UNIVERSITY OF CALIFORNIA SAN DIEGO

Robust Inference and Learning of Multivariate Statistical Models

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

in

Mathematics

by

Linbo Liu

Committee in charge:

Professor Danna Zhang, Chair
Professor Ery Arias-Castro
Professor Dimitris N. Politis
Professor Yixiao Sun
Professor Wenxin Zhou

2022

The Dissertation of Linbo Liu 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 loving parents.

EPIGRAPH

Life is a math equation.
In order to gain the most,
you have to know how to convert negatives into positives.

Anonymous

TABLE OF CONTENTS

Dissertation Approval Page . . . iii

Dedication . . . iv

Epigraph . . . v

Table of Contents . . . vi

List of Figures . . . ix

List of Tables . . . x

Acknowledgements . . . xiii

Vita . . . xv

Abstract of the Dissertation . . . xvi

Chapter 1 A Bernstein-type Inequality for High Dimensional Linear Processes . . . 1
- 1.1 Introduction . . . 1
- 1.2 Bernstein-type Inequality for High Dimensional Linear Processes . . . 4
- 1.3 Robust Estimation of Time Series Regression . . . 11
  - 1.3.1 Estimating Time Series Regression with Correlated Errors . . . 11
  - 1.3.2 Estimating Transition Matrix in VAR Models . . . 15
- 1.4 Simulation Study . . . 19
- 1.5 Concluding Remarks . . . 22
- 1.6 Proofs of Results in Section 1.2 . . . 23
- 1.7 Proofs of Results in Section 1.3 . . . 31
  - 1.7.1 Proofs of Results in Section 1.3.1 . . . 31
  - 1.7.2 Proofs of Results in Section 1.3.2 . . . 34

Chapter 2 Simultaneous Inference of High-dimensional non-Gaussian Vector Autoregressive Models . . . 44
- 2.1 Introduction . . . 44
- 2.2 Main Results . . . 48
  - 2.2.1 Vector autoregressive model . . . 48
  - 2.2.2 De-biased estimator . . . 50
  - 2.2.3 Estimation of the precision matrix . . . 53
  - 2.2.4 Simultaneous inference . . . 55
- 2.3 Estimation Consistency . . . 58
- 2.4 Gaussian Approximation . . . 62
- 2.5 Numerical Experiments . . . 65
- 2.6 Proofs of Results in Section 2.2 . . . 65

2.7 Proofs of Results in Section 2.3 . . . 74
2.8 Proofs of Result in Section 2.4 . . . 79

Chapter 3 Robust Multivariate Time-Series Forecasting: Adversarial Attacks and Defense Mechanisms . . . 85
3.1 Introduction . . . 85
3.2 Related Work . . . 89
3.3 Adversarial Attack Strategies . . . 90
3.3.1 Framework on Sparse and Indirect Adversarial Attack . . . 90
3.3.2 Deterministic Attack . . . 91
3.3.3 Probabilistic Attack . . . 92
3.4 Defense Mechanisms against Adversarial Attacks . . . 95
3.4.1 Randomized Smoothing Defense . . . 95
3.4.2 Mini-max Defense . . . 97
3.5 Experiments . . . 98
3.5.1 Experiment Setups . . . 99
3.5.2 Experiment Results . . . 100
3.5.3 Non-transferrablity between Univariate and Multivariate Cases . . . 101
3.6 Probabilistic Attack Algorithm . . . 103
3.7 Details on the experiment setting . . . 103
3.7.1 Datasets . . . 103
3.7.2 Hyper-parameter choice . . . 103
3.7.3 Metrics . . . 105
3.7.4 More results . . . 105
3.8 Detailed proofs . . . 106
3.9 Non-transferrability of attacks between univariate and multivariate forecasters 110
3.10 Conclusion . . . 111

Chapter 4 Promoting Robustness of Randomized Smoothing: Two Cost-Effective Approaches . . . 112
4.1 Introduction . . . 112
4.2 Related works and backgrounds . . . 114
4.3 Our proposed main methods . . . 117
4.3.1 Approach 1: **AdvMacer** . . . 118
4.3.2 Approach 2: **EsbRs** . . . 121
4.4 Experiments . . . 126
4.5 Full Algorithm of **AdvMacer** . . . 132
4.6 Model configuration . . . 133
4.7 Details of mixed ensemble component . . . 133
4.8 Certified accuracy . . . 134
4.9 Optimal weighted ensemble . . . 134
4.10 More SVHN experiments . . . 134
4.11 Certification with more or fewer samples . . . 134
4.12 ImageNet . . . 134

4.13 Conclusions . . . 134

Chapter 5 Robust Estimation in Linear Regression with both Heavy-tailed Data and Noise . . . 142

5.1 Introduction . . . 142

5.2 Methodology . . . 143

5.2.1 Estimation . . . 143

5.2.2 Main result . . . 144

5.3 Proof . . . 145

5.3.1 Proof of theorem 5.2.2 . . . 145

5.3.2 Proof of theorem 5.2.3 . . . 146

Bibliography . . . 154

LIST OF FIGURES

Figure 1.1. The graph of $\|A^k\|$ for $B = 3, 4$ and $p = 20, 25, 30$. 7

Figure 2.1. The qq-plot of banded design. 66

Figure 2.2. The qq-plot of block diagonal design. 66

Figure 3.1. Illustration figure of sparse indirect attack. 88

Figure 3.2. Plots of average wQL. 99

Figure 3.3. Plots of original time series and the ones under attacks. 102

Figure 4.1. Illustration of the idea behind **AdvMacer** 118

Figure 4.2. Plot of ACR against different number of component models. 127

LIST OF TABLES

Table 1.1. The average of $\|\hat{A}-A\|_F$ based on 1000 repetitions for different methods when $n = 50$ and $p = 10$. 21

Table 1.2. The average of $\|\hat{A}-A\|_F$ based on 1000 repetitions for different methods when $n = 50$ and $p = 30$. 21

Table 1.3. The average of $\|\hat{A}-A\|_F$ based on 1000 repetitions for different methods when $n = 50$ and $p = 100$. 22

Table 3.1. Average wQL on **Electricity** dataset under **deterministic** and **probabilistic** attack. Target time series $\mathcal{I} = \{1\}$ and attacked time stamp $H = \{\tau\}$. Smaller is better. 101

Table 3.2. Average wQL on **Traffic** dataset under **deterministic** and **probabilistic** attack. Target time series $\mathcal{I} = \{1, 5\}$ and attacked time stamp $H = \{\tau - 1, \tau\}$. Smaller is better. 101

Table 3.3. Summary of statistics of the datasets used, including prediction length τ, domain, frequency, dimension, and time steps. 104

Table 3.4. Metrics on **Electricity** dataset under **deterministic** attack. Target time series $\mathcal{I} = \{1\}$ and attacked time stamp $H = \{\tau\}$. Smaller is better. 106

Table 3.5. Metrics on **electricity** dataset under **probabilistic** attack using sparse layer. Target time series $\mathcal{I} = \{1\}$ and attacked time stamp $H = \{\tau\}$. Smaller is better. 107

Table 3.6. Average wQL on **Electricity** dataset under **deterministic** attack. The defense is data augmentation and randomized smoothing with varying $\sigma = 0.1, 0.2, 0.3$. Target time series $\mathcal{I} = \{1\}$ and attacked time stamp $H = \{\tau\}$. Smaller is better. 108

Table 3.7. Average wQL on **Taxi** dataset under **deterministic** attack. Target time series $\mathcal{I} = \{1\}$ and attacked time stamp $H = \{\tau\}$. Smaller is better. 108

Table 3.8. Average wQL on **Taxi** dataset under **probabilistic** attack. Target time series $\mathcal{I} = \{1\}$ and attacked time stamp $H = \{\tau\}$. Smaller is better. 108

Table 3.9. Average wQL on **Wiki** dataset under **deterministic** attack. Target time series $\mathcal{I} = \{1\}$ and attacked time stamp $H = \{\tau\}$. Smaller is better. 108

Table 3.10. Average wQL on **Wiki** dataset under **probabilistic** attack. Target time series $\mathcal{I} = \{1\}$ and attacked time stamp $H = \{\tau\}$. Smaller is better. 109

Table 3.11. Transfer the attack from DeepAR to DeepVAR. Target items $\mathcal{I} = \{1\}$ and time horizon to attack $H = \{\tau\}$. Clean DeepAR and DeepVAR models are used. Averaged wQL is reported. 110

Table 4.1. Cifar-10: ACR of different models on the first 500 test images of Cifar-10 with varing σ. Clean accuracy is reported in parenthesis. Reported models include SmoothAdv, MACER, **AdvMacer** , Esb-RS. $N = 100k$ samples are used in certification unless otherwise specified. 128

Table 4.2. ACR of optimal weighted ensemble. 130

Table 4.3. SVHN: clean accuracy and ACR of different models evaluated at the first 500 test images of SVHN with $\sigma = 0.25$. 131

Table 4.4. Model configuration: main hyper-parameters and training time for Smooth-Adv and **AdvMacer** on Cifar-10 with varing σ. For the additional parameters in **AdvMacer** , we pick $\lambda = 12.0, \gamma = 8.0, \beta = 16.0$. 133

Table 4.5. Component models in mixed ensemble experiment in Figure 4.2. The mixed ensemble with totally N component models uses m **AdvMacer** and n SmoothAdv models. m and n are given as follows. 133

Table 4.6. Certified accuracy: certified accuracy and ACR of the first 500 test images of Cifar-10 with $\sigma = 0.25$. Each column represents the robust accuracy that can be certified at this ℓ_2 radius. 135

Table 4.7. Certified accuracy: certified accuracy and ACR of the first 500 test images of Cifar-10 with $\sigma = 0.50$. Each column represents the robust accuracy that can be certified at this ℓ_2 radius. 136

Table 4.8. Certified accuracy: certified accuracy and ACR of the first 500 test images of Cifar-10 with $\sigma = 1.00$. Each column represents the robust accuracy that can be certified at this ℓ_2 radius. 137

Table 4.9. SVHN: clean accuracy and ACR of different models evaluated at the first 500 test images of SVHN with varing σ. 140

Table 4.10. Cifar-10: ACR of different models on the first 500 test images of Cifar-10 with varing σ. Clean accuracy is reported in parenthesis. Reported models include SmoothAdv, MACER, **AdvMacer** , **EsbRs**. $N = 100,000$ samples are used in certification unless otherwise specified. 141

Table 4.11. ImageNet: ACR on 500 test images of ImageNet. Clean accuracy is reported in parenthesis. 141

ACKNOWLEDGEMENTS

First of all, I would like to express my greatest appreciation to my Ph.D. advisor Professor Danna Zhang, who guided me through the projects and more importantly, taught me how to become an independent researcher. I am very fortunate to be advised by her in my Ph.D. study. She is a very considerate person and cares students as much as she cares the students' work. She provides me with help not only academically, but also mentally; she gives me guidance not only as an advisor, but also as a friend. I really appreciate the opportunity to work with her.

Besides, it's also my great honor to collaborate with Professor Lily Weng from Halıcıoğlu Data Science Institute at UCSD. I have been working with her on computer science related topics for more than one year. Her insights, creativity and passion helped me build the mindset to conduct academic research as a computer scientist. In the collaboration with Lily, I also got the chance to work with other talented industry researchers, Nghia Hoang and Lam Nguyen. Also, I would like to thank Nghia, Lily and Lam for their help in my job search and their advice on my future career.

I am also grateful to Professor Dimitris Politis, Professor Ery Arias-Castro, Professor Wenxin Zhou and Professor Yixiao Sun for serving as my Ph.D. committee and the valuable advice they provided in my advancement talk. It's also important to thank Professor Weijie Su from the Wharton School at University of Pennsylvania. When I was a master student at Penn, it was his encouragement, guidance and advice that helped me make the important decision: pursuing a Ph.D. degree in statistics.

Last but not least, I wish to thank my dear friends at San Diego, Los Angeles, Bay Area and other places. Especially, I would like to thank Xin Liu, Bowen Li, Chloe Lyu, Jason Lin, Chi Zhang, Jinjie Zhang, Jiayuan Gu, Haochen Huang, Yimeng Zhang, Chengzhu Zhang, Muhan Zhao, Tuo Lin, Jiajie Shi, etc. I want to thank my friends for their support, advice, help or simply accompanying me. Finally, I hope to express my deepest gratitude to my parents Xiaojing Wang and Xianpei Liu. Their love and support

encourage and lead me to the completion of Ph.D. study.

Chapter 1, in part, is a reprint of the material in the paper ”A Bernstein-type Inequality for High Dimensional Linear Processes with Applications to Robust Estimation of Time Series Regressions”, Liu, Linbo and Zhang, Danna. This paper is currently under minor revision at *Statistica Sinica*. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, is currently being prepared for submission of the material ”High-dimensional Simultaneous Inference on non-Gaussian VAR Model via De-biased Estimator”, Liu, Linbo and Zhang, Danna. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in part, has been submitted for publication of the material ”Robust Multivariate Time-Series Forecasting: Adversarial Attacks and Defense Mechanisms”, Liu, Linbo, Park, Youngsuk, Hoang, Trong Nghia, Hasson, Hilaf, and Huan, Jun to *International Conference on Learning Representations* and is currently under review. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, has been submitted for publication of the material ”Promoting Robustness of Randomized Smoothing: Two Cost-Effective Approaches”, Liu, Linbo, Trong, Hoang, Nguyen, Lam, and Weng, Tsui-Wei to *Computer Vision and Pattern Recognition Conference* and is currently under review. The dissertation author was the primary investigator and author of this paper.

Chapter 5, in full, is a research project of the material ”Robust Estimation in Linear Regression with both Heavy-tailed Data and Noise”, Liu, Linbo and will be further enhanced for submission. The dissertation author was the primary investigator and author of this project.

VITA

2016	B.S. in Mathmatics and Applied Mathematics, Tongji University
2018	M.A. in Mathematics, University of Pennsylvania
2018–2022	Graduate Teaching and Research Assistant, University of California San Diego
2022	Ph.D. in Mathematics, University of California San Diego

PUBLICATIONS

L. Liu, Y. Park, T. Hoang, H. Hasson and J. Huan, ”Towards Robust Multivariate Time-Series Forecasting: Adversarial Attacks and Defense Mechanisms”, *Preprint*, 2022. `arXiv: 2207.09572`.

L. Liu, T. Hoang, L. Nguyen and T. Weng, ”Promoting Robustness of Randomized Smoothing: Two Cost-Effective Approaches”, *CVPR*, 2023. submitted.

L. Liu and D. Zhang, ”High-dimensional Simultaneous Inference on non-Gaussian VAR Model via De-biased Estimator”, *Preprint*, 2022. `arXiv: 2111.01382`.

L. Liu and D. Zhang, ”A Bernstein-type Inequality for High Dimensional Linear Processes with Applications to Robust Estimation of Time Series Regressions”, *Statistica Sinica*, 2022. minor revision.

ABSTRACT OF THE DISSERTATION

Robust Inference and Learning of Multivariate Statistical Models

by

Linbo Liu

Doctor of Philosophy in Mathematics

University of California San Diego, 2022

Professor Danna Zhang, Chair

Model robustness has become increasingly popular in recent decades. We study multiple aspects of robustness (in the setting of time series, image classification and linear regression) in this dissertation work. First three chapters concerns the time series setting. Specifically, Chapter 1 establishes a novel Bernstein-type inequality for high dimensional linear processes. We then apply it to investigate two high dimensional robust estimation problems: (1) time series regression with fat-tailed and correlated covariates and errors, (2) fat-tailed vector autoregression. As a natural requirement of consistency, the dimension can be allowed to increase exponentially with the sample size under very mild moment and dependence conditions. In Chapter 2, we develop Gaussian approximation theory for

VAR model to derive the asymptotic distribution of the de-biased estimator and propose a multiplier bootstrap-assisted procedure to obtain critical values under very mild moment conditions on the innovations. Chapter 3 studies the threats of adversarial attack on multivariate probabilistic forecasting models and viable defense mechanisms. Our studies discover a new attack pattern that negatively impact the forecasting of a target time series via making strategic, sparse (imperceptible) modifications to the past observations of a small number of other time series. To mitigate the impact of such attack, we also develop two defense strategies. First, we extend a previously developed randomized smoothing technique in classification to multivariate forecasting scenarios. Second, we develop an adversarial training algorithm that learns to create adversarial examples and at the same time optimizes the forecasting model to improve its robustness against such adversarial simulation. In Chapter 4, we improve the robustness of image classifier by enhancing the randomized smoothing technique and model ensemble. Chapter 5 considers the robust estimation of linear regression coefficients under heavy-tailed noise and covariates using a clipping idea.

Chapter 1

A Bernstein-type Inequality for High Dimensional Linear Processes

1.1 Introduction

High dimensional data analysis is increasingly important in the information era with the rapid explosion of massive data sets. High-dimensional linear regression has acquired significant relevance and attention. Consider the linear regression models

$$Y_i = X_i^\top \beta + \xi_i, \quad i = 1, \ldots, n$$

where Y_i, X_i and ξ_i are respectively the response, covariate and error variables. Various regularization methods have been widely used for estimating the p-dimensional regression parameter vector, including [130, 160, 39, 15, 96, 153] and many others; see [19] for a comprehensive overview. In most investigations, covariates X_i (if it is a random design) and errors ξ_i are assumed to be i.i.d. Gaussian or sub-Gaussian random variables, which turns out to be too restrictive in many applications.

On the one hand, serial correlation might occur when the data are collected over time. Linear regression with time series regressors and autoregressive errors are often considered ([55, 131, 120]). On the other hand, many applications involving time series data are concerned with high dimensional objects and fat-tailed distributions, including quantitative finance ([31]), portfolio allocation ([69]), risk management([72]), brain network ([41]) and geophysical dynamic studies ([71]).

Some progress has been made on linear regression with correlated errors. Lasso estimator was studied for linear regression with autoregressive errors by [136] and [150], weakly dependent errors by [50] and long memory errors by [67]. They mainly dealt with the cases where the dimension p is smaller than the sample size n or imposed the Gaussian assumption on the error process. Using the framework of functional dependence measures, [143] and [28] accounted for both dependent covariates and errors in linear regression. As a natural requirement of consistency, p is allowed to increase with n at a polynomial rate; a narrow range is still restricted for the dimension in the presence of non-Gaussian and dependent errors. In contrast, an ultra high dimension can be allowed with i.i.d. well-behaved covariates and fat-tailed errors based on a penalized Huber M-estimator; see, for example, [75, 38, 86, 88] among many others. Other methods for robust linear regression in high dimensions include sparse least trimmed squares ([2]), MM-Lasso ([123]), ESL-Lasso ([137]) and so on. Robust estimation of high dimensional time series regression with fat-tailed and correlated covariates and errors has been rarely considered.

As another closely related topic, vector autoregression is a popular linear model to describe the evolution of a set of variables over time. The past two decades have witnessed a large progress in estimating high-dimensional vector autoregressive models. Inspired from the development in high-dimensional linear regression, [58, 100, 9] considered the Lasso estimator using ℓ_1 penalty. [70] established oracle inequalities for high-dimensional vector autoregressive models. [52] adopted a Dantzig-type penalization. [49] proposed a Bayesian information criterion based on residual sums of least squares estimator to estimate high dimensional banded autoregression. Most work required the Gaussian assumption or the existence of finite exponential moment. In econometric analysis, [122] raised the concern that fat tails in vector autoregressive models can affect the validity of statistical inference and it may result to low degrees of freedom due to the estimation of possibly extremely many parameters. To this end, we shall also investigate robust estimation of high dimensional fat-tailed autoregressive models.

Overall, we will combine all aspects and investigate the linear regression or autoregression with (i) time series covariates, (i) possibly correlated errors, (iii) fat tail and (iv) ultra high dimension. It makes many traditional statistical analysis tools for independent data infeasible and poses great challenge on the developing new tools for high dimensional time series. As one important contribution, we establish a new Bernstein type inequality for the sum of a bounded transformation of high dimensional linear processes. With the help of the newly developed inequality, we can obtain consistency in many estimation problems under the very mild condition of the type $\log p = o(n^c)$ for some $c > 0$.

The rest of the chapter is organized as follows. In Section 2, we introduce the framework of high dimensional linear processes and the useful quantities that can depict the temporal and cross-sectional dependence, then present a new Bernstein type inequality for high dimensional linear processes. In Section 3, we investigate two robust estimation problems: time series linear regression with correlated and fat-tailed covariates and errors, and autoregressive models with fat-tailed errors. Some simulation results are displayed in Section 4 to assess the empirical performance of robust estimators and all of the proofs are relegated to Section 1.6 and Section 1.7.

We introduce some notation. For a vector $\beta = (\beta_1, \ldots, \beta_p)^\top$, let $|\beta|_1 = \sum_i |\beta_i|$, $|\beta|_2 = (\sum_i \beta_i^2)^{1/2}$, $|\beta|_0 = |\{i : \beta_i \neq 0\}|$ and $|\beta|_\infty = \max_i |\beta_i|$. Let $\mathrm{Supp}(\beta)$ be the support of β. For a matrix $A = (a_{ij})_{1 \le i,j \le p} \in \mathbb{R}^{p\times p}$, let λ_i, $i = 1, \ldots, p$, be its eigenvalues and $\lambda_{\max}(A) = \max_i |\lambda_i|$ be the spectral radius, $\lambda_{\min}(A) = \min_i |\lambda_i|$. Let $\kappa(A)$ denote the condition number of A. Denote $|A|_1 = \sum_{i,j} |a_{ij}|$, $\|A\|_1 = \max_j \sum_i |a_{ij}|$, $\|A\|_\infty = \max_i \sum_j |a_{ij}|$, spectral norm $\|A\| = \|A\|_2 = \sup_{|x|_2 \neq 0} |Ax|_2/|x|_2$ and Frobenius norm $\|A\|_F = (\sum_{i,j} a_{ij}^2)^{1/2}$. Moreover, let $\mathrm{tr}(A)$ be the trace of A, $\|A\|_{\max} = \max_{i,j} |a_{ij}|$ be the entry-wise maximum norm, $|A|$ be a matrix after taking absolute value of A, i.e. $|A| = (|a_{ij}|)_{i,j}$. For a random variable X and $q > 0$, define $\|X\|_q = (\mathbb{E}[|X|^q])^{1/q}$. For two real numbers x, y, set $x \vee y = \max(x, y)$. For two sequences of positive numbers $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ if there exists some constant $C > 0$, such that $a_n/b_n \leq C$ as $n \to \infty$, and also write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. We use $c_0, c_1, \ldots$ and $C_0, C_1, \ldots$ to denote some universal positive constants whose values may vary in different context. Throughout the chapter, we consider the high dimensional regime, allowing the dimension p to grow with the sample size n, that is, we assume $p = p_n \to \infty$ as $n \to \infty$.

1.2 Bernstein-type Inequality for High Dimensional Linear Processes

We consider a general framework of p-dimensional stationary linear process

$$X_i = (X_{i1}, \ldots, X_{ip})^\top = \mu + \sum_{k=0}^{\infty} A_k \varepsilon_{i-k} \tag{1.2.1}$$

where $\mu \in \mathbb{R}^p$ is the mean vector, $A_0 = I_p$, A_k, $k \geq 1$, are $p \times p$ coefficient matrices with real entries such that $\sum_{k=0}^{\infty} \operatorname{tr}(A_k^\top A_k) < \infty$, $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{ip})^\top$, and ε_{ij}, $i \in \mathbb{Z}$, $1 \leq j \leq p$, are i.i.d. random variables with zero mean and finite variance. Then by Kolmogorov's three-series theorem, the linear process (1.2.1) is well defined. Many researchers have worked on this model recently; see for example, [11, 12, 64, 84, 25] among others. A special case of (1.2.1) is the stationary Gaussian process. If $A_k = 0$ for $k > d$, it becomes a vector moving average process of order d ([112, 91, 17]). Another important class covered by (1.2.1) is the vector autoregressive (VAR) model, which has been widely used in economics and finance (e.g., [122, 124, 132, 37] etc.).

The linear process (1.2.1) is a flexible multivariate model for correlated data in that the coefficient matrices A_k capture both temporal and cross-sectional (spatial) dependence. We first impose the condition on the decay rate of A_k, which is associated with the dependence strength of the underlying process. We assume that there exist $0 < \rho_p < 1$

and $1 \leq \gamma_p < \infty$ such that

$$\|A_k\| = \sup_{|x|_2 \neq 0} \frac{|A_k x|_2}{|x|_2} \leq \gamma_p \cdot \rho_p^k \tag{1.2.2}$$

for all $k \geq 0$. The condition (1.2.2) implies short-range dependence in the sense that the autocovariance matrices $\mathrm{Cov}(X_0, X_j) = \sum_{k=0}^{\infty} A_k A_{k+j}^\top$ are absolutely summable. Here ρ_p is used to depict the strength of temporal dependence: smaller ρ_p implies faster decay rate and thus weaker temporal dependence. And the magnitude of γ_p naturally quantifies the spatial dependence. As an interesting feature, both quantities γ_p and ρ_p may depend on p in the high dimensional regime. For instance, when p is large, ρ_p may be a large rate close to 1 and it suggests a relatively slow decay speed. In fact, we can always find a proper absolute constant free of p and strictly smaller than 1 so that (1.2.2) can be rephrased as

$$\|A_k\| \leq \gamma_p \cdot \rho_0^{k/\tau_p} \text{ for some } \tau_p \geq 1. \tag{1.2.3}$$

Particularly, we set $\tau_p \equiv 1$ if there exists ρ_0 such that $\rho_p \leq \rho_0 < 1$, and $\tau_p = \log \rho_0 / \log \rho_p$ for ρ_0 satisfying $0 < \rho_0 \leq \rho_p$ if ρ_p is large and increase with p. In the latter case, it could happen that $\tau := \tau_p$ is an unbounded function in terms of the dimension p. The high dimensional vector autoregressive model in Example 1.2.1 illustrates this feature. Thereafter, for notational simplicity, we omit the dimension subscript in γ_p, τ_p, and refer them as γ, τ. And we assume $\tau \leq n$; otherwise there may exist very strong temporal dependence in the sense that $\|A_k\|$ is decaying at a rate no faster than $\rho_0^{1/n}$.

Example 1.2.1 (High Dimensional Vector Autoregressive Models)*.* Consider the VAR(1) model

$$X_i = A X_{i-1} + \varepsilon_i, \tag{1.2.4}$$

where $A \in \mathbb{R}^{p \times p}$ is the transition matrix, and ε_i, $i \in \mathbb{Z}$, are i.i.d. error vectors with mean 0 and covariance matrix I_p. Equivalently it can be represented by the moving average model:

$X_i = \sum_{k=0}^{\infty} A^k \varepsilon_{i-k}$, a special case of (1.2.1) with $A_k = A^k$. The process is stable (and hence stationary) if and only if the spectral radius $\lambda_{\max}(A) < 1$ ([91]). If A is symmetric, as $\lambda_{\max}(A) = \|A\|$, condition (1.2.2) can be easily verified with $\rho_p = \lambda_{\max}(A)$ and $\gamma = 1$. For asymmetric A, it has a better interpretation when we look into condition (1.2.3), and it could happen that τ may increase with the dimension p. Consider the design $A = (a_{ij})_{i,j=1}^p$ with $a_{ij} = \lambda^{j-i+1}\mathbf{1}\{0 \le j-i \le B-1\}$ for some $0 < \lambda < 1$ and $1 \le B \le p$. Here B depicts how many variables at most in X_{i-1} that have spatial effect on X_{ij}. Figure 1.1 delivers the plot of $\|A^k\|$ under the numerical setup $\lambda = 0.55$, $B = 3, 4$ and $p = 20, 25, 30$. As can be seen, $\|A^k\|$ decays truly after a certain lag that is moving forward as p is getting larger. This lag can be defined as τ in condition (1.2.3), so τ increases with p in this design. Additionally the geometric decay (its existence is to be shown later) occurs at a slow speed, viewed as another evidence of large ρ_p (or large τ equivalently). For example, when $B = 3$, $p = 30$, $\|A^k\|$ roughly drops from 1.35 to 0.1 over a broad lag range from 30 to 60. The peak of $\|A^k\|$ before decay is defined as γ, indicating the strength of spatial dependence; we can tell stronger spatial dependence with a larger B results to larger γ by comparing the two plots.

Concentration inequalities play an important role in the study of sums of random variables. A number of inequalities have been derived for independent random variables; see [19] for a review. Bernstein's inequality ([10]) is one of the powerful tools when analyzing the concentration behavior by providing an exponential inequality for sums of independent random variables which are uniformly bounded. To fix the idea, let $Y_1, \ldots, Y_n$ be i.i.d. random variables such that $\mathbb{E}Y_i = 0$, $\mathrm{Var}(Y_i) = \sigma^2 < \infty$, and $|Y_i| \le M$ for all i. Then for any $x > 0$, one has

$$\mathbb{P}\Big(\sum_{i=1}^n Y_i \ge x\Big) \le \exp\Big\{-\frac{x^2}{2n\sigma^2 + 2Mx/3}\Big\}, \tag{1.2.5}$$

which suggests two types of bound for tail probability: sub-Gaussian tail $\exp\{-x^2/(Cn\sigma^2)\}$

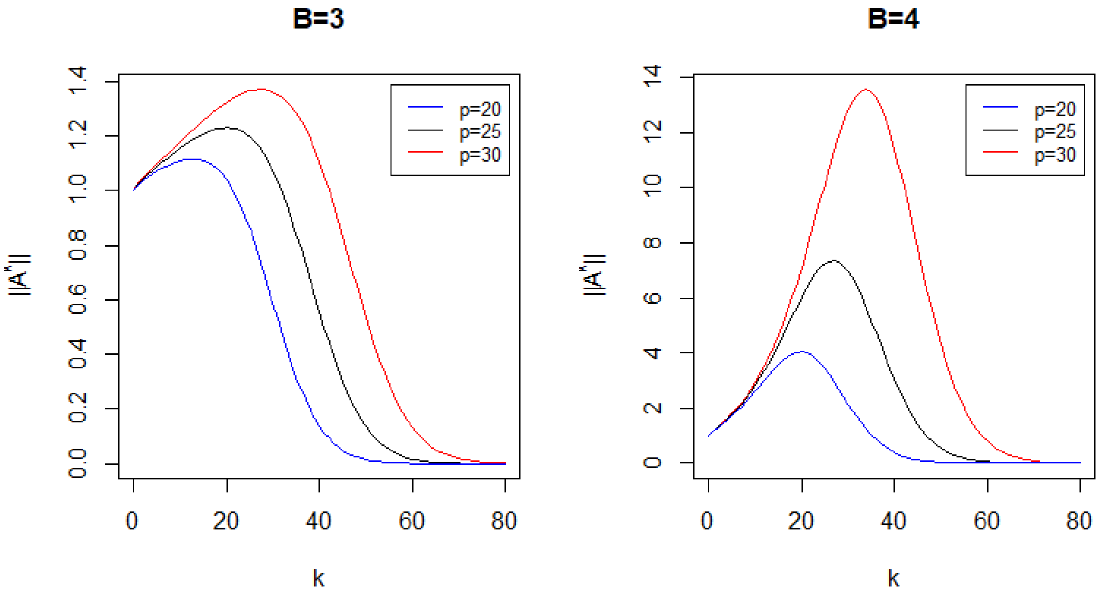


Figure 1.1. The graph of $\|A^k\|$ for $B = 3, 4$ and $p = 20, 25, 30$.

in terms of the variance of $\sum_{i=1}^n Y_i$ and sub-exponential tail $\exp\{-x/(CM)\}$ in terms of the uniform bound M. Bernstein type inequalities have been developed for Markov chains or temporally dependent processes with an additional order ($\log n$ in some constant powers) in the sub-exponential-type tail; see, for example, [1], [97], [53] and [155]. The problem of generalizing to high dimensional time series is quite challenging and very few results have been obtained. Our first goal is to establish a new Bernstein type inequality for the sum of a bounded transformation of the high dimensional linear processes in (1.2.1).

Theorem 1.2.1. *Let X_i be the linear process generated from (1.2.1) with $\mathbb{E}\varepsilon_{ij} = 0$, $\mathbb{E}\varepsilon_{ij}^2 = \sigma^2 < \infty$ and condition (1.2.3) be satisfied. Let $G : \mathbb{R}^p \to \mathbb{R}$ be a function with $|G(u)| \leq M$ for all $u \in \mathbb{R}^p$. Suppose there exists a vector $g = (g_1, \ldots, g_p)^\top$ with $g_i \geq 0$ and $\sum_{i=1}^p g_i = 1$ such that the following Lipschitz condition holds: for all $u = (u_1, \ldots, u_p)^\top$ and $v = (v_1, \ldots, v_p)^\top$,*

$$|G(u) - G(v)| \leq \sum_{i=1}^p g_i |u_i - v_i|. \tag{1.2.6}$$

Then for any $x > 0$, we have

$$\mathbb{P}\Big(\sum_{i=1}^{n} G(X_i) - \mathbb{E}G(X_i) \geq x\Big) \leq 2\exp\Big\{-\frac{x^2}{C_1 n\sigma^2\tau^2\gamma^2 + C_2\tau M x}\Big\}, \qquad (1.2.7)$$

where the constants C_1 and C_2 are given by

$$C_1 = \frac{16\mathrm{e}^2}{\sqrt{2\pi}\rho_0^4[\log(1/\rho_0)]^3}, \quad C_2 = \frac{8\mathrm{e}}{\log(1/\rho_0)}.$$

Remark 1.2.2. In the special case of one dimensional processes $X_i \in \mathbb{R}$, as $\tau = 1$ and γ is also of a constant order, our probability inequality in Theorem 1.2.1 is as sharp as the classical Bernstein's inequality (1.2.5) in view of $\mathrm{Var}(X_i) \asymp \sigma^2\gamma^2$. [97] established an exponential-type concentration with an additional $(\log n)^2$ in the denominator of the exponential inequality:

$$\mathbb{P}\Big(\sum_{i=1}^{n} X_i \geq x\Big) \leq \exp\Big\{-\frac{Cx^2}{nv^2 + M^2 + M(\log n)^2 x}\Big\}, \qquad (1.2.8)$$

where (X_i) is a strong mixing process of mean 0 and upper bounded by M in magnitude, and v^2 is the asymptotic variance of $\sum_{i=1}^{n} X_i/\sqrt{n}$; and [155] also derived a similar bound with the dependence adjusted measure in place of v^2. Compared with (1.2.8), our result is strictly sharper by removing the additional factor $(\log n)^2$, even if the mild order $\nu^2 = O(1)$ is assumed in the last two displays.

The result (1.2.7) can deal with high dimensional dependent processes concerning both temporal dependence and cross-sectional dependence, characterized by τ and γ respectively. We now discuss the conditions of the theorem. The Lipschitz condition (1.2.6) is an essential requirement. It covers the class of linear transforms; particularly, for $G(X_i) = \sum_{j=1}^{p} a_j h_j(X_{ij})$, where $\sum_{j=1}^{n} |a_j| = 1$, $h_j : \mathbb{R} \to \mathbb{R}$ are univariate functions satisfying $|h_j(x)| \leq M$ and $|h_j(x) - h_j(y)| \leq 1$ for any $x, y \in \mathbb{R}$, condition (1.2.6) is satisfied with $g_j = |a_j|$. As a special case, when $G(X_i) = h_j(X_{ij})$, for a fixed $1 \leq j \leq p$, the result then provides a concentration inequality for sums of each component process $(X_{ij})_{i\in\mathbb{Z}}$ after the transformation h_j; see the application of estimating the mean vector of high dimensional linear processes in a robust way at the end of this section. The requirement $|g|_1 = \sum_{i=1}^p g_i = 1$ is not very restrictive, as one can always apply the theorem to the function $G/|g|_1$ to make it satisfied. Condition (1.2.3) requires $\|A_k\|$ geometrically decayed up to the quantity γ and the decay speed is controlled by τ. [25] worked on the same linear model under a weaker condition allowing polynomial decay, namely, $\|A_k\| = O((1 \vee k)^{-\alpha})$ for some $\alpha > 1$, under which, it is noteworthy that an exponential type probability inequality does not hold in general even if it is one dimensional process with a uniform bound. That is to say, if we relax the condition (1.2.2) to a polynomial decay, the concentration inequality delivers an exact rate with algebraic decay for one dimensional linear process; see Theorem 14 in [24].

In Corollary 1.2.2 below, we also provide a tail probability inequality if G satisfies a different Lipschitz condition from (1.2.6). There is an additional $(\log p)^2$ term in the sub-Gaussian-type tail and an additional $\log p$ in the sub-exponential-type tail. In the next section, different formats of G are to be considered in estimating time series regression.

Corollary 1.2.2. *Consider the same setting of the model as in Theorem 1.2.1. Let $G : \mathbb{R}^p \to \mathbb{R}$ be a function with $|G(u)| \leq M$ for all $u \in \mathbb{R}^p$. Assume that*

$$|G(u) - G(v)| \leq |u - v|_2, \text{ for all } u, v \in \mathbb{R}^p.$$

Assume that $\log p > 1$ and $\tau \log p \leq n$. Then for any $x > 0$, we have

$$\begin{aligned} &\mathbb{P}\Big(\sum_{i=1}^n G(X_i) - \mathbb{E}G(X_i) \geq x\Big) \\ \leq\ & 2\exp\Big\{-\frac{x^2}{C_3 n(\sigma^2\gamma^2 + M^2)\tau^2(\log p)^2 + C_4\tau M(\log p)x}\Big\}, \end{aligned} \tag{1.2.9}$$

where the constants C_3 and C_4 depend on ρ_0 only.

In Theorem 1.2.1, the existence of a finite variance of ε_{ij} is assumed. If it is relaxed to the existence of finite exponential moment, a similar bound can be achieved with G not necessarily bounded; see Theorem 1.2.3 below.

Theorem 1.2.3. *In the model (1.2.1), assume that $\mathbb{E}\varepsilon_{ij} = 0$, $\mathbb{E}\exp(c_0|\varepsilon_{ij}|) = \theta < \infty$ for some constant $c_0 > 0$ and condition (1.2.3) is met. Then for G satisfying (1.2.6), it holds that*

$$\mathbb{P}\Big(\sum_{i=1}^{n} G(X_i) - \mathbb{E}G(X_i) \geq x\Big) \leq 2\exp\Big\{ - \frac{x^2}{C_5 n\theta^2\tau^2\gamma^2 + C_6\gamma\tau x}\Big\}, \tag{1.2.10}$$

where the constants C_5 and C_6 depend on ρ_0 and c_0.

One immediate application of Theorem 1.2.1 is to estimate the mean vector for high dimensional fat-tailed linear processes. From an M-estimation viewpoint, we apply Huber's estimator ([60]) of the mean vector, denoted by $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_p)^\top$, with $\hat{\mu}_j$ as the solution of a to the equation

$$\sum_{i=1}^{n} \phi_\nu(X_{ij} - a) = 0,$$

where $\phi_\nu(x) = (x \wedge \nu) \vee (-\nu)$ is the Huber function with the robustification parameter $\nu > 0$.

Theorem 1.2.4. *Let X_i be generated from model (1.2.1) with $\mathbb{E}\varepsilon_{ij} = 0$, $Var(\varepsilon_{ij}) = 1$, $\mu = \mathbb{E}X_i$ and $\max_{1\leq j\leq p} Var(X_{ij}) = \mu_2^2 < \infty$. Choose $\nu \asymp \mu_2\sqrt{n/\log p}$. With probability at least $1 - 4p^{-c}$ for some $c > 0$, it holds that*

$$|\hat{\mu} - \mu|_\infty \leq C(\gamma + \mu_2)\tau\sqrt{\frac{\log p}{n}} \tag{1.2.11}$$

under the scaling condition $(\gamma + \mu_2)\tau\sqrt{\log p/n} \to 0$, where C is a positive constant depending on c and the constants C_1, C_2 in Theorem 1.2.1.

Remark 1.2.3. Theorem 1.2.4 delivers the rate of ℓ_∞ norm convergence for the robust mean estimator $\hat{\mu}$ and it involves a delicate interplay with the cross-sectional dependence, temporal dependence and the variance of the process. If γ, μ_2 and τ are all of a constant order, it follows that

$$|\hat{\mu} - \mu|_\infty = O_{\mathbb{P}}(\sqrt{\log p/n}), \tag{1.2.12}$$

under the scaling condition $\log p/n \to 0$. We shall remark that (1.2.12) is as sharp as the optimal rate provided in Theorem 5 of [38] concerning the concentration of the mean estimation for the i.i.d. case. And it is strictly sharper than the results using existing Bernstein type inequalities for time series such as the ones in [97], [53] and [155].

1.3 Robust Estimation of Time Series Regression

In this section, we shall investigate robust estimation of high dimensional time series linear regression and autoregression with fat-tailed covariates and errors. It is expected that our framework of high dimensional linear processes and these Bernstein type inequalities will be useful in other high-dimensional estimation and inference problems that involve dependent and non-sub-Gaussian random variables.

1.3.1 Estimating Time Series Regression with Correlated Errors

We work on linear regression models with random design that involve time dependent covariates and errors:

$$Y_i = X_i^\top \beta^* + \xi_i, \tag{1.3.1}$$

with more justification provided as follows.

Assumptions.

(A1) X_i is generated from the p-dimensional linear process (1.2.1) with $\mathbb{E}(\varepsilon_{ij}) = 0$ and

$$\mathrm{Var}(\varepsilon_{ij}) = \sigma_\varepsilon^2 < \infty.$$

Condition (1.2.3) is satisfied with γ and τ, which may depend on p.

(A2) $\xi_i = \sum_{k=0}^{\infty} b_k \eta_{i-k}$ is the error process, where η_i are i.i.d. random variables with $\mathbb{E}(\eta_i) = 0$ and $\mathrm{Var}(\eta_i) = \sigma_\eta^2 < \infty$, and $b_k \leq C\rho^k$ for universal constants $0 < \rho < 1$ and $C < \infty$.

(A3) X_i is strictly exogenous in the sense that $(\varepsilon_i)_i$ are independent of $(\eta_i)_i$.

The framework (1.3.1) is quite general as the linear process includes a wide range of commonly used time series models. For linear regression models with dependent errors, earlier work mainly dealt with fixed design or i.i.d. covariates. [136] and [150] considered the case where ξ_i follows an autoregressive process, one type of linear processes. [50] concerned weakly dependent ξ_i introduced by [36] and specifically discussed the AR(1) and ARMA cases. [2] adopted the same format of moving average errors but assumed long memory dependence. More generally, [143] and [28] considered the nonlinear Wold representation with $X_i = g(\ldots, \varepsilon_{i-1}, \varepsilon_i)$ and $\xi_i = h(\ldots, \eta_{i-1}, \eta_i)$.

We form a modified ℓ_1-regularized Huber estimator of β, given by

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \Phi_\nu((Y_i - X_i^\top \beta) w(X_i)) + \lambda |\beta|_1,$$

where Φ_ν is Huber loss function ([60])

$$\Phi_\nu(x) = \begin{cases} x^2/2, & \text{if } |x| \leq \nu, \\ \nu|x| - \nu^2/2, & \text{if } |x| > \nu, \end{cases}$$

defined with respect to the robustification parameter $\nu > 0$. More properties of Huber regression are referred to [59], [148], [93], [125], [38], to name just a few. Motivated by [88], $w(x) : \mathbb{R}^p \to \mathbb{R}$ is a weight function defined by

$$w(x) = \min\left\{1, \frac{b}{|Bx|_2}\right\}$$

where $b \in \mathbb{R}$ is a fixed constant and $B \in \mathbb{R}^{p\times p}$ is a provided positive definite matrix. With such a choice of $w(x)$, it always holds that $|w(x)x|_2 \leq b/\lambda_{\min}(B) =: b_0$. Different from the regular Huber regression concerning well-behaved X_i (e.g., Gaussian or sub-Gaussian), an additional weight function is incorporated on the covariate process to account for the fat tails of X_i. As a popular convention, β^* is assumed to be sparse in the sense that $|\beta^*|_0 = s$. Denote the condition number of B as $\kappa(B) = \lambda_{\max}(B)/\lambda_{\min}(B)$. Theorem 1.3.1 below concerns the estimation consistency of $\hat{\beta}$.

Theorem 1.3.1. *Let Assumptions (A1) (A2) (A3) be satisfied. Assume*

$$b_0(b_0 + \kappa(B)\gamma\sigma_\varepsilon)\tau\sqrt{s}\sqrt{(\log p)^3/n} \to 0. \tag{1.3.2}$$

Choose $\nu \asymp \sigma_\eta(n/\log p)^{1/2}$ *and* $\lambda \asymp b_0\sigma_\eta(\log p/n)^{1/2}$. *With probability at least* $1 - 8p^{-c}$ *for some* $c > 0$, *it holds that that*

$$|\widehat{\beta} - \beta|_2 \leq C\frac{b_0\sigma_\eta}{\lambda_{\min}(\mathbb{E}[\frac{w^2(X_i)}{2}X_iX_i^\top])}\sqrt{\frac{s\log p}{n}}. \tag{1.3.3}$$

The scaling condition (1.3.2) to ensure consistency indicates a subtle interplay with the dimensionality parameters (s, p, n), internal parameters $(\tau, \gamma, \sigma_\varepsilon)$, and the known values b_0 and $\kappa(B)$ associated with the weight function $w(x)$. The convergence rate (1.3.3) scales inversely with the quantity $\lambda_{\min}(\mathbb{E}[\frac{w^2(X_i)}{2}X_iX_i^\top])$ and it suggests that we can not shrink the covariates too aggressively. If X_i is well-behaved with the existence of finite exponential moment, one may eliminate the weight function and replace the factor by the larger quantity $\lambda_{\min}(\mathbb{E}[X_i X_i^\top])$.

In the extensively studied regression setting with i.i.d. covariates, [38] delivered an optimal convergence rate of $|\widehat{\beta} - \beta|_2$ for weakly sparse model under the fat tails (the same as the minimax rate in [110]). In the special exact sparse case, their convergence rate is $\sqrt{s(\log p)/n}$ and it relies on the sub-Gaussian tail assumption for the covariates X_i. [88] allowed broader classes of distributions for X_i by inserting a weight function to control the Euclidean norm of X_i, but required the errors drawn i.i.d. from a symmetric distribution and thus selected ν at a fixed constant order (cf. Theorem 1), while [38] waived the symmetry requirement by allowing ν to diverge in order to reduce the bias induced by the Huber loss when the distribution of ξ_i is asymmetric. We borrow the ideas from both and further account for time dependent covariates and errors. Compared to [88] with i.i.d. covariates and i.i.d. errors, our result requires a stronger scaling condition (1.3.2) in terms of the dependence quantities γ, τ and a larger power of $\log p$, by concerning both dependent covariates and errors.

Applying ℓ_1 regularization in time series regression, [143] (cf. Theorem 5) dealt with correlated covariates and errors and allowed a wider class of stationary processes in a causal form. The linear error process in our consideration falls in the weaker dependence range within their framework. If $\gamma, \tau, \sigma_\eta = O(1)$, $p = o(n^{q-1})$ is required for their regular estimator without accounting for robustness, where $q > 2$ is the order of finite moments for ξ_i. [28] considered the Lasso estimator for a system of time series regression equations with one regression equation as a special case, for which the allowed dimension is still of a polynomial rate to ensure consistency by looking into the performance bound with respect to the prediction norm (cf. Corollary 5.4). In comparison with the above two work, we can allow a much wider range for the dimension p under mild conditions.

The tuning parameter ν plays a key role by adapting to errors with fat tails. In practical applications, the optimal values of the tuning parameters ν and λ can be chosen by a two-dimensional grid search using cross-validation or information-based criterion such as AIC or BIC. We leave theoretical investigation on selecting the tuning parameters as important future work.

1.3.2 Estimating Transition Matrix in VAR Models

To study the evolution of a set of endogenous variables over time, a popular choice is vector autoregression. Interpretations of large vector autoregressive models have been developed in various applications such as policy analysis ([121]), financial systemic risk analysis ([48]), portfolio selection ([77]), functional genomics ([119]) and brain networks ([117]).

As a general VAR model of order d can be reformulated as a VAR(1) model by appropriately redefining the random vectors, much work ([52], [49]) considered the model with lag 1 as given in (2.2.2). Among the work concerning high dimensional vector autoregressive models, most investigations require the Gaussian assumption ([70], [9], [52]) or some structure assumption stronger than the minimal requirement $\lambda_{\max}(A) < 1$; for example, [52] imposed $\|A\| < 1$, and [49] considered banded A with some decay condition on $\|A^k\|$ free of p. For many VAR designs (Example 1.2.1 is one such), it could happen that $\|A\| \geq 1$ and the dimension p, as the size of A, can play a role in measuring the temporal and cross-sectional dependence. [9] proposed stability measures to capture temporal and cross-sectional dependence. From a different viewpoint, we try to fill in the gap between the spectral radius of a matrix and its spectral norm. Intuition can be gained from the proposition below. It provides a sufficient and necessary condition for $\lambda_{\max}(A) < 1$ by relating to the spectral norm.

Proposition 1.3.2. *For any matrix A, it holds that $\lambda_{\max}(A) < 1$ if and only if there exists some finite integer k such that $\|A^k\| \leq \rho_0$ given any universal constant $0 < \rho_0 < 1$.*

Letting $\tau = \min\{k \in \mathbb{Z}^+ : \|A^k\| \leq \rho_0\}$ and $\gamma = \rho_0^{-1} \max_{0 \leq k \leq \tau - 1} \|A^k\|$, condition (1.2.3) holds for the model (2.2.2) without extra requirement. We now introduce the notation. Let $\boldsymbol{a}_{j\cdot}^{\top}$ be the j-th row of A and s_j be the cardinality of the support set of $\boldsymbol{a}_{j\cdot}$, i.e., $s_j = |\text{supp}(\boldsymbol{a}_{j\cdot})| = |\{i : a_{ij} \neq 0\}|$. Denote $s = \max_{1 \le j \le p} s_j$ and $\mathcal{S} = \sum_{i=j}^{p} s_j$. For robustness, we first truncate the data by obtaining $\tilde{X}_i = \phi_\nu(X_i)$, where ν is the truncation parameter and is to be determined later. For notational convenience, we assume X_0 is also observed. Based on the truncated sample $\widetilde{X}_i$ and tuning parameter $\lambda > 0$, we propose to estimate A by solving the following Lasso problem:

$$\widehat{A} = \arg\min_{B \in \mathbb{R}^{p \times p}} \frac{1}{n} \sum_{i=1}^{n} |\widetilde{X}_i - B\widetilde{X}_{i-1}|_2^2 + \lambda |B|_1, \tag{1.3.4}$$

which is equivalent to solving the p sub-problems:

$$\widehat{\boldsymbol{a}}_{j\cdot} = \arg\min_{\boldsymbol{b} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (\widetilde{X}_{ij} - \boldsymbol{b}^\top \widetilde{X}_{i-1})^2 + \lambda |\boldsymbol{b}|_1. \tag{1.3.5}$$

Before proceeding, we state the key assumptions on the process (2.2.2) and some scaling conditions to guarantee consistency of the robust estimator $\hat{A}$.

Assumptions.

(B1) $\mathbb{E}\varepsilon_{ij} = 0$; $\mathbb{E}\varepsilon_{ij}^2 = 1$; $\max_{1 \le j \le p} \|X_{ij}\|_q = \mu_q < \infty$ for some $q > 2$.

(B2) $\lambda_{\min}(\Sigma_0) \ge \kappa$ for some constant $\kappa > 0$, where $\Sigma_0 = \mathbb{E}(X_i X_i^\top)$.

(B3) $\mu_q \gamma \tau s^2 [(\log p)/n]^{(q-2)/(2q-2)} \to 0$.

(B3') $\mu_q \gamma \tau \mathcal{S}^2 [(\log p)/n]^{(q-2)/(2q-2)} \to 0$.

Assumption (B1) imposes polynomial moment conditions for the underlying VAR process. Assumption (B2) requires that the covariance matrix of X_i is well-conditioned. Assumption (B3) (or (B3')) assumes a vanishing scaling property. If μ_q, τ and γ are of a constant order, (B3) is reduced to the scaling condition that involves s (or $\mathcal{S}$), n and p only.

Theorem 1.3.3. *Let Assumptions (B1), (B2) and (B3) be satisfied. Choose the truncation parameter* $\nu \asymp \mu_q (n/\log p)^{1/(2q-2)}$. *Let* $\widehat{A}$ *be the solution of (1.3.4) with* $\lambda \asymp \mu_q \gamma \tau (\|A\|_\infty +$

1)$[(\log p)/n]^{(q-2)/(2q-2)}$. It holds that

$$\|\widehat{A}-A\|_\infty \leq C\mu_q\gamma\tau(\|A\|_\infty+1)s\bigg(\frac{\log p}{n}\bigg)^{\frac{1}{2}-\frac{1}{2q-2}} \tag{1.3.6}$$

with probability at least $1-8p^{-c}$ for some constant $c>0$. If Assumption (B3') is further satisfied, it also holds that

$$\|\widehat{A}-A\|_F \leq C'\mu_q\gamma\tau(\|A\|_\infty+1)\sqrt{\mathcal{S}}\left(\frac{\log p}{n}\right)^{\frac{1}{2}-\frac{1}{2q-2}} \tag{1.3.7}$$

with probability at least $1-8p^{-c}$ for some constant $c>0$.

The obtained rates of convergence are governed by two sets of parameters: (i) dimensionality parameters: the dimension p, sparseness parameter s (or $\mathcal{S}$), and the sample size n; (ii) internal parameters: the moment μ_q, dependence quantities τ, γ, and the maximum absolute row sum $\|A\|_\infty$. If the internal parameters are assumed to be of a constant order, we can get

$$\|\widehat{A}-A\|_F = O_{\mathbb{P}}\Big(\sqrt{\mathcal{S}}\big(\frac{\log p}{n}\big)^{\frac{1}{2}-\frac{1}{2q-2}}\Big).$$

To ensure consistency, the dimension p can be allowed to increase exponentially with n in view of the mild scaling condition. Compared to [49] with the same constant order of internal parameters, they can only allow the narrower range $p=o(n^c)$ for some $0<c<(q-4)/8$ (cf. Condition 4(i)). For Gaussian autoregressive models, proposition 4.1 of [9] suggests the order in terms of dimension parameters as

$$\|\widehat{A}-A\|_F = O_{\mathbb{P}}\Big(\sqrt{\mathcal{S}}\sqrt{\frac{\log p}{n}}\Big).$$

In the presence of fat tails with the existence of finite q-th moment, our result yields a slightly slower convergence rate characterized by the moment order q and it is closer to their bound when q gets larger.

As an alternative method, the idea of Dantzig-type estimation ([23], [22], [52]) can be modified in the robust way. Let Σ_k denote the autocovariance matrix of the process (X_i) at lag k. The celebrated Yule-Walker equation $A = \Sigma_0^{-1}\Sigma_1$ suggests that a good estimate $\widehat{A}$ should have a small error in terms of $\|\Sigma_0\widehat{A} - \Sigma_1\|_{\max}$. Without direct access to the autocovariance matrices Σ_0 and Σ_1, a natural approach is to find nice estimators for them. [52] used sample autocovariance matrices and enjoyed a nice performance bound under Gaussianity. For fat-tailed cases, we consider the robust estimators of the autocovariance matrices based on the truncated sample:

$$\widehat{\Sigma}_k = \frac{1}{n}\sum_{i=1}^{n} \widetilde{X}_{i-k}\widetilde{X}_i^\top, \text{ for } k = 0, 1.$$

The Dantzig- type estimator is then modified to solving the following convex programming:

$$\widehat{A} = \arg\min_{B\in\mathbb{R}^{p\times p}} |B|_1 \quad \text{s.t.} \quad \|\widehat{\Sigma}_0 B - \widehat{\Sigma}_1\|_{\max} \leq \lambda, \tag{1.3.8}$$

where $\lambda > 0$ is a tuning parameter. Observe that problem (1.3.8) can be solved in parallel, namely, (1.3.8) is equivalent to p subproblems:

$$\widehat{\boldsymbol{a}}_{\cdot j} = \arg\min_{\boldsymbol{b}\in\mathbb{R}^p} |\boldsymbol{b}|_1 \quad \text{s.t.} \quad |\widehat{\Sigma}_0\boldsymbol{b} - \widehat{\Sigma}_1 u_j|_\infty \leq \lambda, \quad j = 1, \ldots, p \tag{1.3.9}$$

for any unit vector u_j. Let $\boldsymbol{a}_{\cdot 1}, \boldsymbol{a}_{\cdot 2}, \ldots, \boldsymbol{a}_{\cdot p}$ be columns of A and denote

$$s^* = \max_{1\leq j\leq p} |\mathrm{supp}(\boldsymbol{a}_{\cdot j})|.$$

We can obtain $\widehat{A}$ by simply concatenating all the columns $\widehat{\boldsymbol{a}}_{\cdot j}$, i.e. $\widehat{A} = (\widehat{\boldsymbol{a}}_{\cdot 1}, \widehat{\boldsymbol{a}}_{\cdot 2}, \ldots, \widehat{\boldsymbol{a}}_{\cdot p})$. The next theorem delivers an upper bound on the statistical accuracy.

Theorem 1.3.4. *Let Assumption (B1) be satisfied. Let $\widehat{A}$ be the solution of (1.3.8) with*

$\nu \asymp \mu_q (n/\log p)^{1/(2q-2)}$ and $\lambda \asymp \mu_q \gamma\tau(\|A\|_1 + 1)[(\log p)/n]^{(q-2)/(2q-2)}$. With probability at least $1 - 8p^{-c'}$ for some constant $c' > 0$, it holds that

$$\|\widehat{A} - A\|_{\max} \le C\mu_q \gamma\tau \|\Sigma_0^{-1}\|_1 (\|A\|_1 + 1)\left(\frac{\log p}{n}\right)^{\frac{1}{2}-\frac{1}{2q-2}}, \tag{1.3.10}$$

$$\|\widehat{A} - A\|_1 \le C'\mu_q \gamma\tau \|\Sigma_0^{-1}\|_1 (\|A\|_1 + 1)s^*\left(\frac{\log p}{n}\right)^{\frac{1}{2}-\frac{1}{2q-2}}. \tag{1.3.11}$$

It is interesting to see that the convergence rate of the modified Dantzig-type estimator has a similar form to that of the robust Lasso estimator developed in Theorem 1.3.3, if the included internal parameters for the process are of a constant order. Both methods involve p parallel programming problems with the lasso-based one concerning row-by-row estimation while the Dantzig method concerning column-by-column instead. The situation $\|A\| < 1$ studied by [52] is the special case where $\gamma = 1$ and $\tau = 1$ in our framework. In their paper, a more flexible sparse condition was imposed: the transition matrix A belongs to a class of weakly sparse matrices defined in terms of strong ℓ^r-ball ($0 \le r < 1$), which was also considered by [14], [113], [22], [21] in estimating covariance and precision matrices. For $r = 0$, it is the exact sparse case and Theorem 1 in [52] implies the dimension parameter order

$$\|\widehat{A} - A\|_1 = O_{\mathbb{P}}\Big(s^*\sqrt{\frac{\log p}{n}}\Big),$$

a bit sharper than our result (1.3.11). There is additional cost for fat-tailed processes with robustness absorbed. We shall remark that we can also derive the bound of $\|\widehat{A} - A\|_1$ accordingly for weakly sparse A based on the result (1.3.10) without any technical difficulty.

1.4 Simulation Study

In this section, we evaluate the finite sample performance of both robust Lasso and Dantzig estimators that are proposed in Section 1.3.2 and compare with the traditional

Lasso and Dantzig methods. Simulation on time series linear regression can be conducted similarly. We consider the model (2.2.2), where ε_{ij}'s are i.i.d. standardized Student's t-distributions with $df = 5$ and 10 respectively. Let $s = \lfloor \log p \rfloor$. For the true transition matrix $A = (a_{ij})$, we consider the following designs.

(1) Banded: $A = (\lambda^{|i-j|}\mathbf{1}\{|i-j| \leq s\})$ and $\lambda = 0.5$.

(2) Block diagonal: $A = \text{diag}\{A_i\}$, where each $A_i \in \mathbb{R}^{s \times s}$ follows the structure in Example 1.2.1 with $\lambda_i \sim Unif(-0.8, 0.8)$.

(3) Toeplitz: $A = (\rho^{|i-j|})$ and $\rho = 0.5$.

(4) Random Sparse: $a_{ii} \sim Unif(-0.8, 0.8)$ and $a_{ij} \sim N(0,1)$ for $(i,j) \in C \subset \{(i,j) : i \neq j\}$ where C is randomly selected and $|C| = s^2$.

The designs in (1), (3) and (4) are further scaled by $2\lambda_{\max}(A)$ to ensure that the spectral radius of the transition matrix is smaller than 1. Thus we have a symmetric sparse and weakly sparse matrix in (1) and (3) respectively, while (2) and (4) generate asymmetric matrices. We take the numerical setup of $n = 50$ and $p = 10, 30, 100$.

In each repetition, we generate a process of length $2n$. We run the estimation procedure in (1.3.4) or (1.3.8) based on $\{X_1, \ldots, X_n\}$ by a two-dimensional grid search for the tuning parameters ν and λ. For each (ν, λ) in the grid, denote the estimator by $\hat{A}(\nu, \lambda)$. Then (ν, λ) is chosen such that $n^{-1} \sum_{t=n+1}^{2n} |X_t - \hat{A}(\nu, \lambda) X_{t-1}|_2^2$, the average prediction error on $\{X_{n+1}, \ldots, X_{2n}\}$, is minimized. The following tables (Table 1.1, Table 1.2 and Table 1.3) reports the average $\|\hat{A} - A\|_F$ (estimation error in Frobenius norm) based on 1000 repetitions. As comparisons, we obtain the results for robust methods and the traditional versions (Lasso estimator in [130] and Dantzig-based estimator in [52]) in different designs.

From statistical perspective, the tables indicate that both of our robust estimation methods outperform the regular Lasso and Dantzig, when the innovation vectors have

Table 1.1. The average of $\|\hat{A} - A\|_F$ based on 1000 repetitions for different methods when $n = 50$ and $p = 10$.

$n = 50, p = 10$	Method	Banded	Block	Toeplitz	Random
$\varepsilon_{ij} \sim t_5$	Lasso	0.764	0.716	0.701	0.752
	Robust-Lasso	0.716	0.644	0.669	0.601
	Dantzig	0.809	0.951	0.725	0.662
	Robust-Dantzig	0.718	0.845	0.685	0.620
$\varepsilon_{ij} \sim t_{10}$	Lasso	0.748	0.703	0.691	0.746
	Robust-Lasso	0.724	0.632	0.665	0.652
	Dantzig	0.768	0.930	0.720	1.414
	Robust-Dantzig	0.709	0.881	0.687	1.368

Table 1.2. The average of $\|\hat{A} - A\|_F$ based on 1000 repetitions for different methods when $n = 50$ and $p = 30$.

$n = 50, p = 30$	Method	Banded	Block	Toeplitz	Random
$\varepsilon_{ij} \sim t_5$	Lasso	1.340	1.804	1.182	1.617
	Robust-Lasso	1.052	1.334	1.074	1.382
	Dantzig	1.276	2.337	1.186	2.175
	Robust-Dantzig	1.265	2.109	1.170	2.034
$\varepsilon_{ij} \sim t_{10}$	Lasso	1.262	1.705	1.176	1.564
	Robust-Lasso	1.050	1.635	1.172	1.383
	Dantzig	2.279	2.100	1.178	2.150
	Robust-Dantzig	2.264	2.049	1.172	2.019

Table 1.3. The average of $\|\hat{A} - A\|_F$ based on 1000 repetitions for different methods when $n = 50$ and $p = 100$.

$n = 50, p = 100$	Method	Banded	Block	Toeplitz	Random
$\varepsilon_{ij} \sim t_5$	Lasso	2.138	3.964	2.145	3.113
	Robust-Lasso	1.993	3.260	2.113	2.901
	Dantzig	2.239	4.409	2.149	3.960
	Robust-Dantzig	2.051	3.988	2.014	3.852
$\varepsilon_{ij} \sim t_{10}$	Lasso	2.235	3.881	2.146	3.047
	Robust-Lasso	2.236	3.342	2.144	2.802
	Dantzig	2.238	4.224	2.148	3.975
	Robust-Dantzig	2.139	4.021	2.143	3.971

fat tail and the transition matrix enjoys a sparsity pattern. The differences became less significant if the tail of the innovation distribution becomes lighter. In a nutshell, our robust methods is more advantageous in tackling non-Gaussian time series.

1.5 Concluding Remarks

Time series regression arises in a wide range of disciplines. Conventional tools are inadequate when it involves high dimensional temporal dependent and fat-tailed data. In this chapter, we develop a novel Bernstein inequality for high dimensional linear processes, with the help of which, we have made contributions towards the robust estimation theory of high dimensional time series regression in the presence of fat tails. The convergence rate depends on the strength of temporal and cross-sectional dependence, the moment condition, the dimension and the sample size. We allow the dimension to increase exponentially with the sample size as a natural requirement of consistency. To perform statistical inference of the estimates such as hypothesis testing and construction of confidence intervals, one needs to develop the deeper result in terms of asymptotic distributional theory. The latter is more challenging and we leave it as future work.

1.6 Proofs of Results in Section 1.2

In this section, we provide the proofs of the results presented in Section 1.2.

Proof of Theorem 1.2.1. We first define a filtration $\{\mathcal{F}_i\}$ with σ-field $\mathcal{F}_i = \sigma(\varepsilon_i, \varepsilon_{i-1}, \dots)$, and the projection operator $P_j(\cdot) = \mathbb{E}(\cdot|\mathcal{F}_j) - \mathbb{E}(\cdot|\mathcal{F}_{j-1})$. Conventionally, it follows that $P_j(G(X_i)) = 0$ for $j \geq i+1$. We can write

$$\sum_{i=1}^{n} G(X_i) - \mathbb{E}G(X_i) = \sum_{j=-\infty}^{n} \Big(\sum_{i=1}^{n} P_j(G(X_i)) \Big) =: \sum_{j=-\infty}^{n} L_j,$$

where $L_j = \sum_{i=1}^{n} P_j(G(X_i))$. By the Markov inequality, for any $\lambda > 0$,

$$\begin{aligned} &\mathbb{P}\bigg(\sum_{i=1}^{n} G(X_i) - \mathbb{E}G(X_i) \geq 2x \bigg) \leq \mathbb{P}\bigg(\sum_{j=-\infty}^{0} L_j \geq x \bigg) + \mathbb{P}\bigg(\sum_{j=1}^{n} L_j \geq x \bigg) \\ \leq\ & \mathrm{e}^{-\lambda x} \mathbb{E}\bigg[\exp\bigg\{ \lambda \sum_{j=-\infty}^{0} L_j \bigg\} \bigg] + \mathrm{e}^{-\lambda x} \mathbb{E}\bigg[\exp\bigg\{ \lambda \sum_{j=1}^{n} L_j \bigg\} \bigg]. \end{aligned} \tag{1.6.1}$$

We shall bound the right-hand side of (1.6.1) with a suitable choice of $\lambda > 0$. Observing that $\{L_j\}_{j \leq n}$ is a sequence of martingale differences with respect to $\{\mathcal{F}_j\}$, we firstly seek an upper bound on $\mathbb{E}[\mathrm{e}^{\lambda L_j} | \mathcal{F}_{j-1}]$. By the Lipschitz condition (1.2.6) and the boundedness of G, it follows that

$$\begin{aligned} |L_j| &\leq \sum_{i=1 \vee j}^{n} \min\left\{ \left| \mathbb{E}\left[G(X_i) \middle| \mathcal{F}_j \right] - \mathbb{E}\left[G(X_i) | \mathcal{F}_{j-1} \right] \right|, 2M \right\} \\ &\leq \sum_{i=1 \vee j}^{n} \min\left\{ g^\top |A_{i-j}| \mathbb{E}\left[|\varepsilon_j - \varepsilon_j'| \middle| \mathcal{F}_j \right], 2M \right\}, \end{aligned} \tag{1.6.2}$$

where ε_j' is an i.i.d. copy of ε_j. For notational convenience, we denote $b_i^\top = g^\top |A_i|$ and

$\eta_j = \mathbb{E}(|\varepsilon_j - \varepsilon_j'||\mathcal{F}_j)$. Then we have

$$|L_j| \le 2M \sum_{i=1\vee j}^{n} \mathbb{I}(b_{i-j}^\top \eta_j \ge 2M) + \sum_{i=1\vee j}^{n} b_{i-j}^\top \eta_j \mathbb{I}(b_{i-j}^\top \eta_j \le 2M) =: I_j + II_j.$$

For $j \le 0$ and $k \ge 2$, by the triangle inequality, it holds that

$$\begin{aligned} \mathbb{E}[|L_j|^k|\mathcal{F}_{j-1}] &\le \left[\left(\mathbb{E}[|I_j|^k|\mathcal{F}_{j-1}]\right)^{1/k} + \left(\mathbb{E}[|II_j|^k|\mathcal{F}_{j-1}]\right)^{1/k}\right]^k \\ &\le \left(\|I_j\|_k + \|II_j\|_k\right)^k. \end{aligned} \tag{1.6.3}$$

Moreover,

$$\|I_j\|_k \le 2M \sum_{i=-j}^{\infty} \left\|\mathbb{I}(b_i^\top \eta_j \ge 2M)\right\|_k \le 2M \sum_{i=-j}^{\infty} \left[\mathbb{P}\big((b_i^\top \eta_j)^2 \ge (2M)^2\big)\right]^{1/k}. \tag{1.6.4}$$

Recall the definitions of γ and τ. We have $|b_i|_1 \le \gamma \rho_0^{i/\tau}$, which implies

$$\mathbb{E}[(b_i^\top \eta_j)^2] \le 2\sigma^2 |b_i|_1^2 \le 2\gamma^2\sigma^2\rho_0^{2i/\tau}, \text{ for all } j.$$

By the Markov inequality, we obtain from (1.6.4) that for $k \ge 2$,

$$\|I_j\|_k \le 2M \left(\frac{\gamma\sigma}{\sqrt{2}M}\right)^{2/k} \frac{\rho_0^{-2j/k\tau}}{1-\rho_0^{2/k\tau}}. \tag{1.6.5}$$

In view of the fact $1 - x \ge -x\log x$ for $x \in (0,1)$, we can further relax the bound in (1.6.5). Applying the Stirling formula, for $k \ge 2$, we can obtain

$$\begin{aligned} \|I_j\|_k^k &\le k^k \tau^k \rho_0^{-2/\tau} \left(\frac{M}{\log(1/\rho_0)}\right)^k \left(\frac{\gamma\sigma}{\sqrt{2}M}\right)^2 \rho_0^{-2j/\tau} \\ &\le \frac{1}{2\sqrt{2\pi}} \left(\frac{\gamma\sigma}{\rho_0 M}\right)^2 k!\tau^k \left(\frac{\mathrm{e}M}{\log(1/\rho_0)}\right)^k \rho_0^{-2j/\tau}. \end{aligned}$$

Define the constants

$$C_1 = \frac{1}{2\sqrt{2\pi}}\rho_0^{-2}, \quad \text{and} \quad C_2 = \frac{\mathrm{e}}{\log(1/\rho_0)}.$$

Then we can simply write

$$\|I_j\|_k^k \le C_1 k! \tau^k C_2^k M^{k-2} \gamma^2 \sigma^2 \rho_0^{-2j/\tau}. \tag{1.6.6}$$

Analogously, for $k \ge 2$, we can also get

$$\|II_j\|_k^k \le \Big[\sum_{i=-j}^{\infty} \big\{\mathbb{E}\big[(b_i^\top \eta_j)^2 \, (2M)^{k-2}\big]\big\}^{1/k}\Big]^k \le C_1 k! \tau^k C_2^k M^{k-2} \gamma^2 \sigma^2 \rho_0^{-2j/\tau}. \tag{1.6.7}$$

By (1.6.3), (1.6.6) and (1.6.7), we have

$$\mathbb{E}[|L_j|^k \big| \mathcal{F}_{j-1}] \le C_1 k! \tau^k (C_2')^k M^{k-2} \gamma^2 \sigma^2 \rho_0^{-2j/\tau}, \tag{1.6.8}$$

where $C_2' = 2C_2 = 2\mathrm{e}/\log(1/\rho_0)$. Now we are ready to derive an upper bound for $\mathbb{E}[\mathrm{e}^{\lambda L_j} \big| \mathcal{F}_{j-1}]$. By the Taylor expansion, we have

$$\mathbb{E}[\mathrm{e}^{\lambda L_j} | \mathcal{F}_{j-1}] = 1 + \mathbb{E}[\lambda L_j | \mathcal{F}_{j-1}] + \sum_{k=2}^{\infty} \frac{1}{k!} \mathbb{E}[\lambda^k L_j^k | \mathcal{F}_{j-1}].$$

Notice that $\mathbb{E}[L_j \big| \mathcal{F}_{j-1}] = 0$. For $0 < \lambda < (C_2' M \tau)^{-1}$, we have

$$\begin{aligned} \mathbb{E}[\mathrm{e}^{\lambda L_j} \big| \mathcal{F}_{j-1}] &\le 1 + C_1 M^{-2} \gamma^2 \sigma^2 \rho_0^{-2j/\tau} \sum_{k=2}^{\infty} \Big(C_2' M \tau \lambda\Big)^k \\ &\le \exp\left\{ \frac{C_1' \gamma^2 \sigma^2 \tau^2 \rho_0^{-2j/\tau} \lambda^2}{1 - C_2' M \tau \lambda} \right\}, \end{aligned}$$

where the constant

$$C_1' = C_1 (C_2')^2 = \frac{1}{2\sqrt{2\pi}} \left(\frac{2\mathrm{e}}{\rho_0 \log(1/\rho_0)} \right)^2,$$

Thus, recursively conditioning on $\mathcal{F}_0, \mathcal{F}_{-1}, \ldots$, we have for $0 < \lambda < (C_2'\tau)^{-1}$,

$$\begin{aligned}
\mathbb{P}\Big(\sum_{j=-\infty}^{0} L_j \geq x\Big) &\leq \mathrm{e}^{-\lambda x}\mathbb{E}\Big[\exp\Big\{\lambda \sum_{j=-\infty}^{0} L_j\Big\}\Big] \\
&\leq \mathrm{e}^{-\lambda x}\exp\bigg\{\frac{C_1'\gamma^2\sigma^2\tau^2(1-\rho_0^{2/\tau})^{-1}\lambda^2}{1-C_2'M\tau\lambda}\bigg\}.
\end{aligned}$$

Specifically, choosing $\lambda = x[C_2'M\tau x + 2C_1'\gamma^2\sigma^2\tau^2(1-\rho_0^{2/\tau})^{-1}]^{-1}$ yields

$$\begin{aligned}
\mathbb{P}\Big(\sum_{j=-\infty}^{0} L_j \geq x\Big) &\leq \exp\bigg\{-\frac{x^2}{4C_1'\gamma^2\sigma^2\tau^2(1-\rho_0^{2/\tau})^{-1} + 2C_2'M\tau x}\bigg\} \\
&\leq \exp\bigg\{-\frac{x^2}{2C_1'\gamma^2\sigma^2\rho_0^{-2}(\log(1/\rho_0))^{-1}\tau^3 + 2C_2'M\tau x}\bigg\} \\
&= \exp\bigg\{-\frac{x^2}{C_1''\tau^3\gamma^2\sigma^2 + 2C_2'M\tau x}\bigg\}, \qquad (1.6.9)
\end{aligned}$$

where $C_1'' = 2C_1'\rho_0^{-2}(\log(1/\rho_0))^{-1}$. We can deal with L_j for $j \geq 1$ by similar arguments and obtain

$$\mathbb{E}[\mathrm{e}^{\lambda L_j} \big| \mathcal{F}_{j-1}] \leq \exp\bigg\{\frac{C_1'\gamma^2\sigma^2\tau^2\lambda^2}{1-C_2'M\tau\lambda}\bigg\} \text{ for } j \geq 1.$$

In a similar way as deriving (1.6.9), it follows that

$$\mathbb{P}\left(\sum_{j=1}^{n} L_j \geq x\right) \leq \exp\bigg\{-\frac{x^2}{C_1''\gamma^2\sigma^2\tau^2 n + 2C_2'M\tau x}\bigg\}. \qquad (1.6.10)$$

Combining (1.6.1), (1.6.9) and (1.6.10), we have

$$\mathbb{P}\Big(\sum_{i=1}^{n} G(X_i) - \mathbb{E}[G(X_i)] \geq x\Big) \leq 2\exp\bigg\{-\frac{x^2}{4C_1''\tau^2(\tau \vee n) + 4C_2'M\tau x}\bigg\},$$

which implies (1.2.7) for $\tau \leq n$. □

Proof of Theorem 1.2.3. We follow the starting steps when proving Theorem 1.2.1. Without assuming G bounded, we have

$$|L_j| \le \sum_{i=1\vee j}^{n} g^\top |A_{i-j}| \mathbb{E}\left[|\varepsilon_j - \varepsilon_j'|\big|\mathcal{F}_j\right] = \sum_{i=1\vee j}^{n} b_{i-j}^\top \eta_j =: d_j^\top \eta_j.$$

For $j \le -\tau$, we have

$$|d_j|_1 \le \sum_{i=1}^{n} |b_{i-j}|_1 \le \gamma \frac{\rho_0^{1/\tau}}{1-\rho_0^{1/\tau}} \cdot \rho_0^{-j/\tau} \le (\log(1/\rho_0))^{-1} \gamma\tau \rho_0^{-j/\tau}. \tag{1.6.11}$$

Note that

$$\mathbb{E}[e^{\lambda|L_j|}|\mathcal{F}_{j-1}] \le \mathbb{E}[e^{\lambda d_j^\top \eta_j}|\mathcal{F}_{j-1}] = \mathbb{E}[e^{\lambda d_j^\top \eta_j}] \le \mathbb{E}[e^{\lambda d_j^\top (|\varepsilon_j| + |\varepsilon_j'|)}]. \tag{1.6.12}$$

Let $\lambda^* = c_0(\log(1/\rho_0))(\gamma\tau)^{-1}$ and $Y_j = \lambda^* d_j^\top (|\varepsilon_j| + |\varepsilon_j'|)\rho_0^{j/\tau}$. By (1.6.11) and (1.6.12), it follows that for any $j \le -\tau$, $\mathbb{E}e^{Y_j} \le \theta^2$ and

$$\begin{aligned}
\mathbb{E}[e^{\lambda^*|L_j|} - 1|\mathcal{F}_{j-1}] &\le \mathbb{E}e^{Y_j \rho_0^{-j/\tau}} - 1 = \int_0^\infty \rho_0^{-j/\tau} e^{x\rho_0^{-j/\tau}} \mathbb{P}(Y_j \ge x) dx \\
&\le \int_0^\infty \rho_0^{-j/\tau} e^{x\rho_0^{-j/\tau}} e^{-x} \theta^2 dx \\
&\le \frac{\rho_0^{-j/\tau}\theta^2}{1-\rho_0^{-j/\tau}} \le \frac{\rho_0^{-j/\tau}\theta^2}{1-\rho_0}.
\end{aligned}$$

Since $\mathbb{E}[L_j|\mathcal{F}_j] = 0$, for any $0 < \lambda \le \lambda^*$,

$$\begin{aligned}
\mathbb{E}[e^{\lambda L_j} - 1|\mathcal{F}_{j-1}] &= \mathbb{E}[e^{\lambda L_j} - \lambda L_j - 1|\mathcal{F}_{j-1}] \\
&\le \mathbb{E}[e^{\lambda|L_j|} - \lambda|L_j| - 1|\mathcal{F}_{j-1}] \\
&\le \mathbb{E}[e^{\lambda^*|L_j|} - \lambda^*|L_j| - 1|\mathcal{F}_{j-1}] \cdot \lambda^2/(\lambda^*)^2 \\
&\le \mathbb{E}[e^{\lambda^*|L_j|} - 1|\mathcal{F}_{j-1}] \cdot \lambda^2/(\lambda^*)^2,
\end{aligned}$$

in view of $e^x - x \le e^{|x|} - |x|$ for any x and when $x > 0$, $(e^{\lambda x} - \lambda x - 1)/\lambda^2$ is increasing with $\lambda \in (0, \infty)$. Using $1 + x \le e^x$, we have

$$\begin{aligned} \mathbb{E}[e^{\lambda L_j} | \mathcal{F}_{j-1}] &\le 1 + \mathbb{E}[e^{\lambda^* |L_j|} - 1 | \mathcal{F}_{j-1}] \cdot \lambda^2 / (\lambda^*)^2 \\ &\le 1 + C_1 \rho_0^{-j/\tau} \gamma^2 \tau^2 \theta^2 \lambda^2 \le \exp\left\{ C_1 \rho_0^{-j/\tau} \gamma^2 \tau^2 \theta^2 \lambda^2 \right\}. \end{aligned}$$

where $C = c_0^{-2} (\log(1/\rho_0))^{-2} / (1 - \rho)$, which implies that

$$\mathbb{P}\bigg(\sum_{j=-\infty}^{-\tau} L_j \ge x \bigg) \le \mathrm{e}^{-\lambda x} \mathbb{E}\bigg[\exp\bigg\{ \lambda \sum_{j=-\infty}^{-1} L_j \bigg\} \bigg] \le \mathrm{e}^{-\lambda x} \exp\left\{ C_1 \gamma^2 \tau^3 \theta^2 \lambda^2 \right\}.$$

with $C_1 = C(\log(1/\rho_0))^{-1} (\rho_0)^{-2}$. For the cases when $j > -\tau$, we use the bound $|d_j|_1 \le (\rho_0 \log(1/\rho_0))^{-1} \gamma \tau$ and obtain $\mathbb{E}[e^{\lambda L_j} | \mathcal{F}_{j-1}] \le 1 + C_2 \gamma^2 \tau^2 \theta^2 \lambda^2$ for $C_2 = C / \rho_0^2$ and

$$\mathbb{P}\bigg(\sum_{j=-\tau+1}^{n} L_j \ge x \bigg) \le \exp\left\{ -\lambda x + C_2 (n + \tau) \gamma^2 \tau^2 \theta^2 \lambda^2 \right\}. \tag{1.6.13}$$

Therefore (1.2.10) follows by choosing

$$\lambda = \min\left\{ \lambda^*, \ \frac{x}{2 C_1 \gamma^2 \tau^3 \theta^2}, \ \frac{x}{2 C_2 (n + \tau) \gamma^2 \tau^2 \theta^2}, \right\}.$$

□

By a slight modification of the Lipschitz condition (2.7.2), we can develop some ancillary results in Corollar 1.6.1 and Corollary 1.2.2, that can be useful in estimating time series regression models. The proof follows similarly from that of Theorem 1.2.1 without extra technical difficulty.

Corollary 1.6.1. *Consider the same setting of the model as in Theorem 1.2.1. Let $G : \mathbb{R}^{2p} \to \mathbb{R}$ be a function with $|G(u)| \le M$ for all $u \in \mathbb{R}^{2p}$. Suppose there exists a vector*

$g = (g_1, \ldots, g_{2p})^\top$ with $g_i \geq 0$ for $1 \leq i \leq 2p$ and $\sum_{i=1}^{2p} g_i = 1$ such that

$$|G(u) - G(v)| \leq \sum_{i=1}^{2p} g_i |u_i - v_i|, \text{ for all } u, v \in \mathbb{R}^{2p}.$$

Then for any $x > 0$, we have

$$\mathbb{P}\Big(\sum_{i=1}^{n} G(X_i, X_{i-1}) - \mathbb{E}G(X_i, X_{i-1}) \geq x \Big) \leq 2 \exp\Big\{ - \frac{x^2}{C_1' n \sigma^2 \gamma^2 \tau^2 + C_2' \tau M x} \Big\}. \quad (1.6.14)$$

Proof of Corollary 1.6.1. It follows from the fact that the $(2p)$-dimensional process $(X_i^\top, X_{i-1}^\top)^\top$ is also linear and satisfies the condition (1.2.3) with γ multiplied by a constant depending on ρ_0 only. □

Proof of Corollary 1.2.2. With a different Lipschitz condition on G, the step (2.7.2) becomes

$$|L_j| \leq \sum_{i=1 \vee j}^{n} \min\{|A^{i-j}\eta_j|_2, 2M\} \leq \sum_{i=1 \vee j}^{n} \min\{\gamma \rho_0^{(i-j)/\tau} |\eta_j|_2, 2M\}.$$

Note that $\mathbb{E}|\eta_j|_2^2 \leq 2p\sigma^2$. For $j \leq -n_0$ where $n_0 = \lceil \tau \log p / \log(1/\rho_0) \rceil$, by similar arguments in deriving (1.6.9), it can be obtained that

$$\mathbb{P}\Big(\sum_{j=-\infty}^{-n_0} L_j \geq x \Big) \leq \exp\Big\{ - \frac{x^2}{C_1 \tau^3 + C_2 M \tau x} \Big\}. \quad (1.6.15)$$

For $j > -n_0$, we have

$$|L_j| \leq 2n_0 M + \sum_{i=j+n_0}^{\infty} \min\{\gamma \rho_0^{(i-j)/\tau} |\eta_j|_2, 2M\}.$$

Similarly as (1.6.8), we can get

$$\begin{aligned} \mathbb{E}[|L_j|^k | \mathcal{F}_{j-1}] &\leq 2^k[(2n_0M)^k + C_1' k! \tau^k (C_2')^k M^{k-2} \gamma^2 \sigma^2] \\ &\leq C_3 (C_4 n_0 M)^k k! (1 + M^{-2} \gamma^2 \sigma^2), \end{aligned}$$

which further implies

$$\mathbb{E}\Big[\exp\Big\{\lambda\sum_{j=-s+1}^{n} L_j\Big\}\Big] \leq \exp\Big\{\frac{C_3C_4^2(M^2+\gamma^2\sigma^2)n_0^2(n_0+n)\lambda^2}{1-C_4n_0M\lambda}\Big\},$$

and

$$\mathbb{P}\left(\sum_{j=-n_0+1}^{n} L_j \geq x\right) \leq \exp\left\{-\frac{x^2}{C_3'(M^2+\gamma^2\sigma^2)n_0^2(n_0+n)+C_4'M\tau(\log p)x}\right\}.$$

Then (1.2.9) follows in view of $n_0 \leq C_{\rho_0}n$. □

Proof of Theorem 1.2.11. Let $\hat{\mu}_j$ be the Huber estimator of μ_j. Following similar arguments of proving Theorem 3.1 in [155], for

$$R_{nj}(a) = \sum_{i=1}^{n}[\phi_\nu(X_{ij}-a) - \mathbb{E}\phi_\nu(X_{ij}-a)],$$

it can be obtained that for any $\delta > 0$ with $\nu^{-1}\delta \leq 1/2$,

$$\mathbb{P}(\hat{\mu}_j - \mu_j \geq \delta) \leq \mathbb{P}(R_{nj}(\mu_j+\delta) \geq n(\delta - 4\nu^{-1}\mu_2^2)).$$

By the Lipschitz continuity of the function ϕ_ν and the uniform bound $|\phi_\nu(x)| \leq \nu$, applying Theorem 1.2.1 to $R_{nj}(\mu_j+\delta)$, it follows that

$$\mathbb{P}(R_{nj}(\mu_j+\delta) \geq y) \leq 2\exp\Big\{-\frac{y^2}{2C_1n\tau^2\gamma^2+C_2\tau\nu y}\Big\}.$$

Then it follows that

$$\mathbb{P}(\hat{\mu}_j - \mu_j \geq \delta) \leq 2x$$

by letting $n(\delta - 4\nu^{-1}\mu_2^2) = y = \tau\gamma\sqrt{2C_1n\log(1/x)} + C_2\tau\nu\log(1/x)$ for $0 < x < 1/\mathrm{e}$. The requirement $\nu^{-1}\delta \leq 1/2$ is met if we choose $\nu = \frac{2\mu^*}{\sqrt{C_2}}\sqrt{\frac{n}{\log(1/x)}}$ for any $\mu^* \geq \mu_2$ and impose the condition

$$(\sqrt{2C_1C_2}\gamma/\mu_2 + 4C_2)\tau\log(1/x) \leq n.$$

For $\delta \leq \delta_n = (\sqrt{2C_1}\gamma + 4\sqrt{C_2}\mu^*)\tau\sqrt{\frac{\log(1/x)}{n}}$, we have $\mathbb{P}(\hat{\mu}_j - \mu_j \geq \delta_n) \leq 2x$. It can also be obtained that $\mathbb{P}(\hat{\mu}_j - \mu_j \leq -\delta_n) \leq 2x$ similarly. By letting $x = p^{-c-1}$, for some $c > 0$, it follows that

$$\mathbb{P}\Big(\max_{1\leq j\leq p}|\hat{\mu}_j - \mu_j| \geq \sqrt{c+1}(\sqrt{2C_1}\gamma + 4\sqrt{C_2}\mu^*)\tau\sqrt{\frac{\log p}{n}}\Big) \leq 4p^{-c}.$$

which further implies (1.2.11). □

1.7 Proofs of Results in Section 1.3

This section includes all the proofs for the results on robust estimation of time series regressions presented in Section 1.3

1.7.1 Proofs of Results in Section 1.3.1

Denote $L_n(\beta) = \frac{1}{n}\sum_{i=1}^n \Phi_\nu((Y_i - X_i^\top\beta)w(X_i))$ and $\phi_\nu(\cdot) = \Phi'_\nu(\cdot)$. Recall $b_0 = b/\lambda_{\min}(B)$ and $\kappa(B) = \lambda_{\max}(B)/\lambda_{\min}(B)$.

Lemma 1.7.1 (Deviation bound)**.** *Let Assumptions (A1) (A2) (A3) in Section 1.3.1 be satisfied. Let* $\nu = c\sigma_\eta(n/\log p)^{1/2}$ *and* $\lambda = Cb_0\sigma_\eta(\log p/n)^{1/2}$ *for a sufficiently large* C*, with probability at least* $1 - 4p^{-c_1}$ *for some* $c_1 > 0$*, it holds that* $|\nabla L_n(\beta^*)|_\infty \leq \lambda$.

Proof. Consider the first component $\nabla L_{n1}(\beta^*)$ of $\nabla L_n(\beta^*)$. We have

$$\nabla L_{n1}(\beta^*) = \frac{1}{n}\sum_{i=1}^n \phi_\nu(\xi_i w(X_i))X_{i1}w(X_i).$$

Note that $|\phi_\nu(x) - \phi_\nu(y)| \leq |x - y|$ and $|\phi_\nu(\xi_i w(X_i))X_{i1}w(X_i)| \leq \nu b_0$. Conditioned on

$\{X_i\}_{i=1}^n$, by Theorem 1.2.1, we have

$$\mathbb{P}\big(|\nabla L_{n1}(\beta^*) - \mathbb{E}[\nabla L_{n1}(\beta^*)]| \geq C' b_0 x \ |(X_i)_i\big) \leq 4p^{-c},$$

for $x = \sigma_\eta \sqrt{\log p/n} + \nu \log p/n$ and some constant $c > 1$. Hence by a union bound, with probability at least $1 - 4p^{-c_1}$ for $c_1 > 0$, it holds that

$$|\nabla L_n(\beta^*) - \mathbb{E}[\nabla L_n(\beta^*)]|_\infty \leq C' b_0 x.$$

As $\mathbb{E}|\phi_\nu(\xi_i w(X_i))| = \mathbb{E}[|\xi_i w(X_i)| \mathbf{1}(|\xi_i w(X_i)| > \nu)] \leq C_\rho \sigma_\eta^2 \nu^{-1}$, we have

$$|\mathbb{E}[\nabla L_{n1}(\beta^*)]| \leq \mathbb{E}|\nabla L_{n1}(\beta^*)| \leq C_\rho b_0 \sigma_\eta^2 \nu^{-1}. \tag{1.7.1}$$

Therefore, choosing $\nu = c\sigma_\eta (n/\log p)^{1/2}$ and $\lambda = C b_0 \sigma_\eta \sqrt{\log p/n}$ ensures that $|\nabla L_n(\beta^*)|_\infty \leq \lambda$ with high probability.

□

Lemma 1.7.2 (RSC condition)**.** *Let Assumptions (A1) (A2) (A3) be satisfied. Assume*

$$b_0(b_0 + \kappa(B)\gamma\sigma_\varepsilon)\tau\sqrt{s}\sqrt{(\log p)^3/n} \to 0.$$

We have the following holds uniformly for all β, such that $|\Delta|_2 \leq \nu/(2b_0)$ and $|\Delta_{S^c}|_1 \leq 3|\Delta_S|_1$ with probability no less than $1 - 4p^{-c_2}$ that

$$L_n(\beta) - L_n(\beta^*) - \nabla L_n(\beta^*)^\top(\beta - \beta^*) \geq \frac{1}{2}\lambda_{\min}(\mathbb{E}[\frac{w^2(X_i)}{2} X_i X_i^\top])|\beta - \beta^*|_2^2. \tag{1.7.2}$$

Proof. Denote $S = \mathrm{supp}(\beta^*)$. We will show that with high probability, (1.7.2) holds uniformly over the set

$$\mathcal{C} := \{\beta : |\beta - \beta^*| \leq \frac{\nu}{2b_0}, |\beta_{S^c} - \beta^*_{S^c}|_1 \leq 3|\beta_S - \beta^*_S|_1\},$$

Let $\mathcal{T}(\beta, \beta^*) = L_n(\beta) - L_n(\beta^*) - \nabla L_n(\beta^*)^\top(\beta - \beta^*)$, then it follows the same argument as Appendix B.3 in [88] that

$$\mathcal{T}(\beta, \beta^*) \geq \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}(w(X_i)X_i^\top(\beta - \beta^*))^2 \mathbf{1}_{A_i},$$

where $A_i = \{\xi_i \leq \nu/2\}$. Denote $\Gamma = \frac{1}{n}\sum_{i=1}^{n} \frac{w(X_i)^2}{2} X_i X_i^\top \mathbf{1}_{A_i}$. For any u such that $|u|_2 \leq 1$, we have

$$u^\top \Gamma u = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}(u^\top X_i w(X_i))^2 \mathbf{1}_{A_i}.$$

Notice that $\frac{1}{2}|(u^\top x w(x))^2 - (u^\top y w(y))^2| \leq b_0(\kappa(B) + 1)|x - y|_2$ and $|(u^\top x w(x))^2| \leq b_0^2$. Conditioned on ξ_i, by Corollary 1.2.2 we have

$$\mathbb{P}(|u^\top \Gamma u - \mathbb{E}[u^\top \Gamma u]| \geq t | (\xi_i)_i) \leq 4\exp\{-c_3 s \log p\},$$

where $t = C b_0(b_0 + \kappa(B)\gamma\sigma_\varepsilon)\tau\sqrt{s}\sqrt{(\log p)^3/n}$ for a sufficiently large C such that $c_3 > 4$. Note that $t \to 0$ by assumption. Following the same spirit of the ε-net argument in lemma 15 of [89], we can obtain that

$$\left|v^\top\big(\Gamma - \mathbb{E}\Gamma\big)v\right| \leq t,\ \forall\, v \in \mathbb{R}^p,\ |v|_0 \leq 2s,\ |v|_2 \leq 1,$$

holds with probability at least

$$1 - 4\exp\big\{2s\log 9 + 2s\log p - c_3 s\log p\big\} \geq 1 - 4p^{-c_2},$$

provided that $p \to \infty$ and a sufficiently large c_3. By Lemma 12 in [89], it further implies that

$$|v^\top(\Gamma - \mathbb{E}\Gamma)v| \le 27t\Big(|v|_2^2 + \frac{|v|_1^2}{s}\Big), \ \forall v \in \mathbb{R}^p. \tag{1.7.3}$$

Denote $\Delta = \beta - \beta^*$, then we have

$$\mathcal{T}(\beta, \beta^*) \ge \Delta^\top\Gamma\Delta \ge \mathbb{E}[\Delta^\top\Gamma\Delta] - 27t(|\Delta|_2^2 + \frac{|\Delta|_1^2}{s}). \tag{1.7.4}$$

Moreover, as $\mathbb{E}|\xi_i|^2 \le C_\rho \sigma_\eta^2$ and $\nu \to \infty$,

$$\begin{aligned}\mathbb{E}[\Delta^\top\Gamma\Delta] &= \mathbb{E}[\frac{w^2(X_i)}{2}(\Delta^\top X_i)^2] \cdot \mathbb{P}(|\xi_i| \le \frac{\nu}{2})\\ &\ge \lambda_{\min}(\mathbb{E}[\frac{w^2(X_i)}{2}X_iX_i^\top])|\Delta|_2^2 \cdot \Big(1 - \frac{4\mathbb{E}|\xi_i|^2}{\nu^2}\Big)\\ &\ge \frac{3}{4}\lambda_{\min}(\mathbb{E}[\frac{w^2(X_i)}{2}X_iX_i^\top])|\Delta|_2^2,\end{aligned}$$

Also, for $\beta \in \mathcal{C}$, $|\Delta|_2^2 + \frac{|\Delta|_1^2}{s} \le 17|\Delta|_2^2$. By (1.7.4), we conclude that

$$\begin{aligned}\mathcal{T}(\beta, \beta^*) &\ge \Big(\frac{3}{4}\lambda_{\min}(\mathbb{E}[\frac{w^2(X_i)}{2}X_iX_i^\top]) - 459t\Big)|\Delta|_2^2\\ &\ge \frac{1}{2}\lambda_{\min}(\mathbb{E}[\frac{w^2(X_i)}{2}X_iX_i^\top])|\Delta|_2^2\end{aligned}$$

□

Proof of Theorem 1.3.1. With Lemma 1.7.1 and Lemma 1.7.2, the proof follows the same spirit as Appendix B.1 of [88] without extra technical difficulty. □

1.7.2 Proofs of Results in Section 1.3.2

We shall first prove Proposition 1.3.2.

Proof of Proposition 1.3.2. If $\lambda_{\max}(A) < 1$, for any $\epsilon > 0$, the matrix $B = A/[\lambda_{\max}(A) + \epsilon]$ has spectral radius strictly less than 1. By Theorem 5.6.12 of [44], B is convergent in the sense that $\lim_{k\to\infty} B^k = 0$. Thus, $\|B^k\| \to 0$ as $k \to \infty$ and there exists some

$N = N(\varepsilon, A)$ such that $\|B^k\| < 1$ for all $k \geq N$, which implies $\|A^k\| \leq [\lambda_{\max}(A) + \epsilon]^k$ for all $k \geq N$. Therefore, given the constant $0 < \rho_0 < 1$ and with an arbitrarily small ϵ with $\lambda_{\max}(A) + \epsilon < 1$, there must exist some finite k such that $\|A^k\| \leq \rho_0$. The proof of the converse is easier by the fact that $[\lambda_{\max}(A)]^k = \lambda_{\max}(A^k) \leq \|A^k\|$ for any k. □

To prove Theorem 1.3.3, we introduce some preparatory lemmas. Define $\widetilde{L}_j(\boldsymbol{b}) = n^{-1}\sum_{i=1}^n (\widetilde{X}_{ij} - \boldsymbol{b}^\top \widetilde{X}_{i-1})^2$ for $1 \leq j \leq p$.

Lemma 1.7.3. *Let Assumption (B1) be satisfied. For* $\nu \asymp \mu_q(n/\log p)^{1/2(q-1)}$ *and*

$$\lambda \asymp \tau\gamma\mu_q(\|A\|_\infty + 1)[(\log p)/n]^{1/2-1/2(q-1)},$$

with probability at least $1 - 4p^{-c_1}$ *for some* $c_1 > 0$, *it holds that*

$$\big|\widetilde{L}_j(\boldsymbol{a}_{j\cdot})\big|_\infty \leq \lambda, \textit{ for all } 1 \leq j \leq p. \tag{1.7.5}$$

Proof of Lemma 1.7.3. We consider the first component of $\nabla\widetilde{L}_j(\boldsymbol{a}_{j\cdot})$, denoted by $\nabla\widetilde{L}_{j1}(\boldsymbol{a}_{j\cdot})$. Other components can be manipulated analogously. Let

$$G(X_i, X_{i-1}) = 2(\widetilde{X}_{i1} - \widetilde{X}_{i-1}^\top \boldsymbol{a}_{j\cdot})\widetilde{X}_{(i-1)1}.$$

Then we can write

$$\nabla\widetilde{L}_{j1}(\boldsymbol{a}_{j\cdot}) = \frac{1}{n}\sum_{i=1}^n G(X_i, X_{i-1}).$$

Notice that $|G| \leq 2(\|A\|_\infty + 1)\nu^2$ and $|G(u) - G(v)| \leq g^\top|u - v|$, where $|g|_1 \leq 4(\|A\|_\infty + 1)\nu$. By Corollary 1.6.1, for $x = c'\gamma\tau\sqrt{(\log p)/n}$, we have

$$\mathbb{P}\bigg(\Big|\nabla\widetilde{L}_{j1}(\boldsymbol{a}_{j\cdot}) - \mathbb{E}\big[\nabla\widetilde{L}_{j1}(\boldsymbol{a}_{j\cdot})\big]\Big| \geq 4\nu(\|A\|_\infty + 1)x\bigg) \leq 4\exp\Big\{-\frac{(c')^2\log p}{2C_1}\Big\}. \tag{1.7.6}$$

In view of $\mathbb{E}[\nabla L_n(\boldsymbol{a}_{j\cdot})] = 0$, the triangle inequality and $|\widetilde{X}_{ij}| \le |X_{ij}|$,

$$\begin{aligned} \big|\mathbb{E}\big[\nabla \widetilde{L}_{j1}(\boldsymbol{a}_{j\cdot})\big]\big| &= \big|\mathbb{E}\big[\nabla \widetilde{L}_{j1}(\boldsymbol{a}_{j\cdot})\big] - \mathbb{E}\big[\nabla L_{j1}(\boldsymbol{a}_{j\cdot})\big]\big| \\ &= 2\mathbb{E}\big[\big|(\widetilde{X}_{ij} - \boldsymbol{a}_{j\cdot}^{\top}\widetilde{X}_{i-1})\widetilde{X}_{(i-1)1} - (X_{ij} - \boldsymbol{a}_{j\cdot}^{\top}X_{i-1})X_{(i-1)1}\big|\big] \\ &\lesssim \mathbb{E}\big[\big|X_{(i-1)1}(\widetilde{X}_{ij} - X_{ij})\big|\big] + \mathbb{E}\big[\big|X_{ij}(X_{(i-1)1} - \widetilde{X}_{(i-1)1})\big|\big] \\ &\quad + |\boldsymbol{a}_{j\cdot}|^{\top}\mathbb{E}\big[\big|X_{(i-1)1}(\widetilde{X}_{i-1} - X_{i-1})\big|\big] \\ &\quad + |\boldsymbol{a}_{j\cdot}|^{\top}\mathbb{E}\big[\big|X_{i-1}(\widetilde{X}_{(i-1)1} - X_{(i-1)1})\big|\big]. \end{aligned} \tag{1.7.7}$$

Since $|\widetilde{X}_{ij} - X_{ij}| \le |X_{ij}|\mathbf{1}\{|X_{ij}| \ge \nu\}$, by Hölder's inequality, we have

$$\begin{aligned} \mathbb{E}\big[\big|X_{(i-1)1}(X_{ij} - \widetilde{X}_{ij})\big|\big] &\le \|\tilde{X}_{(i-1)1}\|_q \cdot \|\tilde{X}_{ij} - X_{ij}\|_{q/(q-1)} \\ &\le \mu_q \|\tilde{X}_{ij} - X_{ij}\|_{q/(q-1)}, \end{aligned}$$

where

$$\|\tilde{X}_{ij} - X_{ij}\|_{q/(q-1)}^{q/(q-1)} \le \mathbb{E}|X_{ij}|^{q/(q-1)}\mathbf{1}\{|X_{ij}| \ge \nu\} \le \mu_q^q \nu^{-q(q-2)/(q-1)}.$$

It then follows that $\mathbb{E}\big[\big|X_{(i-1)1}(X_{ij} - \widetilde{X}_{ij})\big|\big] \le \mu_q^q \nu^{2-q}$. Other terms in (1.7.7) can be dealt with similarly. With the choice of ν, we can get $\big|\mathbb{E}\big[\nabla \widetilde{L}_{j1}(\boldsymbol{a}_{j\cdot})\big]\big| \le c\nu(\|A\|_\infty + 1)x$. Letting $\lambda = C\nu(\|A\|_\infty + 1)x$ for a sufficiently large C and $c' > 2\sqrt{C_1}$, it follows from (1.7.6) that

$$\mathbb{P}\Big(\Big|\nabla \widetilde{L}_{j1}(\boldsymbol{a}_{j\cdot})\Big| \ge \lambda\Big) \le 4\exp\Big\{-\frac{(c')^2 \log p}{2C_1}\Big\}.$$

By the Bonferroni inequality, we have

$$\mathbb{P}\Big(\Big|\nabla \widetilde{L}_{j}(\boldsymbol{a}_{j\cdot})\Big|_\infty \ge \lambda, \text{ for all } 1 \le j \le p\Big) \le 4p^{-c_1}$$

where $c_1 = 2^{-1}C_1^{-1}(c')^2 - 2 > 0$. □

Define a cone $C(S) = \{\Delta \in \mathbb{R}^p : |\Delta_{S^c}|_1 \le 3|\Delta_S|_1\}$ for a subset $S \subseteq \{1, 2, \ldots, p\}$.

We shall verify a restricted eigenvalue (RE) condition on the set $C(S)$ in the lemma below.

Lemma 1.7.4. *Let Assumptions (B1), (B2) and (B3) be satisfied. Choose*

$$\nu \asymp \mu_q (n/\log p)^{1/(2q-2)}.$$

Then for all $\Delta \in C(S)$,

$$\Delta^\top \nabla^2 \widetilde{L}_j(\boldsymbol{a}_{j\cdot})\Delta \geq \frac{\kappa}{2}|\Delta|_2^2 \tag{1.7.8}$$

holds with probability at least $1-4p^{-c_2}$ for some constant $c_2>0$.

Proof of Lemma 1.7.4. Denote $\widetilde{X} = (\widetilde{X}_0, \widetilde{X}_1, \dots, \widetilde{X}_{n-1})^\top$. Then $\nabla^2 \widetilde{L}_j(\boldsymbol{a}_{j\cdot}) = 2\widetilde{X}^\top \widetilde{X}/n =: \Gamma$. We shall first show that with probability at least $1-4p^{-c_2}$ for some positive constant c_2, it holds that

$$\big|v^\top\big(\Gamma - \mathbb{E}\Gamma\big)v\big| \leq t,\ \forall\, v \in \mathbb{R}^p,\ |v|_0 \leq 2s,\ |v|_2 \leq 1, \tag{1.7.9}$$

where $t = c_1 \mu_q \gamma\tau s^2 (\log p/n)^{1/2-1/2(q-1)}$. For any $u \in \mathbb{R}^p$ such that $|u|_2 \leq 1$ and $|u|_0 \leq s$ hence $|u|_1 \leq \sqrt{s}$, write

$$u^\top(\Gamma - \mathbb{E}\Gamma)u = 2n^{-1}\sum_{i=0}^{n-1}(u^\top \widetilde{X}_i)^2 - \mathbb{E}(u^\top \widetilde{X}_i)^2 =: n^{-1}\sum_{i=0}^{n-1} G(X_i) - \mathbb{E}[G(X_i)].$$

Thus, for $G(X_i) = (u^\top \widetilde{X}_i)^2$, we have

$$|G(x) - G(y)| \leq 2|u^\top(x+y)\cdot u^\top(x-y)| \leq 4s\nu g^\top |x-y|,$$

where $|g|_1 \leq 1$. Apply Theorem 1.2.1 to function $G(X_i)/(4s\nu)$ and we have for any fixed

u such that $|u|_2 \leq 1$ and $|u|_0 \leq s$,

$$\mathbb{P}\Big(\big|u^\top(\Gamma - \mathbb{E}\Gamma)u\big| \geq t\Big) \leq 4\exp\big\{-c_3 s^2 \log p\big\}.$$

Following the same spirit of the ε-net argument in lemma 15 of [89], we can obtain that (1.7.9) holds with probability at least

$$1 - 4\exp\big\{2s\log 9 + 2s\log p - c_3 s^2 \log p\big\} \geq 1 - 4p^{-c_2},$$

provided that $p \to \infty$ and a sufficiently large c_3 (or equivalently c_1). By Lemma 12 in [89] and (1.7.9), it further implies that with probability greater than $1 - 4p^{-c_2}$,

$$|v^\top(\Gamma - \mathbb{E}\Gamma)v| \leq 27t\bigg(|v|_2^2 + \frac{|v|_1^2}{s}\bigg),\ \forall v \in \mathbb{R}^p. \tag{1.7.10}$$

Note that when $\Delta \in C(S)$,

$$|\Delta|_1 = |\Delta_S|_1 + |\Delta_{S^c}|_1 \leq 4|\Delta_S|_1 \leq 4\sqrt{s}|\Delta_S|_2 \leq 4\sqrt{s}|\Delta|_2. \tag{1.7.11}$$

Furthermore, some algebra delivers that

$$\begin{aligned}\Delta^\top\mathbb{E}\big[\Gamma\big]\Delta = 2\mathbb{E}[(\widetilde{X}_1^\top\Delta)^2] &\geq 2\Big(\Delta^\top\mathbb{E}[X_1X_1^\top]\Delta - \Delta^\top\mathbb{E}[X_1X_1^\top - \widetilde{X}_1\widetilde{X}_1^\top]\Delta\Big) \\ &\geq 2\kappa|\Delta|_2^2 - 2|\Delta|_1^2\big|\mathbb{E}[X_1X_1^\top - \widetilde{X}_1\widetilde{X}_1^\top]\big|_\infty. \end{aligned} \tag{1.7.12}$$

For any $1 \leq j, k \leq p$, by the triangle and Hölder's inequality,

$$|\mathbb{E}\tilde{X}_{ij}\tilde{X}_{ik} - \mathbb{E}X_{ij}X_{ik}| \leq |\mathbb{E}(\tilde{X}_{ij} - X_{ij})\tilde{X}_{ik})| + |\mathbb{E}(\tilde{X}_{ik} - X_{ik})X_{ij})|.$$

We have

$$
\begin{aligned}
|\mathbb{E}(\tilde{X}_{ij} - X_{ij})\tilde{X}_{ik})| &\leq \|\tilde{X}_{ik}\|_q \cdot \|\tilde{X}_{ij} - X_{ij}\|_{q/(q-1)} \\
&\leq \mu_q \|\tilde{X}_{ij} - X_{ij}\|_{q/(q-1)},
\end{aligned}
$$

where

$$
\|\tilde{X}_{ij} - X_{ij}\|_{q/(q-1)}^{q/(q-1)} \leq \mathbb{E}|X_{ij}|^{q/(q-1)} \mathbf{1}\{|X_{ij}| \geq \nu\} \leq \mu_q^q \nu^{-q(q-2)/(q-1)}.
$$

It then follows that $|\mathbb{E}(\tilde{X}_{ij} - X_{ij})\tilde{X}_{ik}| \leq \mu_q^q \nu^{2-q}$. We can also deal with $|\mathbb{E}(\tilde{X}_{ik} - X_{ik})X_{ij})|$ similarly. As a result, we have the bias

$$
|\mathbb{E}[\tilde{X}_{ij}\tilde{X}_{ik} - X_{ij}X_{ik}]| \leq 2\mu_q^q \nu^{2-q} \leq C\mu_q^2 \big(\frac{\log p}{n}\big)^{\frac{1}{2} - \frac{1}{2q-2}}. \tag{1.7.13}
$$

By (1.7.11), (1.7.12) and (1.7.13), it follows that

$$
\Delta^\top \mathbb{E}\big[\Gamma\big]\Delta \geq 2\kappa|\Delta|_2^2 - 16Cs\mu_q^2 \big(\frac{\log p}{n}\big)^{\frac{1}{2} - \frac{1}{2q-2}} |\Delta|_2^2 \geq \kappa|\Delta|_2^2. \tag{1.7.14}
$$

Recall that $t = c_1 \mu_q \gamma \tau s^2 (\log p/n)^{1/2 - 1/q} \to 0$ by Assumption (B3). Combining (1.7.10) and (1.7.14), we can establish the following RE condition

$$
\nabla^2 L_j(\boldsymbol{a}_{j\cdot}) \geq \kappa|\Delta|_2^2 - 27t(|\Delta|_2^2 + |\Delta|_1^2/s) \geq \kappa|\Delta|_2^2 - 459t|\Delta|_2^2 \geq \frac{\kappa}{2}|\Delta|_2^2,
$$

for all $\Delta \in C(S)$ with probability no less than $1 - 4p^{-c_2}$. □

Proof of Theorem 1.3.3. Let $\widehat{\Delta}_j = \widehat{\boldsymbol{a}}_{j\cdot} - \boldsymbol{a}_{j\cdot}$ for $j = 1, \ldots, p$. As the solution of (1.3.5), $\widehat{\boldsymbol{a}}_{j\cdot}$ satisfies

$$
\widetilde{L}_j(\widehat{\boldsymbol{a}}_{j\cdot}) + \lambda|\widehat{\boldsymbol{a}}_{j\cdot}|_1 \leq \widetilde{L}_j(\boldsymbol{a}_{j\cdot}) + \lambda|\boldsymbol{a}_{j\cdot}|_1,
$$

which together with convexity implies,

$$0 \le \widetilde{L}_j(\widehat{\boldsymbol{a}}_{j\cdot}) - \widetilde{L}_j(\boldsymbol{a}_{j\cdot}) - \langle \nabla \widetilde{L}_j(\boldsymbol{a}_{j\cdot}), \widehat{\Delta}_j \rangle \le \lambda(|\boldsymbol{a}_{j\cdot}|_1 - |\widehat{\boldsymbol{a}}_{j\cdot}|_1) + \big|\nabla \widetilde{L}_j(\boldsymbol{a}_{j\cdot})\big|_\infty |\widehat{\Delta}_j|_1. \quad (1.7.15)$$

Denote by A and B the events in Lemma 1.7.3 and Lemma 1.7.4 respectively. Then $\mathbb{P}(A \cap B) = 1 - \mathbb{P}(A^c \cup B^c) \ge 1 - 8p^{-c}$ for $c = \min\{c_1, c_2\}$. Conditioned on the event A, (1.7.15) implies

$$\begin{aligned} 0 &\le |\boldsymbol{a}_{j\cdot,S}|_1 - |\widehat{\boldsymbol{a}}_{j\cdot,S}|_1 - |\widehat{\boldsymbol{a}}_{j\cdot,S^c}|_1 + \frac{1}{2}|\widehat{\Delta}_j|_1 \\ &\le |\widehat{\Delta}_{j,S}|_1 - |\widehat{\Delta}_{j,S^c}|_1 + \frac{1}{2}|\widehat{\Delta}_j|_1 = \frac{3}{2}|\widehat{\Delta}_{j,S}|_1 - \frac{1}{2}|\widehat{\Delta}_{j,S^c}|_1, \end{aligned}$$

which further implies $\widehat{\Delta}_j \in C(S)$ for all $1 \le j \le p$. Conditioned on the event B, by (1.7.5) and the second inequality in (1.7.15), we have

$$\frac{\kappa}{2}|\widehat{\Delta}_j|_2^2 \le \big(\lambda + \big|\nabla L_n(\boldsymbol{a}_{j\cdot})\big|_\infty\big)|\widehat{\Delta}_j|_1 \le 6\sqrt{s}\lambda|\widehat{\Delta}_j|_2. \quad (1.7.16)$$

This immediately shows for all $1 \le j \le p$

$$|\widehat{\Delta}_j|_2 \le \frac{12\sqrt{s}\lambda}{\kappa} \asymp \mu_q \gamma \tau (\|A\|_\infty + 1)\sqrt{s}\left(\frac{\log p}{n}\right)^{\frac{1}{2} - \frac{1}{2q-2}} \quad (1.7.17)$$

as well as

$$|\widehat{\Delta}_j|_1 \lesssim \mu_q \gamma \tau s (\|A\|_\infty + 1)\left(\frac{\log p}{n}\right)^{\frac{1}{2} - \frac{1}{2q-2}}.$$

Hence, (1.3.6) follows in view of $\|\widehat{A} - A\|_\infty = \max_j |\widehat{\Delta}_j|_1$. Moreover, if we consider the estimation of $\mathbf{Vec}(A) = (\boldsymbol{a}_{1\cdot}^\top, \boldsymbol{a}_{2\cdot}^\top, \ldots, \boldsymbol{a}_{p\cdot}^\top)^\top \in \mathbb{R}^{p^2}$ with the sparsity parameter $\mathcal{S} = \sum_{i=j}^p s_j$, by Assumption (B3') and similar arguments of verifying the RE condition in Lemma 1.7.4, (1.7.8) becomes

$$2\Delta^\top \Big[I_p \otimes \Big(\frac{\widetilde{X}^\top \widetilde{X}}{n} \Big) \Big] \Delta \geq \frac{\kappa}{2} |\Delta|_2^2, \quad \text{for all } \Delta \in \mathbb{R}^{p^2}.$$

Thus, similarly as (1.7.17), (1.3.7) follows. □

Next we shall concern the robust Dantzig-type estimator.

Lemma 1.7.5. *Let Assumption (B1) be satisfied. Choose the truncation parameter*

$$\nu \asymp \mu_q (n/\log p)^{1/(2q-2)}.$$

Let $\lambda \asymp \mu_q \gamma \tau (\|A\|_1 + 1)[(\log p)/n]^{(q-2)/(2q-2)}$. *Then with probability at least* $1 - 8p^{-c'}$ *for some constant* $c' > 0$, *it holds that*

$$\|\widehat{\Sigma}_0 - \Sigma_0\|_{\max} \leq \lambda_0 \quad \textit{and} \quad \|\widehat{\Sigma}_1 - \Sigma_1\|_{\max} \leq \lambda_0.$$

Proof of Lemma 1.7.5. Let $\lambda_0 = C\mu_q \tau \gamma [(\log p)/n]^{(q-2)/(2q-2)}$ for a sufficiently large constant C. Applying Theorem 1.2.1 to the (m, l)-th entry of $\widehat{\Sigma}_0$, we have

$$\mathbb{P}\Big(\frac{1}{n} \Big| \sum_{i=1}^n \widetilde{X}_{im} \widetilde{X}_{il} - \mathbb{E} \widetilde{X}_{im} \widetilde{X}_{il} \Big| \geq \lambda_0 \Big) \leq 4 \exp\Big\{ - \frac{c^2 \log p}{2C_1} \Big\} = 4p^{-c^2/(2C_1)}.$$

By (1.7.13) in the proof of Lemma 1.7.4, we see that

$$\Big| \mathbb{E} \widetilde{X}_{im} \widetilde{X}_{il} - \mathbb{E} X_{im} X_{il} \Big| \leq c\mu_q^2 \big(\frac{\log p}{n} \big)^{\frac{1}{2} - \frac{1}{2q-2}} \leq \lambda_0.$$

Therefore,

$$\mathbb{P}\Big(\frac{1}{n} \Big| \sum_{i=1}^n \widetilde{X}_{im} \widetilde{X}_{il} - \mathbb{E}[X_{im} X_{il}] \Big| \geq \lambda_0 \Big)$$

$$\leq \mathbb{P}\Big(\frac{1}{n}\Big|\sum_{i=1}^{n}\widetilde{X}_{im}\widetilde{X}_{il}-\mathbb{E}[\widetilde{X}_{im}\widetilde{X}_{il}]\Big|\geq C_2\lambda_0\Big)\leq 4p^{-C_3}$$

for some $C_3 > 1$. Taking a union bound yields

$$\mathbb{P}(\|\widehat{\Sigma}_0-\Sigma_0\|_{\max}\geq\lambda_0)\leq 4p^{-c'},$$

where $c' = C_3 - 1 > 0$. By Corollary 1.6.1, similar arguments apply to $\widehat{\Sigma}_1$, which delivers $\|\widehat{\Sigma}_1-\Sigma_1\|_{\max}\leq\lambda_0$ with probability at least $1-4p^{-c'}$. In conclusion, it holds simultaneously that $\|\widehat{\Sigma}_0-\Sigma_0\|_{\max}\leq\lambda_0$ and $\|\widehat{\Sigma}_1-\Sigma_1\|_{\max}\leq\lambda_0$ with probability at least $1-8p^{-c'}$. □

Proof of Theorem 1.3.4. We first show that A is feasible to the convex programming (1.3.8) for $\lambda = (\|A\|_1+1)\lambda_0$ with high probability. By the Yule-Walker equation and Lemma 1.7.5, we have

$$\begin{aligned}\|\widehat{\Sigma}_0A-\widehat{\Sigma}_1\|_{\max}&\leq\|\widehat{\Sigma}_0A-\Sigma_1\|_{\max}+\|\Sigma_1-\widehat{\Sigma}_1\|_{\max}\\&\leq\|\widehat{\Sigma}_0-\Sigma_0\|_{\max}\|A\|_1+\|\Sigma_1-\widehat{\Sigma}_1\|_{\max}\leq\lambda,\end{aligned}$$

with probability no less than $1-8p^{-c'}$. Therefore, conditioned on the event in Lemma 1.7.5, we conclude that $|\widehat{\boldsymbol{a}}_{\cdot j}|_1\leq|\boldsymbol{a}_{\cdot j}|_1$ for all $j=1,\ldots,p$ and hence $\|\widehat{A}\|_1\leq\|A\|_1$. Then we have

$$\begin{aligned}\|\widehat{A}-A\|_{\max}&=\|\Sigma_0^{-1}(\Sigma_0\widehat{A}-\widehat{\Sigma}_1+\widehat{\Sigma}_1-\Sigma_1)\|_{\max}\\&\leq\|\Sigma_0^{-1}(\Sigma_0\widehat{A}-\widehat{\Sigma}_0\widehat{A}+\widehat{\Sigma}_0\widehat{A}-\widehat{\Sigma}_1)\|_{\max}+\|\Sigma_0^{-1}(\widehat{\Sigma}_1-\Sigma_1)\|_{\max}\\&\leq\|\Sigma_0^{-1}\|_1\|\Sigma_0-\widehat{\Sigma}_0\|_{\max}\|\widehat{A}\|_1+\|\Sigma_0^{-1}\|_1\|\widehat{\Sigma}_0\widehat{A}-\widehat{\Sigma}_1\|_{\max}\\&\quad+\|\Sigma_0^{-1}\|_1\|\widehat{\Sigma}_1-\Sigma_1\|_{\max}.\end{aligned}$$

By Lemma 1.7.5 and the feasibility of $\widehat{A}$, we have

$$\|\widehat{A} - A\|_{\max} \leq \|\Sigma_0^{-1}\|_1(\lambda_0\|A_1\| + \lambda + \lambda_0) = 2\|\Sigma_0^{-1}\|_1\lambda.$$

Now we shall bound $\|\widehat{A} - A\|_1$ from above. Denote by S_j the support of $\boldsymbol{a}_{\cdot j}$ for $j = 1, \ldots, p$. Then for any $1 \leq j \leq p$, we have

$$\begin{aligned} |\widehat{\boldsymbol{a}}_{\cdot j} - \boldsymbol{a}_{\cdot j}|_1 &= \left|\widehat{\boldsymbol{a}}_{\cdot j,S_j} - \boldsymbol{a}_{\cdot j,S_j}\right|_1 + \left|\widehat{\boldsymbol{a}}_{\cdot j}\right|_1 - \left|\widehat{\boldsymbol{a}}_{\cdot j,S_j}\right|_1 \\ &\leq \left|\widehat{\boldsymbol{a}}_{\cdot j,S_j} - \boldsymbol{a}_{\cdot j,S_j}\right|_1 + \left|\boldsymbol{a}_{\cdot j}\right|_1 - \left|\widehat{\boldsymbol{a}}_{\cdot j,S_j}\right|_1 \\ &\leq 2\left|\widehat{\boldsymbol{a}}_{\cdot j,S_j} - \boldsymbol{a}_{\cdot j,S_j}\right|_1 \leq 4s^*\|\Sigma_0^{-1}\|_1\lambda. \end{aligned} \tag{1.7.18}$$

Since (1.7.18) holds for all $1 \leq j \leq p$, we conclude that

$$\|\widehat{A} - A\|_1 \leq 4s^*\|\Sigma_0^{-1}\|_1\lambda \lesssim \mu_q s^*\gamma\tau\|\Sigma_0^{-1}\|_1(\|A\|_1 + 1)\left(\frac{\log p}{n}\right)^{\frac{1}{2}-\frac{1}{2q-2}}.$$

□

Chapter 1, in part, is a reprint of the material in the paper "A Bernstein-type Inequality for High Dimensional Linear Processes with Applications to Robust Estimation of Time Series Regressions", Liu, Linbo and Zhang, Danna. This paper is currently under minor revision at *Statistica Sinica*. The dissertation author was the primary investigator and author of this paper.

Chapter 2

Simultaneous Inference of High-dimensional non-Gaussian Vector Autoregressive Models

2.1 Introduction

High-dimensional statistics become increasingly important due to the rapid development of information technology in the past decade. In this chapter, we are primarily interested in conducting simultaneous inference via de-biased M-estimator on the transition matrices in a high-dimensional vector autoregressive model with non-Gaussian innovations. An extensive body of work has been proposed on estimation and inference on the coefficient vector in linear regression setting and we refer readers to [19] for an overview of recent development in high-dimensional statistical techniques. M-estimator is one of the most popular tools among them, which has been proved a success in signal estimation ([102]), support recovery ([90]), variable selection ([159]) and robust estimation with heavy-tailed noises using nonconvex loss functions ([86]). As a penalized M-estimator, Lasso ([130]) also plays an important role in estimating transition coefficients in high-dimensional VAR models beyond linear regression; see for example [58], [100], [9] among others. Another line of work is to achieve such estimation tasks by Dantzig selector ([23]). [52] proposed a new approach to estimating the transition matrix via Dantzig-type estimator and solved a linear programming problem. They remarked that this estimation procedure enjoys many advantages including computational efficiency and weaker assumptions on the transition matrix. However, the aforementioned literature mainly discussed the scenario where Gaussian or sub-Gaussian noises are in presence.

To deal with the heavy-tailed errors, regularized robust methods have been widely studied. For instance, [81] proposed an ℓ_1-regularized quantile regression method in low dimensional setting and devised an algorithm to efficiently solve the proposed optimization problem. [145] studied penalized quantile regression from the perspective of variable selection. However, quantile regression and least absolute deviation regression can be significantly different from the mean function, especially when the distribution of noise is asymmetric. To overcome this issue, [38] developed robust approximation Lasso (RA-Lasso) estimator based on penalized Huber loss and proved the feasibility of RA-Lasso in estimation of high-dimensional mean regression. Apart from linear regression setting, [155] also used Huber loss to obtain a consistent estimate of mean vector and covariance matrix for high-dimensional time series. Also, robust estimation of the transition coefficients was studied in [85] via two types of approaches: Lasso-based and Dantzig-based estimator.

Besides estimation, recent research effort also turned to high-dimensional statistical inference, such as performing multiple hypothesis testing and constructing simultaneous confidence intervals, both for regression coefficients and mean vectors of random processes. To tackle the high dimensionality, the idea of low dimensional projection was exploited by numerous popular literature. For instance, [61], [133], [154] constructed de-sparsifying Lasso by inverting the Karush-Kuhn-Tucker (KKT) condition and derived asymptotic distribution for the projection of high-dimensional parameters onto fixed-dimensional space. As an extension of the previous techniques, [87] proposed the asymptotic theory of one-step estimator, allowing the presence of non-Gaussian noises. Employing Gaussian approximation theory ([27]), [157] proposed a bootstrap-assisted procedure to conduct simultaneous statistical inference, which allowed the number of testing to greatly surpass the number of observations as a significant improvement. Although a huge body of work has been completed for the inference of regression coefficients, there have been limited research on the generalization of these theoretical properties to time series, perhaps due to the technical difficulty when generalizing Gaussian approximation results to dependent random variables. [156] adopted the framework of functional dependence measures ([142]) to account for temporal dependency and provided Gaussian approximation results for general time series. They also showed, as an application, how to construct simultaneous confidence intervals for mean vectors of high-dimensional random processes with asymptotically correct coverage probabilities.

In this chapter, we consider simultaneous inference of transition coefficients in possibly non-Gaussian vector autoregressive (VAR) models with lag d:

$$X_i = A_1 X_{i-1} + A_2 X_{i-2} + \cdots + A_d X_{i-d} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $X_i \in \mathbb{R}^p$ is the time series, $A_i \in \mathbb{R}^{p \times p}$, $i = 1, \ldots, d$ are the transition matrices, and $\varepsilon_i \in \mathbb{R}^p$ are the innovation vectors. We allow the dimension p to exceed the number of observations n, or even $\log p = o(n^b)$ for some $b > 0$, as is commonly assumed in high-dimensional regime. Different from many other work, we do not impose Gaussianity or sub-Gaussianity assumptions on the noise terms ε_i.

We are particularly interested in the following simultaneous testing problem:

$$H_0 : A_i = A_i^0, \quad \text{for all } i = 1, \ldots, d$$

versus the alternative hypothesis

$$H_1 : A_i \neq A_i^0, \quad \text{for some } i = 1, \ldots, d.$$

It's worth mentioning that the above problems still have p^2 null hypotheses to verify even if the lag $d = 1$. We propose to build a de-biased estimator $\check{\beta}$ from some consistent pilot estimator $\widehat{\beta}$ (for example, the one provided in [85]). There are a few challenges when we prove the feasibility of de-biased estimator as well as its theoretical guarantees: (i) VAR models display temporal dependency across observations, which makes the majority of probabilistic tools such as classic Bernstein inequality and Gaussian approximation inapplicable. (ii) Fat-tailed innovations ε_i imply fat-tailed x_i in VAR model, while robust methods regarding linear regression can assume ε_i to have heavy-tail but x_i remains sub-Gaussian ([38] and [157]). (iii) We hope our simultaneous inference procedure to work in ultra-high dimensional regime, where p can grow exponentially fast in n. As a result, these challenges inspire us to establish a new Bernstein-type inequality (section 2.3) and Gaussian approximation results (section 2.4) under the framework of VAR model. Also, we will adopt the definition of spectral decay index to capture the dependency among time series data, as in [85].

This chapter is organized as follows. In section 2.2, we first present more details and some preparatory definitions of VAR models and propose the test statistics for simultaneous inference via de-biased estimator, which is constructed through a robust loss function and a weight function on x_i. The main result delivering critical values for such test statistics by multiplier bootstrap is given in section 2.2.4. In section 2.3, we complete the estimation of multiple statistics by establishing a Bernstein inequality. A thorough discussion of Gaussian approximation and its derivation under VAR model are presented in section 2.4. Some numerical experiments are conducted in section 2.5 to assess the empirical performance of the multiplier bootstrap procedure.

Finally, we introduce some notation. For a vector $\beta = (\beta_1, \ldots, \beta_p)^\top$, let $|\beta|_1 = \sum_i |\beta_i|$, $|\beta|_2 = (\sum_i \beta_i^2)^{1/2}$ and $|\beta|_\infty = \max_i |\beta_i|$ be its $\ell_1, \ell_2, \ell_\infty$ norm respectively. For a matrix $A = (a_{ij})_{1 \le i,j \le p}$, let λ_i, $i = 1, \ldots, p$, be its eigenvalues and $\lambda_{\max}(A)$, $\lambda_{\min}(A)$ be its maximum and minimum eigenvalues respectively. Also let $\rho(A) = \max_i |\lambda_i|$ be the spectral radius. Denote $\|A\|_1 = \max_j \sum_i |a_{ij}|$, $\|A\|_\infty = \max_i \sum_j |a_{ij}|$, and spectral norm

$\|A\| = \|A\|_2 = \sup_{|x|_2 \neq 0} |Ax|_2/|x|_2$. Moreover, let $\|A\|_{\max} = \max_{i,j} |a_{ij}|$ be the entry-wise maximum norm. For a random variable X and $q > 0$, define $\|X\|_q = (\mathbb{E}[X^q])^{1/q}$. For two real numbers x, y, set $x \vee y = \max(x, y)$. For two sequences of positive numbers $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ if there exists some constant $C > 0$, such that $a_n/b_n \leq C$ as $n \to \infty$, and also write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. We use $c_0, c_1, \ldots$ and $C_0, C_1, \ldots$ to denote some universal positive constants whose values may vary in different context. Throughout the chapter, we consider the high-dimensional regime, allowing the dimension p to grow with the sample size n, that is, we assume $p = p_n \to \infty$ as $n \to \infty$.

2.2 Main Results

2.2.1 Vector autoregressive model

Consider a VAR(d) model:

$$X_i = A_1 X_{i-1} + A_2 X_{i-2} + \cdots + A_d X_{i-d} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.2.1}$$

where $X_i = (X_{i1}, X_{i2}, \ldots, X_{ip}) \in \mathbb{R}^p$ is the random process of interests, $A_i \in \mathbb{R}^{p \times p}$, $i = 1, \ldots, d$, are the transition matrices and ε_i, $i \in \mathbb{Z}$, are i.i.d. innovation vectors with zero mean and symmetric distribution, i.e. $\varepsilon_i = -\varepsilon_i$ in distribution, for all $i \in \mathbb{Z}$. By a rearrangement of variables, VAR(d) models can be formulated as VAR(1) models (see [85]). Therefore, without loss of generality, we shall work with VAR(1) models:

$$X_i = AX_{i-1} + \varepsilon_i, \quad i = 1, \ldots, n. \tag{2.2.2}$$

This type of random process has a wide range of application, such as finance development ([118]), economy ([65]) and exchange rate dynamics ([144]).

To ensure model stationarity, we assume that the spectral radius $\rho(A) < 1$ throughout the chapter, which is also the sufficient and necessary condition for a VAR(1) model to be stationary. However, a more restrictive condition that $\|A\| < 1$ is always assumed in most of the earlier work. See for example, [52], [89] and [101]. For a non-symmetric matrix A, it could happen that $\|A\| \geq 1$ while $\rho(A) < 1$. To fill the gap between $\rho(A)$ and $\|A\|$, [9] proposed stability measures for high-dimensional time series to capture temporal and cross-section dependence via the spectral density function. In a more recent work, [85] defined spectral decay index to connect $\rho(A)$ with $\|A\|$ from a different point of view. In this chapter, we will adopt the framework of spectral decay index in [85].

Definition 2.2.1. For any matrix $A \in \mathbb{R}^{p\times p}$ such that $\rho(A) < 1$, define the *spectral decay index* as

$$\tau = \min\{t \in \mathbb{Z}^+ : \|A^t\|_\infty < \rho\} \tag{2.2.3}$$

for some constant $0 < \rho < 1$.

Remark 2.2.2. Note that in (2.2.3), we use L_∞ norm, while spectral norm is considered in [85]. However, the spectral decay index shares many properties even if defined in different matrix norms. Some of them are summarized as follows. For any matrix A with $\rho(A) < 1$, finite spectral decay index τ exists. In general, τ may not be of constant order when the dimension p increases. Technically speaking, we need to explicitly write $\tau = \tau_p$ to capture the dependence on p. However, in the rest of the chapter, we simply write τ for ease of notation. For more analysis of spectral decay index, see section 2 of [85].

Next, we are interested in building some estimators of A for which we could establish asymptotic distribution theory. This allows one to conduct statistical inference, such as finding simultaneous confidence interval. There have been some work on the robust estimation only. [85] provides both a Lasso-type estimator and a Dantzig-type estimator to consistently estimate the transition coefficient A given $\{X_i\}$, under very mild moment condition on X_i and ϵ_i. It turns out that both Lasso-type and Dantzig-type estimators are not unbiased for estimating the transition matrix, thus insufficient for tasks like statistical inference. Therefore, one needs to develop more refined method to establish results in terms of asymptotic distributional theory. In the following sections, we will construct a de-biased estimator based on the existing one and derive the limiting distribution for the de-biased estimator.

Unlike many other existing work ([52], [9], etc.), we do not require ε_i to be Gaussian or sub-Gaussian. Instead, it could happen that the innovations ε_i only have some finite moments, which makes the standard techniques for estimation and inference invalid.

2.2.2 De-biased estimator

In this section, we construct a de-biased estimator using the techniques introduced in [13]. To fix the idea, let $a_j^\top$ be the j-th row of A and $\beta^* = \mathrm{Vec}(A) = (a_1^\top, a_2^\top, \ldots, a_p^\top)^\top \in \mathbb{R}^{p^2}$. Suppose we are given a consistent, possibly biased, estimator $\widehat{\beta}$ of β^*, i.e. $|\widehat{\beta} - \beta^*| = o(1)$ (for example, Lasso-type or Dantzig-type estimators in [85]). Define a loss function $L : \mathbb{R}^{p^2} \to \mathbb{R}$ as

$$L_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{p} \ell(X_{ik} - X_{i-1}^\top \beta_k) w(X_{i-1}), \tag{2.2.4}$$

where $\beta = (\beta_1^\top, \ldots, \beta_p^\top)^\top$ with $\beta_k \in \mathbb{R}^p$ for $1 \le k \le p$, the weight function

$$w(x) = \min\left\{1, \frac{T^3}{|x|_\infty^3}\right\}$$

for some threshold $T > 0$ to be determined later, and the robust loss function $\ell(x)$ satisfies:

(i) $\ell(x)$ is a thrice differentiable convex and even function.

(ii) For some constant $C > 0$, $|\ell'|, |\ell''|, |\ell^{(3)}| \le C$.

We give two examples of such loss functions from [105] that satisfy the above conditions.

Examples 2.2.3 (Smoothed huber loss I).

$$\ell(x) = \begin{cases} x^2/2 - |x|^3/6 & \text{if } |x| \le 1, \\ |x|/2 - 1/6 & \text{if } |x| > 1. \end{cases}$$

Examples 2.2.4 (Smoothed huber loss II).

$$\ell(x) = \begin{cases} x^2/2 - x^4/24, & \text{if } |x| \leq \sqrt{2}, \\ \big(2\sqrt{2}/3\big)|x| - 1/2, & \text{if } |x| > \sqrt{2}. \end{cases}$$

Direct calculation shows that everywhere twice differentiable and almost everywhere thrice differentiable. Also, the derivative of first three orders are bounded in magnitude. We mention that generalization to other loss functions that does not satisfy the differentiability conditions (for example, huber loss) may be derived under more refined arguments, but will be omitted in this chapter.

Denote by $\psi(x) = \ell'(x)$ the derivative of $\ell(x)$, then $\psi(x)$ is twice differentiable by condition (i) and $|\psi(x)| \leq C$ for all $x \in \mathbb{R}$ by condition (ii). Let $\mu = (\mu_1, \ldots, \mu_p)^\top \in \mathbb{R}^p$ with $\mu_k = \mathbb{E}[\psi'(\varepsilon_{ik})]$ and $\mu^{-1} = (\mu_1^{-1}, \ldots, \mu_p^{-1})^\top$. Let $\widehat{\mu} = (\widehat{\mu}_1, \ldots, \widehat{\mu}_p)$ be the estimate of μ with $\widehat{\mu}_k = \frac{1}{n}\sum_{i=1}^n \psi'(\widehat{\varepsilon}_{ik})$, where $\widehat{\varepsilon}_{ik} = X_{ik} - X_{i-1}^\top \widehat{\beta}_k$. Let $\Sigma_x = \mathbb{E}[X_i X_i^\top w(X_i)] \in \mathbb{R}^{p\times p}$ be the weighted covariance matrix and $\Omega_x = \Sigma_x^{-1} \in \mathbb{R}^{p\times p}$ be the weighted precision matrix. Denote by $\widehat{\Sigma}_x = n^{-1}\sum_{i=1}^n X_{i-1}X_{i-1}^\top w(X_{i-1})$ the weighted sample covariance. Furthermore, suppose that $\widehat{\Omega}_x$ is a suitable approximation of the weighted precision matrix Ω_x (e.g., CLIME estimator introduced by [22]), as will be discussed in section 2.3. To ensure the validity of such estimator, the sparsity of each row of Ω_x is always assumed due to high dimensionality. Now we introduce a few more notations:

$$\Sigma = \mathrm{diag}(\mu) \otimes \Sigma_x = \begin{bmatrix} \mu_1\Sigma_x & 0 & 0 & \ldots & 0 \\ 0 & \mu_2\Sigma_x & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ldots & 0 \\ 0 & 0 & 0 & 0 & \mu_p\Sigma_x \end{bmatrix} \in \mathbb{R}^{p^2\times p^2}, \tag{2.2.5}$$

and analogously,

$$
\begin{aligned}
&\Omega = \Sigma^{-1} = \mathrm{diag}(\mu^{-1}) \otimes \Omega_x; \\
&\widehat{\Sigma} = \mathrm{diag}(\widehat{\mu}) \otimes \widehat{\Sigma}_x, \quad \widehat{\Omega} = \mathrm{diag}(\widehat{\mu}^{-1}) \otimes \widehat{\Omega}_x. \qquad (2.2.6)
\end{aligned}
$$

Following the one-step estimator in [13], we de-bias $\widehat{\beta}$ by adding an additional term involving the gradient of the loss function L:

$$
\check{\beta} = \widehat{\beta} + \widehat{\Omega}\,\nabla L_n(\widehat{\beta}). \qquad (2.2.7)
$$

To briefly explain the presence of $\widehat{\Omega}$, consider Taylor expansion of $\nabla L_n(\widehat{\beta})$ around $\nabla L_n(\beta^*)$. Write

$$
\begin{aligned}
\sqrt{n}(\check{\beta} - \beta^*) &= \sqrt{n}(\widehat{\beta} - \beta^*) + \sqrt{n}\,\widehat{\Omega}\,\nabla L_n(\beta^*) - \sqrt{n}\,\widehat{\Omega}(\nabla L_n(\widehat{\beta}) - \nabla L_n(\beta^*)) \\
&= \sqrt{n}\,\widehat{\Omega}\,\nabla L_n(\beta^*) + \sqrt{n}\left[(\widehat{\beta} - \beta^*) - \widehat{\Omega}\,\nabla^2 L_n(\beta^*)(\widehat{\beta} - \beta^*) + R\right] \\
&= \underbrace{\sqrt{n}\,\widehat{\Omega}\,\nabla L_n(\beta^*)}_{A} + \underbrace{\sqrt{n}\left[\Big(I_{p^2} - \widehat{\Omega}\,\nabla^2 L_n(\beta^*)(\widehat{\beta} - \beta^*)\right]}_{\Delta} + \sqrt{n}R, \qquad (2.2.8)
\end{aligned}
$$

where the remainder term $\sqrt{n}R = o(1)$ under certain conditions. Moreover, we also hope Δ to be negligible. As will be shown in the following sections,

$$
\Delta \le \sqrt{n}\Big(\|\Omega - \widehat{\Omega}\|_1 \|\Sigma\|_{\max} + \|\nabla^2 L_n(\beta^*) - \Sigma\|_{\max} \|\widehat{\Omega}\|_1\Big)|\widehat{\beta} - \beta^*|_1, \qquad (2.2.9)
$$

To this end, $\widehat{\Omega}$ needs to be a good approximation of the precision matrix Ω, which inspires the construction of such $\widehat{\Omega}$. More rigorous arguments will be presented in the subsequent sections.

Note that the estimator $\check{\beta}$ is closely related to the de-sparsifying Lasso estimator ([133] and [154]), which is employed to conduct simultaneous inference for linear regression models in [157]. $\check{\beta}$ will reduce to de-sparsifying Lasso estimator if the loss $\ell(x)$ in (2.2.4) is squared error loss and the weight $w(x) \equiv 1$. Moreover, [87] uses this one-step estimator to build the limiting distribution of high-dimensional vector restricted to a fixed number of coordinates, and delivers a result that agrees with [13] for low-dimensional robust M-estimators. Different from that, we will derive such conclusions simultaneously for all p^2 coordinates of β^*. In the subsequent sections, we aim at obtaining a limiting distribution for $\check{\beta}$.

2.2.3 Estimation of the precision matrix

In this section, we mainly discuss the validity of having $\widehat{\Omega}$ as an approximation of Ω. By the structure of Ω, we need to first find a suitable estimator of the weighted precision Ω_x.

The estimation of the sparse inverse covariance matrix based on a collection of observations $\{X_i\}$ plays a crucial role in establishing the asymptotic distribution. In high-dimensional regime, one cannot obtain a suitable estimator for the precision matrix by simply inverting the sample covariance, as the sample covariance is not invertible when the number of features exceeds the number of observations. Depending on the purposes, various methodology have been proposed to solve problem of estimating the precision. See for example, graphical Lasso ([151] and [40]) and nodewise regression ([95]). From a different perspective, [22] proposed a CLIME approach to sparse precision estimation, which shall be applied in this chapter. For completeness, we reproduce the CLIME estimator in the following.

Suppose that the sparsity of each row of Ω_x is at most s, i.e., $s = \max_{1\leq i\leq p} |\{j : \Omega_{x,ij} \neq 0\}|$. We first obtain $\widehat{\Theta}$ by solving

$$\widehat{\Theta} = \mathrm{argmin}_{\Theta} \sum_{i,j} \left|\Theta_{ij}\right| \quad \text{subject to:} \qquad \|\widehat{\Sigma}_x \Theta - I_p\|_{\max} \leq \lambda_n,$$

for some regularization parameter $\lambda_n > 0$. Note that the solution Θ may not symmetric. To account for symmetry, the CLIME estimator $\widehat{\Omega}_x$ is defined as

$$\widehat{\Omega}_x = (\widehat{\omega}_{ij}), \quad \text{where } \widehat{\omega}_{ij} = \widehat{\omega}_{ji} = \widehat{\Theta}_{ij}\mathbb{I}\{|\widehat{\Theta}_{ij}| \leq |\widehat{\Theta}_{ji}|\} + \widehat{\Theta}_{ji}\mathbb{I}\{|\widehat{\Theta}_{ij}| > |\widehat{\Theta}_{ji}|\}. \tag{2.2.10}$$

For more analysis of CLIME estimator, see [22]. Next, we present the convergence theorem for CLIME estimator.

Theorem 2.2.1. *Let τ be defined in definition 2.2.1 and $\gamma = \max_{t=0,1\ldots,\tau-1} \|A^t\|$. Choose $\lambda_n \asymp \|\Omega_x\|_1 \gamma \tau^2 T^2 (\log p)^{3/2} n^{-1/2}$, then with probability at least $1 - 4p^{-c_0}$ for some constant $c_0 > 0$,*

$$\|\widehat{\Omega}_x - \Omega_x\|_{\max} \lesssim \|\Omega_x\|_1 \lambda_n \quad \textit{and} \quad \|\widehat{\Omega}_x - \Omega_x\|_1 \lesssim \|\Omega_x\|_1 s \lambda_n.$$

Remark 2.2.5. Theorem 2.2.1 is a direct application of Theorem 6 of [22]. Note that if we assume the eigenvalue condition on Σ_x that $0 \leq c \leq \lambda_{\min}(\Sigma_x) \leq \lambda_{\max}(\Sigma_x) \leq C$, then $\|\Omega_x\|_2 \leq 1/\lambda_{\min}(\Sigma_x) = O(1)$. Therefore, by the sparsity condition on Ω_x, we immediately have that $\|\Omega_x\|_1 = O(\sqrt{s})$. Suppose the scaling condition holds that $s\gamma\tau^2 T^2 (\log p)^{3/2} n^{-1/2} = o(1)$, then the CLIME estimator $\widehat{\Omega}_x$ defined in (2.2.10) is consistent in estimating the weighted precision matrix of the VAR(1) model (2.2.2).

The following theorem shows that $\|\Omega - \widehat{\Omega}\|$ enjoys the same convergence rate as in the previous theorem.

Theorem 2.2.2. *Let $\widehat{\Omega}_x$ be the CLIME estimator defined above. Assume that $\mu_k > c_1 > 0$ for all $1 \leq k \leq p$, then with probability at least $1 - 6p^{-c}$,*

$$\|\Omega - \widehat{\Omega}\|_{\max} \lesssim \|\Omega_x\|_1 \lambda_n \quad \textit{and} \quad \|\Omega - \widehat{\Omega}\|_1 \lesssim \|\Omega_x\|_1 s \lambda_n.$$

The above theorem is built upon two facts: $\widehat{\Omega}_x$ approximates Ω_x and $\widehat{\mu}$ approximates μ. The result regarding the latter approximation will be given in Lemma 2.3.5.

2.2.4 Simultaneous inference

In this section, consider the following hypothesis testing problem:

$$H_0 : A_{ij} = A^0_{ij}, \quad \text{for all } i, j = 1, \ldots, p$$

versus the alternative hypothesis $H_1 : A_{ij} \neq A^0_{ij}$ for some $i, j = 1, \ldots, p$. Equivalently, we can also test for $\beta^*_j = \beta^0_j$, for all $j = 1, \ldots, p$. Instead of projecting the explanatory variables onto a subspace of fixed dimension ([61], [154], [133] and [87]), we allow the number of testings to grow as fast as an exponential order of the sample size n. [157] presented a more related work, where it's also allowed that the testing size to grow as a function of p. However, they conducted such simultaneous inference procedure under linear regression setting with independent random variables.

Employing the de-biased estimator $\check{\beta}$ defined in (2.2.7), we propose to use the test statistics

$$\sqrt{n}|\check{\beta} - \beta^0|_\infty, \tag{2.2.11}$$

where $\check{\beta}$ is defined in (2.2.7). In the next several theorems, we elaborate a multiplier bootstrap method to obtain the critical value of the test statistics, which requires a few scaling and moment assumptions. Recall definition 2.2.1 for τ and theorem 2.2.1 for the definition of γ. Also recall that $s = \max_{1 \le i \le p} |\{j : \Omega_{x,ij} \neq 0\}|$.

Assumptions

(A1) $\sqrt{n}T^3|\widehat{\beta} - \beta^*|_1^2 = o(1)$.

(A2) $\|\Omega_x\|_1^2 s\gamma^2\tau^4 T^4(\log p)^3/\sqrt{n} = o(1)$.

(A3) $s\gamma\tau^2 T^2(\log p)^{3/2}|\widehat{\beta} - \beta^*|_1 = o(1)$.

(A4) $sT^2(\log(pn))^7/n \lesssim n^{-c}$.

(A5) $(\log p)^{3/2}(\log n)^{1/2}T\sqrt{s\tau}\gamma/n^{1/4} = o(1)$.

Additionally, throughout the chapter we assume that for some constant $C > 0$, $\mathbb{E}[X_{ik}^2] \leq C$ and $\mathbb{E}[\varepsilon_{ik}^2] \leq C$ for all $1 \leq k \leq p$. We also suppose that $\|\Sigma_x\|_{\max} = O(1)$ and $0 < c \leq \lambda_{\min}(\Sigma_x) \leq \lambda_{\max}(\Sigma_x) \leq C$. Thus, $\|\Omega_x\|_2 \leq 1/\lambda_{\min}(\Sigma_x) = O(1)$ and $\|\Omega_x\|_1 = O(\sqrt{s})$, where the row sparsity $s = \max_{1\leq i\leq p} |\{j : \Omega_{x,ij} \neq 0\}|$.

Theorem 2.2.3. *Suppose assumptions (A1) – (A3) hold. Define*

$$\zeta_1 = \gamma\tau^2T^2(\log p)^{3/2}|\widehat{\beta} - \beta^*|_1 + \sqrt{n}T^3|\widehat{\beta} - \beta^*|_1^2 + s\gamma^2\tau^4T^4(\log p)^3/\sqrt{n}.$$

Further assume that $\zeta_1\sqrt{1 \vee \log(p/\zeta_1)} = o(1)$. *Then we have*

$$\mathbb{P}\bigg(|\sqrt{n}(\check{\beta} - \beta^*) - \sqrt{n}\Omega\nabla L_n(\beta^*)|_\infty > \zeta_1\bigg) < \zeta_2,$$

where $\zeta_1\sqrt{1 \vee \log(p/\zeta_1)} = o(1)$ *and* $\zeta_2 = o(1)$.

Theorem 2.2.3 rigorously verifies that $\sqrt{n}R = o(1)$ and $\Delta = o(1)$ in (2.2.8) by the proposed construction of $\widehat{\Omega}$ and suggests us to perform further analysis on $\sqrt{n}\Omega\nabla L_n(\beta^*)$. To derive the limiting distribution, we shall use Gaussian approximation technique, since the classic central limit theorem fails in high-dimensional setting.

Gaussian approximation was initially invented for high-dimensional independent random variables in [27] and further generalized to high-dimensional time series in [156]. [157] and [87] applied the GA technique in [27] to the derivation of asymptotic distribution in linear regression setting. However, data generated from VAR model suffers temporal dependence, which makes the aforementioned techniques unavailable. Although [156] established such GA results for general time series using dependence adjusted norm, direct application of their theorems does not yield desirable conclusion in ultra-high dimensional setting. This leads us to derive a new GA theorem with better convergence rate, which is achievable thanks to the structure of VAR model.

The next theorem establishes a Gaussian approximation(GA) result for the term

$$\sqrt{n}\Omega\nabla L_n(\beta^*).$$

For a more detailed description of Gaussian approximation procedure, see Section 2.4.

Theorem 2.2.4. *Denote $D = (D_{jk})_{1\le j,k\le p} \in \mathbb{R}^{p^2\times p^2}$ with*

$$D_{jk} = \frac{\Omega_x \mathbb{E}[\psi(\varepsilon_{ij})\psi(\varepsilon_{ik})]\mathbb{E}[X_i X_i^\top w^2(X_i)]\Omega_x^\top}{\mu_j\mu_k} \in \mathbb{R}^{p\times p}.$$

Under Assumption (A4) and (A5), we have the following Gaussian Approximation result that

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\Big(|\sqrt{n}\Omega\nabla L_n(\beta^*)|_\infty \le t\Big) - \mathbb{P}\Big(\big|\sum_{i=1}^n z_i/\sqrt{n}\big|_\infty \le t\Big)\right| = o(1),$$

where $z_i = (z_{i1}, \dots, z_{ip^2})^\top$ is a sequence of mean zero independent Gaussian vectors with each $\mathbb{E}z_i z_i^\top = D$.

Remark 2.2.6. The above GA results allows the ultra-high dimensional regime, wehere p grows as fast as $O(\mathrm{e}^{n^b})$ for some $0 < b < 1$.

Since the covariance matrix D of the Gaussian analogue z_i is not accessible from the observation $\{X_i\}$, we need to give a suitable estimation of D before further performing multiplier bootstrap. The next theorem delivers a consistent estimator for our purpose.

Theorem 2.2.5.

$$\widehat{D}_{jk} = \frac{\widehat{\Omega}_x\Big(\frac{1}{n}\sum_{i=1}^n \psi(\widehat{\varepsilon}_{ij})\psi(\widehat{\varepsilon}_{ik})\Big)\Big(\frac{1}{n}\sum_{i=1}^n X_i X_i^\top w^2(X_i)\Big)\widehat{\Omega}_x^\top}{\widehat{\mu}_j\widehat{\mu}_k} \in \mathbb{R}^{p\times p}, \tag{2.2.12}$$

where $\widehat{\Omega}_x$ is the CLIME estimator of Ω_x. Under assumptions (A1)–(A5) and additionally assume that $\|\Omega_x\|_1 = O(\sqrt{s})$ and that for all $1 \le k \le p$, $\mu_k > C > 0$ for some constant C, *we have with probability at least $1 - 12p^{-c}$, we have*

$$\|\widehat{D} - D\|_{\max} \lesssim s\gamma\tau^2 T^2 (\log p)^{3/2} n^{-1/2} + |\widehat{\beta} - \beta^*|_1.$$

Indeed, under the scaling assumptions, $\|\widehat{D} - D\|_{\max} = o(1)$. With these preparatory results, we are ready to present the main theorem of this chapter, which describes a procedure to find the critical value of $\sqrt{n}|\check{\beta} - \beta^*|_\infty$ using bootstrap.

Theorem 2.2.6. *Denote*

$$W = |\widehat{D}^{1/2}\eta|_\infty,$$

where $\eta \sim N(0, I_{p^2})$ is independent of $(X_i)_{i=1}^n$ and $\widehat{D}$ is defined in (2.2.12). Let the bootstrap critical value be given by $c(\alpha) = \inf\{t \in \mathbb{R} : \mathbb{P}(W \leq t|\boldsymbol{X}) \geq 1 - \alpha\}$. Let assumptions (A1) — (A5) and the assumptions in theorem 2.2.3 hold. Denote $v = c(s\gamma\tau^2 T^2 (\log p)^{3/2}/\sqrt{n} + |\widehat{\beta} - \beta^*|_1)$ for some constant c. Assume that $\pi(v) = Cv^{1/3}(1 \vee \log(p/v))^{2/3} = o(1)$, then we have

$$\sup_{\alpha \in (0,1)} \left| \mathbb{P}\Big(\sqrt{n}|\check{\beta} - \beta^*|_\infty > c(\alpha)\Big) - \alpha \right| = o(1).$$

This result suggests a way to not only find the asymptotic distribution, but also to provide an accurate critical value $c(\alpha)$ using multiplier bootstrap. Under the null hypothesis H_0, we have $\sqrt{n}|\check{\beta} - \beta^0|_\infty = \sqrt{n}|\check{\beta} - \beta^*|$. This verifies the validity of having (2.2.11) as a test statistics for simultaneous inference.

2.3 Estimation Consistency

Many estimation tasks are needed as preparatory results for proving Theorem 2.2.6. For instance, Theorem 2.2.6 requires an estimation of the theoretical covariance matrix D of the Gaussian analogue Z, as stated in Theorem 2.2.5. Besides, the convergence of CLIME estimator (section 2.3) depends on the convergence of corresponding covariance matrix. Therefore, these problems requires us to develop a new estimation theory that delivers the convergence even in ultra-high dimensional regime.

The success of high-dimensional estimation relies heavily on the application of probability concentration inequality, among which Bernstein-type inequality is especially important. The celebrated Bernstein's inequality ([10]) provides an exponential concentration inequality for sums of independent random variables which are uniformly bounded. Later works relaxed the uniform boundedness condition and extended the validity of Bernstein inequality to independent random variables that have finite exponential moment; see for example, [94] and [134].

Despite the extensive body of work on concentration inequalities for independent random variables, literature remains quiet when it comes to establishing exponential-type tail concentration results for random process. Some related existing work includes Bernstein inequality for sums of strong mixing processes ([97]), Bernstein inequality under functional dependence measures ([155]), etc. In a more recent work, [85] established a sharp Bernstein inequality for VAR model using the definition of spectral decay index, which improved the current rate by a factor of $(\log n)^2$. In this chapter, we will derive another Bernstein inequality for VAR model under slightly different condition from [85]. Before presenting the main results, recall the definition of τ in definition 2.2.1.

Lemma 2.3.1. *Let $\{X_i\}_{i=0}^n$ be generated by a VAR(1) model. Suppose $G : \mathbb{R}^p \to \mathbb{R}$ satisfies that*

$$|G(X) - G(Y)| \leq |X - Y|_\infty, \tag{2.3.1}$$

and that $|G(x)| \leq B$ for all $x \in \mathbb{R}$. Assume that $\mathbb{E}[|\varepsilon_{ij}|^2] \leq \sigma^2$ for all $j = 1, \dots, p$. Then there exists some constants $C_1, C_2, C_3, C_4 > 0$ only depending on ρ and σ, such that

$$\mathbb{P}\left(\Big|\frac{1}{n}\sum_{i=1}^n G(X_{i-1}) - \mathbb{E}[G(X_{i-1})]\Big| \geq x\right) \leq 2\exp\left\{-\frac{nx^2}{C_3 n^{-1}\gamma^2\tau^3 + C_4\tau B x}\right\}$$
$$+2\exp\left\{-\frac{nx^2}{(1+C_1B^{-2})\gamma^2\tau^4B^2(\log p)^2(n^{-1}\tau\log p + 1) + C_2\tau^2B(\log p)x}\right\}.$$

Specifically, under assumption (A2), we see that $\tau(\log p)/n \to 0$*. So for sufficiently large* $B > 0$*, we have*

$$\begin{aligned} &\mathbb{P}\bigg(\Big|\frac{1}{n}\sum_{i=1}^{n} G(X_{i-1}) - \mathbb{E}[G(X_{i-1})]\Big| \geq x\bigg) \\ \leq\ & 4\exp\bigg\{-\frac{nx^2}{C_1'\gamma^2\tau^4B^2(\log p)^2 + C_2'\tau^2B(\log p)x}\bigg\}, \end{aligned} \tag{2.3.2}$$

for some positive constants C_1', C_2' *depending only on* ρ *and* σ*.*

Remark 2.3.1. Note that the Lipschitz condition (2.3.1) is slightly different from that in [85], where instead, they assumed that

$$|G(x) - G(y)| \leq g^\top |x - y|, \tag{2.3.3}$$

for some vector $g \in \mathbb{R}^p$. Since condition (2.3.1) is weaker than (2.3.3), the additional $(\log p)$ appears in the denominator of right-hand side in (2.3.2). For more detailed comparison of different versions of Bernstein inequalities, we refer readers to [85] and the references therein.

With a minor modification of the proof of Lemma 2.3.1, we have the following version of Bernstein inequality which includes a bounded function of the latest innovation ε_i as a multiple.

Corollary 2.3.2. *Let* $\{X_i\}_{i=0}^n$ *be generated by a VAR(1) model. Suppose* $|h(x)| \leq 1$ *and* $G : \mathbb{R}^p \to \mathbb{R}$ *satisfies that*

$$|G(X) - G(Y)| \leq |X - Y|_\infty,$$

and that $|G(x)| \leq B$ *for all* $x \in \mathbb{R}$*. Assume that* $\mathbb{E}[|\varepsilon_{ij}|^2] \leq \sigma^2$ *for all* $j = 1, \ldots, p$*. Then* *there exists some constants $C_1, C_2, C_3, C_4 > 0$ only depending on ρ and σ, such that*

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} h(\varepsilon_i)G(X_{i-1}) - \mathbb{E}[h(\varepsilon_i)G(X_{i-1})]\Big| \geq x\Big) \leq 2\exp\Big\{-\frac{nx^2}{C_3 n^{-1}\gamma^2\tau^3 + C_4\tau Bx}\Big\}$$
$$+ 2\exp\Big\{-\frac{nx^2}{(1+C_1B^{-2})\gamma^2\tau^4B^2(\log p)^2(n^{-1}\tau\log p + 1) + C_2\tau^2B(\log p)x}\Big\}.$$

Specifically, under assumption (A2), we see that $\tau(\log p)/n \to 0$. So for sufficiently large $B > 0$, we have

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} h(\varepsilon_i)G(X_{i-1}) - \mathbb{E}[h(\varepsilon_i)G(X_{i-1})]\Big| \geq x\Big)$$
$$\leq 4\exp\Big\{-\frac{nx^2}{C_1'\gamma^2\tau^4B^2(\log p)^2 + C_2'\tau^2B(\log p)x}\Big\},$$

for some positive constants C_1', C_2' depending only on ρ and σ.

Remark 2.3.2. Since the additional term $h(\varepsilon_i)$ is independent of $G(X_{i-1})$, the proof of Lemma 2.3.1 directly applies without any extra technical difficulty.

Equipped with our new Bernstein inequalities, several estimation results follow immediately. The next theorem regarding the estimation of Σ_x is essential when we prove the convergence rate of CLIME estimator in section 2.3.

Theorem 2.3.3 (Estimation of Σ_x). *Let $\widehat{\Sigma}_x = n^{-1}\sum_{i=1}^{n} X_{i-1}X_{i-1}^\top w(X_{i-1})$ and $\Sigma_x = \mathbb{E}[X_iX_i^\top w(X_i)]$. Then with probability at least $1 - 4p^{-c_0}$ for some constant $c_0 > 0$, it holds that*

$$\|\widehat{\Sigma}_x - \Sigma_x\|_{\max} \lesssim \gamma\tau^2T^2n^{-1/2}(\log p)^{3/2}.$$

We see that the convergence rate of CLIME estimator in Theorem 2.2.1 essentially inherits from the convergence rate in Theorem 2.3.3, with an additional term $\|\Omega_x\|_1$. The following theorem plays an important role in verifying that the Δ defined in (2.2.9) is indeed negligible.

Theorem 2.3.4 (Estimation of Σ by $\nabla^2 L_n(\beta^*)$)**.** *Assume that* $\mathbb{E}[\varepsilon_{ik}^2] \leq \sigma^2$ *for all* $1 \leq k \leq p$. *Then for some constant* $c_1 > 0$*, with probability at least* $1 - 4p^{-c_1}$*, it holds that*

$$\|\nabla^2 L_n(\beta^*) - \Sigma\|_{\max} \lesssim \gamma\tau^2 T^2 n^{-1/2} (\log p)^{3/2}.$$

While the last two theorems make use of Lemma 2.3.1 in this chapter, the next estimation for μ directly applies the concentration inequality in [85] thanks to the stronger assumption that $\widehat{\mu}$ satisfies.

Lemma 2.3.5. *Suppose that* β_k^* *lies in a bounded* ℓ_1 *normed ball for all* $1 \leq k \leq p$ *and that* $\mathbb{E}[X_{ij}^2] \leq C$ *for some constant* $C > 0$ *and for all* $1 \leq j \leq p$*. Then we have*

$$\mathbb{P}\biggl(|\widehat{\mu} - \mu|_\infty \geq \gamma\tau^2 \sqrt{\frac{\log p}{n}} + |\widehat{\beta} - \beta^*|_1\biggr) \leq 2p^{-c},$$

for some positive constant c*.*

2.4 Gaussian Approximation

Conducting simultaneous inference for high-dimensional data is always considered to be a hard task, since central limit theorem fails when the dimension of random vectors can grow as a function of the number of observation n, or even exceeds n. As an alternative to central limit theorem, [27] proposed Gaussian approximation theorem, which states that under certain conditions, the distribution of the maximum of a sum of independent high-dimensional random vectors can be approximated by that of the maximum of a sum of the Gaussian random vectors with the same covariance matrices as the original vectors. Their Gaussian approximation results allow the ultra-high dimensional cases, where the dimension p grows exponentially in n. In the meantime, they also proved that Gaussian multiplier bootstrap method yields a high quality approximation of the distribution of the original maximum and showcased a wide range of application, such as high-dimensional estimation, multiple hypothesis testing, and adaptive specification testing. It is worth noticing that the results from [27] are only applicable when the sequence of random vectors is independent.

[156] generalized Gaussian approximation results to general high-dimensional stationary time series, using the framework of functional dependence measure ([142]). We specifically mention that a direct application of Gaussian approximation from [156] cannot deliver a desired conclusion in ultra-high dimensional regime, due to coarser capture of dependence measure for VAR model. In what follows, we will use refined argument to establish a new Gaussian approximation result for VAR model.

By Theorem 2.2.3, $\sqrt{n}|\check{\beta} - \beta^*|_\infty$ can be approximated by $\sqrt{n}|\Omega\nabla L_n(\beta^*)|_\infty$. Hence, we shall build a GA result for $\sqrt{n}\Omega\nabla L_n(\beta^*)$. Observe that $\sqrt{n}\Omega\nabla L_n(\beta^*) \in \mathbb{R}^{p^2}$ can be written as

$$\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\Omega_x}{\mu_1}\psi_\alpha(\varepsilon_{i1})X_{i-1}^\top w(X_{i-1}), \dots, \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\Omega_x}{\mu_p}\psi_\alpha(\varepsilon_{ip})X_{i-1}^\top w(X_{i-1}),\right)^\top,$$

so it's sufficient to establish GA result for one sub-vector

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\Omega_x}{\mu_k}\psi_\alpha(\varepsilon_{ik})X_{i-1}^\top w(X_{i-1}), \quad k = 1, \dots, p.$$

Fix $1 \leq k \leq p$ and denote $\Theta_k = \Omega_x\mu_k^{-1}$. Let $X_{i,m} = \sum_{l=0}^m A^l\varepsilon_{i-l}$ be the m-approximation of X_i with m to be determined later. Let $Y_i = \psi_\alpha(\varepsilon_{ik})\Theta_k X_{i-1}w(X_{i-1})$ be the quantity that we will establish Gaussian approximation for and denote $T_Y = \sum_{i=1}^n Y_i$. Analogously, let $Y_{i,m} = \psi_\alpha(\varepsilon_{ik})\Theta_k X_{i-1,m}w(X_{i-1,m})$ be the m-approximation of Y_i and write $T_{Y,m} = \sum_{i=1}^n Y_{i,m}$. For simplicity, assume $n = (m + M)w$, where $M \to \infty$, $m \to \infty$, $w \to \infty$ and $m/M \to 0$. Divide the interval $[1, n]$ into alternating large blocks $L_b = [(b-1)(M+m)+1, bM+(b-1)m]$ with M points and small blocks

$S_b = [bM + (b-1)m + 1, b(M+m)]$ with m points, for $1 \le b \le w$. Denote

$$\xi_b = \sum_{i \in L_b} Y_{i,m}/\sqrt{M}, \quad T_{Y,S} = \sum_{b=1}^{w} \sum_{i \in S_b} Y_{i,m}, \quad T_{Y,L} = \sum_{b=1}^{w} \sum_{i \in L_b} Y_{i,m},$$
$$Z \sim N(0, \mu_k^{-2} \mathbb{E}[\psi^2(\varepsilon_{ik})] \Omega_x \mathbb{E}[X_i X_i^\top w^2(X_i)] \Omega_x^\top)$$

Note that the $Y_{i,m}$ from different large blocks L_b are independent, i.e. $\sum_{i \in L_b} Y_{i,m}$ is independent in $b = 1 \ldots, w$. The main result of this section is presented as follow.

Theorem 2.4.1. *Suppose $\mathbb{E}[\varepsilon_{ik}^2] \le \sigma^2$ for all $1 \le k \le p$ and the odd function $\psi(\cdot)$ satisfies that $|\psi(\cdot)| \le C$ and $|\psi'(\cdot)| \le C$. Suppose the scaling condition holds that $sT^2(\log(pn))^7/n \le c_1 n^{-c_2}$. Then for any $\eta > 0$, the Gaussian Approximation holds that*

$$\mathcal{H} := \sup_{t \in \mathbb{R}} \left| \mathbb{P}\big(|T_Y/\sqrt{n}|_\infty \le t\big) - \mathbb{P}\big(|Z|_\infty \le t\big) \right|$$
$$\lesssim f_1(\eta/2, m) + f_2(\eta/2, m) + \eta\sqrt{\log p} + \eta\sqrt{\log(1/\eta)} + cn^{-c'}, \qquad (2.4.1)$$

for some $c, c' > 0$.

This theorem gives an upper bound on the supremum of the difference between the distribution of the maximum of sum of Y_i and that of the maximum of a Gaussian vector Z with the same covariance. Now, we present the outline of the proof of the previous theorem, while we leave the complete proof in the appendix.

First, we show that the sum of $Y_{i,m}$ in the small blocks are negligible, so $T_{Y,m} \approx T_{Y,L}$. Next, we prove that the sum of Y_i can be approximated by its m-approximation, that is, $T_Y \approx T_{Y,m} \approx T_{Y,L}$. Since $T_{Y,L}$ is a sum of independent random vector $\{\sum_{i \in L_b} Y_{i,m}\}_{b=1}^{w}$, the GA theorem from [27] can be applied.

2.5 Numerical Experiments

In this section, we evaluate the performance of the proposed bootstrap-assist procedure in simultaneous inference. We consider the model (2.2.2), where ε_{ij}'s are i.i.d. Student's t-distributions with $df = 5$ or 10. Let $s = \lfloor \log p \rfloor$. We pick $n = 30$ and $p = 10$ in the numerical setup. For the true transition matrix $A = (a_{ij})$, we consider the following designs.

(1) Banded: $A = (\lambda^{|i-j|}\mathbf{1}\{|i-j| \leq s\})$ and $\lambda = 0.5$.

(2) Block diagonal: $A = \text{diag}\{A_i\}$, where each $A_i \in \mathbb{R}^{s\times s}$ has λ_i on the diagonal and λ_i^2 on the superdiagonal with $\lambda_i \sim Unif(-0.8, 0.8)$.

The design in (1) is further scaled by $2\rho(A)$ to ensure that $\rho(A) < 1$. Hence sparse symmetric matrices are generated in (1) and sparse asymmetric matrices are constructed in (2). We draw the qq-plots of the data quantile of $\sqrt{n}|\check{\beta} - \beta^*|_\infty$ versus the data quantile of W defined in Theorem 2.2.6 from $m = 100$ duplicates. The qq-plots are shown in figure 2.1 and figure 2.2 for banded and block diagonal designs respectively.

2.6 Proofs of Results in Section 2.2

Before proceeding with the proofs, we state a helpful lemma that is repeatedly used throughout the chapter and present its proof. This simple lemma is an application of the triangle inequality to the product of two matrices.

Lemma 2.6.1. *Let A, B and $\widehat{A}, \widehat{B}$ be $p \times p$ symmetric matrices and $\|A - \widehat{A}\|_1 = o(1)$. Suppose $\|A\|_1 = O(1)$ and $\|B\|_1 = O(1)$. Then $\|AB - \widehat{A}\widehat{B}\|_{\max} \lesssim \|A - \widehat{A}\|_{\max} + \|B - \widehat{B}\|_{\max}$.*

Proof of Lemma 2.6.1. Since $\|A\|_1 = O(1)$ and $\|A - \widehat{A}\|_1 = o(1)$, $\|\widehat{A}\|_1 \leq \|A - \widehat{A}\|_1 +$

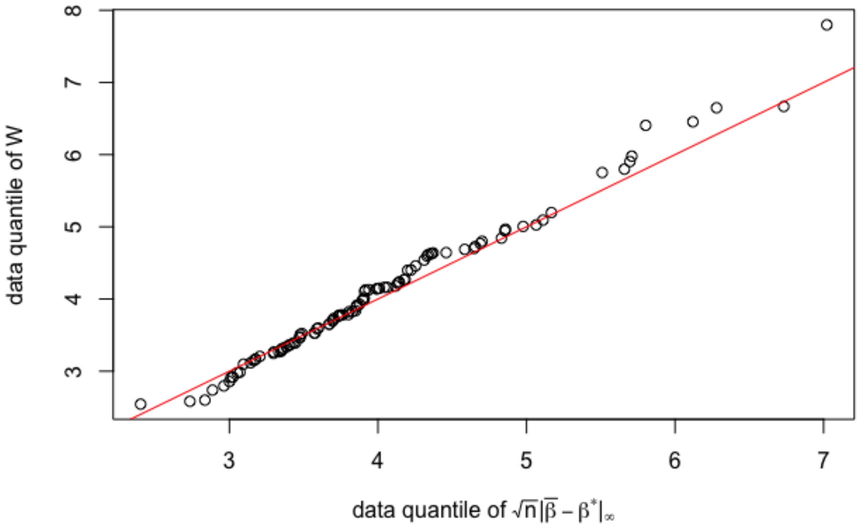


Figure 2.1. The qq-plot of banded design.

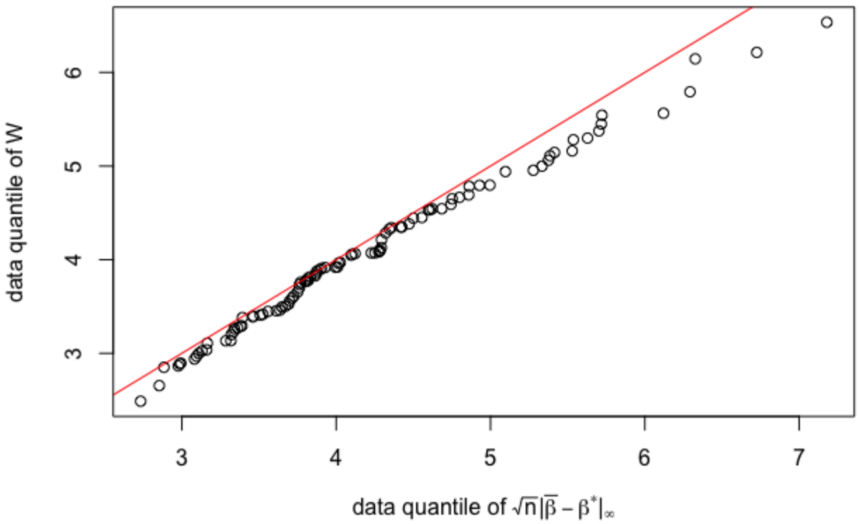


Figure 2.2. The qq-plot of block diagonal design.

$\|A\|_1 = O(1)$. Hence, by triangular inequality,

$$\begin{aligned}\|AB - \widehat{A}\widehat{B}\|_{\max} &\le \|(A-\widehat{A})B\|_{\max} + \|\widehat{A}(B-\widehat{B})\|_{\max}\\ &\le \|B\|_1\|A-\widehat{A}\|_{\max} + \|\widehat{A}\|_1\|B-\widehat{B}\|_{\max}\\ &\lesssim \|A-\widehat{A}\|_{\max} + \|B-\widehat{B}\|_{\max}\end{aligned}$$

□

Proof of Theorem 2.2.1. By Theorem 2.3.3, with probability at least $1-4p^{-c_0}$,

$$\|\widehat{\Sigma}_x - \Sigma_x\|_{\max} \le \lambda_n.$$

By Theorem 6 of [22], we have the desired result. □

Proof of Theorem 2.2.2. Recall that

$$\Omega = \Omega_x \otimes \operatorname{diag}(\mu^{-1}) = \begin{bmatrix} \mu_1^{-1}\Omega_x & 0 & 0 & \dots & 0\\ 0 & \mu_2^{-1}\Omega_x & 0 & \dots & 0\\ \vdots & \vdots & \ddots & \dots & 0\\ 0 & 0 & 0 & 0 & \mu_p^{-1}\Omega_x \end{bmatrix} \tag{2.6.1}$$

and

$$\widehat{\Omega} = \widehat{\Omega}_x \otimes \operatorname{diag}(\widehat{\mu}^{-1}) = \begin{bmatrix} \widehat{\mu}_1^{-1}\widehat{\Omega}_x & 0 & 0 & \dots & 0\\ 0 & \widehat{\mu}_2^{-1}\widehat{\Omega}_x & 0 & \dots & 0\\ \vdots & \vdots & \ddots & \dots & 0\\ 0 & 0 & 0 & 0 & \widehat{\mu}_p^{-1}\widehat{\Omega}_x \end{bmatrix}. \tag{2.6.2}$$

For $1 \leq k \leq p$, consider

$$\begin{aligned}\|\widehat{\Omega}_x \widehat{\mu}_k^{-1} - \Omega_x \mu_k^{-1}\|_{\max} &\leq \|\widehat{\Omega}_x - \Omega_x\|_{\max} |\widehat{\mu}_k^{-1}| + \|\Omega_x\|_{\max} |\widehat{\mu}_k^{-1} - \mu_k^{-1}| \\ &\lesssim \|\Omega_x\|_1 \lambda_n + \|\Omega_x\|_{\max} \frac{|\mu_k - \widehat{\mu}_k|}{\mu_k \widehat{\mu}_k} \\ &\lesssim \|\Omega_x\|_1 \lambda_n,\end{aligned}$$

with probability no less than $1 - 6p^{-c'}$ by theorem 2.2.1 and lemma 2.3.4. Taking a union bound for all k yields

$$\|\Omega - \widehat{\Omega}\|_{\max} = \max_{1 \leq k \leq p} \|\mu_k^{-1} \Omega_x - \widehat{\mu}_k^{-1} \widehat{\Omega}_x\|_{\max} \lesssim \|\Omega_x\|_1 \lambda_n,$$

with probability at least $1 - 6p^{-(c'-1)}$. Replacing max-norm by L_1-norm delivers

$$\|\Omega - \widehat{\Omega}\|_1 \lesssim \|\Omega_x\|_1 s \lambda_n.$$

□

The next lemma provides a high probability bound on $|\nabla L_n(\beta^*)|_\infty$, which will be used in the proof of Theorem 2.2.3.

Lemma 2.6.2. *Suppose that $\mathbb{E}[\varepsilon_{ij}^2] \leq C$ for all $1 \leq j \leq p$. Then it holds that*

$$\mathbb{P}(|\nabla L_n(\beta^*)|_\infty \gtrsim \gamma \tau^2 T (\log p)^{3/2} / \sqrt{n}) \leq 4p^{-c},$$

for some constant $c > 0$.

Proof of Lemma 2.6.2. We shall apply Corollary 2.3.2. Consider the first coordinate $\nabla L_{n1}(\beta^*)$ of $\nabla L_n(\beta^*)$. In Corollary 2.3.2, let $h(\varepsilon_i) = \psi(\varepsilon_{i1})$ and $G(X_i) = X_{i1} w(X_i)$.

Observe $\mathbb{E}[\nabla L_n(\beta^*)] = 0$. By Corollary 2.3.2,

$$\begin{aligned}\mathbb{P}(|\nabla L_{n1}(\beta^*)| \geq x) &= \mathbb{P}(|\nabla L_{n1}(\beta^*) - \mathbb{E}[\nabla L_{n1}(\beta^*)]|_\infty \geq x)\\ &= \mathbb{P}\bigg(\Big|\frac{1}{n}\sum_{i=1}^{n} h(\varepsilon_i)G(X_{i-1}) - \mathbb{E}[h(\varepsilon_i)G(X_{i-1})]\Big| \geq x\bigg)\\ &\leq 4\exp\bigg\{-\frac{nx^2}{C_1\gamma^2\tau^4T^2(\log p)^2 + C_2\tau^2T(\log p)x}\bigg\}.\end{aligned}$$

Choose $x = c'\gamma\tau^2T(\log p)^{3/2}/\sqrt{n}$ and we get

$$\mathbb{P}(|\nabla L_{n1}(\beta^*)| \geq c'\gamma\tau^2T(\log p)^{3/2}/\sqrt{n}) \leq 4p^{-c},$$

for some constant $c > 0$. Take sufficiently large c' such that $c > 1$, so by a union bound we obtain

$$\mathbb{P}(|\nabla L_n(\beta^*)|_\infty \geq c'\gamma\tau^2T(\log p)^{3/2}/\sqrt{n}) \leq 4p^{-c''},$$

where $c'' = c - 1 > 0$. □

Proof of Theorem 2.2.3. By Taylor expansion, we write

$$\begin{aligned}\sqrt{n}(\check{\beta} - \beta^*) &= \sqrt{n}(\widehat{\beta} - \beta^*) + \sqrt{n}\,\widehat{\Omega}\,\nabla L_n(\beta^*) - \sqrt{n}\,\widehat{\Omega}(\nabla L_n(\widehat{\beta}) - \nabla L_n(\beta^*))\\ &= \sqrt{n}\,\widehat{\Omega}\,\nabla L_n(\beta^*) + \sqrt{n}\,\Big[(\widehat{\beta} - \beta^*) - \widehat{\Omega}\,\nabla^2 L_n(\beta^*)(\widehat{\beta} - \beta^*) + R\Big]\\ &= \underbrace{\sqrt{n}\,\widehat{\Omega}\,\nabla L_n(\beta^*)}_{A} + \underbrace{\sqrt{n}\,\Big[\Big(I_{p^2} - \widehat{\Omega}\,\nabla^2 L_n(\beta^*)(\widehat{\beta} - \beta^*)\Big]}_{\Delta} + \sqrt{n}R.\end{aligned}$$

where $z_{ik} = X_i - X_{i-1}^\top\widetilde{\beta}$ for some $\widetilde{\beta}$ lying between β^* and $\widehat{\beta}$. The remainder is denoted by

$$R = \frac{1}{2}\cdot$$
$$\sum_{i=1}^{n}\Big(\psi''(z_{i1})\big(X_{i-1}^\top(\widehat{\beta}_1 - \beta_1^*)\big)^2X_{i-1}^\top w(X_{i-1}), \ldots, \psi''(z_{ip})\big(X_{i-1}^\top(\widehat{\beta}_p - \beta_p^*)\big)^2X_{i-1}^\top w(X_{i-1})\Big)^\top.$$

Now we analyze the above terms A, Δ and R respectively. First we see that $\sqrt{n}|R|_\infty = O_{\mathbb{P}}(\sqrt{n}T^3|\widehat{\beta} - \beta^*|_1^2) = o(1)$ by assumption (A1). To analyze Δ, denote $H = \nabla^2 L_n(\beta^*)$. Then we write

$$\Delta = \sqrt{n}\Big(I_{p^2} - \widehat{\Omega}\, H\Big)(\widehat{\beta} - \beta^*) = \sqrt{n}\Big(\Omega\Sigma - \widehat{\Omega}H\Big)(\widehat{\beta} - \beta^*).$$

Thus, by theorem 2.3.4 and theorem 2.2.2, with probability tending to 1,

$$\begin{aligned}|\Delta|_\infty &\le \sqrt{n}\|\Omega\Sigma - \widehat{\Omega}H\|_{\max}|\widehat{\beta} - \beta^*|_1 \\ &\le \sqrt{n}\Big(\|\Omega - \widehat{\Omega}\|_1\|\Sigma\|_{\max} + \|H - \Sigma\|_{\max}\|\widehat{\Omega}\|_1\Big)|\widehat{\beta} - \beta^*|_1 \\ &\lesssim \sqrt{n}\|\Omega_x\|_1\lambda_n|\widehat{\beta} - \beta^*|_1 \asymp s\gamma\tau^2T^2(\log p)^{3/2}|\widehat{\beta} - \beta^*|_1 = o(1)\end{aligned}$$

by assumption (A3). Finally, by Lemma 2.6.2 and Theorem 2.2.2, with probability tending to 1, it holds that

$$\begin{aligned}|A - \sqrt{n}\Omega\nabla L_n(\beta^*)|_\infty \le \|\widehat{\Omega} - \Omega\|_1|\sqrt{n}\nabla L_n(\beta^*)|_\infty &\le \|\Omega_x\|_1^2 s\gamma^2\tau^4T^4(\log p)^3/\sqrt{n} \\ &\asymp s^2\gamma^2\tau^4T^4(\log p)^3/\sqrt{n}.\end{aligned}$$

Therefore,

$$|\sqrt{n}(\check{\beta} - \beta^*) - \sqrt{n}\Omega\nabla L_n(\beta^*)|_\infty \le |\sqrt{n}(\check{\beta} - \beta^*) - A|_\infty + |A - \sqrt{n}\Omega\nabla L_n(\beta^*)|_\infty \le \zeta_1,$$

where

$$\zeta_1 = s\gamma\tau^2T^2(\log p)^{3/2}|\widehat{\beta} - \beta^*|_1 + \sqrt{n}T^3|\widehat{\beta} - \beta^*|_1^2 + s^2\gamma^2\tau^4T^4(\log p)^3/\sqrt{n}.$$

□

Proof of Theorem 2.2.4. The proof of Theorem 2.4.1 can be easily generalized to p^2 dimensional space, thus it still holds for $|\sqrt{n}\Omega\nabla L_n(\beta^*)|_\infty$. By Theorem 2.4.1, we have for any $\eta > 0$,

$$\sup_{t\in\mathbb{R}}\Big|\mathbb{P}\big(|\sqrt{n}\Omega\nabla L_n(\beta^*)|_\infty \le t\big) - \mathbb{P}\big(|Z|_\infty \le t\big)\Big|$$
$$\lesssim f_1(\eta/2, m) + f_2(\eta/2, m) + \eta\sqrt{\log p} + \eta\sqrt{\log(1/\eta)} + cn^{-c'}, \tag{2.6.3}$$

where

$$f_1(x, m) = \frac{c_1 s p^3 \gamma^2 \rho^{m/\tau}}{x^2},\ f_2(x) = 2p\exp\Big\{ - \frac{nx^2}{2\sqrt{s}TM\sqrt{n}x + 4mwsT^2\sigma^2}\Big\}. \tag{2.6.4}$$

Now choose $\eta \asymp (\log p)T\sqrt{s\tau}\gamma/n^{1/4} = o(1), \omega \asymp n^{1/2}, M \asymp n^{1/2}, m = c\tau\log p$ in (2.6.3) for some constant $c > 0$. For sufficiently large c, basic algebra shows that

$$f_1(\eta/2, m) \lesssim \frac{s\gamma^2}{p^{c-3}\eta^2} \asymp \frac{n^{1/2}}{p^{c-3}T^2\tau(\log p)^2} = o(1), \tag{2.6.5}$$

since the order of p^{c-3} dominates the order of $n^{1/2}$. Moreover,

$$f_2(\eta/2, m) \le 2p\exp\Big\{ - \frac{c_1\gamma^2\log p}{c_2\gamma + c_3}\Big\} \le 2p\exp\{-c_4\log p\} = o(1), \tag{2.6.6}$$

by a proper choice of constant c_1, c_2, c_3. Also, by assumption (A5), $\eta\sqrt{\log p} = o(1)$ and

$$\eta\sqrt{\log(1/\eta)} \lesssim \frac{T\sqrt{s\tau}\gamma\log p}{n^{1/4}}\sqrt{\log n} = o(1).$$

Thus the proof is completed. □

Proof of Theorem 2.2.5. First, we collect several useful results.

(i) With probability at least $1 - 4p^{-c_1}$, $\|\Omega_x - \widehat{\Omega}_x\|_1 \le \|\Omega_x\|_1^2 s\gamma\tau^2T^2(\log p)^{3/2}n^{-1/2}$ and

$\|\Omega_x - \widehat{\Omega}_x\|_{\max} \le \|\Omega_x\|_1^2 \gamma\tau^2 T^2 (\log p)^{3/2} n^{-1/2}$ by Theorem 2.2.1. Therefore, $\|\Omega_x - \widehat{\Omega}_x\|_1 = o(1)$ and $\|\Omega_x - \widehat{\Omega}_x\|_{\max} = o(1)$ by assumption (A2).

(ii) With probability at least $1-2p^{-c_2}$, $|\mu - \widehat{\mu}|_\infty \lesssim \gamma\tau^2\sqrt{\frac{\log p}{n}} + |\widehat{\beta} - \beta^*|_1 = o(1)$ Lemma 2.3.5 and the order comes from assumptions (A1) and (A2).

(iii) Similar to the proof of Lemma 2.3.5, we have with probability at least $1-2p^{-c_3}$,

$$\Big|\frac{1}{n}\sum_{i=1}^n \psi(\widehat{\varepsilon}_{ij})\psi(\widehat{\varepsilon}_{ik}) - \mathbb{E}[\psi(\varepsilon_{ij})\psi(\varepsilon_{ik})]\Big|_\infty \lesssim \gamma\tau^2\sqrt{\frac{\log p}{n}} + |\widehat{\beta} - \beta^*|_1 = o(1).$$

(iv) Similar to the proof of Lemma 2.3.3, we have with probability at least $1-4p^{-c_4}$,

$$\Big\|\frac{1}{n}\sum_{i=1}^n X_i X_i^\top w^2(X_i) - \mathbb{E}[X_i X_i^\top w^2(X_i)]\Big\|_{\max} \lesssim \gamma\tau^2 T^2 (\log p)^{3/2} n^{-1/2} = o(1).$$

Repeatedly using Lemma 2.6.1, we get

$$\begin{aligned}
\|\widehat{D} - D\|_{\max} &\lesssim \max_{1\le j,k\le p}\Big|\frac{1}{\mu_j\mu_k} - \frac{1}{\widehat{\mu}_j\widehat{\mu}_k}\Big| + \Big|\frac{1}{n}\sum_{i=1}^n \psi(\widehat{\varepsilon}_{ij})\psi(\widehat{\varepsilon}_{ik}) - \mathbb{E}[\psi(\varepsilon_{ij})\psi(\varepsilon_{ik})]\Big|_\infty \\
&\quad + 2\|\Omega_x - \widehat{\Omega}_x\|_{\max} + \Big\|\frac{1}{n}\sum_{i=1}^n X_i X_i^\top w^2(X_i) - \mathbb{E}[X_i X_i^\top w^2(X_i)]\Big\|_{\max} \\
&\lesssim \gamma\tau^2\sqrt{\frac{\log p}{n}} + |\widehat{\beta} - \beta^*|_1 + \gamma\tau^2 T^2 (\log p)^{3/2} n^{-1/2} \\
&\lesssim \gamma\tau^2 T^2 (\log p)^{3/2} n^{-1/2} + |\widehat{\beta} - \beta^*|_1
\end{aligned}$$

with probability at least $1-12p^{-c}$, where $c = \min_{1\le i\le 4} c_i$. □

Proof of Theorem 2.2.6. By theorem 2.2.3, we see that

$$\mathbb{P}\Big(|\sqrt{n}(\check{\beta} - \beta^*) - \sqrt{n}\Omega\nabla L_n(\beta^*)|_\infty > \zeta_1\Big) < \zeta_2,$$

where $\zeta_1\sqrt{1 \vee \log(p/\zeta_1)} = o(1)$ and $\zeta_2 = o(1)$. Define $\pi(v) = Cv^{1/3}(1 \vee \log(p/v))^{2/3}$ with

$C_2 > 0$ and

$$\Gamma = \|\widehat{D} - D\|_{\max}.$$

Let $c_z(\alpha) = \inf\{t \in \mathbb{R} : \mathbb{P}(|\sum_{i=1}^n z_i/\sqrt{n}|_\infty \le t) \ge 1 - \alpha\}$, where the sequence $\{z_i\}$ is defined in theorem 2.2.4. From the proof of Lemma 3.2 in [27], we have

$$\mathbb{P}\Big(c(\alpha) \le c_z(\alpha + \pi(v))\Big) \ge 1 - \mathbb{P}(\Gamma > v) \tag{2.6.7}$$

$$\mathbb{P}\Big(c_z(\alpha) \le c(\alpha + \pi(v))\Big) \ge 1 - \mathbb{P}(\Gamma > v) \tag{2.6.8}$$

Therefore, by theorem 2.2.4, (2.6.7) and (2.6.8), we have for every $v > 0$,

$$\begin{aligned}
&\sup_{\alpha\in(0,1)} \left|\mathbb{P}\Big(\sqrt{n}\Omega\nabla L_n(\beta^*) > c(\alpha)\Big) - \alpha\right| \\
&\lesssim \sup_{\alpha\in(0,1)} \left|\mathbb{P}\Big(|\sum_{i=1}^n z_i/\sqrt{n}|_\infty > c(\alpha)\Big) - \alpha\right| + o(1) \\
&\lesssim \pi(v) + \mathbb{P}(\Gamma > v) + o(1)
\end{aligned}$$

Furthermore, following the same spirit as the proof of Theorem 3.2 in [27], we see that

$$\begin{aligned}
&\sup_{\alpha\in(0,1)} \left|\mathbb{P}\Big(\sqrt{n}|\check{\beta} - \beta^*|_\infty > c(\alpha)\Big) - \alpha\right| \\
\lesssim\ & \pi(v) + \mathbb{P}(\Gamma > v) + \zeta_1\sqrt{1 \vee \log(p/\zeta_1)} + \zeta_2 + o(1).
\end{aligned}$$

Now that $\zeta_1\sqrt{1 \vee \log(p/\zeta_1)} = o(1)$ and $\zeta_2 = o(1)$ from Theorem 2.2.3, we only need to choose $v > 0$, such that $\pi(v) = o(1)$ and $\mathbb{P}(\Gamma > v) = o(1)$. Let $v \asymp s\gamma\tau^2T^2(\log p)^{3/2}n^{-1/2} + |\widehat{\beta} - \beta^*|_1$. Then we see that the conditions that $\mathbb{P}(\Gamma > v) = o(1)$ and $\pi(v) = o(1)$ are satisfied by Theorem 2.2.5 and the scaling hypothesis. □

2.7 Proofs of Results in Section 2.3

Proof of Lemma 2.3.1. Define the filtration $\{\mathcal{F}_i\}$ with $\mathcal{F}_i = \sigma(\varepsilon_i, \varepsilon_{i-1}, \dots)$, and let $P_j(\cdot) = \mathbb{E}(\cdot|\mathcal{F}_j) - \mathbb{E}(\cdot|\mathcal{F}_{j-1})$ be a projection. Conventionally it follows that $P_j(G(X_i)) = 0$ for $j \geq i+1$. We can write

$$\sum_{i=1}^{n} G(X_i) - \mathbb{E}G(X_i) = \sum_{j=-\infty}^{n} \left(\sum_{i=1}^{n} P_j(G(X_i)) \right) =: \sum_{j=-\infty}^{n} L_j,$$

where $L_j = \sum_{i=1}^{n} P_j(G(X_i))$. By the Markov inequality, for $\lambda > 0$, we have

$$\begin{aligned} &\mathbb{P}\bigg(\sum_{i=1}^{n} G(X_i) - \mathbb{E}G(X_i) \geq 2x \bigg) \leq \mathbb{P}\bigg(\sum_{j=-\infty}^{-s} L_j \geq x \bigg) + \mathbb{P}\bigg(\sum_{j=-s+1}^{n} L_j \geq x \bigg) \\ &\leq \mathrm{e}^{-\lambda x} \mathbb{E}\bigg[\exp\bigg\{ \lambda \sum_{j=-\infty}^{-s} L_j \bigg\} \bigg] + \mathrm{e}^{-\lambda x} \mathbb{E}\bigg[\exp\bigg\{ \lambda \sum_{j=-s+1}^{n} L_j \bigg\} \bigg], \end{aligned} \tag{2.7.1}$$

for some $s > 0$ to be determined later. We shall bound the right-hand side of (2.7.1) with a suitable choice of $\lambda > 0$. Observing that $\{L_j\}_{j \leq n}$ is a sequence of martingale differences with respect to $\{\mathcal{F}_j\}$, we then seek an upper bound on $\mathbb{E}[\mathrm{e}^{\lambda L_j} \big| \mathcal{F}_{j-1}]$. It follows that

$$\begin{aligned} |L_j| &\leq \sum_{i=1 \vee j}^{n} \min\left\{ \left| \mathbb{E}\left[G(X_i) \big| \mathcal{F}_j \right] - \mathbb{E}\left[G(X_i) | \mathcal{F}_{j-1} \right] \right|, 2B \right\} \\ &\leq \sum_{i=1 \vee j}^{n} \min\left\{ \|A^{i-j}\|_\infty \mathbb{E}\left[|\varepsilon_j - \varepsilon_j'|_\infty \big| \mathcal{F}_j \right], 2B \right\} \\ &\leq \sum_{i=1 \vee j}^{n} \min\left\{ p\rho^{-1} \gamma \rho^{(i-j)/\tau} \eta_j, 2B \right\}, \end{aligned} \tag{2.7.2}$$

where ε_j' is an i.i.d. copy of ε_j and $\eta_j = \mathbb{E}\big[|\varepsilon_{j1} - \varepsilon_{j1}'| \big| \mathcal{F}_j \big]$.

Denote $s = \lfloor \tau \log p / \log(1/\rho) \rfloor + 1$. Note that $s > 0$ is a positive integer. For

$-s < j \leq 0$, we have

$$\begin{aligned}
|L_j| &\leq \sum_{i=0}^{\infty} \min\left\{p\rho^{-1}\gamma\rho^{(i-j)/\tau}\eta_j, 2B\right\} \\
&\leq \sum_{i=0}^{s-1} \min\left\{p\rho^{-1}\gamma\rho^{(i-j)/\tau}\eta_j, 2B\right\} + \sum_{i=s}^{\infty} \min\left\{p\rho^{-1}\gamma\rho^{(i-j)/\tau}\eta_j, 2B\right\} \\
&\leq 2sB + \sum_{i=0}^{\infty} \min\left\{\rho^{-1}\gamma\rho^{i/\tau}\eta_j, 2B\right\}
\end{aligned}$$

For $0 < j \leq n$, we also have

$$|L_j| \leq \sum_{i=j}^{\infty} \min\left\{p\rho^{-1}\gamma\rho^{(i-j)/\tau}\eta_j, 2B\right\} \leq -2sB + \sum_{i=0}^{\infty} \min\left\{\rho^{-1}\gamma\rho^{i/\tau}\eta_j, 2B\right\}$$

Basic algebra shows that

$$\begin{aligned}
\mathbb{E}[|L_j|^k|\mathcal{F}_{j-1}] &\overset{(1)}{\leq} \mathbb{E}\left[\left(2sB + \sum_{i=0}^{\infty} \min\left\{\rho^{-1}\gamma\rho^{i/\tau}\eta_j, 2B\right\}\right)^k\right] \\
&\leq \mathbb{E}\left[2^k\left((2sB)^k + \left(\sum_{i=0}^{\infty} \min\left\{\rho^{-1}\gamma\rho^{i/\tau}\eta_j, 2B\right\}\right)^k\right)\right] \\
&\leq 2^k\left[(2sB)^k + \left(\sum_{i=0}^{\infty} \left\| \min\left\{\rho^{-1}\gamma\rho^{i/\tau}\eta_j, 2B\right\} \right\|_k\right)^k\right], \qquad (2.7.3)
\end{aligned}$$

where (1) comes from the independence of η_j and $\mathcal{F}_{j-1}$. To analyze (2.7.3), we further compute

$$\begin{aligned}
\left\| \min\left\{\rho^{-1}\gamma\rho^{i/\tau}\eta_j, 2B\right\} \right\|_k &= \left\| 2B\mathbb{I}\left(\frac{\gamma}{\rho}\rho^{i/\tau}\eta_j \geq 2B\right) + \frac{\gamma}{\rho}\rho^{i/\tau}\eta_j\mathbb{I}\left(\frac{\gamma}{\rho}\rho^{i/\tau}\eta_j \leq 2B\right) \right\|_k \\
&\leq 2B\left(\mathbb{P}\left(\frac{\gamma}{\rho}\rho^{i/\tau}\eta_j \geq 2B\right)\right)^{1/k} + \mathbb{E}\left[\left(\frac{\gamma}{\rho}\rho^{i/\tau}\eta_j\right)^2(2B)^{k-2}\right]^{1/k} \\
&\leq \left(4\sigma^2\frac{\gamma^2}{\rho^2}\right)^{1/k}\rho^{2i/\tau k}(2B)^{1-2/k} \qquad (2.7.4)
\end{aligned}$$

Plugging (2.7.4) into (2.7.3) yields, for some constant $C_1, C_2 > 0$, that

$$\begin{aligned}
\mathbb{E}[|L_j|^k|\mathcal{F}_{j-1}] &\leq 2^k\Big[(2sB)^k + 4\sigma^2\frac{\gamma^2}{\rho^2}(2B)^{k-2}\Big(\frac{1}{1-\rho^{2/\tau k}}\Big)^k\Big] \\
&\overset{(1)}{\leq} 2^k\Big[(2sB)^k + 4\sigma^2\frac{\gamma^2}{\rho^2}(2B)^{k-2}\Big(\frac{\tau k}{2}\Big)^k\rho^{-2/\tau}\Big(\log(1/\rho)\Big)^k\Big] \\
&\overset{(2)}{\leq} 2^k\Big[(2sB)^k + C_1\gamma^2 B^{-2}C_2^k B^k\tau^k k!\Big] \\
&\leq \gamma^2(Bs\tau)^k k![4 + C_1B^{-2}(2C_2)^k] \\
&\leq \gamma^2(Bs\tau)^k k!(1+C_1B^{-2})(4+2C_2)^k, \qquad (2.7.5)
\end{aligned}$$

where (1) uses the inequality that $1 - x \geq -x\log x$ for $x \in (0,1)$ and (2) uses Stirling formula and the fact that $\rho^{-2/\tau} \leq \rho^{-2}$. Let $\tilde{C}_1 = 1 + C_1B^{-2}$ and $\tilde{C}_2 = 4 + 2C_2$. Then we obtain

$$\begin{aligned}
\mathbb{E}\Big[\mathrm{e}^{\lambda L_j}|\mathcal{F}_{j-1}\Big] &\leq 1 + \sum_{k=2}^{\infty}\Big[\tilde{C}_1\gamma^2(\tilde{C}_2Bs\tau\lambda)^k\Big] = 1 + \frac{\tilde{C}_1\gamma^2\tilde{C}_2^2(Bs\tau)^2\lambda^2}{1-\tilde{C}_2Bs\tau\lambda} \\
&\leq \exp\Big\{\frac{\tilde{C}_1\gamma^2\tilde{C}_2^2(Bs\tau)^2\lambda^2}{1-\tilde{C}_2Bs\tau\lambda}\Big\}. \qquad (2.7.6)
\end{aligned}$$

Furthermore,

$$\mathbb{E}\Big[\exp\Big\{\lambda\sum_{j=s}^{n}L_j\Big\}\Big] \leq \exp\Big\{\frac{\tilde{C}_1\gamma^2\tilde{C}_2^2(Bs\tau)^2(s+n)\lambda^2}{1-\tilde{C}_2Bs\tau\lambda}\Big\}. \qquad (2.7.7)$$

Take $\lambda = x(\tilde{C}_2Bs\tau x + 2\tilde{C}_1\gamma^2\tilde{C}_2^2(Bs\tau)^2(s+n))^{-1}$ and by (2.7.1) we have

$$\begin{aligned}
&\mathbb{P}\Big(\sum_{j=-s+1}^{n}L_j \geq x\Big) \\
\leq\ &\exp\Big\{-\frac{x^2}{(1+C_1B^{-2})\gamma^2B^2\tau^4(\log p)^2(\tau\log p + n) + C_4\tau^2B(\log p)x}\Big\}.
\end{aligned}$$

Similarly, for $j \leq -s$, since $p \leq \rho^{-s/\tau}$,

$$|L_j| \leq \sum_{i=0}^{\infty} \min\left\{\rho^{-1}\gamma\rho^{(i-j-s)/\tau}\eta_j, 2B\right\}.$$

By the same argument, we immediate have

$$\mathbb{P}\bigg(\sum_{j=-\infty}^{-s} L_j \geq x\bigg) \leq \exp\left\{-\frac{x^2}{C_3\gamma^2\tau^3 + C_4\tau B x}\right\}, \tag{2.7.8}$$

where $C_3 = 32\mathrm{e}^2\sigma^2(2\pi)^{-1/2}[\rho^2\log(1/\rho)]^{-3}$ and $C_4 = 8\mathrm{e}[\log(1/\rho)]^{-1}$. By (2.7.8), (2.7.8) and symmetrization argument, we complete the proof. □

Proof of Corollary 2.3.2. It follows from the proof of lemma 2.3.1 without any extra technical difficulty. □

Proof of Theorem 2.3.3. Let $G_{jk} : \mathbb{R}^p \to \mathbb{R}$ be defined as

$$G_{jk}(x) = \left(xx^\top w(x)\right)_{jk} = x_j x_k w(x) \text{ for } j, k = 1, \ldots, p,$$

and hence $|G(x)| \leq T$. Let $u(x) = w^{1/3}(x)$. Observe that

$$\begin{aligned}
& |G_{jk}(x) - G_{jk}(y)| \\
\leq\ & |x_j u(x)\, x_k u(x) - y_j u(y)\, y_k u(y)| u(x) + |y_j u(y)\, y_k u(y)||u(x) - u(y)| \\
\leq\ & |x_i u(x) - y_i u(y)||x_j u(x)| + |x_j u(x) - y_j u(y)||y_i u(y)| + T^2|u(x) - u(y)| \\
\leq\ & 3T|x - y|_\infty.
\end{aligned}$$

By lemma 2.3.1 and taking $x = c\gamma\tau^2 T n^{-1/2}(\log p)^{3/2}$, we have

$$\begin{aligned}
& \mathbb{P}\Big(|\widehat{\Sigma}_{x,jk} - \Sigma_{x,jk}| \geq cTx\Big) \\
=\ & \mathbb{P}\bigg(\Big|\frac{1}{n}\sum_{i=1}^{n} G_{jk}(X_{i-1}) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}G_{jk}(X_{i-1})\Big| \geq cTx\bigg) \leq 4p^{-c_1}.
\end{aligned}$$

A union bound yields

$$\mathbb{P}\bigg(\|\widehat{\Sigma}_x - \Sigma_x\|_{\max} \geq cTx\bigg) \leq 4p^{-c_0},$$

where $c_0 = c_1 - 1 > 0$. □

Proof of Theorem 2.3.4. Denote $H = \nabla^2 L_n(\beta^*)$. We write $\|H - \Sigma\|_{\max}$ as

$$\max_{1\leq k\leq p} \bigg\|\frac{1}{n}\sum_{i=1}^{n} \psi'(\varepsilon_{ik})X_{i-1}X_{i-1}^{\top}w(X_{i-1}) - \mathbb{E}[\psi'(\varepsilon_{ik})X_{i-1}X_{i-1}^{\top}w(X_{i-1})]\bigg\|_{\max}.$$

Using Corollary 2.3.2, it follows from the same argument of the proof of Theoremt 2.3.3 that for some constant $c_1 > 1$, with probability at least $1 - 4p^{-c_1}$,

$$\|H - \Sigma\|_{\max} \lesssim \gamma\tau^2 T^2 n^{-1/2}(\log p)^{3/2}.$$

Finally, a union bound over $1 \leq k \leq p$ yields the conclusion. □

Proof of Lemma 2.3.5. The strategy is to consider each component of $\widehat{\mu}$ and take a union bound. Observe that

$$|\widehat{\mu}_k - \mu_k| \leq \bigg|\frac{1}{n}\sum_{i=1}^{n}\psi'(\widehat{\varepsilon}_{ik}) - \mathbb{E}[\psi'(\widehat{\varepsilon}_{ik})]\bigg| + |\mathbb{E}\psi'(\widehat{\varepsilon}_{ik}) - \mathbb{E}\psi'(\varepsilon_{ik})|, \quad k = 1, 2, \ldots, p.$$

Since $|\psi''|$ is bounded, by the mean value theorem, we have that for some ξ between x and y,

$$|\psi'(x) - \psi'(y)| = |\psi''(\xi)(x - y)| \lesssim |x - y|.$$

So it can be verified that $\psi'(X_{ik} - X_{i-1}^{\top}\widehat{\beta}_k)$ satisfies the conditions in Corollary 2.5 of [85]. By Corollary 2.5 of [85], it holds that

$$\bigg|\frac{1}{n}\sum_{i=1}^{n}\psi'(\widehat{\varepsilon}_{ik}) - \mathbb{E}[\psi'(\widehat{\varepsilon}_{ik})]\bigg| \lesssim \gamma\tau^2\sqrt{\frac{\log p}{n}}$$

with probability at least $1 - 2p^{-c}$ for some positive constant c. Moreover,

$$\max_{1\le k\le p} |\mathbb{E}\widehat{\varepsilon}_{ik} - \mathbb{E}\varepsilon_{ik}| \lesssim \max_{1\le k\le p} \mathbb{E}\big[|X_{i-1}^{\top}(\widehat{\beta}_k - \beta^*)|\big] \lesssim |\widehat{\beta} - \beta^*|_1,$$

where the last inequality comes from the fact that X_{i-1} has bounded second moment. □

2.8 Proofs of Result in Section 2.4

Before proving Theorem 2.4.1, we will first state and prove the corresponding lemmas in the outline listed at the end of section 2.4.

Lemma 2.8.1. *Suppose $\mathbb{E}[\varepsilon_{ik}^2] \le \sigma^2$ for all $1 \le k \le p$ and the odd function $\psi(\cdot)$ satisfies that $|\psi(\cdot)| \le C$ and $|\psi'(\cdot)| \le C$, then we have*

$$\mathbb{P}\bigg(\big|(T_Y - T_{Y,m})/\sqrt{n}\big|_\infty \ge x\bigg) \le \frac{c_1 s p^3 \gamma^2 \rho^{m/\tau}}{x^2} =: f_1(x, m),$$

for some constants $C_1, C_2 > 0$.

Proof of Lemma 2.8.1. Let $D_i = Y_i - Y_{i,m}$. For any $\lambda > 0$, by Markov inequality we have

$$\mathbb{P}\bigg(\sum_{i=1}^n D_{ij}/\sqrt{n} \ge x\bigg) \le \frac{\mathbb{E}\big[\big(\sum_{i=1}^n D_{ij}/\sqrt{n}\big)^2\big]}{x^2}. \tag{2.8.1}$$

Notice that the martingale difference $\{D_{ij}\}_{i=1}^n$ satisfies

$$|D_{ij}| \lesssim \sqrt{s}|X_i - X_{i,m}|_\infty.$$

Thus,

$$\|D_{ij}\|_2 \le \||X_i - X_{i,m}|_\infty\|_2 \le \sum_{l=m+1}^{\infty} \|A^l\|_\infty \||\varepsilon_{i-l}|_\infty\|_2 \lesssim \sqrt{s} p \gamma \rho^{m/\tau}.$$

By Burkholder inequality ([20]), we have

$$\mathbb{E}\Big[\Big(\sum_{i=1}^{n} D_{ij}/\sqrt{n}\Big)^2\Big] \lesssim \mathbb{E}[|D_{ij}|^2] \lesssim sp^2\gamma^2\rho^{2m/\tau} \tag{2.8.2}$$

Hence, by (2.8.1),

$$\mathbb{P}\Big(\sum_{i=1}^{n} D_{ij}/\sqrt{n} \geq x\Big) \leq \frac{c_1' sp^2\gamma^2\rho^{m/\tau}}{x^2}$$

Finally, symmetrization and a union bound give the desired result. □

Lemma 2.8.2. *Under the assumptions in Lemma 2.8.1, it holds that*

$$\mathbb{P}\big(|T_{Y,S}|_\infty/\sqrt{n} \geq x\big) \leq 2p\exp\Big\{-\frac{nx^2}{C_1\sqrt{s}T\sqrt{n}x + C_2 mwsT^2\sigma^2}\Big\} =: f_2(x,m).$$

Proof of Lemma 2.8.2. By the property of $\psi(\cdot)$ and the mean value theorem, we have $|\psi(x)| \leq C|x|$. Consider the first coordinate $(T_{Y,S})_1$ of $T_{Y,S}$. We can write $(T_{Y,S})_1 = \sum_{i=j_1}^{j_r} Y_{i,m,1}$, where $r = m\omega$. Observe that $\{Y_{i,m,1}\}$ is a martingale difference adapted to the filtration $\{\mathcal{F}_i = \sigma(\varepsilon_i, \varepsilon_{i-1}, \dots)\}$ and that $|Y_{i,m,1}| \leq \psi(\varepsilon_{ik})\sqrt{s}T \leq C\sqrt{s}T$. We shall establish a Bernstein-type inequality for the sum of martingale differences $(T_{Y,S})_1$:

$$\mathbb{P}((T_{Y,S})_1 \geq x) \leq \mathrm{e}^{-\lambda x}\mathbb{E}\mathrm{e}^{\lambda\sum_{i=j_1}^{j_r} Y_{i,m,1}}, \quad \text{for any } \lambda > 0. \tag{2.8.3}$$

We now bound $\mathbb{E}\mathrm{e}^{\lambda\sum_{i=j_1}^{j_r} Y_{i,m,1}}$ from above. By the tower property,

$$\begin{aligned}\mathbb{E}\exp\Big\{\lambda\sum_{i=j_1}^{j_r} Y_{i,m,1}\Big\} &= \mathbb{E}\Big[\mathbb{E}\Big[\exp\Big\{\lambda\sum_{i=j_1}^{j_r} Y_{i,m,1}\Big\}\Big|\mathcal{F}_{r-1}\Big]\Big] \\ &= \mathbb{E}\Big[\exp\Big\{\lambda\sum_{i=j_1}^{j_{r-1}} Y_{i,m,1}\Big\}\mathbb{E}[\mathrm{e}^{\lambda Y_{j_r,m,1}}|\mathcal{F}_{j_r-1}]\Big] \end{aligned} \tag{2.8.4}$$

Now, consider

$$\begin{aligned}\mathbb{E}[\mathrm{e}^{\lambda Y_{j_r,m,1}}|\mathcal{F}_{j_r-1}] &= 1+\mathbb{E}\Big[\sum_{t=2}^{\infty}\frac{(\lambda Y_{j_r,m,1})^t}{t!}\Big|\mathcal{F}_{j_r-1}\Big]\\ &\leq 1+\mathbb{E}\Big[\lambda^2T^2s\psi^2(\varepsilon_{j_rk})\sum_{t=0}^{\infty}(\lambda T\sqrt{s}C)^t\Big]\\ &\overset{(1)}{\leq} 1+\frac{C\lambda^2T^2s\sigma^2}{1-C\lambda T\sqrt{s}}\leq \exp\left\{\frac{C\lambda^2T^2s\sigma^2}{1-C\lambda T\sqrt{s}}\right\} \qquad (2.8.5)\end{aligned}$$

where the inequality (1) makes use of the fact that $\psi^2(\varepsilon_{j_r,k})\leq\varepsilon^2_{j_r,k}$. Plug (2.8.5) into (2.8.4) and we obtain

$$\mathbb{E}\exp\Big\{\lambda\sum_{i=j_1}^{j_r}Y_{i,m,1}\Big\}\leq\exp\left\{\frac{C\lambda^2T^2s\sigma^2}{1-C\lambda T\sqrt{s}}\right\}\mathbb{E}\Big[\exp\Big\{\lambda\sum_{i=j_1}^{j_{r-1}}Y_{i,m,1}\Big\}\Big] \qquad (2.8.6)$$

Iterating this procedure yields

$$\mathbb{E}\exp\Big\{\lambda\sum_{i=j_1}^{j_r}Y_{i,m,1}\Big\}\leq\exp\left\{\frac{Cm\omega\lambda^2T^2s\sigma^2}{1-C\lambda T\sqrt{s}}\right\} \qquad (2.8.7)$$

Choose $\lambda=x(CT\sqrt{s}+2CmwT^2s\sigma^2)^{-1}$ and by (2.8.3) we have

$$\mathbb{P}((T_{Y,S})_1\geq x)\leq\exp\left\{-\frac{x^2}{C_1T\sqrt{s}x+C_2m\omega T^2s\sigma^2}\right\}.$$

The symmetrization argument and a union bound deliver the desired result. □

Lemma 2.8.3. *Suppose the scaling condition holds that $sT^2(\log(pn))^7/n\leq c_3n^{-c_4}$. Assume that $\mathbb{E}[X_{ik}]\leq C'$ for all $1\leq k\leq p$. Then we have the following Gaussian Approximation result that*

$$\mathcal{U}:=\sup_{t\in\mathbb{R}}\Big|\mathbb{P}\big(|T_{Y,L}/\sqrt{n}|_\infty\leq t\big)-\mathbb{P}\big(|Z|_\infty\leq t\big)\Big|\leq cn^{-c'}$$

for some constants $c,c'>0$.

Proof of Lemma 2.8.3. Recall that $\xi_b = \sum_{i \in L_b} Y_{i,m}/\sqrt{M}$, thus

$$\mathcal{U} = \sup_{t \in \mathbb{R}} \left| \mathbb{P}\big(|\frac{1}{\sqrt{w}} \sum_{b=1}^{w} \xi_b|_\infty \le t\big) - \mathbb{P}\big(|Z|_\infty \le t\big) \right|.$$

Observe that $\xi_1, \xi_2, \dots, \xi_w$ are independent random variables. We shall apply Corollary 2.1 of [27] by verifying the condition (E.1) therein. For completeness, the conditions are stated below.

(i) $c_1 \le \mathbb{E}[\xi_{bj}^2] \le c_2$ for all $1 \le j \le p$.

(ii) $\max_{k=1,2} \mathbb{E}[|\xi_{bj}|^{2+k}/B_n^k] + \mathbb{E}[\exp(|\xi_{bj}/B_n|)] \le 4$, for some $B_n > 0$ and all $1 \le j \le p$.

(iii) $B_n^2 (\log(pn))^7/n \le c_3 n^{-c_4}$.

To verify condition (i), we see that

$$\begin{aligned} \mathbb{E}[\xi_{bj}^2] &\le c\,\sigma^2 \mathbb{E}[\mathbb{E}[\Omega_{x,j}^\top X_i w(X_i) | \varepsilon_{i-m}, \dots, \varepsilon_i]^2] \\ &\le c\mathbb{E}[(\Omega_{x,j}^\top X_i w(X_i))^2] \le c\,\Omega_{x,j}^\top \Sigma_x \Omega_{x,j} \le c\,\Omega_{x,jj} \end{aligned}$$

where $\Omega_{x,j}$ is the j-th row of Ω_x and $\Omega_{x,jj}$ is the j-th diagonal entry of Ω_x. Now we check condition (ii). By Theorem 3.2 of [20], we have for $k \ge 2$,

$$\mathbb{E}[|\xi_{bj}|^k] \le 18 k^k \mathbb{E}\big[\big|Y_{ij,m}\big|^k\big] \lesssim k! \mathrm{e}^k \mathbb{E}\big[\big|Y_{ij,m}\big|^2\big] (\sqrt{s}T)^{k-2} \lesssim k! \mathrm{e}^k (\sqrt{s}T)^{k-2}.$$

Therefore, take $B_n = C\sqrt{s}T$ for sufficiently large $C > 0$ and we have

$$\mathbb{E}[\exp(|\xi_{bj}/B_n|)] \le 1 + C_1 \sum_{k=1}^{\infty} (\mathrm{e}/C)^k < 2.$$

Moreover, for a suitable choice of $C > 0$,

$$\max_{k=1,2} \mathbb{E}[|\xi_{bj}|^{2+k}/B_n^k] < 2.$$

Hence, condition (ii) is satisfied. Condition (iii) is guaranteed by the scaling assumption.

□

Now, we are ready to give the proof of Theorem 2.4.1.

Proof of Theorem 2.4.1. By triangle inequality,

$$\begin{aligned} \mathcal{H} &\leq \sup_{t\in\mathbb{R}}\Big|\mathbb{P}\big(|T_Y/\sqrt{n}|_\infty \leq t\big) - \mathbb{P}\big(|T_{Y,L}/\sqrt{n}|_\infty \leq t\big)\Big| \\ &\quad + \sup_{t\in\mathbb{R}}\Big|\mathbb{P}\big(|T_{Y,L}/\sqrt{n}|_\infty \leq t\big) - \mathbb{P}\big(|Z|_\infty \leq t\big)\Big| =: I + II. \qquad (2.8.8) \end{aligned}$$

For any $\eta > 0$, elementary calculation shows that

$$\begin{aligned} I &\leq \mathbb{P}\big(\big|(T_Y - T_{Y,L})/\sqrt{n}\big|_\infty > \eta\big) + \sup_{t\in\mathbb{R}}\mathbb{P}\Big(\Big|\big|T_{Y,L}/\sqrt{n}\big|_\infty - t\Big| \leq \eta\Big) \\ &\leq \mathbb{P}\big(\big|(T_Y - T_{Y,m})/\sqrt{n}\big|_\infty > \frac{\eta}{2}\big) + \mathbb{P}\big(\big|T_{Y,S}/\sqrt{n}\big|_\infty > \frac{\eta}{2}\big) \\ &\quad + \sup_{t\in\mathbb{R}}\mathbb{P}\Big(\Big|\big|T_{Y,L}/\sqrt{n}\big|_\infty - t\Big| \leq \eta\Big) \end{aligned}$$

By lemma 2.8.2 and 2.8.1,

$$\mathbb{P}\big(\big|(T_Y - T_{Y,m})/\sqrt{n}\big|_\infty > \frac{\eta}{2}\big) \leq f_1(\eta/2, m), \qquad (2.8.9)$$

$$\mathbb{P}\big(\big|T_{Y,S}/\sqrt{n}\big|_\infty > \frac{\eta}{2}\big) \leq f_2(\eta/2). \qquad (2.8.10)$$

By lemma 2.8.3 and theorem 3 of [26], we obtain that

$$\begin{aligned} \sup_{t\in\mathbb{R}}\mathbb{P}\Big(\Big|\big|T_{Y,L}/\sqrt{n}\big|_\infty - t\Big| \leq \eta\Big) &\leq \sup_{t\in\mathbb{R}}\mathbb{P}\Big(\Big|\big|Z\big|_\infty - t\Big| \leq \eta\Big) + \mathcal{U} \\ &\lesssim \eta\sqrt{\log p} + \eta\sqrt{\log(1/\eta)} + cn^{-c'}, \qquad (2.8.11) \end{aligned}$$

and that

$$II = \mathcal{U} \leq cn^{-c'}. \qquad (2.8.12)$$

By (2.8.8), (2.8.9), (2.8.10), (2.8.11) and (2.8.12), we obtain the inequality stated in the theorem. □

Chapter 2, in full, is currently being prepared for submission of the material "High-dimensional Simultaneous Inference on non-Gaussian VAR Model via De-biased Estimator", Liu, Linbo and Zhang, Danna. The dissertation author was the primary investigator and author of this paper.

Chapter 3

Robust Multivariate Time-Series Forecasting: Adversarial Attacks and Defense Mechanisms

3.1 Introduction

Understanding the robustness for time-series models has been a long-standing issue with applications across many disciplines such as climate change [99], financial market analysis [4, 51], down-stream decision systems in retail [16], resource planning for cloud computing [107, 108], and optimal control of vehicles [68]. In particular, the notion of robustness defines how sensitive the model output is when authentic data is (potentially) perturbed with noises. In practice, as observation data are often corrupted by measurement noises, it is important to develop statistical forecasting models that are less sensitive to such noises [18, 17, 129] or more stable against outliers that might arise from such corruption [30, 43, 85, 135]. However, these approaches have not considered the possibility of adversarial noises which are strategically created to mislead the model rather than being sampled from a known distribution.

As a matter of fact, vulnerabilities against such adversarial noises have been previously pointed out [127, 47] in classification. In practice, it has been shown that human-imperceptible adversarial perturbation can alter classification outcomes of a deep learning (DL) model, revealing a severe threat to many safety-critical systems . As such a risk is associated with the high capacity to fit complex data pattern of DL, we postulate that similar threats might also occur in forecasting where modern DL-based forecasting models [109, 115, 82, 138, 106] have become the dominant approach. For example, to mislead the forecasting of a particular stock, the adversaries might attempt to alter some features external to the stock's financial valuation to maximize the gap between predictions of its values on authentic and altered features. The feasibility of such an adversarial attack has been recently demonstrated with tweet messages [146] on a text-based stock forecasting.

Motivated by these real scenarios, we propose to investigate such adversarial threats on more practical forecasting models whose predictions are based on more precise features, e.g. valuations of other stock indices. Intuitively, rather than releasing adverse information to alter the sentiment about the target stock on social media, the adversaries can instead invest hence change the valuation adversely for a selected subset of stock indices (not including the target stock) which is arguably harder to detect. Interestingly, despite being seemingly plausible given the vast literature on adversarial attack for classification models, formulating such imperceptible attack under a multivariate forecasting setup is not straightforward. This is due to several differences between forecasting and classification, particularly in terms of unique characteristic of time series, e.g., multi-step predictions, correlation over multiple time series, and probabilistic predictions.

These differences open up the question of how adversarial perturbations and robustness should be defined more properly in time series setting. Although there have been a few recent studies in this direction based on randomized smoothing [149], these approaches are all restricted to univariate forecasting where the attack has to make adverse alterations directly to the target time series. Thus, under the less studied scenario of multivariate time-series forecasting setup, it remains unclear whether the attack to a target time series can be made instead via perturbing the other correlated time series;

and whether it is defensible against such adversarial threats. In particular, as illustrated above in the stock forecasting example, there are new regimes of sparse and indirect cross time series attack under multivariate time-series scenarios, which are more effective and realistic than the direct attack in univariate cases.

In order to understand whether such new regimes of attack exists and can be defended against, we raise three questions:

1. **Indirect Attack.** Can we mislead the prediction of some target time series via perturbations on the other time series?

2. **Sparse Attack.** Can such perturbations be sparse and non-deterministic to be less perceptible?

3. **Robust Defense.** Can we defend against those indirect and imperceptible attacks?

Here we summarize our technical contributions by answering the questions above:

Regarding **indirect attack**, we provide general framework of adversarial attack in multivariate time series (see Section 3.3.1). Then, we devise a deterministic attack (see Section 3.3.2) to the state-of-the-art probabilistic multivariate forecasting model. The attack changes the model's prediction on the target time series via adversely perturbing a subset of other time series. This is achieved via formulating the perturbation as solution of an optimization task with packing constraints.

Regarding **sparse attack**, we develop a non-deterministic attack (see Section 3.3.3) that adversely perturbs a stochastic subset of time series related to the target time series, which makes the attack less perceptible. This is achieved via a stochastic and continuous relaxation of the above packing constraint which are shown (see Section 3.5) to be more effective than the deterministic attack in certain cases. Moreover, unlike deterministic attack, its differentiability makes it suitable to be directly integrated as part of a differentiable defense mechanism that can be optimized via gradient descent in an end-to-end fashion, as discussed later in Section 3.4.2.

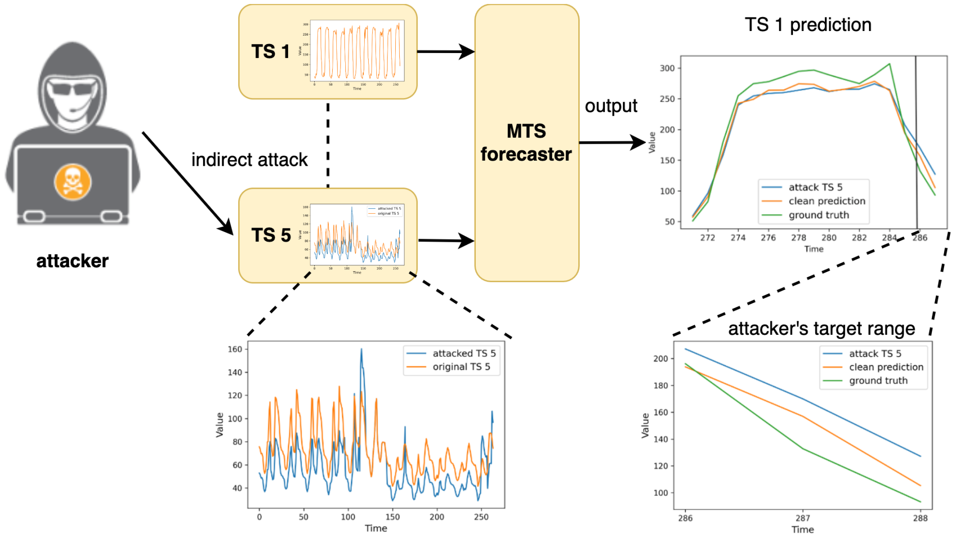


Figure 3.1. Illustration figure: an attacker misleads prediction of time series (TS) 1 at time 288 by indirectly attacking TS 5. Left plot of is authentic (orange) and perturbed (blue) versions of TS 5; right plot is no-attack (orange) and under-attack (blue) predictions for TS 1. Ground truth (green) is also plotted for comparison. No alteration is made to TS 1 but the prediction of TS 1 at the attack time step ($t = 288$) is adversely altered in the under-attack (blue) setting, which can set the prediction of TS 1 significantly away from the ground truth.

Regarding **robust defense**, we propose two defense mechanisms. First, we adapt randomized smoothing to the new multivariate forecasting setup with robust certificate. Second, we devise a defense mechanism (see Section 3.4.2) via solving a mini-max optimization task which minimizes the maximum expected damage caused by the probabilistic attack that continually updates the generation of its adverse perturbations in response to the model updates. Their effectiveness are demonstrated across extensive experiments in Section 3.5.

Furthermore, our experiments in Section 3.5.3 demonstrate that attacks designed for univariate cases cannot be reused as an effective attack to multivariate forecasting models, which highlights the importance and novelty of our studies.

3.2 Related Work

Deep Forecasting Models. The recent decades have witnessed a tremendous progress in DNN-based forecasting models. Given the temporal dependency of time series data, RNN and CNN-based architectures have been proved a success for time series forecasting tasks, see [109, 82, 138, 115] and [104, 8] respectively. To model the uncertainty, various probabilistic models have been proposed from distributional outputs [115, 35, 109] to distribution-free quantile-based outputs [106, 42, 66]. In multivariate cases, [114] generalized DeepAR [115] to multivariate cases and adopted low-rank Gaussian copula process to tackle the high-dimensionality challenge.

Adversarial Attack. Despite its success in various tasks, deep neural network is especially vulnerable to adversarial attacks [127] in the sense that even imperceptible adversarial noise can lead to completely different prediction. In computer vision, many adversarial attack schemes have been proposed. See [47, 92] for attacking image classifiers and [33] for attacking graph structured data. In the field of time series, there is much less literature and even so, most existing studies on adversarial robustness of MTS models [98, 54] are restricted to regression and classification settings. Alternatively, [149] studied both adversarial attacks to probabilistic forecasting models, which is only restricted to univariate settings.

Adversarial Robustness and Certification. Against adversarial attacks, an extensive body of work has been devoted to quantifying model robustness and defense mechanisms. For instance, Fast-Lin/Fast-Lip [141] recursively computes local Lipschitz constant of a neural network; PROVEN [140] certifies robustness in a probabilistic approach. Recently, randomized smoothing has gained increasing popularity as to enhance model robustness, which was proposed by [29, 79] as a defense approach with certification guarantee. To the time series setting, [149] adopted randomized smoothing technique to univariate forecasting models and developed theory therein. However, we are not aware of any prior works on randomized smoothing for multivariate probabilistic models.

3.3 Adversarial Attack Strategies

We provide a generic framework of sparse and indirect adversarial attack under a multivariate setting in Section 3.3.1. Then, a deterministic one to this task is introduced next in Section 3.3.2, followed by a stochastic attack derived in Section 3.3.3.

Notations. Denote d-dimensional multivariate time series $\mathbf{x}_t \in \mathbb{R}^d$ at time t with its observation of i-th time series $x_{i,t} = [\mathbf{x}_t]_i$. We denote $\mathbf{x} = \{\mathbf{x}_t\}_{t=1}^T \in \mathbb{R}^{d\times T}$ and $\mathbf{z} = \{\mathbf{x}_{T+t}\}_{t=1}^{\tau} \in \mathbb{R}^{d\times\tau}$ as recent T historical observations and next τ-step of the future values respectively. Then, probabilistic forecaster p_θ with parameterzation θ takes history $\mathbf{x}$ to predict $\mathbf{z}$, i.e., $\mathbf{z} \sim p_\theta(\cdot \mid \mathbf{x})$. We denote the set $[d] = \{1, \ldots, d\}$ and i-th time series as $\boldsymbol{\delta}^i = ([\boldsymbol{\delta}_t]_i)_{t=1}^T$.

3.3.1 Framework on Sparse and Indirect Adversarial Attack

Given an adversarial prediction target $\mathbf{t}_{\text{adv}}$ and historical input $\mathbf{x}$ to the forecaster $p_\theta(\mathbf{z}|\mathbf{x})$, we design a perturbation matrix $\boldsymbol{\delta}$ such that the perturbed input $\mathbf{x} + \boldsymbol{\delta}$ disturbs a statistic $\chi(\mathbf{z})$ as close as possible to $\mathbf{t}_{\text{adv}}$. That is, we find $\boldsymbol{\delta}$ such that the distance between $\mathbb{E}_{\mathbf{z}|\mathbf{x}+\boldsymbol{\delta}}[\chi(\mathbf{z})]$ and $\mathbf{t}_{\text{adv}}$ is minimized. Here, $\chi(\mathbf{z})$ and $\mathbf{t}_{\text{adv}}$ are any arbitrary function of interest or adversarial target values with the same dimension. We focus on scenarios where the perturbed prediction is far way from original prediction by properly choosing $\chi(\cdot)$ and $\mathbf{t}_{\text{adv}}$.

Thus, suppose the adversaries want to mislead the forecasting of time series in a subset $\mathcal{I} \subset [d]$, denoted as $\mathbf{z}^{\mathcal{I}}$. Let χ be a statistic function of interest that concerns only time series in $\mathcal{I}$, i.e. $\chi(\mathbf{z}) = \chi(\mathbf{z}^{\mathcal{I}})$. To make the attack less perceptible, we impose the following sparse and indirect constraints: First, perturbation $\boldsymbol{\delta}$ cannot be direct to target time series in $\mathcal{I}$ and can be indirectly applied to a small subset of $\mathcal{I}^c = [d] \setminus \mathcal{I}$. In other words, we restrict $\boldsymbol{\delta}^{\mathcal{I}} = \mathbf{0}$ and $s(\boldsymbol{\delta}) = |\{i \in \mathcal{I}^c : \boldsymbol{\delta}^i \neq \mathbf{0}\}| \leq \kappa$ with sparsity level $\kappa \leq d$.

Lastly, to avoid outlier detection, we also cap the energy of the attack such that the value of the perturbation at any coordinates is no more than a pre-defined threshold η. To sum up, the sparse and indirect attack $\boldsymbol{\delta}$ can be found via solving

$$\begin{aligned} \underset{\boldsymbol{\delta}\in\mathbb{R}^{T\times d}}{\text{minimize}} \quad & \left\{F(\boldsymbol{\delta}) \quad \triangleq \quad \left\| \mathbb{E}_{p_\theta(\mathbf{z}|\mathbf{x}+\boldsymbol{\delta})}\left[\chi(\mathbf{z})\right] - \mathbf{t}_{\text{adv}} \right\|_2^2 \right\} \\ \text{subject to} \quad & \quad \|\boldsymbol{\delta}\|_{\max} \leq \eta, \;\; s(\boldsymbol{\delta}) \leq \kappa, \;\; \boldsymbol{\delta}^{\mathcal{I}} = \mathbf{0}, \end{aligned} \tag{3.3.1}$$

where $\|\boldsymbol{\delta}\|_{\max} = \max_{t,i} |[\boldsymbol{\delta}_t]_i|$ is the element-wise maximum norm. As such, small values of κ and η imply a less perceptible attack. However, solving this is intractable due to the discrete cardinality constraint on $s(\boldsymbol{\delta})$. To sidestep this, we develop two approximations in the subsequent sections which correspond to our deterministic and non-deterministic attack strategies.

3.3.2 Deterministic Attack

Here we present an approximated solution. We first get an intermediate solution $\hat{\boldsymbol{\delta}}$ through projected gradient descent (PGD) until it converges,

$$\hat{\boldsymbol{\delta}} \;\leftarrow\; \prod_{\mathcal{B}_\infty(0,\eta)} \left(\hat{\boldsymbol{\delta}} - \alpha \nabla_{\boldsymbol{\delta}} F\left(\hat{\boldsymbol{\delta}}\right)\right), \tag{3.3.2}$$

where $\alpha \geq 0$ is a step size and $\prod_{\mathcal{B}_\infty(0,\eta)}$ is the projection onto the ℓ_∞-norm ball with radius η, allowing a simple element-wise clipping: $\prod_{\mathcal{B}_\infty(0,\eta)}([\hat{\boldsymbol{\delta}}_t]_i) = \text{sign}([\hat{\boldsymbol{\delta}}_t]_i)\,\eta$ if $|[\hat{\boldsymbol{\delta}}_t]_i| > \eta$ else $[\hat{\boldsymbol{\delta}}_t]_i$. With this intermediate non-sparse $\hat{\boldsymbol{\delta}}$, we retrieve for final sparse perturbation $\boldsymbol{\delta}$ via solving

$$\underset{\boldsymbol{\delta}\in\mathbb{R}^{T\times d}}{\text{minimize}} \quad \|\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}\|_{\text{F}} \quad \text{subject to} \quad s(\boldsymbol{\delta}) \;\leq\; \kappa, \;\; \boldsymbol{\delta}^{\mathcal{I}} \;=\; \mathbf{0}. \tag{3.3.3}$$

It turns out (3.3.3) can be solved analytically. Given $\hat{\boldsymbol{\delta}}$, we compute the absolute perturbation added to each row i, $p_i = \sum_{t=1}^{T} |[\hat{\boldsymbol{\delta}}_t]_i|$ for $i \in [d] \setminus \mathcal{I}$ and sort them in descending order

π: $p_{\pi_1} \geq \cdots \geq p_{\pi_d}$. Finally, we construct the solution as $\boldsymbol{\delta}$ with $\boldsymbol{\delta}^{\pi_i} = \hat{\boldsymbol{\delta}}^{\pi_i}$ if $i \leq \kappa$ else $\mathbf{0}$.

Remark. $\nabla_{\boldsymbol{\delta}} F(\boldsymbol{\delta})$ involves the computation of the gradient of an expectation, which doesn't have a closed-form solution. To overcome this intractability, we adopt the re-parameterized sampling approach used in [34] and [149].

Algorithm 1. Deterministic Adversarial Attack

```
input: pre-trained model $p_\theta(\mathbf{z} \mid \mathbf{x})$, observation $\mathbf{x}$ and other parameters:
    • statistic $\chi(\cdot)$, adversarial target $\mathbf{t}_{\text{adv}}$, target set $\mathcal{I} \subset [d]$
    • attack energy $\eta$, sparse constraint $\kappa$, PGD iterations $n$ and step size $\alpha \geq 0$
output: perturbation matrix $\boldsymbol{\delta} \in \mathbb{R}^{T \times d}$ s.t. $\|\boldsymbol{\delta}\|_{\max} \leq \eta,\ s(\boldsymbol{\delta}) \leq \kappa,\ \boldsymbol{\delta}^{\mathcal{I}} = \mathbf{0}$
1. initialize $\boldsymbol{\delta} = \mathbf{0}$
for iteration $1, 2, \ldots, n$ do
    2. compute the expected loss $F(\boldsymbol{\delta})$ using Eq. (3.3.1)
    3. update $\boldsymbol{\delta}$ via PGD in Eq. (3.3.2)
end for
4. for $i \notin \mathcal{I}$, compute $p_i = \sum_{t=1}^{T} |[\boldsymbol{\delta}_t]_i|$
5. sort $p_i$ in a descending order $\pi = (\pi_1, \ldots, \pi_d)$: $p_{\pi_1} \geq p_{\pi_2} \geq \cdots \geq p_{\pi_d}$.
6. set $\boldsymbol{\delta}^{\pi_{\kappa+1}} = \boldsymbol{\delta}^{\pi_{\kappa+2}} = \cdots = \boldsymbol{\delta}^{\pi_d} = \mathbf{0}$ and $\boldsymbol{\delta}^{\mathcal{I}} = \mathbf{0}$. Return $\boldsymbol{\delta}$.
```

3.3.3 Probabilistic Attack

To make the attack even less perceptible, we further show in this section an alternative approximation that results in a probabilistic sparse attack, which makes adverse alterations to a non-deterministic set of coordinates (i.e., time series and time steps). As shown in our experiment, this non-determinism appears to make the attack stronger and harder to detect.

To achieve this, we view the sparse attack vector as a random vector drawn from a distribution with differentiable parameterization. The core challenge is how to configure such a distribution whose support is guaranteed to be within the space of sparse vectors. To achieve this, we propose sparse layer, a distributional output, of a normal standard and a Dirac density combination. The output of this layer satisfied relaxed sparse support condition (see Theorem 2).

Sparse Layer. A sparse layer is defined as a distributional output $q(\boldsymbol{\delta}|\mathbf{x};\beta,\gamma)$ such that its sample (probablistic attack) $\boldsymbol{\delta} \sim q(\boldsymbol{\delta}|\mathbf{x};\beta,\gamma) = \prod_i q_i(\boldsymbol{\delta}^i|\mathbf{x};\beta,\gamma)$ satisfies sparse condition $\mathbb{E}[s(\boldsymbol{\delta})] \leq \kappa$ and $\boldsymbol{\delta}^{\mathcal{I}} = \mathbf{0}$. With $\boldsymbol{\delta}^i$ denoted as the i-th row (time series) of $\boldsymbol{\delta}$ and sparsity level κ, each factor distribution $q_i(\boldsymbol{\delta}^i|\mathbf{x};\beta,\gamma)$ parameterized by β and γ is defined as

$$q_i\Big(\boldsymbol{\delta}^i \mid \mathbf{x};\beta,\gamma\Big) \;\triangleq\; r_i\left(\gamma\right) \cdot q_i'\Big(\boldsymbol{\delta}^i \mid \mathbf{x};\beta\Big) \;+\; \Big(1 - r_i(\gamma)\Big) \cdot D\Big(\boldsymbol{\delta}^i\Big), \qquad (3.3.4)$$

where $r_i(\gamma) \triangleq \kappa\gamma_i^{\frac{1}{2}} \cdot (\sum_{i=1}^{d} \gamma_i)^{-\frac{1}{2}}/\sqrt{d}$, $D(\boldsymbol{\delta}^i) = \mathbb{I}(\boldsymbol{\delta}^i = \mathbf{0})$ is the Dirac density, and $q_i'(\boldsymbol{\delta}^i \mid \mathbf{x};\beta)$ is a Gaussian $\mathbb{N}(\mu(\mathbf{x};\beta), \sigma^2(\mathbf{x};\beta))$.

The combination weight $r_i(\gamma)$ denotes the probability mass of the event $\boldsymbol{\delta}^i = \mathbf{0}$, which is parameterized by γ. Intuitively, this means the choice of $\{r_i(\gamma)\}_{i=1}^{d}$ controls the row sparsity of the random matrix $\boldsymbol{\delta}$, which can be calibrated to enforce that $\mathbb{E}[s(\boldsymbol{\delta})] \leq \kappa$. We will show in Theorem 1 how samples can be drawn from the combined density in (4.3.7).

Theorem 1. Let $\boldsymbol{\delta}^{i'} \sim q_i'(\cdot \mid \mathbf{x};\beta,\gamma)$ and $u_i \sim \mathbb{N}(0,1)$ for $i = 1,\ldots,d$. Define $\boldsymbol{\delta}^i = \boldsymbol{\delta}^{i'} \cdot \mathbb{I}(u_i \leq \Phi^{-1}(r_i(\gamma)))$. Then, $\boldsymbol{\delta}^i \sim q_i(\boldsymbol{\delta}^i \mid \mathbf{x};\beta,\gamma)$.

Here, $q_i(\cdot|\mathbf{x};\beta,\gamma)$ is given in (4.3.7) and Φ^{-1} is the inverse cumulative of the standard normal distribution. We provide the proof in the appendix.

For implementation, we let $q_i'(\cdot \mid \mathbf{x};\beta)$ be a distribution over dense vectors, for example $\mathbb{N}(\mu(\beta), \sigma^2(\beta)\mathbf{I})$, and $u_i \sim \mathbb{N}(0,1)$ for $i \in [d]$. We can construct a binary mask $m_i = \mathbb{I}(u_i \leq \Phi^{-1}(r_i(\gamma)))$, $i \in [d]$, where $r_i(\gamma)$ is defined above. Next, for each $i \in [d]$, we draw $\boldsymbol{\delta}^{i'}$ from $q_i'(\cdot \mid \mathbf{x};\beta)$ and obtain $\boldsymbol{\delta}^i$ by $\boldsymbol{\delta}^i = \boldsymbol{\delta}^{i'} \cdot m_i$ where $\cdot$ denotes the element-wise multiplication. Finally, we set $\boldsymbol{\delta}^{\mathcal{I}} = \mathbf{0}$.

Theorem 2 proves that $\boldsymbol{\delta}$ sampled from (4.3.7) would meet the constraint $\mathbb{E}[s(\boldsymbol{\delta})] \leq \kappa$. Put together, Theorem 1 and Theorem 2 enable differentiable optimization of a sparse attack as desired.

Theorem 2. Let $\boldsymbol{\delta} \sim q(\cdot \mid \mathbf{x}; \beta, \gamma)$. Then, $\mathbb{E}[s(\boldsymbol{\delta})] \leq \kappa$.

Remark. Note that we can also obtain a direct sparse constraint on $s(\boldsymbol{\delta})$ by applying Theorem 2 to a smaller quantity $c\kappa$ for $c \in (0, 1)$. Then, by the Markov inequality, with probability at least $1 - c$, we have $s(\boldsymbol{\delta}) \leq \mathbb{E}[s(\boldsymbol{\delta})]/c = c\kappa/c = \kappa$. We provide the proof of Theorem 2 in Appendix 3.8.

Optimizing Sparse Layer. The differentiable parameterization of the above sparse layer can therefore be optimized for maximum attack impact via minimizing the expected distance between the attacked statistic and adversarial target:

$$\min_{\beta,\gamma} \quad H(\beta, \gamma) \quad \triangleq \quad \mathbb{E}_{\boldsymbol{\delta} \sim q(.|\mathbf{x};\beta,\gamma)} \left\| \mathbb{E}_{\mathbf{z} \sim p_\theta(\mathbf{z}|\mathbf{x}+\boldsymbol{\delta})} \Big[\chi(\mathbf{z}) \Big] - \mathbf{t}_{\text{adv}} \right\|_2^2 \quad . \tag{3.3.5}$$

This attack is probabilistic in two ways. First, the magnitude of the perturbation δ is a random variable from distribution $q(\cdot \mid \mathbf{x})$. Second, the non-zero components of the mask depend on the random Gaussian samples, which brings another degree of non-determinism into the design, making the attack less perceptible and harder to detect. See Algorithm 4 in Section 3.6 for the implementation.

Remark. There are three important advantages of the above probabilistic sparse attack. First, by viewing the attack vector as random variable drawn from a learnable distribution instead of fixed parameter to be optimized, we are able to avoid solving the NP-hard problem (3.3.1) as usually approached in previous literature [32]. Second, our approach introduces multiple degree of non-determinism to the attack vector, apparently making it more stealth and powerful (see Section 3.5). Last, unlike the deterministic attack which has two separate, decoupled approximation stages that cannot be optimized end-to-end due to the non-convex and non-differentiable constraint in (3.3.1), the probabilistic attack model is entirely differentiable. Therefore, it can be directly integrated as part of a differentiable defense mechanism that can be optimized via gradient descent in an end-to-end fashion – see Section 3.4.2 for more details.

3.4 Defense Mechanisms against Adversarial Attacks

The adversarial attack on probabilistic forecasting models was investigated under the univariate time series setting [34, 149]. Beyond basic data augmentation [139], we develop more effective defense mechanism to enhance model robustness via randomized smoothing (in Section 3.4.1) and mini-max defense using sparse layer (in Section 3.4.2).

3.4.1 Randomized Smoothing Defense

Randomized smoothing (RS) [29] is a post-training defense technique. Having never been considered to multivariate setting to the best of our knowledge, we apply RS to our multivariate forecasters $z(\mathbf{x}) \sim p_\theta(\mathbf{z} \mid \mathbf{x})$ which maps $\mathbf{x}$ to a random vector $z(\mathbf{x})$ distributed by $p_\theta(\mathbf{z} \mid \mathbf{x})$. Let $\mathbb{P}_z(z(\mathbf{x}) \preceq \mathbf{r})$ denote the CDF of such random outcome vector where $\preceq$ denotes the element-wise inequality, the RS version

$$g_\sigma(\mathbf{x}) \quad = \quad \mathbb{E}_{\boldsymbol{\epsilon}}\Big[z(\mathbf{x}+\boldsymbol{\epsilon})\Big] \tag{3.4.1}$$

of $z(\mathbf{x})$ with noise level $\sigma > 0$ and $\boldsymbol{\epsilon} \sim \mathbb{N}(0, \sigma^2\mathbf{I})$ is a random vector whose CDF is defined as

$$\mathbb{P}_{g_\sigma}\Big(g_\sigma(\mathbf{x}) \preceq \mathbf{r}\Big) \quad \triangleq \quad \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathbb{N}(\mathbf{0},\sigma^2\mathbf{I})}\Big[\mathbb{P}_z\Big(z(\mathbf{x}+\boldsymbol{\epsilon}) \preceq \mathbf{r}\Big)\Big] \tag{3.4.2}$$

where we abuse the notation $\boldsymbol{\epsilon} \sim \mathbb{N}(0, \sigma^2\mathbf{I})$ to indicate the (scalar) entries of the matrix $\boldsymbol{\epsilon}$ are independently and identically distributed by $\mathbb{N}(0, \sigma^2)$. Computing the output of the smoothed forecaster $g_\sigma(\mathbf{x})$ is intractable in general since the integration of $z(\mathbf{x}+\boldsymbol{\epsilon})$ with $\mathbb{N}(0, \sigma^2\mathbf{I})$ cannot be done analytically. However, it can still be approximated with arbitrarily high accuracy via MC sampling with a sufficiently large number of samples. Check Algorithm 2 for a detailed implementation.

Algorithm 2. Randomized Smoothing

```
input:  pre-trained $\tau$-step forecasting model $p_\theta(\mathbf{z} \mid \mathbf{x})$, observation $\mathbf{x}$ and other parameters
    • number of samples $n$
    • noise level $\sigma$
output:  $n$ sample paths $[\hat{\mathbf{x}}_{T+1:T+\tau}]_i^{(j)}$ for $i = 1, \ldots, d$ and $j = 1, \ldots, n$
for $j = 1, 2, \ldots, n$ do
    1. Sample $\xi_{i,t} \sim \mathbb{N}(0, \sigma^2)$ i.i.d. and compute $\tilde{x}_{i,t} \leftarrow x_{i,t} + \xi_{i,t}$
    2. Sample $[\hat{\mathbf{x}}_{T+1:T+\tau}]_i^{(j)} \sim p_\theta(\mathbf{z} \mid \tilde{\mathbf{x}})$
end for
```

Algorithm 3. Minimax Defense

```
input:  dataset $\mathcal{D}$ of $(\mathbf{x}, \mathbf{z})$ pairs and other parameters:
    • sparse constraint $\kappa$ for $q$ in Eq. (3.4.6)
    • number of optimization iterations $n$
output:  robust forecasting model $p_\theta(\mathbf{z} \mid \mathbf{x})$.
for $1, 2, \ldots, n$ do
    3. Fix $\theta$, minimize $-\sum_{(\mathbf{x},\mathbf{z})\sim\mathcal{D}} \ell_g(\phi; \mathbf{x}, \mathbf{z}, \theta)$ with respect to $\phi$ – see Eq. (3.4.5)
    4. Fix $\phi$, maximize $\sum_{(\mathbf{x},\mathbf{z})\sim\mathcal{D}} \ell_p(\theta; \mathbf{x}, \mathbf{z}, \phi)$ with respect to $\theta$ – see Eq. (3.4.6)
end for
```

For the randomized smoothing version g_σ of the base forecaster $z(\mathbf{x}) \sim p_\theta(\mathbf{z}|\mathbf{x})$, we establish a robustness guarantee or certificate in the following theorem.

Theorem 3 (Robust Certificate). Given an input $\mathbf{x}$, let $g_\sigma(\mathbf{x})$ be defined in Eq. (3.4.1). Let $G(\mathbf{r}) = \mathbb{P}_{g_\sigma}(g_\sigma(\mathbf{x}) \preceq \mathbf{r})$ and $G_{\boldsymbol{\delta}}(\mathbf{r}) = \mathbb{P}_{g_\sigma}(g_\sigma(\mathbf{x}+\boldsymbol{\delta}) \preceq \mathbf{r})$. For any $\boldsymbol{\delta}$, we have

$$\sup_{\mathbf{r}\in\mathbb{R}^d} \Big| G(\mathbf{r}) - G_{\boldsymbol{\delta}}(\mathbf{r}) \Big| \;\leq\; \frac{\sqrt{d}}{\sigma} \cdot \Big\| \boldsymbol{\delta} \Big\|_{\mathrm{F}} \,. \tag{3.4.3}$$

This shows that the difference between the CDFs of the smoothed forecaster on authentic and perturbed input, i.e. $g_\sigma(\mathbf{x})$ and $g_\sigma(\mathbf{x}+\boldsymbol{\delta})$, is guaranteed to be no more than $O(\|\boldsymbol{\delta}\|_{\mathrm{F}})$. We defer the formal proof to Appendix 3.8.

Remark. Different from Theorem 1 in [149] that only applies to univariate cases, our Theorem 3 provides a more general robustness guarantee as it's available for multivariate setting. Also, Theorem 1 in [149] only holds for $\boldsymbol{\delta} \to 0$, but our Theorem 3 holds for any $\boldsymbol{\delta}$.

3.4.2 Mini-max Defense

As discussed in Section 3.3.3, the sparse layer is differentiable, which is suitable to be directly integrated as part of a differentiable defense mechanism that can be optimized via gradient descent in an end-to-end fashion. To fix the idea, with a sparse layer $q(\cdot \mid \mathbf{x}; \phi)$ having parameters $\phi = (\beta, \gamma)$ in Eq. (3.3.4), we propose to train the forecaster by minimizing the worst-case loss caused by $q(\cdot \mid \mathbf{x}; \phi)$:

$$\min_{\phi} \max_{\theta} \sum_{(\mathbf{x},\mathbf{z})\sim\mathcal{D}} \Big[\ell_p(\theta; \mathbf{x}, \mathbf{z}, \phi) - \ell_g(\phi; \mathbf{x}, \mathbf{z}, \theta) \Big]. \tag{3.4.4}$$

Here $\ell_g(\phi; \mathbf{x}, \mathbf{z}, \theta)$ is a function of ϕ conditioned on $(\mathbf{x}, \mathbf{z}, \theta)$ while $\ell_p(\theta; \mathbf{x}, \mathbf{z}, \phi)$ is a function of θ conditioned on $(\mathbf{x}, \mathbf{z}, \phi)$ as follows

$$\ell_g(\phi; \mathbf{x}, \mathbf{z}, \theta) \triangleq \mathbb{E}_{q(\boldsymbol{\delta}|\mathbf{x};\phi)}\left[\mathbb{E}_{p_\theta(\mathbf{z}'|\mathbf{x}+\boldsymbol{\delta})}\left\|\mathbf{z}' - \mathbf{z}\right\|^2\right] \tag{3.4.5}$$

$$\ell_p(\theta; \mathbf{x}, \mathbf{z}, \phi) \triangleq \mathbb{E}_{q(\boldsymbol{\delta}|\mathbf{x};\phi)}\left[\log p_\theta\Big(\mathbf{z} \mid \mathbf{x} + \boldsymbol{\delta}\Big)\right] \tag{3.4.6}$$

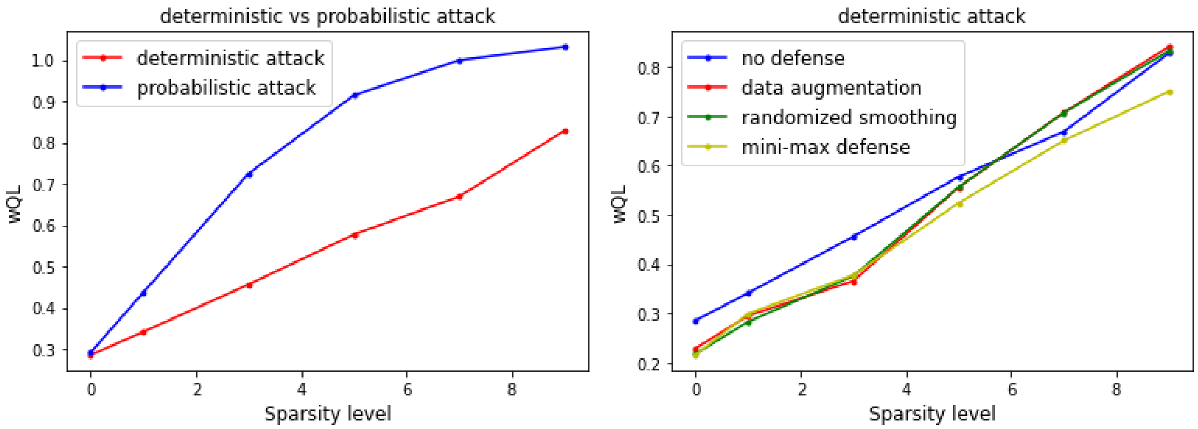
where the expectation is taken over $\boldsymbol{\delta} \sim q(\boldsymbol{\delta}|x;\phi)$ with q given by Eq. (4.3.7), each pair $(\mathbf{x}, \mathbf{z})$ represents a training data point in our dataset with $\mathbf{x} = \{\mathbf{x}_t\}_{t=1}^T$ and $\mathbf{z} = \{\mathbf{x}_{T+t}\}_{t=1}^{\tau}$.

Solving Eq. (3.4.4) therefore means finding a stable state where the model parameter is conditioned to perform best in the worst situation where the adversarial noises are also conditioned to generate the most impact even in the most benign scenario. This can be achieved by alternating between (1) minimizing $-\ell_g$ in Eq. (3.4.5) with respect to (β, γ) and (2) maximizing ℓ_p in Eq. (3.4.6) with respect to θ. We call this defense mechanism a mini-max defense. We note that similar ideas have been previously exploited in deep generative models, such as GAN [46] and WGAN [5]. See Algorithm 3 for a detailed description.

Remark. Unlike the sparse layer used in attack, the sparse layer used to simulate mock attack in our defense strategy does not have access to the actual attack sparsity parameter κ or the set of target time series $\mathcal{I}$. Hence, we need to set the sparsity κ as a tuning parameter and skip the last step of the sparse layer described in Section 3.3.3 where we set $\boldsymbol{\delta}^{\mathcal{I}} = \mathbf{0}$.

3.5 Experiments

We conduct numerical experiments to demonstrate the effectiveness of our proposed indirect sparse attack on a multivariate probabilistic forecasting models and compare various defense mechanisms.



(a) **(b)**

Figure 3.2. Plots of (a) averaged wQL under sparse indirect attack against the sparsity level on electricity dataset. The underlying model is a clean DeepVAR without defense. Target time series $\mathcal{I} = \{1\}$ and attacked time stamp $H = \{\tau\}$; and (b) & (c) averaged wQL under different defense mechanisms on electricity dataset for deterministic & probabilistic attack respectively.

3.5.1 Experiment Setups

Dataset. We include Electricity [6], Traffic [6], Taxi [128], Wiki [74]. See Section 3.7.1 for more information.

Multivariate Forecaster. We consider DeepVAR [115] which is state-of-the-art multivariate probabilistic models with implementation available pytorch-ts [111] with target dimension 10. For more details on the model parameters, see Section 3.7.2.

Data Augmentation (DA) and Randomized Smoothing (RS). Following the convention in [34, 149], we use relative noises in both data augmentation and randomized smoothing. That is, given a sequence of observation $\mathbf{x} = ([\mathbf{x}_t]_i)_{i,t} \in \mathbb{R}^{d \times T}$, we draw i.i.d. noise samples $[\boldsymbol{\epsilon}_t]_i \sim \mathbb{N}(0, \sigma^2)$ and produce noisy input as $[\tilde{\mathbf{x}}_t]_i \leftarrow [\mathbf{x}_t]_i(1 + [\mathbf{x}_t]_i)$. In data augmentation, we train model with noisy input $[\tilde{\mathbf{x}}_t]_i$. In RS, the base model is still trained on noisy input $[\tilde{\mathbf{x}}_t]_i$ with noise level σ. The noise level σ remains the same across DA and RS.

Metrics. We adopt weighted quantile loss (wQL) to measure the performance.

(See Section 3.7.3.)

3.5.2 Experiment Results

Electricity, Traffic, and More Datasets. Averaged wQL loss is reported in Table 3.1 and Table 3.2 for Electricity and Traffic dataset respectively. The attacks include both deterministic and probabilistic ones for both single and multiple target time series and time horizons. Besides, we plot wQL under both attacks against sparsity level to better visualize the effect of different types of attack. See Figure 3.2. More experiment results with error bars on additional datasets can be found in Section 3.7.4.

Message 1: Sparse, Indirect Attack is effective, and becomes more effective as κ increases. In the experiment, we can verify the effectiveness of sparse indirect attack, that is, one can attack the prediction of one time series without directly attacking the history of this time series. For example in Table 3.1, under deterministic attack, the average wQL is increased by 20% by only attacking one out of nine remaining time series (there are totally 10 but the target time series is excluded).

Moreover, attacking half of the time series can increase average wQL by 102%! This observation is even more noticeable under probabilistic attack: average wQL can be increased by 215% with 50% of the time series attacked. Besides, wQL loss increases as attack sparsity κ increases, which is also an evidence that sparse indirect attack is effective.

Message 2: Probabilistic Attack is more effective than Deterministic Attack, especially at low sparsity levels. In general, average wQL increases as sparsity level increases and probabilistic attack appears to be more effective than deterministic one, see Figure 3.2a and Table 3.1. For example, under no defense when $\kappa = 7$, probabilistic attack causes 50% larger wQL loss than deterministic one.

Message 3: Randomized Smoothing (RS) and Mini-Max are more robust than Data Augmentation (DA). As can be seen in Figure 3.2b, Table 3.1 and Table 3.2, all three defense methods can bring robustness to the forecasting model. Data augmentation and randomized smoothing works well under small sparsity and mini-max defense achieves comparable performance as data augmentation and randomized smoothing under small sparsity and outperforms them under large sparsity.

Table 3.1. Average wQL on **Electricity** dataset under **deterministic** and **probabilistic** attack. Target time series $\mathcal{I} = \{1\}$ and attacked time stamp $H = \{\tau\}$. Smaller is better.

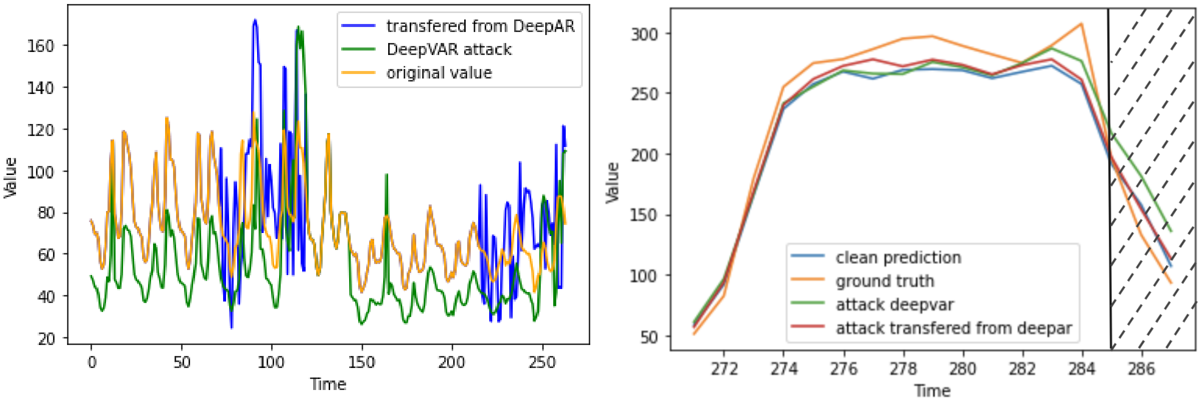
	deterministic attack				probabilistic attack			
sparsity (κ)	no defense	DA	RS	mini-max	no defense	DA	RS	mini-max
no attack	0.2853	0.2288	0.2176	**0.2154**	0.2909	0.2374	**0.2237**	0.2342
1	0.3410	0.2949	**0.2826**	0.2990	**0.4364**	0.5923	0.5940	0.4935
3	0.4559	**0.3655**	0.3757	0.3775	0.7245	0.5738	**0.4581**	0.8079
5	0.5770	0.5554	0.5560	**0.5273**	0.9143	0.8422	0.9276	**0.5265**
7	0.6687	0.7076	0.7072	**0.6506**	0.9991	0.8267	1.0100	**0.6161**
9	0.8282	0.8412	0.8327	**0.7503**	1.0317	0.8139	0.8919	**0.6466**

Table 3.2. Average wQL on **Traffic** dataset under **deterministic** and **probabilistic** attack. Target time series $\mathcal{I} = \{1, 5\}$ and attacked time stamp $H = \{\tau - 1, \tau\}$. Smaller is better.

	deterministic attack				probabilistic attack			
sparsity (κ)	no defense	DA	RS	mini-max	no defense	DA	RS	mini-max
no attack	0.2283	0.1573	**0.1529**	0.1837	0.2283	0.1573	**0.1529**	0.1837
1	0.2190	0.1543	**0.1529**	0.1701	0.2428	0.1807	**0.1796**	0.1904
3	0.2150	0.1884	0.1890	**0.1687**	0.2219	0.2564	0.2467	**0.1714**
5	0.2772	0.2729	0.2648	**0.1688**	0.2719	0.3026	0.3003	**0.1883**
7	0.3620	0.3597	0.3535	**0.1779**	0.3529	0.2893	0.2824	**0.1846**
9	0.4635	0.4058	0.4240	**0.1970**	0.4075	0.3544	0.3376	**0.1911**

3.5.3 Non-transferrablity between Univariate and Multivariate Cases

From the above Section 3.5.2, we verify the effectiveness of sparse indirect attack of multivariate forecasting models. In this subsection, we investigate the transferrability from univariate attack to multivariate attack. To be specific, we study the question whether the



(a) Value of TS 5

(b) Value of TS 1

Figure 3.3. Plots of (a) authentic (orange), DeepAR-attacked (blue) and DeepVAR-attacked (green) versions of time-series (TS) 5; and (b) ground-truth (orange), no-attack (blue), under-DeepAR-attack (red) and under-DeepVAR-attack (green) predictions for TS 1. Shaded area is attacker's target range. Compared to clean prediction, the value of TS 1 at the attack time step ($t = 288$) were adversely altered by DeepVAR-attack (green) but only slightly altered by DeepAR-attack (red). The wQL loss under no attack: 0.288, under DeepAR attack: 0.322, under DeepVAR attack: 0.390.

adversarial perturbation generated by univariate models can be transferred to multivariate models as an indirect attack.

In empirical experiments, we choose sparsity level $\kappa = 1$ and other parameters are the same as what are described in Section 3.5.1. It turns out TS 5 is selected by Algorithm 1 to harm the prediction of TS 1 when $\kappa = 1$. Thus, we use the technique in [34, 149] to generate univariate attack on TS 5 from DeepAR. Note that only the history of TS 5 has been adversely altered. The attacked time series 5 is further fed into DeepVAR model.

Experiment Result. The averaged wQL loss is reported in Table 3.11 in Section 3.9. For a better visualization, the history of TS 5 and prediction of TS 1 are plotted in Figure 3.3a and Figure 3.3b respectively. From the experiment results in Table 3.11, we observe that multivariate attack is 3x more effective than univariate attack, which is also a reason why multivariate attack worth investigation.

3.6 Probabilistic Attack Algorithm

Algorithm 4. Probabilistic Adversarial Attack

```
input: pre-trained model $p_\theta(\mathbf{z} \mid \mathbf{x})$, observation $\mathbf{x}$ and other parameters:
    • statistic $\chi(\cdot)$, adversarial target $\mathbf{t}_{\mathrm{adv}}$, target set $\mathcal{I} \subset [d]$
    • attack energy $\eta$, sparse constraint $\kappa$, number of iterations $n$
output: perturbation matrix $\boldsymbol{\delta} \in \mathbb{R}^{T \times d}$ s.t. $\|\boldsymbol{\delta}\|_{\max} \leq \eta$, $\mathbb{E}[s(\boldsymbol{\delta})] \leq \kappa$, $\boldsymbol{\delta}^{\mathcal{I}} = \mathbf{0}$
1. randomly initialize a sparse layer $q(\cdot|\mathbf{x}; \beta, \gamma)$
for iteration $1, 2, \ldots, n$ do
    2. compute the expected loss $H(\beta, \gamma)$ using Eq. (3.3.5)
    3. update $\beta, \gamma$ via first-order optimization method
end for
4. draw $\boldsymbol{\delta} \sim q(\cdot|\mathbf{x}; \beta, \gamma)$ and return
```

3.7 Details on the experiment setting

3.7.1 Datasets

- Electricity: consists of hourly electricity consumption time series from 370 customers.
- Taxi: traffic time series of New York taxi rides taken at 1214 locations for every 30 minutes from January 2015 to January 2016 and considered to be heterogeneous. We use the taxi-30min dataset provided by GluonTS.
- Traffic: hourly occupancy rate, between 0 and 1, of 963 San Francisco car lanes.
- Wiki: daily page views of 2000 Wikipedia pages used in [42].

3.7.2 Hyper-parameter choice

Electricity & Taxi.

We target at the prediction of the first time series at the last prediction time step, i.e. target time series $\mathcal{I} = \{1\}$ and time horizon to attack $H = \{\tau\}$, so $\chi(\mathbf{z}) = x_{1,T+\tau}$. We choose prediction length $\tau = 24$ and context length $T = 4\tau = 96$, and sparsity level $\kappa = 1, 3, 5, 7, 9$.

Table 3.3. Summary of statistics of the datasets used, including prediction length τ, domain, frequency, dimension, and time steps.

dataset	prediction length τ	domain	frequency	dimension	time steps
Electricity	24	$\mathbb{R}^+$	H	370	5790
Traffic	24	$\mathbb{R}^+$	H	963	10413
Taxi	24	$\mathbb{N}$	30-min	1214	1488
Wiki	30	$\mathbb{N}$	D	2000	792

Traffic.

We target at the prediction of $\chi(\mathbf{z}) = (x_{1,T+\tau-1}, x_{1,T+\tau}, x_{5,T+\tau-1}, x_{5,T+\tau})$. We choose prediction length $\tau = 24$ and context length $T = 4\tau = 96$, and sparsity level $\kappa = 1, 3, 5, 7, 9$.

Wiki.

We target at the prediction of $\chi(\mathbf{z}) = x_{1,T+\tau}$. We choose prediction length $\tau = 30$ and context length $T = 4\tau = 120$, and sparsity level $k = 1, 3, 5, 7, 9$.

For all experiments, we train a DeepVAR with rank 5. The attack energy $\eta = c_1 \max |\mathbf{x}|$, is proportional to the largest element of the past observation in magnitude, where c_1 is set to 0.5. For the adversarial target $\mathbf{t}_{\text{adv}}$, we first draw a prediction $\hat{\mathbf{x}}$ from un-attacked model $p_\theta(\cdot|\mathbf{x})$ and choose $\mathbf{t}_{\text{adv}} = c_2\hat{\mathbf{x}}$ for constants $c_2 = 0.5$ and 2.0. We report the largest error produced by these choices of constants. Unless otherwise stated, the number of sample paths drawn from the prediction distribution $n = 100$ to quantify quantiles $q_{i,t}^{(\alpha)}$. In mini-max defense, the sparsity level of the sparse layer is set to 5 for all cases. For the noise level σ in DA and RS, we select them via a validation set and it turns out no σ is uniformly better than the others across different sparsity level. Thus, $\sigma = 0.1$ is chosen in the empirical evaluation. For an ablation study on the effect of σ, see Table 3.6 in Section 3.7.4.

3.7.3 Metrics

We measure the performance of model under attacks by the popular metric especially for probabilistic forecasting models: weighted quantile loss (wQL), which is defined as

$$\mathrm{wQL}(\alpha) = 2 \cdot \frac{\sum_{i,t}[\alpha \max(x_{i,t} - q_{i,t}^{(\alpha)}, 0) + (1-\alpha)\max(q_{i,t}^{(\alpha)} - x_{i,t}, 0)]}{\sum_{i,t}|x_{i,t}|}$$

where $\alpha \in (0,1)$ is a quantile level. In practical application, under-prediction and over-prediction may cost differently, suggesting wQL should be one's main consideration especially for probabilistic forecasting models. In the subsequent sections, we calculate average wQL over a range of $\alpha = [0.1, 0.2, \ldots, 0.9]$ and evaluate the performance in terms of averaged wQL.

3.7.4 More results

To measure the performance of a forecasting model, other metrics like Weighted Absolute Percentage Error (WAPE) or Weighted Squared Error (WSE) are also considered by a large body of literature. For completeness, we present the definition of WAPE and WSE:

$$\begin{aligned}
\mathrm{WAPE} &= \sum \left|\frac{\text{predicted value}}{\text{true value}} - 1\right| = \frac{1}{|I||H|}\sum_{i\in I, h\in H}\left|\frac{\frac{1}{n}\sum_{j=1}^{n}\hat{x}_{T+h,i}^{j}}{x_{T+h,i}} - 1\right|,\\
\mathrm{WSE} &= \sum \left(\frac{\text{predicted value}}{\text{true value}} - 1\right)^2 = \frac{1}{|I||H|}\sum_{i\in I, h\in H}\left(\frac{\frac{1}{n}\sum_{j=1}^{n}\hat{x}_{T+h,i}^{j}}{x_{T+h,i}} - 1\right)^2,
\end{aligned}$$

where $\hat{x}_{i,j}$ is the predicted values from forecasting model. We report WAPE, WSE and wQL under deterministic and probabilistic attacks on electricity dataset in Table 3.4 and Table 3.5.

Also, Table 3.6 reports the effect of choosing different values for σ in data augmentation and randomized smoothing.

Table 3.4. Metrics on **Electricity** dataset under **deterministic** attack. Target time series $\mathcal{I} = \{1\}$ and attacked time stamp $H = \{\tau\}$. Smaller is better.

Sparsity (κ)	no defense			data augmentation		
	WAPE	WSE	wQL	WAPE	WSE	wQL
0	0.4005±0.2036	0.2360±0.2525	0.2991±0.1684	0.4241±0.2092	0.2596±0.2625	0.3280±0.1497
1	0.4900±0.2488	0.3529±0.3769	0.3745±0.2106	0.4123±0.1829	0.2310±0.1934	0.3019±0.1138
3	0.6382±0.3434	0.6222±0.5886	0.5043±0.2917	0.5654±0.2475	0.4313±0.3707	0.3919±0.1876
5	0.7524±0.3675	0.8123±0.6218	0.6097±0.3218	0.7460±0.3803	0.8201±0.6628	0.5379±0.2833
7	0.8786±0.4171	1.0889±0.7785	0.7432±0.3702	0.8465±0.4014	1.0102±0.6389	0.6425±0.2985
9	1.0134±0.4541	1.4028±0.9685	0.8851±0.4023	0.9093±0.4454	1.1883±0.7720	0.7007±0.3395

Sparsity (κ)	randomized smoothing			mini-max defense		
	WAPE	WSE	wQL	WAPE	WSE	wQL
0	0.3501±0.1630	0.1710±0.1486	0.2751±0.1068	0.3237±0.1379	0.1394±0.0913	0.2342±0.0917
1	0.4209±0.1700	0.2298±0.1683	0.2965±0.1003	0.4498±0.2253	0.2949±0.2276	0.3511±0.1825
3	0.5887±0.2543	0.4644±0.3784	0.3965±0.1797	0.7447±0.3758	0.8120±0.6684	0.6038±0.3358
5	0.7504±0.3607	0.8002±0.5999	0.5619±0.2779	0.9603±0.4190	1.2419±0.8369	0.8182±0.3845
7	0.8353±0.4315	1.0369±0.7496	0.6311±0.3152	1.1056±0.4847	1.6504±1.0591	0.9689±0.4350
9	0.9986±0.5026	1.4574±0.9998	0.7700±0.3717	1.2476±0.4860	1.9870±1.0815	1.1133±0.4306

Next, we report wQL loss of Taxi and Wiki dataset under both types of attacks in Table 3.7, Table 3.8, Table 3.9 and Table 3.10 respectively.

3.8 Detailed proofs

Proof of Lemma 1. We can compute

$$\mathbb{P}(\boldsymbol{\delta}^i \;=\; \mathbf{0}) = 1 \;-\; \mathbb{P}\Big(u_i \;\le\; \Phi^{-1}\Big(r_i(\gamma)\Big)\Big) = 1 - r_i(\gamma). \tag{3.8.1}$$

That is, with probability $1 - r_i(\gamma)$, $\boldsymbol{\delta}^i = 0$. Equivalently, $\boldsymbol{\delta}^i$ is distributed by a degenerated probability measure with Dirac density $D(\boldsymbol{\delta}^i)$ concentrated at 0. On the other hand, with probability $r_i(\gamma)$, $\boldsymbol{\delta}^i$ is distributed as $q_i'(\cdot|\mathbf{x};\beta)$. Combining the two cases, it follows that $\boldsymbol{\delta}^i$ is distributed by a mixture of $q_i'(\cdot|\mathbf{x};\beta)$ and $D(\boldsymbol{\delta}^i)$ with weights $r_i(\gamma)$ and $1 - r_i(\gamma)$ respectively. □

Table 3.5. Metrics on **electricity** dataset under **probabilistic** attack using sparse layer. Target time series $\mathcal{I} = \{1\}$ and attacked time stamp $H = \{\tau\}$. Smaller is better.

Sparsity (κ)	no defense			data augmentation		
	WAPE	WSE	wQL	WAPE	WSE	wQL
0	0.3842±0.2620	0.2162±0.3044	0.2909±0.0748	0.3074±0.1746	0.1250±0.0946	0.2374±0.0764
1	0.6230±0.6324	0.7881±1.1864	0.4364±0.1296	0.7476±0.7240	1.0830±1.8593	0.5923±0.0913
3	1.0540±0.7522	1.6768±1.4810	0.7245±0.2434	0.8484±0.6809	1.1834±1.3998	0.5738±0.1759
5	1.2078±0.7451	2.0139±2.0667	0.9143±0.3235	1.1444±0.6665	1.7538±1.4318	0.8422±0.2945
7	1.3236±0.7310	2.2863±1.8336	0.9991±0.3505	1.1304±0.6522	1.7031±1.4053	0.8267±0.2823
9	1.3656±0.8671	2.6166±2.6679	1.0317±0.3707	1.0912±0.6181	1.5727±1.2081	0.8139±0.2827
Sparsity (κ)	randomized smoothing			mini-max defense		
	WAPE	WSE	wQL	WAPE	WSE	wQL
0	0.2858±0.1547	0.1056±0.0761	0.2237±0.0750	0.3218±0.1429	0.1240±0.0830	0.2342±0.0710
1	0.7683±0.8771	1.3596±2.7290	0.5940±0.1142	0.6990±0.6957	0.9726±1.7182	0.4935±0.1450
3	0.6784±0.5230	0.7337±0.7698	0.4581±0.1301	0.9909±0.7564	1.5540±1.8925	0.8079±0.2838
5	1.2310±0.7025	2.0090±1.6609	0.9276±0.3208	0.6966±0.4554	0.6927±0.8752	0.5265±0.1611
7	1.3496±0.6777	2.2809±1.7240	1.0100±0.3554	0.8424±0.7803	1.3186±1.7286	0.6161±0.1986
9	1.1978±0.6742	1.8894±1.5309	0.8919±0.3072	0.8691±0.7410	1.3043±2.0663	0.6466±0.2054

Proof of Lemma 2. By the construction of $r_i(\gamma)$,

$$\begin{aligned}\mathbb{E}\Big[s(\boldsymbol{\delta})\Big] &= \sum_{i=1}^{d} \mathbb{E}\left[\mathbb{I}\left(u_i \leq \Phi^{-1}\left(r_i(\gamma)\right)\right)\right] = \sum_{i=1}^{d} \mathbb{P}\left(u_i \leq \Phi^{-1}\left(r_i(\gamma)\right)\right) \\ &= \sum_{i=1}^{d} r_i(\gamma) = \frac{\kappa}{\sqrt{d}} \cdot \frac{\sum_{i=1}^{d} \gamma_i^{1/2}}{\left(\sum_{i=1}^{d} \gamma_i\right)^{1/2}} \leq \kappa.\end{aligned}$$

□

Proof of Theorem 3. Denote $p_\sigma(\cdot)$ as the density of $\mathbb{N}(0, \sigma^2\mathbf{I}_d)$ and $p(\cdot)$ as the density of $\mathbb{N}(0, \mathbf{I}_d)$. Let $F_{\mathbf{x}}(\mathbf{r}) \triangleq \mathbb{P}(z(\mathbf{x}) \preceq \mathbf{r})$.

Table 3.6. Average wQL on **Electricity** dataset under **deterministic** attack. The defense is data augmentation and randomized smoothing with varying $\sigma = 0.1, 0.2, 0.3$. Target time series $\mathcal{I} = \{1\}$ and attacked time stamp $H = \{\tau\}$. Smaller is better.

		$\sigma = 0.1$		$\sigma = 0.2$		$\sigma = 0.3$		
Sparsity (κ)	no defense	DA	RS	DA	RS	DA	RS	mini-max
no attack	0.2853	0.2288	0.2176	0.2321	0.2389	0.2999	0.3053	**0.2154**
1	0.3410	0.2949	0.2826	**0.2717**	0.2866	0.2959	0.3456	0.2990
3	0.4559	**0.3655**	0.3757	0.4822	0.4421	0.4323	0.3930	0.3775
5	0.5770	0.5554	0.5560	0.6130	0.5790	0.5998	0.5351	**0.5273**
7	0.6687	0.7076	0.7072	0.6796	0.6677	0.6743	**0.6447**	0.6506
9	0.8282	0.8412	0.8327	0.8243	0.8222	0.7953	**0.7335**	0.7503

Table 3.7. Average wQL on **Taxi** dataset under **deterministic** attack. Target time series $\mathcal{I} = \{1\}$ and attacked time stamp $H = \{\tau\}$. Smaller is better.

Sparsity (κ)	no defense	data augmentation	randomized smoothing	mini-max defense
no attack	1.2135±0.4050	1.2137±0.4091	1.2574±0.4281	**1.0447**±0.3607
1	1.3152±0.4580	1.3455±0.4666	1.3455±0.4627	**1.1222**±0.3960
3	1.6389±0.5810	1.6805±0.5982	1.6503±0.5756	**1.3624**±0.4956
5	2.0317±0.7161	2.0625±0.7290	2.0123±0.7059	**1.6830**±0.6206
7	2.3695±0.8064	2.3712±0.8028	2.3450±0.7978	**1.9750**±0.7033
9	2.5605±0.8531	2.5525±0.8616	2.5422±0.8619	**2.2374**±0.7785

Table 3.8. Average wQL on **Taxi** dataset under **probabilistic** attack. Target time series $\mathcal{I} = \{1\}$ and attacked time stamp $H = \{\tau\}$. Smaller is better.

Sparsity (κ)	no defense	data augmentation	randomized smoothing	mini-max defense
no attack	1.2118±0.4412	1.2526±0.4733	1.2241±0.4531	**1.0481**±0.3840
1	1.4598±0.5315	1.3539±0.5199	1.3512±0.5100	**1.1528**±0.4345
3	1.5659±0.6589	1.5446±0.6197	1.5567±0.5784	**1.3940**±0.5472
5	1.9123±0.7513	1.7824±0.6962	1.8857±0.7441	**1.6897**±0.6829
7	2.2915±0.8954	1.7340±0.7638	1.8370±0.7597	**1.5865**±0.6191
9	2.4815±0.9286	2.1159±0.7515	2.2400±0.7860	**1.4921**±0.5551

Table 3.9. Average wQL on **Wiki** dataset under **deterministic** attack. Target time series $\mathcal{I} = \{1\}$ and attacked time stamp $H = \{\tau\}$. Smaller is better.

Sparsity (κ)	no defense	data augmentation	randomized smoothing	mini-max defense
no attack	0.1645±0.0588	0.0868±0.0232	**0.0796**±0.0272	0.2331±0.1186
1	0.2430±0.0889	0.0775±0.0171	**0.0687**±0.0119	0.1683±0.1097
3	0.2771±0.0807	0.2225±0.1217	0.2260±0.1089	**0.1466**±0.0976
5	0.4260±0.1127	0.3533±0.1602	0.3084±0.1365	**0.1675**±0.0675
7	0.5173±0.1045	0.4290±0.1524	0.4112±0.1420	**0.1973**±0.0632
9	0.6276±0.1178	0.4362±0.1360	0.4451±0.1461	**0.2185**±0.1131

Table 3.10. Average wQL on **Wiki** dataset under **probabilistic** attack. Target time series $\mathcal{I} = \{1\}$ and attacked time stamp $H = \{\tau\}$. Smaller is better.

Sparsity (κ)	no defense	data augmentation	randomized smoothing	mini-max defense
no attack	0.1748±0.1144	0.0837±0.0432	**0.0828**±0.0443	0.2376±0.1510
1	0.3255±0.2132	**0.1647**±0.1126	0.1976±0.1274	0.1834±0.1409
3	0.4080±0.1724	0.3104±0.1550	**0.2255**±0.1322	0.2549±0.1530
5	0.5336±0.2318	0.2759±0.1368	0.1714±0.1348	**0.1299**±0.0852
7	0.6547±0.2940	0.3940±0.1849	0.2656±0.1708	**0.2569**±0.1605
9	0.8463±0.2715	0.5140±0.2195	**0.2745**±0.1513	0.2909±0.1918

Consider

$$
\begin{aligned}
\sup_{\mathbf{r}\in\mathbb{R}^d}\Big|G(\mathbf{r})-G_{\boldsymbol{\delta}}(\mathbf{r})\Big| &= \sup_{\mathbf{r}\in\mathbb{R}^d}\left|\int_{\boldsymbol{\epsilon}\in\mathbb{R}^{d\times T}}\Big(F_{\mathbf{x}+\boldsymbol{\epsilon}}(\mathbf{r})-F_{\mathbf{x}+\boldsymbol{\delta}+\boldsymbol{\epsilon}}(\mathbf{r})\Big)p_\sigma(\boldsymbol{\epsilon})\,\mathrm{d}\boldsymbol{\epsilon}\right| \\
&= \sup_{\mathbf{r}\in\mathbb{R}^d}\left|\int_{\boldsymbol{\epsilon}\in\mathbb{R}^{d\times T}}F_{\boldsymbol{\epsilon}}(\mathbf{r})\Big(p_\sigma(\boldsymbol{\epsilon}-\mathbf{x})-p_\sigma(\boldsymbol{\epsilon}-\mathbf{x}-\boldsymbol{\delta})\Big)\,\mathrm{d}\boldsymbol{\epsilon}\right| \\
&= \sup_{\mathbf{r}\in\mathbb{R}^d}\left|\int_{\boldsymbol{\epsilon}\in\mathbb{R}^{d\times T}}\int_0^1 F_{\boldsymbol{\epsilon}}(\mathbf{r})\nabla p_\sigma(\boldsymbol{\epsilon}-\mathbf{x}-t\boldsymbol{\delta})\boldsymbol{\delta}\,\mathrm{d}t\,\mathrm{d}\boldsymbol{\epsilon}\right| \\
&= \sup_{\mathbf{r}\in\mathbb{R}^d}\left|\int_0^1\int_{\boldsymbol{\epsilon}\in\mathbb{R}^{d\times T}}F_{\boldsymbol{\epsilon}}(\mathbf{r})\,\left(\boldsymbol{\delta}\cdot\frac{\boldsymbol{\epsilon}-\mathbf{x}-t\boldsymbol{\delta}}{\sigma^2}\right)p_\sigma(\boldsymbol{\epsilon}-\mathbf{x}-t\boldsymbol{\delta})\,\mathrm{d}\boldsymbol{\epsilon}\,\mathrm{d}t\right| \\
&= \frac{1}{\sigma}\sup_{\mathbf{r}\in\mathbb{R}^d}\left|\int_0^1\int_{\boldsymbol{\epsilon}\in\mathbb{R}^{d\times T}}F_{\mathbf{x}+t\boldsymbol{\delta}+\boldsymbol{\epsilon}}(\mathbf{r})\,(\boldsymbol{\delta}\cdot\boldsymbol{\epsilon})p(\boldsymbol{\epsilon})\,\mathrm{d}\boldsymbol{\epsilon}\,\mathrm{d}t\right| \\
&\le \frac{1}{\sigma}\int_{\boldsymbol{\epsilon}\in\mathbb{R}^{d\times T}}\Big|\boldsymbol{\delta}\cdot\boldsymbol{\epsilon}\Big|p(\boldsymbol{\epsilon})\,\mathrm{d}\boldsymbol{\epsilon} \\
&\le \frac{\|\boldsymbol{\delta}\|_2}{\sigma}\Big(\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathbb{N}(0,I_d)}\|\boldsymbol{\epsilon}\|_2^2\Big)^{\frac{1}{2}} \;=\; \frac{\sqrt{d}}{\sigma}\|\boldsymbol{\delta}\|_2,
\end{aligned}
$$

which completes the proof. □

3.9 Non-transferrability of attacks between univariate and multivariate forecasters

We study the transferrability from univariate attack to multivariate attack. To be specific, if an attack is generated on the same subset (excluding target time series) of time series using a univariate model and then fed into a multivariate model, can it indirectly harm the prediction of target time series. Next, we report the experiment results of univariate attack and multivariate attack.

Table 3.11. Transfer the attack from DeepAR to DeepVAR. Target items $\mathcal{I}=\{1\}$ and time horizon to attack $H=\{\tau\}$. Clean DeepAR and DeepVAR models are used. Averaged wQL is reported.

No attack	Univariate attack	Multivariate attack
0.288	0.322	0.390

3.10 Conclusion

In this work, we investigate the existence of sparse indirect attack for multivariate time series forecasting models. We propose both deterministic approach and a novel probabilistic approach to finding effective adversarial attack. Besides, we adopt the randomized smoothing technique from image classification and univariate time series to our framework and design another mini-max optimization to effectively defend the attack delivered by our attackers. To the best of our knowledge, this is the first work to study sparse indirect attack on multivariate time series and develop corresponding defense mechanisms, which could inspire a future research direction.

Chapter 3, in part, has been submitted for publication of the material ”Robust Multivariate Time-Series Forecasting: Adversarial Attacks and Defense Mechanisms”, Liu, Linbo, Park, Youngsuk, Hoang, Trong Nghia, Hasson, Hilaf, and Huan, Jun to *International Conference on Learning Representations* and is currently under review. The dissertation author was the primary investigator and author of this paper.

Chapter 4

Promoting Robustness of Randomized Smoothing: Two Cost-Effective Approaches

4.1 Introduction

The existence of adversarial examples of deep neural networks (DNNs) [126, 45] has raised serious concerns to deploy DNNs in real-world systems, especially in the safety critical applications such as self-driving cars and aircraft control systems. Thus, many research efforts have been devoted into developing effective defenses methods to safeguard DNNs. One of the most promising direction is known as *certified defense* via *randomized smoothing*, where the word *certified* means that the defense methods have provable theoretical guarantee as opposed to easily broken heuristic defenses [7], and *randomized smoothing* is a popular technique that allows scalable certified defenses for state-of-the-art DNNs against adversarial examples.

Randomized smoothing is recently proposed by [76, 80, 29] and has achieved state-of-the-art robustness guarantees. Given any classifier f, denoted as a *base classifier*, randomized smoothing predicts the most-likely class on the randomly perturbed input x with Gaussian noises. Following this new prediction rule, randomized smoothing acts like an operator on the original *base classifier* and produce a new *smoothed* classifier which is equipped with provable robustness guarantees under various ℓ_p norm threat models [76, 29, 79].

Unfortunately, without specially-designed training techniques, the robustness certificate of the smoothed classifier is usually very weak [29]. Thus there are a few recent works [116, 152] proposed to design specialized robust training methods to improve the robustness certificate of the smoothed classifier. In [116], the authors propose an adversarial training method called SmoothAdv, which is similar to the PGD training [92] but on the *smoothed* classifier. On the other hand, [152] propose MACER, whose training objective involves a term to maximize the robustness certificate directly. However, SmoothAdv often requires heavy tuning on a number of hyper-parameters for different noise level σ which could be computationally challenging, while MACER needs a much larger number of ($3\times$) training epochs to train (and unfortunately the resulting models often have weaker certificate despite higher clean accuracy).

Motivated by the need of cost-effective robust training methods for randomized smoothing, in this work, we propose two approaches to address the limitations of SmoothAdv and MACER. First, we propose a new robust training method called AdvMacer, which takes the best of both worlds in SmoothAdv and MACER: **AdvMacer** enjoys computational efficiency, gives larger ACR while preserving good accuracy in most settings. Besides, **AdvMacer** attains a universal configuration that works well across different setting with different values of the smoothing noise's variance σ. Second, we propose to equip our **AdvMacer** models with a training-free ensemble method **EsbRs**, which can further enlarge the resulting model's certified radius (by up to 8% compared with SmoothAdv and 15% compared with MACER), hence establishing the new state-of-the-art result on certified radius. Crucially, we present a general theoretical analysis and demonstrate the effect of both *intra*-model ensembles and *mixed*-model ensembles from the theoretical point of view. Grounded by our theoretical findings, an optimal weighted ensemble can be derived analytically where the weights are dependent on the input data.

4.2 Related works and backgrounds

In this section, we first give backgrounds on randomized smoothing and the related certified defense SmoothAdv [116] and MACER [152]. Next, we review recent liteature on applying ensemble methods to randomized smoothing.

Randomized smoothing Consider a neural network classifier $f : \mathbb{R}^d \to \mathcal{Y}$ that maps an input sample $x \in \mathbb{R}^d$ to its predicted label in $\mathcal{Y}$. [29] introduced a randomized smoothing (RS) technique that can turn any base classifier $f(x)$ into a smoothed classifier $g(x)$ with provable robustness guarantees. When taking a sample x, the smoothed classifier g returns the class that the base classifier f is most likely to return under isotropic Gaussian noise perturbation of x:

$$g(x) = \arg\max_{c \in \mathcal{Y}} \mathbb{P}_{\epsilon \sim \mathcal{N}(0,\sigma^2 I)}(f(x+\epsilon) = c),$$

where σ is the noise level that controls the trade-off between clean accuracy and model robustness.

[29] further proved the robustness guarantees of such smoothed classifier in Theorem 4. Let Φ denote the cumulative density function (CDF) of the standard Gaussian distribution. Suppose that under Gaussian perturbation $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, the most likely class c_A is returned with probability p_A and the second most likely (runner-up) class c_B is returned with probability p_B, i.e.

$$\begin{aligned}
&c_A = \arg\max_{c \in \mathcal{Y}} \mathbb{P}(f(x+\epsilon) = c), \\
&c_B = \arg\max_{c \neq c_A} \mathbb{P}(f(x+\epsilon) = c), \\
&p_A = \mathbb{P}(f(x+\epsilon) = c_A),\ p_B = \mathbb{P}(f(x+\epsilon) = c_B).
\end{aligned}$$

Theorem 4 (Theorem 1 of [29]). Assume p_A attains a lower bound $\underline{p_A}$ and p_B attains an upper bound $\bar{p}_B$ with $\underline{p_A} < \bar{p}_B$, then $g(x+\delta) = c_A$ for all $\|\delta\|_2 < R$, where

$$R = \frac{\sigma}{2}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\bar{p}_B)).$$

In practice, Monte Carlo sampling is employed to obtain an estimate of p_A, see [29].

Unfortunately, as reported in [29], the robustness certificate is weak without any specifically-designed training techniques for randomized smoothing. Thus, a few techniques have been developed to enhance the robustness of randomized smoothing. In particular, [116] proposed to train base classifier f on adversarial examples that are generated by PGD [92] applied to soft-RS classifiers. Another line of work [152] considered an attack-free robust training by directly maximizing certified radius of each training sample. We briefly revisit these two methods [116, 152] in the following: Formally, suppose that $F^{\beta} : \mathbb{R}^d \to P(\mathcal{Y})$ is the soft version of classifier f whose last layer is a softmax layer with inverse temperature β and $P(\mathcal{Y})$ is a probability distribution over the label space $\mathcal{Y}$. We omit the superscript β if there is no ambiguity. Consider a smoothed soft classifier

$$G(x) = \mathbb{E}_{\delta \sim \mathcal{N}(0,\sigma^2 I)} F(x+\delta).$$

SmoothAdv [116] introduced SmoothAdv to find adversarial examples by PGD. Denote L_{CE} as the canonical cross entropy loss. Given a labeled data (x, y), SmoothAdv finds a point $\hat{x}$ that maximizes the cross entropy loss of $G(x)$ in the local neighborhood of x:

$$\hat{x} = \underset{\|x'-x\|_2 \le \epsilon}{\arg\max}\, L_{\text{CE}}(G(x'), y) = \underset{\|x'-x\|_2 \le \epsilon}{\arg\max} - \log \mathbb{E}_{\delta \sim \mathcal{N}(0,\sigma^2 I)} F(x'+\delta)_y. \tag{4.2.1}$$

Such optimization problem (4.2.1) is solved by projected gradient descent (PGD). To estimate the gradient of (4.2.1), [116] used Monte Carlo simulation to approximate

$\nabla_{x'} L_{\mathrm{CE}}(G(x'), y)$ by

$$\nabla_{x'}\Big(-\log\Big(\frac{1}{m}\sum_{k=1}^{m} F(x'+\delta_k)_y\Big)\Big),$$

where $\delta_1, \ldots, \delta_m$ are drawn i.i.d. from $\mathcal{N}(0, \sigma^2 I)$.

MACER Since the certified radius is related to the difference between the top probability p_A and the runner-up probability p_B, [152] constructed MACER loss L_{MACER}, which consists of classification loss and robustness loss to both minimize classification error and maximize the certified radius of those correctly classified samples. Specifically,

$$L_{\mathrm{MACER}}(x) = L_{\mathrm{CE}}(G(x), y) + \lambda L_{\mathrm{R}}(G; x, y), \tag{4.2.2}$$

where $\lambda \geq 0$ is a tuning parameter. The loss in (4.2.2) involves the soft smoothed classifier G and [152] proposes to approximate $G(x)$ by Monte Carlo sampling:

$$\begin{aligned} G(x) \approx \hat{z}(x) &= \frac{1}{m}\sum_{k=1}^{m} F(x+\delta_k), \qquad (4.2.3)\\ \hat{g}(x) &= \arg\max_{i\in\mathcal{Y}} \hat{z}_i(x). \end{aligned}$$

where $\delta_1, \ldots, \delta_m$ are drawn i.i.d. from $\mathcal{N}(0, \sigma^2 I)$. Denote by $\widehat{\mathrm{CR}}(x, y)$ the approximated certified radius at x, then

$$\widehat{\mathrm{CR}}(x, y) = \frac{\sigma}{2}(\Phi^{-1}(\hat{z}_y(x)) - \Phi^{-1}(\max_{y'\neq y} \hat{z}_{y'}(x))).$$

Therefore, the robustness loss $L_{\mathrm{R}}(G; x, y)$ can also be approximated by

$$\begin{aligned} L_{\mathrm{R}}(G; x, y) \approx L_{\mathrm{R}}(\hat{z}; x, y) &= \max\{\epsilon + \tilde{\epsilon} - \widehat{\mathrm{CR}}(x, y), 0\}\, \mathbf{1}(\hat{g}(x) = y)\\ &= \frac{\sigma}{2}\max\{\gamma - \hat{\xi}_\theta(x, y), 0\}\, \mathbf{1}(\hat{g}(x) = y), \qquad (4.2.4) \end{aligned}$$

where $\epsilon, \tilde{\epsilon} > 0$ are hyper-parameters in hinge loss, $\gamma = \frac{2}{\sigma}(\epsilon + \tilde{\epsilon})$, and

$$\hat{\xi}_\theta(x, y) = \Phi^{-1}(\hat{z}_y(x)) - \Phi^{-1}(\max_{y' \neq y} \hat{z}_{y'}(x)).$$

Finally, MACER trains a base classifier by minimizing the approximated MACER loss on training dataset. We refer readers to [152] for more details.

Ensemble. Model Ensemble is a popular technique in the machine learning literature to practically improve model performance and reduce generalization errors [3]. Recently, there are a few works investigating the idea of using model ensemble to improve robustness of a randomized smoothed classifier [57, 147]. However, [57] focused on ensemble the same type of models (i.e. models trained from the same process but with different random seeds) and did not study the effect of using different types of models. Although [147] doesn't have any explicit assumptions on model types, they only experimented using same types of model to ensemble. Also, their analysis is based on a model smoothness assumption, which is not easy to verify especially for DNN. In contrast, as will be introduced in Section 3.2, our proposed **EsbRs** is a more general ensemble method where we study the effect of *mixed*-model ensembles and *optimal* weighted ensembles. Although weighted ensemble has also been studied in [57], their model learns the weights from training and cannot justify the weights' optimality. However, in our work, we develop a novel design framework of the optimal weight ensemble based on our theory to best improve the robustness certificate of a randomized smoothed classifier.

4.3 Our proposed main methods

In this section, we propose two novel and cost-effective approaches to improve robustness of a randomized smoothed classifier. First, we introduce a new robust training method **AdvMacer** that aim to maximize the certified radius over adversarial examples, and we present the intuitions, formulations as well as the details of our algorithm in

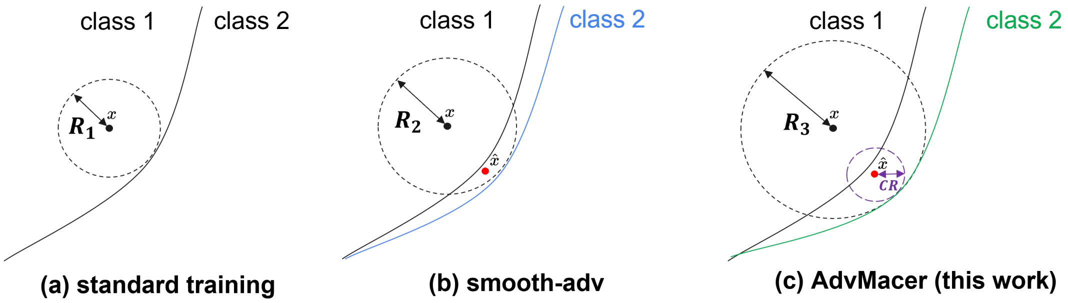


Figure 4.1. The illustration of the idea behind **AdvMacer** : x (black dot) is the original data sample and $\hat{x}$ (red dot) is an adversarial example of x. The solid black line is the original decision boundary. The blue line in (b) is the decision boundary using SmoothAdv and the green line in (c) is the decision boundary after applying **AdvMacer** . SmoothAdv force the classifier to classify $\hat{x}$ correctly to get the red boundary. **AdvMacer** force $\hat{x}$ to not only have correct prediction but also a large margin. Therefore, **AdvMacer** can obtain larger certified radius $R_3 >$ certified radius of smoothadv $R_2 >$ certified radius of the original classifier R_1.

Section 4.3.1. Next, in Section 4.3.2, we propose a novel ensemble method called **EsbRs** with theoretical analysis. Different from the two recent works [147, 57], we provide a more general analysis which does not require individual classifiers to come from the same training method. Our analysis allows the derivation of the optimal weight for individual classifiers, which is the key to promote robustness and the study of optimal weight has not been explored in the prior work.

4.3.1 Approach 1: AdvMacer

Inspired by the prior work SmoothAdv [116] and MACER [152] and to address their limitations, we argue that a smoothed classifier can be trained to have larger certified radius by directly optimizing the certified radius of adversarial examples instead of the clean data points. However, notice that this statement requires adversarial example to be predicted correctly (otherwise, the certified radius of original data point may be actually decreased). The intuition is illustrated in Figure 4.1. Based on the above idea, we propose the following formulation.

Formulation.

Given data x and its label y, we aim to minimize the proposed **AdvMacer** loss consisting of two terms:

$$L_{\text{AdvMacer}}(x) = L_{\text{CE}}(\hat{z}(\hat{x}), y) + \lambda L_{\text{R}}(\hat{z}; \hat{x}, y),$$

where $\hat{z}$ and L_{R} are given in (4.2.3) and (4.2.4) respectively. The 1st term $L_{\text{CE}}(\hat{z}(\hat{x}), y)$ is to encourage adversarial examples $\hat{x}$ to be classified correctly, and the 2nd term

$$L_{\text{R}}(\hat{z}; \hat{x}, y) = \frac{\sigma}{2} \max\{\gamma - \hat{\xi}_\theta(\hat{x}, y), 0\} \, \mathbf{1}(\hat{g}(\hat{x}) = y)$$

is to maximize the certified radius at the adversarial example $\hat{x}$, where

$$\hat{x} = \operatorname*{arg\,max}_{\|x' - x\|_2 \le \epsilon} L_{\text{CE}}(\hat{z}(x'), y).$$

To minimize the $L_{\text{AdvMacer}}(x)$, we generate the adversarial examples $\hat{x}$ via Equation (4.2.1) with T-step PGD using SmoothAdv [116], i.e. in the i-th step, we update

$$x_{i+1} = \prod_{\mathcal{B}(x,\epsilon)} \left(x_i + \nabla_x \Big(-\log \Big(\frac{1}{m} \sum_{k=1}^{m} F(x' + \delta_k)_y \Big) \Big) \Big|_{x = x_i} \right),$$

where $\prod_{\mathcal{S}}(\cdot)$ is the projection onto set $\mathcal{S}$ and we set $\hat{x} = x_T$. The training objective is to minimize $L_{\text{AdvMacer}}(x)$ by first-order optimization method, and a detailed algorithm is presented in the Appendix due to page constraint.

Hyper-parameters Note that there are a few hyper-parameters in **AdvMacer** : σ is the noise level that is introduced when f or F is smoothed; ϵ in (4.2.1) controls the size of the ℓ_2 ball when doing PGD; γ in (4.2.4) is the parameter in hinge loss; λ is the regularization parameter which controls the trade-off between clean accuracy and robustness; m in (4.2.3) is the number of Monte Carlo samples used to estimate $G(x)$; T

is the number of PGD step to generate adversarial samples. Finally, recall that the soft classifier $F = F^{\beta}$, where β is the inverse temperature in softmax layer. The larger β is, the closer the soft classifier F is to the hard classifier f.

Discussion and Comparison.

(I). SmoothAdv [116] adapted adversarial training to defend against the least favorable samples but did not consider certified radius as another metric. MACER [152] used robust training to directly maximize certified radius on clean samples instead of adversarial examples. In contrast, our proposed **AdvMacer** trains a model on adversarial examples while taking certified radius into consideration. Compared with SmoothAdv, **AdvMacer** doesn't bring any additional computational overhead to calculate robust loss as there exist analytic formula for certified radius; in the meantime, compared with MACER, we require much fewer number of epochs (3× smaller) to obtain a robust model with much larger certified radius. From the experiments in Section 4.4, it can be seen that **AdvMacer** outperforms both SmoothAdv and MACER on various dataset. **(II)**. SmoothAdv needs to tune a number of hyper-parameters for different noise level σ, which becomes a significant challenge when the computing resources are limited. Although MACER also has many tuning parameters, empirical experiments showed that most of these parameters don't change across different σ and datset. However, MACER needs more training epochs (440 epochs per [152]) to yield a robust classifier, taking days to train a ResNet-110 [56] on Cifar-10 [73]. Also, MACER usually achieves better clean accuracy but smaller average certified radius (ACR). In contrast, our **AdvMacer** takes the best of both world in Smoothadv and macer: **AdvMacer** enjoys computational efficiency, gives larger ACR while preserves good accuracy in most settings. Besides, **AdvMacer** attains a universal configuration that works well across different σ. Equipped with ensemble method presented in Section 4.3.2, **AdvMacer** also enriches the diversity of component models, making mixed ensemble more robust. For a thorough comparison by experiments, see

Section 4.4 for more details.

4.3.2 Approach 2: EsbRs

Ensemble is a cost-effective post-training technique to enhance model performance and reduce generalization error without spending much additional efforts on re-training the neural networks. By simply averaging the output from several models, ensemble shows remarkable boost in test accuracy and model robustness. Recently, there are a few works investigating the idea of using model ensemble to improve robustness of a randomized smoothed classifier [57, 147]. However, the existing work mainly focused on ensembling similar classifiers with learnable weights. In contrast, we also consider mixed ensemble with component classifiers coming from different training methods and conduct theoretical analysis explaining the success of mixed ensemble in certain cases. Besides, unlike [83] learning the ensemble weights empirically from training set, we develop a novel theoretical framework to design optimal ensemble weights based on our analysis. Empirical experiments verify the superiority of our proposed methods.

Formulation Suppose we have k trained soft classifiers $F^1, \dots, F^k : \mathbb{R}^d \to P(\mathcal{Y})$ and $\mathcal{Y} = \{1, \dots, c\}$. Consider soft-ensemble model H whose output is a weighted average of the logits from $F^1, \dots, F^k$:

$$H(x) = \sum_{l=1}^{k} w_l F^l(x).$$

Suppose the associated hard classifier is

$$h(x) = \arg\max_{c \in \mathcal{Y}} \big(H(x)\big)_c.$$

Then we apply RS to h and get the corresponding smoothed classifier g. Extensive experiments from Section 4.4 show that ensemble-RS classifier g outperforms all component classifiers $F^1, \dots, F^k$ in general, no matter F^l comes from the same or different training methods. Specifically, if F^l comes from more than one training methods, we call g a mixed ensemble.

Theoretical analysis We present some theoretical analysis on how (mixed) ensemble can reduce the variance and hence increase certified radius. We generalize the analysis in [57] to allow mixed ensemble, which provide deeper insights on model ensemble study and in fact motivate a novel design of optimal ensemble as described in **Designing optimal weighted ensemble** paragraph.

For a fixed query point x with a Gaussian perturbation $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, suppose logits vector $\boldsymbol{y}^l \in P(\mathcal{Y})$ is returned by F^l. Without loss of generality, assume 1 is the majority class in RS for all F^l. For simplicity, we can work with classification margin $z_i^l = y_1^l - y_i^l$, for $i \in \mathcal{Y}$. Let $\bar{\boldsymbol{y}} = H(x + \epsilon)$. Therefore, $\bar{\boldsymbol{y}} = \sum_{l=1}^k w_l \boldsymbol{y}^l$. Similarly define $\bar{z}_i = \bar{y}_1 - \bar{y}_i$. Consider $\mathbb{E}[\bar{\boldsymbol{z}}] \in \mathbb{R}^c$ and $\mathrm{Var}(\bar{\boldsymbol{z}}) \in \mathbb{R}^{c \times c}$, where the expectation is taken over the randomness in training process, including random initialization and stochasticity in GD. Then we have

$$\begin{aligned} \mathrm{Var}(\bar{\boldsymbol{z}}) &= \mathrm{Var}\Big(\sum_{l=1}^k w_l \boldsymbol{z}^l\Big) \\ &= \sum_{l=1}^k w_l^2 \mathrm{Var}(\boldsymbol{z}^l) + 2 \sum_{l \neq m} w_l w_m \mathrm{Cov}(\boldsymbol{z}^l, \boldsymbol{z}^m) \end{aligned} \tag{4.3.1}$$

Hence, $\mathrm{Var}(\bar{z}_i) = \mathrm{Var}(\bar{\boldsymbol{z}})_{ii}$. Denote $p_i(w) = \mathrm{Var}(\bar{z}_i)$ as a function of $w = [w_1, \ldots, w_k]^\top$ and

$$\begin{aligned} \alpha_i = \alpha_i(s) &= \max_{1 \le l \le k} \mathrm{Var}(\boldsymbol{z}^l)_{ii} \\ \beta_i = \beta_i(s) &= \max_{l \neq m} \mathrm{Cov}(\boldsymbol{z}^l, \boldsymbol{z}^m)_{ii} \end{aligned}$$

Suppose there are a fixed number of training methods and denote this number by s, so the above maximum is in fact taken over s different classes. As a result, $\alpha_i(s), \beta_i(s) = O(1)$ even as $k \to \infty$.

A special case As a special case, consider $w_l = \frac{1}{k}$ for all $l = 1, \ldots, k$. By (4.3.1), we derive

$$p_i(w) = \mathrm{Var}(\bar{z}_i) \le \frac{k\alpha_i + k(k-1)\beta_i}{k^2} = \beta_i + \frac{\alpha_i - \beta_i}{k}. \tag{4.3.2}$$

These classifiers either come from different training methods, or same training method with different random seeds. Thus, existing work all assumes that the logits from one classifier have larger covariance α_i than the logits from different classifiers β_i. However, as we will see in **Discussion** paragraph, ensemble may harm the performance if the above assumption doesn't hold. For now, let's assume $\alpha_i > \beta_i$. By (4.3.2), we conclude that the upper bound of $\mathrm{Var}(z_i)$ decreases to a constant β_i as $k \to \infty$.

Next, we explain how $\mathrm{Var}(\bar{z}_i)$ affects certified radius. From Theorem 4, we see that $R = \sigma\Phi^{-1}(\underline{p}_A)$ if $\underline{p}_A \ge \frac{1}{2}$, hence we only need to show a lower bound on the top class probability p_A increases as k becomes larger. Since we assume the majority class's number is 1, we see that

$$p_1 = \mathbb{P}(\bar{z}_i > 0, \forall i = 2, \dots, c) \ge 1 - \sum_{i=2}^{c} \mathbb{P}(\bar{z}_i \le 0) \tag{4.3.3}$$

By Chebyshev's inequality, $\mathbb{P}(\bar{z}_i \le 0) \le \mathbb{P}\Big(\big|\bar{z}_i - \mathbb{E}[\bar{z}_i]\big| \ge \mathbb{E}[\bar{z}_i]\Big) \le \frac{\mathrm{Var}(\bar{z}_i)}{\mathbb{E}[\bar{z}_i]^2}$ and let $e_i = e_i(s) = \min_l \mathbb{E}[z_i^l]$, thus from (4.3.3) we have

$$p_1 \ge 1 - \sum_{i=2}^{c} \frac{\mathrm{Var}(\bar{z}_i)}{e_i^2}. \tag{4.3.4}$$

The above equation suggests us to choose the weight w that maximizes the RHS of (4.3.4) to have a larger p_1, hence larger certified radius. Since e_i is independent of the choice of w, we can obtain the optimal weight by solving

$$\min_{w \in \mathbb{R}^k} \sum_{i=2}^{c} a_i p_i(w) \quad \text{s.t.} \sum_{l=1}^{k} w_l = 1, \quad w_l \ge 0, \tag{4.3.5}$$

where $a_i = e_i^{-2}$ are constants. Note that when $w_l = \frac{1}{k}$ for all $l = 1, \dots, k$, we have a lower bound on p_1 by (4.3.2) and (4.3.4):

$$p_1 \geq 1 - \sum_{i=2}^{c} \frac{\beta_i + (\alpha_i - \beta_i)/k}{e_i^2} \to 1 - \sum_{i=2}^{c} \frac{\beta_i}{e_i^2} \text{ as } k \to \infty.$$

This explains why larger k makes p_1 and certified radius larger even in average ensemble.

Discussion

Compared with [57], we generalize their analysis to allow both weighted and mixed ensemble and hence have several new findings. First, if $\alpha_i < \beta_i$, namely the logits from one model have smaller variance than those from different models, the RHS of (4.3.2) becomes an increasing function in k, which implies ensemble does not always work. Second, suppose F^1, F^2 come from model category 1 (for example, SmoothAdv) and F^3 comes from model category 2 (for example, **AdvMacer**). If the logits from different types of models have smaller variance than those from the same type of model, namely $\mathrm{Cov}(F^1, F^3) < \mathrm{Cov}(F^1, F^2)$, β_i will become smaller and makes mixed ensemble work better than single ensemble. This phenomenon is observed in Figure 4.2.

Designing optimal weighted ensemble

The optimization problem in (4.3.5) allows us to design an optimal weight that can maximize the lower bound on p_1. Consider the case where $k = 2$, then (4.3.5) can be solved analytically given the knowledge of $\mathrm{Var}(z^1), \mathrm{Var}(z^2), \mathrm{Cov}(z^1, z^2)$ and a_i. To see this, let $b_i = \mathrm{Var}(z^1)_{ii}, c_i = 2\mathrm{Cov}(z^1, z^2)_{ii}, d_i = \mathrm{Var}(z^2)_{ii}$, then the objective function in (4.3.5) can be re-written as

$$q(w) = \sum_{i=2}^{c} a_i(b_i w_1^2 + c_i w_1 w_2 + d_i w_2^2) \overset{\text{(i)}}{=} \sum_{i=2}^{c} a_i[b_i w_1^2 + c_i w_1(1 - w_1) + d_i(1 - w_1)^2], \quad (4.3.6)$$

where (i) uses the constraint $w_1 + w_2 = 1$ to eliminate w_2. Therefore, the problem (4.3.5) can be further cast as a quadratic optimization with linear constraints:

$$\begin{aligned} \min_{w_1 \in \mathbb{R}} \quad & q(w_1) = Aw_1^2 + Bw_1 + C \\ \text{s.t.} \quad & 0 \le w_1 \le 1, \end{aligned} \tag{4.3.7}$$

where $A = \sum_{i=2}^{c} a_i(b_i + c_i + d_i), B = \sum_{i=2}^{c} -a_i(c_i + 2d_i)$ and $C = \sum_{i=2}^{c} a_i d_i$. Notice that this problem has an analytical solution: if $A > 0$ and $0 \le -\frac{B}{2A} \le 1$, $w_1 = -\frac{B}{2A}$ and $w_2 = 1 + \frac{B}{2A}$; else $q(w_1)$ attains minimum at boundary $w_1 = 0$ or 1.

Next, we aim at giving an estimate of a_i, b_i, c_i, d_i. To account for randomness both from training and Gaussian perturbation ϵ around the input x, we first generate n i.i.d. Gaussian noisy data $x_1, \ldots, x_n$ from $\mathcal{N}(x, \sigma^2 I)$. Second, we incorporate random perturbation for the parameters θ in classifier F to imitate random seeds in training, as this is the cheapest way (without extra training cost). We randomly select $t\%$ parameters from F and add i.i.d. Gaussian noise $\delta \sim \mathcal{N}(0, \tilde{\sigma}^2)$ for each selected parameter. This returns a perturbed model $\hat{F}$ from the base model F. Repeating the above process on F^1 and F^2 for m times gives us $2m$ perturbed models $\hat{F}_1^1, \ldots, \hat{F}_m^1$ and $\hat{F}_1^2, \ldots, F_m^2$.

Now, we pass $x_1, \ldots, x_n$ into $\hat{F}_1^1, \ldots, \hat{F}_m^1$ to get mn output logits vector

$$y^{1,1}, y^{1,2}, \ldots, y^{1,mn}.$$

Also, $y^{2,1}, y^{2,2}, \ldots, y^{2,mn}$ can be obtained similarly by passing n noisy data into m perturbed models of F^2. Compute $z_i^{l,j} = y_1^{l,j} - y_i^{l,j}$ for $1 \le i \le c$ and $l = 1, 2$. Then an estimation of variance and covariance can be their empirical parallel:

$$
\begin{aligned}
b_i &= \mathrm{Var}(z^1)_{ii} = \frac{1}{mn}\sum_{j=1}^{mn}(z^{1,j} - \bar{z}^1)(z^{1,j} - \bar{z}^1)^\top_{ii}, \\
c_i &= 2\mathrm{Cov}(z^1, z^2)_{ii} = \frac{2}{mn}\sum_{j=1}^{mn}(z^{1,j} - \bar{z}^1)(z^{2,j} - \bar{z}^2)^\top_{ii}, \\
d_i &= \mathrm{Var}(z^2)_{ii} = \frac{1}{mn}\sum_{j=1}^{mn}(z^{2,j} - \bar{z}^2)(z^{2,j} - \bar{z}^2)^\top_{ii},
\end{aligned}
$$

where $\bar{z}^l = \frac{1}{mn}\sum_{j=1}^{mn} z^{l,j}$ for $l = 1, 2$. Also obtain $a_i = e_i^{-2} = \min\{\bar{z}_i^1, \bar{z}_i^2\}^{-2}$ Hence, we can solve (4.3.7) by plugging in a_i, b_i, c_i, d_i. A detailed algorithm is given in Algorithm 6 in Section 4.9.

Remark 1. To our best knowledge, we are the first work to develop a practical and theoretical grounded methodology to obtain the optimal weight of the ensemble scheme. We note that the two recent works [147, 57] did not explore this direction.

4.4 Experiments

In this section, we present experimental results that empirically evaluate the performance of our proposed methods, **AdvMacer** and **EsbRs**, on Cifar-10 [73] and SVHN [103] dataset. To make fair comparisons with previous baseline models, we use the same architectures as in [29, 116, 152]: ResNet-110 [56]. We train our models with $\sigma = 0.25, 0.50, 1.00$ on Cifar-10 and $\sigma = 0.25, 0.50$ on SVHN. We train all models on a single NVIDIA V100 GPU and the training time reported below is all from NVIDIA V100 GPU.

Evaluation We mainly evaluate model performance on two metrics: clean accuracy and average certified radius (ACR). Clean accuracy is the classification accuracy when taking the original test images as the input and cannot evaluate model robustness. A more

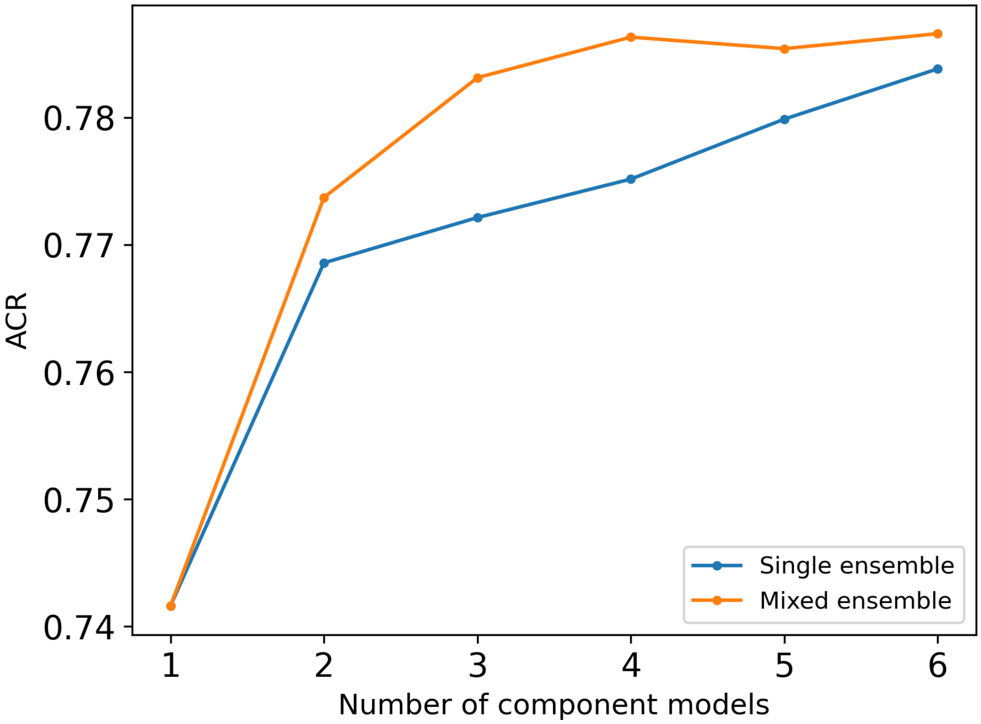


Figure 4.2. The plot of ACR against different number of component models in **EsbRs** on Cifar-10 with $\sigma = 0.50$. Single ensemble uses N **AdvMacer** models. Mixed ensemble with totally N component models uses m **AdvMacer** models and n SmoothAdv models, where m and n are given in Table 4.5 ofSection 4.7.

reasonable metric for evaluating robustness is ACR. We follow the standard evaluation protocol used in [29, 116, 152] for fair comparison: for each test data $(x_i, y_i) \in \mathcal{S}_{\text{test}}$, record the radius R_i that can be certified the by the model g. Set $R_i = 0$ if x_i can't be classified correctly by g. Then ACR $= \frac{1}{|\mathcal{S}_{\text{test}}|} \sum_i R_i$. Since the denominator is the size of the full test set, one cannot obtain large ACR without high accuracy. Thus ACR becomes a popular choice in most of the DL robustness literature. We use CERTIFY algorithm in [29] to obtain certified radius and choose $N_0 = 100, N = 100,000, \alpha = 0.001$ in CERTIFY. We report model performance on the first 500 test images on Cifar-10 and SVHN.

Baseline models Two baseline models are discussed in this section: MACER

Table 4.1. Cifar-10: ACR of different models on the first 500 test images of Cifar-10 with varing σ. Clean accuracy is reported in parenthesis. Reported models include SmoothAdv, MACER, **AdvMacer** , Esb-RS. $N = 100k$ samples are used in certification unless otherwise specified.

	Methods	$\sigma = 0.25$	$\sigma = 0.5$	$\sigma = 1.0$	Ensemble?
Baselines	SmoothAdv [116]	0.541 (74.2%)	0.735 (56.4%)	0.758 (45.8%)	×
	MACER [152]	0.518 (79.4%)	0.682 (63.4%)	0.768 (42.4%)	×
	SmoothMix [62]	0.545 (76.0%)	0.685 (63.8%)	0.626 (48.4%)	×
	SmoothMix+1-step adv [62]	0.533 (72.8%)	**0.743** (62.0%)	0.788 (43.0%)	×
	Consistency [63]	0.535 (78.4%)	0.701 (64.6%)	0.719 (45.8%)	×
	Consistency + SmoothAdv [63]	0.532 (71.8%)	0.733 (53.2%)	**0.834** (42.8%)	×
Ours	**AdvMacer**	**0.554** (76.0%)	**0.742** (58.4%)	**0.794** (47.6%)	×
	EsbRs-AdvMacer ×3	**0.583** (76.4%)	0.772 (58.8%)	0.805 (47.6%)	√
	EsbRs-SmoothAdv×3	0.567 (76.6%)	0.777 (58.4%)	0.801 (46.6%)	√
	EsbRs-AdvMacer ×1+SmoothAdv×2	0.572 (77.2%)	**0.783** (59.4%)	**0.810** (47.2%)	√
	EsbRs-AdvMacer ×2+MACER×1	0.568 (79.8%)	0.728 (63.6%)	0.801 (42.8%)	√
	EsbRs-AdvMacer ×1+MACER×2	0.570 (80.4%)	0.723 (65.0%)	0.760 (44.0%)	√

and SmoothAdv. For MACER, we follow the configurations given by Table 4 in [152]. For SmoothAdv, we pick the best models under different $\sigma = 0.25, 0.50, 1.00$ from the Github repo of [116]. See Table 4.4 in Section 4.6 for more details on hyper-parameter selection of SmoothAdv.

AdvMacer We apply Algorithm 5 to train our **AdvMacer** models. On Cifar-10, we choose $\gamma = 8.0$, $\lambda = 12.0$, $\beta = 16.0$ for all $\sigma = 0.25, 0.50, 1.00$. The choice of T, m, ϵ are summarized in Table 4.4. We follow the same training scheme as [116]. The initial learning rate is 0.1 and decays by a factor of 0.1 every 50 epochs. A batch size of 256 is used in the training. For more details, please refer to [116]. Note that by the choice of hyper-parameters, SmoothAdv and **AdvMacer** have the same training time, which implies the improved performance of **AdvMacer** is not gained from more expensive computation. The experiment results on Cifar-10 are summarized in Table 4.1.

EsbRs We also employ our proposed ensemble technique introduced in Section 4.3.2 to enhance robustness performance. We use the following naming convention to report our result: **EsbRs-**Model1×n+Model2×m represents the ensemble model obtained by n Model1 and m Model2. Each component model's configuration is the same as that of the non-ensemble individual model for fair comparison. For example, **EsbRs-AdvMacer** ×1+SmoothAdv×2 on Cifar-10 with $\sigma = 0.25$ represents the ensemble of one **AdvMacer** model and two SmoothAdv models with configuration from Table 4.4.

Our empirical experiments also verifies the theoretical analysis on the success of mixed ensemble and optimal weighted ensemble in Section 4.3.2. In ensemble experiments, we independently train all **AdvMacer** and SmoothAdv models on Cifar-10 with $\sigma = 0.50$ and the model configuration is given by Table 4.4. For single ensemble, we use 1/2/3/4/5/6 **AdvMacer** models. For mixed ensemble, the number of component model from each category is summarized in Table 4.5 of Section 4.7. In Figure 4.2, we observe that ACR improves as the number of component models increases. The same observation holds for both single ensemble and mixed ensemble. Besides, mixed ensemble gives universally better ACR as shown in Figure 4.2, which is in accordance to the analysis in Section 4.3.2.

Discussion

Different from [57], we allow mixed ensemble that are a mixture of robust models from various training methods. The introduction of **AdvMacer** brings enriched diversity of component models that can be used in model ensemble. In addition, we conduct experiment on optimal weighted ensemble with two component models. The weights w_1, w_2 are computed by Algorithm 6. The experiment is on Cifar-10 with $\sigma = 0.25$ and choose both F^1 and F^2 from **AdvMacer** models. Weighted ensemble model is given by $H = w_1 F^1 + w_2 F^2$. We set $n = 10, m = 10, \sigma = 0.25, \tilde{\sigma} = 0.01, t = 0.3$ and certify each test image x by H using **CERTIFY** algorithm from [29] with $N_0 = 100, N = 100,000, \alpha = 0.001$. The results are given in Table 4.2. It shows that the choice of optimal weight does improve both accuracy and ACR, compared with average weighted ensemble method. It is worth noting that we are the first work to study and propose the optimal design of ensemble weight for randomized smoothing to best improve the robustness.

Additional experiment on SVHN We compare the performance of SmoothAdv,

Table 4.2. Optimal weighted ensemble. ACR and clean accuracy on the first 500 test images of Cifar-10 with $\sigma = 0.25$. All certification has parameters $N_0 = 100, N = 100,000, \alpha = 0.001$. Optimal weights are computed from Algorithm 6 with **AdvMacer** models F^1, F^2, $m = 10, n = 10, t = 0.3, \sigma = 0.25, \tilde{\sigma} = 0.01$.

Model	Accuracy	ACR	Certificate Time
AdvMacer	0.760	0.554	8.9s
Avg weight **EsbRs**	0.760	0.572	18.0s
MME [147]	0.754	0.567	19.6s
Optimal weight **EsbRs**	**0.766**	**0.576**	26.3
max-**EsbRs** ($n = 10^2$)	**0.762**	**0.591**	9.0s
max-**EsbRs** ($n = 10^4$)	**0.766**	**0.597**	10.8s

MACER and **AdvMacer** on SVHN dataset with $\sigma = 0.25, 0.50$. On SVHN with $\sigma = 0.25$, we choose $T = 2$, $m = 4$, $\lambda = 12.0$, $\gamma = 8.0$, $\beta = 16.0$, $\epsilon = 0.5$ and train the model for 150 epochs. On SVHN with $\sigma = 0.50$, we still choose $T = 2$, $m = 4$, $\gamma = 8.0$, $\beta = 16.0$, $\epsilon = 0.5$ but a different $\lambda = 4.0$. The model is also trained for 150 epochs. The initial learning rate is set to 0.01 and drops by a factor of 0.1 every 50 epochs. The other training details follow the same as Cifar-10. For SmoothAdv, take $T = 2, m = 4, \epsilon = 0.5$ when $\sigma = 0.25$ and $T = 2, m = 4, \epsilon = 0.25$ when $\sigma = 0.50$. We train MACER model for 440 epochs whose configuration is given by C.2.2 of [152]. We report the experiment results in Table 4.3 and leave details in Table 4.9 in Section 4.10.

Certified accuracy and Performance We also report the certified accuracy in Section 4.8 for each ℓ_2 radius ranging from 0.25 to 2.00 and increasing by 0.25 under $\sigma = 0.25, 0.50, 1.00$ on Cifar-10. From Table 4.1, Table 4.3 in this section and Table 4.9 in Section 4.10, **AdvMacer** has largest ACR among all non-ensemble models for every σ on both Cifar-10 and SVHN dataset. Training time of **AdvMacer** is the same as SmoothAdv, but significantly less than MACER. Ensemble can boost both accuracy and ACR and **EsbRs-AdvMacer** $\times 3$ achieves the best ACR on Cifar-10 with $\sigma = 0.25$. **AdvMacer** $\times 1$+SmoothAdv$\times 2$ outperforms all the other models on Cifar-10 with $\sigma = 0.50, 1.00$,

Table 4.3. SVHN: clean accuracy and ACR of different models evaluated at the first 500 test images of SVHN with $\sigma = 0.25$.

Model	Accuracy	ACR
SmoothAdv	85.8%	0.560
MACER	86.8%	0.549
AdvMacer	86.6%	**0.569**
EsbRs-SmoothAdv×3	87.8%	0.578
EsbRs-AdvMacer ×3	88.2%	**0.582**
EsbRs-AdvMacer ×1+MACER×2	87.8%	0.559
EsbRs-AdvMacer ×2+MACER×1	88.6%	0.570
EsbRs-AdvMacer ×1+SmoothAdv×2	87.8%	0.577
EsbRs-AdvMacer ×2+SmoothAdv×1	87.6%	**0.582**

suggesting that one may prefer mixed ensemble in particular situations.

Universal configuration As SmoothAdv requires different model configurations for different tasks, unexpectedly long training time becomes a challenging issue especially when the computing resource is limited. Smaller number of Gaussian samples m and fewer PGD steps T in SmoothAdv can reduce the training time significantly but also compromise the robustness significantly. However, **AdvMacer** combines adversarial and robust training and is expected to still performs well even with reduced m and T. We choose universal configuration $m = 2$ and $T = 2$ for all experiment setting on Cifar-10 and compare the performance to SmoothAdv with the same configuration in **??**. For every σ, **AdvMacer** is at least comparable to SmoothAdv and MACER. For example, the ACR of **AdvMacer** is only smaller than the best model by 0.002, but costs 12% of the training time. Moreover, on $\sigma = 0.50$ and $\sigma = 1.00$, **AdvMacer** has noticeably larger ACR than both SmoothAdv and MACER, while still costs only 12% of the training time of MACER. For larger σ ($\sigma = 1.00$), **AdvMacer** even has the best clean acc among all reported models. This characteristic makes **AdvMacer** more scalable and cost-effective.

4.5 Full Algorithm of AdvMacer

```
Algorithm 5. Our AdvMacer $(\sigma, m, T, \lambda, \beta, \gamma)$
  Input: training set $\hat{p}_{\text{data}}$, noise level $\sigma$, number of Gaussian samples $m$, regularization parameter $\lambda$, hinge factor $\gamma$, inverse temperature $\beta$, number of PGD step $T$
  for each iteration do
      1) Sample a mini-batch $(x_1, y_1), \ldots, (x_n, y_n) \sim \hat{p}_{\text{data}}$
      2) For each $(x_i, y_i)$, use $T$-step SmoothAdv to generate adversarial example $\hat{x}_i$
      3) For each $(\hat{x}_i, y_i)$, draw $m$ i.i.d. Gaussian samples $x_{i1}, \ldots, x_{im}$ from $\mathcal{N}(x_i, \sigma^2 I)$
      4) Obtain an estimation of $G_\theta(\hat{x})$ by
            $\hat{z}_\theta(\hat{x}) \leftarrow \frac{1}{m} \sum_{k=1}^{m} F_\theta(\hat{x}_{ik}), \text{ for } i = 1, \ldots, n$
      5) Collect the set of data with correct prediction:
            $\mathcal{S}_\theta = \{i : y_i = \arg\max_c \hat{z}_\theta(\hat{x}_i)_c\}$
      6) For each $i \in \mathcal{S}_\theta$, compute the second most likely class
            $\hat{y}_i \leftarrow \arg\max_{c \neq y_i} \hat{z}_\theta(\hat{x}_i)_c$
      7) For each $i \in \mathcal{S}_\theta$, compute
            $\hat{\xi}(\hat{x}_i, y_i) \leftarrow \Phi^{-1}(\hat{z}_\theta(x)_{y_i}) - \Phi^{-1}(\hat{z}_\theta(x)_{\hat{y}_i})$
      8) Sample $\delta \sim \mathcal{N}(0, \sigma^2 I)$ and update $\theta$ with SGD to minimize
            $-\frac{1}{n} \sum_{i=1}^{n} \log \hat{z}_\theta(\hat{x}_i + \delta)_{y_i}$
                $+ \frac{\lambda\sigma}{2n} \sum_{i \in \mathcal{S}_\theta} \max\{\gamma - \hat{\xi}_\theta(\hat{x}_i + \delta, y_i), 0\}$
  end for
  Output: model parameters $\theta$
```

Table 4.4. Model configuration: main hyper-parameters and training time for Smooth-Adv and **AdvMacer** on Cifar-10 with varing σ. For the additional parameters in **AdvMacer** , we pick $\lambda = 12.0, \gamma = 8.0, \beta = 16.0$.

Models	T	m	ϵ	Epochs	σ	Time
SmoothAdv	2	8	1.0	150	0.25	15.5h
	2	8	2.0	150	0.50	15.5h
	2	4	2.0	150	1.00	8h
AdvMacer	2	8	1.0	150	0.25	15.5h
	2	8	2.0	150	0.50	15.5h
	2	4	2.0	150	1.00	8h

4.6 Model configuration

We summarize the configuration of the SmoothAdv and **AdvMacer** models on Cifar-10 in Table 4.4.

4.7 Details of mixed ensemble component

The details of the component models for mixed ensemble in Figure 4.2 are given in Table 4.5.

Table 4.5. Component models in mixed ensemble experiment in Figure 4.2. The mixed ensemble with totally N component models uses m **AdvMacer** and n SmoothAdv models. m and n are given as follows.

N	m	n
1	1	0
2	1	1
3	1	2
4	2	2
5	3	2
6	3	3

4.8 Certified accuracy

More results on certified accuracy of Cifar-10 is presented in Section 4.8 and Table 4.8.

4.9 Optimal weighted ensemble

We give the detailed algorithm (**ComputeWeight**$(F^1, F^2, \sigma, \tilde{\sigma}, n, m, t, x)$) to compute the optimal weight in ensemble in Algorithm 6.

4.10 More SVHN experiments

See Table 4.9 for more experiments.

4.11 Certification with more or fewer samples

Since ensemble of k models takes k times longer certification time, we also certify ensemble model with N/k samples and single model with kN samples to make certification time comparable. Esb-**AdvMacer** ×3 with $N/3$ samples underperforms **AdvMacer** mainly due to insufficient number of samples ([29] claimed one needs 10^5 samples to achieve significance level $\alpha = 0.001$). Compared SmoothAdv with $3N$ samples to SmoothAdv×3, ensemble model outperforms the base model while $\sigma = 0.50, 1.00$ and is only slightly worse while $\sigma = 0.25$, which showcases the power of ensemble. For a complete table, see Table 4.10 below.

4.12 ImageNet

4.13 Conclusions

In this work, we have proposed two novel and cost-effective approaches to promote robustness of randomized smoothed classifiers. Our first approach **AdvMacer** improve

Table 4.6. Certified accuracy: certified accuracy and ACR of the first 500 test images of Cifar-10 with $\sigma = 0.25$. Each column represents the robust accuracy that can be certified at this ℓ_2 radius.

Model ($\sigma = 0.25$)	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	ACR
SmoothAdv	0.742	0.660	**0.572**	0.45	0.0	0.0	0.0	0.0	0.0	0.541
MACER	**0.794**	**0.678**	0.524	0.400	0.0	0.0	0.0	0.0	0.0	0.518
SmoothMix	0.76	0.688	0.572	0.446	0.0	0.0	0.0	0.0	0.0	0.545
SmoothMix+1-step Adv	0.728	0.634	0.564	0.47	0.0	0.0	0.0	0.0	0.0	0.533
AdvMacer	0.76	0.668	**0.572**	**0.484**	0.0	0.0	0.0	0.0	0.0	**0.554**
EsbRs-AdvMacer ×2 (opt weight)	**0.766**	**0.698**	**0.608**	**0.496**	0.0	0.0	0.0	0.0	0.0	0.576
EsbRs-AdvMacer ×2 (avg weight)	0.76	0.69	0.604	0.49	0.0	0.0	0.0	0.0	0.0	0.572
EsbRs-AdvMacer ×3	0.764	0.700	**0.614**	**0.514**	0.0	0.0	0.0	0.0	0.0	**0.583**
EsbRs-SmoothAdv×3	0.766	0.698	0.600	0.506	0.0	0.0	0.0	0.0	0.0	0.576
EsbRs-AdvMacer ×1+SmoothAdv×2	0.772	0.672	0.594	0.498	0.0	0.0	0.0	0.0	0.0	0.572
EsbRs-AdvMacer ×2+MACER×1	0.798	0.700	0.586	0.472	0.0	0.0	0.0	0.0	0.0	0.568
EsbRs-AdvMacer ×1+MACER×2	**0.804**	**0.714**	0.598	0.462	0.0	0.0	0.0	0.0	0.0	0.570

Table 4.7. Certified accuracy: certified accuracy and ACR of the first 500 test images of Cifar-10 with $\sigma = 0.50$. Each column represents the robust accuracy that can be certified at this ℓ_2 radius.

Model ($\sigma = 0.50$)	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	ACR
SmoothAdv	0.564	0.516	0.468	0.432	0.394	0.328	**0.286**	**0.224**	0.0	0.735
MACER	**0.634**	**0.566**	0.476	0.432	0.346	0.258	0.206	0.126	0.0	0.682
SmoothMix	0.638	0.548	0.48	0.416	0.34	0.274	0.21	0.152	0.0	0.685
SmoothMix+1-step Adv	0.62	0.576	0.504	0.444	0.38	0.314	0.25	0.196	0.0	0.743
AdvMacer	0.584	0.532	**0.486**	**0.442**	**0.398**	**0.334**	0.270	0.216	0.0	**0.742**
EsbRs-AdvMacer ×3	0.588	0.544	0.498	0.448	0.414	0.360	0.288	0.230	0.0	0.772
EsbRs-SmoothAdv×3	0.584	0.530	0.476	**0.454**	0.420	0.362	0.308	**0.254**	0.0	0.777
EsbRs-AdvMacer ×1+SmoothAdv×2	0.594	0.540	0.482	**0.454**	**0.422**	**0.374**	**0.310**	0.238	0.0	**0.783**
EsbRs-AdvMacer ×2+MACER×1	0.636	0.564	0.504	0.446	0.388	0.294	0.240	0.172	0.0	0.728
EsbRs-AdvMacer ×1+MACER×2	**0.650**	**0.568**	**0.506**	0.450	0.370	0.292	0.220	0.158	0.0	0.723

Table 4.8. Certified accuracy: certified accuracy and ACR of the first 500 test images of Cifar-10 with $\sigma = 1.00$. Each column represents the robust accuracy that can be certified at this ℓ_2 radius.

Model ($\sigma = 1.00$)	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	ACR
SmoothAdv	0.458	0.418	0.374	0.312	0.288	0.254	0.234	0.196	0.180	0.158	0.758
MACER	0.424	0.392	0.354	0.328	**0.304**	**0.274**	**0.250**	**0.212**	0.184	0.156	0.768
SmoothMix	0.484	0.42	0.348	0.292	0.244	0.22	0.178	0.15	0.124	0.096	0.626
SmoothMix+1-step Adv	0.43	0.402	0.364	0.332	0.298	0.266	0.248	0.22	0.19	0.172	0.788
AdvMacer	**0.476**	**0.440**	**0.392**	**0.350**	0.302	**0.274**	0.236	**0.212**	**0.186**	**0.164**	**0.794**
EsbRs-AdvMacer ×3	**0.476**	0.426	0.386	0.358	0.302	0.270	0.246	0.220	0.196	0.172	0.805
EsbRs-SmoothAdv×3	0.466	**0.432**	0.388	**0.360**	0.294	0.260	0.240	0.212	0.186	0.170	0.801
EsbRs-AdvMacer ×1+SmoothAdv×2	0.472	**0.432**	**0.394**	0.356	0.304	0.262	0.240	0.214	0.194	**0.174**	**0.810**
EsbRs-AdvMacer ×2+MACER×1	0.428	0.404	0.370	0.342	**0.318**	0.276	**0.256**	**0.230**	**0.198**	0.164	0.801
EsbRs-AdvMacer ×1+MACER×2	0.440	0.404	0.366	0.340	0.304	**0.282**	0.238	0.202	0.184	0.154	0.760

Algorithm 6. Optimal weights of ensemble with 2 models. The function **ComputeWeight** will return the optimal weights

```
**Input:** two base models $F^1, F^2$, number of Gaussian noise $n$, number of perturbed models $m$, noise level $\sigma$, proportion of the parameters to perturb $t$, standard deviation of perturbation on parameters $\tilde{\sigma}$, query point $x$, target label $y$
**function** PerturbModel($F^1, F^2, m, t, \tilde{\sigma}$):
    **for** $l = 1, 2$ **do**
        **for** each $j = 1, \dots, m$ **do**
            **for** each parameter $\theta$ in $F^l$ **do**
                Draw a Bernoulli variable $X$ from Bernoulli($t$)
                **if** $X = 1$ **then**
                    Draw $\delta \sim \mathcal{N}(0, \tilde{\sigma}^2)$ and update $\theta \leftarrow \theta + \delta$
                **end if**
            **end for**
            Store perturbed model $\hat{F}^l_j$
        **end for**
    **end for**
    **Output:** perturbed models $\hat{F}^1_1, \dots, F^1_m$ and $\hat{F}^2_1, \dots, \hat{F}^2_m$.

**function** Estimation($[\hat{F}^1_1, \dots, F^1_m]$, $[\hat{F}^2_1, \dots, \hat{F}^2_m]$, $\sigma, n, x, y$):
    Draw $n$ i.i.d. noisy samples from $\mathcal{N}(x, \sigma^2 I)$ and denote them by $x_1, \dots, x_n$
    **for  do**$l = 1, 2$
        **for** $i = 1, \dots, m$ **do**
            **for** $j = 1, \dots, n$ **do**
                Compute $z^{l,(i-1)n+j} \leftarrow \hat{F}^l_i(x_j)_y \mathbf{1} - \hat{F}^l_i(x_j)$, where $\mathbf{1} = [1, 1, \dots, 1]^\top$ is the vector of all 1's.
            **end for**
        **end for**
    **end for**
    **Output:** estimates of logits $z^{l,j}$ for $l = 1, 2$ and $j = 1, \dots, mn$
```

```
**function** ComputeWeight($F^1, F^2, \sigma, \tilde{\sigma}, n, m, t, x, y$):
    1) $[\hat{F}^1_1, \dots, F^1_m]$, $[\hat{F}^2_1, \dots, \hat{F}^2_m] \leftarrow$ PerturbModel($F^1, F^2, m, t, \tilde{\sigma}$)
    2) $z^{l,j} \leftarrow$ Estimation($[\hat{F}^1_1, \dots, F^1_m]$, $[\hat{F}^2_1, \dots, \hat{F}^2_m]$, $\sigma, n, x, y$) for $l = 1, 2$ and $j = 1, \dots, m$
    3) Compute $\bar{z}^l \leftarrow \frac{1}{mn} \sum_{j=1}^{mn} z^{l,j}$ for $l = 1, 2$ and $a_i \leftarrow \min\{\bar{z}^1_i, \bar{z}^2_i\}^{-2}$ for $i = 1, \dots, c$
    4) Compute
        $$b_i \leftarrow \frac{1}{mn} \sum_{j=1}^{mn} (z^{1,j} - \bar{z}^1)(z^{1,j} - \bar{z}^1)^\top_{ii},$$
        $$c_i \leftarrow \frac{2}{mn} \sum_{j=1}^{mn} (z^{1,j} - \bar{z}^1)(z^{2,j} - \bar{z}^2)^\top_{ii},$$
        $$d_i \leftarrow \frac{1}{mn} \sum_{j=1}^{mn} (z^{2,j} - \bar{z}^2)(z^{2,j} - \bar{z}^2)^\top_{ii}.$$
    5) Compute
        $$A = \sum_{i=2}^{c} a_i(b_i + c_i + d_i), \quad B = \sum_{i=2}^{c} -a_i(c_i + 2d_i), \quad C = \sum_{i=2}^{c} a_i d_i.$$
    **if** $A > 0$ and $0 \leq -\frac{B}{2A} \leq 1$ **then**
        $w_1 \leftarrow -\frac{B}{2A}$ and $w_2 \leftarrow 1 + \frac{B}{2A}$
    **else**
        $w_1 \leftarrow 0, w_2 \leftarrow 1$ if $A + B > 0$ else $w_1 \leftarrow 1, w_2 \leftarrow 0$
    **end if**
    **Output:** $w_1, w_2$
```

Table 4.9. SVHN: clean accuracy and ACR of different models evaluated at the first 500 test images of SVHN with varing σ.

σ	Model	Accuracy	ACR	Training Time
0.25	SmoothAdv	85.8%	0.560	11.4h
	MACER	86.8%	0.549	48.5h
	AdvMacer	86.6%	**0.569**	11.4h
	EsbRs-SmoothAdv×3	87.8%	0.578	NA
	EsbRs-AdvMacer ×3	88.2%	**0.582**	NA
	EsbRs-AdvMacer ×1+MACER×2	87.8%	0.559	NA
	EsbRs-AdvMacer ×2+MACER×1	88.6%	0.570	NA
	EsbRs-AdvMacer ×1+SmoothAdv×2	87.8%	0.577	NA
	EsbRs-AdvMacer ×2+SmoothAdv×1	87.6%	**0.582**	NA
0.50	SmoothAdv	71.2%	0.552	11.4h
	MACER	58.4%	0.535	48.5h
	AdvMacer	67.8%	**0.572**	11.4h
	EsbRs-SmoothAdv×3	71.2%	0.573	NA
	EsbRs-AdvMacer ×3	70.4%	**0.588**	NA
	EsbRs-AdvMacer ×1+MACER×2	62.8%	0.551	NA
	EsbRs-AdvMacer ×2+MACER×1	66.0%	0.564	NA
	EsbRs-AdvMacer ×1+SmoothAdv×2	71.8%	0.577	NA
	EsbRs-AdvMacer ×2+SmoothAdv×1	71.2%	0.583	NA

the robustness by maximizing the certified radius over adversarial example, and our second approach **EsbRs** can further improve **AdvMacer** on both clean accuracy and robustness certificate. We show that we could improve ACR by 15% compared with MACER and 8% compared with the best models of SmoothAdv Moreover, we provided a general theoretical analysis for **EsbRs** and develop a theoretical-grounded methodology to design optimal ensemble scheme, which outperforms prior works.

Chapter 4, in part, has been submitted for publication of the material ”Promoting Robustness of Randomized Smoothing: Two Cost-Effective Approaches”, Liu, Linbo, Trong, Hoang, Nguyen, Lam, and Weng, Tsui-Wei to *Computer Vision and Pattern Recognition Conference* and is currently under review. The dissertation author was the primary investigator and author of this paper.

Table 4.10. Cifar-10: ACR of different models on the first 500 test images of Cifar-10 with varing σ. Clean accuracy is reported in parenthesis. Reported models include SmoothAdv, MACER, **AdvMacer** , **EsbRs**. $N = 100,000$ samples are used in certification unless otherwise specified.

	Methods	$\sigma = 0.25$	$\sigma = 0.5$	$\sigma = 1.0$	Ensemble?
Baselines	SmoothAdv	0.541 (74.2%)	0.735 (56.4%)	0.758 (45.8%)	×
	MACER	0.518 (79.4%)	0.682 (63.4%)	0.768 (42.4%)	×
Ours	**AdvMacer**	**0.554 (76.0%)**	**0.742 (58.4%)**	**0.794 (47.6%)**	×
	EsbRs-AdvMacer ×3	**0.583** (76.4%)	0.772 (58.8%)	0.805 (47.6%)	√
	EsbRs-SmoothAdv×3	0.567 (76.6%)	0.777 (58.4%)	0.801 (46.6%)	√
	EsbRs-AdvMacer ×1+SmoothAdv×2	0.572 (77.2%)	**0.783** (59.4%)	**0.810** (47.2%)	√
	EsbRs-AdvMacer ×2+MACER×1	0.568 (79.8%)	0.728 (63.6%)	0.801 (42.8%)	√
	EsbRs-AdvMacer ×1+MACER×2	0.570 (80.4%)	0.723 (65.0%)	0.760 (44.0%)	√
	EsbRs-AdvMacer ×3 with $N/3$	0.550 (76.4%)	0.736 (58.8%)	0.785 (47.4%)	√
	SmoothAdv with $3N$	0.568 (74.6%)	0.761 (56.8%)	0.771 (45.6%)	×

Table 4.11. ImageNet: ACR on 500 test images of ImageNet. Clean accuracy is reported in parenthesis.

	Methods	$\sigma = 0.25$	$\sigma = 0.5$	$\sigma = 1.0$	Time
Baselines	SmoothAdv	0.519 (61.5%)	0.801 (55.6%)	0.971 (41.4%)	48.4h
	MACER	0.438 (63.2%)	0.628 (52.6%)	0.634 (37.8%)	70h
Ours	**AdvMacer**	**0.537 (63.9%)**	**0.837 (56.2%)**	**0.989 (45.6%)**	48.4h

Chapter 5

Robust Estimation in Linear Regression with both Heavy-tailed Data and Noise

5.1 Introduction

In this big data era, tools for dealing with high dimensionality has been well-studied. The traditional least square technique is not even available in the setting where the dimension of data is only slightly larger than the sample size. The development of popular Lasso [130] provides an alternative to analyze the high-dimensional data, see also [15]. However, Lasso loses its robustness when the error comes from a heavy tail distribution, due to the nature of squared error loss which tends to over-penalize the sample with large error. Modern data collected from various scientific areas reveals a feature of heavy tail, hence more robust methods are to be proposed. [38] proposes to estimate the mean regression by regularized huber loss when the error term is lack of light tail assumption, provided that the underlying data still comes from a sub-Gaussian distribution. Shrinkage techniques are widely used when tacking heavy-tailed features, see, for example, [158].

In this chapter, we aim at estimate mean regression vector by shrinkage and regularized huber loss. To fix the idea, let's consider the following linear model,

$$y = X\beta^* + \varepsilon, \tag{5.1.1}$$

where the design matrix $X \in \mathbb{R}^{n\times p}$ has i.i.d rows x_i for $i = 1, \dots, n$, and the zero-mean noise $\varepsilon \in \mathbb{R}^p$ is independent of X. Throughout the chapter, we assume $\log p/n \to 0$.

The chapter is organized as follows. In section 2, we propose the method of achieving the robust estimation and present the main result of this chapter. All the technical proofs are left in section 3.

5.2 Methodology

5.2.1 Estimation

In order to robustify the estimation, we propose to consider the huber loss

$$\ell_\alpha(x) = \begin{cases} 2\alpha^{-1}|x| - \alpha^{-2}, & \text{if } |x| > \alpha^{-1} \\ x^2, & \text{if } |x| \le \alpha^{-1} \end{cases}$$

and the shrinkage for $\{x_i\}$ and $\{y_i\}$

$$\begin{aligned} \tilde{x}_{ij} &= \min\{|x_{ij}|, T\}\frac{x_{ij}}{|x_{ij}|}, \quad j = 1, \dots, p \\ \tilde{y}_i &= \min\{|y_i|, T\}\frac{y_i}{|y_i|}. \end{aligned}$$

We estimate β^* by solving the following minimization problem:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n}\sum_{i=1}^{n} \ell_\alpha(\tilde{y}_i - \tilde{x}_i^T\beta) + \lambda_n\|\beta\|_1 \tag{5.2.1}$$

The robustness is preserved by the property of huber loss and the nice behavior of the data at the tail is guaranteed by shrinkage operation.

5.2.2 Main result

We shall present the main conclusions in this section. Observe that the signal β^* can be expressed as

$$\beta^* = \arg\min_{\beta} \mathbb{E}[\|y - X\beta\|_2].$$

We decompose the error $\|\beta^* - \hat{\beta}\|_2$ into two terms,

$$\|\beta^* - \hat{\beta}\| \leq \|\beta^*_\alpha - \hat{\beta}\| + \|\beta^* - \beta^*_\alpha\|,$$

where $\beta^*_\alpha = \arg\min_\beta \mathbb{E}[\ell_\alpha(y - x^T\beta)]$, representing the minimizer of the population mean of huber loss. Throughout the chapter, we suppose the vector $\beta^* - \beta^*_\alpha$ lies in some bounded ball in $\mathbb{R}^p$. In the field of sparse recovery, we need to assume some sparsity structure on β^*_α as in many other popular literatures. Here, we suppose that $\text{supp}(\beta^*_\alpha) = S$ with $|S| \leq s$. The sparsity should satisfy $(s \log p)/n \to 0$. For the choice of parameter, we will select $\lambda_n \asymp \sqrt{\log p/n}$ and $T \asymp (n/\log p)^{1/4}$.

Next, we present the separate bounds for both terms.

Assumptions 5.2.1. For some $q > 2$,

(1) $\mathbb{E}[\varepsilon^{2q}] \leq M_{2q}$.

(2) $\mathbb{E}[(x^T v)^{2q}] = \mu$ with $\|v\|_2 = 1$.

(3) $0 \leq \kappa_1 \leq \lambda_{\min}(\mathbb{E}[xx^T]) \leq \lambda_{\max}(\mathbb{E}[xx^T]) \leq \kappa_2 < \infty$.

Theorem 5.2.2. *Under the assumption 5.2.1, it holds that*

$$\|\beta^* - \beta^*_\alpha\|_2 \lesssim \alpha^{q-1}.$$

Now, we aim to bound $\|\hat{\beta} - \beta^*_\alpha\|_2$.

Theorem 5.2.3. *Under assumption 5.2.1 and additionally assume that* $\mathbb{E}[x_{ij}^2 x_{ik}^2] \leq R^4$*, for any* $1 \leq j, k \leq p$*. Then*

$$\|\hat{\beta} - \beta^*_\alpha\|_2 \lesssim \sqrt{\frac{s \log p}{n}}$$

with probability at least $1 - 4p^{-c}$ *for some universal constant* $c > 0$*.*

Put theorem 5.2.2 and theorem 5.2.3 together, we have the following theorem establishing the convergence of our estimator $\hat{\beta}$.

Theorem 5.2.4. *Under the assumptions in theorem 5.2.2 and theorem 5.2.3, we have*

$$\|\hat{\beta} - \beta^*_\alpha\|_2 = O(\alpha^{q-1}) + O(\sqrt{\frac{s \log p}{n}})$$

with probability at least $1 - 4p^{-c}$ *for some universal constant* $c > 0$*.*

Remark 5.2.1. Our result is as sharp as the result in [38] except a multiplicative constant 2 in the power of α. In practice, we always choose α such that $\alpha^{q-1} = O(\sqrt{s \log p/n})$. Hence, by a smaller choice of α, we obtain exactly the same order $O(\sqrt{s \log p/n})$ on the upper bound as in [38], which means the shrinkage estimation is as optimal as the estimation of light-tailed data.

5.3 Proof

In this section, we mainly prove theorem 5.2.2 and theorem 5.2.3.

5.3.1 Proof of theorem 5.2.2

Proof of theorem 5.2.2. It follows from the proof of Theorem 1 in [38] that for some vector $\tilde{\beta}$ lying between β^* and β^*_α, it holds that

$$\kappa_1 \|\beta^*_\alpha - \beta^*\|_2^2 \leq 2(2\alpha)^{q-1} \mathbb{E}\left[\left(M_q^q + |x_i^T(\beta^* - \tilde{\beta})|^q\right) \left|x_i^T(\beta^*_\alpha - \beta^*)\right|\right]. \tag{5.3.1}$$

By hypothesis, we see that

$$\mathbb{E}\left[\left(M_q^q + |x_i^T(\beta^* - \tilde{\beta})|^q\right)^2\right] = O(1).$$

By Cauchy-Schwartz inequality, it follows from (5.3.1) that

$$\kappa_1 \|\beta_\alpha^* - \beta^*\|_2^2 \leq O(\alpha^{q-1}\|\beta_\alpha^* - \beta^*\|_2).$$

Dividing both sides by $\|\beta_\alpha^* - \beta^*\|_2$ completes the proof. □

5.3.2 Proof of theorem 5.2.3

Before we proceed to the proof, it's useful to introduce some notations. Let

$$\begin{aligned} L_n(\beta) &= \frac{1}{n}\sum_{i=1}^n \ell_\alpha(y_i - x_i^T\beta) \\ \tilde{L}_n(\beta) &= \frac{1}{n}\sum_{i=1}^n \ell_\alpha(\tilde{y}_i - \tilde{x}_i^T\beta) \end{aligned}$$

be the average huber loss for the original and shrinkage data respectively. We begin the proof with a preliminary lemma, which explains the choice of λ_n and T.

Lemma 5.3.1. *Choose $\lambda_n \asymp \sqrt{\frac{\log p}{n}}$ and $T \asymp \left(\frac{n}{\log p}\right)^{1/4}$, then it holds that*

$$\hat{\Delta} = \hat{\beta} - \beta_\alpha^* \in C(S) = \{\Delta \in \mathbb{R}^p : \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1\}$$

with probability at least $1 - 2p^{-c_0}$ for some universal constant c_0.

Proof of lemma 5.3.1. By the lemma 1 in [102], $\hat{\Delta} \in C(S)$ whenever $\lambda_n \geq 2\|\nabla \tilde{L}_n(\beta_\alpha^*)\|_\infty$. Therefore, it suffices to show that with the choice of λ_n, the condition that $\lambda_n \geq 2\|\nabla \tilde{L}_n(\beta_\alpha^*)\|_\infty$ holds with high probability.

Direct calculation gives us

$$\nabla \tilde{L}_n(\beta^*_\alpha) = \frac{1}{n}\sum_{i=1}^{n} \frac{2}{\alpha}\psi(\alpha(\tilde{y}_i - \tilde{x}_i^T\beta^*_\alpha))\tilde{x}_i,$$

where $\psi(x) = x$, if $|x| \leq 1$; $\psi(x) = 1$, if $x > 1$; and $\psi(x) = -1$, if $x < -1$. Notice that

$$\frac{1}{\alpha}|\psi(\alpha x)| \leq |x|, \tag{5.3.2}$$

therefore we have

$$\mathrm{Var}\left(\frac{2}{\alpha}\psi(\alpha(\tilde{y}_i - \tilde{x}_i^T\beta^*_\alpha))\tilde{x}_{ij})\right) \leq \mathbb{E}\left[4(|\tilde{y}_i - \tilde{x}_i^T\beta^*_\alpha|\tilde{x}_{ij}))^2\right] \leq 8\mathbb{E}[(\tilde{y}_i\tilde{x}_{ij})^2] + 8\mathbb{E}[(\tilde{x}_i^T\beta^*_\alpha\tilde{x}_{ij})^2]$$

By the hypothesis, one has $\mathbb{E}[\tilde{y}_i^4] \leq \mathbb{E}[y_i^4] < \infty$, and hence by Cauchy-Schwartz inequality $8\mathbb{E}[(\tilde{y}_i\tilde{x}_{ij})^2] \leq L_1$ for some $L_1 > 0$. Consider

$$\begin{aligned}\mathbb{E}[(\tilde{x}_i^T\beta^*_\alpha\tilde{x}_{ij})^2] &\leq \sqrt{\mathbb{E}[(\tilde{x}_i^T\beta^*)^4]\mathbb{E}[\tilde{x}_{ij}^4]} \leq \sqrt{\mathbb{E}[(x_i - (x_i - \tilde{x}_i))^T\beta^*_\alpha]^4}R^2 \\ &\leq 4\sqrt{\mathbb{E}[(x_i^T\beta^*_\alpha)^4] + \mathbb{E}([(x_i - \tilde{x}_i)^T\beta^*_\alpha]^4)}R^2 \\ &\leq L_2,\end{aligned}$$

since we assume that β^*_α lies in a bounded ℓ_2 ball. Combining the above displays delivers

$$\mathrm{Var}\left(\frac{2}{\alpha}\psi(\alpha(\tilde{y}_i - \tilde{x}_i^T\beta^*_\alpha))\tilde{x}_{ij})\right) \leq v,$$

for some finite $v > 0$. Moreover, $|\frac{2}{\alpha}\psi(\alpha(\tilde{y}_i - \tilde{x}_i^T\beta^*_\alpha))\tilde{x}_{ij})| \leq C_1T^2$ for some constant $C_1 > 0$. By Bernstein inequality, taking $T = \xi(n/\log p)^{1/4}$, we obtain

$$\mathbb{P}\left(\nabla\tilde{L}_{nj}(\beta^*_\alpha) - \mathbb{E}[\nabla\tilde{L}_n(\beta^*_\alpha)]_j \geq C_2\sqrt{\frac{\log p}{n}}\right) \leq 2p^{-c'},$$

for some constant $c' > 1$ and $C_2 > 0$. Now we only need to bound $\mathbb{E}[\nabla\tilde{L}_n(\beta^*_\alpha)]_j$. From the optimality condition of β^*_α, we see that $\mathbb{E}[\nabla L_n(\beta^*_\alpha)] = 0$, and hence

$$\mathbb{E}[\nabla\tilde{L}_n] = \mathbb{E}[\nabla\tilde{L}_n - \nabla L_n] + \mathbb{E}[\nabla L_n] \tag{5.3.3}$$

Repeatedly use the observation (5.3.2) and one has $\mathbb{E}[\nabla\tilde{L}_n - \nabla L_n] \asymp \sqrt{\log p/n}$. Therefore by (5.3.3)

$$\mathbb{E}[\nabla\tilde{L}_n(\beta^*_\alpha)] \asymp \sqrt{\frac{\log p}{n}}.$$

Finally, we obtain

$$\mathbb{P}\left(\nabla\tilde{L}_n j(\beta^*_\alpha) \geq C_3\sqrt{\frac{\log p}{n}}\right) \leq 2p^{-c'},$$

for some constant $C_3 > 0$. Taking $\lambda_n = 2C_3\sqrt{\log p/n}$ and a union bound yields

$$\mathbb{P}\left(2\|\nabla\tilde{L}_n(\beta^*_\alpha)\|_\infty \geq \lambda_n\right) \leq 2p^{-(c'-1)}.$$

□

Let's introduce a restricted strong convexity condition, which turns out to be crucial in the proof of theorem 5.2.3. Denote by $\delta L_n(\beta + \Delta, \beta)$ the Taylor remainder if we use first order Taylor expansion to approximate $L_n(\beta + \Delta)$, *i.e.*

$$\delta L_n(\beta + \Delta, \beta) = L_n(\beta + \Delta) - L_n(\beta) - \nabla L_n(\beta)^T\Delta.$$

Now we are ready to state the definition of restricted strong convexity (RSC).

Definition 5.3.1 (Restricted Strong Convexity)**.** The loss function $L_n(\beta)$ satisfy the restricted strong convexity on a set C with curvature κ_L and tolerance τ_L if

$$\delta L_n(\beta + \Delta, \beta) \geq \kappa_L\|\Delta\|_2^2 - \tau_L^2, \quad \text{for all } \Delta \in C.$$

The following lemma establish the RSC condition of $\tilde{L}_n(\beta^*)$ on $C(S) \cap B_2(t)$ with high probability.

Lemma 5.3.2. *The RSC condition holds for $\tilde{L}_n(\beta^*_\alpha)$ on $C(S) \cap B_2(t)$, i.e.*

$$\delta \tilde{L}_n(\beta^*_\alpha + \Delta, \beta^*_\alpha) \geq \kappa \|\Delta\|_2^2 - C_0 t^2 (\sqrt{\frac{\log p}{n}} + \sqrt{\frac{s \log p}{n}}), \quad \forall \, \Delta \in C(S) \cap B_2(t),$$

with probability at least $1 - 2p^{-c_1}$.

Proof of lemma 5.3.2. By Taylor expansion,

$$\delta \tilde{L}_n(\beta^*_\alpha + \Delta, \beta^*_\alpha) = \frac{1}{n} \sum_{i=1}^{n} \psi'(\alpha(\tilde{y}_i - \tilde{x}_i^T \beta^*_\alpha + v \tilde{x}_i^T \beta^*_\alpha))(\tilde{x}_i^T \Delta)^2,$$

where $\psi'(x) = 1$, if $|x| \leq 1$; $\psi'(x) = 0$, if $|x| > 1$. Note that $\psi'(x)$ is not Lipschitz continuous. In order to use the Ledoux Talagrand contraction theorem ([78]), we need to truncate $\psi'(x)$ from below.

Define the truncation function $\phi_m(u)$ by $\phi_m(u) = u^2 \mathbb{I}(|u| < \frac{m}{2}) + (m - u)^2 \mathbb{I}(\frac{m}{2} \leq |u| \leq m)$. Note that $\phi_m(u)$ is bounded by $m^2/4$ with Lipschitz constant at most $2m$. First we claim that

$$\delta \tilde{L}_n(\beta^*_\alpha + \Delta, \beta^*_\alpha) \geq \frac{1}{n} \sum_{i=1}^{n} \phi_{tT_1}(\tilde{x}_i^T \Delta \mathbb{I}(|\tilde{y}_i - \tilde{x}_i^T \beta^*_\alpha| \leq T_2)),$$

for $0 < \alpha \leq 1/(tT_1 + T_2)$, where the thresholds T_1 and T_2 are to be determined later. This result was proved in the proof of Lemma 2 in [38]. Now it suffices to show that

$$\frac{1}{n} \sum_{i=1}^{n} \phi_{tT_1}(\tilde{x}_i^T \Delta \mathbb{I}(|\tilde{y}_i - \tilde{x}_i^T \beta^*_\alpha| \leq T_2)) \geq \kappa \|\Delta\|_2^2 - C_0 t^2 (\sqrt{\frac{\log p}{n}} + \sqrt{\frac{s \log p}{n}}),$$

$$\forall \, \Delta \in C(S) \cap B_2(t),$$

with high probability. We will finish the proof by two steps:

(a) $\mathbb{E}[\phi_{tT_1}(\tilde{x}_i^T\Delta\mathbb{I}(|\tilde{y}_i - \tilde{x}_i^T\beta_\alpha^*| \le T_2))] \ge \frac{\kappa_1}{4}\|\Delta\|_2$.

(b) Define $W(t) = \sup_{\Delta\in C(S)\cap B_2(t)} \frac{1}{n}|\sum\phi_{tT_1} - \mathbb{E}\phi_{tT_1}|$, then

$$W(t) \le C_0t^2(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{s\log p}{n}}) \tag{5.3.4}$$

with high probability

In order to show (a), we observe that for any $1 \le j, k \le p$,

$$\mathbb{E}[|x_{ij}x_{ik} - \tilde{x}_{ij}\tilde{x}_{ik}|] \le \sqrt{\mathbb{E}(x_{ij}x_{ik})^2(\mathbb{P}(|x_{ij}| \ge T) + \mathbb{P}(|x_{ik}| \ge T))} \le \frac{\sqrt{2}R^3}{T^2}.$$

Furthermore, one has $\|\mathbb{E}[|x_{ij}x_{ik} - \tilde{x}_{ij}\tilde{x}_{ik}|]\|_\infty \le \sqrt{2}R^3/T^2$. For notation simplicity, let the event $A_i = \{|\tilde{x}_i^T\Delta| \le tT_1/2\}$ and $B_i = \{|\tilde{y}_i - \tilde{x}_i^T\beta_\alpha^*| \le T_2\}$, and drop the subscript i if there is no ambiguity. Then it's easy to see that

$$\begin{aligned}\mathbb{E}[\phi_{tT_1}(\tilde{x}_i^T\Delta\mathbb{I}(|\tilde{y}_i - \tilde{x}_i^T\beta_\alpha^*| \le T_2))] &\ge \mathbb{E}[(\tilde{x}^T\Delta)\mathbb{I}_{A\cap B}] \\ &\ge \Delta^T\mathbb{E}[xx^T\mathbb{I}_{A\cap B}]\Delta - \Delta^T\mathbb{E}[(xx^T - \tilde{x}\tilde{x}^T)]\Delta \\ &\ge \Delta^T\mathbb{E}[xx^T(1 - \mathbb{I}_{A^c\cup B^c})]\Delta - \frac{\sqrt{2}R^3}{T^2}\|\Delta\|_1^2 \\ &\ge \Delta^T\mathbb{E}[xx^T]\Delta - \mu_4^2\sqrt{\mathbb{P}(A^c) + \mathbb{P}(B^c)}\|\Delta\|_2^2 - C_1s\sqrt{\frac{\log p}{n}}\|\Delta\|_2^2.\end{aligned}$$

It can be shown that

$$\begin{aligned}\mathbb{P}(A^c) \le \frac{4\mathbb{E}(\tilde{x}^T\Delta)^2}{t^2T_1^2} &\le \frac{4\left[\mathbb{E}(x^T\Delta)^2 + \Delta^T\mathbb{E}[\tilde{x}\tilde{x}^T - xx^T]\Delta\right]}{t^2T_1^2} \\ &\le \frac{4\left[\mu_2^2\|\Delta\|_2^2 + C_1s\sqrt{\log p/n}\|\Delta\|_2^2\right]}{t^2T_1^2} \le \frac{C_2}{T_1^2}.\end{aligned}$$

and similarly that $\mathbb{P}(B^c) \le C_3/T_2^2$. Choose sufficiently large T_1 and T_2 of constant order, we have

$$\mathbb{E}[\phi_{tT_1}(\tilde{x}_i^T\Delta\mathbb{I}(|\tilde{y}_i-\tilde{x}_i^T\beta_\alpha^*|\le T_2))]\ge \kappa_1\|\Delta\|_2^2-\frac{\kappa_1}{2}\|\Delta\|_2^2-C_1s\sqrt{\frac{\log p}{n}}\|\Delta\|_2^2\ge\frac{\kappa_1}{4}\|\Delta\|_2^2.$$

Next, we shall finish step (b). Since $W(t)$ is a bounded random variable, by Massart inequality one has for any $w>0$,

$$\mathbb{P}\left(|W(t)-\mathbb{E}W(t)|\ge t^2T_1^2\sqrt{\frac{w}{n}}\right)\le 2\mathrm{e}^{-w/8}. \tag{5.3.5}$$

In order to bound $\mathbb{E}W(t)$, we use symmetrization argument and bound it by Rademacher complexity.

$$\mathbb{E}W(t)\le 2\mathbb{E}\left[\sup_{\Delta}|\frac{1}{n}\sum_{i=1}^n\gamma_i\phi_{tT_1}(\tilde{x}_i^T\Delta\mathbb{I}_{B_i})|\right], \tag{5.3.6}$$

where γ_i are i.i.d. Rademacher variables. By Ledoux-Taragrand contraction theorem ([78]), (5.3.6) can be further bounded by

$$\begin{aligned}\mathbb{E}W(t)&\le 8tT_1\mathbb{E}\left[\sup_{\Delta\in C(S)\cap B_2(t)}|\frac{1}{n}\sum\gamma_i\tilde{x}_i^T\Delta\mathbb{I}_{B_i}|\right]\\&\le 32t^2T_1\sqrt{s}\mathbb{E}\left[\|\frac{1}{n}\sum_{i=1}^n\gamma_i\tilde{x}_i\|_\infty\right]\\&\le C_4t^2T_1\sqrt{\frac{s\log p}{n}},\end{aligned} \tag{5.3.7}$$

Putting together (5.3.5) and (5.3.7) and taking $w=c_1\log p$ for some constant c_1, we get the desired result (5.3.4) with probability at least $1-2p^{-c_1}$. □

With these preparatory lemmas, we are ready to prove theorem 5.2.3.

Proof of theorem 5.2.3. We define an intermediate estimate by

$$\hat{\beta}_\eta = \beta_\alpha^* + \eta(\hat{\beta} - \beta_\alpha^*),$$

where

$$\eta = \begin{cases} 1, & \text{if } \|\hat{\beta} - \beta_\alpha^*\|_2 \leq t, \\ t/\|\hat{\beta} - \beta_\alpha^*\|_2, & \text{if } \|\hat{\beta} - \beta_\alpha^*\|_2 > t. \end{cases}$$

Let $\hat{\Delta}_\eta = \hat{\beta}_\eta - \beta_\alpha^*$, thus $\|\hat{\Delta}_\eta\|_2 \leq t$ by construction. From the optimality of $\hat{\beta}$ and the convexity of $\tilde{L}_n(\beta) + \lambda_n\|\beta\|_1$, one has

$$\tilde{L}_n(\hat{\beta}_\eta) + \lambda_n\|\hat{\beta}_\eta\|_1 \leq \tilde{L}_n(\beta_\alpha^*) + \lambda_n\|\beta_\alpha^*\|_1.$$

Furthermore,

$$\delta\tilde{L}_n(\hat{\beta}_\eta, \beta_\alpha^*) \leq \lambda_n(\|\beta_\alpha^*\|_1 - \|\hat{\beta}_\eta\|_1) - \langle\nabla\tilde{L}_n(\beta_\alpha^*), \hat{\Delta}_\eta\rangle \tag{5.3.8}$$

Denote by E and F the events in lemma 5.3.1 and lemma 5.3.2 respectively. By the Lemma 1 in [102], $\hat{\Delta}_\eta \in C(S)$ whenever $\lambda_n \geq 2\|\nabla\tilde{L}_n(\beta_\alpha^*)\|_\infty$. Therefore, conditioned on the events E and F, we have

$$\begin{aligned}\kappa\|\hat{\Delta}_\eta\|_2^2 - C_0t^2(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{s\log p}{n}}) \leq \delta\tilde{L}_n(\hat{\beta}_\eta, \beta_\alpha^*) &\leq \lambda_n(\|\beta_\alpha^*\|_1 - \|\hat{\beta}_\eta\|_1) - \langle\nabla\tilde{L}_n(\beta_\alpha^*), \hat{\Delta}_\eta\rangle \\ &\leq \lambda_n\|\hat{\Delta}_\eta\|_1 + \|\nabla\tilde{L}_n(\beta_\alpha^*)\|_\infty\|\hat{\Delta}_\eta\|_1 \\ &\leq C\sqrt{\frac{s\log p}{n}}\|\hat{\Delta}_\eta\|_2\end{aligned}$$

Some algebra shows that

$$\|\hat{\Delta}_\eta\|_2 \leq C_1\sqrt{\frac{s\log p}{n}} + tC_2\left(\frac{s\log p}{n}\right)^{\frac{1}{4}} \tag{5.3.9}$$

Choose $t = 2C_1\sqrt{\frac{s\log p}{n}}$. For sufficiently large n and p, we obtain

$$\|\hat{\Delta}_\eta\|_2 < 2C_1\sqrt{\frac{s\log p}{n}} = t$$

By the construction of $\hat{\beta}_\eta$, we see that $\|\hat{\Delta}_\eta\|_2 < t$ happens only if $\hat{\Delta}_\eta = \hat{\Delta}$. Finally we complete the proof by noting that $\mathbb{P}(E \cap F) = 1 - \mathbb{P}(E^c \cup F^c) \geq 1 - 8p^{-c'}$, where $c' = \min\{c_0, c_1\}$. □

Chapter 5, in full, is a research project of the material ”Robust Estimation in Linear Regression with both Heavy-tailed Data and Noise”, Liu, Linbo and will be further enhanced for submission. The dissertation author was the primary investigator and author of this project.

Bibliography

[1] Radoslaw Adamczak et al. A tail inequality for suprema of unbounded empirical processes with applications to markov chains. *Electronic Journal of Probability*, 13:1000–1034, 2008.

[2] Andreas Alfons, Christophe Croux, and Sarah Gelper. Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *Annals of Applied Statistics*, 7(1):226–248, 2013.

[3] Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020.

[4] Torben G Andersen, Tim Bollerslev, Peter Christoffersen, and Francis X Diebold. Volatility forecasting, 2005.

[5] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[6] Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

[7] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pages 274–283. PMLR, 2018.

[8] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[9] Sumanta Basu and George Michailidis. Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567, 2015.

[10] Sergei Bernstein. *The Theory of Probabilities.* Gastehizdat Publishing House, Moscow, 1946.

[11] Monika Bhattacharjee and Arup Bose. Estimation of autocovariance matrices for infinite dimensional vector linear process. *Journal of Time Series Analysis*, 35(3):262–281, 2014.

[12] Monika Bhattacharjee and Arup Bose. Large sample behaviour of high dimensional autocovariance matrices. *The Annals of Statistics*, 44(2):598–628, 2016.

[13] Peter J Bickel. One-step huber estimates in the linear model. *Journal of the American Statistical Association*, 70(350):428–434, 1975.

[14] Peter J Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of statistics*, 36(6):2577–2604, 2008.

[15] Peter J Bickel, Ya'acov Ritov, Alexandre B Tsybakov, et al. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[16] Joos-Hendrik Böse, Valentin Flunkert, Jan Gasthaus, Tim Januschowski, Dustin Lange, David Salinas, Sebastian Schelter, Matthias Seeger, and Yuyang Wang. Probabilistic demand forecasting at scale. *Proceedings of the VLDB Endowment*, 10(12):1694–1705, 2017.

[17] Peter J Brockwell and Richard A Davis. *Time series: theory and methods*. Springer Science & Business Media, 2009.

[18] Robert G Brown. Exponential smoothing for predicting demand. In *Operations Research*, volume 5, pages 145–145. INST OPERATIONS RESEARCH MANAGEMENT SCIENCES 901 ELKRIDGE LANDING RD, STE . . . , 1957.

[19] P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[20] Donald L Burkholder. Distribution function inequalities for martingales. *the Annals of Probability*, pages 19–42, 1973.

[21] T Tony Cai and Harrison H Zhou. Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics*, 40(5):2389–2420, 2012.

[22] T.T. Cai, W. Liu, and X. Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

[23] Emmanuel Candes, Terence Tao, et al. The dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics*, 35(6):2313–2351, 2007.

[24] L. Chen and W.B. Wu. Concentration inequalities for empirical processes of linear time series. *Journal of Machine Learning Research*, 18(231):1–46, 2018.

[25] Xiaohui Chen, Mengyu Xu, and Wei Biao Wu. Regularized estimation of linear functionals of precision matrices for high-dimensional time series. *IEEE Transactions on Signal Processing*, 64(24):6459–6470, 2016.

[26] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Comparison and anti-concentration bounds for maxima of gaussian random vectors. *Probability Theory and Related Fields*, 162(1):47–70, 2015.

[27] Victor Chernozhukov, Denis Chetverikov, Kengo Kato, et al. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819, 2013.

[28] Victor Chernozhukov, Wolfgang Karl Härdle, Chen Huang, and Weining Wang. Lasso-driven inference in time and space. *The Annals of Statistics*, 49(3):1702–1735, 2021.

[29] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019.

[30] J.T. Connor, R.D. Martin, and L.E. Atlas. Recurrent neural networks and robust time series prediction. *IEEE Transactions on Neural Networks*, 5(2):240–254, 1994.

[31] R Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236, 2001.

[32] Francesco Croce and Matthias Hein. Sparse and imperceivable adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4724–4732, 2019.

[33] Hanjun Dai, Hui Li, Tian Tian, Xin Huang, Lin Wang, Jun Zhu, and Le Song. Adversarial attack on graph structured data. In *International conference on machine learning*, pages 1115–1124. PMLR, 2018.

[34] Raphaël Dang-Nhu, Gagandeep Singh, Pavol Bielik, and Martin Vechev. Adversarial attacks on probabilistic autoregressive forecasting models. In *International Conference on Machine Learning*, pages 2356–2365. PMLR, 2020.

[35] Emmanuel de Bézenac, Syama Sundar Rangapuram, Konstantinos Benidis, Michael Bohlke-Schneider, Richard Kurle, Lorenzo Stella, Hilaf Hasson, Patrick Gallinari, and Tim Januschowski. Normalizing kalman filters for multivariate time series analysis. *Advances in Neural Information Processing Systems*, 33:2995–3007, 2020.

[36] Paul Doukhan and Sana Louhichi. A new weak dependence condition and applications to moment inequalities. *Stochastic processes and their applications*, 84(2):313–342, 1999.

[37] J. Fan, J. Lv, and L. Qi. Sparse high dimensional models in economics. *Annual review of economics*, 3:291–317, 2011.

[38] Jianqing Fan, Quefeng Li, and Yuyan Wang. Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):247–265, 2017.

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

[40] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[41] Karl J Friston. Functional and effective connectivity: a review. *Brain connectivity*, 1(1):13–36, 2011.

[42] Jan Gasthaus, Konstantinos Benidis, Yuyang Wang, Syama Sundar Rangapuram, David Salinas, Valentin Flunkert, and Tim Januschowski. Probabilistic forecasting with spline quantile function rnns. In *The 22nd international conference on artificial intelligence and statistics*, pages 1901–1910. PMLR, 2019.

[43] Sarah Gelper, Roland Fried, and Christophe Croux. Robust forecasting with exponential and Holt–Winters smoothing. *Journal of Forecasting*, 29(3):285–300, 2010.

[44] Gene H. Golub and Charles F. Van Loan. *Matrix computations, 4th edition.* Johns Hopkins University Press, 2013.

[45] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

[46] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[47] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[48] C. Gourieroux and J. Jasiak. *Financial Econometrics: Problems, Models, and Methods*. Princeton University Press, 2011.

[49] Shaojun Guo, Yazhen Wang, and Qiwei Yao. High-dimensional and banded vector autoregressions. *Biometrika*, 103(4):889–903, 2016.

[50] Shuva Gupta. A note on the asymptotic distribution of lasso estimator for correlated data. *Sankhya A*, 74(1):10–28, 2012.

[51] David Hallac, Youngsuk Park, Stephen Boyd, and Jure Leskovec. Network inference via the time-varying graphical lasso. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 205–213, 2017.

[52] Fang Han, Huanran Lu, and Han Liu. A direct estimation of high dimensional stationary vector autoregressions. *Journal of Machine Learning Research*, 16:3115–3150, 2015.

[53] Hanyuan Hang and Ingo Steinwart. A bernstein-type inequality for some mixing processes and dynamical systems with an application to learning. *The Annals of Statistics*, 45(2):708–743, 2017.

[54] Samuel Harford, Fazle Karim, and Houshang Darabi. Adversarial attacks on multivariate time series. *arXiv preprint arXiv:2004.00410*, 2020.

[55] Andrew C Harvey. *The econometric analysis of time series.* Mit Press, 1990.

[56] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[57] Miklós Z Horváth, Mark Niklas Müller, Marc Fischer, and Martin Vechev. Boosting randomized smoothing with variance reduced classifiers. *arXiv preprint arXiv:2106.06946*, 2021.

[58] Nan-Jung Hsu, Hung-Lin Hung, and Ya-Mei Chang. Subset selection for vector autoregressive processes using lasso. *Computational Statistics & Data Analysis*, 52(7):3645–3657, 2008.

[59] Peter J Huber. Robust regression: asymptotics, conjectures and monte carlo. *The annals of statistics*, 1(5):799–821, 1973.

[60] P.J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[61] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

[62] Jongheon Jeong, Sejun Park, Minkyu Kim, Heung-Chang Lee, Do-Guk Kim, and Jinwoo Shin. Smoothmix: Training confidence-calibrated smoothed classifiers for certified robustness. *Advances in Neural Information Processing Systems*, 34, 2021.

[63] Jongheon Jeong and Jinwoo Shin. Consistency regularization for certified robustness of smoothed classifiers. *Advances in Neural Information Processing Systems*, 33:10558.10570, 2020.

[64] Baisuo Jin, Chen Wang, ZD Bai, K Krishnan Nair, and Matthew Harding. Limiting spectral distribution of a symmetrized auto-cross covariance matrix. *The Annals of Applied Probability*, 24(3):1199–1225, 2014.

[65] Katarina Juselius. *The cointegrated VAR model: methodology and applications.* Oxford university press, 2006.

[66] Kelvin Kan, François-Xavier Aubet, Tim Januschowski, Youngsuk Park, Konstantinos Benidis, Lars Ruthotto, and Jan Gasthaus. Multivariate quantile function forecaster. In *International Conference on Artificial Intelligence and Statistics*, pages 10603–10621. PMLR, 2022.

[67] Abhishek Kaul. Lasso with long memory regression errors. *Journal of Statistical Planning and Inference*, 153:11–26, 2014.

[68] Jongho Kim, Youngsuk Park, John D Fox, Stephen P Boyd, and William Dally. Optimal operation of a plug-in hybrid vehicle with battery thermal and degradation model. In *2020 American Control Conference (ACC)*, pages 3083–3090. IEEE, 2020.

[69] Y.S. Kim, R. Giacometti, S.T. Rachev, F.J. Fabozzi, and D. Mignacca. Measuring financial risk and portfolio optimization with a non-gaussian multivariate model. *Annals of operations research*, 201(1):325–343, 2012.

[70] Anders Bredahl Kock and Laurent Callot. Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics*, 186(2):325–344, 2015.

[71] Dmitri Kondrashov, S Kravtsov, Andrew W Robertson, and Michael Ghil. A hierarchy of data-based enso models. *Journal of climate*, 18(21):4425–4444, 2005.

[72] Siem Jan Koopman and André Lucas. A non-gaussian panel time series model for estimating and decomposing default risk. *Journal of Business & Economic Statistics*, 26(4):510–525, 2008.

[73] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[74] Lai. Dataset of kaggle competition web traffic time series forecasting, version 3., 2017.

[75] Sophie Lambert-Lacroix and Laurent Zwald. Robust regression through the huber's criterion and adaptive lasso penalty. *Electronic Journal of Statistics*, 5:1015–1053, 2011.

[76] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 656–672, 2019.

[77] Olivier Ledoit and Michael Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of empirical finance*, 10(5):603–621, 2003.

[78] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[79] Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise. *Advances in neural information processing systems*, 32, 2019.

[80] Y. Li, X. Bian, and S. Lyu. Attacking object detectors via imperceptible patches on background. *arXiv preprint arXiv:1809.05966*, 2018.

[81] Youjuan Li and Ji Zhu. L 1-norm quantile regression. *Journal of Computational and Graphical Statistics*, 17(1):163–185, 2008.

[82] Bryan Lim, Stefan Zohren, and Stephen Roberts. Recurrent neural filters: Learning independent bayesian filtering steps for time series prediction. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[83] Chizhou Liu, Yunzhen Feng, Ranran Wang, and Bin Dong. Enhancing certified robustness via smoothed weighted ensembling. *arXiv preprint arXiv:2005.09363*, 2020.

[84] Haoyang Liu, Alexander Aue, and Debashis Paul. On the marčenko–pastur law for linear time series. *The Annals of Statistics*, 43(2):675–712, 2015.

[85] Linbo Liu and Danna Zhang. Robust estimation of high-dimensional vector autoregressive models. *arXiv preprint arXiv:2109.10354*, 2021.

[86] Po-Ling Loh. Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *The Annals of Statistics*, 45(2):866–896, 2017.

[87] Po-Ling Loh. Scale calibration for high-dimensional robust regression. *arXiv preprint arXiv:1811.02096*, 2018.

[88] Po-Ling Loh. Scale calibration for high-dimensional robust regression. *Electronic Journal of Statistics*, 15(2):5933–5994, 2021.

[89] Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637, 2012.

[90] Po-Ling Loh and Martin J Wainwright. Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics*, 45(6):2455–2482, 2017.

[91] Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

[92] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[93] Enno Mammen. Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *The Annals of Statistics*, 17(1):382–400, 1989.

[94] Pascal Massart. *Concentration inequalities and model selection*, volume 6. Springer, 2007.

[95] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 34(3):1436–1462, 2006.

[96] Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The annals of statistics*, 37(1):246–270, 2009.

[97] Florence Merlevède, Magda Peligrad, Emmanuel Rio, et al. Bernstein inequality and moderate deviations under strong mixing conditions. In *High dimensional probability V: the Luminy volume*, pages 273–292. Institute of Mathematical Statistics, 2009.

[98] Gautam Raj Mode and Khaza Anuarul Hoque. Adversarial examples in deep learning for multivariate time series regression. In *2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–10. IEEE, 2020.

[99] Manfred Mudelsee. Trend analysis of climate time series: A review of methods. *Earth-science reviews*, 190:310–322, 2019.

[100] Yuval Nardi and Alessandro Rinaldo. Autoregressive process modeling via the lasso procedure. *Journal of Multivariate Analysis*, 102(3):528–549, 2011.

[101] Sahand Negahban and Martin J Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, pages 1069–1097, 2011.

[102] Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

[103] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[104] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[105] Xiaoou Pan, Qiang Sun, and Wen-Xin Zhou. Iteratively reweighted ℓ1-penalized robust regression. *Electronic Journal of Statistics*, 15(1):3287–3348, 2021.

[106] Youngsuk Park, Danielle Maddix, François-Xavier Aubet, Kelvin Kan, Jan Gasthaus, and Yuyang Wang. Learning quantile functions without quantile crossing for distribution-free time series forecasting. In *International Conference on Artificial Intelligence and Statistics*, pages 8127–8150. PMLR, 2022.

[107] Youngsuk Park, Kanak Mahadik, Ryan A Rossi, Gang Wu, and Handong Zhao. Linear quadratic regulator for resource-efficient cloud services. In *Proceedings of the ACM Symposium on Cloud Computing*, pages 488–489, 2019.

[108] Youngsuk Park, Ryan Rossi, Zheng Wen, Gang Wu, and Handong Zhao. Structured policy iteration for linear quadratic regulator. In *International Conference on Machine Learning*, pages 7521–7531. PMLR, 2020.

[109] Syama Sundar Rangapuram, Matthias W Seeger, Jan Gasthaus, Lorenzo Stella, Yuyang Wang, and Tim Januschowski. Deep state space models for time series forecasting. *Advances in neural information processing systems*, 31, 2018.

[110] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory*, 57(10):6976–6994, 2011.

[111] Kashif Rasul. PytorchTS, 2021.

[112] Gregory C Reinsel. *Elements of multivariate time series analysis.* Springer Science & Business Media, 2003.

[113] Adam J Rothman, Elizaveta Levina, and Ji Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186, 2009.

[114] David Salinas, Michael Bohlke-Schneider, Laurent Callot, Roberto Medico, and Jan Gasthaus. High-dimensional multivariate forecasting with low-rank gaussian copula processes. *Advances in neural information processing systems*, 32, 2019.

[115] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020.

[116] Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. *Advances in Neural Information Processing Systems*, 32, 2019.

[117] K. Sameshima and L. A. Baccala. *Methods in Brain Connectivity Inference Through Multivariate Time Series Analysis.* CRC press, 2014.

[118] Jordan Shan. Does financial development 'lead' economic growth? a vector auto-regression appraisal. *Applied Economics*, 37(12):1353–1367, 2005.

[119] A. Shojaie, S. Basu, and G. Michailidis. Adaptive thresholding for reconstructing regulatory networks from time-course gene expression data. *Statistics in Biosciences*, 4(1):66–83, 2012.

[120] Robert H Shumway, David S Stoffer, and David S Stoffer. *Time series analysis and its applications*, volume 3. Springer, 2000.

[121] C. A. Sims. Interpreting the macroeconomic time series facts: The effects of monetary policy. *European Economic Review*, 36(5):975–1000, 1992.

[122] Christopher A Sims. Macroeconomics and reality. *Econometrica*, 48(1):1–48, 1980.

[123] Ezequiel Smucler and Victor J Yohai. Robust and sparse estimators for linear regression models. *Computational Statistics & Data Analysis*, 111:116–130, 2017.

[124] James H Stock and Mark W Watson. Vector autoregressions. *Journal of Economic perspectives*, 15(4):101–115, 2001.

[125] Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020.

[126] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *ICLR*, 2014.

[127] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[128] NYC Taxi and Limousine Commission. Tlc trip record data. https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page, 2015.

[129] Sean J Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018.

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

[131] Ruey S Tsay. Regression models with time series errors. *Journal of the American Statistical Association*, 79(385):118–124, 1984.

[132] Ruey S Tsay. *Analysis of financial time series.* John wiley & sons, 2005.

[133] Sara Van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

[134] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[135] Di Wang and Ruey S Tsay. Robust estimation of high-dimensional vector autoregressive models. *arXiv preprint arXiv:2107.11002*, 2021.

[136] Hansheng Wang, Guodong Li, and Chih-Ling Tsai. Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1):63–78, 2007.

[137] Xueqin Wang, Yunlu Jiang, Mian Huang, and Heping Zhang. Robust variable selection with exponential squared loss. *Journal of the American Statistical Association*, 108(502):632–643, 2013.

[138] Yuyang Wang, Alex Smola, Danielle Maddix, Jan Gasthaus, Dean Foster, and Tim Januschowski. Deep factors for forecasting. In *International conference on machine learning*, pages 6607–6617. PMLR, 2019.

[139] Qingsong Wen, Liang Sun, Fan Yang, Xiaomin Song, Jingkun Gao, Xue Wang, and Huan Xu. Time series data augmentation for deep learning: A survey. *arXiv preprint arXiv:2002.12478*, 2020.

[140] Lily Weng, Pin-Yu Chen, Lam Nguyen, Mark Squillante, Akhilan Boopathy, Ivan Oseledets, and Luca Daniel. Proven: Verifying robustness of neural networks with a probabilistic approach. In *International Conference on Machine Learning*, pages 6727–6736. PMLR, 2019.

[141] Lily Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Luca Daniel, Duane Boning, and Inderjit Dhillon. Towards fast computation of certified robustness for relu networks. In *International Conference on Machine Learning*, pages 5276–5285. PMLR, 2018.

[142] Wei Biao Wu. Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences*, 102(40):14150–14154, 2005.

[143] Wei-Biao Wu and Ying Nian Wu. Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electronic Journal of Statistics*, 10(1):352–379, 2016.

[144] Yangru Wu and Xing Zhou. Var models: Estimation, inferences, and applications. In *Handbook of Quantitative Finance and Risk Management*, pages 1391–1398. Springer, 2010.

[145] Yichao Wu and Yufeng Liu. Variable selection in quantile regression. *Statistica Sinica*, pages 801–817, 2009.

[146] Yong Xie, Dakuo Wang, Pin-Yu Chen, Jinjun Xiong, Sijia Liu, and Oluwasanmi Koyejo. A word is worth a thousand dollars: Adversarial attack on tweets fools stock prediction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 587–599, Seattle, United States, July 2022. Association for Computational Linguistics.

[147] Zhuolin Yang, Linyi Li, Xiaojun Xu, Bhavya Kailkhura, Tao Xie, and Bo Li. On the certified robustness for ensemble models and beyond. *arXiv preprint arXiv:2107.10873*, 2021.

[148] Victor J Yohai and Ricardo A Maronna. Asymptotic behavior of m-estimators for the linear model. *The Annals of Statistics*, 7(2):258–268, 1979.

[149] TaeHo Yoon, Youngsuk Park, Ernest K Ryu, and Yuyang Wang. Robust probabilistic time series forecasting. In *International Conference on Artificial Intelligence and Statistics*, pages 1336–1358. PMLR, 2022.

[150] Young Joo Yoon, Cheolwoo Park, and Taewook Lee. Penalized regression models with autoregressive error terms. *Journal of Statistical Computation and Simulation*, 83(9):1756–1772, 2013.

[151] Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

[152] Runtian Zhai, Chen Dan, Di He, Huan Zhang, Boqing Gong, Pradeep Ravikumar, Cho-Jui Hsieh, and Liwei Wang. Macer: Attack-free and scalable robust training via maximizing certified radius. In *International Conference on Learning Representations*, 2019.

[153] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.

[154] Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

[155] Danna Zhang. Robust estimation of the mean and covariance matrix for high dimensional time series. *Statistica Sinica*, 31(2):797–820, 2021.

[156] Danna Zhang and Wei Biao Wu. Gaussian approximation for high dimensional time series. *The Annals of Statistics*, 45(5):1895–1919, 2017.

[157] Xianyang Zhang and Guang Cheng. Simultaneous inference for high-dimensional linear models. *Journal of the American Statistical Association*, 112(518):757–768, 2017.

[158] Ziwei Zhu and Wenjing Zhou. Taming heavy-tailed features by shrinkage. *arXiv preprint arXiv:1710.09020*, 2017.

[159] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

[160] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.