ESTIMATION IN AUTOREGRESSIVE PROCESSES WITH PARTIAL OBSERVATIONS

Milind Rao[⋆] *Tara Javidi*[†] *Yonina C. Eldar*[ᵀ] *Andrea Goldsmith*[⋆]

[⋆] Electrical Engineering, Stanford University, Stanford, CA
[†] Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA
[ᵀ] Electrical Engineering Technion, Israel Institute of Technology, Haifa, Israel

ABSTRACT

We consider the problem of estimating the covariance matrix and the transition matrix of vector autoregressive (VAR) processes from partial measurements. This model encompasses settings where there are limitations in the data acquisition of the underlying measurement systems so that data is lost or corrupted by noise. An estimator for the covariance matrix of the observations is first presented. More refined estimators, factoring in structural constraints on the covariance matrix such as sparsity, bandedness, sparsity of the inverse and low-rankness are then introduced that are particularly useful in the high-dimensional regime. These estimates are then used to perform system identification by estimating the state transition matrix with or without further structural assumptions. Non-asymptotic guarantees are presented for all estimators.

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

1. INTRODUCTION

Vector Autoregressive (VAR) models are natural tools for forecasting; they have been used for this purpose in finance, econometrics, and neuroscience, as well as in other applications [1, 2]. VAR models are often used to describe high dimensional data (the dimension n is comparable to or greater than the number of time samples T). In many practical measurement systems, such as wireless sensor networks, there may be communication or energy constraints in collecting measurements. Computation constraints may limit the number of samples a central fusion centre can use at once. A common strategy to deal with this constraint is to sample measurements. Moreover, measurements may be lost or corrupted by noise. These limitations on data acquisition in measurement systems motivate identifying or making inferences from a system with partial noisy observations.

A VAR process is characterized by a finite set of parameters that describes the linear relation between present time vector-valued samples and past vector samples plus independent noise. Specifically, the state vector $x_t \in \mathbf{R}^n$ evolves 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. The stationary VAR process can alternatively be described in terms of its covariance matrix $\Sigma^k = \mathbb{E}[x_t x_{t+k}^\intercal]$ and the quantities are linked by the Yule-Walker equations. The problem of estimating the covariance matrix is of fundamental importance in statistics with applications in Principal Component Analysis, classification and portfolio selection [3]. We have access to a noisy version of each measurement with probability ρ. Our goal is to estimate the covariance matrix and the transition matrix A from these partial and sub-sampled observations of x_t.

This work was funded by the TI Stanford Graduate Fellowship, NSF under CPS Synergy grant 1330081, and NSF Center for Science of Information grant NSF-CCF-0939370.

1.1. Contribution

In this paper, we first propose an estimator for the covariance matrix of a VAR process from partial samples affected by multiplicative and additive noise. We show that the operator norm of the error of the covariance matrix estimate scales as $\frac{1}{\rho}\sqrt{\frac{n \log n}{T}}$ where ρ is the sub-sampling rate or the probability that we view an observation, n is the dimension and T is the number of time samples. If the sampling rate is halved, then the number of time samples needed for estimation of a pre-specified accuracy quadruples. We need T to be proportional to n which may be very high. To overcome this bottleneck, we provide estimators given structural constraints on the covariance matrix such as sparsity, low rank, bandedness and sparsity of the inverse covariance matrix. For these estimators, the operator norm of the error scales proportionally to $\log n$, which reduces the number of time samples needed. These results also apply to partial observations of independent Gaussian samples when $A = 0$. The estimates of the covariance matrix are then used to estimate the transition matrix. In addition, estimators for A given structural constraints of sparsity and low rank on A are presented and we show that factoring in constraints reduces the number of time samples needed for estimation. Finally, extensions to higher order VAR processes are presented and it is shown that the number of time samples needed is increased by the order of the VAR process. Due to space constraints, the proofs are omitted and can be found in [4].

1.2. Related Work

This paper extends the work [5] which considered the problem of VAR system identification from partial samples but did not include higher order VAR processes and structural constraints in covariance matrix estimation. Validation of the estimators via simulation can be found in this work. Earlier, [6] presented the asymptotic analysis of covariance estimation for the simpler autoregressive process identification with sampling.

With complete observations, it has been shown that the naive empirical covariance estimator is consistent only in the low dimensional regime ($n \ll T$) [7]. This paper extends prior work on structured covariance matrix estimation to include constraints of sparsity, bandedness and precision matrix sparsity. Prior works [8, 9, 10] deal with each of these constraints, respectively, but under noiseless and full observations. [11] provides estimators and their analysis for high-dimensional linear regression with partial *iid* observations; their estimators for covariance matrix estimation are applied to dependent data in this work and the focus of our analysis in on convergence as opposed to proving restricted eigenvalue properties. Low rank covariance matrix estimation with missing iid data was considered in [12] and lower bounds on covariance estimation error were provided. For iid samples with partial samples, [13] analyses an estimator with a sparse inverse covariance matrix assumption. This paper extends these results to partially observed VAR processes. We also *recover* these results in the case of independent Gaussian samples with full observations ($\rho = 1$) in terms of scaling with n, T and structural constraint constants. Structured covariance estimation

with full observations has been looked at in stationary processes [3] but not for VAR processes.

In the full observation and noiseless scenario, [14] first proposed the least-squares procedure for estimating A. In order to enforce identifiability when the dimension n is larger than the number of time samples T, structural assumptions are required. Methods for sparse estimation of A in VAR processes include a 2-stage approach for fitting sparse models [15], lasso regularization [16] and a dantzig estimator for weakly sparse matrix estimation [17]. In [18], authors consider the question of sparse transition matrix estimation for a continuous time VAR process. [19] showed how spectral density functions influence the rate of convergence. Low-rank transition matrix estimation was considered in [20, 21]. These works do not consider partial observations. We again recover bounds of [17, 21] in the full observation case in terms of scaling in n, T and structural constraint constants.

The rest of this paper is organized as follows: the problem description is provided in Section 2. Section 3 presents algorithms and convergence analysis to estimate the covariance matrix given structural constraints. In Section 4, it is shown how these estimators for the covariance matrices can be used to obtain the transition matrix. The algorithms are generalized to higher order VAR processes in Section 5.

Notation Trace inner product $\mathrm{Tr}(A^\intercal B)$ is denoted by $\langle A, B\rangle$. 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. For order, $f_n = \mathcal{O}(g_n)$ denotes that there exists $c > 0$ such that $|f_n| \le c|g_n|$. Similarly $f_n = \Omega(g_n)$ implies there exists $c > 0$ such that $|f_n| > c|g_n|$. The set $[q] = \{1, 2. \ldots, q\}$.

2. 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 $\|A\|_2 = \sigma_{\max} < 1$ to ensure the spectral radius of A is bounded by 1 and the VAR process is stable. 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 with x_0 satisfying $\|x_0\|_2 = o(T^{-1/2})$. This ensures that the initial effects die down. At each time instant we observe

$$z_t = P_t(x_t + v_t),$$

where $v_t \overset{\text{iid}}{\sim} \mathcal{N}(0, Q_v)$ is an additive observation noise and the covariance matrix Q_v is assumed to be known and P_t is a random measurement matrix of the form $P_t = \mathrm{diag}(p_t)$ with p_t denoting an $n-$dimensional random vector. This vector is independently sampled from a distribution $\mathcal{P}$ on *bounded* non-negative support, where it is assumed that the first and second order statistics of p_t are known. Let $\theta(k)_{ij}$ denote the average scaling due to the multiplicative noise observed in the i^{th} element of an observation and j^{th} element of an observation k instants later or $\theta(k) = \mathbb{E}[p_t p_{t+k}^\intercal]$. A quantity that will be of use later is $\theta(k)_* = \min_{ij}\theta(k)_{ij}$ which denotes the minimum scaling of an element in Σ^k.

This model can encompass several scenarios. **Scenario 1:** Each observation sample is seen independently with probability ρ ($(p_t)_i \overset{\text{iid}}{\sim} \mathcal{B}(\rho)$). This is the random sampling case and is analyzed specifically throughout the paper. Complete observations would correspond to the case where $\rho = 1$. **Scenario 2:** Each observation sample is seen with different probability ($(p_t)_i \overset{\text{indep}}{\sim} \mathcal{B}(\rho_i)$) which can model the case where it is more expensive to get observations from certain sensors. **Scenario 3:** The observation vector is seen with probability ρ ($P_t = \mathbf{I}$ w.p. ρ, $\mathbf{0}$ o.w.). This removes the independence assumption across observation samples and has been the model studied in the intermittent Kalman filtering literature [22, 23, 24]; **Scenario 4:** Observations could have bounded multiplicative noise modelling fading that occurs in wireless sensor networks. Here $0 \le p_l \le (p_t)_i \le p_u < \infty$.

To reiterate, we assume that (1) $\|A\|_2 = \sigma_{\max} < 1$; (2) the first two moments of p_t are known; (3) the covariance matrix Q_v of the additive observation noise is known; (4) the innovation noise process w_t, the additive noise process v_t and the multiplicative noise process p_t are independent from each other and across time instants. 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. We also note that because of the equivalent representation, the estimation of A and of Σ^k are closely related.

3. ESTIMATING THE COVARIANCE MATRICES

In this section, we focus on estimators of the k correlation matrix $\Sigma^k \triangleq \mathbb{E}[x_t x_{t+k}^\intercal]$ for a stationary VAR process. We note that in particular, this can be used to estimate the subspace that signals lie in by taking the top eigenvectors. The error is proportional to the error in estimating the covariance matrix through the Davis-Kahn Theorem [25].

3.1. General Case

We first consider the empirical covariance matrix S^k of the T observations $\{z_t\}_{t\in[T]}$:

$$S^k = \frac{1}{T-k}\sum_{t=1}^{T-k} z_t z_{t+k}^\intercal.$$

We observe that $\mathbb{E}[S^k] = \mathbb{E}[z_t z_t^\intercal] = \theta(k) \circ (\Sigma^k + Q_v \mathbf{1}(k=0))$. As our estimate of Σ^k, we use

$$(\hat{\Sigma}^k)_{ij} = (S^k)_{ij}/\theta(k)_{ij} - (Q_v)_{ij}\mathbf{1}(k=0). \tag{3}$$

The estimator $\hat{\Sigma}^k$ is unbiased if x_0 is initialized from the stationary distribution of x_t. In the random sampling case where observations are seen with probability ρ, this reduces to

$$\hat{\Sigma}^k = \frac{1}{\rho^2}S^k - \left(\frac{1-\rho}{\rho^2}S^k \circ I_n + Q_v\right)\mathbf{1}(k=0).$$

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

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

$$\begin{aligned}\|\Delta\Sigma^k\|_{\max} &\le \gamma(\delta)\\ \|\Delta\Sigma^k\|_2 &\le \gamma_2(\delta) = 4\sqrt{n}\gamma(\delta),\end{aligned} \tag{4}$$

where up to order $(T-k)^{-1/2}$,

$$\gamma(\delta) = \sqrt{\frac{8\log(6n^2/\delta)}{(T-k)\theta(k)_*}}p_u \max\left(\frac{\|Q_w\|_2}{(1-\sigma_{\max})^2}\|, Q_v\|_2\right).$$

Theorem 1 implies that the number of time samples needed for $\|\Delta\Sigma^k\|_2 \leq \epsilon$ in the random sampling case (Scenario 1) is $\mathcal{O}(\frac{n \log n}{\rho^2 \epsilon^2 (1-\sigma_{\max})^4})$. Time samples needed are proportional to the dimension which can be large. As the sampling rate is halved, the number of time samples needed increases four-fold. As $\sigma_{\max} \to 1$, we need more samples and this is intuitive as the samples present less new information if there is strong dependency.

3.2. Refined estimates of the covariance matrix

We next observe how structural constraints on the covariance estimates can improve estimation accuracy. Before we proceed, we note that structural properties of matrix Σ^k has implications for the structure of A and vice versa. For instance, when A is block-diagonal and $Q_w = \mathbf{I}$ or the VAR process can be decomposed to decoupled sub-systems (for instance, $(x_t)_{[1,3]}$ could evolve by a 3-dimensional VAR process), Σ^0 can be shown to be banded. On the other hand, when the decoupled structure of A is not known, it might be written as a permutation of a block-diagonal matrix. The covariance matrix would be sparse in this case. Finally, if A and Q_w are low rank, Σ^0 can be low rank as well. This is practical motivation for not only considering the general problem of estimating Σ^k but also the special cases when Σ^k satisfies additional structural constraints.

3.2.1. Bandedness

We consider the case where Σ^k is banded and follow the analysis of [9]. Let us suppose that the covariance matrix (and its transpose) belongs to the following well-conditioned tapering class:

$$\mathcal{V} = \{\Sigma : \sum_j |\Sigma_{i,j}| \mathbf{1}(|i-j| \geq s) \leq C s^{-\alpha}, \forall i\}.$$

The higher the value of α, the more banded the matrix. We use the following operation

$$B_s(\Sigma) = [\Sigma_{i,j} \mathbf{1}(|i-j| \leq s)]_{i,j \in [n]},$$

to construct a refined estimate of the covariance matrix.

Theorem 2. *If we choose a banding factor $s = \gamma(\delta)^{-1/(\alpha+1)}$, we have with probability at least $1-\delta$ that $\|B_s(\hat{\Sigma^k}) - \Sigma^k\|_2 = \mathcal{O}([\gamma(\delta)]^{\alpha/(\alpha+1)})$. Additionally, if $\lambda_{\min}(\Sigma^0) > 0$, then $\|[B_s(\hat{\Sigma^0})]^{-1} - [\Sigma^0]^{-1}\|_2 = \mathcal{O}(s\gamma(\delta))$.*

This theorem can be extended to soft banding operations as well. As α increases, the band s we consider decreases. We find that the number of time samples needed for $\|\Delta\Sigma^k\|_2 \leq \epsilon$ is s^2/n times that of the naive estimator in Eq. (3). We have a weaker s^2 dependence because we have bounded $\|\Delta\Sigma^k\|_2^2 \leq \|\Delta\Sigma^k\|_1 \|\Delta\Sigma^k\|_\infty$ and $\|\Delta\Sigma^k\|_1, \|\Delta\Sigma^k\|_\infty = \mathcal{O}(s)$.

3.2.2. Sparsity

We next consider the case where the covariance matrix is sparse. Following the analysis of [8], let us suppose that the covariance matrix belongs to the class of sparse positive definite matrices $\mathcal{U}$ defined as

$$\mathcal{U} = \{\Sigma : \sum_j |\Sigma_{i,j}|^q \leq s, \sum_j |\Sigma_{j,i}|^q \leq s \ \ \forall i\}.$$

This is a class of well conditioned sparse covariance matrices. When $q = 0$, then there are s non-zero values in each row (and column). We process our empirical covariance matrix by thresholding the entries. In other words,

$$U_u(\Sigma) = [\Sigma_{i,j} \mathbf{1}(|\Sigma_{i,j}| \geq u)]_{i,j \in [n]}. \tag{5}$$

Theorem 3. *When the thresholding factor is $u = 2\gamma(\delta)$, we have with probability at least $1-\delta$ that $\|U_u(\hat{\Sigma^k}) - \Sigma^k\|_2 = \mathcal{O}(s[\gamma(\delta)]^{1-q})$. Additionally, if $\lambda_{\min}(\Sigma^0) > 0$, we have $\|[U_u(\hat{\Sigma^0})]^{-1} - [\Sigma^0]^{-1}\|_2 = \Omega(s[\gamma(\delta)]^{1-q})$.*

In the random sampling case (Scenario 1) with s non-zero elements in each row and column, we would need $\mathcal{O}(\frac{s^2 \log n}{\rho^2 (1-\sigma_{\max})^4 \epsilon^2})$ time samples for $\|\Delta\Sigma^k\|_2 \leq \epsilon$ error. We need s^2/n fraction of time samples compared to estimator in (3) which can be significantly smaller than 1 in the sparse setting.

3.2.3. Sparsity of the Inverse

A popular regularization assumption in the case of covariance estimation from independent samples is the sparsity of the precision (inverse covariance matrix $\Theta^0 = (\Sigma^0)^{-1}$) assumption. This assumption is valid in the independent case ($\sigma_{\max} = 0$) where estimating the sparse precision matrix maps to Gaussian Markov random model selection. Also, when the covariance matrix is block-diagonal, we expect the inverse to be block-diagonal and sparse.

In this section, we follow the convergence analysis of [10]. 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 $\mathcal{S} = \mathcal{E}(\Theta) \cup \{(i,i) | i \in [n]\}$ consists of $\mathcal{E}(\Theta)$ and the diagonal elements. Also, d is the maximum row cardinality which is the maximum number of non-zero elements in any row of the inverse covariance matrix.

The estimator for the empirical inverse covariance matrix is obtained from the Bregman divergence on the log determinant function [26]. Consider $g(\Theta) = -\log|\Theta|$. We now find the symmetric positive definite matrix Θ that minimizes $D_g(\Theta^0 || \Theta)$. We obtain the final estimator by replacing the unknown Σ^0 with its empirical estimate and a regularization term which is the ℓ_1 sum of off-diagonal elements $\|\Theta\|_{1,\text{off}} = \sum_{i,j \ i \neq j} |\Theta_{ij}|$:

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

This is a convex optimization problem and can be efficiently solved.

Theorem 4. *For the choice of regularization parameter $\lambda_n = \Omega(\gamma(\delta))$ and when the number of time samples $T = \Omega(\frac{d^2 \log n}{\theta(0)_*^2 (1-\sigma_{\max})^4})$, we have with probability at least $1-\delta$ that $\|\hat{\Theta}^0 - \Theta^0\|_2 = \mathcal{O}(\min(\sqrt{s+n}, d)\gamma(\delta))$ and $\|[\hat{\Theta}]^{-1} - \Sigma_0\|_2 = \mathcal{O}(d\gamma(\delta))$.*

The number of time samples required for the ℓ_2 norm of the covariance matrix error estimate to be within ϵ is $\mathcal{O}(d^2/n)$ times the number needed in the naive estimator.

3.2.4. Low Rank

Finally, we consider the assumption that the rank $r(\Sigma^k) \ll n$ which is valid if the system is evolving in a small subspace. We follow [12] and refine our estimate as

$$\bar{\Sigma} = \text{argmin}_\Sigma \, \|\Sigma - \hat{\Sigma}^0\|_F^2 + \lambda_n \|\Sigma\|_*.$$

Theorem 5. *Using regularization factor $\lambda_n = 4\gamma_2(\delta) = 16\sqrt{n}\gamma(\delta)$, we obtain with probability at least $1-\delta$ that*

$$\frac{1}{\sqrt{n}} \|\bar{\Sigma}^k - \Sigma^k\|_F = \mathcal{O}(\gamma(\delta)\sqrt{r}).$$

We need a fraction $\frac{r}{n}$ time samples for the Frobenius norm of the estimator error to be lower than ϵ compared to the naive estimator in (3). Analysis in [12] indicates that for the iid case, this scaling in dimension and rank is also present in the lower bound.

4. ESTIMATION OF THE TRANSITION MATRIX

Once Σ^k is estimated, we use it to estimate the transition matrix as the quantities are related through the Yule-Walker equation $\Sigma^1 = \Sigma^0 A^\intercal$. The error analysis also allows us to analyse the error in future predictions given the current state value as

$$\|x_{t+1} - \hat{A}x_t\| = \|\Delta A\|\|x_t\| + \|w_t\|.$$

We present three estimators depending on structural assumptions.

4.0.1. General Case

For dense A, our estimate $\hat{A}$ is given by

$$\hat{A}^\intercal = \hat{\Sigma}^{0\dagger}\hat{\Sigma}^1. \tag{6}$$

The error in the estimate of the transition matrix $\Delta A = \hat{A} - A$ is bounded in the following theorem.

Theorem 6. *Let $\sigma_{\min}$ be the minimal singular value of Σ^0. When the number of samples $T = \Omega(\frac{\log n}{\theta(0)_*(1-\sigma_{\max})^4})$, with probability at least $1-\delta$, we have that*

$$\|\hat{A} - A\|_2 = \mathcal{O}\left(\frac{\sigma_{\max}\gamma_2(\delta/2)\|Q_w\|_2}{\sigma_{\min}^2(1-\sigma_{\max}^2)}\right).$$

For $\|\hat{A} - A\|_2 \leq \epsilon$, we need $\mathcal{O}(\frac{n}{(1-\sigma_{\max})^5\rho^2\epsilon^2})$ samples, which can be very large in the high-dimensional case. If we make assumptions on Σ^k, we can use the estimators of Section 3.2 to reduce the number of time samples needed.

4.0.2. Sparsity

When A is sparse, we follow [17] and assume $A, A^\intercal \in \mathcal{A}(q, s, A_1)$, where

$$\mathcal{A}(q, s, A_1) = \left\{B \in \mathbf{R}^{n\times n} : \max_{j\in[n]} \sum_{i=1}^{n} |B_{i,j}^q| \leq s, \|B\|_1 \leq A_1\right\}.$$

Note that $q = 0$ indicates we have an s sparse matrix with bounded ℓ_1 induced norm. The scalar $A_1 \in [0, \sqrt{n}\sigma_{\max}]$ restricts the size of the class of transition matrices from which the estimate $\hat{A}$ is obtained. This is the weakly sparse case and need not exclude the case of Σ^0 being dense. We use:

$$\hat{A}^\intercal = \operatorname{argmin}_{M\in\mathbf{R}^{n\times n}} \sum_{i,j} |M_{i,j}|$$
$$\text{s.t.} \quad \|\hat{\Sigma}^1 - \hat{\Sigma}^0 M\|_{\max} \leq \lambda. \tag{7}$$

This estimate is of the form of the Dantzig selector and amounts to selecting the sparsest matrix A that satisfies the constraints. It reduces to solving parallel linear programs and can be efficiently computed.

Theorem 7. *Let $A, A^\intercal \in \mathcal{A}(q, s, A_1)$ and $\lambda_n = (1 + A_1)\gamma(\delta/2)$, With probability greater than $1-\delta$, $\|\hat{A}-\hat{A}\|_2 = \mathcal{O}\left(s(\lambda_n\|\Sigma^{0\dagger}\|_1)^{1-q}\right)$.*

In the random sampling case (Scenario 1), we find that the number of time samples for estimation to a specified accuracy is a fraction $\frac{s^2}{n}$ of the time samples of the naive estimator (6).

4.0.3. Low Rank

We assume the rank of the transition matrix A is $r \ll n$, and Σ^0 is full rank. We use the following estimator:

$$\hat{A} = \operatorname{argmin}_M \langle M^\intercal, \hat{\Sigma}^0 M^\intercal - 2\hat{\Sigma}^1\rangle + \lambda_n\|M\|_*. \tag{8}$$

We obtain this estimator from nuclear norm regularized minimization of $\mathbb{E}[\|x_{t+1} - Mx_t\|_2^2]$. With a sufficiently large enough number of samples, $\hat{\Sigma}^0 \succeq 0$ and (8) becomes a convex optimization problem.

Theorem 8. *Let the number of time samples $T = \Omega(\frac{d^2\log n}{\theta(0)_*(1-\sigma_{\max})^4})$. When $\lambda_n \geq 4(1+\sigma_{\max})\gamma_2(\delta/2)$, we have $\|\hat{A} - A\|_F = \mathcal{O}(\lambda_n\sqrt{r})$ with probability at least $1-\delta$.*

The low rank estimator reduces $\|\Delta A\|_F$ by a factor of n/r which is significant in the high dimensional regime.

5. EXTENSIONS TO GENERAL MODELS

We can extend our results to estimators for VAR(p) processes. In this case,

$$x_{t+1} = A_1x_t + A_2x_{t-1} + \dots A_px_{t-p+1} + w_t.$$

Using $\underline{x}_t = [x_t x_{t-1} \dots x_{t-p+1}]$, we can write

$$\underline{x}_{t+1} = \begin{bmatrix} A_1 & \dots & A_p \\ I & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & I \end{bmatrix} \underline{x}_t + \begin{bmatrix} w_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$
$$\Rightarrow \underline{x}_{t+1} = \underline{A}\underline{x}_t + \underline{w}_t,$$

which is now a VAR(1) process. It is assumed that $\|\underline{A}\|_2 = \sigma_{\max} < 1$. This is a sufficient condition for the system to be stable.

Rewriting $\underline{z}_t = [z_t z_{t-1} \dots z_{t-p+1}]$, $\underline{v}_t = [v_t v_{t-1} \dots v_{t-p+1}]$, and $\underline{P}_t = \operatorname{diag}(P_t, P_{t-1}, \dots, P_{t-p+1})$, we get $\underline{z}_t = \underline{P}_t(\underline{x}_t + \underline{v}_t)$. The key difference between the VAR(p) and the VAR(1) model considered earlier is that $\underline{v}_t$ and $\underline{P}_t$ are not independent across time. The latter matrix does not have independent diagonal matrices either.

Nonetheless, we can extend our earlier estimator. Defining matrices $\underline{Q}_v = \mathbb{E}[\underline{v}_t\underline{v}_t^\intercal], \underline{\theta}(k) = \mathbb{E}[\operatorname{diag}(\underline{P}_t)\operatorname{diag}(\underline{P}_t)^\intercal]$, we have

$$\underline{S}^k = \frac{1}{T-k-p+1}\sum_{t=p}^{T-k} \underline{z}_t\underline{z}_{t+k}^\intercal$$
$$(\hat{\underline{\Sigma}}^k)_{ij} = \underline{S}_{ij}^k/\underline{\theta}(k)_{ij} - (\underline{Q}_v)_{ij}.$$

Taking $p = 1$, we retrieve our old estimator.

We have the following concentration theorem for VAR(p) processes.

Theorem 9. *With probability greater than $1-\delta$*

$$\|\Delta\underline{\Sigma}^k\|_{\max} = \mathcal{O}\left(\sqrt{\frac{\log(np/\delta)}{\log(n/\delta)}}\gamma(\delta)\right)$$
$$\|\Delta\underline{\Sigma}^k\|_2 = \mathcal{O}\left(\sqrt{pn}\gamma(\delta)\right)$$

After obtaining $\underline{\hat{A}}$ from $\hat{\underline{\Sigma}}^k$ for $k = 0, 1$, we have $\hat{A}_i = (\underline{\hat{A}})_{[n]\times\{p(i-1)+1,\dots,pi\}}$. The main conclusion from the theorem is that covariance estimation and hence system identification of VAR(p) processes is approximately p times slower than identification of VAR(1) processes.

6. CONCLUSION

We have considered the problem of estimating the covariance matrix and performing system identification of large structured vector autoregressive processes with partial observations which could be corrupted by multiplicative or additive noise. This problem is motivated by the limitations of data acquisition in high-dimensional measurement systems.

An estimator of the covariance matrices of the process that can start from an arbitrary state is first described. Refined estimators for the case where the covariance matrix has structural constraints such as sparsity, low rank, bandedness, and sparsity in the inverse covariance matrix are then described and analysed. These were subsequently used to obtain the transition matrix in both the general case and one with structural constraints of sparsity and low-rank. The number of time samples required for a pre-specified accuracy scales with the dimension and inversely with the squared sampling rate in the general case. As structural constraints are factored in, the number of time samples scales logarithmically in the dimension.

7. 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, "Estimation in autoregressive processes with partial observations: Proofs," http://stanford.edu/~milind/reports/system_id_icassp_proof.pdf, accessed: 2016-09-12.

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

[6] 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.

[7] 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.

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

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

[10] 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

[11] P.-L. Loh and M. J. Wainwright, "High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2011, pp. 2726–2734.

[12] 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

[13] M. Kolar and E. Xing, "Estimating sparse precision matrices from data with missing values," in *Proceedings of ICML*, 2012.

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

[15] 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.

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

[17] 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.

[18] 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

[19] 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

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

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

[22] X. Liu and A. Goldsmith, "Kalman filtering with partial observation losses," in *Decision and Control, 2004. CDC. 43rd IEEE Conference on*, vol. 4, Dec 2004, pp. 4180–4186 Vol.4.

[23] E. Rohr, D. Marelli, and M. Fu, "Kalman filtering with intermittent observations: Bounds on the error covariance distribution," in *2011 50th IEEE Conference on Decision and Control and European Control Conference*, Dec 2011, pp. 2416–2421.

[24] B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, M. I. Jordan, and S. S. Sastry, "Kalman filtering with intermittent observations," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1453–1464, Sept 2004.

[25] 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.

[26] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.