Lasso Guarantees for Dependent Data

by

Kam Chung Wong

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2017

Doctoral Committee:

Associate Professor Ambuj Tewari, Chair
Assistant Professor Laura Kathryn Balzano
Professor Edward Ionides
Assistant Professor Po-Ling Loh

Kam Chung Wong
kamwong@umich.edu
ORCID iD:0000-0002-9269-9845

TABLE OF CONTENTS

ABSTRACT . . . . **v**

CHAPTER

I. Introduction 1

1.1 Recent Work on High Dimensional Time Series 4

II. Lasso Guarantees for β-Mixing Heavy Tailed Time Series 6

2.1 Introduction 6
2.1.1 Organization of the Chapter 8
2.2 Preliminaries 9
2.2.1 Notation 10
2.2.2 Sparsity, Stationarity and Zero Mean Assumptions 11
2.2.3 A Master Theorem 12
2.2.4 A Brief Overview of Mixing Conditions 14
2.3 Gaussian Processes under α-Mixing 16
2.3.1 Estimation and Prediction Errors 18
2.3.2 Examples 19
2.4 Subweibull Random Vectors under β-Mixing 21
2.4.1 Subweibull Random Variables and Vectors 22
2.4.2 Estimation and Prediction Errors 26
2.4.3 Examples 27
2.5 Proof of Master Theorem 29
2.6 Proofs for Gaussian Processes under α-Mixing 31
2.6.1 Hanson-Wright Inequality 40
2.7 Proofs for Subweibull Random Vectors under β-Mixing 41
2.7.1 Proof Related to Subweibull Properties 41
2.7.2 Subweibull Norm Under Linear Transformations 43
2.7.3 Concentration Inequality for Sums of β-Mixing Subweibull Random Variables 43

2.7.4 Proofs of Deviation and RE Bounds 47

2.8 Verification of Assumptions for the Examples 52

2.8.1 VAR 52

2.8.2 VAR with Misspecification 55

2.8.3 ARCH 57

III. Lasso Guarantees for Time Series Estimation Under Subgaussian Tails and β-Mixing 60

3.1 Introduction 60

3.2 Preliminaries 62

3.3 Main Results 65

3.3.1 Assumptions 67

3.3.2 High Probability Guarantees for the Lower Restricted Eigenvalue and Deviation Bound Conditions 69

3.3.3 Estimation and Prediction Errors 70

3.4 Examples 71

3.5 Simulations 75

3.6 Supplement 75

3.6.1 Sub-Gaussian Constants for Random Vectors 75

3.6.2 Proof of Master Theorem 77

3.6.3 Proofs for Sub-Gaussian Random Vectors under β-Mixing 80

3.6.4 Bernstein's Concentration Inequality 89

3.6.5 Verification of Assumptions for the Examples 89

IV. Lasso Guarantees for Gaussian Vector Autoregressive Processes with Missing or Noisy Data 97

4.1 Introduction 97

4.2 Preliminaries 100

4.2.1 Matrix and Vector Notations 100

4.2.2 VAR 101

4.2.3 Lasso Estimation Procedure For Transition Matrix 104

4.3 Theoretical Guarantees 106

4.3.1 Lasso Consistency and Sufficient Conditions 106

4.3.2 Missing and Corrupted Data 108

4.3.3 High Probabilistic Guarantees for Additive Data Corruption Case 109

4.3.4 High Probabilistic Guarantees for Multiplicative Noise Case 110

4.4 Proofs of Corrupted Data Case 111

4.4.1 Variant of Hanson-Wright Inequality 111

4.4.2 Proof of Restricted Eigenvalue Guarantees 112
4.4.3 Deviation Bound 117
4.5 Proofs of Missing Data Case 122

V. Future Directions 124

Bibliography **127**

ABSTRACT

Serially correlated high dimensional data are prevalent in the big data era. In order to predict and learn the complex relationship among the multiple time series, high dimensional modeling has gained importance in various fields such as control theory, statistics, economics, finance, genetics and neuroscience. We study a number of high dimensional statistical problems involving different classes of mixing processes. For example, given a sequence (X_t), one might be interested in predicting X_t using the past observations $X_{t-1}, \cdots, X_{t-d}$.

The vector autoregressive (VAR) models naturally admit a linear autoregressive representation that allows us to study the joint evolution of the time series. In the high-dimensional setting, where both the number of components of the time series and the order of the model are allowed to grow with sample size, consistent estimation is impossible without structural assumptions on the transition matrices. One such structural assumption is that of sparsity. The lasso program is a celebrated method for sparse estimation. The majority of theoretical and empirical results on lasso, however, assume iid data. In addition, it is common for real data sets to contain missing and/or corrupted data. In the autoregressive scenario, both the independent and dependent variables are affected, and hence requires careful consideration. We study the problem based upon the framework proposed in Loh and Wainwright [2012]

In addition, many theoretical results on estimation of high dimensional time series require specifying an underlying data generating model (DGM). Instead, we assume only (strict) stationarity and mixing conditions to establish finite sample consistency of lasso for data coming from various families of distributions. When the DGM is nonlinear, the lasso estimate corresponds to that of a best linear predictor which we assume is sparse. We provide results for three different combinations of dependence and tail behavior of the time series: α-mixing and Gaussian, β-mixing and subgaussian tails, and β-mixing and subweilbull tails. To prove our results for the second set, we derive a novel Hanson-Wright type concentration inequality for β-mixing subgaussian random vectors that may be of independent interest. Together, applications of these results extend to non-Gaussian, non-Markovian and non-linear times series models as the examples we provide demonstrate.

CHAPTER I

Introduction

This thesis establishes high probability guarantees for the restricted eigenvalue (RE) and deviation bound (DB) conditions for a wide range of geometrically mixing processes with subweibull observations. To the best of our knowledge, similar guarantees have not been given in the literature before. We also provide the RE and DB guarantees for noisy and corrupted data under the modified lasso framework(Loh and Wainwright [2012]). The RE and DB conditions, in addition to the sparsity assumption, allow (via a standard result) us to give high probabilistic non-asymptotic lasso guarantees in the respective scenarios.

The information age and scientific advances have led to explosions in large data sets as well as new statistical methods and algorithms aimed at extracting valuable information in them. On the data side, it is common to see massive dependent data collected from, for example, micro-array experiments, social networks, mobile phone usage, high frequency stock market trading, daily grocery sales, etc. At the same time, the high speed Internet makes big data warehouses readily accessible.

The surge in big data has stimulated exciting developments in statistical models and algorithms to do prediction and/or gain scientific understanding in the data. Among the plethora of methods, the linear parametric models remain highly popular thanks to its superiority in interpretability, computational efficiency, and the rich and sophisticated theoretical literature. For example, given a sequence of observations (X_t), one might be interested in predicting X_t using linear combinations of the past observations $X_{t-1}, \cdots, X_{t-d}$.

The vector autoregressive (VAR) models are a linear parametric family that allows researchers to model interrelationships among variables that exhibit temporal dependence. The transition matrices in a VAR capture the co-evolutionary dynamics of the system of variables and are the central objects of estimation in this thesis. Many theoretical results on estimation of high dimensional time series require specifying an underlying data generating model (DGM) which in reality is almost never known. Instead, we view the VAR models trained on the data as a *predictive* tool and do not assume any parametric DGM. When the underlying DGM is really VAR, the lasso estimates correspond to those of the transition matrices; otherwise, they are estimates of a best sparse linear predictor. Alternative to their DGM, dependent data can be characterized by (1) their dependence structure, and (2) the probability distribution of marginal observations. We provide lasso guarantees for (strictly) stationary time series data that satisfy some dependence and tail behavior conditions.

There are three popular approaches to measuring dependence: physical and predictive dependence measures [Wu, 2005], spectral analysis [Priestley, 1981, Stoica and Moses, 1997, Basu and Michailidis, 2015], and mixing coefficients [Bradley, 2005]. We adopt the mixing coefficients route. One of the advantages of this measure is that the mixing coefficients with respect to a process is "*invariant*" under measurable transformation of the original process. This allows us to study dependence structure of more complicated processes from simpler ones. We consider a full spectrum of (geometric) mixing processes – from independent to identical data.

The other dimension in analysis of a time series is the marginal distribution of observations. In particular, the tail behavior of the random variable has direct impact on the dimension and sample size scalings of the lasso bounds. Because of that, we study and define a class of *subweibull(γ)* (with $\gamma > 0$) random variables characterized by tails decaying in the same order of magnitude as that of a Weibull(γ) one. The parameter γ serves as a measure of how heavy the tail is. The subweibull family subsumes the well-known subgaussian ($\gamma = 2$) and subexponential ($\gamma = 1$) classes. It is common in the literature (eg, see Foss et al. [2011]) to call a random variable *heavy-tailed* if its tail decays slower than exponential. This is natural because their moment generation functions fail to exist at any point. As such, the subweibull family also includes some heavy tailed random variables.

The crux of this thesis consists of non-asymptotic lasso guarantees for full spectrum of (geometrically) mixing processes with subweibull(γ) observations and are discussed in details in Sections II and III. Specifically, we present results for three different combinations of dependence and tail behavior: α-mixing and Gaussian (Section II), β-mixing and subgaussian tails (Section III), and β-mixing and subweilbull tails (Section II). To prove our results for the second set, we derive a novel Hanson-Wright type concentration inequality for β-mixing subgaussian random vectors that may be of independent interest. To illustrate the applicability of the theory, we give examples on nonlinear time series (autoregressive conditionally heteroscedastic model), misspecified and non-Markovian model (VAR with endogenous variable left out), heavy-tailed time series (subweibull VAR). Justification of the mixing and subweibull assumptions for these examples are provided. This concludes the part on general mixing processes.

On the other hand, when the true DGM is VAR, the lasso estimates correspond to those of the transition matrices which we assume to be sparse. Despite the relatively simpler scenarios, in real data applications, we are faced with issues such as noisy or missing data. We consider the simple Gaussian data corruption and missing completely at random problems in Section IV and provide the associated lasso guarantees. It is based on the modified lasso framework proposed by Loh and Wainwright [2012].

The thesis is concluded with future directions of research in Section V.

1.1 Recent Work on High Dimensional Time Series

While we mention a few related work in the introductions of the ensuing chapters as motivation, we wish to emphasize that several other researchers have recently published work on statistical analysis of high dimensional time series. Song and Bickel [2011], Wu and Wu [2015] and Alquier et al. [2011] give theoretical guarantees assuming that RE conditions hold. As Basu and Michailidis [2015] pointed out, it takes a fair bit of work to actually establish RE conditions in the presence of dependence. Chudik and Pesaran [2011, 2013, 2014] use high dimensional time series for global macroeconomic modeling. Alternatives to lasso that have been explored include quantile based methods for heavy-tailed data [Qiu et al., 2015], quasi-likelihood approaches [Uematsu, 2015], two-stage estimation techniques [Davis et al., 2012] and the Dantzig selector [Han and Liu, 2013, Han et al., 2015]. Both Han and Liu [2013] and Han et al. [2015] studied the stable Gaussian VAR models while this chapter covers wider classes of processes as our examples demonstrate. Fan et al. [2016] considered the case of multiple sequences of univariate α-mixing heavy-tailed dependent data. Under a stringent condition on the auto-covariance structure (please refer to Chapter 2.7 for details), the paper established finite sample ℓ_2 consistency in the real support for penalized least squares estimators. In addition, under mutual incoherence type assumption, it provided sign and ℓ_∞ consistency. An AR(1) example was given as an illustration. Both Uematsu [2015] as well as Kock and Callot [2015] establish oracle inequalities for the lasso applied to time series prediction. Uematsu [2015] provided results not just for lasso but also for estimators using penalties such as the SCAD penalty. Also, instead of assuming Gaussian errors, it is only assumed that fourth moments of the errors exist. Kock and Callot [2015] provided non-asymptotic lasso error and prediction error bounds for stable Gaussian VARs. Both Sivakumar et al. [2015] and Medeiros and Mendes [2016] considered subexponential designs. Sivakumar et al. [2015] studied lasso on iid subexponential designs and provide finite sample bounds. Medeiros and Mendes [2016] studied adaptive lasso for linear time series models and provide sign consistency results. Wang et al. [2007] provided theoretical guarantees for lasso in linear regression models with autoregressive errors. Other structured penalties beyond the ℓ_1 penalty have also been considered [Nicholson et al., 2014, 2015, Guo et al., 2015, Ngueyep and Serban, 2014]. Zhang and Wu [2015], McMurry and Politis [2015], Wang et al. [2013] and Chen et al. [2013] consider estimation of the covariance (or precision) matrix of high dimensional time series. McMurry and Politis [2015] and Nardi and Rinaldo [2011] both highlight that autoregressive (AR) estimation, even in univariate time series, leads to high dimensional parameter estimation problems if the lag is allowed to be unbounded.

CHAPTER II

Lasso Guarantees for β-Mixing Heavy Tailed Time Series

2.1 Introduction

High dimensional statistics is a vibrant area of research in modern statistics and machine learning [Bühlmann and Van De Geer, 2011, Hastie et al., 2015]. The interplay between computational and statistical aspects of estimation in high dimensions has led to a variety of efficient algorithms with statistical guarantees including methods based on convex relaxation (see, e.g., Chandrasekaran et al. [2012], Negahban et al. [2012]) and methods using iterative optimization techniques (see, e.g., Beck and Teboulle [2009], Agarwal et al. [2012], Donoho et al. [2009]). However, the bulk of existing theoretical work focuses on iid samples. The extension of theory and algorithms in high dimensional statistics to time series data, where dependence is the norm rather than the exception, is just beginning to occur. We briefly summarize some recent work in Section 1.1 below.

Our focus in this chapter is to give guarantees for ℓ_1-regularized least squares estimation, or lasso [Hastie et al., 2015], that hold even when there is temporal dependence in data. The recent work of Basu and Michailidis [2015] took a major step forward in providing guarantees for lasso in the time series setting. They considered Gaussian Vector Auto-Regressive (VAR) models with finite lag (see Example 1) and defined a measure of stability using the spectral density, which is the Fourier transform of the autocovariance function of the time series. Then they showed that one can derive error bounds for lasso in terms of their measure of stability. Their bounds are an improvement over previous work [Negahban and Wainwright, 2011, Loh and Wainwright, 2012, Han and Liu, 2013] that assumed operator norm bounds on the transition matrix. These operator norm conditions are restrictive even for VAR models with a lag of 1 and never hold (Please see pp. 11–13 in the Supplement of Basu and Michailidis [2015] for details) if the lag is strictly larger than 1! Therefore, the results of Basu and Michailidis [2015] hold in greater generality than previous work. But they do have limitations.

A key limitation is that Basu and Michailidis [2015] assume that the VAR model is the true data generating mechanism (DGM). Their proof techniques rely heavily on having the VAR representation of the stationary process available. The VAR model assumption, while popular in many areas, can be restrictive since the VAR family is not closed under linear transformations: if Z_t is a VAR process then CZ_t may not expressible as a finite lag VAR [Lütkepohl, 2005]. We later provides examples (Examples 2 and 4) of VAR processes where leaving out a single variable breaks down the VAR assumption. What if we do not assume that Z_t is a finite lag VAR process but simply that it is stationary? Under stationarity (and finite 2nd moment conditions), the best linear predictor of Z_t in terms of $Z_{t-d}, \ldots, Z_{t-1}$ is well defined even if Z_t is not a lag d VAR. If we assume that this best linear predictor involves sparse coefficient matrices, can we still guarantee consistent parameter estimation? This chapter provides an affirmative answer to this important question.

We provide finite sample parameter estimation and prediction error bounds for lasso in two cases: (a) for stationary Gaussian processes with suitably decaying α-mixing coefficients (Section 2.3), and (b) for stationary processes with subweibull marginals and geometrically decaying β-mixing coefficients (Section 2.4). The class of subweibull random variables that we introduce includes subgaussian and subexponential random variables but also includes random variables with tails heavier than an exponential. We also show that, for Gaussian processes, the β-mixing condition can be relaxed to summability of the α-mixing coefficients. Our work provides an alternative proof of the consistency of the lasso for sparse Gaussian VAR models. But the applicability of our results extends to non-Gaussian and non-linear times series models as the examples we provide demonstrate.

It is well known that guarantees for lasso follow if one can establish restricted eigenvalue (RE) conditions and provide deviation bounds (DB) for the correlation of noise with the regressors (see the Master Theorem in Section 2.2.3 below for a precise statement). Therefore, the bulk of the technical work in this chapter boils down to establishing, with high probability, that RE and DB conditions hold under the Gaussian α-mixing Propositions II.3 and II.2) and the subweibull β-mixing assumptions respectively (Propositions II.7 and II.8). Note that RE conditions were previously shown to hold under the *iid assumption* by Raskutti et al. [2010] for Gaussian random vectors and by Rudelson and Zhou [2013] for subgaussian random vectors.

2.1.1 Organization of the Chapter

Section 2.2 introduces our notation, presents the common assumptions and states some useful facts needed later. Then we present two sets of high probability guarantees for the lower restricted eigenvalue and deviation bound conditions in Sections 2.3 and 2.4 respectively. Section 2.3 covers α-mixing Gaussian time series. Note that α-mixing is a weaker notion than β-mixing and all the parameter dependences are explicit. It is followed by Section 2.4 which covers β-mixing time series with subweibull observations and we make the dependence on the subweibull norm explicit.

We present five examples, two involving α-mixing Gaussian processes and three β-mixing subweibull vectors. They are presented along with the corresponding theoretical results to illustrate applicability of the theory. Examples 1 and 2 concern applications of the results in Section 2.3. We consider VAR models with Gaussian innovations when the model is correctly or incorrectly specified. In Examples 3, 4, and 5, we focus on the case of subweibull random vectors. We consider VAR models with subweibull innovations when the model is correctly or incorrectly specified (Examples 3 and 4). In addition, we go beyond linear models and introduce non-linearity in the DGM in Example 5.

These examples serve to illustrate that our theoretical results for lasso on high dimensional dependent data estimation extend beyond the classical linear Gaussian setting and provides guarantees potentially in the presence of one or more of the following scenarios: model mis-specification, heavy tailed non-Gaussian innovations and nonlinearity in the DGM.

2.2 Preliminaries

Consider a stochastic process of pairs $(X_t, Y_t)_{t=1}^{\infty}$ where $X_t \in \mathbb{R}^p$, $Y_t \in \mathbb{R}^q$, $\forall t$. One might be interested in predicting Y_t given X_t. In particular, given a dependent sequence $(Z_t)_{t=1}^{T}$, one might want to forecast the present Z_t using the past $(Z_{t-d}, \ldots, Z_{t-1})$. A linear predictor is a natural choice. To put it in the regression setting, we identify $Y_t = Z_t$ and $X_t = (Z_{t-d}, \ldots, Z_{t-1})$. The pairs (X_t, Y_t) defined as such are no longer iid. Assuming strict stationarity, the parameter matrix of interest

$\Theta^\star \in \mathbb{R}^{p\times q}$ is

$$\Theta^\star = \operatorname*{arg\,min}_{\Theta\in\mathbb{R}^{p\times q}} \mathbb{E}[\|Y_t - \Theta' X_t\|_2^2]. \tag{2.2.1}$$

Note that $\Theta^\star$ is independent of t owing to stationarity. Because of high dimensionality ($pq \gg T$), consistent estimation is impossible without regularization. We consider the lasso procedure. The ℓ_1-penalized least squares estimator $\widehat{\Theta} \in \mathbb{R}^{p\times q}$ is defined as

$$\widehat{\Theta} = \operatorname*{arg\,min}_{\Theta\in\mathbb{R}^{p\times q}} \frac{1}{T}\| \operatorname{vec}(\mathbf{Y} - \mathbf{X}\Theta)\|_2^2 + \lambda_T \left\|\operatorname{vec}(\Theta)\right\|_1 . \tag{2.2.2}$$

where

$$\mathbf{Y} = (Y_1, Y_2, \ldots, Y_T)' \in \mathbb{R}^{T\times q} \qquad \mathbf{X} = (X_1, X_2, \ldots, X_T)' \in \mathbb{R}^{T\times p}. \tag{2.2.3}$$

The following matrix of true residuals is not available to an estimator but will appear in our analysis:

$$\mathbf{W} := \mathbf{Y} - \mathbf{X}\Theta^\star. \tag{2.2.4}$$

2.2.1 Notation

For scalars a and b, define shorthands $a \wedge b := \min\{a, b\}$ and $a \vee b := \max\{a, b\}$. For a symmetric matrix $\mathbf{M}$, let $\lambda_{\max}(\mathbf{M})$ and $\lambda_{\min}(\mathbf{M})$ denote its maximum and minimum eigenvalues respectively. For any matrix let $\mathbf{M}$, $r(\mathbf{M})$, $|||\mathbf{M}|||$, $|||\mathbf{M}|||_\infty$, and $|||\mathbf{M}|||_F$ denote its spectral radius $\max_i \{|\lambda_i(\mathbf{M})|\}$, operator norm $\sqrt{\lambda_{\max}(\mathbf{M}'\mathbf{M})}$, entry-wise ℓ_∞ norm $\max_{i,j} |\mathbf{M}_{i,j}|$, and Frobenius norm $\sqrt{\operatorname{tr}(\mathbf{M}'\mathbf{M})}$ respectively. For any vector $v \in \mathbb{R}^p$, $\|v\|_q$ denotes its ℓ_q norm $(\sum_{i=1}^p |v_i|^q)^{1/q}$. Unless otherwise specified, we shall use $\|\cdot\|$ to denote the ℓ_2 norm. For any vector $v \in \mathbb{R}^p$, we use $\|v\|_0$ and $\|v\|_\infty$ to denote $\sum_{i=1}^p \mathbb{1}\{v_i \neq 0\}$ and $\max_i\{|v_i|\}$ respectively. Similarly, for any matrix $\mathbf{M}$, $|||\mathbf{M}|||_0 = \|\operatorname{vec}(\mathbf{M})\|_0$ where $\operatorname{vec}(\mathbf{M})$ is the vector obtained from $\mathbf{M}$ by concatenating the rows of M. We say that matrix $\mathbf{M}$ (resp. vector v) is *s-sparse* if $\left|\left|\left|\mathbf{M}\right|\right|\right|_0 = s$ (resp. $\|v\|_0 = s$). We use v' and $\mathbf{M}'$ to denote the transposes of v and $\mathbf{M}$ respectively. When we index a matrix, we adopt the following conventions. For any matrix $\mathbf{M} \in \mathbb{R}^{p \times q}$, for $1 \leq i \leq p$, $1 \leq j \leq q$, we define $\mathbf{M}[i,j] \equiv \mathbf{M}_{ij} := e_i' \mathbf{M} e_j$, $\mathbf{M}[i,:] \equiv \mathbf{M}_{i:} := e_i' \mathbf{M}$ and $\mathbf{M}[:,j] \equiv \mathbf{M}_{:j} := \mathbf{M} e_j$ where e_i is the vector with all 0s except for a 1 in the ith coordinate. The set of integers is denoted by $\mathbb{Z}$.

For a lag $l \in \mathbb{Z}$, we define the auto-covariance matrix w.r.t. $(X_t, Y_t)_t$ as $\Sigma(l) = \Sigma_{(X;Y)}(l) := \mathbb{E}[(X_t; Y_t)(X_{t+l}; Y_{t+l})']$. Note that $\Sigma(-l) = \Sigma(l)'$. Similarly, the auto-covariance matrix of lag l w.r.t. $(X_t)_t$ is $\Sigma_X(l) := \mathbb{E}[X_t X_{t+l}']$, and w.r.t. $(Y_t)_t$ is $\Sigma_Y(l) := \mathbb{E}[Y_t Y_{t+l}']$. At lag 0, we often simplify the notation as $\Sigma_X \equiv \Sigma_X(0)$ and $\Sigma_Y \equiv \Sigma_Y(0)$. The cross-covariance matrix at lag l is $\Sigma_{X,Y}(l) := \mathbb{E}[X_t Y_{t+l}']$. Note the difference between $\Sigma_{(X;Y)}(l)$ and $\Sigma_{X,Y}(l)$: the former is a $(p+q) \times (p+q)$ matrix, the latter is a $p \times q$ matrix. Thus, $\Sigma_{(X;Y)}(l)$ is a matrix consisting of four sub-matrices. Using Matlab-like notation, $\Sigma_{(X;Y)}(l) = [\Sigma_X, \Sigma_{X,Y}; \Sigma_{Y,X}, \Sigma_Y]$. As per our convention, at lag 0, we omit the lag argument l. For example, $\Sigma_{X,Y}$ denotes $\Sigma_{X,Y}(0) = \mathbb{E}[X_t Y_t']$. Finally, let $\hat{\Gamma} := \frac{\mathbf{X}'\mathbf{X}}{T}$ be the empirical covariance matrix.

2.2.2 Sparsity, Stationarity and Zero Mean Assumptions

The following assumptions are maintained throughout; we will make additional assumptions specific to each of the subweibull and Gaussian scenarios. Our goal is to provide finite sample bounds on the error $\widehat{\Theta} - \Theta^\star$. We shall present theoretical guarantees on the ℓ_2 parameter estimation error $\|\operatorname{vec}(\widehat{\Theta} - \Theta^\star)\|_2$ and also the associated (in-sample) prediction error $\left|\left|\left|(\widehat{\Theta} - \Theta^\star)'\hat{\Gamma}(\widehat{\Theta} - \Theta^\star)\right|\right|\right|_F$.

Assumption 1. The matrix $\Theta^\star$ is s-sparse, i.e., $\|\operatorname{vec}(\Theta^\star)\|_0 \leq s$.

Assumption 2. The process (X_t, Y_t) is strictly stationary: i.e., $\forall m, \tau,\ n \geq 0$,

$$((X_m, Y_m), \cdots, (X_{m+n}, Y_{m+n})) \ \overset{d}{=}\ ((X_{m+\tau}, Y_{m+\tau}), \cdots, (X_{m+n+\tau}, Y_{m+n+\tau})).$$

where "$\overset{d}{=}$" denotes equality in distribution.

Assumption 3. The process (X_t, Y_t) is centered; i.e., $\forall t,\ \mathbb{E}(X_t) = 0_{p\times 1}$, and $\mathbb{E}(Y_t) = 0_{q\times 1}$.

2.2.3 A Master Theorem

We shall start with what we call a "master theorem" that provides non-asymptotic guarantees for lasso estimation and prediction errors under two well-known conditions, viz. the restricted eigenvalue (RE) and the deviation bound (DB) conditions. Note that in the classical linear model setting (see, e.g., Hayashi [2000, Ch 2.3]) where sample size is larger than the dimensions $(n > p)$, the conditions for consistency of the ordinary least squares(OLS) estimator are as follows: (a) the empirical covariance matrix $\mathbf{X}'\mathbf{X}/T \xrightarrow{P} Q$ and Q invertible, i.e., $\lambda_{\min}(Q) > 0$, and (b) the regressors and the noise are asymptotically uncorrelated, i.e., $\mathbf{X}'\mathbf{W}/T \to \mathbf{0}$.

In high-dimensional regimes, Bickel et al. [2009], Loh and Wainwright [2012] and Negahban and Wainwright [2012] have established similar consistency conditions for lasso. The first one is the *restricted eigenvalue* (RE) condition on $\mathbf{X}'\mathbf{X}/T$ (which is a special case, when the loss function is the squared loss, of the *restricted strong convexity* (RSC) condition). The second is the *deviation bound* (DB) condition on $\mathbf{X}'\mathbf{W}/T$. The following lower RE and DB definitions are modified from those given by Loh and Wainwright [2012].

Definition 1 (Lower Restricted Eigenvalue)**.** A symmetric matrix $\Gamma \in \mathbb{R}^{p\times p}$ satisfies a lower restricted eigenvalue condition with curvature $\alpha > 0$ and tolerance $\tau(T, p) > 0$

if,

$$\forall v \in \mathbb{R}^p, \; v'\Gamma v \geq \alpha \left\| v \right\|_2^2 - \tau(T,p) \left\| v \right\|_1^2 .$$

Definition 2 (Deviation Bound)**.** Consider the random matrices $\mathbf{X} \in \mathbb{R}^{T \times p}$ and $\mathbf{W} \in \mathbb{R}^{T \times q}$ defined in (2.2.3) and (2.2.4) above. They are said to satisfy the deviation bound condition if there exist a deterministic multiplier function $\mathbb{Q}(\mathbf{X}, \mathbf{W}, \Theta^\star)$ and a rate of decay function $\mathbb{R}(p, q, T)$ such that,

$$\frac{1}{T} |||\mathbf{X}'\mathbf{W}|||_\infty \leq \mathbb{Q}(\mathbf{X}, \mathbf{W}, \Theta^\star)\mathbb{R}(p, q, T).$$

We now present a master theorem that provides guarantees for the ℓ_2 parameter estimation error and the (in-sample) prediction error. The proof, given in Chapter 2.5 builds on existing result of the same kind [Bickel et al., 2009, Loh and Wainwright, 2012, Negahban and Wainwright, 2012] and we make no claims of originality for either the result or for the proof.

Theorem II.1 (Estimation and Prediction Errors)**.** *Consider the lasso estimator* $\widehat{\Theta}$ *defined in* (2.2.2)*. Suppose Assumption 1 holds. Further, suppose that* $\hat{\Gamma} := \boldsymbol{X}'\boldsymbol{X}/T$ *satisfies the lower RE*(α, τ) *condition with* $\alpha \geq 32 s\tau$ *and* $\boldsymbol{X}'\boldsymbol{W}$ *satisfies the deviation bound. Then, for any* $\lambda_T \geq 4\mathbb{Q}(\boldsymbol{X}, \boldsymbol{W}, \Theta^\star)\mathbb{R}(p, q, T)$*, we have the following guarantees:*

$$\left\| \mathrm{vec}(\widehat{\Theta} - \Theta^\star) \right\| \leq 4\sqrt{s}\lambda_T/\alpha, \tag{2.2.5}$$

$$\left|\left|\left| (\widehat{\Theta} - \Theta^\star)'\hat{\Gamma}(\widehat{\Theta} - \Theta^\star) \right|\right|\right|_F^2 \leq \frac{32\lambda_T^2 s}{\alpha}. \tag{2.2.6}$$

With this master theorem at our disposal, we just need to establish the validity of the restricted eigenvalue (RE) condition and deviation bound (DB) conditions for stationary time series by making appropriate assumptions. We shall do that *without* assuming any parametric form of the data generating mechanism. Instead, we will impose appropriate tail conditions on the random vectors X_t, Y_t and also assume that they satisfy some type of mixing condition. Specifically, in Section 2.3, we consider α-mixing Gaussian random vectors. Next, in Section 2.4, we consider β-mixing subweibull random vectors (we define subweibull random vectors below in Section 2.4.1). Classically, mixing conditions were introduced to generalize classic limit theorems in probability beyond the case of iid random variables [Rosenblatt, 1956]. Recent work on high dimensional statistics has established the validity of RE conditions in the iid Gaussian [Raskutti et al., 2010] and iid subgaussian cases [Rudelson and Zhou, 2013]. One of the main contributions of our work is to extend these results in high dimensional statistics from the iid to the mixing case.

2.2.4 A Brief Overview of Mixing Conditions

Mixing conditions [Bradley, 2005] are well established in the stochastic processes literature as a way to allow for dependence in extending results from the iid case. The general idea is to first define a measure of dependence between two random variables X, Y (that can vector-valued or even take values in a Banach space) with associated sigma algebras $\sigma(X), \sigma(Y)$. For example,

$$\alpha(X,Y) = \sup\{|P(A \cap B) - P(A)P(B)| \ : \ A \in \sigma(X), B \in \sigma(Y)\}.$$

Then for a stationary stochastic process $(X_t)_{t=-\infty}^{\infty}$, one defines the mixing coefficients, for $l \geq 1$,

$$\alpha(l) = \alpha(X_{-\infty:t}, X_{t+l:\infty}).$$

We say that that the process is mixing, in the sense just defined, when $\alpha(l) \to 0$ as $l \to \infty$. The particular notion we get using the α measure of dependence above is called “α-mixing”. It was first used by Rosenblatt [1956] to extend the central limit theorem to dependent random variables. There are other, stronger notions of mixing, such as ρ-mixing and β-mixing that are defined using the dependence measures:

$$\rho(X,Y) = \sup\{\mathrm{Cov}(f(X),g(Y)) \,:\, \mathbb{E}f = \mathbb{E}g = 0, \mathbb{E}f^2 = \mathbb{E}g^2 = 1\}$$

$$\beta(X,Y) = \sup \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J} |P(S_i \cap T_j) - P(S_i)P(T_j)|$$

where the last supremum is over all pairs of partitions $\{A_1,\ldots,A_I\}$ and $\{B_1,\ldots,B_I\}$ of the sample space Ω such that $A_i \in \sigma(X), B_j \in \sigma(Y)$ for all i,j. The ρ-mixing and β-mixing conditions do not imply each other but each, by itself, implies α-mixing [Bradley, 2005]. For stationary gaussian processes, ρ-mixing is equivalent to α-mixing (see Fact 3 below).

The β-mixing condition has been of interest in statistical learning theory for obtaining finite sample generalization error bounds for empirical risk minimization [Vidyasagar, 2003, Sec. 3.4] and boosting [Kulkarni et al., 2005] for dependent samples. There is also work on estimating β-mixing coefficients from data [Mcdonald et al., 2011]. The usefulness of β-mixing lies in the fact that by using a simple blocking technique, that goes back to the work of Yu [1994], one can often reduce the situation to the iid setting. At the same time, many interesting processes such as Markov and hidden Markov processes satisfy a β-mixing condition [Vidyasagar, 2003, Sec. 3.5]. To the best of our knowledge, however, there are no results showing that RE and DB conditions holds under mixing conditions. Next we fill this gap in the literature. Before we continue, we note an elementary but useful fact about mixing conditions, viz. they persist under arbitrary measurable transformations of the original stochastic process.

Fact 1. The range of values that the α, β and ρ-mixing coefficients can take on are bounded(see e.g. Bradley [2005]): Consider the probability space $(\Omega, \mathcal{F}, P)$, for any two sigma fields $\mathcal{A}, \mathcal{B} \in \mathcal{F}$, we have

$$0 \leq \alpha(\mathcal{A}, \mathcal{B}) \leq 1/4, \qquad 0 \leq \beta(\mathcal{A}, \mathcal{B}) \leq 1, \qquad 0 \leq \rho(\mathcal{A}, \mathcal{B}) \leq 1$$

Fact 2. Suppose a stationary process $\{U_t\}_{t=1}^T$ is α, ρ, or β-mixing. Then the stationary sequence $\{f(U_t)\}_{t=1}^T$, for any measurable function $f(\cdot)$, also is mixing in the same sense with its mixing coefficients bounded by those of the original sequence.

2.3 Gaussian Processes under α-Mixing

Here we will study Gaussian processes under the α-mixing condition which is weaker than that of the β-mixing. We make the following additional assumption.

Assumption 4 (Gaussianity)**.** The process (X_t, Y_t) is a Gaussian process.

Assume $(X_t, Y_t)_{t=1}^T$ satisfies Assumptions 2, 3, and 4. Note that $X_t \sim \mathcal{N}(0, \Sigma_X)$ and $Y_t \sim \mathcal{N}(0, \Sigma_Y)$. To control dependence over time, we will assume α-mixing, the weakest notion among α, ρ and β-mixing.

Assumption 5 (α-Mixing)**.** The process (X_t, Y_t) is an α-mixing process. Let $S_\alpha(T) := \sum_{l=0}^T \alpha(l)$. If $\alpha(l)$ is summable, we let $\tilde{\alpha} := \lim_{T\to\infty} S_\alpha(T) < \infty$.

We will use the following useful fact [Ibragimov and Rozanov, 1978, p. 111] in our analysis.

Fact 3. For any stationary Gaussian process, the α and ρ-mixing coefficients are related as follows:

$$\forall l \geq 1, \ \alpha(l) \leq \rho(l) \leq 2\pi\alpha(l).$$

Proposition II.2 (Deviation Bound, Gaussian Case)**.** *Suppose Assumptions 2–5 hold. Then, there exists a deterministic positive constant $\tilde{c}$, and a free parameter*

$b > 0$, such that, for $T \geq \sqrt{\frac{b+1}{\tilde{c}}} \log(pq)$, we have

$$\mathbb{P}\left[\left|\!\left|\!\left|\frac{\boldsymbol{X}'\boldsymbol{W}}{T}\right|\!\right|\!\right|_\infty \leq \mathbb{Q}(\boldsymbol{X},\, \boldsymbol{W}, \Theta^\star)\mathbb{R}(p,q,T)\right] \geq 1 - 8\exp(-b\log(pq))$$

where

$$\mathbb{Q}(\boldsymbol{X},\, \boldsymbol{W}, \Theta^\star) = 8\pi\sqrt{\frac{(b+1)}{\tilde{c}}}\left(|\!|\!|\Sigma_X|\!|\!|\left(1 + \max_{1\leq i\leq p}\|\Theta^\star_{:i}\|_2^2\right) + |\!|\!|\Sigma_Y|\!|\!|\right)$$
$$\mathbb{R}(p,q,T) = S_\alpha(T)\sqrt{\frac{\log(pq)}{T}}.$$

Remark 1. Note that the free parameter b serves to trade-off between the success probability on the one hand and the sample size threshold and multiplier function $\mathbb{Q}$ on the other. A large b increases the success probability but worsen the sample size threshold and the multiplier function.

Proposition II.3 (RE, Gaussian Case)**.** *Suppose Assumptions 2–5 hold. There exists some universal constant $c > 0$, such that for sample size $T \geq \frac{42e\log(p)}{c\min\{1,\eta^2\}}$, we have, with probability at least $1-2\exp\left(-\frac{c}{2}T\min\{1,\eta^2\}\right)$ that for every vector $v \in \mathbb{R}^p$,*

$$|v'\hat{\Gamma}v| > \alpha\|v\|_2^2 - \tau(T,p)\|v\|_1^2, \tag{2.3.1}$$

where

$$\alpha = \frac{1}{2}\lambda_{\min}(\Sigma_X), \qquad \tau(T,p) = \alpha/\lceil c\frac{T}{4\log(p)}\min\{1,\eta^2\}\rceil, \qquad \textit{and}$$
$$\eta = \frac{\lambda_{\min}(\Sigma_X)}{108\pi S_\alpha(T)\lambda_{\max}(\Sigma_X)}.$$

Remark 2. Note that, in Theorem II.1, it is advantageous to have a large α and a smaller τ so that the convergence rate is fast and the initial sample threshold for the result to hold is small. The result above, therefore, clearly shows that is advantageous to have a well-conditioned Σ_X.

2.3.1 Estimation and Prediction Errors

Substituting the RE and DB constants from Propositions II.2-II.3 into Theorem II.1 immediately yields the following guarantees.

Corollary II.4 (Lasso Guarantees for Gaussian Vectors under α-Mixing)**.** *Suppose Assumptions 2–5 hold. Let $c, \tilde{c}$ be fixed constants and b be free parameter defined as in Propositions II.2 and II.3. Then, for sample size*

$$\begin{aligned} T &\geq \max\left\{\frac{\log(p)}{c\min\{1,\eta^2\}}\max\left\{42e, 128s\right\}, \log(pq)\sqrt{\frac{b+1}{\tilde{c}}}\right\} \\ \textit{where } \eta &= \frac{\lambda_{\min}(\Sigma_X)}{108\pi S_\alpha(T)\lambda_{\max}(\Sigma_X)} \end{aligned}$$

we have, with probability at least $1 - 2\exp\left(-\frac{c}{2}T\min\{1,\eta^2\}\right) - 8\exp(-b\log(pq))$*, that the lasso error bounds* (2.2.5) *and* (2.2.6) *hold with*

$$\begin{aligned} \alpha &= \frac{1}{2}\lambda_{\min}(\Sigma_X) \\ \lambda_T &= 4\mathbb{Q}(\boldsymbol{X}, \boldsymbol{W}, \Theta^\star)\mathbb{R}(p,q,T) \end{aligned}$$

where

$$\begin{aligned} \mathbb{Q}(\boldsymbol{X}, \boldsymbol{W}, \Theta^\star) &= 8\pi\sqrt{\frac{(b+1)}{\tilde{c}}}\left(|||\Sigma_X|||\left(1 + \max_{1\leq i\leq p}\|\Theta^\star_{:i}\|_2^2\right) + |||\Sigma_Y|||\right), \\ \mathbb{R}(p,q,T) &= S_\alpha(T)\sqrt{\frac{\log(pq)}{T}}. \end{aligned}$$

Remark 3. If the α-mixing coefficients are summable, i.e., $S_\alpha(T) \leq \tilde{\alpha} < \infty$, $\forall T$, then we get the usual convergence rate of $O(\sqrt{\frac{\log(pq)}{T}})$. Also, the threshold sample size is $O\left(s\log(pq)\right)$. This is in agreement with what is happens in the iid Gaussian case. When $\alpha(l)$ is not summable then both the initial sample threshold required for the guarantee to be valid as well as the rate of error decay deteriorate. The latter becomes $O(\sqrt{\frac{S_\alpha(T)\log(pq)}{T}})$. We see that as long as $S_\alpha(T) \in o\left(\sqrt{T}\right)$, we still have consistency. In the finite order stable Gaussian VAR case considered by Basu and Michailidis [2015], the α-mixing coefficients are geometrically decaying and hence summable (see Example 1 for details).

2.3.2 Examples

We illustrate applicability of our theory in Section 2.3 using the examples below.

Example 1 (Gaussian VAR)**.** Transition matrix estimation in sparse stable VAR models has been considered by several authors in recent years [Davis et al., 2015, Han and Liu, 2013, Song and Bickel, 2011]. The lasso estimator is a natural choice for the problem.

Formally a finite order Gaussian VAR(d) process is defined as follows. Consider a sequence of serially ordered random vectors $(Z_t)_{t=1}^{T+d}$, $Z_t \in \mathbb{R}^p$ that admits the following auto-regressive representation:

$$Z_t = \mathbf{A}_1 Z_{t-1} + \cdots + \mathbf{A}_d Z_{t-d} + \mathcal{E}_t \tag{2.3.2}$$

where each $\mathbf{A}_k, k = 1, \ldots, d$ is a sparse non-stochastic coefficient matrix in $\mathbb{R}^{p \times p}$ and innovations $\mathcal{E}_t$ are p-dimensional random vectors from $\mathcal{N}(0, \Sigma_\epsilon)$ with $\lambda_{\min}(\Sigma_\epsilon) > 0$ and $\lambda_{\max}(\Sigma_\epsilon) < \infty$.

Assume that the VAR(d) process is *stable*; i.e. $\det\left(\mathbf{I}_{p\times p} - \sum_{k=1}^d \mathbf{A}_k z^k\right) \neq 0, \forall\, |z| \leq 1$. Now, we identify $X_t := (Z_t', \cdots, Z_{t-d+1}')'$ and $Y_t := Z_{t+d}$ for $t = 1, \ldots, T$.

We can verify (see Section 2.8.1 for details) that Assumptions 1–5 hold. Note that $\Theta^\star = (\mathbf{A}_1, \ldots, \mathbf{A}_d)' \in \mathbb{R}^{dp \times p}$. As a result, Propositions II.2 and II.3, and thus Corollary II.4 follow and hence we have all the high probabilistic guarantees for lasso on Example 1. This shows that our theory covers the stable Gaussian VAR models for which Basu and Michailidis [2015] provided lasso errors bounds.

We state the following convenient fact because it allows us to study any finite order VAR model by considering its equivalent VAR(1) representation. See Section 2.8.1 for details.

Fact 4. Every VAR(d) process can be written in VAR(1) form (see e.g. [Lütkepohl, 2005, Ch 2.1]).

Therefore, without loss of generality, we can consider VAR(1) model in the ensuing Examples.

Example 2 (Gaussian VAR with Omitted Variable)**.** We study OLS estimator of a VAR(1) process when there are endogenous variables omitted. This arises naturally when the underlying DGM is high-dimensional but not all variables are available/observable/measurable to the researcher to do estimation/prediction. This also happens when the researcher mis-specifies the scope of the model.

Notice that the system of the retained set of variables is no longer a finite order VAR(and thus non-Markovian). There is model mis-specification and this example also serves to illustrate that our theory is applicable to models beyond the finite order VAR setting.

Consider a VAR(1) process $(Z_t, \Xi_t)_{t=1}^{T+1}$ such that each vector in the sequence is generated by the recursion below:

$$(Z_t; \Xi_t) = \mathbf{A}(Z_{t-1}; \Xi_{t-1}) + (\mathcal{E}_{Z,t-1}; \mathcal{E}_{\Xi,t-1})$$

where $Z_t \in \mathbb{R}^p$, $\Xi_t \in \mathbb{R}$, $\mathcal{E}_{Z,t} \in \mathbb{R}^p$, and $\mathcal{E}_{\Xi,t} \in \mathbb{R}$ are partitions of the random vectors (Z_t, Ξ_t) and $\mathcal{E}_t$ into p and 1 variables. Also,

$$\mathbf{A} := \begin{bmatrix} \mathbf{A}_{ZZ} & \mathbf{A}_{Z\Xi} \\ \mathbf{A}_{\Xi Z} & \mathbf{A}_{\Xi\Xi} \end{bmatrix}$$

is the coefficient matrix of the VAR(1) process with $\mathbf{A}_{Z\Xi}$ 1-sparse, $\mathbf{A}_{ZZ}$ p-sparse and $r(\mathbf{A}) < 1$. $\mathcal{E}_t := (\mathcal{E}_{X,t-1}; \mathcal{E}_{Z,t-1})$ for $t = 1, \ldots, T+1$ are iid draws from a Gaussian white noise process.

We are interested in the OLS 1-lag estimator of the system restricted to the set of variables in Z_t. Recall that

$$\Theta^\star := \operatorname*{arg\,min}_{\mathbf{B} \in \mathbb{R}^{p \times p}} \mathbb{E}\left(\|Z_t - \mathbf{B}' Z_{t-1}\|_2^2 \right)$$

Now, set $X_t := Z_t$ and $Y_t := Z_{t+1}$ for $t = 1, \ldots, T$. It can be shown that $(\Theta^\star)' = \mathbf{A}_{ZZ} + \mathbf{A}_{Z\Xi}\Sigma_{\Xi Z}(0)(\Sigma_Z)^{-1}$. We can verify that Assumptions 1–5 hold. See Section 2.8.2 for details. As a result, Propositions II.2 and II.3, and thus Corollary II.4 follow and hence we have all the high probabilistic guarantees for lasso on this non-Markovian example.

2.4 Subweibull Random Vectors under β-Mixing

Existing analyses of lasso mostly assume data have subgaussian or subexponential tails. These assumptions ensure that the moment generating function exists, at least for some values of the free parameter. Non-existence of the moment generating function is often taken as the definition of having a heavy tail [Foss et al., 2011]. We now introduce a family of random variables that subsumes subgaussian and subexponential random variables. In addition, it includes some heavy tailed distributions.

2.4.1 Subweibull Random Variables and Vectors

Among the several equivalent definitions of the subgaussian and subexponential random variables, we recall the ones that are based on the growth behavior of moments. Recall that a subgaussian (resp. subexponential) random variable X can be defined as one for which $\mathbb{E}(|X|^p)^{1/p} \leq K\sqrt{p}$, $\forall p \geq 1$ for some constant K (resp. $\mathbb{E}(|X|^p)^{1/p} \leq Kp$, $\forall p \geq 1$). A natural generalization of these definitions that allows for heavier tails is as follows. Fix some $\gamma > 0$, and require

$$\|X\|_p := (\mathbb{E}|X|^p)^{1/p} \leq Kp^{1/\gamma},\ \forall p \geq 1 \wedge \gamma$$

There are a few different equivalent ways to imposing the condition above including a tail condition that says that the tail is no heavier than that of a Weibull random variable with parameter γ. That is the reason why we call this family "subweibull(γ)".

Lemma II.5. *(Subweibull properties) Let X be a random variable. Then the following statements are equivalent for every $\gamma > 0$. The constants K_1, K_2, K_3 differ from each other at most by a constant depending only on γ.*

1. *The tails of X satisfies*

$$\mathbb{P}\left(|X| > t\right) \leq 2\exp\left\{-(t/K_1)^\gamma\right\},\ \forall t \geq 0.$$

2. *The moments of X satisfy,*

$$\|X\|_p := (\mathbb{E}|X|^p)^{1/p} \leq K_2 p^{1/\gamma},\ \forall p \geq 1 \wedge \gamma.$$

3. *The moment generating function of $|X|^\gamma$ is finite at some point; namely*

$$\mathbb{E}\left[\exp\left(|X|/K_3\right)^\gamma\right] \leq 2.$$

Remark 4. A similar tail condition is called “Condition C0” by Tao and Vu [2013]. However, to the best of our knowledge, this family has not been systematically introduced. The equivalence above is related to the theory of Orlicz spaces (see, for example, Lemma 3.1 in the lecture notes of Pisier [2016]).

Definition 3. (Subweibull(γ) Random Variable and Norm). A random variable X that satisfies any property in Lemma II.5 is called a subweibull(γ) random variable. The subweibull(γ) norm associated with X, denoted $\|X\|_{\psi_\gamma}$, is defined to be the smallest constant such that the moment condition in definition Lemma II.5 holds. In other words, for every $\gamma > 0$,

$$\|X\|_{\psi_\gamma} := \sup_{p \geq 1} (\mathbb{E}|X|^p)^{1/p} p^{-1/\gamma}.$$

It is easy to see that $\|\cdot\|_{\psi_\gamma}$, being a pointwise supremum of norms, is indeed a norm on the space of subweibull(γ) random variables.

Remark 5. It is common in the literature (see, for example Foss et al. [2011]) to call a random variable *heavy-tailed* if its tail decays slower than that of an exponential random variable. This way of distinguishing between light and heavy tails is natural because the moment generating function for a heavy-tailed random variable thus defined fails to exist at any point. Note that, under such a definition, subweibull(γ) random variables with $\gamma < 1$ include heavy-tailed random variables.

In our theoretical analysis, we will often be dealing with squares of random variables. The next lemma tells us what happens to the subweibull parameter γ and the associated constant, under squaring.

Lemma II.6. *For any $\gamma \in (0, \infty)$, if a random variable X is subweibull(2γ) then X^2 is subweibull(γ). Moreover,*

$$\|X^2\|_{\psi_\gamma} \leq 2^{1/\gamma} \|X\|_{\psi_{2\gamma}}^2.$$

We now define the subweibull norm of a random vector to capture dependence among its coordinates. It is defined using one dimensional projections of the random vector in the same way as we define subgaussian and subexponential norms of random vectors.

Definition 4. Let $\gamma \in (0, \infty)$. A random vector $X \in \mathbb{R}^p$ is said to be a subweibull(γ) random vector if all of its one dimensional projections are subweibull(γ) random variables. We define the subweibull(γ) norm of a random vector as,

$$\|X\|_{\psi_\gamma} := \sup_{v \in S^{p-1}} \|v'X\|_{\psi_\gamma}$$

where S^{p-1} is the unit sphere in $\mathbb{R}^p$.

Having introduced the subweibull family, we present the assumptions required for the lasso guarantees. In proving our results, we need measures that control the amount of dependence in the observations across time as well as within a given time period.

Assumption 6. The process (X_t, Y_t) is geometrically β-mixing; i.e., there exist constants $c > 0$ and $\gamma_1 > 0$ such that

$$\beta(n) \leq 2\exp(-c \cdot n^{\gamma_1}), \ \forall n \in \mathbb{N}.$$

Assumption 7. Each random vector in the sequences (X_t) and (Y_t) follows a subweibull(γ_2) distribution with $\|X_t\|_{\psi_{\gamma_2}} \leq K_X$, $\|Y_t\|_{\psi_{\gamma_2}} < K_Y$ for $t = 1, \cdots, T$.

Finally, we make an joint assumption on the allowed pairs γ_1, γ_2.

Assumption 8. Assume $\gamma < 1$ where

$$\gamma := \left(\frac{1}{\gamma_1} + \frac{2}{\gamma_2}\right)^{-1}.$$

Remark 6. Note that the parameters γ_1 and γ_2 defines a difficulty landscape with smaller values of γ_1, γ_2 corresponding to harder problems. The "easy case" where $\gamma_1 \geq 1$ and $\gamma_2 \geq 2$ are already addressed in the literature (see, e.g., Wong et al. [2017]). This chapter serves to provide theoretical guarantees for the difficult scenario when the tail probability decays slowly ($\gamma_2 < 2$) and/or data exhibit strong temporal dependence ($\gamma_1 < 1$) and hence extends the literature to the entire spectrum of possibilities, i.e., all positive values of γ_1 and γ_2.

Now, we are ready to provide high probability guarantees for the deviation bound and restricted eigenvalue conditions.

Proposition II.7 (Deviation Bound, β-Mixing Subweibull Case)**.** *Suppose Assumptions 1-3 and 6-8 hold. Let $c' > 0$ be a universal constant and let K be defined as*

$$K := 2^{2/\gamma_2} \left(K_Y + K_X \left(1 + \|\!|\Theta^\star|\!\|\right)\right)^2.$$

Then with sample size $T \geq C_1(\log(pq))^{\frac{2}{\gamma}-1}$, we have

$$\mathbb{P}\left(\frac{1}{T}\|\!|\boldsymbol{X}'\boldsymbol{W}|\!\|_\infty > C_2 K\sqrt{\frac{\log(pq)}{T}}\right) \leq 2\exp(-c'\log(pq))$$

where the constants C_1, C_2 depend only on c' and the parameters γ_1, γ_2, c appearing in Assumptions 6 and 7.

Proposition II.8 (RE, β-Mixing Subweibull Case)**.** *Suppose Assumptions 1-3 and 6-8 hold. Let*

$$K := 2^{2/\gamma_2} K_X^2.$$

Then for sample size

$$T \geq \max\left\{ \frac{54K\left(2C_1\log(p)\right)^{1/\gamma}}{\lambda_{\min}(\Sigma_X)}, \left(\frac{54K}{\lambda_{\min}(\Sigma_X)}\right)^{\frac{2-\gamma}{1-\gamma}}\left(\frac{C_2}{C_1}\right)^{\frac{1}{1-\gamma}}\right\}$$

we have with probability at least

$$1 - 2T\exp\left\{-\tilde{c}T^{\gamma}\right\}, \ \textit{where } \tilde{c} = \frac{(\lambda_{\min}(\Sigma_X))^{\gamma}}{(54K)^{\gamma}2C_1},$$

that for all $v \in \mathbb{R}^p$,

$$\frac{1}{T}\left\|\boldsymbol{X}v\right\|_2^2 \geq \alpha\left\|v\right\|_2^2 - \tau\left\|v\right\|_1^2.$$

where $\alpha = \frac{1}{2}\lambda_{\min}(\Sigma_X)$ and $\tau = \frac{\alpha}{2\tilde{c}} \cdot \left(\frac{\log(p)}{T^{\gamma}}\right)$. Note that the constants C_1, C_2 depend only on the parameters γ_1, γ_2, c appearing in Assumptions 6 and 7.

2.4.2 Estimation and Prediction Errors

Substituting the RE and DB constants from Propositions II.7-II.8 into Theorem II.1 immediately yields the following guarantee.

Corollary II.9 (Lasso Guarantees for Subweibull Vectors under β-Mixing)**.** *Suppose Assumptions 1-3 and 6-8 hold. Let $c', C_1, C_2, \tilde{c}$ be constants as defined in Propositions II.7-II.8, and let $K := 2^{2/\gamma_2}\left(K_Y + K_X\left(1 + \|\!|\Theta^{\star}|\!\|\right)\right)^2$.*

Then for sample size

$$T \geq \max\Bigg\{ C_1(\log(pq))^{\frac{2}{\gamma}-1}, \\ \frac{54K\left[2\max\{8s/\tilde{c}, C_1\}\log(p)\right]^{1/\gamma}}{\lambda_{\min}(\Sigma_X)}, \left(\frac{54K}{\lambda_{\min}(\Sigma_X)}\right)^{\frac{2-\gamma}{1-\gamma}}\left(\frac{C_2}{C_1}\right)^{\frac{1}{1-\gamma}}\Bigg\}$$

we have with probability at least

$$1 - 2T\exp\left\{-\tilde{c}T^{\gamma}\right\} - 2\exp(-c'\log(pq))$$

that the lasso error bounds (2.2.5) *and* (2.2.6) *hold with*

$$\alpha = \frac{1}{2}\lambda_{\min}(\Sigma_X)$$

$$\lambda_T = 4\mathbb{Q}(\boldsymbol{X}, \boldsymbol{W}, \Theta^\star)\mathbb{R}(p, q, T)$$

where

$$\mathbb{Q}(\boldsymbol{X}, \boldsymbol{W}, \Theta^\star) = C_2 K,$$

$$\mathbb{R}(p, q, T) = \sqrt{\frac{\log(pq)}{T}}.$$

Remark 7. The impact of mixing behavior is limited to the initial sample size and the probability with which the error bounds hold. The parameter error bound itself resembles the bounds obtained in the iid case but with an additional multiplicative factor that depends on the 'effective condition number" $K/\lambda_{\min}(\Sigma_X)$.

2.4.3 Examples

We explore applicability of our theory in Section 2.4 beyond just linear Gaussian processes using the examples below. Together, these demonstrate that the high probabilistic guarantees for lasso cover cases of heavy tailed subweibull data, presence of model mis-specification, and/or nonlinearity.

Example 3 (Subweibull VAR)**.** We study a generalization of the VAR, one that has subweibull(γ_2) realizations. Consider a VAR(1) model defined as in Example 1 except that we replace the Gaussian white noise innovations with iid random vectors from some subweibull(γ_2) distribution with a non-singular covariance matrix Σ_ϵ. Now, consider a sequence $(Z_t)_t$ generated according to the model. Then, each Z_t will be a mean zero subweibull random vector.

Now, we identify $X_t := (Z_t', \cdots, Z_{t-d+1}')'$ and $Y_t := Z_{t+d}$ for $t = 1, \ldots, T$. Assuming that $\mathbf{A}_i$'s are sparse, $|||\mathbf{A}||| < 1$ and $(Z_t)_t$ is stable, we can verify (see Section 2.8.1

for details) that Assumptions 1-3 and 6-8 hold. Note that $\Theta^\star = (\mathbf{A}_1, \ldots, \mathbf{A}_d)' \in \mathbb{R}^{dp \times p}$. As a result, Propositions II.7 and II.8 follow and hence we have all the high probability guarantees for lasso on Example 3. This shows that our theory covers DGMs beyond just the stable Gaussian processes.

Example 4 (VAR with Subweibull Innovations and Omitted Variable)**.** Using the same setup as in Example 2 except that we replace the Gaussian white noise innovations with iid random vectors from some subweibull(γ_2) distribution with a non-singular covariance matrix Σ_ϵ. Now, consider a sequence $(Z_t)_t$ generated according to the model. Then, each Z_t will be a mean zero subweibull random vector.

Now, set $X_t := Z_t$ and $Y_t := Z_{t+1}$ for $t = 1, \ldots, T$. Assume $|||\mathbf{A}||| < 1$. It can be shown that $(\Theta^\star)' = \mathbf{A}_{ZZ} + \mathbf{A}_{Z\Xi}\Sigma_{\Xi Z}(0)(\Sigma_Z)^{-1}$. We can verify that Assumptions 1-3 and 6-7 hold. See Section 2.8.2 for details. Therefore, Propositions II.7 and II.8 and thus Corollary II.9 follow and and hence we have all the high probabilistic guarantees for subweibull random vectors from a non-Markovian model.

Example 5 (Multivariate ARCH)**.** We explore the generality of our theory by considering a multivariate nonlinear time series model with subweibull innovations. A popular nonlinear multivariate time series model in econometrics and finance is the vector autoregressive conditionally heteroscedastic (ARCH) model. We choose the following specific ARCH model just for convenient validation of the geometric β-mixing property of the process; it may potentially be applicable to a larger class of multivariate ARCH models.

Let $(Z_t)_{t=1}^{T+1}$ be random vectors defined by the following recursion, for any constants $c > 0$, $m \in (0, 1)$, $a > 0$, and $\mathbf{A}$ sparse with $|||\mathbf{A}||| < 1$:

$$\begin{aligned} Z_t &= \mathbf{A} Z_{t-1} + \Sigma(Z_{t-1}) \mathcal{E}_t \\ \Sigma(z) &:= c \cdot \operatorname{clip}_{a,b}\left(\|z\|^m\right) \mathbf{I}_{p \times p} \end{aligned} \tag{2.4.1}$$

where $\mathcal{E}_t$ are iid random vectors from some subweibull(γ_2) distribution with a non-singular covariance matrix Σ_ϵ, and $\operatorname{clip}_{a,b}(x)$ clips the argument x to stay in the interval $[a, b]$. Consequently, each Z_t will be a mean zero subweibull random vector. Note that $\Theta^* = \mathbf{A}'$, the transpose of the coefficient matrix $\mathbf{A}$ here.

Now, set $X_t := Z_t$ and $Y_t = Z_{t+1}$ for $t = 1, \ldots, T$. We can verify (see Section 2.8.3 for details) that Assumptions 1-3 and 6-7 hold. Therefore, Propositions II.7 and II.8, and thus Corollary II.9 follow and and hence we have all the high probabilistic guarantees for lasso on nonlinear models with subweibull innovations.

2.5 Proof of Master Theorem

Proof of Theorem II.1. We will break down the proof in steps.

1. Since $\widehat{\Theta}$ is optimal for 2.2.2 and $\Theta^\star$ is feasible,
$$\frac{1}{T}\left|\left|\left| Y - \mathbf{X}\widehat{\Theta} \right|\right|\right|_F^2 + \lambda_T \left\| \operatorname{vec}(\widehat{\Theta}) \right\|_1 \leq \frac{1}{T}|||Y - \mathbf{X}\Theta^\star|||_F^2 + \lambda_T \left\|\operatorname{vec}(\Theta^\star)\right\|_1$$

2. Let $\hat{\Delta} := \widehat{\Theta} - \Theta^\star \in \mathbb{R}^{p \times q}$
$$\frac{1}{T}\left|\left|\left| \mathbf{X}\hat{\Delta} \right|\right|\right|_F^2 \leq \frac{2}{T}\operatorname{tr}(\hat{\Delta}'\mathbf{X}'\mathbf{W}) + \lambda_T \left(\|\operatorname{vec}(\Theta^\star)\|_1 - \left\|\operatorname{vec}(\widehat{\Theta})\right\|_1 \right)$$
Note that
$$\begin{aligned} \left\|\operatorname{vec}(\Theta^\star + \hat{\Delta})\right\|_1 - \|\operatorname{vec}(\Theta^\star)\|_1 \geq & \{\|\operatorname{vec}(\Theta_S^\star)\|_1 - \left\|\operatorname{vec}(\hat{\Delta}_S)\right\|_1\} \\ & + \left\|\operatorname{vec}(\hat{\Delta}_{S^c})\right\|_1 - \|\operatorname{vec}(\Theta^\star)\|_1 \\ = & \left\|\operatorname{vec}(\hat{\Delta}_{S^c})\right\|_1 - \left\|\operatorname{vec}(\hat{\Delta}_S)\right\|_1 \end{aligned}$$

where S denote the support of $\Theta^\star$.

3. With RE constant α and tolerance τ, deviation bound constant $\mathbb{Q}(\Sigma_X, \Sigma_W)$ and $\lambda_T \geq 2\mathbb{Q}(\Sigma_X, \Sigma_W)\sqrt{\frac{\log(q)}{T}}$, we have

$$\begin{aligned}
\alpha\left|\!\left|\!\left|\hat{\Delta}\right|\!\right|\!\right|_F^2 - \tau\|\operatorname{vec}(\hat{\Delta})\|_1^2 & \\
&\overset{RE}{\leq} \frac{1}{T}|\!|\!|\mathbf{X}\Delta|\!|\!|_F^2 \\
&\leq \frac{2}{T}\mathrm{tr}(\hat{\Delta}'\mathbf{X}'\mathbf{W}) + \lambda_T\{\left\|\operatorname{vec}(\hat{\Delta}_S)\right\|_1 - \left\|\operatorname{vec}(\hat{\Delta}_{S^c})\right\|_1\} \\
&\leq \frac{2}{T}\sum_{k=1}^{q}\|\hat{\Delta}_{:k}\|_1\|(\mathbf{X}'\mathbf{W})_{:k}\|_\infty + \lambda_T\{\left\|\operatorname{vec}(\hat{\Delta}_S)\right\|_1 - \left\|\operatorname{vec}(\hat{\Delta}_{S^c})\right\|_1\} \\
&\leq \frac{2}{T}\|\operatorname{vec}(\hat{\Delta})\|_1|\!|\!|\mathbf{X}'\mathbf{W}|\!|\!|_\infty + \lambda_T\{\left\|\operatorname{vec}(\hat{\Delta}_S)\right\|_1 - \left\|\operatorname{vec}(\hat{\Delta}_{S^c})\right\|_1\} \\
&\overset{DB}{\leq} 2\|\operatorname{vec}(\hat{\Delta})\|_1\mathbb{Q}(\Sigma_X, \Sigma_W)\mathbb{R}(p, q, T) + \lambda_T\{\left\|\operatorname{vec}(\hat{\Delta}_S)\right\|_1 - \left\|\operatorname{vec}(\hat{\Delta}_{S^c})\right\|_1\} \\
&\leq \|\operatorname{vec}(\hat{\Delta})\|_1\lambda_N/2 + \lambda_T\{\left\|\operatorname{vec}(\hat{\Delta}_S)\right\|_1 - \left\|\operatorname{vec}(\hat{\Delta}_{S^c})\right\|_1\} \\
&\leq \frac{3\lambda_T}{2}\left\|\operatorname{vec}(\hat{\Delta}_S)\right\|_1 - \frac{\lambda_T}{2}\left\|\operatorname{vec}(\hat{\Delta}_{S^c})\right\|_1 \\
&\leq 2\lambda_T\left\|\operatorname{vec}(\hat{\Delta})\right\|_1
\end{aligned}$$

4. In particular, this says that $3\left\|\operatorname{vec}(\hat{\Delta}_S)\right\|_1 \geq \left\|\operatorname{vec}(\hat{\Delta}_{S^c})\right\|_1$
   So $\left\|\operatorname{vec}(\hat{\Delta})\right\|_1 \leq 4\left\|\operatorname{vec}(\hat{\Delta}_S)\right\|_1 \leq 4\sqrt{s}\left\|\operatorname{vec}(\hat{\Delta})\right\|$

5. Finally, with $\alpha \geq 32s\tau$,

$$\begin{aligned}
\frac{\alpha}{2}\left\|\operatorname{vec}(\hat{\Delta})\right\|_F^2 &\leq (\alpha - 16s\tau)\left\|\operatorname{vec}(\hat{\Delta})\right\|_F^2 \\
&\leq \alpha\left\|\operatorname{vec}(\hat{\Delta})\right\|_F^2 - \tau\|\operatorname{vec}(\hat{\Delta})\|_1^2 \\
&\leq 2\lambda_T\|\operatorname{vec}(\hat{\Delta})\|_1 \\
&\leq 2\sqrt{s}\lambda_T\|\hat{\Delta}\|_F
\end{aligned}$$

6.

$$\left\|\operatorname{vec}(\hat{\Delta})\right\|_F \leq \frac{4\lambda_T\sqrt{s}}{\alpha}$$

7. From step 4, we have

$$\frac{1}{T}\left|\!\left|\!\left|\mathbf{X}\hat{\Delta}\right|\!\right|\!\right|_F^2 \leq 8\lambda_T\sqrt{s}\left\|\text{vec}(\hat{\Delta})\right\|$$

Then, from step 6

$$\frac{1}{T}\left|\!\left|\!\left|\mathbf{X}\hat{\Delta}\right|\!\right|\!\right|_F^2 \leq 8\lambda_T\sqrt{s}\left\|\text{vec}(\hat{\Delta})\right\| \leq 32\lambda_T^2 s/\alpha$$

□

2.6 Proofs for Gaussian Processes under α-Mixing

We will also need the following result to control spectral/operator norms of matrices.

Fact 5 (Schur Test). For any matrix $\mathbf{M}$, we have

$$|\!|\!|\mathbf{M}|\!|\!|^2 \leq \max_i \left\|\mathbf{M}_{i:}\right\|_1 \cdot \max_j \left\|\mathbf{M}_{:j}\right\|_1 .$$

Therefore, for any *symmetric* matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$, $|\!|\!|\mathbf{M}|\!|\!| \leq \max_{1\leq i\leq n}\left\|\mathbf{M}_{i:}\right\|_1$.

Claim 1. For any random vectors $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^n$, we have

$$|\!|\!|\mathbb{E}\left[XY'\right]|\!|\!| = |\!|\!|\mathbb{E}\left[YX'\right]|\!|\!| \leq \frac{|\!|\!|\Sigma_X|\!|\!| + |\!|\!|\Sigma_Y|\!|\!|}{2}$$

Proof. We have,

$$\begin{aligned}
|\!|\!|\mathbb{E}\left[XY'\right]|\!|\!| = |\!|\!|\mathbb{E}\left[YX'\right]|\!|\!| & \\
&:= \sup_{\|u\|\leq 1,\, \|v\|\leq 1} \mathbb{E}\left[u'YX'v\right] \\
&= \sup_{\|u\|\leq 1,\, \|v\|\leq 1} \mathbb{E}\left[(Y'u)(X'v)\right] \\
&\leq \sup_{\|u\|\leq 1,\, \|v\|\leq 1} \sqrt{\mathbb{E}\left[(Y'u)^2\right]}\sqrt{\mathbb{E}\left[(X'v)^2\right]} \qquad \text{by CauchySchwarz ineq.} \\
&= \sup_{\|u\|\leq 1} \sqrt{\mathbb{E}\left[(Y'u)^2\right]} \sup_{\|v\|\leq 1} \sqrt{\mathbb{E}\left[(X'v)^2\right]} \\
&= \sqrt{|\!|\!|\mathbb{E}\left[XX'\right]|\!|\!|}\sqrt{|\!|\!|\mathbb{E}\left[YY'\right]|\!|\!|} \\
&\leq \frac{|\!|\!|\mathbb{E}\left[XX'\right]|\!|\!| + |\!|\!|\mathbb{E}\left[YY'\right]|\!|\!|}{2}.
\end{aligned}$$

□

The proof of Proposition II.3 relies on the following result.

Lemma II.10. *For a second order stationary ρ-mixing sequence of random vectors $\{X_t\}_t$, their l-th auto-covariance matrix can be bounded as follows:*

$$|||\Sigma_X(l)||| \leq \rho(l)|||\Sigma_X(0)|||, \, \forall\, l \in \mathbb{Z}.$$

Proof. Recall the definition of ρ-mixing, for random vectors X and Y on the probability space $(\Omega, \mathcal{F}, \mathcal{P})$, let $\mathcal{A} := \sigma(X)$ and $\mathcal{B} := \sigma(Y)$.

$$\begin{aligned}\rho(X,Y) &:= \sup\{\mathrm{cor}(f(X), g(Y)) \mid f \in \mathcal{L}^2(\mathcal{A}), g \in \mathcal{L}^2(\mathcal{B})\} \\ &\geq \mathrm{cor}(u'X,\, v'Y) && \forall \text{ fixed } u,\, v \\ &= \frac{|\mathbb{E}[u'Xv'Y]|}{\sqrt{\mathbb{E}(u'X)^2\mathbb{E}(v'Y)^2}} && u,\, v \,\text{non-zero}\end{aligned}$$

Hence, $\forall u,\, v$ fixed,

$$\begin{aligned}|u'\mathbb{E}[XY']v| &\leq \rho(X,Y)\sqrt{\mathbb{E}(u'X)^2}\sqrt{\mathbb{E}(v'Y)^2} \\ \sup_{u,v}|u'\mathbb{E}[XY']v| &\leq \rho(X,Y)\sqrt{\sup_u \mathbb{E}(u'X)^2}\sqrt{\sup_v \mathbb{E}(v'Y)^2}\end{aligned}$$

But,

$$|||\mathbb{E}[XY']||| \equiv \sup_{u,v}|u'\mathbb{E}[XY']v| \leq \rho(X,Y)\sqrt{\sup_u \mathbb{E}(u'X)^2}\sqrt{\sup_v \mathbb{E}(v'Y)^2}$$

For a stationary time series $\{X_t\}$, recall that $\forall t, l$

$$\Sigma_X(l) := \mathbb{E}[X_t X_{t+l}'].$$

By stationarity, $\forall t, l$

$$\rho(X_t, X_{t+l}) = \rho(l).$$

Hence,

$$|||\Sigma_X(l)||| \leq \rho(l)|||\Sigma_X(0)|||.$$

□

Proof of Proposition II.2. Note that by Fact 2 α and ρ-mixing are equivalent for stationary Gaussian processes. The proof will operate via arguments in ρ-mixing coefficients.

Recall $|||\mathbf{X}'\mathbf{W}|||_\infty = \max_{1\leq i\leq p, 1\leq j\leq q} |[\mathbf{X}'\mathbf{W}]_{i,j}| = \max_{1\leq i\leq p, 1\leq j\leq q} |\mathbf{X}'_{:i}\mathbf{W}_{:j}|$.

By Assumption (3), we have

$$\begin{aligned}&\mathbb{E}\mathbf{X}_{:i} = 0, \forall i \quad \text{and}\\ &\mathbb{E}\mathbf{Y}_{:j} = 0, \forall j\end{aligned}$$

By first order optimality of the optimization problem in (2.2.1), we have

$$\mathbb{E}\mathbf{X}'_{:i}(\mathbf{Y} - \mathbf{X}\Theta^\star) = 0, \forall i \Rightarrow \mathbb{E}\mathbf{X}_{:i}{}'\mathbf{W}_{:j} = 0, \forall i, j$$

We know $\forall i, j$

(2.6.1)

$$\begin{aligned}|\mathbf{X}'_{:i}\mathbf{W}_{:j}| &= |\mathbf{X}'_{:i}\mathbf{W}_{:j} - \mathbb{E}[\mathbf{X}'_{:i}\mathbf{W}_{:j}]|\\ &= \frac{1}{2}\left|\left(\|\mathbf{X}_{:i} + \mathbf{W}_{:j}\|^2 - \mathbb{E}[\|\mathbf{X}_{:i} + \mathbf{W}_{:j}\|^2]\right) - \left(\|\mathbf{X}_{:i}\|^2 - \mathbb{E}[\|\mathbf{X}_{:i}\|^2]\right) - \left(\|\mathbf{W}_{:j}\|^2 - \mathbb{E}[\|\mathbf{W}_{:j}\|^2]\right)\right|\\ &\leq \frac{1}{2}\left|\|\mathbf{X}_{:i} + \mathbf{W}_{:j}\|^2 - \mathbb{E}[\|\mathbf{X}_{:i} + \mathbf{W}_{:j}\|^2]\right| + \frac{1}{2}\left|\|\mathbf{X}_{:i}\|^2 - \mathbb{E}[\|\mathbf{X}_{:i}\|^2]\right| + \frac{1}{2}\left|\|\mathbf{W}_{:j}\|^2 - \mathbb{E}[\|\mathbf{W}_{:j}\|^2]\right|\end{aligned}$$

Therefore,

$$
\begin{aligned}
&\mathbb{P}\left(\frac{1}{T}\left|\mathbf{X}_{:i}'\mathbf{W}_{:j}\right| > 3t\right) \\
&\leq \mathbb{P}\left(\frac{1}{2T}\left|\|\mathbf{X}_{:i}+\mathbf{W}_{:j}\|^2 - \mathbb{E}[\|\mathbf{X}_{:i}+\mathbf{W}_{:j}\|^2]\right| > t\right) + \mathbb{P}\left(\frac{1}{2T}\left|\|\mathbf{X}_{:i}\|^2 - \mathbb{E}[\|\mathbf{X}_{:i}\|^2]\right| > t\right) \\
&\quad + \mathbb{P}\left(\frac{1}{2T}\left|\|\mathbf{W}_{:j}\|^2 - \mathbb{E}[\|\mathbf{W}_{:j}\|^2]\right| > t\right)
\end{aligned}
$$

This suggests proof strategy via controlling tail probability on each of the terms $|\|\mathbf{X}_{:i}\|^2 - \mathbb{E}[\|\mathbf{X}_{:i}\|^2]|$, $|\|\mathbf{W}_{:j}\|^2 - \mathbb{E}[\|\mathbf{W}_{:j}\|^2]|$ and $|\|\mathbf{X}_{:i}+\mathbf{W}_{:j}\|^2 - \mathbb{E}[\|\mathbf{X}_{:i}+\mathbf{W}_{:j}\|^2]|$. Assuming the conditions in Proposition II.2, we can apply the Hanson-Wright inequality (Lemma II.11) on each of them because we know that

$\forall i, j$

$$\mathbf{X}_{:i} \sim N(0, \Sigma_{\mathbf{X}_{:i}}),\ i = 1, \cdots, p,\ \text{and}$$

$$\mathbf{W}_{:j} := \mathbf{Y}_{:j} - [\mathbf{X}\Theta^\star]_{:j} \sim N(0, \Sigma_{\mathbf{W}_{:j}}),\ j = 1, \cdots, q$$

since both $\{X_t\}_{t=1}^T$ and $\{Y_t\}_{t=1}^T$ are centered Gaussian vectors.

So,

$$\mathbf{X}_{:i} + \mathbf{W}_{:j} \sim N(0, \Sigma_{\mathbf{X}_{:i}} + \Sigma_{\mathbf{W}_{:j}} + \Sigma_{\mathbf{W}_{:j},\mathbf{X}_{:i}} + \Sigma_{\mathbf{X}_{:i},\mathbf{W}_{:j}})$$

We are ready to apply the tail bound on each term on the RHS of (2.6.1). By Lemma (II.11), $\exists$ constant $c > 0$ such that $\forall t \geq 0$

$$
\begin{aligned}
\frac{\|\mathbf{X}_{:i}\|^2 - \mathbb{E}\|\mathbf{X}_{:i}\|^2}{T|||\Sigma_{\mathbf{X}_{:i}}|||} &\leq t \qquad \text{w.p. at least } 1 - 2\exp(-cT\min\{t, t^2\}) \\
\frac{\|\mathbf{W}_{:j}\|^2 - \mathbb{E}\|\mathbf{W}_{:j}\|^2}{T|||\Sigma_{\mathbf{W}_{:j}}|||} &\leq t \qquad \text{w.p. at least } 1 - 2\exp(-cT\min\{t, t^2\}) \\
\frac{\|\mathbf{X}_{:i}+\mathbf{W}_{:j}\|^2 - \mathbb{E}\|\mathbf{X}_{:i}+\mathbf{W}_{:j}\|^2}{T|||\Sigma_{\mathbf{X}_{:i}+\mathbf{W}_{:j}}|||} &\leq t \qquad \text{w.p. at least } 1 - 2\exp(-cT\min\{t, t^2\})
\end{aligned}
$$

With Claim 1, the third inequality implies, for some $\tilde{c} > 0$, that w.p. at least

$$1 - 8\exp(-\tilde{c}T\min\{t, t^2\})$$

the following holds

$$\frac{\|\mathbf{X}_{:i} + \mathbf{W}_{:j}\|^2 - \mathbb{E}\|\mathbf{X}_{:i} + \mathbf{W}_{:j}\|^2}{2T(|||\Sigma_{\mathbf{X}_{:i}}||| + |||\Sigma_{\mathbf{W}_{:j}}|||)} \leq \frac{\|\mathbf{X}_{:i} + \mathbf{W}_{:j}\|^2 - \mathbb{E}\|\mathbf{X}_{:i} + \mathbf{W}_{:j}\|^2}{T|||\Sigma_{\mathbf{X}_{:i}+\mathbf{W}_{:j}}|||} \leq t$$

Therefore,

$$\frac{\mathbf{X}'_{:i}\mathbf{W}_{:j}}{3T(|||\Sigma_{\mathbf{X}_{:i}}||| + |||\Sigma_{\mathbf{W}_{:j}}|||)} \leq 3t \text{ w.p. at least } 1 - 8\exp(-\tilde{c}T\min\{t, t^2\})$$

Appealing to the union bound over all $i \in [1 \cdots p]$ and $j \in [1 \cdots q]$, for any Δ

$$\mathbb{P}[\max_{1\leq i\leq p, 1\leq j\leq q} \mathbf{X}'_{:i}\mathbf{W}_{:j} \geq \Delta] \leq pq\mathbb{P}[\mathbf{X}'_{:i}\mathbf{W}_{:j} \geq \Delta]$$

We can conclude that

$$\mathbb{P}[\max_{1\leq i\leq p, 1\leq j\leq q} \frac{\mathbf{X}'_{:i}\mathbf{W}_{:j}}{3T(|||\Sigma_{\mathbf{X}_{:i}}||| + |||\Sigma_{\mathbf{W}_{:j}}|||)} \leq 3t] \geq 1 - 8pq\exp(-\tilde{c}T\min\{t, t^2\})$$

Now, for a free parameter $b > 0$, choose$t = \sqrt{\frac{(b+1)\log(pq)}{\tilde{c}T}}$, for $T \geq \frac{(b+1)\log(pq)}{\tilde{c}}$ we have

$$\mathbb{P}\left[\left|\left|\left|\frac{\mathbf{X}'\mathbf{W}}{T}\right|\right|\right|_\infty \leq \sqrt{\frac{(b+1)\log(pq)}{\tilde{c}T}} \max_{1\leq i\leq p, 1\leq j\leq q}(|||\Sigma_{\mathbf{X}_{:i}}||| + |||\Sigma_{\mathbf{W}_{:j}}|||)\right]$$
$$\geq 1 - 8\exp[-b\log(pq)]$$

Let's find out what $|||\Sigma_{\mathbf{W}_{:j}}|||$ and $|||\Sigma_{\mathbf{X}_{:i}}|||$ are. Recall $\mathbf{W} = \mathbf{Y} - \mathbf{X}\Theta^\star$. So,

$$\Sigma_{\mathbf{W}_{:i}} = \Sigma_{\mathbf{Y}_{:i}} + \Sigma_{\mathbf{X}\Theta^\star_{:i}} + \Sigma_{\mathbf{Y}_{:i}, \mathbf{X}\Theta^\star_{:i}} + \Sigma_{\mathbf{X}\Theta^\star_{:i}, \mathbf{Y}_{:i}} \tag{2.6.2}$$

By Claim 1, we have

$$|||\Sigma_{\mathbf{W}_{:i}}||| \leq 2|||\Sigma_{\mathbf{Y}_{:i}}||| + 2|||\Sigma_{\mathbf{X}\Theta^\star_{:i}}||| \tag{2.6.3}$$

Let's figure out each of the summands on the RHS of equation (2.6.3) above.

$$\begin{aligned}
\Sigma_{\mathbf{X}\Theta^{\star}_{:i}}[l,k] &:= \mathbb{E}\left[(\mathbf{X}\Theta^{\star}_{:i})(\mathbf{X}\Theta^{\star}_{:i})'\right][l,k] \\
&= \mathbb{E}\left[e_l'(\mathbf{X}\Theta^{\star}_{:i})(\mathbf{X}\Theta^{\star}_{:i})'e_k\right] \\
&= \mathbb{E}\left[(\mathbf{X}_{l,:}'\Theta^{\star}_{:i})((\Theta^{\star}_{:i})'\mathbf{X}_{k,:})\right] \\
&= \mathbb{E}\left[(\Theta^{\star}_{:i})'\mathbf{X}_{l,:}\mathbf{X}_{k,:}'\Theta^{\star}_{:i}\right] \\
&= (\Theta^{\star}_{:i})'\left[\mathbb{E}\mathbf{X}_{l,:}\mathbf{X}_{k,:}'\right]\Theta^{\star}_{:i}
\end{aligned}$$

With the equality above,

$$\begin{aligned}
\left|\left|\left|\Sigma_{\mathbf{X}\Theta^{\star}_{:i}}\right|\right|\right| &\le \max_{1\le k\le T}\left\|(\Sigma_{\mathbf{X}\Theta^{\star}_{:i}})[k,:]\right\|_1 && \text{by Fact 5} \\
&\le 2\sum_{l=0}^{T}\rho(l)\,\|\Theta^{\star}_{:i}\|_2^2\,|||\Sigma_X(0)||| && \text{by Lemma II.10}
\end{aligned}$$

Therefore,

$$\max_{1\le i\le p}\left|\left|\left|\Sigma_{\mathbf{X}\Theta^{\star}_{:i}}\right|\right|\right| \le \max_{1\le i\le p} 2\sum_{l=0}^{T}\rho(l)\,\|\Theta^{\star}_{:i}\|_2^2\,|||\Sigma_X(0)||| = 2|||\Sigma_X(0)|||\sum_{l=0}^{T}\rho(l)\max_{1\le i\le p}\|\Theta^{\star}_{:i}\|_2^2$$

Similarly,

$$\max_{1\le i\le p}|||\Sigma_{\mathbf{Y}_{:i}}||| \le 2|||\Sigma_Y(0)|||\sum_{l=0}^{T}\rho(l)$$

and

$$\max_{1\le i\le p}|||\Sigma_{\mathbf{X}_{:i}}||| \le 2|||\Sigma_X(0)|||\sum_{l=0}^{T}\rho(l)$$

So,. by inequality (2.6.3)

$$\max_{1\le i\le q}|||\Sigma_{W_{:i}}||| \le 4\sum_{l=0}^{T}\rho(l)\left(|||\Sigma_X|||\max_{1\le i\le p}\|\Theta^{\star}_{:i}\|_2^2 + |||\Sigma_Y|||\right)$$

Therefore,

$$\max_{1\le i\le p,1\le j\le q}\left(|||\Sigma_{\mathbf{X}_{:i}}||| + \left|\left|\left|\Sigma_{\mathbf{W}_{:j}}\right|\right|\right|\right) \le 4\sum_{l=0}^{T}\rho(l)\left(|||\Sigma_X|||\left(1+\max_{1\le i\le p}\|\Theta^{\star}_{:i}\|_2^2\right) + |||\Sigma_Y|||\right)$$

Finally, to state the final result:

For a free parameter $b > 0$, choose$t = \sqrt{\frac{(b+1)\log(pq)}{\tilde{c}T}}$, for $T \geq \frac{(b+1)\log(pq)}{\tilde{c}}$ we have with probability at least

$$1 - 8\exp[-b\log(pq)]$$

that

$$\left\vert\!\left\vert\!\left\vert \frac{\mathbf{X}'\mathbf{W}}{T} \right\vert\!\right\vert\!\right\vert_{\infty} \leq \sqrt{\frac{(b+1)\log(pq)}{\tilde{c}T}} 4\sum_{l=0}^{T} \rho(l) \left(|||\Sigma_X||| \left(1 + \max_{1\leq i\leq p} \|\Theta^{\star}_{:i}\|_2^2 \right) + |||\Sigma_Y||| \right)$$

Also, because of Fact 3, we have

$$\left\vert\!\left\vert\!\left\vert \frac{\mathbf{X}'\mathbf{W}}{T} \right\vert\!\right\vert\!\right\vert_{\infty} \leq \sqrt{\frac{(b+1)\log(pq)}{\tilde{c}T}} 8\pi S_{\alpha}(T) \left(|||\Sigma_X||| \left(1 + \max_{1\leq i\leq p} \|\Theta^{\star}_{:i}\|_2^2 \right) + |||\Sigma_Y||| \right)$$

□

Proof of Proposition II.3. Note that, by Fact 3, α and ρ-mixing are equivalent for stationary Gaussian processes. The proof will operate via arguments involving ρ-mixing coefficients.

For a fixed unit test vector $v \in \mathbb{R}^p$, $\|v\|_2 = 1$, consider the Gaussian vector $\mathbf{X}v \in \mathbb{R}^T$. To apply the Hanson-Wright inequality (Lemma (II.11)), we have to upper bound the operator norm of the covariance matrix $\mathbf{Q}$ of $\mathbf{X}v$.

$\mathbf{Q}$ takes the form

$$\mathbf{Q} = \begin{bmatrix} v'\mathbb{E}X_1X_1'v & \cdots & v'\mathbb{E}X_1X_j'v & \cdots & v'\mathbb{E}X_1X_T'v \\ \vdots & \ddots & & & \vdots \\ v'\mathbb{E}X_tX_1'v & & v'\mathbb{E}X_tX_t'v & & v'\mathbb{E}X_tX_T'v \\ \vdots & & & \ddots & \vdots \\ v'\mathbb{E}X_TX_1'v & \cdots & v'\mathbb{E}X_TX_1'v & \cdots & v'\mathbb{E}X_TX_T'v \end{bmatrix}$$

We can thus use Fact 5 and Lemma II.10 to upper bound $|||\mathbf{Q}|||$ by

$$\sum_{t=0}^{T} \rho(l) |||\Sigma_X(0)|||$$

Now, we can apply Lemma II.11 on any fixed unit test vector $v \in \mathbb{R}^p$, $\|v\|_2 = 1$. Recall $\hat{\Gamma} := \frac{\mathbf{X'X}}{T} \in \mathbb{R}^{p\times p}$. Using Lemma II.11, we have, $\forall \eta > 0$

$$\mathbb{P}[|v'(\hat{\Gamma} - \Sigma_X(0))v > \eta|||\mathbf{Q}|||\,|] \leq 2\exp\{-cT\min(\eta, \eta^2)\} \Rightarrow$$

$$\mathbb{P}[v'(\hat{\Gamma} - \Sigma_X(0))v > \eta \sum_{t=0}^{T} \rho(l) |||\Sigma_X(0)|||] \leq 2\exp\{-cT\min(\eta, \eta^2)\}$$

Using Lemma F.2 in Basu and Michailidis [2015], for any integer $k > 0$, we extend it to all vectors in $\mathbb{J}(2k) := \{v \in \mathbb{R}^p : \|v\| \leq 1, \|v\|_0 \leq 2k\}$:

$$\mathbb{P}\left[\sup_{v\in\mathbb{J}(2k)} |v'(\hat{\Gamma} - \Sigma_X(0))v| > \eta \sum_{t=0}^{T} \rho(l) |||\Sigma_X(0)|||\right]$$
$$\leq 2\exp\{-cT\min\{\eta, \eta^2\} + 2k\min\{\log(p), \log(\frac{21ep}{2k}))\}$$

By Lemma 12 in Loh and Wainwright [2012], we further extend the bound to all $\forall v \in \mathbb{R}^p$,

$$|v'(\hat{\Gamma} - \Sigma_X(0))v| > 27\eta \sum_{t=0}^{T} \rho(l) |||\Sigma_X(0)||| \left(\|v\|_2^2 + \frac{1}{k}\|v\|_1^2 \right)$$

$$\text{w.p.} \leq 2\exp\{-cT\min(\eta, \eta^2) + 2k\min(\log(p), \log(\frac{21ep}{2k}))\}$$

$$\Updownarrow$$

$$|v'(\hat{\Gamma} - \Sigma_X(0))v| \leq 27\eta \sum_{t=0}^{T} \rho(l) |||\Sigma_X(0)||| \left(\|v\|_2^2 + \frac{1}{k}\|v\|_1^2 \right)$$

$$\text{w.p.} > 1 - 2\exp\{-cT\min(\eta, \eta^2) + 2k\min(\log(p), \log(\frac{21ep}{2k}))\}$$

$$\Downarrow$$

$$|v'(\hat{\Gamma})v| > -27\eta \sum_{t=0}^{T} \rho(l) |||\Sigma_X(0)||| \left(\|v\|_2^2 + \frac{1}{k}\|v\|_1^2 \right) + \lambda_{\min}(\Sigma_X(0))\|v\|_2^2$$

$$\text{w.p.} > 1 - 2\exp\{-cT\min(\eta, \eta^2) + 2k\min(\log(p), \log(\frac{21ep}{2k}))\}$$

Intuitively, we know the quadratic form of a Hermitian matrix should have its magnitude bounded from below by its minimum eigenvalue. To achieve that, pick $\eta = \frac{\lambda_{\min}(\Sigma_X(0))}{54\sum_{t=0}^{T}\rho(l)\lambda_{\max}(\Sigma_X(0))}$. So, we have

$$|v'\hat{\Gamma}v| > \frac{1}{2}\lambda_{\min}(\Sigma_X(0))\|v\|_2^2 - \frac{\lambda_{\min}(\Sigma_X(0))}{2k}\|v\|_1^2$$

w.p.

$$\geq 1 - 2\exp\{-cT\min(1, \eta^2) + 2k\min(\log(p), \log(\frac{21ep}{2k}))\}$$

because $\min(1, \eta^2) \leq \min(\eta, \eta^2)$.

Now, we choose k to make sure the first component in the exponential dominates. For now, assume $p \geq \frac{21ep}{2k}$. Let $k = \lceil c\frac{T}{4\log(p)}\min\{1, \eta^2\}\rceil$. Now, choose T such that $k \geq \frac{21e}{2}$. Let $T \geq \frac{42e\log(p)}{c\min\{1,\eta^2\}}$, where s is the sparsity.

Finally, we have, for $T \geq s\frac{42e\log(p)}{c\min\{1,\eta^2\}}$, with probability at least

$$1 - 2\exp\{-T\frac{c}{2}\min\{1,\eta^2\}\}$$

the following holds

$$|v'\hat{\Gamma}v| > \frac{1}{2}\lambda_{\min}(\Sigma_X(0))\|v\|_2^2 - \frac{\lambda_{\min}(\Sigma_X(0))}{2k}\|v\|_1^2$$

Also, let $\tilde{\eta} := \frac{\lambda_{\min}(\Sigma_X(0))}{108\pi S_\alpha(T)\lambda_{\max}(\Sigma_X(0))}$ we can bound η with $\tilde{\eta}$ by Fact 3. □

2.6.1 Hanson-Wright Inequality

The general statement of the Hanson-Wright inequality can be found in the paper by [Rudelson and Vershynin, 2013, Theorem 1.1]. We use a form of the inequality which is derived in the proof of Proposition 2.4 of Basu and Michailidis [2015] as an easy consequence of the general result. We state the modified form of the inequality and the proof below for completeness.

Lemma II.11 (Variant of Hanson-Wright Inequality). *If $Y \sim N(0_{n\times 1}, \boldsymbol{Q}_{n\times n})$, then there exists universal constant $c > 0$ such that for any $\eta > 0$,*

$$\mathbb{P}\left[\frac{1}{n}\left|\|Y\|_2^2 - \mathbb{E}\|Y\|_2^2\right| > \eta|||\boldsymbol{Q}|||\right] \leq 2\exp\left[-cn\min\left\{\eta,\eta^2\right\}\right]. \tag{2.6.4}$$

Proof. The lemma easily follows from Theorem 1.1 in Rudelson and Vershynin [2013]. Write $Y = \mathbf{Q}^{1/2}X$, where $X \sim \mathcal{N}(0,\mathbf{I})$ and $(\mathbf{Q}^{1/2})'(\mathbf{Q}^{1/2}) = \mathbf{Q}$. Note that each component X_i of X is independent $\mathcal{N}(0,1)$, so that $\|X_i\|_{\psi_2} \leq 1$. Then, by the above theorem,

$$\begin{aligned}\mathbb{P}\left[\frac{1}{n}\left|\|Y\|_2^2 - \mathrm{Tr}(\mathbf{Q})\right| > \eta|||\mathbf{Q}|||\right] &= \mathbb{P}\left[\frac{1}{n}\left|X'\mathbf{Q}X - \mathbb{E}X'\mathbf{Q}X\right| > \eta|||\mathbf{Q}|||\right] \\ &\leq 2\exp\left[-c\min\left\{\frac{n^2\eta^2|||\mathbf{Q}|||}{|||\mathbf{Q}|||_F^2}, \frac{n\eta|||\mathbf{Q}|||}{|||\mathbf{Q}|||}\right\}\right] \\ &\leq 2\exp\left[-c\min\left\{\eta,\eta^2\right\}\right] \qquad \text{since } |||\mathbf{Q}|||_F^2 \leq n|||\mathbf{Q}|||^2\end{aligned}$$

Lastly, note that $\mathrm{Tr}(\mathbf{Q}) = \mathrm{Tr}(\mathbb{E}YY') = \mathbb{E}\,\mathrm{Tr}(YY') = \mathbb{E}\,\mathrm{Tr}(Y'Y) = \mathbb{E}\,\mathrm{Tr}\,\|Y\|^2 = \mathbb{E}\,\|Y\|^2$. □

2.7 Proofs for Subweibull Random Vectors under β-Mixing

2.7.1 Proof Related to Subweibull Properties

Proof. (of Lemma II.5) *Property 1 $\Rightarrow$ Property 2:* Since we can scale X by K_1, without loss of generality, we can assume $K_1 = 1$. Then we have, for $p \geq \gamma$,

$$\begin{aligned}
\mathbb{E}\,|X|^p &= \int_0^\infty \mathbb{P}\left(|X|^p \geq u\right) du \\
&= \int_0^\infty \mathbb{P}\left(|X| \geq t\right) pt^{p-1} dt && \text{using change of variable } u = t^p \\
&\leq \int_0^\infty 2e^{-t^\gamma} pt^{p-1} dt && \text{by Property 1} \\
&= \frac{2p}{\gamma} \int_0^\infty e^{-v} \cdot v^{\frac{p-1}{\gamma}} v^{\frac{1-\gamma}{\gamma}} dv && \text{using change of variable } v = t^\gamma \\
&= \frac{2p}{\gamma} \int_0^\infty e^{-v} \cdot v^{p/\gamma - 1} dv \\
&= \frac{2p}{\gamma} \cdot \Gamma\left(\frac{p}{\gamma}\right) \\
&\leq \frac{2p}{\gamma} \left(\frac{p}{\gamma}\right)^{p/\gamma} && \text{since } \Gamma(x) \leq x^x,\ \forall x \geq 1.
\end{aligned}$$

Therefore, for $p \geq \gamma$,

$$\left(\mathbb{E}\,|X|^p\right)^{1/p} \leq 2^{1/p} (1/\gamma)^{1/p} p^{1/p} \left(p/\gamma\right)^{1/\gamma} \leq C_\gamma \cdot p^{1/\gamma}$$

where $C_\gamma = 4(1/\gamma \vee 1)(1/\gamma)^{1/\gamma}$. If $\gamma \leq 1$, this covers all $p \geq 1$. If $\gamma > 1$, we have, for $p = 1, \ldots, \lceil \gamma \rceil - 1$,

$$\left(\mathbb{E}\,|X|^p\right)^{1/p} \leq 2^{1/p} (1/\gamma)^{1/p} p^{1/p} \max_{i=1,\ldots,\lceil \gamma \rceil - 1} \Gamma(i/\gamma)^{1/i} \leq C'_\gamma,$$

where $C'_\gamma = 4(1/\gamma \vee 1) \max_{i=1,\ldots,\lceil \gamma \rceil - 1} \Gamma(i/\gamma)^{1/i}$. Therefore, for all p,

$$\left(\mathbb{E}\,|X|^p\right)^{1/p} \leq (C_\gamma \vee C'_\gamma) \cdot p^{1/\gamma}.$$

Property 2 ⇒ Property 3: Without loss of generality, we can assume $K_2 = 1$. Using Taylor series expansion of $\exp(\cdot)$, for some positive λ,

$$\begin{aligned}
\mathbb{E}\exp\left[(\lambda\left|X\right|)^{\gamma_2}\right] &= \mathbb{E}\left[1+\sum_{p=1}^{\infty}\frac{\mathbb{E}\left[((\lambda\left|X\right|)^{\gamma})^p\right]}{p!}\right] \\
&\le 1+\sum_{p=1}^{\infty}\frac{(\lambda^{\gamma}\gamma p)^p}{(p/e)^p} && \text{by Property 2 and Stirling's approx.} \\
&= \sum_{p=0}^{\infty}(e\gamma\lambda^{\gamma})^p = \frac{1}{1-e\gamma\lambda^{\gamma}} \le 2,
\end{aligned}$$

where the last inequality holds for any λ satisfying $e\gamma\lambda^{\gamma} \le 1/2$, i.e., $\lambda \le (2e\gamma)^{-1/\gamma}$. Therefore Property 2 holds with $K_3 = (2e\gamma)^{1/\gamma}$.

Property 3 ⇒ Property 1: Without loss of generality, we can assume $K_3 = 1$. For all $t > 0$,

$$\begin{aligned}
\mathbb{P}\left(\left|X\right| > t\right) &= \mathbb{P}\left(\exp\left(\left|X\right|^{\gamma_2}\right) \ge \exp\left(t^{\gamma_2}\right)\right) \\
&\le \exp\left(-\left(t^{\gamma_2}\right)\right)\mathbb{E}\exp\left(\left|X\right|^{\gamma_2}\right) && \text{by Markov's inequality} \\
&\le 2\exp\left(-\left(t^{\gamma_2}\right)\right) && \text{by Property 3}
\end{aligned}$$

□

Proof. (of Lemma II.6) By definition,

$$\begin{aligned}
\left\|X^2\right\|_{\psi_\gamma} &= \sup_{p\ge 1} p^{-1/\gamma}\left(\mathbb{E}\left|X^2\right|^p\right)^{1/p} \\
&= \sup_{p\ge 1}\left(p^{-1/(2\gamma)}\left(\mathbb{E}\left|X\right|^{2p}\right)^{1/2p}\right)^2
\end{aligned}$$

Now we make a change of variables $\tilde{p} := 2p$. Then, we have,

$$\begin{aligned}
\left\|X^2\right\|_{\psi_\gamma} &= 2^{1/\gamma} \sup_{\tilde{p}\geq 2} \left(\tilde{p}^{-1/(2\gamma)} \left(\mathbb{E}\left|X\right|^{\tilde{p}}\right)^{1/\tilde{p}}\right)^2 \\
&\leq 2^{1/\gamma} \sup_{\tilde{p}\geq 1} \left(\tilde{p}^{-1/(2\gamma)} \left(\mathbb{E}\left|X\right|^{\tilde{p}}\right)^{1/\tilde{p}}\right)^2 \\
&= 2^{1/\gamma} \left(\sup_{\tilde{p}\geq 1} \tilde{p}^{-1/(2\gamma)} \left(\mathbb{E}\left|X\right|^{\tilde{p}}\right)^{1/\tilde{p}}\right)^2 \\
&= 2^{1/\gamma} \left\|X\right\|_{\psi_{2\gamma}}^2 . \quad \square
\end{aligned}$$

2.7.2 Subweibull Norm Under Linear Transformations

We will need the following result about changes to the subweibull norm under linear transformations.

Lemma II.12. *Let X be a random vector and $\boldsymbol{A}$ be a fixed matrix. We have,*

$$\left\|\boldsymbol{A}X\right\|_{\psi_\gamma} \leq \left|\!\left|\!\left|\boldsymbol{A}\right|\!\right|\!\right| \cdot \left\|X\right\|_{\psi_\gamma}$$

Proof. We have,

$$\begin{aligned}
\left\|\mathbf{A}X\right\|_{\psi_\gamma} &= \sup_{\|v\|_2\leq 1} \left\|v'\mathbf{A}X\right\|_{\psi_\gamma} = \sup_{\|v\|_2\leq 1} \left\|(\mathbf{A}'v)'X\right\|_{\psi_\gamma} \\
&\leq \sup_{\|u\|_2\leq |\!|\!|\mathbf{A}|\!|\!|} \left\|u'X\right\|_{\psi_\gamma} \\
&= |\!|\!|\mathbf{A}|\!|\!| \sup_{\|u\|_2\leq 1} \left\|u'X\right\|_{\psi_\gamma} = |\!|\!|\mathbf{A}|\!|\!| \left\|X\right\|_{\psi_\gamma} .
\end{aligned}$$

□

2.7.3 Concentration Inequality for Sums of β-Mixing Subweibull Random Variables

We will state and prove a modified form of Theorem 1 of Merlevède et al. [2011]. This concentration result will be used to prove the high probability guarantees on the deviation bound (Lemma II.7) and lower restricted eigenvalue (Lemma II.8) conditions.

Lemma II.13. *Let $(X_j)_{j=1}^T$ be a strictly stationary sequence of zero mean random variables that are subweibull(γ_2) with subweibull constant K. Denote their sum by S_T. Suppose their β-mixing coefficients satisfy $\beta(n) \leq 2\exp(-cn^{\gamma_1})$. Let γ be a parameter given by*

$$\frac{1}{\gamma} = \frac{1}{\gamma_1} + \frac{1}{\gamma_2}.$$

Further assume $\gamma < 1$. Then for $T > 4$, and any $t > 1/T$,

$$\mathbb{P}\left\{\left|\frac{S_T}{T}\right| > t\right\} \leq T\exp\left\{-\frac{(tT)^\gamma}{K^\gamma C_1}\right\} + \exp\left\{-\frac{t^2 T}{K^2 C_2}\right\} \tag{2.7.1}$$

where the constants C_1, C_2 depend only on γ_1, γ_2 and c.

Proof. Note that, in this proof, constants $C, C_1, C_2, \ldots$ can depend on c, γ_1 and γ_2 and C_1, C_2 in the proof are not the same as the eventual constants C_1, C_2 that appear in the lemma statement.

Further, we will assume that $K = 1$. The general form then follows by scaling the random variables by $1/K$ and applying the lemma with t replaced by t/K. The proof consists of two parts. First, we will state a concentration inequality of Merlevède et al. [2011] and bound a certain parameter V appearing in their inequality using the β-mixing assumption. Second, we will simplify the expression that we get directly from their concentration inequality to get a more convenient form.

Step 1: Controlling the V parameter using β-mixing coefficients First, recall that Theorem 1 of Merlevède et al. [2011], under the condition of our lemma, gives

$$\begin{aligned} \mathbb{P}\left\{|S_T| > u\right\} \leq T\exp\left\{-\frac{u^\gamma}{C_1}\right\} + \exp\left\{-\frac{u^2}{C_2(1+TV)}\right\} \\ + \exp\left\{-\frac{u^2}{C_3 T}\exp\left\{\frac{1}{C_4}\left(\frac{u^{(1-\gamma)}}{\log(u)}\right)^\gamma\right\}\right\} \end{aligned} \tag{2.7.2}$$

First of all, we need to control the quantity V that appears in the denominator of the second term of (2.7.2). V is a worst case measure of the partial sum of the auto-covariances on the clipped dependent sequence $(X_t)_{t=1}^T$. It is increasing in time horizon T and related to dimension p and sparsity s and hence not an absolute constant. To the best of our knowledge, V is not controllable under the weaker α-mixing condition. As Merlevède et al. [2011] mention in their Section 2.1.1, using results of Viennet [1997], we have, for any β-mixing strictly stationary sequence (Y_t) with geometrically decaying β-mixing coefficients; i.e.,

$$\beta(k) \leq 2\exp\left\{-ck^{\gamma_1}\right\} \text{ for any positive } k$$

the associated quantity V can be upper bounded as

$$V \leq \mathbb{E}X_1^2 + 4\sum_{k\geq 0}\mathbb{E}(B_k X_1^2)$$

for some sequence (B_k) with values in $[0,1]$ satisfying $\mathbb{E}(B_k) \leq \beta(k)$. In our case, (X_t) is stationary and we know that its finite moments exist because of Assumption 7. Then,

$$\begin{aligned}
V &\leq \mathbb{E}X_1^2 + 4\sum_{k\geq 0}\mathbb{E}(B_k X_1^2) \\
&\leq \mathbb{E}X_1^2 + 4\sum_{k\geq 0}\sqrt{\mathbb{E}(B_k^2)\mathbb{E}(X_1^4)} && \text{Cauchy-Schwarz ineqequality} \\
&= \mathbb{E}X_1^2 + 4\sqrt{\mathbb{E}(X_1^4)}\sum_{k\geq 0}\sqrt{\mathbb{E}(B_k^2)} && \text{all finite moments of } X_1 \text{ exist} \\
&\leq \mathbb{E}X_1^2 + 4\sqrt{\mathbb{E}(X_1^4)}\sum_{k\geq 0}\sqrt{\mathbb{E}(B_k)} \\
&\leq C,
\end{aligned}$$

where the second to last inequality follows because $B_k \in [0,1] \Rightarrow B_k^2 \leq B_k$. The last inequality comes from the fact that $\sqrt{\mathbb{E}(B_k)} \leq \sqrt{\beta(k)} \leq \sqrt{2}\exp\left\{-\frac{1}{2}ck^{\gamma_1}\right\} \Rightarrow$

$(\sqrt{\mathbb{E}(B_k)})$ summable. Moreover since X_1 is subweibull(γ_2) with constant 1, both $\mathbb{E}X_1^2$ and $\mathbb{E}(X_1^4)$ are bounded with constants depending only on γ_2. Note that C depends on c, γ_1 and γ_2.

Step 2: Deriving a Convenient form Eventually we will apply the concentration inequality above with $u = tT$, and we will choose t such that $u = tT > 1$. Under the condition that $u > 1$, we will now show that the term appearing in the exponent in the third term in (2.7.2),

$$\frac{(u)^{(1-\gamma)}}{(\log(u))}, \tag{2.7.3}$$

is larger than a γ-dependent constant. Along with the fact that V is a constant in the second term, the second and third terms in (2.7.2) can then be combined into one.

Let $u > 1$. Note that the expression (2.7.3) remains positive and blows up to infinity as u approaches 1 from above. Taking derivative with respect to u, we obtain

$$\frac{d}{du}\frac{u^{(1-\gamma)}}{(\log(u))} = \frac{u^{-\gamma}}{\log(u)}\left[(1-\gamma) - \frac{1}{\log(u)}\right]$$

Observe that the derivative is negative when $u < u^* = e^{\frac{1}{1-\gamma}}$; for $u > u^*$, it becomes positive again. Hence, the expression (2.7.3) reaches its minimum at u^*, where its value is,

$$\frac{\left(e^{\frac{1}{1-\gamma}}\right)^{1-\gamma}}{\frac{1}{1-\gamma}} = e(1-\gamma),$$

which is positive since $\gamma < 1$. □

2.7.4 Proofs of Deviation and RE Bounds

Proof. (of Proposition II.7) Note that constants $C_1, C_2, \ldots$ can change from line to line and depend only on γ_1, γ_2, c appearing in Assumption 6 and Assumption 7, and on the constant c' appearing in the high probability guarantee.

Recall that $\mathbf{W} \; := \mathbf{Y} - \mathbf{X}\Theta^\star$, and

$$\begin{aligned} |||\mathbf{X}'\mathbf{W}|||_\infty &= \max_{1\le i\le p, 1\le j\le q} |[\mathbf{X}'\mathbf{W}]_{i,j}| \\ &= \max_{1\le i\le p, 1\le j\le q} |(\mathbf{X}_{:i})'\mathbf{W}_{:j}| \end{aligned}$$

By Assumption 3, we have

$$\begin{aligned} &\mathbb{E}\mathbf{X}_{:i} = 0, \forall i = 1, \cdots, p \quad \text{and} \\ &\mathbb{E}\mathbf{Y}_{:j} = 0, \forall j = 1, \cdots, q \end{aligned}$$

By first order optimality of the optimization problem in (2.2.1), we have

$$\mathbb{E}(\mathbf{X}_{:i})'(\mathbf{Y} - \mathbf{X}\Theta^\star) = 0, \forall i \Rightarrow \mathbb{E}(\mathbf{X}_{:i})'\mathbf{W}_{:j} = 0, \forall i, j$$

We know $\forall i, j$

$$\begin{aligned} &|(\mathbf{X}_{:i})'\mathbf{W}_{:j}| \\ =\;& |(\mathbf{X}_{:i})'\mathbf{W}_{:j} - \mathbb{E}[(\mathbf{X}_{:i})'\mathbf{W}_{:j}]| \\ =\;& \frac{1}{2}\left|\left(\|\mathbf{X}_{:i} + \mathbf{W}_{:j}\|^2 - \mathbb{E}[\|\mathbf{X}_{:i} + \mathbf{W}_{:j}\|^2]\right) - \left(\|\mathbf{X}_{:i}\|^2 - \mathbb{E}[\|\mathbf{X}_{:i}\|^2]\right) - \left(\|\mathbf{W}_{:j}\|^2 - \mathbb{E}[\|\mathbf{W}_{:j}\|^2]\right)\right| \\ \le\;& \frac{1}{2}\left|\|\mathbf{X}_{:i} + \mathbf{W}_{:j}\|^2 - \mathbb{E}[\|\mathbf{X}_{:i} + \mathbf{W}_{:j}\|^2]\right| + \frac{1}{2}\left|\|\mathbf{X}_{:i}\|^2 - \mathbb{E}[\|\mathbf{X}_{:i}\|^2]\right| + \frac{1}{2}\left|\|\mathbf{W}_{:j}\|^2 - \mathbb{E}[\|\mathbf{W}_{:j}\|^2]\right| \end{aligned}$$

Therefore,

$$\begin{aligned} &\mathbb{P}\left(\frac{1}{T}|(\mathbf{X}_{:i})'\mathbf{W}_{:j}| > 3t\right) \\ \le\;& \mathbb{P}\left(\frac{1}{2T}\left|\|\mathbf{X}_{:i} + \mathbf{W}_{:j}\|^2 - \mathbb{E}[\|\mathbf{X}_{:i} + \mathbf{W}_{:j}\|^2]\right| > t\right) + \mathbb{P}\left(\frac{1}{2T}\left|\|\mathbf{X}_{:i}\|^2 - \mathbb{E}[\|\mathbf{X}_{:i}\|^2]\right| > t\right) \\ &+ \mathbb{P}\left(\frac{1}{2T}\left|\|\mathbf{W}_{:j}\|^2 - \mathbb{E}[\|\mathbf{W}_{:j}\|^2]\right| > t\right) \end{aligned}$$

We will control the tail probabilities for each of the terms $|\|\mathbf{X}_{:i}\|^2 - \mathbb{E}[\|\mathbf{X}_{:i}\|^2]|$, $|\|\mathbf{W}_{:j}\|^2 - \mathbb{E}[\|\mathbf{W}_{:j}\|^2]|$ and $|\|\mathbf{X}_{:i} + \mathbf{W}_{:j}\|^2 - \mathbb{E}[\|\mathbf{X}_{:i} + \mathbf{W}_{:j}\|^2]|$. Before we apply Lemma II.13, we have to figure out their subweibull norms. We will first calculate the subweibull(γ_2) norm of $\mathbf{X}_{ti}, \mathbf{W}_{tj}$ and $\mathbf{X}_{ti}+\mathbf{W}_{tj}$. This will immediate yield control of the subweibull($\gamma_2/2$) norms of their squares via Lemma II.6.

Recall that

$$\begin{aligned} \mathbf{W}_{t:} &= \mathbf{Y}_{t:} - (\mathbf{X}\Theta^\star)_{t:} \\ &= \mathbf{Y}_{t:} - \mathbf{X}_{t:}\Theta^\star \end{aligned}$$

Therefore, we have,

$$\begin{aligned} \|\mathbf{W}_{tj}\|_{\gamma_2} &\leq \|\mathbf{W}_{t:}\|_{\gamma_2} && \text{by Definition 4} \\ &= \|\mathbf{Y}_{t:} - \mathbf{X}_{t:}\Theta^\star\|_{\gamma_2} \\ &\leq \|\mathbf{Y}_{t:}\|_{\gamma_2} + \|\mathbf{X}_{t:}\Theta^\star\|_{\gamma_2} && \|\cdot\|_{\gamma_2} \text{ is a norm} \\ &\leq \|\mathbf{Y}_{t:}\|_{\gamma_2} + \|\mathbf{X}_{t:}\|_{\gamma_2} |||\Theta^\star||| && \text{by Lemma } II.12 \\ &\leq K_Y + |||\Theta^\star||| K_X && \text{by Assumption 7.} \end{aligned}$$

We also have,

$$\begin{aligned} \|\mathbf{X}_{ti} + \mathbf{W}_{tj}\|_{\gamma_2} &\leq \|\mathbf{X}_{ti}\|_{\gamma_2} + \|\mathbf{W}_{tj}\|_{\gamma_2} && \|\cdot\|_{\gamma_2} \text{ is a norm} \\ &\leq K_Y + K_X\left(1 + |||\Theta^\star|||\right). \end{aligned}$$

Using Lemma II.6, we know that the subweibull($\gamma_2/2$) constants of the squares of $\mathbf{X}_{ti}, \mathbf{W}_{tj}$ and $\mathbf{X}_{ti} + \mathbf{W}_{tj}$ are all bounded by

$$K = 2^{2/\gamma_2}\left(K_Y + K_X\left(1 + |||\Theta^\star|||\right)\right)^2.$$

We now apply Lemma II.13 three times with γ_2 replaced by $\gamma_2/2$, to get, for any $t > 1/2T$,

$$\mathbb{P}\left(\frac{1}{T}\left|(\mathbf{X}_{:i})'\mathbf{W}_{:j}\right| > 3t\right) \leq 3T\exp\left\{-\frac{(2tT)^{\gamma}}{K^{\gamma}C_1}\right\} + 3\exp\left\{-\frac{4t^2T}{K^2C_2}\right\},$$

where $\gamma = (1/\gamma_1 + 2/\gamma_2)^{-1}$ is less than 1 by Assumption 8.

Now, taking a union bound over the pq possible values of i, j, gives us

$$\mathbb{P}\left(\frac{1}{T}|||\mathbf{X}'\mathbf{W}|||_{\infty} > 3t\right) \leq 3Tpq\exp\left\{-\frac{(2tT)^{\gamma}}{K^{\gamma}C_1}\right\} + 3pq\exp\left\{-\frac{4t^2T}{K^2C_2}\right\}.$$

If we set,

$$t = K\max\left\{C_2\sqrt{\frac{\log(3pq)}{T}}, \frac{C_1}{T}\left(\log(3Tpq)\right)^{1/\gamma}\right\}$$

then the probability of the large deviation event above is at most

$$2\exp(-c'\log(3pq)).$$

Note that the constant c' can be made arbitrarily large but affects the constants C_1, C_2 above.

In the expression for t above, we want to ensure that two conditions are met. First, the $1/\sqrt{T}$ term should dominate. That is, we want,

$$\sqrt{\frac{\log(3pq)}{T}} \geq \frac{C_1}{T}\left(\log(3Tpq)\right)^{1/\gamma},$$

which, in turn, is implied by

$$\sqrt{\frac{\log(3pq)}{T}} \geq \frac{C_2}{T}\left(\log(3T)\right)^{1/\gamma} \quad \text{and} \quad \sqrt{\frac{\log(3pq)}{T}} \geq \frac{C_2}{T}\left(\log(pq)\right)^{1/\gamma}$$

Both of these are met if $T \geq C_3(\log(pq))^{\frac{2}{\gamma}-1}$.

Finally, the condition $t > 1/2T$ should be met.That is,

$$C_2\sqrt{\frac{\log(3pq)}{T}} > \frac{1}{2T}$$

which happens as soon as $T \geq C_2^2/4$. □

Proof. (of Lemma II.8) Recall that $X_1, \cdots, X_t \in \mathbb{R}^p$ are subweibull random variables forming a β-mixing and stationary sequence.

Step I: Concentration for a fixed vector Now, fix a unit vector $v \in \mathbb{R}^p$, $\|v\|_2 = 1$. Define real valued random variables $Z_t = v'X_t$, $t = 1, \cdots, T$. Note that the β-mixing rate of (Z_t) is bounded by the same of (X_t) by Fact 2. From Lemma II.6, we know that $\|Z_t^2\|_{\psi_{\gamma_2/2}} \leq 2^{2/\gamma_2} \|Z_t\|_{\psi_{\gamma_2}}^2$. Moreover, $\|Z_t\|_{\psi_{\gamma_2}} \leq \|X_t\|_{\psi_{\gamma_2}}$. Therefore, we can invoke Lemma II.13 for the sum $S_T(v) = \sum_{t=1}^T (Z_t^2 - \mathbb{E}Z_t^2)$ with γ_2 replaced by $\gamma_2/2$, $\gamma = (1/\gamma_1 + 2/\gamma_2)^{-1}$ and $K = 2^{2/\gamma_2} K_X^2$ to get the following bound, for $T > 4$ and $t > 1/T$,

$$\mathbb{P}\left\{\left|\frac{S_T(v)}{T}\right| > t\right\} \leq T \exp\left\{-\frac{(tT)^\gamma}{K^\gamma C_1}\right\} + \exp\left\{-\frac{t^2 T}{K^2 C_2}\right\}$$

Step II: Uniform concentration over all vectors Let $\mathbb{J}(2k)$ denote the set of $2k$-sparse vector with Euclidean norm at most 1. Then, using union bound arguments similar to those in Lemma F.2 of Basu and Michailidis [2015], we have

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{v \in \mathbb{J}(2k)} \left|\frac{S_T(v)}{T}\right| > 3t\right\} \\
&\qquad \leq \exp\left\{\log(T) - \frac{(tT)^\gamma}{K^\gamma C_1} + k\log(p)\right\} + \exp\left\{-\frac{t^2 T}{K^2 C_2} + k\log(p)\right\}.
\end{aligned}$$

From the $2k$-sparse set, we will extend our bound to all $v \in \mathbb{R}^p$. To do so, we will apply Lemma 12 in Loh and Wainwright [2012]. For $k \geq 1$, with probability at least

$$1 - \exp\left\{\log(T) - \frac{(tT)^\gamma}{K^\gamma C_1} + k\log(p)\right\} - \exp\left\{-\frac{t^2 T}{K^2 C_2} + k\log(p)\right\} \tag{2.7.4}$$

the following holds uniformly for all $v \in \mathbb{R}^p$

$$\frac{1}{T}|S_T(v)| \geq 27t\left(\|v\|_2^2 + \frac{1}{k}\|v\|_1^2\right).$$

Let $\hat{\Sigma}_T(v) := \frac{1}{T}\left\|\mathbf{X}v\right\|_2^2$ and note that $\mathbb{E}\hat{\Sigma}_T(v) = v'\Sigma_X(0)v$. Therefore, $\frac{1}{T}S_T = \hat{\Sigma}_T(v) - \mathbb{E}\hat{\Sigma}_T(v)$. Using these notations, the above inequality implies that

$$\begin{aligned}
\hat{\Sigma}_T(v) &\geq v'\left(\Sigma_X(0)\right)v - 27\cdot t\left(\left\|v\right\|_2^2 + \frac{1}{k}\left\|v\right\|_1^2\right) \\
&\geq \lambda_{\min}(\Sigma_X(0))\left\|v\right\|_2^2 - 27\cdot t\left(\left\|v\right\|_2^2 + \frac{1}{k}\left\|v\right\|_1^2\right) \\
&= \left\|v\right\|_2^2\left(\lambda_{\min}(\Sigma_X(0)) - 27t\right) - \frac{27t}{k}\left\|v\right\|_1^2 \\
&= \left\|v\right\|_2^2\frac{1}{2}\lambda_{\min}(\Sigma_X(0)) - \frac{\lambda_{\min}(\Sigma_X(0))}{2k}\left\|v\right\|_1^2,
\end{aligned}$$

where the last line follows by picking $t = \frac{1}{54}\lambda_{\min}(\Sigma_X(0))$.

Step III: Selecting parameters The only thing left is to set the parameter k appropriately. We want to set it so that

$$2k\log p = \min\left\{\frac{(tT)^\gamma}{K^\gamma C_1}, \frac{t^2T}{K^2C_2}\right\}$$

so that the failure probability in (2.7.4) is at most $1 - 2T\exp(-k\log p)$. We want the minimum above to be attained at the first term which means we want

$$T \geq \left(\frac{K}{t}\right)^{\frac{2-\gamma}{1-\gamma}}\left(\frac{C_2}{C_1}\right)^{\frac{1}{1-\gamma}}$$

Under this condition, we have

$$k = \frac{(tT)^\gamma}{2K^\gamma C_1\log p}.$$

To ensure that $k \geq 1$, we need

$$T \geq \frac{54K\left(2C_1\log(p)\right)^{1/\gamma}}{\lambda_{\min}(\Sigma_X(0))}$$

To conclude, we have the following RE guarantee. For sample size

$$T \geq \max\left\{\frac{54K\left(2C_1\log(p)\right)^{1/\gamma}}{\lambda_{\min}(\Sigma_X(0))}, \left(\frac{54K}{\lambda_{\min}(\Sigma_X(0))}\right)^{\frac{2-\gamma}{1-\gamma}}\left(\frac{C_2}{C_1}\right)^{\frac{1}{1-\gamma}}\right\}$$

we have with probability at least

$$1 - 2T\exp\left\{-c'T^{\gamma}\right\}, \text{ where } c' = \frac{(\lambda_{\min}(\Sigma_X(0)))^{\gamma}}{(54K)^{\gamma}2C_1}$$

we have, for all $v \in \mathbb{R}^p$,

$$\hat{\Sigma}_T(v) \geq \alpha \|v\|_2^2 - \tau \|v\|_1^2$$

where

$$\alpha = \frac{1}{2}\lambda_{\min}(\Sigma_X(0)), \qquad \tau = \frac{\alpha}{2c'} \cdot \left(\frac{\log(p)}{T^{\gamma}}\right).$$

□

2.8 Verification of Assumptions for the Examples

2.8.1 VAR

Note that every VAR(d) process has an equivalent VAR(1) representation (see e.g. [Lütkepohl, 2005, Ch 2.1]) as

$$\tilde{Z}_t = \tilde{\mathbf{A}}\tilde{Z}_{t-1} + \tilde{\mathcal{E}}_t \tag{2.8.1}$$

where

(2.8.2)

$$\tilde{Z}_t := \begin{bmatrix} Z_t \\ Z_{t-1} \\ \vdots \\ Z_{t-d+1} \end{bmatrix}_{(pd\times 1)} \qquad \tilde{\mathcal{E}}_t := \begin{bmatrix} \mathcal{E}_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{(pd\times 1)} \qquad \text{and} \quad \tilde{\mathbf{A}} := \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 & \cdots & \mathbf{A}_{d-1} & \mathbf{A}_d \\ \mathbf{I}_p & 0 & 0 & 0 & 0 \\ 0 & \mathbf{I}_p & & 0 & 0 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \mathbf{I}_p & 0 \end{bmatrix}_{(dp\times dp)}$$

Because of this equivalence, justification of Assumptions 5(Gaussian case) and 6 (subweibull case) will operate through this corresponding augmented VAR(1) representation.

For both Gaussian and subweibull VARs, Assumption 3 is true since the sequences (Z_t) is centered. Second, $\Theta^\star = (\mathbf{A}_1, \cdots, \mathbf{A}_d)$. So Assumption 1 follows from construction.

For the remaining assumptions, we will consider the Gaussian and subweibull cases separately.

Gaussian VAR (Z_t) satisfies Assumption 4 by model assumption.

To show that (Z_t) is α-mixing with summable coefficients, we use the following facts together with the equivalence between (Z_t) and $(\tilde{Z}_t)$ and Fact 2.

Since $(\tilde{Z}_t)$ is stable, the spectral radius of $\tilde{\mathbf{A}}$, $r(\tilde{\mathbf{A}}) < 1$, hence Assumption 2 holds. Also the innovations $\tilde{\mathcal{E}}$ has finite first absolute moment and positive support everywhere. Then, according to Tjøstheim [1990, Theorem 4.4], $(\tilde{Z}_t)$ is *geometrically ergodic*. Note here that Gaussianity is *not* required here. Hence, it also applies to innovations from mixture of Gaussians.

Next, we present a standard result (see e.g. [Liebscher, 2005, Proposition 2]).

Fact 6. A stationary Markov chain $\{Z_t\}$ is geometrically ergodic implies $\{Z_t\}$ is *absolutely regular* (or β-mixing) with

$$\beta(n) = O(\gamma^n),\ \gamma \in (0, 1)$$

By the fact that β-mixing implies α-mixing (see Section 2.2.4) for a random process, we know that α-mixing coefficients decay geometrically and hence is summable. So, Assumption 5 holds.

Subweibull VAR To show that (Z_t) satisfies Assumptions 2 and 6, we establish that (Z_t) is geometrically ergodic. To show the latter, we use Propositions 1 and 2 in Liebscher [2005] together with the equivalence between (Z_t) and $(\tilde{Z}_t)$ and Fact 2.

To apply Proposition 1 in Liebscher [2005], we check the three conditions one by one. Condition (i) is immediate with $m = 1$, $E = \mathbb{R}^p$, and μ is the Lebesgue measure. For condition (ii), we set $E = \mathbb{R}^p$, μ to be the Lebesgue measure, and $\bar{m} = \lceil \inf_{u \in C, v \in A} \|u - v\|_2 \rceil$ the minimum "distance" between the sets C and A. Because C is bounded and A Borel, $\bar{m}$ is finite. Lastly, for condition (iii), we again let $E = \mathbb{R}^p$, μ to be the Lebesgue measure, and now the function $Q(\cdot) = \|\cdot\|$ and the set $K = \{x \in \mathbb{R}^p : \|x\| \leq \frac{2\mathbb{E}\|\tilde{\mathcal{E}}_t\|}{c\epsilon}\}$ where $c = 1 - |||\tilde{\mathbf{A}}|||$. Then,

- Recall from model assumption that $|||\tilde{\mathbf{A}}||| < 1$; hence,

  $$\mathbb{E}\left[\left\|\tilde{Z}_{t+1}\right\| \middle| \tilde{Z}_t = z\right] < |||\tilde{\mathbf{A}}||| \, \|z\| + \mathbb{E}(\left\|\tilde{\mathcal{E}}_{t+1}\right\|) \leq \left(1 - \frac{c}{2}\right) \|z\| - \epsilon,$$

  for all $z \in E \backslash K$

- For all $z \in K$,

  $$\mathbb{E}\left[\left\|\tilde{Z}_{t+1}\right\| \middle| \tilde{Z}_t = z\right] < |||\tilde{\mathbf{A}}||| \, \|z\| + \mathbb{E}(\left\|\tilde{\mathcal{E}}_{t+1}\right\|) \leq |||\tilde{\mathbf{A}}||| \frac{2\mathbb{E}\left\|\tilde{\mathcal{E}}_t\right\|}{c\epsilon}$$

- For all $z \in K$,

  $$0 \leq \|z\| \leq \frac{2\mathbb{E}\left\|\tilde{\mathcal{E}}_t\right\|}{c\epsilon}$$

Now, by Proposition 1 in Liebscher [2005], $(\tilde{Z}_t)$ is geometrically ergodic; hence $(\tilde{Z}_t)$ will be stationary. Once it reaches stationarity, by Proposition 2 in the same paper, the sequence will be β-mixing with geometrically decaying mixing coefficients. Therefore, Assumptions 2 and 6 hold.

We are left with checking Assumption 7. Let γ be the subweibull parameter associated with $(\mathcal{E}_t)$.

Assume that the spectral radius of A is smaller than 1; i.e. $r(A) < 1$. This is an equivalent notion of stability of VAR process.

By the definition of the spectral radius,

$$\lim_{m \to \infty} |||A^m|||^{1/m} = r(A) < 1$$

In other words, there exists a positive integer $k < \infty$ such that $\left|\left|\left|A^k\right|\right|\right| < 1$.

By the recursive nature of the time series,

$$\|Z_t\|_{\psi_\gamma} \leq \left|\left|\left|A^k\right|\right|\right| \|Z_{t-1}\|_{\psi_\gamma} + \sum_{i=1}^{k} \left|\left|\left|A^{k-i}\right|\right|\right| \|\mathcal{E}_{t-k+i}\|_{\psi_\gamma}$$

To simplify notation, let $C_i := \left|\left|\left|A^{k-i}\right|\right|\right|$. Using stationarity, we have the following

$$\|Z_t\|_{\psi_\gamma} \leq \frac{\|\epsilon_t\|_{\psi_\gamma}}{1 - |||A^k|||} \left(\sum_{i=1}^{k} c_i \right) < \infty$$

The last inequality follows because $C_i < \infty$, $\forall i = 1, \cdots, k$.

Thus, the sequence (Z_t) satisfies Assumption 7.

2.8.2 VAR with Misspecification

Assumption 3 is immediate from model definitions. By the same arguments as in Chapter 2.8.1, (Z_t, Ξ_t) are stationary and so is the sub-process (Z_t); Assumption 2 holds. Again, (Z_t, Ξ_t) satisfy Assumption 5 (for Example 2) and Assumption 6 (for Example 4) according to Chapter 2.8.1. By Fact 2, we have the same Assumptions hold for the respective sub-processes (Z_t) in the respective cases.

To show that $(\Theta^\star)' = \mathbf{A}_{ZZ} + \mathbf{A}_{Z\Xi}\Sigma_{\Xi Z}(0)(\Sigma_Z(0))^{-1}$, consider the following arguments. By Assumption 2, we have the auto-covariance matrix of the whole system (Z_t, Ξ_t) as

$$\Sigma_{(Z,\Xi)} = \begin{bmatrix} \Sigma_X(0) & \Sigma_{X\Xi}(0) \\ \Sigma_{\Xi X}(0) & \Sigma_{\Xi}(0) \end{bmatrix}$$

Recall our $\Theta^\star$ definition from Eq. (2.2.1)

$$\Theta^\star := \underset{B\in\mathbb{R}^{p\times p}}{\arg\min}\, \mathbb{E}\left(\|Z_t - B'Z_{t-1}\|_2^2\right)$$

Taking derivatives and setting to zero, we obtain

$$(\Theta^\star)' = \Sigma_Z(-1)(\Sigma_Z)^{-1} \tag{2.8.3}$$

Note that

$$\begin{aligned}\Sigma_Z(-1) &= \Sigma_{(Z,\Xi)}(-1)[1:p_1, 1:p_1] \\ &= \mathbb{E}\left(\mathbf{A}_{ZZ}Z_{t-1} + \mathbf{A}_{Z\Xi}\Xi_{t-1} + \mathcal{E}_{Z,t-1}\right)Z'_{t-1} \\ &= \mathbb{E}\left(\mathbf{A}_{ZZ}Z_{t-1}Z'_{t-1} + \mathbf{A}_{Z\Xi}\Xi_{t-1}Z'_{t-1} + \mathcal{E}_{Z,t-1}Z'_{t-1}\right) \\ &= \mathbf{A}_{ZZ}\Sigma_Z(0) + \mathbf{A}_{Z\Xi}\Sigma_{\Xi Z}(0)\end{aligned}$$

by Assumption 2 and the fact that the innovations are iid.

Naturally,

$$(\Theta^\star)' = \mathbf{A}_{ZZ}\Sigma_Z(0)(\Sigma_Z(0))^{-1} + \mathbf{A}_{Z\Xi}\Sigma_{\Xi Z}(0)(\Sigma_Z(0))^{-1} = \mathbf{A}_{ZZ} + \mathbf{A}_{Z\Xi}\Sigma_{\Xi Z}(0)(\Sigma_Z(0))^{-1}$$

Remark 8. Notice that $\mathbf{A}_{Z\Xi}$ is a column vector and suppose it is 1-sparse, and $\mathbf{A}_{ZZ}$ is p-sparse, then $\Theta^\star$ is at most $2p$-sparse. So Assumption 1 can be built in by model construction.

Remark 9. We gave an explicit model here where the left out variable Ξ was univariate. That was only for convenience. In fact, whenever the set of left-out variables Ξ

affect only a small set of variables Ξ in the retained system Z, the matrix $\Theta^\star$ is guaranteed to be sparse. To see that, suppose $\Xi \in \mathbb{R}^q$ and $\mathbf{A}_{Z\Xi}$ has at most s_0 non-zero rows (and let $\mathbf{A}_{ZZ}$ to be s-sparse as always), then $\Theta^\star$ is at most $(s_0 p + s)$-sparse.

Lastly, for Example 2, the sub-process (Z_t) is Gaussian because is obtained from a linear transformation of (Z_t, Ξ_t) which is Gaussian; we have Assumption 4. For Example 4, note that $Z_t = \mathbf{M}(Z_t, \Xi_t)$ where $\mathbf{M} = [\mathbf{I}_p, 0; 0', 0]$ is a sub-setting matrix that selects the first p entries of a $(p + 1)$-dimensional vector. Hence, the fact that Z_t is subweibull follows from the same arguments in Chapter 2.8.1 pertaining to establishing the subweibull property in conjunction with applying Lemma II.12 on $Z_t = \mathbf{M}(Z_t, \Xi_t)$; so, Assumption 7 holds.

Remark 10. Any VAR(d) process has an equivalent VAR(1) representation (Lutkepohl 2005). Our results extend to any VAR(d) processes.

2.8.3 ARCH

Verifying the Assumptions. To show that Assumption 6 hold for a process defined by Eq. (2.4.1) we leverage on Theorem 2 from Liebscher [2005]. Note that the original ARCH model in Liebscher [2005] assumes the innovations to have positive support everywhere. However, this is just a convenient assumption to establish the first two conditions in Proposition 1 (on which proof of Theorem 2 relies) from the same paper. ARCH model with innovations from more general distributions (e.g. uniform) also satisfies the first two conditions of Proposition 1 by the same arguments in the *Subweibull* paragraph of Chapter 2.8.1.

Theorem 2 tells us that for our ARCH model, if it satisfies the following conditions, it is guaranteed to be absolutely regular with geometrically decaying β-coefficients.

- $\mathcal{E}_t$ has positive density everywhere on $\mathbb{R}^p$ and has identity covariance by construction.
- $\Sigma(z) = o(\|z\|)$ because $m \in (0,1)$.
- $|||\Sigma(z)^{-1}||| \leq 1/(ac)$, $|\det(\Sigma(z))| \leq bc$
- $r(\mathbf{A}) \leq |||\mathbf{A}||| < 1$

So, Assumption 6 is valid here. We check other assumptions next.

Mean 0 is immediate, so we have Assumption 3. When the Markov chain did not start from a stationary distribution, geometric ergodicity implies that the sequence is approaching the stationary distribution exponentially fast. So, after a burning period, we will have Assumption 2 approximately valid here.

The subweibull constant of $\Sigma(Z_{t-1})\mathcal{E}_t$ given $Z_{t-1} = z$ is bounded as follows: for every z,

$$\begin{aligned} \|\Sigma(z)\mathcal{E}_t\|_{\psi_\gamma} &\leq |||\Sigma(z)|||\, \|\mathcal{E}_t\|_{\psi_\gamma} && \text{by Lemma II.12} \\ &\leq K_e cb =: K_E \end{aligned}$$

where $K_e := \sup_t \|\mathcal{E}_t\|_{\psi_\gamma}$

By the recursive relationship of $(Z_t)_t$, we have

$$\|Z_t\|_{\psi_\gamma} \leq |||\mathbf{A}|||\, \|Z_{t-1}\|_{\psi_\gamma} + K_E.$$

which yields the bound $\|Z_t\|_{\psi_\gamma} \leq K_E/(1 - |||\mathbf{A}|||) < \infty$. Hence Assumption 7 holds.

We will show below that $\Theta^\star = \mathbf{A}'$. Hence, sparsity (Assumption 1) can be built in when we construct our model 2.4.1.

Recall Eq. 2.8.3 from Chapter 2.8.2 that

$$\Theta^{\star} = \Sigma_Z(-1)(\Sigma_Z)^{-1}$$

Now,

$$\begin{aligned}
\Sigma_Z(-1) &= \mathbb{E} Z_t Z'_{t-1} && \text{by stationarity} \\
&= \mathbb{E}\left(\mathbf{A} Z_{t-1} + \Sigma(Z_{t-1})\mathcal{E}_t\right) Z'_{t-1} && \text{Eq. (2.4.1)} \\
&= \mathbf{A}\mathbb{E} Z_{t-1} Z'_{t-1} + \mathbb{E}\Sigma(Z_{t-1})\mathcal{E}_t Z'_{t-1} \\
&= \mathbf{A}\Sigma_Z + \mathbb{E}[c\,\mathrm{clip}_{a,b}\left(\|Z_{t-1}\|^m\right)\mathcal{E}_t Z'_{t-1}] \\
&= \mathbf{A}\Sigma_Z + \mathbb{E}[c\mathcal{E}_t Z'_{t-1}\mathrm{clip}_{a,b}\left(\|Z_{t-1}\|^m\right)] \\
&= \mathbf{A}\Sigma_Z + c\mathbb{E}\left[\mathcal{E}_t\right]\mathbb{E}\left[Z'_{t-1}\mathrm{clip}_{a,b}\left(\|Z_{t-1}\|^m\right)\right] && \text{i.i.d. innovations} \\
&= \mathbf{A}\Sigma_Z && \mathcal{E}_t \text{ mean 0,}
\end{aligned}$$

where $\mathrm{clip}_{a,b}\left(x\right) := \min\{\max\{x, a\}, b\}$ for $b > a$.

Since Σ_Z is invertible, we have $(\Theta^{\star})' = \Sigma_Z(-1)(\Sigma_Z)^{-1} = \mathbf{A}$.

CHAPTER III

Lasso Guarantees for Time Series Estimation Under Subgaussian Tails and β-Mixing

3.1 Introduction

Efficient estimation methods in high dimensional statistics [Bühlmann and Van De Geer, 2011, Hastie et al., 2015] include methods based on convex relaxation (see, e.g., Chandrasekaran et al. [2012], Negahban et al. [2012]) and methods using iterative optimization techniques (see, e.g., Beck and Teboulle [2009], Agarwal et al. [2012], Donoho et al. [2009]). A lot of work in the past decade has improved our understanding of the theoretical properties of these algorithms. However, the bulk of existing theoretical work focuses on *iid samples*. The extension of theory and algorithms in high dimensional statistics to *time series data* is just beginning to occur as we briefly summarize in Section **??** below. Note that, in time series applications, *dependence among samples* is the norm rather than the exception. So the development of high dimensional statistical theory to handle dependence is a pressing concern in time series estimation.

In this chapter, we give guarantees for ℓ_1-regularized least squares estimation, or Lasso [Hastie et al., 2015], that hold even when there is temporal dependence in data. Recently, Basu and Michailidis [2015] took a step forward in providing guarantees for

Lasso in the time series setting. They considered Gaussian VAR models with finite lag (see Example 6) and defined a measure of stability using the spectral density, which is the Fourier transform of the autocovariance function of the time series. Then they showed that one can derive error bounds for Lasso in terms of their measure of stability. Their bounds are an improvement over previous work [Negahban and Wainwright, 2011, Loh and Wainwright, 2012, Han and Liu, 2013] that assumed operator norm bounds on the transition matrix. These operator norm conditions are restrictive even for VAR models with a lag of 1 and never hold if the lag is strictly larger than 1! Therefore, the results of Basu and Michailidis [2015] are very interesting. But they do have limitations.

A key limitation is that Basu and Michailidis [2015] assume that the VAR model is the true data generating mechanism (DGM). Their proof techniques rely heavily on having the VAR representation of the stationary process available. The VAR model assumption, though popular, can be restrictive. The VAR family is not closed under linear transformations: if Z_t is a VAR process then CZ_t may not expressible as a finite lag VAR [Lütkepohl, 2005]. We later provides an example (Example 8) of VAR processes where omitting a single variable breaks down the VAR assumption. What if we do not assume that Z_t is a finite lag VAR process but simply that it is stationary? Under stationarity (and finite 2nd moment conditions), the best linear predictor of Z_t in terms of $Z_{t-d}, \ldots, Z_{t-1}$ is well defined even if Z_t is not a lag d VAR. If we assume that this best linear predictor involves sparse coefficient matrices, can we still guarantee consistent parameter estimation? This chapter provides an affirmative answer to this important question.

We provide finite sample parameter estimation and prediction error bounds for Lasso in stationary processes with subgaussian marginals and geometrically decaying β-

mixing coefficients (Corollary III.4). It is well known that guarantees for Lasso follow if one can establish restricted eigenvalue (RE) conditions and provide deviation bounds (DB) for the correlation of the noise with the regressors (see Theorem IV.2 in Section 3.6.2 below for a precise statement). Therefore, the bulk of the technical work in this chapter boils down to establishing, with high probability, that the RE and DB conditions hold under the subgaussian β-mixing assumptions. (Propositions III.2, III.3). Note that RE conditions were previously shown to hold under the *iid assumption* by Raskutti et al. [2010] for Gaussian random vectors and by Rudelson and Zhou [2013] for subgaussian random vectors. Our results rely on novel concentration inequality (Lemma III.1) for β-mixing subgaussian random variables that may be of independent interest. The inequality is proved by applying a blocking trick to Bernstein's concentration inequality for iid random variables. All proofs are deferred to the appendix.

To illustrate potential applications of our results, we present four examples. Example 6 considers a vanilla Gaussian VAR. Example 7 considers VAR models with subgaussian innovations. Examples 8 is concerned with subgaussian VAR models when the model is mis-specified. Lastly, we go beyond linear models and introduce non-linearity in the DGM in Example 9. To summarize, our theory for Lasso in high dimensional time series estimation extends beyond the classical linear Gaussian settings and provides guarantees potentially in the presence of model mis-specification, subgaussian innovations and/or nonlinearity in the DGM.

3.2 Preliminaries

Lasso Estimation Procedure for Dependent Data Consider a stochastic process of pairs $(X_t, Y_t)_{t=1}^{\infty}$ where $X_t \in \mathbb{R}^p$, $Y_t \in \mathbb{R}^q$, $\forall t$. One might be interested in predicting

Y_t given X_t. In particular, given a depedndent sequence $(Z_t)_{t=1}^T$, one might want to forecast the present Z_t using the past $(Z_{t-d}, \dots, Z_{t-1})$. A linear predictor is a natural choice. To put it in the regression setting, we identify $Y_t = Z_t$ and $X_t = (Z_{t-d}, \dots, Z_{t-1})$. The pairs (X_t, Y_t) defined as such are no longer iid. Assuming strict stationarity, the parameter matrix of interest $\Theta^\star \in \mathbb{R}^{p \times q}$ is

$$\Theta^\star = \underset{\Theta \in \mathbb{R}^{p \times q}}{\arg\min}\, \mathbb{E}[\|Y_t - \Theta' X_t\|_2^2]. \tag{3.2.1}$$

Note that $\Theta^\star$ is independent of t owing to stationarity. Because of high dimensionality ($pq \gg T$), consistent estimation is impossible without regularization. We consider the Lasso procedure. The ℓ_1-penalized least squares estimator $\widehat{\Theta} \in \mathbb{R}^{p \times q}$ is defined as

$$\widehat{\Theta} = \underset{\Theta \in \mathbb{R}^{p \times q}}{\arg\min} \frac{1}{T} \| \operatorname{vec}(Y - \mathbf{X}\Theta)\|_2^2 + \lambda_T \left\|\operatorname{vec}(\Theta)\right\|_1 . \tag{3.2.2}$$

where

$$Y = (Y_1, Y_2, \dots, Y_T)' \in \mathbb{R}^{T \times q} \qquad \mathbf{X} = (X_1, X_2, \dots, X_T)' \in \mathbb{R}^{T \times p}. \tag{3.2.3}$$

The following matrix of true residuals is not available to an estimator but will appear in our analysis:

$$\mathbf{W} := \mathbf{Y} - \mathbf{X}\Theta^\star. \tag{3.2.4}$$

Matrix and Vector Notation For a symmetric matrix $\mathbf{M}$, let $\lambda_{\max}(\mathbf{M})$ and $\lambda_{\min}(\mathbf{M})$ denote its maximum and minimum eigenvalues respectively. For any matrix let $\mathbf{M}$, $r(\mathbf{M})$, $|||\mathbf{M}|||$, $|||\mathbf{M}|||_\infty$, and $|||\mathbf{M}|||_F$ denote its spectral radius $\max_i \{|\lambda_i(\mathbf{M})|\}$, operator norm $\sqrt{\lambda_{\max}(\mathbf{M}'\mathbf{M})}$, entrywise ℓ_∞ norm $\max_{i,j} |\mathbf{M}_{i,j}|$, and Frobenius norm $\sqrt{\operatorname{tr}(\mathbf{M}'\mathbf{M})}$ respectively. For any vector $v \in \mathbb{R}^p$, $\left\|v\right\|_q$ denotes its ℓ_q norm $(\sum_{i=1}^p |v_i|^q)^{1/q}$.

Unless otherwise specified, we shall use $\|\cdot\|$ to denote the ℓ_2 norm. For any vector $v \in \mathbb{R}^p$, we use $\|v\|_0$ and $\|v\|_\infty$ to denote $\sum_{i=1}^p \mathbb{1}\{v_i \neq 0\}$ and $\max_i\{|v_i|\}$ respectively. Similarly, for any matrix $\mathbf{M}$, $|||\mathbf{M}|||_0 = \|\text{vec}(\mathbf{M})\|_0$ where $\text{vec}(\mathbf{M})$ is the vector obtained from $\mathbf{M}$ by concatenating the rows of M. We say that matrix $\mathbf{M}$ (resp. vector v) is *s-sparse* if $|||\mathbf{M}|||_0 = s$ (resp. $\|v\|_0 = s$). We use v' and $\mathbf{M}'$ to denote the transposes of v and $\mathbf{M}$ respectively. When we index a matrix, we adopt the following conventions. For any matrix $\mathbf{M} \in \mathbb{R}^{p\times q}$, for $1 \leq i \leq p$, $1 \leq j \leq q$, we define $\mathbf{M}[i,j] \equiv \mathbf{M}_{ij} := e_i'\mathbf{M}e_j$, $\mathbf{M}[i,:] \equiv \mathbf{M}_{i:} := e_i'\mathbf{M}$ and $\mathbf{M}[:,j] \equiv \mathbf{M}_{:j} := \mathbf{M}e_j$ where e_i is the vector with all 0s except for a 1 in the ith coordinate. The set of integers is denoted by $\mathbb{Z}$.

For a lag $l \in \mathbb{Z}$, we define the auto-covariance matrix w.r.t. $(X_t, Y_t)_t$ as $\Sigma(l) = \Sigma_{(X;Y)}(l) := \mathbb{E}[(X_t; Y_t)(X_{t+l}; Y_{t+l})']$. Note that $\Sigma(-l) = \Sigma(l)'$. Similarly, the auto-covariance matrix of lag l w.r.t. $(X_t)_t$ is $\Sigma_X(l) := \mathbb{E}[X_t X_{t+l}']$, and w.r.t. $(Y_t)_t$ is $\Sigma_Y(l) := \mathbb{E}[Y_t Y_{t+l}']$. The cross-covariance matrix at lag l is $\Sigma_{X,Y}(l) := \mathbb{E}[X_t Y_{t+l}']$. Note the difference between $\Sigma_{(X;Y)}(l)$ and $\Sigma_{X,Y}(l)$: the former is a $(p+q)\times(p+q)$ matrix, the latter is a $p \times q$ matrix. Thus, $\Sigma_{(X;Y)}(l)$ is a matrix consisting of four sub-matrices. Using Matlab-like notation, $\Sigma_{(X;Y)}(l) = [\Sigma_X, \Sigma_{X,Y}; \Sigma_{Y,X}, \Sigma_Y]$. As per our convention, at lag 0, we omit the lag argument l. For example, $\Sigma_{X,Y}$ denotes $\Sigma_{X,Y}(0) = \mathbb{E}[X_t Y_t']$.

A Brief Introduction to the β-Mixing Condition Mixing conditions [Bradley, 2005] are well established in the stochastic processes literature as a way to allow for dependence in extending results from the iid case. The general idea is to first define a measure of dependence between two random variables X, Y (that can vector-valued or even take values in a Banach space) with associated sigma algebras $\sigma(X), \sigma(Y)$.

In particular,

$$\beta(X, Y) = \sup \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} |P(A_i \cap B_j) - P(A_i)P(B_j)|$$

where the last supremum is over all pairs of partitions $\{A_1, \ldots, A_I\}$ and $\{B_1, \ldots, B_I\}$ of the sample space Ω such that $A_i \in \sigma(X), B_j \in \sigma(Y)$ for all i, j. Then for a stationary stochastic process $(X_t)_{t=-\infty}^{\infty}$, one defines the mixing coefficients, for $l \geq 1$,

$$\beta(l) = \beta(X_{-\infty:t}, X_{t+l:\infty}).$$

The β-mixing condition has been of interest in statistical learning theory for obtaining finite sample generalization error bounds for empirical risk minimization [Vidyasagar, 2003, Sec. 3.4] and boosting [Kulkarni et al., 2005] for dependent samples. There is also work on estimating β-mixing coefficients from data [Mcdonald et al., 2011]. At the same time, many interesting processes such as Markov and hidden Markov processes satisfy a β-mixing condition [Vidyasagar, 2003, Sec. 3.5]. Before we continue, we note an elementary but useful fact about mixing conditions, viz. they persist under arbitrary measurable transformations of the original stochastic process.

Fact 7. Suppose a stationary process $\{U_t\}_{t=1}^{T}$ is β-mixing. Then the stationary sequence $\{f(U_t)\}_{t=1}^{T}$, for any measurable function $f(\cdot)$, also is mixing in the same sense with its mixing coefficients bounded by those of the original sequence.

3.3 Main Results

We start with introducing two well-known sufficient conditions that enable us to provide non-asymptotic guarantees for Lasso estimation and prediction errors – the restricted eigenvalue (RE) and the deviation bound (DB) conditions. Note that in the classical linear model setting (see, e.g., Chap. 2.3 in Hayashi [2000]) where sample size is larger than the dimensions ($n > p$), the conditions for consistency of the ordinary least squares (OLS) estimator are as follows: (a) the empirical covariance matrix $\mathbf{X}'\mathbf{X}/T \overset{P}{\to} Q$ and Q invertible, i.e., $\lambda_{\min}(Q) > 0$, and (b) the regressors and the noise are asymptotically uncorrelated, i.e., $\mathbf{X}'\mathbf{W}/T \to 0$.

In high-dimensional regimes, Bickel et al. [2009], Loh and Wainwright [2012] and Negahban and Wainwright [2012] have established similar consistency conditions for Lasso. The first one is the *restricted eigenvalue* (RE) condition on $\mathbf{X}'\mathbf{X}/T$ (which is a special case, when the loss function is the squared loss, of the *restricted strong convexity* (RSC) condition). The second is the *deviation bound* (DB) condition on $\mathbf{X}'\mathbf{W}$. The following lower RE and DB definitions are modified from those given by Loh and Wainwright [2012].

Definition 5 (Lower Restricted Eigenvalue)**.** A symmetric matrix $\Gamma \in \mathbb{R}^{p\times p}$ satisfies a lower restricted eigenvalue condition with curvature $\alpha > 0$ and tolerance $\tau(T, p) > 0$ if

$$\forall v \in \mathbb{R}^p,\ v'\Gamma v \geq \alpha \left\|v\right\|_2^2 - \tau(T,p) \left\|v\right\|_1^2 .$$

Definition 6 (Deviation Bound)**.** Consider the random matrices $\mathbf{X} \in \mathbb{R}^{T\times p}$ and $\mathbf{W} \in \mathbb{R}^{T\times q}$ defined in (3.2.3) and (3.2.4) above. They are said to satisfy the deviation bound condition if there exist a deterministic multiplier function $\mathbb{Q}(\mathbf{X}, \mathbf{W}, \Theta^\star)$ and a rate of decay function $\mathbb{R}(p, q, T)$ such that:

$$\frac{1}{T}|||\mathbf{X}'\mathbf{W}|||_\infty \leq \mathbb{Q}(\mathbf{X}, \mathbf{W}, \Theta^\star)\mathbb{R}(p, q, T).$$

We will show that, with high probability, the RE and DB conditions hold for dependent data that satisfy Asumptions 9–13 described below. We shall do that *without* assuming any parametric form of the data generating mechanism. Instead, we will assume a subgaussian tail condition on the random vectors X_t, Y_t and that they satisfy the geometrically β-mixing condition.

3.3.1 Assumptions

Assumption 9 (Sparsity)**.** The matrix $\Theta^\star$ is s-sparse, i.e. $\|\text{vec}(\Theta^\star)\|_0 \leq s$.

Assumption 10 (Stationarity)**.** The process (X_t, Y_t) is strictly stationary: i.e., $\forall t, \tau,\, n \geq 0$,

$$((X_t, Y_t), \cdots, (X_{t+n}, Y_{t+n})) \overset{d}{=} ((X_{t+\tau}, Y_{t+\tau}), \cdots, (X_{t+\tau+n}, Y_{t+\tau+n})).$$

where "$\overset{d}{=}$" denotes equality in distribution.

Assumption 11 (Centering)**.** We have, $\forall t,\ \mathbb{E}(X_t) = 0_{p\times 1}$, and $\mathbb{E}(Y_t) = 0_{q\times 1}$.

The thin tail property of the Gaussian distribution is desirable from the theoretical perspective, so we would like to keep that but at the same time allow for more generality. The subgaussian distributions are a nice family characterized by having tail probabilities of the same as or lower order than the Gaussian. We now focus on subgaussian random vectors and present high probabilistic error bounds with all parameter dependences explicit.

Assumption 12 (Subgaussianity)**.** The subgaussian constants of X_t and Y_t are bounded above by $\sqrt{K_X}$ and $\sqrt{K_Y}$ respectively. (Please see Section 3.6.1 for a detailed introduction to subgaussian random vectors.)

Classically, mixing conditions were introduced to generalize classic limit theorems in probability beyond the case of iid random variables [Rosenblatt, 1956]. Recent work on high dimensional statistics has established the validity of RE conditions in the iid Gaussian [Raskutti et al., 2010] and iid Subgaussian cases [Rudelson and Zhou, 2013]. One of the main contributions of our work is to extend these results in high dimensional statistics from the iid to the mixing case.

Assumption 13 (β-Mixing)**.** The process $((X_t, Y_t))_t$ is geometrically β-mixing, i.e., there exists some constant $c_\beta > 0$ such that $\forall l \geq 1$, $\beta(l) \leq \exp(-c_\beta l)$,

The β-mixing condition allows us to apply the independent block technique developed by Yu [1994]. For examples of large classes of Markov and hidden Markov processes that are geometrically β-mixing, see Theorem 3.11 and Theorem 3.12 of Vidyasagar [2003]. In the independent blocking technique, we construct a new set of *independent* blocks such that each block has the same distribution as that of the corresponding block from the original sequence. Results of Yu [1994] provide upper bounds on the difference between probabilities of events defined using the independent blocks versus the same event defined using the original data. Classical probability theory tools for independent data can then be applied on the constructed independent blocks. In Section 3.6.3, we apply the independent blocking technique to Bernstein's inequality to get the following concentration inequality for β-mixing random variables.

Lemma III.1 (Concentration of β-Mixing Subgaussian Random Variables)**.** *Let $Z = (Z_1, \ldots, Z_T)$ consist of a sequence of mean-zero random variables with exponentially decaying β-mixing coefficients as in 13. Let K be such that $\max_{t=1}^T \|Z_t\|_{\psi_2} \leq \sqrt{K}$. Choose a block length $a_T \geq 1$ and let $\mu_T = \lfloor T/(2a_T) \rfloor$. We have, for any $t > 0$,*

$$\mathbb{P}[\frac{1}{T}|\|Z\|_2^2 - \mathbb{E}[\|Z\|_2^2]| > t] \leq 4\exp\left(-C_B \min\left\{\frac{t^2\mu_T}{K^2}, \frac{t\mu_T}{K}\right\}\right) + 2(\mu_T - 1)\exp\left(-c_\beta a_T\right) + \exp\left(\frac{-2t\mu_T}{K}\right).$$

In particular, for $0 < t < K$,

$$\begin{aligned}\mathbb{P}\left[\frac{1}{T}|\|Z\|_2^2 - \mathbb{E}[\|Z\|_2^2]| > t\right] \leq & 4\exp\left(-C_B\frac{t^2\mu_T}{K^2}\right) \\ & + 2(\mu_T - 1)\exp\left(-c_\beta a_T\right) + \exp\left(\frac{-2t\mu_T}{K}\right).\end{aligned}$$

Here C_B is the universal constant appearing in Bernstein's inequality (Proposition III.7).

Remark 11. The three terms in the bound above all have interpretations: the first is a concentration term with a rate that depends on the "effective sample size" μ_T, the number of blocks; the second is a dependence penalty accounting for the fact that the blocks are not exactly independent; and the third is a remainder term coming from the fact that $2a_T$ may not exactly divide T. The key terms are the first two and exhibit a natural trade-off: increasing a_T worsens the first term since μ_T decreases, but it improves the second term since there is less dependence at larger lags.

3.3.2 High Probability Guarantees for the Lower Restricted Eigenvalue and Deviation Bound Conditions

We show that both lower RE and DB conditions hold, with high probability, under our assumptions.

Proposition III.2 (RE)**.** *Suppose Assumptions 9–13 hold. Let C_B be the Berstein's inequality constant, $C = \min\{C_B, 2\}$, $b = \min\{\frac{1}{54K_X}\lambda_{\min}(\Sigma_X), 1\}$ and $c = \frac{1}{6}\max\{c_\beta, Cb^2\}$. Then for $T \geq \left(\frac{1}{c}\log(p)\right)^2$, with probability at least $1 - 5\exp\left(-CT^{\frac{1}{2}}\right) - 2(T^{\frac{1}{2}} - 1)\exp\left(-c_\beta T^{\frac{1}{2}}\right)$, we have for every vector $v \in \mathbb{R}^p$,*

$$v'\hat{\Gamma}v \geq \alpha_2\left\|v\right\|^2 - \tau_2(T,p)\left\|v\right\|_1^2,$$

where $\alpha_2 = \frac{1}{2}\lambda_{\min}(\Sigma_X)$, and $\tau_2(T,p) = 27bK_X\log(p)/cT^{\frac{1}{2}}$.

Proposition III.3 (Deviation Bound)**.** *Suppose Assumptions 9–13 hold. Let* $K = \sqrt{K_Y} + \sqrt{K_X}\,(1 + |||\Theta^\star|||)$ *and* $\xi \in (0,1)$ *be a free parameter. Then, for sample size*

$$T \geq \max\left\{ \left(\log(pq) \max\left\{ \frac{K^4}{2C_B}, K^2 \right\} \right)^{\frac{1}{1-\xi}}, \left[\frac{2}{c_\beta} \log(pq) \right]^{\frac{1}{\xi}} \right\},$$

we have

$$\mathbb{P}\left[\frac{1}{T} |||\boldsymbol{X}'\boldsymbol{W}|||_\infty \leq \mathbb{Q}(\boldsymbol{X}, \boldsymbol{W}, \Theta^\star) \mathbb{R}(p,q,T) \right]$$
$$\geq 1 - 15 \exp\left(-\frac{1}{2} \log(pq) \right) - 6(T^{1-\xi} - 1) \exp\left(-\frac{1}{2} c_\beta T^\xi \right)$$

where

$$\mathbb{Q}(\boldsymbol{X}, \boldsymbol{W}, \Theta^\star) = \sqrt{\frac{2K^4}{C_B}}, \qquad \mathbb{R}(p,q,T) = \sqrt{\frac{\log(pq)}{T^{1-\xi}}}.$$

Remark 12. Since $\xi \in (0,1)$ is a free parameter, we choose it to be arbitrarily close to zero so that $\mathbb{R}(p,q,T)$ scales at a rate arbitrarily close to $\sqrt{\frac{\log(pq)}{T}}$. However, there is a price to pay for this: both the initial sample threshold and the success probability worsen as we make ξ very small.

3.3.3 Estimation and Prediction Errors

The guarantees below follow easily from plugging the RE and DB constants from Propositions III.2 and III.3 into a "master theorem" (Theorem III.5 in Section 3.6.2). The "master theorem", in various forms, is well-known in the literature (e.g., see Bickel et al. [2009], Loh and Wainwright [2012], Negahban and Wainwright [2012]).

Corollary III.4 (Lasso Guarantee under Subgaussian Tails and β-Mixing)**.** *Suppose Assumptions 9–13 hold. Let* C_B, C, c, b *and* K *be as defined in Propositions III.2 and*

III.3 and $\tilde{C} := \min\{C, c_\beta\}$*. Let* $\xi \in (0,1)$ *be a free parameter. Then, for sample size*

$$T \geq \max\left\{ \left(\frac{\log(p)}{c}\right)^2 \max\left\{ \left(\frac{1728 sbK_X}{\lambda_{\min}(\Sigma_X)}\right)^2, 1 \right\}, \right.$$
$$\left. \left(\log(pq) \max\left\{\frac{K^4}{2C_B}, K^2\right\}\right)^{\frac{1}{1-\xi}}, \left[\frac{2}{c_\beta}\log(pq)\right]^{\frac{1}{\xi}} \right\}$$

we have with probability at least

$$1 - 15\exp\left(-\frac{1}{2}\log(pq)\right) - 6(T^{1-\xi} - 1)\exp\left(-\frac{1}{2}c_\beta T^{\xi}\right) - 5(T^{\frac{1}{2}} - 1)\exp\left(-\tilde{C}T^{\frac{1}{2}}\right)$$

the Lasso estimation and (in-sample) prediction error bounds

$$\left\|\operatorname{vec}(\widehat{\Theta} - \Theta^\star)\right\| \leq 4\sqrt{s}\lambda_T/\alpha, \tag{3.3.1}$$

$$\left|\left|\left|(\widehat{\Theta} - \Theta^\star)'\hat{\Gamma}(\widehat{\Theta} - \Theta^\star)\right|\right|\right|_F^2 \leq \frac{32\lambda_T^2 s}{\alpha}. \tag{3.3.2}$$

hold with

$$\alpha = \frac{1}{2}\lambda_{\min}(\Sigma_X), \qquad \lambda_T = 4\mathbb{Q}(\boldsymbol{X}, \boldsymbol{W}, \Theta^\star)\mathbb{R}(p, q, T)$$

where

$$\hat{\Gamma} := \boldsymbol{X}'\boldsymbol{X}/T, \qquad \mathbb{Q}(\boldsymbol{X}, \boldsymbol{W}, \Theta^\star) = \sqrt{\frac{2K^4}{C_B}}, \qquad \mathbb{R}(p, q, T) = \sqrt{\frac{\log(pq)}{T^{1-\xi}}}.$$

Remark 13. The condition number of Σ_X plays an important part in the literature of Lasso error guarantees [Loh and Wainwright, 2012, e.g.]. Here, we see that the role of the condition number $\lambda_{\max}(\Sigma_X)/\lambda_{\min}(\Sigma_X)$ is replaced by $K_X/\lambda_{\min}(\Sigma_X)$ that now serves as the "effective condition number."

3.4 Examples

We explore applicability of our theory beyond just linear Gaussian processes using the examples below. In the following examples, we identify $X_t := Z_t$ and $Y_t := Z_{t+1}$

for $t = 1, \ldots, T$. For the specific parameter matrix $\Theta^\star$ in each Example below, we can verify that Assumptions 9–13 hold (see Section 3.6.5) for details. Therefore, Propositions III.2 and III.3 and Corollary III.4 follow. Hence we have all the high probabilistic guarantees for Lasso on data generated from DGM potentially involving subgaussianity, model mis-specification, and/or nonlinearity.

Example 6 (Gaussian VAR)**.** Transition matrix estimation in sparse stable VAR models has been considered by several authors in recent years [Davis et al., 2015, Han and Liu, 2013, Song and Bickel, 2011]. The Lasso estimator is a natural choice for the problem.

We state the following convenient fact because it allows us to study any finite order VAR model by considering its equivalent VAR(1) representation. See Section 3.6.5 for details.

Fact 8. Every VAR(d) process can be written in VAR(1) form (see e.g. [Lütkepohl, 2005, Ch 2.1]).

Therefore, without loss of generality, we can consider VAR(1) model in the ensuing Examples.

Formally a first order Gaussian VAR(1) process is defined as follows. Consider a sequence of serially ordered random vectors (Z_t), $Z_t \in \mathbb{R}^p$ that admits the following auto-regressive representation:

$$Z_t = \mathbf{A} Z_{t-1} + \mathcal{E}_t \tag{3.4.1}$$

where $\mathbf{A}$ is a non-stochastic coefficient matrix in $\mathbb{R}^{p \times p}$ and innovations $\mathcal{E}_t$ are p-dimensional random vectors from $\mathcal{N}(0, \Sigma_\epsilon)$ with $\lambda_{\min}(\Sigma_\epsilon) > 0$ and $\lambda_{\max}(\Sigma_\epsilon) < \infty$.

Assume that the VAR(1) process is *stable*; i.e. $\det\left(\mathbf{I}_{p\times p} - \mathbf{A}z\right) \neq 0, \forall\,|z| \leq 1$. Also, assume $\mathbf{A}$ is s-sparse. In here, $\Theta^\star = \mathbf{A}' \in \mathbb{R}^{p\times p}$.

Example 7 (VAR with Subgaussian Innovations)**.** Consider a VAR(1) model defined as in Example 6 except that we replace the Gaussian white noise innovations with subgaussian ones and assume $|||\mathbf{A}||| < 1$.

For example, take iid random vectors from the uniform distribution; i.e. $\forall t$, $\mathcal{E}_t \overset{iid}{\sim} U\left(\left[-\sqrt{3}, \sqrt{3}\right]^p\right)$. These $\mathcal{E}_t$ will be independent centered isotropic subgaussian random vectors, giving us we a VAR(1) model with subgaussian innovations. If we take a sequence $(Z_t)_{t=1}^{T+1}$ generated according to the model, each element Z_t will be a mean zero subgaussian random vector. Note that $\Theta^\star = A'$.

Example 8 (VAR with subgaussian Innovations and Omitted Variable)**.** We will study estimation of a VAR(1) process when there are endogenous variables omitted. This arises naturally when the underlying DGM is high-dimensional but not all variables are available/observable/measurable to the researcher to do estimation/prediction. This also happens when the researcher mis-specifies the scope of the model.

Notice that the system of the retained set of variables is no longer a finite order VAR (and thus non-Markovian). This example serves to illustrate that our theory is applicable to models beyond the finite order VAR setting.

Consider a VAR(1) process $(Z_t, \Xi_t)_{t=1}^{T+1}$ such that each vector in the sequence is generated by the recursion below:

$$(Z_t; \Xi_t) = \mathbf{A}(Z_{t-1}; \Xi_{t-1}) + (\mathcal{E}_{Z,t-1}; \mathcal{E}_{\Xi,t-1})$$

where $Z_t \in \mathbb{R}^p$, $\Xi_t \in \mathbb{R}$, $\mathcal{E}_{Z,t} \in \mathbb{R}^p$, and $\mathcal{E}_{\Xi,t} \in \mathbb{R}$ are partitions of the random vectors (Z_t, Ξ_t) and $\mathcal{E}_t$ into p and 1 variables. Also,

$$\mathbf{A} := \begin{bmatrix} \mathbf{A}_{ZZ} & \mathbf{A}_{Z\Xi} \\ \mathbf{A}_{\Xi Z} & \mathbf{A}_{\Xi\Xi} \end{bmatrix}$$

is the coefficient matrix of the VAR(1) process with $\mathbf{A}_{Z\Xi}$ 1-sparse, $\mathbf{A}_{ZZ}$ p-sparse and $|||\mathbf{A}||| < 1$. $\mathcal{E}_t := (\mathcal{E}_{X,t-1}; \mathcal{E}_{Z,t-1})$ for $t = 1, \ldots, T+1$ are iid draws from a subgaussian distribution; in particular we consider the subgaussian distribution described in Example 7.

We are interested in the OLS 1-lag estimator of the system restricted to the set of variables in Z_t. Recall that

$$\Theta^\star := \underset{\mathbf{B} \in \mathbb{R}^{p \times p}}{\arg\min}\, \mathbb{E}\left(\|Z_t - \mathbf{B}' Z_{t-1}\|_2^2\right)$$

We show in the chapter that $(\Theta^\star)' = \mathbf{A}_{ZZ} + \mathbf{A}_{Z\Xi}\Sigma_{\Xi Z}(0)(\Sigma_Z)^{-1}$ is sparse.

Example 9 (Multivariate ARCH)**.** We will explore the generality of our theory by considering a multivariate nonlinear time series model with subgaussian innovations. A popular nonlinear multivariate time series model in econometrics and finance is the vector autoregressive conditionally heteroscedastic (ARCH) model. We chose the following specific ARCH model for convenient validation of the geometric β-mixing property; it may potentially be applicable to a larger class of multivariate ARCH models. Consider a sequence of random vector $(Z_t)_{t=1}^{T+1}$ generated by the following recursion. For any constants $c > 0$, $m \in (0,1)$, $a > 0$, and $\mathbf{A}$ sparse with $|||\mathbf{A}||| < 1$:

$$\begin{aligned} Z_t &= \mathbf{A} Z_{t-1} + \Sigma(Z_{t-1})\mathcal{E}_t \\ \Sigma(z) &:= c \cdot \operatorname{clip}_{a,b}\left(\|z\|^m\right) \mathbf{I}_{p\times p} \end{aligned} \tag{3.4.2}$$

where $\mathcal{E}_t$ are iid random vectors from some subgaussian distribution and $\operatorname{clip}_{a,b}(x)$ clips the argument x to stay in the interval $[a, b]$. We can take innovations $\mathcal{E}_t$ to be iid random vectors from uniform distribution as described in Example 7. Consequently, each Z_t will be a mean zero subgaussian random vector. Note that $\Theta^\star = \mathbf{A}'$, the transpose of the coefficient matrix $\mathbf{A}$ here.

3.5 Simulations

Corollary III.4 in Section 3.3 makes a precise prediction for the ℓ_2 parameter error $\left|\left|\left|\Theta^* - \hat{\Theta}\right|\right|\right|_F$. We report scaling simulations for Examples 6–9 to confirm the sharpness of the bounds.

Sparsity is always $s = \sqrt{p}$, noise covariance matrix $\Sigma_\epsilon = I_p$, and the operator norm of the driving matrix set to $|||\mathbf{A}||| = 0.9$. The problem dimensions are $p \in \{50, 100, 200, 300\}$. Top left, top right, bottom left and bottom right sub-figures in Figure 3.1 correspond to simulations of Examples 6, 7, 8 and 9 respectively.

In all combinations of the four dimensions and Examples, the error decreases to zero as the sample size n increases, showing consistency of the method. In each sub-figure, the ℓ_2 parameter error curves align when plotted against a suitably rescaled sample size ($\frac{T}{s\log(p)}$) for different values of dimension p. We see the error scaling agrees nicely with theoretical guarantees provided by Corollary III.4.

3.6 Supplement

3.6.1 Sub-Gaussian Constants for Random Vectors

The sub-Gaussian and sub-Exponential constants have various equivalent definitions, we adopt the following from Rudelson and Vershynin [2013].

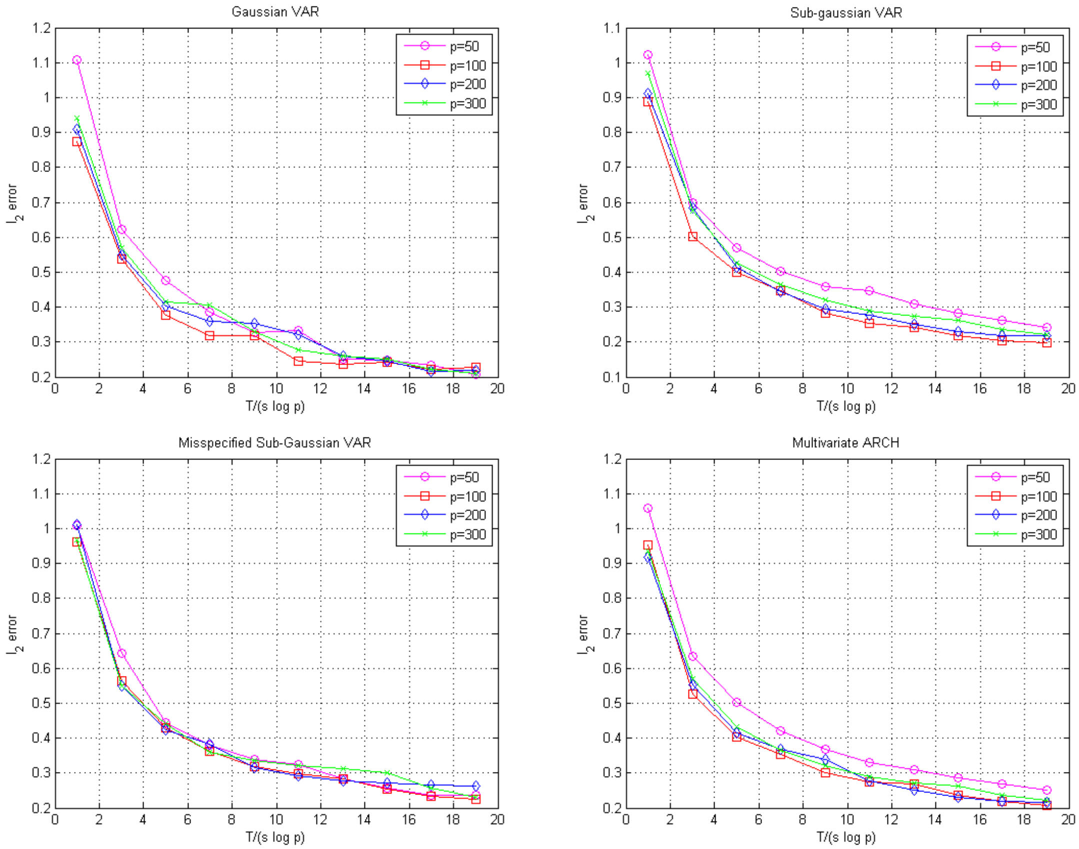

Figure 3.1: ℓ_2 estimation error of lasso against rescaled sample size for Examples 6–9.

Definition 7 (Sub-Gaussian Norm and Random Variables/Vectors)**.** A random variable U is called sub-Gaussian with sub-Gaussian constant K if its sub-Gaussian norm

$$\|U\|_{\psi_2} := \sup_{p\geq 1} p^{-\frac{1}{2}} (\mathbb{E}\,|U|^p)^{1/p}$$

satisfies $\|U\|_{\psi_2} \leq K$.

A random vector $V \in \mathbb{R}^n$ is called sub-Gaussian if all of its one-dimensional projections are sub-Gaussian and we define

$$\|V\|_{\psi_2} := \sup_{v\in\mathbb{R}^n:\|v\|\leq 1} \|v'V\|_{\psi_2}$$

.

Definition 8 (Sub-exponential Norm and Random Variables/Vectors)**.** A random variable U is called sub-exponential with sub-exponential constant K if its sub-exponential norm

$$\|U\|_{\psi_1} := \sup_{p\geq 1} p^{-1} (\mathbb{E}\,|U|^p)^{1/p}$$

satisfies $\|U\|_{\psi_1} \leq K$.

A random vector $V \in \mathbb{R}^n$ is called sub-exponential if all of its one-dimensional projections are sub-exponential and we define

$$\|U\|_{\psi_1} := \sup_{v\in\mathbb{R}^n:\|v\|\leq 1} \|v'V\|_{\psi_1}$$

Fact 9*.* A random variable U is sub-Gaussian iff U^2 is sub-exponential with $\|U\|_{\psi_2}^2 = \|U^2\|_{\psi_1}$.

3.6.2 Proof of Master Theorem

We present a master theorem that provides guarantees for the ℓ_2 parameter estimation error and for the (in-sample) prediction error. The proof builds on existing result of the same kind [Bickel et al., 2009, Loh and Wainwright, 2012, Negahban and Wainwright, 2012] and we make no claims of originality for either the result or for the proof.

Theorem III.5 (Estimation and Prediction Errors)**.** *Consider the Lasso estimator $\widehat{\Theta}$ defined in* (3.2.2)*. Suppose Assumption 9 holds. Further, suppose that $\hat{\Gamma} := \boldsymbol{X}'\boldsymbol{X}/T$ satisfies the lower RE(α, τ) condition with $\alpha \geq 32s\tau$ and $\boldsymbol{X}'\boldsymbol{W}$ satisfies the deviation bound. Then, for any $\lambda_T \geq 4\mathbb{Q}(\boldsymbol{X}, \boldsymbol{W}, \Theta^\star)\mathbb{R}(p, q, T)$, we have the following guarantees:*

$$\left\|\text{vec}(\widehat{\Theta} - \Theta^\star)\right\| \leq 4\sqrt{s}\lambda_T/\alpha, \tag{3.6.1}$$

$$\left\|\left|(\widehat{\Theta} - \Theta^\star)'\hat{\Gamma}(\widehat{\Theta} - \Theta^\star)\right\|\right|_F^2 \leq \frac{32\lambda_T^2 s}{\alpha}. \tag{3.6.2}$$

Proof of Theorem IV.2. We wil break down the proof in steps.

1. Since $\widehat{\Theta}$ is optimal for 3.2.2 and $\Theta^\star$ is feasible,

$$\frac{1}{T}\left\|\left|Y - \mathbf{X}\widehat{\Theta}\right\|\right|_F^2 + \lambda_T \left\|\text{vec}(\widehat{\Theta})\right\|_1 \leq \frac{1}{T}|||Y - \mathbf{X}\Theta^\star|||_F^2 + \lambda_T \left\|\text{vec}(\Theta^\star)\right\|_1$$

2. Let $\hat{\Delta} := \widehat{\Theta} - \Theta^\star \in \mathbb{R}^{p\times q}$

$$\frac{1}{T}\left\|\left|\mathbf{X}\hat{\Delta}\right\|\right|_F^2 \leq \frac{2}{T}\text{tr}(\hat{\Delta}'\mathbf{X}'\mathbf{W}) + \lambda_T\left(\|\text{vec}(\Theta^\star)\|_1 - \left\|\text{vec}(\widehat{\Theta})\right\|_1\right)$$

Note that

$$\begin{aligned}\left\|\text{vec}(\Theta^\star + \hat{\Delta})\right\|_1 - \|\text{vec}(\Theta^\star)\|_1 \geq& \{\|\text{vec}(\Theta^\star_S)\|_1 - \left\|\text{vec}(\hat{\Delta}_S)\right\|_1\} \\ &+ \left\|\text{vec}(\hat{\Delta}_{S^c})\right\|_1 - \|\text{vec}(\Theta^\star)\|_1 \\ &= \left\|\text{vec}(\hat{\Delta}_{S^c})\right\|_1 - \left\|\text{vec}(\hat{\Delta}_S)\right\|_1\end{aligned}$$

where S denote the support of $\Theta^\star$.

3. With RE constant α and tolerance τ, deviation bound constant $\mathbb{Q}(\Sigma_X, \Sigma_W)$ and $\lambda_T \geq 2\mathbb{Q}(\Sigma_X, \Sigma_W)\sqrt{\frac{\log(q)}{T}}$, we have

$$\begin{aligned}
\alpha \left|\!\left|\!\left| \hat{\Delta} \right|\!\right|\!\right|_F^2 - \tau \| \operatorname{vec}(\hat{\Delta}) \|_1^2 & \\
&\overset{RE}{\leq} \frac{1}{T} |||\mathbf{X}\Delta|||_F^2 \\
&\leq \frac{2}{T} \operatorname{tr}(\hat{\Delta}'\mathbf{X}'\mathbf{W}) + \lambda_T \{ \left\| \operatorname{vec}(\hat{\Delta}_S) \right\|_1 - \left\| \operatorname{vec}(\hat{\Delta}_{S^c}) \right\|_1 \} \\
&\leq \frac{2}{T} \sum_{k=1}^{q} \| \hat{\Delta}_{:k} \|_1 \| (\mathbf{X}'\mathbf{W})_{:k} \|_\infty + \lambda_T \{ \left\| \operatorname{vec}(\hat{\Delta}_S) \right\|_1 - \left\| \operatorname{vec}(\hat{\Delta}_{S^c}) \right\|_1 \} \\
&\leq \frac{2}{T} \| \operatorname{vec}(\hat{\Delta}) \|_1 |||\mathbf{X}'\mathbf{W}|||_\infty + \lambda_T \{ \left\| \operatorname{vec}(\hat{\Delta}_S) \right\|_1 - \left\| \operatorname{vec}(\hat{\Delta}_{S^c}) \right\|_1 \} \\
&\overset{DB}{\leq} 2 \| \operatorname{vec}(\hat{\Delta}) \|_1 \mathbb{Q}(\Sigma_X, \Sigma_W) \mathbb{R}(p, q, T) + \lambda_T \{ \left\| \operatorname{vec}(\hat{\Delta}_S) \right\|_1 - \left\| \operatorname{vec}(\hat{\Delta}_{S^c}) \right\|_1 \} \\
&\leq \| \operatorname{vec}(\hat{\Delta}) \|_1 \lambda_N / 2 + \lambda_T \{ \left\| \operatorname{vec}(\hat{\Delta}_S) \right\|_1 - \left\| \operatorname{vec}(\hat{\Delta}_{S^c}) \right\|_1 \} \\
&\leq \frac{3\lambda_T}{2} \left\| \operatorname{vec}(\hat{\Delta}_S) \right\|_1 - \frac{\lambda_T}{2} \left\| \operatorname{vec}(\hat{\Delta}_{S^c}) \right\|_1 \\
&\leq 2\lambda_T \left\| \operatorname{vec}(\hat{\Delta}) \right\|_1
\end{aligned}$$

4. In particular, this says that $3 \left\| \operatorname{vec}(\hat{\Delta}_S) \right\|_1 \geq \left\| \operatorname{vec}(\hat{\Delta}_{S^c}) \right\|_1$
   So $\left\| \operatorname{vec}(\hat{\Delta}) \right\|_1 \leq 4 \left\| \operatorname{vec}(\hat{\Delta}_S) \right\|_1 \leq 4\sqrt{s} \left\| \operatorname{vec}(\hat{\Delta}) \right\|$

5. Finally, with $\alpha \geq 32s\tau$,

$$\begin{aligned}
\frac{\alpha}{2} \left\| \operatorname{vec}(\hat{\Delta}) \right\|_F^2 &\leq (\alpha - 16s\tau) \left\| \operatorname{vec}(\hat{\Delta}) \right\|_F^2 \\
&\leq \alpha \left\| \operatorname{vec}(\hat{\Delta}) \right\|_F^2 - \tau \| \operatorname{vec}(\hat{\Delta}) \|_1^2 \\
&\leq 2\lambda_T \| \operatorname{vec}(\hat{\Delta}) \|_1 \\
&\leq 2\sqrt{s}\lambda_T \| \hat{\Delta} \|_F
\end{aligned}$$

6.

$$\left\| \operatorname{vec}(\hat{\Delta}) \right\|_F \leq \frac{4\lambda_T\sqrt{s}}{\alpha}$$

7. From step 4, we have

$$\frac{1}{T}\left|\left|\left|\mathbf{X}\hat{\Delta}\right|\right|\right|_F^2 \leq 8\lambda_T\sqrt{s}\left\|\text{vec}(\hat{\Delta})\right\|$$

Then, from step 6

$$\frac{1}{T}\left|\left|\left|\mathbf{X}\hat{\Delta}\right|\right|\right|_F^2 \leq 8\lambda_T\sqrt{s}\left\|\text{vec}(\hat{\Delta})\right\| \leq 32\lambda_T^2 s/\alpha$$

□

3.6.3 Proofs for Sub-Gaussian Random Vectors under β-Mixing

Proof of Lemma III.1.

Following the description in Yu [1994], we divide the stationary sequence of real valued random variables $\{Z_t\}_{t=1}^T$ into $2\mu_T$ blocks of size a_T with a remainder block of length $T - 2\mu_T a_T$. Let H and T be sets that denote the indices in the odd and even blocks respectively, and let Re to denote the indices in the remainder block. To be specific,

$$O = \cup_{j=1}^{\mu_T} O_j \text{ where } O_j := \{i : 2(j-1)a_T + 1 \leq i \leq (2j-1)a_T\}, \forall j$$

$$E := \cup_{j=1}^{\mu_T} E_j \text{ where } E_j := \{i : (2j-1)a_T + 1 \leq i \leq (2j)a_T\}, \forall j$$

Let $Z_o := \{Z_t : t \in O\}$ be a collection of the random vectors in the odd blocks. Similarly, $Z_e := \{Z_t : t \in E\}$ is a collection of the random vectors in the even blocks, and $Z_r := \{Z_t : t \in Re\}$ a collection of the random vectors in the remainder block. Lastly, $Z := Z_O \cup Z_e \cup Z_r$

Now, take a sequence of i.i.d. blocks $\{\tilde{Z}_{O_j} : j = 1, \cdots, \mu_t\}$ such that each $\tilde{Z}_{O_j}$ is independent of $\{Z_t\}_{t=1}^T$ and each $\tilde{Z}_{O_j}$ has the same distribution as the corresponding block from the original sequence $\{Z_j : j \in O_j\}$. We construct the even and remainder blocks in a similar way and denote them $\{\tilde{Z}_{E_j} : j = 1, \cdots, \mu_t\}$ and $\tilde{Z}_{Re}$ respectivey. $\tilde{Z}_O := \cup_{j=1}^{\mu_T} \tilde{Z}_{O_j} (\tilde{Z}_E := \cup_{j=1}^{\mu_T} \tilde{Z}_{E_j})$ denote the union of the odd(even) blocks.

For the odd blocks: $\forall\, t > 0$,

$$\begin{aligned}
&\mathbb{P}[\frac{2}{T}|\|Z_o\|_2^2 - \mathbb{E}(\|Z_o\|_2^2)| > t] \\
&= \mathbb{E}[\mathbb{1}\{\frac{2}{T}|\|Z_o\|_2^2 - \mathbb{E}(\|Z_o\|_2^2)|\} > t\}] \\
&\le \mathbb{E}[\mathbb{1}\{\frac{2}{T}|\|\tilde{Z}_o\|_2^2 - \mathbb{E}(\|\tilde{Z}_o\|_2^2)|\} > t\}] + (\mu_{a_T} - 1)\beta(a_T) \\
&= \mathbb{P}[\frac{2}{T}|\|\tilde{Z}_o\|_2^2 - \mathbb{E}(\|\tilde{Z}_o\|_2^2)| > t\}] + (\mu_{a_T} - 1)\beta(a_T) \\
&= \mathbb{P}[\frac{1}{\mu_T}|\sum_{i=1}^{\mu_T} \|\tilde{Zo}_i\|_2^2 - \mathbb{E}(\|\tilde{Z}_{o_i}\|_2^2)| > ta_T] + (\mu_{a_T} - 1)\beta(a_T) \\
&\le 2\exp\left\{-C_B \min\left\{\frac{t^2\mu_T}{K^2}, \frac{t\mu_T}{K}\right\}\right\} + (\mu_{a_T} - 1)\beta(a_T)
\end{aligned}$$

Where the first inequality follows from [Yu, 1994, Lemma 4.1] with $M = 1$. By Fact (9), the corresponding sub-exponential constant of each $\left\|\tilde{Z}_{o_i}\right\|^2 \le a_T K$ where K is the sub-exponential norm because of fact 9. With this, the second inequality follows from the Bernstein's inequality (Proposition (III.7)) with some constant $C_B > 0$.

Then

$$\begin{aligned}
&2\exp\left\{-C_B \min\left\{\frac{t^2\mu_T}{K^2}, \frac{t\mu_T}{K}\right\}\right\} + (\mu_{a_T} - 1)\beta(a_T) \\
&\quad \le 2\exp\left\{-C_B \min\left\{\frac{t^2\mu_T}{K^2}, \frac{t\mu_T}{K}\right\}\right\} + (\mu_T - 1)\exp\{-c_\beta a_T\}
\end{aligned}$$

So,

$$\mathbb{P}[\frac{2}{T}|\|Z_o\|_2^2 - \mathbb{E}(\|Z_o\|_2^2)| > t] \le 2\exp\left\{-C_B \min\left\{\frac{t^2\mu_T}{K^2}, \frac{t\mu_T}{K}\right\}\right\} + (\mu_T - 1)\exp\{-c_\beta a_T\}$$

Taking the union bound over the odd and even blocks,

$$\mathbb{P}[\frac{1}{T}|\|Z\|_2^2-\mathbb{E}(\|Z\|_2^2)| > t] \leq 4\exp\left\{-C_B\min\left\{\frac{t^2\mu_T}{K^2},\frac{t\mu_T}{K}\right\}\right\}+2(\mu_T-1)\exp\{-c_\beta a_T\}$$

For $0 < t < K$, it reduces to

$$\mathbb{P}[\frac{1}{T}|\|Z\|_2^2-\mathbb{E}(\|Z\|_2^2)| > t] \leq 4\exp\left\{-C_B\frac{t^2\mu_T}{K^2}\right\}+2(\mu_T-1)\exp\{-c_\beta a_T\}$$

For the remainder block, since $\|Z_r\|_2^2$ has sub-exponential constant at most $a_TK \leq KT/(2\mu_T)$, we have

$$\mathbb{P}\left[\frac{1}{T}|\|Z_r\|_2^2-\mathbb{E}(\|Z_r\|_2^2)| > t\right] \leq \exp\left(\frac{-tT}{a_TK}\right) \leq \exp\left(\frac{-2t\mu_T}{K}\right)$$

Together, by union bound

$$\mathbb{P}\left[\frac{1}{T}|\|Z\|_2^2-\mathbb{E}(\|Z\|_2^2)| > t\right] \leq 4\exp\{-C_B\frac{t^2\mu_T}{K^2}\}+2(\mu_T-1)\exp\{-c_\beta a_T\}+\exp\{\frac{-2t\mu_T}{K}\}$$

□

Proof of Proposition III.2. Recall that the sequence $X_1,\cdots,X_T \in \mathbb{R}^p$ form a β-mixing and stationary sequence.

Now, fix a unit vector $v \in \mathbb{R}^p$, $\|v\|^2 = 1$.

Define real valued random variables $Z_t = X_t'v,\ t = 1,\cdots,T$. Note that the β mixing rate of $\{Z_t\}_{t=1}^T$ is bounded by the same of $\{X_t\}_{t=1}^T$ by Fact 7. We suppress the X subscript of the sub-Gaussian constant $\sqrt{K_X}$ here, and refer it as $\sqrt{K}$.

We can apply Lemma III.1 on $Z := \{Z_t\}_{t=1}^T$. Set $t = bK$. We have,

$$\begin{aligned}\mathbb{P}\left[\frac{1}{T}|\|Z\|_2^2-\mathbb{E}(\|Z\|_2^2)| > bK\right] &\leq 4\exp\left\{-C_Bb^2\mu_T\right\}+2(\mu_t-1)\exp\{-c_\beta a_t\}+\exp\{-b\mu_T\}\\ &\leq 5\exp\{-\min\{C_B,2\}b^2\mu_T\}+2(\mu_t-1)\exp\{-c_\beta a_t\}\end{aligned}$$

Using Lemma F.2 in Basu and Michailidis [2015], we extend the inequality to hold for all vectors $\mathbb{J}(2k)$, the set of unit norm $2s$-sparse vectors. We have

$$\mathbb{P}\left[\sup_{v\in\mathbb{J}(2k)} \frac{1}{T}|\|Z\|_2^2 - \mathbb{E}(\|Z\|_2^2)| > bK\right]$$
$$\leq 5\exp\{-Cb^2\mu_T + 3k\log(p)\} + 2(\mu_t - 1)\exp\{-c_\beta a_t + 3k\log(p)\}$$

The constant C is defined as $C := \min\{C_B, 2\}$.

Recall $\hat{\Gamma} := \frac{\mathbf{X}'\mathbf{X}}{T}$, the above concentration can be equivalently expressed as

$$\mathbb{P}\left[\sup_{v\in\mathbb{J}(2k)} \left|v'\left(\hat{\Gamma} - \Sigma_X(0)\right)v\right| \leq bK\right]$$
$$\geq 1 - 5\exp\{-Cb^2\mu_T + 3k\log(p)\} - 2(\mu_t - 1)\exp\{-c_\beta a_t + 3k\log(p)\}$$

Finally, we will extend the concentration to all $v \in \mathbb{R}^p$ to establish the lower-RE result. By Lemma 12 of Loh and Wainwright [2012], for parameter $k \geq 1$, w.p. at least

$$1 - 5\exp\{-Cb^2\mu_T + 3k\log(p)\} - 2(\mu_t - 1)\exp\{-c_\beta a_t + 3k\log(p)\}$$

we have

$$\left|v'\left(\hat{\Gamma} - \Sigma_X(0)\right)v\right| \leq 27Kb\left[\|v\|^2 + \frac{1}{k}\|v\|_1^2\right]$$

This implies that

$$v'\hat{\Gamma}v \geq \|v\|^2\left[\lambda_{\min}(\Sigma_X(0)) - 27bK\right] - \frac{27bK}{k}\|v\|_1^2$$

w.p. $1 - 5\exp\{-Cb^2\mu_T + 3k\log(p)\} - 2(\mu_t - 1)\exp\{-c_\beta a_t + 3k\log(p)\}$.

Now, choose set $k = \frac{1}{6\log(p)} \min\{Cb^2\mu_t, c_\beta a_T\}$. Let's choose that, for some $\xi \in (0,1)$, $a_t = T^\xi$ and $\mu_T = T^{1-\xi}$. Then,

$$k = c\frac{1}{\log(p)}\min\{a_T, \mu_T\} = c\frac{1}{\log(p)}\min\{T^\xi, T^{1-\xi}\}$$

Where $c = \frac{1}{6}\max\{c_\beta, Cb^2\}$. To ensure $k \geq 1$, we require $T \geq \left(\frac{1}{c}\log(p)\right)^{\min\left\{\frac{1}{\xi}, \frac{1}{1-\xi}\right\}}$

With these specifications, We have for probability at least

$$1 - 5\exp\{-Cb^2T^{\frac{1}{2}}\} - 2(T^{\frac{1}{2}} - 1)\exp\{-c_\beta T^{\frac{1}{2}}/2\}$$

that

$$v'\hat{\Gamma}v \geq \|v\|^2\left[\lambda_{\min}(\Sigma_X(0)) - 27bK\right] - \frac{27bK\log(p)}{c\min\{T^\xi, T^{1-\xi}\}}\|v\|_1^2.$$

Now, choose $\xi = \frac{1}{2}$ since it optimizes the rate of decay in the tolerance parameter. Also, choose $b = \min\{\frac{1}{54K}\lambda_{\min}(\Sigma_X(0)), 1\}$; this ensures that $\lambda_{\min}(\Sigma_X(0)) - 27bK \geq \frac{1}{2}\lambda_{\min}(\Sigma_X(0))$.

In all, for $T \geq \left(\frac{1}{c}\log(p)\right)^2$ w.p. at least

$$1 - 5\exp\{-Cb^2T^{\frac{1}{2}}\} - 2(T^{\frac{1}{2}} - 1)\exp\{-c_\beta T^{\frac{1}{2}}/2\}$$

$$v'\hat{\Gamma}v \geq \|v\|^2\frac{1}{2}\lambda_{\min}(\Sigma_X(0)) - \frac{27bK\log(p)}{cT^{\frac{1}{2}}}\|v\|_1^2.$$

□

Proof of Proposition III.3.

Recall $|||\mathbf{X}'\mathbf{W}|||_\infty = \max_{1\leq i\leq p, 1\leq j\leq q}|[\mathbf{X}'\mathbf{W}]_{i,j}| = \max_{1\leq i\leq p, 1\leq j\leq q}|\mathbf{X}'_{:i}\mathbf{W}_{:j}|$.

By lemma condition (11), we have

$$\begin{aligned}
&\mathbb{E}\mathbf{X}_{:i} = 0, \forall i \quad \text{and} \\
&\mathbb{E}\mathbf{Y}_{:j} = 0, \forall j
\end{aligned}$$

By first order optimality of the optimization problem in (3.2.1), we have

$$\mathbb{E}\mathbf{X}_{:i}'(\mathbf{Y} - \mathbf{X}\Theta^\star) = 0, \forall i \Rightarrow \mathbb{E}\mathbf{X}_{:i}{}'\mathbf{W}_{:j} = 0, \forall i, j$$

We know $\forall i, j$

$$\begin{aligned}
|\mathbf{X}_{:i}'\mathbf{W}_{:j}| &= |\mathbf{X}_{:i}'\mathbf{W}_{:j} - \mathbb{E}[\mathbf{X}_{:i}'\mathbf{W}_{:j}]| \\
&= \frac{1}{2}| \left(\|\mathbf{X}_{:i} + \mathbf{W}_{:j}\|^2 - \mathbb{E}[\|\mathbf{X}_{:i} + \mathbf{W}_{:j}\|^2]\right) \\
&\quad - \left(\|\mathbf{X}_{:i}\|^2 - \mathbb{E}[\|\mathbf{X}_{:i}\|^2]\right) - \left(\|\mathbf{W}_{:j}\|^2 - \mathbb{E}[\|\mathbf{W}_{:j}\|^2]\right) | \\
&\le \frac{1}{2}\left|\|\mathbf{X}_{:i} + \mathbf{W}_{:j}\|^2 - \mathbb{E}[\|\mathbf{X}_{:i} + \mathbf{W}_{:j}\|^2]\right| \\
&\quad + \frac{1}{2}\left|\|\mathbf{X}_{:i}\|^2 - \mathbb{E}[\|\mathbf{X}_{:i}\|^2]\right| + \frac{1}{2}\left|\|\mathbf{W}_{:j}\|^2 - \mathbb{E}[\|\mathbf{W}_{:j}\|^2]\right|
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{T}|\mathbf{X}_{:i}'\mathbf{W}_{:j}| > 3t\right) \\
&\le \mathbb{P}\left(\frac{1}{2T}\left|\|\mathbf{X}_{:i} + \mathbf{W}_{:j}\|^2 - \mathbb{E}[\|\mathbf{X}_{:i} + \mathbf{W}_{:j}\|^2]\right| > t\right) + \mathbb{P}\left(\frac{1}{2T}\left|\|\mathbf{X}_{:i}\|^2 - \mathbb{E}[\|\mathbf{X}_{:i}\|^2]\right| > t\right) \\
&\quad + \mathbb{P}\left(\frac{1}{2T}\left|\|\mathbf{W}_{:j}\|^2 - \mathbb{E}[\|\mathbf{W}_{:j}\|^2]\right| > t\right)
\end{aligned}$$

This suggests proof strategy via controlling tail probability on each of the terms $|\|\mathbf{X}_{:i}\|^2 - \mathbb{E}[\|\mathbf{X}_{:i}\|^2]|$, $|\|\mathbf{W}_{:j}\|^2 - \mathbb{E}[\|\mathbf{W}_{:j}\|^2]|$ and $|\|\mathbf{X}_{:i} + \mathbf{W}_{:j}\|^2 - \mathbb{E}[\|\mathbf{X}_{:i} + \mathbf{W}_{:j}\|^2]|$. Assuming the conditions in lemma III.3, we can apply lemma III.1 on each of them. We have to figure out their sub-Gaussian constants.

Let's define

$$K_{\mathbf{W}} := \sup_{1\leq t\leq T, 1\leq j\leq q} \|\mathbf{W}_{tj}\|_{\psi_2}$$

and

$$K_{\mathbf{X}+\mathbf{W}} := \sup_{1\leq t\leq T, 1\leq j\leq q, 1\leq i\leq p} \|\mathbf{X}_{ti} + \mathbf{W}_{tj}\|_{\psi_2}$$

. We have to figure out the constants $K_{\mathbf{W}}$ and $K_{\mathbf{W}+\mathbf{X}}$.

Now,

$$\begin{aligned} \sup_{1\leq t\leq T} \sup_{1\leq i\leq q} \|\mathbf{W}_{ti}\|_{\psi_2} &\leq \sup_{1\leq t\leq T} \|\mathbf{W}_{t:}\|_{\psi_2} && \text{by definition of sub-Gaussian random vector} \\ &= \|\mathbf{W}_{1:}\|_{\psi_2} && \text{by stationarity} \end{aligned}$$

Let's figure out $\|\mathbf{W}_{1:}\|_{\psi_2}$,

$$\begin{aligned} \mathbf{W}_{1:} &= \mathbf{Y}_{1:} - (\mathbf{X}\Theta^{\star})_{1:} \\ &= \mathbf{Y}_{1:} - \mathbf{X}_{1:}\Theta^{\star} \end{aligned}$$

Thus,

$$\begin{aligned} \|\mathbf{W}_{1:}\|_{\psi_2} &\leq \|\mathbf{Y}_{1:}\|_{\psi_2} + \|\mathbf{X}_{1:}\Theta^{\star}\|_{\psi_2} && \text{since } \|\cdot\|_{\psi_2} \text{ is a norm} \\ &\leq \|\mathbf{Y}_{1:}\|_{\psi_2} + \|\mathbf{X}_{1:}\|_{\psi_2} |||\Theta^{\star}||| && \text{by lemma III.6} \\ &= \sqrt{K_Y} + |||\Theta^{\star}|||\sqrt{K_X} && \text{by stationarity} \end{aligned}$$

Therefore,

$$K_{\mathbf{W}} \leq \sqrt{K_Y} + |||\Theta^{\star}|||\sqrt{K_X} \tag{3.6.3}$$

Similarly,

$$\begin{aligned}\sup_{1\le i\le p, 1\le j\le q, 1\le t\le T} \|\mathbf{X}_{ti}+\mathbf{W}_{tj}\|_{\psi_2} &\le \sup_{1\le i\le p, 1\le t\le T}\|\mathbf{X}_{ti}\|_{\psi_2} + \sup_{1\le j\le q, 1\le t\le T}\|\mathbf{W}_{tj}\|_{\psi_2}\\ &\le \|\mathbf{X}_{1:}\|_{\psi_2} + \|\mathbf{W}_{1:}\|_{\psi_2}\\ &\le \sqrt{K_Y} + \sqrt{K_X}\,(1+|||\Theta^\star|||)\end{aligned}$$

where the last inequality follows from equation (3.6.3).

Therefore,

$$K_{\mathbf{X}+\mathbf{W}} \le \sqrt{K_Y} + \sqrt{K_X}\,(1+|||\Theta^\star|||) \tag{3.6.4}$$

Take

$$K := \max\{K_{\mathbf{X}}, K_{\mathbf{W}}, K_{\mathbf{X}+\mathbf{W}}\} \le \sqrt{K_Y} + \sqrt{K_X}\,(1+|||\Theta^\star|||) \tag{3.6.5}$$

For $\xi \in [0,1]$, set $a_T = T^\xi$ and $\mu_T = T^{1-\xi}$. Applying lemma III.1 three times with sub-Gaussian constant K, we have

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{T}\left|\mathbf{X}'_{:i}\mathbf{W}_{:j}\right| > 3t\right)\\ &\le \mathbb{P}\left(\frac{1}{2T}\left|\|\mathbf{X}_{:i}+\mathbf{W}_{:j}\|^2 - \mathbb{E}[\|\mathbf{X}_{:i}+\mathbf{W}_{:j}\|^2]\right| > t\right)\\ &\quad + \mathbb{P}\left(\frac{1}{2T}\left|\|\mathbf{X}_{:i}\|^2 - \mathbb{E}[\|\mathbf{X}_{:i}\|^2]\right| > t\right)\\ &\quad + \mathbb{P}\left(\frac{1}{2T}\left|\|\mathbf{W}_{:j}\|^2 - \mathbb{E}[\|\mathbf{W}_{:j}\|^2]\right| > t\right)\\ &\le \ 4\exp\{-C_B\frac{4t^2T^{1-\xi}}{K^4}\} + 2(T^{1-\xi}-1)\exp\{-c_\beta T^\xi\} + \exp\{-\frac{2}{K^2}tT^{1-\xi}\}\}\\ &\quad + 4\exp\{-C_B\frac{4t^2T^{1-\xi}}{K^4}\} + 2(T^{1-\xi}-1)\exp\{-c_\beta T^\xi\} + \exp\{-\frac{2}{K^2}tT^{1-\xi}\}\}\\ &\quad + 4\exp\{-C_B\frac{4t^2T^{1-\xi}}{K^4}\} + 2(T^{1-\xi}-1)\exp\{-c_\beta T^\xi\} + \exp\{-\frac{2}{K^2}tT^{1-\xi}\}\}\end{aligned}$$

By union bound,

$$\begin{aligned}
\mathbb{P}[\frac{1}{T}\|\!|\mathbf{X}'\mathbf{W}\|\!|_{\infty} > 3t] &= \mathbb{P}[\max_{1\leq i\leq p,\, 1\leq j\leq q} \frac{1}{T}|\mathbf{X}'_{:i}\mathbf{W}_{:j}| > 3t] \\
&\leq 3pq\left\{4\exp\{-C_B\frac{4t^2T^{1-\xi}}{K^4}\} + 2(T^{1-\xi}-1)\exp\{-c_\beta T^{\xi}\} + \exp\{-\frac{2}{K^2}tT^{1-\xi}\}\right\} \\
&= 12\exp\{-C_B\frac{4t^2T^{1-\xi}}{K^4} + \log\{pq\}\} \\
&\quad + 6(T^{1-\xi}-1)\exp\{-c_\beta T^{\xi} + \log\{pq\}\} \\
&\quad + 3\exp\{-\frac{2}{K^2}tT^{1-\xi} + \log\{pq\}\}
\end{aligned}$$

To ensure proper decay in the probability, we require

$$T \geq \max\left\{\left(\log(pq)\max\left\{\frac{K^4}{2C_B}, K^2\right\}\right)^{\frac{1}{1-\xi}}, \left[\frac{2}{c_\beta}\log(pq)\right]^{\frac{1}{\xi}}, \right\}$$

With

$$t := \sqrt{\frac{K^4\log(pq)}{2T^{1-\xi}C_B}}$$

$$\begin{aligned}
\mathbb{P}\left[\frac{1}{T}\|\!|\mathbf{X}'\mathbf{W}\|\!|_{\infty} > \sqrt{\frac{72K^4\log(pq)}{T^{1-\xi}C_B}}\right] &\leq 15\exp\left\{-\frac{1}{2}\log(pq)\right\} \\
&\quad + 6(T^{1-\xi}-1)\exp\left\{-\frac{1}{2}c_\beta T^{\xi}\right\}
\end{aligned}$$

where $K = \sqrt{K_{\mathbf{Y}}} + \sqrt{K_X}\,(1 + \|\!|\Theta^\star\|\!|)$

□

Lemma III.6. *For any sub-Gaussian random vector X and non-stochastic matrix $\boldsymbol{A}$. We have*

$$\|\boldsymbol{A}X\|_{\psi_2} \leq \|\!|\boldsymbol{A}\|\!|\,\|X\|_{\psi_2}$$

Proof. We have,

$$\begin{aligned}\|\mathbf{A}X\|_{\psi_2} &= \sup_{\|v\|_2\le 1} \|v'\mathbf{A}X\|_{\psi_2}\\ &= \sup_{\|v\|_2\le 1} \|(\mathbf{A}'v)'X\|_{\psi_2}\\ &\le \sup_{\|u\|_2\le |||\mathbf{A}|||} \|u'X\|_{\psi_2}\\ &= |||\mathbf{A}||| \sup_{\|u\|_2\le 1} \|u'X\|_{\psi_2}\\ &= |||\mathbf{A}|||\, \|X\|_{\psi_2}\,.\end{aligned}$$

□

3.6.4 Bernstein's Concentration Inequality

We state the Bernstein's inequality [Vershynin, 2010, Proposition 5.16] below for completeness.

Proposition III.7 (Bernstein's Inequality)**.** *Let $X_1, \cdots, X_N$ be independent centered sub-exponential random variables, and $K = \max_i \|X_i\|_{\psi_1}$. Then for every $a = (a_1, \cdots, a_N) \in \mathbb{R}^N$ and every $t \ge 0$, we have*

$$\mathbb{P}\left\{\left|\sum_{i=1}^N a_i X_i\right| \ge t\right\} \le 2\exp\left[-C_B \min\left(\frac{t^2}{K^2\|a\|_2^2}, \frac{t}{K\|a\|_\infty}\right)\right]$$

where $C_B > 0$ is an absolute constant.

3.6.5 Verification of Assumptions for the Examples

VAR

Formally a finite order Gaussian VAR(d) process is defined as follows. Consider a sequence of serially ordered random vectors $(Z_t)_{t=1}^{T+d}$, $Z_t \in \mathbb{R}^p$ that admits the following auto-regressive representation:

$$Z_t = \mathbf{A}_1 Z_{t-1} + \cdots + \mathbf{A}_d Z_{t-d} + \mathcal{E}_t \tag{3.6.6}$$

where each $\mathbf{A}_k, k = 1, \ldots, d$ is a non-stochastic coefficient matrix in $\mathbb{R}^{p\times p}$ and innovations $\mathcal{E}_t$ are p-dimensional random vectors from $\mathcal{N}(0, \Sigma_\epsilon)$. Assume $\lambda_{\min}(\Sigma_\epsilon) > 0$ and $\lambda_{\max}(\Sigma_\epsilon) < \infty$.

Note that every VAR(d) process has an equivalent VAR(1) representation (see e.g. [Lütkepohl, 2005, Ch 2.1]) as

$$\tilde{Z}_t = \tilde{\mathbf{A}}\tilde{Z}_{t-1} + \tilde{\mathcal{E}}_t \tag{3.6.7}$$

where

(3.6.8)

$$\tilde{Z}_t := \begin{bmatrix} Z_t \\ Z_{t-1} \\ \vdots \\ Z_{t-d+1} \end{bmatrix}_{(pd\times 1)} \quad \tilde{\mathcal{E}}_t := \begin{bmatrix} \mathcal{E}_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{(pd\times 1)} \quad \text{and} \quad \tilde{A} := \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 & \cdots & \mathbf{A}_{d-1} & \mathbf{A}_d \\ \mathbf{I}_p & 0 & 0 & 0 & 0 \\ 0 & \mathbf{I}_p & & 0 & 0 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \mathbf{I}_p & 0 \end{bmatrix}_{(dp\times dp)}$$

Because of this equivalence, justification of Assumption 13 will operate through this corresponding augmented VAR(1) representation.

For both Gaussian and sub-Gaussian VARs, Assumption 11 is true since the sequences (Z_t) is centered. Second, $\Theta^\star = (\mathbf{A}_1, \cdots, \mathbf{A}_d)$. So Assumption 9 follows from construction.

For the remaining Assumptions, we will consider the Gaussian and sub-Gaussian cases separately.

Gaussian VAR (Z_t) satisfies Assumption 12 by model assumption.

To show that (Z_t) is β-mixing with geometrically decaying coefficients, we use the following facts together with the equivalence between (Z_t) and $(\tilde{Z}_t)$ and Fact 7.

Since $(\tilde{Z}_t)$ is stable, the spectral radius of $\tilde{A}$, $r(\tilde{A}) < 1$, hence Assumption 10 holds. Also the innovations $\tilde{\mathcal{E}}$ has finite first absolute moment and positive support everywhere. Then, according to Theorem 4.4 in Tjøstheim [1990], $(\tilde{Z}_t)$ is *geometrically ergodic*. Note here that Gaussianity is *not* required here. Hence, it also applies to innovations from mixture of Gaussians.

Next, we present a standard result (see e.g. [Liebscher, 2005, Proposition 2]).

Fact 10. A stationary Markov chain $\{Z_t\}$ is geometrically ergodic implies $\{Z_t\}$ is *absolutely regular*(a.k.a. β-mixing) with

$$\beta(n) = O(\gamma^n), \ \gamma^n \in (0, 1)$$

So, Assumption 13 holds.

Sub-Gaussian VAR When the innovations are random vectors from the uniform distribution, they are sub-Gaussian. That (Z_t) are sub-Gaussian follows from arguments as in Chapter 3.6.5 with $\Sigma(\cdot)$ set to be the identity operator in this case. So, Assumption 12 holds.

To show that (Z_t) satisfies Assumptions 10 and 13, we establish that (Z_t) is geometrically ergodic. To show the latter, we use Propositions 1 and 2 in Liebscher [2005] together with the equivalence between (Z_t) and $(\tilde{Z}_t)$ and Fact 7.

To apply Proposition 1 in Liebscher [2005], we check the three conditions one by one. Condition (i) is immediate with $m = 1$, $E = \mathbb{R}^p$, and μ is the Lebesgue measure. For condition (ii), we set $E = \mathbb{R}^p$, μ to be the Lebesgue measure, and $\bar{m} = \lceil \inf_{u \in C, v \in A} \|u - v\|_2 \rceil$ the minimum "distance" between the sets C and A. Because C is bounded and A Borel, $\bar{m}$ is finite. Lastly, for condition (iii), we again let

$E = \mathbb{R}^p$, μ to be the Lebesgue measure, and now the function $Q(\cdot) = \|\cdot\|$ and the set $K = \{x \in \mathbb{R}^p : \|x\| \le \frac{2\mathbb{E}\left\|\tilde{\mathcal{E}}_t\right\|}{c\epsilon}\}$ where $c = 1 - \left|\left|\left|\tilde{A}\right|\right|\right|$. Then,

- Recall from model assumption that $\left|\left|\left|\tilde{A}\right|\right|\right| < 1$; hence,

  $$\mathbb{E}\left[\left\| \tilde{Z}_{t+1}\right\| \middle| \tilde{Z}_t = z\right] < \left|\left|\left|\tilde{A}\right|\right|\right| \|z\| + \mathbb{E}(\left\|\tilde{\mathcal{E}}_{t+1}\right\|) \le \left(1 - \frac{c}{2}\right) \|z\| - \epsilon,$$

  for all $z \in E \backslash K$

- For all $z \in K$,

  $$\mathbb{E}\left[\left\| \tilde{Z}_{t+1}\right\| \middle| \tilde{Z}_t = z\right] < \left|\left|\left|\tilde{A}\right|\right|\right| \|z\| + \mathbb{E}(\left\|\tilde{\mathcal{E}}_{t+1}\right\|) \le \left|\left|\left|\tilde{A}\right|\right|\right| \frac{2\mathbb{E}\left\|\tilde{\mathcal{E}}_t\right\|}{c\epsilon}$$

- For all $z \in K$,

  $$0 \le \|z\| \le \frac{2\mathbb{E}\left\|\tilde{\mathcal{E}}_t\right\|}{c\epsilon}$$

Now, by Proposition 1 in Liebscher [2005], $(\tilde{Z}_t)$ is geometrically ergodic; hence $(\tilde{Z}_t)$ will be stationary. Once it reaches stationarity, by Proposition 2 in the same paper, the sequence will be β-mixing with geometrically decaying mixing coefficients. Therefore, Assumptions 10 and 13 hold.

VAR with Misspecification

Assumptions: Assumption 11 is immediate from model definitions. By the same arguments as in Chapter 3.6.5, (Z_t, Ξ_t) are stationary and so is the sub-process (Z_t); Assumption 10 holds. Again, (Z_t, Ξ_t) satisfy Assumption 13 according to Chapter 3.6.5. By Fact 7, we have the same Assumptions hold for the respective sub-processes (Z_t) in both cases. Assumption 12 holds by the same reasoning as in Chapter 3.6.5.

To show that $(\Theta^\star)' = A_{ZZ} + A_{Z\Xi}\Sigma_{\Xi Z}(0)(\Sigma_Z(0))^{-1}$, consider the following arguments. By Assumption 10, we have the auto-covariance matrix of the whole system (Z_t, Ξ_t) as

$$\Sigma_{(Z,\Xi)} = \begin{bmatrix} \Sigma_X(0) & \Sigma_{X\Xi}(0) \\ \Sigma_{\Xi X}(0) & \Sigma_{\Xi}(0) \end{bmatrix}$$

Recall our $\Theta^\star$ definition from Eq. (3.2.1)

$$\Theta^\star := \operatorname*{arg\,min}_{B \in \mathbb{R}^{p\times p}} \mathbb{E}\left(\|Z_t - B'Z_{t-1}\|_2^2 \right)$$

Taking derivatives and setting to zero, we obtain

$$(\Theta^\star)' = \Sigma_Z(-1)(\Sigma_Z)^{-1} \tag{3.6.9}$$

Note that

$$\begin{aligned} \Sigma_Z(-1) &= \Sigma_{(Z,\Xi)}(-1)[1:p_1, 1:p_1] \\ &= \mathbb{E}\left(A_{ZZ}Z_{t-1} + A_{Z\Xi}\Xi_{t-1} + \mathcal{E}_{Z,t-1}\right) Z_{t-1}' \\ &= \mathbb{E}\left(A_{ZZ}Z_{t-1}Z_{t-1}' + A_{Z\Xi}\Xi_{t-1}Z_{t-1}' + \mathcal{E}_{Z,t-1}Z_{t-1}'\right) \\ &= A_{ZZ}\Sigma_Z(0) + A_{Z\Xi}\Sigma_{\Xi Z}(0) \end{aligned}$$

by Assumptions 10 and the fact that the innovations are iid.

Naturally,

$$(\Theta^\star)' = A_{ZZ}\Sigma_Z(0)(\Sigma_Z(0))^{-1} + A_{Z\Xi}\Sigma_{\Xi Z}(0)(\Sigma_Z(0))^{-1} = A_{ZZ} + A_{Z\Xi}\Sigma_{\Xi Z}(0)(\Sigma_Z(0))^{-1}$$

Remark 14. Notice that $A_{Z\Xi}$ is a column vector and suppose it is 1-sparse, and A_{ZZ} is p-sparse, then $\Theta^\star$ is at most $2p$-sparse. So Assumption 9 can be built in by model construction.

Remark 15. We gave an explicit model here where the left out variable Ξ was univariate. That was only for convenience. In fact, whenever the set of left-out variables Ξ

affect only a small set of variables Ξ in the retained system Z, the matrix $\Theta^\star$ is guaranteed to be sparse. To see that, suppose $\Xi \in \mathbb{R}^q$ and $A_{Z\Xi}$ has at most s_0 non-zero rows (and let A_{ZZ} to be s-sparse as always), then $\Theta^\star$ is at most $(s_0 p + s)$-sparse.

Remark 16. Any VAR(d) process has an equivalent VAR(1) representation (Lutkepohl 2005). Our results extend to any VAR(d) processes.

ARCH

Verifying the Assumptions. To show that Assumption 13 holds for a process defined by Eq. (3.4.2) we leverage on Theorem 2 from Liebscher [2005]. Note that the original ARCH model in Liebscher [2005] assumes the innovations to have positive support everywhere. However, this is just a convenient assumption to establish the first two conditions in Proposition 1 (on which proof of Theorem 2 relies) from the same paper. Our example ARCH model with innovations from the uniform distribution also satisfies the first two conditions of Proposition 1 by the same arguments in the *Sub-Gaussian* paragraph of Chapter 3.6.5.

Theorem 2 tells us that for our ARCH model, if it satisfies the following conditions, it is guaranteed to be absolutely regular with geometrically decaying β-coefficients.

- $\mathcal{E}_t$ has positive density everywhere on $\mathbb{R}^p$ and has identity covariance by construction.
- $\Sigma(z) = o(\|z\|)$ because $m \in (0, 1)$.
- $|||\Sigma(z)^{-1}||| \le 1/(ac)$, $|\det(\Sigma(z))| \le bc$
- $r(A) \le |||A||| < 1$

So, Assumption 13 is valid here. We check other assumptions next.

Mean 0 is immediate, so we have Assumption 11. When the Markov chain did not start from a stationary distribution, geometric ergodicity implies that the sequence is approaching the stationary distribution exponentially fast. So, after a burning period, we will have Assumption 10 approximately valid here.

The sub-Gaussian constant of $\Sigma(Z_{t-1})\mathcal{E}_t$ given $Z_{t-1} = z$ is bounded as follows: for every z,

$$\begin{aligned} \|\Sigma(z)\mathcal{E}_t\|_{\psi_2} &\leq |||\Sigma(z)||| \, \|\mathcal{E}_t\|_{\psi_2} && \text{by Lemma III.6} \\ &\leq C|||\Sigma(z)||| \cdot \|e_1'\mathcal{E}_t\|_{\psi_2} \\ &\leq C|||\Sigma(z)||| \cdot \left\| U\left(-\sqrt{3}, \sqrt{3}\right) \right\|_{\psi_2} \\ &\leq C'cb =: K_E \end{aligned}$$

The second inequality follows since $\mathcal{E}_t \overset{iid}{\sim} U\left(\left[-\sqrt{3}, \sqrt{3}\right]^p\right)$ and a standard result that

Fact 11. Let $X = (X_1, \cdots, X_p) \in \mathbb{R}^p$ be a random vector with independent, mean zero, sub-Gaussian coordinates X_i. Then X is a sub-Gaussian random vector, and there exists a positive constant C for which

$$\|X\|_{\psi_2} \leq C \cdot \max_{i \leq p} \|X_i\|_{\psi_2}$$

The forth inequality follows since the sub-Gaussian norm of a bounded random variable is also bounded.

By the recursion for Z_t, we have

$$\|Z_t\|_{\psi_2} \leq |||A||| \, \|Z_{t-1}\|_{\psi_2} + K_E.$$

which yields the bound $\|Z_t\|_{\psi_2} \leq K_E/(1 - |||A|||) < \infty$. Hence Assumption 12 holds.

We will show below that $\Theta^\star = A'$. Hence, sparsity (Assumption 9) can be built in when we construct our model 3.4.2.

Recall Eq. 3.6.9 from Chapter 3.6.5 that

$$\Theta^\star = \Sigma_Z(-1)(\Sigma_Z)^{-1}$$

Now,

$$\begin{aligned}
\Sigma_Z(-1) &= \mathbb{E}Z_t Z_{t-1}' && \text{by stationarity} \\
&= \mathbb{E}\left(AZ_{t-1} + \Sigma(Z_{t-1})\mathcal{E}_t\right) Z_{t-1}' && \text{Eq. (3.4.2)} \\
&= A\mathbb{E}Z_{t-1}Z_{t-1}' + \mathbb{E}\Sigma(Z_{t-1})\mathcal{E}_t Z_{t-1}' \\
&= A\Sigma_Z + \mathbb{E}[c\,\mathrm{clip}_{a,b}\left(\|Z_{t-1}\|^m\right)\mathcal{E}_t Z_{t-1}'] \\
&= A\Sigma_Z + \mathbb{E}[c\mathcal{E}_t Z_{t-1}'\mathrm{clip}_{a,b}\left(\|Z_{t-1}\|^m\right)] \\
&= A\Sigma_Z + c\mathbb{E}\left[\mathcal{E}_t\right]\mathbb{E}\left[Z_{t-1}'\mathrm{clip}_{a,b}\left(\|Z_{t-1}\|^m\right)\right] && \text{i.i.d. innovations} \\
&= A\Sigma_Z && \mathcal{E}_t \text{ mean } 0,
\end{aligned}$$

where $\mathrm{clip}_{a,b}\left(x\right) := \min\{\max\{x, a\}, b\}$ for $b > a$.

Since Σ_Z is invertible, we have $(\Theta^\star)' = \Sigma_Z(-1)(\Sigma_Z)^{-1} = A$.

CHAPTER IV

Lasso Guarantees for Gaussian Vector Autoregressive Processes with Missing or Noisy Data

4.1 Introduction

The information age and scientific advances have led to explosions in large data sets as well as new statistical methods and algorithms aimed at extracting valuable information in them. On the data side, it is common to see massive dependent data collected from, for example, micro-array experiments, social networks, mobile phone usage, high frequency stock market trading, daily grocery sales, etc. At the same time, the high speed Internet makes big data warehouses readily accessible.

The surge in big data has stimulated exciting developments in statistical models and algorithms to exploit these data sets. Among the plethora of methods, the linear parametric models remain highly popular thanks to its superiority in interpretability, computational efficiency and the rich and sophisticated theoretical literature. The vector autoregressive (VAR) models are a linear parametric family that allows researchers to model interrelationships among variables that exhibit temporal dependence. When we assume the innovations in VAR are Gaussian, observations generated according to a VAR are also Gaussian. There are in general two objects of interest in the estimation – inverse covariance matrix and transition matrices. The former amounts to estimating contemporaneous correlations between components and the latter directed correlations between components at different time points. We focus on the latter in this paper.

The main difficulty in high dimensional estimation of the VAR lies in the fact that the number of parameters far out number that of the data points we usually have at our disposal. As a result, consistent estimation is impossible without imposing restrictions on the parameter space. Sparsity is a common and natural assumption to make when we believe that evolution of variables do not depend on all the others. This corresponds to an ℓ_0 constrain on the parameters which is non-convex and hence poses serious challenges on the computational side. Researchers often relax it to the closest convex ℓ_1 constraint. The ℓ_1 penalty on the parameters together with the square error loss constitutes the (Lagrangian form) celebrated lasso procedure.

The literature of lasso, however, mainly focuses on the iid sample scenario, leaving that of the dependent data case relatively sparse. This paper serves to provide theoretical guarantees for lasso on VAR data when the data are either corrupted or missing completely at random. The analysis crucially depends on the results from the work of Basu and Michailidis [2015] and Loh and Wainwright [2012]. The set of lasso guarantees presented here extends the previous work in two ways: (1) we remove the restriction on operator norm of transition matrix being smaller than 1 as in Loh and Wainwright [2012] and thus generalize the guarantees to any stationary VAR(d) models, and (2) by incorporating the modified lasso framework from Loh and Wainwright [2012], we generalize the results in Basu and Michailidis [2015] to addressing cases of corrupted and missing data.

We give a brief review of historical development of the lasso theory in VAR estimation below.

Brief Review of Lasso Theory on VAR Several researchers [Bickel et al., 2009, Loh and Wainwright, 2012, Negahban et al., 2012] have analyzed the lasso and established consistency of it under sufficient conditions – restricted strong convexity(RSC) or restricted eigenvalue (RE) for square error losses on the gram matrix ${}^{\mathbf{X}'\mathbf{X}}/_{N}$ and deviation bound (DB) on correlation between design matrix and error ${}^{\mathbf{X}'\mathbf{E}}/_{N}$ (more details in Section 4.2). Unfortunately, on fixed design matrices, it is in general NP-hard to check RE-type assumptions.

To get around that Raskutti et al. [2010], Rudelson and Zhou [2013] have established high probability guarantees of these conditions on random iid subgaussian data.

Negahban and Wainwright [2011], Loh and Wainwright [2012] provided a more in-depth analysis and showed high probability validity for the RSC and DB conditions for a VAR(1) process $X_t = \mathbf{A}X_{t-1} + \epsilon_t$ under the assumption that operator norm of transition matrix $|||\mathbf{A}||| < 1$.

However, as pointed out by [Basu and Michailidis, 2015, see pg 11-13 in supplementary], $|||\mathbf{A}||| < 1$ is a stringent assumption and in general does not hold for any VAR(d) model for $d > 1$. Basu and Michailidis [2015] took the spectral density route in analyzing the VAR and provided the RE and DB guarantees in terms of stability measure and fundamental properties of the process.

We build upon the work of Basu and Michailidis [2015] and Loh and Wainwright [2012] to give high probabilistic lasso error bounds for general stable Gaussian VAR with missing or corrupted data.

4.2 Preliminaries

4.2.1 Matrix and Vector Notations

For a symmetric matrix $\mathbf{M}$, let $\lambda_{\max}(\mathbf{M})$ and $\lambda_{\min}(\mathbf{M})$ denote its maximum and minimum eigenvalues respectively. For any matrix let $\mathbf{M}$, $r(\mathbf{M})$, $|||\mathbf{M}|||$, $|||\mathbf{M}|||_\infty$, and $|||\mathbf{M}|||_F$ denote its spectral radius $\max_i \{|\lambda_i(\mathbf{M})|\}$, operator norm $\sqrt{\lambda_{\max}(\mathbf{M}'\mathbf{M})}$, entrywise ℓ_∞ norm $\max_{i,j} |\mathbf{M}_{i,j}|$, and Frobenius norm $\sqrt{\operatorname{tr}(\mathbf{M}'\mathbf{M})}$ respectively. For any vector $v \in \mathbb{R}^p$, $\|v\|_q$ denotes its ℓ_q norm $(\sum_{i=1}^p |v_i|^q)^{1/q}$. Unless otherwise specified, we shall use $\|\cdot\|$ to denote the ℓ_2 norm. For any vector $v \in \mathbb{R}^p$, we use $\|v\|_0$ and $\|v\|_\infty$ to denote $\sum_{i=1}^p \mathbb{1}\{v_i \neq 0\}$ and $\max_i\{|v_i|\}$ respectively. Similarly, for any matrix $\mathbf{M}$, $|||\mathbf{M}|||_0 = \|\operatorname{vec}(\mathbf{M})\|_0$ where $\operatorname{vec}(\mathbf{M})$ is the vector obtained from $\mathbf{M}$ by concatenating the rows of M. We say that matrix $\mathbf{M}$ (resp. vector v) is s-*sparse* if $|||\mathbf{M}|||_0 = s$ (resp. $\|v\|_0 = s$). We use v' and $\mathbf{M}'$ to denote the transposes of v and $\mathbf{M}$ respectively. For a given vector $v \in \mathbb{R}^p$, its i-th element is denoted $v[i]$. When we index a matrix, we adopt the following conventions. For any matrix $\mathbf{M} \in \mathbb{R}^{p\times q}$, for $1 \leq i \leq p$, $1 \leq j \leq q$, we define $\mathbf{M}[i,j] \equiv \mathbf{M}_{ij} := e_i'\mathbf{M}e_j$, $\mathbf{M}[i,:] \equiv \mathbf{M}_{i:} := e_i'\mathbf{M}$ and $\mathbf{M}[:,j] \equiv \mathbf{M}_{:j} := \mathbf{M}e_j$ where e_i is the vector with all 0s except for a 1 in the ith coordinate. The set of integers is denoted by $\mathbb{Z}$. Realizations x, y, x are written in lower case letters, random variables and vectors X, Y, Z upper case letters, and random matrices $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ *bold* upper case letters.

For a lag $l \in \mathbb{Z}$, we define the auto-covariance matrix w.r.t. $(X_t, Y_t)_t$ as $\Sigma(l) = \Sigma_{(X;Y)}(l) := \mathbb{E}[(X_t; Y_t)(X_{t+l}; Y_{t+l})']$. Note that $\Sigma(-l) = \Sigma(l)'$. Similarly, the auto-covariance matrix of lag l w.r.t. $(X_t)_t$ is $\Sigma_X(l) := \mathbb{E}[X_t X_{t+l}']$, and w.r.t. $(Y_t)_t$ is $\Sigma_Y(l) := \mathbb{E}[Y_t Y_{t+l}']$. The cross-covariance matrix at lag l is $\Sigma_{X,Y}(l) := \mathbb{E}[X_t Y_{t+l}']$. Note the difference between $\Sigma_{(X;Y)}(l)$ and $\Sigma_{X,Y}(l)$: the former is a $(p+q) \times (p+q)$

matrix, the latter is a $p \times q$ matrix. Thus, $\Sigma_{(X;Y)}(l)$ is a matrix consisting of four sub-matrices. Using Matlab-like notation, $\Sigma_{(X;Y)}(l) = [\Sigma_X, \Sigma_{X,Y}; \Sigma_{Y,X}, \Sigma_Y]$. As per our convention, at lag 0, we omit the lag argument l. For example, $\Sigma_{X,Y}$ denotes $\Sigma_{X,Y}(0) = \mathbb{E}[X_t Y_t']$.

4.2.2 VAR

For a p-dimensional vector-valued stationary time series (X_t), a vector autoregressive model of order d (VAR(d)) with independent Gaussian innovations admits the following representation

$$X_t = \mathbf{A}_1 X_{t-1} + \cdots + \mathbf{A}_d X_{t-d} + \epsilon_t, \text{ where}$$
$$\epsilon_t \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \Sigma_\epsilon),\ \operatorname{Cov}(\epsilon_t, \epsilon^s) = \mathbf{0}\ \forall\, t \neq s$$

The class of VAR models provide a systemic way to model temporal and cross-sectional correlations between variables. One way to interpret a Gaussian VAR model is one via the directed graph. A non-zero entry in the transition matrix $\mathbf{A}_k[i,j]$ can be understood as a directed link from $X_t[j]$ to $X_{t+k}[i]$. An undirected link between $X_t[i]$ and $X_t[j]$ indicates contemporary correlation between the variables and is represented by the non-zero (i,j)th entry in Σ_ϵ^{-1} since the errors are Gaussian.

We note a useful fact: any VAR(d) process has an equivalent VAR(1) representation. Therefore, without loss of generality, we can focus on VAR(1). Consider writing a VAR(d) model

$$X_t = \mathbf{A}_1 X_{t-1} + \cdots + \mathbf{A}_d X_{t-d} + \epsilon_t$$

as a VAR(1):

$$\tilde{X}_t = \tilde{\mathbf{A}}\tilde{X}^{t-1} + \tilde{\epsilon}_t$$

where

$$\tilde{X}_t = \begin{pmatrix} X_t \\ X_{t-1} \\ \vdots \\ X_{t-d+1} \end{pmatrix}_{dp \times 1} \qquad \tilde{\mathbf{A}} = \begin{pmatrix} \mathbf{A}_1 & \mathbf{A}_2 & \cdots & \mathbf{A}_{d-1} & \mathbf{A}_d \\ \mathbf{I}_p & 0 & \cdots & 0 & 0 \\ 0 & \mathbf{I}_p & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \mathbf{I}_p & 0 \end{pmatrix}_{dp \times dp} \qquad \tilde{\epsilon}_t = \begin{pmatrix} \epsilon_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{dp \times 1}$$

In particular, a process (X_t) is stable if and only if its equivalent representation $(\tilde{X}_t)$ is stable.

We denote the cross-covariance matrix with respect to $(\tilde{X}_t)$ as $\Gamma^N_{\tilde{X}}$, such that $\forall i, j \in \{d+1, \cdots, N\}$,

$$\Gamma^N_{\tilde{X}}[i,j] := \mathbb{E}\left[\tilde{X}_i\left(\tilde{X}_j\right)'\right]$$

In particular, if we assume stationarity, the variance matrix of $(\tilde{X}_t)$ is $\Sigma_{\tilde{\mathbf{X}}} := \mathbb{E}\left[\tilde{X}_t(\tilde{X}_t)'\right] \equiv \Gamma^N_{\tilde{X}}(i,i)$.

Stability of a stochastic process is an important concept in time series. Much of the literature focus on the stable time series.

To gain some intuition about stability, note that the a VAR(1) process can be equivalently represented, via backward substitution, as

$$\tilde{X}_t = \tilde{\mathbf{A}}^{j+1}\tilde{X}^{t-j-1} + \sum_{i=0}^{j} \tilde{\mathbf{A}}^{i}\tilde{\epsilon}^{t-i}$$

Stability requires that the infinite sum

$$\sum_{i=1}^{\infty} \tilde{\mathbf{A}^i} \epsilon^{\tilde{t-i}}$$

exists in mean square.

Intuitively, we want $|\lambda_i(\tilde{\mathbf{A}})| < 1$ for all i. Equivalently, it can be shown that the VAR(d) process is stable iff all the eigenvalues of the matrix valued reverse characteristic polynomial $\mathcal{A}(z) = \mathbf{I}_p - \sum_{t=1}^{d} \mathbf{A}_t z_t$ are non-zero on the unit circle $\{z \in \mathbb{C} : |z| = 1\}$. I.e.

$$\det(\mathcal{A}(z)) = \det(\mathbf{I}_p - \sum_{t=1}^{d} \mathbf{A}_t z_t) \neq 0, \ \forall\, |z| \leq 1.$$

As such, stability is a qualitative, all-or-none concept. Basu and Michailidis [2015] defined a quantitative measures of stability for a VAR process:

Stability of the process guarantees that all the eigenvalues of the Hermitian matrix $\mathcal{A}^*(z)\mathcal{A}(z)$ are positive, whenever $|z| = 1$. Hence, we can take the minimum eigenvalue of $\mathcal{A}^*(z)\mathcal{A}(z)$ evaluated at $|z| = 1$ as a measure of stability of the VAR(d) process:

$$\mu_{\min}(\mathcal{A}) = \min_{|z|=1} \Lambda_{\min}(\mathcal{A}^*(z)\mathcal{A}(z))$$

And, a related quantity:

$$\mu_{\max}(\mathcal{A}) = \max_{|z|=1} \Lambda_{\max}(\mathcal{A}^*(z)\mathcal{A}(z))$$

Note that $\mu_{\min}(\mathcal{A})$ and $\mu_{\max}(\mathcal{A})$ are well-defined because of the continuity of eigenvalues and compactness of the unit circle $\{z \in \mathbb{C} : |z| = 1\}$. Further, these are positive if the corresponding process is stable.

This is useful because it plays a central role in bounding the eigenvalues of the auto-covariance matrices of a VAR(d) process. Using these quantities, Basu and Michailidis [2015] established bounds of the auto-covariance matrices (Proposition IV.1) in terms of the fundamental properties of the underlying process. This result together with application of the Hansen-Wright inequality constitute the crux in proving our lasso guarantees.

Proposition IV.1 (Eigenvalues of a Block Toeplitz Matrix)**.** *Consider a p-dimensional stable Gaussian VAR(d) process* $\{X_t\}$ *with characteristic polynomial* $\mathcal{A}(z)$ *and error covariance matrix* Σ_ϵ*. Denote the covariance matrix of the np-dimensional random vector* $\{(X_n)', (X_{n-1})', \cdots, (X_1)'\}$ *by* $\Gamma_n^{\mathcal{A}} = [\Gamma(r-s)_{p\times p}]_{1\leq r,s\leq n}$*, where* $\Gamma(h) = \mathbb{E}[X_t(X_{t+h})']$ *is the auto-covariance matrix of order h of the process* $\{X_t\}$*. Then*

$$\frac{\Lambda_{\min}(\Sigma_\epsilon)}{\mu_{\max}(\mathcal{A})} \leq \Lambda_{\min}(\Gamma_n^{\mathcal{A}}) \leq \Lambda_{\max}(\Gamma_n^{\mathcal{A}}) \leq \frac{\Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\mathcal{A})}$$

This result is provides a bound on the dependence across observations at different time points of a VAR process. This is crucial in our analysis.

4.2.3 Lasso Estimation Procedure For Transition Matrix

We focus on estimating the transition matrices. In high dimensional settings, the number of parameters scales as p^2d. Therefore, consistent estimation is impossible without regularization. In particular, we assume that $\{\mathbf{A}_1, \cdots, \mathbf{A}_d\}$ are sparse and impose ℓ_1 penalty in our estimation.

Regression Notation We are interested in estimating the transition matrices via minimizing a least squares objective. Before stating the optimization problem, we collect the matrix representations of the data here for reader's convenience:

For a sequence of samples $\{Z_t\}_{t=1}^{M}$, we can write the autoregressive of order d relationship in regression form

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}$$

where

$$\mathbf{Y} = \begin{pmatrix} (Z_M)' \\ \vdots \\ (Z_d)' \end{pmatrix}_{(M-d+1)\times p} \qquad \mathbf{X} = \begin{pmatrix} (Z_{M-1})' & \cdots & (Z_{M-d})' \\ \vdots & \ddots & \vdots \\ (Z_{d-1})' & \cdots & (Z_0)' \end{pmatrix}_{(M-d+1)\times(pd)}$$

$$\mathbf{B} = \begin{pmatrix} \mathbf{A}_1' \\ \vdots \\ \mathbf{A}_d' \end{pmatrix}_{(pd)\times p} \qquad \mathbf{E} = \begin{pmatrix} (\epsilon^M)' \\ \vdots \\ (\epsilon^d)' \end{pmatrix}_{(M-d+1)\times p}$$

Modified Lasso Framework We introduce the modified lasso framework proposed by [Loh and Wainwright, 2012] here.

Given a sequence of duples (X_t, Y_t), the target parameter matrix we are after is

$$\mathbf{B}^* = \underset{\mathbf{B}\in\mathbb{R}^{pd\times p}}{\arg\min}\, \mathbb{E}\|Y_t - X_t'\mathbf{B}\|_F^2$$

The corresponding estimator from the samples is

$$\hat{\mathbf{B}} = \underset{\mathbf{B}\in\mathbb{R}^{pd\times p}}{\arg\min}\, \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2 \tag{4.2.1}$$

It can be written equivalently, upon expanding the squares,as

$$\begin{aligned} \hat{\mathbf{B}} &= \underset{\mathbf{B}\in\mathbb{R}^{pd\times p}}{\arg\min}\, \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2 \\ &= \underset{\mathbf{B}\in\mathbb{R}^{pd\times p}}{\arg\min} - \mathrm{Tr}(2\mathbf{Y}'\mathbf{X}\mathbf{B}) + \mathrm{Tr}(\mathbf{B}'\mathbf{X}'\mathbf{X}\mathbf{B}) \end{aligned}$$

If we believe that $\mathbf{B}^*$ is sparse, we can encourage sparsity in $\hat{\mathbf{B}}$ by penalizing the least squares objective with ℓ_1 norm of the parameters and obtain the penalized version of the lasso:

$$\begin{aligned}\hat{\mathbf{B}} &= \underset{\mathbf{B}\in\mathbb{R}^{pd\times p}}{\arg\min} \|\mathbf{Y}-\mathbf{XB}\|_F^2 + \lambda_N \|\operatorname{vec}(\mathbf{B})\|_1 \\ &= \underset{\mathbf{B}\in\mathbb{R}^{pd\times p}}{\arg\min} - \operatorname{Tr}(2\mathbf{Y}'\mathbf{XB}) + \operatorname{Tr}(\mathbf{B}'\mathbf{X}'\mathbf{XB}) + \lambda_N \|\operatorname{vec}(\mathbf{B})\|_1\end{aligned}$$

Observe that if we replace the cross products $\mathbb{E}\mathbf{X}'\mathbf{X}$ and $\mathbb{E}\mathbf{Y}'\mathbf{X}$ by their unbiased and consistent estimators $\hat{\Gamma}$ and $\hat{\gamma}$, the above program remains consistent for $\mathbf{B}^*$. This gives rise to the modified lasso framework proposed by Loh and Wainwright [2012]:

$$\hat{\mathbf{B}} = \underset{\mathbf{B}\in\mathbb{R}^{pd\times p}}{\arg\min} - \operatorname{Tr}(2\hat{\gamma}\mathbf{B}) + \operatorname{Tr}(\mathbf{B}'\hat{\Gamma}\mathbf{B}) + \lambda_N \|\operatorname{vec}(\mathbf{B})\|_1 \tag{4.2.2}$$

This opens the door to designing $\hat{\Gamma}$ and $\hat{\gamma}$ when we are faced with real data application issues such as missing data or data corruption. Denote the "*effective*" sample size as $T = M - d + 1$.

4.3 Theoretical Guarantees

4.3.1 Lasso Consistency and Sufficient Conditions

We shall start with what we call a "master theorem" that provides non-asymptotic guarantees for lasso estimation and prediction errors under two well-known conditions, viz. the restricted eigenvalue (RE) and the deviation bound (DB) conditions. Note that in the classical linear model setting (see, e.g., Hayashi [2000, Ch 2.3]) where sample size is larger than the dimensions ($T > p$), the conditions for consistency of the ordinary least squares(OLS) estimator are as follows: (a) the empirical covariance matrix $\mathbf{X}'\mathbf{X}/T \xrightarrow{P} \mathbf{Q}$ and $\mathbf{Q}$ invertible, i.e., $\lambda_{\min}(\mathbf{Q}) > 0$, and (b) the regressors and the noise are asymptotically uncorrelated, i.e., $\mathbf{X}'\mathbf{W}/T \to \mathbf{0}$.

In high-dimensional regimes, Bickel et al. [2009], Loh and Wainwright [2012] and Negahban and Wainwright [2012] have established similar consistency conditions for lasso. The first one is the restricted eigenvalue (RE) condition on $\mathbf{X}'\mathbf{X}/T$ (which is a special case, when the loss function is the squared loss, of the restricted strong convexity (RSC) condition). The second is the deviation bound (DB) condition on $\mathbf{X}'\mathbf{W}/T$. The following lower RE and DB definitions are adopted from Loh and Wainwright [2012].

Definition 9 (Lower Restricted Eigenvalue)**.** A symmetric matrix $\Gamma \in \mathbb{R}^{p\times p}$ satisfies a lower restricted eigenvalue condition with curvature $\alpha > 0$ and tolerance $\tau(T,p) > 0$ if,

$$\forall v \in \mathbb{R}^p, \;\; v'\Gamma v \geq \alpha \left\|v\right\|_2^2 - \tau(T,p) \left\|v\right\|_1^2 .$$

Definition 10 (Deviation Bound)**.** Consider the random matrices $\hat{\Gamma} \in \mathbb{R}^{pd\times pd}$, $\hat{\gamma} \in \mathbb{R}^{p\times pd}$, and $\mathbf{B}^* \in \mathbb{R}^{p\times pd}$ defined above. They are said to satisfy the deviation bound condition if there exist a deterministic multiplier function $\mathbb{Q}(\mathbf{X},\mathbf{B}^*)$ and a rate of decay function $\mathbb{R}(p,d,T)$ such that,

$$\frac{1}{T}\left|\!\left|\!\left|\hat{\gamma} - (\mathbf{B}^*)'\hat{\Gamma}\right|\!\right|\!\right|_\infty \leq \mathbb{Q}(\mathbf{X},\Theta^\star)\mathbb{R}(p,d,T).$$

We now present a master theorem that provides guarantees for the ℓ_2 parameter estimation error and the (in-sample) prediction error. The proof, given in Appendix A of Wong et al. [2017], builds on existing result of the same kind [Bickel et al., 2009, Loh and Wainwright, 2012, Negahban and Wainwright, 2012] and we make no claims of originality for either the result or for the proof.

Theorem IV.2 (Estimation and Prediction Errors)**.** *Consider the lasso estimator $\hat{\boldsymbol{B}}$ defined in* (4.2.2)*. Suppose that $\hat{\Gamma}$ satisfies the lower RE(α, τ) condition with $\alpha \geq 32s\tau$ and $(\hat{\Gamma}, \hat{\gamma})$ satisfies the deviation bound. Then, for any $\lambda_T \geq 4\mathbb{Q}(\boldsymbol{X}, \boldsymbol{B}^*)\mathbb{R}(p, d, T)$, we have the following guarantees:*

$$\left\|\text{vec}(\hat{\boldsymbol{B}} - \boldsymbol{B}^*)\right\| \leq 4\sqrt{s}\lambda_T/\alpha, \tag{4.3.1}$$

$$\left|\left|\left|(\hat{\boldsymbol{B}} - \boldsymbol{B}^*)'\hat{\Gamma}(\hat{\boldsymbol{B}} - \boldsymbol{B}^*)\right|\right|\right|_F^2 \leq \frac{32\lambda_T^2 s}{\alpha}. \tag{4.3.2}$$

With this master theorem at our disposal, we just need to establish the validity of the restricted eigenvalue (RE) condition and deviation bound (DB) conditions for VAR processes when there are (1) data corruption or (2) missing data.

4.3.2 Missing and Corrupted Data

We study lasso estimation on stable Gaussian VAR data under two simple noisy and missing data scenarios. We state the corresponding $\hat{\Gamma}$ and $\hat{\gamma}$ in each case below

Data Corruption Instead of observing the complete data matrix $\mathbf{X}$, we see $\mathbf{Z} = \mathbf{X} + \mathbf{W}$ where $\mathbf{W}$ is a random matrix independent of $\mathbf{X}$, with each row $\mathbf{W}_{i:}$ sampled i.i.d. from $\mathcal{N}(\mathbf{0}, \Sigma_W)$. Assume $\Sigma_{\mathbf{W}}$ is known.

Define the matrices of noises as

$$\mathbf{W} = \begin{pmatrix} (W^{M-1})' & \dots & (W^{M-d})' \\ \vdots & \ddots & \vdots \\ (W^{d-1})' & \dots & (W^0)' \end{pmatrix}_{T \times dp} \qquad \omega = \begin{pmatrix} (W^M)' \\ \vdots \\ (W^d)' \end{pmatrix}_{T \times p}$$

Define

$$\hat{\Gamma} := \frac{(\mathbf{X} + \mathbf{W})'(\mathbf{X} + \mathbf{W})}{T} - \Sigma_{\mathbf{W}}, \text{ and}$$

$$\hat{\gamma} := (1/T)(\mathbf{X} + \mathbf{W})'(\mathbf{Y} + \omega)$$

Note that $\hat{\Gamma}$ defined this way is negative definite since $\frac{(\mathbf{X}+\mathbf{W})'(\mathbf{X}+\mathbf{W})}{T}$ is positive semi-definite containing zero eigenvalue(s) and $\Sigma_{\mathbf{W}}$ is positive definite.

Missing Data Instead of observing complete data matrix $\mathbf{X}$, we have $\mathbf{Z} = \mathbf{X} \odot \mathbf{W}$ where $\mathbf{W}$ is a random matrix independent of $\mathbf{X}$, with each element $\mathbf{W}_{ij}$ sampled i.i.d. from Ber(ρ), for some positive parameter ρ.

Define

$$\hat{\Gamma} := \frac{(\mathbf{X} \odot \mathbf{W})'(\mathbf{X} \odot \mathbf{W}) \oslash \mathbf{M}}{T}, \text{ and}$$
$$\hat{\gamma} := \frac{(\mathbf{X} \odot \mathbf{W})'(\mathbf{Y} \odot \omega)}{T\rho^2}$$

where $\mathbf{M} := E(\mathbf{W}'\mathbf{W})$ satisfies

$$\mathbf{M}_{ij} = \begin{cases} \rho^2, & \text{if } i \neq j \\ \rho, & \text{othewrwise .} \end{cases}$$

We can check (see the proof in the Appendix) that the $\hat{\Gamma}$ and $\hat{\gamma}$ defined as such are unbiased and consistent for $\mathbf{X}'\mathbf{X}$ and $\mathbf{Y}'\mathbf{X}$ respectively. Next, we state the high probability guarantees for the RE and DB conditions below under each scenario.

4.3.3 High Probabilistic Guarantees for Additive Data Corruption Case

Consider a stable Gaussian VAR process. For the $\hat{\Gamma}$ and $\hat{\gamma}$ defined in Section 4.3.2 pertaining to the corrupted data scenario, we have the following guarantees

Lemma IV.3 (RE Bound for Corrupted Data)**.** *There exists constant $c > 0$ such that, for sample size $T \geq \frac{42e\cdot k\log(pd)}{c\cdot\min\{\eta,\eta^2\}}$, with probability at least* $1-\exp\left\{-\frac{cT}{2}\min\{\eta,\eta^2\}\right\}$, *we have*

$$\hat{\Gamma} \sim \mathrm{RE}(\alpha, \tau)$$

where

$$\begin{aligned} \alpha &= \frac{\Lambda_{\min}(\Sigma_\epsilon)}{2\mu_{\max}(\mathcal{A})}, \qquad \tau = \frac{2\log(pd)\alpha}{T\min\{\eta,\eta^2\}}, \qquad \textit{and} \\ \eta &= \frac{1}{54}\frac{\Lambda_{\min}(\Sigma_\epsilon)}{\mu_{\max}(\mathcal{A})}\left(\frac{\Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\tilde{\mathcal{A}})} + \Lambda_{\max}(\Sigma_{\tilde{W}})\right)^{-1} \end{aligned}$$

Lemma IV.4 (DB Bound for Corrupted Data)**.** *There exist constant $C > 0$ such that, for sample size $T \geq \log(d^2p^2)$, with probability at least $1 - 14(dp)^{-C}$, we have*

$$\frac{1}{T}\left\|\left\|\hat{\gamma} - (\boldsymbol{B}^*)'\hat{\Gamma}\right\|\right\|_{\max} \leq \sqrt{\frac{\log(d^2p^2)}{T}}\mathcal{Q}(B^*, \Sigma_{\tilde{W}}, \Sigma_\epsilon)$$

where

$$\begin{aligned} \mathcal{Q}(B^*, \Sigma_{\tilde{W}}, \Sigma_\epsilon) = &\Bigg[\Lambda_{\max}(\Sigma_{\tilde{W}})\left(3 + 2\max_j \left\|B^*_{:j}\right\|_1 + |||B^*|||^2\right) \\ &+ \left(\Lambda_{\max}(\Sigma_\epsilon)\left(2 + \frac{2}{\mu_{\min}(\tilde{A})} + \frac{\mu_{\max}(A)}{\mu_{\min}(\tilde{A})}\right)\right)\Bigg] \end{aligned}$$

4.3.4 High Probabilistic Guarantees for Multiplicative Noise Case

Consider a stable Gaussian VAR process. For the $\hat{\Gamma}$ and $\hat{\gamma}$ defined in Section 4.3.2 pertaining to the missing data scenario, we have the following guarantees.

Lemma IV.5 (RE Bound for Missing Data)**.** *There exist constants c_i such that, with probability at least $1 - c_1\exp\{-c_2T\min\{1, \frac{\Lambda_{\min}(\Sigma_\epsilon)^2}{\xi^4}\}\}$ we have*

$$\hat{\Gamma} \sim \mathrm{RE}(\alpha, \tau)$$

where

$$\begin{aligned}
\alpha &= \frac{\Lambda_{\min}(\Sigma_\epsilon)}{2\mu_{\max}(\mathcal{A})},\\
\tau &= c_0 \frac{\log(pd)}{T}\alpha \cdot \max\left\{\left(\frac{1}{\rho}\right)^4 \frac{\lambda_{\max}(\Sigma_\epsilon)\mu_{\max}(A)}{\lambda_{\min}(\Sigma_\epsilon)\mu_{\min}(\tilde{A})}, 1\right\}, \qquad \textit{and}\\
\xi^2 &= \frac{1}{\rho^2}\frac{\Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\tilde{A})}
\end{aligned}$$

Lemma IV.6 (DB Bound for Missing Data)**.** *There exist constants c_i such that, with probability at least* $1 - c_1 \exp\{-c_2 \log(pd)\}$

$$\frac{1}{T}\left\|\left\|\hat{\gamma} - (\boldsymbol{B}^*)'\hat{\Gamma}\right\|\right\|_{\max} \le c_0 \xi^2 \max_i \|\boldsymbol{B}^*_{:i}\|_1^2 \sqrt{\frac{\log(pd)}{T}}$$

where

$$\xi^2 = \frac{1}{\rho^2}\frac{\Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\tilde{A})}$$

4.4 Proofs of Corrupted Data Case

To better differentiate the $\hat{\Gamma}$ and $\hat{\gamma}$ in the corrupted and missing data case in the proofs, we match the notation $\hat{\Gamma} \equiv \mathbf{G}^{\text{cor}}_{\tilde{Z}}$ and $\hat{\gamma} \equiv \mathbf{R}^{\text{cor}}$. Also, $N \equiv T$ denote the sample size. Let $\mathbf{Z} = \mathbf{X} + \mathbf{W}$. We begin by stating a variant of the Hansen-Weight inequality which serves as the basis of the proofs.

4.4.1 Variant of Hanson-Wright Inequality

The general statement of the Hanson-Wright inequality can be found in the paper by [Rudelson and Vershynin, 2013, Theorem 1.1]. We use a form of the inequality which is derived in the proof of Proposition 2.4 of Basu and Michailidis [2015] as an easy consequence of the general result. We state the modified form of the inequality and the proof below for completeness.

Lemma IV.7 (Variant of Hanson-Wright Inequality)**.** *If* $Y \sim \mathcal{N}(0_{n\times 1}, \boldsymbol{Q}_{n\times n})$*, then there exists universal constant* $c > 0$ *such that for any* $\eta > 0$*,*

$$\mathbb{P}\left[\frac{1}{n}\left|\|Y\|_2^2 - \mathbb{E}\,\|Y\|_2^2\right| > \eta |\!|\!| \boldsymbol{Q} |\!|\!|\right] \leq 2\exp\left[-cn\min\left\{\eta,\eta^2\right\}\right]. \tag{4.4.1}$$

Proof. The lemma easily follows from Theorem 1.1 in Rudelson and Vershynin [2013]. Write $Y = \mathbf{Q}^{1/2}X$, where $X \sim \mathcal{N}(0,\mathbf{I})$ and $(\mathbf{Q}^{1/2})'(\mathbf{Q}^{1/2}) = \mathbf{Q}$. Note that each component X_i of X is independent $\mathcal{N}(0,1)$, so that $\|X_i\|_{\psi_2} \leq 1$. Then, by the above theorem,

$$\begin{aligned}
&\mathbb{P}\left[\frac{1}{n}\left|\|Y\|_2^2 - \mathrm{Tr}(\mathbf{Q})\right| > \eta |\!|\!|\mathbf{Q}|\!|\!|\right] \\
&\quad = \mathbb{P}\left[\frac{1}{n}\left|X'\mathbf{Q}X - \mathbb{E}X'\mathbf{Q}X\right| > \eta|\!|\!|\mathbf{Q}|\!|\!|\right] \\
&\quad \leq 2\exp\left[-c\min\left\{\frac{n^2\eta^2|\!|\!|\mathbf{Q}|\!|\!|}{|\!|\!|\mathbf{Q}|\!|\!|_F^2}, \frac{n\eta|\!|\!|\mathbf{Q}|\!|\!|}{|\!|\!|\mathbf{Q}|\!|\!|}\right\}\right] \\
&\quad \leq 2\exp\left[-c\min\left\{\eta,\eta^2\right\}\right] \qquad \text{since } |\!|\!|\mathbf{Q}|\!|\!|_F^2 \leq n|\!|\!|\mathbf{Q}|\!|\!|^2
\end{aligned}$$

Lastly, note that $\mathrm{Tr}(\mathbf{Q}) = \mathrm{Tr}(\mathbb{E}YY') = \mathbb{E}\,\mathrm{Tr}(YY') = \mathbb{E}\,\mathrm{Tr}(Y'Y) = \mathbb{E}\,\mathrm{Tr}\,\|Y\|^2 = \mathbb{E}\,\|Y\|^2$. □

4.4.2 Proof of Restricted Eigenvalue Guarantees

Goal: Show $\underset{pd\times pd}{\mathbf{G}_{\tilde{Z}}^{\mathrm{cor}}} := \mathbf{Z}'\mathbf{Z}/N - \Sigma_{\tilde{W}} \sim RE(\alpha,\tau)$ for some α, $\tau > 0$.

1. Show $\mathbf{G}_{\tilde{Z}}^{\mathrm{cor}}$ is unbiased for $\Sigma_{\tilde{X}} := \mathbf{X}'\mathbf{X}/N$

   $\mathbb{E}[\mathbf{G}_{\tilde{Z}}^{\mathrm{cor}}] = \mathbb{E}[\frac{(\mathbf{X}+\mathbf{W})'(\mathbf{X}+\mathbf{W})}{N}] - \Sigma_{\tilde{W}}$. But $\mathbf{X} \perp\!\!\!\perp \mathbf{W}$ and $\mathbb{E}(\mathbf{X}) = \mathbb{E}(\mathbf{W}) = \mathbf{0}$. Therefore,

   $$\mathbb{E}[\mathbf{G}_{\tilde{Z}}^{\mathrm{cor}}] = \frac{1}{N}\mathbb{E}\left[\mathbf{X}'\mathbf{X} + \mathbb{E}\mathbf{W}'\mathbf{W}\right] - \Sigma_{\tilde{W}} = \Sigma_{\tilde{X}}.$$

2. Show the result for one fixed $\mu \in \mathbb{R}^{pd}$

Let $\mathbf{S}_Z := \frac{\mathbf{Z}'\mathbf{Z}}{N}$. Then, $\mathbf{G}^{\mathrm{cor}}_{\tilde{Z}} - \Sigma_{\tilde{X}} = \mathbf{S}_Z - \Sigma_{\tilde{Z}}$. Our goal is to control for all fixed test vector $\mu \in \mathbb{R}^q$,

$$\mu(\mathbf{S}_z - \Sigma_{\tilde{Z}})\mu = \frac{1}{N}[(\mathbf{Z}\mu)'(\mathbf{Z}\mu) - N\mu'\Sigma_{\tilde{Z}}\mu]$$

to be positive for the set of "sparse" vectors.

Essentially, we aim to show that the sample covariance matrix of the vectors $\mathbf{Z}\mu \in \mathbb{R}^N$ satisfies the "RE" condition for a single vector in the unit ℓ_2 ball.

First of all, note that $\mathbf{Z}\mu \sim N(0, Q)$, where $Q \triangleq \mathbb{E}[(\mathbf{Z}\mu)(\mathbf{Z}\mu)'] \in \mathbb{R}^{N\times N}$.

Now,

$$\begin{aligned}
Q_{ij} &= \mathbb{E}[(\mathbf{Z}\mu)(\mathbf{Z}\mu)']_{ij} \\
&= \mathbb{E}((\mathbf{Z}_{i:})\mu)'((\mathbf{Z}_{j:})\mu) \\
&= \mu'(\mathbb{E}\tilde{Z}_i\tilde{Z}_j')\mu \\
&= \mu'(\mathbb{E}\tilde{X}_i\tilde{X}_j')\mu + \mu(\mathbb{E}\tilde{W}_i\tilde{W}_j')\mu \\
&= \mu'\Gamma^N_{\tilde{X}}[i,j]\mu + \mu\Gamma^N_{\tilde{W}}[i,j]\mu \\
&= \mu'\Gamma^N_{\tilde{Z}}[i,j]\mu
\end{aligned}$$

To apply Lemma IV.7, we need to take a closer look at the two quantities (1) $\mathrm{Tr}(Q)$ and (2) $|||Q|||$.

It is easy to see that $\mathrm{Tr}(Q) = \mu'\sum_{i=1}^N \Gamma^N_{\tilde{Z}}(i,i)\mu = N\mu'\Sigma_{\tilde{Z}}\mu$.

To control $\vert\!\vert\!\vert Q \vert\!\vert\!\vert$, we will invoke IV.1. Fix $u \in \mathbb{R}^N$, with $\|u\|_2 = 1$

$$
\begin{aligned}
u'Qu &= \sum_{r=1}^{N}\sum_{s=1}^{N} u_r u_s Q_{rs} \\
&= \sum_{r=1}^{N}\sum_{s=1}^{N} u_r u_s \mu' \Gamma_{\tilde{Z}}^{N}[s,r]\mu \\
&= (\mu \otimes u)' \Gamma_{\tilde{Z}}^{N} (\mu \otimes u) \\
&\leq \Lambda_{\max}(\Gamma_{\tilde{Z}}^{N}) \qquad\qquad \text{since } \|\mu \otimes u\| = 1
\end{aligned}
\tag{4.4.2}
$$

It remains to bound $\Lambda_{\max}(\Gamma_{\tilde{Z}}^{N})$.

But

$$
\begin{aligned}
\Lambda_{\max}(\Gamma_{\tilde{Z}}^{N}) &= \Lambda_{\max}(\Gamma_{\tilde{X}}^{N} + \Gamma_{\tilde{W}}^{N}) \qquad\qquad \text{since } \tilde{X}^t \perp\!\!\!\perp \tilde{W}^s,\ \forall t, s \\
&\leq \Lambda_{\max}(\Gamma_{\tilde{X}}^{N}) + \Lambda_{\max}(\Gamma_{\tilde{W}}^{N})
\end{aligned}
$$

Recall that $\{\tilde{X}^n\}_{n=1}^{N}$ is a VAR(1) process in $\mathbb{R}^{dp}$; i.e. $\tilde{X}^{n+1} = \tilde{A}\tilde{X}^n + \widetilde{\epsilon^n}$. Also, $\{\tilde{X}^n\}_{n=1}^{N}$ is stable since is stable. So, $\{\tilde{X}^n\}_{n=d+1}^{N}$ has the reverse characteristic polynomial, for any test vector $u \in \mathbb{R}^{dp}$

$$
\tilde{\mathcal{A}}(u) = \underset{dp\times dp}{I} - \underset{dp\times dp}{\tilde{A}}\underset{dp\times 1}{u}
$$

Similarly for $\{\tilde{W}^n\}_{n=d+1}^{N}$.

Therefore, by Proposition IV.1,

$$
\Lambda_{\max}\left(\Gamma_{\tilde{Z}}^{N}\right) \leq \frac{\Lambda_{\max}\Sigma_{\tilde{\epsilon}}}{\mu_{\min}(\tilde{\mathcal{A}})} + \frac{\Lambda_{\max}\Sigma_{\tilde{W}}}{\mu_{\min}(I)}
$$

Because of the structure of $\tilde{\epsilon}$ and $\tilde{W}$, $\Lambda_{\max}(\Sigma_{\tilde{\epsilon}}) = \Lambda_{\max}(\Sigma_{\epsilon})$ and $\Lambda_{\max}(\Sigma_{\tilde{W}}) = \Lambda_{\max}(\Sigma_W)$. Also, $\mu_{\min}(I) = 1$. Thus, we can simply the above bound as

$$
\Lambda_{\max}(\Gamma_{Z}^{N}) \leq \frac{\Lambda_{\max}(\Sigma_{\epsilon)}}{\mu_{\min}(\tilde{\mathcal{A}})} + \Lambda_{\max}(\Sigma_{\tilde{W}})
$$

Now, we are ready to apply Lemma IV.7, $\forall \eta > 0$,

$$\begin{aligned}
&\mathbb{P}[|\mu'(\mathbf{G}^{\text{cor}}_{\tilde{Z}} - \Sigma_{\tilde{X}})\mu| > \eta(\frac{\Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\tilde{\mathcal{A}})} + \Lambda_{\max}(\Sigma_{\tilde{W}}))] \\
&= \mathbb{P}[|\mu'(\mathbf{S}_Z - \Sigma_{\tilde{Z}})\mu| > \eta(\frac{\Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\tilde{\mathcal{A}})} + \Lambda_{\max}(\Sigma_{\tilde{W}}))] \\
&\leq 2\exp\left[-cN \min\{\eta, \eta^2\}\right]
\end{aligned}$$

3. Extending the result from one fixed $\mu \in \mathbb{R}^{pd}$ to a uniform bound over a $2s-$sparse net.

   Define $\kappa(2s) := \{\mu \in \mathbb{R}^{pd} : \|\mu\|_2 \leq 1, \|\mu\|_0 \leq 2s\}$, for some integer $2s \geq 1$. Applying Lemma F.2 in Basu and Michailidis [2015], for some constant $C > 0$,

$$\begin{aligned}
&\mathbb{P}\left[\sup_{\mu\in\kappa(2s)} |\mu'(\mathbf{G}^{\text{cor}}_{\tilde{Z}} - \Sigma_{\tilde{X}})\mu| > \eta(\frac{\Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\tilde{\mathcal{A}})} + \Lambda_{\max}(\Sigma_{\tilde{W}}))\right] \\
&\leq 2\exp\left(-CN\ \min\{\eta, \eta^2\} + 2s\ \min\left(\log(pd), \log\left(\frac{21epd}{2s}\right)\right)\right)
\end{aligned}$$

4. Getting the lower-RE condition

   To simplify notation, let

   $\Delta := \left(\frac{\Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\tilde{\mathcal{A}})} + \Lambda_{\max}(\Sigma_{\tilde{W}})\right)$; $\Xi := -CN\ \min\{\eta, \eta^2\} + 2s\ \min\left(\log(pd), \log\left(\frac{21epd}{2s}\right)\right)$

   We can establish, for all natural number $s \geq 1$

$$\begin{aligned}
&\sup_{\nu\in\kappa(2s)} |\nu'(\mathbf{G}^{\text{cor}}_{\tilde{Z}} - \Sigma_{\tilde{X}})\nu| \geq \eta\Delta \quad \text{w.p.} \leq 2\exp\Xi \\
&\longrightarrow \sup_{\nu\in\kappa(2s)} |\nu'(\mathbf{G}^{\text{cor}}_{\tilde{Z}} - \Sigma_{\tilde{X}})\nu| \leq \eta\Delta \quad \text{w.p.} \geq 1 - 2\exp\Xi \\
&\longrightarrow |\nu'(\mathbf{G}^{\text{cor}}_{\tilde{Z}} - \Sigma_{\tilde{X}})\nu| \leq 27\eta\Delta\left(\|\nu\|_2^2 + \frac{1}{s}\|v\|_1^2\right)\ \forall s > 1,\ \forall \nu,\ \ \text{w.p.} \geq 1 - 2\exp\Xi \\
&\longrightarrow \nu'\mathbf{G}^{\text{cor}}_{\tilde{Z}}\nu \geq [\Lambda_{\min}(\Sigma_{\tilde{X}}) - 27\eta\Delta]\|\nu\|_2^2 - \frac{27\eta\Delta}{s}\|\nu\|_1^2,\ \forall \nu,\ \text{w.p.} \geq 1 - 2\exp\Xi
\end{aligned}$$

   where the second last inequality follows from Lemma 12 in supplement of Loh and Wainwright [2012].

5. Choose η and s

Now, we want

$$\Lambda_{\min}(\Sigma_{\tilde{X}}) - 27\eta\Delta \geq 0$$

Choose η such that

$$27\eta\Delta \leq \Lambda_{\min}(\Sigma_{\tilde{X}}) \quad \longrightarrow \quad \eta \leq \frac{1}{27}\Lambda_{\min}(\Sigma_{\tilde{X}})\frac{1}{\Delta}$$

Hence, we can choose $\eta \leq \frac{1}{54}\Lambda_{\min}(\Sigma_{\tilde{X}})\frac{1}{\Delta}$. Note that $\eta > 0$ since Σ_ϵ is positive definite by assumption. Again by Proposition IV.1, we can set $\eta = \frac{1}{54\Delta}\frac{\Lambda_{\min}(\Sigma_\epsilon)}{\mu_{\max}(\mathcal{A})}$.

Now, we have, by result 2, w.p. $\geq 2\exp\left(-CN\ \min\{\eta,\eta^2\} + 2s\ \min\left(\log(pd), \log\left(\frac{21epd}{2s}\right)\right)\right)$.

$$\nu\mathbf{G}_{\tilde{Z}}^{\text{cor}}\nu \geq \alpha\left\|\nu\right\|_2^2 - \tau\left\|\nu\right\|_1^2$$

where $\alpha = \frac{\Lambda_{\min}(\Sigma_\epsilon)}{2\mu_{\max}(\mathcal{A})}$, $\tau = \frac{\alpha}{s}$

To obtain the right coverage probability, we choose

$$s = \frac{Nc\min\{\eta,\eta^2\}}{2\log(pd)}$$

6. The final guarantee on RE.

This gives us that for sample size

$$N \geq \frac{42e \cdot k\log(pd)}{c \cdot \min\{\eta,\eta^2\}}$$

with probability at least

$$1 - \exp\left\{-\frac{cN}{2}\min\{\eta,\eta^2\}\right\}$$

that

$$\mathbf{G}_{\tilde{Z}}^{\text{cor}} \sim \text{RE}(\alpha,\tau)$$

where

$$\alpha = \frac{\Lambda_{\min}(\Sigma_\epsilon)}{2\mu_{\max}(\mathcal{A})}, \quad \tau = \frac{2\log(pd)\alpha}{N\min\{\eta, \eta^2\}}, \quad \text{and } \eta = \frac{1}{54}\frac{\Lambda_{\min}(\Sigma_\epsilon)}{\mu_{\max}(\mathcal{A})}\left(\frac{\Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\tilde{\mathcal{A}})} + \Lambda_{\max}(\Sigma_{\tilde{W}})\right)^{-1}$$

4.4.3 Deviation Bound

Recall that

$$\mathbf{G}^{\text{cor}}_{\tilde{Z}} := \frac{\mathbf{Z}'\mathbf{Z}}{N} - \Sigma_{\tilde{W}}; \quad \mathbf{R}^{\text{cor}} := \frac{1}{N}\mathbf{Z}'(\mathbf{Y} + \mathcal{E})$$

First, we will establish unbiasedness of $\mathbf{R}^{\text{cor}}$. Due to independence of $\mathbf{W}$ and $\mathbf{X}$

$$\mathbb{E}[\mathbf{G}^{\text{cor}}_{\tilde{Z}} B^* - \mathbf{R}^{\text{cor}}] = \Sigma_{\tilde{X}} B^* - \Sigma_{\tilde{X}} B^* = 0$$

To establish the deviation bound boils down to controlling, with high probability, $\left|\!\left|\!\left|\mathbf{R}^{\text{cor}} - \mathbf{G}^{\text{cor}}_{\tilde{Z}} B^*\right|\!\right|\!\right|_{\max}$.

$$\mathbf{R}^{\text{cor}} - \mathbf{G}^{\text{cor}}_{\tilde{Z}} B^* = \frac{1}{N}\{\mathbf{W}'(\mathbf{E} + \mathcal{E}) + \mathbf{X}'(\mathcal{E} - \mathbf{W}B^*) + \mathbf{X}'\mathbf{E} + (N\Sigma_{\tilde{W}} - \mathbf{W}'\mathbf{W})B^*\}$$

We will control each term separately, and then apply the triangle inequality. The general strategies:

- $(N\Sigma_{\tilde{W}} - W'W)B^*$, apply Lemma IV.7.
- $\mathbf{X}'\mathbf{E}$ taken care of as in Basu and Michailidis [2015].
- $\mathbf{W} \perp\!\!\!\perp \mathbf{E}$, $\mathbf{X} \perp\!\!\!\perp \mathcal{E}$ and $\mathbf{X} \perp\!\!\!\perp \mathbf{W}$. We can handle these terms similarly.

Control $\|\|(N\Sigma_{\tilde{W}} - W'W)B^*\|\|_{\max}$

Define $\mathbf{A} := N\Sigma_{\tilde{W}} - \mathbf{W}'\mathbf{W} \in \mathbb{R}^{pd \times pd}$. For all i, j, let $e_i \in \mathbb{R}^{dp}$ and $\tilde{e_j} \in \mathbb{R}^p$ denote the canonical basis vectors. We have,

$$\begin{aligned}
\|\|(N\Sigma_{\tilde{W}} - \mathbf{W}'\mathbf{W})B^*\|\|_{\max} &= \|\|\mathbf{A}B^*\|\|_{\max} \\
&= \max_{i,j} |e_i'\mathbf{A}B\tilde{e_j}| \\
&\leq \max_j \left\|B^*_{:j}\right\|_1 (\max_{i,j} |e_i'\mathbf{A}e_j|) \\
&\leq \frac{1}{2} \max_j \left\|B^*_{:j}\right\|_1 \max_{i,j}\{|(e_i + e_j)'\mathbf{A}(e_i + e_j)| + |e_i'\mathbf{A}e_i| + |e_i'\mathbf{A}e_j|\}
\end{aligned}$$

Note each of the three terms above are of the form $v'\mathbf{A}v$. So, we will derive a general recipe to control such a quadratic form. This is accomplished in two steps – first using concentration bound, followed by applying the union bound and lastly choosing $\eta = C\sqrt{\frac{\log(p)}{N}}$, for some constant C.

1. First, control $v'\mathbf{A}v$ using Lemma IV.7. For a unit vector $v \in \mathbb{R}^{pd}$

$$\mathbf{A} = N\Sigma_{\tilde{W}} - \mathbf{W}'\mathbf{W}$$
$$v'\mathbf{A}v = N(v'\Sigma_{\tilde{W}}v - (\mathbf{W}v)'(\mathbf{W}v))$$

Let $Y = \mathbf{W}v \in \mathbb{R}^N$. We have Y follows a normal distribution with mean $\mathbb{E}(Y) = 0$. Define $Q := \mathbb{E}YY'$

Then, using similar arguments as in Chapter 4.4.2, we know

$$\text{Tr}(Q) = Nv'\Sigma_{\tilde{W}}v$$

Also,

$$\begin{aligned}
\|Q\|_{op} &\leq \Lambda_{\max}(\Gamma^N_{\tilde{W}}) \\
&\leq \frac{\Lambda_{\max}(\Sigma_{\tilde{W}})}{\mu_{\min}(I)} = \Lambda_{\max}(\Sigma_{\tilde{W}}) \qquad \text{by Proposition IV.1}
\end{aligned}$$

Now,

$$\|Y\|_2^2 = Y'Y = (\mathbf{W}v)'(\mathbf{W}v)$$

$$\mathrm{Tr}(Y) = Nv'\Sigma_{\tilde{W}}v$$

So, $|\ \|Y\|_2^2 - \mathrm{Tr}(Q)| = |v'(\mathbf{W}'\mathbf{W} - N\Sigma_{\tilde{W}})v\,| = |v'\mathbf{A}v|$.

2. Apply Lemma IV.7

For any $\eta > 0$ and constant $C > 0$,

$$\mathbb{P}\{\frac{1}{N}|v'\mathbf{A}v| > \eta |\!|\!| Q |\!|\!|\} \le 2\exp(-cN\ \min\{\eta,\eta^2\})$$
$$\Rightarrow \mathbb{P}\{\frac{1}{N}|v'\mathbf{A}v| > \eta\Lambda_{\max}(\Sigma_{\tilde{W}})\} \le 2\exp(-cN\ \min\{\eta,\eta^2\})$$
$$\Rightarrow \mathbb{P}\{\frac{1}{N}|e_i'\mathbf{A}e_j| > \eta\Lambda_{\max}(\Sigma_{\tilde{W}})\} \le 2\exp(-cN\ \min\{\eta,\eta^2\})$$
$$\Rightarrow \mathbb{P}\{\frac{1}{N}\max_j \left\|B^*_{:j}\right\|_1 |e_i'\mathbf{A}e_j| > 2\max_j \left\|B^*_{:j}\right\|_1 \eta\Lambda_{\max}(\Sigma_{\tilde{W}})\} \le \exp(-cN\ \min\{\eta,\eta^2\})$$

Let $\eta = \sqrt{\frac{\log(p^2d^2)}{N}}$. For $N \ge 2(\log(p) + \log(d),\ 0 < \eta < 1$ and $\min\{\eta,\eta^2\} = \eta^2$.

Therefore,

(4.4.3)
$$\mathbb{P}\{\frac{1}{N}\max_j \left\|B^*_{:j}\right\|_1 |e_i'\mathbf{A}e_j| > \max_j \left\|B^*_{:j}\right\|_1 \sqrt{\frac{\log(p^2d^2)}{N}}\Lambda_{\max}(\Sigma_{\tilde{W}})\} \le 2\exp(-cN\,\eta^2)$$

3. Take union bound over $1 \le i,j \le pd$.

$$\mathbb{P}\left(\frac{1}{N}\max_j \left\|B^*_{:j}\right\|_1 \max_{i,j} |e_i'\mathbf{A}e_j| > \max_j \left\|B^*_{:j}\right\|_1 \sqrt{\frac{\log(p^2d^2)}{N}}\Lambda_{\max}(\Sigma_{\tilde{W}})\right) \le 2\frac{p^2d^2}{(p^2d^2)^c}$$

In all, for $N > \log(p^2d^2)$

$$\mathbb{P}\left(\frac{1}{N}\left\|N\Sigma_{\tilde{W}} - \mathbf{W}'\mathbf{W})B^*\right\|_{\max} \le \sqrt{\frac{\log(p^2d^2)}{N}}\mathbb{Q}_2(B^*,\mathbf{W})\right) \ge 1 - 2(pd)^{2C}$$

where $\mathbb{Q}_2(B^*,\mathbf{W}) = \max_j \left\|B^*_{:j}\right\|_1 \Lambda_{\max}(\Sigma_{\tilde{W}})$

Controlling interaction $\left\lvert\!\left\lvert\!\left\lvert \mathbf{X}'\mathbf{E} \right\rvert\!\right\rvert\!\right\rvert_{\max}$

This follows from Eq.(2.11) in Basu and Michailidis [2015]. With positive constants c_1–c_2, we have with probability at least $1 - 6\exp\{-c_1 \log(pd)\}$, $\forall e_i, e_j$

$$e_i' \left|\mathbf{X}'\mathbf{E}\right| e_j \leq c_2 \sqrt{\frac{\log(pd)}{N}} \left[\Lambda_{\max}(\Sigma_\epsilon)\left(1 + \frac{1}{\mu_{\min}(\mathcal{A})} + \frac{\mu_{\max}(\mathcal{A})}{\mu_{\min}(\mathcal{A})}\right)\right] \tag{4.4.4}$$

Controlling maximum norm of interaction between independent matrices

We first derive a concentration for interaction between Gaussian random vectors. We isolate it as a lemma.

Lemma IV.8. *Given Gaussian random matrices $\boldsymbol{C}, \boldsymbol{D} \in \mathbb{R}^{N\times q}$ such that $\boldsymbol{C} \perp\!\!\!\perp \boldsymbol{D}$. Then, we have the following guarantees: $\forall \eta > 0$*

$$\mathbb{P}\left[\left|\frac{1}{N}(Cu)'(Dv)\right| > \eta\left(\left\lvert\!\left\lvert\!\left\lvert\Sigma_{Cu}\right\rvert\!\right\rvert\!\right\rvert + \left\lvert\!\left\lvert\!\left\lvert\Sigma_{Dv}\right\rvert\!\right\rvert\!\right\rvert\right)\right] \leq 6\exp\left(-cN\min\{\eta, \eta^2\}\right)$$

Proof of Lemma IV.8. For unit test vectors u, v of the right dimensions, because $C \perp\!\!\!\perp D$,

$$\begin{aligned}(Cu)'(Dv) &= \frac{1}{2}\left[\left(\|Cu + Dv\|^2 - N\mathbb{E}\,\|Cu + Dv\|^2\right)\right.\\ &\quad - \left(\|Cu\|^2 - N\mathbb{E}\,\|Cu\|^2\right)\\ &\quad \left.- \left(\|Dv\|^2 - N\mathbb{E}\,\|Dv\|^2\right)\right]\\ &= \frac{1}{2}\left[\left(\|Cu + Dv\|^2 - \mathrm{Tr}(\Sigma_{Cu+Dv})\right)\right.\\ &\quad - \left(\|Cu\|^2 - \mathrm{Tr}(\Sigma_{Cu})\right)\\ &\quad \left.- \left(\|Dv\|^2 - \mathrm{Tr}(\Sigma_{Dv})\right)\right]\end{aligned}$$

We apply Lemma IV.7 on each of the three summand above with their corresponding $\left\lvert\!\left\lvert\!\left\lvert\Sigma\right\rvert\!\right\rvert\!\right\rvert$. Then we invoke Proposition IV.1 to get an upper bound on the corresponding

$\lVert\!|\Sigma|\!\rVert$.

$$\mathbb{P}\left[\left|\frac{1}{N}(Cu)'(Dv)\right| > \eta\left(\lVert\!|\Sigma_{Cu}|\!\rVert + \lVert\!|\Sigma_{Dv}|\!\rVert\right)\right] \le 6\exp\left(-cN\min\{\eta,\eta^2\}\right)$$

We apply Lemma IV.8 to each of $\mathbf{W},\mathbf{E}$, $\mathbf{X},\mathcal{E}$ and $\mathbf{X},\mathbf{W}$ pairs followed by Proposition IV.1 to obtain $\forall e_i, e_j$

$$\mathbb{P}\left[\frac{1}{N}\left|e_i'\mathbf{W}'(\mathbf{E}+\mathcal{E})e_j\right| > \eta\left(\Lambda_{\max}(\Sigma_\epsilon) + 2\Lambda_{\max}(\Sigma_{\tilde{W}})\right)\right] \le 4\exp\left(-cN\min\{\eta,\eta^2\}\right) \tag{4.4.5}$$

We know

$$\lVert\!|\Sigma_{\tilde{X}}|\!\rVert \le \frac{\Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\mathcal{A})},$$

and

$$\begin{aligned}\lVert\!|\Sigma_{\mathcal{E}-\mathbf{W}B^*}|\!\rVert &\le \lVert\!|\Sigma_{\tilde{W}}|\!\rVert + \lVert\!|\Sigma_{\tilde{W}B^*}|\!\rVert \\ &\le \Lambda_{\max}(\Sigma_{\tilde{W}}) + \lVert\!|B^*|\!\rVert^2\Lambda_{\max}(\Sigma_{\tilde{W}}) \quad \text{operator norm sub-multiplicative} \\ &= \Lambda_{\max}(\Sigma_{\tilde{W}})\left(1+\lVert\!|B^*|\!\rVert^2\right)\end{aligned}$$

Hence,

$$\begin{aligned}&\mathbb{P}\left[\frac{1}{N}\left|e_i'\mathbf{X}'(\mathcal{E}-\mathbf{W}B^*)e_j\right| > \eta\left(\Lambda_{\max}(\Sigma_{\tilde{W}})\left(1+\lVert\!|B^*|\!\rVert^2\right) + \frac{\Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\mathcal{A})}\right)\right] \\ &\le 4\exp\left(-cN\min\{\eta,\eta^2\}\right)\end{aligned} \tag{4.4.6}$$

Combinging the inequalities

Combining equations (4.4.3),(4.4.4), (4.4.5) and (4.4.6), we have wp less than

$$14\exp\{-cN\min\{\eta,\eta^2\}\}$$

$$
\begin{aligned}
e_i' \left|\mathbf{R}^{\mathrm{cor}} - \mathbf{G}_{\tilde{Z}}^{\mathrm{cor}} B^*\right| e_j & \\
> \eta \Bigg[& \Lambda_{\max}(\Sigma_{\tilde{W}}) \left(1 + |||B^*|||^2\right) + \frac{\Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\mathcal{A})} \\
& + \left(\Lambda_{\max}(\Sigma_\epsilon) + 2\Lambda_{\max}(\Sigma_{\tilde{W}})\right) \\
& + \left(\Lambda_{\max}(\Sigma_\epsilon) \left(1 + \frac{1}{\mu_{\min}(\mathcal{A})} + \frac{\mu_{\max}(\mathcal{A})}{\mu_{\min}(\mathcal{A})}\right)\right) \\
& + 2 \max_j \left\|B^*_{:j}\right\|_1 \Lambda_{\max}(\Sigma_{\tilde{W}}) \} \Bigg] \\
= \eta \Bigg[& \Lambda_{\max}(\Sigma_{\tilde{W}}) \left(3 + 2 \max_j \left\|B^*_{:j}\right\|_1 + |||B^*|||^2\right) \\
& + \left(\Lambda_{\max}(\Sigma_\epsilon) \left(2 + \frac{2}{\mu_{\min}(\mathcal{A})} + \frac{\mu_{\max}(\mathcal{A})}{\mu_{\min}(\mathcal{A})}\right)\right) \Bigg]
\end{aligned}
$$

Choose $\eta = \sqrt{\frac{\log(d^2p^2)}{N}}$. Then, for $N \geq \log(d^2p^2)$, for constant $C > 0$, with probability at least $1 - 14(dp)^{-C}$ that

$$
\left|\left|\left|\mathbf{G}_{\tilde{Z}}^{\mathrm{cor}} - (\mathbf{B}^*)'\mathbf{R}^{\mathrm{cor}}\right|\right|\right|_{\max} \leq \sqrt{\frac{\log(d^2p^2)}{N}} \mathcal{Q}(B^*, \Sigma_{\tilde{W}}, \Sigma_\epsilon)
$$

where

$$
\begin{aligned}
\mathcal{Q}(B^*, \Sigma_{\tilde{W}}, \Sigma_\epsilon) = \Bigg[& \Lambda_{\max}(\Sigma_{\tilde{W}}) \left(3 + 2 \max_j \left\|B^*_{:j}\right\|_1 + |||B^*|||^2\right) \\
& + \left(\Lambda_{\max}(\Sigma_\epsilon) \left(2 + \frac{2}{\mu_{\min}(\mathcal{A})} + \frac{\mu_{\max}(\mathcal{A})}{\mu_{\min}(\mathcal{A})}\right)\right) \Bigg]
\end{aligned}
$$

□

4.5 Proofs of Missing Data Case

We will modify Lemma 17 in Loh and Wainwright [2012] for the missing data case with a different bound on the operator norm(Eq. (G.6) in Loh and Wainwright [2012]). The rest of the proofs for RE and DB bounds follow similarly as in Loh and Wainwright [2012].

Given a realization of $\mathbf{U} = U$, we write the random matrix $\mathbf{X}^U := \mathbf{X} \odot U$. Similarly, denote $\tilde{Z}_i^U = \tilde{Z}_i \odot \tilde{U}_i$. For any unit test vector μ, $\mathbf{X}^U\mu$ is a zero mean mixture of Gaussians. Its covariance matrix is such that its (i,j)th component is

$$\begin{aligned}
Q_{ij} &= \mathbb{E}[(\mathbf{X}^U\mu)(\mathbf{X}^U\mu)']_{ij} \\
&= \mathbb{E}((\mathbf{X}^U_{i:})\mu)'((\mathbf{X}^U_{j:})\mu) \\
&= \mathbb{E}\mu'\tilde{Z}_i^U\mu'\tilde{Z}_j^U \\
&= \mathbb{E}\mu_1'\tilde{Z}_i\mu_2'\tilde{Z}_j \\
&= \mu_1'\mathbb{E}\left[\tilde{Z}_i\tilde{Z}_j\right]\mu_2 \\
&= \mu_1'\Gamma^N_{\tilde{X}}[i,j]\mu_2
\end{aligned}$$

where μ_1 and μ_2 are vectors μ with 0's in the positions corresponding to those of $\tilde{Z}_i^U$ and $\tilde{Z}_j^U$ respectively. Note that their magnitudes are at most 1.

Hence, by a similar argument as in Eq.(4.4.2) and using Proposition IV.1 on $\Gamma^N_{\tilde{X}}$, we arrive at

$$|||Q||| \leq \frac{\Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\mathcal{A})}$$

The rest will follow from similar logic as in the Proof of Corollary 4 in Loh and Wainwright [2012].

CHAPTER V

Future Directions

The previous chapters have laid out the theoretical guarantees of the RE and DB conditions for data satisfying either (1) the VAR model, or (2) geometrically mixing processes. In the former, we have provided lasso consistency for data corrupted with Gaussian noise or when we have data missing completely at random. In the latter, we have results for full spectrum of geometrically α and β-mixing processes with subweibull observations. As illustrations of the theory, we have also given examples satisfying the subweibull and geometrically mixing assumptions. These include nonlinear time series (autoregressive conditionally heteroscedastic model), misspecified and non-Markovian model (VAR with endogenous variable left out), heavy-tailed time series (subweibull VAR).

I conclude my thesis by listing a few plans for future directions of research here.

Lasso guarantees for general mixing processes observed with noise and/or missingness

The key mathematical tool in the analysis of Section IV is the Hansen-Wright inequality for Gaussian vectors. The proof scheme follows discretization of the parameter space, extension to a sparse net, and finally to any vectors in the space. The same proof strategies can be employed to extend the lasso guarantees to α and

β-mixing subweibull stationary time series as in Sections II-III. In general, given a concentration inequality, and any unbiased consistent estimators for $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{W}$, we can establish high probabilistic guarantees for the RE and DB conditions pertaining to the modified lasso.

Provably efficient streaming estimator for time series data In the large data era, when we receive new data points sequentially in time, it is computationally costly to update the batch estimator upon each new sample. For example, even with the simple OLS estimator, we have to invert an $N \times N$ gram matrix each time we recalculate it. It will be expensive if the sample size N is, say, in the order of millions.

One way around it is to update the estimator incrementally with the new sample instead of the whole set of data. We call this an online, aka streaming, estimator. This is useful if we can establish that, when sample size approaches infinity, the online estimator converges to a limit which is close to that of its batch counterpart within accuracy of statistical error.

In particular, the online mirror descent (OMD) algorithm, with appropriate choice of Bregman divergence, can achieve regret error bounds with graceful dimension scaling (logarithmic). We can hope to obtain an average loss (less that evaluated at optimal) with respect to the iterate average from the OMD algorithm to be roughly in the order of the lasso. Following the work of Duchi et al. [2012, 2010], we can contemplate a "slow rate" of lasso convergence under a specific setting of AR model with respect to the squared losses.

Guarantees for Other Regularized Estimators on Mixing Processes A large body of the literature in high dimensional statistics rely on establishing sufficient conditions similar to the RE(or RSC) and DB ones. Most of the theoretical analysis has been done under the iid assumption. We expect similar proof strategies can be employed to establish consistency guarantees for regularized estimating procedures under the context of geometrically mixing subweibull stationary time series. For example, Loh and Wainwright [2013] has established that under a suitable RSC condition and proper scaling on the sup norm of the gradient of the loss function, the estimation error bounds of SCAD and MCP scale roughly in the same order as the lasso. Negahban et al. [2012] has provided a unified framework for some class of decomposable penalties under the sufficient conditions of appropriate forms of RSC on the loss and DB on the gradient.

Bibliography

Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics*, 40(5):2452–2482, 2012.

Pierre Alquier, Paul Doukhan, et al. Sparsity considerations for dependent variables. *Electronic journal of statistics*, 5:750–774, 2011.

Sumanta Basu and George Michailidis. Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567, 2015.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

Richard Bradley. Basic properties of strong mixing conditions. a survey and some open questions. *Probability surveys*, 2(2):107–144, 2005.

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

Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12(6):805–849, 2012.

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

Alexander Chudik and M Hashem Pesaran. Infinite-dimensional VARs and factor models. *Journal of Econometrics*, 163(1):4–22, 2011.

Alexander Chudik and M Hashem Pesaran. Econometric analysis of high dimensional VARs featuring a dominant unit. *Econometric Reviews*, 32(5-6):592–649, 2013.

Alexander Chudik and M Hashem Pesaran. Theory and practice of GVAR modelling. *Journal of Economic Surveys*, 2014.

Richard A Davis, Pengfei Zang, and Tian Zheng. Sparse vector autoregressive modeling. *arXiv preprint arXiv:1207.0520*, 2012.

Richard A Davis, Pengfei Zang, and Tian Zheng. Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics*, (just-accepted):1–53, 2015.

David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45): 18914–18919, 2009.

John C Duchi, Shai Shalev-Shwartz, Yoram Singer, and Ambuj Tewari. Composite objective mirror descent. In *COLT*, pages 14–26, 2010.

John C Duchi, Alekh Agarwal, Mikael Johansson, and Michael I Jordan. Ergodic mirror descent. *SIAM Journal on Optimization*, 22(4):1549–1578, 2012.

JianQing Fan, Lei Qi, and Xin Tong. Penalized least squares estimation with weakly dependent data. *Science China Mathematics*, 59(12):2335–2354, 2016.

Sergey Foss, Dmitry Korshunov, Stan Zachary, et al. *An introduction to heavy-tailed and subexponential distributions*, volume 6. Springer, 2011.

Shaojun Guo, Yazhen Wang, and Qiwei Yao. High dimensional and banded vector autoregressions. *arXiv preprint arXiv:1502.07831*, 2015.

Fang Han and Han Liu. Transition matrix estimation in high dimensional time series. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 172–180, 2013.

Fang Han, Huanran Lu, and Han Liu. A direct estimation of high dimensional stationary vector autoregressions. *Journal of Machine Learning Research*, 16:3115–3150, 2015.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations.* CRC Press, 2015.

Fumio Hayashi. *Econometrics.* Princeton University Press, 2000.

Ild́ar Abdulovič Ibragimov and Yurii Antolevich Rozanov. *Gaussian random processes.* Springer, 1978.

Anders Bredahl Kock and Laurent Callot. Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics*, 186(2):325–344, 2015.

Sanjeev Kulkarni, Aurelie C Lozano, and Robert E Schapire. Convergence and consistency of regularized boosting algorithms with stationary β-mixing observations. In *Advances in neural information processing systems*, pages 819–826, 2005.

Eckhard Liebscher. Towards a unified approach for proving geometric ergodicity and mixing properties of nonlinear autoregressive processes. *Journal of Time Series Analysis*, 26(5):669–689, 2005.

Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40 (3):1637–1664, 2012.

Po-Ling Loh and Martin J Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems*, pages 476–484, 2013.

Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

Daniel J Mcdonald, Cosma R Shalizi, and Mark J Schervish. Estimating beta-mixing coefficients. In *International Conference on Artificial Intelligence and Statistics*, pages 516–524, 2011.

Timothy L McMurry and Dimitris N Politis. High-dimensional autocovariance matrices and optimal linear prediction. *Electronic Journal of Statistics*, 9:753–788, 2015.

Marcelo C Medeiros and Eduardo F Mendes. ℓ_1-regularization of high-dimensional time-series models with non-gaussian and heteroskedastic errors. *Journal of Econometrics*, 191(1):255–271, 2016.

Florence Merlevède, Magda Peligrad, and Emmanuel Rio. A bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, 151(3-4):435–474, 2011.

Yuval Nardi and Alessandro Rinaldo. Autoregressive process modeling via the lasso procedure. *Journal of Multivariate Analysis*, 102(3):528–549, 2011.

Sahand Negahban and Martin J Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, pages 1069–1097, 2011.

Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1):1665–1697, 2012.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Rodrigue Ngueyep and Nicoleta Serban. Large vector auto regression for multi-layer spatially correlated time series. *Technometrics*, 2014.

William Nicholson, David Matteson, and Jacob Bien. VARX-L: Structured regularization for large vector autoregressions with exogenous variables. *arXiv preprint arXiv:1508.07497*, 2015.

William B Nicholson, Jacob Bien, and David S Matteson. Hierarchical vector autoregression. *arXiv preprint arXiv:1412.5250*, 2014.

Gilles Pisier. Subgaussian sequences in probability and fourier analysis, 2016. arXiv preprint arXiv:1607.01053v3.

Maurice Bertram Priestley. Spectral analysis and time series. 1981.

Huitong Qiu, Sheng Xu, Fang Han, Han Liu, and Brian Caffo. Robust estimation of transition matrices in high dimensional heavy-tailed vector autoregressive processes. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1843–1851, 2015.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.

Murray Rosenblatt. A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences of the United States of America*, 42 (1):43, 1956.

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18:1–9, 2013.

Mark Rudelson and Shuheng Zhou. Reconstruction from anisotropic random measurements. *Information Theory, IEEE Transactions on*, 59(6):3434–3447, 2013.

Vidyashankar Sivakumar, Arindam Banerjee, and Pradeep K Ravikumar. Beyond sub-gaussian measurements: High-dimensional structured estimation with subexponential designs. In *Advances in Neural Information Processing Systems*, pages 2206–2214, 2015.

Song Song and Peter J Bickel. Large vector auto regressions. *arXiv preprint arXiv:1106.3915*, 2011.

Petre Stoica and Randolph L Moses. *Introduction to spectral analysis*, volume 1. Prentice hall Upper Saddle River, NJ, 1997.

Terence Tao and Van Vu. Random matrices: Sharp concentration of eigenvalues. *Random Matrices: Theory and Applications*, 2(03):1350007, 2013.

Dag Tjøstheim. Non-linear time series and markov chains. *Advances in Applied Probability*, pages 587–611, 1990.

Yoshimasa Uematsu. Penalized likelihood estimation in high-dimensional time series models and uts application. *arXiv preprint arXiv:1504.06706*, 2015.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Mathukumalli Vidyasagar. *Learning and generalisation: with applications to neural networks*. Springer Science & Business Media, second edition, 2003.

Gabrielle Viennet. Inequalities for absolutely regular sequences: application to density estimation. *Probability theory and related fields*, 107(4):467–492, 1997.

Hansheng Wang, Guodong Li, and Chih-Ling Tsai. Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1):63–78, 2007.

Zhaoran Wang, Fang Han, and Han Liu. Sparse principal component analysis for high dimensional vector autoregressive models. *arXiv preprint arXiv:1307.0164*, 2013.

Kam Chung Wong, Zifan Li, and Ambuj Tewari. Lasso guarantees for time series estimation under subgaussian tails and β-mixing, 2017.

W. B. Wu and Y. N. Wu. High-dimensional linear models with dependent observations, 2015. under review as per `http://www.stat.ucla.edu/ ywu/papers.html`. Accessed: October, 2015.

Wei Biao Wu. Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences of the United States of America*, 102(40): 14150–14154, 2005.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116, 1994.

Danna Zhang and Wei Biao Wu. Gaussian approximation for high dimensional time series. *arXiv preprint arXiv:1508.07036*, 2015.