Fundamental Estimation in Autoregressive Processes with Compressive Measurements: Proofs

Milind Rao, Tara Javidi, Yonina Eldar, Andrea Goldsmith

Appendix A

In this section, we focus on showing how VAR processes can be reconstructed from low dimensional random projections.

At each point in time for state x_t, we are given two random low dimensional projections $z_t^i = \Psi_t^i x_t,\ i \in \{1,2\}$ where the rank of matrix Ψ_t^i is $m \ll n$.

We saw that we could write $z_t^1 = \eta_t x_t + \sqrt{\eta_t(1-\eta_t)} R_t^1 x_t$ and $z_t^2 = \omega_t x_t + \sqrt{\omega_t(1-\omega_t)} R_t^2 x_t$. Here R_t^i are rotation matrices that are uniformly distributed on the hypersphere and perpendicular to x_t. $\eta_t, \omega_t \overset{\text{iid}}{\sim} \text{Beta}\left(\frac{m}{2}, \frac{n-m}{2}\right)$.

Consider the estimate of the covariance matrix,

$$\begin{aligned}
\hat{\Sigma}^k &= \frac{n^2}{(T-k)m^2} \sum_{t=1}^{T-k} z_t^1 z_{t+k}^{2\intercal} \\
&= \frac{n^2}{(T-k)m^2} \sum_{t=1}^{T-k} \eta_t \omega_{t+k} x_t x_{t+k}^\intercal + \\
&\sqrt{\eta_t \omega_{t+k}(1-\omega_{t+k})(1-\eta_t)} R_t^1 x_t x_{t+k} R_{t+k}^{2\intercal} + \\
&\sqrt{\eta_t(1-\eta_t)}\omega_{t+k} R_t^1 x_t x_{t+k}^\intercal + \\
&\sqrt{\omega_{t+k}(1-\omega_{t+k})}\eta_t x_t x_{t+k}^\intercal R_{t+k}^{2\intercal} \\
&= P_1 + P_2 + P_3 + P_4
\end{aligned}$$

It can be seen that,

$$\begin{aligned}
\mathbb{E}[P_1] &= \mathbb{E}\left[\frac{1}{T-k}\sum_{t=1}^{T-k} x_t x_{t+k}^\intercal\right] \\
\mathbb{E}[P_2] &= \mathbb{E}[P_3] = \mathbb{E}[P_4] = 0
\end{aligned}$$

The former is because $\mathbb{E}[\eta_t] = \mathbb{E}[\omega_t] = m/n$ and the latter is because R_t^i is a symmetric random rotation matrix.

The difference between the mean of term P_1 and the true covariance matrices is bounded as

$$\|\mathbb{E}[P_1] - \Sigma^k\|_2 \leq \frac{\sigma_{\max}^k}{(1-\sigma_{\max}^2)(T-k)} \frac{\|Q_w\|_2}{(1-\sigma_{\max}^2)}.$$

This is because $\Sigma^k = \mathbb{E}[x_t x_{t+k}^\intercal] = \left(\sum_{i=0}^\infty A^i Q_w A^{i\intercal}\right) A^{k\intercal}$.

$$\begin{aligned}
\mathbb{E}[P_1] &= \mathbb{E}\left[\frac{1}{T-k}\sum_{t=1}^{T-k} x_t x_{t+k}^\intercal\right] \\
&= \frac{1}{T-k}\sum_{t=1}^{T-k}\sum_{i=0}^{t-1} A^i Q_w A^{i+k\intercal}
\end{aligned}$$

$$\begin{aligned}
\|\mathbb{E}[P_1] - \Sigma^k\|_2 &\leq \frac{1}{T-k}\sum_{t=1}^{T-k}\sum_{i=t}^{\infty} \|Q_w\|_2 \sigma_{\max}^{2i+k} \\
&\leq \frac{\|Q_w\|_2 \sigma_{\max}^k}{(1-\sigma_{\max}^2)(T-k)} \sum_{t=1}^{T-k} \sigma_{\max}^{2t} \\
&\leq \frac{\sigma_{\max}^k}{(1-\sigma_{\max}^2)(T-k)} \frac{\|Q_w\|_2}{(1-\sigma_{\max}^2)}
\end{aligned}$$

This quantity is zero if the system is initiated from the stationary distribution.

We create stacked vectors of noise $W = [w_0|w_1|\ldots|w_T]$. Consider $\Phi \in \mathbf{R}^{nT \times n(T)}, \Gamma_i \in \mathbf{R}^{T \times nT}$

$$\Phi = \begin{bmatrix} \mathbf{I} & \ldots & \mathbf{0} \\ A & \ldots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ A^{T-1} & \ldots & \mathbf{I} \end{bmatrix}$$

$$\Gamma_t = \begin{bmatrix} \mathbf{0}_{n\times n} & \ldots & I & \ldots & \mathbf{0}_{n\times n} \end{bmatrix}$$

We can define binary matrices $\{J_l\}_{l\in[T]} \in$ of dimension $T \times T$. J_l denotes locations in block matrix Φ where A^l is present. J_l has at most 1 non-zero entry in each row. Hence, $\|J_l\|_2 \leq 1$.

$$\Phi = \sum_{l=0}^{T} J_l \otimes A^l \qquad \text{[Kronecker product]}$$

$$\Rightarrow \|\Phi\|_2 \leq \sum_{l=0}^{\infty} \|J_l\|_2 \|A^l\|_2 \qquad \text{[Norm over } \otimes\text{]}$$

$$\Rightarrow \|\Phi\|_2 \leq \sum_{l=0}^{\infty} \sigma_{\max}^l = \frac{1}{(1-\sigma_{\max})}$$

Using these, we can write $x_t = \Gamma_t \Phi W$. Writing

$$P_i = \frac{n^2}{(T-k)m^2}\sum_{t=1}^{T-k} p_t^i R_{i,t}^{1\intercal} x_t x_{t+k}^\intercal R_{i,t+k}^{2\intercal}, \tag{1}$$

M. Rao and A. Goldsmith are with the Dept. of Electrical Engineering, Stanford University, Stanford, CA 94305, USA (e-mail: milind@stanford.edu, andrea@ee.stanford.edu).

T. Javidi is with the Dept. of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093, USA (e-mail: tjavidi@ucsd.edu).

Y. Eldar is with the Dept. of Electrical Engineering Technion, Israel Institute of Technology, Haifa 32000, Israel (email:yonina@ee.technion.ac.il)

P_i	$R^1_{i,t}$	$R^2_{i,t+k}$	p^i_t
P_1	I	I	$\eta_t\omega_{t+k}$
P_2	R^1_t	R^2_{t+k}	$\sqrt{\eta_t(1-\eta_t)\omega_{t+k}(1-\omega_{t+k})}$
P_3	R^1_t	I	$\sqrt{\eta_t(1-\eta_t)}\omega_{t+k}$
P_4	I	R^2_{t+k}	$\eta_t\sqrt{\omega_{t+k}(1-\omega_{t+k})}$

TABLE I: Values of the terms in (1).

where terms are detailed in Table I.

We observe that $(p^i_t)^2 \le \eta_t\omega_{t+k}$. We now note for $\alpha, \beta \in \mathbf{R}^n, \|\alpha\|_2 = \|\beta\|_2 = 1$.

$$\begin{aligned}
\alpha^\intercal P_i\beta &= \frac{n^2}{(T-k)m^2}\sum_{t=1}^{T-k}\alpha^\intercal p^i_t R^1_{i,t}x_t x^\intercal_{t+k}R^{2\intercal}_{i,t+k}\beta \\
&= W^\intercal\Phi^\intercal\left(\frac{n^2}{m^2(T-k)}\sum_{t=1}^{T-k}p^i_t\Gamma^\intercal_{t+k}R^{2\intercal}_{i,t+k}\beta\alpha^\intercal R^{1\intercal}_{i,t}\Gamma_t\right)\Phi W \\
&= W^\intercal\Phi^\intercal B\Phi W
\end{aligned}$$

$W = Q_w^{1/2}z$ where $z \sim \mathcal{N}(0, I)$. Using this, $\alpha^\intercal P_i\beta = z^\intercal L z$.

$$\begin{aligned}
\|L\|_F^2 &= \|Q_w^{1/2\intercal}\Phi^\intercal B\Phi Q_w^{1/2}\|_F^2 \\
&\le \|Q_w\|_2^2\|\Phi\|_2^4\frac{n^4}{(T-k)^2m^4}\sum_{t=1}^{T-k}\eta_t\omega_{t+k}\|R^{2\intercal}_{i,t+k}\beta\alpha^\intercal R^{1\intercal}_{i,t}\|_F^2 \\
&\le \frac{\|Q_w\|_2^2 n^2}{(1-\sigma_{\max})^4m^2(T-k)} + o(T^{-1}) \qquad (2)
\end{aligned}$$

The final step is by using the Hoeffding bound for the convergence of $\frac{n^2}{m^2(T-k)}\sum_{t=1}^{T-k}\eta_t\omega_{t-k}$. Each term in the summation is bounded by $[0,1]$ and is subgaussian $1/4$. By Hoeffding bound with probability $> 1-\delta/5$, $\frac{1}{(T-k)}\sum_{t=1}^{T-k}\eta_t\omega_{t-k} \le \mathbb{E}[\eta_t\omega_{t+k}] + \sqrt{\frac{\log(5/\delta)}{2(T-k)}} \le m^2/n^2 + \mathcal{O}(T^{-1/2})$. Let this event be Err^c.

For the concentration result, consider eigenvalues of symmetric matrix $L^s = \frac{L+L^\intercal}{2}$ be λ_i. We have $\sum_i\lambda_i^2 = \|L^s\|_F^2 \le L_F^2$. Diagonalizing L^s and because of the circularly symmetric nature of standard gaussian vector

$$\begin{aligned}
z^\intercal Lz - \mathbb{E}[z^\intercal Lz] &= \sum_i\lambda_i(z_i^2-1) \\
\Pr\Big(\sum_i\lambda_i(z_i^2-1)\ge\epsilon\Big) &\le \mathrm{e}^{-t\epsilon}\prod_i\mathbb{E}[\exp\left(t\lambda_i(z_i^2-1)\right)] \\
&\le \exp(-t\epsilon)\prod_i\frac{e^{-t\lambda_i}}{\sqrt{1-2t\lambda_i}} \\
&\le \exp\left(-t\epsilon + 2t^2\sum_i\lambda_i^2\right)
\end{aligned}$$

The first inequality holds when $t \ge 0$. The second holds using MGF of χ^2 random variable when $t\lambda_i \le \frac{1}{2}$. The last inequality holds as $\log(1-x) \ge -x-x^2$ when $x \le \frac{1}{2}$ or whenever $t\lambda_i \le \frac{1}{4}$. We take $t = \frac{\epsilon}{4L_F^2}$ to obtain that conditioned on Err^c, with probability $> 1-\delta/5$,

$$|\alpha^\intercal(P_i - \mathbb{E}[P_i])\beta| \le \sqrt{\frac{8\log(10/\delta)}{T-k}}\frac{n\|Q_w\|_2}{m(1-\sigma_{\max})^2} + o(T^{-1/2}). \qquad (3)$$

We now present the proof of Theorem 3 which combines the above results.

Proof. **Max norm bound** Observe

$$\begin{aligned}
\|\Sigma^k - \hat\Sigma^k\|_{\max} &\le \|\hat\Sigma^k - \mathbb{E}[\hat\Sigma^k]\|_{\max} + \|\mathbb{E}[\hat\Sigma^k] - \Sigma^k\|_{\max} \\
&\le \sum_{i=1}^4\|P_i - \mathbb{E}[P_i]\|_{\max} + \mathcal{O}(T^{-1}).
\end{aligned}$$

We use (3) to get

$$\begin{aligned}
&\alpha^\intercal(\hat\Sigma^k - \Sigma^k)\beta \\
&\le 4\sqrt{\frac{8\log(10/\delta)}{T-k}}\frac{n\|Q_w\|_2}{m(1-\sigma_{\max})^2} + o(T^{-1/2})
\end{aligned}$$

when $\|\alpha\|_2, \|\beta\|_2 \le 1$.

Now using $\alpha = e_i$ and $\beta = e_j$ we obtain the convergence result for each element $|\hat\Sigma^k_{ij} - \Sigma^k_{ij}|$ and taking union bound over the n^2 choices, we obtain the result for the max bound.

ℓ_2 **norm bound** Let us define $\Delta\Sigma^k = \hat\Sigma^k - \Sigma^k$. We consider a covering set $\mathcal{A}$ such that for any $\alpha \in \mathbf{R}^n$ such that $\|\alpha\|_2 \le 1$, there exists $\alpha' \in \mathcal{A}$ with $\|\alpha'\|_2 \le 1, \|\alpha - \alpha'\|_2 \le \epsilon$. From covering set theory, we can construct such a set with $|\mathcal{A}| \le (3/\epsilon)^n$. Applying union bound, we find

$$\begin{aligned}
&\max_{\alpha,\beta\in\mathcal{A}}\alpha^\intercal\Delta\Sigma^k\beta \le 4\sqrt{\frac{8(2n\log(3/\epsilon)+\log(6/\delta))}{(T-k)}}\times \\
&\frac{n\|Q_w\|_2}{m(1-\sigma_{\max})^2} + o((T-k)^{-1/2})
\end{aligned}$$

Now, we see

$$\begin{aligned}
\|\Delta\Sigma^k\|_2 &= \max_{\alpha,\beta}\alpha^\intercal\Delta\Sigma^k\beta \\
&\le \max_{\alpha',\beta'\in\mathcal{A}}\alpha'^\intercal\Delta\Sigma^k\beta' + (\alpha-\alpha')^\intercal\Delta\Sigma^k\beta' \\
&\quad + \alpha^\intercal\Delta\Sigma^k(\beta-\beta') \\
&\le \max_{\alpha',\beta'\in\mathcal{A}}\alpha'^\intercal\Delta\Sigma^k\beta' + 2\epsilon\|\Delta\Sigma^k\|_2 \\
\Rightarrow \|\Delta\Sigma^k\|_2 &\le \frac{1}{1-2\epsilon}\max_{\alpha',\beta'\in\mathcal{A}}\alpha'^\intercal\Delta\Sigma^k\beta'
\end{aligned}$$

We use $\epsilon = 1/4$ to obtain the final result. ☐

Subsampling case The above proof has been derived for the compressive measurement case but it also holds for the subsampling case. Here $z^i_t = \Psi^i_t x_t, i \in \{1,2\}$ where Ψ^i_t is a binary matrix with m ones and $n-m$ zeros.

$$\begin{aligned}
\hat\Sigma^k &= \frac{n^2}{m^2(T-k)}\sum_{t=1}^{T-k}\Psi^1_t x_t x^\intercal_{t+k}\Psi^{2\intercal}_{t+k} \\
\alpha^\intercal\hat\Sigma^k\beta & \\
&= W^\intercal\Phi^\intercal\left(\frac{n^2}{m^2(T-k)}\sum_{t=1}^{T-k}\Gamma^\intercal_{t+k}\Psi^{2\intercal}_{t+k}\beta\alpha^\intercal\Psi^1_t\Gamma_t\right)\Phi W \\
&= W^\intercal\Phi^\intercal B\Phi W
\end{aligned}$$

Like in the earlier case, we need to bound $\|B\|_F^2$

$$\|B\|_F^2 \leq \frac{n^4}{(T-k)^2 m^4} \sum_{i,j} \beta_i^2 \alpha_j^2 \sum_{t=1}^{T-k} (\Psi_{t+k}^2)_{ii} (\Psi_t^1)_{jj}$$

From Hoeffding bound, with probability $> 1-\delta/5$ for all values of i, j,

$$\frac{1}{T-k} \sum_{t=1}^{T-k} (\Psi_{t+k}^2)_{ii} (\Psi_t^1)_{jj} \leq \mathbb{E}[(\Psi_{t+k}^2)_{ii} (\Psi_t^1)_{jj}] + \mathcal{O}(\frac{\log n/\delta}{\sqrt{T-k}})$$
$$\leq \frac{m^2}{n^2} + \mathcal{O}(T^{-1/2} \log n)$$

The rest of the proof is the same as upper bound (2) holds.

Appendix B

In this section, we estimate the transition matrix and covariance matrix under various constraints.

We derive convergence guarantees for the covariance matrix under structural assumptions.

Sparsity Let the set $\mathcal{U} = \{\Sigma : \sum_j |\Sigma_{ij}|^q \leq s \forall i\}$. We assume $\Sigma^k \in \mathcal{U}$. First we suppose $U_u(\hat{\Sigma}^k - \Sigma^k$ is symmetric.

Consider the thresholding operation $U_u(\cdot)$ defined as

$$(U_u(\Sigma))_{ij} = \Sigma_{ij} \mathbf{1}(|\Sigma_{ij}| \geq u).$$

We observe,

$$\|U_u(\hat{\Sigma}^k) - \Sigma^k\|_2 \leq \|U_u(\hat{\Sigma}^k) - U_u(\Sigma^k)\|_2 + \|U_u(\Sigma^k) - \Sigma^k\|_2$$

The second term can be bounded as

$$\begin{aligned}\|U_u(\Sigma^k) - \Sigma^k\|_2 &\leq \max_i \sum_j |\Sigma_{ij}^k| \mathbf{1}(|\Sigma_{ij}^k| \leq u) \\ &\leq \max_i u \sum_j |\Sigma_{ij}^k/u|^q \mathbf{1}(|\Sigma_{ij}^k| \leq u) \\ &\leq u^{1-q} s \qquad (4)\end{aligned}$$

The first term needs a more detailed analysis as

$$\begin{aligned}&\|U_u(\hat{\Sigma}^k) - U_u(\Sigma^k)\|_2 \\ &\leq \max_i \sum_j |(U_u(\hat{\Sigma}^k) - U_u(\Sigma^k))_{ij}| \\ &\leq \max_i \sum_j |\Sigma_{ij}^k - \hat{\Sigma}_{ij}^k| \mathbf{1}(|\Sigma_{ij}^k| \geq u, |\hat{\Sigma}_{ij}^k| \geq u) \\ &+ \max_i \sum_j |\Sigma_{ij}^k| \mathbf{1}(|\Sigma_{ij}^k| \geq u, |\hat{\Sigma}_{ij}^k| \leq u) \\ &+ \max_i \sum_j |\hat{\Sigma}_{ij}^k| \mathbf{1}(|\Sigma_{ij}^k| \leq u, |\hat{\Sigma}_{ij}^k| \geq u) \\ &= \mathrm{I} + \mathrm{II} + \mathrm{III}\end{aligned}$$

I can be bounded with high probability as,

$$\begin{aligned}\mathrm{I} &\leq \|\Delta\Sigma^k\|_{\max} \max_i \sum_j \mathbf{1}(|\Sigma_{ij}^k| \geq u) \\ &\leq \gamma(\delta) \max_i \sum_j (\Sigma_{ij}^k/u)^q \mathbf{1}(|\Sigma_{ij}^k| \geq u) \qquad (5) \\ &\leq \gamma(\delta) s u^{-q}\end{aligned}$$

For term II, we have,

$$\begin{aligned}\mathrm{II} &\leq \max_i \sum_j \left(|\Delta\Sigma_{ij}^k| + |\hat{\Sigma}_{ij}^k|\right) \mathbf{1}(|\Sigma_{ij}^k| \geq u, |\hat{\Sigma}_{ij}^k| \leq u) \\ &\leq (\gamma(\delta) + u) k u^{-q}\end{aligned}$$

where we have used the bound in (5) and recognised that each term in the second summation is bounded by u.

Term III can be written in two parts

$$\begin{aligned}\mathrm{III} &\leq \max_i \sum_j [|\Delta\Sigma_{ij}^k| + |\Sigma_{ij}^k|] \mathbf{1}(|\Sigma_{ij}^k| \leq u, |\hat{\Sigma}_{ij}^k| \geq u) \\ &\leq \max_i \sum_j |\Delta\Sigma_{ij}^k| \mathbf{1}(|\Sigma_{ij}^k| \leq u, |\hat{\Sigma}_{ij}^k| \geq u) + s u^{1-q} \\ &\leq \gamma(\delta) \max_i \sum_j \mathbf{1}(|\Sigma_{ij}^k| \geq u - \gamma(\delta)) + s u^{1-q} \\ &\leq \gamma(\delta) \frac{u^{-q}}{(1 - \gamma(\delta)/u)^q} + s u^{1-q}\end{aligned}$$

where (4) has been used.

We now use $u = 2\gamma(\delta)$ to obtain the bound. if Σ^k is not symmetric. We bound $\|\Delta\Sigma^k\|_1, \|\Delta\Sigma^k\|_\infty$ as above and use $\|\Delta\Sigma^k\|_2^2 \leq \|\Delta\Sigma^k\|_1 \|\Delta\Sigma^k\|_\infty$.

Additionally, if $\lambda_{min}(\Sigma^k) \geq \epsilon_0$, we obtain the result for the inverse as well as $\|(U_u(\hat{\Sigma}^k))^{-1} - (\Sigma^k)^{-1}\|_2 = \Omega\left(\|U_u(\hat{\Sigma}^k) - \Sigma^k\|_2\right)$

Dense Transition Matrix

With probability greater than $1-2\delta$ both, maximum value of $\Delta\Sigma^0 = \hat{\Sigma}^0 - \Sigma^0$ and $\Delta\Sigma^1 = \hat{\Sigma}^1 - \Sigma^1$ are less than γ. We have also seen that $\|\Delta\Sigma^0\|_2, \|\Delta\Sigma^1\|_2 \leq \mathcal{O}(\sqrt{n}\gamma)$. As mentioned in [1], we get

$$\|\Delta\Sigma^{0\dagger}\|_2 \leq \|\Sigma^{0\dagger}\|_2^2 \|\Delta\Sigma^0\|_2 \leq \frac{4\sqrt{n}\gamma}{\sigma_{\min}^2}.$$

This is true when $\|\Delta\Sigma^0\|_2 < \lambda_{\min}(\Sigma^0)$ and Σ^0 is invertible.

The error is given by,

$$\begin{aligned}\|\hat{A} - A\|_2 &\leq \|\hat{\Sigma}^{1\intercal}\hat{\Sigma}^{0\dagger} - \Sigma^{1\intercal}\hat{\Sigma}^{0\dagger} + \Sigma^{1\intercal}\hat{\Sigma}^{0\dagger} - \Sigma^{1\intercal}\Sigma^{0\dagger}\|_2 \\ &\leq (\|\Delta\Sigma^{0\dagger}\|_2 + \|\Sigma^{0\dagger}\|_2)\|\Delta\Sigma^1\|_2 + \|\Sigma^1\|_2 \|\Delta\Sigma^{0\dagger}\|_2 \\ &\leq \frac{4\sigma_{\max}\sqrt{n}\gamma\|Q_w\|_2}{\sigma_{\min}^2(1 - \sigma_{\max}^2)},\end{aligned}$$

completing the proof.

Sparse Transition Matrix

We now obtain results with sparse A. This proof is described in [2] for getting performance bounds on estimate A using the Dantzig selector algorithm with our estimates of Σ^0, Σ^1.

Let γ be the maximum deviation of empirical covariance matrices as earlier.

We show that $A^\intercal = \Sigma^{0\dagger}\Sigma^1$ is a feasible solution with high probability.

$$\begin{aligned}\|\hat{\Sigma}^0 A^\intercal - \hat{\Sigma}^1\|_{\max} &\leq \|(\hat{\Sigma}^0 - \Sigma^0)A\|_{\max} + \|(\hat{\Sigma}^1 - \Sigma^1)\|_{\max} \\ &\leq \gamma(\|A\|_1 + 1) = \lambda\end{aligned}$$

Clearly, $\|\hat{A}\|_1 \leq \|A\|_1$ with high probability. We also obtain,

$$\|\hat{A}-A\|_{\max} = \|\Sigma^{0\dagger}(\Sigma^0\hat{A}^\intercal - \Sigma^1)\|_{\max}$$
$$= \|\Sigma^{0\dagger}\left(\Sigma^0\hat{A}^\intercal - \hat{\Sigma}^0\hat{A}^\intercal + \hat{\Sigma}^0\hat{A}^\intercal - \hat{\Sigma}^1 + \hat{\Sigma}^1 - \Sigma^1\right)\|_{\max}$$
$$\leq 2\lambda\|\Sigma^{0\dagger}\|_1 = \lambda_1$$

We can use λ_1 as a threshold level for sparsity. We consider each column j separately. Define set $\mathcal{T} = \{i \in [n] | A_{ij}| \geq \lambda_1\}$. For convenience, we denote column j of matrix A as a and matrix $\hat{A}$ as $\hat{a}$. We can write

$$\begin{aligned}\|\hat{a}-a\|_1 &\leq \|\hat{a}_{\mathcal{T}^c}\|_1 + \|a_{\mathcal{T}^c}\|_1 + \|\hat{a}_{\mathcal{T}} - a_{\mathcal{T}}\|_1\\ &\leq \|a\|_1 + \|a_{\mathcal{T}^c}\|_1 - \|\hat{a}_{\mathcal{T}}\|_1 + \|\hat{a}_{\mathcal{T}} - a_{\mathcal{T}}\|_1\\ &\leq 2\|a_{\mathcal{T}^c}\|_1 + (\|a_{\mathcal{T}}\|_1 - \|\hat{a}_{\mathcal{T}}\|_1) + \|\hat{a}_{\mathcal{T}} - a_{\mathcal{T}}\|_1\\ &\leq 2\left(\|a_{\mathcal{T}^c}\|_1 + \|a_{\mathcal{T}} - \hat{a}_{\mathcal{T}}\|_1\right)\end{aligned}$$

Consider sum

$$\begin{aligned}s_a &= \sum_i \min(\frac{|a_i|}{\lambda_1}, 1)\\ &\leq \lambda_1^{-q}\sum_i |a_i|^q = s\lambda_1^{-q}\end{aligned}$$

Now, $\|a_{\mathcal{T}^c}\|_1 \leq \lambda_1 s_a = s\lambda_1^{1-q}$. Also, $\|a_{\mathcal{T}} - \hat{a}_{\mathcal{T}}\|_1 \leq \lambda_1|T_j| \leq \lambda_1 s_a = s\lambda_1^{1-q}$. Substituting these, we get the bound $\|\hat{A} - A\|_1 \leq 4s\lambda_1^{1-q}$.

Low Rank Transition Matrix

We assume the rank of the transition matrix A is $r \ll n$. We use the following estimator

$$\hat{A} = \mathrm{argmin}_B\langle A^\intercal, \hat{\Sigma}^0 A^\intercal - 2\hat{\Sigma}^1\rangle + \lambda_n\|A\|_*$$

For the analysis, we again denote $\hat{\Delta} = \hat{A} - A$. From the optimality conditions and some algebra,

$$\begin{aligned}\langle\bar{\Delta}^\intercal, \hat{\Sigma}^0\bar{\Delta}^\intercal\rangle &\leq 2\langle\bar{\Delta}^\intercal, \hat{\Sigma}^1 - \hat{\Sigma}^0 A^\intercal\rangle + \lambda_n(\|A\|_* - \|\hat{A}\|_*)\\ &\leq (2\|\hat{\Sigma}^1 - \hat{\Sigma}^0 A^\intercal\|_2 + \lambda_n)\|\bar{\Delta}\|_*\\ &\leq (2(\|\Delta\Sigma^1\|_2 + \sigma_{\max}\|\Delta\Sigma^0\|_2) + \lambda_n)\|\|\bar{\Delta}\|_*\end{aligned}$$

As shown in appendix earlier, we get $\|\hat{\Delta}\|_* \leq 4\sqrt{2r}\|\hat{\Delta}\|_F$ when $\lambda_n \geq 4(\|\Delta\Sigma^1\|_2 + \sigma_{\max}\|\Delta\Sigma^0\|_2) = 4(1+\sigma_{\max})\gamma_2$.

Now the optimization problem is convex when $\hat{\Sigma}^0 \succ \mathbf{0}$ and a sufficient condition is when $\|\Delta\Sigma^0\|_2 \leq \gamma_2 < \lambda_{\min}(\Sigma^0)/2$. This happens when we have large enough number of time samples $T = \Omega(\frac{128n^3\log 1/\delta}{\lambda_{\min}^2 m^2}\frac{\|Q_w\|_2^2}{(1-\sigma_{\max})^4}$. Now $\langle\bar{\Delta}^\intercal, \hat{\Sigma}^0\bar{\Delta}^\intercal\rangle \geq \frac{\lambda_{\min}(\Sigma^0)}{2}\|\bar{\Delta}\|_F^2$ which leads to the bound $\|\bar{\Delta}\|_F \leq 12\lambda_n\sqrt{2r}$.

APPENDIX C

In this appendix, we prove fundamental lower bounds on the estimation of the parameters of the autoregressive process.

1) Covariance Matrix: We consider a class of $n-$dimensional autoregressive processes with $A = 0$ and Σ^0 arising from a class $\mathcal{B}$ of symmetric $s-$sparse matrices (that have at most s elements in each row and column) detailed below

$$\mathcal{B} = \left\{\gamma\sum_{1\leq i<jn}\varepsilon_{i,j}(e_ie_j^\intercal + e_je_i^\intercal)\mathbf{1}_{(k-1)s\leq i<j\leq(k-1)(s+1),k\in[n/s])}\right.$$
$$\left. + I, \varepsilon\in\{0,1\}^{n(s-1)/2}\right\}.$$

This is the class of symmetric block-diagonal matrices. For convenience, we assume that s divides n but this assumption can be relaxed. Here $\gamma = c(m^2T/n^2)^{-1/2}$ is a parameter which we set.

Consider any $\Sigma_\varepsilon \in \mathcal{B}$. Observe that Σ_0 with $\varepsilon = 0$ is also a member. We observe that $\|\Sigma_\varepsilon - \Sigma_0\|_2 \leq s\gamma$. This quantity would be less than 1 guaranteeing that $\Sigma_\varepsilon \succeq 0$ if $T = \Omega(s^2n^2/m^2)$.

The Gilbert-Varshamov bound states that there exists a set $\mathcal{E}$ of $n(s-1)/2-$dimensional binary vectors of size $|\mathcal{E}| > 2^{\frac{n(s-1)}{16}}$ such that for any $\varepsilon, \varepsilon' \in \mathcal{E}$, $\|\varepsilon - \varepsilon'\|_1 > \frac{n(s-1)}{16}$. Using this, there exists a subset $\mathcal{B}_\mathcal{E}$, $|\mathcal{B}_\mathcal{E}| > 2^{n(s-1)/16}$, and for any $\Sigma_\varepsilon, \Sigma_{\varepsilon'}$, we have that

$$\|\Sigma_\varepsilon - \Sigma_{\varepsilon'}\|_F^2 \geq \frac{\gamma^2 n(s-1)}{8} > \frac{\gamma^2 ns}{16}$$
$$\Rightarrow \|\Sigma_\varepsilon - \Sigma_{\varepsilon'}\|_2 \geq \gamma\sqrt{\frac{s}{4}}$$

At each point in time, we observe $Z_t = \Psi_t X_t$. Alternatively, we could observe $Y_t = M_t X_t \in \mathbf{R}^m$. In the subsampling case, M_t is Ψ_t with all the zero rows removed. In the *orthogonal compressive measurement scenario*, M_t has rows that are uniformly sampled from the $n-$dimensional hypersphere and are orthogonal to one another. To reiterate, $\Psi_t = M_t' M_t$ in this case. Now we can observe that $Y_t \sim \mathbb{P}'_{\Sigma^0} = \mathcal{N}(0, M_t\Sigma^0 M_t^\intercal)$. Also define, $\mathbb{P}_{t,\Sigma}(Z_t) = \mathbb{P}(M_t)\mathbb{P}'_{t,\Sigma}(Y_t)$. As an example, we see that $\mathbb{P}'_{t,\Sigma_0} = \mathcal{N}(0, I_m)$. It follows from independence ($A = 0$) that $\mathbb{P}_\Sigma(Z_1^T) = \prod_{t=1}^T \mathbb{P}_{t,\Sigma}(Z_t)$.

We now find an upper bound for $D_{KL}(\mathbb{P}_{\Sigma_\varepsilon}\|\mathbb{P}_{\Sigma_0})$. We see,

$$\begin{aligned}D_{KL}(\mathbb{P}_{\Sigma_\varepsilon}\|\mathbb{P}_{\Sigma_0}) &= \mathbb{E}_{M_1^T}\mathbb{E}\left[\log\left(\frac{\mathbb{P}_{\Sigma_\varepsilon}(Z_1^T)}{\mathbb{P}_{\Sigma_0}}\right)|M_1^T\right]\\ &= \sum_{t=1}^T \mathbb{E}_{M_t}\left[D_{KL}(\mathbb{P}'_{t,\Sigma_\varepsilon}\|\mathbb{P}'_{t,\Sigma_0})\right]\end{aligned}$$

We use the KL divergence between absolutely continuous normal distributions to note

$$D_{KL}(\mathbb{P}'_{t,\Sigma_\varepsilon}\|\mathbb{P}'_{t,\Sigma_0}) = \frac{1}{2}\mathrm{Tr}(M_t\Sigma_\varepsilon M_t^\intercal) - \frac{1}{2}\log|M_t\Sigma_\varepsilon M_t^\intercal| - \frac{m}{2}$$
$$\begin{aligned}M_t\Sigma_\varepsilon M_t^\intercal &= I_m + \gamma\sum_{i\neq j} M_t\varepsilon_{i,j}e_ie_j^\intercal M_t^\intercal\\ &= I_m + Q_t\end{aligned}$$

Q_t has zero for its diagonal elements in expectation. To see this,

$$\mathbb{E}[(M_t e_i e_j^\intercal M_t^\intercal)_{kk}] = \mathbb{E}[(M_t)_{k,i}(M_t)_{k,j}] \\ = 0 \text{ when } i \neq j. \quad (6)$$

This is because row $(M_t)_k$ is a uniformly chosen unit vector with $(M_t)_{k,i} = \frac{u_i}{\sqrt{\sum_{i=1}^n u_i^2}}, u_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Symmetry dictates (6). Denote its eigenvalues by $\lambda_i, i \in [n]$. We see that $\mathbb{E}[\text{Tr}(Q_t)] = \sum_{i=1}^r \lambda_i = 0$. Also,

$$D_{KL}(\mathbb{P}'_{t,\Sigma_\varepsilon} \| \mathbb{P}'_{t,\Sigma_0}) = -\frac{1}{2}\log|I_m + Q_t| \\ = -\frac{1}{2}\sum_{i=1}^r \log(1+\lambda_i) \leq \frac{1}{4}\sum_{i=1}^r \lambda_i^2 - 2\lambda_i$$

$$\Rightarrow \mathbb{E}[\frac{1}{4}\sum_{i=1}^r \lambda_i^2 - 2\lambda_i] \leq \frac{1}{4}\mathbb{E}[\|Q_t\|_F^2] \leq \frac{\gamma^2 n(s-1)m^2}{2n^2}$$

For the last step, we use (7) and (8) detailed below.

$$\mathbb{E}[\|Q_t\|_F^2] \leq \gamma^2 \sum_{a,b\in[m]} \mathbb{E}\left[\left(\sum_{i\neq j}\varepsilon_{i,j}(M_t)_{a,i}(M_t)_{b,j}\right)^2\right] \\ \leq \gamma^2 \sum_{a,b\in[m]}\sum_{i\neq j}\varepsilon_{i,j}^2 \mathbb{E}[(M_t)_{a,i}^2(M_t)_{b,j}^2] \\ \leq \gamma^2 n(s-1)m^2 \mathbb{E}[(M_t)_{a,i}^2(M_t)_{b,j}^2] \quad (7)$$

where we have used $\mathbb{E}[(M_t)_{a,i}(M_t)_{b,j}] = 0$. Now, $(M_t)_{a,i}^2 \sim \text{Beta}(\frac{1}{2}, \frac{n-1}{2})$. Using this and cauchy inequality,

$$\mathbb{E}[(M_t)_{a,i}^2(M_t)_{b,j}^2] \leq \mathbb{E}[(M_t)_{a,i}^4] \\ \leq \frac{2}{n^2}. \quad (8)$$

Putting everything together,

$$D_{KL}(\mathbb{P}_{\Sigma_\varepsilon} \| \mathbb{P}_{\Sigma_0}) \leq \frac{\gamma^2 Tn(s-1)}{2n^2} \\ \leq c\frac{n(s-1)}{16} = c\log|\mathcal{B}_\mathcal{E}|$$

A. Transition Matrix

We consider a class $\mathcal{A}$ of transition matrices that are block diagonal with each block being $s \times s$. The noise matrix $Q_w = I$. Again, for convenience, we assume s divides n but the proof can easily be extended to relax this assumption. The transition matrix comes from class:

$$\mathcal{A} = \\ \left\{\gamma \sum_{i,j\in[n]} \varepsilon_{i,j} e_i e_j^\intercal \mathbf{1}_{(k-1)s\leq i<j\leq(k-1)(s+1), k\in[n/s]}) \varepsilon \in \{0,1\}^{ns}\right\}$$

Here, $\gamma = cn/m\sqrt{T}$. We require that $\|A_\varepsilon\|_2 \leq \sigma_{\max} < 1$ for the VAR process to be stable as described in Section **??**. Seeing $\|A_\varepsilon\|_2 \leq \|A_\varepsilon\|_1 \leq s\gamma$, we require that $T = \Omega(ns/m)$.

Any matrix $A_\varepsilon \in \mathcal{A}$ is indexed by an $ns-$dimensional binary vector ε. From the Gilbert-Varshamov theorem, we can come up with a subset $\mathcal{A}_\mathcal{E} \subset \mathcal{A}$ with $|\mathcal{A}_E| \geq 2^{ns/8}$ such that for any $A_\varepsilon, A_{\varepsilon'} \in \mathcal{A}_\mathcal{E}$, we have,

$$\|A_\varepsilon - A_{\varepsilon'}\|_F^2 \geq \frac{ns\gamma^2}{8} \Rightarrow \|A_\varepsilon - A_{\varepsilon'}\|_2 \geq \gamma\sqrt{\frac{s}{8}}$$

Observe that stacked states

$$X_1^T \sim \mathcal{N}\left(0, \begin{bmatrix} I & A & A^2 & \dots & A^{T-1} \\ A^\intercal & I & A & \dots & A^{T-2} \\ \vdots & & \ddots & & \vdots \\ A^{T-1\intercal} & A^{T-2\intercal} & A^{T-3\intercal} & \dots & I \end{bmatrix}\right) \\ \sim \mathcal{N}(0, \Phi_A)$$

Retaining notation Y_1^T and stacking matrices M_t diagonally to form M, we get that $Y_1^T \sim \mathbb{P}'_A = \mathcal{N}(0, M\Phi M^\intercal)$ and $\mathbb{P}_A(Z_1^T) = \mathbb{P}(M_1^T)\mathbb{P}'_A(Y_1^T)$. We seek to bound $D_{KL}(\mathbb{P}_{A_\varepsilon}\|\mathbb{P}_{A_0})$.

$$D_{KL}(\mathbb{P}_{A_\varepsilon}\|\mathbb{P}_{A_0}) = \mathbb{E}_{M_1^T}\left[D_{KL}(\mathbb{P}'_{A_\varepsilon}\|\mathbb{P}'_{A_0})\right] \\ = \mathbb{E}\left[\frac{1}{2}\text{Tr}(M\Phi_\varepsilon M^\intercal) - \frac{1}{2}\log|M\Phi_\varepsilon M^\intercal| - \frac{Tm}{2}\right] \\ = \mathbb{E}\left[-\frac{1}{2}\log|I_{Tm} + Q|\right] \\ \leq \mathbb{E}\left[\frac{1}{4}\|Q\|_F^2\right]$$

Where $M\Phi_A M^\intercal = I_{Tm} + Q$. Now,

$$\mathbb{E}[\|Q\|_F^2] \\ \leq \sum_{t_1\neq t_2\in[T]} \mathbb{E}\left[\|M_{t_1}A^{|t_2-t_1|}M_{t_2}^\intercal\|_F^2\right] \\ \leq \sum_{t_1\neq t_2\in[T];a,b\in[m]} \mathbb{E}\left[\left(\sum_{i,j}(A^{|t_2-t_1|})_{i,j}(M_{t_1})_{a,i}(M_{t_2})_{b,j}\right)^2\right] \\ \leq \sum_{t_1\neq t_2\in[T]} \frac{2m^2\gamma^2\sigma_{\max}^{2|t_1-t_2|}ns}{n^2} \\ \leq \frac{4Tm^2\gamma^2 ns}{n^2(1-\sigma_{\max}^2)} \\ \Rightarrow D_{KL}(\mathbb{P}_{A_\varepsilon}\|\mathbb{P}_{A_0}) \leq c'\frac{ns}{8} < c'\log|\mathcal{A}_\mathcal{E}|$$

A fact used here is that $|(A^l)_{i,j}| \leq \gamma\sigma_{\max}^{l-1}$.

Low rank Transition Matrices

1) Low-Rank Transition Matrix: We consider the family $\mathcal{A}$ of rank r transition matrices (with $Q_w = I$). For convenience, assume r divides n.

$$\mathcal{A} = \left\{\mathbf{1}_{n/r} \otimes \bar{A}_\varepsilon, \bar{A}_\varepsilon \in \mathbf{R}^{r\times n}, (\bar{A}_\varepsilon)_{i,j} = \gamma\varepsilon_{i,j}, \varepsilon \in \{0,1\}^{nr}\right\}$$

Here $\gamma = c\sqrt{rn/Tm^2}$. For any $A \in \mathcal{A}$, we require stability, or $n\gamma \leq \sigma_{\max} < 1$, which implies a requirement of $T = \Omega(n^3 r/m^2)$. From the Gilbert-Varshamov theorem, we know that there exists $\mathcal{A}_\mathcal{E} \subset \mathcal{A}$ with $|\mathcal{A}_\mathcal{E}| \geq 2^{nr/8}$ and for $A_\varepsilon, A_{\varepsilon'} \in \mathcal{A}_\mathcal{E}$,

$$\|A_\varepsilon - A_{\varepsilon'}\|_F^2 \geq \frac{\gamma^2 n^2}{8}$$

If we write out the KL divergence, it is almost identical to the previous case. We obtain

$$D_{KL}(\mathbb{P}_{A_\varepsilon}\|\mathbb{P}_{A_0}) \leq \frac{2Tm^2\gamma^2n^2}{n^2(1-\sigma_{\max}^2)} \leq \frac{c'nr}{8} = c'\log|\mathcal{A}_\mathcal{E}|$$

Appendix D

A. Sparse Covariance Matrix

In this section, we prove a tighter lower bound for the rate of convergence of sparse covariance matrices.

We follow the analysis of [3] and consider a class of covariance matrices that are sparse. The analysis follows a modified version of Assouad's lemma.

We consider the class of symmetric covariance matrices defined as

$$\mathcal{S} = \left\{\Sigma | \max_{j\leq n}\sum_{i\neq j}|\Sigma_{ij}|^q \leq s\right\}$$

When $q=0$, we see that there are at most s non-zero non-diagonal elements in each column and by symmetry, each row.

Our constructed parameter set is as follows:

1) Consider $r=\lfloor n/2\rfloor$, approximately half the size of the dimension. We consider a matrix of dimension $r\times r$ that has exactly s non-zero elements in each row and at most $2s$ non-zero elements in each column. We call this set Λ. To be more precise,

$$\Lambda = \left\{M\in\mathbf{R}^{r\times r} | \forall i\in[r], \sum_j |M_{i,j}|^0 = s, \forall j\in[r]\sum_i |M_{i,j}|^0 \leq 2s, M_{i,j}\in\{0,\nu\}\right\}$$

2) Further consider set Γ, the set of all binary sequences of length r. This set would express whether a row of a matrix in Λ is seen.
3) For any $\lambda\in\Lambda$, let λ_i represent row i. Now we define matrix $L(\lambda_i)$ as follows. Consider $\lambda'_i\in\mathbf{R}^{1\times n}$ where $\lambda'_{i,j}=\lambda_{i,j-\lceil n/2\rceil}\mathbf{1}(j\geq\lceil n/2\rceil)$. Now, $L(\lambda_i)=\lambda_i^{'\intercal}\lambda'_i$. This means that the i^{th} row of $L(\lambda_i)$ has the r elements of λ_i as its right-most elements. By symmetry, the last r elements of the i^{th} column also arise from here.
4) Consider the parameter set $\Theta=(\Gamma,\Lambda)$ with elements $\theta=(\gamma,\lambda)$. We now define the class of covariance matrices we consider as

$$\mathcal{S}_1 = \left\{\Sigma(\theta) = I + \nu\sum_{i=1}^r\gamma_i L(\lambda_i), \theta\in\Theta\right\}$$

First we note that $\|\Sigma(\theta)\|_2\geq 1-2s\nu$. Taking $\nu=\mathcal{O}(c\sqrt{\frac{\log n}{T}})$, when $s=\mathcal{O}(\sqrt{\frac{T}{\log n}})$, we see that $\Sigma(\theta)$ is psd. To reiterate, we note that the number of non-zero elements in each row and column does not exceed $2s$.

In this case, we assume that $A=0$ and $X_t\overset{\text{iid}}{\sim}\mathcal{N}(0,\Sigma(\theta))$. Let $\mathbb{P}_\theta$ denote the probability of observing Z_1^T. We see $Z_t=M_tX_t$, and thus $\mathbb{P}_{t,\theta}(Z_t)=\mathbb{P}(M_t)\mathbb{P}'_{t,\theta}(Z_t)$ where $\mathbb{P}'_{t,\theta}=\mathcal{N}(0,M_t\Sigma_\theta m_t^\intercal)$. We borrow some notation from earlier and write.

$$\mathbb{P}_\theta(Z_1^T) = \prod_{t=1}^T\mathbb{P}_{t,\theta}(Z_t) = \prod_{t=1}^T\mathbb{P}(M_t)\mathbb{P}'_{t,\theta}(Z_t)$$

Upon observing Z_1^T, an estimator comes up with an estimate $\Sigma_{\hat\theta}$. Observe the following sequence

$$\begin{aligned}\max_\theta\mathbb{E}[\|\Sigma_{\hat\theta}-\Sigma_\theta\|_2] &\geq \frac{1}{2^r|\Lambda|}\sum_\theta\mathbb{E}[\|\Sigma_{\hat\theta}-\Sigma_\theta\|_2] \\ &\geq \frac{1}{2^r|\Lambda|}\sum_\theta\mathbb{E}[\frac{\|\Sigma_{\hat\theta}-\Sigma_\theta\|_2}{\rho(\hat\gamma,\gamma)\wedge 1}\rho(\hat\gamma,\gamma)] \\ &\geq \min_{\rho(\hat\gamma,\gamma)\geq 1}\frac{\|\Sigma_{\hat\theta}-\Sigma_\theta\|_2}{\rho(\hat\gamma,\gamma)}\frac{1}{2^r|\Lambda|}\sum_\theta\mathbb{E}[\rho(\hat\gamma,\gamma)]\end{aligned}$$

Now we show for $\rho(\hat\gamma,\gamma)\geq 1$,

$$\frac{\|\Sigma_{\hat\theta}-\Sigma_\theta\|_2^2}{\rho(\hat\gamma,\gamma)} \geq \frac{\|(\Sigma_{\hat\theta}-\Sigma_\theta)v\|_2^2}{\rho(\hat\gamma,\gamma)\|v\|_2^2} \geq \frac{s^2\nu^2}{n}$$

The choice of v here is $v_j=\mathbf{1}(j\geq\lceil n/2\rceil)$.

We now focus on the other term and see that

$$\begin{aligned}&\frac{1}{2^r|\Lambda|}\sum_\theta\mathbb{E}[\rho(\hat\gamma,\gamma)] \\ &\geq \frac{1}{2^r|\Lambda|}\mathbb{E}_{M_t}\left[\sum_{i=1}^r\sum_{\theta:\gamma_i=0}\mathbb{E}[\hat\gamma_i|M_t] + \sum_{\theta:\gamma_i=1}\mathbb{E}[1-\hat\gamma_i|M_t]\right] \\ &\geq \frac{1}{2}\sum_{i=1}^r\mathbb{E}_{M_t}\left[\int\hat\gamma_i\sum_{\gamma_i=0}\frac{d\mathbb{P}'_\theta}{2^{r-1}|\Lambda|} + (1-\hat\gamma_i)\sum_{\gamma_i=1}\frac{d\mathbb{P}'_\theta}{2^{r-1}|\Lambda|}\right] \\ &\geq \frac{1}{2}\sum_{i=1}^r\mathbb{E}_{M_t}\left[1-D_{TV}(\bar{\mathbb{P}}'_{\theta,\gamma_i=0},\bar{\mathbb{P}}'_{\theta,\gamma_i=1})\right]\end{aligned}$$

Here $\bar{\mathbb{P}}'_{\theta,\gamma_i=0}=\frac{1}{2^{r-1}|\Lambda|}\sum_{\theta:\gamma_i=0}\mathbb{P}'_\theta$. D_{TV} is the total variation distance.

It is easy to see that that the total variation distance between mixture distributions is less than the total variation distance between constituents leading to

$$\begin{aligned}&D_{TV}(\bar{\mathbb{P}}'_{\gamma_i=0},\bar{\mathbb{P}}'_{\gamma_i=1}) \\ &\leq \frac{1}{2^{r-1}|\Lambda_{-i}|}\sum_{\theta:\gamma_{-i},\lambda_{-i}}D_{TV}(\mathbb{P}'_{\gamma_i=0,\gamma_{-i},\lambda_{-i}},\bar{\mathbb{P}}'_{\gamma_i=1,\gamma_{-i},\lambda_{-i}}) \\ &\leq \min_{\gamma_{-i},\lambda_{-i}}D_{TV}(\mathbb{P}'_{\gamma_i=0,\gamma_{-i},\lambda_{-i}},\bar{\mathbb{P}}'_{\gamma_i=1,\gamma_{-i},\lambda_{-i}})\end{aligned}$$

We now use the following relation between distances between measures

$$D_{TV}(\mathbb{P}_a,\mathbb{P}_b) \leq \sqrt{D_{\chi^2}(\mathbb{P}_a,\mathbb{P}_b)} = \mathbb{E}_{\mathbb{P}_b}[(d\mathbb{P}_a/d\mathbb{P}_b)^2-1]$$

We now study what the distributions we are considering look like. $\mathbb{P}'_{\gamma_1=0,\gamma_{-1},\lambda_{-1}} = \prod_t \mathbb{P}'_{t,\gamma_1=0,\gamma_{-1},\lambda_{-1}}$, the latter is a single multivariate distribution with the covariance matrix,

$$\Sigma_0 = \begin{cases} \begin{bmatrix} 1 & 0 \\ 0 & M_{t,-1} S M_{t,-1}^\intercal \end{bmatrix} & M_{t,1} = e_1, \\ M_t S' M_t^\intercal & e_1 \notin M_t \end{cases},$$

where $M_t = [M_{t,1}; M_{t,-1}]$ and S is a symmetric matrix dependent on $(\lambda_{-1}, \gamma_{-1})$ with the property for $i \leq j$

$$S_{ij} = \begin{cases} 1 & i = j \\ \nu & \gamma_i = \lambda_{ij} = 1 \\ 0 & \text{else} \end{cases}$$

We can see that $\bar{\mathbb{P}}'_{t,\gamma_1=1,\gamma_{-1},\lambda_{-1}}$ is a mixture of distributions of a number of Gaussians. Suppose $n_{\lambda_{-1}}$ is the number of columns in λ_{-1} with elements equal to $2s$. From $n_{\lambda_{-1}} 2s \leq rs$, we see that $n_{\lambda_{-1}} \leq r/2$. Thus the number of distributions is given by the number of non-zero elements in the first row λ_1 that are not in these $n_{\lambda_{-1}}$ positions. The maximum number is given by $\binom{r/2}{s} = \binom{n/4}{s}$. Each of these distributions has this form

$$\Sigma_i = \begin{cases} \begin{bmatrix} 1 & r^\intercal M_{t,-1}^\intercal \\ M_{t,-1} r & M_{t,-1} S M_{t,-1}^\intercal \end{bmatrix} & M_{t,1} = e_1, \\ M_t S' M_t^\intercal & e_1 \notin M_t \end{cases},$$

We see that if $e_1 \notin M_t$, distributions $\mathbb{P}'_{t,\gamma_1=0,\gamma_{-1},\lambda_{-1}} = \bar{\mathbb{P}}'_{t,\gamma_1=1,\gamma_{-1},\lambda_{-1}}$ and the distance between them is 0. Since we seek to find an upper bound to the distance, we can assume that $e_1 \in M_t$.

We use the following useful lemma relating to chi-squared distances between normal distributions $g_i = \mathcal{N}(0, \Sigma_i)$:

$$\int \frac{g_1 g_2}{g_0} = |I - \Sigma_0^{-2}(\Sigma_1 - \Sigma_0)(\Sigma_2 - \Sigma_0)|^{-1/2}$$

Let's denote

$$R(t, \gamma_{-1}, \lambda_{-1}, \lambda_1, \lambda_1') = |I - \Sigma_0^{-2}(\Sigma_{\lambda_1} - \Sigma_0)(\Sigma_{\lambda_1'} - \Sigma_0)|^{-1/2}$$

We can now write

$$\mathbb{E}_{\gamma_{-1},\lambda_{-1}} \left[\int \left(\frac{\bar{\mathbb{P}}_{\gamma_1=1,\gamma_{-1},\lambda_{-1}}}{\bar{\mathbb{P}}_{\gamma_1=0,\gamma_{-1},\lambda_{-1}}} \right)^2 d\bar{\mathbb{P}}_{\gamma_1=0,\gamma_{-1},\lambda_{-1}} - 1 \right] \leq$$
$$\mathbb{E}_{\lambda_1,\lambda_1'} \mathbb{E}_{\gamma_{-1},\lambda_{-1}|\lambda_1,\lambda_1'} \left[\prod_{t=1}^T R(t, \gamma_{-1}, \lambda_{-1}, \lambda_1, \lambda_1') - 1 \right]$$

Here is an observation:

$$R(t, \gamma_{-1}, \lambda_{-1}, \lambda_1, \lambda_1')$$
$$= R'(t, \gamma_{-1}, \lambda_{-1}, \lambda_1, \lambda_1') |I - ((\Sigma_{\lambda_1} - \Sigma_0)(\Sigma_{\lambda_1'} - \Sigma_0))|^{-1/2}$$

As proven in Lemma 11 of [3],

$$\mathbb{E}_{\lambda_1,\lambda_1'|J} \mathbb{E}_{\gamma_{-1},\lambda_{-1}|\lambda_1,\lambda_1'} \prod_{t=1}^T R'(t, \gamma_{-1}, \lambda_{-1}, \lambda_1, \lambda_1') \leq 1.5$$

Let's focus on the matrix $(\Sigma_{\lambda_1} - \Sigma_0)(\Sigma_{\lambda_1'} - \Sigma_0)$. It can be written as

$$(\Sigma_{\lambda_1} - \Sigma_0)(\Sigma_{\lambda_1'} - \Sigma_0) = \begin{bmatrix} r_1^\intercal M_{t,-1}^\intercal M_{t,-1} r_2 & 0 \\ 0 & M_{t,-1} r_1 r_2^\intercal M_{t,-1}^\intercal \end{bmatrix}$$

This can be seen to be a rank-2 matrix as it is of the form $\begin{bmatrix} \alpha^\intercal \beta & 0 \\ 0 & \alpha \beta^\intercal \end{bmatrix}$ and the identical eigenvalues are $|r_1^\intercal M_{t,-1}^\intercal M_{t,-1} r_2|$. Thus,

$$|I - ((\Sigma_{\lambda_1} - \Sigma_0)(\Sigma_{\lambda_1'} - \Sigma_0))|^{-1/2}$$
$$= (1 - |r_1^\intercal M_{t,-1}^\intercal M_{t,-1} r_2|)^{-1}$$

Let the rows of $M_{t,-1}$ be $m_{t,i}, i \in [m-1]$. We had assumed that $m_{t,i}$ are orthogonal and from the unit sphere. Suppose that r_1 is non-zero in indices I_1 and r_2 is non-zero in indices I_2. Let the number of overlapping indices be J. We note that

$$r_1^\intercal M_{t,-1}^\intercal M_{t,-1} r_2 \leq \sum_{l=1}^m \sum_{i \in I_1, j \in I_2} \nu^2 m_{t,l,i} m_{t,l,j}$$
$$\leq s^2 \nu^2 < 1,$$

with appropriate choice of constant in ν. We can conclude

$$|I - ((\Sigma_{\lambda_1} - \Sigma_0)(\Sigma_{\lambda_1'} - \Sigma_0))|^{-1/2}$$
$$\leq 1 + 2|r_1^\intercal M_{t,-1}^\intercal M_{t,-1} r_2|$$

As described in [3], J arises from a hypergeometric distribution and is bounded by $\left(\frac{s^2}{n/4-1-s}\right)^j$

Putting all of this together,

$$\mathbb{E}_{M_t} \mathbb{E}_{\gamma_{-1},\lambda_{-1}} \left[\int \left(\frac{\bar{\mathbb{P}}_{\gamma_1=1,\gamma_{-1},\lambda_{-1}}}{\bar{\mathbb{P}}_{\gamma_1=0,\gamma_{-1},\lambda_{-1}}} \right)^2 d\bar{\mathbb{P}}_{\gamma_1=0,\gamma_{-1},\lambda_{-1}} - 1 \right]$$
$$\leq \sum_j \left(\frac{s^2}{n/4 - 1 - s} \right)^j \left\{ \mathbb{E}_{M_t} \prod_{t=1}^T (1 + 2|r_1^\intercal M_{t,-1}^\intercal M_{t,-1} r_2|) \frac{3}{2} - 1 \right\}$$
$$\leq\leq \sum_j \left(\frac{s^2}{n/4 - 1 - s} \right)^j \left\{ \prod_{t=1}^T (1 + 2j\nu^2 \frac{m^2}{n^2}) \frac{3}{2} - 1 \right\}$$

References

[1] J. Demmel, "The componentwise distance to the nearest singular matrix," *SIAM Journal on Matrix Analysis and Applications*, vol. 13, no. 1, pp. 10–19, 1992. [Online]. Available: http://dx.doi.org/10.1137/0613003

[2] F. Han, H. Lu, and H. Liu, "A direct estimation of high dimensional stationary vector autoregressions," *Journal of Machine Learning Research*, vol. 16, pp. 3115–3150, 2015.

[3] T. T. Cai, H. H. Zhou *et al.*, "Optimal rates of convergence for sparse covariance matrix estimation," *The Annals of Statistics*, vol. 40, no. 5, pp. 2389–2420, 2012.