The Annals of Statistics
2020, Vol. 48, No. 2, 1124–1142
https://doi.org/10.1214/19-AOS1840

LASSO GUARANTEES FOR β-MIXING HEAVY-TAILED TIME SERIES

BY KAM CHUNG WONG[1,*], ZIFAN LI[2] AND AMBUJ TEWARI[1,**]

[1]*Department of Statistics, University of Michigan,* [*]*kamwong@umich.edu;* [**]*tewaria@umich.edu*
[2]*Department of Statistics, Yale University, zifan.li@yale.edu*

Many theoretical results for lasso require the samples to be i.i.d. Recent work has provided guarantees for lasso assuming that the time series is generated by a sparse Vector Autoregressive (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 nonasymptotic 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 sub-Weibull random vectors. The class of sub-Weibull random variables that we introduce includes sub-Gaussian 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 nonlinear 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., Agarwal, Negahban and Wainwright (2012), Beck and Teboulle (2009), Donoho, Maleki and Montanari (2009)). However, the bulk of existing theoretical work focuses on i.i.d. 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 Autoregressive (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 (Han and Liu (2013), Loh and Wainwright (2012), Negahban and Wainwright (2011)) that assumed operator norm bounds on the transition matrix. These operator norm conditions are restrictive even

Received August 2017; revised March 2019.
MSC2010 subject classifications. 60K35.
Key words and phrases. Time series, mixing, high-dimensional estimation, lasso.

for VAR models with a lag of 1 and never hold (Please see pages 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 $Z_{t-d}, \ldots, Z_{t-1}$ is well defined even if Z_t is not a lag d VAR. If we assume that this best linear predictor involves sparse coefficient matrices, can we still guarantee consistent parameter estimation? 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 sub-Weibull 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 sub-Weibull β-mixing assumptions, respectively (Propositions 7 and 8). Note that the RE conditions were previously shown to hold under the i.i.d. assumption by Raskutti, Wainwright and Yu (2010) for Gaussian random vectors and for sub-Gaussian 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 and Doukhan (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 autocovariance structure (please refer to Appendix D for details (Wong, Li and Tewari (2020)), 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 (Guo, Wang and Yao (2016), Ngueyep and Serban (2015), Nicholson, Bien and Matteson (2014), Nicholson, Matteson and Bien (2017)). Zhang and Wu (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 sub-Weibull observations and we make the dependence on the sub-Weibull 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 sub-Weibull 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 sub-Weibull random vectors. We consider VAR models with sub-Weibull innovations when the model is correctly or incorrectly specified (Examples 3 and 4). In addition, we go beyond linear models and present a nonlinear 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 misspecification, 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 $(Z_{t-d}, \ldots, Z_{t-1})$. A linear predictor is a convenient choice. To frame it as a regression problem, we identify $Y_t = Z_t$ and $X_t = (Z_{t-d}, \ldots, Z_{t-1})$. The pairs (X_t, Y_t) defined as such are no longer i.i.d. Assuming strict stationarity, the parameter matrix of interest $\Theta^\star \in \mathbb{R}^{p\times q}$ is

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

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

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

where

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

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 shorthand $a \wedge b := \min\{a, b\}$ 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 $\lambda_i(M)$, $i = 1, \ldots, d$ 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 $\sqrt{\lambda_{\max}(\mathbf{M}'\mathbf{M})}$, entry-wise ℓ_∞ norm $\max_{i,j} |\mathbf{M}_{i,j}|$, and Frobenius norm $\sqrt{\operatorname{tr}(\mathbf{M}'\mathbf{M})}$, respectively. For any vector $v \in \mathbb{R}^p$, $\|v\|_q$ denotes its ℓ_q norm $(\sum_{i=1}^p |v_i|^q)^{1/q}$. Unless otherwise specified, we shall use $\|\cdot\|$ to denote the ℓ_2 norm. For any vector $v \in \mathbb{R}^p$, we use $\|v\|_0$ and $\|v\|_\infty$ to denote $\sum_{i=1}^p \mathbb{1}\{v_i \neq 0\}$ and $\max_i\{|v_i|\}$, respectively. Similarly, for any matrix $\mathbf{M}$, $|||\mathbf{M}|||_0 = \|\operatorname{vec}(\mathbf{M})\|_0$ where $\operatorname{vec}(\mathbf{M})$ is the vector obtained from $\mathbf{M}$ by concatenating the rows of M. We say that matrix $\mathbf{M}$ (resp., vector v) is *s-sparse* if $|||\mathbf{M}|||_0 = s$ (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 $1 \le i \le p$, $1 \le j \le q$, 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 autocovariance 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 autocovariance 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 submatrices 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 sub-Weibull and Gaussian scenarios. Our goal is to provide finite sample bounds on the error $\widehat{\Theta} - \Theta^\star$. We shall present theoretical guarantees on the ℓ_2 parameter estimation error $\|\operatorname{vec}(\widehat{\Theta} - \Theta^\star)\|_2$ and also the associated (in-sample) prediction error $|||(\widehat{\Theta} - \Theta^\star)'\hat{\Gamma}(\widehat{\Theta} - \Theta^\star)|||_F$.

ASSUMPTION 1. The matrix $\Theta^\star$ is s-sparse; that is, $\|\mathrm{vec}(\Theta^\star)\|_0 = s$.

ASSUMPTION 2. The process (X_t, Y_t) is strictly stationary; that is, $\forall m, \tau, n \geq 0$,

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

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

ASSUMPTION 3. The process (X_t, Y_t) is centered; that is, $\forall t$, $\mathbb{E}(X_t) = \mathbf{0}_{p\times 1}$, 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 nonasymptotic guarantees for lasso estimation and prediction errors under two well-known conditions, namely, the restricted eigenvalue (RE) and the deviation bound (DB) conditions. Note that in the classical linear model setting (see, e.g., Hayashi (2000), Chapter 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; that is, $\lambda_{\min}(\mathbf{Q}) > 0$, and (b) the regressors and the noise are asymptotically uncorrelated; that is, $\mathbf{X}'\mathbf{W}/T \to \mathbf{0}$.

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

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

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

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

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(α_C, τ) 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*:

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

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

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 sub-Weibull random vectors (we define sub-Weibull random vectors below in Section 4.1). Historically, mixing conditions were introduced to generalize the classic limit theorems in probability beyond the case of i.i.d. random variables (Rosenblatt (1956)). Recent work on high-dimensional statistics has established the validity of RE conditions in the i.i.d. Gaussian (Raskutti, Wainwright and Yu (2010)) and i.i.d. sub-Gaussian cases (Rudelson and Zhou (2013)). One of the main contributions of our work is to extend these results in high-dimensional statistics from the i.i.d. 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 sub-Weibull.

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

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

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

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

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

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

Fix i, j and write

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

The Hanson–Wright inequality (Lemma 11) takes care of the Gaussian process case. For the independent sub-Gaussian 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 *sub-Gaussian*.

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 *sub-Weibull* 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 i.i.d. 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,

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

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:

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

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

where the last supremum is over all pairs of partitions $\{A_1,\dots,A_I\}$ and $\{B_1,\dots,B_I\}$ 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), Section 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 i.i.d. setting. At the same time, many interesting processes such as Markov and hidden Markov processes satisfy a β-mixing condition (Vidyasagar (2003), Section 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, namely, 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:

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

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*

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

where

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

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

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

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

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

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 a free parameter defined as in Propositions 2 and 3. Then, for sample size,*

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

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

we have, with probability at least $1 - 2\exp(-\frac{c}{2}T\min\{1,\eta^2\}) - 8\exp(-b\log(pq))$, that the lasso error bounds (2.5) and (2.6) hold with

$$\alpha_C = \frac{1}{2}\lambda_{\min}(\Sigma_X), \quad \text{and}$$

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

where

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

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

REMARK 3. If the α-mixing coefficients are summable, that is, $S_\alpha(T) \leq \tilde{\alpha} < \infty$, $\forall T$, then we get the usual convergence rate of $O(\sqrt{\frac{\log(pq)}{T}})$. Also, the threshold sample size is $O(s\log(pq))$. This is in agreement with what happens in the i.i.d. 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 $S_\alpha(T) \in o(\sqrt{T})$, 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 autoregressive representation:

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

where each $\mathbf{A}_k, k = 1, \ldots, d$ is a sparse nonstochastic 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*; that is, $\det(\mathbf{I}_{p\times p} - \sum_{k=1}^{d} \mathbf{A}_k z^k) \neq 0$, $\forall |z| \leq 1$. Now, we identify $X_t := (Z_t', \ldots, Z_{t-d+1}')'$ 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), Chapter 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 misspecifies 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 misspecification and this example also serves to illustrate that our theory is applicable to models beyond the finite order VAR setting.

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

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

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

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

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

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

Note that Z_t is not necessarily a finite-rder 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. Sub-Weibull random vectors under β-mixing. Existing analyses of lasso mostly assume data have sub-Gaussian or subexponential tails. These assumptions ensure that the moment generating function exists, at least for some values of the free parameter. Nonexistence of the moment generating function is often taken as a definition of having a heavy tail (Foss, Korshunov and Zachary (2011)). We now introduce a family of random variables that subsumes sub-Gaussian and subexponential random variables. In addition, it includes some heavy-tailed distributions.

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

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

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 "sub-Weibull(γ)."

LEMMA 5 (Sub-Weibull 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*

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

2. *The moments of X satisfy*

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

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

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

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, e.g., Lemma 3.1 in the lecture notes of Pisier (2016)).

DEFINITION 3 (Sub-Weibull(γ) random variable and norm). A random variable X that satisfies any property in Lemma 5 is called a sub-Weibull(γ) random variable. The sub-Weibull(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 $\gamma > 0$,

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

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

REMARK 5. It is common in the literature (see, e.g., Foss, Korshunov and Zachary (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, sub-Weibull(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 sub-Weibull parameter γ and the associated constant, under squaring.

LEMMA 6. *For any $\gamma \in (0, \infty)$, if a random variable X is sub-Weibull(2γ) then X^2 is sub-Weibull(γ). Moreover,*

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

We now define the sub-Weibull 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 sub-Gaussian 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 sub-Weibull(γ) random vector if all of its one-dimensional projections are sub-Weibull(γ) random variables. We define the sub-Weibull(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 sub-Weibull 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; that is, there exist constants $c > 0$ and $\gamma_1 > 0$ such that

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

ASSUMPTION 7. Each random vector in the sequences (X_t) and (Y_t) follows a sub-Weibull(γ_2) distribution with $\|X_t\|_{\psi_{\gamma_2}} \le K_X$, $\|Y_t\|_{\psi_{\gamma_2}} \le K_Y$ for $t = 1, \ldots, T$.

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 $\gamma_1 \ge 1$ and $\gamma_2 \ge 2$ 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, that is, 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 sub-Weibull case). *Suppose Assumptions 1–3 and 6–8 hold. Let $c' > 0$ be a universal constant and let K be defined as*

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

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

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

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 sub-Weibull case). *Suppose Assumptions 1–3 and 6–8 hold. Let*

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

Then for sample size

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

we have with probability at least

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

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

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

where $\alpha_C = \frac{1}{2}\lambda_{\min}(\Sigma_X)$ and $\tau = \frac{\alpha_C}{2\tilde{c}} \cdot (\frac{\log(p)}{T^\gamma})$. 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 sub-Weibull 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 $K := 2^{2/\gamma_2}(K_Y + K_X(1 + |||\Theta^\star|||))^2$.*

Then for sample size

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

we have with probability at least

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

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 i.i.d. 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 apply even in the presence of one or more of the following: heavy-tailed sub-Weibull data, model misspecification and nonlinearity in the DGM.

EXAMPLE 3 (Sub-Weibull VAR). We study a generalization of the VAR, one that has sub-Weibull(γ_2) realizations. Consider a VAR(1) model defined as in Example 1 except that we replace the Gaussian white noise innovations with i.i.d. random vectors from some sub-Weibull(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 sub-Weibull random vector.

Now, we identify $X_t := (Z_t', \ldots, Z_{t-d+1}')'$ 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 sub-Weibull innovations and omitted variable). Using the same setup as in Example 2 except that we replace the Gaussian white noise innovations with i.i.d. random vectors from some sub-Weibull(γ_2) distribution with a nonsingular 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 sub-Weibull 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 sub-Weibull 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 sub-Weibull 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 $r(\mathbf{A}) < 1$:

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

where $\mathcal{E}_t$ are i.i.d. random vectors from some sub-Weibull(γ_2) distribution with a nonsingular covariance matrix Σ_ϵ, and $\mathrm{clip}_{a,b}(x)$ clips the argument x to stay in the interval $[a, b]$; that is,

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

Consequently, each Z_t will be a mean zero sub-Weibull 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 nonlinear 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 sub-Weibull 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 a s-sparse $p \times p$ matrix with spectral radius c, X_t is a $p \times 1$ vector and ϵ_t is a sub-Weibull random vector where each entry is i.i.d. 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 nonzero entry i.i.d. 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 1.

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 nonlinear, 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, that is, $Y_t = Y_{t-1}$, $X_t = X_{t-1}$. With probability $1 - \rho$, (X_t, ϵ_t) are fresh independent samples,

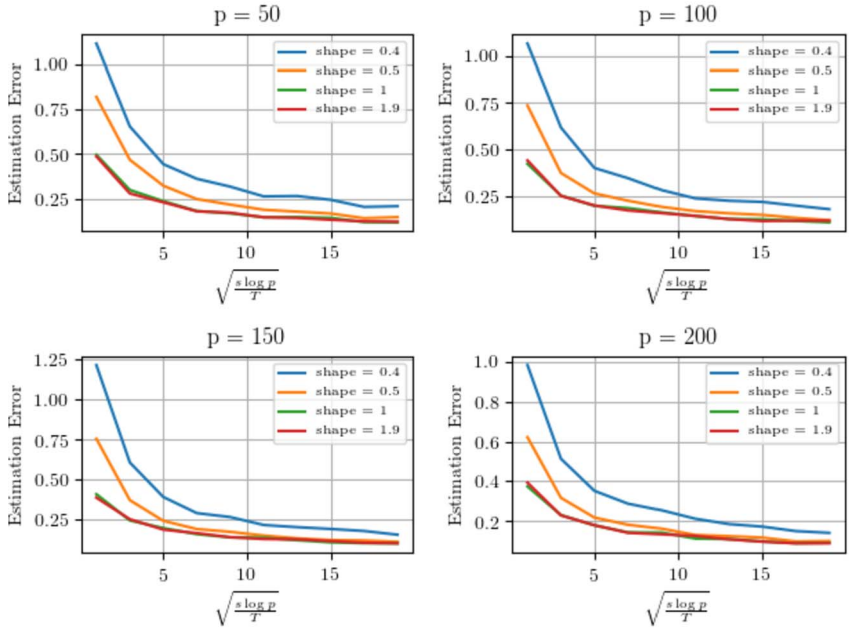

FIG. 1. *Effect of heavy tails on lasso estimation error. A smaller shape parameter corresponds to heavier tails of the noise.*

$X_t \sim \mathcal{N}(0, I_{p\times p})$, $Y_t = \mathbf{A}X_t + c\epsilon_t$. The settings of $\mathbf{A}$ and ϵ_t are exactly the same as above in Section 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 2. For

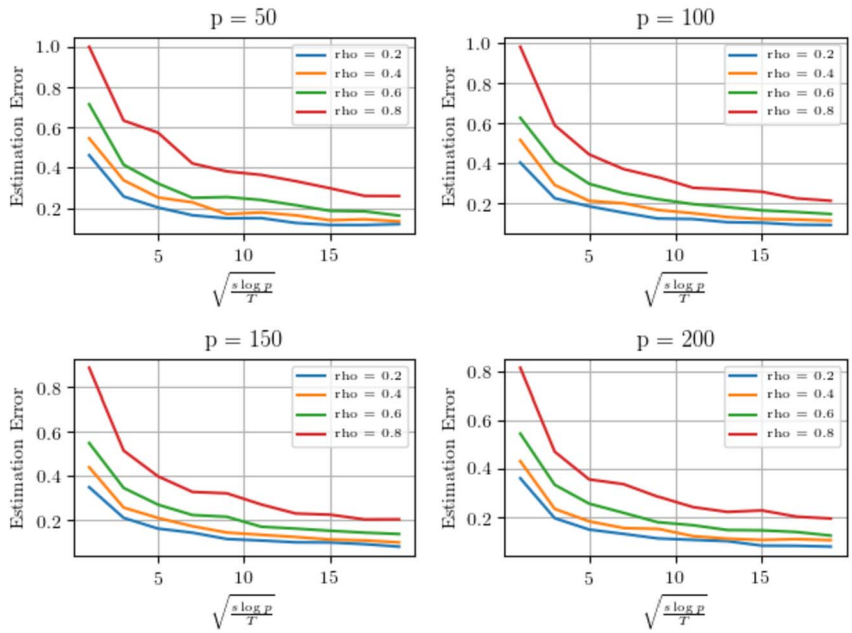

FIG. 2. *Effect of dependence on lasso estimation error. A larger ρ parameter corresponds to more dependence in the generating process.*

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.

6. Conclusion. One way to interpret the main results of this paper is that *the lasso procedure is robust to deviations from idealized conditions*, for example, independence and sub-Gaussian 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 sub-Weibull 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 nonexhaustive 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 considerably improved this paper.

SUPPLEMENTARY MATERIAL

Supplement to "Lasso guarantees for β-mixing heavy-tailed time series" (DOI: 10.1214/19-AOS1840SUPP; .pdf). Supplementary information.

REFERENCES

AGARWAL, A., NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Fast global convergence of gradient methods for high-dimensional statistical recovery. *Ann. Statist.* **40** 2452–2482. MR3097609 https://doi.org/10.1214/12-AOS1032

ALQUIER, P. and DOUKHAN, P. (2011). Sparsity considerations for dependent variables. *Electron. J. Stat.* **5** 750–774. MR2824815 https://doi.org/10.1214/11-EJS626

BASU, S. and MICHAILIDIS, G. (2015). Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.* **43** 1535–1567. MR3357870 https://doi.org/10.1214/15-AOS1315

BECK, A. and TEBOULLE, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2** 183–202. MR2486527 https://doi.org/10.1137/080716542

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469 https://doi.org/10.1214/08-AOS620

BRADLEY, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surv.* **2** 107–144. MR2178042 https://doi.org/10.1214/154957805100000104

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. *Springer Series in Statistics*. Springer, Heidelberg. MR2807761 https://doi.org/10.1007/978-3-642-20192-9

CHANDRASEKARAN, V., RECHT, B., PARRILO, P. A. and WILLSKY, A. S. (2012). The convex geometry of linear inverse problems. *Found. Comput. Math.* **12** 805–849. MR2989474 https://doi.org/10.1007/s10208-012-9135-7

CHEN, X., XU, M. and WU, W. B. (2013). Covariance and precision matrix estimation for high-dimensional time series. *Ann. Statist.* **41** 2994–3021. MR3161455 https://doi.org/10.1214/13-AOS1182

CHUDIK, A. and PESARAN, M. H. (2011). Infinite-dimensional VARs and factor models. *J. Econometrics* **163** 4–22. MR2803662 https://doi.org/10.1016/j.jeconom.2010.11.002

CHUDIK, A. and PESARAN, M. H. (2013). Econometric analysis of high dimensional VARs featuring a dominant unit. *Econometric Rev.* **32** 592–649. MR3041089 https://doi.org/10.1080/07474938.2012.740374

CHUDIK, A. and PESARAN, M. H. (2016). Theory and practice of GVAR modelling. *J. Econ. Surv.* **30** 165–197. https://doi.org/10.1111/joes.12095

DAVIS, R. A., ZANG, P. and ZHENG, T. (2016). Sparse vector autoregressive modeling. *J. Comput. Graph. Statist.* **25** 1077–1096. MR3572029 https://doi.org/10.1080/10618600.2015.1092978

DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2009). Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci. USA* **106** 18914–18919.

FAN, J., QI, L. and TONG, X. (2016). Penalized least squares estimation with weakly dependent data. *Sci. China Math.* **59** 2335–2354. MR3578960 https://doi.org/10.1007/s11425-016-0098-x

FOSS, S., KORSHUNOV, D. and ZACHARY, S. (2011). *An Introduction to Heavy-Tailed and Subexponential Distributions. Springer Series in Operations Research and Financial Engineering*. Springer, New York. MR2810144 https://doi.org/10.1007/978-1-4419-9473-8

GUO, S., WANG, Y. and YAO, Q. (2016). High-dimensional and banded vector autoregressions. *Biometrika* **103** 889–903. MR3620446 https://doi.org/10.1093/biomet/asw046

HAN, F. and LIU, H. (2013). Transition matrix estimation in high dimensional time series. In *Proceedings of the* 30*th 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. *J. Mach. Learn. Res.* **16** 3115–3150. MR3450535

HASTIE, T., TIBSHIRANI, R. and WAINWRIGHT, M. (2015). *Statistical Learning with Sparsity*: *The Lasso and Generalizations*. *Monographs on Statistics and Applied Probability* **143**. CRC Press, Boca Raton, FL. MR3616141

HAYASHI, F. (2000). *Econometrics*. Princeton Univ. Press, Princeton, NJ. MR1881537

IBRAGIMOV, I. D. A. and ROZANOV, Y. A. (1978). *Gaussian Random Processes*. *Applications of Mathematics* **9**. Springer, New York–Berlin. MR0543837

KOCK, A. B. and CALLOT, L. (2015). Oracle inequalities for high dimensional vector autoregressions. *J. Econometrics* **186** 325–344. MR3343790 https://doi.org/10.1016/j.jeconom.2015.02.013

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. and TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927. MR3485948 https://doi.org/10.1214/15-AOS1371

LIEBSCHER, E. (2005). Towards a unified approach for proving geometric ergodicity and mixing properties of nonlinear autoregressive processes. *J. Time Ser. Anal.* **26** 669–689. MR2188304 https://doi.org/10.1111/j.1467-9892.2005.00412.x

LOH, P.-L. and WAINWRIGHT, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Statist.* **40** 1637–1664. MR3015038 https://doi.org/10.1214/12-AOS1018

LÜTKEPOHL, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer, Berlin. MR2172368 https://doi.org/10.1007/978-3-540-27752-1

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. *Electron. J. Stat.* **9** 753–788. MR3331856 https://doi.org/10.1214/15-EJS1000

MEDEIROS, M. C. and MENDES, E. F. (2016). ℓ_1-regularization of high-dimensional time-series models with non-Gaussian and heteroskedastic errors. *J. Econometrics* **191** 255–271. MR3434446 https://doi.org/10.1016/j.jeconom.2015.10.011

MERLEVÈDE, F., PELIGRAD, M. and RIO, E. (2011). A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Probab. Theory Related Fields* **151** 435–474. MR2851689 https://doi.org/10.1007/s00440-010-0304-9

NARDI, Y. and RINALDO, A. (2011). Autoregressive process modeling via the Lasso procedure. *J. Multivariate Anal.* **102** 528–549. MR2755014 https://doi.org/10.1016/j.jmva.2010.10.012

NEGAHBAN, S. and WAINWRIGHT, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.* **39** 1069–1097. MR2816348 https://doi.org/10.1214/10-AOS850

NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *J. Mach. Learn. Res.* **13** 1665–1697. MR2930649

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. *Statist. Sci.* **27** 538–557. MR3025133 https://doi.org/10.1214/12-STS400

NGUEYEP, R. and SERBAN, N. (2015). Large-vector autoregression for multilayer spatially correlated time series. *Technometrics* **57** 207–216. MR3369677 https://doi.org/10.1080/00401706.2014.902775

NICHOLSON, W. B., BIEN, J. and MATTESON, D. S. (2014). Hierarchical vector autoregression. Preprint. Available at arXiv:1412.5250.

NICHOLSON, W. B., MATTESON, D. S. and BIEN, J. (2017). VARX-l: Structured regularization for large vector autoregressions with exogenous variables. *Int. J. Forecast.* **33** 627–651.

PISIER, G. (2016). Subgaussian sequences in probability and Fourier analysis. *Grad. J. Math.* **1** 60–80. MR3850765

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* 32*nd 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. *J. Mach. Learn. Res.* **11** 2241–2259. MR2719855

ROSENBLATT, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Natl. Acad. Sci. USA* **42** 43–47. MR0074711 https://doi.org/10.1073/pnas.42.1.43

RUDELSON, M. and VERSHYNIN, R. (2013). Hanson-Wright inequality and sub-Gaussian concentration. *Electron. Commun. Probab.* **18** no. 82, 9. MR3125258 https://doi.org/10.1214/ECP.v18-2865

RUDELSON, M. and ZHOU, S. (2013). Reconstruction from anisotropic random measurements. *IEEE Trans. Inform. Theory* **59** 3434–3447. MR3061256 https://doi.org/10.1109/TIT.2013.2243201

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. Preprint. Available at arXiv:1106.3915.

TAO, T. and VU, V. (2013). Random matrices: Sharp concentration of eigenvalues. *Random Matrices Theory Appl.* **2** 1350007, 31. MR3109424 https://doi.org/10.1142/S201032631350007X

TJØSTHEIM, D. (1990). Nonlinear time series and Markov chains. *Adv. in Appl. Probab.* **22** 587–611. MR1066965 https://doi.org/10.2307/1427459

UEMATSU, Y. (2015). Penalized likelihood estimation in high-dimensional time series models and uts application. Preprint. Available at arXiv:1504.06706.

VIDYASAGAR, M. (2003). *Learning and Generalization*: *With Applications to Neural Networks*, 2nd ed. *Communications and Control Engineering Series*. Springer, London. MR1938842 https://doi.org/10.1007/978-1-4471-3748-1

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. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 63–78. MR2301500 https://doi.org/10.1111/j.1467-9868.2007.00577.x

WONG, K. C., LI, Z. and TEWARI, A. (2016). Lasso guarantees for time series estimation under subgaussian tails and β-mixing. Preprint. Available at arXiv:1602.04265.

WONG, K. C., LI, Z. and TEWARI, A. (2020). Supplement to "Lasso guarantees for β-mixing heavy-tailed time series." https://doi.org/10.1214/19-AOS1840SUPP.

WU, W. B. and WU, Y. N. (2016). High-dimensional linear models with dependent observations. *Electron. J. Stat.* **10** 352–379.

YU, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences. *Ann. Probab.* **22** 94–116. MR1258867

ZHANG, D. and WU, W. B. (2017). Gaussian approximation for high dimensional time series. *Ann. Statist.* **45** 1895–1919. MR3718156 https://doi.org/10.1214/16-AOS1512

ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449