alpha}(x+\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon^*)-\widehat{\alpha}(x))| \\
&\leq\sqrt{n}\mathbf{P}^*(|\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon^*|>\delta)+\xi\leq\xi+\frac{\widehat{\sigma}^2}{\sqrt{n}\delta^2}\mathscr{X}_f^T\left(\frac{\mathscr{X}^T\mathscr{X}}{n}\right)^{-1}\mathscr{X}_f
\end{aligned} \tag{C.1.67}$$

$\widehat{\sigma}^2$ is defined in(4.14). Also notice that

$$\sqrt{n}\mathbf{E}^*(\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon^*)^2=\frac{\widehat{\sigma}^2}{\sqrt{n}}\mathscr{X}_f^T\left(\frac{\mathscr{X}^T\mathscr{X}}{n}\right)^{-1}\mathscr{X}_f \tag{C.1.68}$$

Since $\mathbf{E}\widehat{\sigma}^2\leq 4\sigma^2+\frac{4\sigma^2M^2}{n}\|(\frac{\mathscr{X}^T\mathscr{X}}{n})^{-1}\|_2+2\mathbf{E}\widehat{\lambda}^2=O(1)$, combine with (C.1.39) we have $\forall\xi>0$,

$$\mathbf{P}\left(\sup_{x\in[r,s]}|\mathscr{S}(x)-\left(\widetilde{M}_m(\frac{x+m}{2m})-\widetilde{M}_m^-(\frac{-x+m}{2m})\right)|>\xi\right)\to 0 \tag{C.1.69}$$

Finally, from (C.1.3) and corollary 3, $\forall \delta > 0$,

$$\begin{aligned}
&\sup_{x\in[r,s],y\in\mathbf{R}} |\mathbf{P}(\mathscr{S}(x)\le y) - \Phi\left(\frac{y}{\sqrt{\mathscr{U}(x)}}\right)| \\
&\le \sup_{x\in[r,s]} \mathbf{P}\left(|\mathscr{S}(x) - \widetilde{\mathscr{M}_m}(\frac{x+m}{2m}) - \widetilde{\mathscr{M}_m^-}(\frac{-x+m}{2m})| > \delta\right) \\
&+3 \sup_{x\in[r,s],y\in\mathbf{R}} |\mathbf{P}\left(\widetilde{\mathscr{M}_m}(\frac{x+m}{2m}) - \widetilde{\mathscr{M}_m^-}(\frac{-x+m}{2m}) \le y\right) - \Phi\left(\frac{y}{\sqrt{\mathscr{U}(x)}}\right)| \\
&+ \sup_{x\in[r,s],y\in\mathbf{R}} \left(\Phi\left(\frac{y+\delta}{\sqrt{\mathscr{U}(x)}}\right) - \Phi\left(\frac{y-\delta}{\sqrt{\mathscr{U}(x)}}\right)\right)
\end{aligned} \tag{C.1.70}$$

From assumption 4, we prove (4.20). □

C.2 Proofs of theorems in section 4.6

The Wasserstein distance can be used to quantify the difference between two probability distributions. We refer chapter 6, Villani [2009] for a detail introduction. Lemma C.2.1 below bounds the Wasserstein distance between the distribution $T(x) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\varepsilon_i - \overline{\varepsilon} \le x}$ and $F(x) = \mathbf{P}(\varepsilon_1 \le x)$, $x \in \mathbf{R}$. Here $\overline{\varepsilon} = \frac{1}{n}\sum_{i=1}^n \varepsilon_i$.

Lemma C.2.1. *Suppose assumption 1 and 2, then*

$$\lim_{n\to\infty} \inf_{X,Y} \mathbf{E}^*|X-Y|^2 = 0 \text{ almost surely} \tag{C.2.1}$$

The infimum is taken over all random variables $(X,Y) \in \mathbf{R}^2$ *such that* $\mathbf{P}^*(X \le x) = T(x)$ *and* $\mathbf{P}^*(Y \le x) = F(x)$.

Proof. From assumption 1, Gilvenko-Cantelli theorem, and the strong law of large number(e.g.,

theorem 1.13 in Shao [2003]),

$$\begin{aligned}\lim_{n\to\infty}\sup_{x\in\mathbf{R}}|T(x)-F(x)| \leq \lim_{n\to\infty}\sup_{x\in\mathbf{R}}|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\varepsilon_i\leq x}-F(x)| \\ +\lim_{n\to\infty}\sup_{x\in\mathbf{R}}|F(x+\overline{\varepsilon})-F(x)|=0 \text{ almost surely}\end{aligned} \tag{C.2.2}$$

From the strong law of large number, $\lim_{n\to\infty}\int_{\mathbf{R}}x^2dT=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2-\lim_{n\to\infty}\overline{\varepsilon}^2=\sigma^2$ almost surely. Choose $x_0=0$ in definition 6.8, Villani [2009]. From proposition 5.7, page 112 in Çinlar [2011] and theorem 6.9, Villani [2009], we prove (C.2.1). □

Recall (4.27) of the paper, the stochastic process $\widehat{\mathscr{S}}(x)$ is defined as

$$\begin{aligned}\widehat{\mathscr{M}}(x)=\sqrt{n}\widehat{F}\left(x+\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^Te^*-\frac{1}{n}\sum_{j=1}^{n}e_j^*\right)-\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\mathbf{1}_{e_j^*\leq x} \\ \text{and } \widehat{\mathscr{S}}(x)=\widehat{\mathscr{M}}(x)-\widehat{\mathscr{M}}^-(-x)\end{aligned} \tag{C.2.3}$$

Lemma C.2.2 ensures that $\widehat{\mathscr{S}}$ (defined in (4.27) of the paper) has the same asymptotic distribution as $\mathscr{S}$ (defined in (4.19), also see theorem 9).

Lemma C.2.2. *Suppose assumption 1 to 4 hold true. Then for any given* $0<r<s<\infty,\xi>0$*,*

$$\lim_{n\to\infty}\mathbf{P}\left(\sup_{x\in[r,s]}\sup_{y\in\mathbf{R}}|\mathbf{P}^*\left(\widehat{\mathscr{S}}(x)\leq y\right)-\Phi\left(\frac{y}{\sqrt{\mathscr{U}(x)}}\right)|>\xi\right)=0 \tag{C.2.4}$$

here Φ *is the cumulative distribution function of a standard normal random variable*

Recall that $\Phi^{-1}(\alpha)$ is the $\alpha-$th quantile of Φ. For any given $0<r<s<\infty$ and $\xi>0$,

lemma C.2.2 implies with probability tending to 1, $\forall 2\xi < 1-\gamma < 1-\xi, r \le x \le s$,

$$\begin{aligned}
\mathbf{P}^*\left(\widehat{\mathscr{S}}(x) \le \sqrt{\mathscr{U}(x)} \times \Phi^{-1}(1-\gamma-2\xi)\right) - (1-\gamma-2\xi) \le \xi \\
\Rightarrow d^*_{1-\gamma}(x) \ge \sqrt{\mathscr{U}(x)} \times \Phi^{-1}(1-\gamma-2\xi) \\
\mathbf{P}^*\left(\widehat{\mathscr{S}}(x) \le \sqrt{\mathscr{U}(x)} \times \Phi^{-1}(1-\gamma+\xi)\right) - (1-\gamma+\xi) \ge -\xi \\
\Rightarrow d^*_{1-\gamma}(x) \le \sqrt{\mathscr{U}(x)} \times \Phi^{-1}(1-\gamma+\xi)
\end{aligned} \tag{C.2.5}$$

see (4.28) for the definition of $d^*_{1-\gamma}(x)$.

Suppose the integer $m > s+1$. In (C.1.61) we show the stochastic process $\widetilde{M}_m\left(\frac{x+m}{2m}\right) - \widetilde{M}^-_m\left(\frac{-x+m}{2m}\right)$(defined in (C.1.1)) has an asymptotic distribution $\Phi\left(\cdot/\sqrt{\mathscr{U}(x)}\right)$. So the remaining problem involves approximating the distribution of $\widehat{\mathscr{S}}(x)$ by the distribution of $\widetilde{M}_m\left(\frac{x+m}{2m}\right) - \widetilde{M}^-_m\left(\frac{-x+m}{2m}\right)$.

Proof of lemma C.2.2. From lemma C.2.1, almost surely for $\forall\ 1/4 > \delta > 0$, $\exists\ N > 0$ such that $\forall\ n \ge N$, there exists a random vector $(e^\dagger_1, \varepsilon^\dagger_1) \in \mathbf{R}^2$ such that $\mathbf{P}^*(e^\dagger_1 \le x) = T(x)$ (defined in lemma C.2.1) and $\mathbf{P}^*(\varepsilon^\dagger_1 \le x) = F(x)$. Moreover, $\mathbf{E}^*(\varepsilon^\dagger_1 - e^\dagger_1)^2 < \delta^9$. We generate n i.i.d. observations $(e^\dagger_i, \varepsilon^\dagger_i),\ i = 1,2,...,n$ and define $e^\dagger = (e^\dagger_1,...,e^\dagger_n)^T$ as well as $\varepsilon^\dagger = (\varepsilon^\dagger_1,...,\varepsilon^\dagger_n)^T$. Suppose $m > s+1$ and define

$$\begin{aligned}
\widetilde{\mathscr{M}}^\dagger_m(x) = \sqrt{n}F'(x_m)\left(\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T e^\dagger - \frac{1}{n}\sum_{j=1}^n e^\dagger_j\right) \\
-\frac{1}{\sqrt{n}}\sum_{j=1}^n(\mathbf{1}_{e^\dagger_j \le x_m} - T(x_m)) \\
\widetilde{M}^\dagger_m(x) = \sqrt{n}F'(x_m)\left(\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T \varepsilon^\dagger - \frac{1}{n}\sum_{j=1}^n \varepsilon^\dagger_j\right) \\
-\frac{1}{\sqrt{n}}\sum_{j=1}^n(\mathbf{1}_{\varepsilon^\dagger_j \le x_m} - F(x_m))
\end{aligned} \tag{C.2.6}$$

here $x \in [0,1]$, $x_m = 2mx - m$. With this definition we have $\mathbf{P}^*(\widetilde{M}^\dagger_m(x) \le y) = \mathbf{P}(\widetilde{M}_m(x) \le y)$ for

any x, y. For any given $1/4 > \xi > 0$ and $x \in (\frac{1}{2}, 1]$,

$$\begin{aligned}
&\mathbf{P}^*\left(|(\widetilde{\mathscr{M}}_m^{\dagger}(x) - \widetilde{\mathscr{M}}_m^{\dagger -}(1-x)) - (\widetilde{M}_m^{\dagger}(x) - \widetilde{M}_m^{\dagger -}(1-x))| > 3\xi\right) \\
&\leq \mathbf{P}^*\left(\sqrt{n}|F^{'}(x_m) - F^{'}(-x_m)| \times |\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T(\varepsilon^{\dagger} - e^{\dagger})| > \xi\right) \\
&\quad + \mathbf{P}^*\left(|F^{'}(x_m) - F^{'}(-x_m)| \times |\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\varepsilon_i^{\dagger} - e_i^{\dagger})| > \xi\right) \\
&\quad + \mathbf{P}^*(\frac{1}{\sqrt{n}}|\sum_{i=1}^{n}\mathbf{1}_{-x_m \leq e_i^{\dagger} \leq x_m} - T(x_m) + T^-(-x_m) - \mathbf{1}_{-x_m \leq \varepsilon_i^{\dagger} \leq x_m} \\
&\quad + F(x_m) - F(-x_m)| > \xi) \\
&\leq \frac{(F^{'}(x_m) - F^{'}(-x_m))^2\mathbf{E}^*(\varepsilon_1^{\dagger} - e_1^{\dagger})^2}{\xi^2} \times \left(\mathscr{X}_f^T\left(\frac{\mathscr{X}^T\mathscr{X}}{n}\right)^{-1}\mathscr{X}_f + 1\right) \\
&\quad + \frac{4}{\xi^2}\mathbf{E}^*(\mathbf{1}_{e_1^{\dagger} \leq x_m} - T(x_m) - \mathbf{1}_{\varepsilon_1^{\dagger} \leq x_m} + F(x_m))^2 \\
&\quad + \frac{4}{\xi^2}\mathbf{E}^*(\mathbf{1}_{e_1^{\dagger} < -x_m} - T^-(-x_m) - \mathbf{1}_{\varepsilon_1^{\dagger} < -x_m} + F(-x_m))^2
\end{aligned} \tag{C.2.7}$$

Notice that

$$\begin{aligned}
\mathbf{E}^*(\mathbf{1}_{e_1^{\dagger} \leq x_m} - T(x_m) - \mathbf{1}_{\varepsilon_1^{\dagger} \leq x_m} + F(x_m))^2 &\leq 2\mathbf{E}^*(\mathbf{1}_{e_1^{\dagger} \leq x_m} - \mathbf{1}_{\varepsilon_1^{\dagger} \leq x_m})^2 \\
&\quad + 2\sup_{x \in \mathbf{R}}|T(x) - F(x)|^2
\end{aligned} \tag{C.2.8}$$

from (C.1.3),

$$\begin{aligned}
\mathbf{E}^*|\mathbf{1}_{e_1^{\dagger} \leq x_m} - \mathbf{1}_{\varepsilon_1^{\dagger} \leq x_m}| &\leq \mathbf{P}^*(|e_1^{\dagger} - \varepsilon_1^{\dagger}| > \xi) + F(x_m + \xi) - F(x_m - \xi) \\
&\leq \frac{\delta^9}{\xi^2} + \sup_{x \in \mathbf{R}}(F(x) - F(x - 2\xi))
\end{aligned} \tag{C.2.9}$$

From dominated convergence theorem

$$\begin{aligned}
&\mathbf{E}^*(\mathbf{1}_{e_1^\dagger<-x_m}-T^-(-x_m)-\mathbf{1}_{\varepsilon_1^\dagger<-x_m}+F(-x_m))^2\\
=&\lim_{h\to\infty}\mathbf{E}^*(\mathbf{1}_{e_1^\dagger\le-x_m-\frac{1}{h}}-T(-x_m-\frac{1}{h})-\mathbf{1}_{\varepsilon_1^\dagger\le-x_m-\frac{1}{h}}+F(-x_m-\frac{1}{h}))^2\\
\le&\frac{2\delta^9}{\xi^2}+2\sup_{x\in\mathbf{R}}(F(x)-F(x-2\xi))+2\sup_{x\in\mathbf{R}}|T(x)-F(x)|^2
\end{aligned}\tag{C.2.10}$$

therefore, from (C.1.3), assumption 4 and (C.1.61)

$$\begin{aligned}
&\sup_{x\in[\frac{r+m}{2m},\frac{s+m}{2m}],y\in\mathbf{R}}|\mathbf{P}^*\left(\widetilde{\mathscr{M}}_m^\dagger(x)-\widetilde{\mathscr{M}}_m^{\dagger-}(1-x)\le y\right)-\Phi\left(\frac{y}{\sqrt{\mathscr{U}(x_m)}}\right)|\\
\le&\sup_{x\in[\frac{r+m}{2m},\frac{s+m}{2m}],y\in\mathbf{R}}\mathbf{P}^*\left(|(\widetilde{\mathscr{M}}_m^\dagger(x)-\widetilde{\mathscr{M}}_m^{\dagger-}(1-x))-(\widetilde{M}_m^\dagger(x)-\widetilde{M}_m^{\dagger-}(1-x))|>3\xi\right)\\
&+3\sup_{x\in[\frac{r+m}{2m},\frac{s+m}{2m}],y\in\mathbf{R}}|\mathbf{P}^*\left(\widetilde{M}_m^\dagger(x)-\widetilde{M}_m^{\dagger-}(1-x)\le y\right)-\Phi\left(\frac{y}{\sqrt{\mathscr{U}(x_m)}}\right)|\\
&+\sup_{x\in[\frac{r+m}{2m},\frac{s+m}{2m}],y\in\mathbf{R}}\left(\Phi\left(\frac{y+3\xi}{\sqrt{\mathscr{U}(x_m)}}\right)-\Phi\left(\frac{y-3\xi}{\sqrt{\mathscr{U}(x_m)}}\right)\right)\\
\Rightarrow&\lim_{n\to\infty}\sup_{x\in[\frac{r+m}{2m},\frac{s+m}{2m}],y\in\mathbf{R}}|\mathbf{P}^*\left(\widetilde{\mathscr{M}}_m^\dagger(x)-\widetilde{\mathscr{M}}_m^{\dagger-}(1-x)\le y\right)-\Phi\left(\frac{y}{\sqrt{\mathscr{U}(x_m)}}\right)|\\
&=0\text{ almost surely}
\end{aligned}\tag{C.2.11}$$

Define a random variable $(e_1^*,e_1^\dagger)\in\mathbf{R}^2$ which has probability mass $1/n$ on $(\widehat{\varepsilon}_i,\varepsilon_i-\overline{\varepsilon}),i=1,2,...,n$. We generate independent random variables $(e_i^*,e_i^\dagger),i=1,2,...,n$ having the same distribution as $(e_1^*,e_1^\dagger)$. Define $e^*=(e_1^*,...,e_n^*)^T,e^\dagger=(e_1^\dagger,...,e_n^\dagger)^T$. With this definition $e_1^\dagger$ still has the cumulative distribution function $T(x)$. Define the stochastic process

$$\widetilde{\mathscr{M}}_m^*(x)=\sqrt{n}F^{'}(x_m)\left(\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^Te^*-\frac{1}{n}\sum_{i=1}^n e_i^*\right)-\frac{1}{\sqrt{n}}\sum_{i=1}^n(\mathbf{1}_{e_i^*\le x_m}-\widehat{F}(x_m))\tag{C.2.12}$$

here $x_m=2mx-m$. This process uses the same mechanism for generating residuals e^* as in

RBUG(defined in algorithm 5). We have

$$\begin{aligned}
&\mathbf{P}^*\left(|\widetilde{\mathscr{M}_m^*}(x)-\widetilde{\mathscr{M}_m^{*-}}(1-x)-\widetilde{\mathscr{M}_m^{\dagger}}(x)+\widetilde{\mathscr{M}_m^{\dagger-}}(1-x)|>3\xi\right)\\
\leq\ &\frac{|F^{'}(x_m)-F^{'}(-x_m)|^2\mathbf{E}^*(e_1^*-e_1^{\dagger})^2}{\xi^2}\times\left(\mathscr{X}_f^T\left(\frac{\mathscr{X}^T\mathscr{X}}{n}\right)^{-1}\mathscr{X}_f+1\right)\\
&+\frac{4}{\xi^2}\mathbf{E}^*(\mathbf{1}_{e_1^*\leq x_m}-\widehat{F}(x_m)-\mathbf{1}_{e_1^{\dagger}\leq x_m}+T(x_m))^2\\
&+\frac{4}{\xi^2}\mathbf{E}^*(\mathbf{1}_{e_1^*<-x_m}-\widehat{F}^-(-x_m)-\mathbf{1}_{e_1^{\dagger}<-x_m}+T^-(-x_m))^2
\end{aligned}\tag{C.2.13}$$

Recall $\overline{\mathscr{X}}_n=\frac{1}{n}\sum_{i=1}^n\mathscr{X}_i$,

$$\begin{aligned}
\mathbf{E}^*(e_1^*-e_1^{\dagger})^2&=\frac{1}{n}\sum_{i=1}^n(\widehat{\varepsilon}_i-\varepsilon_i+\overline{\varepsilon})^2=\frac{1}{n}\sum_{i=1}^n\left((\mathscr{X}_i-\overline{\mathscr{X}}_n)^T(\widehat{\beta}-\beta)\right)^2\\
\mathbf{E}\left((\mathscr{X}_i-\overline{\mathscr{X}}_n)^T(\widehat{\beta}-\beta)\right)^2&=\sigma^2(\mathscr{X}_i-\overline{\mathscr{X}}_n)^T(\mathscr{X}^T\mathscr{X})^{-1}(\mathscr{X}_i-\overline{\mathscr{X}}_n)
\end{aligned}\tag{C.2.14}$$

Assumption 3 implies $\mathbf{E}^*(e_1^*-e_1^{\dagger})^2=O_p(1/n)$. From assumption 3 and Cauchy inequality

$$\begin{aligned}
\widehat{F}(x)=\frac{1}{n}\sum_{i=1}^n\mathbf{1}_{\varepsilon_i-\overline{\varepsilon}\leq x+(\mathscr{X}_i-\overline{\mathscr{X}}_n)^T(\widehat{\beta}-\beta)}&\leq T(x+2M\|\widehat{\beta}-\beta\|_2)\\
\text{and }\widehat{F}(x)&\geq T(x-2M\|\widehat{\beta}-\beta\|_2)
\end{aligned}\tag{C.2.15}$$

therefore

$$\begin{aligned}
\sup_{x\in\mathbf{R}}|\widehat{F}(x)-T(x)|&\leq\sup_{x\in\mathbf{R}}|T(x+2M\|\widehat{\beta}-\beta\|_2)-T(x)|\\
&\quad+\sup_{x\in\mathbf{R}}|T(x-2M\|\widehat{\beta}-\beta\|_2)-T(x)|\\
&\leq 4\sup_{x\in\mathbf{R}}|F(x)-T(x)|+2\sup_{x\in\mathbf{R}}|F(x+2M\|\widehat{\beta}-\beta\|_2)-F(x)|\\
&\quad+2\sup_{x\in\mathbf{R}}|F(x-2M\|\widehat{\beta}-\beta\|_2)-F(x)|
\end{aligned}\tag{C.2.16}$$

Since

$$\begin{aligned}
&\mathbf{E}^*(\mathbf{1}_{e_1^*\le x_m}-\widehat{F}(x_m)-\mathbf{1}_{e_1^\dagger\le x_m}+T(x_m))^2\\
&\le \frac{2\mathbf{E}^*(e_1^*-e_1^\dagger)^2}{\xi^2}+2\sup_{x\in\mathbf{R}}(T(x+\xi)-T(x-\xi))+2\sup_{x\in\mathbf{R}}|\widehat{F}(x)-T(x)|^2
\end{aligned}\tag{C.2.17}$$

The dominated convergence theorem implies

$$\begin{aligned}
&\mathbf{E}^*(\mathbf{1}_{e_1^*<-x_m}-\widehat{F}^-(-x_m)-\mathbf{1}_{e_1^\dagger<-x_m}+T^-(-x_m))^2\\
&=\lim_{h\to\infty}\mathbf{E}^*(\mathbf{1}_{e_1^*\le -x_m-\frac{1}{h}}-\widehat{F}(-x_m-\frac{1}{h})-\mathbf{1}_{e_1^\dagger\le -x_m-\frac{1}{h}}+T(-x_m-\frac{1}{h}))^2\\
&\le \frac{2\mathbf{E}^*(e_1^*-e_1^\dagger)^2}{\xi^2}+2\sup_{x\in\mathbf{R}}(T(x+\xi)-T(x-\xi))+2\sup_{x\in\mathbf{R}}|\widehat{F}(x)-T(x)|^2
\end{aligned}\tag{C.2.18}$$

and

$$\begin{aligned}
&\sup_{x\in[\frac{r+m}{2m},\frac{s+m}{2m}],y\in\mathbf{R}}|\mathbf{P}^*\left(\widetilde{\mathscr{M}_m^*}(x)-\widetilde{\mathscr{M}_m^{*-}}(1-x)\le y\right)-\Phi\left(\frac{y}{\sqrt{\mathscr{U}(x_m)}}\right)|\\
\le&\sup_{x\in[\frac{r+m}{2m},\frac{s+m}{2m}],y\in\mathbf{R}}\mathbf{P}^*\left(|(\widetilde{\mathscr{M}_m^*}(x)-\widetilde{\mathscr{M}_m^{*-}}(1-x))-(\widetilde{\mathscr{M}_m^\dagger}(x)-\widetilde{\mathscr{M}_m^{\dagger-}}(1-x))|>3\xi\right)\\
&+3\sup_{x\in[\frac{r+m}{2m},\frac{s+m}{2m}],y\in\mathbf{R}}|\mathbf{P}^*\left(\widetilde{\mathscr{M}_m^\dagger}(x)-\widetilde{\mathscr{M}_m^{\dagger-}}(1-x)\le y\right)-\Phi\left(\frac{y}{\sqrt{\mathscr{U}(x_m)}}\right)|\\
&+\sup_{x\in[\frac{r+m}{2m},\frac{s+m}{2m}],y\in\mathbf{R}}\left(\Phi\left(\frac{y+3\xi}{\sqrt{\mathscr{U}(x_m)}}\right)-\Phi\left(\frac{y-3\xi}{\sqrt{\mathscr{U}(x_m)}}\right)\right)
\end{aligned}\tag{C.2.19}$$

(C.2.2), (C.2.11), and (C.2.16) imply for $\forall\xi>0$

$$\begin{aligned}
\lim_{n\to\infty}\mathbf{P}(&\sup_{x\in[\frac{r+m}{2m},\frac{s+m}{2m}],y\in\mathbf{R}}|\mathbf{P}^*\left(\widetilde{\mathscr{M}_m^*}(x)-\widetilde{\mathscr{M}_m^{*-}}(1-x)\le y\right)\\
&-\Phi\left(\frac{y}{\sqrt{\mathscr{U}(x_m)}}\right)|>\xi)=0
\end{aligned}\tag{C.2.20}$$

Finally, we adopt the notations in lemma C.1.2. Recall (4.27) (or (C.2.3) in this section), define

$x_m = 2mx - m$

$$\begin{aligned}\sup_{x\in[0,1]} |\widehat{\mathscr{M}}(x_m) - \widetilde{\mathscr{M}}^*_m(x)| &\le \frac{\sqrt{n}\sup_{x\in\mathbf{R}}|F''(x)|}{2} \\ &\times \left(\mathscr{X}_f^T (\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T e^* - \frac{1}{n}\sum_{i=1}^n e_i^* \right)^2 \\ + \sup_{x\in[0,1]} |\widehat{\alpha}(x_m + \mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T e^* - \frac{1}{n}\sum_{i=1}^n e_i^*) - \widehat{\alpha}(x_m)|\end{aligned} \tag{C.2.21}$$

and

$$\begin{aligned}&\sup_{x\in[r,s]} |\widehat{\mathscr{S}}(x) - \left(\widetilde{\mathscr{M}}^*_m(\frac{x+m}{2m}) - \widetilde{\mathscr{M}}^{*-}_m(\frac{-x+m}{2m}) \right)| \le \sup_{x\in[r,s]} |\widehat{\mathscr{M}}(x) - \widetilde{\mathscr{M}}^*_m(\frac{x+m}{2m})| \\ &+ \sup_{x\in[r,s]} \lim_{h\to\infty} |\widehat{\mathscr{M}}(-x-\frac{1}{h}) - \widetilde{M}^*_m(\frac{-x+m}{2m} - \frac{1}{2hm})| \le 2 \sup_{x\in[0,1]} |\widehat{\mathscr{M}}(x_m) - \widetilde{\mathscr{M}}^*_m(x)|\end{aligned} \tag{C.2.22}$$

$\forall \xi > 0$, (C.1.3) implies

$$\begin{aligned}\sup_{x\in[r,s],y\in\mathbf{R}} |\mathbf{P}^*\left(\widehat{\mathscr{S}}(x) \le y\right) - \Phi\left(\frac{y}{\sqrt{\mathscr{U}(x)}}\right)| &\le \mathbf{P}^*\left(\sup_{x\in[0,1]} |\widehat{\mathscr{M}}(x_m) - \widetilde{\mathscr{M}}^*_m(x)| > \xi\right) \\ &+ \sup_{x\in[r,s],y\in\mathbf{R}} \left(\Phi\left(\frac{y+2\xi}{\sqrt{\mathscr{U}(x)}}\right) - \Phi\left(\frac{y-2\xi}{\sqrt{\mathscr{U}(x)}}\right)\right) \\ +3 \sup_{x\in[\frac{r+m}{2m},\frac{s+m}{2m}],y\in\mathbf{R}} &|\mathbf{P}^*\left(\widetilde{\mathscr{M}}^*_m(x) - \widetilde{\mathscr{M}}^{*-}_m(1-x) \le y\right) - \Phi\left(\frac{y}{\sqrt{\mathscr{U}(x_m)}}\right)|\end{aligned} \tag{C.2.23}$$

From lemma C.1.2, for any given $\xi > 0$, $\exists \frac{1}{4} > \xi_2 > 0, N > 0$ such that for any $n \ge N$, $\mathbf{P}\left(\sup_{x,y\in[-m-1,m+1],|x-y|<\xi_2} |\widehat{\alpha}(x) - \widehat{\alpha}(y)| > \frac{\xi}{4}\right) < \xi$. If

$\sup_{x,y\in[-m-1,m+1],|x-y|<\xi_2} |\widehat{\alpha}(x)-\widehat{\alpha}(y)| \leq \frac{\xi}{4}$, then

$$\begin{aligned}
&\mathbf{P}^*\left(\sup_{x\in[0,1]}|\widehat{\mathscr{M}}(x_m)-\widetilde{\mathscr{M}^*_m}(x)|>\xi\right)\\
\leq\ &\mathbf{P}^*\left(\frac{\sqrt{n}\sup_{x\in\mathbf{R}}|F''(x)|}{2}\times\left(\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T e^*-\frac{1}{n}\sum_{i=1}^n e_i^*\right)^2>\frac{\xi}{2}\right)\\
&+\mathbf{P}^*\left(|\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T e^*-\frac{1}{n}\sum_{i=1}^n e_i^*|\geq\xi_2\right)\\
\leq\ &\left(\frac{\sqrt{n}\sup_{x\in\mathbf{R}}|F''(x)|}{\xi}+\frac{1}{\xi_2^2}\right)\mathbf{E}^*\left(\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T e^*-\frac{1}{n}\sum_{i=1}^n e_i^*\right)^2
\end{aligned} \tag{C.2.24}$$

Since $\mathbf{E}^*\left(\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T e^*-\frac{1}{n}\sum_{i=1}^n e_i^*\right)^2 \leq \frac{2\widehat{\sigma}^2}{n}\left(\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X}/n)^{-1}\mathscr{X}_f+1\right)$.
From (C.2.19) we prove (C.2.4). □

In lemma C.2.3, we define

$$G^*(x)=\mathbf{P}^*\left(|\mathscr{Y}_f^*-\mathscr{X}_f^T\widehat{\beta}^*|\leq x\right),\ x\in\mathbf{R} \tag{C.2.25}$$

See algorithm 5 for the meaning of $\mathscr{Y}_f^*$ and $\widehat{\beta}^*$

Lemma C.2.3. *Suppose assumption 1 to 4. Then* $\forall -\infty<r<s<\infty, \zeta>0$, $\exists\delta>0$ *such that*

$$\limsup_{n\to\infty}\mathbf{P}\left(\sup_{x\in[r,s]}\sqrt{n}\left(G^*(x)-G^*(x-\frac{\delta}{\sqrt{n}})\right)\geq\zeta\right)<\zeta \tag{C.2.26}$$

Proof. We adopt the notations in lemma C.1.2 and recall $\mathscr{Y}_f^*=\mathscr{X}_f^T\widehat{\beta}+\xi^*$. By conditioning on

ε^*,

$$
\begin{aligned}
G^*(x) &= \mathbf{E}^*\mathbf{P}^*\left(|\xi^* - \mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon^*| \le x \middle| \varepsilon^*\right) \\
&= \mathbf{E}^*\widehat{F}(x+\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon^*) - \mathbf{E}^*\widehat{F}^-(-x+\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon^*) \\
&= (F(x)-F(-x)) + \mathbf{E}^*\left(F(x+\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon^*) - F(x)\right) \\
&-\mathbf{E}^*\left(F(-x+\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon^*) - F(-x)\right) + \frac{(\widehat{\alpha}(x)-\widehat{\alpha}^-(-x))}{\sqrt{n}} \\
&+\frac{1}{\sqrt{n}}\mathbf{E}^*\left(\widehat{\alpha}(x+\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon^*) - \widehat{\alpha}(x)\right) \\
&-\frac{1}{\sqrt{n}}\mathbf{E}^*\left(\widehat{\alpha}^-(-x+\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon^*) - \widehat{\alpha}^-(-x)\right)
\end{aligned}
\tag{C.2.27}
$$

Therefore, for $\forall \frac{1}{4} > \delta > 0$,

$$
\begin{aligned}
\sup_{x\in[r,s]}\sqrt{n}\left(G^*(x) - G^*(x-\frac{\delta}{\sqrt{n}})\right) &\le 2\delta \sup_{x\in[-s-1,s+1]}|F^{'}(x)| \\
&+\frac{2\sup_{x\in\mathbf{R}}|F^{''}(x)|\widehat{\sigma}^2}{\sqrt{n}}\left(\mathscr{X}_f^T\left(\frac{\mathscr{X}^T\mathscr{X}}{n}\right)^{-1}\mathscr{X}_f\right) \\
&+2\sup_{x,y\in[-s-1,s+1],|x-y|\le\frac{\delta}{\sqrt{n}}}|\widehat{\alpha}(x)-\widehat{\alpha}(y)| \\
+4\sup_{x\in[-s-1,s+1]}\mathbf{E}^*|\widehat{\alpha}(x+\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon^*) - \widehat{\alpha}(x)|&
\end{aligned}
\tag{C.2.28}
$$

From lemma C.1.2 and (C.1.67), we prove (C.2.26). □

Suppose assumption 1 to 4, from (C.2.27), (C.2.28), (C.2.16), and (C.2.2),

$$
\begin{aligned}
\sup_{x>0}|G^*(x) - (F(x)-F(-x))| &\le 2\sup_{x\in\mathbf{R}}|\widehat{F}(x)-F(x)| \\
&+\frac{\sup_{x\in\mathbf{R}}|F^{''}(x)|\widehat{\sigma}^2}{n}\left(\mathscr{X}_f^T\left(\frac{\mathscr{X}^T\mathscr{X}}{n}\right)^{-1}\mathscr{X}_f\right)
\end{aligned}
\tag{C.2.29}
$$

which implies $\forall\xi > 0$, $\lim_{n\to\infty}\mathbf{P}(\sup_{x>0}|G^*(x) - (F(x)-F(-x))| > \xi) = 0$. If $\sup_{x>0}|G^*(x) -$

$(F(x)-F(-x))| \leq \xi$, by defining $c_{1-\alpha}$ such that $F(c_{1-\alpha})-F(-c_{1-\alpha}) = 1-\alpha$,

$$\begin{aligned} G^*(c_{1-\alpha+2\xi}) \geq 1-\alpha+\xi,\ G^*(c_{1-\alpha-2\xi}) \leq 1-\alpha-\xi \\ \Rightarrow c_{1-\alpha-2\xi} \leq c^*_{1-\alpha} \leq c_{1-\alpha+2\xi},\ \forall 2\xi < \alpha < 1-2\xi \end{aligned} \tag{C.2.30}$$

proof of theorem 10. Recall $\Phi^{-1}(\alpha)$ is the $\alpha-$th quantile of the standard normal distribution. We choose r,s in lemma C.2.2 as $c_{(1-\alpha)/4}, c_{1-\alpha/4}$, here $F(c_z)-F(-c_z) = z, \forall z \in (0,1)$. From (C.2.5), (C.2.29) and (C.2.30), for sufficiently small $\xi > 0$, with probability tending to 1,

$$\begin{aligned} d^*_{1-\gamma}(c^*_{1-\alpha}) &\leq \sup_{x\in[c_{1-\alpha-2\xi},c_{1-\alpha+2\xi}]} d^*_{1-\gamma}(x) \\ &\leq \sup_{x\in[c_{1-\alpha-2\xi},c_{1-\alpha+2\xi}]} \sqrt{\mathscr{U}(x)} \times \Phi^{-1}(1-\gamma+\xi) \\ &\text{and } d^*_{1-\gamma}(c^*_{1-\alpha}) \\ \geq \inf_{x\in[c_{1-\alpha-2\xi},c_{1-\alpha+2\xi}]} d^*_{1-\gamma}(x) &\geq \inf_{x\in[c_{1-\alpha-2\xi},c_{1-\alpha+2\xi}]} \sqrt{\mathscr{U}(x)} \times \Phi^{-1}(1-\gamma-2\xi) \end{aligned} \tag{C.2.31}$$

Define

$$\begin{aligned} \overline{d} &= \sup_{x\in[c_{1-\alpha-2\xi},c_{1-\alpha+2\xi}]} \sqrt{\mathscr{U}(x)} \times \Phi^{-1}(1-\gamma+\xi) \\ \text{and } \underline{d} &= \inf_{x\in[c_{1-\alpha-2\xi},c_{1-\alpha+2\xi}]} \sqrt{\mathscr{U}(x)} \times \Phi^{-1}(1-\gamma-2\xi) \end{aligned} \tag{C.2.32}$$

From (C.2.29), with probability tending to 1,

$$\begin{aligned} c^*(1-\alpha,1-\gamma) \leq c^*_{1-\alpha+\frac{\overline{d}}{\sqrt{n}}} \leq c_{1-\alpha+\frac{\overline{d}}{\sqrt{n}}+2\xi} &\text{ and } c^*(1-\alpha,1-\gamma) \geq c^*_{1-\alpha+\frac{d}{\sqrt{n}}} \\ &\geq c_{1-\alpha+\frac{d}{\sqrt{n}}-2\xi} \end{aligned} \tag{C.2.33}$$

Define $\overline{c} = c_{1-\alpha+\frac{\overline{d}}{\sqrt{n}}+2\xi}$, and $\underline{c} = c_{1-\alpha+\frac{d}{\sqrt{n}}-2\xi}$. From assumption 1 and 4, c_α is continuous in

$\alpha \in (0,1)$; and $\mathscr{U}(x)$ is continuous in $\mathbf{R}$. Define $\mathscr{S}$ as in (4.19). Then

$$\begin{aligned}
&\sqrt{n}\left(\mathbf{P}^*\left(|\mathscr{Y}_f - \mathscr{X}_f^T\widehat{\beta}| \le c^*(1-\alpha,1-\gamma)\right) - (1-\alpha)\right) \\
&= \mathscr{S}(c_{1-\alpha}) + (\mathscr{S}(c^*(1-\alpha,1-\gamma)) - \mathscr{S}(c_{1-\alpha})) \\
&+\sqrt{n}\left(G^*(c^*(1-\alpha,1-\gamma)) - (1-\alpha+\frac{d^*_{1-\gamma}(c^*_{1-\alpha})}{\sqrt{n}})\right) \\
&+\sqrt{\mathscr{U}(c_{1-\alpha})}\times\Phi^{-1}(1-\gamma) + \left(d^*_{1-\gamma}(c^*_{1-\alpha}) - \sqrt{\mathscr{U}(c_{1-\alpha})}\times\Phi^{-1}(1-\gamma)\right)
\end{aligned} \tag{C.2.34}$$

we choose $r = \underline{c}$ and $s = \overline{c}$ in lemma C.2.3. With probability tending to 1

$$\begin{aligned}
&|\sqrt{n}\left(G^*(c^*(1-\alpha,1-\gamma)) - (1-\alpha+\frac{d^*_{1-\gamma}(c^*_{1-\alpha})}{\sqrt{n}})\right)| \\
&\le \sqrt{n}\left(G^*(c^*(1-\alpha,1-\gamma)) - G^*\left(c^*(1-\alpha,1-\gamma) - \frac{1}{n}\right)\right) < \xi
\end{aligned} \tag{C.2.35}$$

We choose a positive integer $m > \overline{c}+1$. From (C.1.69) and lemma C.1.1, with probability tending to 1,

$$\begin{aligned}
&|\mathscr{S}(c^*(1-\alpha,1-\gamma)) - \mathscr{S}(c_{1-\alpha})| \le \sup_{x\in[\underline{c},\overline{c}]}|\mathscr{S}(x) - \mathscr{S}(c_{1-\alpha})| \\
&\le 2\sup_{x\in[\underline{c},\overline{c}]}|\mathscr{S}(x) - \left(\widetilde{M}_m(\frac{x+m}{2m}) - \widetilde{M}_m^-(\frac{-x+m}{2m})\right)| \\
&+2\sup_{y,z\in[0,1],|y-z|\le\frac{\overline{c}-\underline{c}}{2m}}|\widetilde{M}_m(y) - \widetilde{M}_m(z)| \\
&\Rightarrow \text{for } \forall\xi>0,\ \limsup_{n\to\infty}\mathbf{P}(|\mathscr{S}(c^*(1-\alpha,1-\gamma)) - \mathscr{S}(c_{1-\alpha})| > \xi) < \xi
\end{aligned} \tag{C.2.36}$$

For $\mathscr{U}$ is continuous and

$$\begin{aligned}
&|d^*_{1-\gamma}(c^*_{1-\alpha}) - \sqrt{\mathscr{U}(c_{1-\alpha})}\times\Phi^{-1}(1-\gamma)| \\
&\le |\overline{d} - \sqrt{\mathscr{U}(c_{1-\alpha})}\Phi^{-1}(1-\gamma)| + |\underline{d} - \sqrt{\mathscr{U}(c_{1-\alpha})}\Phi^{-1}(1-\gamma)|
\end{aligned} \tag{C.2.37}$$

with probability tending to 1, we have for $\forall \xi > 0$,

$$\begin{aligned}\lim_{n\to\infty} \mathbf{P}(|\sqrt{n}\left(\mathbf{P}^*\left(|\mathscr{Y}_f - \mathscr{X}_f^T\widehat{\beta}| \le c^*(1-\alpha,1-\gamma)\right) - (1-\alpha)\right) \\ - \left(\mathscr{S}(c_{1-\alpha}) + \sqrt{\mathscr{U}(c_{1-\alpha})} \times \Phi^{-1}(1-\gamma)\right)| > \xi) = 0\end{aligned} \tag{C.2.38}$$

On one hand, from theorem 9, $\forall \xi > 0$, we choose $Z > 0$ such that $\Phi\left(\frac{Z}{\sqrt{\mathscr{U}(c_{1-\alpha})}}\right) - \Phi\left(-\frac{Z}{\sqrt{\mathscr{U}(c_{1-\alpha})}}\right) > 1-\xi$, we have $\lim_{n\to\infty}\mathbf{P}(|\mathscr{S}(c_{1-\alpha})| \le Z) > 1-\xi$. On the other hand, for any given $\xi \in \mathbf{R}$,

$$\begin{aligned}\lim_{n\to\infty} \mathbf{P}\left(\mathscr{S}(c_{1-\alpha}) + \sqrt{\mathscr{U}(c_{1-\alpha})} \times \Phi^{-1}(1-\gamma) + \xi \ge 0\right) \\ = 1 - \Phi\left(-\Phi^{-1}(1-\gamma) - \frac{\xi}{\sqrt{\mathscr{U}(c_{1-\alpha})}}\right)\end{aligned} \tag{C.2.39}$$

Combine with (C.2.38), we prove theorem 10. □

Proof of corollary 2. From theorem 10.1 in Seber and Lee [2003] and assumption 3, define $\widetilde{\varepsilon}_i$ and $\widetilde{r}_i$ as in (4.13) and (4.18), $\widetilde{r}_i = \widetilde{\varepsilon}_i/(1-h_i)$ with $h_i = \mathscr{X}_i^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}_i$, and $\exists C > 0$ such that $h_i \le C/n$ for $i = 1,2,...,n$. From Cauchy inequality, for sufficiently large n

$$\begin{aligned}\widehat{r}_i &= \frac{\widehat{\varepsilon}_i}{1-h_i} + \frac{1}{n}\sum_{j=1}^{n}\frac{(h_i-h_j)\widetilde{\varepsilon}_j}{(1-h_i)(1-h_j)} \\ \Rightarrow \sum_{i=1}^{n}(\widehat{r}_i - \widehat{\varepsilon}_i)^2 &\le \sum_{i=1}^{n}\frac{2h_i^2\widehat{\varepsilon}_i^2}{(1-h_i)^2} + \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{(h_i-h_j)^2}{(1-h_i)^2(1-h_j)^2}\sum_{j=1}^{n}\widetilde{\varepsilon}_j^2 \\ &\le \frac{4C^2}{n^2}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2 + \frac{16C^2}{n^2}\sum_{j=1}^{n}\widetilde{\varepsilon}_j^2 \\ \Rightarrow \mathbf{E}\sum_{i=1}^{n}(\widehat{r}_i - \widehat{\varepsilon}_i)^2 &\le \frac{4C^2}{n^2}\sum_{i=1}^{n}\mathbf{E}\widehat{\varepsilon}_i^2 + \frac{16C^2}{n^2}\sum_{j=1}^{n}\mathbf{E}\widetilde{\varepsilon}_j^2 \\ &\le \frac{20C^2}{n^2} \times (2n\sigma^2 + 2\sigma^2\sum_{i=1}^{n}\mathscr{X}_i^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}_i)\end{aligned} \tag{C.2.40}$$

We define a random vector $(\varepsilon_1^*, r_1^*) \in \mathbf{R}^2$ having probability mass $1/n$ on $(\widehat{\varepsilon}_i, \widehat{r}_i), i = 1,2,...,n$. We generate i.i.d. random variables $(\varepsilon_i^*, r_i^*), i = 1,2,...,n$ and (ε_f^*, ξ^*) with the same distribution as

(ε_1^*, r_1^*). We denote $\varepsilon^* = (\varepsilon_1^*, ..., \varepsilon_n^*)^T$ and $r^* = (r_1^*, ..., r_n^*)^T$. For any given $0 < r < s < \infty, \xi > 0$, we choose $\delta = C/n^{3/4}$ in (C.1.3) with C a constant. Then define

$$\mathscr{G}^*(x) = \mathbf{P}^*\left(|\mathscr{Y}_f^* - \mathscr{X}_f^T \widehat{\beta}^*| \le x\right),\ x \in \mathbf{R} \tag{C.2.41}$$

here we choose $\widehat{\tau} = \widehat{r}$ in algorithm 5. In other words, $\mathscr{G}^*$ plays the same role as G^*, and the only difference is the mechanism for generating bootstrapped random variables.

$$\begin{aligned}
&\sup_{x\in[r,s]} |G^*(x) - \mathscr{G}^*(x)| \\
\le\ &\mathbf{P}^*\left(\left|\, |\varepsilon_f^* - \mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon^*| - |r_f^* - \mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T r^*| \,\right| > \frac{C}{n^{3/4}}\right) \\
&+ \sup_{x\in[r,s]} \mathbf{P}^*\left(x - \frac{C}{n^{3/4}} < |\varepsilon_f^* - \mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T\varepsilon^*| \le x + \frac{C}{n^{3/4}}\right) \\
&\le \frac{4\sqrt{n}}{C^2}\sum_{i=1}^n (\widehat{\varepsilon}_i - \widehat{r}_i)^2 \\
+\ &\frac{4\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X}/n)^{-1}\mathscr{X}_f}{C^2\sqrt{n}} \times \sum_{i=1}^n (\widehat{\varepsilon}_i - \widehat{r}_i)^2 + \sup_{x\in[r,s]}\left(G^*(x + \frac{C}{n^{3/4}}) - G^*(x - \frac{C}{n^{3/4}})\right)
\end{aligned} \tag{C.2.42}$$

and for sufficiently large n,

$$\begin{aligned}
&\sup_{x\ge r} |\mathscr{G}^*(x) - (F(x) - F(-x))| \\
\le\ &\frac{4\sqrt{n}}{C^2}\sum_{i=1}^n (\widehat{\varepsilon}_i - \widehat{r}_i)^2 + \frac{4\mathscr{X}_f^T(\mathscr{X}^T\mathscr{X}/n)^{-1}\mathscr{X}_f}{C^2\sqrt{n}} \times \sum_{i=1}^n (\widehat{\varepsilon}_i - \widehat{r}_i)^2 \\
+\ &3\sup_{x>0} |G^*(x) - (F(x) - F(-x))| + \sup_{x\ge r}\left(F(x + \frac{C}{n^{3/4}}) - F(x - \frac{C}{n^{3/4}})\right) \\
&+ \sup_{x\ge r}\left(F\left(-x + \frac{C}{n^{3/4}}\right) - F\left(-x - \frac{C}{n^{3/4}}\right)\right)
\end{aligned} \tag{C.2.43}$$

Lemma C.2.3 and (C.2.29) imply $\lim_{n\to\infty} \mathbf{P}\left(\sqrt{n}\sup_{x\in[r,s]} |G^*(x) - \mathscr{G}^*(x)| > \xi\right) = 0$; and $\lim_{n\to\infty} \mathbf{P}\left(\sup_{x\ge r} |\mathscr{G}^*(x) - (F(x) - F(-x))| > \xi\right) = 0$.

We define $\widehat{\mathscr{F}}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\widehat{r}_i \le x}$, and $\widehat{\alpha}(x)$ as in lemma C.1.2. For any given $-\infty < r < s < \infty, \xi > 0$, and sufficiently large n, lemma C.1.2 implies

$$\begin{aligned}
&\sup_{x\in[r,s]} |\widehat{\mathscr{F}}(x) - \widehat{F}(x)| \le \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{|\widehat{r}_i - \widehat{\varepsilon}_i| > \frac{C}{n^{3/4}}} + \sup_{x\in[r,s]} \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{x-\frac{C}{n^{3/4}} < \widehat{\varepsilon}_i \le x+\frac{C}{n^{3/4}}} \\
&\Rightarrow \mathbf{P}\left(\sqrt{n}\sup_{x\in[r,s]} |\widehat{\mathscr{F}}(x) - \widehat{F}(x)| > \xi\right) \le \frac{2}{\sqrt{n}\xi}\sum_{i=1}^{n} \mathbf{P}\left(|\widehat{\varepsilon}_i - \widehat{r}_i| > \frac{C}{n^{3/4}}\right) \\
&\qquad + \mathbf{P}\left(\sup_{x\in[r-1,s+1]} |\widehat{\alpha}(x) - \widehat{\alpha}(x - \frac{2C}{n^{3/4}})| > \frac{\xi}{4}\right) \\
&+\mathbf{P}\left(\sup_{x\in[r-1,s+1]} F^{'}(x) \times \frac{2C}{n^{1/4}} > \frac{\xi}{4}\right) \Rightarrow \lim_{n\to\infty} \mathbf{P}\left(\sqrt{n}\sup_{x\in[r,s]} |\widehat{\mathscr{F}}(x) - \widehat{F}(x)| > \xi\right) = 0
\end{aligned} \tag{C.2.44}$$

Here C is an arbitrary large positive constant.

We define $\widehat{\mathscr{M}}$ and $\widehat{\mathscr{S}}$ as in (4.27); define $\Lambda(z) = \mathscr{X}_f^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}^T z - \frac{1}{n}\sum_{i=1}^{n} z_i, \forall z = (z_1,...,z_n)^T \in \mathbf{R}^n$; and define

$$\widehat{\mathscr{N}}(x) = \sqrt{n}\widehat{\mathscr{F}}(x + \Lambda(u^*)) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathbf{1}_{u_i^* \le x}, \quad \widehat{\mathscr{T}}(x) = \widehat{\mathscr{N}}(x) - \widehat{\mathscr{N}^-}(-x) \tag{C.2.45}$$

Here $u^* = (u_1^*,...,u_n^*)^T$ are i.i.d. random variables generated by drawing from $\widehat{r}$ with replacement.

For any given $0 < r < s < \infty$ and $\xi > 0$,

$$\begin{aligned}
\mathbf{P}^*\left(\sup_{x\in[r,s]}|\widehat{\mathscr{S}}(x)-\widehat{\mathscr{T}}(x)|>4\xi\right) &\le \mathbf{P}^*\left(\sup_{x\in[r,s]}|\widehat{\mathscr{M}}(x)-\widehat{\mathscr{N}}(x)|>2\xi\right)\\
&\quad+\mathbf{P}^*\left(\sup_{x\in[r,s]}|\widehat{\mathscr{M}^-}(-x)-\widehat{\mathscr{N}^-}(-x)|>2\xi\right)\\
&\le \mathbf{P}^*\left(\sup_{x\in[r,s]}\sqrt{n}|\widehat{F}(x+\Lambda(\varepsilon^*))-\widehat{\mathscr{F}}(x+\Lambda(u^*))|>\xi\right)\\
&\quad+\mathbf{P}^*\left(\sup_{x\in[r,s]}|\frac{1}{\sqrt{n}}\sum_{i=1}^n\mathbf{1}_{e_i^*\le x}-\frac{1}{\sqrt{n}}\sum_{i=1}^n\mathbf{1}_{u_i^*\le x}|>\xi\right)\\
&\quad+\mathbf{P}^*\left(\sup_{x\in[r/2,s+1]}\sqrt{n}|\widehat{F}(-x+\Lambda(\varepsilon^*))-\widehat{\mathscr{F}}(-x+\Lambda(u^*))|>\xi\right)\\
&\quad+\mathbf{P}^*\left(\sup_{x\in[r/2,s+1]}|\frac{1}{\sqrt{n}}\sum_{i=1}^n\mathbf{1}_{e_i^*\le -x}-\frac{1}{\sqrt{n}}\sum_{i=1}^n\mathbf{1}_{u_i^*\le -x}|>\xi\right)
\end{aligned} \tag{C.2.46}$$

If $\sup_{x\in[-s-2,s+2]}\sqrt{n}|\widehat{\mathscr{F}}(x)-\widehat{F}(x)|<\xi/4$ and $\sup_{x,y\in[-s-2,s+2],|x-y|<\delta}|\widehat{\alpha}(x)-\widehat{\alpha}(y)|\le\xi/8$

with $0<\delta<1/8$,

$$
\begin{aligned}
&\mathbf{P}^*\left(\sup_{x\in[r,s]}\sqrt{n}|\widehat{F}(x+\Lambda(\varepsilon^*))-\widehat{\mathscr{F}}(x+\Lambda(u^*))|>\xi\right),\\
&\mathbf{P}^*\left(\sup_{x\in[r/2,s+1]}\sqrt{n}|\widehat{F}(-x+\Lambda(\varepsilon^*))-\widehat{\mathscr{F}}(-x+\Lambda(u^*))|>\xi\right)\\
&\leq\mathbf{P}^*\left(\sup_{x\in[-s-1,s+1]}\sqrt{n}|\widehat{F}(x+\Lambda(\varepsilon^*))-\widehat{F}(x+\Lambda(u^*))|>\xi/2\right)\\
&+\mathbf{P}^*\left(\sup_{x\in[-s-1,s+1]}\sqrt{n}|\widehat{F}(x+\Lambda(u^*))-\widehat{\mathscr{F}}(x+\Lambda(u^*))|>\xi/2\right)\\
&\leq\mathbf{P}^*\left(\sup_{x\in[-s-1,s+1]}|\widehat{\alpha}(x+\Lambda(\varepsilon^*))-\widehat{\alpha}(x+\Lambda(u^*))|>\xi/4\right)\\
&+\mathbf{P}^*\left(\sup_{x\in[-s-1,s+1]}\sqrt{n}|F(x+\Lambda(\varepsilon^*))-F(x+\Lambda(u^*))|>\xi/4\right)\\
&+\mathbf{P}^*\left(|\Lambda(u^*)|>1\right)\leq 2\mathbf{P}^*(|\Lambda(\varepsilon^*)|>\delta/4)+3\mathbf{P}^*(|\Lambda(u^*)|>\delta/4)\\
&+\mathbf{P}^*\left(\sup_{x\in[-s-2,s+2]}\sqrt{n}|F^{'}(x)|\times|\Lambda(\varepsilon^*)-\Lambda(u^*)|>\xi/4\right)
\end{aligned}
\tag{C.2.47}
$$

For

$$
\begin{aligned}
&\mathbf{E}^*\Lambda(\varepsilon^*)^2\leq\frac{2\widehat{\sigma}^2}{n}\left(\mathscr{X}_f^T\left(\frac{\mathscr{X}^T\mathscr{X}}{n}\right)^{-1}\mathscr{X}_f+1\right)\\
\text{and }&\mathbf{E}^*(\Lambda(u^*)-\Lambda(\varepsilon^*))^2\leq\frac{2}{n}\left(\mathscr{X}_f^T\left(\frac{\mathscr{X}^T\mathscr{X}}{n}\right)^{-1}\mathscr{X}_f+1\right)\times\frac{1}{n}\sum_{i=1}^{n}(\widehat{\varepsilon}_i-\widehat{r}_i)^2
\end{aligned}
\tag{C.2.48}
$$

(C.1.5), (C.2.40), (C.2.44) and lemma C.1.2 imply
$\lim_{n\to\infty}\mathbf{P}\left(\mathbf{P}^*\left(\sup_{x\in[r,s]}\sqrt{n}|\widehat{F}(x+\Lambda(\varepsilon^*))-\widehat{\mathscr{F}}(x+\Lambda(u^*))|>\xi\right)>\xi\right)=0$;
and $\lim_{n\to\infty}\mathbf{P}\left(\mathbf{P}^*\left(\sup_{x\in[r/2,s+1]}\sqrt{n}|\widehat{F}(-x+\Lambda(\varepsilon^*))-\widehat{\mathscr{F}}(-x+\Lambda(u^*))|>\xi\right)>\xi\right)=0$. On

the other hand, we define $\widetilde{\alpha}^*$ as in lemma C.1.2; from (C.1.3), for any $0 < \delta < 1/4$,

$$\begin{aligned}
&\mathbf{P}^*\left(\sup_{x\in[r,s]}|\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{1}_{e_i^*\le x} - \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{1}_{u_i^*\le x}| > \xi\right),\\
&\mathbf{P}^*\left(\sup_{x\in[r/2,s+1]}|\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{1}_{e_i^*\le -x} - \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{1}_{u_i^*\le -x}| > \xi\right)\\
&\le \mathbf{P}^*\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{1}_{|e_i^*-u_i^*|>\frac{\delta}{\sqrt{n}}} > \xi/2\right)\\
&+\mathbf{P}^*\left(\sup_{x\in[-s-1,s+1]}\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{1}_{x-\frac{\delta}{\sqrt{n}}<e_i^*\le x+\frac{\delta}{\sqrt{n}}} > \xi/2\right)\\
\le\ &\frac{2\sqrt{n}}{\xi}\mathbf{P}^*\left(|e_1^*-u_1^*| > \frac{\delta}{\sqrt{n}}\right) + \mathbf{P}^*\left(\sup_{x\in[-s-2,s+2]}|\widetilde{\alpha}^*(x) - \widetilde{\alpha}^*(x-\frac{2\delta}{\sqrt{n}})| > \xi/4\right)\\
&+\mathbf{P}^*\left(\sup_{x\in[-s-1,s+1]}\sqrt{n}|\widehat{F}(x) - \widehat{F}(x-\frac{2\delta}{\sqrt{n}})| > \xi/4\right)
\end{aligned} \tag{C.2.49}$$

Since $\mathbf{P}^*\left(|e_1^*-u_1^*| > \frac{\delta}{\sqrt{n}}\right) \le \frac{\sum_{i=1}^n(\widehat{\varepsilon}_i-\widehat{r}_i)^2}{\delta^2}$, (C.2.40) and lemma C.1.2 imply
$\lim_{n\to\infty}\mathbf{P}\left(\mathbf{P}^*\left(\sup_{x\in[r,s]}|\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{1}_{e_i^*\le x} - \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{1}_{u_i^*\le x}| > \xi\right) > \xi\right) = 0;$
and $\lim_{n\to\infty}\mathbf{P}\left(\mathbf{P}^*\left(\sup_{x\in[r/2,s+1]}|\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{1}_{e_i^*\le -x} - \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{1}_{u_i^*\le -x}| > \xi\right) > \xi\right) = 0.$ In particular, $\forall \xi > 0$,

$$\lim_{n\to\infty}\mathbf{P}\left(\mathbf{P}^*\left(\sup_{x\in[r,s]}|\widehat{\mathscr{S}}(x) - \widehat{\mathscr{T}}(x)| > \xi\right) > \xi\right) = 0 \tag{C.2.50}$$

For $\forall \xi > 0$,

$$\begin{aligned}
\sup_{x\in[r,s],y\in\mathbf{R}}|\mathbf{P}^*\left(\widehat{\mathscr{T}}(x)\le y\right) - \Phi\left(\frac{y}{\sqrt{\mathscr{U}(x)}}\right)| &\le \mathbf{P}^*\left(\sup_{x\in[r,s]}|\widehat{\mathscr{S}}(x) - \widehat{\mathscr{T}}(x)| > \xi\right)\\
&+3\sup_{x\in[r,s],y\in\mathbf{R}}|\mathbf{P}^*\left(\widehat{\mathscr{S}}(x)\le y\right) - \Phi\left(\frac{y}{\sqrt{\mathscr{U}(x)}}\right)|\\
&+\sup_{x\in[r,s],y\in\mathbf{R}}\left(\Phi\left(\frac{y+\xi}{\sqrt{\mathscr{U}(x)}}\right) - \Phi\left(\frac{y-\xi}{\sqrt{\mathscr{U}(x)}}\right)\right)
\end{aligned} \tag{C.2.51}$$

Lemma C.2.2 implies $\lim_{n\to\infty}\mathbf{P}\left(\sup_{x\in[r,s],y\in\mathbf{R}}|\mathbf{P}^*\left(\widehat{\mathscr{T}}(x)\le y\right) - \Phi\left(\frac{y}{\sqrt{\mathscr{U}(x)}}\right)| > \xi\right) = 0.$

Suppose $\frac{1}{8}\min(\alpha,1-\alpha)>\xi>0$. From (C.2.5) and (C.2.51), with probability tending to 1, $\forall 2\xi<1-\gamma<1-\xi, r\le x\le s$, $\sqrt{\mathscr{U}(x)}\times\Phi^{-1}(1-\gamma-2\xi)\le D^*_{1-\gamma}(x)\le\sqrt{\mathscr{U}(x)}\times\Phi^{-1}(1-\gamma+\xi)$. We define $c_z, z\in(0,1)$ and $\underline{d},\overline{d}$ as in the proof of theorem 10. We choose $r=c_{(1-\alpha)/8}>0$ in (C.2.43), with probability tending to 1, $c_{\tau-2\xi}\le C^*_\tau\le c_{\tau+\xi}, \forall(1-\alpha)/8+2\xi<\tau<1-2\xi$. In particular, this implies $c_{1-\alpha-2\xi}\le C^*_{1-\alpha}\le c_{1-\alpha+\xi}$, and $\underline{d}\le D^*_{1-\gamma}(C^*_{1-\alpha})\le\overline{d}$. We choose $r=c_{(1-\alpha)/8}$ and $s=c_{1-\alpha+4\xi}$ in (C.2.42) and lemma C.2.3, $C^*(1-\alpha,1-\gamma)\le C^*_{1-\alpha+\frac{\overline{d}}{\sqrt{n}}}\le c^*_{1-\alpha+\frac{\overline{d}+2\xi}{\sqrt{n}}}$; and $C^*(1-\alpha,1-\gamma)\ge C^*_{1-\alpha+\frac{\underline{d}}{\sqrt{n}}}\ge c^*_{1-\alpha+\frac{\underline{d}-3\xi}{\sqrt{n}}}$. We define $\mathscr{S}$ and $\mathscr{U}$ as in (4.19) and (4.15), since

$$\begin{aligned}
&|\sqrt{n}\left(\mathbf{P}^*\left(|\mathscr{Y}_f-\mathscr{X}_f^T\widehat{\beta}|\le C^*(1-\alpha,1-\gamma)\right)-(1-\alpha)\right)\\
&\qquad-\left(\mathscr{S}(c_{1-\alpha})+\sqrt{\mathscr{U}(c_{1-\alpha})}\times\Phi^{-1}(1-\gamma)\right)|\\
&\le|\sqrt{n}\left(\mathbf{P}^*\left(|\mathscr{Y}_f-\mathscr{X}_f^T\widehat{\beta}|\le c^*_{1-\alpha+\frac{\overline{d}+2\xi}{\sqrt{n}}}\right)-(1-\alpha)\right)\\
&\qquad-\left(\mathscr{S}(c_{1-\alpha})+\sqrt{\mathscr{U}(c_{1-\alpha})}\times\Phi^{-1}(1-\gamma)\right)|\\
&+|\sqrt{n}\left(\mathbf{P}^*\left(|\mathscr{Y}_f-\mathscr{X}_f^T\widehat{\beta}|\le c^*_{1-\alpha+\frac{\underline{d}-3\xi}{\sqrt{n}}}\right)-(1-\alpha)\right)\\
&\qquad-\left(\mathscr{S}(c_{1-\alpha})+\sqrt{\mathscr{U}(c_{1-\alpha})}\times\Phi^{-1}(1-\gamma)\right)|
\end{aligned}\tag{C.2.52}$$

Replace $c^*(1-\alpha,1-\gamma)$ in (C.2.34) to (C.2.36) by $c^*_{1-\alpha+\frac{\overline{d}+2\xi}{\sqrt{n}}}$ and $c^*_{1-\alpha+\frac{\underline{d}-3\xi}{\sqrt{n}}}$, and set $\xi\to0$, we prove (4.31). □

C.3 Results used in the paper

This paper uses many results from the stochastic process and some results from the optimal transport. Statisticians may not be familiar with them. To make the paper self-contained, this section quotes the frequently used theorems from textbooks and papers. However, we cannot explain the background of each theorem in detail. So we encourage the readers to look through those materials if possible.

Lemma C.3.1 (theorem 13.5, Billingsley [1999]). *Suppose that*

$$(X_{t_1}^n, ..., X_{t_k}^n) \to_{\mathscr{L}} (X_{t_1}, ..., X_{t_k}) \tag{C.3.1}$$

for any points t_i; that

$$X_1 - X_{1-\delta} \to_{\mathscr{L}} 0 \text{ as } \delta \to 0, \delta > 0 \tag{C.3.2}$$

and that for any $r \leq s \leq t,\ n \geq 1,\ \lambda > 0$,

$$\mathbf{P}[|X_s^n - X_r^n| \wedge |X_t^n - X_s^n| \geq \lambda] \leq \frac{1}{\lambda^{4\beta}}[F(t) - F(r)]^{2\alpha} \tag{C.3.3}$$

where $\beta \geq 0$, $\alpha > 1/2$ and F is a non-decreasing, continuous function on $[0,1]$. Then $X^n \to_{\mathscr{L}} X$

Lemma C.3.2 (theorem 13.6, Billingsley [1999]). *There exists in* **D***(see section 4.4) a random element with finite-dimensional distributions $\mu_{t_1 \ldots t_k}$, provided these distributions are consistent(i.e., satisfy the consistency conditions of Kolmogorov's existence theorem); provided that, for $t_1 \leq t \leq t_2$,*

$$\mu_{t_1 t t_2}[(u_1, u, u_2) : |u - u_1| \wedge |u_2 - u| \geq \lambda] \leq \frac{1}{\lambda^{4\beta}}(F(t_2) - F(t_1))^{2\alpha} \tag{C.3.4}$$

where $\beta \geq 0$, $\alpha > 1/2$, and F is a non-decreasing, continuous function on $[0,1]$; and provided that

$$\lim_{h \to 0, h > 0} \mu_{t,t+h}[(u_1, u_2) : |u_2 - u_1| \geq \varepsilon] = 0,\ 0 \leq t \leq 1 \tag{C.3.5}$$

Lemma C.3.3 (theorem 2.3 in Hahn [1977]). *Let f be a nonnegative function on $[0,1]$ which is nondecreasing in a neighborhood of 0. Let $X(t)$ be a stochastic process such that for some $r \geq 1$, $\mathbf{E}|X(t) - X(s)|^r \leq f(|t-s|)$. If*

$$\int_0 y^{-(r+1)/r} f^{1/r}(y) dy < \infty \tag{C.3.6}$$

then there exists a nondecreasing functin ϕ on $[0,1]$ with $\phi(0)=0$, which depends only on f, and a random variable A such that $\mathbf{E}|A|^r < \infty$ and

$$|\widetilde{X}(s)-\widetilde{X}(t)| \leq A\phi(|t-s|) \tag{C.3.7}$$

Moreover, $\|A\|_r$ is bounded above by a constant depending only on f and ϕ. Here $\widetilde{X}$ is a separable version of X.

Lemma C.3.4 (3.8, page 348 in Jacod and Shiryaev [2003])**.** *Assume that $X^n \to_{\mathscr{L}} X$ and that $\mathbf{P}(X \in C)=1$, where C is the continuity set of the function $h: E \to E^{'}$. Then*

i. If $E^{'}=\mathbf{R}$ and h is bounded, then $\mathbf{E}h(X^n) \to \mathbf{E}h(X)$;

ii. If $E^{'}$ is Polish, then $h(X^n) \to_{\mathscr{L}} h(X)$.

Lemma C.3.5 (theorem 3.1 in Ranga Rao [1962])**.** *Let $\mathscr{A}$ be a class of continuous functions possessing the following properties: 1. $\mathscr{A}$ is uniformly bounded, i.e., $\exists$ a constant $M>0$ such that $|f(x)| \leq M$ for all $f \in \mathscr{A}$ and all x; 2. $\mathscr{A}$ is equi-continuous. If μ_n, μ satisfies $\mu_n \to_{\mathscr{L}} \mu$, then*

$$\lim_{n\to\infty} \sup_{f\in\mathscr{A}} |\int f d\mu_n - \int f d\mu| = 0 \tag{C.3.8}$$

Lemma C.3.6 (theorem 6.2.1 in Koul [2002])**.** *Suppose that the model $\mathscr{Y} = \mathscr{X}\beta + \varepsilon$ holds true. In addition suppose $(\mathscr{X}^T\mathscr{X})^{-1}$ exists, $\max_{i=1,\ldots,n} \mathscr{X}_i^T(\mathscr{X}^T\mathscr{X})^{-1}\mathscr{X}_i = o(1)$ and F has uniform continuous density f. Suppose $\widehat{\beta}$ is an estimator of β satisfying*

$$|A^{-1}(\widehat{\beta}-\beta)|_2 = O_p(1) \tag{C.3.9}$$

then

$$\sup_{t\in[0,1]} |W_1(t,\widehat{\beta}) - W_1(t,\beta) - q_0(t)\sqrt{n} \times \overline{\mathscr{X}}_n^T A A^{-1}(\widehat{\beta}-\beta)| = o_p(1) \tag{C.3.10}$$

here $q_0(t) = f(F^{-1}(t))$, $W_1(t,s) = \sqrt{n}(H_n(F^{-1}(t),s)-t)$, $H_n(y,s) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\mathscr{Y}_i \leq y + \mathscr{X}_i^T s}$ and $A = (\mathscr{X}^T\mathscr{X})^{-1/2}$. $|\cdot|_2$ is the vector 2 norm in the Euclidean space.

Lemma C.3.7 (theorem 6.9, Villani [2009]). *Let $(\mathscr{X},d)$ be a Polish space, and $p \in [1,\infty)$. Define $P_p(\mathscr{X})$ as the Borel probability measure on $\mathscr{X}$ with finite moments of order p. Then the Wasserstein distance W_p metrizes the weak convergence. In other words, if $(\mu_k)_{k\in\mathbf{N}} \subset P_p(\mathscr{X})$ is a sequence of measures and $\mu \in P(\mathscr{X})$ is another Borel probability measure on $\mathscr{X}$, then the statement μ_k converges weakly in $P_p(\mathscr{X})$ to μ and $W_p(\mu_k,\mu) \to 0$ are equivalent. Here $W_p(\mu_k,\mu)$ is the Wasserstein distance(see lemma C.2.1). The weakly convergence in $P_p(\mathscr{X})$ means $\exists x_0 \in \mathscr{X}$ such that*

$$\mu_k \to_{\mathscr{L}} \mu \text{ and } \int d(x_0,x)^p d\mu_k \to \int d(x_0,x)^p d\mu \tag{C.3.11}$$

Bibliography

S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40 – 79, 2010.

Z. D. Bai and Y. Q. Yin. Limit of the smallest eigenvalue of a large-dimensional sample covariance matrix. *Ann. Probab.*, 21(3):1275–1294, 1993.

R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani. The limits of distribution-free conditional predictive inference. *Inf. Inference*, 10(2):455–482, 2021.

S. Basu and G. Michailidis. Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.*, 43(4):1535 – 1567, 2015.

S. Bates, T. Hastie, and R. Tibshirani. Cross-validation: what does it estimate and how well does it do it? *arXiv e-prints*, art. arXiv:2104.00673, Apr. 2021.

R. Beran. Calibrating prediction regions. *J. Amer. Statist. Assoc.*, 85(411):715–723, 1990.

P. J. Bickel and D. A. Freedman. Some asymptotic theory for the bootstrap. *Ann. Statist.*, 9(6): 1196–1217, 1981.

P. J. Bickel, Y. acov Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705 – 1732, 2009.

P. Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, Inc., New York, second edition, 1999. ISBN 0-471-19745-9. A Wiley-Interscience Publication.

P. J. Brockwell and R. A. Davis. *Time series: theory and methods*. Springer Series in Statistics. Springer-Verlag, New York, second edition, 1991. ISBN 0-387-97429-6.

P. Bühlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4): 1212–1242, 2013.

P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011. ISBN 978-3-642-20191-2.

D. L. Burkholder, B. J. Davis, and R. F. Gundy. Integral inequalities for convex functions of operators on martingales. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, pages 223–240, Berkeley, Calif., 1972. University of California Press.

E. Çinlar. *Probability and Stochastics*. Springer-Verlag New York, 1 edition, 2011. ISBN 978-0-387-87858-4.

M. Celentano, A. Montanari, and Y. Wei. The lasso with general gaussian designs with applications to hypothesis testing. *arXiv preprint arXiv:2007.13716*, 2020.

A. Chatterjee and S. N. Lahiri. Asymptotic properties of the residual bootstrap for Lasso estimators. *Proc. Amer. Math. Soc.*, 138(12):4497–4509, 2010.

A. Chatterjee and S. N. Lahiri. Bootstrapping lasso estimators. *J. Amer. Statist. Assoc.*, 106(494): 608–625, 2011.

X. Chen and W.-X. Zhou. Robust inference via multiplier bootstrap. *Ann. Statist.*, 48(3): 1665–1691, 06 2020.

Z. Chen, J. Fan, and R. Li. Error variance estimation in ultrahigh-dimensional additive models. *J. Amer. Statist. Assoc.*, 113(521):315–327, 2018.

G. Cheng and J. Z. Huang. Bootstrap consistency for general semiparametric M-estimation. *Ann. Statist.*, 38(5):2884–2915, 2010.

V. Chernozhukov, D. Chetverikov, and K. Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.*, 41(6): 2786–2819, 2013.

V. Chernozhukov, D. Chetverikov, and K. Kato. Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probab. Theory Related Fields*, 162(1-2):47–70, 2015.

V. Chernozhukov, K. Wüthrich, and Y. Zhu. Distributional conformal prediction. *Proc. Natl. Acad. Sci. USA*, 118(48):Paper No. e2107794118, 9, 2021.

T. Conley, S. Gonçalves, M. S. Kim, and B. Perron. Bootstrap inference under cross sectional dependence. *unpublished*, 2019.

R. Dahlhaus. Fitting time series models to nonstationary processes. *Ann. Statist.*, 25(1):1–37, 1997.

R. Dahlhaus and S. Subba Rao. Statistical inference for time-varying ARCH processes. *Ann. Statist.*, 34(3):1075–1114, 2006.

R. Dahlhaus, S. Richter, and W. B. Wu. Towards a general theory for nonlinear locally stationary processes. *Bernoulli*, 25(2):1013–1044, 2019.

L. Dai, K. Chen, Z. Sun, Z. Liu, and G. Li. Broken adaptive ridge regression and its asymptotic properties. *J. Multivariate Anal.*, 168:334 – 351, 2018.

A. S. Dalalyan, M. Hebiri, and J. Lederer. On the prediction performance of the Lasso. *Bernoulli*, 23(1):552 – 581, 2017.

S. Das and D. N. Politis. Predictive inference for locally stationary time series with an application to climate data. *J. Amer. Statist. Assoc.*, 116(534):919–934, 2021.

S. De Gryze, I. Langhans, and M. Vandebroek. Using the correct intervals for prediction: A tutorial on tolerance intervals for ordinary least-squares regression. *Chemometrics and Intelligent Laboratory Systems*, 87(2):147–154, 2007.

H. Dette and W. Wu. Prediction in Locally Stationary Time Series. *J. Bus. Econom. Statist.*, 40 (1):370–381, 2022.

R. Dezeure, P. Bühlmann, and C.-H. Zhang. High-dimensional simultaneous inference with the bootstrap. *TEST*, 26(4):685–719, Dec 2017. ISSN 1863-8260. doi: 10.1007/s11749-017-0554-2. URL https://doi.org/10.1007/s11749-017-0554-2.

E. Dobriban and S. Wager. High-dimensional asymptotics of prediction: ridge regression and classification. *Ann. Statist.*, 46(1):247–279, 2018.

J. J. Dolado and H. Lütkepohl. Making wald tests work for cointegrated var systems. *Econometric Rev.*, 15(4):369–386, 1996.

J. Fan. Statistical foundations of data science. https://fan.princeton.edu/DataScience/, Accessed: 2022.

J. Fan and I. Gijbels. Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Roy. Statist. Soc. Ser. B*, 57(2):371–394, 1995.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360, 2001.

J. Fan, R. Li, C.-H. Zhang, and H. Zou. *Statistical Foundations of Data Science*. Chapman&Hall/CRC /Taylor & Francis Group, first edition, 2020. ISBN 9781466510845.

W. Fithian, D. Sun, and J. Taylor. Optimal inference after model selection. 2017. Preprint. arXiv:1410.2597.

G. B. Folland. *Real analysis*. Pure and Applied Mathematics (New York). John Wiley & Sons, Inc., New York, second edition, 1999. ISBN 0-471-31716-0. Modern techniques and their applications, A Wiley-Interscience Publication.

D. A. Freedman. Bootstrapping regression models. *Ann. Statist.*, 9(6):1218–1228, 1981.

S. Geisser. *Predictive inference*, volume 55 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, New York, 1993. ISBN 0-412-03471-9.

L. Giraitis, G. Kapetanios, and T. Yates. Inference on stochastic time-varying coefficient models. *J. Econometrics*, 179(1):46–65, 2014.

S. Gonçalves and T. J. Vogelsang. Block bootstrap hac robust tests: the sophistication of the naive bootstrap. *Econometric Theory*, 27(4):745 – 791, 2011. doi: 10.1017/S0266466610000496.

E. Greenshtein and Y. Ritov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988, 2004.

M. G. Hahn. Conditions for sample-continuity and the central limit theorem. *Ann. Probability*, 5 (3):351–360, 1977.

P. Hall and L.-S. Huang. Nonparametric kernel regression subject to monotonicity constraints. *Ann. Statist.*, 29(3):624–647, 2001.

Y. Han and R. S. Tsay. High-dimensional linear regression for dependent data with applications to nowcasting. *Statist. Sinica*, 30(4):1797–1827, 2020.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, second edition, 2009. ISBN 978-0-387-84857-0. Data mining, inference, and prediction.

T. Hayfield and J. S. Racine. Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5), 2008.

R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, second edition, 2013. ISBN 978-0-521-54823-6.

J. Jacod and A. N. Shiryaev. *Limit theorems for stochastic processes*, volume 288 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, second edition, 2003. ISBN 3-540-43932-3.

A. Javanmard and H. Javadi. False discovery rate control via debiased lasso. *Electron. J. Statist.*, 13(1):1212 – 1253, 2019.

A. Javanmard and A. Montanari. Debiasing the lasso: Optimal sample size for gaussian designs. *Ann. Statist.*, (6A):2593–2622, 12 .

H. H. Kelejian and I. R. Prucha. Hac estimation in a spatial framework. *Journal of Econometrics*, 140(1):131–154, 2007.

M. S. Kim and Y. Sun. Spatial heteroskedasticity and autocorrelation consistent estimation of covariance matrix. *J. Econometrics*, 160(2):349–371, 2011.

H. L. Koul. *Weighted empirical processes in dynamic nonlinear models*, volume 166 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 2002. ISBN 0-387-95476-7. Second edition of ıt Weighted empiricals and linear models [Inst. Math. Statist., Hayward, CA, 1992; MR1218395 (95c:62061)].

J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor. Exact post-selection inference, with application to the lasso. *Ann. Statist.*, 44(3):907–927, 2016.

J. Lei and L. Wasserman. Distribution-free prediction bands for non-parametric regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 76(1):71–96, 2014.

J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman. Distribution-free predictive inference for regression. *J. Amer. Statist. Assoc.*, 113(523):1094–1111, 2018.

G. J. Lieberman and R. G. Miller, Jr. Simultaneous tolerance intervals in regression. *Biometrika*, 50:155–168, 1963.

H. Liu and B. Yu. Asymptotic properties of Lasso+mLS and Lasso+Ridge in sparse high-dimensional linear regression. *Electron. J. Stat.*, 7:3124–3169, 2013.

R. Y. Liu. Bootstrap Procedures under some Non-I.I.D. Models. *Ann. Statist.*, 16(4):1696 – 1708, 1988.

W.-Y. Loh. Calibrating confidence coefficients. *J. Amer. Statist. Assoc.*, 82(397):155–162, 1987.

W.-Y. Loh. Bootstrap calibration for confidence interval construction and selection. *Statist. Sinica*, 1(2):477–491, 1991.

M. Lopes. A residual bootstrap for high-dimensional regression with near low-rank designs. In *Advances in Neural Information Processing Systems 27*, pages 3239–3247, 2014.

E. Mammen. Bootstrap and Wild Bootstrap for High Dimensional Linear Models. *Ann. Statist.*,

21(1):255 – 285, 1993.

E. Mammen. Empirical process of residuals for high-dimensional linear models. *Ann. Statist.*, 24(1):307 – 335, 1996.

T. S. McElroy and D. N. Politis. *Time Series: A First Course with Bootstrap Starter*. CRC Press, first edition, 2020. ISBN 978-1-4398-7561-0.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 2006.

N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.*, 37(1):246–270, 2009.

B. Øksendal. *Stochastic Differential Equations*. Springer-Verlag Berlin Heidelberg, 2003. ISBN 978-3-642-14394-6.

L. Pan and D. N. Politis. Bootstrap prediction intervals for linear, nonlinear and nonparametric autoregressions. *Journal of Statistical Planning and Inference*, 177:1 – 27, 2016a.

L. Pan and D. N. Politis. Bootstrap prediction intervals for Markov processes. *Comput. Statist. Data Anal.*, 100:467–494, 2016b.

E. Paparoditis and D. N. Politis. The local bootstrap for kernel estimators under general dependence conditions. *Ann. Inst. Statist. Math.*, 52(1):139–159, 2000.

M. A. Petersen. Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches. *The Review of Financial Studies*, 22(1):435–480, 06 2008.

D. N. Politis. Adaptive bandwidth choice. *J. Nonparametr. Stat.*, 15(4-5):517–533, 2003.

D. N. Politis. Model-free model-fitting and predictive distributions. *TEST*, 22(2):183–221, 2013.

D. N. Politis. *Model-free prediction and regression*. Frontiers in Probability and the Statistical Sciences. Springer, Cham, 2015. ISBN 978-3-319-21346-0; 978-3-319-21347-7. A transformation-based approach to inference.

D. N. Politis and H. White. Automatic block-length selection for the dependent bootstrap. *Econometric Rev.*, 23(1):53–70, 2004.

D. N. Politis, J. P. Romano, and M. Wolf. *Subsampling*. Springer Series in Statistics. Springer-Verlag, New York, 1999. ISBN 0-387-98854-8.

M. B. Priestley. *Nonlinear and nonstationary time series analysis*. Academic Press, Inc.

[Harcourt Brace Jovanovich, Publishers], London, 1988. ISBN 0-12-564910-X.

R. Ranga Rao. Relations between weak and uniform convergence of measures with applications. *Ann. Math. Statist.*, 33:659–680, 1962.

G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated Gaussian designs. *J. Mach. Learn. Res.*, 11:2241–2259, 2010.

G. C. Reinsel. *Elements of multivariate time series analysis*. Springer Series in Statistics. Springer-Verlag, New York, 1993. ISBN 0-387-94063-4.

Y. Romano, E. Patterson, and E. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, pages 3543–3553. Curran Associates, Inc.

Y. Romano, M. Sesia, and E. Candès. Deep knockoffs. *J. Amer. Statist. Assoc.*, 115(532): 1861–1872, 2020.

G. A. F. Seber and A. J. Lee. *Linear regression analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2003. ISBN 0-471-41540-5.

M. Sesia and E. J. Candès. A comparison of some conformal quantile regression methods. *Stat*, 9(1):e261.

G. Shafer and V. Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

J. Shao. Bootstrap model selection. *J. Amer. Statist. Assoc.*, 91(434):655–665, 1996.

J. Shao. *Mathematical statistics*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2003. ISBN 0-387-95382-5.

J. Shao and X. Deng. Estimation in high-dimensional linear models with deterministic design matrices. *Ann. Statist.*, 40(2):812–831, 2012.

X. Shao. The dependent wild bootstrap. *J. Amer. Statist. Assoc.*, 105(489):218–235, 2010. With supplementary material available online.

L. Steinberger and H. Leeb. Leave-one-out prediction intervals in linear regression models with many variables. *arXiv preprint arXiv:1602.05801*, 2016.

R. A. Stine. Bootstrap prediction intervals for regression. *J. Amer. Statist. Assoc.*, 80(392): 1026–1031, 1985.

T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 09 2012.

Y. Sun. A heteroskedasticity and autocorrelation robust f test using an orthonormal series variance estimator. *Econom. J.*, 16(1):1–26.

Y. Sun. Robust trend inference with series variance estimator and testing-optimal smoothing parameter. *J. Econometrics*, 164(2):345 – 366, 2011.

Y. Sun and X. Wang. An asymptotically f-distributed chow test in the presence of heteroscedasticity and autocorrelation. *Econometric Reviews*, 0(0):1–35, 2021.

L. A. Thombs and W. R. Schucany. Bootstrap prediction intervals for autoregression. *J. Amer. Statist. Assoc.*, 85(410):486–492, 1990.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1): 267–288, 1996.

R. Tibshirani. Regression shrinkage and selection via the lasso: a retrospective. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 73(3):273–282, 2011.

R. Tibshirani, N. Jeliazkova, and E. LeDell. Conformal inference r project. https://github.com/ryantibs/conformal, 2021.

R. J. Tibshirani, A. Rinaldo, R. Tibshirani, and L. Wasserman. Uniform asymptotic inference and the bootstrap after model selection. *Ann. Statist.*, 46(3):1255–1287, 2018.

S. van de Geer. On the asymptotic variance of the debiased Lasso. *Electron. J. Statist.*, 13(2): 2970 – 3008, 2019.

S. van de Geer, P. Bühlmann, and S. Zhou. The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). *Electron. J. Statist.*, 5(none):688 – 749, 2011.

S. A. van de Geer. High-dimensional generalized linear models and the lasso. *Ann. Statist.*, 36 (2):614–645, 2008.

C. Villani. *Optimal transport*, volume 338 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. ISBN 978-3-540-71049-3.

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

V. Vovk. Conditional validity of inductive conformal predictors. In S. C. H. Hoi and W. Buntine,

editors, *Proceedings of the Asian Conference on Machine Learning*, volume 25 of *Proceedings of Machine Learning Research*, pages 475–490, Singapore Management University, Singapore, 04–06 Nov 2012. PMLR.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*. Springer, New York, 2005. ISBN 978-0387-00152-4; 0-387-00152-2.

W. A. Wallis. Tolerance intervals for linear regression. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*, pages 43–51. University of California Press, Berkeley and Los Angeles, Calif., 1951.

Y. Wang and D. N. Politis. Model-free bootstrap and conformal prediction in regression: Conditionality, conjecture testing, and pertinent prediction intervals. *arXiv preprint arXiv:2109.12156*, 2021.

H. White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.

P. Whittle. Bounds for the moments of linear and quadratic forms in independent variables. *Theory Probab. Appl.*, 5(3):302–305, 1960.

C. F. J. Wu. Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. *Ann. Statist.*, 14(4):1261 – 1295, 1986.

W. B. Wu. Nonlinear system theory: another look at dependence. *Proc. Natl. Acad. Sci. USA*, 102(40):14150–14154, 2005.

W.-B. Wu and Y. N. Wu. Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electron. J. Stat.*, 10(1):352–379, 2016.

W. B. Wu and Z. Zhou. Gaussian approximations for non-stationary multiple time series. *Statist. Sinica*, 21(3):1397–1413, 2011.

H. Xie and J. Huang. SCAD-penalized regression in high-dimensional partially linear models. *Ann. Statist.*, 37(2):673–696, 2009.

M. Xu, D. Zhang, and W. B. Wu. Pearson's chi-squared statistics: approximation theory and beyond. *Biometrika*, 106(3):716–723, 2019.

P. Yuan and X. Guo. High-dimensional inference for linear model with correlated errors. *Metrika*.

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 76(1):217–242, 2014.

D. Zhang and W. B. Wu. Gaussian approximation for high dimensional time series. *Ann. Statist.*, 45(5):1895–1919, 10 2017.

D. Zhang and W. B. Wu. Convergence of covariance and spectral density estimates for high-dimensional locally stationary processes. *Ann. Statist.*, 49(1):233–254, 2021.

X. Zhang and G. Cheng. Simultaneous inference for high-dimensional linear models. *J. Amer. Statist. Assoc.*, 112(518):757–768, 2017.

Y. Zhang and D. N. Politis. Ridge regression revisited: Debiasing, thresholding and bootstrap. *arXiv preprint arXiv:2009.08071*, 2020.

Y. Zhang and D. N. Politis. Bootstrap prediction intervals with asymptotic conditional validity and unconditional guarantees, 2021a. Preprint. arXiv:2005.09145.

Y. Zhang and D. N. Politis. Statistical inference on $ar(p)$ models with non-iid innovations. *arXiv preprint arXiv:2110.14067*, 2021b.

P. Zhao and B. Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006.

Z. Zhou. Inference of weighted V-statistics for nonstationary time series and its applications. *Ann. Statist.*, 42(1):87–114, 2014.

H. Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2):301–320, 2005.