SUPPLEMENTARY TO “REGULARIZED ESTIMATION IN SPARSE HIGH-DIMENSIONAL TIME SERIES MODELS”

BY SUMANTA BASU[*] AND GEORGE MICHAILIDIS[*]

University of Michigan [*]

APPENDIX A: RESULTS FOR STOCHASTIC REGRESSION

PROOF OF PROPOSITION 3.1. Let us recall that $S = \mathcal{X}'\mathcal{X}/n$, $J = supp(\beta^*)$ with $|J| = k$, $\mathcal{C}(J,\kappa) = \{v \in \mathbb{R}^p : \|v_{J^c}\|_1 \le \kappa \|v_J\|_1\}$ and $\mathcal{K}(s) = \mathbb{B}_0(s) \cap \mathbb{B}_2(1)$, for any $s \ge 1$. We need a positive lower bound on $v'Sv/\|v\|^2$, uniformly over all $v \in \mathcal{C}(J,3)\backslash\{0\}$, that holds with high probability. Assuming $\|v\| = 1$ does not result in any loss of generality since $v \in \mathcal{C}(J,3)\backslash\{0\}$ if and only if $v/\|v\| \in \mathcal{C}(J,3)\backslash\{0\}$. If $\mathfrak{m}(f_X) > 0$, the lower bound in Proposition 2.3 ensures that

$$\inf_{v \in \mathcal{C}(J,3),\, \|v\|=1} v'\Gamma_X(0)v \ge 2\pi\mathfrak{m}(f_X) > 0 \tag{A.1}$$

So it remains to show that $v'(S - \Gamma_X(0))v$ is sufficiently small, uniformly for all $v \in \mathcal{C}(J,3)$ with $\|v\| = 1$. We start with the single deviation bound of (2.7) with a $2k$-sparse v, $\|v\| = 1$:

$$\mathbb{P}\left[\left|v'\left(S - \Gamma_X(0)\right)v\right| > 2\pi\mathcal{M}(f_X,2k)\eta\right] \le 2\exp\left[-cn\min\{\eta,\eta^2\}\right]$$

Using a discretization argument presented in Lemma F.2, we can extend it to the following uniform lower bound on all $2k$-sparse vectors v of unit norm:

$$\begin{aligned} &\mathbb{P}\left[\sup_{v \in \mathcal{K}(2k)} \left|v'\left(S - \Gamma_X(0)\right)v\right| > 2\pi\mathcal{M}(f_X,2k)\eta\right] \\ \le\ & 2\exp\left[-cn\min\{\eta,\eta^2\} + 2k\min\{\log p,\, \log(21ep/2k)\}\right] \end{aligned} \tag{A.2}$$

In the next step, we use Lemma F.1 to conclude that the set $\mathcal{C}(J,3) \cap \mathbb{B}_2(1)$ is contained in a closed, convex hull of k-sparse vectors $5cl\{conv\{\mathcal{K}(k)\}\}$. This, together with the approximation of Lemma F.3, leads to the following upper bound

$$\begin{aligned} \sup_{v \in \mathcal{C}(J,3),\, \|v\|=1} \left|v'(S - \Gamma_X(0))v\right| &\le \sup_{v \in 5cl\{conv\{\mathcal{K}(k)\}\}} \left|v'(S - \Gamma_X(0))v\right| \\ &= 25 \sup_{v \in cl\{conv\{\mathcal{K}(k)\}\}} \left|v'(S - \Gamma_X(0))v\right| \\ &\le 75 \sup_{v \in \mathcal{K}(2k)} \left|v'(S - \Gamma_X(0))v\right| \end{aligned}$$

Using the deviation bound of (A.2) and $\min\{\eta, \eta^2\} \geq \min\{1, \eta^2\}$, we have

$$\begin{aligned} &\mathbb{P}\left[\sup_{v \in \mathcal{C}(J,3),\, \|v\|=1} \left|v'(S - \Gamma_X(0))v\right| > 150\pi\mathcal{M}(f_X, 2k)\eta\right] \\ \leq\quad & 2\exp\left[-cn\min\{1,\eta^2\} + 2k\min\{\log p, \log(21ep/2k)\}\right] \end{aligned}$$

Setting $\eta = \mathfrak{m}(f_X)/150\mathcal{M}(f_X, 2k)$ and combining this deviation bound with (A.1), we obtain the final result.

Note that similar lower bounds can be derived if, instead of assuming $\mathfrak{m}(f_X) > 0$, one assumes $\Lambda_{\min}(\Gamma_X(0)) > 0$, or $\alpha := \inf_{v \in \mathcal{C}(J,3)\backslash\{0\}} v'\Gamma_X(0)v/\|v\|^2 > 0$. In these cases, $2\pi\mathfrak{m}(f_X)$ is replaced by $\Lambda_{\min}(\Gamma_X(0))$ or α in (A.1). □

PROOF OF PROPOSITION 3.2. We need an upper bound on $\|\mathcal{X}'E/n\|_\infty$ that holds with high probability. To this end, note that for any $j \in \{1, \ldots, p\}$, the deviation bound (2.10) applied on the processes $\{X_j^t\}$ and $\{\epsilon^t\}$ with $u = v = 1$ ensures

$$\mathbb{P}\left[|\mathcal{X}_j'E/n| > 2\pi(\mathcal{M}(f_{X_j}) + \mathcal{M}(f_\epsilon))\eta\right] \leq 6\exp\left[-cn\min\{\eta, \eta^2\}\right]$$

The cross-spectrum term vanishes since $\{X^t\}$ and $\{\epsilon^t\}$ are independent.

Taking an union bound over all j, we have:

$$\mathbb{P}\left[\max_{1\leq j\leq p}\frac{1}{n}\left|\mathcal{X}_j'E\right| > 2\pi\eta\left[\mathcal{M}(f_X,1) + \mathcal{M}(f_\epsilon)\right]\right] \leq 6p\exp\left[-cn\min\{\eta^2, \eta\}\right]$$

Setting $\eta = c_0\sqrt{\frac{\log p}{n}}$ and using the fact that $n \gtrsim \log p$, we have the required result. □

PROOF OF PROPOSITION 3.3. The events of Propositions 3.1 and 3.2 hold with probability $1 - c_1\exp[-c_2\log p]$ for some $c_i > 0$, under the assumptions on n and λ_n. Denote $v = \hat{\beta} - \beta^*$ and $J = supp(\beta^*)$ so that $|J| = k$. Then we have,

$$\frac{1}{n}\|Y - \mathcal{X}\hat{\beta}\|^2 + \lambda_n\|\hat{\beta}\|_1 \leq \frac{1}{n}\|Y - \mathcal{X}\beta^*\|^2 + \lambda_n\|\beta^*\|_1$$

After some algebra, this reduces to

$$v'Sv - \frac{2}{n}v'\left(\mathcal{X}'E\right) \leq \lambda_n\|\beta^*\|_1 - \lambda_n\|\beta^* + v\|_1$$

With the proposed choice of λ_n, we have

$$\begin{aligned} 0 \le v'Sv &\le \frac{\lambda_n}{2}\|v\|_1 + \lambda_n\|\beta^*\|_1 - \lambda_n\|\beta^* + v\|_1 \\ &\le \frac{\lambda_n}{2}\|v\|_1 + \lambda_n\left(\|\beta^*_J\|_1 - \|\beta^*_J + v_J\|_1 - \|v_{J^c}\|_1\right), \text{ since } \beta^*_{J^c} = 0 \\ &\le \frac{\lambda_n}{2}\left(\|v_J\|_1 + \|v_{J^c}\|_1\right) + \lambda_n\left(\|v_J\|_1 - \|v_{J^c}\|_1\right), \text{ by triangle inequality} \\ &\le \frac{3\lambda_n}{2}\|v_J\|_1 - \frac{\lambda_n}{2}\|v_{J^c}\|_1 \end{aligned}$$

This ensures $\|v_J\|_1 \le 3\|v_{J^c}\|_1$, i.e., $v \in \mathcal{C}(J,3)$ and $v'Sv \le 2\lambda_n\|v_J\|_1 \le 2\lambda_n\sqrt{k}\|v\|$. Using RE condition, we have

$$\alpha_{RE}\|v\|^2 \le v'Sv \le 2\lambda_n\sqrt{k}\|v\|$$

This implies

$$\begin{aligned} \|v\| &\le \frac{2\lambda_n\sqrt{k}}{\alpha_{RE}} \\ \|v\|_1 &\le 4\|v_J\|_1 \le 4\sqrt{k}\|v_J\| \le \frac{8\lambda_n k}{\alpha_{RE}} \\ \|v'Sv\| &\le \frac{4\lambda_n^2 k}{\alpha_{RE}} \end{aligned}$$

To derive the upper bound on the number of false positives selected by the thresholded lasso, note that

$$\begin{aligned} \left|supp(\tilde{\beta})\backslash supp(\beta^*)\right| &= \sum_{j\notin J}\mathbf{1}_{\{|\hat{\beta}_j|>\lambda_n\}} \le \sum_{j\notin J}\left|\hat{\beta}_j\right|/\lambda_n \\ &\le \frac{1}{\lambda_n}\sum_{j\notin J}|v_j| \le \frac{3}{\lambda_n}\sum_{j\in J}|v_j| \le \frac{3\,\|v\|_1}{\lambda_n} \le \frac{24k}{\alpha_{RE}} \end{aligned}$$

□

APPENDIX B: RESULTS ON VAR ESTIMATION

Proof of Proposition 4.1. Since $\hat{\beta}$ is a minimizer of (4.6), for all $\beta \in \mathbb{R}^q$ we have

$$-2\hat{\beta}'\hat{\gamma} + \hat{\beta}'\hat{\Gamma}\hat{\beta} + \lambda_N\|\hat{\beta}\|_1 \le -2\beta'\hat{\gamma} + \beta'\hat{\Gamma}\beta + \lambda_N\|\beta\|_1$$

For $\beta = \beta^*$, the above inequality reduces to

$$v'\hat{\Gamma}v \le 2v'(\hat{\gamma} - \hat{\Gamma}\beta^*) + \lambda_N\left\{\|\beta^*\|_1 - \|\beta^* + v\|_1\right\} \tag{B.1}$$

where $v = \hat{\beta} - \beta^*$.

The first term on the right hand side of (B.1) is at most $2\|v\|_1 \mathbb{Q}(\beta^*, \Sigma_\epsilon)\sqrt{\log q/N}$. The second term, by triangle inequality, is at most $\lambda_N\{\|v_J\|_1 - \|v_{J^c}\|_1\}$, where J denotes the support of β^*. Together with the proposed choice of λ_N, this leads to the following inequality

$$\begin{aligned} 0 \le v'\hat{\Gamma}v &\le \frac{\lambda_N}{2}\left\{\|v_J\|_1 + \|v_{J^c}\|_1\right\} + \lambda_N\left\{\|v_J\|_1 - \|v_{J^c}\|_1\right\} \\ &\le \frac{3\lambda_N}{2}\|v_J\|_1 - \frac{\lambda_N}{2}\|v_{J^c}\|_1 \le 2\lambda_N\|v\|_1 \end{aligned}$$

In particular, this ensures $\|v_{J^c}\|_1 \le 3\|v_J\|_1$ so that $\|v\|_1 \le 4\|v_J\|_1 \le 4\sqrt{k}\|v\|$. From the restricted eigenvalue assumption and the upper bound on $k\tau(N,q)$, we have

$$v'\hat{\Gamma}v \ge \alpha\,\|v\|^2 - \tau(N,q)\|v\|_1^2 \ge (\alpha - 16k\tau(N,q))\|v\|^2 \ge \frac{\alpha}{2}\|v\|^2$$

Together, the upper and lower bounds on $v'\hat{\Gamma}v$ guarantee that

$$\frac{\alpha}{4}\|v\|^2 \le \lambda_N\|v\|_1 \le 4\sqrt{k}\lambda_N\|v\|$$

This implies

$$\begin{aligned} \|v\| &\le 16\sqrt{k}\lambda_N/\alpha \\ \|v\|_1 \le 4\sqrt{k}\lambda_N\|v\| &\le 64k\lambda_N/\alpha \\ v'\hat{\Gamma}v \le 2\lambda_N\|v\|_1 &\le 128k\lambda_N^2/\alpha \end{aligned}$$

To derive the upper bound on the number of false positives selected by thresholded lasso, note that

$$\begin{aligned} \left|supp(\tilde{\beta})\backslash supp(\beta^*)\right| &= \sum_{j\notin J}\mathbf{1}_{\{|\hat{\beta}_j|>\lambda_N\}} \le \sum_{j\notin J}\left|\hat{\beta}_j\right|/\lambda_N \\ &\le \frac{1}{\lambda_N}\sum_{j\notin J}|v_j| \le \frac{3}{\lambda_N}\sum_{j\in J}|v_j| \le \frac{3\,\|v\|_1}{\lambda_N} \le \frac{192k}{\alpha} \end{aligned}$$

□

Proof of Proposition 4.2. Note that for ℓ_1-LS and ℓ_1-LL, the matrix $\hat{\Gamma}$ takes the form $I_p \otimes (\mathcal{X}'\mathcal{X}/N)$ and $\Sigma_\epsilon^{-1} \otimes (\mathcal{X}'\mathcal{X}/N)$, respectively. To prove that $\hat{\Gamma}$ satisfies RE, we first show that the random matrix $S = \mathcal{X}'\mathcal{X}/N$ satisfies RE(α, τ) with high probability, for some $\alpha > 0$, $\tau > 0$. Then we invoke Lemma B.1 to extend the result to $\hat{\Gamma}$.

To prove that $S = \mathcal{X}'\mathcal{X}/N$ satisfies RE condition, note that the rows of the design matrix $\mathcal{X}$ are generated according to a stable VAR(1) process $\{\tilde{X}^t\}$, as defined in (4.2). In particular, each row of $\mathcal{X}$ is centered Gaussian with covariance $\Gamma_{\tilde{X}}(0)$. Now $\Gamma_{\tilde{X}}(0) = \Upsilon_1^{\tilde{X}} = \Upsilon_d^X$, where Υ_n^X is the covariance of the vectorized data matrix containing n observations generated according to the process $\{X^t\}$, as defined in Section 2.3. Hence, from Proposition 2.3 and the bounds in (4.1), we have

$$\Lambda_{\min}\left(\Gamma_{\tilde{X}}(0)\right) \geq \frac{\Lambda_{\min}(\Sigma_\epsilon)}{\mu_{\max}(\mathcal{A})}$$

Also, from Proposition 2.4 and (4.1), we have, for any $v \in \mathbb{R}^{dp}$, $\|v\| \leq 1$, and any $\eta > 0$,

$$\mathbb{P}\left[\left|v'\left(S - \Gamma_{\tilde{X}}(0)\right)v\right| > \eta\frac{\Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\tilde{\mathcal{A}})}\right] \leq 2\exp\left[-cn\min\{\eta, \eta^2\}\right] \tag{B.2}$$

The next step is to extend the deviation bound (B.2) for a single v to an appropriate set of sparse vectors $\mathcal{K}(2s) := \{v \in \mathbb{R}^{dp} : \|v\| \leq 1, \|v\|_0 \leq 2s\}$, for an integer $s \geq 1$ to be specified later. Using the discretization argument of Lemma F.2, we have,

$$\mathbb{P}\left[\sup_{v \in \mathcal{K}(2s)}\left|v'\left(S - \Gamma_{\tilde{X}}(0)\right)v\right| > \eta\frac{\Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\tilde{\mathcal{A}})}\right]$$

is at most $2\exp\left[-cN\min\{\eta, \eta^2\} + 2s\min\{\log(dp), \log(21e\,dp/2s)\}\right]$.

Next, we set $\eta = \omega^{-1}$ with $c_3 = 54$ and note that $\min\{\eta, \eta^2\} \geq \min\{1, \eta^2\}$. Applying Supplementary Lemma 12 in Loh and Wainwright [2012] with $\delta = \Lambda_{\min}(\Sigma_\epsilon)/54\mu_{\max}(\mathcal{A})$ and $\Gamma = S - \Gamma_{\tilde{X}}(0)$, we have

$$v'Sv \geq \alpha\|v\|^2 - \frac{\alpha}{s}\|v\|_1^2, \quad \text{for all } v \in \mathbb{R}^{dp}$$

with probability at least $1 - 2\exp\left[-cN\min\{\omega^{-2}, 1\} + 2s\log(dp)\right]$.

Finally, we set $s = \lceil cN\,\min\{\omega^{-2}, 1\}/4\log(dp)\rceil$ [note that $s \geq 1$ with the required choice of N] to conclude that $S \sim RE(\alpha, \tau)$ with high probability. □

Lemma B.1 (RE condition for $\hat{\Gamma}$). *If $\mathcal{X}'\mathcal{X}/N \sim RE(\alpha, \tau)$, then so does $I_p \otimes \mathcal{X}'\mathcal{X}/N$. Further, if Σ_ϵ^{-1} satisfies $\bar{\sigma}_\epsilon^i := \sigma_\epsilon^{ii} - \sum_{j \neq i}\sigma_\epsilon^{ij} > 0$, for $i = 1, \ldots, p$, then*

$$\Sigma_\epsilon^{-1} \otimes \mathcal{X}'\mathcal{X}/N \sim RE\left(\alpha\,\min_i \bar{\sigma}_\epsilon^i, \tau\,\max_i \bar{\sigma}_\epsilon^i\right)$$

Proof. $S = \mathcal{X}'\mathcal{X}/N \sim RE(\alpha, \tau)$. Consider $\hat{\Gamma} = I_p \otimes S$. For any $\theta \in \mathbb{R}^{dp^2}$ with $\theta' = (\theta_1', \ldots, \theta_p')'$, each $\theta_i \in \mathbb{R}^{dp}$, we have

$$\theta'(I_p \otimes S)\theta = \sum_{r=1}^{p} \theta_r' S \theta_r \geq \alpha \sum_{r=1}^{p} \|\theta_r\|^2 - \tau \sum_{r=1}^{p} \|\theta_r\|_1^2 \geq \alpha\|\theta\|^2 - \tau\|\theta\|_1^2$$

proving the first part. To prove the second part, note that

$$\theta'(\Sigma_\epsilon^{-1} \otimes S)\theta = \sum_{r,s=1}^{p} \sigma_\epsilon^{rs} \theta_r' S \theta_s = \sum_{r=1}^{p} \sigma_\epsilon^{rr} \theta_r' S \theta_r + \sum_{r \neq s}^{p} \sigma_\epsilon^{rs} \theta_r' S \theta_s$$

Since the matrix S is non-negative definite, $\theta_r' S \theta_s \geq -\frac{1}{2}(\theta_r' S \theta_r + \theta_s' S \theta_s)$ for every $r \neq s$. This implies

$$\begin{aligned}
\theta'(\Sigma_\epsilon^{-1} \otimes S)\theta &\geq \sum_{r=1}^{p} \sigma_\epsilon^{rr} \theta_r' S \theta_r - \sum_{r<s} \sigma_\epsilon^{rs} (\theta_r' S \theta_r + \theta_s' S \theta_s) \\
&= \sum_{r=1}^{p} \left(\sigma_\epsilon^{rr} - \sum_{r \neq s} \sigma_\epsilon^{rs} \right) \theta_r' S \theta_r = \sum_{r=1}^{p} \bar{\sigma}_\epsilon^{r} \theta_r' S \theta_r \\
&\geq \alpha \sum_{r=1}^{p} \bar{\sigma}_\epsilon^{r} \|\theta_r\|^2 - \tau \sum_{r=1}^{p} \bar{\sigma}_\epsilon^{r} \|\theta_r\|_1^2 \\
&\geq \left(\alpha \min_i \bar{\sigma}_\epsilon^{i} \right) \|\theta\|^2 - \left(\tau \max_i \bar{\sigma}_\epsilon^{i} \right) \|\theta\|_1^2
\end{aligned}$$

□

Proof of Proposition 4.3. We want to establish an upper bound on $\|\hat{\gamma} - \hat{\Gamma}\beta^*\|_\infty$ that holds with high probability. To this end, first note that in the context of (4.6),

$$\begin{aligned}
\hat{\gamma} &= \left(W \otimes \mathcal{X}'\right)\left(I_p \otimes \mathcal{X}\right)\beta^*/N + \left(W \otimes \mathcal{X}'\right) vec(E)/N \\
\hat{\Gamma}\beta^* &= \left(W \otimes \mathcal{X}'\mathcal{X}/N\right)\beta^*
\end{aligned}$$

which implies $\hat{\gamma} - \hat{\Gamma}\beta^* = (W \otimes \mathcal{X}')\, vec(E)/N = vec(\mathcal{X}'EW)/N$.

Define, for every $h = 1, \ldots, d$, $\mathcal{X}_{(h)} = \left[X^{T-h} : \cdots : X^{d-h}\right]'$. We need to concentrate

$$\left\| \hat{\gamma} - \hat{\Gamma}\beta^* \right\|_\infty = \max_{\substack{1 \leq h \leq d \\ 1 \leq i, j \leq p}} \left| e_i' \left(\mathcal{X}_{(h)}' EW/N \right) e_j \right| \tag{B.3}$$

For ℓ_1-LS, $W = I$ and $EW = E$ is a data matrix from the process $\{\epsilon^t\}$. So, for a given h, i, j, the single deviation term on the right hand side of (B.3) can be concentrated using (2.11) with $u = e_i$, $v = e_j$. Setting $\eta = c_0\sqrt{\log q/N}$ and taking a union bound over the $q = dp^2$ possible choices of h, i, j yield the final result.

For ℓ_1-LL, $W = \Sigma_\epsilon^{-1}$ and we repeat the above argument for each of the dp^2 terms. Note that in this case, EW can be viewed as a data matrix from the process $\bar{\epsilon}^t := \Sigma_\epsilon^{-1}\epsilon^t$, a Gaussian white noise process with $\Gamma_{\bar{\epsilon}}(0) = \Sigma_\epsilon^{-1}$. As in the proof of (2.11) in Proposition 2.4, we will apply (2.10). To do this, we will need upper bounds on $\mathcal{M}(f_X)$, $\mathcal{M}(f_{\bar{\epsilon}})$ and $\mathcal{M}(f_{X^{t-h},\bar{\epsilon}^t})$. $\mathcal{M}(f_X)$ is at most $\Lambda_{\max}(\Sigma_\epsilon)/\mu_{\min}(\mathcal{A})$ and $\mathcal{M}(f_{\bar{\epsilon}})$ is $\Lambda_{\max}(\Sigma_\epsilon^{-1}) = 1/\Lambda_{\min}(\Sigma_\epsilon)$. To derive an upper bound on the third term, note that

$$f_{X^{t-h},\bar{\epsilon}^t}(\theta) = f_{X^{t-h},\epsilon^t}(\theta)\Sigma_\epsilon^{-1}$$

Therefore,

$$\mathcal{M}(f_{X^{t-h},\bar{\epsilon}^t})(\theta) \le \mathcal{M}(f_{X^{t-h},\epsilon^t})(\theta)/\Lambda_{\min}(\Sigma_\epsilon) \le \mathcal{M}(f_X)\mu_{\max}(\mathcal{A})/\Lambda_{\min}(\Sigma_\epsilon)$$

where the last inequality follows from the upper bound on $\mathcal{M}(f_{X^{t-h},\epsilon^t})$ derived in the proof of (2.11).

Applying (2.10) with $\eta = c_0\sqrt{\log q/N}$ and taking a union bound over the dp^2 choices of h, i, j yield the final deivation bound for ℓ_1-LL.

□

APPENDIX C: IMPLEMENTATION OF PENALIZED VAR

In this section, we discuss fast implementation strategies for the penalized VAR estimates proposed in Section 4. For ℓ_1-LL, one important step is the estimation of the error covariance Σ_ϵ. In the absence of any structural information, the residuals from an initial ℓ_1-LS fit can be used to estimate Σ_ϵ. However, it is also possible to incorporate *a priori* available structural information (sparsity, block structure etc.) on the residual network structure $\Sigma_\epsilon^{-1} = \left(\left(\sigma^{ij}\right)\right)$ into the model. For instance, the following optimization problem (henceforth refered as ℓ_1-ML)

$$\left(\hat{\beta}, \hat{\Sigma_\epsilon}^{-1}\right) = \underset{\beta\in\mathbb{R}^q, \Theta\in\mathbb{R}^{p\times p}}{\text{argmin}} \frac{1}{N}\left(Y - Z\beta\right)'\left(\Theta \otimes I\right)\left(Y - Z\beta\right) - \log\det(\Theta)$$
$$+\lambda_N \|\beta\|_1 + P_\gamma(\Theta)$$

maximizes the joint likelihood with respect to $(\beta, \Sigma_\epsilon^{-1})$ under an ℓ_1-regularization on β and some penalty P_γ on Σ_ϵ^{-1}. The second penalty can be chosen appropriately to incorporate some relevant structural assumption on Σ_ϵ^{-1}.

The optimization problem ℓ_1-LS in (4.3) can be expressed as p *separate* penalized regression problems:

$$\begin{aligned} &\underset{\beta\in\mathbb{R}^q}{\operatorname{argmin}} \frac{1}{N}\left\|Y - Z\beta\right\|^2 + \lambda_N \left\|\beta\right\|_1 \\ \equiv\ &\underset{B_1,\ldots,B_p}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^{p}\left\|\mathcal{Y}_i - \mathcal{X}\,B_i\right\|^2 + \lambda_N \sum_{i=1}^{p}\left\|B_i\right\|_1 \end{aligned}$$

This amounts to running p separate lasso programs, each with dp predictors: $\mathcal{Y}_i \sim \mathcal{X}$, $i = 1,\ldots,p$. For large d and p, the p programs can be solved in parallel.

In the optimization problem ℓ_1-LL, the above regressions are coupled through Σ_ϵ^{-1}. One way to solve the problem, as mentioned in Davis, Zang and Zheng [2012], is to reformulate it into a single penalized regression problem:

$$\begin{aligned} &arg \min_{\beta\in\mathbb{R}^q} \frac{1}{N}\left(Y - Z\beta\right)'\left(\Sigma_\epsilon^{-1}\otimes I\right)\left(Y - Z\beta\right) + \lambda_N\left\|\beta\right\|_1 \\ \equiv\ &arg \min_{\beta\in\mathbb{R}^q} \frac{1}{N}\left\|\left(\Sigma_\epsilon^{-1/2}\otimes I\right)Y - \left(\Sigma_\epsilon^{-1/2}\otimes\mathcal{X}\right)\beta\right\|^2 + \lambda_N\left\|\beta\right\|_1 \end{aligned}$$

This amounts to running a single lasso program with dp^2 predictors: $\left(\Sigma_\epsilon^{-1/2}\otimes I\right)Y \sim \Sigma_\epsilon^{-1/2}\otimes\mathcal{X}$. This is computationally expensive for large d and p. Unlike ℓ_1-LS, this algorithm is not parallelizable.

We propose an alternative algorithm based on block-wise coordinate descent to estimate the ℓ_1-LL and ℓ_1-ML coefficients. To this end, we first observe that the objective function in (4.4) can be simplified to

$$\frac{1}{N}\sum_{i=1}^{p}\sum_{j=1}^{p}\sigma_\epsilon^{ij}\left(\mathcal{Y}_i - \mathcal{X}B_i\right)'\left(\mathcal{Y}_j - \mathcal{X}B_j\right) + \lambda_N\sum_{k=1}^{p}\left\|B_k\right\|_1$$

Minimizing the above objective function cyclically with respect to each B_i leads to the following algorithm for ℓ_1-LL and ℓ_1-ML (the last step is used only for ℓ_1-ML):

1. pre-select d. Run ℓ_1-LS to get $\hat{B}$, $\hat{\Sigma}_\epsilon^{-1}$.
2. iterate till convergence:
   (a) For $i = 1,\ldots,p$,
      - set $r_i := (1/2\,\hat{\sigma}_\epsilon^{ii})\sum_{j\neq i}\hat{\sigma}_\epsilon^{ij}\left(\mathcal{Y}_j - \mathcal{X}\hat{B}_j\right)$
      - update $\hat{B}_i = \underset{B_i}{\operatorname{argmin}} \dfrac{\hat{\sigma}_\epsilon^{ii}}{N}\left\|(\mathcal{Y}_i + r_i) - \mathcal{X}B_i\right\|^2 + \lambda_N\left\|B_i\right\|_1$

(b) *[only for ℓ_1-ML]:* update $\hat{\Sigma}_\epsilon^{-1}$ as

$$\underset{\Theta \in \mathbb{R}^{p\times p}}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{p} \sum_{j=1}^{p} \Theta^{ij} \left(\mathcal{Y}_i - \mathcal{X}\hat{B}_i\right)' \left(\mathcal{Y}_j - \mathcal{X}\hat{B}_j\right) \\ - \log\det(\Theta) + P_\gamma(\Theta)$$

In this algorithm, a single iteration amounts to running p *separate* lasso programs, each with dp predictors: $\mathcal{Y}_i + r_i \sim \mathcal{X}$, $i = 1, \ldots, p$. As in ℓ_1-LS, these p programs can be solved in parallel. Further, for solving ℓ_1-ML, one can incorporate structural information about Σ_ϵ^{-1} using the penalty $P_\gamma(.)$. For instance, sparsity on Σ_ϵ^{-1} can be incorporated using $P_\gamma(\Theta) = \sum_{i\neq j} |\Theta|$, which amounts to running the popular graphical lasso procedure in step 2(b). *a priori* available knowledge of block structure on Σ_ϵ can be incorporated by choosing large penalties on the appropriate off-diagonal blocks.

APPENDIX D: RESULTS FOR COVARIANCE ESTIMATION

Proof of Proposition 5.1. The sample covariance matrix $\hat{\Gamma}(0)$ can be expressed as $\hat{\Gamma}(0) = S - \bar{X}\bar{X}'$ where $S = \mathcal{X}'\mathcal{X}/n$ and $\bar{X} = \mathcal{X}'\mathbf{1}/n$, $\mathbf{1}_{n\times 1} = (1, 1, \ldots, 1)'$. First, we derive element-wise concentration bound for $\hat{\Gamma}(0)$ around $\Gamma(0)$. To this end, note that for any $i, j \in \{1, \ldots, p\}$,

$$\left|\hat{\Gamma}_{ij}(0) - \Gamma_{ij}(0)\right| \leq |S_{ij} - \Gamma_{ij}(0)| + \left|\bar{X}_i\bar{X}_j\right|$$

Taking maximum over all i, j, we have

$$\max_{1\leq i,j\leq p} \left|\hat{\Gamma}_{ij}(0) - \Gamma_{ij}(0)\right| \quad \leq \quad \max_{1\leq i,j\leq p} |S_{ij} - \Gamma_{ij}(0)| + \max_{1\leq i\leq p} \left|\bar{X}_i\right|^2$$

Equation (2.9) provides a concentration bound on the first term. To concentrate the second term, note that $\bar{X}_i = \mathbf{1}'\mathcal{X}e_i/n$. Set $Y = \mathcal{X}e_i$. Then $Y_{n\times 1}$ can be viewed as the data matrix consisting of n observations from the i^{th} subprocess of $\{X^t\}$. Thus, $Y \sim N(0, Q)$ with $\|Q\| \leq 2\pi\mathcal{M}(f_X, 1)$, using Proposition 2.3. Now, for $Z = \mathbf{1}'Y/\sqrt{n}$, we have $\text{Var}(Z) = u'Qu \leq 2\pi\mathcal{M}(f_X, 1)$, since $u = \mathbf{1}/\sqrt{n}$ is an unit norm vector. Using this upper bound on $\text{Var}(Z)$ together with the standard Gaussian tail bound, we have, for any $\eta \geq 0$,

$$\begin{aligned}\mathbb{P}\left(|\bar{X}_i|^2 > 4\pi\mathcal{M}(f_X, 1)\eta\right) \quad &\leq \quad \mathbb{P}\left(|Z| > \sqrt{4\pi\mathcal{M}(f_X, 1)\eta}\sqrt{n}\right) \\ &\leq \quad 2\exp\left[-\frac{4\pi\mathcal{M}(f_X, 1)\eta n}{2\text{Var}(Z)}\right] \leq 2\exp\left[-n\min\{\eta, \eta^2\}\right]\end{aligned}$$

Combining the concentration bounds for the two terms and setting $\eta = \sqrt{\frac{\log p}{n}} = o_P(1)$, we conclude

$$\max_{i,j} \left| \hat{\Gamma}_{ij}(0) - \Gamma_{ij}(0) \right| = O_P\left(\mathcal{M}(f_X, 1)\sqrt{\frac{\log p}{n}} \right)$$

This provides element-wise concentration bounds similar to equation (12) in Bickel and Levina [2008]. The remainder of the proof follows exactly along the lines of Theorems 1 and 2 in that paper. □

APPENDIX E: MEASURE OF STABILITY

In this section, we discuss the connection of the proposed dependence condition 2.1 with existing ones and establish some properties of the stability measure introduced in Section 2.

Quantifying Dependence using Spectral Density. The assumption 2.1 is motivated by the connection of the time domain and spectral domain representations of stationary processes. Meaningful statistical inference with stationary processes in time domain is generally undertaken under some decay conditions on the temporal dependence. An analogous treatment for univariate processes in the frequency domain requires some nice property (boundedness, continuity, smoothness etc.) of the underlying spectra [Giraitis, Koul and Surgailis, 2012]. For analyzing multivariate processes in high-dimension, assumption 2.1 requires existence and boundedness of the spectra.

As shown in Section 2, this assumption is satisfied by the popular vector-valued ARMA models. More generally, consider a (possibly non-causal) general linear process $X^t = \sum_{l=-\infty}^{\infty} K^l \epsilon^{t-l}$, $t \in \mathbb{Z}$ with a white noise process ϵ^t and a transfer function $\mathcal{K}(z) = \sum_{l=-\infty}^{\infty} K^l z^l$, $z \in \mathbb{C}$. This process satisfies assumption 2.1 as long as $\sum_{l=-\infty}^{\infty} \|K^l\| < \infty$, i.e., the transfer function is stable.

Mixing Condition. For univariate processes, Bradley et al. [2005] provide an excellent review of the popular mixing conditions and their implications in many commonly used stochastic processes. In particular, for a strongly mixing stationary, centered Gaussian process, Theorem 7.1 in the above paper asserts that the spectral density has the following form

$$f(\theta) = \left| p(e^{i\theta}) \right|^2 \exp\left[u(e^{i\theta}) + \tilde{v}(e^{i\theta}) \right]$$

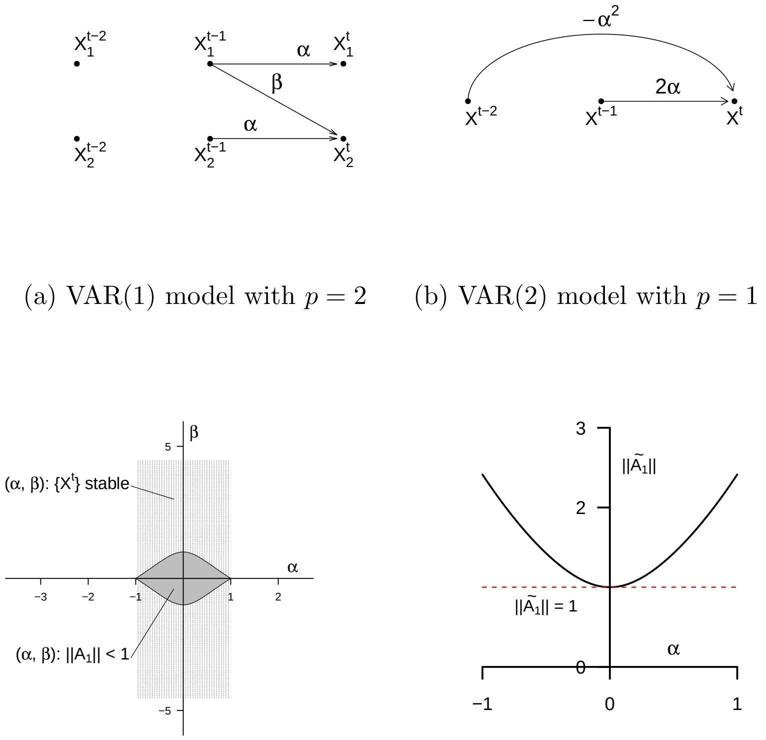
where u and v are continuous real functions and $\tilde{v}$ is the harmonic conjugate (Hilbert transform) of v. For a more detailed discussion of different necessary and sufficient conditions for strongly mixing Gaussian processes, we refer the reader to the above paper and the references therein. In short, these conditions impose some continuity or smoothness assumption on the underlying spectra, but do not require any boundedness. For instance, [Ibragimov, 1965] shows that a strongly mixing Gaussian process can not have jump discontinuities in its spectra. On the other hand, our assumption allows such discontinuities, but it is violated if the spectrum is unbounded while mixing conditions do not impose such a restriction.

For multivariate processes, theoretical results connecting spectral properties with the mixing properties of stationary processes are sparse in the literature. We refer the reader to [Cheng and Pourahmadi, 1993] for some results connecting strong mixing with smoothness of the spectrum under an assumption of a bounded spectrum similar to Assumption 2.1.

Functional Dependence Measure. [Wu, 2005, 2011] demonstrate how the functional and predictive dependence measures behave for some commonly used classes of stationary processes. Chen et al. [2013] derive asymptotic theory assuming a short range dependence condition of absolute summability of the functional dependence measure. For the class of linear Gaussian processes, this is similar to the assumption of absolute summability of autocovariance functions, under which our dependence condition 2.1 is also satisfied.

For multivariate AR(1) processes, the above papers show that the dependence measures scale as $O(\rho(A))$. For multivariate stationary linear processes, the decay assumption of the functional dependence measure required for theory [Chen et al., 2013] is verified under another decay assumption on the transition matrices. Our assumption relies on the stability of the process. For stable AR(1) and linear processes, this assumption is automatically always satisfied.

The assumption $\|A\| < 1$ for Gaussian VAR(1) processes. The assumption $\|A\| < 1$ was used in the recent literature [Loh and Wainwright, 2012, Negahban and Wainwright, 2011] to measure the temporal and cross-sectional dependence of Gaussian VAR(1). We show that the assumption $\|A_1\| < 1$ guarantees stability of the process, but not the other way. If, however, the transition matrix A_1 is symmetric, the assumption $\|A_1\| < 1$ is necessary for stability. We also show that this assumption is violated for all stable VAR(d) models, whenever $d > 1$. We conclude the section with the proof of Proposition 2.2, where we derive upper and lower bounds on the quantities $\mu_{\min}(\mathcal{A})$ and $\mu_{\max}(\mathcal{A})$. Finally, we illustrate the stringency of the assumption $\|A\| < 1$ on two separate classes of stable VAR models in Figure 1.



(a) VAR(1) model with $p = 2$ (b) VAR(2) model with $p = 1$

(c) Stability and $\|A_1\| < 1$ (d) Stability and $\|\tilde{A}_1\| < 1$

Fig 1: In the left panel, we consider a VAR(1) model with $p = 2$, $X^t = A_1 X^{t-1} + \epsilon^t$, where $A_1 = [\alpha\ 0; \beta\ \alpha]$. The unbounded set (dotted) denotes the values of (α, β) for which the process is stable. The bounded region (solid) represents the VAR models that satisfy $\|A_1\| < 1$. In the right panel, we consider a VAR(2) model with $p = 1$, $X^t = 2\alpha X^{t-1} - \alpha^2 X^{t-2} + \epsilon^t$. Equivalent formulation of this model as VAR(1) is: $Y^t = \tilde{A}_1 Y^{t-1} + \tilde{\epsilon}^t$, where $Y^t = [X^t, X^{t-1}]'$, $\tilde{A}_1 = [2\alpha\ \ -\alpha^2; 1\ 0]$, and $\tilde{\epsilon}^t = [\epsilon^t, 0]'$. The model is stable whenever $|\alpha| < 1$ but $\|\tilde{A}_1\|$ is always greater than or equal to 1.

LEMMA E.1. *A VAR(1) process is stable if $\|A_1\| < 1$. If A_1 is symmetric, then a VAR(1) process is stable only if $\|A_1\| < 1$.*

PROOF. If $\|A_1\| < 1$, then all the eigenvalues of A_1 lie inside the unit circle $\{z \in \mathbb{C} : |z| \leq 1\}$. So the process is stable.

If the process is stable, then all the eigenvalues of A_1 lie inside the unit circle. In addition, if A_1 is symmetric, then this implies that $\|A_1\| = \sqrt{\Lambda_{\max}(A_1' A_1)} < 1$. □

LEMMA E.2. *Consider the VAR(1) representation of a VAR(d) process in (4.2). $\|\tilde{A}_1\| \nless 1$ whenever $d > 1$.*

PROOF.

$$\tilde{A}_1 \tilde{A}_1' = \begin{bmatrix} \sum_{t=1}^d A_t A_t' & A_1 & \dots & A_{d-1} \\ A_1' & I_p & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ A_{d-1}' & \mathbf{0} & \dots & I_p \end{bmatrix}_{dp \times dp}$$

So for any $v \in \mathbb{R}^{dp}$ with $v' = (v_1', \dots, v_d')$, each $v_t \in \mathbb{R}^p$, we have

$$v' \tilde{A}_1 \tilde{A}_1' v = v_1' \left(\sum_{t=1}^d A_t A_t' \right) v_1 + 2 v_1' \sum_{t=2}^d A_{t-1} v_t + \sum_{t=2}^d v_t^2$$

This implies

$$\Lambda_{\max}\left(\tilde{A}_1 \tilde{A}_1'\right) = \max_{\|v\|=1} v' \tilde{A}_1 \tilde{A}_1' v \geq \max_{\substack{\|v\| = 1 \\ v_1 = \mathbf{0}}} v' \tilde{A}_1 \tilde{A}_1' v = \max_{\substack{\|v\| = 1 \\ v_1 = \mathbf{0}}} \sum_{t=2}^d v_t^2 = 1$$

□

PROOF OF PROPOSITION 2.2. $\mathcal{A}(z) = I_p - A_1 z - A_2 z^2 - \dots - A_d z^d$

(i) Using $|z| = 1$ together with the matrix norm inequality $\|A\| \leq \|A\|_1 \|A\|_\infty$, we have

$$\begin{aligned} \sqrt{\mu_{\max}(\mathcal{A})} &= \max_{|z|=1} \left\| I - A_1 z - \dots - A_d z^d \right\| \\ &\leq 1 + \sum_{h=1}^d \|A_h\| \leq 1 + \sum_{h=1}^d \sqrt{\|A_h\|_1 \|A_h\|_\infty} \\ &\leq 1 + \sum_{h=1}^d \left(\max_{1 \leq i \leq p} \sum_{j=1}^p |A_{h,ij}| + \max_{1 \leq j \leq p} \sum_{i=1}^p |A_{h,ij}| \right) /2 \end{aligned}$$

(ii) For $d = 1$, $\mathcal{A}(z) = I_p - A_1 z$. First note that

$$\mu_{\min}(\mathcal{A}) \;=\; \min_{|z|=1} \Lambda_{\min}\left((I - A_1 z)^*(I - A_1 z)\right) = \min_{|z|=1} \Lambda_{\min}\left((zI - A_1)^*(zI - A_1)\right)$$

If A_1 is diagonalizable with eigenvalues $\lambda_1, \ldots, \lambda_p$ and corresponding eigenvectors $w_1, \ldots, w_p$, we have the decomposition $A_1 = PDP^{-1}$, where D is a diagonal matrix with entries λ_i and $P = [w_1 : \ldots : w_p]$. So, $zI - A_1 = PD_zP^{-1}$, where D_z is diagonal with entries $(z - \lambda_i)$, $i = 1, \ldots, p$. The condition $det(\mathcal{A}(z)) \neq 0$ ensures all the eigenvalues of A_1 are inside the unit circle $\{z \in \mathbb{C} : |z| = 1\}$. This implies D_z is invertible, for all $|z| = 1$ and the eigenvalues of $D_z^* D_z$ are $|z - \lambda_i|^2 \geq (1 - \rho(A_1))^2$, for all $|z| = 1$ and $i = 1, \ldots, p$. Hence,

$$\begin{aligned} \mu_{\min}(\mathcal{A}) &= \min_{|z|=1} \left[\left\|PD_z^{-1}P^{-1}(P')^{-1}(D_z^*)^{-1}P'\right\|\right]^{-1} \\ &\geq \|P\|^{-2}\|P^{-1}\|^{-2}\left(1 - \rho(A_1)\right)^{-2} \end{aligned}$$

□

APPENDIX F: AUXILIARY LEMMAS

LEMMA F.1 (Approximating cone sets by sparse sets). *For any $S \subset \{1, \ldots, p\}$ with $|S| = s$ and $\kappa > 0$,*

$$\mathcal{C}(S, \kappa) \cap \mathbb{B}_2(1) \subseteq \mathbb{B}_1((\kappa + 1)\sqrt{s}) \cap \mathbb{B}_2(1) \subseteq (\kappa + 2)cl\{conv\{\mathcal{K}(s)\}\}$$

PROOF. The first inequality follows from the fact that for any $v \in \mathcal{C}(S, \kappa)$,

$$\|v\|_1 = \|v_S\|_1 + \|v_{S^c}\|_1 \leq (\kappa + 1)\|v_S\|_1 \leq (\kappa + 1)\sqrt{s}\|v_S\| \leq (\kappa + 1)\sqrt{s}$$

Both $A := \mathbb{B}_1((\kappa+1)\sqrt{s}) \cap \mathbb{B}_2(1)$ and $B := (\kappa+2)cl\{conv\{\mathcal{K}(s)\}\}$ are closed convex sets. We will show that the support function of A is dominated by the support function of B.

The support function of A is $\phi_A(z) = \sup_{\theta \in A}\langle \theta, z\rangle$. For a given $z \in \mathbb{R}^p$, let J denote the set of coordinates of z with the s largest absolute values, so that $||z_{J^c}||_\infty \leq ||z_J||_1/s \leq ||z_J||/\sqrt{s}$. Also note that for any $\theta \in A$, $\|\theta_{J^c}\|_1 \leq (\kappa + 1)\sqrt{s}$. Then we have, for any $\theta \in A$, $z \in \mathbb{R}^p$,

$$\langle \theta, z\rangle = \sum_{i \in J^c} \theta_i z_i + \sum_{i \in J} \theta_i z_i \leq ||z_{J^c}||_\infty \|\theta_{J^c}\|_1 + \|z_J\|\|\theta_J\| \leq (\kappa + 1)\|z_J\| + \|z_J\|$$

so that $\phi_A(z) \leq (\kappa + 2)\|z_J\|$.

On the other hand, $\phi_B(z) := \sup_{\theta \in B}\langle \theta, z\rangle = \sup_{|U|=s} \sum_{i \in U} \theta_i z_i = (\kappa + 2)\|z_J\|$. □

LEMMA F.2. *Consider a symmetric matrix $D_{p\times p}$. If, for any vector $v \in \mathbb{R}^p$ with $\|v\| \le 1$, and any $\eta \ge 0$,*

$$\mathbb{P}\left[\left|v'Dv\right| > C\eta\right] \le 2\exp\left[-cn\min\{\eta,\eta^2\}\right]$$

then, for any integer $s \ge 1$, we have

$$\mathbb{P}\left[\sup_{v\in\mathcal{K}(s)}\left|v'Dv\right| > C\eta\right] \le 2\exp\left[-cn\min\{\eta,\eta^2\} + s\min\{\log p,\, \log\left(21ep/s\right)\}\right]$$

PROOF. Choose $U \subset \{1,\ldots,p\}$ with $|U| = s$. Define $S_U = \{v \in \mathbb{R}^p : \|v\| \le 1,\, supp(v) \subseteq U\}$. Then $\mathcal{K}(s) = \cup_{|U|\le s} S_U$. Choose $\mathcal{A} = \{u_1,\ldots,u_m\}$, a $1/10$-net of S_U. By Lemma 3.5 of Vershynin [2009], $|\mathcal{A}| \le 21^s$. For every $v \in S_U$, there exists some $u_i \in \mathcal{A}$ such that $\|\Delta v\| \le 1/10$, where $\Delta v = v - u_i$. Then we have,

$$\gamma := \sup_{v\in S_U}\left|v'Dv\right| \le \max_i\left|u_i'Du_i\right| + 2\sup_{v\in S_U}\left|\max_i u_i'D(\Delta v)\right| + \sup_{v\in S_U}\left|(\Delta v)'D(\Delta v)\right|$$

Since $10(\Delta v) \in S_U$, the third term is bounded above by $\gamma/100$. The second term is bounded above by $6\gamma/10$, as shown below:

$$\begin{aligned}2\sup_{v\in S_U}\left|\max_i u_i'D(\Delta v)\right| &\le \frac{1}{10}\sup_{v\in S_U}\left|(u_i+10\Delta v)'D(u_i+10\Delta v)\right| \\ &\quad + \frac{1}{10}\sup_{v\in S_U}\left|u_i'Du_i\right| + \frac{1}{10}\sup_{v\in S_U}\left|(10\Delta v)'D(10\Delta v)\right| \\ &\le \frac{4\gamma}{10} + \frac{\gamma}{10} + \frac{\gamma}{10}\end{aligned}$$

Readjusting, we have $\gamma \le 3\max_i\left|u_i'Du_i\right|$. Taking an union bound over all $u_i \in \mathcal{A}$, we have

$$\mathbb{P}\left[\sup_{v\in S_U}\left|v'Dv\right| > 3C\eta\right] \le 2\exp\left[-cn\min\{\eta,\eta^2\} + s\log(21)\right]$$

Taking another union bound over $\binom{p}{s} \le \min\{p^s, (ep/s)^s\}$ choices of U, we obtain the required result. □

LEMMA F.3.

$$\sup_{v\in cl\{conv\{\mathcal{K}(s)\}\}}\left|v'Dv\right| \le 3\sup_{v\in\mathcal{K}(2s)}\left|v'Dv\right|$$

PROOF. Let $v \in conv\{\mathcal{K}(s)\}$. Then $v = \sum_{i=1}^{k} \alpha_i v_i$, for some $k \geq 1$, $v_i \in \mathcal{K}(s)$ and $0 \leq \alpha_i \leq 1$, for all $1 \leq i \leq k$, such that $\sum_i \alpha_i = 1$. Then

$$\begin{aligned} 2\left|v'Dv\right| &\leq 2\sum_{i,j=1}^{k} \alpha_i\alpha_j \left|v_i'Dv_j\right| \\ &\leq \sum_{i,j=1}^{k} \alpha_i\alpha_j \left[\left|(v_i+v_j)'D(v_i+v_j)\right| + \left|v_i'Dv_i\right| + \left|v_j'Dv_j\right|\right] \\ &\leq 6\sum_{i,j=1}^{k} \alpha_i\alpha_j \sup_{v \in \mathcal{K}(2s)} \left|v'Dv\right| \end{aligned}$$

By the continuity of quadratic forms, the result follows. □

REFERENCES

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008 (2010b:62197)

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

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

Cheng, R. and Pourahmadi, M. (1993). The mixing rate of a stationary multivariate process. *Journal of Theoretical Probability* **6** 603–617.

Davis, R. A., Zang, P. and Zheng, T. (2012). Sparse Vector Autoregressive Modeling. *ArXiv e-prints.*

Giraitis, L., Koul, H. L. and Surgailis, D. (2012). *Large sample inference for long memory processes.* Imperial College Press London.

Ibragimov, I. (1965). On The Spectrum Of Stationary Gaussian Sequences Satisfying the Strong Mixing Condition I. Necessary Conditions. *Theory of Probability & Its Applications* **10** 85-106.

Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity. *Ann. Stat.* **40** 1637-1664.

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.* **39** 1069–1097. MR2816348 (2012f:62046)

Vershynin, R. (2009). *Lectures in Geometric Functional Analysis.* available at http://www-personal.umich.edu/ romanv/papers/GFA-book/GFA-book.pdf.

Wu, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences of the United States of America* **102** 14150-14154.

Wu, W. B. (2011). Asymptotic theory for stationary processes. *Statistics and Its Interface,* *0* 1–20.

Sumanta Basu and George Michailidis
Department of Statistics
University of Michigan
Ann Arbor MI 48109
E-mail: sumbose@umich.edu
gmichail@umich.edu