**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 $z_i^t = \Psi_i x_t$, $i \in \{1, 2\}$ where the rank of matrix $\Psi_i$ is $m < n$.

We saw that we could write $z_i^t = \eta_i x_t + \sqrt{\eta_i(1 - \eta_i)} R_i^t x_t$ and $z_i^t = \omega_i x_t + \sqrt{\omega_i(1 - \omega_i)} R_i^t x_t$. Here $R_i^t$ are rotation matrices that are uniformly distributed on the hypersphere and perpendicular to $x_t$. $\eta_i, \omega_i \sim \text{Beta}\left(\frac{m}{2}, \frac{n-m}{2}\right)$.

Consider the estimate of the covariance matrix,

$$\hat{\Sigma}^k = \frac{n^2}{(T-k)m^2} \sum_{t=k}^{T-k} z_i^t z_i^{T+k}$$

This is because $\Sigma^k = \mathbb{E}[x_t x_t^\top] = \left(\sum_{i=0}^{\infty} \mathbb{A}_i^T \mathbb{A}_i^\top\right) A^k$.  

$$\mathbb{E}[P_i] = \mathbb{E}\left[ \frac{1}{T-k} \sum_{t=1}^{T-k} x_t x_t^\top \right]$$

$$\mathbb{E}[P_2] = \mathbb{E}[P_3] = \mathbb{E}[P_4] = 0$$

The former is because $\mathbb{E}[\eta_i] = \mathbb{E}[\omega_i] = m/n$ and the latter is because $R_i^t$ is a symmetric random rotation matrix.

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

$$\|\mathbb{E}[P_1] - \Sigma^k\|_2 \leq \frac{\sigma^\text{max}_o}{(1 - \sigma^\text{max}_o)(T-k)} \|Q_o\|_2$$

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)
where terms are detailed in Table I.

We observe that $(p^2_k) \leq \eta \omega_{t+k}$. We now note for $\alpha, \beta \in \mathbb{R}^n$, $\|\alpha\|_2 = \|\beta\|_2 = 1$.

$$
\alpha^T P_1 \beta = \frac{n^2}{(T-k)m^2} \sum_{t=1}^{T-k} \alpha^T p_t i x_t x_{t+k} R_{t+k}^2 \beta
$$

$$
= W^T \Phi \left( \frac{n^2}{m^2(T-k)} \sum_{t=1}^{T-k} \eta \omega_{t+k} \| R_{t+k}^2 \beta \|_{1}^2 \right) \Phi W
$$

$$
W = Q_w^{1/2} \text{ where } z \sim N(0, I). \text{ Using this, } \alpha^T P_1 \beta = z^T L_z.
$$

$$
\|L\|_F^2 = \|Q_w^{1/2} \Phi B \Phi Q_w^{1/2} \|_F^2
$$

$$
\leq \|Q_w\|_2^2 |\Phi|_2^4 \sum_{t=1}^{T-k} \eta \omega_{t+k} \| R_{t+k}^2 \beta \|_{1}^2 \| \alpha^T R_{t+k}^2 \beta \|_2^2
$$

$$
\leq \frac{\|Q_w\|_2^2 n^2}{(1 - \sigma_{max}^2)4m^2(T-k)} + o(T^{-1})
$$

(2)

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

For the concentration result, consider eigenvalues of symmetric matrix $L^* = L + L^T/2$ be $\lambda_i$. We have $\lambda_i^2 = \| L \|_F^2 \leq L^2_f$. Diagonalizing $L^*$ and because of the circularly symmetric nature of standard gaussian vector $z^T L z = \sum_{i} \lambda_i (z^2_i - 1)$

$$
\Pr(\sum_{i} \lambda_i (z^2_i - 1) \geq \epsilon) \leq e^{-\epsilon^2 \sum_{i} \lambda_i^2 - 1})
$$

$$
\leq \exp(-te - \epsilon t \lambda_i / \sqrt{1 - 2t \lambda_i})
$$

$$
\leq \exp(-te - 2t^2 \sum_{i} \lambda_i^2)
$$

The first inequality holds when $t \geq 0$. The second holds using MGF of $\chi^2$ 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 $t = \frac{1}{4}$ to obtain that conditioned on $\mathcal{C}$, with probability $> 1 - \delta/5$,

$$
\alpha^T (P_1 - \mathbb{E}[P_1]) \beta \leq \frac{8 \log(10/\delta)}{T - k} \sum_{t=1}^{T-k} \eta \omega_{t+k} \| p_t \|_2 + o(T^{-1/2}).
$$

(3)

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

**Proof.** Max norm bound** Observe

$$
\|\Sigma^k - \hat{\Sigma}^k\|_{\text{max}} \leq \|\Sigma^k - \mathbb{E}[\Sigma^k]\|_{\text{max}} + \mathbb{E}[\hat{\Sigma}^k] - \Sigma^k\|_{\text{max}}
$$

$$
\leq \sum_{i=1}^{k} \|P_i - \mathbb{E}[P_i]\|_{\text{max}} + O(T^{-1}).
$$

We use (3) to get

$$
\alpha^T (\hat{\Sigma}^k - \Sigma^k) \beta
$$

$$
\leq 4 \frac{\sqrt{8 \log(10/\delta)}}{T - k} \frac{\|Q_w\|_2}{m(1 - \sigma_{max}^2)^2} + o(T^{-1/2})
$$

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

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

**$\ell_2$ norm bound** Let us define $\hat{\Sigma}^k = \Sigma^k - \hat{\Sigma}^k$. We consider a covering set $\mathcal{A}$ such that for any $\alpha \in \mathbb{R}^n$ 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

$$
\max_{\alpha, \beta \in \mathcal{A}} \alpha^T \Delta \Sigma^k \beta \leq \frac{4 \sqrt{8 \log(3/\epsilon) + \log(6/\delta)}}{(T - k)} \sum_{t=1}^{T-k} \mathbb{E} \left[ \Psi_{t,i} \right] + o((T - k)^{-1/2})
$$

Now, we see

$$
\|\Delta \Sigma^k\|_2 \leq \max_{\alpha, \beta \in \mathcal{A}} \alpha^T \Delta \Sigma^k \beta
$$

$$
\leq \max_{\alpha', \beta' \in \mathcal{A}} \alpha'^T \Delta \Sigma^k \beta' + (\alpha - \alpha')^T \Delta \Sigma^k \beta'
$$

$$
\leq \max_{\alpha', \beta' \in \mathcal{A}} \alpha'^T \Delta \Sigma^k \beta' + 2\epsilon \|\Delta \Sigma^k\|_2
$$

$$
\Rightarrow \|\Delta \Sigma^k\|_2 \leq \frac{1}{1 - 2\epsilon} \max_{\alpha', \beta' \in \mathcal{A}} \alpha'^T \Delta \Sigma^k \beta'
$$

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 $z^*_t = \Psi_{t,i}, i \in \{1, 2\}$ where $\Psi_{t,i}$ is a binary matrix with $m$ ones and $n - m$ zeros.

$$
\hat{\Sigma}^k = \frac{n^2}{m^2(T - k)} \sum_{t=1}^{T-k} \Psi_{t,i} x_{t,i} x_{t+k} \Psi_{t+k}^2
$$

$$
\alpha^T \hat{\Sigma}^k \beta
$$

$$
= \frac{n^2}{m^2(T - k)} \sum_{t=1}^{T-k} \Psi_{t,i} x_{t,i} x_{t+k} \Psi_{t+k}^2 \alpha \mathbb{E}[\Psi_{t,i} \Lambda_i] \Phi W
$$

$$
= \frac{n^2}{m^2(T - k)} \sum_{t=1}^{T-k} \Gamma_{t+k} \Psi_{t+k}^2 \alpha \mathbb{E}[\Psi_{t,i}] \Phi W
$$

$$
= \frac{n^2}{m^2(T - k)} \sum_{t=1}^{T-k} \Gamma_{t+k} \Psi_{t+k}^2 \alpha \mathbb{E}[\Psi_{t,i}] \Phi W
$$
Like in the earlier case, we need to bound \( ||B||^2_F \):
\[
||B||^2_F \leq \frac{n^4}{(T-k)^2} m \sum_i \beta_i^2 \alpha_j^2 \sum_{t=1}^{T-k} (\Psi^2_{t+k})_{ii}(\Psi^1_{t})_{jj}
\]
From Hoeffding bound, with probability \( 1 - \delta/5 \) for all values of \( i,j \),
\[
\frac{1}{T-k} T_{t=1}^{T-k} (\Psi^2_{t+k})_{ii}(\Psi^1_{t})_{jj} \leq \mathbb{E}[(\Psi^2_{t+k})_{ii}(\Psi^1_{t})_{jj}] + O(\frac{\log n/\delta}{\sqrt{T-k}})
\]
\[
\leq \frac{n^2}{n^2} + O(T^{-1/2} \log n)
\]
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 \( U = \{ \Sigma : \sum_j |\Sigma_{ij}|^q \leq s \forall i \}. \) We assume \( \Sigma^k \in U \). First we suppose \( U_n(\hat{\Sigma}^k) - \Sigma^k \) is symmetric.

Consider the thresholding operation \( U_n(\cdot) \) defined as \( (U_n(\Sigma))_{ij} = \Sigma_{ij} 1(|\Sigma_{ij}| \geq u) \).

We observe,
\[
||U_n(\hat{\Sigma}^k) - \Sigma^k||_2 \leq ||U_n(\hat{\Sigma}^k) - U_n(\Sigma^k)||_2 + ||U_n(\Sigma^k) - \Sigma^k||_2
\]
The second term can be bounded as
\[
||U_n(\Sigma^k) - \Sigma^k||_2 \leq \max_i \sum_j |\Sigma^k_{ij}| 1(|\Sigma^k_{ij}| \leq u)
\]
\[
\leq \max_i u \sum_j |\Sigma^k_{ij}|/u^q 1(|\Sigma^k_{ij}| \leq u)
\]
\[
\leq u^{1-q} s\text{.} \tag{4}
\]
The first term needs a more detailed analysis as
\[
||U_n(\hat{\Sigma}^k) - U_n(\Sigma^k)||_2
\]
\[
\leq \max_i \sum_j ||(U_n(\hat{\Sigma}^k) - U_n(\Sigma^k))_{ij}||
\]
\[
\leq \max_i \sum_j |\Sigma^k_{ij} - \hat{\Sigma}^k_{ij}| 1(|\Sigma^k_{ij}| \geq u, |\hat{\Sigma}^k_{ij}| \geq u)
\]
\[
+ \max_i \sum_j |\Sigma^k_{ij} - \hat{\Sigma}^k_{ij}| 1(|\Sigma^k_{ij}| \geq u, |\hat{\Sigma}^k_{ij}| \leq u)
\]
\[
+ \max_i \sum_j |\hat{\Sigma}^k_{ij} - \Sigma^k_{ij}| 1(|\Sigma^k_{ij}| \leq u, |\hat{\Sigma}^k_{ij}| \leq u)
\]
\[
= I + II + III
\]
I can be bounded with high probability as,
\[
I \leq ||\Delta \Sigma^k|| \max_i \sum_j 1(|\Sigma^k_{ij}| \geq u)
\]
\[
\leq \gamma(\delta) \max_i \sum_j (\Sigma^k_{ij}/u)^q 1(|\Sigma^k_{ij}| \geq u)
\]
\[
\leq \gamma(\delta) su^{-q}\tag{5}
\]
For term II, we have,
\[
II \leq \max_i \sum_j \left( |\Delta \Sigma^k_{ij}| + |\hat{\Sigma}^k_{ij}| \right) 1(|\Sigma^k_{ij}| \geq u, |\hat{\Sigma}^k_{ij}| \leq u)
\]
\[
\leq \gamma(\delta) u k^{u-\gamma}
\]
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,
\[
III \leq \max i \sum j |\Delta \Sigma^k_{ij}| + |\hat{\Sigma}^k_{ij}| 1(|\Sigma^k_{ij}| \leq u, |\hat{\Sigma}^k_{ij}| \geq u)
\]
\[
\leq \gamma(\delta) \max i \sum j 1(|\Sigma^k_{ij}| \geq u - \gamma(\delta)) + su^{1-q}
\]
\[
\leq \gamma(\delta) \frac{u^{-q}}{1 - \gamma(\delta)/u} + su^{1-q}
\]
where (4) has been used.

We now use \( u = 2\gamma(\delta) \) to obtain the bound. If \( \Sigma^k \) 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
\[
||U_n(\Sigma^k)^{-1} - (\Sigma^k)^{-1}||_2 \leq \Omega \left( ||U_n(\Sigma^k) - \Sigma^k||_2 \right)
\]

**Dense Transition Matrix**

With probability greater than \( 1 - 2\delta \) both, maximum value of \( \Delta \Sigma^0 = \Sigma^0 - \Sigma^0 \) and \( \Delta \Sigma^1 = \Sigma^1 - \Sigma^1 \) are less than \( \gamma \). We have also seen that \( ||\Delta \Sigma^0||_2, ||\Delta \Sigma^1||_2 \leq O(\sqrt{n}) \).

As mentioned in [1], we get
\[
||\Delta \Sigma^0||_2 \leq ||\Sigma^0||_2^2 ||\Delta \Sigma^0||_2 \leq \frac{4\sqrt{n}}{\sigma^2_{\min}}
\]
This is true when \( ||\Delta \Sigma^0||_2 \leq \lambda_{\min}(\Sigma^0) \) and \( \Sigma^0 \) is invertible.

The error is given by,
\[
||\hat{A} - A||_2 \leq ||\hat{\Sigma}^1 \hat{\Sigma}^0 - \Sigma^1 \Sigma^0||_2 + ||\Sigma^1 \hat{\Sigma}^0 - \Sigma^1 \Sigma^0||_2
\]
\[
\leq (||\Delta \Sigma^0||_2 + ||\Sigma^1||_2) ||\Delta \Sigma^1||_2 + ||\Sigma^1||_2 ||\Delta \Sigma^0||_2
\]
\[
\leq \frac{4\sigma_{\max}(\Sigma^0) ||\Delta \Sigma^1||_2}{\sigma_{\min}(1 - \sigma_{\max}^2)}
\]
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 \( \Sigma^0, \Sigma^1 \).

Let \( \gamma \) be the maximum deviation of empirical covariance matrices as earlier.

We show that \( A^T = \Sigma^0 \Sigma^1 \) is a feasible solution with high probability.
\[
||\Sigma^0 A^T - \Sigma^1||_{\max} \leq ||(\Sigma^0 - \Sigma^0)A||_{\max} + ||(\Sigma^1 - \Sigma^1)||_{\max}
\]
\[
\leq \gamma(||A||_1 + 1) = \lambda
\]
Clearly, $||\hat{A}||_1 \leq ||A||_1$ with high probability. We also obtain,

$$\|\hat{A} - A\|_{\text{max}} = \|\Sigma^0_1 (\Sigma^0 A^\top - \Sigma^1_1)\|_{\text{max}}$$

$$= \|\Sigma^0 (\Sigma^0 A^\top - \Sigma^0_1 A^\top + \Sigma^0_1 A^\top - \Sigma^1_1 + \Sigma^1_1 - \Sigma^1_1)\|_{\text{max}}$$

$$\leq 2\lambda\|\Sigma^0\|_1 = \lambda_1$$

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

$$\|a - a\|_1 \leq \|\hat{a} - T\|_1 + \|\hat{a} - a\|_1 + \|\hat{a} - a\|_1$$

$$\leq 2\|\hat{a} - T\|_1 + 2\|\hat{a} - a\|_1 \leq 2\|\hat{a} - T\|_1 + 2\|\hat{a} - a\|_1$$

Consider sum

$$s_n = \sum_i \min\left(\frac{|a_i|}{\lambda_1}, 1\right)$$

$$\leq \lambda_1^{-q} \sum_i |a_i|^q = s\lambda_1^{-q}$$

Now, $\|a - a\|_1 \leq \lambda_1 s_n = s\lambda_1^{-q}$. Also, $\|a - \hat{a} - a\|_1 \leq \lambda_1 |T_j| \leq \lambda_1 s_n = s\lambda_1^{-q}$. Substituting these, we get the bound $\|\hat{A} - A\|_1 \leq 4s\lambda_1^{-q}$.

**Low Rank Transition Matrix**

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

$$\hat{A} = \arg\min_B \langle A^\top, \hat{\Sigma}^0 A^\top - 2\Sigma^1_1 \rangle + \lambda_n \|A\|_*$$

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

$$\langle \Delta^\top, \hat{\Sigma}^0 \Delta^\top \rangle \leq 2\langle \Delta^\top, \Sigma^1_1 - \hat{\Sigma}^0 A^\top \rangle + \lambda_n (\|A\|_* - \|\hat{A}\|_*)$$

$$\leq (2\|\Delta^\top - \hat{\Sigma}^0 A^\top\|_2 + \lambda_n) ||\hat{\Delta}||_*$$

As shown in appendix earlier, we get $||\hat{\Delta}||_* \leq 4\sqrt{2r}||\Delta||_F$ when $\lambda_n \geq 4(\|\Delta^\top\|_2 + \sigma_{\max}(\Delta^\top)\|\Delta^\top\|)$. Now, $\langle \Delta^\top, \hat{\Sigma}^0 \Delta^\top \rangle \leq \lambda_{\min}(\hat{\Sigma}^0)/2$ and a sufficient condition is when $\|\Delta^\top\|_2 \leq \gamma_2 < \lambda_{\min}(\Sigma^0)/2$. This happens when we have large enough number of time samples $T = \Omega(\frac{128n^3}{\lambda_{\min}^2 m^2} 1/\delta \|Q\|_2^2).$ Now $\langle \Delta^\top, \hat{\Sigma}^0 \Delta^\top \rangle \geq \lambda_{\min}(\hat{\Sigma}^0)/2 \|\hat{\Delta}\|_2^2$, which leads to the bound $||\hat{\Delta}||_F \leq 12\lambda_n \sqrt{2r}$.

**1) Covariance Matrix**

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

$$B = \{ \gamma \sum_{1 \leq i < j < n} \varepsilon_{ij} (e_i e_j^\top + e_j e_i^\top) 1_{(k-1)s \leq i < j \leq (k-1)(s+1), k \in [n/s]} + I, \varepsilon \in \{0, 1\}^{(n-1)/2} \}.$$  

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

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

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

$$\|\Sigma_e - \Sigma_{\varepsilon'}\|_2^2 \geq \frac{\gamma^2 n(s-1)}{16} > \frac{\gamma^2 n s}{4}$$

At each point in time, we observe $Z_t = \Psi_t X_t$. Alternatively, we could observe $Y_t = M_t X_t \in \mathbb{R}^m$. In the subsampling case, $M_t$ is $\Psi_t$ 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, $\Psi_t = M_t \cdot M_t^\top$ in this case. Now we can observe that $Y_t \sim \mathcal{N}(0, M_t \Sigma_0 M_t^\top)$. Also define, $\mathbb{P}_{t, \Sigma_t}(Z_t) = \mathbb{P}(M_t) \mathbb{P}_{t, \Sigma_t}(Y_t)$. As an example, we see that $\mathbb{P}_{t, \Sigma_t}(Z_t) = N(0, I_m)$. It follows from independence ($A = 0$) that $\mathbb{P}_{t, \Sigma_t}(Z_t) = \prod_{t=1}^T \mathbb{P}_{t, \Sigma_t}(Z_t)$.

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

$$D_{KL}(\mathbb{P}_{t, \Sigma_t} || \mathbb{P}_{\Sigma_0}) = \mathbb{E}_{M_t} \log \left( \frac{\mathbb{P}_{\Sigma_t}(Z_t)}{\mathbb{P}_{\Sigma_0}} \right) |M_t^\top|$$

$$= \sum_{t=1}^T \mathbb{E}_{M_t} [D_{KL}(\mathbb{P}_{t, \Sigma_t} || \mathbb{P}_{t, \Sigma_0})]$$

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

$$D_{KL}(\mathbb{P}_{t, \Sigma_t} || \mathbb{P}_{t, \Sigma_0}) = \frac{1}{2} \text{Tr}(M_t \Sigma_t M_t^\top) - \frac{1}{2} \log |M_t \Sigma_t M_t^\top| |M_t^\top|$$

$$= I_m + \gamma \sum_{i \neq j} M_t \varepsilon_{ij} e_i e_j^\top M_t^\top$$

$$= I_m + Q_t$$
where we have used
This is because row 
Here, 
Putting everything together,
For the V AR process to be stable as described in Section
\[ E[M_t e_i e_j^T M_t^T]_{kk} = E[M_t k, i (M_t)_{k,j}] \]
\[ = 0 \text{ when } i \neq j. \] (6)
\[ A \] is a uniformly chosen unit vector
\[ \text{Symmetry dictates} \]
\[ D_{KL}(P_{t,S} || P_{t,S_0}) = -\frac{1}{2} \log |I_m + Q_t| \]
\[ = -\frac{1}{2} \sum_{i=1}^r \log(1 + \lambda_i) \leq \frac{1}{4} \sum_{i=1}^r \lambda_i^2 - 2\lambda_i \]
\[ \Rightarrow E[\frac{1}{4} \sum_{i=1}^r \lambda_i^2 - 2\lambda_i] \leq \frac{1}{4} E[\|Q_t\|_F^2] \leq \frac{\gamma^2 n(s-1)m^2}{2n^2} \]
For the last step, we use (7) and (8) detailed below.
\[ E[\|Q_t\|_F^2] \leq \gamma^2 \sum_{a,b \in [m]} E \left[ \left( \sum_{i \neq j} \varepsilon_{i,j} (M_{a,i} M_{b,j}) \right)^2 \right] \]
\[ \leq \gamma^2 \sum_{a,b \in [m]} \varepsilon_{a,b}^2 E[(M_{a,b})^2] \]
\[ \leq \gamma^2 n(s-1)m^2 E[(M_{a,b})^2] \]
(7)
where we have used \[ E[(M_{a,b})^2] = 0. \] Now, \[ (M_{a,b})^2 \sim \text{Beta}(\frac{n}{2}, \frac{n-1}{2}). \] Using this and cauchy inequality,
\[ E[(M_{a,b})^2] \leq E[(M_{a,b})] \]
\[ \leq \frac{2}{n^2}. \] (8)
Putting everything together,
\[ D_{KL}(P_{t,S} || P_{t,S_0}) \leq \frac{\gamma^2 T n(s-1)}{2n^2} \]
\[ \leq \frac{n(s-1)}{16} = c \log |B_\gamma| \]
A. Transition Matrix
We consider a class \( A \) 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:
\[ A = \left\{ \gamma \sum_{i,j \in [n]} \varepsilon_{i,j} e_i e_j^T (k-1)_{s \leq i < j \leq (k-1)(s+1), k \in [n/s]} \varepsilon \in \{0,1\}^{ns} \right\} \]
Here, \( \gamma = cn/m \sqrt{T} \). We require that \( \|A_c\|_F \leq \sigma_{\max} < 1 \) for the VAR process to be stable as described in Section 2.5.2. Seeing \( \|A_c\|_F \leq \|A_c\|_2 \leq \sigma_{\max} < 1 \), we require that \( T = \Omega(n^3/m^2) \). From the Gilbert-Varshamov theorem, we know that there exists \( A_c \subseteq A \) with \( |A_c| \geq 2^{ns/8} \) and for \( A_c, A_{c'} \in A_c \),
\[ \|A_c - A_{c'}\|_F^2 \geq \frac{\gamma^2 n^2}{8} \]
If we write out the KL divergence, it is almost identical to the previous case. We obtain
\[
D_{KL}(P_{\lambda}, P_{A_t}) \leq \frac{2TM^2\gamma^2n^2}{n^2(1 - \sigma^2_{\text{max}})}
\leq \frac{c't\text{tr}}{8} = c' \log |\mathcal{A}|.
\]

**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
\[
\mathcal{S} = \left\{(\sum_{i \leq s} |\Sigma_{ij}|^q \leq s \right\}
\]

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 \( \Lambda \). To be more precise,
\[
\Lambda = \left\{ M \in \mathbb{R}^{r \times r} | \forall i \in [r], \sum_j |M_{i,j}|^0 = s, \forall j \in [r] \sum_i |M_{i,j}|^0 \leq 2s, M_{i,j} \notin \{0, \nu\} \right\}
\]

2) Further consider set \( \Gamma \), the set of all binary sequences of length \( r \). This set would express whether a row of a matrix \( A \) is seen.

3) For any \( \lambda \in \Lambda \), let \( \lambda_i \) represent row \( i \). Now we define matrix \( L(\lambda_i) \) as follows. Consider \( \lambda_i' \in \mathbb{R}^{1 \times n} \) where \( \lambda_i' j = \lambda_{i,j} - \lfloor n/2 \rfloor \mathbbm{1}(j \geq \lfloor n/2 \rfloor) \). Now, \( L(\lambda_i) = \lambda_i' \lambda_i' \). This means that the \( i^{th} \) row of \( L(\lambda_i) \) has the \( r \) elements of \( \lambda_i \) 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 \( \nu = \mathcal{O}(c\sqrt{\log n}) \), when \( s = \mathcal{O}(\sqrt{T \log r}) \), 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 \( X_t \sim \mathcal{N}(0, \Sigma(\theta)) \). Let \( P_{\theta} \) denote the probability of observing \( Z_t^T \). We see \( Z_t = M_t X_t \), and thus \( P_{t,\theta}(Z_t) = P(M_t) P_{t,\theta}(Z_t) \) where \( P_{t,\theta}(Z_t) = \mathcal{N}(0, M_t \Sigma_\theta m_t^T) \). We borrow some notation from earlier and write.
\[
P_{\theta}(Z_t^T) = \prod_{t=1}^{T} P_{t,\theta}(Z_t)
\]

Upon observing \( Z_t^T \), an estimator comes up with an estimate \( \hat{\Sigma}_\theta \). Observe the following sequence
\[
\max_{\theta} \mathbb{E}[\|\hat{\Sigma}_\theta - \Sigma_\theta\|_2] \geq \frac{1}{2r^2|A|} \sum_{\theta} \mathbb{E}[\|\hat{\Sigma}_\theta - \Sigma_\theta\|_2]
\]
\[
\geq \frac{1}{2r^2|A|} \sum_{\theta} \mathbb{E}[\|\hat{\Sigma}_\theta - \Sigma_\theta\|_2 / \rho(\hat{\gamma}, \gamma) \land 1 \rho(\hat{\gamma}, \gamma)]
\]
\[
\geq \min_{\rho(\hat{\gamma}, \gamma) \geq 1} \frac{\|\Sigma_\theta - \Sigma_\theta\|_2}{\rho(\hat{\gamma}, \gamma)} \frac{1}{2r^2|A|} \sum_{\theta} \mathbb{E}[\rho(\hat{\gamma}, \gamma)]
\]

Now we show for \( \rho(\hat{\gamma}, \gamma) \geq 1 \),
\[
\frac{\|\Sigma_\theta - \Sigma_\theta\|_2}{\rho(\hat{\gamma}, \gamma)} \Rightarrow \frac{\|\Sigma_\theta - \Sigma_\theta\|_2}{\rho(\hat{\gamma}, \gamma)} \Rightarrow \frac{1}{2r^2|A|} \sum_{\theta} \mathbb{E}[\rho(\hat{\gamma}, \gamma)]
\]

The choice of \( v \) here is \( v_j = 1(j \geq \lfloor n/2 \rfloor) \).

We now focus on the other term and see that
\[
\frac{1}{2r^2|A|} \sum_{\theta} \mathbb{E}[\rho(\hat{\gamma}, \gamma)]
\]
\[
\geq \frac{1}{2r^2|A|} \sum_{\theta} \sum_{i=1}^{r} \mathbb{E}[\gamma_i|M_t] + \sum_{\theta : \gamma_i = 1} \mathbb{E}[1 - \gamma_i|M_t]
\]
\[
\geq \frac{1}{2} \sum_{i=1}^{r} \mathbb{E}[M_t] \left[ \int \gamma_i \sum_{\gamma_i = 0}^{\gamma_i = 1} \frac{d\mathbb{P}_\theta}{\gamma_i} \right] + \sum_{\theta : \gamma_i = 1} \mathbb{E}[1 - \gamma_i|M_t]
\]
\[
\geq \frac{1}{2} \sum_{i=1}^{r} \mathbb{E}[M_t] \left[ 1 - D_{TV}(\mathbb{P}_\theta, \mathbb{P}_\theta, \gamma_i = 0, \gamma_i = 1) \right]
\]

Here \( D_{TV}(\mathbb{P}_\theta, \gamma_i = 0, \gamma_i = 1) \) is the total variation distance.

It is easy to see that the total variation distance between mixture distributions is less than the total variation distance between constituents leading to
\[
D_{TV}(\mathbb{P}_\theta, \gamma_i = 0, \gamma_i = 1) \leq \sum_{\theta : \gamma_i = 1} D_{TV}(\mathbb{P}_\theta, \gamma_i = 0, \gamma_i = 1, \lambda_i = 1)
\]
\[
\leq \min_{\gamma_i = 1} D_{TV}(\mathbb{P}_\theta, \gamma_i = 0, \gamma_i = 1, \lambda_i = 1)
\]

We now use the following relation between distances between measures
\[
D_{TV}(\mathbb{P}_a, \mathbb{P}_b) \leq \sqrt{D_{TV}(\mathbb{P}_a, \mathbb{P}_b)} = \mathbb{E}_{\mathbb{P}_b}(d\mathbb{P}_a/d\mathbb{P}_b)^2 - 1
\]
We now study what the distributions we are considering look like. \( P'_{t, \gamma_1=0, \gamma_1, \lambda_1} = \prod_t P'_{t, \gamma_1=0, \gamma_1, \lambda_1} \), the latter is a single multivariate distribution with the covariance matrix,

\[
\Sigma_0 = \begin{cases} 
1 & i = j \\
0 & e_1 \notin M_t \\
M_t S'M_t' & e_1 \notin M_t \\
M_t S'M_t' & \end{cases}
\]

where \( M_t = [M_{t,1}; M_{t,-1}] \) and \( S \) is a symmetric matrix dependent on \((\lambda_1, \gamma_1, -1)\) 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 \( P'_{t, \gamma_1=1, \gamma_1, \lambda_1} \) is a mixture of distributions of a number of Gaussians. Suppose \( n_{\lambda_1} \) is the number of columns in \( \lambda_1 \) 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 \( \lambda_1 \) that are not in these \( n_{\lambda_1} \) positions. The maximum number is given by \((r/2s) = (n/4s)\). Each of these distributions has this form

\[
\Sigma_i = \begin{cases} 
1 & i = j \\
0 & e_1 \notin M_t \\
M_t S'M_t' & e_1 \notin M_t \\
M_t S'M_t' & \end{cases}
\]

We see that if \( e_1 \notin M_t \), distributions \( P'_{t, \gamma_1=0, \gamma_1, \lambda_1} = P'_{t, \gamma_1=1, \gamma_1, \lambda_1} \) 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

\[
R(t, \gamma_1, \lambda_1, \lambda'_1) = |I - \Sigma_0^{-2}(\Sigma_{\lambda_1} - \Sigma_0)(\Sigma_{\lambda_1'} - \Sigma_0)|^{-1/2}
\]

We can now write

\[
\mathbb{E}_{\gamma_1, \lambda_1} \left[ \int \left( \frac{P_{t, \gamma_1=1, \gamma_1, \lambda_1}}{P_{t, \gamma_1=0, \gamma_1, \lambda_1}} \right)^2 dP_{t, \gamma_1=0, \gamma_1, \lambda_1} - 1 \right] \leq \mathbb{E}_{\lambda_1, \lambda'_1} \mathbb{E}_{\gamma_1, \lambda_1} |R(t, \gamma_1, \lambda_1, \lambda_1') - 1|
\]

Here is an observation:

\[
R(t, \gamma_1, \lambda_1, \lambda_1') = R'(t, \gamma_1, \lambda_1, \lambda_1') |I - ((\Sigma_{\lambda_1} - \Sigma_0)(\Sigma_{\lambda_1'} - \Sigma_0))|^{-1/2}
\]

As proven in Lemma 11 of [3],

\[
\mathbb{E}_{\lambda_1, \lambda'_1} |R(t, \gamma_1, \lambda_1, \lambda_1')| \leq 1.5
\]