Fundamental Estimation Limits in Autoregressive Processes with Compressive Measurements

Milind Rao *, Tara Javidi†, Yonina C. Eldar‡, and Andrea Goldsmith*
* Electrical Engineering, Stanford University, Stanford, CA
† Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA
‡ Electrical Engineering Technion, Israel Institute of Technology, Haifa, Israel
E-mail: milind@stanford.edu, tjavidi@ucsd.edu, yonina@ee.technion.ac.il, andrea@wsl.stanford.edu

***Abstract*—We consider the problem of estimating the parameters of a vector autoregressive (VAR) process from low-dimensional random projections of the observations. This setting covers the cases where we take compressive measurements of the observations or have limits in the data acquisition process associated with the measurement system and are only able to subsample. We first present fundamental bounds on the convergence of any estimator for the covariance or state-transition matrices with and without considering structural constraints of sparsity and low-rankness. We then construct an estimator for these matrices or the parameters of the VAR process and show that it is order optimal.**

***Index Terms*—system identification, covariance estimation, autoregressive processes, high-dimensional analysis, robust estimation, minimax theory.**

I. Introduction

A Vector Autoregressive (VAR) process is characterized by a finite set of parameters that describe the linear relation between present and future values of a state vector $x_t \in \mathbf{R}^n$ as

$$x_{t+1} = Ax_t + w_t \quad w_t \overset{\text{iid}}{\sim} \mathcal{N}(0, Q_w),\ 1 \le t \le T, \qquad (1)$$

where A is the *state-transition* matrix. VAR processes have been used as models in finance, econometrics, neuroscience,and bioinformatics among other areas [1], [2] because of their expressive and predictive power. A central problem in these applications is to identify the state-transition matrix A or equivalently the stationary covariance matrices $\Sigma^k = \mathbb{E}[x_t x_{t+k}^\intercal]$ as quickly as possible with limited samples. This is because measurement systems attempting to capture such models, e.g. in wireless sensor networks, may have communication, energy, or hardware constraints in collecting or communicating measurements to a central fusion center. This is especially relevant if we assume the measurements from a large number of sensors arise from a high-dimensional VAR process. One strategy to deal with this constraint is to sample measurements or take compressive measurements of the state vector x_t. This sampling or compression procedure may or may not be under our control. These limitations on data acquisition motivate identifying or making inferences from a system with partial or compressed observations. The problem of estimating the covariance matrix is also of fundamental importance in statistics with applications in principal component analysis, classification and portfolio selection [3].

This research is funded in part by the Texas Instruments Stanford Graduate Fellowship, NSF Center for Science of Information grant NSF-CCF-0939370, and NSF under CPS Synergy grant 1330081. The authors would like to thank the anonymous reviewers for their very helpful comments

We assume that we have access to m-dimensional projections ($m \ll n$) of the state vector x_t, which models the scenario where we have access to a few components or a few compressive measurements. Our goal is to use them to estimate the stationary covariance matrix and the transition matrix A.

A. Contribution

We first analyse the fundamental rates of convergence of any estimator of the covariance and transition matrices from $m-$dimensional projections. We show that if $T = o(n^3/(m^2\epsilon^2))$, then for any estimator, there exists processes for which the error in the estimate (of A or Σ^0) exceeds ϵ with non-negligible probability. This implies that the minimum number of time samples needed to get within a given estimation error scales with the dimension n and inverse quadratically with the number of dimensions we see at each point in time. We can remove this dependence on dimension by imposing a structural constraint of sparsity or low-rankness on the matrix we are estimating to reduce the number of minimum time samples required for estimation. Finally, we construct estimators of these matrices and present refinements when we have structural assumptions. We show that the accuracy of these estimators is optimal and matches the fundamental bound up to logarithmic factors in the order of n, T, m, and constants describing structural constraints. This provides an upper bound guarantee on the number of time samples required for a given estimation accuracy. We note that these results extend to the case with iid observations that is widely studied in the literature by taking $A = 0$. Proofs are omitted and can be found in [4].

B. Related Work

Subsampling measurements is a common approach to dealing with high-dimensional systems [5]. The use of compressive measurements (or low-dimensional projections) was introduced in [6] to estimate the subspace of *iid, bounded* observations. We use this compressive measurement approach to estimate parameters of our VAR process.

The estimation of parameters of VAR processes with complete observations is well studied: a least-squares procedure for estimating A was first proposed in [7]. These results were obtained in the regime where the number of time samples T is larger than the dimension n of the process. When n is larger than T, structural assumptions on A are required for identifiability. When A is sparse, many estimation methods have been proposed, including a 2-stage approach for fitting sparse models [8], lasso regularization [9] and a Dantzig estimator for weakly sparse matrix estimation [10]. In [11], the authors

consider the question of sparse transition matrix estimation for a continuous time VAR process. The work [12] showed how spectral density functions influence the rate of convergence. The case where A is low-rank and measurements are fully available was considered in [13], [14]. We recover bounds of [10], [14] in the full observation case in terms of scaling in n, T and structural constraint constants. Asymptotic analysis of covariance estimation for the simpler scalar autoregressive process with sampling is presented in [15]. Works [16], [17] considered the problem of VAR system identification from subsampled observations, which included multiplicative and additive noise. The estimators were used in joint system-state identification and validated via simulation.

We obtain estimates of the covariance matrix of the observations in this work. With complete observations, it has been shown that the empirical covariance is an asymptotically consistent estimator only in the low dimensional regime ($n \ll T$) [18]. Sparse covariance matrix estimation from *iid* observations was considered in [19], [20]. We recover the results of [19] in the case of independent Gaussian samples ($A = 0$) with full observations ($m = n$) in terms of scaling with n, T and structural constraint constants. We obtain fundamental bounds for sparse covariance matrices by following principles laid out in [21] and match these for the full observation case. Structured covariance estimation with full observations has been considered in stationary processes [3] but not for VAR processes. Low rank covariance matrix estimation with missing iid data was treated in [22] and lower bounds on covariance estimation error were provided.

This work extends [16], [17] by considering the case of compressive measurements. This paper also extends the structured covariance results to VAR processes observed with compressive measurements. Finally, fundamental performance bounds are provided that follow the principles laid out in [22], [23], [21].

The rest of this paper is organized as follows: the problem description is provided in Section II. In Section III, minimax lower bounds on the rates of convergence are presented for any estimator of Σ^k and A. Section IV presents an achievable estimator along with performance guarantees.

Notation Operator norm is denoted by $\|\cdot\|_2$, Frobenius norm by $\|\cdot\|_F$, maximum element $\max_{i,j}|A|_{ij}$ by $\|A\|_{\max}$, nuclear norm by $\|\cdot\|_*$, the ℓ_1 to ℓ_1 norm is denoted by $\|A\|_1$ which is also $\max_j \sum_i |A_{ij}|$, the ℓ_∞ to ℓ_∞ norm which is also $\|A^\intercal\|_1$ is denoted by $\|A\|_\infty$. The zero matrix or vector is denoted by $\mathbf{0}$ and the subscript when provided denotes size. Term $\mathbf{1}(\cdot)$ evaluates to one if the condition in the parenthesis is true and zero otherwise. The indicator vector is e_i where $(e_i)_j = \mathbf{1}(i = j)$. Kronecker product is $\otimes$ and $\circ$ denotes the Schur or elementwise product. Consider functions f_n, g_n of variables $n = (n_1, n_2, \ldots, n_m)$. $f_n = \mathcal{O}(g_n)$ denotes that there exists constant $c > 0$ independent of n such that $f_n \leq cg_n$. Similarly $f_n = \Omega(g_n)$ implies there exists constant $c > 0$ such that $f_n > cg_n$. The set $[q] = \{1, 2. \ldots, q\}$.

II. Problem Description

Consider a vector autoregressive process with state vector $x_t \in \mathbf{R}^n$ evolving as (1) where the noise vector w_t is a zero-mean normally distributed variable. The transition matrix A and covariance matrix Q_w are unknown. It is assumed that $\|A\|_2 = \sigma_{\max} < 1$. This is a sufficient condition to ensure the spectral radius of A is bounded by 1 and the VAR process is stable [10], [11]. Note that if $\sigma_{\max} = 0$, then the observations are independent across time.

Alternatively, the stationary VAR process can be viewed as a Gauss-Markov vector valued stochastic process with covariance matrix $\Sigma^k = \mathbb{E}[x_t x_{t+k}^\intercal]$ satisfying the Yule-Walker equations:

$$\begin{aligned} \Sigma^0 &= A\Sigma^0 A^\intercal + Q_w \\ \Sigma^{k+1} &= \Sigma^k A^\intercal. \end{aligned} \tag{2}$$

The system is initiated at $x_0 = 0$.

In our model, we observe $m-$dimensional projections of the state vector at each point in time. These projections may be randomly chosen using a process that is independent from the innovation noise process w_t. In other words as $x_{t+1} = Ax_t + w_t$, we observe $z_t = \Psi_t x_t$, where $z_t \in \mathbf{R}^n$ and $\Psi_t \in \mathbf{R}^{n\times n}$ is an $m-$dimensional projection matrix. The data model may represent two scenarios:

1) Subsampling: In the simplest case, the low-dimensional projection could be viewing m components out of n. In this case Ψ_t is a diagonal binary matrix with m of the diagonal components uniformly and randomly chosen to be one. In other words, consider a subset of m indices $\{\psi_1, \ldots, \psi_m\} \in [n]^m$ uniformly chosen at each time instant, and let

$$\begin{aligned} \Psi_t &= \operatorname{diag}\left(\sum_{j=1}^m e_{\psi_j}\right) \\ z_t &= \Psi_t x_t = \operatorname{diag}\left(\sum_{j=1}^m e_{\psi_j}\right) x_t. \end{aligned} \tag{3}$$

Here, the measurement and communication costs at each instant are $\mathcal{O}(m)$.

2) Compressive Measurements: Suppose we take m compressive measurements at each instant in time. The compressive measurement is $y_t = M_t x_t$ where $y_t \in \mathbf{R}^m$, $M_t \in \mathbf{R}^{m\times n}$ and the rows $(M_t)_i$ are picked uniformly and independently from the surface of an $n-$dimensional sphere. An alternate way of characterizing y_t is through

$$z_t = M_t^\intercal (M_t M_t^\intercal)^{-1} y_t = \Psi_t x_t, \tag{4}$$

where z_t is what we observe using the random m-dimensional projection $\Psi_t = M_t^\intercal (M_t M_t^\intercal)^{-1} M_t$. In this case, the measurement cost at time t is $\mathcal{O}(m)$. The communication cost is $\mathcal{O}(m)$ if the central fusion centre has access to M_t or $\mathcal{O}(n)$ otherwise.

To summarize, our goal is to propose and analyse algorithms to estimate the transition matrix A and the stationary covariance matrix $\Sigma^0 = \mathbb{E}[x_t x_t^\intercal]$ from a finite number of samples z_t and find achievable and fundamental bounds on $\|\hat{A} - A\|_2, \|\hat{\Sigma}^0 - \Sigma^0\|_2$. We investigate the case where we do not make any structural assumptions on A, Σ^0 as well as the setting where A or Σ^0 are $s-$sparse, meaning they have at most s non-zero components in each row or they are rank-r matrices.

III. Fundamental Bounds in Estimation

In this section, we focus on obtaining lower bounds on estimator error for the estimation of the transition matrix A and the stationary covariance matrix Σ^0 from partial measurements. This will give us the minimum number of time samples we need to observe below which any estimate is inaccurate with non-negligible probability for some systems. We focus on the case

where we see a random m components of the n dimensional state vector x_t at each point in time $z_t = \Psi_t x_t$ where Ψ_t arises in the sampling case as in (3) or the compressive measurements scenario (4) but with M_t having orthogonal rows. We term the latter observations as orthogonal compressive measurements.

We first have a theorem on the error of any estimator of the covariance matrix.

Theorem 1. *Consider observations $z_t = \Psi_t x_t$ which model either subsampling in* (3) *or orthogonal compressive measurements* (4)*. Let $\mathcal{B}_1$ represent the class of $s-$sparse covariance matrices where $s = \mathcal{O}(m\sqrt{T}/n)$. Then there exists constants $b, c > 0$ such that for any estimator $\hat{\Sigma}^0$*

$$\inf_{\hat{\Sigma}^0} \sup_{\Sigma^0 \in \mathcal{B}_1, A, Q_w} \mathbb{P}_{\Sigma^0, A, Q_w} \left(\|\hat{\Sigma}^0 - \Sigma^0\|_2 \geq c \frac{n}{m} \sqrt{\frac{s}{T}} \right) \geq b.$$

Moreover, let the covariance matrix belong to a class of rank r positive semidefinite matrices $\mathcal{B}_2$. When $r = \mathcal{O}(m\sqrt{T}/n)$,

$$\inf_{\hat{\Sigma}^0} \sup_{\Sigma^0 \in \mathcal{B}_2, A, Q_w} \mathbb{P}_{\Sigma^0, A, Q_w} \left(\|\hat{\Sigma}^0 - \Sigma^0\|_F \geq c \frac{n}{m} \sqrt{\frac{nr}{T}} \right) \geq b.$$

This theorem states that with non-negligible probability b, any estimator will incur an error of at least $\Omega\left(\frac{n}{m}\sqrt{\frac{s}{T}}\right)$ for some values of sparse Σ^0. We can further improve the bound in terms of s as $\Omega(s)$ instead of $\Omega(\sqrt{s})$ as done in Appendix D of [4] using the proof technique of [21]. This implies that if the number of observations T scales as $o(\frac{n^2 s}{m^2 \epsilon^2})$, the error incurred by any estimator for some Σ^0 is greater than ϵ. Similarly, the theorem states that the number of observations cannot be $o(\frac{n^3 r}{m^2 \epsilon^2})$ if we require that $\|\hat{\Sigma}^0 - \Sigma^0\|_F \leq \epsilon$ for some estimator $\hat{\Sigma}^0$ and all values of Σ^0. One observation we make is that the minimum number of time instants required scales inverse quadratically with the number of low-dimensional projections m at each time instant; i.e. if m halves, T increases by a factor of 4. We can also see the gain of applying structural constraints of sparsity or low-rankness; the error of the optimal estimator is reduced by a factor of at least n/s or n/r in either case.

We next consider estimation of the transition matrix with various constraints. We note that the covariance matrix is not completely independent from the transition matrix and assumptions on one have an impact on the other. For instance, a block-diagonal and $s-$sparse A and innovation noise matrix Q_w would lead to a block-diagonal and $s-$sparse covariance matrix. However, it is possible to have sparse covariance matrices for full-rank and dense transition matrices and vice versa. In different circumstances, we may be compelled to make structural assumptions on either the covariance matrix or the transition matrix.

Theorem 2. *Consider the same observation process as in Theorem 1. Let $\mathcal{A}_1$ represent the class of transition matrices A which are $s-$sparse and the innovation noise process is $Q_w = I$. For $s = \mathcal{O}(m\sqrt{T}/n)$, there exist constants $c, b > 0$ such that for any estimator $\hat{A}$*

$$\inf_{\hat{A}} \sup_{A \in \mathcal{A}_1, Q_w} \mathbb{P}_{A, Q_w} \left(\|\hat{A} - A\|_2 \geq c \frac{n}{m} \sqrt{\frac{s}{T}} \right) \geq b.$$

Moreover, let $\mathcal{A}_2$ denote the class of rank r transition matrices such that $r = \mathcal{O}(m^2 T/n^3)$. Then

$$\inf_{\hat{A}} \sup_{A \in \mathcal{A}_2, Q_w} \mathbb{P}_{A, Q_w} \left(\|\hat{A} - A\|_F \geq c \frac{n}{m} \sqrt{\frac{nr}{T}} \right) \geq b.$$

An immediate consequence of the theorem is that for an optimal estimator that matches this lower bound, the minimum number of time samples required is at least $\Omega(n^2 s/(m^2 \epsilon^2))$ for obtaining $\|\hat{A} - A\|_2 \leq \epsilon$ for the sparse case. This result, as shown later, is sharp in n, m, and T. The estimation error of the optimal estimator falls by a factor of n/s when the structural constraint of sparsity is applied and n/r when a low-rank transition matrix is considered.

Proof Outline

We use the general framework introduced by Tsybakov [23] to prove most lower bounds.

Let the probability distribution for some observations be indexed by parameters $\theta \in \Theta$. In this case, the parameters could refer to any combination of the transition matrix A, the innovation noise matrix Q_w, and the covariance matrices Σ^k. Let d represent the distance between parameters. In our case, it could refer to the 2-norm or the Frobenius norm. We focus on minimax lower bounds and proceed as follows:

1) We construct a finite set of parameters $\theta_i \in \Theta, i \in \{0, 1, \ldots, M\}$ that are adequately distant from each other

$$I = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{P}_\theta(d(\hat{\theta}, \theta) \geq \delta) \geq \inf_{\hat{\theta}} \max_{i \in [M]} \mathbb{P}_{\theta_i}(d(\hat{\theta}, \theta_i) \geq \delta).$$

If quantity $I > 0$, then there is a non-negligible probability that $d(\hat{\theta}, \theta) \geq \delta$, i.e. any estimate $\hat{\theta}$ is δ distant from the optimal value. Further, if $d(\theta_i, \theta_j) \geq 2\delta$, then we use the closest-distance estimator $\Upsilon^* = \operatorname{argmin}_{\theta_i} d(\hat{\theta}, \theta_i)$ to conclude that for any estimator Υ,

$$I \geq \inf_{\Upsilon} \max_i \mathbb{P}_{\theta_i}(\Psi \neq \theta_i) = p_e.$$

2) We show that the KL divergence between the probabilities indexed by θ_0 and $\theta_i, i \in [M]$, is bounded by:

$$\frac{1}{M} \sum_{i=1}^{M} D_{KL}(\mathbb{P}_{\theta_i} \| \mathbb{P}_{\theta_0}) \leq \frac{\log M}{8}. \tag{5}$$

As shown in [23], this implies that $p_e > 0$.

We now list the subset of the parameter class that we include to prove bounds with various structural assumptions. In Appendix C, we show that bound (5) is met in each of these cases.

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 rank r matrices detailed below

$$\mathcal{B} = \left\{ \gamma \sum_{1 \leq i < j \leq r} \varepsilon_{i,j} (e_i e_j^\intercal + e_j e_i^\intercal) + \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}, \varepsilon \in \{0, 1\}^{\frac{r(r-1)}{2}} \right\}.$$

This class of matrices is non-zero only in the upper left corner. Here $\gamma = c(m^2 T/n^2)^{-1/2}$ is a parameter.

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

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

$$\|\Sigma_\varepsilon - \Sigma_{\varepsilon'}\|_F^2 \geq \frac{\gamma^2 r(r-1)}{8} > \frac{\gamma^2 r^2}{16}.$$

This is a subset of the class of $s-$sparse matrices if $s = r$. Observe additionally that $\|\Sigma_\varepsilon - \Sigma_{\varepsilon'}\|_2 \geq \frac{\gamma\sqrt{s}}{4}$

2) Sparse 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 $Q_w = I$. For convenience, we assume s divides n but the proof can easily be extended to relax this assumption. The transition matrix comes from the 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 II. Noting that $\|A_\varepsilon\|_2 \leq \|A_\varepsilon\|_1 \leq s\gamma$, we require that $T = \Omega(n^2 s^2/m^2)$.

Any matrix $A_\varepsilon \in \mathcal{A}$ is indexed by an $ns-$dimensional binary vector ε. From the Gilbert-Varshamov theorem, we construct 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,

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

3) 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. Now,

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

with $\gamma = c\sqrt{rn/Tm^2}$. 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 $|\mathcal{A}_\mathcal{E}| \geq 2^{nr/8}$ such that for $A_\varepsilon, A_{\varepsilon'} \in \mathcal{A}_\mathcal{E}$,

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

IV. Achievable Estimation Error Bounds

In this section, we construct an estimator of the k correlation matrix $\Sigma^k \triangleq \mathbb{E}[x_t x_{t+k}^\intercal]$ for a stationary VAR process and then use these estimates for finding the transition matrix A. We find the non-asymptotic error bounds and show that our estimate is optimal in an order sense. We assume that we have 2 independent views z_t^1 and z_t^2 of the state vector ($z_t^i = \Psi_t^i x_t$) at each time instant. This could be considered equivalent to having a view of $2m$ low-dimensional samples at each time instant. These views could model subsampling (3) or compressive measurements (4).

We construct an estimate of the covariance matrix as

$$\hat{\Sigma}^k = \frac{n^2}{m^2(T-k)} \sum_{t=1}^{T-k} z_t^1 z_{t+k}^{2\intercal}. \tag{6}$$

Clearly, in the sub-sampling case, this scaled version of the empirical covariance matrix would be unbiased as the probability of observing the i^{th} component in z_t^1 and the j^{th} component of z_t^2 that allows for the estimation of Σ_{ij}^k is m^2/n^2.

To see that this is the case for the random low-dimensional projection, any low rank projection matrix can be written as $UU^\intercal$, where $U = \Omega \begin{bmatrix} I_m \\ 0 \end{bmatrix}$ for Ω a random rotation matrix including permutations. Now, Ωx_t can be considered as a random rotation of x_t with magnitude $\|x_t\|_2$. A vector uniformly

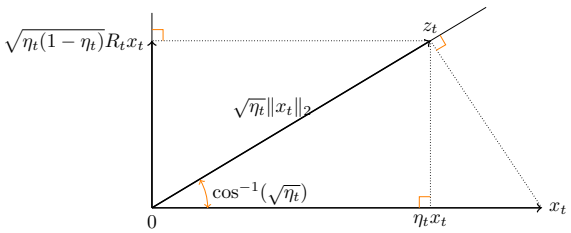

Fig. 1: Geometric depiction of projection z_t.

located on a spherical surface can be generated by a normalized Gaussian vector with iid components. Thus, we see that

$$z_t^1 = \Omega^\intercal \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \frac{\|x_t\|_2}{\sqrt{\sum_{i=1}^n u_i^2}} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

$$\Rightarrow \|z_t^1\|_2^2 = \|x_t\|_2^2 \frac{\sum_{i=1}^m u_i^2}{\sum_{i=1}^n u_i^2} = \|x_t\|_2^2 \eta_t.$$

Here $u_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ are the components of a uniform vector on a hypersphere. The variable $\eta_t \overset{\text{iid}}{\sim} \text{Beta}\left(\frac{m}{2}, \frac{n-m}{2}\right)$ is the ratio of chi-squared distributions $\frac{\sum_{i=1}^m u_i^2}{\sum_{i=1}^n u_i^2}$.

The angle θ between x_t and z_t^1 is given by $\cos^{-1}\left(\frac{\|z_t^1\|_2}{\|x_t\|_2}\right) = \cos^{-1}(\sqrt{\eta_t})$. Thus we can write,

$$\begin{aligned} z_t^1 &= \langle z_t^1, \frac{x_t}{\|x_t\|_2} \rangle \frac{x_t}{\|x_t\|_2} + \langle z_t^1, R_t^1 \frac{x_t}{\|x_t\|_2} \rangle \frac{x_t}{\|x_t\|_2} \\ &= x_t \eta_t + \sqrt{\eta_t(1-\eta_t)} R_t^1 x_t. \end{aligned}$$

Here R_t^1 is a rotation matrix and $R_t^1 \frac{x_t}{\|x_t\|_2}$ is distributed on the hypersphere and orthogonal to x_t. Also, $z_t^2 = \omega_t x_t + \sqrt{\omega_t(1-\omega_t)} R_t^2 x_t$. This is depicted in Fig. 1. The variable $\eta_t, \omega_t \overset{\text{iid}}{\sim} \text{Beta}\left(\frac{m}{2}, \frac{n-m}{2}\right)$. From the fact that $\mathbb{E}[\eta_t \omega_t] = m^2/n^2$ and in expectation the cross-terms (eg. $x_t x_t^\intercal R_t^1$) are zero by symmetry, we see that the estimator is unbiased.

The following theorem presents error bounds for $\Delta\Sigma^k \triangleq \hat{\Sigma}^k - \Sigma^k$ which we term the error in the estimate of the covariance estimate.

Theorem 3. *With probability at least $1-\delta$ we have*

$$\|\Delta\Sigma^k\|_{\max} \leq \gamma = \mathcal{O}\left(\sqrt{\frac{\log(n^2/\delta)}{(T-k)}} \frac{n\|Q_w\|_2}{m(1-\sigma_{\max})^2} \right)$$

$$\|\Delta\Sigma^k\|_2 \leq \gamma_2 = \mathcal{O}(\sqrt{n}\gamma). \tag{7}$$

Theorem 3 implies that the number of time samples needed for $\|\Delta\Sigma^k\|_2 \leq \epsilon$ is $\mathcal{O}(\frac{n^3 \log n}{m^2 \epsilon^2 (1-\sigma_{\max})^4})$. The main point to note is that this achievable error bound is off only by a factor of $\log n$ compared to the optimal estimator of Section III. Also, as $\sigma_{\max} \to 1$, we need more samples and this can be understood as the samples present less new information if there is strong dependency. In Fig. 2, we see the performance of this estimator with low-dimensional projections when $n = 10$ averaged over 10 trials. It can be seen that error depends inversely on $T^{-1/2}m$.

Table I, collating results from [17], shows how we can refine this estimate of the covariance matrix with a sparsity constraint as well as estimating A with or without constraints of sparsity and low-rank. In the sparse Σ estimator, observe that if $q = 0$, there are s non-zero values in each row and column. We would

Structural Constraint	**Refinement**	**Convergence Result** w.h.p.
Sparse Covariance Matrix $\sum_j \lvert\Sigma_{i,j}\rvert^q \leq s, \sum_j \lvert\Sigma_{j,i}\rvert^q \leq s \ \ \forall i$	Thresholding $U(\Sigma) = [\Sigma_{i,j}\mathbf{1}(\lvert\Sigma_{i,j}\rvert \geq 2\gamma)]_{i,j\in[n]}$	$\lVert\Delta\Sigma\rVert_2 = \mathcal{O}(s[\gamma]^{1-q})$
Dense Transition Matrix Σ^0 is full rank	$\hat{A}^\intercal = \hat{\Sigma}^{0\dagger}\hat{\Sigma}^1$ when $T = \Omega(n^2 \log n/m^2)$	$\lVert\Delta A\rVert_2 = \mathcal{O}(\gamma_2)$
Low Rank Covariance Matrix, Σ^0 is rank$-r$	Nuclear-norm regularization $\hat{\Sigma}^0 = \min_{\Sigma\succeq 0} \lVert\Sigma - \hat{\Sigma}^0\rVert_F^2 + \lambda_n\lVert\Sigma\rVert_*$, when $\lambda_n = \mathcal{O}(\gamma_2)$	$\lVert\Delta\Sigma\rVert_F = \mathcal{O}(\gamma\sqrt{nr})$
Sparse Transition Matrix $\max_{j\in[n]} \sum_{i=1}^n \lvert A_{i,j}^q\rvert \leq s, \max_{i\in[n]} \sum_{j=1}^n \lvert A_{i,j}^q\rvert \leq s, \lVert A\rVert_1 \leq A_1$	Dantzig estimator $\hat{A}^\intercal = \operatorname{argmin}_{M\in\mathbf{R}^{n\times n}} \sum_{i,j} \lvert M_{i,j}\rvert$ s.t. $\lVert\hat{\Sigma}^1 - \hat{\Sigma}^0 M\rVert_{\max} \leq A_1\gamma$	$\lVert\Delta A\rVert_2 = \mathcal{O}\left(s(A_1\gamma\lVert\Sigma^{0\dagger}\rVert_1)^{1-q}\right)$.
Low Rank Transition Matrix A is rank$-r$ and Σ^0 is full-rank	Nuclear-norm regularization $\hat{A} = \operatorname{argmin}_M \langle M^\intercal, \hat{\Sigma}^0 M^\intercal - 2\hat{\Sigma}^1\rangle + \lambda_n\lVert M\rVert_*$. when $T = \Sigma(n^3 \log n/m^2)$ and $\lambda_n = \mathcal{O}(\gamma_2)$	$\lVert\Delta A\rVert_F = \mathcal{O}(\gamma\sqrt{nr})$

TABLE I: Refinements for estimators of the covariance and transition matrix with different structural constraints.

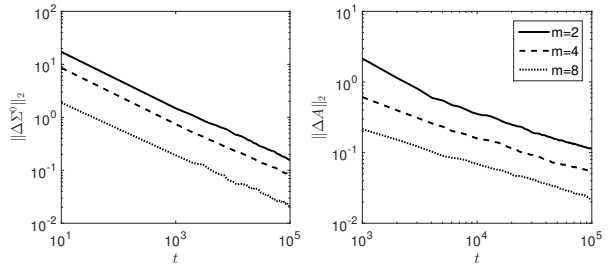

Fig. 2: Parameter estimation with low-dimensional projections, $n = 10$.

then need $\mathcal{O}(\frac{s^2n^2\log n}{m^2\epsilon^2})$ time samples for $\|\Delta\Sigma^k\|_2 \leq \epsilon$; this estimator is of the same order as the optimal estimator but for a factor of $s\log n$. In dense A estimation, for $\|\hat{A} - A\|_2 \leq \epsilon$, we need $\mathcal{O}(\frac{n^3\log n}{m^2\epsilon^2})$ samples which is again off by a $\log n$ factor from the ideal estimator. In sparse A estimation, when $q = 0$, the class being considered is $s-$sparse. The convergence rate agrees with the ideal rate up to a factor of $s\log n$. Finally, in low-rank A estimation, this estimator matches the fundamental bound up to a logarithmic order $\log n$.

V. Conclusion

We considered the problem of estimating the parameters of large vector autoregressive processes from low-dimensional random projections of the underlying state vector. This models a sub-sampling strategy where we have access to a limited number of components of the state vector as well as compressive measurement strategies. We presented minimax lower bounds on the convergence rate of any estimator of the covariance and state-transition matrices. It was seen that the number of time samples required for any estimator to get the estimation error within a required bound scales proportionally to the dimension (this was removed when structural assumptions were added) and inverse quadratically with the number of components seen. An estimator for the covariance matrix was then constructed. A refinement based on an assumption of sparsity, suitable for high-dimensional processes, was described as well as estimates for the transition matrix with and without further structural assumptions of sparsity or low-rank. We show that our estimators are optimal up to logarithmic factors in the dimension.

References

[1] C. A. Sims, "Macroeconomics and reality," *Econometrica*, vol. 48, no. 1, pp. 1–48, 1980.

[2] H. Lütkepohl, *Vector Autoregressive Models*, M. Lovric, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-04898-2_609

[3] X. Chen, M. Xu, W. B. Wu *et al.*, "Covariance and precision matrix estimation for high-dimensional time series," *The Annals of Statistics*, vol. 41, no. 6, pp. 2994–3021, 2013.

[4] M. Rao, T. Javidi, Y. C. Eldar, and A. Goldsmith, "Fundamental estimation bounds in autoregressive processes with compressive measurements: Proofs," http://stanford.edu/\~milind/reports/system_id_isit_proof.pdf, accessed: 2017-01-19.

[5] Y. C. Eldar and G. Kutyniok, "Compressed sensing: theory and applications," 2012.

[6] M. Azizyan, A. Krishnamurthy, and A. Singh, "Subspace learning from extremely compressed measurements," in *2014 48th Asilomar Conference on Signals, Systems and Computers*, Nov 2014, pp. 311–315.

[7] J. D. Hamilton, *Time series analysis*. Princeton university press, Princeton, 1994, vol. 2.

[8] R. A. Davis, P. Zang, and T. Zheng, "Sparse vector autoregressive modeling," *Journal of Computational and Graphical Statistics*, vol. 0, no. ja, pp. 1–53, 2015.

[9] S. Song and P. J. Bickel, "Large vector auto regressions," *arXiv preprint arXiv:1106.3915*, 2011.

[10] 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.

[11] J. Pereira, M. Ibrahimi, and A. Montanari, "Learning networks of stochastic differential equations," in *Advances in Neural Information Processing Systems 23*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds. Curran Associates, Inc., 2010, pp. 172–180. [Online]. Available: http://papers.nips.cc/paper/4055-learning-networks-of-stochastic-differential-equations.pdf

[12] S. Basu and G. Michailidis, "Regularized estimation in sparse high-dimensional time series models," *Ann. Statist.*, vol. 43, no. 4, pp. 1535–1567, 08 2015. [Online]. Available: http://dx.doi.org/10.1214/15-AOS1315

[13] S. Negahban and M. J. Wainwright, "Estimation of (near) low-rank matrices with noise and high-dimensional scaling," *The Annals of Statistics*, vol. 39, no. 2, pp. 1069–1097, 2011.

[14] F. Han, S. Xu, and H. Liu, "Rate optimal estimation of high dimensional time series," *Preprint*, 2016.

[15] Y. Rosen and B. Porat, "The second-order moments of the sample covariances for time series with missing observations," *IEEE Transactions on Information Theory*, vol. 35, no. 2, pp. 334–341, Mar 1989.

[16] M. Rao, A. Kipnis, T. Javidi, Y. C. Eldar, and A. Goldsmith, "Performing system identification from partial samples: Non-asymptotic analysis," in *CDC 2016*, 2016.

[17] M. Rao, T. Javidi, Y. C. Eldar, and A. Goldsmith, "Estimation in autoregressive processes from partial observations," in *ICASSP 2016, accepted*, 2016.

[18] V. A. Marčenko and L. A. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Mathematics of the USSR-Sbornik*, vol. 1, no. 4, p. 457, 1967.

[19] P. Bickel and E. Levina, "Covariance regularization by thresholding," *The Annals of Statistics*, vol. 36, no. 6, pp. 2577–2604, 2008.

[20] P. Bickel and E. Leniva, "Regularized estimation of large covariance matrices," *The Annals of Statistics*, vol. 36, no. 1, pp. 199–227, 2008.

[21] 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.

[22] K. Lounici, "High-dimensional covariance matrix estimation with missing observations," *Bernoulli*, vol. 20, no. 3, pp. 1029–1058, 08 2014. [Online]. Available: http://dx.doi.org/10.3150/12-BEJ487

[23] A. B. Tsybakov, "Introduction to nonparametric estimation. revised and extended from the 2004 french original. translated by vladimir zaiats," 2009.