Estimation in Autoregressive Processes with Partial Observations: Proofs

Milind Rao, Tara Javidi, Yonina Eldar, Andrea Goldsmith

Appendix A

In this appendix, we prove the convergence of $\|\hat{\Sigma}^0 - \Sigma^0\|_2$. In order to do this, we use a covering net argument. First, we prove convergence for any $\alpha, \beta \in \mathbf{R}^n$ such that $\|\alpha\|_2, \|\beta\|_2 \leq 1$.

We assume that the process begins $T_p \geq 0$ time units before observations take place. In other words, $x_{-T_p} = x_S$. We provide some definitions and rewrite a few expressions.

Consider $\Phi \in \mathbf{R}^{nT\times n(T_p+T)}, \Gamma_i \in \mathbf{R}^{T\times nT}, \Lambda_k \in \mathbf{R}^{T\times T}$

$$\Phi = \begin{bmatrix} A^{T_p} & \dots & A & \mathbf{I} & \dots & \mathbf{0} \\ A^{T_p+1} & \dots & A^2 & A & \dots & \mathbf{0} \\ \vdots & & & \ddots & & \vdots \\ A^{T_p+T-1} & \dots & A^T & A^{T-1} & \dots & \mathbf{I} \end{bmatrix}$$

$$\Gamma_i = \begin{bmatrix} e_i^\intercal \\ e_{n+i}^\intercal \\ \vdots \\ e_{n(T-1)+i}^\intercal \end{bmatrix}$$

$$\Lambda_k = \begin{bmatrix} \mathbf{0}_{T-k\times k} & \mathbf{I}_{T-k\times T-k} \\ \mathbf{0}_{k\times k} & \mathbf{0}_{k\times T-k} \end{bmatrix}$$

Lemma 1. *We have these properties:*

1) $\|\Phi\|_2 \leq (1-\sigma_{\max})^{-1}$
2) $\Lambda_k^\intercal \Gamma_i \Gamma_j^\intercal \Lambda_k = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{T-k\times T-k} \end{bmatrix} \mathbf{1}(i=j)$

Proof. We can define binary matrices $\{J_l\}_{l\in[T_p+T]} \in$ of dimension $T \times T_p + T$. 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, $\|J_l\|_2 \leq 1$.

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

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

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

The second point is self-evident by definition. □

Let $\mathbf{0}_l$ be an l-dimensional vector of zeros. We create stacked vectors of noise $W = [w_{-T_p+1}|\dots|w_0|w_1|\dots|w_T]$, the initial conditions of the same dimension $X_S = [x_S|\mathbf{0}((T+T_p-1)n)]$, and the observational noise $V = [v_1|\dots|v_T]$. Let the stacked vector of observations of position i with delay k be the T-dimensional vector $Z(k)_i = [z_{1+k,i}|z_{2+k,i}|\dots|z_{T,i}|\mathbf{0}_k]$. We recall that $P_{t,i}$ is 1 if the i^{th} position of noisy observation of x_t is observed in the sampling case or is the multiplicative noise otherwise. We create the T-diagonal matrix $P(k)_i = \text{diag}([P_{1+k,i}|\dots|P_{T,i}|\mathbf{0}(k)])$ and denote with $P(k)_{i,j} = P(0)_i P(k)_j$. Finally, $\theta(k)_{i,j} = \mathbb{E}[P_{t,i}P_{t+k,j}]$.

First, we prove a lemma about the impact of multiplicative noise or sampling.

Lemma 2. *With bounded multiplicative noise, we have with probability at most $\delta/3$, event* Err *occurs where*

$$\text{Err} = \left\{ \max_{i,j} \frac{\text{Tr}(P^2(k)_{i,j}}{(T-k)\theta(k)_{i,j}} - 1 \geq \sqrt{\frac{(k+1)(p_u^4 - p_l^4)\log(3n^2(k+1)/\delta)}{2(T-2k)\theta(k)_*^2}} \right\}$$

Proof. To bound $\text{Tr}(P^2(k)_{i,j})$, we need to bound the sum $\sum_{t=1}^{T-k} P_{t,i}^2 P_{t+k,j}^2$. We break this up into $k+1$ with the number of terms being at least $\lceil T - 2k/k+1 \rceil$ independent terms. The m^{th} such series is bounded by $S_m = p_u^2 \sum_{t=1}^{\lceil T-k-m+1/k+1 \rceil} P_{(k+1)t+m-1,i} P_{(k+1)t+m-1+k,j}$.

First consider the case where $P_{t,i}$ is bounded between $[p_l, p_u]$. Each of the terms in the sum is $(p_u^4 - p_l^4)^2/4$ subgaussian. By Hoeffding inequality,

$$\Pr\left(S_m \geq \theta(k)_{i,j}\lceil T-k-m+1/k+1 \rceil (1+p_\rho)\right) \leq \exp\left(-\frac{2\theta(k)_{i,j}^2 p_\rho^2 \lceil T-2k/k+1 \rceil}{(p_u^4 - p_l^4)^2}\right)$$

We re-arrange and use union bound over these $k+1$ sums as well as the n^2 number of i,j terms and rearrange to complete the proof. □

From earlier definitions, we have

$$Z(k)_i = P(k)_i \Lambda_k \Gamma_i (\Phi(W + X_S) + V)$$

$$\alpha^\intercal \hat{\Sigma}_{ij}^k \beta = \sum_{i,j} \alpha_i \beta_j \big[\frac{1}{(T-k)\theta(k)_{i,j}} Z(0)_i^\intercal Z(k)_j - (Q_v)_{i,j} \mathbf{1}(k=0) \big].$$

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)

We can split $\alpha^\intercal \hat{\Sigma}^k_{ij} \beta$ into these three terms -

$$
\begin{aligned}
T_1 &= (W^\intercal \Phi^\intercal + V^\intercal) A_T (\Phi W + V) \\
&\quad - \alpha^\intercal Q_v \beta \mathbf{1}(k=0) \\
T_2 &= X_S^\intercal (A_T + A_T^\intercal)(\Phi W + V) \\
T_3 &= X_S^\intercal \Phi^\intercal A_T \Phi X_S \\
\hat{\Sigma}^k_{i,j} &= T_1 + T_2 + T_3 \\
A_T &= \sum_{i,j} \alpha_i \beta_j \Gamma_i^\intercal \frac{P(k)_{i,j}}{(T-k)\theta(k)_{i,j}} \Lambda_k \Gamma_j
\end{aligned}
$$

Lemma 3. *Conditioned on the event that* Err *does not occur, we have*

$$
\begin{aligned}
&\Pr\left(|T_1 - \mathbb{E}[T_1]| \geq \epsilon\right) \\
&\leq 2\exp\left(-\frac{\epsilon^2 (T-k)\theta(k)_*}{8\max(\|Q_v\|_2^2, \frac{\|Q_w\|_2^2}{(1-\sigma_{\max})^4})}\right) \qquad (1) \\
&\Pr\left(|T_2| \geq \epsilon\right) \\
&\leq 2\exp\left(-\frac{\epsilon^2 (T-k)^2 \theta(k)_*^2}{8 p_u^4 \|x_S\|_2^2 (\|Q_w\|_2 (1-\sigma_{\max})^{-2} + \|Q_v\|_2)}\right) \qquad (2)
\end{aligned}
$$

$$
|T_3| \leq \frac{p_u^2 \sigma_{\max}^{2T_p} \|x_S\|_2^2}{(T-k)\theta(k)_* (1-\sigma_{\max})^2} \qquad (3)
$$

$$
\begin{aligned}
&|T_3 - \mathbb{E}[T_3]| \\
&\leq \frac{(\frac{p_u^2}{\theta(k)_*} + 1)\sigma_{\max}^{2T_p} \|x_S\|_2^2}{(T-k)(1-\sigma_{\max})^2} \qquad (4)
\end{aligned}
$$

Proof. **Term** T_1**:**

W can be written as $Q_W^{1/2} z_w$ where $Q_W = \mathbb{E}[WW^\intercal] = Q_w \otimes \mathbf{I}_{T+T_p \times T+T_p}$ and $z_w \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{n\times n})$. Similarly $V = Q_V^{1/2} z_v$. It can be seen that $\|Q_W\|_2 \leq \|Q_w\|_2$, $\|Q_V\|_2 \leq \|Q_v\|_2$.

$$
\begin{aligned}
T_1 &= \begin{bmatrix} z_W \\ z_V \end{bmatrix}^\intercal L_1 \begin{bmatrix} z_W \\ z_V \end{bmatrix} - \alpha^\intercal Q_v \beta \mathbf{1}(k=0) \\
L_1 &= B_T^\intercal A_T B_T \\
B_T &= \begin{bmatrix} \Phi Q_W^{1/2} & \mathbf{0} \\ \mathbf{0} & Q_V^{1/2} \end{bmatrix} \\
\Rightarrow \|L_1\|_F^2 &\leq \|B_T\|_2^4 \|A_T\|_F^2
\end{aligned}
$$

Norm of B_T can be bounded as

$$
\|B_T\|_2^4 \leq \max(\|Q_v\|_2^2, \frac{\|Q_w\|_2^2}{(1-\sigma_{\max})^4})
$$

We employ lemma **??** and **??** to now bound A_T with high probability as

$$
\begin{aligned}
\|A_T\|_F^2 &= \sum_{i,j} \frac{\alpha_i^2 \beta_j^2}{(T-k)^2 \theta(k)_{i,j}} \|P(k)_{i,j} \Lambda_k\|_F^2 \\
&\leq \sum_{i,j} \frac{\alpha_i^2 \beta_j^2}{(T-k)^2 \theta(k)_{i,j}} \mathrm{Tr}(P(k)_{i,j}^2) \\
&\leq \frac{1}{(T-k)\theta(k)_*} (\sum_i \alpha_i^2)(\sum_j \beta_j^2) \\
&\leq \frac{1}{(T-k)\theta(k)_*}
\end{aligned}
$$

For the concentration result, consider eigenvalues of symmetric matrix $L^s = \frac{L_1 + L_1^\intercal}{2}$ be λ_i. We have $\sum_i \lambda_i^2 = \|L^s\|_F^2 \leq L_F^2$. Diagonalizing L^s and because of the circularly symmetric nature of standard gaussian vector

$$
\begin{aligned}
z^\intercal L_1 z - \mathbb{E}[z^\intercal L_1 z] &= \sum_i \lambda_i (z_i^2 - 1) \\
\Pr(\sum_i \lambda_i (z_i^2 - 1) \geq \epsilon) &\leq \mathrm{e}^{-t\epsilon} \prod_i \mathbb{E}[\exp\left(t\lambda_i (z_i^2 - 1)\right)] \\
&\leq \exp\left(-t\epsilon\right) \prod_i \frac{e^{-t\lambda_i}}{\sqrt{1 - 2t\lambda_i}} \\
&\leq \exp\left(-t\epsilon + 2t^2 \sum_i \lambda_i^2\right)
\end{aligned}
$$

The first inequality holds when $t \geq 0$. The second holds using MGF of χ^2 random variable when $t\lambda_i \leq \frac{1}{2}$. The last inequality holds as $\log(1-x) \geq -x - x^2$ when $x \leq \frac{1}{2}$ or whenever $t\lambda_i \leq \frac{1}{4}$. We take $t = \frac{\epsilon}{4L_F^2}$ to obtain the bound.

Term T_2

We can write

$$
\begin{aligned}
T_2 &= l_2^\intercal \begin{bmatrix} z_w \\ z_v \end{bmatrix} \\
l_2 &= X_S^\intercal \Phi^\intercal (A_T + A_T^\intercal) \begin{bmatrix} \Phi Q_W^{1/2} & Q_V^{1/2} \end{bmatrix}
\end{aligned}
$$

$$
\Rightarrow \|l_2\|_2^2 \leq \frac{4}{(T-k)^2} \|x_S\|_2^2 \|A_T\|_2^2 [(1-\sigma_{\max})^{-2}) \|Q_w\|_2 + \|Q_v\|_2]
$$

We now bound $\|A_T\|_2^2$ as

$$
\begin{aligned}
\|A_T\|_2^2 &\leq \sum_{i,j,i',j'} \alpha_i \beta_j \alpha_{i'} \beta_{j'} \|\Gamma_j^\intercal \Lambda_k \frac{P(k)_{i,j}}{\theta(k)_{i,j}} \Gamma_i \Gamma_{i'}^\intercal \frac{P(k)_{i',j'}}{\theta(k)_{i',j'}} \Lambda_k \Gamma_{j'}\|_2 \\
&\leq \frac{p_u^4}{\theta(k)_*^2} \sum_{i,j} \alpha_i^2 \beta_j^2
\end{aligned}
$$

where the last inequality is by applying lemma **??** and observing that $\Gamma_j^\intercal \Lambda_k^\intercal P^2 \Lambda_k \Gamma_{j'}$ is zero when $j \neq j'$ as P is a diagonal matrix. We now apply Hoeffding bound to arrive at the answer.

Term T_3

We use the bound on $\|A_T\|_2$ and submultiplicative property of the ℓ_2 bound to prove the bound. Also, $|T_3 - \mathbb{E}[T_3]| \leq |T_3| + |\mathbb{E}[T_3]|$. $\square$

Lemma 4. *The difference between the mean of the sample covariance and the true covariance matrices is bounded as*

$$\|\mathbb{E}[\hat{\Sigma}^k] - \Sigma^k\|_2 \leq \frac{\sigma_{\max}^{2T_p+k}}{(1-\sigma_{\max}^2)(T-k)} \times \left[\frac{\|Q_w\|_2}{(1-\sigma_{\max}^2)} + \frac{p_u^2\|x_S\|_2^2}{\min_{i,j}\theta(k)_{i,j}}\right].$$

Proof. We have $\Sigma^k = \mathbb{E}[x_t x_{t+k}^\intercal] = \left(\sum_{i=0}^{\infty} A^i Q_w A^{i\intercal}\right) A^{k\intercal}$. Now we can split the empirical covariance into two terms - the first due to a start from origin and the second due to the exponential decay of the initial state captured in T_3.

$$\mathbb{E}[\hat{\Sigma}^k] = \mathbb{E}\left[\frac{1}{T-k}\sum_{t=1}^{T-k} x_t x_{t+k}^\intercal \mid x_{-T_p} = x_S\right] \succeq \frac{1}{T-k}\sum_{t=1}^{T-k}\sum_{i=0}^{T_p+t-1} A^i Q_w A^{i+k\intercal} + |T_3| I$$

$$\begin{aligned}\|\mathbb{E}[\hat{\Sigma}^k] - \Sigma^k\|_2 &\leq \frac{1}{T-k}\sum_{t=1}^{T-k}\sum_{i=T_p+t}^{\infty}\|Q_w\|_2\sigma_{\max}^{2i+k} + |T_3| \\ &\leq \frac{\|Q_w\|_2\sigma_{\max}^k}{(1-\sigma_{\max}^2)(T-k)}\sum_{t=1}^{T-k}\sigma_{\max}^{2(T_p+t)} + |T_3| \\ &\leq \frac{\sigma_{\max}^{2T_p+k}}{(1-\sigma_{\max}^2)(T-k)}\left[\frac{\|Q_w\|_2}{(1-\sigma_{\max}^2)} + \frac{p_u^2\|x_S\|_2^2}{\min_{i,j}\theta(k)_{i,j}}\right]\end{aligned}$$

We complete the proof by observing that for any $M \times M$ matrix L, $\|L\|_{\max} = \max_{i,j\in[M]}|e_i^\intercal L e_j| \leq \|L\|_2$. ☐

We now present the proof of Theorem 1 which combines the above results.

Proof. **Max norm bound** Conditioned on event Err^c, using Lemma **??** and Lemma **??**, we see that with probability larger than $1-\delta/3$,

$$|T_1 - \mathbb{E}[T_1]| \leq \sqrt{\frac{8\log(6/\delta)}{(T-k)\theta(k)_*}}\max\left(\frac{\|Q_w\|_2}{(1-\sigma_{\max})^2}, \|Q_v\|_2\right) + o((T-k)^{-0.5}).$$

Similarly, for T_2 we find that with probability larger than $1-\delta/3$,

$$|T_2| \leq \frac{p_u^2\|x_S\|_2}{(T-k)\theta(k)_*} \times \sqrt{8\log(6/\delta)\left(\frac{\|Q_w\|_2}{(1-\sigma_{\max})^2} + \|Q_v\|_2\right)}$$

which is $o((T-k)^{-0.5})$.

Finally,

$$\begin{aligned}\|\Sigma^k - \hat{\Sigma}^k\|_{\max} &\leq \|\hat{\Sigma}^k - \mathbb{E}[\hat{\Sigma}^k]\|_{\max} + \|\mathbb{E}[\hat{\Sigma}^k] - \Sigma^k\|_{\max} \\ &\leq |T_1 - \mathbb{E}[T_1]| + |T_2| \\ &\quad + |T_3 - \mathbb{E}[T_3]| + \|\mathbb{E}[\hat{\Sigma}^k] - \Sigma^k\|_{\max}\end{aligned}$$

We use Lemma **??** to get

$$\alpha^\intercal(\hat{\Sigma}^k - \Sigma^k)\beta \leq \sqrt{\frac{8\log(6/\delta)}{(T-k)\theta(k)_*}}\max\left(\frac{\|Q_w\|_2}{(1-\sigma_{\max})^2}, \|Q_v\|_2\right) + o((T-k)^{-1/2})$$

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

Now using $\alpha = e_i$ and $\beta = e_j$ we obtain the convergence result for each element $|\hat{\Sigma}_{ij}^k - \Sigma_{ij}^k|$ 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 $\|\alpha\|_2 \leq 1$, there exists $\alpha' \in \mathcal{A}$ with $\|\alpha'\|_2 \leq 1, \|\alpha - \alpha'\|_2 \leq \epsilon$. From covering set theory, we can construct such a set with $|\mathcal{A}| \leq (3/\epsilon)^n$. Applying union bound, we find

$$\max_{\alpha,\beta\in\mathcal{A}}\alpha^\intercal\Delta\Sigma^k\beta \leq \sqrt{\frac{8(2n\log(\epsilon/3) + \log(6/\delta))}{(T-k)\theta(k)_*}} \times \max\left(\frac{\|Q_w\|_2}{(1-\sigma_{\max})^2}, \|Q_v\|_2\right) + o((T-k)^{-1/2})$$

Now, we see

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

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

APPENDIX B

In this appendix, 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

$$\begin{aligned}\|U_u(\Sigma^k) - \Sigma^k\|_2 &\leq \max_i\sum_j|\Sigma_{ij}^k|\mathbf{1}(|\Sigma_{ij}^k| \leq u) \\ &\leq \max_i u\sum_j|\Sigma_{ij}^k/u|^q\mathbf{1}(|\Sigma_{ij}^k| \leq u) \\ &\leq u^{1-q}s \qquad (5)\end{aligned}$$

The first term needs a more detailed analysis as

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

I can be bounded with high probability as,

$$\begin{aligned}\mathrm{I} &\le \|\Delta\Sigma^k\|_{\max} \max_i \sum_j \mathbf{1}(|\Sigma^k_{ij}| \ge u) \\ &\le \gamma(\delta) \max_i \sum_j (\Sigma^k_{ij}/u)^q \mathbf{1}(|\Sigma^k_{ij}| \ge u) \\ &\le \gamma(\delta) s u^{-q}\end{aligned} \tag{6}$$

For term II, we have,

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

where we have used the bound in Eq. **??** and recognised that each term in the second summation is bounded by u.

Term III can be written in two parts

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

where Eq. **??** 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 $\|\Delta\Sigma^k\|_2^2 \le \|\Delta\Sigma^k\|_1 \|\Delta\Sigma^k\|_\infty$.

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

Bandedness It is assumed that $\Sigma^k \in \mathcal{V} = \{\Sigma : \max_i \sum_j |\Sigma^k_{ij}| \mathbf{1}(|i-j| > s) \le C s^{-q} \forall k, i\}$.

We consider the banding operation $B_s(\cdot)$ defined as

$$B_s(\Sigma)_{ij} = \Sigma_{ij} \mathbf{1}(|i-j| \le s)$$

As earlier, we observe,

$$\begin{aligned}\|B_s(\hat{\Sigma}^k) - \Sigma^k\|_2 &\le \|B_s(\hat{\Sigma}^k) - B_s(\Sigma^k)\|_2 + \|B_s(\Sigma^k) - \Sigma^k\|_2 \\ &\le 2s\gamma(\delta) + C s^{-\alpha}\end{aligned}$$

We use $s = \gamma^{-1/(\alpha+1)}(\delta)$ to obtain the final answer $\mathcal{O}(\gamma^{\alpha/(\alpha+1)}(\delta))$. The inverse can be obtained in a similar manner to the sparse case by additionally assuming that the minimum eigenvalue of Σ^k is above ϵ_0.

Sparsity of the Inverse

Here we make the assumption that the inverse covariance matrix $\Theta^0 = (\Sigma^0)^{-1}$ is sparse. Let $\mathcal{E}(\Theta^0) = \{(i,j) | i \neq j, \Theta^0_{ij} \neq 0\}$ be the set of off-diagonal non-zero elements in the inverse covariance matrix. Define $s = |\mathcal{E}(\Theta^0)|$ as the size of this set. Set $\mathcal{S} = \mathcal{E}(\Theta) \cup \{(i,i) | i \in [n]\}$ includes the diagonals. Also, d is the maximum row cardinality which is the maximum number of non-zero elements in any row of the inverse covariance matrix.

We define $\Gamma = (\Theta^0)^{-1} \otimes (\Theta^0)^{-1}$ which is the Hessian of the log-determinant determinant function. We characterize the convergence in terms of quantities $\kappa_\Sigma = \|\Sigma^0\|_\infty, \kappa_\Gamma = \|\Gamma\|_\infty$. Another important assumption being made is an irrepresentability condition given by $\|\Gamma_{\mathcal{S}^c\mathcal{S}}(\Gamma_{\mathcal{S}\mathcal{S}})^{-1}\|_\infty \le 1 - \alpha$.

The estimator for the empirical inverse covariance matrix is obtained from the Bregman divergence on the log determinant function. Consider $g(\Theta) = -\log|\Theta|$. We now find symmetric positive definite matrix Θ which minimizes $D_g(\Theta^0 || \Theta)$ which leads to

$$\hat{\Theta}^0 = \mathrm{argmin}_{\Theta \succ 0} \mathrm{Tr}(\Theta^\intercal \Sigma^0) - \log|\Theta| + \lambda_n \|\Theta\|_{1,\mathrm{off}}$$

We obtain the final estimator by replacing unknown Σ^0 with its empirical estimate and a regularization term which is the ℓ_1 sum of off-diagonal elements $\|\Theta\|_{1,\mathrm{off}} = \sum_{i,j\ i \neq j} |\Theta_{ij}|$.

For $T \ge 288 \log \frac{6n^2}{\delta} d^2 \max(\frac{\|Q_w\|_2^2}{(1-\sigma_{\max})^4}, \|Q_v\|_2^2) \max(\kappa_\Gamma^2 \kappa_\Sigma^2, \kappa_\Gamma^4 \kappa_\Sigma^6)(1 + \frac{8}{\alpha})^2 \theta(0)_*^{-1}$, with probability at least $\|\Delta\Sigma^0\|_{\max} \le \gamma(\delta) \le \frac{1}{6(1+8/\alpha) d \max(\kappa_\Gamma \kappa_\Sigma, \kappa_\Gamma^2 \kappa_\Sigma^3)}$. Following Theorem 1 and corollary 3 of [**?**], we see with high probability and upto order $T^{-1/2}$

$$\begin{aligned}\|\hat{\Theta}^0 - \Theta^0\|_{\max} &\le 2\kappa_\Gamma (1 + \frac{8}{\alpha}) \gamma(\delta) \\ \|\hat{\Theta}^0 - \Theta^0\|_F &\le 2\kappa_\Gamma (1 + \frac{8}{\alpha}) \sqrt{s+n} \gamma(\delta) \\ \|\hat{\Theta}^0 - \Theta^0\|_2 &\le 2\kappa_\Gamma (1 + \frac{8}{\alpha}) \min(\sqrt{s+n}, d) \gamma(\delta) \\ \|\hat{\hat{\Sigma}}^0 - \Sigma^0\|_2 &\le 2\kappa_\Sigma^2 \kappa_\Gamma (1 + \frac{8}{\alpha}) d\gamma(\delta) + 6\kappa_\Sigma^3 \kappa_\Gamma^2 (1 + \frac{8}{\alpha})^2 d^2 \gamma^2(\delta)\end{aligned}$$

Low rank matrix We assume the rank of the matrix Σ^k is $r \ll n$. We employ the following estimator to obtain a low rank matrix approximation

$$\bar{\Sigma}^k = \mathrm{argmin}_\Sigma \|\Sigma - \hat{\Sigma}^k\|_F^2 + \lambda_n \|\Sigma\|_*$$

Define $\bar{\Delta} = \bar{\Sigma}^k - \Sigma^k$. We now observe,

$$\begin{aligned}&\|\bar{\Sigma}^k - \hat{\Sigma}^k\|_F^2 + \lambda_n \|\bar{\Sigma}\|_* \le \|\Sigma^k - \hat{\Sigma}^k\|_F^2 + \lambda_n \|\Sigma^k\|_* \\ &\Rightarrow \|\bar{\Delta}\|_F^2 - 2\langle \bar{\Delta}, \Delta\Sigma^k \rangle \le \lambda_n \|\bar{\Delta}\|_* \\ &\Rightarrow \|\bar{\Delta}\|_F^2 \le (2\|\Delta\Sigma^k\|_2 + \lambda_n) \|\bar{\Delta}\|_* \\ &\Rightarrow \|\bar{\Delta}\|_F^2 \le \frac{3}{2} \lambda_n \|\bar{\Delta}\|_*\end{aligned} \tag{7}$$

where in the final step, we have used the fact that $\lambda_n \geq 4\|\Delta\Sigma^k\|_2$ and $\|A\|_* \leq \sqrt{r}\|A\|_F$.

We now bound $\|\bar{\Delta}\|_*$. We define subspace $\mathcal{A}$ to span the first r singular vectors of Σ^k and $\mathcal{B}$ the remaining singular vectors. We use $\Pi_{\mathcal{A}}$ to denote the euclidean projection operation onto subspace $\mathcal{A}$. Clearly, $\Sigma^k = \Pi_{\mathcal{A}}(\Sigma^k) + \Pi_{\mathcal{B}}(\Sigma^k)$.

We now define $\bar{\Delta}_2 = \Pi_{\mathcal{B}}(\bar{\Delta})$ and $\bar{\Delta}_1 = \bar{\Delta} - \bar{\Delta}_2$. Consider the SVD of $\Sigma^k = UDV^\intercal$. We can write

$$\begin{aligned}\bar{\Delta} &= U \begin{bmatrix}\nu_{11} & \nu_{12}\\ \nu_{21} & \nu_{22}\end{bmatrix} V^\intercal\\ \Rightarrow \bar{\Delta}_1 &= U \begin{bmatrix}\nu_{11} & \nu_{12}\\ \nu_{21} & \mathbf{0}\end{bmatrix} V^\intercal\\ &= U \left(\begin{bmatrix}\nu_{11}/2 & \mathbf{0}\\ \nu_{21} & \mathbf{0}\end{bmatrix} + \begin{bmatrix}\nu_{11}/2 & \nu_{12}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\right) V^\intercal\end{aligned}$$

where $\nu_{11} \in \mathbf{R}^{r\times r}$. Clearly, $\operatorname{rank}(\bar{\Delta}_1) \leq 2r$ as it can be written as a sum of 2 matrices with r non-zero rows or columns in each.

We can write

$$\begin{aligned}\|\bar{\Sigma}^k\|_* &= \|\Pi_{\mathcal{A}}(\Sigma^k) + \bar{\Delta}_2 + \Pi_{\mathcal{B}}(\Sigma^k) + \bar{\Delta}_1\|_*\\ &\geq \|\Pi_{\mathcal{A}}(\Sigma^k) + \bar{\Delta}_2\|_* - \|\Pi_{\mathcal{B}}(\Sigma^k) + \bar{\Delta}_1\|_*\\ &\geq \|\Pi_{\mathcal{A}}(\Sigma^k)\|_* + \|\bar{\Delta}_2\|_* - \|\Pi_{\mathcal{B}}(\Sigma^k)\|_* - \|\bar{\Delta}_1\|_* \quad (8)\end{aligned}$$

From optimal solution of optimization problem, we have

$$\begin{aligned}&0 \leq \|\bar{\Delta}\|_F^2/\lambda_n\\ &\leq \frac{1}{2}\bar{\Delta}\|_* + \|\Sigma^k\|_* - \|\bar{\Sigma}^k\|_*\\ &2\|\Pi_{\mathcal{B}}(\Sigma^k)\|_* + \frac{3}{2}\|\bar{\Delta}_1\|_* - \frac{1}{2}\|\bar{\Delta}_2\|_*\\ &\Rightarrow \|\bar{\Delta}_2\|_* \leq 3\|\bar{\Delta}_1\|_* + 4\|\Pi_{\mathcal{B}}(\Sigma^k)\|_*,\end{aligned}$$

where we have used Eq. **??** in the third inequality. We conclude

$$\begin{aligned}\|\bar{\Delta}\|_* &\leq 4\|\bar{\Delta}_1\|_*\\ &\leq 4\sqrt{2r}\|\bar{\Delta}\|_F\end{aligned}$$

We substitute this in the Eq. **??** to obtain $\|\bar{\Delta}\|_F \leq 6\lambda_n\sqrt{2r}$

Appendix C

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

Dense Transition Matrix

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

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

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

The error is given by,

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

completing the proof.

Sparse Transition Matrix

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

Let $\gamma(\delta/2)$ 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.

$$\begin{aligned}\|\hat{\Sigma}^0 A^\intercal - \hat{\Sigma}^1\|_{\max} &\leq \|(\hat{\Sigma}^0 - \Sigma^0)A\|_{\max} + \|(\hat{\Sigma}^1 - \Sigma^1)\|_{\max}\\ &\leq \gamma(\delta/2)(\|A\|_1 + 1) = \lambda\end{aligned}$$

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

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

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

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

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 $\lambda_n \geq 4(\|\Delta\Sigma^1\|_2 + \sigma_{\max}\|\Delta\Sigma^0\|_2) = 4(1 + \sigma_{\max})\gamma_2(\delta/2)$.

Now the optimization problem is convex when $\hat{\Sigma}^0 \succ \mathbf{0}$ and a sufficient condition is when

$\|\Delta\Sigma^0\|_2 \leq \gamma_2(\delta/2) < \lambda_{\min}(\Sigma^0)/2$. This happens when we have large enough number of time samples $T \geq \frac{128n\log 1/\delta}{\lambda_{\min}^2\theta(0)_*}\max\left(\frac{\|Q_w\|_2^2}{(1-\sigma_{\max})^4}, \|Q_v\|_2^2\right)$. 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 D

In this section, we prove the analogue of Theorem 1 for higher order VAR processes.

The proof from section **??** goes through with a few modifications. $Q_{\underline{V}} = \mathbb{E}[\underline{V}\underline{V}^\intercal] = Q_v \otimes J_V$ where J_V is a binary matrix with at most p ones in each row. Thus $\|Q_{\underline{V}}\|_2 \leq p\|Q_v\|_2$.

The other difference is the term $\mathrm{Tr}(\underline{P}^2(k)_{i,j})$. It can be observed that

$$\mathrm{Tr}(\underline{P}^2(k)_{i,j}) = \mathrm{Tr}\left(P^2\left(\left|\lfloor\frac{j-1}{n}\rfloor - \lfloor\frac{i-1}{n}\rfloor + k\right|\right)_{i_p,j_p}\right)$$

$$(i_p, j_p) = \begin{cases}(i-1 \bmod n+1, j-1 \bmod n+1) \\ \quad \lfloor\frac{j-1}{n}\rfloor - \lfloor\frac{i-1}{n}\rfloor + k \geq 0 \\ (j-1 \bmod n+1, i-1 \bmod n) \quad \text{o.w.}\end{cases}.$$

Thus earlier convergence result holds with union bound taken over $(np)^2$ choices of i, j.

We also now take the union bound over $(np)^2$ choices for the max bound and correspondingly larger set for the 2 norm. $|\mathcal{A}| \leq (3/\epsilon)^{np}$ to get the final answer.

REFERENCES

[1] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu, "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence," *Electron. J. Statist.*, vol. 5, pp. 935–980, 2011. [Online]. Available: http://dx.doi.org/10.1214/11-EJS631

[2] 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

[3] 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.