JOURNAL OF TIME SERIES ANALYSIS
J. Time Ser. Anal. **43**: 532–557 (2022)
Published online 26 October 2021 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1111/jtsa.12627

ORIGINAL ARTICLE

REGULARIZED ESTIMATION OF HIGH-DIMENSIONAL VECTOR AUTOREGRESSIONS WITH WEAKLY DEPENDENT INNOVATIONS

RICARDO P. MASINI,[a] MARCELO C. MEDEIROS[b] AND EDUARDO F. MENDES[c*]

[a]*Center for Statistics and Machine Learning, Princeton University, Princeton, NJ, USA*
[b]*Department of Economics, Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil*
[c]*School of Applied Mathematics (EMAp), Getulio Vargas Foundation, Rio de Janeiro (FGV), Rio de Janeiro, RJ, Brazil*

There has been considerable advance in understanding the properties of sparse regularization procedures in high-dimensional models. In time series context, it is mostly restricted to Gaussian autoregressions or mixing sequences. We study oracle properties of LASSO estimation of weakly sparse vector-autoregressive models with heavy tailed, weakly dependent innovations. In contrast to current literature, our innovation process satisfy an L^1 mixingale type condition on the centered conditional covariance matrices. This condition covers L^1-NED sequences and strong (α-) mixing sequences as particular examples.

Received 30 October 2020; Accepted 01 October 2021

Keywords: high-dimensional time series, LASSO, VAR, mixing

JEL. C32; C55; C58.

MOS subject classification: 62M10.

1. INTRODUCTION

Modeling multivariate time series data is an important and vibrant area of research. Applications range from economics and finance, as in Sims (1980), Bauer and Vornik (2011), Chiriac and Voev (2011), or Ramey (2016), to air pollution and ecological studies (Hoek *et al.*, 2013; Ensor *et al.*, 2013; Schweinberger *et al.*, 2017). Among alternatives, the vector autoregressive (VAR) model is certainly one of the most successful in modeling temporal evolution of vectors, networks, and matrices. See Lütkepohl (1991) or Wilson *et al.* (2015) for comprehensive textbook introductions.

The advances in data collection and storage have created data sets with large numbers of time series (*Big Data*), where the number of model parameters to be estimated may exceed the number of available data observations. A common approach to dealing with high-dimensional data is to impose additional structure in the form of (approximate) sparsity and estimate the parameters by some shrinkage method. Examples of estimation techniques range from Bayesian estimation with 'spike-and-slab' priors to sparsity-inducing shrinkage, such as the least absolute and shrinkage estimator (LASSO) and its many extensions. See Miranda-Agrippino and Ricco (2019) for a nice survey on Bayesian VARs or Kock *et al.* (2020) for a review on penalized regressions applied to time-series models.

1.1. Our Contributions

In this article, we study non-asymptotic properties of high-dimensional VAR models and their parameter estimates using equation-wise (row-wise or node-wise) LASSO. We show that, with high probability, estimated and

*Correspondence to: Eduardo F. Mendes, School of Applied Mathematics (EMAp), Getulio Vargas Foundation, Praia de Botafogo, 190, Rio de Janeiro (FGV), RJ, Brazil. E-mail: eduardo.mendes@fgv.br

population parameter vectors are close to each other in the Euclidean norm and discuss restrictions on the rate which the number of parameters can increase as the sample size diverges.

The importance of our results relies on the fact that our non-asymptotic guarantees serve as a fundamental ingredient for the derivation of asymptotic properties of penalized estimators in high-dimensional VAR models, as in Adamek *et al.* (2020). In particular, our results apply with minimal restrictions on the conditional variance model, allowing, for instance, large-dimensional multivariate linear processes in the variance. Moreover, auxiliary results proved in this article are of independent interested and can, for instance, be used to derive finite bounds for other type of penalization such as group/structured lasso, elastic-net, SCAD or non-convex penalties.

The data are assumed to be generated from a covariance-stationary and weakly sparse VAR model, where the innovations are martingale difference with sub-Weibull tails and conditional covariance matrix satisfying a L^1 mixingale assumption. An important feature is that the resulting process $\{\mathbf{y}_t\}$ is not necessarily mixing. Mixing assumptions can be notoriously difficult to show and we avoid it in this article. Nevertheless, it follows that our conditions cover strong mixing innovations as a particular case.

1.2. Literature Review

Some consistency results on model estimation and selection of high-dimensional VAR processes were obtained by Song and Bickel (2011), though under much stronger assumptions, such as Gaussianity. Loh and Wainwright (2012), Basu and Michailidis (2015) and Kock and Callot (2015) developed powerful concentration inequalities that enabled them to establish consistency under weaker conditions and prove that these conditions hold with high probability. In particular, Basu and Michailidis (2015) established consistency of ℓ_1-penalized least squares and maximum likelihood estimators of the coefficients of high-dimensional Gaussian VAR processes and related the estimation and prediction error to the complex dependence structure of VAR processes. Other estimation approaches, including Bayesian approaches, are discussed by Davis *et al.* (2016). Miao *et al.* (2020) proposed a factor-augmented large dimensional VAR and studied finite sample properties and provide estimation results. However, they assume independent and identically distributed errors. More recently, Wong *et al.* (2020) derived finite-sample guarantees for the LASSO in a misspecified VAR model. Authors assume the series is either β-mixing process with sub-Weibull marginal distributions or α-mixing Gaussian processes. Finally, Adamek *et al.* (2020) develop theoretical results for point estimation and inference in near epoch dependent time series using desparsified lasso under high level conditions on the generating process.

1.3. Organization of the Article

The article is organized as follows. In Section 2, we define the model and the main assumptions in the article. In Section 3, we discuss examples of applications of our results. The theoretical results are presented in Section 4, while in Section 5, we provide a discussion of our findings and conclude the article. All technical proofs are relegated to the Appendix.

1.4. Notation

Throughout the article we use the following notation. For a vector $\mathbf{b} = (b_1, \dots, b_k)' \in \mathbb{R}^k$ and $p \in [1, \infty]$, $|\mathbf{b}|_p$ denotes its ℓ_p norm, that is, $|\mathbf{b}|_p = \left(\sum_{i=1}^k |b_i|^p\right)^{1/p}$ for $p \in [1, \infty)$ and $|\mathbf{b}|_\infty = \max_{1\le i\le k} |b_i|$. We also define $|\mathbf{b}|_0 = \sum_{i=1}^k I(b_i \neq 0)$. For a random variable X, $\|X\|_p = (\mathbb{E}|X|^p)^{1/p}$ for $p \in [1, \infty)$ and $\|X\|_\infty = \inf\{a \in \mathbb{R} : \Pr(|X| \ge a) = 0\}$. For a $m \times n$ matrix $\boldsymbol{A}$ with elements a_{ij}, we denote $|||\boldsymbol{A}|||_1 = \max_{1\le j\le n} \sum_{i=1}^m |a_{ij}|$, $|||\boldsymbol{A}|||_\infty = \max_{1\le i\le m} \sum_{j=1}^n |a_{ij}|$, the induced ℓ_∞ and ℓ_1 norms respectively, and the maximum elementwise norm $|||\boldsymbol{A}|||_{\max} = \max_{i,j} |a_{ij}|$. Also $\Lambda_{\min}(\boldsymbol{A})$ and $\Lambda_{\max}(\boldsymbol{A})$ denotes the minimum and maximum eigenvalues of the square matrix $\boldsymbol{A}$ respectively.

DOI: 10.1111/jtsa.12627

2. MODEL SETUP AND ASSUMPTIONS

Let $\left\{\mathbf{y}_t = \left(y_{t,1}, \ldots, y_{t,n}\right)' : t \in \mathbb{Z}\right\}$ be a vector stochastic process defined in some fixed probability space taking values on $\mathbb{R}^n$ given by

$$\mathbf{y}_t = \boldsymbol{A}_1 \mathbf{y}_{t-1} + \cdots + \boldsymbol{A}_p \mathbf{y}_{t-p} + \mathbf{u}_t, \tag{1}$$

where $\mathbf{u}_t = \left(u_{t,1}, \ldots, u_{t,n}\right)'$ is a zero-mean vector of innovations and $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_p$, are $n \times n$ parameter matrices. The dimension $n := n_T$ and order $p := p_T$ of the process are allowed to increase with the number of observations T. Write the vector-autoregressive (VAR) process (1) using its first-order representation:

$$\tilde{\mathbf{y}}_t = \boldsymbol{F}_T \tilde{\mathbf{y}}_{t-1} + \tilde{\mathbf{u}}_t, \tag{2}$$

where $\tilde{\mathbf{y}}_t = \left(\mathbf{y}_t', \ldots, \mathbf{y}_{t-p+1}'\right)'$, $\tilde{\mathbf{u}}_t = \left(\mathbf{u}_t', \mathbf{0}', \ldots, \mathbf{0}'\right)'$, and

$$\boldsymbol{F}_T = \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{A}_2 & \cdots & \boldsymbol{A}_{p-1} & \boldsymbol{A}_p \\ \boldsymbol{I}_n & \mathbf{0}_n & \cdots & \mathbf{0}_n & \mathbf{0}_n \\ \mathbf{0}_n & \boldsymbol{I}_n & & \mathbf{0}_n & \mathbf{0}_n \\ \vdots & & \ddots & \vdots & \vdots \\ \mathbf{0}_n & \mathbf{0}_n & & \boldsymbol{I}_n & \mathbf{0} \end{bmatrix}.$$

In high-dimensional statistics the dimension of the model is relatively larger the number of observations T. In practical terms, it is mathematically modeled as the rate in which the dimension may increase as a function of the number of observations available, hence the dependence of n and p on T. In the following assumptions we need the data generating process to satisfy bound conditions that must hold for *any sample size* T. Although particular constants can vary with T, they have to be uniformly bounded. To simplify exposition, we simply state that constants involved do not depend on T. As our results hold for finite trajectories $X_1, \ldots, X_T$ we formulate Assumptions (A1)–(A3) in a way that they must hold '*…uniformly in* $1 \leq t \leq T$ *and* $T \in \mathbb{N}$'. Assumptions (A4) and (A5) involve quantities that are likely to change with the sample size and may affect regularization rate. These sequences appear explicitly in the probability bounds and rates.

Assumption (A1)**.** All roots of the reverse characteristic polynomial $\boldsymbol{\mathcal{A}}(z) = \boldsymbol{I}_n - \sum_{i=1}^{p} \boldsymbol{A}_j z^j$ lie outside the unit disk for each $p, n \in \mathbb{N}$, and there exist $\bar{c}_\Phi > 0$, $c_\phi > 0$ and $0 < \gamma_1 \leq 1$ such that for all $m \in \mathbb{N}$

$$\sum_{k=m}^{\infty} \left|\boldsymbol{\phi}_{k,i}\right|_1 \leq \bar{c}_\Phi e^{-c_\phi m^{\gamma_1}}, \tag{3}$$

uniformly in $1 \leq i \leq n$, where $\boldsymbol{\Phi}_k := \boldsymbol{J}' \boldsymbol{F}_T^k \boldsymbol{J} = \left(\boldsymbol{\phi}_{k,1}, \ldots, \boldsymbol{\phi}_{k,n}\right)'$, $\boldsymbol{F}_T$ denote the companion matrix and $\boldsymbol{J} = \left(\boldsymbol{I}_n, \mathbf{0}_n, \ldots, \mathbf{0}_n\right)'$.

Assumption (A2)**.** The sequence $\left\{\left(\mathbf{u}_t, \mathcal{F}_t\right)\right\}_t$ is a covariance stationary (for each $T \in \mathbb{N}$) martingale difference process where the filtration $\{\mathcal{F}_t\}_t$ includes the natural filtration of $\{\mathbf{u}_t\}$. The smallest and largest eigenvalues of $\boldsymbol{\Sigma} := \mathbb{E}\left(\mathbf{u}_1 \mathbf{u}_1'\right)$ are bounded away from 0 and ∞ respectively, uniformly in $T \in \mathbb{N}$. Furthermore, for all $\mathbf{b}_1, \mathbf{b}_2 \in \left\{\boldsymbol{v} \in \mathbb{R}^n : |\boldsymbol{v}|_1 \leq 1\right\}$ and $m \in \mathbb{N}$,

$$\mathbb{E}\left|\mathbb{E}\left[\mathbf{b}_1' \left(\mathbf{u}_t \mathbf{u}_t' - \boldsymbol{\Sigma}\right) \mathbf{b}_2 | \mathcal{F}_{t-m}\right]\right| \leq a_1 e^{-a_2 m^{\gamma_2}},$$

for some $a_1, a_2 > 0$ and $0 < \gamma_2 \leq 1$, uniformly in $1 \leq t \leq T$ and $T \in \mathbb{N}$.

DOI: 10.1111/jtsa.12627

Assumption (A3)**.** For all $\mathbf{b} \in \left\{ v \in \mathbb{R}^n : |v|_1 \leq 1 \right\}$ and all $0 < x < \infty$, $\Pr\left(|\mathbf{b}'\mathbf{u}_t| > x\right) \leq 2e^{-|x/c_\alpha|^\alpha}$ for some $\alpha > 0$, $0 < c_\alpha < \infty$, uniformly in $1 \leq t \leq T$ and $T \in \mathbb{N}$.

Assumption (A1) requires that the VAR process is stable and admits an infinite-order vector moving average, VMA(∞), representation for all n and p as

$$\mathbf{y}_t = \sum_{i=0}^{\infty} \boldsymbol{J}' \boldsymbol{F}_T^i \boldsymbol{J} \mathbf{u}_{t-i} = \sum_{i=0}^{\infty} \boldsymbol{\Phi}_i \mathbf{u}_{t-i}. \tag{4}$$

Furthermore, the coefficients of the MA(∞) representations of each $\{y_{i,t}\}$, $i = 1, \ldots, n$, are absolutely summable with exponentially decaying rate. This condition is satisfied in standard VAR(p) models, where n and p are fixed. In models that n is large, Lemma 4 in Appendix B.1 shows that condition (3) is satisfied if $\sum_{k=1}^{p} \left\| \left| \boldsymbol{A}_k \right| \right\|_\infty < 1$ and further regularity conditions on the size of the coefficients. Finally, notice that under (A1) it is also true that $\max_{k,i} |\boldsymbol{\phi}_{k,i}|_\infty \leq \bar{c}_\Phi$, which means that the coefficients $\{\boldsymbol{\Phi}_k\}$ are uniformly upper bounded under the maximum entry-wise norm.

Assumption (A2) requires the error process to be a martingale difference process and satisfy a very weak dependence condition on its conditional variance. The former restricts the model to be correctly specified in the mean. Nevertheless, this assumption is standard in the literature and we are able to derive results covering a broad range of data generating processes and conditional dependence measures. The latter is the L^1 projective dependence measure appearing in Dedecker *et al.* (2007, section 2.2.4). Note that (1) strong mixing (or α-mixing) sequences with exponential decay of the mixing coefficient satisfy this condition (Davidson, 1994, Theorem 14.2); and (2) uniform mixing sequences (ϕ-mixing) and β-mixing sequences are also strong mixing, but the converse is not true (Bradley, 2005, Eqs. (1.11)–(1.18)). If we denote the centered outer product series $\boldsymbol{v}_t = \text{vech}\left(\mathbf{u}_t\mathbf{u}_t' - \boldsymbol{\Sigma}\right)$, this assumptions requires that $\{\boldsymbol{v}_t\}$ is L^1 mixingale. It means that stochastic process with L^r bounded, L^1 near-epoch dependent, centered outer product series $\boldsymbol{v}_t$ are also contemplated in this setting Andrews (1988). Finally, Assumptions (A1) and (A2) combined ensure that $\{\mathbf{y}_t\}$ is second-order stationary for each n and p (Lütkepohl, 2006, Ch. 2).

Condition (A3) imposes restrictions on the tail behavior of the innovation process $\{\mathbf{u}_t\}$ that are shared by $\{\mathbf{y}_t\}$. More precisely, we impose moment conditions on all linear combinations $\mathbf{b}'\mathbf{u}_t$. Lemma 3, in the appendix, shows that each $\{y_{i,t}\}$ ($i = 1, \ldots, n$) also share the same tail properties of $\{\mathbf{u}_t\}$. This condition is essential for defining the rate in which n and p increase with T. We focus on the case the tail decays at rate $O\left(e^{-cx^\alpha}\right)$ for some $\alpha > 0$, that is, $\{\mathbf{b}'\mathbf{u}_t\}$ is sub-Weibull with parameter α studied in Wong *et al.* (2020, Section 4.1). Note that when $\alpha \geq 1$ and $\alpha \geq 2$ we have the sub-exponential and sub-Gaussian tails respectively. However, when $\alpha \in (0, 1)$ the moment generating function does not exist at any point and and these variables are usually called *heavy tailed*.

It is convenient to write the model in stacked form. Let $\mathbf{x}_t = \left(\mathbf{y}_{t-1}', \ldots, \mathbf{y}_{t-p}'\right)'$ be the $np \times 1$ vector of regressors and $\mathbf{X} = \left(\mathbf{x}_1, \ldots, \mathbf{x}_T\right)'$ the $T \times np$ matrix of covariates. Let $\mathbf{Y}_i = \left(y_{i,1}, \ldots, y_{i,T}\right)'$ be the $T \times 1$ vector of observations for the ith element of $\mathbf{y}_t$, and $\mathbf{U}_i = \left(u_{i,1}, \ldots, u_{i,T}\right)'$ the corresponding vector of innovations. Denote $\boldsymbol{\beta}_i$ the $np \times 1$ vector of coefficients corresponding to equation i. Then, model (1) is equivalent to

$$\mathbf{Y}_i = \mathbf{X}\boldsymbol{\beta}_i + \mathbf{U}_i, \quad i = 1, \ldots, n. \tag{5}$$

We now make additional assumptions concerning model (5).

Assumption (A4)**.** The true parameter vectors $\boldsymbol{\beta}_i$, $i = 1, \ldots, n$, satisfy $\sum_{j=1}^{np} |\beta_{i,j}|^q \leq R_q$ for some $0 \leq q < 1$ and $0 < R_q < \infty$ where $R_q := R_{q,T}$ is allowed to depend on the sample size T.

Assumption (A5)**.** For each $T \in \mathbb{N}$, the smallest eigenvalue of $\boldsymbol{\Gamma} := T^{-1}\mathbb{E}(\mathbf{X}'\mathbf{X})$ is greater than a positive constant σ_Γ^2 that might depend on T.

DOI: 10.1111/jtsa.12627

Assumption (A4) imposes *weak sparsity* of the coefficients, in a sense that most of them are small. This condition is slightly stronger than we need in a sense that we may have distinct q_i and $R_{q,i}$ for each equation. In the case $q = 0$ we have sparsity in the standard sense, meaning that $R_0 = s$, the number of non-zero coefficients. In practice, we estimate a sparse model that truncates all coefficients close to zero. This assumption is standard for *weak sparsity*, see Negahban *et al.* (2012, section 4.3) and Han and Tsay (2020, Assumption 1) for an application in time series setting.

Assumption (A5) is often used in the sparse estimation literature (e.g. Kock and Callot, 2015; Medeiros and Mendes, 2016; Han and Tsay, 2020). Basu and Michailidis (2015, Proposition 2.3) derived bounds for $\Lambda_{\min}(\mathbf{\Gamma})$ and $\Lambda_{\max}(\mathbf{\Gamma})$ using properties of the block Toeplitz matrix $\mathbf{\Gamma}$ and its generating function, the cross-spectral density of the generating VAR(p) process:

$$\frac{\Lambda_{\min}(\mathbf{\Sigma})}{\max_{|z|=1}\Lambda_{\max}\left(\mathcal{A}^*(z)\mathcal{A}(z)\right)} \leq \Lambda_{\min}(\mathbf{\Gamma}) \leq \Lambda_{\max}(\mathbf{\Gamma}) \leq \frac{\Lambda_{\max}(\mathbf{\Sigma})}{\max_{|z|=1}\Lambda_{\min}\left(\mathcal{A}^*(z)\mathcal{A}(z)\right)}, \tag{6}$$

where $\mathcal{A}^*$ is the conjugate transpose of $\mathcal{A}$, the reverse characteristic polynomial, defined in Assumption (A1). Basu and Michailidis (2015)[Proposition 2.2] shows that under (A1),

$$\max_{|z|=1}\Lambda_{\max}\left(\mathcal{A}^*(z)\mathcal{A}(z)\right) < \left[1 + \frac{\sum_{k=1}^{p}\left(|||\boldsymbol{A}_k|||_1 + |||\boldsymbol{A}_k|||_\infty\right)}{2}\right]^2.$$

Hence, (A5) is satisfied if, for instance, $\Lambda_{\min}(\mathbf{\Sigma}) > 0$, $\sum_{k=1}^{p}|||\boldsymbol{A}_k|||_1 < \infty$ and $\sum_{k=1}^{p}|||\boldsymbol{A}_k|||_\infty < \infty$.

3. EXAMPLES

We illustrate processes satisfying Assumptions (A2) and (A3). In the first two examples we discuss sufficient conditions involving mixing and near epoch dependent sequences, traditionally found in the literature. In the final two examples, we discuss variance process admitting an AR(∞) representation. In all examples, the innovation vector $\mathbf{u}_t$ has dimension n, which is a function of T, but bounds are independent on the dimension of the problem or sample size. Hence, following examples hold *for each n*, as in the assumptions.

Example 1 (Strong mixing sequences)**.** Let $\{\mathbf{u}_t\}$ denote a martingale difference, strong mixing sequence with coefficients $\alpha_m < b_1\exp\left(-b_2 m^{\gamma_2}\right)$ and common covariance matrix $\mathbf{\Sigma}$ with eigenvalues bounded away from zero and infinity, uniformly in n. It follows that $r_t = \mathbf{b}_1'\mathbf{u}_t\mathbf{u}_t'\mathbf{b}_2$ is also strong mixing of same size and, from Davidson (1994, Theorem 14.2), $\mathbb{E}\left[r_t - \mathbb{E}(r_t)|\mathcal{F}_{t-m}\right] \leq a_1\exp\left(-a_2 m^{\gamma_2}\right)$, for constants a_1 and a_2.

Example 2 (L^1 near-epoch dependent process)**.** Let $\{\mathbf{u}_t\}$ denote a weakly stationary, martingale difference sequence. Suppose $\mathbf{b}'\boldsymbol{v}_t = \mathbf{b}'\text{vech}\left(\mathbf{u}_t\mathbf{u}_t' - \mathbf{\Sigma}\right)$ is a centered, L^1-NED sequence on $\mathcal{F}_t = \sigma\langle\epsilon_t, \epsilon_{t-1}, \ldots\rangle$, where $\{\epsilon_t\}$ is α-mixing with coefficients $\alpha_m \leq c_1\exp\left(c_2 m^{\gamma_1}\right)$, for all $\mathbf{b} \in \left\{\mathbf{b} \in \mathbb{R}^{n(n+1)/2} : |\mathbf{b}|_1 \leq 1\right\}$. It means that there are finite constants $\{d_t\}$ and $\{\psi_m\}$ such that

$$\mathbb{E}\left|\mathbf{b}'\left(\boldsymbol{v}_t - \mathbb{E}\left[\boldsymbol{v}_t|\mathcal{F}_{t-m:t}\right]\right)\right| \leq d_t\psi_m,$$

where $\mathcal{F}_{t-m:t} = \sigma\langle\epsilon_t, \ldots, \epsilon_{t-m}\rangle$ and $\psi_m \leq \exp\left(c_3 m^{\gamma_2}\right)$. Under Assumption (A3), it follows from Wong *et al.* (2020, Lemma 5) and Hölder inequality that for any $r < \infty$

$$\|\mathbf{b}'\boldsymbol{v}_t\|_r \leq |\mathbf{b}|_1^r \max_{1\leq i\leq j\leq n}\|u_{it}u_{jt}\|_r \leq \max_{1\leq i\leq n}\|u_{it}\|_{2r} \leq c_4 r^{1/\alpha}.$$

Finally, it follows from Andrews (1988, Example 6) that Assumption (A2) holds with $a_1 \geq \left(2\max_t d_t + c_4 r^{1/\alpha}\right)\left(e^{c_3/2^{\gamma_2}} + 6c_1 e^{c_2(r-1)/r2^{\gamma_2}}\right)$ and $a_2 \leq \left(c_3 \wedge c_2(r-1)/r\right)/2^{\gamma_2}$.

DOI: 10.1111/jtsa.12627

Example 3 (Linear process in the variance)**.** Let $\{\boldsymbol{v}_t, \boldsymbol{\epsilon}_t\}$ denote a sequence of centered independently and identically distributed, sub-Weibull random variables with parameter (at least) 2α, taking values in $\mathbb{R}^{2n}$ with identity covariance matrix. Let $\mathbf{u}_t = \boldsymbol{H}_t^{1/2}\boldsymbol{v}_t$ where $\boldsymbol{H}_t^{1/2}$ is the lower diagonal Cholesky decomposition of $\boldsymbol{H}_t$ and

$$\boldsymbol{h}_t = \text{vech}\left(\boldsymbol{H}_t\right) = \boldsymbol{c} + \sum_{j=1}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{\eta}_{t-j}.$$

Here, $\text{vech}(\boldsymbol{M})$ stacks the lower diagonal elements of matrix $\boldsymbol{M}$, $\boldsymbol{c}$ is a vector of constants and $\boldsymbol{\eta}_t = \text{vech}\left(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t'\right)$. For all $\tilde{\mathbf{b}} \in \left\{\mathbf{b} \in \mathbb{R}^{n(n+1)/2} : |\mathbf{b}|_1 \leq 1\right\}$, $\{\boldsymbol{\Psi}_j\}$ satisfy $\sum_{j=m}^{\infty} |\tilde{\mathbf{b}}'\boldsymbol{\Psi}_j|_1 \lesssim e^{-a_2 m^{\gamma_2}}$.

We first show $\{\mathbf{u}_t\}$ is weakly stationary martingale difference with respect to $\mathcal{F}_{t-1} = \sigma\left\langle \left(\boldsymbol{v}_{t-j}, \boldsymbol{\epsilon}_{t-j}\right) : j = 1, 2, \ldots \right\rangle$. The process $\{\mathbf{u}_t\}$ is $\mathcal{F}_t$ measurable and satisfy $\mathbb{E}\left[\mathbf{u}_t|\mathcal{F}_{t-1}\right] = \boldsymbol{H}_t^{1/2}\mathbb{E}\left[\boldsymbol{v}_t|\mathcal{F}_{t-1}\right] = \mathbf{0}$. Its covariance matrix is

$$\mathbb{E}\left[\mathbf{u}_t\mathbf{u}_t'\right] = \mathbb{E}\left[\boldsymbol{H}_t^{1/2}\mathbb{E}\left(\boldsymbol{v}_t\boldsymbol{v}_t'|\mathcal{F}_{t-1}\right)\left(\boldsymbol{H}_t^{1/2}\right)'\right] = \mathbb{E}\left[\boldsymbol{H}_t\right].$$

Now, $\mathbb{E}[\boldsymbol{h}_t] = \boldsymbol{c} + \sum_{j=1}^{\infty} \boldsymbol{\Psi}_j \mathbb{E}\boldsymbol{\eta}_{t-j} = \boldsymbol{c} + \sum_{j=1}^{\infty} \boldsymbol{\Psi}_j \text{vech}(\boldsymbol{I}_n) = \boldsymbol{\Sigma}$, where $\mathbb{E}(\boldsymbol{\eta}_t) = \text{vech}\left(\mathbb{E}\left(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t'\right)\right) = \text{vech}(\boldsymbol{I}_n)$ for all t.

For constant vectors $\mathbf{b}_1, \mathbf{b}_2 \in \{\mathbf{b} \in \mathbb{R}_n : |\mathbf{b}| \leq 1\}$,

$$\begin{aligned}
\mathbb{E}\left[\mathbf{b}_1'\left(\mathbf{u}_t\mathbf{u}_t' - \boldsymbol{\Sigma}\right)\mathbf{b}_2|\mathcal{F}_{t-1}\right] &= \mathbf{b}_1'\left(\boldsymbol{H}_t - \mathbb{E}\boldsymbol{H}_t\right)\mathbf{b}_2 \\
&= \tilde{\mathbf{b}}'\left(\boldsymbol{h}_t - \mathbb{E}\boldsymbol{h}_t\right) \\
&= \sum_{j=1}^{\infty} \tilde{\mathbf{b}}'\boldsymbol{\Psi}_j\left(\boldsymbol{\eta}_{t-j} - \mathbb{E}\boldsymbol{\eta}_{t-j}\right),
\end{aligned}$$

where $\tilde{\mathbf{b}} \in \left\{\mathbf{b} \in \mathbb{R}^{n(n+1)/2} : |\mathbf{b}|_1 \leq 1\right\}$. It follows that

$$\begin{aligned}
\mathbb{E}\left|\mathbb{E}\left[\mathbf{b}_1'\left(\mathbf{u}_t\mathbf{u}_t' - \boldsymbol{\Sigma}\right)\mathbf{b}_2|\mathcal{F}_{t-m}\right]\right| &= \mathbb{E}\left|\sum_{j=1}^{\infty} \tilde{\mathbf{b}}'\boldsymbol{\Psi}_j\mathbb{E}\left(\boldsymbol{\eta}_{t-j} - \mathbb{E}\boldsymbol{\eta}_{t-j}|\mathcal{F}_{t-m}\right)\right| \\
&= \left\|\sum_{j=m}^{\infty} \tilde{\mathbf{b}}'\boldsymbol{\Psi}_j\left(\boldsymbol{\eta}_{t-j} - \mathbb{E}\boldsymbol{\eta}_{t-j}\right)\right\|_1 \\
&\leq 2\left(\sum_{j=m}^{\infty} |\tilde{\mathbf{b}}'\boldsymbol{\Psi}_j|_1\right)\max_{|\mathbf{b}|_1 \leq 1} \|\mathbf{b}'\boldsymbol{\epsilon}_t\|_2^2,
\end{aligned}$$

where in the last line we use the same arguments of Lemma 3 in the appendix, followed by the triangle inequality. Then, Assumption (A2) is satisfied under the condition that $\sum_{j=m}^{\infty} |\tilde{\mathbf{b}}'\boldsymbol{\Psi}_j|_1 \lesssim e^{-a_2 m^{\gamma_2}}$ and $\|\mathbf{b}'\boldsymbol{\epsilon}_t\|_2 \leq c_2 < \infty$.

It follows from Wong *et al.* (2020, Lemma 5) that $\{\boldsymbol{u}_t\}$ is sub-Weibull with parameter α if $\sup_{d\geq_1} d^{-1/\alpha}\|\mathbf{b}'\mathbf{u}_t\|_d \leq c_\alpha < \infty$. For any $d \geq 1$,

$$\begin{aligned}
d^{-1/\alpha}\|\mathbf{b}'\mathbf{u}_t\|_d &= d^{-1/\alpha}\|\mathbf{b}'\mathbf{u}_t\mathbf{u}_t'\mathbf{b}\|_{d/2}^{1/2} \\
&= d^{-1/\alpha}\left\|\mathbf{b}'\boldsymbol{H}_t^{1/2}\boldsymbol{v}_t\boldsymbol{v}_t'\left(\boldsymbol{H}_t^{1/2}\right)\mathbf{b}\right\|_{d/2}^{1/2} \\
&= d^{-1/\alpha}\left\|\mathbf{b}'\boldsymbol{H}_t\mathbf{b} \times \frac{\mathbf{b}'\boldsymbol{H}_t^{1/2}\boldsymbol{v}_t\boldsymbol{v}_t'\left(\boldsymbol{H}_t^{1/2}\right)\mathbf{b}}{\mathbf{b}'\boldsymbol{H}_t\mathbf{b}}\right\|_{d/2}^{1/2}
\end{aligned}$$

DOI: 10.1111/jtsa.12627

$$
\begin{aligned}
&\leq d^{-1/\alpha}\left\{\mathbb{E}\left(\left|\mathbf{b}'\boldsymbol{H}_t\mathbf{b}\right|^{d/2}\mathbb{E}\left[\left|\frac{\mathbf{b}'\boldsymbol{H}_t^{1/2}\boldsymbol{v}_t\boldsymbol{v}_t'\left(\boldsymbol{H}_t^{1/2}\right)\mathbf{b}}{\mathbf{b}'\boldsymbol{H}_t\mathbf{b}}\right|^{d/2}\Big|\mathcal{F}_{t-1}\right]\right)\right\}^{1/d}\\
&\leq d^{-1/\alpha}\left\{\mathbb{E}\left(\left|\mathbf{b}'\boldsymbol{H}_t\mathbf{b}\right|^{d/2}\sup_{\boldsymbol{\delta}'\boldsymbol{\delta}=1}\mathbb{E}\left[\left(\boldsymbol{\delta}'\boldsymbol{v}_t\boldsymbol{v}_t'\boldsymbol{\delta}\right)^{d/2}|\mathcal{F}_{t-1}\right]\right)\right\}^{1/d}\\
&= d^{-1/2\alpha}\left\|\mathbf{b}'\boldsymbol{H}_t\mathbf{b}\right\|_{d/2}^{1/2}\sup_{\boldsymbol{\delta}'\boldsymbol{\delta}=1}d^{-1/\alpha}\|\boldsymbol{\delta}'\boldsymbol{v}_t\|_d,\\
&\leq\left(d^{-1/\alpha}\left\|\mathbf{b}'\boldsymbol{H}_t\mathbf{b}\right\|_{d/2}\right)^{1/2}c_{2\alpha}
\end{aligned}
$$

where the two last lines follow because $\{\boldsymbol{v}_t\}$ is independent and sub-Weibull process with parameter 2α. There is a $\tilde{\mathbf{b}}\in\left\{\mathbf{b}\in\mathbb{R}^{n(n-1)/2}:|\mathbf{b}|_1\leq 1\right\}$ such that

$$
\begin{aligned}
d^{-1/\alpha}\left\|\mathbf{b}'\boldsymbol{H}_t\mathbf{b}\right\|_{d/2} &= d^{-1/\alpha}\|\tilde{\mathbf{b}}'\boldsymbol{h}_t\|_{d/2}\\
&= d^{-1/\alpha}\left\|\tilde{\mathbf{b}}'\boldsymbol{c}+\sum_{j=1}^{\infty}\tilde{\mathbf{b}}'\boldsymbol{\Psi}_j\boldsymbol{\eta}_{t-j}\right\|_{d/2}\\
&\leq d^{-1/\alpha}\left|\tilde{\mathbf{b}}'\boldsymbol{c}\right|_{d/2}+d^{-1/\alpha}\left\|\sum_{j=1}^{\infty}\tilde{\mathbf{b}}'\boldsymbol{\Psi}_j\boldsymbol{\eta}_{t-j}\right\|_{d/2}\\
&\leq d^{-1/\alpha}\left|\tilde{\mathbf{b}}'\boldsymbol{c}\right\|t_{d/2}+2\left(\sum_{j=1}^{\infty}|\tilde{\mathbf{b}}'\boldsymbol{\Psi}_j|_1\right)\max_{|\mathbf{b}|_1\leq 1}\left(d^{-1/2\alpha}\|\mathbf{b}'\boldsymbol{\epsilon}_t\|_d\right)^2\\
&\leq\left|\tilde{\mathbf{b}}'\boldsymbol{c}\right|_{d/2}+c_{2\alpha}^2.
\end{aligned}
$$

Combining these bounds, process $\{\mathbf{u}_t\}$ is sub-Weibull with parameter α, satisfying Condition (A3).

Example 4 (Stochastic covariance)**.** Let

$$
\mathbf{y}_t=\sum_{i=1}^{p}\boldsymbol{A}_i\mathbf{y}_{t-i}+\boldsymbol{H}_t^{1/2}\boldsymbol{v}_t,\quad \boldsymbol{H}_{t+1}=\boldsymbol{C}_0+\boldsymbol{\Psi}\boldsymbol{H}_t\boldsymbol{\Psi}'+\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t',
$$

where $\boldsymbol{v}_t\overset{\text{i.i.d.}}{\sim}\mathsf{N}(\mathbf{0},\boldsymbol{I}_n)$, $\boldsymbol{\epsilon}_t\overset{\text{i.i.d.}}{\sim}(\mathbf{0},\boldsymbol{I}_n)$ is sub-Weibull with parameter 2α, and processes $\{\boldsymbol{v}_t\}$ and $\{\boldsymbol{\epsilon}_t\}$ are independent. Let $\mathcal{F}_t=\sigma\left\langle\left(\boldsymbol{v}_{t-j},\boldsymbol{\epsilon}_{t-j}\right):j=0,1,2,\ldots\right\rangle$.

Process $\{\boldsymbol{H}_t\}$ is a matrix process characterizing the stochastic covariance of $\boldsymbol{u}_t=\boldsymbol{H}_t^{1/2}\boldsymbol{v}_t$ evolves according to a matrix autoregressive process. The *intercept* $\boldsymbol{C}_0$ is symmetric and positive definite matrix and the eigenvalues of $\boldsymbol{\Psi}$ are inside the unity circle. We also assume $\|\|\boldsymbol{\Psi}\|\|_1\|\|\boldsymbol{\Psi}\|\|_\infty<1$, which implies that the largest singular value of $\boldsymbol{\Psi}$ is smaller than one.

Under these conditions, the process $\{\boldsymbol{H}_t\}$ is ensured to be positive definite and stationary. To see the latter, vectorize the process to obtain $\boldsymbol{h}_t=\mathsf{vec}\left(\boldsymbol{H}_t\right)$, $\boldsymbol{c}_0=\mathsf{vec}\left(\boldsymbol{C}_0\right)$, $\boldsymbol{\eta}_t=\mathsf{vec}\left(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t'\right)$, $\bar{\boldsymbol{\Psi}}=\boldsymbol{\Psi}\otimes\boldsymbol{\Psi}'$ and

$$
\left(\boldsymbol{I}-\bar{\boldsymbol{\Psi}}L\right)\boldsymbol{h}_{t+1}=\boldsymbol{c}_0+\boldsymbol{\eta}_t\Leftrightarrow\boldsymbol{h}_{t+1}=\sum_{j=0}^{\infty}\bar{\boldsymbol{\Psi}}^j\boldsymbol{c}_0+\sum_{j=0}^{\infty}\bar{\boldsymbol{\Psi}}^j\boldsymbol{\eta}_{t-j}.
$$

DOI: 10.1111/jtsa.12627

The process is stationary because eigenvalues of $\bar{\mathbf{\Psi}}$ are products of eigenvalues of $\mathbf{\Psi}$, which is also inside the unity circle. We are exactly in the setting of previous example.

We have to show $\sum_{j=m+1}^{\infty} |\mathbf{b}'\bar{\mathbf{\Psi}}^j|_1 \lesssim e^{-a_2 m^{\gamma_2}}$ for all $\mathbf{b} \in R^{n^2} : |\mathbf{b}|_1 = 1$. The term on the left-hand side is bounded by $\sum_{j>m} \left|\left|\left|\bar{\mathbf{\Psi}}^j\right|\right|\right|_1$ and each $\left|\left|\left|\bar{\mathbf{\Psi}}^j\right|\right|\right|_1 \leq \left|\left|\left|\mathbf{\Psi}^j\right|\right|\right|_1 \left|\left|\left|\mathbf{\Psi}^j\right|\right|\right|_\infty \leq \left(|||\mathbf{\Psi}|||_1 |||\mathbf{\Psi}|||_\infty\right)^j$. Under the assumption that $|||\mathbf{\Psi}|||_1 |||\mathbf{\Psi}|||_\infty = c_{\max} < 1$, $\sum_{j=m+1}^{\infty} |\mathbf{b}'\bar{\mathbf{\Psi}}^j|_1 \leq \frac{c_{\max}}{1-c_{\max}} e^{-m\log\left(1/c_{\max}\right)}$ meaning that condition is satisfied with $\gamma_2 = 1$ and $a_2 = \log\left(1/c_{\max}\right)$.

Finally, the unconditional covariance of $\mathbf{u}_t$ is

$$\mathbf{\Sigma} := \mathbb{E}[\boldsymbol{H}_t] = \sum_{j=0}^{\infty} \mathbf{\Psi}^j \left(\boldsymbol{C}_0 + \boldsymbol{I}_n\right) \mathbf{\Psi}'^j.$$

The smallest eigenvalue $\min_{\delta'\delta=1} \boldsymbol{\delta}'\mathbf{\Sigma}\boldsymbol{\delta} \geq \Lambda_{\min}\left(\boldsymbol{C}_0\right) + 1$ and largest eigenvalue of $\max_{\delta'\delta=1} \boldsymbol{\delta}'\mathbf{\Sigma}\boldsymbol{\delta} \leq \left(\rho\left(\boldsymbol{C}_0\right) + 1\right)\left(1 - \rho(\mathbf{\Psi})^2\right)^{-1}$, where $\Lambda_{\min}(\boldsymbol{A})$ is the smallest eigenvalue of $\boldsymbol{A}$ and $\rho(\boldsymbol{A})$ is the spectral radius of $\boldsymbol{A}$.

4. LASSO ESTIMATION BOUNDS

Let $\mathcal{L}_T\left(\boldsymbol{\beta}_i\right) = \frac{1}{T}|\mathbf{Y}_i - \mathbf{X}\boldsymbol{\beta}_i|_2^2$ denote the empirical squared risk, for each $i = 1, \ldots, n$. We estimate $\boldsymbol{\beta}_i$, $i = 1, \ldots, n$, equation-wise using the LASSO procedure

$$\widehat{\boldsymbol{\beta}}_i \in \arg\min_{\boldsymbol{\beta}_i \in \mathbb{R}^{np}} \left\{\mathcal{L}_T\left(\boldsymbol{\beta}_i\right) + \lambda_i |\boldsymbol{\beta}_i|_1\right\}, \quad i = 1, \ldots, n, \tag{7}$$

where λ_i are positive regularization parameters. For ease of exposition we assume $\lambda_1 = \cdots = \lambda_n = \lambda$. It is well known that $\boldsymbol{\beta}_i^* = \arg\min_{\boldsymbol{\beta}_i} \mathbb{E}\left\{\mathcal{L}_T\left(\boldsymbol{\beta}_i\right)\right\}$ are the population parameters in (5), under stated conditions.

We follow the steps in Negahban *et al.* (2012) to derive error bounds for the equation-wise LASSO estimator. First define the pair of subspaces $\mathcal{M}(S) = \left\{\boldsymbol{u} \in \mathbb{R}^{np} | u_i = 0, i \in S^c\right\}$ and its orthogonal complement $\mathcal{M}^{\perp}(S) = \left\{\boldsymbol{u} \in \mathbb{R}^{np} | u_i = 0, i \in S\right\}$, where $S \subseteq \{1, \ldots, np\}$. Set $\boldsymbol{u}_{\mathcal{M}}$ and $\boldsymbol{u}_{\mathcal{M}^{\perp}}$ the projection of $\boldsymbol{u}$ on $\mathcal{M}(S)$ and $\mathcal{M}^{\perp}(S)$ respectively. Clearly, for any $\boldsymbol{u} \in \mathbb{R}^{np}$, $|u|_1 = |\boldsymbol{u}_{\mathcal{M}}|_1 + |\boldsymbol{u}_{\mathcal{M}^{\perp}}|_1$. We say $|\cdot|_1$ is decomposable with respect to the pair $\left(\mathcal{M}(S), \mathcal{M}^{\perp}(S)\right)$ for any set $S \subset \{1, \ldots, np\}$.

We have to show two conditions to obtain a finite sample estimation error bound for the parameter vectors. The first condition is known as *restricted strong convexity* (RSC) and restricts the geometry of the loss function around the optimum $\boldsymbol{\beta}^*$ and is related to the Restricted Eigenvalue (Van De Geer and Bühlmann, 2009). The second condition is known as *deviation bound* and restricts the size of the sup-norm of the gradient $\nabla\mathcal{L}_T\left(\boldsymbol{\beta}^*\right)$. These conditions are shown to be satisfied in a set of large probability defined in Propositions 1 and 2.

Definition 1 (Deviation bound (DB))**.** The *deviation bound* condition holds when the regularization parameter λ satisfies $\left\{\lambda \geq 2|\mathbf{X}'\boldsymbol{U}_i/T|_\infty\right\}$ for all $i = 1, \ldots, n$.

Note that one may adopt individual λ_is for each equation, in which above definition should be modified adequately.

Definition 2 (Restricted Strong Convexity (RSC))**.** Define $\mathbb{C}\left(\boldsymbol{\beta}^*, \mathcal{M}, \mathcal{M}^{\perp}\right) = \left\{\boldsymbol{\Delta} \in \mathbb{R}^{np} | |\boldsymbol{\Delta}_{\mathcal{M}^{\perp}}|_1 \leq 3|\boldsymbol{\Delta}_{\mathcal{M}}|_1 + 4|\boldsymbol{\beta}^*_{\mathcal{M}^{\perp}}|_1\right\}$. The *restricted strong convexity* holds for parameters $\kappa_{\mathcal{L}}$ and $\tau_{\mathcal{L}}$ if for any $\boldsymbol{\Delta} \in \mathbb{C}$,

$$\frac{\boldsymbol{\Delta}'\mathbf{X}'\mathbf{X}\boldsymbol{\Delta}}{T} \geq \kappa_{\mathcal{L}}|\boldsymbol{\Delta}|_2^2 - \tau_{\mathcal{L}}^2(\boldsymbol{\beta}^*).$$

DOI: 10.1111/jtsa.12627

Negahban *et al.* (2012, Section 4) show these conditions are satisfied by many loss functions and penalties. Basu and Michailidis (2015) show that both DB and RSC are satisfied by Gaussian VAR(p) models in high dimensions.

If both DB and RSC hold with large probability, Negahban *et al.* (2012, Theorem 1) provides an ℓ_2 estimation bound for $\widehat{\boldsymbol{\beta}}_i$. Our goal is to show that the error bounds are valid for each $\boldsymbol{\Delta}_i = \widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^*$, $i = 1, \ldots, n$ at the same time.

Lemma 1 characterizes the solutions of the optimization program in (7). We require further notation. Define $\mathbb{C}_i := \mathbb{C}\left(\boldsymbol{\beta}_i^*, \mathcal{M}_{i,\eta}, \mathcal{M}_{i,\eta}^{\perp}\right)$ for a pair of subsets $\mathcal{M}_{i,\eta} = \mathcal{M}\left(S_{i,\eta}\right)$ and $\mathcal{M}_{i,\eta}^{\perp} = \mathcal{M}^{\perp}\left(S_{i,\eta}\right)$, where $S_{i,\eta} = \left\{j \in \{1, \ldots, pn\} | |\beta_{i,j}| > \eta\right\}$ and $S_{i,\eta}^c = \{j \in \{1, \ldots, pn\} | |\beta_{i,j}| \leq \eta\}$. These sets represent the *active parameters* under weak sparsity. In Theorem 1 we set $\eta = \lambda/\sigma_{\Gamma}^2$ to derive our results.

Lemma 1. Suppose $\{\mathbf{y}_t\}$ is generated from (1) and Assumptions (A1)–(A3) are satisfied. Set

$$\lambda > \tau^* \left(\epsilon + \log(Tn^2p)\right)^{2/\alpha} \sqrt{\frac{\epsilon + \log(n^2p)}{T}}, \tag{8}$$

where $\epsilon > 0$ and $\tau^* > 0$ depends on τ, α and $\bar{c}_{\Phi}$. Then, if $T > \epsilon + \log(n^2p)$, the event $\left\{\forall i = 1, \ldots, n : \widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}^* \in \mathbb{C}_i\right\}$ holds with probability at least $1 - 10e^{-\epsilon}$.

Lemma 1 shows that under restrictions on λ the solutions to the optimization program in (7) lie inside the star-shaped sets $\mathbb{C}_i$ with high probability, as the sample size increases. It restricts the directions in which we should control the variation of our estimators. Next result shows the deviation bound holds with high probability for appropriate choice of λ. To formalize the idea, let

$$\mathcal{D}_i(\lambda) = \left\{\lambda \geq 2\left|\frac{1}{T}\boldsymbol{X}'\mathbf{U}_i\right|_{\infty}\right\}, \quad i = 1, \ldots, n, \tag{9}$$

denote the event '*DB holds for equation i with regularization parameter* λ'.

Proposition 1 (Deviation bound)**.** Suppose that $\{\mathbf{y}_t\}$ is generated from (1), Assumptions (A1)–(A3) are satisfied and $T > \epsilon + \log(n^2p)$ for some $\epsilon > 0$. Set the penalty parameters λ as in (8). Then,

$$\Pr\left(\bigcup_{i=1}^{n} \mathcal{D}_i^c(\lambda)\right) \leq \pi_1(\epsilon) := 10e^{-\epsilon}.$$

Suppose $\epsilon = \log(np)$, $n^2p > T$. The regularization parameter λ satisfies

$$\lambda \gtrsim [\log(np)]^{2/\alpha} \sqrt{\frac{\log(np)}{T}},$$

and $\pi_1(\lambda) \propto 1/n^2p$. This regularization parameter is $O\left([\log(np)]^{2/\alpha}\right)$ larger, in rate, than one obtained in Wong *et al.* (2020, Proposition 7). Their results relied heavily in $\mathbf{y}_t$ being a β-mixing sequence in a sense that the concentration inequality derived in Merlevède *et al.* (2011) depends on it. In our case, the dependence is characterized by the conditional variance of the innovation process and coefficients $\Phi_1, \Phi_2, \ldots$, and we are not aware of 'tight' concentration inequalities that hold under these assumptions. Nevertheless, for fixed n, it is possible to show that the concentration inequality for sub-Weibull martingales in Lemma 5 is tight (Fan *et al.*, 2012b).

DOI: 10.1111/jtsa.12627

Let $\mathbf{\Gamma}_T = \mathbf{X}'\mathbf{X}/T$ denote the scaled Gram matrix and $\mathbf{\Gamma}$ its expected value. We show that if each element in $\mathbf{\Gamma}_T$ is sufficiently close to its expectation, and Assumptions (A4) and (A5) hold, then RSC is satisfied with high probability.

Lemma 2 (Restricted strong convexity)**.** Suppose Assumptions (A4) and (A5) hold and that $\left|\left|\left|\mathbf{\Gamma}_T - \mathbf{\Gamma}\right|\right|\right|_{\max} \leq \frac{\sigma_\Gamma^2 \eta^q}{64 R_q}$. Then, for any $\mathbf{\Delta}_i \in \mathbb{C}_i$ for $1 \leq i \leq n$

$$\mathbf{\Delta}_i' \mathbf{\Gamma}_T \mathbf{\Delta}_i \geq \frac{\sigma_\Gamma^2}{2} |\mathbf{\Delta}_i|_2^2 - \frac{\sigma_\Gamma^2}{2} R_q \eta^{2-q}. \tag{10}$$

To show RSC holds with high probability for all $i = 1, \ldots, n$ at the same time, we have to bound the event

$$\mathcal{B}(a) = \left\{ \left|\left|\left|\mathbf{\Gamma}_T - \mathbf{\Gamma}\right|\right|\right|_{\max} \leq a \right\}. \tag{11}$$

where $a = \frac{\sigma_\Gamma^{2(1-q)} \lambda^q}{64 R_q}$. If we assume distinct $R_{q,i}$ and q_i for each equation, we should work with $\cap_i \mathcal{B}_i$ and $\mathcal{B}_i$ defined accordingly.

Proposition 2. Suppose Assumptions (A1)–(A3) hold. If

$$p < \frac{T^{\gamma_1 \wedge \gamma_2}}{\left(\frac{2}{\gamma_1 \wedge \gamma_2 + 1}\right)\left(2 + \frac{1.4}{2\gamma_1 c_\phi \wedge a_2}\right)},$$

and

$$a \geq \sqrt{\frac{2(1+\xi)^{1+2/\alpha} \tau^2 [\log(npT)]^{1+2/\alpha}}{T}},$$

for some $\xi > 0$, then $\Pr(\mathcal{B}^c(a)) \leq \pi_2(a)$, where

$$\pi_2(a) := \frac{2}{(np)^\xi T^{1+\xi}} + \frac{8}{(np)^\xi T^\xi} + \frac{n^2}{a}\left(b_1 e^{-c_\phi \wedge a_2 (T/2)^{\gamma_2 \wedge \gamma_1}} + b_8 e^{-2\gamma_1 c_\phi (T/2)^{\gamma_1}}\right).$$

This bound controls the proximity between the empirical and population covariance matrices. Similar concentration inequalities were derived by Kock and Callot (2015, Lemma 9), Loh and Wainwright (2012, Lemma 14) and Medeiros and Mendes (2017). Their results, however, cannot be applied in our setting. Explicit expressions for the constants b_1, b_8 and τ in Proposition 2 are found in Lemma 6. Also, one may replace ϵ by its lower bound to remove dependence.

This concentration guided the choice of dependence condition used in this work. Traditionally one uses either a Hanson–Wright inequality or a Bernstein or Hoeffding type inequality to bound the empirical covariance around its mean. We write the centered Gram matrix $\mathbf{\Gamma}_T - \mathbf{\Gamma}$ as a sum of martingales and a dependence term. The martingales are handled using a Bernstein type bound and the dependence term is handled using both assumptions (A1) and (A2). Combined, they imply a sub-Weibull type decay on expected value of dependence term.

Finally, we use the bounds $\pi_1(\cdot)$ and $\pi_2(\cdot)$ in Proposition 1 and Proposition 2 respectively, to derive an upper bound for the prediction error and for the difference between the lasso parameter estimates and the true parameters in the ℓ_2 norm.

DOI: 10.1111/jtsa.12627

Theorem 1. Suppose assumptions (A1) – (A3) hold. Under conditions of Proposition 1, there exists $T_0 > 0$ such that for all $T \geq T_0$,

$$\left| \boldsymbol{X} \left(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^* \right) / T \right|_2^2 \leq 12 |\boldsymbol{\beta}_i^*|_1 \lambda, \quad i = 1, \ldots, n,$$

with probability at least $1 - \pi_1(\epsilon)$ for $\epsilon > 0$. Suppose further that assumptions (A4) and (A5) hold. Set $\eta = \lambda / \sigma_\Gamma^2$. Under conditions of Propositions 1 and 2, there exists $T_0 > 0$ such that for all $T \geq T_0$,

$$|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^*|_2^2 \leq (44 + 2\lambda) R_q \left(\frac{\lambda}{\sigma_\Gamma^2} \right)^{2-q}, \quad i = 1, \ldots, n,$$

with probability at least $1 - \pi_1(\epsilon) - \pi_2 \left(\frac{\sigma_\Gamma^{2(1-q)} \lambda^q}{64 R_q} \right)$.

Theorem 1 states that, with high probability, estimated and population parameter vectors are close to each other in the Euclidean norm. It requires that Propositions 1 and 2 hold jointly, meaning that λ, R_q and σ_Γ^2 must satisfy rate conditions. We show that if the size of 'small' coefficients and smallest eigenvalue σ_Γ^2 of Γ are restricted, then the rate of λ in Proposition 1 is unaffected. For $\epsilon = \log(np)$ and $T < np^2$ Proposition 1 requires, after simplification,

$$\lambda \geq \tau^* \log(np)^{2/\alpha} \sqrt{\frac{4 \log(np)}{T}},$$

for some constant τ^*. Replacing a by $\frac{\sigma_\Gamma^{2(1-q)} \lambda^q}{64 R_q}$ in Proposition 2 we obtain

$$\lambda^q \gtrsim \left(\log(np)^{2/\alpha} \sqrt{\frac{\log(np)}{T}} \right) \times \left(\frac{R_q}{\sigma_\Gamma^{2(1-q)} \log(np)^{1/\alpha}} \right).$$

However, it is not necessarily a constraint in the rate of λ. Propositions 1 and 2 will hold jointly for T sufficiently large for $0 \leq q < 1$ if

$$\frac{R_q}{\sigma_\Gamma^{2(1-q)}} = o \left(\log(np)^{(2q-1)/\alpha} \left(\frac{T}{\log(np)} \right)^{(1-q)/2} \right).$$

In other words, if the *small* parameters are not too large and smallest eigenvalue of $\boldsymbol{\Sigma}$ is not too small as a function of T.

4.1. Simulation

To evaluate the finite-sample performance of the LASSO estimator in large VARs with stochastic volatility, we consider the following model:

$$\begin{aligned} \mathbf{y}_t &= \boldsymbol{A}_0 + \boldsymbol{A}_1 \mathbf{y}_{t-1} + \boldsymbol{A}_4 \mathbf{y}_{t-4} + \boldsymbol{H}_t^{1/2} \boldsymbol{v}_t, \quad \boldsymbol{v}_t \sim \mathsf{N}(\mathbf{0}, \boldsymbol{I}) \\ \boldsymbol{H}_{t+1} &= \boldsymbol{C}_0 + \boldsymbol{\Psi} \boldsymbol{H}_t \boldsymbol{\Psi}' + \boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t', \quad \boldsymbol{\epsilon}_t \sim \mathsf{N}(\mathbf{0}, \boldsymbol{I}). \end{aligned} \tag{12}$$

We consider 1000 replications of model (12) with $T = 100, 300$ observations and the number of series is a function of the sample size, that is, $n = c \times T$, where $c = \{1, 2, 3\}$. $\boldsymbol{A}_1$ and $\boldsymbol{A}_4$ are two block-diagonal matrices with

DOI: 10.1111/jtsa.12627

Table I. Simulation results

T	$n = T$	$n = 2T$	$n = 3T$
Panel (a): Mean squared estimation error			
100	0.1000	0.1926	0.2871
300	0.0288	0.0576	0.0914
Panel (b): Mean squared forecasting error			
100	1.0900	1.1456	1.2891
300	1.0150	1.0248	1.0345

Note: The table reports the simulation results of the LASSO estimation of model (12) for different combinations of sample size and number of variables. The penalty parameter of the LASSO is selected by the BIC. Panel (a) reports the mean squared error of the estimation of the VAR parameters. Panel (b) reports the ratio of the mean squared out-of-sample forecasting error of the LASSO with respect to the Oracle forecast.

blocks of dimension 5×5 and entries equal to 0.15 and -0.1 respectively. $\boldsymbol{C}_0$ and $\boldsymbol{\Psi}$ are two diagonal matrices with diagonal elements all equal to 1e$-$5 and 0.8, respectively.

This model satisfies Assumptions (A1)–(A5). First, Assumption (A1) is satisfied because the model is block diagonal, with fixed block size. Assumptions (A2) and (A3) follows as in Example 4 given diagonal specification of $\boldsymbol{C}_0$ and $\boldsymbol{\Psi}$. Assumption (A4) is satisfied with $q = 0$ and $R = 10$. Finally, assumption (A5) holds because lower bound in (6), is strictly positive $\left(\sum_{i=1}^{4} \left(|||\boldsymbol{A}_i|||_1 + |||\boldsymbol{A}_i|||_\infty\right) < \infty \text{ and } \Lambda_{\min}(\boldsymbol{\Sigma}) > 0\right.$, independently of n and $p\big)$.

Table I reports the simulation results of the LASSO estimation of model (12) for different combinations of sample size and number of variables. The penalty parameter of the LASSO is selected by the BIC. Panel (a) reports the mean squared error (MSE) of the estimation of the VAR parameters. Panel (b) reports the ratio of the mean squared out-of-sample forecasting error (MSFE) of the LASSO with respect to the Oracle forecast. A couple of findings emerge from the table. First, as expected the performance of the LASSO improves with the sample size. Second, as the number of candidate variables grows, the MSE increases. This is also expected. Finally, the out-of-sample forecasts of the LASSO get quite close to the ones from the Oracle when $T = 300$, even when $n = 3T$.

5. DISCUSSION

This work provides finite sample ℓ_2 error bounds for the equation-wise LASSO parameters estimates of a weakly sparse, high-dimensional, VAR(p) model, with dependent and heavy tailed innovation process. It covers a large collection of specifications as illustrated in Section 3.

A distinctive feature of this work is that the dependence structure of the innovations are characterized by a very weak projective dependence condition that is naturally verifiable in settings where one is interested in the conditional variance of the process. The series of innovations is not necessarily mixing nor the resulting time series $\{\mathbf{y}_t\}$.

Our bounds hold under a heavy tailed setting in a sense that we do not require the moment generating function to exist. Despite the tails in $\{\mathbf{y}_t\}$ being sub-Weibull as in Wong *et al.* (2020), we are not able to recover the same rates and lower bound for the regularization parameter λ. The reason is that Wong *et al.* (2020) bounds rely heavily on the concentration inequality for mixing sequences in Merlevède *et al.* (2011). Given the weak projective dependence adopted, we chose to use a martingale concentration and overcome all together the issue of using the dependence metric for deriving the concentration bound. Nevertheless, we believe the loss in efficiency is minimal. Close inspection of proof of Lemma 5 shows that the loss of efficiency is concentrated in bounding the tail. It amounts to an extra $\log(T)$ term, which does not change the rates under assumption that $T < n^2 p$, eventually.

A limitation of this work is the restriction that the model is correctly specified in the mean, in a sense that innovations are martingale differences. Nevertheless, this assumption is standard in the literature and we are able

DOI: 10.1111/jtsa.12627

to derive results covering a broad range of data generating processes and conditional dependence measures. The martingale difference condition cannot be relaxed at this moment as our deviation bound depends on it. Furthermore, we do not require strong sparsity in a sense that near zero coefficients are effectively treated as zero as long as they are concentrated in some slowly increasing ℓ_q ball ($0 \leq q < 1$) around the origin.

Results in this article can be extended to polynomial tails. The strategy is to replace the martingale concentration in Lemma 4, used to prove Propositions 1 and 2 by

$$\Pr\left(\max_{1\leq i\leq n}\left|\sum_{t=1}^{T}\xi_{it}\right| > Ta\right) \leq \frac{nK}{\left(a\sqrt{T}\right)^{d}},$$

whenever $\|\xi_{it}\|_d < \infty$. If available under our dependence conditions, one could employ a Fuk–Nagaev type inequality. Nevertheless, it follows that under appropriate changes to concentration rates, equation-wise LASSO estimators also admit oracle bounds. A direct consequence is that moments conditions on Carrasco and Chen (2002) and Hafner and Preminger (2009a,b) are directly applicable.

Despite working with a relatively simple structure and estimation model, the machinery can be applied to more complex settings. The key points are showing that the empirical covariance concentrates around its mean in terms of its maximum entry-wise norm and the concentration inequality for large dimensional, sub-Weibull martingales. Following development of Negahban *et al.* (2012), the results may be naturally extended to structured regularization with node-wise regression and replacing using the Frobenius norm for system estimation. Finally, the high-dimensional VAR specification encompasses large dimensional vector-panels among other models.

In Adamek *et al.* (2020), authors consider a near epoch dependent time series. This condition covers misspecification and non-Gaussian, conditionally heteroskedastic models, such as GARCH innovations. The main focus of the authors is on inference using the desparsified lasso, but estimation error bounds are also derived. Assumptions are in the same line of Medeiros and Mendes (2016), where concentration bounds are assumed to hold in probability. In contrast, we focus on finite sample error bounds with relatable dependence condition on the conditional covariance. The heavy lifting in our article is to derive the required concentration inequalities. Effectively, one could use our results to provide theoretical justification for the concentration bounds under particular model specifications. On the other hand, Adamek *et al.* (2020) provides theoretical justification for desparsified inference in some large dimensional models discussed in our article.

ACKNOWLEDGEMENTS

The authors are grateful to Anders B. Kock, Giuseppe Cavaliere, Etienne JJ Wijler, editor (Robert Taylor) and co-editor of JTSA, and two anonymous referees. Their comments and suggestions led to a much improved version of this manuscript. M. C. Medeiros acknowledges partial support from CNPq/Brazil and CAPES.

DATA AVAILABILITY STATEMENT

The code that supports the simulation experiment in this study is available from the corresponding author on request.

REFERENCES

Adamek R., Smeekes S., Wilms I. 2020. *Lasso inference for high-dimensional time series*. arXiv preprint arXiv:2007.10952.

Andrews DW. 1988. Laws of large numbers for dependent non-identically distributed random variables. *Econometric Theory* **4**(3): 458–467.

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

Bauer G, Vornik K. 2011. Forecasting multivariate realized stock market volatility. *Journal of Econometrics* **160**: 93–101.

DOI: 10.1111/jtsa.12627

Bradley RC. 2005. Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys* **2**(2): 107–144.
Carrasco M, Chen X. 2002. Mixing and moment properties of various GARCH and stochastic volatility models. *Econometric Theory* **18**: 17–39.
Chiriac R, Voev V. 2011. Modelling and forecasting multivariate realized volatility. *Journal of Applied Econometrics* **26**: 922–947.
Davidson J. 1994. *Stochastic Limit Theory*. Oxford: Oxford University Press.
Davis R, Zang P, Zhang T. 2016. Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics* **25**: 1077–1096.
Dedecker J, Doukhan P, Lang G, León J, Louhichi S, Prieur C. 2007. *Weak Dependence with Examples and Applications*. Berlin: Springer.
Ensor K, Raun L, Persse D. 2013. A case-crossover analysis of out-of-hospital cardiac arrest and air pollution. *Circulation* **127**: 1192–1199.
Fan X, Grama I, Liu Q. 2012a. Hoeffding's inequality for supermartingales. *Stochastic Processes and their Applications* **122**(10): 3545–3559.
Fan X, Grama I, Liu Q. 2012b. Large deviation exponential inequalities for supermartingales. *Electronic Communications in Probability* **17**.
Hafner CM, Preminger A. 2009a. Asymptotic theory for a factor GARCH model. *Econometric Theory* **25**(02): 336–363.
Hafner CM, Preminger A. 2009b. On asymptotic theory for multivariate GARCH models. *Journal of Multivariate Analysis* **100**(9): 2044–2054.
Han Y, Tsay RS. 2020. High-dimensional linear regression for dependent data with applications to nowcasting. *Statistica Sinica* **30**: 1797–1827.
Hoek G, Krishnan R, Beelen R, Peters A, Ostro B, Brunekreef B, Kaufman J. 2013. Long-term air pollution exposure and cardio-respiratory mortality: a review. *Environmental Health* **12**: 43.
Kock A, Callot L. 2015. Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics* **186**: 325–344.
Kock A., Medeiros M., Vasconcelos G. 2020. Penalized regressions. In *Macroeconomic Forecasting in the Era of Big Data: Theory and Practice. Advanced Studies in Theoretical and Applied Econometrics*, Vol. 52. Berlin: Springer; 193–228.
Loh P-L, Wainwright M. 2012. High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity. *The Annals of Statistics* **40**: 1637–1664.
Lütkepohl H. 1991. *Introduction to Multiple Time Series Analysis*. Berlin: Springer-Verlag.
Lütkepohl H. 2006. *New Introduction to Multiple Time Series Analysis*. Berlin: Springer-Verlag.
Medeiros M., Mendes E. 2016. ℓ_1 regularization of high-dimensional time-series models with non-Gaussian and heteroskedastic innovations. *Journal of Econometrics* **191**: 255–271.
Medeiros M, Mendes E. 2017. Adaptive LASSO estimation for ARDL models with GARCH innovations. *Econometric Reviews* **622–637**(6–9).
Merlevède F, Peligrad M, Rio E. 2011. A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields* **151**(3–4): 435–474.
Miao K, Phillips PC, Su L. 2020. *High-dimensional VARs with common factors*, Cowles Foundation Discussion Papers 2252, Cowles Foundation for Research in Economics, Yale University.
Miranda-Agrippino S, Ricco G. 2019. *Bayesian vector autoregressions: Estimation*, Oxford Research Encyclopedia of Economics and Finance.
Negahban SN, Ravikumar P, Wainwright MJ, Yu B. 2012. A unified framework for high-dimensional analysis of *m*-estimators with decomposable regularizers. *Statistical Science* **27**(4): 538–557.
Ramey V. 2016. Macroeconomic shocks and their propagation. In *Handbook of Macroeconomics* Amsterdam: Elsevier; 71–162.
Schweinberger M, Babkin S, Ensor K. 2017. High-dimensional multivariate time series with additional structure. *Journal of Computational and Graphical Statistics* **26**: 610–622.
Sims C. 1980. Macroeconomics and reality. *Econometrica* **48**: 1–48.
Song S., Bickel P. J. 2011. *Large vector auto regressions*, arXiv preprint arXiv:1106.3915.
Van De Geer SA, Bühlmann P. 2009. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics* **3**: 1360–1392.
Wilson G, Reale M, Haywood J. 2015. *Developments in Multivariate Time Series Modeling*. Boca Raton: CRC Press.
Wong K, Li Z, Tewari A. 2020. Lasso guarantees for β-mixing heavy tailed time series. *Annals of Statistics* **48**(2): 1124–1142.

DOI: 10.1111/jtsa.12627

APPENDIX A. PROOF OF MAIN RESULTS

A.1. Proof of Lemma 1

We apply Negahban *et al.* (2012, Lemma 1). The empirical loss $\mathcal{L}_T\left(\boldsymbol{\beta}_i\right)$ is convex for each i. Proposition 1 ensures each (9) hold with desired probabilities. □

A.2. Proof of Proposition 1

Write the event $\mathcal{A}_i = \left\{\max_j |\mathbf{u}_i'\mathbf{x}_j| < T\lambda_0/2\right\}$. We shall derive probability bounds for $\Pr\left(\cap_{i=1}^n \mathcal{A}_i\right) \geq 1 - \Pr\left(\max_{i,j} |\mathbf{u}_i'\mathbf{x}_j| \geq T\lambda_0/2\right)$. We bound the probability using Corollary 1.

Under Assumption (A2), $\mathbf{u}_i'\mathbf{x}_j = \sum_{t=p}^T u_{ti}y_{t-s,j}$ ($s = 1, \ldots, p$ and $i,j = 1, \ldots, n$) is a martingale and each $u_{ti}y_{t-s,j}$ is a martingale difference process. Hence, we follow by applying Corollary 1. Conditions on T and $\lambda_0/2$ are already satisfied. We need to show that $u_{ti}y_{t-s,j}$ is sub-Weibull. For each $d \geq 1$, $\|u_{ti}y_{t-s,j}\|_d \leq \|u_{ti}\|_{2p}^{1/2}\|y_{t-s,j}\|_{2p}^{1/2} \leq \bar{c}_\Phi \max_{|\mathbf{b}|_1\leq 1} \|\mathbf{b}'\mathbf{u}\|_{2p}$, by Lemma 3. Then, it follows from Wong *et al.* (2020, Lemmas 5 and 6) and Assumption (A3) that $u_{ti}y_{t-s,j}$ is sub-Weibull with parameter $\alpha/2$. Hence, there is some constant τ^* depending on τ, $\bar{c}_\Phi$ and α such that $\Pr\left(|u_{ti}y_{t-s,j}| > x\right) \leq 2\exp\left(-|x/\tau^*|^{\alpha/2}\right)$. Result follows.

A.3. Proof of Lemma 2

For notational simplicity, write $\|\boldsymbol{\Gamma}_T - \boldsymbol{\Gamma}\|_{\max} \leq \delta \leq \sigma_\Gamma^2/64\psi^2\left(\mathcal{M}_{i,\eta}\right)$ where $\psi\left(\mathcal{M}_{i,\eta}\right) = \sup_{u\in\mathbb{C}_i} |u|_1/|u|_2 = \sqrt{|S_{i,\eta}|}$. Using the arguments in (Negahban *et al.*, 2012, section 4.3), $|\boldsymbol{\beta}_{i,\mathcal{M}_{i,\eta}^\perp}|_1 \leq \sum_{j\in\mathcal{S}_{i,\eta}^c} |\beta_{i,j}|^q|\beta_{i,j}|^{1-q} \leq \eta^{1-q}R_q$, and $R_q \geq \sum_{j\in S_{i\eta}} |\beta_{i,j}|^q \geq |S_{i,\eta}|\eta^q$. Hence $\frac{\sigma_\Gamma^2\eta^q}{64R_q} \leq \frac{\sigma_\Gamma^2}{64\psi^2\left(\mathcal{M}_{i,\eta}\right)}$. It follows that

$$
\begin{aligned}
\boldsymbol{\Delta}_i'\boldsymbol{\Gamma}_T\boldsymbol{\Delta}_i &= \boldsymbol{\Delta}_i'\Gamma\boldsymbol{\Delta}_i + \boldsymbol{\Delta}_i'[\boldsymbol{\Gamma}_T - \boldsymbol{\Gamma}]\boldsymbol{\Delta}_i \\
&\geq |\boldsymbol{\Delta}_i|_2^2 \inf_{\mathbf{u}\in\mathbb{C}_i\setminus\{0\}} \frac{\mathbf{u}'\boldsymbol{\Gamma}\mathbf{u}}{\mathbf{u}'\mathbf{u}} - |\boldsymbol{\Delta}_i|_1|[\boldsymbol{\Gamma}_T - \boldsymbol{\Gamma}]\boldsymbol{\Delta}_i|_\infty \\
&\geq \sigma_\Gamma^2|\boldsymbol{\Delta}_i|_2^2 - |\boldsymbol{\Delta}_i|_1^2\left\|\left|\boldsymbol{\Gamma}_T - \boldsymbol{\Gamma}\right|\right\|_{\max} \\
&\geq \sigma_\Gamma^2|\boldsymbol{\Delta}_i|_2^2 - \delta|\boldsymbol{\Delta}_i|_1^2 \\
&\geq \sigma_\Gamma^2|\boldsymbol{\Delta}_i|_2^2 - \delta\left(4|\boldsymbol{\Delta}_{i,\mathcal{M}_{i,\eta}}|_1 + 4|\boldsymbol{\beta}_{i,\mathcal{M}_{i,\eta}^\perp}|_1\right)^2 \\
&\geq |\boldsymbol{\Delta}_i|_2^2\left(\sigma_\Gamma^2 - 32\delta\psi\left(\mathcal{M}_{i,\eta}\right)^2\right) + 32\delta|\boldsymbol{\beta}_{i,\mathcal{M}_{i,\eta}^\perp}|_1^2 \\
&\geq |\boldsymbol{\Delta}_i|_2^2\frac{\sigma_\Gamma^2}{2} - \frac{\sigma_\Gamma^2}{2\psi^2\left(\mathcal{M}_{i,\eta}\right)}|\boldsymbol{\beta}_{i,\mathcal{M}_{i,\eta}^\perp}|_1^2 \\
&\geq |\boldsymbol{\Delta}_i|_2^2\frac{\sigma_\Gamma^2}{2} - \frac{\sigma_\Gamma^2}{2R_q\eta^{-q}}\eta^{2(1-q)}R_q^2 \\
&= |\boldsymbol{\Delta}_i|_2^2\frac{\sigma_\Gamma^2}{2} - \frac{\sigma_\Gamma^2}{2}\eta^{2-q}R_q,
\end{aligned}
$$

proving the result. □

A.4. Proof of Proposition 2

The proof consists on a trivial application of Lemmas 6 setting $\epsilon = \sigma_\Gamma^{2(1-q)}\lambda^q/64R_q$. □

DOI: 10.1111/jtsa.12627

A.5. Proof of Theorem 1

We apply Negahban *et al.* (2012, Theorem 1). Lemma 1 ensures λ is selected accordingly, $\mathcal{L}_T\left(\boldsymbol{\beta}_i\right)$ is a convex function of $\boldsymbol{\beta}_i$, Lemma 2 ensures RSC is satisfied with $\kappa_{\mathcal{L}} = \sigma_\Gamma^2/2$ and $\tau_{\mathcal{L}}^2\left(\boldsymbol{\beta}_i\right) = \frac{\sigma_\Gamma^2 \eta^{2-q} R_q}{2}$. Define $\psi\left(\mathcal{M}_{i,\eta}\right)$ as in the proof of Lemma 2 and recall $|S_{i,\eta}| \leq R_q \eta^{-q}$ and $|\boldsymbol{\beta}_{i,\mathcal{M}_{i,\eta}^\perp}|_1 \leq R_q \eta^{1-q}$, and that $\eta = \lambda/\sigma_\Gamma^2$. For each i,

$$
\begin{aligned}
|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}^*|_2^2 &\leq 9\frac{\lambda}{\kappa_{\mathcal{L}}^2}\psi^2\left(\mathcal{M}_{i,\eta}\right) + \frac{\lambda}{\kappa_{\mathcal{L}}}\left[2\tau_{\mathcal{L}}^2\left(\boldsymbol{\beta}_i^*\right) + 4|\boldsymbol{\beta}_{i,\mathcal{M}_{i,\eta}^\perp}|_1\right] \\
&\leq 36\frac{\lambda}{\sigma_\Gamma^4}R_q\eta^{-q} + 2\frac{\lambda}{\sigma_\Gamma^2}\left[R_q\eta^{2-q}\sigma_\Gamma^2 + 4R_q\eta^{1-q}\right] \\
&\leq 36\frac{\lambda}{\sigma_\Gamma^4}R_q\left(\frac{\lambda}{\sigma_\Gamma^2}\right)^{-q} + 2\frac{\lambda}{\sigma_\Gamma^2}\left[R_q\left(\frac{\lambda}{\sigma_\Gamma^2}\right)^{2-q}\sigma_\Gamma^2 + 4R_q\left(\frac{\lambda}{\sigma_\Gamma^2}\right)^{1-q}\right] \\
&\leq 36R_q\left(\frac{\lambda}{\sigma_\Gamma^2}\right)^{2-q} + 2\lambda R_q\left(\frac{\lambda}{\sigma_\Gamma^2}\right)^{2-q} + 8R_q\left(\frac{\lambda}{\sigma_\Gamma^2}\right)^{2-q} \\
&= (44 + 2\lambda)R_q\left(\frac{\lambda}{\sigma_\Gamma^2}\right)^{2-q}.
\end{aligned}
$$

□

APPENDIX B. AUXILIARY LEMMATA

B.1. Properties of $\mathbf{y}_t$

We will derive properties of the process $\{\mathbf{y}_t\}$ described in (1)

Lemma 3. Suppose that for some norm $\|\cdot\|_\psi$ we have

$$\max_t \max_{|\mathbf{b}|_1 \leq 1} \|\mathbf{b}'\mathbf{u}_t\|_\psi \leq c_\psi,$$

for some constant $c_\psi < \infty$ that only depends on the norm $\|\cdot\|_\psi$. Then, under conditions (A1) and (A2), for all t and $i \in \{1, \ldots, n\}$,

$$\|y_{i,t}\|_\psi \leq c_\Phi \times \sum_{j=0}^{\infty} |\boldsymbol{e}_i'\boldsymbol{\Phi}_j|_1.$$

Proof. Under assumption (A1) the VAR model in (1) admits the VMA(∞) representation (4) for all n and p. Let $\{\boldsymbol{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)', i = 1, \ldots, n\}$ the canonical basis vectors. Then, for all i, $y_{i,t} = \boldsymbol{e}_i'\mathbf{y}_t$ and

$$
\begin{aligned}
\|\boldsymbol{e}_i'\mathbf{y}_t\|_\psi &= \left\|\sum_{j=0}^{\infty} \boldsymbol{e}_i'\boldsymbol{\Phi}_j\mathbf{u}_{t-j}\right\|_\psi \\
&= \left\|\sum_{j=0}^{\infty}\sum_{k=1}^{n} \boldsymbol{e}_i'\boldsymbol{\Phi}_j\boldsymbol{e}_k u_{k,t-j}\right\|_\psi
\end{aligned}
$$

DOI: 10.1111/jtsa.12627

$$
\begin{aligned}
&= \left\| \sum_{j=0}^{\infty} |\boldsymbol{e}_i' \boldsymbol{\Phi}_j|_1 \sum_{k=1}^{n} \frac{\boldsymbol{e}_i' \boldsymbol{\Phi}_j \boldsymbol{e}_k}{|\boldsymbol{e}_i' \boldsymbol{\Phi}_j|_*} u_{k,t-j} \right\|_\psi \\
&\leq \left(\sum_{j=0}^{\infty} |\boldsymbol{e}_i' \boldsymbol{\Phi}_j|_1 \right) \max_t \max_{|\mathbf{b}|_1 \leq 1} \left\| \mathbf{b}' \mathbf{u}_t \right\|_\psi \\
&\leq \sum_{j=0}^{\infty} |\boldsymbol{e}_i' \boldsymbol{\Phi}_j|_1 \times c_\psi,
\end{aligned}
$$

where $|\cdot|_* := |\cdot|_1 I(|\cdot| > 0) + I(|\cdot|_1 = 0)$. □

Due stability condition (A1), for each n and p, there exists $\bar{c}_\Phi$ such that $\sum_{i=0}^{\infty} |\boldsymbol{\phi}_{i,j}|_1 \leq \bar{c}_\Phi$ for all $j = 1, \ldots, n$. Let $\|\cdot\|_\psi$ be the Orlicz norm,

$$\|\cdot\|_\psi = \inf\{c > 0 : \psi(|\cdot|/c) \leq 1\},$$

where $\psi(\cdot) : \mathbb{R}^+ \mapsto \mathbb{R}^+$ is convex, increasing function with $\psi(0) = 0$ and $\psi(x) \to \infty$ as $x \to \infty$. Traditional choices of $\psi(\cdot)$ are (a) $\psi(x) = x^p$, $p \geq 1$, (b) $\psi(x) = \exp(x^a) - 1$, $a > 1$, and (c) $\psi(x) = (ae)^{1/a} x I\left(x \leq a^{-1/a}\right) + \exp(x^a) I\left(x > a^{-1/a}\right)$. These choices contemplate sub-Gaussian and sub-exponential tails, as well as process with heavy-tails, such as sub-Weibull and polynomial tails. Note that by combining this result with Wong *et al.* (2020, Lemmas 5 and 6) if $\left\{\mathbf{b}' \mathbf{u}_t\right\}$ are sub-Weibull, so are $\left\{\mathbf{b}' \mathbf{y}_t\right\}$.

Assumption (A1) is satisfied under restrictions on the parameter space. The stability assumption is standard in the literature whereas the tail sum (3) requires further constraints on the parameter matrices. Lemma 4 presents a sufficient set of restrictions on the sparse parameter matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_p$ so that (3) is satisfied.

Lemma 4. Suppose that for all n and p, there exists some $\rho > 0$ such that

$$\sum_{k=1}^{p} \left\| \left| \boldsymbol{A}_k \right| \right\|_\infty = \sum_{k=1}^{p} \max_{j=1,\ldots,n} |\boldsymbol{a}_{k,j}|_1 \leq e^{-\rho},$$

where $\boldsymbol{A}_k = \left[\boldsymbol{a}_{k,1} : \cdots : \boldsymbol{a}_{k,n}\right]'$. Then for every $h = 1, \ldots, n$,

(i) $|\boldsymbol{\phi}_{k,h}|_1 \leq \sum_{j=1}^{p \wedge k} \left\| \left| \boldsymbol{A}_j \right| \right\|_\infty |\boldsymbol{\phi}_{k-j,h}|_1$, $k = 1, 2, \ldots$
(ii) $\sum_{k=m}^{\infty} |\boldsymbol{\phi}_{k,h}|_1 \leq c_0 \, e^{-m\rho}$, $m \geq 1$, provided that for all p,

$$\max_{h=1,\ldots,n} \max_{k=1,\ldots,p} e^{k\rho} \times \sum_{j=1}^{k} \tilde{\alpha}_j |\boldsymbol{\phi}_{j,h}|_1 \leq (1 - e^{-\rho}) \, c_0, \tag{B1}$$

where $\alpha_i = e^\rho \left\| \left| \boldsymbol{A}_i \right| \right\|_\infty$ and $\tilde{\alpha}_i = \sum_{\boldsymbol{i} : |\boldsymbol{i}| = k-p+j} \prod_{l=1}^{k-p} \alpha_{i_l}$ where $\boldsymbol{i} = \left(i_1, \ldots, i_{k-p}\right)$ is a multi-index.

Proof. Starting from the recursive definition of $\Phi_k = \sum_{j=1}^{p \wedge k} \Phi_{k-j} \boldsymbol{A}_j$,

$$|\boldsymbol{\phi}_{k,h}|_1 = |\boldsymbol{e}_h' \Phi_k|_1 = \left| \sum_{j=1}^{p \wedge k} \boldsymbol{e}_h' \Phi_{k-j} \boldsymbol{A}_j \right|_1 \leq \sum_{j=1}^{p \wedge k} |\boldsymbol{\phi}_{k-j,h} \boldsymbol{A}_j|_1 \leq \sum_{j=1}^{p \wedge k} |\boldsymbol{\phi}_{k-j,h}|_1 \left\| \left| \boldsymbol{A}_j \right| \right\|_\infty.$$

DOI: 10.1111/jtsa.12627

Suppose $k \geq p$, let $\alpha_j = e^{\rho} \left|\!\left|\!\left| \boldsymbol{A}_j \right|\!\right|\!\right|_{\infty}$ and verify that $0 \leq \sum_{j=1}^{p} \alpha_j \leq 1$. Iterating on the previous argument $s \leq k-p$ times yields

$$
\begin{aligned}
|\boldsymbol{\phi}_{k,h}|_1 &\leq \sum_{j_1=1}^{p} \cdots \sum_{j_s=1}^{p} \left(\prod_{l=1}^{s} |\boldsymbol{A}_{j_l}|_{\infty} \right) |\boldsymbol{\phi}_{k-\sum_{l=1}^{s} j_l,h}|_1 \\
&= e^{-s\rho} \sum_{j_1=1}^{p} \cdots \sum_{j_s=1}^{p} \left(\prod_{l=1}^{s} \alpha_{j_l} \right) |\boldsymbol{\phi}_{k-\sum_{l=1}^{s} j_l,h}|_1 \\
&= \cdots \\
&= e^{-\rho(k-p)} \sum_{j=1}^{p} \left(\sum_{\boldsymbol{i}:|\boldsymbol{i}|_1=k-p+j} \prod_{l=1}^{k-p} \alpha_{i_l} \right) |\boldsymbol{\phi}_{p-j,h}|_1,
\end{aligned}
$$

where $\boldsymbol{i} = \left(i_1, \ldots, i_{k-p}\right)$ is a multi-index and the summation is over all combinations satisfying $|\boldsymbol{i}|_1 = k-p+j$. The term inside parentheses is $\tilde{\alpha}_j$ and under the conditions of the lemma

$$
|\boldsymbol{\phi}_{k,h}|_1 \leq e^{-\rho k} \times \left[e^{\rho p} \sum_{j=1}^{p} \tilde{\alpha}_j |\boldsymbol{\phi}_{p-j,h}|_1 \right] \leq (1-e^{-\rho})\, c_0\, e^{-\rho k}.
$$

The same result follows trivially for $k < p$ under the assumptions of the lemma.

Summing over all values of $k \geq m$,

$$
\sum_{k=m}^{\infty} |\boldsymbol{\phi}_{k,h}|_1 \leq c_0 (1-e^{-\rho}) \sum_{k=m}^{\infty} e^{-\rho k} = c_0 e^{-m\rho} \frac{\sum_{k=0}^{\infty} e^{-\rho k}}{(1-e^{-\rho})^{-1}} = c_0 e^{-m\rho}.
$$

□

B.2. Concentration Inequality for Martingales

We derive concentration bounds for martingales. In the firs theorem we consider martingales with at most d finite moments, whereas in the second we allow the tails of the marginal distributions to decrease at a sub-Weibull, sub-exponential or, even sub- and super-Gaussian rate.

Lemma 5 (Concentration bounds for high dimensional martingales)**.** Let $\{\boldsymbol{\xi}_t\}_{t=1,\ldots,T}$ denote a multivariate martingale difference process with respect to the filtration $\mathcal{F}_t$ taking values on $\mathbb{R}^n$ and assume $\mathbb{E}\left(\xi_{it}^2\right)$ is finite for all $1 \leq i \leq t$ and $1 \leq t \leq T$. Then,

$$
\Pr\left(\left| \sum_{t=1}^{T} \boldsymbol{\xi}_t \right|_{\infty} > Tx \right) \leq 2n \exp\left(-\frac{Tx^2}{2M^2 + xM} \right) + 4 \Pr\left(\max_{1 \leq t \leq T} |\xi_t|_{\infty} > M \right),
$$

for all $M > 0$.

Proof. Write $\boldsymbol{\xi}_t = \left(\xi_{1t}, \ldots, \xi_{nt}\right)'$. The proof follows after application of Fan *et al.* (2012a, Corollary 2.3).

DOI: 10.1111/jtsa.12627

Write $V_k^2(M) = \max_{1\le i\le n}\sum_{t=1}^{k}\mathbb{E}\left[\xi_{it}^2 I(\xi_{it} < M)|\mathcal{F}_t\right]$, $X_{ik} = \sum_{t=1}^{k}\xi_{it}$ and $X'_{ik}(M) = \sum_{t=1}^{k}\xi_{it}I(\xi_{it}\le M)$. It follows that for $v > 0$ and $x > 0$,

$$\begin{aligned}
\Pr\left(|\mathbf{X}_T|_\infty > x\right) &\le \Pr\left(\exists i,k\,:\, X_{ik} > x \cap V_k^2(M) \le v^2\right) + \Pr\left(V_T^2(M) > v^2\right)\\
&\le \Pr\left(\exists i,k\,:\, X'_{ik}(M) > x \cap V_k^2(M) \le v^2\right) + \Pr\left(V_T^2(M) > v^2\right)\\
&\quad + \Pr\left(\max_{1\le i\le n}\sum_{t=1}^{k}\xi_{it}I(\xi_{it} > M) > 0\right)\\
&\overset{(1)}{\le} n\exp\left(-\frac{(Tx/M)^2}{2\left((v/M)^2 + \frac{T}{3}x/M\right)}\right) + \Pr\left(V_n^2(M) > v^2\right)\\
&\quad + \Pr\left(\max_{1\le t\le T}|\boldsymbol{\xi}_t|_\infty > M\right)\\
&\overset{(2)}{\le} n\exp\left(-\frac{Tx^2}{2M^2 + Mx}\right) + 2\Pr\left(\max_{1\le t\le T}|\boldsymbol{\xi}_t|_\infty > M\right).
\end{aligned}$$

In (1) we use union bound and Fan *et al.* (2012a, Theorem 2.1) and in (2) we set $v^2 = T\left(M^2 + \frac{1}{6T}Mx\right)$ and the following:

$$\begin{aligned}
\Pr\left(V_T^2(M) > v^2\right) &\le \Pr\left(\max_{1\le i\le n}\sum_{t=1}^{T}\mathbb{E}\left[\xi_{it}^2 I\left(|\xi_{it}|\le M\right)|\mathcal{F}_t\right] \ge v^2\right)\\
&\quad + \Pr\left(\max_{1\le i\le n}\sum_{t=1}^{T}\mathbb{E}\left[\xi_{it}^2 I(\xi_{it} < -M)|\mathcal{F}_t\right] > 0\right)\\
&\le \Pr\left(\max_{1\le i\le n}\sum_{t=1}^{T}\mathbb{E}\left[\xi_{it}^2 I\left(|\xi_{it}|\le M\right)|\mathcal{F}_t\right] \ge T\left(M^2 + \frac{1}{6T}Mx\right)\right)\\
&\quad + \Pr\left(\max_{1\le t\le T}|\boldsymbol{\xi}_t|_\infty > M\right)\\
&\le \Pr\left(\max_{1\le t\le T}|\boldsymbol{\xi}_t|_\infty > M\right),
\end{aligned}$$

where in the last line we note that $\sum_{t=1}^{T}\mathbb{E}\left[\xi_{it}^2 I\left(|\xi_{it}|\le M\right)|\mathcal{F}_t\right] \le TM^2$.

Finally, write $\Pr\left(|\mathbf{X}_T|_\infty \ge Tx\right) = \Pr\left(\max_{i\le n}X_{iT} \ge Tx\right) + \Pr\left(\max_{i\le n}(-X_{iT}) \ge Tx\right)$ and apply above development in both terms. □

Corollary 1. Let $\left\{\boldsymbol{\xi}_t = \left(\xi_{1t},\dots,\xi_{nt}\right)'\right\}_{t\ge1}$ denote a multivariate martingale difference process with respect to the filtration $\mathcal{F}_t$ taking values on $\mathbb{R}^n$. Suppose that for each $\max_{i,t}\Pr\left(|\xi_{it}| > x\right) \le 2e^{-(x/\tau)^\alpha}$, for all $x > 0$, some $\alpha > 0$ and $\tau > 0$ Then,

$$\Pr\left(\left|\sum_{t=1}^{T}\boldsymbol{\xi}_t\right|_\infty > Tx\right) \le 2n\exp\left(-\frac{Tx^2}{2M^2 + xM}\right) + 8nT\exp\left(-\frac{M^\alpha}{\tau^\alpha}\right).$$

DOI: 10.1111/jtsa.12627

In particular, if $x > \tau\left(\epsilon + \log(nT)\right)^{1/\alpha}\sqrt{\epsilon + \log n}/\sqrt{T}$ and $T > (\epsilon + \log n)$ for any $\epsilon > 0$,

$$\Pr\left(\left|\sum_{t=1}^{T}\boldsymbol{\xi}_t\right|_{\infty} > Tx\right) \leq 10e^{-\epsilon}.$$

Proof. The first part we combine the union bound with assumption on ξ_{it}. In the second part, we will need the following bound. Let $0 < a < b/4 < \infty$. Then $\sqrt{a+b} - \sqrt{a} \geq \sqrt{b}\left(1 - 2\sqrt{a/b}\right)^{1/2}$. To verify that, first note that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, then $\left(\sqrt{a+b} - \sqrt{a}\right)^2 = 2a + b - 2\sqrt{a^2+ab} \geq b - 2\sqrt{ab} = b\left(1 - 2\sqrt{a/b}\right)$. Now, let $a = 1/T$ and $b = 8/(\epsilon + \log n)$ and verify that the choice $M = x\sqrt{T}/\sqrt{\epsilon + \log n}$ satisfy $\log n - \frac{Tx^2}{2M^2 + Mx} < -\epsilon$, then replace M and x to obtain the bound. □

B.3. Concentration Bound for Empirical Covariance Matrix

We derive concentration bound for $\|\boldsymbol{\Gamma}_T - \boldsymbol{\Gamma}\|_{\max}$, where $\boldsymbol{\Gamma}_T = \mathbf{X}'\mathbf{X}/T$ and $\boldsymbol{\Gamma} = \mathbb{E}\boldsymbol{\Gamma}_T$. We first split the problem into a sum of martingales and a tail dependence term. Then, we bound both individually.

Lemma 6. Suppose Assumptions (A1)–(A3) hold and

$$p < \frac{T^{\gamma_1 \wedge \gamma_2}}{\left(\frac{2}{\gamma_1 \wedge \gamma_2 + 1}\right)\left(2 + \frac{1.4}{2\gamma_1 c_\phi \wedge a_2}\right)}.$$

If for some $\xi > 0$

$$\epsilon^2 \geq \frac{2(1+\xi)^{1+2/\alpha}\tau^2[\log(npT)]^{1+2/\alpha}}{T},$$

then

$$\Pr\left(\|\boldsymbol{\Gamma}_T - \boldsymbol{\Gamma}\|_{\max} \geq \epsilon\right) \leq \frac{2}{(np)^\xi T^{1+\xi}} + \frac{8}{(np)^\xi T^\xi} + \frac{n^2}{\epsilon}\left(b_1 e^{-c_\phi \wedge a_2 (T/2)^{\gamma_2 \wedge \gamma_1}} + b_8 e^{-2\gamma_1 c_\phi (T/2)^{\gamma_1}}\right) \tag{B2}$$

where b_1, b_5 and τ are constants not depending on T.

Proof. Use the union bound to rewrite our probability bound in terms of $\mathbf{y}_{t-s}$:

$$\Pr\left(\left\|\boldsymbol{\Gamma}_T - \boldsymbol{\Gamma}\right\|_{\max} > \epsilon\right) \leq 2\sum_{r=0}^{p}\sum_{s=0}^{p-r}\Pr\left(\left\|\sum_{t=p+1}^{T}\mathbf{y}_{t-r}\mathbf{y}'_{t-r-s} - \mathbb{E}\left[\mathbf{y}_{t-r}\mathbf{y}'_{t-r-s}\right]\right\|_{\max} > T\epsilon\right).$$

Now, use a telescopic expansion of $\mathbf{y}_t\mathbf{y}'_{t-s}$ to obtain a sum of martingales and a dependence term:

$$\sum_{t=p+1}^{T}\mathbf{y}_t\mathbf{y}'_{t-s} - \mathbb{E}\left[\mathbf{y}_t\mathbf{y}'_{t-s}\right] = \underbrace{\sum_{t=p+1}^{T}\sum_{l=1}^{m}\mathbb{E}\left[\mathbf{y}_t\mathbf{y}'_{t-s}|\mathcal{F}_{t-l+1}\right] - \mathbb{E}\left[\mathbf{y}_t\mathbf{y}'_{t-s}|\mathcal{F}_{t-l}\right]}_{I_1}$$

DOI: 10.1111/jtsa.12627

$$+ \underbrace{\sum_{t=p+1}^{T} \mathbb{E}\left[\mathbf{y}_t \mathbf{y}'_{t-s} | \mathcal{F}_{t-m}\right] - \mathbb{E}\left[\mathbf{y}_t \mathbf{y}'_{t-s}\right]}_{I_2}$$
$$= I_1 + I_2.$$

Here,

$$I_1 = \sum_{l=1}^{m} \sum_{t=p+1}^{T} V_{l,t}^{(s)} \quad \text{and} \quad I_2 = \sum_{t=p+1}^{T} \mathbb{E}\left[\mathbf{y}_t \mathbf{y}'_{t-s} | \mathcal{F}_{t-m}\right] - \mathbb{E}\left[\mathbf{y}_t \mathbf{y}'_{t-s}\right]$$

where $\left\{V_{l,t}^{(s)}\right\}_t$, $l = 1, \ldots, m$, are sequences of martingale differences. The same decomposition holds for all terms $\mathbf{y}_{t-s}\mathbf{y}_{t-r-s}$. Then,

$$\begin{aligned}
\Pr\left(\left\|I_1 + I_2\right\|_{\max} > T\epsilon\right) &\leq \sum_{l=1}^{m} \Pr\left(\left\|\sum_{t=p+1}^{T} V_{l,t}^{(s)}\right\|_{\max} > \frac{T\epsilon}{2m}\right) \\
&\quad + \Pr\left(\left\|\sum_{t=p+1}^{T} \mathbb{E}\left[\mathbf{y}_t \mathbf{y}'_{t-s} | \mathcal{F}_{t-m}\right] - \mathbb{E}\left[\mathbf{y}_t \mathbf{y}'_{t-s}\right]\right\|_{\max} > \frac{T\epsilon}{2}\right) \\
&\leq 2mn^2 \exp\left(-\frac{T\epsilon^2}{2M^2 + M\epsilon}\right) \\
&\quad + 4mn^2 T \max_{l,t} \Pr\left(|V_{l,t}^{(s)}| > M\right) \qquad \text{(B3)} \\
&\quad + \frac{2}{T\epsilon}\mathbb{E}\left|\max_{1\leq i,j<n} \mathbb{E}|\sum_{t=p+1}^{T} \mathbb{E}\left[e_i' \mathbf{y}_t \mathbf{y}_{t-s} \boldsymbol{e}_j | \mathcal{F}_{t-m}\right] - \boldsymbol{e}_i' \mathbb{E}\left[\mathbf{y}_t' \mathbf{y}_{t-s}\right] \boldsymbol{e}_j\right|, \qquad \text{(B4)}
\end{aligned}$$

where $\boldsymbol{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)'$ is the ith canonical basis vector in $\mathbb{R}^n$.

Bounding the tail (B3):

The martingale differences $\left\{V_{t,l}^{(s)}\right\}$ ($l = 1, \ldots, n$ and $s = 0, \ldots, p$) are sub-Weibull with parameter $\alpha/2$. For any random variables (X, Y) and σ-algebras $\mathcal{F}$ and $\mathcal{G}$,

$$\|\mathbb{E}[XY|\mathcal{F}] - \mathbb{E}[XY|\mathcal{G}]\|_p \leq 2\|XY\|_p \leq 2\|Y^2\|_p^{1/2}\|X^2\|_p^{1/2}.$$

Therefore, it follows from Wong *et al.* (2020, Lemmas 5 and 6) that if both X and Y are sub-Weibull with parameter α, then XY is sub-Weibull with parameter $\alpha/2$. Therefore, there exists some τ^* such that $\Pr\left(|V_{t,l}^{(s)}| > s\right) \leq 2\exp\left(-|x/\tau^*|^{\alpha/2}\right)$, bounding (B3).

Bounding covariances (B4):

Now we move toward bounding the dependence term (B4).Write

$$\mathbf{y}_t \mathbf{y}'_{t-s} = \sum_{j=0}^{s-1} \boldsymbol{\Phi}_j \mathbf{u}_{t-j} \sum_{j=0}^{\infty} \mathbf{u}'_{t-s-j} \boldsymbol{\Phi}'_j + \sum_{j=0}^{\infty} \boldsymbol{\Phi}_{s+j} \mathbf{u}_{t-s-j} \sum_{j=0}^{\infty} \mathbf{u}'_{t-s-j} \boldsymbol{\Phi}'_j$$

DOI: 10.1111/jtsa.12627

$$
\begin{aligned}
&= \sum_{j=0}^{s-1} \mathbf{\Phi}_j \mathbf{u}_{t-j} \sum_{j=0}^{\infty} \mathbf{u}'_{t-s-j} \mathbf{\Phi}'_j \\
&+ \sum_{j=0}^{\infty} \mathbf{\Phi}_{j+s} \mathbf{u}_{t-s-j} \mathbf{u}_{t-s-j} \mathbf{\Phi}_j \\
&+ \sum_{k=1}^{\infty} \sum_{j=0}^{\infty} \mathbf{\Phi}_{j+s} \mathbf{u}_{t-s-j} \mathbf{u}_{t-s-j-k} \mathbf{\Phi}_{j+k} \\
&+ \sum_{k=1}^{\infty} \sum_{j=0}^{\infty} \mathbf{\Phi}_{j+s+k} \mathbf{u}_{t-s-j-k} \mathbf{u}_{t-s-j} \mathbf{\Phi}_j.
\end{aligned}
$$

It follows that $\mathbb{E}\left[\mathbf{y}_t \mathbf{y}'_{t-s}\right] = \sum_{j=0}^{\infty} \mathbf{\Phi}_{j+s} \mathbf{\Sigma} \mathbf{\Phi}_j$. Recall that $\mathcal{F}_{t-m} = \sigma\langle \mathbf{u}_{t-i} : i = m, m+1, \ldots \rangle$, then, for $m > s$,

$$
\begin{aligned}
\mathbb{E}\left[\mathbf{y}_t \mathbf{y}_{t-s} | \mathcal{F}_{t-m}\right] - \mathbb{E}\left[\mathbf{y}_t \mathbf{y}'_{t-s}\right] &= \sum_{j=0}^{m-s-1} \mathbf{\Phi}_j \mathbb{E}\left[\mathbf{u}_{t-s-j} \mathbf{u}'_{t-s-j} - \mathbf{\Sigma} | \mathcal{F}_{t-m}\right] \mathbf{\Phi}'_{j+s} \\
&+ \sum_{j=0}^{\infty} \mathbf{\Phi}_{m+j} \left(\mathbf{u}_{t-m-j} \mathbf{u}_{t-m-j} - \mathbf{\Sigma}\right) \mathbf{\Phi}_{m-s+j} \\
&+ \sum_{k=1}^{\infty} \sum_{j=0}^{\infty} \mathbf{\Phi}_{m+j} \mathbf{u}_{t-m-j} \mathbf{u}'_{t-m-j-k} \mathbf{\Phi}_{m-s+j+k} \\
&+ \sum_{k=1}^{\infty} \sum_{j=0}^{\infty} \mathbf{\Phi}_{m+j+k} \mathbf{u}_{t-m-j-k} \mathbf{u}'_{t-m-j} \mathbf{\Phi}_{m-s+j} \\
&= A_1(t,s,m) + A_2(t,s,m) + A_3(t,s,m) + A_4(t,s,m). \qquad \text{(B5)}
\end{aligned}
$$

We shall bound $\mathbb{E}\left|\sum_{r=0}^{p} \sum_{s=0}^{p-r} \boldsymbol{e}'_k A_i(t-r,s,m) \boldsymbol{e}_l\right|$ individually, for all $\{\boldsymbol{e}_i, i = 1, \ldots, n\}$ the canonical basis vector in $\mathbb{R}^n$.

(a) Bounding $\mathbb{E}\left|\sum_{r=0}^{p} \sum_{s=0}^{p-r} A_1(t-r,s,m)\right|$:

It follows from Assumption (A2) that for all $\mathbf{b}_1, \mathbf{b}_2 \in \left\{\mathbf{b} \in \mathbb{R}^n : |\mathbf{b}|_1 = 1\right\}$

$$
\max_t \mathbb{E}\left|\mathbb{E}\left[\mathbf{b}'_1 \left(\mathbf{u}_t \mathbf{u}'_t - \mathbf{\Sigma}\right)' \mathbf{b}_2 | \mathcal{F}_{t-m}\right]\right| \leq a_1 \exp(-a_2 m).
$$

Set $\{\boldsymbol{e}_i, i = 1, \ldots, n\}$ the canonical basis vector in $\mathbb{R}^n$. It follows from Assumptions (A1) and (A2) that for $j \leq m-s-1$:

$$
\begin{aligned}
&\mathbb{E}\left|\boldsymbol{e}'_k \mathbf{\Phi}_j \mathbb{E}\left[\left(\mathbf{u}_{t-s-j} \mathbf{u}'_{t-s-j} - \mathbf{\Sigma}\right) | \mathcal{F}_{t-m}\right] \mathbf{\Phi}'_{j+s} \boldsymbol{e}_l\right| \\
&\quad \leq |\boldsymbol{\phi}_{j,k}|_1 \left| |\boldsymbol{\phi}_{j+s,l}|_1 \max_t \mathbb{E}\mathbb{E}\left[\mathbf{b}'_1 \left(\mathbf{u}_{t-s-j} \mathbf{u}'_{t-s-j} - \mathbf{\Sigma}\right) \mathbf{b}_2 | \mathcal{F}_{t-m}\right]\right| \\
&\quad \leq \bar{c}_{\Phi} e^{-c_\phi (j+s)^{\gamma_1}} e^{-2c_\phi j^{\gamma_1}} \left[a_1 e^{-a_2 (m-j-s)^{\gamma_2}}\right]
\end{aligned}
$$

Let $0 < \gamma \leq 1$ and $\frac{m^\gamma}{p} > \left(\frac{2}{\gamma+1}\right)\left(2 + \frac{1.4}{c}\right)$ then $c(m-p)^\gamma - \log(p+1) \geq c(m/2)^\gamma$. Rewriting the inequality, we have to show that $(m/p-1)^\gamma - (m/2p)^\gamma > 2\log(p+1)/cp^\gamma$. In the LHS, a second order Taylor series expansion

DOI: 10.1111/jtsa.12627

yields $(2a-1)^\gamma - a^\gamma \geq \gamma\frac{a-1}{a}a^\gamma\left(1-\frac{1-\gamma}{2}\frac{a-1}{a}\right) \geq \frac{\gamma+1}{2}\frac{a-1}{a}a^\gamma$, for $a>1$. Set $a=m/2p$, so that $\frac{\gamma+1}{2}\left(\frac{m-2p}{m}\right)\left(\frac{m}{2p}\right)^\gamma \geq \log(p+1)/cp^\gamma$. As for the RHS, $\log(p+1)/p < 0.7$. Combining bounds above we show our claim.

Note that for $a,b \geq 0$ and $0<\gamma\leq 1$ $a^\gamma+b^\gamma = (a+b)^\gamma\,[x^\gamma+(1-x)^\gamma]$ where $x=a/(a+b)$, and $1 \leq [x^\gamma+(1-x)^\gamma] \leq 2$ for $x\in[0,1]$. Now, let $0<\gamma\leq 1$ and $c>0$, then $\sum_{j=0}^n e^{-cj^\gamma} \leq \int_0^n e^{-cx^\gamma}\mathrm{d}x = c^{1/\gamma}/\gamma\Gamma(1/\gamma,n) \uparrow c^{1/\gamma}\Gamma(1/\gamma+1) < \infty$, as $n\to\infty$, where $\Gamma(a,n)=\int_0^n x^{a-1}e^{-x}\mathrm{d}x \uparrow \int_0^\infty x^{a-1}e^{-x}\mathrm{d}x = \Gamma(a)$ are the incomplete gamma function and gamma function respectively.

Then,

$$\begin{aligned}
\mathbb{E}\left|\sum_{r=0}^{p}\sum_{s=0}^{p-r}\boldsymbol{e}_k'A_1(t-r,s,m)\boldsymbol{e}_l\right| &= \sum_{r=0}^{p}\sum_{s=0}^{p-r}\sum_{j=0}^{m-r-s-1}\mathbb{E}\left|\boldsymbol{e}_k'\boldsymbol{\Phi}_j\mathbb{E}\left[\mathbf{u}_{t-r-s-j}\mathbf{u}_{t-r-s-j}'-\boldsymbol{\Sigma}|\mathcal{F}_{t-m}\right]\boldsymbol{\Phi}_{j+s}'\boldsymbol{e}_l\right| \\
&\leq \bar{c}_\Phi a_1\sum_{r=0}^{p}\sum_{s=0}^{p-r}\sum_{j=0}^{m-r-s-1} e^{-c_\phi(j+s)^{\gamma_1}}e^{-2c_\phi j^{\gamma_1}}e^{-a_2(m-r-j-s)^{\gamma_2}} \\
&\leq \bar{c}_\Phi a_1\sum_{r=0}^{p} e^{-\left(c_\phi\wedge a_2\right)(m-r)^{\gamma_2\wedge\gamma_1}}\sum_{s=0}^{p-r}\sum_{j=0}^{m-r-s-1} e^{-c_\phi j^{\gamma_1}} \\
&\leq \bar{c}_\Phi a_1\frac{c_\phi^{1/\gamma_1}\Gamma\left(1/\gamma_1\right)}{2\gamma_1}(p+1)^2e^{-\left(c_\phi\wedge a_2\right)(m-p)^{\gamma_2\wedge\gamma_1}} \\
&\leq b_1e^{-\left(c_\phi\wedge a_2\right)(m/2)^{\gamma_2\wedge\gamma_1}}
\end{aligned}$$

(b) Bounding $\mathbb{E}\left|\sum_{r=0}^{p}\sum_{s=0}^{p-r}\boldsymbol{e}_k'A_2(t-r,s,m)\boldsymbol{e}_l\right|$:

Let $\max_{\mathbf{b}_1,\mathbf{b}_2,t}\mathbb{E}\left|\mathbf{b}_1'\left(\mathbf{u}_t\mathbf{u}_t'-\boldsymbol{\Sigma}\right)\mathbf{b}_2\right| \leq 2\Lambda_{\max}(\boldsymbol{\Sigma})$ where $\mathbf{b}_1,\mathbf{b}_2\in\left\{\mathbf{b}\in\mathbb{R}^n : |\mathbf{b}|_1=1\right\}$. It follows from Lemma 8 after rearranging terms:

$$\begin{aligned}
\mathbb{E}\left|\sum_{r=0}^{p}\sum_{s=0}^{p-r}\boldsymbol{e}_k'A_2(t-r,s,m)\boldsymbol{e}_l\right| &\leq \sum_{r=0}^{p}\sum_{s=0}^{p-r}\sum_{j=0}^{\infty}\mathbb{E}\left|\boldsymbol{e}_k'\boldsymbol{\Phi}_{m+j}\left(\mathbf{u}_{t-r-m-j}\mathbf{u}_{t-r-m-j}'-\boldsymbol{\Sigma}\right)\boldsymbol{\Phi}_{m-s+j}'\boldsymbol{e}_l\right| \\
&\leq \sum_{r=0}^{p}\sum_{s=0}^{p-r}\sum_{j=m}^{\infty}|\boldsymbol{\phi}_{j,k}|_1|\boldsymbol{\phi}_{j-s,l}|_1\max_{\mathbf{b}_1,\mathbf{b}_2,t}\mathbb{E}\left|\mathbf{b}_1'\left(\mathbf{u}_t\mathbf{u}_t'-\boldsymbol{\Sigma}\right)\mathbf{b}_2\right| \\
&\leq 2\Lambda_{\max}(\boldsymbol{\Sigma})\bar{c}_\Phi^2\sum_{r=0}^{p}\sum_{s=0}^{p-r}\sum_{j=m}^{\infty}e^{-c_\phi j^{\gamma_1}}e^{-c_\phi(j-r)^{\gamma_1}} \\
&= 2b_2\left(\sum_{r=0}^{p}\sum_{s=0}^{p-r}1\right)\sum_{j=m-p}^{\infty}e^{-2c_\phi(j-r)^{\gamma_1}} \\
&= 2b_2\left[(p+1)(m-p)^{1/2}e^{-c_\phi(m-p)^{\gamma_1}}\right]^2 \\
&\leq 2b_2e^{-c_\phi\left(\frac{m}{2}\right)^{\gamma_1}},
\end{aligned}$$

were $b_2=\Lambda_{\max}(\boldsymbol{\Sigma})\bar{c}_\Phi^2$. In the last line, recall that $\left(c_\phi/2\right)(m-p)^{\gamma_1}-\log(p+1) \geq (c_\phi/2)(m/2)^{\gamma_1}$, then

$$\begin{aligned}
(p+1)(m-p)^{1/2}e^{-c_\phi(m-p)^{\gamma_1}} &= (p+1)e^{\frac{1}{2}\log(m-p)-c_\phi(m-p)^{\gamma_1}} \\
&\leq e^{\log(p+1)-\frac{c_\phi}{2}\left(\frac{m}{2}\right)^{\gamma_1}} \leq e^{-\frac{c_\phi}{2}\left(\frac{m}{2}\right)^{\gamma_1}}.
\end{aligned}$$

DOI: 10.1111/jtsa.12627

(c) Bounding $\sum_{r=0}^{p}\sum_{s=0}^{p-r}A_j(t-r,s,m)$ $(j=3,4)$:

Under Assumption (A1)-(A3), for all $\mathbf{b}\in\mathbb{R}^n$ with $|\mathbf{b}|_1=1$,

$$\begin{aligned}\mathbb{E}|\boldsymbol{e}_r'\boldsymbol{\Phi}_{m+j}\mathbf{u}_{t-m-j}\mathbf{u}_{t-m-j-k}'\boldsymbol{\Phi}_{m-s+j+k}\boldsymbol{e}_s| &\le |\boldsymbol{\phi}_{m+j,r}|_1|\boldsymbol{\phi}_{m-s+j+k,s}|_1\max_t\|\mathbf{b}'\mathbf{u}_t\|_2^2\\ &\le b_2e^{-c_\phi(m+j)^{\gamma_1}-c_\phi(m-s+j+k)^{\gamma_1}},\end{aligned}$$

where $b_2=\bar{c}_\Phi^2\Lambda_{\max}(\boldsymbol{\Sigma})$. As before, if we have $\mathbf{u}_{t-r-s-j}$ we must replace m my $m-r$. It follows from Lemma 8, $\frac{m^{\gamma_1}}{p}>\left(\frac{2}{\gamma_1+1}\right)\left(2+\frac{1.4}{2\gamma_1c_\phi}\right)$ and m sufficiently large:

$$\begin{aligned}&\sum_{r=0}^{p}\sum_{j=0}^{\infty}e^{-c_\phi(m-r+j)^{\gamma_1}}\sum_{s=0}^{p-r}\sum_{k=0}^{\infty}e^{-c_\phi(m-r+j+k-s)^{\gamma_1}}\\ &\le\left(\sum_{r=0}^{p}\sum_{s=0}^{p-r}1\right)\sum_{j=0}^{\infty}e^{-c_\phi(m-p+j)^{\gamma_1}}\sum_{k=0}^{\infty}e^{-c_\phi(m-p+j+k)^{\gamma_1}}\\ &\le\frac{1}{2}(p+1)^2\sum_{j=m-p}^{\infty}\sum_{k=0}^{\infty}e^{-c_\phi j^{\gamma_1}-c_\phi(j+k)^{\gamma_1}}\\ &\le b_3\left[(p+1)(m-p)e^{-c_\phi(m-p)^{\gamma_1}}\right]^2\\ &\le b_3\left[e^{\log(p+1)-\frac{c_\phi}{2}(m-p)^{\gamma_1}}\right]^2\\ &\le b_3e^{-c_\phi\left(\frac{m}{2}\right)^{\gamma_1}},\end{aligned}$$

where $b_3=\left(1+\frac{2}{\gamma_1}\right)/2$.

Then, it follows that

$$\sum_{r=0}^{p}\sum_{s=0}^{p-r}\sum_{k=1}^{\infty}\sum_{j=0}^{\infty}\mathbb{E}|\boldsymbol{e}_i'\boldsymbol{\Phi}_{m+j}\mathbf{u}_{t-m-j}\mathbf{u}_{t-m-j-k}'\boldsymbol{\Phi}_{m-s+j+k}\boldsymbol{e}_l|\le b_3e^{-c_\phi\left(\frac{m}{2}\right)^{\gamma_1}},\tag{B6}$$

where $b_4=b_2b_3=\bar{c}_\Phi^2\Lambda_{\max}(\boldsymbol{\Sigma})\left(1+\frac{2}{\gamma_1}\right)/2$.

(d) Combining bounds:

Finally combining the three bounds above and setting m satisfying $\frac{m^{\gamma_1\wedge\gamma_2}}{p}>\left(\frac{2}{\gamma_1\wedge\gamma_2+1}\right)\left(2+\frac{1.4}{2\gamma_1c_\phi\wedge a_2}\right)$:

$$\begin{aligned}&\sum_{r=0}^{p}\sum_{s=0}^{p-r}\mathbb{E}\left|\max_{1\le i,j<n}\frac{1}{T}\mathbb{E}|\sum_{t=p+1}^{T}\mathbb{E}\left[\boldsymbol{e}_i'\mathbf{y}_{t-r}\mathbf{y}_{t-r-s}'\boldsymbol{e}_j|\mathcal{F}_{t-m}\right]-\mathbb{E}\left[\boldsymbol{e}_i'\mathbf{y}_{t-r}\mathbf{y}_{t-r-s}'\boldsymbol{e}_j\right]\right|\\ &\le n^2\left(b_1e^{-(c_\phi\wedge a_2)(m/2)^{\gamma_2\wedge\gamma_1}}+b_5e^{-c_\phi(m/2)^{\gamma_1}}\right),\end{aligned}\tag{B7}$$

where $b_5=2b_2+b_4=\bar{c}_\Phi^2\Lambda_{\max}(\boldsymbol{\Sigma})\left[2+\left(1+\frac{2}{\gamma_1}\right)/2\right]$

DOI: 10.1111/jtsa.12627

Combining tail and covariance:

Set $m = T$, then we require that

$$p < \frac{T^{\gamma_1 \wedge \gamma_2}}{\left(\frac{2}{\gamma_1 \wedge \gamma_2 + 1}\right)\left(2 + \frac{1.4}{2\gamma_1 c_\phi \wedge a_2}\right)}.$$

Then, combining bounds:

$$\begin{aligned}\Pr\left(\|\mathbf{\Gamma}_T - \mathbf{\Gamma}\|_{\max} \geq \epsilon\right) \leq\ & 2\frac{1}{T}\exp\left(2\log(npT) - \frac{T\epsilon^2}{2M^2 + M\epsilon}\right) \\ & + 8e^{2\log(npT) - M^\alpha/\tau^\alpha} \\ & + \frac{n^2}{\epsilon}\left(b_1 e^{-a_2 \wedge c_\phi (T/2)^{\gamma_2 \wedge \gamma_1}} + b_5 e^{-c_\phi (T/2)^{\gamma_1}}\right),\end{aligned}$$

where τ, b_1 and b_5 are as above. Let $\xi > 0$ and $M^\alpha = (2 + \xi)\tau^\alpha \log(npT)$, by assumption $T\epsilon^2 \geq 2(1 + \xi)^{1+2/\alpha}\tau^2[\log(npT)]^{1+2/\alpha}$ which implies that $2\log(npT) - \frac{T\epsilon^2}{2M^2 + M\epsilon} \leq -\xi \log(npT)$. Finally,

$$\Pr\left(\|\mathbf{\Gamma}_T - \mathbf{\Gamma}\|_{\max} \geq \epsilon\right) \leq \frac{2}{(np)^\xi T^{1+\xi}} + \frac{8}{(np)^\xi T^\xi} + \frac{n^2}{\epsilon}\left(b_1 e^{-a_2 \wedge c_\phi (T/2)^{\gamma_2}} + b_5 e^{-2\gamma_1 c_\phi (T/2)^{\gamma_1}}\right).$$

□

Lemma 7. Let X denote a positive random variable with $P(X \geq u) \leq e^{-cu^\alpha}$ for positive constants $c > 0$ and $\alpha > 0$. Then, for $M > 0$ and $p \geq 1$,

$$M^p P(X \geq M) \leq \mathbb{E}\left[X^p I(X \geq M)\right] \leq \left(1 + \frac{p}{\alpha}\right) M^p e^{-cM^\alpha}. \tag{B8}$$

In particular, let $P(X \geq u) = e^{-cu^\alpha}$. Then,

$$M^p e^{-cM^\alpha} \leq \mathbb{E}\left[X^p I(X \geq M)\right] \leq \left(1 + \frac{p}{\alpha}\right) M^p e^{-cM^\alpha}.$$

Proof. The lower bound follows trivially: $\mathbb{E}\left[X^p I(X \geq M)\right] \geq M^p P(X \geq M)$. Now, verify that $P(XI(X > M) >\geq u) = I(u < M^p)P(X \geq M) + I(u \geq M^p)P(X^p \geq u)$. We have

$$\begin{aligned}\mathbb{E}\left[X^p I(X \geq M)\right] &= \int_0^\infty P(XI(X > M) >\geq u)\,\mathrm{d}u \\ &= \int_0^{M^p} du P(X \geq M) + \int_{M^p}^\infty P\left(X^p \geq u\right)\mathrm{d}u \\ &\leq M^p e^{-cM^\alpha} + \int_{M^p}^\infty e^{-cu^{\alpha/p}}\mathrm{d}u \\ &= M^p e^{-cM^\alpha} + \frac{p}{\alpha}c^{-p/\alpha}\Gamma\left(p/\alpha, cM^\alpha\right) \\ &\leq \left(1 + \frac{p}{\alpha}\right) M^p e^{-cM^\alpha},\end{aligned}$$

where $\Gamma(s, x) = \int_x^\infty t^{s-1}e^{-t}\mathrm{d}t$ is the upper incomplete gamma function. □

DOI: 10.1111/jtsa.12627

Lemma 8. Let $0 < a \leq 1$ $b > 0$, $n \geq 1$. Then

$$\sum_{i=n}^{\infty}\sum_{j=0}^{\infty} e^{-bi^a - b(i+j)^a} \leq \left(2 + \frac{1}{a}\right) n^2 e^{-2bn^a}, \tag{B9}$$

and

$$\sum_{i=n}^{\infty} e^{-bi^a} \leq 2ne^{-bn^a}. \tag{B10}$$

Proof. Let V denote a Weibull$(a, 2b)$ random variable.

$$\begin{aligned}
\sum_{i=n}^{\infty}\sum_{j=0}^{\infty} e^{-bi^\alpha - b(i+j)^\alpha} &= \sum_{j=n}^{\infty}(j-n+1)e^{-2bj^\alpha} \\
&= \sum_{j=n}^{\infty}(j-n+1)\mathbb{E}[I(V \geq j)] \\
&= \sum_{j=n}^{\infty}(j-n+1)^2\mathbb{E}[I(j \leq V < j+1)] \\
&\leq \mathbb{E}[(V-n+1)^2 I(V \geq n)].
\end{aligned}$$

Expanding squares and using Lemma 7,

$$\begin{aligned}
\mathbb{E}\left[(V-n+1)^2 I(V \geq n)\right] &= \mathbb{E}[V^2 I(V \geq n)] - 2(n-1)\mathbb{E}[VI(V \geq n)] + (n-1)^2 P(V \geq n) \\
&\leq (1 + 2/a)n^2 e^{-2bn^a} - 2(n-1)ne^{-2bn^a} + (n-1)e^{-2bn^a} \\
&= (2/a)e^{-2bn^a} n^2 + e^{-2bn^a}\left[n - (n-1)\right]^2 \\
&\leq \left(1 + \frac{2}{a}\right) n^2 e^{-2bn^a}.
\end{aligned}$$

Similarly, let V be a Weibull(a, b) random variable,

$$\begin{aligned}
\sum_{i=n}^{\infty} e^{-bi^a} &\leq \mathbb{E}[(V-n+1)I(V > n)] \\
&\leq ne^{-bn^a} + (n-1)e^{-bn^a} \\
&< 2ne^{-bn^a}.
\end{aligned}$$

□

DOI: 10.1111/jtsa.12627