THE UNIVERSITY OF CHICAGO

ROBUST ESTIMATION OF HIGH DIMENSIONAL TIME SERIES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY

YUEFENG HAN

CHICAGO, ILLINOIS
AUGUST 2019

To my parents, Guozhu Han and Zhuzhen Su

"In this world, there is only one protagonist in one million people. The protagonists are the ones who can achieve perfection, but they have to get to the right stage first." – *Project Gutenberg.*

TABLE OF CONTENTS

LIST OF FIGURES vii

LIST OF TABLES viii

ACKNOWLEDGMENTS x

ABSTRACT xi

1 INTRODUCTION 1

2 HIGH-DIMENSIONAL LINEAR REGRESSION WITH APPLICATIONS TO NOWCASTING 5
- 2.1 Introduction 5
- 2.2 High Dimensional Time Series 7
- 2.3 Convergence Rate of the Lasso estimator 11
- 2.4 Model Selection Consistency 16
- 2.5 Simulation Study 23
- 2.6 Empirical Analysis 29
  - 2.6.1 Predicting GDP growth 29
  - 2.6.2 Nowcasting $PM_{2.5}$ 35
- 2.7 Deferred Proofs 37
  - 2.7.1 Lemmas 37
  - 2.7.2 A general theorem of estimation error for weak sparsity 39
  - 2.7.3 Proof of Theorem 2.3.1 40
  - 2.7.4 Proof of Theorem 2.3.2 43
  - 2.7.5 Proof of Theorem 2.4.1 45
  - 2.7.6 Proof of Proposition 2.4.1 48

3 HIGH DIMENSIONAL GENERALIZED LINEAR MODELS 50
- 3.1 Introduction 50
- 3.2 The Model 52
  - 3.2.1 Generalized Linear Models and the Loss Function 52
  - 3.2.2 Robust Lasso Estimator 54
- 3.3 Asymptotic properties 56
  - 3.3.1 Definitions and Assumptions 57
  - 3.3.2 Rate of Convergence 59
- 3.4 Application to Linear Regression 67
- 3.5 Simulation Study 70
  - 3.5.1 Statistical error 70
  - 3.5.2 Convergence speed 76
- 3.6 Real Data Analysis 77
- 3.7 Inequalities for Empirical Processes of High-dimensional Time Series 81
- 3.8 Deferred Proofs 96

REFERENCES 106

LIST OF FIGURES

2.1 Results of Lasso regression for the two AR(3) series in (2.4.7) and (2.4.8) via the glmnet package of R. 22
2.2 Panel (a): Cumulative absolute errors. Panel (b): Cumulative squared errors. MIDAS-A represents the MIDAS regression model using only monthly all-employees total payrolls as the explanatory variable. MIDAS-B represents the MIDAS regression model with seven regressors $z_{1,\cdot}, \cdots, z_{7,\cdot}$, where $z_{6,\cdot}$ and $z_{7,\cdot}$ are aggregated into weekly data. 33
2.3 Panel (a): Cumulative absolute errors. Panel (b) Cumulative squared errors. MIDAS-D represents the MIDAS regression model with seven regressors $z_{1,\cdot}, ..., z_{7,\cdot}$. MIDAS-C represents the MIDAS regression model with nine regressors $z_{1,\cdot}, ..., z_{9,\cdot}$. Now-casting 1 and Now-casting 2 represent predicting quarterly GDP growth rate when the first month and the first two months data are available, respectively. . 36

LIST OF TABLES

2.1 Accuracy in Parameter Estimation of Lasso Regression and Mixed-Frequency Data Sampling Regression. The results are based on 10,000 repetitions, where AE and RMSE denote the average of mean absolute errors and average of root mean square errors over Monte Carlo repetitions and parameters. In the table, s, p, and n denote the number of non-zero parameters, the dimension of regressors, and sample size, respectively. 27

2.2 Performance of Out-of-sample predictions of Lasso regression and mixed frequency data sampling regression (MIDAS). The results are based on 10 one-step ahead predictions and 10,000 iterations, where AFE and RMSFE denote the average absolute forecast errors and root mean squared forecast errors, respectively, and s, p, and n are the number of non-zero parameters, the dimension of regressors, and sample size. For MIDAS, the maximum p is fixed at 100. 28

2.3 Accuracy in Model Selection and Parameter Estimation of Lasso Estimator and Dantzig Estimator for Linear Regression. The results are based on 10,000 repetitions, where RMSE denote the average of root mean square errors over Monte Carlo repetitions and parameters. In the table, s, p, and n denote the number of non-zero parameters, the dimension of regressors, and sample size, respectively. 30

2.4 Results of out-of-sampling prediction of U.S. quarterly real GDP growth rate. The data span is from 1980 to February 2017, but the forecast origins start from the second quarter of 2013 to the first quarter of 2017. All measurements are multiplied by 10^3. In the table, MAD, MAE, and RMSE are the median absolute error, mean absolute error, and root mean squared error, respectively. 32

2.5 Comparison between forecasting and now-casting in predicting the U.S. quarterly real GDP growth rate. The data span is from 1980 to February 2017, but the forecast origins are from the second quarter of 2013 to the first quarter of 2017. All measurements are multiplied by 10^3. In the table, MAD, MAE, RMSE are the median absolute deviation, mean absolute error, and root mean squared error, respectively. 35

2.6 Comparison between forecasting and now-casting in predicting the daily maximum of $PM_{2.5}$. The data span is from 2006 to 2015, and the forecast origins are from 2013 to the end of 2015.(Feb 29 was dropped). In the table, MAE, RMSE are the mean absolute error, and root mean squared error for one-step ahead predictions, respectively. 38

3.1 Simulation results of Lasso and robust Lasso for logistic regression ($p = 400, \rho = 0.5$, where n is the sample size. The results are based on 5000 replications. . . . 73

3.2 Simulation results of Lasso and robust Lasso for linear regression ($p = 400, \rho = 0.5$), where n is the sample size and the results are based on 5000 replications . 75

3.3 Simulated values of the tail probability ratios $\Lambda_1(t)$ and $\Lambda_2(t)$ of Equation (3.5.3) for Lasso and robust Lasso (R-Lasso) procedures. 1000000 replications are used to evaluate the probabilities. 77

3.4 Forecast tabulation for the ordered Probit model 80

3.5 Forecast tabulation for the 27-10-1 feedforward neural network 80

3.6 Forecast tabulation for the standard Lasso method 81
3.7 Forecast tabulation for the robust Lasso method 81

ACKNOWLEDGMENTS

First and foremost, I am very grateful to my advisors, Wei Biao Wu and Ruey S. Tsay, for their careful supervision and valuable suggestions of my academic studies in the past five years. They have greatly improved my understanding of the discipline of Statistics and taught me a lot of specific research skills. Their keen and vigorous academic observation enlightens me not only in this thesis but also in my future academic career.

I would like to express my heartfelt gratitude to my committee member Chao Gao and Per Mykland for their deep insight, strong support and helpful feedback on my work. I am greatly indebted to the professors and teachers at the Department of Statistics, especially Stephen Stigler, Peter McCullagh, Michael Stein, Steven Lalley, Dan Nicolae, Matthew Stephens and Mei Wang, who have instructed and helped me a lot to develop the fundamental and essential academic competence. I would also like to thank staff and my fellow students who I have met in Chicago, especially Mengyin Lu, Likai Chen, Zhipeng Lou, Pinhan Chen, Yuancheng Zhu, Qinqing Zheng, Mengyu Xu, Si Tang and Luyi Yang.

Special thanks should go to my best friend, Danna Zhang. Thanks for giving me care and support, because of her, the university life is colorful. She looks like a bright light in the dark that illuminates my life. Without her impressive kindness and encouragement, I wouldn't be who I am today.

Last but not least, my thanks would go to my beloved family, my parents Guozhu Han and Zhuzhen Su, for their loving considerations and great confidence in me all through these years. They are always my most strong shield.

ABSTRACT

In recent years, extensive research has focused on the ℓ_1 penalized least squares (Lasso) estimators of high-dimensional regression when the number of covariates p is considerably larger than the sample size n. However, there is limited attention paid to the properties of the estimators when the errors and/or the covariates are serially dependent and/or heavy tailed.

This thesis concerns the theoretical properties of the Lasso estimators for linear regression with random design and weak sparsity under serially dependent and/or non-sub-Gaussian errors and covariates. In contrast to the traditional case in which the errors are independent and identically distributed (i.i.d.) and have finite exponential moments, we show that p can be at most a power of n if the errors have only finite polynomial moments. In addition, the rate of convergence becomes slower due to the serial dependence in errors and the covariates. We also consider sign consistency for model selection via Lasso when there are serial correlations in the errors or the covariates or both. Adopting the framework of functional dependence measure, we provide a detailed description on how the rates of convergence and the selection consistency of the estimators depend on the dependence measures and moment conditions of the errors and the covariates. We apply the results obtained for the Lasso method to now-casting with mixed-frequency data for which serially correlated errors and a large number of covariates are common. The empirical results show the superiority of Lasso procedure in both forecasting and now-casting.

This thesis also proposes a new robust M-estimator for generalized linear models. We investigate properties of the proposed robust procedure and the classical Lasso procedure both theoretically and numerically. As an extension, we also introduce robust estimator for linear regression. We show that the proposed robust estimator for linear model will achieve the optimal rate which is the same as the one for i.i.d sub-Gaussian data. Simulation results show that the proposed method performs well numerically in terms of heavy-tailed and serially dependent covariates and/or errors, and it significantly outperforms the classical Lasso

method. For applications, we demonstrate the regularized robust procedure via analyzing high-frequency trading data in finance. We also provide new Bousquet type inequalities for high-dimensional time series, which could be quite useful in empirical process of dependent data.

CHAPTER 1

INTRODUCTION

During the past decade, there has been a well-developed theory for regularized estimation for high dimensional regression. However, most literature deals with the case that the samples are i.i.d. High dimensional time series analysis has gained its credibility recently in finance, signal processing, neuroscience, meteorology, seismology and many other areas. In many applications, we also face the challenge of heavy tails.

Generalized linear models (GLM, McCullagh and Nelder [1989]) are a flexible generalization of the ordinary linear regression by allowing researchers to model the relationship between the predictors and a function of the mean of the response variable, which can follow a continuous or discrete distribution. In a variety of applications, the observed response consists of count data for which GLM is especially useful. This chapter deals with the Lasso penalty for GLM applied to high dimensional time series data. Under the independent and identically distributed (i.i.d.) setting, there exists a substantial literature on the Lasso methods for high-dimensional GLM. For instance, van de Geer [2008] showed non-asymptotic oracle inequalities for the empirical risk minimizer with Lasso penalty for high-dimensional GLMs with a Lipschitz loss function. Kong and Nan [2014] extended the approach of van de Geer [2008] to the Cox proportional hazards regression. Gaïffas et al. [2012] considered a quadratic loss function in place of a negative log-likelihood function in an additive hazards model with Lasso penalty. Huang et al. [2013] studied Lasso estimator in the Cox proportional hazards regression when the covariates are time dependent, and established oracle inequalities for prediction and estimation errors. A number of papers analyzed penalized methods beyond Lasso. Meier et al. [2008] applied group Lasso to high-dimensional logistic regression, proposed an efficient algorithm, and showed consistency of the estimator. Negahban et al. [2012] studied penalized M-estimators with a general class of regularization methods, including an ℓ_2 error bound for the Lasso in GLM. Huang and Zhang [2012] studied weighted absolute penalty and its adaptive, multistage application in

GLM. Fan and Lv [2013] investigated asymptotic equivalence of Lasso and other concave regularized methods in a thresholded parameter space. Ivanoff et al. [2016] studied adaptive Lasso and group-Lasso for the functional Poisson regression.

Despite the extensive research in GLMs for i.i.d data, very limited work focused on theoretical properties of the regularized estimates when the observations are dependent. Basu and Michailidis [2015] investigated theoretical properties of Lasso estimators with a random design for high-dimensional Gaussian processes. Wu and Wu [2016] analyzed Lasso estimator with a fixed design matrix and Dantzig selector under random design. Hall et al. [2016] studied Lasso estimators of high-dimensional autoregressive generalized linear models. Zhou and Raskutti [2018] further extended the study to non-parametric sparse additive model. Han and Tsay [2017] extended the Lasso estimator to random design and weakly sparse time series with application in now-casting.

The phenomenon of heavy-tails is widely observed in time series data. It is one of the stylized facts in financial econometrics that financial returns and macroeconomic variables have high excess kurtosis. Large scale imaging data in biology, such as neural spike recordings (see, for example, Brown et al. [2004] and Pillow et al. [2008]), are often corrupted by non-Gaussian noises. The usual regression estimator may fare poorly or even be inconsistent when the observations are heavy tailed and/or contaminated by outliers in the predictors and/or the response variable. Therefore, it is important to study effective principles for dealing with heavy-tailed or noisy time series data.

The origin of robust statistics dates back to the fundamental works of John Tukey (Tukey [1960, 1962]), Peter Huber (Huber [1964, 1967]) and Frank Hampel (Hampel [1971, 1974]). In general, robustness can be defined in two ways; model misspecification and outliers. For example, Tukey's work (Tukey [1960]) is about robustness to a misspecification of the Gaussian model, while Hodges' work (Hodges Jr [1967]) is robustness to contamination of the dataset by extreme outliers or robustness to heavy-tailed distributions in the model that lead to the appearance of some aberrant data.

We aim to establish convergence rate and sign consistency for high-dimensional linear regression in Chapter 2. We first define the high-dimensional dependence measure, adopting the concept of Wu [2005]. Under weak sparsity condition and heavy tailed covariates and errors, we investigate rates of convergence of Lasso estimators for high dimensional time series. We show that p can be at most a power of n if the errors have only finite polynomial moments. In addition, the rate of convergence becomes slower due to the serial dependence in errors and the covariates. We also study model selection consistency of Lasso estimators. As an application, we apply the results obtained for the Lasso method to now-casting with mixed-frequency data for which serially correlated errors and a large number of covariates are common. The empirical results show the superiority of Lasso method in both forecasting and now-casting.

The primary goal of Chapter 3 is to construct robust M-estimators of generalized linear models for serially dependent data and lay a theoretical foundation for estimation consistency. We first introduce the standard Lasso procedure and the robust procedure for generalized linear model. To provide a solid theoretical guarantee, we derive convergence rates of both robust and non-robust estimators which depend on the sample size, the dimension, the moment condition and the dependence of the underlying processes. In particular, we study the usual linear regression with robust estimator and time series data. We show that the proposed robust estimator for linear model will achieve the optimal rate which is the same as the one for i.i.d sub-Gaussian data. We also provide new Bousquet type inequalities for high-dimensional time series, which could be quite useful in empirical process of dependent data.

We now introduce some notations. For the matrix $A = (a_{ij}) \in \mathbb{R}^{p\times q}$, denote $|A|_\infty = \max_{i,j} |a_{ij}|$, the spectral norm $\rho(A) = \sup_{|x|\leq 1} |Ax|_2$ and the Frobenius norm $|A|_F = (\sum_{ij} a_{ij}^2)^{1/2}$. For a vector $x = (x_1, ..., x_p)'$, define $|x|_q = (x_1^q + ... + x_p^q)^{1/q}$ and $|x|_\infty = \max\{|x_1|, ..., |x_p|\}$. We write a random variable $\xi \in \mathcal{L}^m$, $m \geq 1$, if $\|\xi\|_m = (\mathbb{E}|\xi|^m)^{1/m} < \infty$. For simplicity, write $\|\xi\| = \|\xi\|_2 = (\mathbb{E}|\xi|^2)^{1/2}$. For a set S, write $|S|_0$ or $|S|$ as its cardi-

nality. For two sequences of real numbers $\{a_n\}$ and $\{b_n\}$, write $a_n = O(b_n)$ if there exists a constant C such that $|a_n| \leq C|b_n|$ holds for all sufficiently large n, write $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$, and write $a_n \asymp b_n$ if there are positive constants c and C such that $c \leq a_n/b_n \leq C$ for all sufficiently large n. Denote $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. We use $C, C_1, C_2, \cdots$ to denote positive constants whose values may differ from place to place. A constant with a symbolic subscript is used to emphasize the dependence of the value on the subscript. We assume $p = p_n \to \infty$ as $n \to \infty$.

CHAPTER 2

HIGH-DIMENSIONAL LINEAR REGRESSION WITH APPLICATIONS TO NOWCASTING

2.1 Introduction

During the past two decades, there have been significant developments in high-dimensional linear regression analysis. Consider the linear regression for the response variable Y_i and the covariate vector X_i,

$$Y_i = X_i^T \beta + e_i, \quad 1 \leq i \leq n, \tag{2.1.1}$$

where $\beta \in \mathbb{R}^p$ consists of unknown coefficients, e_i is the error term, and X_i^T denotes the transpose of the covariate vector X_i. Denote the dimension of X_i by p. In matrix form, we can write the model as $\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{e}$, where $\boldsymbol{Y}$ is the $n \times 1$ response vector, $\boldsymbol{X}$ is the $n \times p$ design matrix, and $\boldsymbol{e}$ is the $n \times 1$ error vector. Under certain sparsity conditions on β, a great deal of attention has been focused on the ℓ_1 penalized least squares (Lasso) estimator of β when the number of variables p can be much larger than the sample size n; see Efron et al. [2004], Zhao and Yu [2006], and Meinshausen and Yu [2009], among others. Other related approaches include the Dantzig-selector of Candes and Tao [2007], the adaptive Lasso of Zou [2006], the Group Lasso by Yuan and Lin [2006] and the SCAD estimator of Fan and Li [2001], among others. Theoretical properties of those estimators have been established in the literature under the independence assumption; see, for example, Bickel et al. [2009] and Bühlmann and Van De Geer [2011]. Here we focus on the Lasso estimator defined as

$$\hat{\beta} = \arg\min_{\beta} \left(\frac{1}{2} |\boldsymbol{Y} - \boldsymbol{X}\beta|_2^2 + \lambda |\beta|_1 \right), \tag{2.1.2}$$

where $\lambda \geq 0$ is a tuning parameter, controlling the level of sparsity in $\hat{\beta}$.

Much of the available research dedicated to the Lasso problem deals with the case of large p and small n when the design matrix is static and the errors are independent and identically distributed (i.i.d.) random variables. On the other hand, in many real applications, X_i consists of stochastic random variables that might be dynamically dependent or e_i is serially dependent or both. Despite a considerable amount of recent work on Lasso estimators, there has been limited research on theoretical properties of the estimates when the observations are dependent. Wang et al. [2007] proposed a Lasso estimator for the regression model with autoregressive errors. Gupta [2012] investigated Lasso estimator for weakly dependent errors. Both papers concentrate on the case when n is greater than p. More recently, Basu and Michailidis [2015] investigated theoretical properties of Lasso estimators with a random design for high-dimensional Gaussian processes. Kock and Callot [2015] established oracle inequalities of the Lasso for Gaussian errors in stationary vector autoregressive models. Wu and Wu [2016] analyzed Lasso estimator with a fixed design matrix and assumed that a restricted eigenvalue condition is satisfied. Medeiros and Mendes [2016] studied the asymptotic properties of the adaptive Lasso when the errors are non-Gaussian and may be conditionally heteroskedastic. The goal of this chapter is to investigate the limiting properties of Lasso estimators of Model (2.1.1) in the presence of serial dependence in both the covariate vector X_i and the errors. We establish rate of convergence of the lasso estimator under weak sparsity condition, and provide sign consistency of lasso regression. Our results extend beyond the fixed design and exact sparsity time series, and we do not assume the restricted eigenvalue condition on either the sample or the population covariance matrix.

In practice, many important macroeconomic variables are not sampled at the same frequency. For example, gross domestic product (GDP) data are available quarterly, industrial production data are monthly, and most interest rate data are available daily. Analyzing such data jointly is referred to as the mixed-frequency data analysis. In the econometrics literature, Ghysels et al. [2004] proposed a mixed-data sampling (MIDAS) approach to analyze mixed-frequency data. In particular, they use newly available high-frequency data to

improve the prediction of a lower-frequency macroeconomic variable of interest and refer to such predictions as *now-casting*. Consider, for example, the problem of predicting quarterly GDP growth rate y_{n+1} at the forecast origin $i = n$. Here the time interval is a quarter. Traditional forecasting methods employ quarterly data available at $i = n$ to build a model, then use the fitted model to perform prediction. In practice, some monthly and daily data become available during the quarter $i = n + 1$. Now-casting is to make use of such newly available monthly and daily data to update the prediction of y_{n+1}. Therefore, the term now-casting means taking advantages of high-frequency data within a given quarter to update the prediction of GDP growth rate of that quarter. In short, the basic principle of now-casting is the exploitation of the information which is published at higher frequencies than the target variable of interest in order to obtain an improved prediction before the official lower-frequency data becomes available. Since many high-frequency data are available, a large number of covariates are common in now-casting. Therefore, Model (2.1.1) with dependent covariates and errors is applicable to now-casting, and the Lasso method is highly relevant. The mixed-data sampling approach of Ghysels et al. [2004] has proven useful for various forecasting and now-casting purposes. We compare the performance of Lasso regression with MIDAS regression and autoregressive model with exogenous variables (ARX) in this chapter. To the best of our knowledge, this is the first results to apply lasso regression to nowcasting. Both simulation studies and empirical studies show that Lasso estimator outperforms the existing MIDAS regression and ARX model.

2.2 High Dimensional Time Series

Let $\varepsilon_i, i \in \mathbb{Z}$, be i.i.d. random vectors and σ-field $\mathcal{F}_i = (\cdots, \varepsilon_{i-1}, \varepsilon_i)$. In our random-design setting, we assume that in Model (2.1.1) the covariate process $(X_i, i = 1, ..., n)$ is high-dimensional and weakly stationary in the form

$$X_i = (g_1(\mathcal{F}_i), ..., g_p(\mathcal{F}_i))^T, \tag{2.2.1}$$

the error e_i satisfies

$$e_i = g_e(\mathcal{F}_i), \tag{2.2.2}$$

and the response Y_i assumes the form

$$Y_i = g_y(\mathcal{F}_i), \tag{2.2.3}$$

where $g_1(\cdot), \ldots, g_p(\cdot)$ and $g_e(\cdot), g_y(\cdot)$ are measurable functions in $\mathbb{R}$ such that X_i, e_i and Y_i are well-defined. In the scalar case with $p = 1$, (2.2.1) and (2.2.2) include a very general class of stationary process (c.f. Wiener [1958], Rosenblatt [1971], Priestley [1988], Tong [1990], Tsay [2005], Wu [2005]). They also allow models with homogeneous or heteroscedastic errors; see Example 2.3.1. In the homogeneous case, the covariate process (X_i) and the errors (e_i) can be independent of each other.

Following Wu [2005], we define the functional dependence measure

$$\begin{aligned}
\delta_{i,q,j} &= \|X_{ij} - X_{ij}^*\|_q = \|g_j(\mathcal{F}_i) - g_j(\mathcal{F}_i^*)\|_q, && (2.2.4)\\
\delta_{i,q,e} &= \|e_i - e_i^*\|_q = \|g_e(\mathcal{F}_i) - g_e(\mathcal{F}_i^*)\|_q, && (2.2.5)\\
\delta_{i,q,y} &= \|Y_i - Y_i^*\|_q = \|g_y(\mathcal{F}_i) - g_y(\mathcal{F}_i^*)\|_q, && (2.2.6)
\end{aligned}$$

where the coupled process $X_{ij}^* = g_j(\mathcal{F}_i^*)$, $e_i^* = g_e(\mathcal{F}_i^*)$ and $Y_i^* = g_y(\mathcal{F}_i^*)$ with $\mathcal{F}_i^* = (..., \varepsilon_{-1}, \varepsilon_0', \varepsilon_1, ..., \varepsilon_{i-1}, \varepsilon_i)$ and $\varepsilon_0', \varepsilon_l, l \in \mathbb{Z}$, being i.i.d. random variables. We assume short-

range dependence so that

$$\Delta_{m,q,j} := \sum_{i=m}^{\infty} \delta_{i,q,j} < \infty, \tag{2.2.7}$$

$$\Delta_{m,q,e} := \sum_{i=m}^{\infty} \delta_{i,q,e} < \infty, \tag{2.2.8}$$

$$\Delta_{m,q,y} := \sum_{i=m}^{\infty} \delta_{i,q,y} < \infty. \tag{2.2.9}$$

Then for fixed m, $\Delta_{m,q,j}$, $\Delta_{m,q,e}$ and $\Delta_{m,q,y}$ measure the cumulative effect of ε_0 on $(X_{ij})_{i\geq m}$, $(e_i)_{i\geq m}$ and $(Y_i)_{i\geq m}$. We introduce the following dependence adjusted norm (DAN)

$$\|x_{\cdot j}\|_{q,\alpha} = \sup_{m\geq 0}(m+1)^{\alpha}\Delta_{m,q,j}, \quad \alpha \geq 0. \tag{2.2.10}$$

$$\|e_{\cdot}\|_{q,\alpha} = \sup_{m\geq 0}(m+1)^{\alpha}\Delta_{m,q,e}, \quad \alpha \geq 0. \tag{2.2.11}$$

$$\|Y_{\cdot}\|_{q,\alpha} = \sup_{m\geq 0}(m+1)^{\alpha}\Delta_{m,q,y}, \quad \alpha \geq 0. \tag{2.2.12}$$

It can happen that, due to dependence, $\|e_{\cdot}\|_{q,\alpha} = \infty$ while $\|e_i\|_q < \infty$. Since $e_0 = \sum_{l=-\infty}^{0}(\mathbb{E}(e_0|\mathcal{F}_l) - \mathbb{E}(e_0|\mathcal{F}_{l-1}))$, we have

$$\|e_0\|_q \leq \sum_{l=0}^{\infty}\|\mathbb{E}(e_0|\mathcal{F}_{-l}) - \mathbb{E}(e_0|\mathcal{F}_{-l-1})\|_q = \sum_{l=0}^{\infty}\|\mathbb{E}(e_l - e_l^*|\mathcal{F}_0)\|_q \leq \sum_{l=0}^{\infty}\|e_l - e_l^*\|_q = \|e_{\cdot}\|_{q,0}, \tag{2.2.13}$$

by stationarity. If e_i, $i \in \mathbb{Z}$, are i.i.d., the dependence adjusted norm $\|e_{\cdot}\|_{q,\alpha}$ and the $\mathcal{L}^q$ norm $\|e_0\|_q$ are equivalent in the sense that $\|e_0\|_q \leq \|e_{\cdot}\|_{q,\alpha} \leq 2\|e_0\|_q$.

To account for the cross-sectional dependence of the p-dimensional stationary process (X_i), we define the $\mathcal{L}^{\infty}$ functional dependence measure and its corresponding dependence

adjusted norm (c.f. Chen et al. [2013], Zhang and Wu [2017])

$$\omega_{i,q} = \| \max_{1\le j\le p} |X_{ij} - X^*_{ij}| \|_q,$$
$$\| |X_\cdot|_\infty \|_{q,\alpha} = \sup_{m\ge 0} (m+1)^\alpha \Omega_{m,q}, \quad \alpha \ge 0, \quad \text{and } \Omega_{m,q} = \sum_{i=m}^{\infty} \omega_{i,q}.$$

Additionally, we define

$$\Psi_{q,\alpha} = \max_{1\le j\le p} \|X_{\cdot j}\|_{q,\alpha} \quad \text{and} \quad \Upsilon_{q,\alpha} = \left(\sum_{j=1}^{p} \|X_{\cdot j}\|^q_{q,\alpha} \right)^{1/q},$$

where $\Psi_{q,\alpha}$ and $\Upsilon_{q,\alpha}$ can be viewed as the uniform and the overall dependence adjusted norms of (X_i). Clearly, $\Psi_{q,\alpha} \le \| |X_\cdot|_\infty \|_{q,\alpha} \le \Upsilon_{q,\alpha}$.

We give an example of high-dimensional time series to illustrate how the univariate and multivariate dependence adjusted norms scale.

Example 2.2.1. *Let $\varepsilon_{ij}, i, j \in \mathbb{Z}$, be i.i.d. random variables with mean 0, variance 1, and having finite qth moments, $q > 2$, and let A_i, $i \ge 0$, be $p \times d$ coefficient matrices with real entries such that $\sum_{i=0}^{\infty} tr(A_i A_i^T) < \infty$. Write $\varepsilon_i = (\varepsilon_{i1}, ..., \varepsilon_{id})^T$. Then by Kolmogorov's three-series theorem the linear process*

$$X_i = \sum_{l=0}^{\infty} A_l \varepsilon_{i-l} \tag{2.2.14}$$

exists. Denote $A_l = (a_{l;jk})_{1\le j\le p, 1\le k\le d}$, $A_{l,j\cdot}$ the jth row of A_l. By Burkholder's inequality, $\|A_{l,j\cdot}\varepsilon_0\|_q \le \sqrt{q-1}|A_{l,j\cdot}|_2 \|\varepsilon_{00}\|_q$. We assume that the linear process satisfies the decay condition

$$\max_{j\le p} |A_{l,j\cdot}|_2 \le K_1 (1 \vee l)^{-\theta} \tag{2.2.15}$$

for all $l \ge 0$, where $\theta > 1/2$ and $K_1 > 0$. If $\theta > 1$, (2.2.15) implies short-range dependence

(SRD) since the auto-covariance matrices $\Sigma_k = \sum_{l=0}^{\infty} A_l A_{l+k}^T$ *are absolutely summable. On the other hand, if* $1 > \theta > 1/2$*, then* (X_i) *in (2.2.14) may not have summable auto-covariance matrices, thus allowing long-range dependence (LRD). The classical literature on LRD primarily focuses on the univariate case* $p = 1$*. Then under the SRD case, the dependence adjusted norms have the following bounds*

$$\Psi_{q,\alpha} = \max_{1\le j\le p} \|X_{\cdot j}\|_{q,\alpha} = \max_j \sup_{m\ge 0} (m+1)^{\alpha} \sum_{i=m}^{\infty} \|A_{i,j\cdot}\varepsilon_0\|_q \le K_1 K_2 \|\varepsilon_{00}\|_q, \tag{2.2.16}$$

$$\||X_{\cdot}|_{\infty}\|_{q,\alpha} = \sup_{m\ge 0} (m+1)^{\alpha} \sum_{i=m}^{\infty} \| \max_j |A_{i,j\cdot}\varepsilon_0|\|_q \le K_1 K_2 p^{1/q} \|\varepsilon_{00}\|_q, \tag{2.2.17}$$

where $\alpha = \theta - 1$ *and the constant* K_2 *only depends on* θ *and* q*.*

In this chapter, we use dependence adjusted norms $\||X_{\cdot}|_{\infty}\|_{q,\alpha}$, $\Psi_{q,\alpha}$, and $\Upsilon_{q,\alpha}$ to study the limiting properties of Lasso estimators in the presence of serial dependence. These adjusted norms are more convenient than the commonly used mixing conditions for handling serial dependence in high-dimensional time series.

2.3 Convergence Rate of the Lasso estimator

In this section, we present the main results on convergence rate of the Lasso estimator for dependent data. In the low-dimensional case, the consistency of $\hat{\beta}$ relies on the assumption that the sample covariance matrix converges to the population covariance matrix. In the high-dimensional case ($n \ll p$), it requires that $|\boldsymbol{X}(\hat{\beta} - \beta)|_2$ is small only when $|\hat{\beta} - \beta|_2$ is small. Let $\hat{\Sigma} = (\hat{\sigma}_{jk})_{1\le j,k\le p} = n^{-1} \sum_{i=1}^{n} X_i X_i^T$ be the sample covariance. Typically, researchers assume with high probability, the following Restricted Strong Convexity condition holds,

$$u'\hat{\Sigma}u \ge \kappa_1 |u|_2^2 - \kappa_2 g(n,p) |u|_1^2, \tag{2.3.1}$$

for all $u \in \mathbb{R}^p$, where κ_1, κ_2 are positive constants and $g(n,p)$ is a function of the sample size n and ambient dimension p. It can be viewed as an analogous sufficient condition in the high-dimensional case. As shown in the proof of Theorem 2.3.1, the Restricted Strong Convexity condition for the sample covariance matrix holds with high probability under certain conditions.

To establish our theoretical results, we first impose a weak sparsity condition:

Assumption 2.3.1. *There exits some $0 \leq \theta < 1$, with a uniform radius K_θ such that*

$$\sum_{j=1}^{p} |\beta_j|^\theta \leq K_\theta. \tag{2.3.2}$$

The following theorem shows the L_2 and L_1 convergence rates of $\hat{\beta}$ to β depend on the moment condition and the temporal and cross-sectional dependence conditions.

Theorem 2.3.1. *Denote the population covariance matrix by $\Sigma = (\sigma_{jk}) = [Cov(X_{ij}, X_{ik})]$. Suppose the minimum eigenvalue of Σ satisfies $\lambda_{\min}(\Sigma) \geq \kappa > 0$. Assume that $\Psi_{\gamma,\alpha_X} = \max_j \|X_{\cdot j}\|_{\gamma,\alpha_X} = M_X < \infty$ and $\|e_\cdot\|_{q,\alpha_e} = M_e < \infty$, where $q > 2, \gamma > 4$ and $\alpha_X, \alpha_e > 0$. Define*

$$\nu = \begin{cases} 1 & \text{if } \alpha_X \geq 1/2 - 2/\gamma, \\ \gamma/4 - \alpha_X\gamma/2 & \text{if } \alpha_X < 1/2 - 2/\gamma. \end{cases}$$

Assume $\tau = q\gamma/(q+\gamma) > 2$ and let $\alpha = \min(\alpha_X, \alpha_e)$. Define

$$\rho = \begin{cases} 1 & \text{if } \alpha \geq 1/2 - 1/\tau, \\ \tau/2 - \alpha\tau & \text{if } \alpha < 1/2 - 1/\tau. \end{cases}$$

Denote $\omega = \sqrt{\log p/n} M_X^2 + n^{2\nu/\gamma - 1} (\log p)^{3/2} \| |X_\cdot|_\infty \|_{\gamma,\alpha_X}^2$. Suppose Assumption 2.3.1 holds. Then for any λ such that

$$\lambda \gtrsim \sqrt{\log p/n} M_e M_X + n^{\rho/\tau - 1} (\log p)^{3/2} M_e \| |X_\cdot|_\infty \|_{\gamma,\alpha_X},$$

and $K_\theta \omega \lambda^{-\theta} \leq C$ *for some positive constant* C*, any Lasso solution* $\hat{\beta}$ *satisfies,*

$$|\hat{\beta} - \beta|_2 \lesssim \sqrt{K_\theta} \left(\frac{\lambda}{\kappa}\right)^{1-\theta/2}, \tag{2.3.3}$$

$$|\hat{\beta} - \beta|_1 \lesssim K_\theta \left(\frac{\lambda}{\kappa}\right)^{1-\theta}. \tag{2.3.4}$$

with probability at least $1 - C_1(\log p)^{-\gamma/2} - C_2 p^{-C_3} - C_4(\log p)^{-\tau}$*, where* $C_1, ..., C_4$ *are positive constants.*

In the special case $\theta = 0$, the quantity of weak sparsity corresponds to an exact sparsity constraint–that is, β has at most $s := K_0$ nonzero entries. The following theorem shows the convergence rate of $\hat{\beta}$ and the prediction error $|\boldsymbol{X}(\hat{\beta} - \beta)|_2^2$ for the exact sparsity case.

Theorem 2.3.2. *Suppose the same conditions of Theorem 2.3.1 hold. If* $|\beta|_0 = s$ *and* $\kappa \asymp 1$*,*

$$n \gtrsim M_X^4 s^2 \log p + s^{1/(1-2\nu/\gamma)} (\log p)^{3/(2-4\nu/\gamma)} \| |X_\cdot|_\infty \|_{\gamma,\alpha_X}^{2/(1-2\nu/\gamma)},$$

then for any λ *such that*

$$\lambda \gtrsim \sqrt{\log p/n} M_e M_X + n^{\rho/\tau - 1} (\log p)^{3/2} M_e \| |X_\cdot|_\infty \|_{\gamma,\alpha_X},$$

any Lasso solution $\hat{\beta}$ *satisfies,*

$$|\hat{\beta} - \beta|_2 \lesssim \lambda\sqrt{s}/\kappa, \tag{2.3.5}$$

$$|\hat{\beta} - \beta|_1 \lesssim \lambda s/\kappa, \tag{2.3.6}$$

$$|X(\hat{\beta} - \beta)|_2^2/n \lesssim \lambda^2 s/\kappa, \tag{2.3.7}$$

with probability at least $1 - C_1(\log p)^{-\gamma/2} - C_2 p^{-C_3} - C_4(\log p)^{-\tau}$*.*

Remark 2.3.1. *In the exact sparsity case, instead of the condition* $\lambda_{\min}(\Sigma) \geq \kappa > 0$*, we may require that the restricted eigenvalue assumption RE(s,3) of Bickel et al. [2009] holds*

for the population covariance matrix Σ, namely

$$\kappa := \min_{J\subseteq\{1,\ldots,p\},|J|_0\leq s} \min_{u\neq 0,|u_{J^c}|_1\leq 3|u_J|_1} u'\Sigma u/|u|_2^2 > 0, \tag{2.3.8}$$

where J^c is the complement of the set J, i.e., *$J^c = \{1,2,...,p\}\backslash J$, u_J is defined as a modification of u by setting its elements outside J to zero. All the bounds (2.3.5), (2.3.6) and (2.3.7) still hold with high probability.*

Remark 2.3.2. *The best known convergence rate of Lasso estimators for i.i.d sub-Gaussian data requires that $K_\theta(\log p/n)^{1-\theta/2} \leq C$ for some positive constant C. Our theorems require that $K_\theta\omega\lambda^{-\theta} \leq C$, where*

$$\omega = \sqrt{\log p/n}M_X^2 + n^{2\nu/\gamma-1}(\log p)^{3/2}\||X_\cdot|_\infty\|_{\gamma,\alpha_X}^2,$$

and

$$\lambda \gtrsim \sqrt{\log p/n}M_e M_X + n^{\rho/\tau-1}(\log p)^{3/2}M_e\||X_\cdot|_\infty\|_{\gamma,\alpha_X}.$$

The second terms in ω and λ are introduced by the heavy tails, and thus are unavoidable. In other words, under heavy tailed distributions, sometimes, the allowed dimension p for Lasso methods can be at most a power of the sample size n.

In the exact sparsity case, we require

$$n \gtrsim M_X^4 s^2\log p + s^{1/(1-2\nu/\gamma)}(\log p)^{3/(2-4\nu/\gamma)}\||X_\cdot|_\infty\|_{\gamma,\alpha_X}^{2/(1-2\nu/\gamma)}.$$

One may argue the first term $M_X^4 s^2\log p$ can be further improved to $M_X^4 s\log p$ for short range temporal dependence data, in agreement with i.i.d. sub-Gaussian data. However, we cannot achieve it because even the optimal Bernstein type inequality for nonlinear weakly dependent data is still an open problem. The best known result is proposed by Merlevède et al. [2009].

Remark 2.3.3. *Based on Theorem 2.3.2, we have the following cases: Assume* $M_X \asymp 1$ *and* $M_e \asymp 1$. *Under the weak cross-sectional dependence* $\||X_\cdot|_\infty\|_{\gamma,\alpha_X} \asymp p^{1/\gamma}$, *which holds if the* p *components* x_{ij} $(1 \le j \le p)$ *are nearly independent, then the required sample size for exact sparsity is* $n \gtrsim s^2 \log p + s^{1/(1-2\nu/\gamma)} (\log p)^{3/(2-4\nu/\gamma)} p^{2/(\gamma-2\nu)}$ *and regularization parameter satisfies* $\lambda \gtrsim \sqrt{\log p/n} + n^{\rho/\tau-1} (\log p)^{3/2} p^{1/\gamma}$. *In comparison, Bonferroni Inequality and Lemma 1 in the Appendix would result in* $n \gtrsim s^2 \log p + s^{1/(1-2\nu/\gamma)} p^{4/(\gamma-2\nu)}$ *and* $\lambda \gtrsim \sqrt{\log p/n} + n^{\rho/\tau-1} p^{1/\tau}$.

In addition, under the strong cross-sectional dependence $\||X_\cdot|_\infty\|_{\gamma,\alpha_X} \asymp 1$, *which holds if the* p *components* x_{ij} $(1 \le j \le p)$ *are linear combinations of fixed random variables, the required sample size for exact sparsity is* $n \gtrsim s^2 \log p + s^{1/(1-2\nu/\gamma)} (\log p)^{3/(2-4\nu/\gamma)}$ *and the regularization parameter satisfies* $\lambda \gtrsim \sqrt{\log p/n} + n^{\rho/\tau-1} (\log p)^{3/2}$.

Next, we give an example for which the results of Theorem 2.3.1 apply.

Example 2.3.1. *Consider the autoregressive model with exogenous variables, that is, the ARX(*a, b*) model:*

$$Y_i = \sum_{l=1}^{a} \phi_l Y_{i-l} + \sum_{l=0}^{b} \psi_l' \mathbf{z}_{i-l} + e_i = \beta' X_i + e_i, \tag{2.3.9}$$

where a *and* b *are nonnegative integers,* e_i *follows a GARCH(1,1) model defined below, and* $\mathbf{z}_i$ *is a linear process defined by*

$$\mathbf{z}_i = \sum_{l=0}^{\infty} A_l \varepsilon_{i-l}, \tag{2.3.10}$$

where the random variables ε_{ij} *and coefficient matrices* A_l *are given in Example 1 with* $E|\varepsilon_{ij}|^\gamma < \infty$ *and* $\gamma > 2$. *Assume the roots of the polynomial* $1 - \sum_{l=1}^{a} \phi_l B^l$ *are outside the unit circle, which ensures stationarity of the autoregressive part of the model. Also assume the population covariance matrix* $\Sigma = \mathbb{E} X_i X_i'$ *is positive definite.*

Let

$$e_i = \sqrt{h_i}\eta_i, \qquad h_i = \pi_0 + \pi_1 e_{i-1}^2 + \pi_2 h_{i-1}, \tag{2.3.11}$$

with $\pi_0 > 0$, $\pi_1 \geq 0$, $\pi_2 \geq 0$ *and* $\mathbb{E}(\pi_1 + \pi_2\eta_{i-1}^2)^{q/2} < \infty$, $q > 4$. *Then it is easy to show* $\|e_{\cdot}\|_{q,\alpha_e} < \infty$.

Again, by Burkholder's inequality, $\|A_{l,j\cdot}\varepsilon_0\|_\gamma \leq \sqrt{\gamma - 1}|A_{l,j\cdot}|_2\|\varepsilon_{00}\|_\gamma$. *If there exist constants* $K_1 > 1$ *and* $\alpha_Z > 0$ *such that* $\max_{j\leq p}|A_{l,j\cdot}|_2 \leq K_1(l+1)^{-1-\alpha_Z}$ *holds for all* $l \geq 0$, *then we have* $\max_j \|z_{\cdot j}\|_{\gamma,\alpha_Z} \leq K_1K_2\|\varepsilon_{00}\|_\gamma$, *where the constant* K_2 *only depends on* α_Z *and* γ. *Together with the assumption that the roots of the polynomial* $1 - \sum_{l=1}^a \phi_l B^l$ *are outside the unit circle, we ensure* $\max_j \|X_{\cdot j}\|_{\gamma,\alpha_Z} < \infty$.

2.4 Model Selection Consistency

In this section, we extend the asymptotic properties of sign consistency for model selection via the Lasso to the dependent setting. The sign consistency of Lasso was first introduced by Zhao and Yu [2006]. Without loss of generality, write $\beta = (\beta_1, ..., \beta_s, ..., \beta_p)'$, where $\beta_j \neq 0$ if $j \leq s$ and $\beta_j = 0$ if $j > s$. That is, the first s predictors are relevant variables. Denote $\beta = (\beta_{(1)}', \beta_{(2)}')'$, where $\beta_{(1)}$ is a $s \times 1$ vector. Correspondingly, for any i, denote $X_i = (X_{i(1)}', X_{i(2)}')'$ and $\boldsymbol{X} = (X_1, ..., X_n)' = (\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)})$, where $\boldsymbol{X}_{(1)}$ is the $n \times s$ sub-matrix consisting of the relevant variables, and $\boldsymbol{X}_{(2)}$ is the $n \times (p-s)$ sub-matrix with the irrelevant ones. Similarly, consider the partition of the covariance matrix as

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where $\Sigma_{11} = \mathbb{E}X_{i(1)}X_{i(1)}'$ is a $s \times s$ sub-matrix associated with the relevant variables.

We impose the following assumptions.

Assumption 2.4.1. *For any* $1 \leq i \leq n$, $\mathbb{E}(X_{ik}|\boldsymbol{X}_{(1)}, \boldsymbol{e}) = \Sigma_{2k,1}\Sigma_{11}^{-1}X_{i(1)}$, *where* $\Sigma_{2k,1}$ *is the* k*-th row of* Σ_{21}.

Define $z_{ik} = X_{ik} - \mathbb{E}(X_{ik}|\boldsymbol{X}_{(1)}, \boldsymbol{e})$ for $s+1 \leq k \leq p$, and $\mathbf{z}_i = (z_{i,s+1}, ..., z_{i,p})'$.

Assumption 2.4.2. *There exists* $L > 0$ *such that* $\min_{1 \leq j \leq s} |\beta_j| \geq L$.

Assumption 2.4.3. *There exists a constant* $N_1 > 0$ *such that*

$$\inf_{|\zeta|_2=1} \zeta'\Sigma_{11}\zeta = N_1.$$

Assumption 2.4.4. *There exists a positive constant* $\eta \in (0, 1)$ *such that*

$$|\Sigma_{21}\Sigma_{11}^{-1} sign(\beta_{(1)})|_\infty \leq 1 - \eta. \tag{2.4.1}$$

Assumption 2.4.1 explicitly defines how the irrelevant variables depend on the relevant variables and the errors. Note that $\text{Cov}(\Sigma_{2k,1}\Sigma_{11}^{-1}X_{i(1)}, X_{ik} - \Sigma_{2k,1}\Sigma_{11}^{-1}X_{i(1)}) = 0$ always holds, for all $s+1 \leq k \leq p$. That is, $\Sigma_{2k,1}\Sigma_{11}^{-1}X_{i(1)}$ and $X_{ik} - \Sigma_{2k,1}\Sigma_{11}^{-1}X_{i(1)}$ are mutually uncorrelated. We further assume they are independent. Intuitively, $\mathbf{z}_i$ can be viewed as the unique part of irrelevant variables that cannot be explained by the relevant variables. Thus, for irrelevant variables, $\mathbf{z}_i$ is more representative than $X_{i(2)}$. Assumption 2.4.2 controls the lower bound of the non-zero parameters; see, for example, Bühlmann and Van De Geer [2011]. Assumption 2.4.3 imposes a lower bound, N_1, on the minimal eigenvalue of the covariance matrix of relevant variables. In practice, quantifying the rate under which N_1 decreases is difficult and problem specific, and it is frequently assumed constant, e.g., Medeiros and Mendes [2016] and Kock and Callot [2015]. Assumption 2.4.4 employs the strong irrepresentable condition of population covariance, which is similar to the condition in Zhao and Yu [2006].

To account for the cross-sectional dependence of the stationary process $(X_{i(1)})$ and $(\mathbf{z}_i)$, we also define the $\mathcal{L}^\infty$ functional dependence measure and its corresponding dependence

adjusted norm

$$\omega_{i,q,1} = \| \max_{1\le j\le s} |X_{ij} - X^*_{ij}| \|_q,$$

$$\| |X_{\cdot(1)}|_\infty \|_{q,\alpha} = \sup_{m\ge 0} (m+1)^\alpha \Omega_{m,q,1}, \quad \alpha \ge 0, \quad \text{and } \Omega_{m,q,1} = \sum_{i=m}^{\infty} \omega_{i,q,1}.$$

Additionally, we define

$$\Psi_{q,\alpha,1} = \max_{1\le j\le s} \|X_{\cdot j}\|_{q,\alpha} \quad \text{and} \quad \Upsilon_{q,\alpha,1} = \left(\sum_{j=1}^{s} \|X_{\cdot j}\|^q_{q,\alpha} \right)^{1/q}.$$

For $(\mathbf{z}_i)$, the quantities $\| |\mathbf{z}_\cdot|_\infty \|_{q,\alpha}$, $\Psi_{q,\alpha,2}$ and $\Upsilon_{q,\alpha,2}$ can be similarly defined. Clearly, $\Psi_{q,\alpha,1} \le \| |X_{\cdot(1)}|_\infty \|_{q,\alpha} \le \Upsilon_{q,\alpha,1}$ and $\Psi_{q,\alpha,2} \le \| |\mathbf{z}_\cdot|_\infty \|_{q,\alpha} \le \Upsilon_{q,\alpha,2}$.

Let $\sigma = \mathbb{E} e_i^2$. Define

$$\begin{aligned}
\delta_*(\lambda, N_1, \sigma) &= \frac{\lambda^2 s}{2nN_1} + \frac{2\sigma}{n}, \\
M(\delta_*, \eta, \iota, \gamma) &= \eta^{-1}\sqrt{\delta_* \log p} + \eta^{-1} n^{(\iota-1)/\gamma} \delta_*^{1/2} (\log p)^{3/2} \| |\mathbf{z}_\cdot|_\infty \|_{\gamma,\alpha_X}, \\
Q(\rho, \tau) &= \sqrt{n \log s} + n^{\rho/\tau} (\log s)^{3/2} \| |X_{\cdot(1)}|_\infty \|_{\gamma,\alpha_X}, \\
V_1(N_1) &= \frac{s^2 \log s}{N_1}, \\
V_2(N_1) &= \frac{1}{N_1} s (\log s)^{3/2} \| |X_{\cdot(1)}|_\infty \|^2_{\gamma,\alpha_X}.
\end{aligned}$$

These quantities are used in the following theorem.

Theorem 2.4.1 below extends the results of Zhao and Yu [2006] to random design linear model with dependent errors. In comparison, Medeiros and Mendes [2016] derived asymptotic properties of sign consistency for the adaptive Lasso, while our results apply to the original Lasso and do not need any assumptions on weights. Note that even for heavy-tail variables, our results show that if the dependence among $\mathbf{z}_i$ is strong, the allowed dimension p can be as large as some exponential of the sample size n; see Remark 2 for more details.

Theorem 2.4.1. *Suppose Assumptions 2.4.1, 2.4.2, 2.4.3 and 2.4.4 hold. Assume that* $\max_{1\le j\le p}\|X_{\cdot j}\|_{\gamma,\alpha_X} < C_\gamma < \infty$, *and* $\|e_\cdot\|_{q,\alpha_e} < C_q < \infty$, *where* $q,\gamma > 4$, $\alpha_X, \alpha_e > 0$, *constants* C_γ, C_q *only depend on* γ, q. *Define*

$$\nu = \begin{cases} 1 & \text{if } \alpha_X > 1/2 - 2/\gamma, \\ \gamma/4 - \alpha_X\gamma/2 & \text{if } \alpha_X < 1/2 - 2/\gamma, \end{cases}$$

and

$$\iota = \begin{cases} 1 & \text{if } \alpha_X > 1/2 - 1/\gamma, \\ \gamma/2 - \alpha_X\gamma & \text{if } \alpha_X < 1/2 - 1/\gamma. \end{cases}$$

Let $\alpha = \min(\alpha_X, \alpha_e)$. *Assume* $\tau = q\gamma/(q+\gamma) > 2$ *and define*

$$\rho = \begin{cases} 1 & \text{if } \alpha > 1/2 - 1/\tau, \\ \tau/2 - \alpha\tau & \text{if } \alpha < 1/2 - 1/\tau. \end{cases}$$

Furthermore, suppose $s = o(n)$. *Then, for any* λ *and the sample size* n *such that*

$$n \gtrsim V_1(N_1), \tag{2.4.2}$$

$$n^{1-2\nu/\gamma} \gtrsim V_2(N_2), \tag{2.4.3}$$

$$M(\delta_*, \eta, \iota, \gamma) + Q(\rho, \tau) \lesssim \lambda \le \frac{nN_1L}{4\sqrt{s}}, \tag{2.4.4}$$

the consistency probability $\mathbb{P}(\hat\beta =_s \beta)$ *is at least*

$$1 - C_1(\log p)^{-\gamma} - C_2(\log s)^{-\gamma/2} - C_3(\log s)^{-\tau} - C_4 p^{-C_5} - C_6 s^{-C_7} - \frac{\|e_\cdot\|_{q,\alpha_e}^q}{n^{q-1}\sigma^q} - \exp\left(-\frac{n\sigma^2}{\|e_\cdot\|_{2,\alpha_e}^2}\right). \tag{2.4.5}$$

Remark 2.4.1. *In particular, assume* $N_1 \asymp 1$, $\eta \asymp 1$. *Also assume the weak temporal dependence case* $\alpha_X > 1/2 - 1/\gamma$ *and* $\alpha > 1/2 - 1/\tau$. *If the dependence measure* $\||X_{\cdot(1)}|_\infty\|_{\gamma,\alpha_X} \asymp s^{1/\gamma}$ *and* $\||\mathbf{z}_\cdot|_\infty\|_{\gamma,\alpha_X} \asymp p^{1/\gamma}$, *which would hold if all the components* X_{ij}

($1 \le j \le s$) and z_{ik} ($s+1 \le k \le p$) are nearly independent, then (2.4.2), (2.4.3) and (2.4.4) reduce to

$$n \gtrsim s^2 \log s + s^{\frac{1+2/\gamma}{1-2/\gamma}} (\log s)^{\frac{3}{2-4/\gamma}} + sp^{2/\gamma} (\log p)^3$$

and

$$\sqrt{n \log s} + n^{1/\tau} s^{1/\tau} (\log s)^{3/2} \lesssim \lambda \lesssim \frac{nL}{\sqrt{s}}.$$

Additionally, if $s = O(n^{c_1})$ for some $c_1 < \min\{1/2, (\gamma-2)/(\gamma+2)\}$, then the valid regularization parameter λ has the range $n^{1/2} + n^{1/\tau + c_1/\gamma} \ll \lambda \ll n^{1-c_1/2} L$. The dimension p satisfies that $p \ll n^{\gamma(1-c_1)/2}$.

On the other hand, assume $\||X_{\cdot(1)}|_\infty\|_{\gamma,\alpha_X} \asymp s^{1/\gamma}$ and $\||\mathbf{z}_\cdot|_\infty\|_{\gamma,\alpha_X} \asymp 1$, that is, all the components z_{ik} ($s+1 \le k \le p$) are strongly dependent. Let $s = O(n^{c_1})$ for some $c_1 < \min\{1/2, (\gamma-2)/(\gamma+2)\}$, then the existence of regularization parameter λ requires $n^{1/2} + n^{1/\tau + c_1/\gamma} \ll \lambda \ll n^{1-c_1/2} L$. The dimension p satisfies $p \ll \exp\{n^{(1-c_1)/3}\}$.

Furthermore, if $\||X_{\cdot(1)}|_\infty\|_{\gamma,\alpha_X} \asymp 1$ and $\||\mathbf{z}_\cdot|_\infty\|_{\gamma,\alpha_X} \asymp 1$, $s = O(n^{c_1})$ for some $c_1 < 1/2$, then the existence of regularization parameter λ requires $n^{1/2} \ll \lambda \ll n^{1-c_1/2} L$, and the dimension p satisfies $p \ll \exp\{n^{(1-c_1)/3}\}$.

In summary, the allowed dimension p varies from $n^{\gamma(1-c_1)/2}$ to $\exp\{n^{(1-c_1)/3}\}$ depending on the cross-sectional dependence of $z_{ik}, s+1 \le k \le p$.

Note that if the assumptions in Example 2.3.1 hold, together with the Strong Irrepresentable Condition, the results of Theorem 2.4.1 continue to apply. In general, the Strong Irrepresentable Condition is non-trivial, particularly since we do not know $\text{sign}(\beta)$ a priori. Then, we need the Strong Irrepresentable Condition to hold for every possible combination of different signs and placement of zeros. We give a simple example below in which the Strong Irrepresentable Condition is guaranteed. All diagonal elements of Σ are assumed to be 1 which is equivalent to normalizing all covariates in the model to the same scale since

Strong Irrepresentable Condition is invariant under any common scaling of Σ.

Example 2.4.1. *Consider the following autoregressive model with exogenous variables:*

$$Y_i = \sum_{l=1}^{a} \phi_l Y_{i-l} + \psi \mathbf{z}_i + e_i = \beta' X_i + e_i, \tag{2.4.6}$$

where a is nonnegative finite integer, $\mathbf{z}_i$ is independent of e_i, and the errors e_i are homogeneous. Assume the roots of the polynomial $1 - \sum_{l=1}^{a} \phi_l B^l$ are outside the unity circle, which ensures stationarity of the autoregressive part of the model. Also assume $\Sigma = \mathbb{E} X_i X_i'$ is positive definite.

Furthermore, suppose β has s nonzero entries. Similar to Corollary 2 in Zhao and Yu [2006], Σ has 1's on the diagonal and bounded correlation $|\sigma_{jk}| \leq c/(2s-1)$ for a constant $0 < c < 1$ then Strong Irrepresentable Condition holds. In this case, we need autocorrelation of Y_i to be weak, and all the covariates $\mathbf{z}_i$ are slightly correlated.

Remark 2.4.2. *Lasso may fail in the presence of strong serial dependence. Consider two scalar Gaussian autoregressive, AR(3), models:*

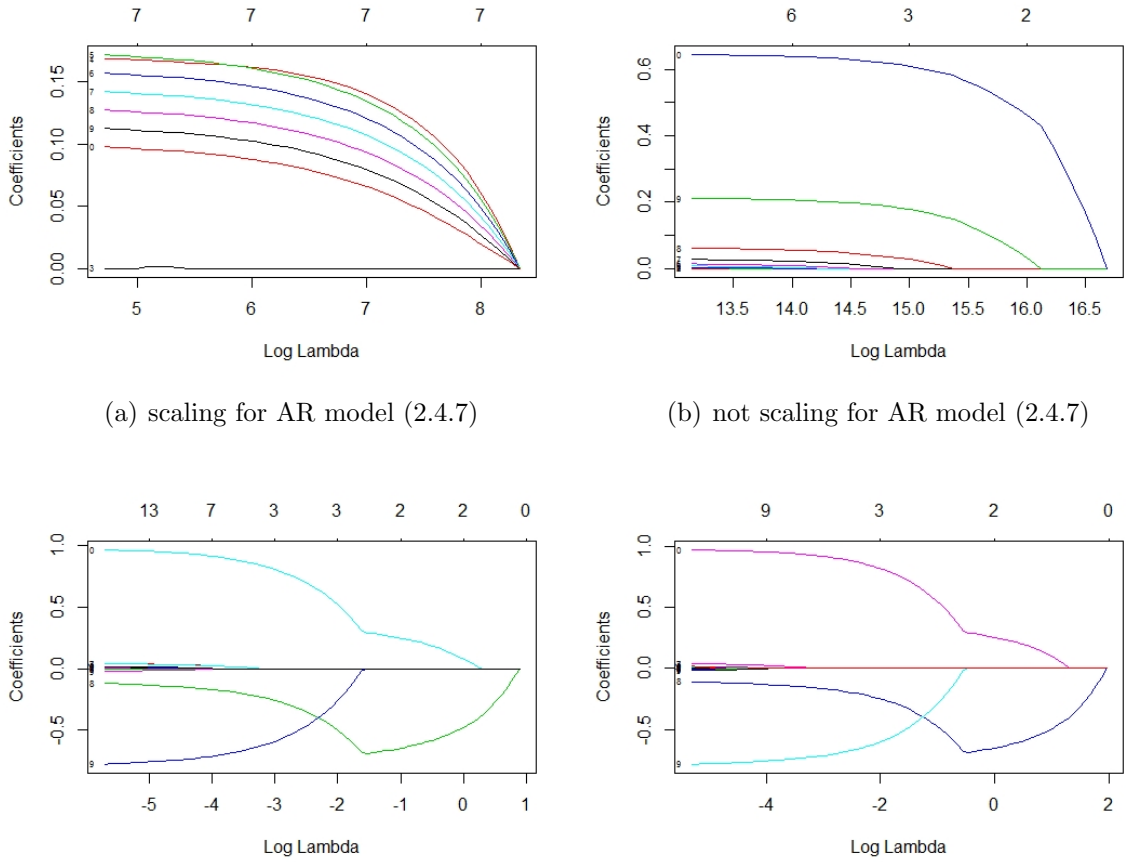
$$y_i = 1.9 y_{i-1} - 0.8 y_{i-2} - 0.1 y_{i-3} + e_i, \tag{2.4.7}$$

and

$$y_i = y_{i-1} - 0.8 y_{i-2} - 0.1 y_{i-3} + e_i, \tag{2.4.8}$$

where e_i follows the standard normal distribution. Then AR(3) model (2.4.7) is unit-root nonstationary, but model (2.4.8) is stationary. We generate 2000 observations from each of the two models. We choose $y_{i-10}, y_{i-9}, ..., y_{i-1}$, and $x_{1i}, ..., x_{10,i}$ as regressors, where x_{li} are i.i.d. standard normal. Figure 2.1 shows the model selection results for scaling vs. not scaling the predictors.

The default Lasso procedure standardizes each variable in y_i. For unit-root non-stationary



(a) scaling for AR model (2.4.7)

(b) not scaling for AR model (2.4.7)

(c) scaling for AR model (2.4.8)

(d) not scaling for AR model (2.4.8)

Figure 2.1: Results of Lasso regression for the two AR(3) series in (2.4.7) and (2.4.8) via the glmnet package of R.

time series, standardization might wash out the dependence of the stationary part; see Parts (a) and (b) of Figure 2.1. In this section, we only consider stationary time series for which scaling the predictors does not affect the estimation consistency of the Lasso estimates; see Parts (c) and (d) of Figure 2.1.

The following proposition shows a necessary and sufficient condition for a stationary AR(2) model under which the Strong Irrepresentable Condition (Assumption 2.4.4) holds. Similar results also hold for the general stationary AR(d) model.

Proposition 2.4.1. *Consider the stationary AR(2) model,*

$$y_i = \phi_1 y_{i-1} + \phi_2 y_{i-2} + e_i,$$

where e_i are i.i.d. random variates with mean zero and finite variance. We also normalize y_i such that the variance of y_i is 1. Then, the Strong Irrepresentable Condition (Assumption 2.4.4) holds if and only if

$$|\phi_1| + |\phi_2| < 1. \tag{2.4.9}$$

2.5 Simulation Study

In this section, we use simulation to demonstrate the performance of Lasso regression for dependent data in finite samples and to compare its efficacy with the mixed-frequency data sampling regression (MIDAS) commonly used in the econometric literature; see Ghysels et al. [2004]. In addition, we also compare the model selection consistency and parameter estimation of Lasso estimator and Dantzig estimator for dependent data in finite samples.

We first consider the following data generating process,

$$
\begin{aligned}
y_i &= \phi y_{i-1} + X_{i-1,1}^T \beta_s + e_i, \\
X_i &= \begin{bmatrix} X_{i,1} \\ X_{i,2} \end{bmatrix} = \sum_{j=1}^{m} A_j \begin{bmatrix} X_{i-j,1} \\ X_{i-j,2} \end{bmatrix} + \boldsymbol{\eta}_i,
\end{aligned} \tag{2.5.1}
$$

where $\phi = 0.6$ and each element of β_s is given by $\beta_{s,j} = \frac{1}{\sqrt{s}}(-1)^j$, $X_{i,1}$ is a $s \times 1$ vector of relevant variables. Let $\beta = (\beta_s, \beta_{s^c})$, where $\beta_{s^c} = \mathbf{0}$ is a $(p-s) \times 1$ vector. The errors e_i and $\boldsymbol{\eta}_{ij}$ are i.i.d random variables of Student-t distribution with 5 degrees of freedom, and e_i and $\boldsymbol{\eta}_i$ are all mutually uncorrelated. The explanatory variable process X_i, which has $p - s$ irrelevant variables, follows a vector autoregressive, VAR(m), model. The following two choices of X_i are considered, denoted as Model 1 and Model 2, respectively.

(a). **Model 1**: The explanatory process X_i is a VAR(4) process, where A_1 and A_4 assume a block-diagonal structure and $A_2 = A_3 = 0$. In particular, the first two and the last two blocks are 5×5 matrices with all entries of the blocks of A_1 equal to 0.15 and all entries of the blocks of A_4 equal to -0.1. The other blocks are 10×10 matrices with all elements of the blocks of A_1 equal to 0.075 and all elements of the blocks of A_4 equal to -0.05. This structure could be motivated by a model built for mixed-frequency data with some quarterly time series often encountered in macroeconomic analysis.

(b). **Model 2**: The explanatory process X_i follows a VAR(1) model, where A_1 is block-diagonal with the same block structure given by Model 1. The (j,k)th entry of the block is $(-1)^{|j-k|}\rho^{|j-k|+1}$ with $\rho = 0.4$. Hence, the entries decrease exponentially fast with their distances from the diagonal.

We employ sample sizes $n = 50, 100, 200$ with different choices of p and s. We set $p = 100, 200, 400$ and $s = 5, 10, 20$. For comparison, we also simulate a response series from a MIDAS model. In Model (2.5.1), for $s = 5, 10, 20$, let $\beta_s = \beta(1)$, $(\beta(1)^T, \beta(2)^T)^T$ or

$(\beta(1)^T, \beta(2)^T, \beta(3)^T)^T$ respectively, with

$$\beta_j(l) = \frac{\exp(\delta_1 j + \delta_2 j^2)}{\sum_{k=1}^{|\beta(l)|_0} \exp(2\delta_1 k + 2\delta_2 k^2)} \tag{2.5.2}$$

where $\beta(1)$ and $\beta(2)$ have 5 variables, $\beta(3)$ has 10 variables, and $\delta = (\delta_1, \delta_2)' = (0.5, -1)'$. All the other settings are the same as before. The two choices of X_i as in Models 1 and 2 are used, and we denote the resulting MIDAS models as Models 3 and 4, respectively. The models estimated by Lasso are with λ selected by the BIC; see Bühlmann and Van De Geer [2011]. The consistency of Lasso estimator selected by BIC was first proved by Zou et al. [2007] under the case $p < n$. Then, Tibshirani and Taylor [2012] studied the effective degrees of freedom of the Lasso when $p > n$. It is interesting to investigate the theoretical justification of the consistency of the BIC criterion for Lasso under the time series setting. We leave this to future work. We also employed models with λ selected by cross validation but found that cross-validation does not improve the results while being considerably much slower in computation. For the models estimated by MIDAS, we only consider Exponential Almon lag polynomial weighting scheme (see (2.5.2)) for the first 100 variables and impute the true values as initial values.

Table 2.1 shows the average of absolute error (AE), the average of root mean squared error (RMSE) for the Lasso estimators and MIDAS estimators over the 10,000 Monte Carlo simulations for the data generating processes used. The AE and the RMSE are defined as,

$$\begin{aligned} \text{AE} &= \frac{1}{MC} \sum_{l=1}^{MC} |(\hat{\phi}; \hat{\beta}) - (\phi; \beta)|_1, \\ \text{RMSE} &= \sqrt{\frac{1}{MC} \sum_{l=1}^{MC} |(\hat{\phi}; \hat{\beta}) - (\phi; \beta)|_2^2}, \end{aligned}$$

where MC denotes the number of Monte Carlo repetitions. From the table, it is clear that both the AE and RMSE measures show that the Lasso regression provides substantially more

accurate parameter estimation than the mixed-frequency data sampling regression (MIDAS) in the presence of irrelevant variables. Also, as expected, the AE and the RMSE of the estimators decrease with n, but increase with s and p.

To evaluate the performance of out-of-sample forecasts, we use the estimated parameters to compute one-step-ahead forecasts and consider a total of 10 out-of-sample predictions, denoted by $y_{n+1}, \ldots, y_{n+10}$. Table 2.2 shows the average absolute forecast error (AFE) and the average root mean squared forecast error (RMSFE) over the 10,000 Monte Carlo simulations, which are calculated as

$$\begin{aligned} \text{AFE} &= \frac{1}{10MC} \sum_{l=1}^{MC} \sum_{k=1}^{10} |\hat{y}_{n+k} - y_{n+k}|, \\ \text{RMSFE} &= \sqrt{\frac{1}{10MC} \sum_{l=1}^{MC} \sum_{k=1}^{10} |\hat{y}_{n+k} - y_{n+k}|^2}. \end{aligned}$$

The forecasting results in Table 2.2 show that the Lasso regression has smaller AE and RMSFE than the MIDAS in all settings. Furthermore, the results show clearly that the performance of the Lasso regression and the MIDAS improves with the sample size, but deteriorates as the number of relevant variables s increases. Finally, both AE and RMSFE of the Lasso regression decrease faster than those of MIDAS as the sample size n increases. As a matter of fact, the AE and the RMSFE of the MIDAS remain high even when $n = 200$. Since we only fit MIDAS through the first 100 variables, the performance of the MIDAS does not change as p increases. Overall, in the presence of irrelevant variables, the Lasso regression significantly outperforms the MIDAS regression.

Next, we compare the model selection and parameter estimation of Lasso estimator and Dantzig estimator for dependent data. We use the same data generating process (2.5.1), where $\phi = 0.6$ and $X_{i,1}$ is a $s \times 1$ vector of relevant variables. Here, we set each element of β_s by $\beta_{s,j} = 3(-1)^j$. Model 1 and Model 2 defined before are chosen for X_i. Table 2.3 shows the number of noise covariates that are selected (False Positive), the number of signal

Table 2.1: Accuracy in Parameter Estimation of Lasso Regression and Mixed-Frequency Data Sampling Regression. The results are based on 10,000 repetitions, where AE and RMSE denote the average of mean absolute errors and average of root mean square errors over Monte Carlo repetitions and parameters. In the table, s, p, and n denote the number of non-zero parameters, the dimension of regressors, and sample size, respectively.

s	n	Absolute Error (AE) $\times 10^2$						Root Mean Square Error (RMSE) $\times 10^2$					
		Lasso			MIDAS			Lasso			MIDAS		
		p											
		100	200	400	100	200	400	100	200	400	100	200	400
Model 1													
	50	2.44	2.64	2.84	6.63	6.64	6.67	3.08	3.75	4.55	3.73	4.44	5.29
5	100	1.89	2.07	2.22	6.24	6.26	6.28	2.79	3.44	4.21	3.64	4.33	5.15
	200	1.27	1.49	1.70	5.91	5.91	5.94	2.28	2.92	3.71	3.56	4.23	5.04
	50	4.60	4.99	5.30	8.26	8.31	8.32	3.66	4.45	5.38	4.09	4.88	5.80
10	100	3.69	4.11	4.39	7.86	7.88	7.90	3.36	4.17	5.10	4.02	4.78	5.69
	200	2.28	2.74	3.29	7.50	7.55	7.56	2.65	3.39	4.42	3.96	4.71	5.60
	50	7.83	8.81	8.93	10.76	10.82	10.83	4.08	5.00	6.00	4.42	5.26	6.26
20	100	6.56	7.33	7.70	10.38	10.43	10.44	3.84	4.75	5.77	4.35	5.18	6.16
	200	4.69	5.55	6.56	10.08	10.12	10.15	3.31	4.21	5.40	4.30	5.12	6.09
Model 2													
	50	0.95	1.14	1.38	4.95	4.97	4.99	2.02	2.53	3.19	3.31	3.94	4.70
5	100	0.54	0.60	0.67	4.55	4.58	4.58	1.56	1.92	2.36	3.18	3.79	4.50
	200	0.34	0.36	0.38	4.20	4.21	4.22	1.26	1.53	1.87	3.06	3.64	4.33
	50	1.91	2.40	2.92	5.46	5.46	5.46	2.54	3.26	4.18	3.53	4.20	4.99
10	100	1.06	1.24	1.46	5.03	5.07	5.08	1.92	2.41	3.04	3.39	4.04	4.81
	200	0.65	0.71	0.79	4.60	4.63	4.65	1.52	1.87	2.31	3.25	3.87	4.61
	50	3.19	4.21	4.94	6.12	6.15	6.18	2.95	3.85	4.96	3.76	4.48	5.34
20	100	1.75	2.14	2.59	5.68	5.69	5.70	2.23	2.85	3.64	3.63	4.32	5.15
	200	1.07	1.21	1.38	5.26	5.27	5.29	1.77	2.20	2.74	3.51	4.18	4.98
Model 3													
	50	1.71	2.05	2.43	6.57	6.60	6.62	2.62	3.29	4.11	3.74	4.46	5.30
5	100	0.93	1.06	1.21	6.27	6.31	6.33	2.03	2.54	3.18	3.65	4.35	5.18
	200	0.57	0.63	0.69	6.17	6.20	6.21	1.62	2.02	2.50	3.61	4.30	5.11
	50	3.74	4.47	5.07	8.41	8.44	8.46	3.34	4.17	5.16	4.13	4.92	5.86
10	100	2.06	2.52	3.00	8.20	8.24	8.25	2.59	3.32	4.25	4.08	4.85	5.77
	200	1.20	1.38	1.58	8.10	8.14	8.16	2.00	2.52	3.18	4.05	4.82	5.73
	50	7.23	8.77	9.38	11.02	11.07	11.09	3.90	4.88	5.95	4.47	5.32	6.32
20	100	4.45	5.81	7.01	10.92	10.97	11.00	3.22	4.16	5.32	4.43	5.28	6.28
	200	2.53	2.93	3.50	10.87	10.93	10.95	2.49	3.11	3.97	4.42	5.26	6.26
Model 4													
	50	1.39	1.58	1.78	5.14	5.16	5.16	2.49	3.10	3.83	3.47	4.13	4.90
5	100	0.96	1.05	1.12	4.58	4.59	4.59	2.12	2.63	3.22	3.31	3.95	4.69
	200	0.71	0.77	0.83	4.22	4.23	4.25	1.83	2.28	2.81	3.23	3.85	4.58
	50	2.40	2.79	3.14	6.03	6.07	6.10	2.90	3.64	4.54	3.80	4.53	5.39
10	100	1.67	1.86	2.02	5.50	5.53	5.55	2.47	3.08	3.80	3.69	4.39	5.23
	200	1.23	1.38	1.50	5.13	5.15	5.16	2.11	2.65	3.30	3.62	4.31	5.13
	50	3.68	4.58	4.97	6.88	6.93	6.93	3.22	4.09	5.14	4.06	4.84	5.75
20	100	2.43	2.77	3.06	6.38	6.42	6.45	2.71	3.41	4.25	3.96	4.72	5.62
	200	1.78	2.00	2.22	6.04	6.08	6.08	2.32	2.91	3.66	3.91	4.66	5.55

Table 2.2: Performance of Out-of-sample predictions of Lasso regression and mixed frequency data sampling regression (MIDAS). The results are based on 10 one-step ahead predictions and 10,000 iterations, where AFE and RMSFE denote the average absolute forecast errors and root mean squared forecast errors, respectively, and s, p, and n are the number of non-zero parameters, the dimension of regressors, and sample size. For MIDAS, the maximum p is fixed at 100.

s	n	Absolute Error (AE) $\times 10^2$						Root Mean Square Forecast Error (RMSFE) $\times 10^2$					
		Lasso			MIDAS			Lasso			MIDAS		
		p											
		100	200	400	100	200	400	100	200	400	100	200	400
Model 1													
	50	120.0	125.8	130.3	169.2	161.5	162.3	147.2	153.8	158.8	206.0	197.4	198.0
5	100	102.7	106.6	110.7	162.2	156.7	156.4	127.7	132.3	136.9	197.7	191.7	191.3
	200	86.9	90.6	95.4	156.8	152.4	153.2	109.7	114.1	119.4	191.4	186.7	187.4
	50	151.6	159.6	166.4	185.0	178.8	179.9	185.2	194.3	202.0	225.9	218.6	219.9
10	100	125.6	133.9	141.7	177.6	171.6	172.3	155.1	164.9	173.8	216.9	210.1	211.4
	200	96.0	101.9	112.2	171.9	167.7	168.2	120.3	127.2	139.3	210.2	206.1	206.3
	50	177.7	188.8	195.0	205.0	200.0	199.9	216.1	229.2	236.2	250.2	244.5	244.1
20	100	150.2	162.2	170.0	195.8	191.7	191.1	184.0	198.5	207.7	239.5	235.0	234.4
	200	118.6	128.7	145.1	190.1	185.9	188.2	146.7	159.2	178.4	232.4	228.4	230.5
Model 2													
	50	96.4	101.8	107.2	147.3	148.8	148.5	119.7	125.5	131.7	179.9	181.5	180.9
5	100	84.1	85.7	88.1	142.1	142.7	142.9	106.2	108.1	110.4	173.3	174.0	174.0
	200	78.4	79.6	80.6	138.6	137.6	138.8	99.9	101.5	102.3	169.0	168.1	169.2
	50	114.1	125.7	140.0	171.5	164.2	163.9	139.9	153.7	169.8	208.0	199.6	199.5
10	100	90.9	95.3	100.8	156.7	157.9	158.0	114.1	118.7	124.9	190.6	191.9	191.9
	200	81.7	83.1	85.3	151.4	151.1	151.9	103.7	105.1	107.6	184.1	183.7	184.5
	50	126.9	144.5	167.8	178.2	173.1	173.5	155.4	175.9	202.9	216.5	211.1	211.6
20	100	97.7	105.1	113.7	169.7	164.3	164.7	121.6	130.1	139.9	206.5	200.4	200.9
	200	85.3	87.9	91.9	161.7	157.1	158.0	107.8	110.5	115.1	196.9	191.9	192.9
Model 3													
	50	117.4	128.7	140.5	152.9	153.1	153.3	143.0	155.5	168.3	187.5	187.6	187.9
5	100	89.4	92.8	97.1	144.7	145.0	145.0	112.2	116.0	120.4	144.7	178.2	178.1
	200	80.6	81.2	82.8	142.3	141.0	141.1	102.3	103.0	105.0	174.7	173.6	173.5
	50	154.4	172.5	188.9	178.9	179.1	179.8	185.9	206.0	224.6	218.4	218.8	219.7
10	100	103.1	112.9	124.3	171.2	171.3	170.6	127.9	138.7	152.2	209.7	209.5	209.1
	200	84.7	88.2	91.0	166.9	168.2	167.4	107.1	111.1	114.3	204.7	205.9	205.3
	50	197.3	224.5	244.3	206.9	205.0	205.3	236.1	266.5	288.4	251.6	249.7	249.8
20	100	130.9	150.8	172.2	196.6	197.6	196.4	160.2	182.9	207.6	240.2	240.9	239.3
	200	97.0	101.2	109.0	193.1	193.8	193.7	121.2	125.9	134.8	236.5	237.4	237.1
Model 4													
	50	103.0	108.7	113.2	131.7	131.9	130.9	126.8	133.3	138.2	162.6	162.9	161.7
5	100	88.4	90.4	92.9	121.6	122.0	121.5	110.9	113.0	115.8	150.9	151.5	150.8
	200	81.3	82.6	83.4	118.0	117.1	116.8	103.3	104.4	105.4	147.0	145.7	145.2
	50	117.6	126.6	136.2	148.6	148.5	148.5	144.1	154.4	165.7	183.7	183.0	183.0
10	100	95.8	99.8	103.4	139.4	139.5	139.3	119.8	124.2	128.2	172.5	172.6	172.3
	200	84.9	87.5	89.7	134.2	135.0	134.7	107.3	110.1	112.7	166.5	167.3	167.0
	50	132.2	148.5	162.3	163.8	164.7	163.7	161.3	180.2	196.3	201.7	202.7	201.6
20	100	102.4	108.9	115.4	154.2	154.0	154.7	127.2	134.8	142.3	190.6	190.7	190.9
	200	88.6	92.1	96.2	150.0	149.8	150.2	111.7	115.7	120.2	185.8	185.4	185.7

covariates that are not selected (False Negative), the average of root mean squared error (RMSE) for the Lasso estimators and the Dantzig estimators over the 10,000 Monte Carlo simulations for the data generating processes used. As expected, False Positive and RMSE decrease with n, but increase with s and p. False Negative for both methods are almost the same. In terms of False Negative and RMSE, Lasso estimator substantially outperforms the Dantzig selector. Dantzig selector might be more sensitive to heavy tails and outliers, since Dantzig selector uses L_∞ norm. The rate of convergence for Lasso estimator in this chapter is faster than that for Dantzig selector in Wu and Wu [2016]. They built L_∞ type rate of convergence for Dantzig estimator, which is related to the unknown L_1 norm of true coefficients and matrix L_1 norm of population matrix. In this section, we overcome this weakness and achieve the same bounds for Lasso regression under i.i.d. data, but with different requirements for regularization parameter λ and sample size n.

2.6 Empirical Analysis

2.6.1 Predicting GDP growth

We consider the problem of predicting the growth rate of U.S. quarterly gross domestic product (GDP). In addition, nine (9) macroeconomic variables with different sampling frequencies are also available. The data are obtained from the St. Louis Federal Reserve Economic Data website. The predictive regression used is

$$y_i = \phi_0 + \phi_1 y_{i-1} + \cdots + \phi_a y_{i-a} + \sum_{l=1}^{9} \sum_{b=0}^{B_l} \beta_{l,b} z_{l,i\times m_l - b} + e_i \tag{2.6.1}$$

where a and B_l are nonnegative integers, Y_i is the growth rate (first difference of natural logarithm) of U.S. quarterly seasonally adjusted real GDP and $z_{l,\cdot}$'s are the high-frequency covariates with frequency m_l, e.g., $m_l = 3$ for monthly data. The nine covariates considered in this study are: $z_{1,\cdot}$ is the change of monthly civilian unemployment rates, $z_{2,\cdot}$ is the

Table 2.3: Accuracy in Model Selection and Parameter Estimation of Lasso Estimator and Dantzig Estimator for Linear Regression. The results are based on 10,000 repetitions, where RMSE denote the average of root mean square errors over Monte Carlo repetitions and parameters. In the table, s, p, and n denote the number of non-zero parameters, the dimension of regressors, and sample size, respectively.

s	n	Model 1						Model 2					
		Lasso			Dantzig			Lasso			Dantzig		
		p											
		100	200	400	100	200	400	100	200	400	100	200	400
False Negative													
	50	0.077	0.20	0.67	0.072	0.28	0.73	0	0	0.003	0	0	0.002
5	100	0	0	0.01	0	0	0.04	0	0	0	0	0	0
	200	0	0	0	0	0	0	0	0	0	0	0	0
	50	0.67	2.07	4.28	0.81	2.50	4.04	0.011	0.14	0.79	0.045	0.17	0.95
10	100	0	0.004	0.11	0	0.006	0.15	0	0	0	0	0	0
	200	0	0	0	0	0	0	0	0	0	0	0	0
	50	4.65	7.45	9.61	4.83	7.29	10.1	1.94	5.70	7.18	2.48	5.45	8.34
20	100	0	0.21	2.27	0.03	0.28	2.14	0	0	0.029	0	0.002	0.052
	200	0	0	0	0	0	0	0	0	0	0	0	0
False Positive													
	50	10.9	15.0	22.5	15.85	24.20	32.0	5.25	8.01	11.5	6.45	13.3	21.7
5	100	5.30	8.97	13.6	8.03	14.11	19.8	2.05	3.02	4.00	4.14	7.43	9.91
	200	1.63	3.01	4.95	3.66	6.25	9.52	0.49	0.79	1.31	2.88	3.56	5.03
	50	13.7	23.2	29.1	19.2	32.5	37.5	10.8	18.2	23.5	13.8	23.4	32.3
10	100	8.69	15.2	23.8	12.4	23.6	30.9	4.60	7.46	9.52	9.01	13.5	18.9
	200	2.12	4.96	7.24	4.02	8.17	10.6	1.08	2.05	3.68	3.37	6.09	10.75
	50	17.6	26.9	31.8	28.9	37.4	39.8	16.5	25.3	30.0	21.3	28.8	37.3
20	100	12.0	23.1	25.0	16.6	30.2	30.5	8.21	16.2	23.9	14.0	24.7	31.1
	200	3.99	7.01	10.6	5.84	10.2	15.1	2.49	4.05	8.46	8.22	11.3	16.2
RMSE													
	50	1.78	2.63	3.88	2.06	2.76	4.07	0.80	0.98	1.17	0.88	1.04	1.21
5	100	0.87	1.04	1.19	0.95	1.02	1.27	0.44	0.48	0.54	0.50	0.52	0.57
	200	0.69	0.64	0.70	0.61	0.80	0.83	0.33	0.33	0.34	0.42	0.39	0.41
	50	4.53	7.49	9.22	5.50	7.79	9.29	1.83	3.02	5.67	2.43	3.74	6.39
10	100	1.52	1.76	2.78	1.59	2.00	2.41	0.76	0.84	0.96	0.90	0.96	1.09
	200	0.97	1.01	1.09	0.94	1.21	1.24	0.55	0.56	0.58	0.69	0.70	0.68
	50	10.6	13.3	14.4	11.1	13.3	14.5	8.48	12.8	15.1	9.58	13.0	15.3
20	100	2.61	4.06	8.25	3.53	5.55	8.95	1.46	1.81	2.71	2.21	2.67	3.98
	200	1.46	1.65	1.78	1.69	1.76	1.91	0.91	0.94	1.00	1.13	1.22	1.31

growth rate of monthly all employees total payrolls, $z_{3,\cdot}$ is the growth rate of monthly industrial production total index, $z_{4,\cdot}$ is the growth rate of monthly consumer price index, $z_{5,\cdot}$ is the growth rate of monthly Moody's Seasoned Baa Corporate Bond Yields, $z_{6,\cdot}$ is the change of daily 3-Month Treasury Bill Secondary Market Rate, $z_{7,\cdot}$ is the change of daily 10-Year Treasury Constant Maturity Rate, $z_{8,\cdot}$ is the change of daily NASDAQ Composite Index, and $z_{9,\cdot}$ is the change of daily Wilshire 5000 Total Market Full Cap Index. The transformations of all variables are based on those of Stock and Watson [2002]. Note that all data are seasonally adjusted if necessary, and the explanatory variables are monthly or daily data. For daily variables $z_{6,\cdot}$ and $z_{7,\cdot}$, we only use data of the first 16 trading days in a month. For daily variables $z_{8,\cdot}$ and $z_{9,\cdot}$, we only use data of the first 15 trading days. The sampling period was from January 1980 to February 2017, but the prediction origin started with the second quarter of 2013 and ended with the first quarter of 2017. There was no trading activities during weekends and holidays, and there exist some missing data in the trading activities. Trading days for each month varies. We choose the first 15 or 16 trading days simply because they are the minimum number of trading days available for each month (mainly February).

Two types of empirical analysis are entertained. First, we consider a linear model with all explanatory variables and estimated by the Lasso procedure. For comparison, we include a model with all explanatory variables except the NASDAQ Composite Index and Wilshire 5000 Total Market Full Cap Index, estimated by the MIDAS regression (denoted by MIDAS-B model), a model with monthly all-employees total payrolls as the only explanatory variable, also estimated by MIDAS (denoted by MIDAS-A model), and a simple ARMA model of the GDP growth rates (denoted by ARMA model). We use BIC to select the number of autoregressive lags (a) and the lags (B_l) of explanatory variables. The Lasso tuning parameter λ is also chosen by the BIC; see Bühlmann and Van De Geer [2011]. Here we aggregate daily explanatory variables z_6 and z_7 to weekly frequency for the MIDAS regression.

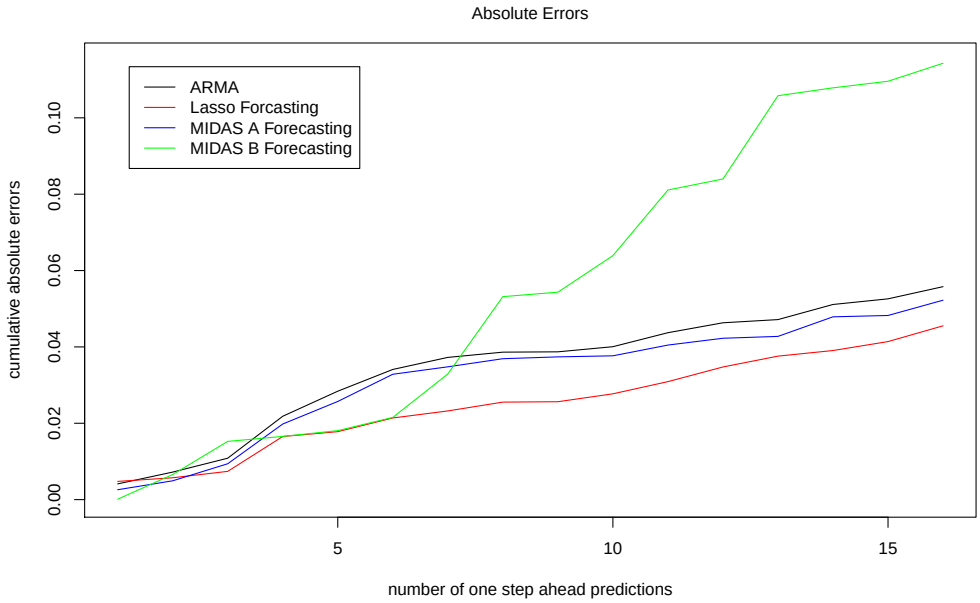
Table 2.4 shows the median absolute deviation (MAD), the mean absolute error (MAE), and the root mean squared error (RMSE) for the prediction period. From the table, it is clear that the Lasso based model outperforms all the other models in this particular instance. The poor performance of MIDAS-B is likely due to using too many explanatory variables with multiple sampling frequencies.

Figure 2.2 displays the cumulative absolute errors and the cumulative squared errors for different models in predicting the GDP growth rate. It shows clearly that the Lasso model is the best one. The MIDAS-A model also improves the prediction errors over the simple ARMA model. However, the MIDAS-B model fares poorly. Consequently, unlike the Lasso model, the MIDAS regression is not robust to the presence of irrelevant regressors. In fact, the MIDAS regression is also sensitive to the weighting schemes and the starting points of its optimization program.

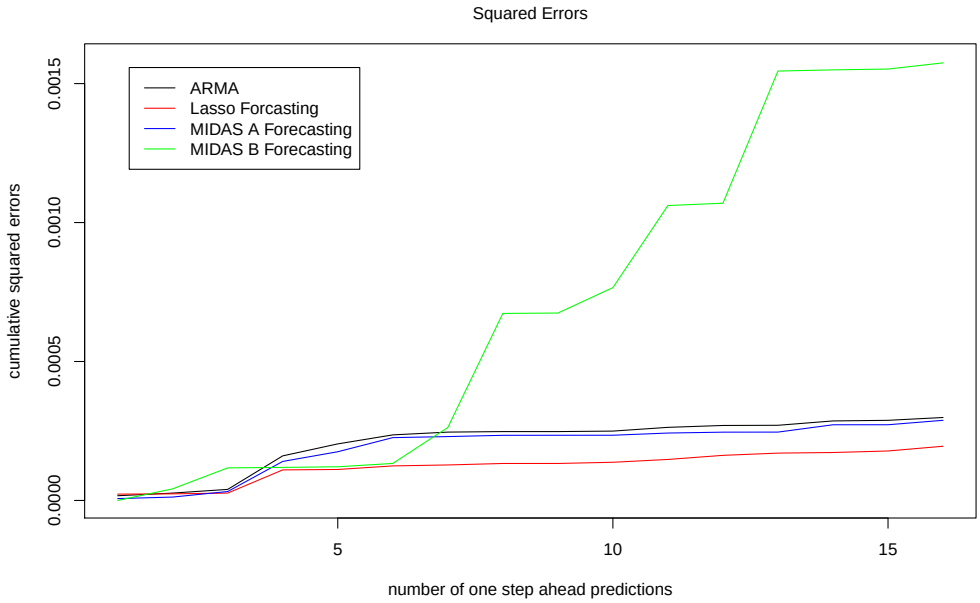
Table 2.4: Results of out-of-sampling prediction of U.S. quarterly real GDP growth rate. The data span is from 1980 to February 2017, but the forecast origins start from the second quarter of 2013 to the first quarter of 2017. All measurements are multiplied by 10^3. In the table, MAD, MAE, and RMSE are the median absolute error, mean absolute error, and root mean squared error, respectively.

Model	MAD	MAE	RMSE
ARMA	3.175	3.486	4.319
Lasso	2.328	2.845	3.491
MIDAS-A	2.463	3.264	4.245
MIDAS-B	4.089	7.143	9.920

Turn to comparison between forecasting and now-casting. Recall that the goal of now-casting is to take advantages of available high-frequency data to improve the prediction of lower-frequency variables of interest. For the quarterly GDP growth rate, during the quarter of interest, some monthly macroeconomic variables and even some daily economic variables become available, now-casting attempts to update the GDP prediction by incorporating those newly available high-frequency explanatory variables. In this exercise, we consider now-casting with the first month data within the quarter available and the first two months



(a)



(b)

Figure 2.2: Panel (a): Cumulative absolute errors. Panel (b): Cumulative squared errors. MIDAS-A represents the MIDAS regression model using only monthly all-employees total payrolls as the explanatory variable. MIDAS-B represents the MIDAS regression model with seven regressors $z_{1,\cdot}, \cdots, z_{7,\cdot}$, where $z_{6,\cdot}$ and $z_{7,\cdot}$ are aggregated into weekly data.

data available.

For comparison purpose, we employ an autoregressive (AR) model

$$y_i = \phi_0 + \phi_1 y_{i-1} + \cdots + \phi_a y_{i-a} + \epsilon_i, \tag{2.6.2}$$

as a benchmark for prediction. The AR order is selected by the BIC in the modeling subsample and is assumed to be fixed in the forecasting subsample. The AR model in Equation (2.6.2) is estimated by two ways. First, it is estimated by the ordinary least squares method and we denote the model by AR-OLS. Second, assuming sparsity, we estimate the AR model via Lasso method with the tuning parameter λ also selected by BIC. The forecasting result of this model is denoted by AR-Lasso. These two models represent the performance of forecasting.

For now-casting, we augment the AR model in Equation (2.6.2) with all explanatory variables available in the first month of the quarter and denote the results by Now-casting 1. Similarly, if we augment the AR model with all explanatory variables available in the first two months of the quarter, then the results are denoted by Now-casting 2. Specifically, for now-casting, we employ the model

$$y_i = \phi_0 + \phi_1 y_{i-1} + \cdots + \phi_a y_{i-a} + \beta^T X_i + \epsilon_i,$$

where X_i denotes the available high-frequency explanatory variables. For Now-casting 1, X_i consists of data of the first month into a given quarter whereas for Now-casting 2, it consists of data of the first two months into a given quarter. In this exercise, we use all monthly and daily high-frequency variables $z_{1,.}, \cdots, z_{9,.}$. We denote the results for MIDAS regression as MIDAS-C Now-casting 1 and MIDAS-C Now-casting 2, respectively. Finally, we also employ a MIDAS regression that only uses explanatory variables $z_{1,.}, \cdots, z_{7,.}$ in the now-casting and denote the results as MIDAS-D.

Table 2.5 summarizes the performance of now-casting in predicting U.S. quarterly GDP

growth rates in the forecasting period. From the table, we make the following observations. First, as expected, now-casting fares better than forecasting. The only exception is MIDAS-D now-casting. Second, also as expected, Now-casting 2 shows some improvement over Now-casting 1 for a given model. Keep in mind, however, Now-casting 1 is available one month into a quarter whereas Now-casting 2 needs to wait for an additional month. Third, from the performance of MIDAS-C and MIDAS-D, the stock market indexes do not seem to be helpful in predicting the GDP growth rate. In real applications, there exist many high-frequency explanatory variables, but their contributions to predicting the low-frequency variable of interest in unknown a priori. In this situation, the results obtained in this section suggest that the Lasso regression could be helpful.

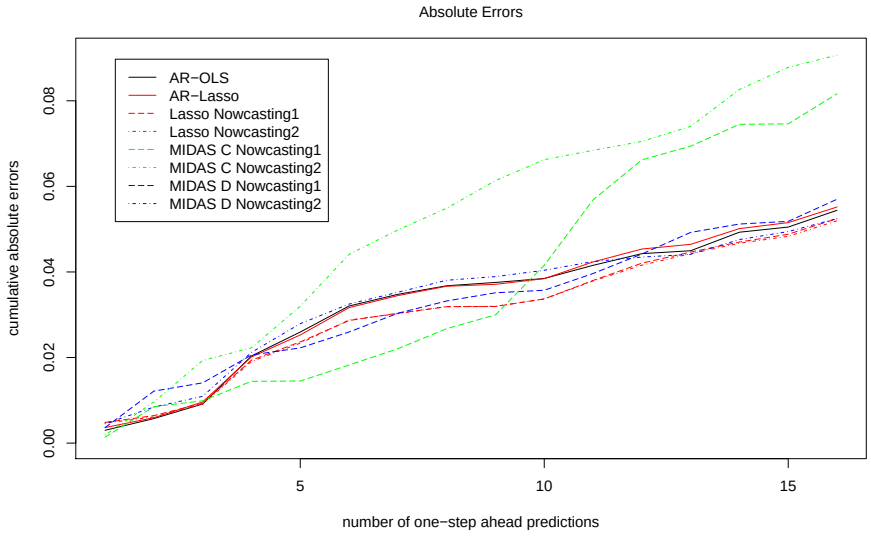
Figure 2.3 shows both the Lasso model and the MIDAS-B model improve the prediction via now-casting. But when irrelevant variables exist, MIDAS regression might encounter some difficulties.

Table 2.5: Comparison between forecasting and now-casting in predicting the U.S. quarterly real GDP growth rate. The data span is from 1980 to February 2017, but the forecast origins are from the second quarter of 2013 to the first quarter of 2017. All measurements are multiplied by 10^3. In the table, MAD, MAE, RMSE are the median absolute deviation, mean absolute error, and root mean squared error, respectively.

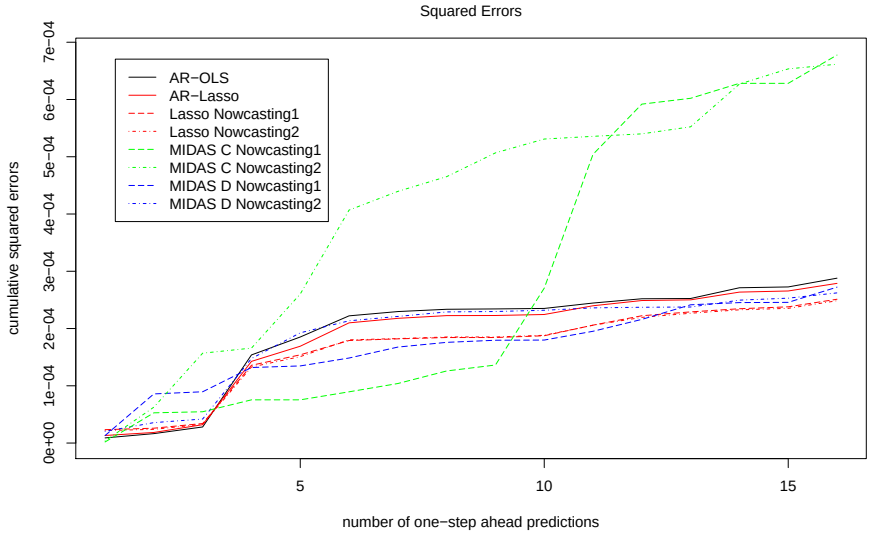
Model	MAD	MAE	RMSE
AR-OLS	2.865	3.400	4.242
AR-Lasso	3.327	3.448	4.174
Lasso Now-casting 1	2.731	3.278	3.962
Lasso Now-casting 2	2.834	3.247	3.941
MIDAS-C Now-casting 1	4.181	5.102	6.507
MIDAS-C Now-casting 2	5.108	5.666	6.430
MIDAS-D Now-casting 1	3.670	3.561	4.125
MIDAS-D Now-casting 2	2.784	3.279	4.048

2.6.2 Nowcasting $PM_{2.5}$

Consider next the prediction of $PM_{2.5}$. The response y is the square-root transformed daily maximum of $PM_{2.5}$. Hourly data of a monitoring station in the southern part of Taiwan



(a)



(b)

Figure 2.3: Panel (a): Cumulative absolute errors. Panel (b) Cumulative squared errors. MIDAS-D represents the MIDAS regression model with seven regressors $z_{1,.}, ..., z_{7,.}$. MIDAS-C represents the MIDAS regression model with nine regressors $z_{1,.}, ..., z_{9,.}$. Now-casting 1 and Now-casting 2 represent predicting quarterly GDP growth rate when the first month and the first two months data are available, respectively.

are used. To see the nowcasting effects, we consider adding 6 covariates, which are the first 6 hourly $PM_{2.5}$ readings of the same day, starting from midnight. The time period is from 2006 to 2015 so that there are 3650 observations. (Feb 29 was dropped.) We reserve the last 730 data points (2 years) for one-step ahead out-of-sample forecasts.

For comparison purpose, we first consider the square-root $PM_{2.5}$ (i.e. response y) as a pure time series. An AR(22) model is selected. Thus, the baseline model is a univariate AR(22). We denote the model by AR-OLS. For now-casting, we augment the AR model by the first 6 hourly readings. If we augment the AR model with the first hourly $PM_{2.5}$ reading, then the results are denoted by Now-casting 1. Similarly, if we augment the AR model with the first two hourly $PM_{2.5}$ readings, then the results are denoted by Now-casting 2, so on so forth. We denote the results for autoregressive model with exogenous variables as ARX Now-casting 1, ARX Now-casting 2, etc. We use BIC to select the number of autoregressive lags. The Lasso tuning parameter λ is also chosen by the BIC.

Table 2.6 summarizes the performance of now-casting in predicting daily maximum of $PM_{2.5}$. From the table, we make the following observations. First, as expected, now-casting outperforms forecasting. Second, also as expected, for a given model, Now-casting 2 shows some improvement over Now-casting 1, Now-casting 3 shows some improvement over Now-casting 2, so on so forth. Third, of most interest, Lasso estimator significantly outperforms the ARX model and the benchmark model. In short, Lasso regression seems to be helpful in applying now-casting to $PM_{2.5}$.

2.7 Deferred Proofs

2.7.1 Lemmas

We start with some lemmas that are useful in deriving the main results of the paper.

Lemma 2.7.1. *Assume that* $\|e_\cdot\|_{q,\alpha} < \infty$, *where* $q > 2$ *and* $\alpha > 0$, $\sum_{i=1}^n w_i^2 = n$. *Let* $w = (w_1, ..., w_n)$, $\varsigma_n = 1$ *(resp.* $(\log n)^{1+2q}$ *or* $n^{q/2-1-\alpha q}$*) if* $\alpha > 1/2 - 1/q$ *(resp.* $\alpha = 0$ *or*

Table 2.6: Comparison between forecasting and now-casting in predicting the daily maximum of $PM_{2.5}$. The data span is from 2006 to 2015, and the forecast origins are from 2013 to the end of 2015.(Feb 29 was dropped). In the table, MAE, RMSE are the mean absolute error, and root mean squared error for one-step ahead predictions, respectively.

Model	MAE	RMSE
AR-OLS	1619.6	73.71
ARX Now-casting 1	975.9	46.54
ARX Now-casting 2	940.9	44.92
ARX Now-casting 3	904.2	43.40
ARX Now-casting 4	879.7	42.31
ARX Now-casting 5	850.6	41.24
ARX Now-casting 6	835.3	40.31
Lasso Now-casting 1	659.3	31.74
Lasso Now-casting 2	628.4	30.58
Lasso Now-casting 3	623.2	30.72
Lasso Now-casting 4	600.7	29.63
Lasso Now-casting 5	595.0	29.49
Lasso Now-casting 6	576.3	28.47

$\alpha < 1/2 - 1/q$). *Then for all* $x > 0$, $S_n = \sum_{i=1}^n w_i e_i$,

$$\mathbb{P}(|S_n| \geq x) \leq K_1 \frac{\varsigma_n |w|_q^q \|e_\cdot\|_{q,\alpha}^q}{x^q} + K_2 \exp\left(-\frac{K_3 x^2}{n\|e_\cdot\|_{2,\alpha}^2}\right)$$

where K_1, K_2, K_3 *are constants that depend only on* q *and* α.

Proof. See Wu and Wu [2016] Theorem 2. □

Lemma 2.7.2. *Assume* $\||\mathbf{x}_\cdot|_\infty\|_{q,\alpha} < \infty$, *where* $q > 2$ *and* $\alpha > 0$, *and* $\Psi_{2,\alpha} < \infty$, $\sum_{i=1}^n w_i^2 = n$. *Let* $w = (w_1, ..., w_n)$ *and* $T_n = \sum_{i=1}^n w_i \mathbf{x}_i$. *(i) If* $\alpha > 1/2 - 1/q$, *then for* $x \gtrsim \sqrt{n \log p}\Psi_{2,\alpha} + |w|_q (\log p)^{3/2} \||\mathbf{x}_\cdot|_\infty\|_{q,\alpha}$,

$$\mathbb{P}(|T_n|_\infty \geq x) \leq \frac{K_{q,\alpha}|w|_q^q (\log p)^{q/2} \||\mathbf{x}_\cdot|_\infty\|_{q,\alpha}^q}{x^q} + K_{q,\alpha} \exp\left(-\frac{K_{q,\alpha} x^2}{n\Psi_{2,\alpha}^2}\right). \tag{2.7.1}$$

(ii) If $0 < \alpha < 1/2 - 1/q$, *then for* $x \gtrsim \sqrt{n \log p}\Psi_{2,\alpha} + n^{1/2-\alpha-1/q}|w|_q (\log p)^{3/2} \||\mathbf{x}_\cdot|_\infty\|_{q,\alpha}$,

$$\mathbb{P}(|T_n|_\infty \geq x) \leq \frac{K_{q,\alpha} n^{q/2-1-\alpha q}|w|_q^q (\log p)^{q/2} \||\mathbf{x}_\cdot|_\infty\|_{q,\alpha}^q}{x^q} + K_{q,\alpha} \exp\left(-\frac{K_{q,\alpha} x^2}{n\Psi_{2,\alpha}^2}\right), \tag{2.7.2}$$

where $K_{q,\alpha}$ is a constant that depends on q and α only.

Proof. The lemma can be shown following similar arguments as those in the proof of Zhang and Wu [2017] Theorem 6.2. Details are omitted. □

Lemma 2.7.3. *Let A and B denote two positive semi-definite, s-dimensional square matrices. If $\max_{1\le j,k\le s}\left|A_{jk}-B_{jk}\right|\le\delta$, then $\inf_{|\zeta|_2=1}\zeta' B\zeta > \inf_{|\zeta|_2=1}\zeta' A\zeta - s\delta$.*

Proof. See Lemma 3 of Medeiros and Mendes [2016]. □

Lemma 2.7.4. *For linear model $Y = X\beta+e$, assume that the matrix $X_{(1)}^T X_{(1)}$ is invertible. Then for any given $\lambda>0$, and any noise term $e\in\mathbb{R}^n$, there exists a Lasso estimator $\hat\beta(\lambda)$ which satisfies $\hat\beta(\lambda)=_s\beta$, if and only if the following two conditions hold*

$$sign\left(\beta_{(1)}+(\frac{1}{n}X_{(1)}^TX_{(1)})^{-1}\left[\frac{1}{n}X_{(1)}^Te-\lambda sign(\beta_{(1)})\right]\right)=sign(\beta_{(1)}),$$
$$\left|X_{(2)}^TX_{(1)}(X_{(1)}^TX_{(1)})^{-1}\left[\frac{1}{n}X_{(1)}^Te-\lambda sign(\beta_{(1)})\right]-\frac{1}{n}X_{(2)}^Te\right|\le\lambda,$$

where the vector inequality and equality are taken elementwise, $\beta_{(1)}$ and $\beta_{(2)}$ denote the first s and last $p-s$ entries of β respectively.

Proof. See Wainwright [2009]. □

2.7.2 A general theorem of estimation error for weak sparsity

Lemma 2.7.5. *Define $\hat\Delta=\hat\beta-\beta$, where β satisfies weakly sparsity condition (Assumption 1),* i.e., *$\sum_{j=1}^p|\beta_j|^\theta\le K_\theta$ for $0\le\theta<1$. Suppose $\hat\Delta\hat\Sigma\hat\Delta\ge\kappa|\hat\Delta|_2^2$, where κ is a positive constant that does not depend on $\hat\Delta$. Choose $\lambda\ge 2|n^{-1}\sum_{i=1}^n\mathbf{x}_ie_i|_\infty$. Then we have for some constants C_1, C_2,*

$$|\hat\Delta|_2^2\le C_1K_\theta\left(\frac{\lambda}{\kappa}\right)^{2-\theta}, \tag{2.7.3}$$

$$|\hat\Delta|_1\le C_2K_\theta\left(\frac{\lambda}{\kappa}\right)^{1-\theta}. \tag{2.7.4}$$

This result is deterministic and non-asymptotic. The statistical performance of $\hat{\beta}$ relies on the restricted eigenvalue condition properties of sample covariance $\hat{\Sigma}$.

Proof. This result is just a simple application of the theoretical framework established in Negahban et al. [2012], for the sake of brevity, we omitted the detailed proof here. □

2.7.3 Proof of Theorem 2.3.1

Proof. Recall $\hat{\Sigma} = (\hat{\sigma}_{jk})_{1\leq j,k\leq p} = 1/n \sum_{i=1}^{n} x_i x_i^T = n^{-1} X^T X$, $\Sigma = (\sigma_{jk})_{1\leq j,k\leq p}$. Define the events

$$\mathcal{A} = \{|\hat{\Sigma} - \Sigma|_\infty \leq a\} = \{\max_{j,k} |\hat{\sigma}_{jk} - \sigma_{jk}| \leq a\}, \tag{2.7.5}$$

$$\mathcal{B} = \{n^{-1} \left| X^T e \right|_\infty \leq \lambda/2\}. \tag{2.7.6}$$

The first step is to control the probability $\mathbb{P}(\mathcal{A}^c)$ and $\mathbb{P}(\mathcal{B}^c)$. By Hölder's inequality, we have for $m \geq 0$ that

$$\begin{aligned} \sum_{l=m}^{\infty} \|x_{lj} e_l - x_{lj}^* e_l^*\|_\tau &\leq \sum_{l=m}^{\infty} \left(\|x_{lj}(e_l - e_l^*)\|_\tau + \|(x_{lj} - x_{lj}^*) e_l^*\|_\tau \right) \\ &= \sum_{l=m}^{\infty} \left(\|x_{lj}\|_\gamma \|e_l - e_l^*\|_q + \|x_{lj} - x_{lj}^*\|_\gamma \|e_l^*\|_q \right). \end{aligned}$$

Since $\alpha = \min(\alpha_X, \alpha_e)$, the dependence adjusted norm satisfies

$$\|x_{.j} e_.\|_{\tau,\alpha} \leq \|x_{.j}\|_{\gamma,0} \|e_.\|_{q,\alpha_e} + \|x_{.j}\|_{\gamma,\alpha_X} \|e_.\|_{q,0} \leq 2\|x_{.j}\|_{\gamma,\alpha_X} \|e_.\|_{q,\alpha_e}. \tag{2.7.7}$$

Similarly, we have

$$\|x_{.j} x_{.k} - \sigma_{jk}\|_{\gamma/2,\alpha_X/2} \leq 2\|x_{.j}\|_{\gamma,\alpha_X} \|x_{.k}\|_{\gamma,\alpha_X}, \tag{2.7.8}$$

Hence,

$$\max_{1\le j\le p}\|x_{\cdot j}e_{\cdot}\|_{\tau,\alpha}\le 2M_eM_X, \tag{2.7.9}$$

$$\max_{1\le j,k\le p}\|x_{\cdot j}x_{\cdot k}-\sigma_{jk}\|_{\gamma/2,\alpha_X/2}\le 2M_X^2. \tag{2.7.10}$$

Employing a similar derivation, we can show that,

$$\|\max_{1\le j\le p}|x_{\cdot j}e_{\cdot}|\|_{\tau,\alpha}\le 2\||\mathbf{x}_{\cdot}|_\infty\|_{\gamma,\alpha_X}M_e, \tag{2.7.11}$$

$$\|\max_{1\le j,k\le p}|x_{\cdot j}x_{\cdot k}-\sigma_{jk}|\|_{\gamma/2,\alpha_X/2}\le 2\||\mathbf{x}_{\cdot}|_\infty\|^2_{\gamma,\alpha_X}. \tag{2.7.12}$$

Note that $M_X\le\||\mathbf{x}_{\cdot}|_\infty\|_{\gamma,\alpha_X}\le\Upsilon_{\gamma,\alpha_X}$.

If $\tau>2$, for $\lambda\gtrsim\sqrt{\log p/n}M_eM_X+n^{\rho/\tau-1}(\log p)^{3/2}M_e\||\mathbf{x}_{\cdot}|_\infty\|_{\gamma,\alpha_X}$, adopting (2.7.9), (2.7.11) and Lemma 2.7.2, we have,

$$\mathbb{P}(\mathcal{B}^c)=C_4\frac{n^\rho(\log p)^{\tau/2}\||\mathbf{x}_{\cdot}|_\infty\|^\tau_{\gamma,\alpha_X}M_e^\tau}{(n\lambda)^\tau}+C_5e^{-C_6n\lambda^2/(M_X^2M_e^2)}.$$

Under our choice of λ, if $\tau>2$, $\mathbb{P}(\mathcal{B}^c)=C_4(\log p)^{-\tau}+C_5p^{-C_6}$. Similarly, we can prove, if $na\gtrsim\sqrt{n\log p}M_X^2+n^{2\nu/\gamma}(\log p)^{3/2}\||\mathbf{x}_{\cdot}|_\infty\|^2_{\gamma,\alpha_X}$, $\mathbb{P}(\mathcal{A}^c)=C_1(\log p)^{-\gamma/2}+C_2p^{-C_3}$.

Denote $\omega=\sqrt{\log p/n}M_X^2+n^{2\nu/\gamma-1}(\log p)^{3/2}\||\mathbf{x}_{\cdot}|_\infty\|^2_{\gamma,\alpha_X}$. Then for some constant $\eta_1>0$, we have

$$\mathbb{P}\left(\forall\Delta\in\mathbb{R}^p,\Delta'\hat{\Sigma}\Delta\ge\Delta'\Sigma\Delta-\eta_1\omega|\Delta|_1^2\right)\ge 1-C_1(\log p)^{-\gamma/2}-C_2p^{-C_3}. \tag{2.7.13}$$

In other words, with high probability $1-\mathbb{P}(\mathcal{A}^c)$, the Restricted Strong Convexity condition $\Delta'\hat{\Sigma}\Delta\ge\kappa|\Delta|_2^2-\eta_1\omega|\Delta|_1^2$ holds.

Denote $\hat{\Delta}=\hat{\beta}-\beta$. For a threshold $\delta>0$, we choose

$$d=\#\{j\in\{1,2,...,p\}||\beta_j|\ge\delta\}.$$

Let $S = \{j : |\beta_j| \geq \delta\}$ and $S^c = \{j : |\beta_j| < \delta\}$. Applying Lemma 1 in Negahban et al. [2012], if $\lambda \geq 2|n^{-1}\sum_{i=1}^n x_i e_i|_\infty$, it holds that,

$$|\hat{\Delta}_{S^c}|_1 \leq 3|\hat{\Delta}_S|_1 + 4\sum_{j\in S^c}|\beta_j|.$$

We thus have

$$|\hat{\Delta}|_1 \leq |\hat{\Delta}_S|_1 + |\hat{\Delta}_{S^c}|_1 \leq 4|\hat{\Delta}_S|_1 + 4\sum_{j\in S^c}|\beta_j| \leq 4\sqrt{d}|\hat{\Delta}_S|_2 + 4\sum_{j\in S^c}|\beta_j|.$$

If follows that

$$\sum_{j\in S^c}|\beta_j| \leq \delta\sum_{j\in S^c}\left(\frac{|\beta_j|}{\delta}\right)^\theta \leq \delta^{1-\theta}K_\theta. \tag{2.7.14}$$

Thus

$$|\hat{\Delta}|_1 \leq 4\sqrt{d}|\hat{\Delta}_S|_2 + 4\delta^{1-\theta}K_\theta.$$

On the other hand, we have

$$d \leq \sum_{j\in S^c}\left(\frac{|\beta_j|}{\delta}\right)^\theta \leq \delta^{-\theta}K_\theta. \tag{2.7.15}$$

Suppose $|\hat{\Delta}|_2 \geq c_1\sqrt{K_\theta}(\lambda/\kappa)^{1-\theta/2}$ for some constant $c_1 > 0$. Then by (2.7.14) and (2.7.15), setting $\delta = \lambda/\kappa$,

$$\begin{aligned}
|\hat{\Delta}|_1 &\leq 4\sqrt{d}|\hat{\Delta}_S|_2 + 4\delta^{1-\theta}K_\theta \\
&\leq 4\sqrt{K_\theta}\left(\frac{\lambda}{\kappa}\right)^{-\theta/2}|\hat{\Delta}|_2 + 4\left(\frac{\lambda}{\kappa}\right)^{1-\theta}K_\theta \\
&\leq 4(1+c_1^{-1})\sqrt{K_\theta}\left(\frac{\lambda}{\kappa}\right)^{-\theta/2}|\hat{\Delta}|_2.
\end{aligned}$$

Recall $\lambda_{\min}(\Sigma) \geq \kappa > 0$. If $32(1+c_1^{-1})^2\eta_1 K_\theta \omega \lambda^{-\theta} \leq \kappa^{1-\theta}$, we will have,

$$\mathbb{P}\left(\hat{\Delta}'\hat{\Sigma}\hat{\Delta} \geq \frac{1}{2}\kappa|\hat{\Delta}|_2^2\right) \geq 1 - C_1(\log p)^{-\gamma/2} - C_2 p^{-C_3}.$$

An application of Lemma 2.7.5 shows that for constants $c_2, c_3 > 0$, if $\lambda \geq 2|n^{-1}\sum_{i=1}^n x_i e_i|_\infty$, with probability at least $1 - C_1(\log p)^{-\gamma/2} - C_2 p^{-C_3}$,

$$\begin{aligned} |\hat{\Delta}|_2 &\leq c_2\sqrt{K_\theta}\left(\frac{\lambda}{\kappa}\right)^{1-\theta/2}, \\ |\hat{\Delta}|_1 &\leq c_3 K_\theta \left(\frac{\lambda}{\kappa}\right)^{1-\theta}. \end{aligned}$$

When $|\hat{\Delta}|_2 \leq c_1\sqrt{K_\theta}(\lambda/\kappa)^{1-\theta/2}$ for some constant $c_1 > 0$. Then by (2.7.14) and (2.7.15), setting $\delta = \lambda/\kappa$, we can still obtain

$$\begin{aligned} |\hat{\Delta}|_1 &\leq 4\sqrt{d}|\hat{\Delta}_S|_2 + 4\delta^{1-\theta}K_\theta \\ &\leq 4\sqrt{K_\theta}\left(\frac{\lambda}{\kappa}\right)^{-\theta/2}|\hat{\Delta}|_2 + 4\left(\frac{\lambda}{\kappa}\right)^{1-\theta}K_\theta \\ &\leq 4(1+c_1)K_\theta\left(\frac{\lambda}{\kappa}\right)^{1-\theta}. \end{aligned}$$

Therefore, with probability at least $1 - C_1(\log p)^{-\gamma/2} - C_2 p^{-C_3} - C_4(\log p)^{-\tau}$, we have bounds (2.3.3) and (2.3.4). □

2.7.4 Proof of Theorem 2.3.2

Proof. Applying Theorem 2.3.1 with $\theta = 0$, with probability at least $1 - C_1(\log p)^{-\gamma/2} - C_2 p^{-C_3} - C_4(\log p)^{-\tau}$, we have

$$\begin{aligned} |\hat{\beta} - \beta|_2 &\lesssim \sqrt{s}\lambda/\kappa, \\ |\hat{\beta} - \beta|_1 &\lesssim s\lambda/\kappa. \end{aligned}$$

Since $s = K_\theta$, $s\omega \lesssim 1$ implies that

$$n \gtrsim M_X^4 s^2 \log p + s^{1/(1-2\nu/\gamma)} (\log p)^{3/(2-4\nu/\gamma)} \| |\mathbf{x}_\cdot|_\infty \|_{\gamma,\alpha_X}^{2/(1-2\nu/\gamma)}.$$

Recall the events

$$\begin{aligned} \mathcal{A} &= \{|\hat{\Sigma} - \Sigma|_\infty \leq a\} = \{\max_{j,k} |\hat{\sigma}_{jk} - \sigma_{jk}| \leq a\}, \\ \mathcal{B} &= \{n^{-1} \left| X^T e \right|_\infty \leq \lambda/2\}. \end{aligned}$$

Since $\hat{\beta}$ minimizes equation (2.1.2), we have

$$\frac{1}{2}|Y - X\hat{\beta}|_2^2 + \lambda|\hat{\beta}|_1 \leq \frac{1}{2}|Y - X\beta|_2^2 + \lambda|\beta|_1. \tag{2.7.16}$$

After some algebra, this reduces to

$$(\hat{\beta} - \beta)\hat{\Sigma}(\hat{\beta} - \beta) + \lambda|\hat{\beta}|_1 \leq 2e^T X(\hat{\beta} - \beta)/n + \lambda|\beta|_1 \tag{2.7.17}$$

On the event $\mathcal{B}$, the above inequality implies that

$$0 \leq (\hat{\beta} - \beta)\hat{\Sigma}(\hat{\beta} - \beta) \leq \frac{3}{2}\lambda|\hat{\beta}_J - \beta_J|_1 - \frac{1}{2}\lambda|\hat{\beta}_{J^c}|_1 \tag{2.7.18}$$

Then inequality (2.7.18) implies that

$$\frac{1}{2}\lambda|\hat{\beta} - \beta|_1 + (\hat{\beta} - \beta)\hat{\Sigma}(\hat{\beta} - \beta) \leq 2\lambda|\hat{\beta}_J - \beta_J|_1 \leq 2\lambda\sqrt{s}|\hat{\beta}_J - \beta_J|_2 \tag{2.7.19}$$

So (2.3.7) follow on the event $\mathcal{A} \cap \mathcal{B}$. □

2.7.5 Proof of Theorem 2.4.1

Proof. Reall $|\Sigma_{11}^{-1}|_2 = 1/N_1$ and let $|\hat{\Sigma}_{11}^{-1}|_2 = 1/N_2$. Without loss of generality, let $J = support(\beta) = \{1, ..., s\}$. Let $\boldsymbol{X} = (X_1, ..., X_n)'$ and denote by $\boldsymbol{X}_{(1)}$ and $\boldsymbol{X}_{(2)}$ the first s and last $p-s$ columns of X. Denote $W_n = \sum_{i=1}^n X_i e_i$ and $W_n(1)$, $X_{i,(1)}$, $\beta_{(1)}$ and $W_n(2)$, $X_{i,(2)}$, $\beta_{(2)}$ the first s and last $p-s$ entries of W_n, X_i and β, respectively. Define $b =$sign$(\beta_{(1)})$. Let

$$\begin{aligned} B &= (\frac{1}{n}X_{(1)}^T X_{(1)})^{-1}\left[\frac{1}{n}X_{(1)}^T e - \lambda b\right], \\ D_k &= X_{(2),k}^T\left\{X_{(1)}(X_{(1)}^T X_{(1)})^{-1}\lambda b - \left[X_{(1)}(X_{(1)}^T X_{(1)})^{-1}X_{(1)}^T - I\right]\frac{e}{n}\right\}, \end{aligned}$$

where $X_{(2),k} = (X_{1k}, ..., X_{nk})^T$ denote the k-th columns of $\boldsymbol{X}$ and $s+1 \le k \le p$. Denote the j-th element of B as B_j.

By rearranging terms, it is easy to see that the events

$$\mathcal{B} = \{\max_{1\le j\le s}|B_j| < L\}, \tag{2.7.20}$$

$$\mathcal{D} = \{\max_{s+1\le k\le p}|D_k| < \lambda\}, \tag{2.7.21}$$

are sufficient to guarantee that conditions in Lemma 2.7.4 hold. Then $\mathbb{P}(\hat{\beta} \neq_s \beta) \le \mathbb{P}(\mathcal{B}^c) + \mathbb{P}(\mathcal{D}^c)$.

We first analyze the event $\mathcal{D}$. Recall $\mathbb{E}(X_{ik}|X_{(1)}, e) = [\Sigma_{21}\Sigma_{11}^{-1}X_{i,(1)}]_k$ and $z_{ik} = X_{ik} - \mathbb{E}(X_{ik}|X_{(1)}, e)$ for $s+1 \le k \le p$. Let $\omega_1 = X_{(1)}(X_{(1)}^T X_{(1)})^{-1}\lambda b$, $\omega_2 = [I - X_{(1)}(X_{(1)}^T X_{(1)})^{-1}X_{(1)}^T]e/n$ and $\omega = \omega_1 + \omega_2$. Denote $Z_k = (z_{1k}, ..., z_{nk})^T$, $U_k = Z_k^T\omega$ and $\mu_k = \mathbb{E}(X_{(2),k}^T\omega|X_{(1)}, e)$. Note that $\mathbb{E}Z_k = 0$ and $\omega_1^T\omega_2 = 0$. Then by the irrepresentable

condition,

$$\begin{aligned}\max_{s+1\le k\le p}|D_k| &= \max_{s+1\le k\le p}|\mu_k+U_k| \\ &\le \max_{s+1\le k\le p}[|\mu_k|+|U_k|] \\ &\le (1-\eta)\lambda+\max_{s+1\le k\le p}|U_k|.\end{aligned}$$

From this inequality, we have

$$\{\max_{s+1\le k\le p}|U_k|<\eta\lambda\}\subset\{\max_{s+1\le k\le p}|D_k|<\lambda\}.$$

Define the events

$$\mathcal{A}_1 = \{|\hat{\Sigma}_{11}-\Sigma_{11}|_\infty\le a\}=\{\max_{1\le j,k\le s}|\hat{\sigma}_{jk}-\sigma_{jk}|\le a\}, \tag{2.7.22}$$

$$\mathcal{A}_2 = \{n^{-1}e_i^2\le 2\sigma\}, \tag{2.7.23}$$

$$\mathcal{T} = \{|\omega|_2^2\le\delta_*\}. \tag{2.7.24}$$

By Lemma 2.7.3, on the event $\mathcal{A}_1$ with $a=N_1/(2s)$,

$$N_2=\inf_{|\zeta|_2=1}\zeta^T\hat{\Sigma}_{11}\zeta>\inf_{|\zeta|_2=1}\zeta^T\Sigma_{11}\zeta-sa=\frac{N_1}{2}.$$

By Lemma 2.7.1,

$$\mathbb{P}\left(\left|\sum_{i=1}^n(e_i^2-\sigma)\right|\ge n\sigma\right)\le\frac{n\|e.\|_{q,\alpha_e}^q}{n^q\sigma^q}+\exp\left(-\frac{n\sigma^2}{\|e.\|_{2,\alpha_e}^2}\right):=P_2$$

Denote $P_1=\mathbb{P}(\mathcal{A}^c)$ with $a=N_1/(2s)$. We know

$$\omega_1^T\omega_1 = \lambda^2 b^T(X_{(1)}^TX_{(1)})^{-1}b\le\frac{\lambda^2 s}{nN_2},$$

and

$$\omega_2^T \omega_2 \;\; \leq \;\; \frac{e^T e}{n^2}.$$

Thus, we have

$$\mathbb{P}(\mathcal{T}^c) \leq \mathbb{P}\left(\omega_1^T \omega_1 \geq \frac{2\lambda^2 s}{nN_1}\right) + \mathbb{P}\left(\omega_2^T \omega_2 \geq 2n\sigma\right) \leq P_1 + P_2.$$

By Lemma 2.7.2, if $\eta\lambda \gtrsim \sqrt{\delta_* \log p}\Psi_{2,\alpha_X,(2)} + n^{(\iota-1)/\gamma}\delta_*^{1/2}(\log p)^{3/2}\||Z_\cdot|_\infty\|_{\gamma,\alpha_X}$,

$$\mathbb{P}\left(\max_{s+1\leq k\leq p} |U_k| \geq \eta\lambda \mid \mathcal{T}\right) \;\; \leq \;\; C_1(\log(p-s))^{-\gamma} + C_2(p-s)^{-C_3} := P_3.$$

By the total probability rule, we have

$$\mathbb{P}(\mathcal{D}^c) \;\; \leq \;\; \mathbb{P}\left(\max_{s+1\leq k\leq p} |U_k| \geq \eta\lambda \mid \mathcal{T}\right) + \mathbb{P}(\mathcal{T}^c) \leq P_1 + P_2 + P_3.$$

Now we analyze the event $\mathcal{B}$. Note that $|\hat{\Sigma}_{11}^{-1}b|_\infty \leq \sqrt{s}|\hat{\Sigma}_{11}^{-1}|_2 = \sqrt{s}/N_2$. Recall $\lambda \leq nN_1L/(4\sqrt{s})$. On the event $\mathcal{A}$, $nL - \lambda|[\hat{\Sigma}_{11}^{-1}b]_j| \geq nL(1 - N_1/(4N_2)) \geq \sqrt{n}L/2$ for all $1 \leq j \leq s$. Simple application of the Cauchy inequality shows that

$$\sup_{|\zeta|_2=1} \zeta^T \hat{\Sigma}_{11}^{-1} W_n(1) \leq \frac{1}{N_2}\sqrt{\sum_{j=1}^{s}(\sum_{i=1}^{n} x_{ij}e_i)^2}.$$

This yields

$$\begin{aligned}
\mathcal{B} &= \bigcap_{j=1}^{s}\{|[\hat{\Sigma}_{11}^{-1}W_n(1)]_j| < \frac{1}{2}nL\} \\
&= \{\sup_{|\zeta|_2=1}\zeta^T\hat{\Sigma}_{11}^{-1}W_n(1) < \frac{1}{2}nL\} \\
&\supseteq \left\{\sqrt{\sum_{j=1}^{s}(\sum_{i=1}^{n}x_{ij}e_i)^2} < \frac{1}{2}nLN_2\right\} \\
&\supseteq \left\{\max_{1\le j\le s}\left|\sum_{i=1}^{n}x_{ij}e_i\right| < \lambda\right\}\bigcap\left\{|\hat{\Sigma}_{11}-\Sigma_{11}|_\infty \le \frac{N_1}{2s}\right\}.
\end{aligned}$$

Thus,

$$\mathbb{P}(\mathcal{B}^c) \le \mathbb{P}(|W_n(1)|_\infty \ge \lambda) + P_1.$$

By carrying out similar procedures as those in the proof of Theorem 2.3.1, we can control the probability P_1 and $\mathbb{P}(|W_n(1)|_\infty \ge \lambda)$. Then (2.4.5) follows. □

2.7.6 Proof of Proposition 2.4.1

Proof. Let $\gamma_l = \mathbb{E}y_i y_{i-l}$. Set the candidate lags of this AR(2) model as d. Since $\gamma_0 = 1$, we have

$$\Sigma_{11} = \begin{pmatrix} 1 & \gamma_1 \\ \gamma_1 & 1 \end{pmatrix},$$

and

$$\Sigma_{21} = \begin{pmatrix} \gamma_2 & \gamma_1 \\ \cdots & \cdots \\ \gamma_{d-1} & \gamma_{d-2} \end{pmatrix}.$$

Basic calculation shows that

$$\Sigma_{11}^{-1} = \begin{pmatrix} \frac{1}{1-\gamma_1^2} & -\frac{\gamma_1}{1-\gamma_1^2} \\ -\frac{\gamma_1}{1-\gamma_1^2} & \frac{1}{1-\gamma_1^2} \end{pmatrix},$$

and

$$\begin{aligned} \gamma_1 &= \frac{\phi_1}{1-\phi_2}, \\ \gamma_l &= \phi_1\gamma_{l-1} + \phi_2\gamma_{l-2}, \end{aligned}$$

for $2 \leq l \leq d$.

We first consider the case $\phi_1 > 0$ and $\phi_2 > 0$. Then the Strong Irrepresentable Condition

$$|\Sigma_{21}\Sigma_{11}^{-1}\text{sign}(\beta_{(1)})|_\infty = \max_{2\leq j\leq d-1} \frac{\gamma_j}{1-\gamma_1^2} - \frac{\gamma_{j-1}\gamma_1}{1-\gamma_1^2} - \frac{\gamma_j\gamma_1}{1-\gamma_1^2} + \frac{\gamma_{j-1}}{1-\gamma_1^2} < 1$$

For $j = 2$, it can be shown that

$$\frac{\gamma_j}{1-\gamma_1^2} - \frac{\gamma_{j-1}\gamma_1}{1-\gamma_1^2} - \frac{\gamma_j\gamma_1}{1-\gamma_1^2} + \frac{\gamma_{j-1}}{1-\gamma_1^2} < 1$$

is equivalent to $\phi_1 + \phi_2 < 1$. Then $\gamma_1 < 1$ and $\gamma_j < \gamma_{j-1}$ for all $j \geq 1$. Thus, we have, $|\Sigma_{21}\Sigma_{11}^{-1}\text{sign}(\beta_{(1)})|_\infty < 1$ is equivalent to $\phi_1 + \phi_2 < 1$.

Similarly, we can prove the cases $\phi_1 > 0, \phi_2 \leq 0$ and $\phi_1 \leq 0, \phi_2 > 0$ and $\phi_1 \leq 0, \phi_2 \leq 0$.

□

CHAPTER 3

HIGH DIMENSIONAL GENERALIZED LINEAR MODELS

3.1 Introduction

In recent years, information technology has made high-dimensional time series data increasingly common. The demand for modelling and forecasting such data arises naturally from market analysis in finance, panel studies in economics, environmental studies, and communication engineering, among others. In many applications, we often face the challenge of dealing with a large number of complicated issues such as missing values or heavy tails. The Lasso regularized method, originally introduced by Tibshirani [1996] and subsequently investigated by many others, is a popular technique for high-dimensional linear regression models with sparse coefficients. As a matter of fact, the ℓ_1-type penalty of the Lasso can also be applied to other models in high dimension, including, for example, logistic regression (Lokhorst [1999]; Roth [2004]; Shevade and Keerthi [2003]; Genkin et al. [2007], among others), multinomial logistic regression (Krishnapuram et al. [2005]) or Cox regression (Tibshirani [1997]) by replacing the ℓ_2 loss function by the corresponding negative log-likelihood function. See also, Li et al. [2018], Dou and Liang [2019], Dou and Anitescu [2019], Liu and Gao [2017], Liu and Barber [2018], Ha et al. [2018], Liu [2019], among others.

It is well known that if the covariates and/or the errors deviate more wildly from the sub-Gaussian distribution, the linear regression estimator based on the least squares loss no longer converges at the optimal rates. Intuitively, an outlier in the covariates may cause the corresponding M-estimator to behave arbitrarily badly. This motivates the use of generalized M-estimators that downweight high-leverage observations. In the classical theory of robust regression in low dimensions, many weighting functions are introduced, such as Mallows estimator (Mallows [1975]), Hill-Ryan estimator (Hill [1977]), and Schweppe estimator (Merrill and Schweppe [1971]). In this chapter, we focus on heavy-tailed covariates and heavy-tailed errors for generalized linear model. We also extend the robust M-estimator

to high-dimensional time series.

Driven by a wide range of contemporary scientific applications, robust regression of high-dimensional data is of substantial research interest. Indeed, several papers have shed new light on high dimensional robust M-estimator when the population distribution is heavy tailed or noisy. Catoni et al. [2012] considered estimation of the mean of heavy-tailed distributions via a robust empirical loss, which is insensitive to extreme values. Cantoni's mean estimator is further extended in Brownlees et al. [2015] to empirical risk minimization. Fan et al. [2016] introduces a simple principle for robust high-dimensional low rank matrix recovery via an appropriate shrinkage on the data. Fan et al. [2017] developed estimation bounds for penalized robust regression with the Huber loss function. Loh [2017] gave a general framework for robust regularized M-estimators under both convex and non-convex loss functions. However, all prior works focused on the setting where samples are i.i.d. To the best of our knowledge, existing procedures cannot be readily applied to high-dimensional time series data. The second goal of our study is to provide a solid theoretical guarantee on the robustness of generalized linear models for serially dependent data

Following the work of Fan et al. [2016], we propose to appropriately shrink the feature variables before calculating the M-estimator to achieve the robustness for high-dimensional time series regression. Let X_i be a p-dimensional vector of covariates. If X_i is heavy-tailed, the basic idea is to truncate each feature $X_{ij}(1 \leq j \leq p)$ to a predetermined threshold level τ. We show that the regularized robust regression functions continue to enjoy good behavior. Our first contribution is to provide the asymptotic behavior of the estimated GLM coefficients and the excess risk of the Lasso penalized method for both the original time series data and shrinkage heavy-tailed data. It is shown that an appropriate truncation does not induce significant bias. Under only bounded moment conditions for either noise or covariates, our robust estimator can nearly achieve the error bound for i.i.d. sub-Gaussian data, modulo a price for temporal dependence. The allowed dimension p can be as large as $\exp(n^c)$, where n is the sample size and $0 < c < 1$. This means that shrinkage not only overcomes heavy-

tailed corruption, but also mitigates the curse of dimensionality. Furthermore, unlike the usual robust quasi-likelihood estimators in low dimension, which is non-convex, our method still maintains convexity, thus has certain computational advantages.

In addition, our robust estimator can also be applied to the usual linear regression setting for high-dimensional time series. For weakly temporal dependence and heavy-tailed data, our robust method achieves the minimax optimal rate of ℓ_2 norm established by Raskutti et al. [2011] for i.i.d sub-Gaussian data. However, the difference lies in the scaling requirements on p, n and the sparsity condition. It is also worth noting that we provide new concentration inequalities, which extend Bousquet's inequality (Bousquet [2002]) to high-dimensional time series. This extension is of independent interest.

Besides the theoretical properties, we also study the numerical performance of the proposed robust procedure using both simulated and real data. Section 3.5 considers the simulation studies and shows that our robust procedure performs well numerically in the presence of both symmetric and asymmetric heavy-tailed covariates and/or errors. In particular, the robust procedure significantly outperforms the standard Lasso method, especially in L_1 and L_2 losses of the GLM coefficients. The simulation study also shows that the proposed procedure improves the convergence speed of the coefficient estimators to the true ones with heavy-tailed time series data. We also illustrate our procedure with an application to high-frequency stock trading for predicting price changes in consecutive transactions via a multinomial logistic regression. Our method leads to marked improvements in prediction compared with the existing methods in financial econometrics.

3.2 The Model

3.2.1 Generalized Linear Models and the Loss Function

Consider n observations $\{(X_i, Y_i)\}_{i=1}^n$, where $X_i \in \mathcal{X} \subset \mathbb{R}^p$ is a p-dimensional vector of covariate variables, and $Y_i \in \mathcal{Y} \subset \mathbb{R}$ is a response variable. We model the dependence of the

mean of Y_i on X_i via the linear function $f_{\beta^0}(X_i) = X_i^T\beta^0$, where β^0 is a vector of unknown coefficients and X_i^T is the transpose of X_i. The goal is to estimate β^0. In a high dimensional model, the number of covariates p can be much larger than the number of observations n. Let $R : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a loss function.

We consider the following estimator of empirical risk minimization with Lasso penalty

$$\hat{\beta} := \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} R(f_\beta(X_i), Y_i) + \lambda \left|\beta\right|_1 \right\}, \tag{3.2.1}$$

where $f_\beta(X_i) = X_i^T\beta$. Denote the best linear approximation of the *theoretical risk* by

$$\beta^0 := \arg\min_{\beta} \mathbb{E}[R(f_\beta(X), Y)], \quad \text{and} \quad f^0 = f_{\beta^0}. \tag{3.2.2}$$

The *excess risk* is

$$\mathcal{E}(f_\beta) := \mathbb{E}[R(f_\beta(X), Y)] - \mathbb{E}[R(f_{\beta^0}(X), Y)]. \tag{3.2.3}$$

Note that by definition, $\mathcal{E}(f_\beta) \geq 0$ for all β.

Assume the response Y is from an exponential family with the probability density function taking the canonical form

$$h_Y(y; \mu) = \exp\left[y\mu - r(\mu) + b(y)\right]$$

for some known functions $r(\cdot)$, $b(\cdot)$ and unknown function μ. The function μ is usually called the canonical or natural parameter. The mean response is $r'(\mu)$, the first derivative of $r(\mu)$ with respect to μ. The generalized linear model assumes the form:

$$\mathbb{E}(Y|X) = r'(\mu(X)) = r'(X^T\beta^0).$$

The canonical link function is thus defined as $g := (r')^{-1}$. Let $z = \mu(X)$. The maximum

marginal (log-)likelihood loss function is then

$$R(z, y) = -yz + r(z), \quad y \in \mathcal{Y}, \quad z \in \mathbb{R}. \tag{3.2.4}$$

More generally, the quasi-(log)likelihood function is

$$R(z, y) := -\int_y^{H(z)} \frac{y - u}{\mathcal{V}(u)} du, \quad y \in \mathcal{Y}, \quad z \in \mathbb{R},$$

where $\mathcal{V} : \mathbb{R} \to (0, \infty)$ is a given variance function, and H is the inverse link function; see also McCullagh and Nelder [1989]. The canonical link function (up to an additive constant) is

$$g(t) := \int_{y_0}^{t} \frac{1}{\mathcal{V}(u)} du, \quad t \in \mathcal{Y},$$

where y_0 is an arbitrary but fixed constant. Let

$$r(z) := \int_{y_0}^{H(z)} \frac{u}{\mathcal{V}(u)} du, \quad z \in \mathbb{R}.$$

Then the loss function is $R(z, y) = -yg(H(z)) + r(z)$. In this chapter, we define the loss function

$$R(z, y) = -yh(z) + r(z), \tag{3.2.5}$$

and assume h and r satisfy some uniform continuity conditions (see Assumptions 3.3.2 and 3.3.3 below).

3.2.2 Robust Lasso Estimator

Inspired by the theory on robust estimation for linear regression (see Fan et al. [2016]), we study regularized versions of high-dimensional robust GLM estimators and establish statis-

tical guarantees. In order to deal with heavy-tailed data, we propose the robust estimator to be used in (3.2.1) by the simple and classical principle of truncation, or more generally shrinkage. Our approach is simple: we truncate or shrink appropriately the heavy-tailed covariates or/and the response variable. Intuitively, shrinkage reduces sensitivity of the estimator to data corruption caused by the heavy-tailed distributions. However, shrinkage leads to bias. We shall find an appropriate shrinkage level to balance the induced bias and the statistical error rate. The resulting estimator is then defined as follows:

$$\hat{\beta} := \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} R_{\tau}(f_{\beta}(X_i), Y_i) + \lambda \left|\beta\right|_1 \right\}, \tag{3.2.6}$$

where τ is a predetermined threshold level,

$$R_{\tau}(f_{\beta}(X_i), Y_i) := R(f_{\beta}(\widetilde{X}_i), Y_i),$$

and $\widetilde{X}_i$ is a truncated version of the covariates X_i if they are heavy-tailed and equals the original covariates (infinite truncation threshold) if they are light-tailed. When the covariates X_i are heavy-tailed, we choose

$$\widetilde{X}_{ij} = sgn(X_{ij})(|X_{ij}| \wedge \tau), \quad 1 \leq j \leq p,$$

where $a \wedge b = \min(a, b)$. In Section 3.3, we focus on the case when the data generating distribution of the response is indeed the model distribution so that we only need to trim the covariates. If the response is also corrupted by random noises, we shall also truncate the response. For the linear regression model with least square loss in Section 3.4, if the errors have heavy tails, we also need to truncate the response to achieve robustness.

With the aforementioned data robustfications, the proposed methodology yields an estimator that, under a bounded moment condition on the covariates or/and the response, has the similar statistical rate as that of the estimator available in the literature for sub-Gaussian

distributions. Our study gives a formal theoretical consideration of both the original estimator in (3.2.1) and the robust one in (3.2.6).

The first and most important advantage of our robust method is to maintain the convexity of (negative) log-likelihood loss. There are several alternatives to robust estimation in the context of low dimensional generalized linear model. For example, Cantoni and Ronchetti [2001] proposed a class of robust quasi-likelihood loss function. Bianco and Yohai [1996] constructed robust estimator for the logistic regression by bounded deviance, which was further extended to other generalized linear models. Zhang et al. [2014] introduced a class of robust estimators for generalized linear models motivated by the Bregman divergence. However, most of these robust estimators in the low dimensional case are non-convex.

Our robust method is also much easier to implement than many existing ones, as it only needs to truncate or shrink the data before applying the standard method to the transformed data. The tuning parameter τ plays a key role by adapting to covariates and/or errors with different shapes and tails. In practice, the optimal values of tuning parameters τ and λ can be chosen by a two-dimensional grid search using an information-based criterion or cross-validation, e.g., the Akaike information criterion or Bayesian information criterion. Specifically, we may partition a rectangle in the scale of $(\log(\tau), \log(\lambda))$ to form the search grid. Then the optimal values are achieved by the combination of the two parameters that minimizes the cross-validated measurement, the Akaike information criterion or Bayesian information criterion.

3.3 Asymptotic properties

We now consider the properties of the standard Lasso method (3.2.1) and the robust Lasso method (3.2.6). We first show the asymptotic behavior of the estimated GLM coefficients and the excess risk for the original time series data and then demonstrate that the convergence rates of robust estimator for shrinkage heavy-tailed data are significantly improved.

3.3.1 Definitions and Assumptions

We begin by introducing some additional definitions and technical assumptions. Besides the definitions of dependence measures in Chapter 2.2, we introduce the following dependence adjusted norms

$$\|X_{\cdot j}\|_{q,\alpha} = \sup_{m\geq 0}(m+1)^{\alpha}\Delta_{m,q,j}, \quad \alpha \geq 0, \tag{3.3.1}$$

$$\|Y_{\cdot}\|_{q,\alpha} = \sup_{m\geq 0}(m+1)^{\alpha}\Delta_{m,q,y}, \quad \alpha \geq 0, \tag{3.3.2}$$

$$\|Y_{\cdot}\|_{\psi_\nu} = \sup_{q\geq 2} q^{-\nu}\Delta_{0,q,y}, \quad \alpha \geq 0. \tag{3.3.3}$$

Here $\|Y_{\cdot}\|_{\psi_\nu}$ can be naturally interpreted as the ψ_ν Orlicz norm in the dependence case. Hence $\|\cdot\|_{q,\alpha}$ and $\|\cdot\|_{\psi_\nu}$ represent polynomial decay and exponential decay dependence adjusted norm, respectively.

In this chapter, we use the dependence adjusted norms $\||X_{\cdot}|_\infty\|_{q,\alpha}$, $\||X_{\cdot}|_\infty\|_{\psi_\nu}$, $\|Y_{\cdot}\|_{q,\alpha}$ and $\|Y_{\cdot}\|_{\psi_\nu}$ to study the limiting properties of Lasso estimators in the presence of serial dependence.

To prove theoretical properties of our method, we require the following conditions:

Assumption 3.3.1 (Convex Loss)**.** *Throughout this chapter, the map*

$$z \mapsto R(z, y)$$

is convex for all $y \in \mathcal{Y}$.

This assumption is important from a computational perspective; see e.g. van de Geer and Müller [2012] and Loh [2017]. It also plays a crucial role in our theory, as it allows us to prove that the estimator $\hat{\beta}$ is in a l_1-neighborhood of β^0.

Assumption 3.3.2 (Lipschitz Property)**.** *For the function r, it holds that*

$$|r(u) - r(u^*)| \leq C\,|u - u^*|\,,$$

where C is a positive constant.

Assumption 3.3.3. *For the function r, it also holds that*

$$\left|r(x_1^T\beta) - r(x_1^T\beta^*) - r(x_2^T\beta) + r(x_2^T\beta^*)\right| \leq C\left|(x_1 - x_2)^T(\beta - \beta^*)\right|,$$

where C is a positive constant.

Assumption 3.3.3 is needed to deal with the temporal dependence of $\{X_i\}_{i=1}^n$. If the observed data are i.i.d., then we only require Lipschitz property (Assumption 3.3.2) of the loss function. Both Assumptions 3.3.2 and 3.3.3 describe some uniform continuity conditions. They allow one to apply concentration inequalities in the Section 3.7. Elementary calculation shows that the loss function for logistic regression and multinomial logistic regression satisfy both assumptions. Consider the following smoothed version of hinge loss:

$$R_h(z,y) = \begin{cases} \frac{1}{2} - zy & \text{if } zy \leq 0, \\ \frac{1}{2}(1-zy)^2 & \text{if } 0 < zy \leq 1, \\ 0 & \text{if } 1 \leq zy. \end{cases}$$

It can be shown that $R_h(X_i^T\beta, Y_i)$ also satisfies Assumptions 3.3.1, 3.3.2 and 3.3.3.

Assumption 3.3.4 (Margin Condition)**.** *There exist $M_0 > 0$ and a strictly convex function $J(\cdot)$ such that for all β with $|\beta - \beta^0|_1 \leq M_0$, one has*

$$\mathcal{E}(f_\beta) \geq J\left(\|f_\beta - f_{\beta^0}\|\right).$$

By Assumption 3.3.1, the stochastic part of the problem needs only to be studied locally, near the target. Moreover, the margin condition is only needed locally.

The convex conjugate of the function J is defined as G. Hence, by definition, for any positive u and v, we have $uv \leq J(u) + G(v)$. Indeed, in a typical case, J can be chosen as

a quadratic function, for example, $J(u) = u^2/2$, then $G(v) = v^2/2$. The margin condition requires that in the neighborhood $|\beta - \beta^0|_1 \leq M_0$, the excess risk is bounded from below by a strictly convex function. This assumption is similar to the one in van de Geer [2008] and Bühlmann and Van De Geer [2011].

Assumption 3.3.5 (Compatibility Condition)**.** *Let $S \subset \{1, 2, ..., p\}$ and S^c be the complimentary set of S. For a constant $\kappa(S) > 0$, if for all β satisfying $|\beta_{S^c}|_1 \leq 3\,|\beta_S|_1$, it holds that*

$$|\beta_{S^c}|_1^2 \leq \left\|f_\beta\right\|^2 |S| \,/\kappa^2(S).$$

The compatibility condition we discuss here follows from van de Geer [2008]. It is closely related to (and weaker than) the restricted eigenvalue condition as given in Bickel et al. [2009].

3.3.2 Rate of Convergence

In the context of generalized linear model, following van de Geer [2008] and Bühlmann and Van De Geer [2011], one can define the *oracle* β^* as

$$\beta^* := \arg \min_{\beta:\ S_\beta \in \mathcal{T}} \left\{ 3\mathcal{E}(f_\beta) + 2G\left(\frac{4\lambda\sqrt{s_\beta}}{\kappa(S_\beta)}\right)\right\}. \tag{3.3.4}$$

where $\mathcal{T}$ is a large collection of index sets, $S_\beta := \{j : \beta_j \neq 0\}$ and $|S_\beta| = s_\beta$. And the associated minimum risk is denoted as

$$\mathcal{E}^* := \frac{3}{2}\mathcal{E}(f_{\beta^*}) + G\left(\frac{4\lambda\sqrt{s_{\beta^*}}}{\kappa(S_{\beta^*})}\right). \tag{3.3.5}$$

The second term in (3.3.5) depends only on the set of nonzero coefficients in β^*, and we can refer to it as "estimation error" in a generic sense. Note that in the typical case $G(v) = v^2/2$, it is up to a constant equal to $\lambda^2 s_{\beta^*}$. The first term in (3.3.5) will be referred to as

"approximation error". Then β^* balances "approximation error" and "estimation error", we refer to it as the "oracle"; see van de Geer [2008] and Bühlmann and Van De Geer [2011]) for detailed interpretation of β^*.

We are now ready to state our main results of Lasso method for the original time series data. For any Lasso solution $\hat{\beta}$ of the problem (3.2.1), the following theorem provides the rate of convergence of $|\hat{\beta}-\beta|_1$ and the excess risk $\mathcal{E}(f_{\hat{\beta}})$ by the moment and the dependence conditions.

Theorem 3.3.1. *Suppose Assumptions 3.3.1, 3.3.4 and 3.3.5 hold. Suppose the loss function has the form (3.2.5) and the functions $h(\cdot), r(\cdot)$ therein satisfy Assumptions 3.3.2 and 3.3.3, or the loss function itself satisfies Assumptions 3.3.2 and 3.3.3. Let β^* and $\mathcal{E}^*$ be given in (3.3.4) and (3.3.5).*

(i). Assume that $\||X_\cdot|_\infty\|_{\gamma,\alpha_X} < \infty$ and $\|Y_\cdot\|_{q,\alpha_Y} < \infty$, where $q, \gamma > 2$ and $\alpha_X, \alpha_Y > 0$. Define

$$\nu = \begin{cases} 1 & \text{if } \alpha_X \geq 1/2 - 1/\gamma, \\ \gamma/2 - \alpha_X\gamma & \text{if } \alpha_X < 1/2 - 1/\gamma. \end{cases}$$

Assume $\chi = q\gamma/(q+\gamma) > 2$ and let $\alpha = \min(\alpha_X, \alpha_Y)$. Define

$$\rho = \begin{cases} 1 & \text{if } \alpha \geq 1/2 - 1/\chi, \\ \chi/2 - \alpha\chi & \text{if } \alpha < 1/2 - 1/\chi. \end{cases}$$

Suppose that

$$\begin{aligned} \lambda_0 \gtrsim\ & p^{1/(2(\gamma\wedge\chi)+2)}(\log(pn))^{1/2}n^{-1/2}\,\||X_\cdot|_\infty\|_{2,\alpha_X} \\ & + n^{\nu/\gamma-1}p^{3/(2(\gamma\wedge\chi)+2)}(\log(pn))^{3/2}\,\||X_\cdot|_\infty\|_{\gamma,\alpha_X} \\ & + p^{1/(2(\gamma\wedge\chi)+2)}(\log(pn))^{1/2}n^{-1/2}\||X_\cdot|_\infty\|_{\gamma,\alpha_X}\|Y_\cdot\|_{q,\alpha_Y} \\ & + n^{\rho/\chi-1}p^{3/(2(\gamma\wedge\chi)+2)}(\log(pn))^{3/2}\||X_\cdot|_\infty\|_{\gamma,\alpha_X}\|Y_\cdot\|_{q,\alpha_Y}. \end{aligned}$$

Assume $|\beta^* - \beta^0|_1 \leq M_0/2$ *and* $\mathcal{E}^*/\lambda_0 \leq M_0/2$. *If* $\lambda \geq (20/3)\lambda_0$, *then we have with probability at least* $1-C_1(\log(pn))^{-(\gamma\wedge\chi)}$, *any Lasso solution* $\hat{\beta}$ *of the problem (3.2.1) satisfies*

$$\lambda|\hat{\beta} - \beta^*|_1 \leq \frac{10}{3}\mathcal{E}^*, \tag{3.3.6}$$

$$\mathcal{E}(f_{\hat{\beta}}) \leq \frac{8}{3}\mathcal{E}^*, \tag{3.3.7}$$

where the positive constant C_1 *only depends on* q, γ, α_X *and* α_Y.

(ii). Assume that $\||X_\cdot|_\infty\|_{\psi_\iota} < \infty$ *and* $\|Y_\cdot\|_{\psi_\nu} < \infty$, *where* $\iota, \nu \geq 0$. *Suppose that for some positive constants* A *and* B,

$$\begin{aligned}\lambda_0 = {} & A((\log p)^{1+\iota} + (\log p)^{(1+2\iota)/2}(\log n)^{1/2})n^{-1/2}\||X_\cdot|_\infty\|_{\psi_\iota} \\ & + B((\log p)^{1+\iota+\nu} + (\log p)^{(1+2\iota+2\nu)/2}(\log n)^{1/2})n^{-1/2}\||X_\cdot|_\infty\|_{\psi_\iota}\|Y_\cdot\|_{\psi_\nu}.\end{aligned}$$

Assume $|\beta^* - \beta^0|_1 \leq M_0/2$ *and* $\mathcal{E}^*/\lambda_0 \leq M_0/2$. *If* $\lambda \geq (20/3)\lambda_0$, *then with probability at least*

$$1 - C_2 p^{-C_3 A^{\frac{2}{1+2\iota}}} - C_4 p^{-C_5 B^{\frac{2}{1+2\iota+2\nu}}},$$

for any Lasso solution $\hat{\beta}$ *of the problem (3.2.1), we have bounds (3.3.6) and (3.3.7), where the positive constants* $C_2, ..., C_5$ *only depend on* ι *and* ν.

If the loss function is the marginal (log-)likelihood loss function, we can further achieve some sharper results.

Corollary 3.3.2. *Assume the loss function is the maximum marginal (log-)likelihood loss function in (3.2.4). Assume* $|\beta^* - \beta^0|_1 \leq M_0/2$ *and* $\mathcal{E}^*/\lambda_0 \leq M_0/2$. *Then (i), suppose that*

$$\begin{aligned}\lambda_0 \gtrsim & p^{1/(2\gamma+2)}(\log(pn))^{1/2}n^{-1/2}\,\||X_\cdot|_\infty\|_{2,\alpha_X} + n^{\nu/\gamma-1}p^{3/(2\gamma+2)}(\log(pn))^{3/2}\,\||X_\cdot|_\infty\|_{\gamma,\alpha_X} \\ & + (\log p)^{1/2}n^{-1/2}\||X_\cdot|_\infty\|_{\gamma,\alpha_X}\|Y_\cdot\|_{q,\alpha_Y} + n^{\rho/\chi-1}(\log p)^{3/2}\||X_\cdot|_\infty\|_{\gamma,\alpha_X}\|Y_\cdot\|_{q,\alpha_Y}.\end{aligned}$$

If $\lambda \geq (20/3)\lambda_0$, *then under the conditions of Theorem 3.3.1.(i), we have, with probability*

at least $1 - C_1(\log(pn))^{-\gamma} - C_2 p^{-C_3} - C_4(\log p)^{-\chi}$, *the bounds (3.3.6) and (3.3.7).*

(ii). Suppose that

$$\begin{aligned}\lambda_0 =& A((\log p)^{1+\iota} + (\log p)^{(1+2\iota)/2}(\log n)^{1/2})n^{-1/2}\||X_\cdot|_\infty\|_{\psi_\iota} \\ &+ B^{(1+2\iota+2\nu)/2}(\log p/n)^{1/2}\||X_\cdot|_\infty\|_{\psi_\iota}\|Y_\cdot\|_{\psi_\nu}.\end{aligned}$$

If $\lambda \geq (20/3)\lambda_0$, *then under the conditions of Theorem 3.3.1.(ii), with probability at least*

$$1 - C_5 p^{-C_6 A^{\frac{2}{1+2\iota}}} - C_7 e^{-C_8 B},$$

we have bounds (3.3.6) and (3.3.7).

The assumption that $|\beta^* - \beta^0|_1 \leq M_0/2$ and $\mathcal{E}^*/\lambda_0 \leq M_0/2$ is a technical condition. See Example 6.4 in Bühlmann and Van De Geer [2011] for detailed verification in the case of logistic regression. Here we also assume that the short-range dependence (SRD) condition holds, that is,

$$\||X_\cdot|_\infty\|_{\gamma,\alpha_X} < \infty \quad \text{and} \quad \|Y_\cdot\|_{q,\alpha_Y} < \infty.$$

If it fails, the processes (X_i) and (Y_i) may exhibit some long-range dependence, and the asymptotic behavior can be quite different.

Theorem 3.3.1 states the "oracle rate" for the excess risk and the ℓ_1 norm, where β^* is the oracle. This terminology "oracle" is mainly chosen for ease in reference, and allows much flexibility, in the sense that the choice of the collection $\mathcal{T}$ is left unspecified. In order to improve the bound for the excess risk, one may choose to minimize over a larger collection $\mathcal{T}$. See Bühlmann and Van De Geer [2011] Chapter 6 for detailed discussions. In general, β^* may not be the parameter of interest and is not easy to interpret. The following corollary provides the rate of convergence of $|\hat{\beta} - \beta^0|_1$ and the excess risk $\mathcal{E}(f_{\hat{\beta}})$ with respect to λ and s_0.

Corollary 3.3.3. *Suppose that* $S_0 := S_{\beta^0}$ *(resp.* $s_0 := |S_0|$*) satisfies Assumption 3.3.5, the*

compatibility condition, and $\mathcal{T} = \{S_0\}$. *Then under the conditions of Theorem 3.3.1,*

$$\lambda|\hat{\beta} - \beta^0|_1 \leq \frac{10}{3} G\left(\frac{4\lambda\sqrt{s_0}}{\kappa(S_0)}\right), \tag{3.3.8}$$

$$\mathcal{E}(f_{\hat{\beta}}) \leq \frac{8}{3} G\left(\frac{4\lambda\sqrt{s_0}}{\kappa(S_0)}\right). \tag{3.3.9}$$

Theorem 3.3.1 and Corollary 3.3.2 describe how the rate of convergence depends on the sample size n, the dimension p, the oracle excess risk $\mathcal{E}(f_{\beta^*})$, the margin condition quantified by the function $G(\cdot)$ and the moment conditions and strength of dependence which are characterized by q, γ, α_X, α_Y, ι and ν, respectively. Theorem 3.3.1(ii) and Corollary 3.3.2(ii) suggest that, under short-range dependence with exponential decay rate (similar to exponential moment conditions for i.i.d data), we can take $\lambda \asymp (\log p)^{c_1}/n^{c_2}$ for some positive constants c_1, c_2. Then the allowed dimension p can be as large as $\log p = o(n^{c_3})$ with $c_3 = c_2/c_1$. Theorem 3.3.1(i) and Corollary 3.3.2(i) show the results under short-range dependence with polynomial decay rate (similar to bounded moment conditions for i.i.d. data). Roughly speaking, we can take $\lambda \asymp p^{c_4}/n^{c_5}$ for some positive constants c_4 and c_5.

Assume that $\||X_\cdot|_\infty\|_{\gamma,\alpha_X} \asymp 1$, $\|Y_\cdot\|_{q,\alpha_Y} \asymp 1$, $\kappa(S_0) \asymp 1$ and $s_0 = |S_{\beta^0}|$. Consider the maximum marginal (log)-likelihood loss. Assume the typical quadratic margin behavior, *i.e.,* $J(u) = cu^2$, then $G(v) = v^2/(4c)$. We generally can take

$$\lambda \asymp \lambda_0 \asymp p^{1/(2\gamma+2)}(\log(pn))^{1/2} n^{-1/2} + p^{3/(2\gamma+2)}(\log(pn))^{3/2} n^{\nu/\gamma-1} + (\log p)^{3/2} n^{\rho/\chi-1}.$$

By (3.3.8), the estimation error behaves like λs_0, *i.e.,* of the order

$$s_0 p^{1/(2\gamma+2)}(\log(pn))^{1/2} n^{-1/2} + s_0 p^{3/(2\gamma+2)}(\log(pn))^{3/2} n^{\nu/\gamma-1} + s_0(\log p)^{3/2} n^{\rho/\chi-1}. \tag{3.3.10}$$

Similarly, under Corollary 3.3.2.(ii), assume that $\||X_\cdot|_\infty\|_{\psi_\iota} \asymp 1$ and $\|Y_\cdot\|_{\psi_\nu} \asymp 1$. Consider the maximum marginal (log)-likelihood loss and the typical quadratic margin behavior, we

can take

$$\lambda \asymp \lambda_0 \asymp ((\log p)^{1+\iota} + (\log p)^{(1+2\iota)/2}(\log n)^{1/2})n^{-1/2}.$$

The estimation error behaves like λs_0, *i.e.,* of the order

$$s_0((\log p)^{1+\iota} + (\log p)^{(1+2\iota)/2}(\log n)^{1/2})n^{-1/2}. \tag{3.3.11}$$

For i.i.d. data, the asymptotic behavior of Lasso estimator for GLM was studied in van de Geer [2008]. Many other papers investigated high dimensional robust M-estimator, such as Fan et al. [2017] and Loh [2017]. A key technique in these articles is Bousquet's inequality (Bousquet [2002]) or other similar inequalities for empirical process. These methods cannot be used for time dependent data. It is noteworthy that the proof of Theorem 3.3.1 requires new concentration inequalities for high dimensional time series. We establish Bousquet-type inequality for time dependent data, which is shown in the Appendix.

To tackle the problem of heavy-tailed data, we propose to use the robust regularization method in Section 3.2.2. We now turn to analyze the robust estimator. Under the same assumptions, the next theorem shows the rate of convergence of $|\hat{\beta} - \beta|_1$ and the excess risk $\mathcal{E}(f_{\hat{\beta}})$ by the moment condition and the dependence condition. In Theorem 3.3.4, the response is also assumed to be generated from a particular distribution in an exponential family. In other words, there does not exists any model misspecification issue, and the response has sub-exponential tails.

Theorem 3.3.4. *Suppose Assumptions 3.3.1, 3.3.4 and 3.3.5 hold. Suppose the loss function $R(z, y)$ has the form (3.2.5) and the functions $h(\cdot), r(\cdot)$ therein satisfy Assumptions 3.3.2 and 3.3.3, or the loss function itself satisfies Assumptions 3.3.2 and 3.3.3. Let β^* and $\mathcal{E}^*$ be given in (3.3.4) and (3.3.5).*

Assume that $\max_{1\le j\le p} \|X_{ij}\|_q^q < C < \infty$ *and* $\|Y_\cdot\|_{\psi_\nu} < \infty$, *where* $\nu \ge 0$ *and* $q \ge 2$. *Assume that* $|\beta^* - \beta^0|_1 \le M_0/2$ *and* $\mathcal{E}^*/\lambda_0 \le M_0/2$. *Suppose that for some positive constants A and*

B,

$$\begin{aligned}\lambda_0 &= A\left((\log p)^{1-\frac{4}{3q}} + (\log p)^{\frac{1}{2}(1-\frac{4}{3q})}(\log n)^{\frac{1}{2}(1-\frac{4}{3q})}\right) n^{-\frac{1}{2}(1-\frac{4}{3q})} \\ &\quad + B\left((\log p)^{(1+\nu)(1-\frac{4}{3q})} + (\log p)^{\frac{1+2\nu}{2}(1-\frac{4}{3q})}(\log n)^{\frac{1}{2}(1-\frac{4}{3q})}\right) n^{-\frac{1}{2}(1-\frac{4}{3q})} \|Y_\cdot\|_{\psi_\nu}^{1-\frac{4}{3q}},\end{aligned}$$

and $\tau \asymp \lambda_0^{-4/(3q-4)}$. *If* $\lambda \geq (22/3)\lambda_0$, *then with probability at least*

$$1 - C_1 p^{-C_2 A^{\frac{6q}{3q-4}}} - C_3 p^{-C_4 B^{\frac{6q}{(1+2\nu)(3q-4)}}},$$

any Lasso solution $\hat{\beta}$ *of the problem (3.2.6) satisfies*

$$(\lambda - 2\lambda_0)|\hat{\beta} - \beta^*|_1 \leq \frac{10}{3}\mathcal{E}^*, \tag{3.3.12}$$

$$\mathcal{E}(f_{\hat{\beta}}) \leq \frac{8}{3}\mathcal{E}^*, \tag{3.3.13}$$

where the constants $C_1, ..., C_4$ *only depend on* q *and* ν.

Specifically, if $\max_i |Y_i| < C < \infty$, *let*

$$\begin{aligned}\lambda_0 &= A\left((\log p)^{1-\frac{1}{q}} + (\log p)^{\frac{1}{2}(1-\frac{1}{q})}(\log n)^{\frac{1}{2}(1-\frac{1}{q})}\right) n^{-\frac{1}{2}(1-\frac{1}{q})} \\ &\quad + B\left((\log p)^{(1+\nu)(1-\frac{1}{q})} + (\log p)^{\frac{1+2\nu}{2}(1-\frac{1}{q})}(\log n)^{\frac{1}{2}(1-\frac{1}{q})}\right) n^{-\frac{1}{2}(1-\frac{1}{q})},\end{aligned}$$

and $\tau \asymp \lambda_0^{-1/(q-1)}$. *If* $\lambda \geq (22/3)\lambda_0$, *with probability at least*

$$1 - C_1 p^{-C_2 A^{\frac{2q}{q-1}}} - C_3 p^{-C_4 B^{\frac{2q}{(1+2\nu)(q-1)}}},$$

for any Lasso solution $\hat{\beta}$ *of the problem (3.2.6), we have bounds (3.3.12) and (3.3.13).*

Similarly, if the loss function is the marginal (log-)likelihood loss function, sharper results can be effectively established.

Corollary 3.3.5. *Assume the robust loss function is corresponding to the maximum marginal (log-)likelihood loss function in (3.2.4). Assume that* $|\beta^* - \beta^0|_1 \leq M_0/2$ *and* $\mathcal{E}^*/\lambda_0 \leq M_0/2$. *Suppose that*

$$\begin{aligned}\lambda_0 =& A\left((\log p)^{1-\frac{4}{3q}} + (\log p)^{\frac{1}{2}(1-\frac{4}{3q})}(\log n)^{\frac{1}{2}(1-\frac{4}{3q})}\right) n^{-\frac{1}{2}(1-\frac{4}{3q})} \\ &+ B(\log p)^{\frac{1}{2}(1-\frac{4}{3q})} n^{-\frac{1}{2}(1-\frac{4}{3q})} \|Y_\cdot\|_{\psi_\nu}^{1-\frac{4}{3q}},\end{aligned}$$

and $\tau \asymp \lambda_0^{-4/(3q-4)}$. *If* $\lambda \geq (22/3)\lambda_0$, *then under the conditions of Theorem 3.3.4, with probability at least*

$$1 - C_1 p^{-C_2 A^{\frac{6q}{3q-4}}} - C_3 e^{-C_4 B^{\frac{6q}{(1+2\nu)(3q-4)}}},$$

we have bounds (3.3.12) and (3.3.13).

Furthermore, if $\max_i |Y_i| < C < \infty$, *let*

$$\lambda_0 = A\left((\log p)^{1-\frac{1}{q}} + (\log p)^{\frac{1}{2}(1-\frac{1}{q})}(\log n)^{\frac{1}{2}(1-\frac{1}{q})}\right) n^{-\frac{1}{2}(1-\frac{1}{q})} + B(\log p)^{\frac{1}{2}(1-\frac{1}{q})} n^{-\frac{1}{2}(1-\frac{1}{q})},$$

and $\tau \asymp \lambda_0^{-1/(q-1)}$. *If* $\lambda \geq (22/3)\lambda_0$, *with probability at least*

$$1 - C_1 p^{-C_2 A^{\frac{2q}{q-1}}} - C_3 e^{-C_4 B^{\frac{2q}{(1+2\nu)(q-1)}}},$$

we have bounds (3.3.12) and (3.3.13).

From the theorem, our robust regularization method in (3.2.6) can handle the dimensionality

$$\log p = o(n^{c_4}),$$

for some $0 < c_4 < 1$, which guarantees that the "estimation error" in Theorem 3.3.4 converges to zero as long as the tuning parameters τ and λ are properly chosen. Note that the rate in Theorem 3.3.4 is similar to that in Theorem 3.3.1(ii). It also reveals that the estimation

error vanishes faster if higher-order moments of the covariates exist.

Assume that $\max_{1\le j\le p}\|X_{ij}\|_q^q < C$, $\max_i |Y_i| < C < \infty$, $\kappa(S_0) \asymp 1$ and $s_0 = |S_{\beta^0}|$. Consider the maximum marginal (log)-likelihood loss. Assume the typical quadratic margin behavior, *i.e.,* $J(u) = cu^2$, then $G(v) = v^2/(4c)$. We generally can take

$$\lambda \asymp \lambda_0 \asymp \left((\log p)^{1-\frac{1}{q}} + (\log p)^{\frac{1}{2}(1-\frac{1}{q})}(\log n)^{\frac{1}{2}(1-\frac{1}{q})}\right) n^{-\frac{1}{2}(1-\frac{1}{q})}.$$

By (3.3.8), the estimation error behaves like λs_0, *i.e.,* of the order

$$s_0 \left((\log p)^{1-\frac{1}{q}} + (\log p)^{\frac{1}{2}(1-\frac{1}{q})}(\log n)^{\frac{1}{2}(1-\frac{1}{q})}\right) n^{-\frac{1}{2}(1-\frac{1}{q})}. \tag{3.3.14}$$

It is clear that the estimation rate achieved in (3.3.14) is better than the one achieved in (3.3.10).

3.4 Application to Linear Regression

Robust estimation of linear regression can be regarded as a generalized linear model with quadratic loss. In this special case, although Lipschitz property does not hold, we still have an explicit concentration result for our robust estimator. Consider the usual linear regression setup for the response variable Y_i and the covariate vector X_i,

$$Y_i = X_i^T\beta^0 + e_i, \tag{3.4.1}$$

where $\beta^0 \in \mathbb{R}^p$ are unknown coefficients to be estimated, e_i is the error term. Let the loss function $R(f_\beta(X_i), Y_i) = (Y_i - f_\beta(X_i))^2/2$. Different from Theorem 3.3.4 in Section 3.3, if the error has heavy tails, the response Y_i also has heavy tails. In other words, $\|Y_\cdot\|_{\psi_\nu} = \infty$. This motivates us to truncate both the heavy tailed covariates X_i and the response Y_i under L_2 loss. Then similar to (3.2.6), we propose to use the following M-estimator of β^0 with the

generalized ℓ_2 loss to robustify the estimation:

$$\hat{\beta} := \arg\min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (\widetilde{Y}_i - f_\beta(\widetilde{X}_i))^2 + \lambda \left|\beta\right|_1 \right\}, \tag{3.4.2}$$

where

$$f_\beta(\widetilde{X}_i) = \widetilde{X}_i^T \beta. \tag{3.4.3}$$

Here, we choose $\widetilde{Y}_i = \widetilde{Y}_i(\tau_1) = sgn(Y_i)(|Y_i| \wedge \tau_1)$ and $\widetilde{X}_{ij} = sgn(X_{ij})(|X_{ij}| \wedge \tau_2)$ for all $1 \leq j \leq p$, where τ_1, τ_2 are both predetermined thresholds.

To establish the convergence rate of $\hat{\beta}$, we shall adopt the geometric moment contraction (GMC) condition for temporal dependence.

Assumption 3.4.1 (Geometric Moment Contraction (GMC))**.** *Assume that for all i and j, X_{ij} has finite q-th moment, $q > 2$. We say that (X_{ij}) is Geometric Moment Contraction (GMC; see Wu and Shao [2004]) if there exist $a_1 \in (0,1)$, $C_1 = C_1(a_1) > 0$ such that*

$$\|X_{ij} - g_j(..., \varepsilon'_{-1}, \varepsilon'_0, \varepsilon_1, ..., \varepsilon_i)\|_q \leq C_1 a_1^i. \tag{3.4.4}$$

It is easily seen that (3.4.4) is equivalent to

$$\|X_{ij} - g_j(\mathcal{F}_i^*)\|_q \leq C_2 a_2^i, \tag{3.4.5}$$

for some $a_2 \in (0,1)$ and $C_2 = C_2(a_2) > 0$.

Many nonlinear time series models satisfy GMC (cf. Shao and Wu [2007]). Instead of geometric β-mixing, we employ the GMC as an underlying assumption for our asymptotic theory.

To be solvable in the high dimensional regression setting, β^0 is usually assumed to be sparse or weakly sparse, *i.e.,* many elements of β^0 are 0 or small. In particular, we impose

the following weakly sparsity condition.

Assumption 3.4.2 (Weakly Sparsity Condition)**.** *There exits some* $0 \leq \vartheta \leq 1$*, with a uniform radius* K_ϑ *such that*

$$\sum_{j=1}^{p} |\beta_j^*|^\vartheta \leq K_\vartheta. \tag{3.4.6}$$

Note that K_ϑ *might depend on* p*. In the special case* $\vartheta = 0$*, this quantity corresponds to an exact sparsity constraint–that is,* β^0 *has at most* K_0 *nonzero entries.*

Finally, both assumptions together lead to the following main result. We show that $|\hat{\beta} - \beta^0|_2$ and $|\hat{\beta} - \beta^0|_1$ are upper bounded by the same rate as the optimal rate under light tails and i.i.d. data so long as the tuning parameters τ_1, τ_2 and λ are properly chosen.

Theorem 3.4.1. *Assume Assumptions 3.4.1 and 3.4.2 hold. Assume* $|\beta^0|_1 \leq C_\beta < \infty$*. Also assume* $\mathbb{E}|X_{ij}|^4 \leq C < \infty$*, for any* $1 \leq j \leq p$*, and* $\mathbb{E}|Y_i|^4 \leq C < \infty$*. Choose* $\tau_1, \tau_2 \asymp n^{1/4}/(\log p)^{3/4}$ *and* $\lambda = 2c_1\sqrt{\eta \log p/n}$*, where* $\eta > 1$*. Then as long as* $K_\vartheta(\log p/n)^{(1-\vartheta)/2} \leq c_2$*, we have*

$$\mathbb{P}\left(\|\hat{\beta} - \beta^0\|_2^2 > c_3 K_\vartheta \left(\frac{\eta \log p}{n} \right)^{1-\frac{\vartheta}{2}} \right) \leq 3p^{1-\eta}, \tag{3.4.7}$$

and

$$\mathbb{P}\left(\|\hat{\beta} - \beta^0\|_1 > c_4 K_\vartheta \left(\frac{\eta \log p}{n} \right)^{1-\vartheta} \right) \leq 3p^{1-\eta} \tag{3.4.8}$$

for some positive constants $c_1, ..., c_4$*.*

Theorem 3.4.1 reveals that $|\hat{\beta} - \beta^0|_2$ has a convergence rate $K_\vartheta^{1/2}(\log(p)/n)^{1/2-\vartheta/4}$. This rate is the same as the minimax rate (Raskutti et al. [2011]) for weakly sparsity models under light tails and i.i.d data setting. In a special case that $\vartheta = 0$, $|\hat{\beta} - \beta^0|_2$ converges at rate

$K_0^{1/2}(\log(p)/n)^{1/2}$, where K_0 is the number of non-zero elements of β^0. It suggests that our robust method does not lose much information for heavy-tailed data with geometrically decaying temporal dependence. Fan et al. [2016] considered the special setting with i.i.d. data vectors. Our Theorem 3.4.1 is a significant improvement by relaxing the independence assumption, by presenting a Bernstein type inequality for weakly dependent data in the Appendix. In addition, to achieve the optimal statistical rate, we need $K_\vartheta(\log p/n)^{(1-\vartheta)/2} \leq c_2$.

3.5 Simulation Study

In this section, we expound upon concrete instances of our theoretical results and provide some simulation results.

3.5.1 Statistical error

In the first set of simulation, we assess the finite sample performance of the robust procedure and compare it with the standard Lasso procedure in logistic regression and linear regression. The implementation of our robust procedure is simple: truncate or shrink the data appropriately, then apply the standard procedure to the transformed data. The simulations are based on 5000 independent Monte Carlo replications.

We first specify the parameters in the logistic regression. We generate data from independent AR(1) processes, say

$$Z_{ij} = \phi_j Z_{i-1,j} + a_{ij}, \quad 1 \leq j \leq p, \tag{3.5.1}$$

where $\phi_j \sim U[0.2, 0.6]$ or $\phi_j \sim U[-0.6, -0.2]$, and the innovations a_{ij} is given below. To generate cross-dependence, let

$$\Sigma = (\sigma_{jk}) = (\rho^{|j-k|}), \quad 0 < \rho < 1,$$

and $\Sigma^{1/2}$ be the square-root matrix of Σ. We use $X_i = \Sigma^{1/2} Z_i$, $1 \leq i \leq n$, where $Z_i = (Z_{i1}, ..., Z_{ip})'$. We choose the true regression coefficient vector as

$$\beta^* = (3, ..., 3, 0, ..., 0)',$$

where the first 20 elements are all 3 and the rest are all 0. Let $\theta_i = 1/(1 + \exp(-X_i'\beta^*))$ be the probability of success of the Bernoulli distribution of Y_i. Thus, Y_i is a random draw from Bern(θ_i). We run simulations for sample sizes $n = 200, 300, 400, 500, 800, 1000, 2000, 3000$, and choose the number of parameters p to be 400, dependence parameter $\rho = 0.5$. For each case, additional 500 observations are generated and used for out-of-sample predictions. To entertain various shapes of covariate distributions, we consider the following two scenarios for a_{ij} of (3.5.1):

(a). the Student-t distribution with degrees of freedom 5 times 1/10, i.e. $0.1t_5$;

(b). a log-normal distribution with parameters 0 and 0.25^2 times 1/5, i.e. $0.2\,\text{logNormal}(0, 0.5^2)$.

They represent heavy-tailed symmetric and asymmetric distributions, respectively. To meet the model assumptions, the covariates are standardized to have *mean 0*. The constants used are chosen to ensure appropriate signal-to-noise ratio and θ_i not trivially equal to either 0 or 1 for better presentation. The numerical performance of the robust procedure and standard Lasso procedure under the two scenarios is evaluated by the following five measurements.

(a). L_2 error, which is defined as $|\hat{\beta} - \beta^*|_2$;

(b). L_1 error, which is defined as $|\hat{\beta} - \beta^*|_1$;

(c). the number of false positive results, FP, which is the number of noise covariates that are selected;

(d). the number of false negative results, FN, which is the number of signal covariates that are not selected;

(e). one-step-ahead forecast errors of a total of 500 out-of-sample observations, FE, which is the misclassification rate.

For the robust Lasso, we choose the optimal tuning parameters λ and τ on the basis of 100 independent validation data sets. For each case, we run a two-dimensional grid search to find the best (λ, τ) pair that minimizes the misclassification rate of the 100 validation data sets. Then the optimal pair is used in the simulation. Similar methods are applied in choosing the tuning optimal parameters in other models. The means of the five performance measures are summarized in Table 3.1.

The results of Table 3.1 show that our robust Lasso method has certain advantages over the standard Lasso method when the covariates are heavy-tailed. The results are in agreement with the theorems. As the sample size increases, the performance measures improve. In both symmetric and asymmetric covariates cases, our robust method has smaller L_1 and L_2 errors. The advantage of our robust method is more pronounced when the sample size is large. As expected, FP increases slightly with sample size n, but FN approaches zero as n increases.

We also investigate the empirical properties of the proposed method in linear regression. We again generate data from independent AR(1) model (3.5.1). Analogous to the logistic regression, we set $X_i = \Sigma^{1/2} Z_i$. For response, we generate time series process Y_i from the model,

$$Y_i = X_i' \beta^* + e_i, \tag{3.5.2}$$

where e_i is given below. The following two scenarios for a_{ij} are considered:

(a). the Student-t distribution with degrees of freedom 5, t_5;

(b). a log-normal distribution with parameters 0 and 0.25^2, logNormal$(0, 0.25^2)$.

For the distributions of error e_i, we choose:

Table 3.1: Simulation results of Lasso and robust Lasso for logistic regression ($p = 400, \rho = 0.5$, where n is the sample size. The results are based on 5000 replications.

Scenario		Student t_5		LogNormal$(0, 0.25)$	
		Lasso	robust Lasso	Lasso	robust Lasso
$n = 200$	L_2 loss	11.53	11.32	11.59	11.39
	L_1 loss	50.14	48.38	50.30	48.60
	FP	0.96	0.88	0.88	0.79
	FN	11.30	10.85	11.84	11.41
	FE	28.11%	27.24%	29.08%	28.50%
$n = 300$	L_2 loss	10.22	9.85	10.28	9.94
	L_1 loss	43.69	40.89	43.80	41.15
	FP	1.69	1.51	1.61	1.44
	FN	5.98	5.37	6.59	5.92
	FE	22.02%	20.82%	23.09%	22.00%
$n = 400$	L_2 loss	9.46	8.95	9.48	9.05
	L_1 loss	40.21	36.62	40.10	36.85
	FP	2.28	2.04	2.24	1.97
	FN	3.53	3.08	4.08	3.44
	FE	20.33%	19.24%	21.34%	20.23%
$n = 500$	L_2 loss	8.87	8.28	8.88	8.34
	L_1 loss	37.56	33.51	37.38	33.61
	FP	2.64	2.32	2.54	2.23
	FN	2.23	1.84	2.66	2.14
	FE	19.44%	18.37%	20.40%	19.33%
$n = 800$	L_2 loss	7.66	6.82	7.67	6.93
	L_1 loss	32.39	27.29	32.15	27.52
	FP	3.28	2.94	3.12	2.71
	FN	0.56	0.41	0.79	0.53
	FE	18.18%	17.17%	19.22%	18.13%
$n = 1000$	L_2 loss	7.15	6.20	7.11	6.26
	L_1 loss	30.27	24.71	29.85	24.82
	FP	3.49	3.13	3.38	2.96
	FN	0.24	0.17	0.35	0.22
	FE	17.83%	16.82%	18.83%	17.77%
$n = 2000$	L_2 loss	5.67	4.42	5.61	4.43
	L_1 loss	24.15	17.48	23.69	17.51
	FP	3.82	3.50	3.69	3.34
	FN	0.0018	0.0008	0.0060	0.0014
	FE	17.19%	16.22%	18.12%	17.05%
$n = 3000$	L_2 loss	4.92	3.53	4.84	3.52
	L_1 loss	21.07	13.87	20.50	13.90
	FP	3.90	3.56	3.79	3.41
	FN	0	0	0	0
	FE	16.94%	15.98%	17.97%	16.93%

(a). the Student-t distribution with degrees of freedom 3 times 20, i.e. $20t_3$, the standard deviation of which is about 34.64;

(b). a log-normal distribution with parameters 0 and 0.5^2 times 20, i.e. $20\,\text{logNormal}(0, 0.5^2)$, the standard deviation of which is about 12.08.

Again, the covariates and the errors are standardized to have *mean 0*, and the constants used are chosen to ensure appropriate signal-to-noise ratio for better presentation. Set $n = 50$, 100, 200, 300, 400, 800, 1000, 2000. We use the root mean squared forecast error (RMSE) to measure one-step-ahead forecasts of a total of 200 out-of-sample predictions. The results are reported in Table 3.2.

The results of Table 3.2 are also in agreement with the theorem. In particular, as expected, the RMSE approaches the standard deviation of e_i as the sample size increases. In general, similar to the logistic regression, the robust estimator outperforms the non-robust one. This is particularly so for the case of heavy-tailed noises $e_i \sim 20t_3$. But as the sample size increases, the difference between the robust procedure and the standard procedure gradually decreases. The out-of-sample predictions seem to work well when the sample size $n \geq p$. For log-normal errors, FN seems to be higher than FP. Both are sizable when the sample size is small.

In conclusion, our robust method is more flexible than the standard Lasso. The above two cases evidently show that the robust procedure outperforms the standard procedure under the setting with heavy-tailed covariates and errors. The truncation parameter enables the robust method to render consistently satisfactory results under all scenarios considered in our simulation.

Table 3.2: Simulation results of Lasso and robust Lasso for linear regression ($p = 400, \rho = 0.5$), where n is the sample size and the results are based on 5000 replications

Scenario		Student t		LogNormal	
		Lasso	robust Lasso	Lasso	robust Lasso
$n = 50$	L_2 loss	24.01	22.31	35.11	31.90
	L_1 loss	148.01	140.39	218.33	201.09
	FP	44.26	44.12	47.52	47.31
	FN	12.23	12.16	15.12	14.94
	RMSE	49.86	48.10	16.46	15.78
$n = 100$	L_2 loss	25.29	23.44	40.86	37.04
	L_1 loss	200.83	187.76	331.69	302.13
	FP	51.31	50.79	57.69	57.12
	FN	8.05	7.73	11.52	11.18
	RMSE	49.72	47.66	17.19	16.39
$n = 200$	L_2 loss	12.46	11.04	24.52	22.78
	L_1 loss	66.12	52.68	205.78	187.80
	FP	11.89	7.83	33.56	31.73
	FN	7.19	5.85	11.06	10.64
	RMSE	40.84	39.13	15.10	14.73
$n = 300$	L_2 loss	9.67	8.03	12.64	10.36
	L_1 loss	39.98	36.83	55.58	52.13
	FP	3.49	3.34	7.81	7.91
	FN	5.78	4.47	10.30	10.25
	RMSE	38.25	37.37	13.34	13.26
$n = 400$	L_2 loss	8.77	8.04	11.72	9.24
	L_1 loss	37.53	32.60	50.27	47.80
	FP	2.99	2.82	1.50	1.71
	FN	3.67	2.67	10.27	9.51
	RMSE	37.04	36.18	13.13	13.01
$n = 800$	L_2 loss	6.42	5.78	9.44	8.95
	L_1 loss	25.13	22.34	38.17	36.04
	FP	2.56	2.52	1.37	1.61
	FN	0.86	0.41	5.94	4.67
	RMSE	35.11	34.65	12.58	12.51
$n = 1000$	L_2 loss	5.79	5.15	8.75	8.28
	L_1 loss	22.53	19.84	34.74	32.82
	FP	2.58	2.49	0.71	0.90
	FN	0.45	0.16	4.29	3.29
	RMSE	34.82	34.42	12.46	12.40
$n = 2000$	L_2 loss	4.11	3.56	6.69	6.22
	L_1 loss	15.85	12.60	25.65	23.94
	FP	2.22	2.13	0.39	0.49
	FN	0.0376	0.0014	1.10	0.67
	RMSE	34.39	34.14	12.26	12.23

3.5.2 Convergence speed

To study how the dependence and the heavy tails affect the convergence speed of the estimators, we consider the two tail probability ratios for linear regression

$$\Lambda_1(t) = \frac{\mathbb{P}(|\hat{\beta} - \beta^*|_1 \geq t)}{\mathbb{P}(|\hat{\beta}^\dagger - \beta^*|_1 \geq t)}, \tag{3.5.3}$$

$$\Lambda_2(t) = \frac{\mathbb{P}(|\hat{\beta} - \beta^*|_2 \geq t)}{\mathbb{P}(|\hat{\beta}^\dagger - \beta^*|_2 \geq t)}, \tag{3.5.4}$$

where $\hat{\beta}^\dagger$ is the Lasso estimator of β^* in the same models as in Section 3.5.1 with $\phi_j = 0$, a_{ij} being i.i.d. standard normal random variables and e_i being i.i.d. $20N(0,1)$ random variables, and $\hat{\beta}$ is the Lasso estimator of the models in (3.5.1) and (3.5.2) with different serial dependence and tail conditions. We choose $\phi_j = -0.9, -0.8, ..., 0.9$ for all $1 \leq j \leq p$, a_{ij} being t_5 and e_i being $20t_3$. The denominator in (3.5.3) can be viewed as benchmark probabilities. By cross validation, the optimal threshold value λ for benchmark light-tail model is around 3. Hence in our simulations, we use $\lambda = 3$, $n = 100$, $p = 400$ and $\rho = 0.5$. For the benchmark, based on 10^6 repetitions, the 99% and 99.9% quantiles of $|\hat{\beta}^\dagger - \beta^*|_2$ (resp. $|\hat{\beta}^\dagger - \beta^*|_1$) are 15.327, 16.441 (resp. 84.865 and 91.677). For each ϕ_j, the tail probability $\mathbb{P}(|\hat{\beta} - \beta^*|_2 \geq t)$ (resp. $\mathbb{P}(|\hat{\beta} - \beta^*|_1 \geq t)$) is estimated by 10^6 repetitions of $|\hat{\beta} - \beta^*|_2$ (resp. $|\hat{\beta} - \beta^*|_1$) that are larger than the critical value t, respectively.

Table 3.3 reports the results of $\Lambda_2(t)$ with $t = 15.327$ and 16.441 and $\Lambda_1(t)$ with $t = 84.865$ and 91.677. They correspond to the ratio between $\mathbb{P}(|\hat{\beta} - \beta^*|_2 \geq t)$ or $\mathbb{P}(|\hat{\beta} - \beta^*|_1 \geq t)$ under various serial dependence and moment conditions and the benchmark tail probabilities 0.01 and 0.001, respectively. As expected from our theoretical results, Table 3.3 shows that the upper tail probabilities $\mathbb{P}(|\hat{\beta} - \beta^*|_1 \geq t)$ and $\mathbb{P}(|\hat{\beta} - \beta^*|_2 \geq t)$ with larger t are affected more than the one with smaller t; heavy tails can lead to larger $\mathbb{P}(|\hat{\beta} - \beta^*|_1 \geq t)$ and $\mathbb{P}(|\hat{\beta} - \beta^*|_2 \geq t)$, thus inflating the tail probability ratios. Furthermore, compared with the standard Lasso method, the robust Lasso procedure significantly improves the convergence speed as the ratios $\Lambda_1(t)$ and $\Lambda_2(t)$ are much smaller. In addition, the tail probability ratios

decrease as the serial dependence increases (larger $|\phi_j|$). But if the dependence strength is too strong (ϕ_j is close to 1), the tail probability ratios increase again.

Table 3.3: Simulated values of the tail probability ratios $\Lambda_1(t)$ and $\Lambda_2(t)$ of Equation (3.5.3) for Lasso and robust Lasso (R-Lasso) procedures. 1000000 replications are used to evaluate the probabilities.

	$\Lambda_1(t)$				$\Lambda_2(t)$			
	$t = 84.865$		$t = 91.677$		$t = 15.327$		$t = 16.441$	
ϕ_j	Lasso	R-Lasso	Lasso	R-Lasso	Lasso	R-Lasso	Lasso	R-Lasso
-0.9	16.4	6.4	92.1	15.9	20.5	10.0	107.3	25.8
-0.8	26.0	6.3	163.1	16.1	25.3	7.5	148.4	17.6
-0.7	37.5	8.6	252.8	25.1	32.7	8.3	202.2	20.6
-0.6	48.3	12.5	346.5	43.2	40.0	10.7	260.3	29.1
-0.5	57.5	17.5	430.7	66.4	46.5	13.8	312.7	41.9
-0.4	64.7	22.8	501.5	98.9	51.9	17.6	362.2	60.0
-0.3	70.0	28.1	558.9	129.1	56.6	21.5	403.0	77.9
-0.2	73.3	32.2	595.2	158.0	59.5	24.9	429.4	97.8
-0.1	75.7	35.2	625.0	182.0	61.9	27.5	454.2	112.5
0	76.6	36.3	633.8	189.9	63.0	28.5	464.5	120.2
0.1	76.0	35.8	628.2	184.8	62.8	28.5	462.3	119.2
0.2	74.8	33.8	612.2	171.4	62.0	27.0	454.8	112.1
0.3	71.8	30.4	577.7	147.6	59.8	25.1	436.1	100.2
0.4	67.8	26.3	534.8	119.9	57.5	22.9	417.1	90.9
0.5	61.7	22.0	474.0	92.4	54.5	21.5	388.7	85.7
0.6	55.0	18.4	405.8	72.1	52.0	21.8	374.1	96.4
0.7	47.8	16.8	338.4	64.8	52.2	26.7	384.6	143.7
0.8	42.8	20.1	292.3	86.5	57.5	40.1	455.4	273.6
0.9	50.3	40.1	365.0	247.5	73.7	66.2	655.9	561.5

3.6 Real Data Analysis

In this section, we use a real dataset to illustrate the performance of the Lasso procedures. We consider a high frequency financial dataset that was first studied by Tsay and Chen [2018]. The data consist of the high-frequency trading of Walgreens stock on February 6, 2017. The data are available from the TAQ database of the New York Stock Exchange. Let y_i^* be the observed price change of the ith trade during the normal trading hours between 9:30AM to 4:00PM, Eastern Time. Due to the discreteness of y_i^*, as suggested by Tsay and

Chen [2018] Example 4.2, we divide the price changes into 7 categories, namely,

$$< -0.02, \quad [-0.02, -0.01), \quad [-0.01, 0), \quad 0, \quad (0, 0.01], \quad (0.01, 0.02], \quad > 0.02,$$

where the unit is one U.S. dollar. The category associated with y_i^* is thus defined as Y_i. If $y_i^* < -0.02$, we have $Y_i = 1$, if $-0.02 \leq y_i^* < -0.01$, $Y_i = 2$, and so on. We define t_i to be the duration between $(i-1)$th and ith transactions, which is measured in seconds. Let s_i be the normalized size of the transaction, which is the trading volume (number of shares) of the ith trade divided by 100. We also define six dummy variables for the price changes. Specifically, let

$$z_{i,j} = \begin{cases} 1 & \text{if } Y_i = j, \\ 0 & \text{if } Y_i \neq j, \end{cases} \quad j = 2, ..., 7. \tag{3.6.1}$$

Denote $z_i = (z_{i,2}, ..., z_{i,7})'$. Then in our study, we employ the following $9d$ input variables,

$$X_i := \{z_{i-l}, y_{i-l}^*, t_{i-l}, s_{i-l} | l = 1, 2, ..., d\},$$

where d denotes lag in the time series. For this dataset, we want to predict trade-by-trade price change. On February 6, 2017, there were 29275 transactions available for the Walgreens stock. We use the first 27275 observations as the training subsample and reserve the last 2000 observations for out-of-sample prediction for comparison.

The well known ordered probit model (Hausman et al. [1992]) with $d = 3$ is used as benchmark. Setting $d = 3$, Tsay and Chen [2018] compare the benchmark with several network models. In this particular instance, the 27-10-1 (feedforward) neural network appears to perform the best among the network models considered. The prediction results for both models are reported in Tables 3.4 and 3.5, respectively. In comparison, we apply Lasso methods with multinomial logistic regression to the data. Besides the main effects X_i, we also add two-way interactions between $y_{i-l}^*, t_{i-l}, s_{i-l}$ and z_{i-l}, $l = 1, 2, ..., d$. That is, there

are a total of $27d$ input variables. Note adding two-way interactions does not improve the predictions of the benchmark. In both standard and robust Lasso procedures, we choose $d = 16$. The optimal values of tuning parameters are chosen by a two-dimensional grid search using BIC; see also Section 3.2.2. The prediction results are summarized in Tables 3.6 and 3.7.

Table 3.4 shows that the ordered probit model does not perform well in prediction. As a matter of fact, the model predicts no price change for all of the last 2000 transactions. This is not surprising as the probability of no price change in the training subsample is 71.3%. The forecasting results in Table 3.5 show that the 27-10-1 neural network is able to correctly predict 3, 47, 1389, and 11 times for Categories 1, 3, 4, and 5, respectively. Its misclassification rate is 27.5%. In comparison, the standard Lasso procedure for multinomial logistic regression correctly predicts 2, 5, 35, 1378, 20, and 3 times for Categories 1, 2, 3, 4, 5, and 6, respectively. The corresponding misclassification rate is 27.85%. And the proposed robust Lasso procedure for multinomial logistic regression correctly predicts 3, 6, 32, 1378, 20, 3 and 1 times for Categories 1 to 7, respectively. Its misclassification rate is also 27.85%. The standard Lasso procedure and the robust Lasso procedure perform almost the same for the misclassification rate but the latter improves the predictions in most categories.

In some scenarios of multiclass classification, researchers assign different costs for classifying certain classes (see Duchi et al. [2016]); for example, it may be less costly to misclassify a benign tumor as cancerous than the opposite. In our particular example, the classes are unbalanced and we are more interested in big price changes. To this end, we use a cost matrix $W = (w_{jk})_{j,k=1}^{7} \in \mathbb{R}_{+}^{7\times 7}$, where $w_{jk} \geq 0$ is the cost (weights) for classifying an observation of class j as class k. We assume $w_{jj} = 0$ for each j. Then the weighted empirical error is defined as

$$err_W(g) = \frac{1}{n}\sum_{i\leq n}\sum_{j,k=1}^{7} w_{jk}\mathbf{1}_{\{g(X_i)=j, Y_i=k\}}, \tag{3.6.2}$$

where g is a classifier. We further define $w_{jk}(k \neq j)$ as the reciprocal of the proportion of class j among all the 29275 observations, *i.e.,*

$$w_{jk} = \left(\frac{1}{29275} \sum_{i=1}^{29275} \mathbf{1}_{\{Y_i=j\}} \right)^{-1}, \quad k \neq j \tag{3.6.3}$$

Then, the weighted empirical errors for the 27-10-1 network, standard Lasso procedure and robust Lasso procedure are 4.09, 4.11, 3.99, respectively. Therefore, both the standard Lasso procedure and robust Lasso procedure are compatible to neural networks. From the weighted empirical error perspective, our robust Lasso procedure fares best. This example demonstrates that the standard Lasso procedure and robust Lasso procedure can be helpful in modeling trade-by-trade price changes in financial market.

Table 3.4: Forecast tabulation for the ordered Probit model

		Predicted Categories						
	Y_i	1	2	3	4	5	6	7
Real Categories	1	0	0	0	4	0	0	0
	2	0	0	0	100	0	0	0
	3	0	0	0	180	0	0	0
	4	0	0	0	1437	0	0	0
	5	0	0	0	174	0	0	0
	6	0	0	0	100	0	0	0
	7	0	0	0	5	0	0	0

Table 3.5: Forecast tabulation for the 27-10-1 feedforward neural network

		Predicted Categories						
	Y_i	1	2	3	4	5	6	7
Real Categories	1	3	0	0	1	0	0	0
	2	0	0	2	98	0	0	0
	3	0	0	47	129	4	0	0
	4	2	0	38	1389	8	0	0
	5	0	0	11	152	11	0	0
	6	0	0	5	95	0	0	0
	7	0	0	0	5	0	0	0

Table 3.6: Forecast tabulation for the standard Lasso method

		Predicted Categories						
	Y_i	1	2	3	4	5	6	7
Real Categories	1	2	1	0	1	0	0	0
	2	0	5	2	91	1	1	0
	3	0	0	35	134	11	0	0
	4	1	7	23	1378	20	6	2
	5	0	1	6	146	20	1	0
	6	0	1	0	94	2	3	0
	7	1	0	0	4	0	0	0

Table 3.7: Forecast tabulation for the robust Lasso method

		Predicted Categories						
	Y_i	1	2	3	4	5	6	7
Real Categories	1	3	0	0	1	0	0	0
	2	0	6	2	90	1	1	0
	3	0	0	32	141	7	0	0
	4	0	7	24	1378	22	5	1
	5	0	1	5	147	20	1	0
	6	0	1	0	96	0	3	0
	7	1	0	0	3	0	0	1

3.7 Inequalities for Empirical Processes of High-dimensional Time Series

The proof of the main theorem requires additional new concentration inequalities for empirical processes of time series. Analogous to Bousquet's inequality (Bousquet [2002]) for i.i.d data, we present concentration inequalities for both heavy-tailed and light-tailed high dimensional time series under the functional dependence measure framework. The result may be of independent interest. To simplify the notation, denote $r(f_\beta(X_i)) = (r(f_{\beta(1)}(X_i)), ..., r(f_{\beta(m)}(X_i)))$, $r(f_{\beta^*}(X_i)) = (r(f_{\beta^*}(X_i)), ..., r(f_{\beta^*}(X_i)))$ (m-dimensional vector) and $\bar{r}(\cdot) = r(\cdot) - \mathbb{E}r(\cdot)$. Denote $\mathcal{F}_i^l = \sigma(\varepsilon_l, ..., \varepsilon_i)$ with $l \leq i$, $\mathcal{F}_i = \sigma(\cdots, \varepsilon_{i-1}, \varepsilon_i)$. Write $\mathcal{P}_l(\cdot) = \mathbb{E}(\cdot|\mathcal{F}_l) - \mathbb{E}(\cdot|\mathcal{F}_{l-1})$.

Theorem 3.7.1. *Let $|\beta - \beta^*|_0 = s$, $t = s(\log p + \log n)$, and $C_{q,\alpha}$ be a constant depending on q and α. Assume $\||X.|_\infty\|_{q,\alpha} < \infty$, where $q > 2$ and $\alpha > 0$. Suppose that Assumptions 3.3.2*

and 3.3.3 hold. (i) If $\alpha > 1/2 - 1/q$*, then for* $x \geq \sqrt{nt}\,\||X.|_\infty\|_{2,\alpha} + n^{1/q}t^{3/2}\,\||X.|_\infty\|_{q,\alpha}$*,*

$$\mathbb{P}\left(\sup_{\substack{|\beta-\beta^*|_1\leq M,\\ |\beta-\beta^*|_0=s}}\left|\sum_{i=1}^{n}[\bar{r}(f_\beta(X_i)) - \bar{r}(f_{\beta^*}(X_i))]\right| \geq Mx\right) \leq \frac{C_{q,\alpha}nt^{q/2}\,\||X.|_\infty\|_{q,\alpha}^q}{x^q} + C_{q,\alpha}\exp\left(-\frac{C_{q,\alpha}x^2}{n\,\||X.|_\infty\|_{2,\alpha}^2}\right). \tag{3.7.1}$$

(ii) If $0 < \alpha < 1/2 - 1/q$*, then for* $x \geq \sqrt{nt}\,\||X.|_\infty\|_{2,\alpha} + n^{1/2-\alpha}t^{3/2}\,\||X.|_\infty\|_{q,\alpha}$*,*

$$\mathbb{P}\left(\sup_{\substack{|\beta-\beta^*|_1\leq M,\\ |\beta-\beta^*|_0=s}}\left|\sum_{i=1}^{n}[\bar{r}(f_\beta(X_i)) - \bar{r}(f_{\beta^*}(X_i))]\right| \geq Mx\right) \leq \frac{C_{q,\alpha}n^{q/2-\alpha q}t^{q/2}\,\||X.|_\infty\|_{q,\alpha}^q}{x^q} + C_{q,\alpha}\exp\left(-\frac{C_{q,\alpha}x^2}{n\,\||X.|_\infty\|_{2,\alpha}^2}\right). \tag{3.7.2}$$

Proof. Denote $\mathcal{A} := \{\beta : |\beta - \beta^*|_1 \leq M, |\beta - \beta^*|_0 = s\}$. We divide each coordinate of $\beta - \beta^*$ into D segments. Define set $\mathcal{A}_m := \{\beta^{(1)}, ..., \beta^{(m)}\} \subset \mathcal{A}$, such that for any $\beta \in \mathcal{A}, \min_{\beta^{(j)}\in\mathcal{A}_m} |\beta - \beta^{(j)}|_1 \leq M/D$. The number of ϵ-balls that is required to cover a k-dimensional unit diamond is bounded by $(4/\epsilon)^k$. Since $\beta - \beta^*$ is a p-dimensional vector with at most s nonzero coordinates, the covering number we require is at most $(4pD)^s$. Then

$m = |\mathcal{A}_m| \asymp p^s D^s$. Let $\beta^h := \arg\min_{\beta^{(j)} \in \mathcal{A}_m} |\beta - \beta^{(j)}|_1$. Then,

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{\substack{|\beta-\beta^*|_1 \le M, \\ |\beta-\beta^*|_0 = s}} \left| \sum_{i=1}^n [\bar{r}(f_\beta(X_i)) - \bar{r}(f_{\beta^*}(X_i))] \right| \ge Mx \right) \\
&\le \mathbb{P}\left(\sup_{\substack{|\beta-\beta^*|_1 \le M, \\ |\beta-\beta^*|_0 = s}} \left| \sum_{i=1}^n [\bar{r}(f_{\beta^h}(X_i)) - \bar{r}(f_{\beta^*}(X_i))] \right| \ge \frac{Mx}{2} \right) \\
&\quad + \mathbb{P}\left(\sup_{\substack{|\beta-\beta^*|_1 \le M, \\ |\beta-\beta^*|_0 = s}} \left| \sum_{i=1}^n [\bar{r}(f_\beta(X_i)) - \bar{r}(f_{\beta^h}(X_i))] \right| \ge \frac{Mx}{2} \right) \\
&= \mathbb{P}\left(\sup_{\beta^{(j)} \in \mathcal{A}_m} \left| \sum_{i=1}^n [\bar{r}(f_{\beta^{(j)}}(X_i)) - \bar{r}(f_{\beta^*}(X_i))] \right| \ge \frac{Mx}{2} \right) \\
&\quad + \mathbb{P}\left(\sup_{\substack{|\beta-\beta^*|_1 \le M, \\ |\beta-\beta^*|_0 = s}} \left| \sum_{i=1}^n [\bar{r}(f_\beta(X_i)) - \bar{r}(f_{\beta^h}(X_i))] \right| \ge \frac{Mx}{2} \right) \\
&:= \mathrm{I}_1 + \mathrm{I}_2.
\end{aligned}
$$

We first consider I_2. By the Lipschitz property, Assumption 3.3.2, $\left| r(f_\beta(X_i)) - r(f_{\beta^h}(X_i)) \right| \le |(\beta - \beta^h)^T X_i|$. Some basic calculations show that

$$
\|(\beta - \beta^h)^T X_i\|_q \le M\Omega_{0,q}/D \le M\||X_\cdot|_\infty\|_{q,\alpha}/D.
$$

Then,

$$
\begin{aligned}
\mathrm{I}_2 &\le \mathbb{P}\left(2n \max_i \sup_{|\beta-\beta^*|_1 \le M} |(\beta - \beta^h)^T X_i| \ge \frac{Mx}{2} \right) \\
&\le \frac{C_{q,2} n^{q+1} \||X_\cdot|_\infty\|_{q,\alpha}^q}{D^q x^q}.
\end{aligned}
$$

Note $\log m \asymp s \log p + s \log D$. We can set $\log m = s(\log p + \log n)$ and $D \asymp n$. Thus,

$$\mathrm{I}_2 \quad \leq \quad \frac{C_q n \| |X_\cdot|_\infty \|_{q,\alpha}^q}{x^q}. \tag{3.7.3}$$

It remains to consider I_1. Let $t = 1 \vee \log m$, where $\log m = s(\log p + \log n)$. Note that for any vector $u = (u_1, ..., u_m)^T$, $|u|_\infty \leq |u|_t \leq m^{1/t}|u|_\infty$. Let $L = \lfloor \log n / \log 2 \rfloor$, $\tau_l = 2^l$ if $1 \leq l \leq L$, $\tau_L = n$ and $\tau_0 = 0$. To simplify the notation, denote $r(f_\beta(X_i)) = (r(f_{\beta^{(1)}}(X_i)), ..., r(f_{\beta^{(m)}}(X_i)))$, $r(f_{\beta^*}(X_i)) = (r(f_{\beta^*}(X_i)), ..., r(f_{\beta^*}(X_i)))$. Let $W_n = \sum_{i=1}^n [\bar{r}(f_\beta(X_i)) - \bar{r}(f_{\beta^*}(X_i))]$, $W_{n,k} = \sum_{i=1}^n \mathbb{E}([\bar{r}(f_\beta(X_i)) - \bar{r}(f_{\beta^*}(X_i))]|\varepsilon_{i-k}, ..., \varepsilon_i)$, $X_{i,k} = \mathbb{E}(X_i|\varepsilon_{i-k}, ..., \varepsilon_i)$. Define $Q_{n,l} = W_{n,\tau_l} - W_{n,\tau_{l-1}}$ for $1 \leq l \leq L$ and write

$$W_n = W_{n,0} + W_n - W_{n,n} + \sum_{l=1}^{L} Q_{n,l}. \tag{3.7.4}$$

Note that $W_{n,n} - W_{n,0} = \sum_{l=1}^L Q_{n,l}$. By Lemma 3.7.7 and Jensen's inequality,

$$\begin{aligned}
&\| |W_{n,k+1} - W_{n,k}|_\infty \|_q \\
&= \left\| \left| \sum_{i=1}^n [\mathbb{E}(r(f_\beta(X_i)) - r(f_{\beta^*}(X_i))|\mathcal{F}_i^{i-k-1}) - \mathbb{E}(r(f_\beta(X_i)) - r(f_{\beta^*}(X_i))|\mathcal{F}_i^{i-k})] \right|_t \right\|_q \\
&\leq C_q \sqrt{nt} \left\| \left| \mathbb{E}(r(f_\beta(X_i)) - r(f_{\beta^*}(X_i))|\mathcal{F}_i^{i-k-1}) - \mathbb{E}(r(f_\beta(X_i)) - r(f_{\beta^*}(X_i))|\mathcal{F}_i^{i-k}) \right|_t \right\|_q \\
&\leq C_q \sqrt{nt} \left\| \left| r(f_\beta(X_i)) - r(f_{\beta^*}(X_i)) - r(f_\beta(X_{i,i-k-1})) + r(f_{\beta^*}(X_{i,i-k-1})) \right|_t \right\|_q .
\end{aligned}$$

Under Assumption 3.3.2,

$$\| |W_{n,k+1} - W_{n,k}|_\infty \|_q \leq C_q \sqrt{nt} M \left\| |X_i - X_{i,i-k-1}|_t \right\|_q \leq C_q \sqrt{nt} M \omega_{k+1,q}.$$

Then

$$\||W_n - W_{n,n}|_\infty\|_q \le \sum_{k=n}^{\infty} \||W_{n,k+1} - W_{n,k}|_\infty\|_q \le M \sum_{k=n}^{\infty} C_q \sqrt{nt}\omega_{k+1,q} = C_q M \sqrt{nt}\Omega_{n+1,q}.$$

By Markov's inequality, we have

$$\mathbb{P}(|W_n - W_{n,n}|_t \ge Mx) \le \frac{\||W_n - W_{n,n}|_t\|_q^q}{M^q x^q} \le \frac{C_q (nt)^{q/2} \Omega_{n+1,q}^q}{x^q}. \tag{3.7.5}$$

Note that $\Omega_{n+1,q} \le \||X_\cdot|_\infty\|_{q,\alpha} n^{-\alpha}$.

Recall $X_{i,0} = \mathbb{E}(X_i|\varepsilon_i)$ and $W_{n,0} = \sum_{i=1}^n \mathbb{E}[\bar{r}(f_\beta(X_i)) - \bar{r}(f_{\beta^*}(X_i))|\varepsilon_i]$. Under Assumption 3.3.3,

$$\begin{aligned} |\mathbb{E}[r(f_{\beta^{(j)}}(X_i)) - r(f_{\beta^*}(X_i))|\varepsilon_i]| &\le |\mathbb{E}[(\beta^{(j)} - \beta^*)X_i|\varepsilon_i]| \le |\beta^{(j)} - \beta^*|_1 |\mathbb{E}(X_i|\varepsilon_i)|_\infty \\ &\le M|X_{i,0}|_\infty. \end{aligned}$$

Note that $\mathbb{E}[\bar{r}(f_\beta(X_i)) - \bar{r}(f_{\beta^*}(X_i))|\varepsilon_i]$ are independent for different i. By Fuk-Nagaev inequality in Lemma A.2 of Chernozhukov et al. [2013], we have

$$\begin{aligned} &\mathbb{P}(|W_{n,0}|_\infty - 2\mathbb{E}|W_{n,0}|_\infty \ge Mx) \\ &\le \frac{C_{q,\alpha} \sum_{i=1}^n \mathbb{E} \max_{1\le j\le m} \left|\mathbb{E}[r(f_{\beta^{(j)}}(X_i)) - r(f_{\beta^*}(X_i))|\varepsilon_i]\right|^q}{M^q x^q} \\ &+ \exp\left(-\frac{M^2 x^2}{3 \max_{1\le j\le m} \sum_{i=1}^n \mathbb{E}|\mathbb{E}[r(f_{\beta^{(j)}}(X_i)) - r(f_{\beta^*}(X_i))|\varepsilon_i]|^2}\right) \\ &\le \frac{C_{q,\alpha} n \||X_{i,0}|_\infty\|_q^q}{x^q} + \exp\left(-\frac{x^2}{3n\||X_{i,0}|_\infty\|_2^2}\right). \end{aligned}$$

Then, by Lemma A.3 of Chernozhukov et al. [2013],

$$\begin{aligned}\mathbb{E}|W_{n,0}|_\infty &\lesssim \sqrt{t}\sqrt{\max_j \sum_{i=1}^n \mathbb{E}(\mathbb{E}[r(f_{\beta(j)}(X_i)) - r(f_{\beta^*}(X_i))|\varepsilon_i])^2} \\ &\quad + t\sqrt{\mathbb{E}\max_i \max_j |\mathbb{E}[r(f_{\beta(j)}(X_i)) - r(f_{\beta^*}(X_i))|\varepsilon_i]|^2} \\ &\lesssim \sqrt{nt}M\||X_{i,0}|_\infty\|_2 + tn^{1/q}\left[\mathbb{E}\max_j |\mathbb{E}[r(f_{\beta(j)}(X_i)) - r(f_{\beta^*}(X_i))|\varepsilon_i]|^q\right]^{1/q} \\ &\lesssim \sqrt{nt}M\||X_{i,0}|_\infty\|_2 + tn^{1/q}M\||X_{i,0}|_\infty\|_q.\end{aligned}$$

Hence $\mathbb{E}|W_{n,0}|_\infty \lesssim Mx$, which implies that

$$\mathbb{P}(|W_{n,0}|_\infty \geq Mx) \;\leq\; \frac{C_1 n\||X_{i,0}|_\infty\|_q^q}{x^q} + C_2 \exp\left(-\frac{x^2}{n\||X_{i,0}|_\infty\|_2^2}\right). \tag{3.7.6}$$

Finally, for each $1 \leq l \leq L$ and $1 \leq i \leq \lfloor n/\tau_l \rfloor$, define

$$Z_{i,l} = \sum_{k=(i-1)\tau_l+1}^{i\tau_l \wedge n} \mathbb{E}(r(f_\beta(X_i)) - r(f_{\beta^*}(X_i))|\mathcal{F}_k^{k-\tau_l}) - \mathbb{E}(r(f_\beta(X_i)) - r(f_{\beta^*}(X_i))|\mathcal{F}_k^{k-\tau_{l-1}});$$

$$U_{n,l}^e = \sum_{i \text{ is even}} Z_{i,l} \quad \text{and} \quad U_{n,l}^o = \sum_{i \text{ is odd}} Z_{i,l}.$$

Let $c = q/2 - 1 - \alpha q$; let $\lambda_l = l^2/(\pi^2/3)$ if $1 \leq l \leq L/2$ and $\lambda_l = (L+1-l)^{-2}/(\pi^2/3)$ if $L/2 < l \leq L$. Then $\sum_{l=1}^L \lambda_l < 1$. Since $Z_{i,l}$ and $Z_{i',l}$ are independent for $|i - i'| > 1$, by Lemma 3.7.8

$$\mathbb{P}(|U_{n,l}^e|_t - 2\mathbb{E}|U_{n,l}^e|_t \geq \lambda_l Mx) \leq \frac{C_q \sum_{i \text{ is even}} \mathbb{E}|Z_{i,l}|_t^q}{(\lambda_l Mx)^q} + \exp\left(-\frac{(\lambda_l Mx)^2}{3\sum_{i \text{ is even}} |\sigma_{Z_{i,l}}|_t^2}\right),$$

where $\sigma_{Z_{i,l}} = (\|Z_{i1,l}\|_2, ..., \|Z_{im,l}\|_2)^T$.

Similarly, by Lemma 3.7.7, we can obtain

$$\||Z_{i,l}|_t\|_q \le MC_q(\tau_l t)^{1/2}\tilde{\omega}_{l,q}, \quad \text{where } \tilde{\omega}_{l,q} = \sum_{\tau_{l-1}+1}^{\tau_l} \omega_{k,q} \le \tau_{l-1}^{-\alpha}\||X_\cdot|_\infty\|_{q,\alpha}.$$

For $1 \le j \le m$, by Theorem 3.2 of Burkholder [1973] and Jensen's inequality,

$$\begin{aligned}
\|Z_{ij,l}\|_2 &\le \sqrt{\tau_l} \sum_{k=\tau_{l-1}+1}^{\tau_l} \Big\| \mathbb{E}(r(f_{\beta^{(j)}}(X_i)) - r(f_{\beta^*}(X_i))|\mathcal{F}_i^{i-k}) \\
&\quad -\mathbb{E}(r(f_{\beta^{(j)}}(X_i)) - r(f_{\beta^*}(X_i))|\mathcal{F}_i^{i-k+1}) \Big\|_2 \\
&\le \sqrt{\tau_l} \sum_{k=\tau_{l-1}+1}^{\tau_l} \||(\beta^{(j)} - \beta^*)^T(X_{i,k} - X_{i,k-1})|\|_2 \\
&\le M\sqrt{\tau_l} \sum_{k=\tau_{l-1}+1}^{\tau_l} \||X_{i,k} - X_{i,k-1}|_\infty\|_2 \\
&\le M\sqrt{\tau_l} \sum_{k=\tau_{l-1}+1}^{\tau_l} \omega_{k,2} \\
&= M\sqrt{\tau_l}\tilde{\omega}_{l,2}.
\end{aligned}$$

This implies that $|\sigma_{Z_{i,l}}|_t \lesssim M\sqrt{\tau_l}\tau_{l-1}^{-\alpha}\||X_\cdot|_\infty\|_{2,\alpha}$. So we obtain

$$\begin{aligned}
&\mathbb{P}(|U_{n,l}^e|_t - 2\mathbb{E}|U_{n,l}^e|_t \ge \lambda_l Mx) \\
\le\ & \frac{C_q n t^{q/2}\tau_l^{q/2-1}\tilde{\omega}_{l,q}^q}{\lambda_l^q x^q} + \exp\left(-\frac{C_q\lambda_l^2 x^2 \tau_{l-1}^{2\alpha}}{n\||X_\cdot|_\infty\|_{2,\alpha}}\right) \\
\le\ & \frac{C_q n t^{q/2}}{x^q} \cdot \frac{\tau_l^{q/2-1}\tau_{l-1}^{-\alpha q}\||X_\cdot|_\infty\|_{q,\alpha}}{\lambda_l^q} + \exp\left(-\frac{C_q x^2\lambda_l^2\tau_{l-1}^{2\alpha}}{n\||X_\cdot|_\infty\|_{2,\alpha}}\right).
\end{aligned}$$

By Lemma 8 in Chernozhukov et al. [2014],

$$\begin{aligned}
\mathbb{E}|U^e_{n,l}|_t &\lesssim \sqrt{t}\sqrt{\max_j \sum_{i \text{ is even}} \mathbb{E}Z^2_{ij,l}} + t\sqrt{\mathbb{E}\max_i \max_j Z^2_{ij,l}} \\
&\lesssim \sqrt{t}\sqrt{\max_j \sum_{i \text{ is even}} \mathbb{E}Z^2_{ij,l}} + t(n/\tau_l)^{1/q}\||Z_{i,l}|_t\|_q \\
&\lesssim M\sqrt{nt}\tilde{\omega}_{l,2} + Mn^{1/q}t^{3/2}\tau_l^{1/2-1/q}\tilde{\omega}_{l,q} \\
&\lesssim M\sqrt{nt}\tau_l^{-\alpha}\||X_\cdot|_\infty\|_{2,\alpha} + Mn^{1/q}t^{3/2}\tau_l^{-c/q}\||X_\cdot|_\infty\|_{q,\alpha}.
\end{aligned}$$

Notice that $\lambda_l^{-1}\tau_l^{c/q} \lesssim n^{c/q}$ for $c > 0$ and $\min_{c\geq 0} \lambda_l \tau_l^{-c/q} > 1$ for $c < 0$ and $\min_{l\geq 0} \lambda_l \tau_l^{\alpha} > 1$. Hence $\mathbb{E}|U^e_{n,l}|_t \lesssim M\lambda_l x$ always holds. Therefore,

$$\mathbb{P}(|U^e_{n,l}|_t \geq \lambda_l M x) \leq \frac{C_3 n t^{q/2}}{x^q} \cdot \frac{\tau_l^{q/2-1}\tau_{l-1}^{-\alpha q}\||X_\cdot|_\infty\|_{q,\alpha}}{\lambda_l^q} + \exp\left(-\frac{C_4 x^2 \lambda_l^2 \tau_{l-1}^{2\alpha}}{n\||X_\cdot|_\infty\|_{2,\alpha}}\right). \quad (3.7.7)$$

A similar inequality holds for $U^o_{n,l}$. Let

$$A = \sum_{l=1}^{L} \frac{\tau_l^c}{\lambda_l^q} \quad \text{and} \quad B = \sum_{l=1}^{L} \exp\left\{-\frac{C_5 x^2 \lambda_l^2 \tau_l^{2\alpha}}{n\||X_\cdot|_\infty\|^2_{2,\alpha}}\right\}.$$

Since $\sum_{l=1}^{L} \lambda_l \leq 1$ and $|Q_{n,l}|_t \leq |U^e_{n,l}|_t + |U^o_{n,l}|_t$, by (3.7.7),

$$\begin{aligned}
\mathbb{P}(|\sum_{l=1}^{L} Q_{n,l}|_t \geq 2Mx) &\leq \sum_{l=1}^{L} \mathbb{P}(|Q_{n,l}|_t \geq 2\lambda_l M x) \\
&\leq \sum_{l=1}^{L} [\mathbb{P}(|U^e_{n,l}|_t \geq \lambda_l M x) + \mathbb{P}(|U^o_{n,l}|_t \geq \lambda_l M x)] \\
&\leq \frac{C_6 n t^{q/2}\||X_\cdot|_\infty\|^q_{q,\alpha}}{x^q} A + C_7 B. \quad (3.7.8)
\end{aligned}$$

Let $\psi := \min_{l\geq 1} \lambda_l^2 \tau_l^{2\alpha} > 0$. By the definition of τ_l and λ_l and by elementary calculations,

there exists a constant $C_8 > 1$ such that for all $y \geq 1$,

$$\sum_{l=1}^{L} \exp\left\{-C_5 y \lambda_l^2 \tau_l^{2\alpha}\right\} \leq C_8 \exp\{-C_5 y \psi\}. \tag{3.7.9}$$

We apply (3.7.9) with $y = x^2/(n|||X.|_\infty\|_{2,\alpha}^2)$. If $c > 0$, it can be obtained that $A \leq C_9 \lambda_l^c \leq C_9 n^c$. If $c < 0$, then $A \leq C_{10}$. Hence, combining (3.7.4), (3.7.5), (3.7.6), (3.7.8) and (3.7.9), if $c > 0$ and $x \gtrsim \sqrt{nt}|||X.|_\infty\|_{2,\alpha} + n^{1/q+c/q} t^{3/2} |||X.|_\infty\|_{q,\alpha}$,

$$\mathbb{P}(|W_n|_t \geq Mx) \leq \exp\left\{-\frac{C_{q,\alpha} x^2}{n|||X.|_\infty\|_{2,\alpha}^2}\right\} + \frac{C_{q,\alpha} n^{c+1} t^{q/2} |||X.|_\infty\|_{q,\alpha}^q}{x^q}, \tag{3.7.10}$$

if $c < 0$ and $x \gtrsim \sqrt{nt}|||X.|_\infty\|_{2,\alpha} + n^{1/q} t^{3/2} |||X.|_\infty\|_{q,\alpha}$,

$$\mathbb{P}(|W_n|_t \geq Mx) \leq \exp\left\{-\frac{C_{q,\alpha} x^2}{n|||X.|_\infty\|_{2,\alpha}^2}\right\} + \frac{C_{q,\alpha} n t^{q/2} |||X.|_\infty\|_{q,\alpha}^q}{x^q}. \tag{3.7.11}$$

By (3.7.3), (3.7.10) and (3.7.11), both cases with $c < 0$ and $c > 0$ of Theorem 3.7.1 follow.

□

Theorem 3.7.2. *Assume $|||X.|_\infty\|_{\psi_\nu} < \infty$, where $\nu \geq 0$. Let $|\beta - \beta^*|_0 = s$, $t = s(\log p + \log n)$ and $\alpha = 2/(1+2\nu)$, then under Assumptions 3.3.2 and 3.3.3, there exists a constant $C_\nu > 0$ depending on ν such that*

$$\mathbb{P}\left(\sup_{\substack{|\beta-\beta^*|_1 \leq M, \\ |\beta-\beta^*|_0 = s}} \left|\sum_{i=1}^{n} \bar{r}(f_\beta(X_i)) - \bar{r}(f_{\beta^*}(X_i))\right| \geq Mx\right) \leq C_\nu \exp\left(-\frac{x^\alpha}{2e\alpha(nt)^{\alpha/2} |||X.|_\infty\|_{\psi_\nu}^\alpha}\right). \tag{3.7.12}$$

Proof. We shall use the argument in the proof of Theorem 3.7.1. Denote $\mathcal{A} := \{\beta : |\beta - \beta^*|_1 \leq M, |\beta - \beta^*|_0 = s\}$. Define set $\mathcal{A}_m := \{\beta^{(1)}, ..., \beta^{(m)}\} \subset \mathcal{A}$, such that for any $\beta \in \mathcal{A}, \min_{\beta^{(j)} \in \mathcal{A}_m} |\beta - \beta^{(j)}|_1 \leq M/D$. Then $m = |\mathcal{A}_m| \asymp p^s D^s$. Let $\beta^h := \arg\min_{\beta^{(j)} \in \mathcal{A}_m} |\beta -$

$\beta^{(j)}|_1$. Similar to the proof of Theorem 3.7.1, we can obtain

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\substack{|\beta-\beta^*|_1\le M,\\ |\beta-\beta^*|_0=s}}\left|\sum_{i=1}^n[\bar r(f_\beta(X_i))-\bar r(f_{\beta^*}(X_i))]\right|\ge 2Mx\right)\\
&\le \mathbb{P}\left(\sup_{\beta^{(j)}\in\mathcal{A}_m}\left|\sum_{i=1}^n[\bar r(f_{\beta^{(j)}}(X_i))-\bar r(f_{\beta^*}(X_i))]\right|\ge Mx\right)\\
&\quad+\mathbb{P}\left(\sup_{\substack{|\beta-\beta^*|_1\le M,\\ |\beta-\beta^*|_0=s}}\left|\sum_{i=1}^n[\bar r(f_\beta(X_i))-\bar r(f_{\beta^h}(X_i))]\right|\ge Mx\right)\\
&:= \mathrm{I}_1+\mathrm{I}_2.
\end{aligned}$$

Let $u_0=(e\alpha\||X.|_\infty\|_{\psi_\nu}^\alpha)^{-1}$ and $t=1\vee\log m$. Note that for any vector $u=(u_1,...,u_m)^T$, $|u|_\infty\le|u|_t\le m^{1/t}|u|_\infty$. Let $L=\lfloor\log n/\log 2\rfloor$, $\tau_l=2^l$ if $1\le l\le L$, $\tau_L=n$ and $\tau_0=0$. Denote $r(f_\beta(X_i))=(r(f_{\beta^{(1)}}(X_i)),...,r(f_{\beta^{(m)}}(X_i)))$, $r(f_{\beta^*}(X_i))=(r(f_{\beta^*}(X_i)),...,r(f_{\beta^*}(X_i)))$. Let $W_n=\sum_{i=1}^n[\bar r(f_\beta(X_i))-\bar r(f_{\beta^*}(X_i))]$,
$W_{n,k}=\sum_{i=1}^n\mathbb{E}([\bar r(f_\beta(X_i))-\bar r(f_{\beta^*}(X_i))]|\varepsilon_{i-k},...,\varepsilon_i)$, $X_{i,k}=\mathbb{E}(X_i|\varepsilon_{i-k},...,\varepsilon_i)$. Define $Q_{n,l}=\sum_{i=1}^n\mathcal{P}_{i-l}[\bar r(f_\beta(X_i))-\bar r(f_{\beta^*}(X_i))]$. Then, $Q_{n,l}$ is a martingale. By Assumption 3.3.3, Lemma 3.7.7 and Jensen's inequality,

$$\begin{aligned}
&\||Q_{n,l}|_\infty\|_q\\
&\le C\sqrt{qnt}\left\|\left|\mathbb{E}(r(f_\beta(X_i))-r(f_{\beta^*}(X_i))|\mathcal{F}_{i-l})-\mathbb{E}(r(f_\beta(X_i))-r(f_{\beta^*}(X_i))|\mathcal{F}_{i-l-1})\right|_t\right\|_q\\
&\le C\sqrt{qnt}\left\|\left|r(f_\beta(X_i))-r(f_{\beta^*}(X_i))-r(f_\beta(X_{i,i-l}))+r(f_{\beta^*}(X_{i,i-l}))\right|_t\right\|_q\\
&\le C\sqrt{qnt}M\left\|\left|X_i-X_{i,i-l}\right|_t\right\|_q\\
&= C\sqrt{qnt}M\omega_{l,q}.
\end{aligned}$$

Then

$$\||W_n|_t\|_q \le \sum_{l=0}^{\infty} \||Q_{n,l}|_t\|_q \le M \sum_{l=0}^{\infty} C\sqrt{qnt}\omega_{l,q} = CM\sqrt{qnt}\Omega_{0,q}.$$

Let $Z_n = W_n/(M\sqrt{nt})$. Then $\||Z_n|_t\|_q \le C\sqrt{q}\Omega_{0,q}$. Write the negative binomial expansion $(1-s)^{-1/2} = 1+\sum_{k=1}^{\infty} a_k s^k$, where $|s|<1$ and $a_k = (2k)!/(2^{2k}(k!)^2)$. By Stirling's formula, as $k\to\infty$, $a_k \sim (k\pi)^{-1/2}$. Hence $k! \sim \sqrt{2}(k/e)^k a_k^{-1}$, and there exists absolute constants $c_1, c_2 > 0$ such that $c_1(k/e)^k a_k^{-1} \le k! \le c_2(k/e)^k a_k^{-1}$ holds for all $k \ge 1$. If $\alpha k > 2$, we have $\Omega_{0,\alpha k} \le (\alpha k)^{\nu}\||X.|_\infty\|_{\psi_\nu}$. Hence, by elementary manipulations,

$$\frac{u^k \||Z_n|_t^\alpha\|_k^k}{k!} \le \frac{u^k(\alpha k)^{\alpha k/2}\Omega_{0,\alpha k}^{\alpha k}}{c_1(k/e)^k a_k^{-1}} \le \frac{a_k u^k}{c_1 u_0^k} \tag{3.7.13}$$

If $\alpha k < 2$, then $\||Z_n|_t\|_{\alpha k} \le \||Z_n|_t\|_2 \le 2^{\nu}\||X.|_\infty\|_{\psi_\nu}$. Using $e^x = \sum_{k=0}^{\infty} x^k/k!$, we obtain,

$$\begin{aligned}
\sup_n \mathbb{E}\exp\{u|Z_n|_t^\alpha\} &\le 1 + \sum_{1\le k<2/\alpha} \frac{u^k(2^\nu\||X.|_\infty\|_{\psi_\nu})^{\alpha k}}{k!} + \sum_{k\ge 2/\alpha} \frac{a_k u^k}{c_1 u_0^k} \\
&\le 1 + c'_\alpha \sum_{k=1}^{\infty} a_k \frac{u^k}{u_0^k} \\
&\le 1 + c_\alpha \frac{u/u_0}{(1-u/u_0)^{1/2}},
\end{aligned}$$

where constants $c_\alpha, c'_\alpha > 0$ only depend on α. Let $u = u_0/2$, then $\sup_n \mathbb{E}\exp\{u|Z_n|_t^\alpha\} \le 1 + c_\alpha/\sqrt{2}$. Hence,

$$\mathrm{I}_1 = \mathbb{P}(|W_n|_t \ge Mx) = \mathbb{P}(|Z_n|_t \ge \frac{x}{\sqrt{nt}}) \le (1+c_\alpha/\sqrt{2})\exp\left\{-\frac{x^\alpha}{2e\alpha(nt)^{\alpha/2}\||X.|_\infty\|_{\psi_\nu}^\alpha}\right\}. \tag{3.7.14}$$

Basic calculation shows that $\|(\beta-\beta^h)^T X_i\|_q \le M\Omega_{0,q}/D$. Employing similar arguments,

we can obtain

$$\mathrm{I}_2 \;\leq\; (1+c_\alpha/\sqrt{2})n\exp\left\{-\frac{x^\alpha D^\alpha}{2e\alpha n^\alpha \||X.|_\infty\|_{\psi_\nu}^\alpha}\right\}. \tag{3.7.15}$$

Note $\log m \asymp s\log p + s\log D$. We can set $\log m = s(\log p + \log n)$ and $D \asymp n$. Thus,

$$\mathrm{I}_2 \;\leq\; (1+c_\alpha/\sqrt{2})\exp\left\{-\frac{x^\alpha}{2e\alpha \||X.|_\infty\|_{\psi_\nu}^\alpha} + \log n\right\}. \tag{3.7.16}$$

Clearly, Theorem 3.7.2 follows from (3.7.14) and (3.7.16).

□

Theorem 3.7.3. *Assume $\Omega_{0,q} < \infty$. Denote $\widetilde{X}_{ij} = sgn(X_{ij})(|X_{ij}| \wedge \tau)$ and $\widetilde{X}_i = (\widetilde{X}_{i1}, ..., \widetilde{X}_{ip})^T$. Let $|\beta - \beta^*|_0 = s$, $t = s(\log p + \log n)$, then under Assumptions 3.3.2 and 3.3.3,*

$$\mathbb{P}\left(\sup_{\substack{|\beta-\beta^*|_1\leq M,\\ |\beta-\beta^*|_0=s}}\left|\sum_{i=1}^n \bar{r}(f_\beta(\widetilde{X}_i)) - \bar{r}(f_{\beta^*}(\widetilde{X}_i))\right| \geq Mx\right) \leq C_q\exp\left(-\frac{x^2}{4ent\tau^2}\right). \tag{3.7.17}$$

Proof. It can be carried out following the same routes as those in the proof of Theorem 3.7.2 by setting $\alpha = 2$ therein. □

Lemma 3.7.4. *Assume $\||X.|_\infty\|_{q,\alpha} < \infty$, where $q > 2$ and $\alpha > 0$, $C_{q,\alpha}$ is a constant depending on q and α. (i) If $\alpha > 1/2 - 1/q$, then for $x \geq \sqrt{n\log p}\,\||X.|_\infty\|_{2,\alpha}$ $+n^{1/q}(\log p)^{3/2}\,\||X.|_\infty\|_{q,\alpha}$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right|_\infty \geq x\right) \leq \frac{C_{q,\alpha}n(\log p)^{q/2}\,\||X.|_\infty\|_{q,\alpha}^q}{x^q} + C_{q,\alpha}\exp\left(-\frac{C_{q,\alpha}x^2}{n\,\||X.|_\infty\|_{2,\alpha}^2}\right). \tag{3.7.18}$$

(ii) If $0 < \alpha < 1/2 - 1/q$, *then for* $x \geq \sqrt{n \log p}\, \| |X_\cdot|_\infty \|_{2,\alpha} + n^{1/2-\alpha} (\log p)^{3/2} \| |X_\cdot|_\infty \|_{q,\alpha}$,

$$\mathbb{P}\left(\left| \sum_{i=1}^n X_i \right|_\infty \geq x \right) \leq \frac{C_{q,\alpha} n^{q/2-\alpha q} (\log p)^{q/2} \| |X_\cdot|_\infty \|_{q,\alpha}^q}{x^q} + C_{q,\alpha} \exp\left(- \frac{C_{q,\alpha} x^2}{n \| |X_\cdot|_\infty \|_{2,\alpha}^2} \right). \tag{3.7.19}$$

Proof. Use the same argument as that in Theorem 6.2 of Zhang and Wu [2017]. □

Lemma 3.7.5. *Assume* $\| |X_\cdot|_\infty \|_{\psi_\nu} < \infty$, *where* $\nu \geq 0$. *Let* $\alpha = 2/(1+2\nu)$, *then there exists a constant* $C_\nu > 0$ *depending on* ν *such that*

$$\mathbb{P}\left(\left| \sum_{i=1}^n X_i \right|_\infty \geq x \right) \quad \leq \quad C_\nu \exp\left(- \frac{x^\alpha}{2e\alpha (n \log p)^{\alpha/2} \| |X_\cdot|_\infty \|_{\psi_\nu}^\alpha} \right). \tag{3.7.20}$$

Proof. The proof is similar to that of Theorem 3.7.2, and is omitted. □

Theorem 3.7.6. *Let* (X_i) *be a stationary process in the form* $X_i = g(..., \varepsilon_{i-1}, \varepsilon_i)$, *where* ε_i, $i \in \mathbb{Z}$, *be i.i.d. random variables. Assume that* (X_i) *is geometric moment contraction (see Assumption 3.4.1), that is, (3.4.5) holds with q-moment and constant* $C_2 > 0$, $0 < a_2 < 1$. *If* $\mathbb{E}X_i = 0$ *and* $|X_i| \leq M$, *then we have,*

$$\mathbb{P}\left(\left| \sum_{i=1}^n X_i \right| \geq x \right) \leq \exp\left(- \frac{x^2}{16 n\nu + c_{q,1} M \log(n/x) + c_{q,2} \sqrt{M} x^{3/2}} \right), \tag{3.7.21}$$

where $\nu = 2C_2^2 (1-a_2)^{-1} (1-a_2^2)^{-1}$ *and* $c_{q,1}, c_{q,2}$ *are constants depending on* q *and* a_2.

Proof. Denote $X_{i,m} = \mathbb{E}(X_i | \varepsilon_{i-m}, ..., \varepsilon_i)$ and $X_{i,\{l\}} = g(..., \varepsilon_{l-1}, \varepsilon_l', \varepsilon_{l+1}, ..., \varepsilon_i)$. Then,

$$\mathbb{P}\left(\left| \sum_{i=1}^n X_i \right| \geq x \right) \quad \leq \quad \mathbb{P}\left(\left| \sum_{i=1}^n (X_i - X_{i,m}) \right| \geq \frac{x}{2} \right) + \mathbb{P}\left(\left| \sum_{i=1}^n X_{i,m} \right| \geq \frac{x}{2} \right) := \mathrm{I}_1 + \mathrm{I}_2.$$

We first consider I_1. By Markov inequality,

$$\mathrm{I}_1 \leq \mathbb{P}\left(\max_i \left| (X_i - X_{i,m}) \right| \geq \frac{x}{2n} \right) \leq 2^{-q} n^{q+1} x^{-q} \mathbb{E} |X_i - X_{i,m}|^q.$$

By GMC (3.4.5), we obtain

$$\begin{aligned}\|X_i - X_{i,m}\|_q &\leq \sum_{l=m}^{\infty} \|X_{i,l+1} - X_{i,l}\|_q \leq \sum_{l=m}^{\infty} \|X_{l+1} - X_{l+1,\{0\}}\|_q \leq \sum_{l=m}^{\infty} C_2 a_2^{l+1} \\ &\leq C_2(1-a_2)^{-1} a_2^{m+1}.\end{aligned}$$

Thus,

$$\mathrm{I}_1 \leq C_2^q (1-a_2)^{-q} 2^{-q} n^{q+1} x^{-q} a_2^{q(m+1)} \leq c_1 e^{-m(q\log(a_2^{-1}))-q\log x+(q+1)\log n} \quad (3.7.22)$$

It remains to bound I_2. Let $L = \lfloor n/m \rfloor$. Define $Z_{l,m} = \sum_{i=m(l-1)+1}^{(ml)\wedge n} X_{i,m}$ for $1 \leq l \leq L$. Then, we have

$$\mathrm{I}_2 \leq \mathbb{P}\left(\left|\sum_{l \text{ is odd}} Z_{l,m}\right| \geq \frac{x}{4}\right) + \mathbb{P}\left(\left|\sum_{l \text{ is even}} Z_{l,m}\right| \geq \frac{x}{4}\right) := \mathrm{I}_{2,1} + \mathrm{I}_{2,2}.$$

Recall $\mathcal{P}_i(\cdot) = \mathbb{E}(\cdot|\mathcal{F}_i) - \mathbb{E}(\cdot|\mathcal{F}_{i-1})$. Some basic calculation shows that

$$\begin{aligned}|\mathrm{Cov}(X_{i,m}, X_{i+k,m})| &= |\sum_{h=0}^{\infty} \mathbb{E}(\mathcal{P}_{-h}X_{0,m})(\mathcal{P}_{-h}X_{k,m})| \leq \sum_{h=0}^{\infty} |\mathbb{E}(\mathcal{P}_{-h}X_{0,m})(\mathcal{P}_{-h}X_{k,m})| \\ &\leq \sum_{h=0}^{\infty} \|X_h - X_{h,\{0\}}\|_2 \|X_{h+k} - X_{h+k,\{0\}}\|_2 \leq C_2^2 \sum_{h=0}^{\infty} a_2^{2h+k}.\end{aligned}$$

Thus,

$$\mathbb{E}Z_{1,m}^2 = \mathbb{E}|\sum_{i=1}^{m} X_{i,m}|^2 \leq mC_2^2(\sum_{h=0}^{\infty} a_2^{2h} + 2\sum_{k=1}^{m}\sum_{h=0}^{\infty} a_2^{2h+k}) \leq 2mC_2^2(\sum_{k=0}^{\infty}\sum_{h=0}^{\infty} a_2^{2h+k}) \leq m\nu,$$

where $\nu = 2C_2^2(1-a_2)^{-1}(1-a_2^2)^{-1}$. Note that $Z_{l,m} \leq mM$. Since $Z_{l,m}$ and $Z_{l',m}$ are independent for $|l-l'| > 1$, by Bernstein inequality, we obtain

$$\mathrm{I}_{2,1} \leq \exp\left(-\frac{(x/4)^2}{(x/4)\cdot mM + n/m \cdot m\nu}\right) = \exp\left(-\frac{x^2}{4xmM + 16n\nu}\right). \quad (3.7.23)$$

Taking $m = \sqrt{x/M} + (q+1)\log(n/x)$, we have

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| \geq x\right) \quad \leq \quad \exp\left(-c_2\sqrt{\frac{x}{M}} - c_3\frac{x}{M\log(n/x)} - \frac{x^2}{16n\nu}\right),$$

for some constants c_2, c_3 depending on q, a_2. Then (3.7.21) follows. □

Remark 3.7.1. *Under the geometric moment contraction, our Theorem 3.7.6 is sharper than Theorem 3 in Wu and Wu [2016].*

Lemma 3.7.7. *Let D_i, $1 \leq i \leq n$, be p-dimensional martingale difference vectors with respect to the σ-field $\mathcal{G}_i$. Let $s > 1$ and $q \geq 2$. Then*

$$\||D_1 + \ldots + D_n|_s\|_q \leq c\left\{q\|\sup_i |D_i|_s\|_q + \sqrt{q(s-1)}\left\|\left[\sum_{i=1}^{n}\mathbb{E}(|D_i|_s^2|\mathcal{G}_{i-1})\right]^{1/2}\right\|_q\right\},$$

where c is an absolute constant.

Lemma 3.7.7 provides a Rosenthal-Burkholder type bound on moments of Banach-spaced martingales and follows from Theorem 4.1 of Pinelis [1994].

Lemma 3.7.8. *Assume $s > 1$. Let $X_1, \ldots, X_n$ be p-dimensional independent random vectors with mean zero such that for some $q > 2$, $\||X_i|_s\|_q < \infty$, $1 \leq i \leq n$. Let $T_n = \sum_{i=1}^n X_i$ and $\sigma_i = (\|X_{i1}\|_2, \ldots, \|X_{ip}\|_2)^\top$. Then for any $y > 0$,*

$$\mathbb{P}\left(|T_n|_s \geq 2\mathbb{E}|T_n|_s + y\right) \leq C_q y^{-q}\sum_{i=1}^{n}\mathbb{E}|X_i|_s^q + \exp\left(-\frac{y^2}{3\sum_{i=1}^n |\sigma_i|_s^2}\right), \tag{3.7.24}$$

where C_q is a positive constant depending only on q.

Proof. See Lemma C.6 in Zhang and Wu [2017]. □

3.8 Deferred Proofs

Throughout this section, without loss of generality, define

$$V_n(\beta) = \sum_{i=1}^{n} [R(f_\beta(X_i), Y_i) - \mathbb{E}R(f_\beta(X_i), Y_i)].$$

Set $M := \mathcal{E}^*/\lambda_0$. With the newly defined β^*, let

$$T_n := \sup_{|\beta - \beta^*|_1 \leq M} |V_n(\beta) - V_n(\beta^*)|. \tag{3.8.1}$$

Proof of Theorem 3.3.1. We use the short hand notation $S_* = S_{\beta^*}$, $s_* = s_{\beta^*}$ and $\kappa_* = \kappa(S_{\beta^*})$. The proof is along the lines of Theorem 6.4 in Bühlmann and Van De Geer [2011]. Let

$$t := \frac{M}{M + |\hat{\beta} - \beta^*|_1},$$

and

$$\tilde{\beta} := t\hat{\beta} + (1-t)\beta^*.$$

Then

$$|\tilde{\beta} - \beta^*|_1 = \frac{M|\hat{\beta} - \beta^*|_1}{M + |\hat{\beta} - \beta^*|_1} \leq M.$$

So if we show that $|\tilde{\beta} - \beta^*|_1 \leq M/2$, then $|\hat{\beta} - \beta^*|_1 \leq M$.

Since $\hat{\beta}$ minimizes (3.2.1), by the assumed convexity of $z \mapsto R(z, y)$, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} R(f_{\tilde{\beta}}(X_i), Y_i) + \lambda|\tilde{\beta}|_1 \\
&\leq t\left\{\frac{1}{n}\sum_{i=1}^{n} R(f_{\hat{\beta}}(X_i), Y_i) + \lambda|\hat{\beta}|_1\right\} + (1-t)\left\{\frac{1}{n}\sum_{i=1}^{n} R(f_{\beta^*}(X_i), Y_i) + \lambda|\beta^*|_1\right\} \\
&\leq \frac{1}{n}\sum_{i=1}^{n} R(f_{\beta^*}(X_i), Y_i) + \lambda|\beta^*|_1.
\end{aligned} \tag{3.8.2}$$

On the event

$$\mathcal{T} := \{T_n \leq n\lambda_0 M\} = \{T_n \leq n\mathcal{E}^*\}, \tag{3.8.3}$$

inequality (3.8.2) implies that

$$\mathcal{E}(f_{\tilde{\beta}}) + \lambda|\tilde{\beta}|_1 \leq -\frac{1}{n}\left[V_n(\tilde{\beta}) - V_n(\beta^*)\right] + \mathcal{E}(f_{\beta^*}) + \lambda|\beta^*|_1 \leq \lambda_0 M + \mathcal{E}(f_{\beta^*}) + \lambda|\beta^*|_1.$$

Now, for any β,

$$\beta = \beta_{S_*} + \beta_{S_*^c}.$$

Note that $\beta^*_{S_*} = \beta^*$ and $\beta^*_{S_*^c} = 0$. So, we have,

$$\mathcal{E}(f_{\tilde{\beta}}) + \lambda|\tilde{\beta}_{S_*^c}|_1 \leq \mathcal{E}^* + \mathcal{E}(f_{\beta^*}) + \lambda|\tilde{\beta}_{S_*} - \beta^*_{S_*}|_1 \leq \frac{5}{3}\mathcal{E}^* + \lambda|\tilde{\beta}_{S_*} - \beta^*_{S_*}|_1.$$

Case a) If $\lambda|\tilde{\beta} - \beta^*|_1 > 10\mathcal{E}^*/3$, we find

$$\mathcal{E}(f_{\tilde{\beta}}) + \frac{1}{2}\lambda|\tilde{\beta}_{S_*^c}|_1 \leq \frac{3}{2}\lambda|\tilde{\beta}_{S_*} - \beta^*_{S_*}|_1.$$

It follows that $|\tilde{\beta}_{S_*^c}|_1 \leq 3|\tilde{\beta}_{S_*} - \beta^*_{S_*}|_1$. This means that we can apply the compatibility condition. Thus, invoking the margin condition,

$$\begin{aligned}
\mathcal{E}(f_{\tilde{\beta}}) + \lambda|\tilde{\beta} - \beta^*|_1 &\leq \mathcal{E}^* + \mathcal{E}(f_{\beta^*}) + 2\lambda|\tilde{\beta}_{S_*} - \beta^*_{S_*}|_1 \\
&\leq \mathcal{E}^* + \mathcal{E}(f_{\beta^*}) + 2\lambda\sqrt{s_*}\|f_{\tilde{\beta}} - f_{\beta^*}\|/\kappa_* \\
&\leq \mathcal{E}^* + \mathcal{E}(f_{\beta^*}) + G\left(\frac{4\lambda\sqrt{s_*}}{\kappa_*}\right) + \frac{1}{2}\mathcal{E}(f_{\tilde{\beta}}) + \frac{1}{2}\mathcal{E}(f_{\beta^*})
\end{aligned}$$

It follows that

$$\frac{1}{2}\mathcal{E}(f_{\tilde{\beta}}) + \lambda|\tilde{\beta} - \beta^*|_1 \leq 2\mathcal{E}^*.$$

But then $\lambda|\tilde{\beta} - \beta^*|_1 \leq 2\mathcal{E}^*$, which is a contradiction.

Case b) If $\lambda|\tilde{\beta} - \beta^*|_1 \leq 10\mathcal{E}^*/3$, this implies

$$|\tilde{\beta} - \beta^*|_1 \leq \frac{10\lambda_0 M}{3\lambda} \leq \frac{1}{2}M,$$

and

$$\mathcal{E}(f_{\tilde{\beta}}) + \lambda|\tilde{\beta}_{S_*^c}|_1 \leq 5\mathcal{E}^*.$$

We immediately have $|\hat{\beta} - \beta^*|_1 \leq M$.

Hence, in both Case a) and Case b), the conclusion is that $|\hat{\beta} - \beta^*|_1 \leq M$. We can now use the same argument with $\tilde{\beta}$ replaced by $\hat{\beta}$ to establish that in fact on the event $\mathcal{T}$,

$$\lambda|\hat{\beta} - \beta^*|_1 \leq \frac{10}{3}\mathcal{E}^*.$$

Next, we repeat the above argument in Case a) $\lambda|\hat{\beta}_{S_*} - \beta^*_{S_*}|_1 > \mathcal{E}^*$ and Case b) $\lambda|\hat{\beta}_{S_*} - \beta^*_{S_*}|_1 \leq \mathcal{E}^*$. This gives

$$\mathcal{E}(f_{\hat{\beta}}) \leq \frac{8}{3}\mathcal{E}^*.$$

So Theorem 3.3.1 follows if we can control the probability $\mathbb{P}(\mathcal{T})$.

We now prove (i). By Hölder's inequality, we have for $m \geq 0$ that

$$\begin{aligned}
& \sum_{l=m}^{\infty} \| \max_j |X_{lj} Y_l - X_{lj,\{0\}} Y_{l,\{0\}}| \|_\chi \\
\leq & \sum_{l=m}^{\infty} \left(\| \max_j |X_{lj}(Y_l - Y_{l,\{0\}})| \|_\chi + \| \max_j |(X_{lj} - X_{lj,\{0\}}) Y_{l,\{0\}}| \|_\chi \right) \\
= & \sum_{l=m}^{\infty} \left(\| \max_j |X_{lj}| \|_\gamma \|Y_l - Y_{l,\{0\}}\|_q + \| \max_j |X_{lj} - X_{lj,\{0\}}| \|_\gamma \|Y_{l,\{0\}}\|_q \right).
\end{aligned}$$

Since $\alpha = \min(\alpha_X, \alpha_Y)$, the dependence adjusted norm satisfies

$$\begin{aligned} \|\max_j |X_{\cdot j} Y_\cdot|\|_{\tau,\alpha} &\leq \|\max_j |X_{\cdot j}|\|_{\gamma,0} \|Y_\cdot\|_{q,\alpha_Y} + \|\max_j |X_{\cdot j}|\|_{\gamma,\alpha_X} \|Y_\cdot\|_{q,0} \\ &\leq 2\||X_\cdot|_\infty\|_{\gamma,\alpha_X} \|Y_\cdot\|_{q,\alpha_Y}. \end{aligned} \tag{3.8.4}$$

For constant $s_0 < p$, create a partition of $\{1, 2, ..., p\}$ as

$$I_1 = \{1, 2, ..., s_0\}, I_2 = \{s_0 + 1, s_0 + 2, ..., 2s_0\}, ...$$

where I_l, $l = 1, 2, ...,$ has cardinality s_0, except the last set which may have cardinality smaller than s_0. Write $\delta_0 = \beta^*$, $\delta_{\lceil p/s_0 \rceil} = \beta$ and $\delta_l = \beta^* + \sum_{j=1}^{l} (\beta - \beta^*)_{I_j} = (\beta_1, ..., \beta_{ls_0}, \beta^*_{ls_0+1}, ..., \beta^*_p)$ for all $1 \leq l \leq \lceil p/s_0 \rceil$. Then this partition leads to the following decomposition

$$\frac{1}{n} \sum_{i=1}^{n} [\bar{r}(f_\beta(X_i)) - \bar{r}(f_{\beta^*}(X_i))] = \sum_{l=1}^{\lceil p/s_0 \rceil} \frac{1}{n} \sum_{i=1}^{n} [\bar{r}(f_{\delta_l}(X_i)) - \bar{r}(f_{\delta_{l-1}}(X_i))], \tag{3.8.5}$$

$$\frac{1}{n} \sum_{i=1}^{n} [\bar{h}(f_\beta(X_i))Y_i - \bar{h}(f_{\beta^*}(X_i))Y_i] = \sum_{l=1}^{\lceil p/s_0 \rceil} \frac{1}{n} \sum_{i=1}^{n} [\bar{h}(f_{\delta_l}(X_i))Y_i - \bar{h}(f_{\delta_{l-1}}(X_i))Y_i]. \tag{3.8.6}$$

Denote

$$\mathcal{T}_{l1} = \left\{ \sup_{|\beta - \beta^*|_1 \leq M} \left| \frac{1}{n} \sum_{i=1}^{n} [\bar{r}(f_{\delta_l}(X_i)) - \bar{r}(f_{\delta_{l-1}}(X_i))] \right| \leq \frac{1}{2} \lambda_0 |\delta_l - \delta_{l-1}|_1 \right\}, \tag{3.8.7}$$

$$\mathcal{T}_{l2} = \left\{ \sup_{|\beta - \beta^*|_1 \leq M} \left| \frac{1}{n} \sum_{i=1}^{n} [\bar{h}(f_{\delta_l}(X_i))Y_i - \bar{h}(f_{\delta_{l-1}}(X_i))Y_i] \right| \leq \frac{1}{2} \lambda_0 |\delta_l - \delta_{l-1}|_1 \right\}. \tag{3.8.8}$$

For $\lambda_0 \gtrsim \sqrt{s_0 \log(pn)/n} \||X_\cdot|_\infty\|_{\gamma,\alpha_X} \|Y_\cdot\|_{q,\alpha_Y} + n^{\rho/\chi - 1} (s_0 \log(pn))^{3/2} \||X_\cdot|_\infty\|_{\gamma,\alpha_X} \|Y_\cdot\|_{q,\alpha_Y}$,

adopting (3.8.4) and Theorem 3.7.1, we have,

$$\begin{aligned}\mathbb{P}\left(\mathcal{T}_{l2}^{c}\right) &\leq \frac{C_4 n(s_0\log(pn))^{\chi/2}\||X_\cdot|_\infty\|_{\gamma,\alpha_X}^{\chi}\|Y_\cdot\|_{q,\alpha_Y}^{\chi}}{(n\lambda_0)^{\chi}} + C_5\exp\left(-\frac{C_6(n\lambda_0)^2}{n\||X_\cdot|_\infty\|_{\gamma,\alpha_X}^2\|Y_\cdot\|_{q,\alpha_Y}^2}\right)\\ &\leq C_4(s_0\log(pn))^{-\chi} + C_5(pn)^{-C_6 s_0}. \qquad (3.8.9)\end{aligned}$$

Similarly, applying Theorem 3.7.1, for

$$\lambda_0 \gtrsim \sqrt{s_0\log(pn)/n}\,\||X_\cdot|_\infty\|_{2,\alpha_X} + n^{\nu/\gamma-1}(s_0\log(pn))^{3/2}\,\||X_\cdot|_\infty\|_{\gamma,\alpha_X},$$

$$\mathbb{P}\left(\mathcal{T}_{l1}^{c}\right) \;\leq\; C_1(s_0\log(pn))^{-\gamma} + C_2(pn)^{-C_3 s_0}.$$

Hence, by (3.8.5) and (3.8.6), setting $s_0 = p^{1/(\gamma\wedge\chi+1)}$, $T_n \leq n\lambda_0 M$ holds with probability at least,

$$\begin{aligned}1 - (\lceil p/s_0\rceil)(C_1(s_0\log(pn))^{-\gamma} + C_2(pn)^{-C_3 s_0} + C_4(s_0\log(pn))^{-\chi} + C_5(pn)^{-C_6 s_0})\\ \geq 1 - C_1(\log(pn))^{-(\gamma\wedge\chi)}.\end{aligned}$$

Next, we prove (ii). Let $\gamma = \tau(1+\iota/\nu)$ and $q = \tau(1+\nu/\iota)$. Then

$$\begin{aligned}&\sum_{l=0}^{\infty}\|\max_j|X_{lj}Y_l - X_{lj,\{0\}}Y_{l,\{0\}}|\|_\tau\\ \leq\;& \sum_{l=0}^{\infty}\left(\|\max_j|X_{lj}|\|_\gamma\|Y_l - Y_{l,\{0\}}\|_q + \|\max_j|X_{lj}-X_{lj,\{0\}}|\|_\gamma\|Y_{l,\{0\}}\|_q\right)\\ \leq\;& 2\Omega_{0,\gamma}\Delta_{0,q}.\end{aligned}$$

By the definition of γ and q, we have

$$\sup_{\tau\geq 2}\frac{\Omega_{0,\gamma}\Delta_{0,q}}{\tau^{\iota+\nu}} \leq \||X_\cdot|_\infty\|_{\psi_\iota}\|Y_\cdot\|_{\psi_\nu}\sup_{\tau\geq 2}\frac{\gamma^\iota q^\nu}{\tau^{\iota+\nu}} = \||X_\cdot|_\infty\|_{\psi_\iota}\|Y_\cdot\|_{\psi_\nu}C_7,$$

where $C_7 = (1+\iota/\nu)^{\iota}(1+\nu/\iota)^{\nu}$.

Write $\delta_0 = \beta^*$, $\delta_p = \beta$ and $\delta_l = \beta^* + \sum_{j=1}^{l}(\beta - \beta^*)_j = (\beta_1, ..., \beta_l, \beta^*_{l+1}, ..., \beta^*_p)$ for all $1 \le l \le p$. Again, denote

$$\begin{aligned}
\mathcal{T}_{l1} &= \left\{ \sup_{|\beta-\beta^*|_1 \le M} \left| \frac{1}{n}\sum_{i=1}^{n} [\bar{r}(f_{\delta_l}(X_i)) - \bar{r}(f_{\delta_{l-1}}(X_i))] \right| \le \frac{1}{2}\lambda_0 |\delta_l - \delta_{l-1}|_1 \right\}, \\
\mathcal{T}_{l2} &= \left\{ \sup_{|\beta-\beta^*|_1 \le M} \left| \frac{1}{n}\sum_{i=1}^{n} [\bar{h}(f_{\delta_l}(X_i))Y_i - \bar{h}(f_{\delta_{l-1}}(X_i))Y_i] \right| \le \frac{1}{2}\lambda_0 |\delta_l - \delta_{l-1}|_1 \right\}.
\end{aligned}$$

Then by Theorem 3.7.2,

$$\mathbb{P}(\mathcal{T}_{l2}^c) \le C_9 e^{-C_{10} B^{2/(1+2\iota+2\nu)} \log p}. \tag{3.8.10}$$

Similarly, by Theorem 3.7.2, we have

$$\mathbb{P}(\mathcal{T}_{l1}^c) \le C_7 e^{-C_8 A^{2/(1+2\iota)} \log p}. \tag{3.8.11}$$

Hence, $T_n \le n\lambda_0 M$ holds with probability at least,

$$\begin{aligned}
1 - p \cdot C_7 p^{-C_8 A^{2/(1+2\iota)}} - p \cdot C_9 p^{-C_{10} B^{2/(1+2\iota+2\nu)}} &\ge 1 - C_{11} p^{-C_{12} A^{2/(1+2\iota)}} \\
&\quad - C_{13} p^{-C_{14} B^{2/(1+2\iota+2\nu)}}.
\end{aligned}$$

□

Proof of Theorem 3.3.4. We use the short hand notation $S_* = S_{\beta^*}$, $s_* = s_{\beta^*}$ and $\kappa_* = \kappa(S_{\beta^*})$. The proof is similar to that of Theorem 3.3.1. Let

$$t := \frac{M}{M + |\hat{\beta} - \beta^*|_1},$$

and

$$\tilde{\beta} := t\hat{\beta} + (1-t)\beta^*.$$

Define $\widetilde{V}_n(\beta) = \sum_{i=1}^n [R(f_\beta(\widetilde{X}_i), Y_i) - \mathbb{E}R(f_\beta(\widetilde{X}_i), Y_i)]$. Let

$$\widetilde{T}_n := \sup_{|\beta - \beta^*|_1 \le M} |\widetilde{V}_n(\beta) - \widetilde{V}_n(\beta^*)|. \tag{3.8.12}$$

Since $\hat{\beta}$ minimizes (3.2.6), we have

$$\frac{1}{n}\sum_{i=1}^n R(f_{\tilde{\beta}}(\widetilde{X}_i), Y_i) + \lambda|\tilde{\beta}|_1 \le \frac{1}{n}\sum_{i=1}^n R(f_{\beta^*}(\widetilde{X}_i), Y_i) + \lambda|\beta^*|_1. \tag{3.8.13}$$

On the event

$$\tilde{\mathcal{T}} := \{\widetilde{T}_n \le n\lambda_0 M\} = \{\widetilde{T}_n \le n\mathcal{E}^*\}, \tag{3.8.14}$$

inequality (3.8.13) implies that

$$\begin{aligned}
&\mathcal{E}(f_{\tilde{\beta}}) + \lambda|\tilde{\beta}|_1 \\
&\le \lambda_0 M + \mathbb{E}R(f_{\beta^*}(\widetilde{X}_i), Y_i) - \mathbb{E}R(f_{\tilde{\beta}}(\widetilde{X}_i), Y_i) - \mathbb{E}R(f_{\beta^*}(X_i), Y_i) + \mathbb{E}R(f_{\tilde{\beta}}(X_i), Y_i) \\
&\qquad + \mathcal{E}(f_{\beta^*}) + \lambda|\beta^*|_1 \\
&\le \lambda_0 M + |\mathbb{E}(X_i - \widetilde{X}_i)^T(\tilde{\beta} - \beta^*)| + |\mathbb{E}[Y_i(X_i - \widetilde{X}_i)^T(\tilde{\beta} - \beta^*)]| + \mathcal{E}(f_{\beta^*}) + \lambda|\beta^*|_1 \\
&\le \lambda_0 M + \max_j E|X_{ij}\mathbf{1}_{\{|X_{ij}|>\tau\}}| \cdot |\tilde{\beta} - \beta^*|_1 + \max_j E|Y_i X_{ij}\mathbf{1}_{\{|X_{ij}|>\tau\}}| \cdot |\tilde{\beta} - \beta^*|_1 \\
&\qquad + \mathcal{E}(f_{\beta^*}) + \lambda|\beta^*|_1 \\
&\le \lambda_0 M + \lambda_C|\tilde{\beta} - \beta^*|_1 + \mathcal{E}(f_{\beta^*}) + \lambda|\beta^*|_1,
\end{aligned}$$

where $\lambda_C = C_1\tau^{-3q/4+1}$. If $\max_i |Y_i| < C < \infty$ for some constant C, then define $\lambda_C = C_1\tau^{-q+1}$. We set $\lambda_C = \lambda_0$. Now, for any β,

$$\beta = \beta_{S_*} + \beta_{S_*^c}.$$

Note that $\beta^*_{S_*} = \beta^*$ and $\beta^*_{S_*^c} = 0$. So, we have,

$$\begin{aligned}\mathcal{E}(f_{\tilde{\beta}}) + (\lambda - \lambda_C)|\tilde{\beta}_{S_*^c}|_1 &\leq \mathcal{E}^* + \mathcal{E}(f_{\beta^*}) + (\lambda + \lambda_C)|\tilde{\beta}_{S_*} - \beta^*_{S_*}|_1 \\ &\leq \frac{5}{3}\mathcal{E}^* + (\lambda + \lambda_C)|\tilde{\beta}_{S_*} - \beta^*_{S_*}|_1.\end{aligned}$$

Case a) If $(\lambda - 2\lambda_C)|\tilde{\beta} - \beta^*|_1 > 10\mathcal{E}^*/3$, we find

$$\mathcal{E}(f_{\tilde{\beta}}) + \frac{1}{2}\lambda|\tilde{\beta}_{S_*^c}|_1 \leq \frac{3}{2}\lambda|\tilde{\beta}_{S_*} - \beta^*_{S_*}|_1.$$

It follows that $|\tilde{\beta}_{S_*^c}|_1 \leq 3|\tilde{\beta}_{S_*} - \beta^*_{S_*}|_1$. This means that we can apply the compatibility condition. Thus, invoking the margin condition,

$$\begin{aligned}\mathcal{E}(f_{\tilde{\beta}}) + (\lambda - \lambda_C)|\tilde{\beta} - \beta^*|_1 &\leq \mathcal{E}^* + \mathcal{E}(f_{\beta^*}) + 2\lambda|\tilde{\beta}_{S_*} - \beta^*_{S_*}|_1 \\ &\leq \mathcal{E}^* + \mathcal{E}(f_{\beta^*}) + 2\lambda\sqrt{s_*}\|f_{\tilde{\beta}} - f_{\beta^*}\|/\kappa_* \\ &\leq \mathcal{E}^* + \mathcal{E}(f_{\beta^*}) + G\left(\frac{4\lambda\sqrt{s_*}}{\kappa_*}\right) + \frac{1}{2}\mathcal{E}(f_{\tilde{\beta}}) + \frac{1}{2}\mathcal{E}(f_{\beta^*})\end{aligned}$$

It follows that

$$\frac{1}{2}\mathcal{E}(f_{\tilde{\beta}}) + (\lambda - \lambda_C)|\tilde{\beta} - \beta^*|_1 \leq 2\mathcal{E}^*.$$

But then $(\lambda - 2\lambda_C)|\tilde{\beta} - \beta^*|_1 \leq 2\mathcal{E}^*$, which is a contradiction.

Case b) If $(\lambda - 2\lambda_C)|\tilde{\beta} - \beta^*|_1 \leq 10\mathcal{E}^*/3$, then

$$|\tilde{\beta} - \beta^*|_1 \leq \frac{10\lambda_0 M}{3(\lambda - 2\lambda_C)} \leq \frac{1}{2}M,$$

and

$$\mathcal{E}(f_{\tilde{\beta}}) + \lambda|\tilde{\beta}_{S_*^c}|_1 \leq 5\mathcal{E}^*.$$

Note that $\lambda_C = \lambda_0$. We immediately have $|\hat{\beta} - \beta^*|_1 \leq M$.

Hence, in both Case a) and Case b), the conclusion is that $|\hat{\beta} - \beta^*|_1 \leq M$. Following the same procedure, on the event $\tilde{\mathcal{T}}$, we can obtain bounds (3.3.12) and (3.3.13). So Theorem 3.3.4 follows if we can control the probability $\mathbb{P}(\tilde{\mathcal{T}})$.

Write $\delta_0 = \beta^*$, $\delta_p = \beta$ and $\delta_l = \beta^* + \sum_{j=1}^{l}(\beta - \beta^*)_j = (\beta_1, ..., \beta_l, \beta^*_{l+1}, ..., \beta^*_p)$ for all $1 \leq l \leq p$. Again, denote

$$\begin{aligned}
\mathcal{T}_{l1} &= \left\{ \sup_{|\beta-\beta^*|_1 \leq M} \left| \frac{1}{n} \sum_{i=1}^{n} [\bar{r}(f_{\delta_l}(\widetilde{X}_i)) - \bar{r}(f_{\delta_{l-1}}(\widetilde{X}_i))] \right| \leq \frac{1}{2}\lambda_0 |\delta_l - \delta_{l-1}|_1 \right\}, \\
\mathcal{T}_{l2} &= \left\{ \sup_{|\beta-\beta^*|_1 \leq M} \left| \frac{1}{n} \sum_{i=1}^{n} [\bar{h}(f_{\delta_l}(\widetilde{X}_i))Y_i - \bar{h}(f_{\delta_{l-1}}(\widetilde{X}_i))Y_i] \right| \leq \frac{1}{2}\lambda_0 |\delta_l - \delta_{l-1}|_1 \right\}.
\end{aligned}$$

Then by Theorem 3.7.2, for

$$\lambda_0 = A(\log p)^{1/2}(\log(pn)/n)^{1/2}\tau + B(\log p)^{(1+2\nu)/2}(\log(pn)/n)^{1/2}\tau \|Y_\cdot\|_{\psi_\nu},$$

$$\mathbb{P}(\mathcal{T}_{l2}^c) \leq C_9 e^{-C_{10}B^{2/(1+2\nu)} \log p}. \tag{3.8.15}$$

Similarly, by Theorem 3.7.3, we have

$$\mathbb{P}(\mathcal{T}_{l1}^c) \leq C_7 e^{-C_8 A^2 \log p}. \tag{3.8.16}$$

Hence, $\widetilde{T}_n \leq n\lambda_0 M$ holds with probability at least,

$$\left(1 - C_2 p^{-C_3 A^2} - C_4 p^{-C_5 B^{2/(1+2\nu)}}\right)^p \geq 1 - C_6 p^{-C_7 A^2} - C_8 p^{-C_9 B^{2/(1+2\nu)}}.$$

By setting $\lambda_0 = \lambda_C$, we are able to obtain τ. □

Lemma 3.8.1. *Assume $\|\beta^0\|_1 \leq C_\beta < \infty$. Also assume $\mathbb{E}|X_{ij}|^4 \leq C < \infty$, for any $1 \leq j \leq p$, and $\mathbb{E}|Y_i|^4 \leq C < \infty$. Choose $\tau_1, \tau_2 \asymp n^{1/4}/(\log p)^{3/4}$. For any $\eta_1 > 1$, it holds*

that

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_i(\widetilde{Y}_i-\widetilde{X}_i^T\beta^0)\right|_\infty > c_1\sqrt{\frac{\eta_1\log p}{n}}\right)\leq p^{1-\eta_1}, \tag{3.8.17}$$

where c_1 is a universal constant.

Proof. Applying Theorem 3.7.6, it can be carried out following the same routes as those in the proof of Lemma 1 Fan et al. [2016]. □

Lemma 3.8.2. *Assume $\mathbb{E}|X_{ij}|^4 \leq C < \infty$, for any $1 \leq j,k \leq p$. Denote $\hat{\Sigma}_{XX}(\tau_2) = \sum_{i=1}^n \widetilde{X}_i\widetilde{X}_i^T/n$, and $\Sigma_{XX} = \mathbb{E}X_iX_i^T$. Choose $\tau_2 \asymp n^{1/4}/(\log p)^{3/4}$. For any $\eta_2 > 2$, it holds that for some constant $c_2 > 0$,*

$$\mathbb{P}\left(\beta^T\hat{\Sigma}_{XX}(\tau_2)\beta \geq \beta^T\Sigma_{XX}\beta - c_2\eta_2\sqrt{\frac{\log p}{n}}\|\beta\|_1^2, \quad \forall\beta\in\mathbb{R}^p\right)\leq p^{2-\eta_2}. \tag{3.8.18}$$

Proof. Applying Theorem 3.7.6, it can be carried out following the same routes as those in the proof of Lemma 2 Fan et al. [2016]. □

Proof of Theorem 3.4.1. Employing the same arguments as those in the proof of Theorem 2 Fan et al. [2016] with Bernstein's inequality replaced by Theorem 3.7.6, Theorem 3.4.1 then follows from Theorem 3.7.6, Lemma 3.8.1 and Lemma 3.8.2. □

REFERENCES

Sumanta Basu and George Michailidis. Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.*, 43(4):1535–1567, 08 2015.

Ana M Bianco and Víctor J Yohai. Robust estimation in the logistic regression model. In *Robust statistics, data analysis, and computer intensive methods*, pages 17–34. Springer, 1996.

Peter J. Bickel, Yaacov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 08 2009.

Olivier Bousquet. A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495 – 500, 2002.

Emery N Brown, Robert E Kass, and Partha P Mitra. Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature neuroscience*, 7(5):456, 2004.

Christian Brownlees, Emilien Joly, Gábor Lugosi, et al. Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics*, 43(6):2507–2536, 2015.

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

D.L. Burkholder. Distribution function inequalities for martingales. *The Annals of Probability*, 1(1):19–42, 02 1973.

Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 12 2007.

Eva Cantoni and Elvezio Ronchetti. Robust inference for generalized linear models. *Journal of the American Statistical Association*, 96(455):1022–1030, 2001.

Olivier Catoni et al. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 48, pages 1148–1185. Institut Henri Poincaré, 2012.

Xiaohui Chen, Mengyu Xu, and Wei Biao Wu. Covariance and precision matrix estimation for high-dimensional time series. *Ann. Statist.*, 41(6):2994–3021, 12 2013.

V. Chernozhukov, D. Chetverikov, and K. Kato. Testing many moment inequalities. *arXiv preprint arXiv:1312.7614*, 2013.

V. Chernozhukov, D. Chetverikov, and K. Kato. Comparison and anti-concentration bounds for maxima of gaussian random vectors. *Probability Theory and Related Fields*, 162(1-2): 47–70, 2014.

Xialiang Dou and Mihai Anitescu. Distributionally robust optimization with correlated data from vector autoregressive processes. *Operations Research Letters*, 47(4):294–299, 2019.

Xialiang Dou and Tengyuan Liang. Training neural networks as learning data-adaptive kernels: Provable representation and approximation benefits. *arXiv preprint arXiv:1901.07114*, 2019.

John C Duchi, Khashayar Khosravi, and Feng Ruan. Multiclass classification, information, divergence, and surrogate risk. *arXiv preprint arXiv:1603.00126*, 2016.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 04 2004.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360, 2001.

Jianqing Fan, Weichen Wang, and Ziwei Zhu. A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. *arXiv preprint arXiv:1603.08315*, 2016.

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.

Yingying Fan and Jinchi Lv. Asymptotic equivalence of regularization methods in thresholded parameter space. *Journal of the American Statistical Association*, 108(503):1044–1061, 2013.

Stéphane Gaïffas, Agathe Guilloux, et al. High-dimensional additive hazards models and the lasso. *Electronic Journal of Statistics*, 6:522–546, 2012.

Alexander Genkin, David D Lewis, and David Madigan. Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304, 2007.

Eric Ghysels, Pedro Santa-Clara, and Rossen Valkanov. The midas touch: Mixed data sampling regression models. *Finance*, 2004.

Shuva Gupta. A note on the asymptotic distribution of lasso estimator for correlated data. *Sankhyā: The Indian Journal of Statistics, Series A (2008-)*, 74(1):10–28, 2012.

Wooseok Ha, Haoyang Liu, and Rina Foygel Barber. An equivalence between stationary points for rank constraints versus low-rank factorizations. *arXiv preprint arXiv:1812.00404*, 2018.

Eric C Hall, Garvesh Raskutti, and Rebecca Willett. Inference of high-dimensional autoregressive generalized linear models. *arXiv preprint arXiv:1605.02693*, 2016.

Frank R Hampel. A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, pages 1887–1896, 1971.

Frank R Hampel. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393, 1974.

Yuefeng Han and Ruey S Tsay. High-dimensional linear regression for dependent observations with application to nowcasting. *arXiv preprint arXiv:1706.07899*, 2017.

Jerry A. Hausman, Andrew W. Lo, and A.Craig MacKinlay. An ordered probit analysis of transaction stock prices. *Journal of Financial Economics*, 31(3):319 – 379, 1992.

Richard Walter Hill. *Robust regression when there are outliers in the carriers.* PhD thesis, Harvard University, 1977.

Joseph L Hodges Jr. Efficiency in normal samples and tolerance of extreme values for some estimates of location. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 163–186, 1967.

Jian Huang and Cun-Hui Zhang. Estimation and selection via absolute penalized convex minimization and its multistage adaptive applications. *Journal of Machine Learning Research*, 13(Jun):1839–1864, 2012.

Jian Huang, Tingni Sun, Zhiliang Ying, Yi Yu, and Cun-Hui Zhang. Oracle inequalities for the lasso in the cox model. *Ann. Statist.*, 41(3):1142–1165, 06 2013.

Peter J Huber. Robust estimation of a location parameter. *The annals of mathematical statistics*, pages 73–101, 1964.

Peter J Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233. Berkeley, CA, 1967.

Stéphane Ivanoff, Franck Picard, and Vincent Rivoirard. Adaptive lasso and group-lasso for functional poisson regression. *Journal of Machine Learning Research*, 17(55):1–46, 2016.

Anders Bredahl Kock and Laurent Callot. Oracle inequalities for high dimensional vector autoregressions. *J Econom*, 186(2):325 – 344, 2015.

Shengchun Kong and Bin Nan. Non-asymptotic oracle inequalities for the high-dimensional cox regression via lasso. *Statistica Sinica*, 24(1):25, 2014.

B. Krishnapuram, L. Carin, M. A. T. Figueiredo, and A. J. Hartemink. Sparse multinomial logistic regression: fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):957–968, 2005.

Shuo Li, Xialiang Dou, Ruiqi Gao, Xinzhou Ge, Minping Qian, and Lin Wan. A remark on copy number variation detection methods. *PloS one*, 13(4), 2018.

Haoyang Liu. Exact high-dimensional asymptotics for support vector machine. *arXiv preprint arXiv:1905.05125*, 2019.

Haoyang Liu and Rina Foygel Barber. Between hard and soft thresholding: optimal iterative thresholding algorithms. *arXiv preprint arXiv:1804.08841*, 2018.

Haoyang Liu and Chao Gao. Density estimation with contaminated data: Minimax rates and theory of adaptation. *arXiv preprint arXiv:1712.07801*, 2017.

Po-Ling Loh. Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *Ann. Statist.*, 45(2):866–896, 04 2017.

Justin Lokhorst. The lasso and generalised linear models. *Honors Project, The University of Adelaide, Australia*, 1999.

Colin L Mallows. On some topics in robustness. *Unpublished memorandum, Bell Telephone Laboratories, Murray Hill, NJ*, 1975.

P McCullagh and John A Nelder. *Generalized Linear Models*, volume 37. CRC Press, 1989.

Marcelo C. Medeiros and Eduardo F. Mendes. L1-regularization of high-dimensional time-series models with non-gaussian and heteroskedastic errors. *J Econom*, 191(1):255 – 271, 2016.

Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

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

Florence Merlevède, Magda Peligrad, and Emmanuel Rio. *Bernstein inequality and moderate deviations under strong mixing conditions*, volume 5 of *Collections*, pages 273–292. Institute of Mathematical Statistics, 2009.

Hyde M Merrill and Fred C Schweppe. Bad data suppression in power system static state estimation. *IEEE Transactions on Power Apparatus and Systems*, (6):2718–2725, 1971.

Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statist. Sci.*, 27(4):538–557, 11 2012.

Jonathan W Pillow, Jonathon Shlens, Liam Paninski, Alexander Sher, Alan M Litke, EJ Chichilnisky, and Eero P Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995, 2008.

I. Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pages 1679–1706, 1994.

M.B. Priestley. *Non-linear and Non-stationary Time Series Analysis*. Academic Press, 1988.

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.

M. Rosenblatt. *Markov Processes: Structure and Asymptotic Behavior*. Springer, 1971.

Volker Roth. The generalized lasso. *IEEE transactions on neural networks*, 15(1):16–28, 2004.

Xiaofeng Shao and Wei Biao Wu. Asymptotic spectral theory for nonlinear time series. *Ann. Statist.*, 35(4):1773–1801, 08 2007.

Shirish Krishnaj Shevade and S Sathiya Keerthi. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19(17):2246–2253, 2003.

James H. Stock and Mark W. Watson. Forecasting using principal components from a large number of predictors. *J. Amer. Statist. Assoc.*, 97(460):1167–1179, 2002.

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

Robert Tibshirani. The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16(4):385–395, 1997.

Ryan J. Tibshirani and Jonathan Taylor. Degrees of freedom in lasso problems. *Ann. Statist.*, 40(2):1198–1232, 04 2012.

H. Tong. *Non-linear Time Series: A Dynamical System Approach*. Oxford University Press, 1990.

R.S. Tsay. *Analysis of Financial Time Series*, volume 543. John Wiley & Sons, 2005.

R.S. Tsay and R. Chen. *Nonlinear Time Series Analysis*. Wiley Series in Probability and Statistics. Wiley, 2018.

John W Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pages 448–485, 1960.

John W Tukey. The future of data analysis. *The Annals of Mathematical Statistics*, 33(1): 1–67, 1962.

Sara van de Geer and Patric Müller. Quasi-likelihood and/or robust estimation in high dimensions. *Statistical Science*, pages 469–480, 2012.

Sara A. van de Geer. High-dimensional generalized linear models and the lasso. *Ann. Statist.*, 36(2):614–645, 04 2008.

Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Trans. Inf. Theory*, 55(5): 2183–2202, 2009.

Hansheng Wang, Guodong Li, and Chih-Ling Tsai. Regression coefficient and autoregressive order shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, 69(1):63–78, 2007. ISSN 1467-9868.

N. Wiener. *Nonlinear Problems in Random Theory*. Wiley, New York, 1958.

Wei Biao Wu. Nonlinear system theory: Another look at dependence. *Proc. Natl. Acad. Sci. U. S. A.*, 102(40):14150–14154, 2005.

Wei Biao Wu and Xiaofeng Shao. Limit theorems for iterated random functions. *Journal of Applied Probability*, 41(2):425–436, 2004.

Wei-Biao Wu and Ying Nian Wu. Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electron. J. Statist.*, 10(1):352–379, 2016.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B*, 68(1):49–67, 2006. ISSN 1467-9868.

Chunming Zhang, Xiao Guo, Chen Cheng, and Zhengjun Zhang. Robust-bd estimation and inference for varying-dimensional general linear models. *Statistica Sinica*, pages 653–673, 2014.

Danna Zhang and Wei Biao Wu. Gaussian approximation for high dimensional time series. *Ann. Statist.*, 45(5):1895–1919, 2017.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7: 2541–2563, December 2006. ISSN 1532-4435.

Hao Zhou and Garvesh Raskutti. Non-parametric sparse additive auto-regressive network models. *arXiv preprint arXiv:1801.07644*, 2018.

Hui Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476): 1418–1429, 2006.

Hui Zou, Trevor Hastie, and Robert Tibshirani. On the degrees of freedom of the lasso. *Ann. Statist.*, 35(5):2173–2192, 10 2007.