Submitted to the Annals of Statistics
arXiv: 1708.01505

LASSO GUARANTEES FOR β-MIXING HEAVY TAILED TIME SERIES[*]

By Kam Chung Wong[†], Zifan Li and Ambuj Tewari

University of Michigan and Yale University

Many theoretical results for lasso require the samples to be iid. Recent work has provided guarantees for lasso assuming that the time series is generated by a sparse Vector Auto-Regressive (VAR) model with Gaussian innovations. Proofs of these results rely critically on the fact that the true data generating mechanism (DGM) is a finite-order Gaussian VAR. This assumption is quite brittle: linear transformations, including selecting a subset of variables, can lead to the violation of this assumption. In order to break free from such assumptions, we derive non-asymptotic inequalities for estimation error and prediction error of lasso estimate of the best linear predictor without assuming any special parametric form of the DGM. Instead, we rely only on (strict) stationarity and geometrically decaying β-mixing coefficients to establish error bounds for lasso for subweibull random vectors. 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 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.

1. Introduction.

High-dimensional statistics is a vibrant area of research in modern statistics and machine learning (Bühlmann and Van De Geer, 2011; Hastie, Tibshirani and Wainwright, 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, Negahban and Wainwright (2012);

[*] Footnote to the title with the 'thankstext' command.

[†] Footnote to the first author with the 'thankstext' command.

MSC 2010 subject classifications: Primary 60K35, 60K35; secondary 60K35

Keywords and phrases: time series, mixing, high-dimensional estimation, lasso

Donoho, Maleki and Montanari (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 paper is to give guarantees for ℓ_1-regularized least squares estimation, or lasso (Hastie, Tibshirani and Wainwright, 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) assumed 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 and C is a linear transformation then CZ_t may not be expressible as a finite lag VAR (Lütkepohl, 2005). We later provide 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 $Z_{t-d}, \dots, Z_{t-1}$ 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? Our paper 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 3), and (b) for stationary processes with subweibull marginals and geometrically decaying β-mixing coefficients (Section 4). It is well known that guarantees for lasso follow if one can establish the restricted eigenvalue (RE) conditions and provide deviation bounds (DB) for the correlation between noise and the regressors (see the Master Theorem in Section 2.3 below for a precise statement). Therefore, the bulk of the technical work in this paper boils down to establishing, with high probability, that the DB and RE conditions hold under the Gaussian α-mixing (Propositions 2 and 3) and the subweibull β-mixing assumptions respectively (Propositions 7 and 8). Note that the RE conditions were previously shown to hold under the *iid assumption* by Raskutti, Wainwright and Yu (2010) for Gaussian random vectors and for subgaussian random vectors by Rudelson and Zhou (2013). We also include some simulations (Section 5) to study the effect of VAR dimension, tail behavior, and temporal dependence on the estimation error decay rate as a function of the sample size.

1.1. *Summary of Recent Work on High-Dimensional Time Series.* While we discussed the work of Basu and Michailidis (2015) – since ours is closely related to theirs – 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 (2016) and Alquier et al. (2011) gave theoretical guarantees assuming that the RE conditions hold. As Basu and Michailidis (2015) pointed out, it takes a fair bit of work to actually establish the RE conditions in the presence of dependence. Chudik and Pesaran (2011, 2013, 2014) used 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, Zang and Zheng, 2016) and the Dantzig selector (Han and Liu, 2013; Han, Lu and Liu, 2015). Both Han and Liu (2013) and Han, Lu and Liu (2015) studied the stable Gaussian VAR models while our paper covers wider classes of processes as our examples demonstrate. Fan, Qi and Tong (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 Appendix D for details), the paper established finite sample ℓ_2 consistency in the real support for penalized least squares estimators. In addition, under a mutual incoherence type assumption, it provided sign and ℓ_∞ consistency. An AR(1) example was given as an illustration. Uematsu (2015) and Kock and Callot (2015) established oracle inequalities for 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, the author only required the existence of the fourth moments of the errors. Kock and Callot (2015) provided non-asymptotic lasso estimation and prediction error bounds for stable Gaussian VARs. Both Sivakumar, Banerjee and Ravikumar (2015) and Medeiros and Mendes (2016) considered subexponential designs. Sivakumar, Banerjee and Ravikumar (2015) studied lasso on iid subexponential designs and provided finite sample bounds. Medeiros and Mendes (2016) studied adaptive lasso for linear time series models and provided sign consistency results. Wang, Li and Tsai (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, Bien and Matteson, 2014; Nicholson, Matteson and Bien, 2017; Guo, Wang and Yao, 2016; Ngueyep and Serban, 2014). Zhang et al. (2017), McMurry and Politis (2015), Wang, Han and Liu (2013) and Chen, Xu and Wu (2013) considered estimation of the covariance (or precision) matrix of high-dimensional time series. McMurry and Politis (2015) and Nardi and Rinaldo (2011) both highlighted that autoregressive (AR) estimation, even in univariate time series, leads to high-dimensional parameter estimation problems if the lag is allowed to be unbounded.

1.2. *Organization of the Paper.* Section 2 introduces our notation, presents the assumptions used to derive our key results, 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 3 and 4 respectively. Section 3 deals with α-mixing Gaussian time series. Note that α-mixing is a weaker notion than β-mixing and all the parameter dependences are explicit. Section 4 deals with β-mixing time series with subweibull observations and we make the dependence on the subweibull norm explicit. Section 5 presents two simulation results: one where we vary the heaviness of the tail of the random vectors in the time series and another one where we vary the degree of temporal dependence in the time series.

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 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 present a non-linear 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 provide guarantees potentially in one or more of the following scenarios: model mis-specification, heavy tailed non-Gaussian innovations and nonlinearity in the DGM.

2. Preliminaries.

Consider a stochastic process of pairs $(X_t, Y_t)_{t=1}^{\infty}$ where $\forall t$, $X_t \in \mathbb{R}^p$, $Y_t \in \mathbb{R}^q$. 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 $(Z_{t-d}, \dots, Z_{t-1})$. A linear predictor is a convenient choice. To frame it as a regression problem, 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

$$\Theta^\star = \operatorname*{arg\,min}_{\Theta \in \mathbb{R}^{p\times q}} \mathbb{E}[\left\|Y_t - \Theta' X_t\right\|_2^2]. \tag{2.1}$$

Note that $\Theta^\star$ is independent of t due 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

$$\widehat{\Theta} = \operatorname*{arg\,min}_{\Theta \in \mathbb{R}^{p\times q}} \frac{1}{T} \|\operatorname{vec}(\mathbf{Y} - \mathbf{X}\Theta)\|_2^2 + \lambda_T \left\|\operatorname{vec}(\Theta)\right\|_1 . \tag{2.2}$$

where

$$\mathbf{Y} = (Y_1, Y_2, \dots, Y_T)' \in \mathbb{R}^{T\times q} \qquad \mathbf{X} = (X_1, X_2, \dots, X_T)' \in \mathbb{R}^{T\times p}. \tag{2.3}$$

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

$$\mathbf{W} := \mathbf{Y} - \mathbf{X}\Theta^\star. \tag{2.4}$$

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 square matrix $\mathbf{M}$ with rank d, let $\lambda_i(M)$, $i = 1, \dots, d$ denote its eigenvalues. Then, $r(\mathbf{M})$ denotes its spectral radius $\max_i \{|\lambda_i(\mathbf{M})|\}$. For any matrix $\mathbf{M}$, let $|||\mathbf{M}|||$, $|||\mathbf{M}|||_\infty$, and $|||\mathbf{M}|||_F$ denote its 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 $\mathbf{M}$, $|||\mathbf{M}|||_0 = \|\text{vec}(\mathbf{M})\|_0$ 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 $\mathbb{Z}$. Note that Σ and Γ are matrices but we will not use bold font for them in this paper.

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 whereas the latter is a $p \times q$ matrix. Thus, $\Sigma_{(X;Y)}(l)$ is a matrix consisting of four sub-matrices with the following block structure:

$$\Sigma_{(X;Y)}(l) = \begin{bmatrix} \Sigma_X(l) & \Sigma_{X,Y}(l) \\ \Sigma_{Y,X}(l) & \Sigma_Y(l) \end{bmatrix}.$$

Let $\mathbf{1}$ and $\mathbf{0}$ denote vectors consisting of ones and zeros respectively with dimensionality indicated in a subscript (if it is not clear from the context). We adopt the convention that, 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. *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 $\|\text{vec}(\widehat{\Theta} - \Theta^\star)\|_2$ 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., $\|\mathrm{vec}(\Theta^\star)\|_0 = s$.

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., $\forall t,\ \mathbb{E}(X_t) = \mathbf{0}_{p\times 1}$, and $\mathbb{E}(Y_t) = \mathbf{0}_{q\times 1}$.

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 dimensionality , 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, Ritov and Tsybakov (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 slight modifications of 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 $\alpha_C > 0$ and tolerance $\tau(T,p) > 0$ if,

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

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 (2.3) and (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,

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

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 Appendix A, builds on existing results of the same kind (Bickel, Ritov and Tsybakov, 2009; Loh and Wainwright, 2012; Negahban and Wainwright, 2012) and we make no claims of originality for either the result or the proof.

THEOREM 1 (Estimation and Prediction Errors). *Consider the lasso estimator* $\widehat{\Theta}$ *defined in* (2.2). *Suppose Assumption 1 holds. Further, suppose that* $\hat{\Gamma} := \boldsymbol{X}'\boldsymbol{X}/T$ *satisfies the lower* $RE(\alpha_C, \tau)$ *condition with* $\alpha_C \geq 32s\tau$ *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:*

$$\left\|\mathrm{vec}(\widehat{\Theta} - \Theta^\star)\right\| \leq 4\sqrt{s}\lambda_T/\alpha_C, \tag{2.5}$$

$$\left|\!\left|\!\left|(\widehat{\Theta} - \Theta^\star)'\hat{\Gamma}(\widehat{\Theta} - \Theta^\star)\right|\!\right|\!\right|_F^2 \leq \frac{32\lambda_T^2 s}{\alpha_C}. \tag{2.6}$$

With this master theorem at our disposal, we just need to establish the validity of the restricted eigenvalue (RE) 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 3, we consider α-mixing Gaussian random vectors. Next, in Section 4, we consider β-mixing subweibull random vectors (we define subweibull random vectors below in Section 4.1). Historically, mixing conditions were introduced to generalize the 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, Wainwright and Yu, 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.

2.4. *Proof Strategies for the RE and DB bounds.*

The key ingredients in establishing both the DB and RE conditions are concentration inequalities. The general strategy is to discretize the vector space, apply the concentration inequality, and use the union bound. This occurs in the proofs of the DB and RE conditions in both cases: α-mixing Gaussian and β-mixing subweibull.

A brief sketch of the proof of the DB condition via concentration goes like this: Consider a fixed vector $v \in \mathbb{R}^p$ and let $\Sigma_X = \mathbb{E}[X_t X_t^T]$. Use concentration inequality to show that

$$v'\mathbf{X}'\mathbf{X}v/T - v'\Sigma_X v = \sum (1/T) \sum_{t=1}^{T} \left(\|X_t' v\|_2^2 - \mathbb{E}[\|X_t' v\|_2^2] \right)$$

is sufficiently small. Then apply the union bound over a set of sparse v.

The arguments to show the RE condition via concentration proceed as follows. Note that

$$|||\mathbf{X}'\mathbf{W}|||_\infty = \max_{1\le i\le p, 1\le j\le q} |[\mathbf{X}'\mathbf{W}]_{i,j}| = \max_{1\le i\le p, 1\le j\le q} \left|(\mathbf{X}_{:i})'\mathbf{W}_{:j}\right| .$$

At the population level, there is no correlation between $\mathbf{W}$ and $\mathbf{X}$. Therefore,

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

Fix i, j and write,

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

The Hanson-Wright inequality (Lemma 11) takes care of the Gaussian process case. For the independent subgaussian case, the classical Bernstein's concentration inequality will allow us to prove lasso guarantees. However, applying the Bernstein's inequality requires the random vectors to satisfy *independence* and *subexponential* tail assumptions. Since a random variable is subgaussian if and only if its square is subexponential, the set of conditions required for the original stochastic process translate into *independence* and *subgaussian*.

Often times, real time series data exhibit large tail behavior in addition to being dependent. Therefore, the analysis of lasso for real life time series data requires the arguments to deal with the two complications. As a result, we need ways to quantify dependence and heavy tailed behavior. In addition, we need concentration inequalities that hold under weaker conditions. Next, we quantify *dependence* using *mixing coefficients*. Also, we quantify *tail behavior* using the notion of *subweibull* random variables. The concentration inequality we use here is Lemma 13 which we derive in Appendix D.3 building on the work of Merlevède, Peligrad and Rio (2011).

2.5. *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 be vector-valued or even take values in a Banach space) with associated sigma algebras $\sigma(X), \sigma(Y)$. For example,

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

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 a 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 generalize 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:

$$\rho(X,Y) = \sup\{\mathrm{Cov}(f(X), g(Y)) \, : \, \mathbb{E}f = \mathbb{E}g = 0, \mathbb{E}f^2 = \mathbb{E}g^2 = 1\}$$
$$\beta(X,Y) = \sup \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} |P(A_i \cap B_j) - P(A_i)P(B_j)|$$

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. 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 2 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, Lozano and Schapire, 2005) for dependent samples. There is also work on estimating β-mixing coefficients from data (Mcdonald, Shalizi and Schervish, 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 the 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. 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.

3. Gaussian Processes under α-Mixing.

Here we will study Gaussian processes under the α-mixing condition which is a weaker one than the β-mixing. We make the following additional assumptions.

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 2. 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 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*

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

where

$$\mathbb{Q}(\boldsymbol{X}, \boldsymbol{W}, \Theta^\star) = 8\pi\sqrt{\frac{(b+1)}{\tilde{c}}}\left(\lvert\lvert\lvert\Sigma_X\rvert\rvert\rvert\left(1 + \max_{1\leq i\leq p}\|\Theta^\star_{:i}\|_2^2\right) + \lvert\lvert\lvert\Sigma_Y\rvert\rvert\rvert\right), \textit{and}$$

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

Remark 1. Note that the free parameter b serves as a 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 value of b increases the success probability but worsens the sample size threshold and the multiplier function.

Proposition 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 $v \in \mathbb{R}^p$,*

$$|v'\hat{\Gamma}v| > \alpha_C\|v\|_2^2 - \tau(T,p)\|v\|_1^2, \tag{3.1}$$

where

$$\alpha_C = \frac{1}{2}\lambda_{\min}(\Sigma_X), \qquad \tau(T,p) = \alpha_C/\lceil c\frac{T}{4\log(p)}\min\{1,\eta^2\}\rceil, \quad \textit{and}$$

$$\eta = \frac{\lambda_{\min}(\Sigma_X)}{108\pi S_\alpha(T)\lambda_{\max}(\Sigma_X)}.$$

Remark 2. Note that, in Theorem 1, it is advantageous to have a large α_C 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 it is advantageous to have a well-conditioned Σ_X.

3.1. *Estimation and Prediction Errors.*

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

Corollary 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 2 and 3. Then, for sample size*

$$T \geq \max\left\{\frac{\log(p)}{c\min\{1,\eta^2\}}\max\{42e, 128s\}, \log(pq)\sqrt{\frac{b+1}{\tilde{c}}}\right\}$$

$$\textit{where } \eta = \frac{\lambda_{\min}(\Sigma_X)}{108\pi S_\alpha(T)\lambda_{\max}(\Sigma_X)}$$

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* (2.5) *and* (2.6) *hold with*

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

where

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

REMARK 3. If the α-mixing coefficients are summable, i.e., $S_\alpha(T)\le\tilde{\alpha}<\infty$, $\forall T$, 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 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 $O(S_\alpha(T)\sqrt{\frac{\log(pq)}{T}})$. We see that as long as $S_\alpha(T)\in o\left(\sqrt{T}\right)$, we still have consistency. For 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).

3.2. *Examples.*

We illustrate applicability of our theory developed in this section 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, Zang and Zheng, 2016; 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:

$$Z_t = \mathbf{A}_1 Z_{t-1}+\cdots+\mathbf{A}_d Z_{t-d}+\mathcal{E}_t \tag{3.2}$$

where each $\mathbf{A}_k, k=1,\dots,d$ 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 Appendix E.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 2 and 3, and thus Corollary 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 error 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 Appendix E.1 for details.

FACT 3. Every VAR(d) process can be written in a VAR(1) form (see e.g. (Lütkepohl, 2005, Ch 2.1)).

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

EXAMPLE 2 (Gaussian VAR with Omitted Variable). We study lasso estimation for 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 (e.g., it is impossible to observe them or perhaps very costly to measure them) to the researcher to perform estimation and prediction. Such a situation can also arise 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). As we describe below, the target of estimation is still the best linear predictor (in the least squares sense) of the future given the past. 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,

$$\mathbf{A} := \left[\begin{array}{cc} \mathbf{A}_{ZZ} & \mathbf{A}_{Z\Xi} \\ \mathbf{A}_{\Xi Z} & \mathbf{A}_{\Xi\Xi} \end{array} \right]$$

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 best (in the least squares sense) 1-lag predictor of Z_t as a function of Z_{t-1}. Recall that

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

Note that Z_t is not necessarily a finite order VAR process. 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 Appendix E.2 for details. As a result, Propositions 2 and 3, and thus Corollary 4 follow and hence we have all the high probabilistic guarantees for lasso on this non-Markovian example.

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 a 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.

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 $\mathbb{E}(|X|^p)^{1/p} \le K\sqrt{p}$, $\forall p \ge 1$ for some constant K (resp. $\mathbb{E}(|X|^p)^{1/p} \le Kp$, $\forall p \ge 1$). A natural generalization of these definitions that allows for heavier tails is as follows. Fix some $\gamma > 0$, and require

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

There are a few different equivalent ways to impose the condition above. One of them requires 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 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 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 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 6. *For any $\gamma \in (0, \infty)$, if a random variable X is subweibull(2γ) then X^2 is subweibull(γ). Moreover,*

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

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 $\|X_t\|_{\psi_{\gamma_2}} \leq K_X$, $\|Y_t\|_{\psi_{\gamma_2}} \leq K_Y$ 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, Li and Tewari (2016)). This paper 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 the theoretical results available in 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 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 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 $v \in \mathbb{R}^p$,

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

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

4.2. *Estimation and Prediction Errors.* Substituting the RE and DB constants from Propositions 7-8 into Theorem 1 immediately yields the following guarantee.

Corollary 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 7-8, and let* $K := 2^{2/\gamma_2}\left(K_Y + K_X\left(1 + \left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert \Theta^{\star} \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert\right)\right)^2$.

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 (2.5) and (2.6) hold with

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

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" $K/\lambda_{\min}(\Sigma_X)$.

4.3. *Examples.*

We explore applicability of our theory in Section 4 beyond just linear Gaussian processes using the examples below. Together, these demonstrate that our high probability guarantees for lasso can applied even in the presence of one or more of the following: heavy tailed subweibull data, model mis-specification, and nonlinearity in the DGM.

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, $r(\mathbf{A}) < 1$, we can verify (see Appendix E.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 7 and 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 $r(\mathbf{A}) < 1$. 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 6-8 hold. See Appendix E.2 for details. Therefore, Propositions 7 and 8 and thus Corollary 9 follow 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 $r(\mathbf{A}) < 1$:

$$\begin{aligned} Z_t &= \mathbf{A}Z_{t-1} + \Sigma(Z_{t-1})\mathcal{E}_t \\ \Sigma(z) &:= c \cdot \operatorname{clip}_{a,b}\left(\|z\|^m\right)\mathbf{I}_{p\times p} \end{aligned} \tag{4.1}$$

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 $[a,b]$; i.e.,

$$\operatorname{clip}_{a,b}(x) = \begin{cases} b & \text{if } x \geq 0 \\ x & \text{if } a < x < b \\ a & \text{otherwise.} \end{cases}$$

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 Appendix E.3 for details) that Assumptions 1-3 and 6-8 hold. Therefore, Propositions 7 and 8, and thus Corollary 9 follow and hence we have all the high probabilistic guarantees for lasso for a nonlinear models with subweibull innovations. Our example is admittedly contrived, but we hope that our techniques and results will allow other researchers to consider more compelling non-linear models.

5. Simulations.

In this section we report simulation results to study the effect of heavy tails and temporal dependence on the estimation error of lasso.

5.1. *Effect of Heavy-Tailedness.*

We conducted a simulation experiment to investigate the effect of heavy-tailedness via a subweibull VAR (Example 3). Consider a standard VAR Model

$$X_{t+1} = \mathbf{A}X_t + c\epsilon_t$$

where the underlying parameter matrix $\mathbf{A}$ is an s-sparse $p \times p$ matrix with spectral radius c, X_t is a $p \times 1$ vector, and ϵ_t is a subweibull random vector where each entry is iid Weibull random variable with shape parameter α and scale parameter 1. Moreover, ϵ_t is independent across time. For the simulation, $\mathbf{A}$ is generated by first randomly choosing s positions with non-zero entries and then sampling each non-zero entry iid from Uniform(0,1). Finally, $\mathbf{A}$ is rescaled so that its spectral radius is $c = 0.5$. Now, let $s = \sqrt{p}$, $p \in \{50, 100, 150, 200\}$, $\alpha \in \{0.4, 0.5, 1.0, 1.9\}$, and $T = m \times s \log(p)$ where multiples $m \in \{1, 3, 5, 7, 9, 11, 13, 15, 17, 19\}$. The average estimation error (Frobenius norm of the difference between true parameter matrix $\mathbf{A}$ and its estimated counterpart $\hat{\mathbf{A}}$) over 10 repetitions plotted against $\sqrt{\frac{s \log(p)}{T}}$ is shown in Figure 5.1.

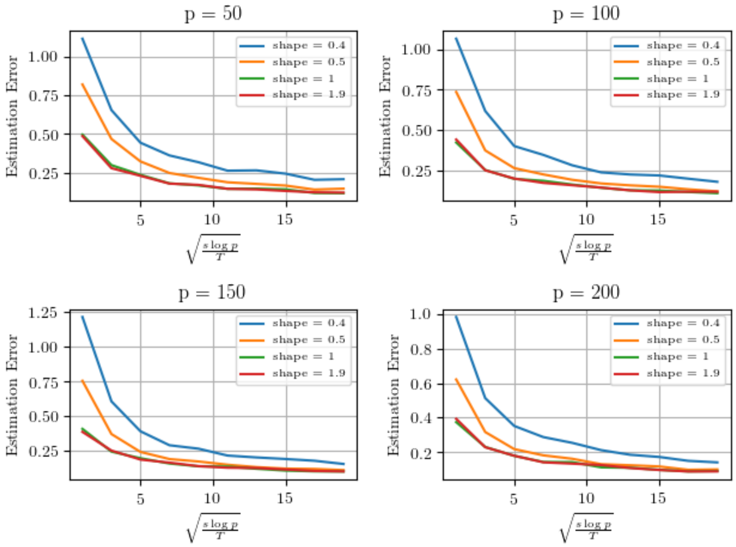

FIG 1. *Effect of Heavy Tails on Lasso Estimation Error. A smaller shape parameter corresponds to heavier tails of the noise.*

The relation between shape parameter and the estimation error is quite clear. Smaller shape parameter means a heavier tail, resulting in larger estimation error, which is indeed what we observe here. The unequal spacing in the choice of shape parameter is due to the fact that the relation of shape parameter and estimation error is highly non-linear, and using equal spacing would make the differences between plots less easy to see.

5.2. *Effect of Dependence.* Next we set up a simulation to study the effect of dependence on lasso estimation error. At time 0, we set $X_0 \sim \mathcal{N}(0, I_{p\times p})$, $Y_0 = \mathbf{A}X_0 + c\epsilon_0$. For $t \geq 1$, with probability ρ, (X_t, Y_t) is just a copy of the previous observations, i.e., $Y_t = Y_{t-1}$, $X_t = X_{t-1}$. With probability $1-\rho$, (X_t, ϵ_t) are fresh independent samples, $X_t \sim \mathcal{N}(0, I_{p\times p})$, $Y_t = \mathbf{A}X_t + c\epsilon_t$. The settings of $\mathbf{A}$ and ϵ_t are exactly the same as above in Subsection 5.1. We fix shape parameter $\alpha = 1$. Now, let $s = \sqrt{p}$, $p \in \{50, 100, 150, 200\}$, $\rho \in \{0.2, 0.4, 0.6, 0.8\}$, and $T = m \times s\log(p)$ where $m \in \{1, 3, 5, 7, 9, 11, 13, 15, 17, 19\}$. The average estimation error over 10 repetitions plotted against $\sqrt{\frac{s\log(p)}{T}}$ is shown in Figure 5.2. For all choices of the dimension p, the results confirm the intuition that higher ρ leads to higher estimation error as higher ρ implies more dependence in data.

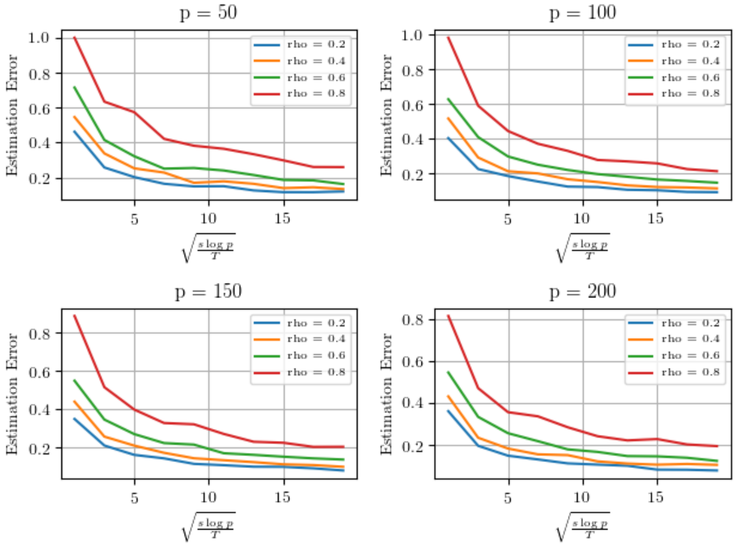

FIG 2. *Effect of Dependence on Lasso Estimation Error. A larger ρ parameter corresponds to more dependence in the generating process.*

6. Conclusion. One way to interpret the main results of this paper is that *the lasso procedure is robust to deviations from idealized conditions*, e.g., independence and subgaussian tails. Our work shows that lasso continues to enjoy theoretical guarantees even in the presence of dependence and heavy tails.

A key theoretical question we left largely unaddressed is that of *lower bounds*. Our simulations suggest that the performance of lasso does deteriorate as tails of random variable get heavier and there is more temporal dependence. It will be good to derive lower bounds on estimation and prediction errors in terms of subweibull and mixing parameters. If we can derive matching lower and upper bounds, then it will enhance our theoretical understanding of lasso.

We also note that there are several related topics that were outside the scope of the present paper but that merit further attention. A non-exhaustive list includes low-dimensional structures other than sparsity (and associated regularization penalties) (see, e.g., Negahban et al. (2012)), model selection consistency (see, e.g., Zhao and Yu (2006)), and post-selection inference (see, e.g., Lee et al. (2016)).

Acknowledgments. We thank Sumanta Basu and George Michailidis for helpful discussions, and Roman Vershynin for pointers to the literature. We acknowledge the support of NSF via a regular (DMS-1612549) and a CAREER grant (IIS-1452099). We would like to thank the anonymous reviewers whose many comments improved this paper considerably.

References.

Agarwal, A., Negahban, S. and Wainwright, M. J. (2012). Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics* **40** 2452–2482.

Alquier, P., Doukhan, P. et al. (2011). Sparsity considerations for dependent variables. *Electronic journal of statistics* **5** 750–774.

Basu, S. and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics* **43** 1535–1567.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences* **2** 183–202.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics* 1705–1732.

Bradley, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probability surveys* **2** 107–144.

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

Chandrasekaran, V., Recht, B., Parrilo, P. A. and Willsky, A. S. (2012). The convex geometry of linear inverse problems. *Foundations of Computational mathematics* **12** 805–849.

Chen, X., Xu, M. and Wu, W. B. (2013). Covariance and precision matrix estimation for high-dimensional time series. *The Annals of Statistics* **41** 2994–3021.

Chudik, A. and Pesaran, M. H. (2011). Infinite-dimensional VARs and factor models. *Journal of Econometrics* **163** 4–22.

Chudik, A. and Pesaran, M. H. (2013). Econometric analysis of high dimensional VARs featuring a dominant unit. *Econometric Reviews* **32** 592–649.

Chudik, A. and Pesaran, M. H. (2014). Theory and Practice of GVAR Modelling. *Journal of Economic Surveys*.

Davis, R. A., Zang, P. and Zheng, T. (2016). Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics* **25** 1077–1096.

Donoho, D. L., Maleki, A. and Montanari, A. (2009). Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences* **106** 18914–18919.

Fan, J., Qi, L. and Tong, X. (2016). Penalized least squares estimation with weakly dependent data. *Science China Mathematics* **59** 2335–2354.

Foss, S., Korshunov, D., Zachary, S. et al. (2011). *An introduction to heavy-tailed and subexponential distributions* **6**. Springer.

Guo, S., Wang, Y. and Yao, Q. (2016). High-dimensional and banded vector autoregressions. *Biometrika* asw046.

Han, F. and Liu, H. (2013). Transition matrix estimation in high dimensional time series. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)* 172–180.

Han, F., Lu, H. and Liu, H. (2015). A direct estimation of high dimensional stationary vector autoregressions. *The Journal of Machine Learning Research* **16** 3115–3150.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC Press.

Hayashi, F. (2000). Econometrics Princeton University Press.

Ibragimov, I. A. and Rozanov, Y. A. (1978). *Gaussian random processes*. Springer.

Kock, A. B. and Callot, L. (2015). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics* **186** 325–344.

Kulkarni, S., Lozano, A. C. and Schapire, R. E. (2005). Convergence and Consistency of Regularized Boosting Algorithms with Stationary β-Mixing Observations. In *Advances in neural information processing systems* 819–826.

Lee, J. D., Sun, D. L., Sun, Y., Taylor, J. E. et al. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics* **44** 907–927.

Liebscher, E. (2005). Towards a unified approach for proving geometric ergodicity and mixing properties of nonlinear autoregressive processes. *Journal of Time Series Analysis* **26** 669–689.

Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics* **40** 1637–1664.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.

Mcdonald, D. J., Shalizi, C. R. and Schervish, M. J. (2011). Estimating beta-mixing coefficients. In *International Conference on Artificial Intelligence and Statistics* 516–524.

McMurry, T. L. and Politis, D. N. (2015). High-dimensional autocovariance matrices and optimal linear prediction. *Electronic Journal of Statistics* **9** 753–788.

Medeiros, M. C. and Mendes, E. F. (2016). ℓ_1-regularization of high-dimensional time-series models with non-Gaussian and heteroskedastic errors. *Journal of Econometrics* **191** 255–271.

Merlevède, F., Peligrad, M. and Rio, E. (2011). A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields* **151** 435–474.

Nardi, Y. and Rinaldo, A. (2011). Autoregressive process modeling via the lasso procedure. *Journal of Multivariate Analysis* **102** 528–549.

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics* 1069–1097.

Negahban, S. and Wainwright, M. J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research* **13** 1665–1697.

Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012). A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers. *Statistical Science* **27** 538–557.

Ngueyep, R. and Serban, N. (2014). Large Vector Auto Regression for Multi-Layer Spatially Correlated Time Series. *Technometrics.*

Nicholson, W. B., Bien, J. and Matteson, D. S. (2014). Hierarchical Vector Autoregression. *arXiv preprint arXiv:1412.5250.*

Nicholson, W. B., Matteson, D. S. and Bien, J. (2017). VARX-L: Structured regularization for large vector autoregressions with exogenous variables. *International Journal of Forecasting* **33** 627–651.

Pisier, G. (2016). Subgaussian sequences in probability and Fourier analysis. arXiv preprint arXiv:1607.01053v3.

Qiu, H., Xu, S., Han, F., Liu, H. and Caffo, B. (2015). 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)* 1843–1851.

Raskutti, G., Wainwright, M. J. and Yu, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *The Journal of Machine Learning Research* **11** 2241–2259.

Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences of the United States of America* **42** 43.

Rudelson, M. and Vershynin, R. (2013). Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability* **18** 1–9.

Rudelson, M. and Zhou, S. (2013). Reconstruction from anisotropic random measurements. *Information Theory, IEEE Transactions on* **59** 3434–3447.

Sivakumar, V., Banerjee, A. and Ravikumar, P. K. (2015). Beyond Sub-Gaussian Measurements: High-Dimensional Structured Estimation with Sub-Exponential Designs. In *Advances in Neural Information Processing Systems* 2206–2214.

Song, S. and Bickel, P. J. (2011). Large vector auto regressions. *arXiv preprint arXiv:1106.3915.*

Tao, T. and Vu, V. (2013). Random matrices: Sharp concentration of eigenvalues. *Random Matrices: Theory and Applications* **2** 1350007.

Tjøstheim, D. (1990). Non-linear time series and Markov chains. *Advances in Applied Probability* 587–611.

Uematsu, Y. (2015). Penalized Likelihood Estimation in High-Dimensional Time Series Models and uts Application. *arXiv preprint arXiv:1504.06706.*

Vidyasagar, M. (2003). *Learning and generalisation: with applications to neural networks*, second ed. Springer Science & Business Media.

Wang, Z., Han, F. and Liu, H. (2013). Sparse principal component analysis for high dimensional multivariate time series. In *Artificial Intelligence and Statistics* 48–56.

Wang, H., Li, G. and Tsai, C.-L. (2007). Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69** 63–78.

Wong, K. C., Li, Z. and Tewari, A. (2016). Lasso Guarantees for Time Series Estimation

Under Subgaussian Tails and β-Mixing. *arXiv preprint arXiv:1602.04265*.

Wu, W. B. and Wu, Y. N. (2016). High-dimensional linear models with dependent observations. *Electronic Journal of Statistics* **10** 352–379.

Yu, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability* 94–116.

Zhang, D., Wu, W. B. et al. (2017). Gaussian approximation for high dimensional time series. *The Annals of Statistics* **45** 1895–1919.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine learning research* **7** 2541–2563.

APPENDIX A: PROOF OF MASTER THEOREM

Proof of Theorem 1. The proof follows from optimality of $\widehat{\Theta}$ and the definitions of the RE and DB conditions.

1. Since $\widehat{\Theta}$ is optimal for optimization problem (2.2) and $\Theta^\star$ is feasible,
(A.1)
$$\frac{1}{T}\left|\!\left|\!\left|\mathbf{Y}-\mathbf{X}\widehat{\Theta}\right|\!\right|\!\right|_F^2+\lambda_T\left\|\mathrm{vec}(\widehat{\Theta})\right\|_1\leq\frac{1}{T}|\!|\!|\mathbf{Y}-\mathbf{X}\Theta^\star|\!|\!|_F^2+\lambda_T\left\|\mathrm{vec}(\Theta^\star)\right\|_1$$

2. Let $\hat{\Delta}:=\widehat{\Theta}-\Theta^\star\in\mathbb{R}^{p\times q}$. Then, rearranging the inequality (A.1) above yields
$$\text{(A.2)}\quad \frac{1}{T}\left|\!\left|\!\left|\mathbf{X}\hat{\Delta}\right|\!\right|\!\right|_F^2\leq\frac{2}{T}\mathrm{tr}(\hat{\Delta}'\mathbf{X}'\mathbf{W})+\lambda_T\left(\|\mathrm{vec}(\Theta^\star)\|_1-\left\|\mathrm{vec}(\widehat{\Theta})\right\|_1\right)$$

3. Let S denote the support of $\Theta^\star$. For any matrix $\mathbf{M}$, let $\mathbf{M}_S$ be the components of $\mathbf{M}$ restricted to the support S and similarly for $\mathbf{M}_{SC}$; therefore $\Theta^\star=\Theta^\star_S+\Theta^\star_{SC}$. With these, note that
$$\begin{aligned}\left\|\mathrm{vec}(\Theta^\star+\hat{\Delta})\right\|_1-\|\mathrm{vec}(\Theta^\star)\|_1\geq&\{\|\mathrm{vec}(\Theta^\star_S)\|_1-\left\|\mathrm{vec}(\hat{\Delta}_S)\right\|_1\}\\&+\left\|\mathrm{vec}(\hat{\Delta}_{S^c})\right\|_1-\|\mathrm{vec}(\Theta^\star)\|_1\\&=\left\|\mathrm{vec}(\hat{\Delta}_{S^c})\right\|_1-\left\|\mathrm{vec}(\hat{\Delta}_S)\right\|_1\end{aligned}$$

4. Recall the *RE* condition with parameters α and τ, and deviation bound condition with constant $\mathbb{Q}(\Sigma_X,\Sigma_W)$. For tuning parameter $\lambda_T\geq$

$2\mathbb{Q}(\Sigma_X, \Sigma_W)\sqrt{\frac{\log(q)}{T}}$, we have

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

5. 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, we have a lower bound
$$\left\|\operatorname{vec}(\hat{\Delta})\right\|_1 \leq 4\left\|\operatorname{vec}(\hat{\Delta}_S)\right\|_1 \leq 4\sqrt{s}\left\|\operatorname{vec}(\hat{\Delta})\right\|$$
6. Finally, with $\alpha \geq 32s\tau$,
$$
\begin{aligned}
\frac{\alpha}{2}\left\|\operatorname{vec}(\hat{\Delta})\right\|_2^2 &\leq (\alpha - 16s\tau)\left\|\operatorname{vec}(\hat{\Delta})\right\|_2^2 \\
&\leq \alpha\left\|\operatorname{vec}(\hat{\Delta})\right\|_2^2 - \tau\|\operatorname{vec}(\hat{\Delta})\|_1^2 \\
&\leq 2\lambda_T\|\operatorname{vec}(\hat{\Delta})\|_1 \\
&\leq 2\sqrt{s}\lambda_T\|\hat{\Delta}\|_2
\end{aligned}
$$
Thus, we have the upper bound
$$\left\|\operatorname{vec}(\hat{\Delta})\right\|_2 \leq \frac{4\lambda_T\sqrt{s}}{\alpha}$$
7. From steps (4) and (5), we have
$$\frac{1}{T}\left|\!\left|\!\left|\mathbf{X}\hat{\Delta}\right|\!\right|\!\right|_F^2 \leq 8\lambda_T\sqrt{s}\left\|\operatorname{vec}(\hat{\Delta})\right\|_2$$

Together with step (6), we obtain that

$$\frac{1}{T}\left|\!\left|\!\left|\mathbf{X}\hat{\Delta}\right|\!\right|\!\right|_F^2 \leq 8\lambda_T\sqrt{s}\left\|\mathrm{vec}(\hat{\Delta})\right\|_2 \leq 32\lambda_T^2 s/\alpha$$

□

APPENDIX B: PROOFS FOR GAUSSIAN PROCESSES UNDER α-MIXING

We will also need the following result to control operator norms of matrices in terms of ℓ_1 norms of the rows and columns.

FACT 4 (Schur Test). For any matrix $\mathbf{M}$, we have

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

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

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

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

PROOF. We have,

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

The proof of Proposition 3 relies on the following result.

LEMMA 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 $\mathcal{B} := \sigma(Y)$, and $\mathcal{L}^2(\mathcal{C})$ denote the space of square-integrable, C-measurable (real-valued) random variables.

$$\rho(X,Y) := \sup\{\mathrm{cor}(f(X), g(Y)) \mid f \in \mathcal{L}^2(\mathcal{A}), g \in \mathcal{L}^2(\mathcal{B})\}$$

Then, in particular, we consider the one-dimensional projections and obtain that

$$\begin{aligned} \rho(X,Y) &\geq \mathrm{cor}(u'X,\, v'Y) && \forall \text{ fixed } u,\, v \\ &= \frac{|\mathbb{E}[u'Xv'Y]|}{\sqrt{\mathbb{E}(u'X)^2\mathbb{E}(v'Y)^2}} && u,\, v \text{ non-zero} \end{aligned}$$

Re-arranging, $\forall u,\, v$ 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,

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

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,

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

Proof of Proposition 2. Note that by Fact 1 α and ρ-mixing are equivalent for stationary Gaussian processes. The proof will operate via arguments in ρ-mixing coefficients.

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

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 (2.1), 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 $\forall i, j$

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

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 a proof strategy where we control each of the three tail probabilities above. Assuming the conditions in Proposition 2, we can apply the Hanson-Wright inequality (Lemma 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 (B.1). By Lemma (11), there exists a constant $c > 0$ such that $\forall t \geq 0$

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

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

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

Therefore,

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

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

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

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

$$\Sigma_{\mathbf{W}_{:i}} = \Sigma_{\mathbf{Y}_{:i}} + \Sigma_{\mathbf{X}\Theta^\star_{:i}} + \Sigma_{\mathbf{Y}_{:i}, \mathbf{X}\Theta^\star_{:i}} + \Sigma_{\mathbf{X}\Theta^\star_{:i}, \mathbf{Y}_{:i}} \tag{B.2}$$

By Claim 1, we have

$$|\!|\!|\Sigma_{\mathbf{W}_{:i}}|\!|\!| \leq 2|\!|\!|\Sigma_{\mathbf{Y}_{:i}}|\!|\!| + 2|\!|\!|\Sigma_{\mathbf{X}\Theta^\star_{:i}}|\!|\!| \tag{B.3}$$

Let us figure out each of the summands on the RHS of equation (B.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,

$$\begin{aligned}
|\!|\!|\Sigma_{\mathbf{X}\Theta^\star_{:i}}|\!|\!| &\leq \max_{1\leq k\leq T}\left\|(\Sigma_{\mathbf{X}\Theta^\star_{:i}})[k,:]\right\|_1 && \text{by Fact 4} \\
&\leq 2\sum_{l=0}^{T}\rho(l)\left\|\Theta^\star_{:i}\right\|_2^2 |\!|\!|\Sigma_X(0)|\!|\!| && \text{by Lemma 10}
\end{aligned}$$

Therefore,

$$\max_{1\leq i\leq p}|\!|\!|\Sigma_{\mathbf{X}\Theta^\star_{:i}}|\!|\!| \leq \max_{1\leq i\leq p} 2\sum_{l=0}^{T}\rho(l)\left\|\Theta^\star_{:i}\right\|_2^2|\!|\!|\Sigma_X(0)|\!|\!| = 2|\!|\!|\Sigma_X(0)|\!|\!|\sum_{l=0}^{T}\rho(l)\max_{1\leq i\leq p}\left\|\Theta^\star_{:i}\right\|_2^2.$$

Similarly,

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

and

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

So, by inequality (B.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,

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

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

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

that

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

Also, because of Fact 2, we have

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

□

PROOF OF PROPOSITION 3. Note that, by Fact 2, 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 (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 4 and Lemma 10 to upper bound $|||\mathbf{Q}|||$ by

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

Now, we can apply Lemma 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 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 $\forall v \in \mathbb{R}^p$,

$$\mathbb{P}\left\{|v'(\hat{\Gamma} - \Sigma_X(0))v| > 27\eta \sum_{t=0}^{T} \rho(l) |||\Sigma_X(0)||| \left(\|v\|_2^2 + \frac{1}{k}\|v\|_1^2\right)\right\}$$
$$\leq 2 \exp\{-cT \min(\eta, \eta^2) + 2k \min(\log(p), \log(\frac{21ep}{2k}))\}$$
$$\Updownarrow$$
$$\mathbb{P}\left\{|v'(\hat{\Gamma} - \Sigma_X(0))v| \leq 27\eta \sum_{t=0}^{T} \rho(l) |||\Sigma_X(0)||| \left(\|v\|_2^2 + \frac{1}{k}\|v\|_1^2\right)\right\}$$
$$> 1 - 2 \exp\{-cT \min(\eta, \eta^2) + 2k \min(\log(p), \log(\frac{21ep}{2k}))\}$$
$$\Downarrow$$
$$\mathbb{P}\left\{|v'(\hat{\Gamma})v| > -27\eta \sum_{t=0}^{T} \rho(l) |||\Sigma_X(0)||| \left(\|v\|_2^2 + \frac{1}{k}\|v\|_1^2\right) + \lambda_{\min}(\Sigma_X(0))\|v\|_2^2\right\}$$
$$> 1 - 2 \exp\{-cT \min(\eta, \eta^2) + 2k \min(\log(p), \log(\frac{21ep}{2k}))\}$$

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 2. □

APPENDIX C: HANSON-WRIGHT INEQUALITY

The general statement of the Hanson-Wright inequality can be found in the paper by Rudelson and Vershynin (2013) (see their 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 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 $\eta > 0$,*

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

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,

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

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

APPENDIX D: PROOFS FOR SUBWEIBULL RANDOM VECTORS UNDER β-MIXING

D.1. Proof Related to Subweibull Properties.

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

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

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 $\gamma \le 1$, this covers all $p \ge 1$. If $\gamma > 1$, we have, for $p = 1, \dots, \lceil \gamma \rceil - 1$,

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

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

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

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,

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

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 $t > 0$,

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

Proof. (of Lemma 6) By definition,

$$\begin{aligned}
\left\|X^2\right\|_{\psi_\gamma} &= \sup_{p\ge1} p^{-1/\gamma}\left(\mathbb{E}\left|X^2\right|^p\right)^{1/p} \\
&= \sup_{p\ge1}\left(p^{-1/(2\gamma)}\left(\mathbb{E}\,|X|^{2p}\right)^{1/2p}\right)^2
\end{aligned}$$

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

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

D.2. Subweibull Norm Under Linear Transformations. We will need the following result about changes to the subweibull norm under linear transformations.

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

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

PROOF. We have,

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

D.3. Concentration Inequality for Sums of β-Mixing Subweibull Random Variables. We will state and prove a modified form of Theorem 1 of Merlevède, Peligrad and Rio (2011). This concentration result will be used to prove the high probability guarantees on the deviation bound (Proposition 7) and lower restricted eigenvalue (Proposition 8) conditions.

LEMMA 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 $t > 1/T$,

$$\mathbb{P}\left\{\left|\frac{S_T}{T}\right| > t\right\} \le T\exp\left\{-\frac{(tT)^\gamma}{K^\gamma C_1}\right\} + \exp\left\{-\frac{t^2 T}{K^2 C_2}\right\} \tag{D.1}$$

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, Peligrad and Rio (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, Peligrad and Rio (2011), under the condition of our lemma, gives

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

First of all, we need to control the quantity V that appears in the denominator of the second term of (D.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, Peligrad and Rio (2011) mention in their Section 2.1.1, using results of **?**, we have, for any β-mixing strictly stationary sequence (Y_t) with geometrically decaying β-mixing coefficients; i.e.,

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

the associated quantity V can be upper bounded as

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

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

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

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

$$\frac{(u)^{(1-\gamma)}}{(\log(u))}, \tag{D.3}$$

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 (D.2) can then be combined into one.

Let $u > 1$. Note that the expression (D.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 (D.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 $\gamma < 1$. □

D.4. Proofs of Deviation and RE Bounds.

Proof. (of Proposition 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 $|||\mathbf{X}'\mathbf{W}|||_\infty = \max_{1\le i\le p, 1\le j\le q} |[\mathbf{X}'\mathbf{W}]_{i,j}| = \max_{1\le i\le p, 1\le j\le q} |(\mathbf{X}_{:i})'\mathbf{W}_{:j}|$.

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 (2.1), 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 $\forall i, j$

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

Therefore,

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

We will now control each of the three the tail probabilities above. Before we apply Lemma 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 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,

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

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 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 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 Proposition 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 1. From Lemma 6, we know that $\left\|Z_t^2\right\|_{\psi_{\gamma_2/2}} \leq 2^{2/\gamma_2} \|Z_t\|_{\psi_{\gamma_2}}^2$. Moreover, $\|Z_t\|_{\psi_{\gamma_2}} \leq \|X_t\|_{\psi_{\gamma_2}}$. Therefore, we can invoke Lemma 13 for the sum $S_T(v) = \sum_{t=1}^T \left(Z_t^2 - \mathbb{E}Z_t^2\right)$ 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

$$1 - \exp\left\{\log(T) - \frac{(tT)^\gamma}{K^\gamma C_1} + k\log(p)\right\} - \exp\left\{-\frac{t^2 T}{K^2 C_2} + k\log(p)\right\} \tag{D.4}$$

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

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

Let $\hat{\Sigma}_T(v) := \frac{1}{T}\|\mathbf{X}v\|_2^2$ 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

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

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

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

so that the failure probability in (D.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

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

APPENDIX E: VERIFICATION OF ASSUMPTIONS FOR THE EXAMPLES

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

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

where

(E.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 1.

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 5. 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 2.5) 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 1. It will be useful to note the fact that spectral radius $r(\tilde{A}) < 1$ implies that $\exists k \in \mathbf{Z}$ such that $\left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert \tilde{A}^k \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert < 1$.

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 then set $K_c = \{x \in \mathbb{R}^p : \|x\| \le \frac{4C_{A\epsilon}}{c}\}$ where $c = 1 - \left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert \tilde{A}^k \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert$ and $C_{A\epsilon} := \sum_{i=1}^{k+1} \left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert \tilde{A}^{k-i} \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert \mathbb{E} \|\epsilon_{t-k+1}\|$. Then, since spectral radius $r(\tilde{A}) < 1$,

- For all $z \in E \backslash K_c$; i.e. z such that $\|z\| > \frac{4C_{A\epsilon}}{c}$,

$$\begin{aligned}\mathbb{E}\left[\left\| \tilde{Z}_{t+1} \right\| \middle| \tilde{Z}_t = z\right] &\le \left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert \tilde{A}^k \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert \|z\| + \sum_{i=1}^{k+1} \left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert \tilde{A}^{k-i} \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert \mathbb{E} \|\epsilon_{t-k+1}\| \\ &\equiv (1 - c)\|z\| + C_{A\epsilon} \\ &< \left(1 - \frac{c}{2}\right)\|z\| - C_{A\epsilon}.\end{aligned}$$

- For all $z \in K_c$,

$$\mathbb{E}\left[\left\| \tilde{Z}_{t+1} \right\| \middle| \tilde{Z}_t = z\right] < \left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert \tilde{A} \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert \|z\| + C_{A\epsilon} \le \frac{4C_{A\epsilon}(1-c)}{c} + C_{A\epsilon}$$

- For all $z \in K_c$,

$$0 \le \|z\| \le \frac{4C_{A\epsilon}}{c}.$$

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 $\left|\left|\left|\tilde{A}^k\right|\right|\right| < 1$. By the recursive nature of the time series,

$$\|Z_t\|_{\psi_\gamma} \leq \left|\left|\left|\tilde{A}^k\right|\right|\right| \|Z_{t-1}\|_{\psi_\gamma} + \sum_{i=1}^{k+1} \left|\left|\left|\tilde{A}^{k-i}\right|\right|\right| \|\mathcal{E}_{t-k+i}\|_{\psi_\gamma} \tag{E.3}$$

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

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

The last inequality follows because $C_i < \infty, \forall i = 1, \cdots, k$. Thus, the sequence (Z_t) satisfies Assumption 7.

E.2. VAR with Misspecification. Assumption 3 is immediate from model definitions. By the same arguments as in Appendix E.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 Appendix E.1. By Fact 1, 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. (2.1)

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

Taking derivatives and setting to zero, we obtain

$$(\Theta^\star)' = \Sigma_Z(-1)(\Sigma_Z)^{-1} \tag{E.4}$$

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 Appendix E.1 pertaining to establishing the subweibull property in conjunction with applying Lemma 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 Lütkepohl (2005). Our results extend to any VAR(d) processes.

E.3. ARCH.

Verifying the Assumptions. To show that Assumption 6 hold for a process defined by Eq. (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 Appendix E.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 $m \in (0,1)$.
- $\left|\!\left|\!\left|\Sigma(z)^{-1}\right|\!\right|\!\right| \leq 1/(ac)$, $|\det(\Sigma(z))| \leq bc$
- $r(\mathbf{A}) < 1$

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,

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

where $K_e := \sup_t \left\|\mathcal{E}_t\right\|_{\psi_\gamma}$. By the same arugments as in Equation E.3, we have that 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 4.1.

Recall Eq. E.4 from Appendix E.2 that

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

Now,

$$\begin{aligned}
\Sigma_Z(-1) &= \mathbb{E} Z_t Z_{t-1}' && \text{by stationarity} \\
&= \mathbb{E}\left(\mathbf{A} Z_{t-1} + \Sigma(Z_{t-1})\mathcal{E}_t\right) Z_{t-1}' && \text{Eq. (4.1)} \\
&= \mathbf{A}\mathbb{E} Z_{t-1} Z_{t-1}' + \mathbb{E}\Sigma(Z_{t-1})\mathcal{E}_t Z_{t-1}' && \\
&= \mathbf{A}\Sigma_Z + \mathbb{E}[c\,\mathrm{clip}_{a,b}\left(\|Z_{t-1}\|^m\right)\mathcal{E}_t Z_{t-1}'] && \\
&= \mathbf{A}\Sigma_Z + \mathbb{E}[c\mathcal{E}_t Z_{t-1}' \mathrm{clip}_{a,b}\left(\|Z_{t-1}\|^m\right)] && \\
&= \mathbf{A}\Sigma_Z + c\mathbb{E}\left[\mathcal{E}_t\right]\mathbb{E}\left[Z_{t-1}'\mathrm{clip}_{a,b}\left(\|Z_{t-1}\|^m\right)\right] && \text{i.i.d. innovations} \\
&= \mathbf{A}\Sigma_Z && \mathcal{E}_t \text{ mean } 0,
\end{aligned}$$

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

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

Address of the authors,
E-mail: kamwong@umich.edu; zifan.li@yale.edu; tewaria@umich.edu