System Identification from Partial Samples: Non-Asymptotic Analysis

Milind Rao*, Alon Kipnis*, 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,kipnisal}@stanford.edu, tjavidi@ucsd.edu, yonina@ee.technion.ac.il, andrea@wsl.stanford.edu

***Abstract*—The problem of learning the parameters of a vector autoregressive (VAR) process from partial random measurements is considered. This setting arises due to missing data or data corrupted by multiplicative bounded noise. We present an estimator of the covariance matrix of the evolving state-vector from its partial noisy observations. We analyze the non-asymptotic behavior of this estimator and provide an upper bound for its convergence rate. This expression shows that the effect of partial observations on the first order convergence rate is equivalent to reducing the sample size to the average number of observations viewed, implying that our estimator is order-optimal. We then present and analyze two techniques to recover the VAR parameters from the estimated covariance matrix applicable in dense and in sparse high-dimensional settings. We demonstrate the applicability of our estimation techniques in joint state and system identification of a stable linear dynamic system with random inputs.**

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

I. Introduction

Vector Autoregressive (VAR) models were first introduced by Sims [1] as a tool in macroeconometric analysis. These models are natural tools for forecasting since the model implies that current values of variables depend on past values through a joint generation mechanism [2] and find application in finance, econometrics, and neuroscience. Often VAR models are fit on high dimensional data, in which the costs of collecting, communicating, storing or computing may be prohibitively high. One example is in wireless sensor networks measuring an underlying VAR phenomenon. To conserve battery power, measurements are typically collected from only a few of the sensors at each epoch. In addition, data collected using a sensor array may be missing or noisy due to additive receiver noise or multiplicative fading noise. These limitations on data acquisition motivate identifying or making inferences from a system with partial observations. Under such restrictions on the availability of the data, it is important to design and analyze estimation procedures that are robust to missing data.

An auto-regressive (AR) process is characterized by a finite set of parameters that describe the linear relation between present time vector-valued samples and past vector samples plus independent noise. A VAR process extends the definition of an AR process by assuming that each coefficient in this linear combination is a matrix. Specifically, random vectors $x_t \in \mathbf{R}^n$ evolve as

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

where A is the *state-transition* matrix. Our goal is to infer the matrix A from partial observations of x_t. Specifically, suppose that we observe each element of x_t with probability ρ. We achieve this by first estimating the sample covariance matrix for several time lags and then using these estimates to approximate A.

A. Contribution

The main result of this paper is an estimator for the covariance matrix of a first order VAR process from its partial samples which are corrupted by multiplicative and additive noise. We analyze the non-asymptotic behavior of the error under this covariance estimator and provide an upper bound for its convergence rate. For large time horizon T, we show that the operator norm of the covariance estimator error vanishes as $1/\rho\sqrt{n/T}$ where $\rho \in (0,1]$ is the effective rate of missing data, T is the number of samples, and n is the dimension. This implies that the effect of missing data translates to a ρ^2 reduction in efficiency of estimation. In addition, we analyse two methods to identify the transition matrix of the VAR: the first approach may be applied to all matrices, while the second one is particularly suited for sparse matrices. We compare these two techniques by simulations which confirm the theoretical convergence predictions. Finally, we demonstrate our proposed VAR parameter estimation techniques in estimating the parameters of a stable linear dynamical system from its partial state measurements. Once the system is identified, Kalman filtering and Rauch-Tung-Striebel (RTS) smoothing are used to jointly identify the state.

B. Related Work

While estimating the parameters of scalar auto-regressive (AR) processes as well as their statistics is a well-established

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.

field in signal processing and control [3], [4], some of the counterparts of these problems for VAR estimation have still not been studied. Recovering autoregressive model parameters from partial observations is considered in [5]. In that work, the authors quantify the asymptotic variance of the sample covariances which indicate how quickly their estimator of the covariance converges. The autoregressive model parameters can then be obtained from the covariance.

The classic solution for estimating the state-transition matrix A from full noiseless observations is done by using least-squares [6]. When dimension n is larger than T, structural assumptions on the transition matrix are imposed to enforce identifiability. The work in [7] proposes a 2-stage approach for fitting sparse VAR models where non-zero coefficients of A are selected in the first stage and estimated in the second. The authors of [8] and [9] impose lasso and nuclear norm regularization procedures to encourage sparsity, low-rank in the transition matrices. Non-asymptotic performance analysis was performed assuming that certain stringent restricted strong convexity properties held. Basu [10] showed how spectral density influence the rate of convergence of sparse VAR transition matrices.

In this work, methods from the following two papers on VAR parameter estimation are adapted in our problem formulation. In Bento et al. [11], attention is paid to recovering the sparse support of the state-transition matrix of a VAR process with complete observations. They consider a continuous-time process model, but they break down the proof on the convergence rate of the estimator for the state-transition matrix by first considering the discrete-time case. The result in this paper regarding the convergence rate of the covariance matrix estimator is an extension of the discrete-time case of [11]. Our results reduce to these in the special case of full and noiseless observations. With full noiseless observations, work by Han and Liu [12] analyses the problem of estimating weakly sparse transition matrices without making the restricted eigenvalue or irrepresentable condition approximations of prior analyses. We apply their algorithm for estimating sparse transition matrices from covariance estimates and obtain the same rate of convergence in the case with complete noiseless observations.

The paper [13] is especially relevant to the problem at hand. There the authors provide guarantees for high-dimensional linear regression where the observations may have added noise, missing data and may be dependent. This is done by showing that the estimators for the covariance matrix satisfy restricted strong convexity conditions. The estimators of the covariance matrix from partial observations used in this work are similar to the ones we use.

The problem of performing system identification from partial samples is closely related to the subspace learning problem with partial information. In the subspace learning problem, iid samples are given from a high dimensional distribution over a smaller subspace. The subspace in which the observations lie is determined from its covariance matrix. Like in our VAR parameter estimation problem, it is of interest to determine the covariance matrix from partial measurements in the subspace identification problem. The algorithm GROUSE of [14] and PETRELS of [15] are online algorithms without guarantees in which the subspace is learned from partial observations. Theoretical sample guarantees for this problem are obtained in [16]. In that work, the convergence of an estimator of the covariance matrix is first analysed after which bandit principle component analysis (PCA) as well as other algorithms are used to obtain the subspace. We obtain the same scaling in the covariance estimate error. Similar in spirit, [17] uses a few compressive linear measurements instead of partial samples and seeks to learn the covariance matrix and the subspace of the observations. The key difference is that the samples in our problem are not iid but are dependent, which leads to slower convergence of the covariance matrix.

The rest of this paper is organized as follows: the estimation problem is presented in Section II. Section III states the main results with respect to the estimators of covariance and state-transition matrices and their convergence rates, where sketches of the proofs are given in Section IV. Section V presents two examples where our estimation techniques may apply, as well as numerical simulations of the performance.

II. Problem Description

Consider a linear dynamical system with state vector $x_t \in \mathbf{R}^n$ evolving as (1). The transition matrix A is unknown and it is assumed that $\|A\|_2 = \sigma_{\max} < 1$ A is stable; this implies that the spectral radius is bounded by one and that A is stable. It is observed that if $\sigma_{\max} = 0$, then the observations are independent across time. Innovation process statistic Q_w is unknown. At each time instant we observe

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

where $v_t \overset{\text{iid}}{\sim} \mathcal{N}(0, Q_v)$ and the matrix Q_v is assumed to be known. The matrix P_t is a random measurement matrix of the form

$$P_t = \text{diag}(p_t),$$

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. The system is initiated such that x_0 is not assumed to be known but $\|x_0\|_2$ is bounded by $\mathcal{O}(T^{-1/2})$.

The above scenario may correspond to any of the following observation models:

(i) **Independent and homogenous random sampling of observations:** When $\mathcal{P} \equiv \mathcal{B}^n(\rho)$, each observation is viewed independently with fixed probability ρ. If $\rho \approx 1$, we have full observations and if $\rho \approx 0$, we have limited observations. Prior work [11], [12] has focussed on the case where $\rho \approx 1$.

(ii) **Independent and heterogenous random sampling of observations:** In this setting $\mathcal{P} \equiv \prod_{i=1}^{n} \mathcal{B}(\rho_i)$, so that each observation is viewed independently with differing probabilities. This could model a scenario where communication from some sensors is costlier or noisier and hence observations from these sensors are made less often.

(iii) **Intermittent observations:** When

$$\mathcal{P} \equiv \begin{cases} I & \text{w.p. } \rho \\ 0 & \text{else} \end{cases},$$

we see all the observations with probability ρ or see no observations at all at any time instant. This is the scenario considered in the intermittent Kalman filtering literature [18], [19], [20].

(iv) **Multiplicative noise:** The case in which $\mathcal{P}$ is an arbitrary distribution on non-negative bounded support $[p_l, p_u]^n, 0 \leq p_l \leq p_u < \infty$ could model independent multiplicative noise. Multiplicative noise could arise from independent fading in the wireless sensor network setting. Note that the multiplicative noise is independent across time but the noise affecting each dimension of the observation may be correlated.

We seek to find an algorithm to estimate A and bound the number of samples T such that $\|\hat{A}_T - A\| \leq \epsilon$ for a suitable norm.

III. Algorithm and Results

A. Estimation Algorithm

Consider the k cross-correlation matrix

$$\Sigma^k \triangleq \mathbb{E}[x_t x_{t+k}^\intercal].$$

From the Yule-Walker equations,

$$\Sigma^1 = \Sigma^0 A^\intercal,$$

so that

$$A = \Sigma^{1\intercal} \Sigma^{0\dagger}. \tag{2}$$

Our approach is to form an estimate of Σ^k and then use that to estimate A.

Since Σ^k is unknown, we first consider the empirical covariance matrix S^k of observations z_t:

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

Let $\theta(k)_{ij}$ denote the average scaling due to the multiplicative noise observed in the ij^{th} element of S^k. We have

$$\theta(k) = \mathbb{E}[p_t p_{t+k}^\intercal].$$

We can observe that $\mathbb{E}[S^k] = \mathbb{E}[z_t z_t^\intercal] = \theta(k) \circ (\Sigma^k + Q_v \mathbf{1}(k = 0))$ where $\circ$ denotes the entrywise or Hadamard product, Q_v is the covariance of the additive noise, and $\mathbf{1}(\cdot)$ is 1 if the condition inside evaluates to true and is zero otherwise. We define $\theta(k)_* = \min_{i,j} \theta(k)_{i,j}$ to denote the minimum scaling due to the multiplicative noise.

As our estimate of Σ^k, we use

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

The estimator $\hat{\Sigma}^k$ is unbiased if the observations are taken when the system is stationary (x_0 arises from the stationary distribution of the state vector).

Remark 1. *In the special case of independently sampled observations with probability ρ, matrix $\theta(k)_{ij}$ is the probability that both the i^{th} element of x_t and the j^{th} element of x_{t+k} are observed. The estimator can then be rewritten as*

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

where $\circ$ denotes the Hadamard or entrywise product of matrices. The matrix $S^k \circ I_n$ is diagonal with entries that are the diagonal entries of matrix S^k.

Once Σ^k is estimated, we use it to estimate the transition matrix in two ways, depending on the matrix properties:

1) **Dense Matrix:**
For dense A, our estimate $\hat{A}$ is given by,

$$\hat{A}^\intercal = \hat{\Sigma}^{0\dagger} \hat{\Sigma}^1. \tag{4}$$

2) **Sparse Matrix:**
The estimation (4) is not cognizant of special structural properties of A such as sparsity. Sparsity is especially relevant in high-dimensional regimes where the number of samples is not much larger or even smaller than the dimension. When A is sparse, the following estimate from [12] is used:

$$\hat{A}^\intercal = \underset{M \in \mathbf{R}^{n \times n}}{\operatorname{argmin}} \sum_{i,j} |M_{i,j}| \quad \text{s.t.} \quad \|\hat{\Sigma}^1 - \hat{\Sigma}^0 M\|_{\max} \leq \lambda \tag{5}$$

The estimate (5) is similar to the Dantzig selector which selects the sparsest choice of transition matrix satisfying the constraints given for an appropriate choice of λ. This form (5) suggests that we can recover other structured high-dimensional transition matrices by using appropriate regularization. An example would be to obtain a low rank transition matrix by using the nuclear norm. This will be investigated in future work.

B. Key Results

In this section, we obtain high probability performance guarantees for the estimators described in the previous section.

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

$$\begin{aligned} \|\Delta\Sigma^k\|_{\max} &\leq \gamma(\delta) \\ \|\Delta\Sigma^k\|_2 &\leq \gamma_2(\delta) = 4\sqrt{n}\gamma(\delta), \end{aligned} \tag{6}$$

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

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

Theorem 1 implies that the maximum deviation of the sample covariance matrix from the expected covariance matrix is proportional to the logarithm of the dimension. We could have exponentially more dimensions than we have samples and still expect to see low maximum deviation. The maximum deviation is also penalized by the innovation noise and observation noise although the impact of the former is scaled up by $\sigma_{\max}$, which represents the dependency factor. As $\sigma_{\max} \to 1$, we need more samples to estimate the covariance matrix to a given accuracy. This is intuitive as the samples present less new information if there is strong dependency.

Corollary 1. *In the independent sampling case with probability ρ and with noiseless observations, we obtain with probability greater than $1 - \delta$*

$$\|\hat{\Sigma}^k - \Sigma^k\|_{\max} \le \frac{\|Q_w\|_2}{\rho(1-\sigma_{\max})^2}\sqrt{\frac{8\log(6n^2/\delta)}{(T-k)}} + o((T-k)^{-0.5}). \quad (7)$$

The key fact to observe here is that the error is inversely proportional to the square root of the number of samples and inversely proportional to the sampling probability ρ. For example, this implies that if the sampling ratio is halved, then the number of time samples to maintain the same error increases four-fold. This phenomena can be intuitively understood since the probability of observing an element in the sample covariance matrix is proportional to ρ^2. For $\|\hat{\Sigma}^0 - \Sigma^0\|_{\max} \le \epsilon$, we need $\mathcal{O}(\frac{\log n}{\rho^2\epsilon^2(1-\sigma_{\max})^4})$ time samples. For $\|\hat{\Sigma}^0 - \Sigma^0\|_2 \le \epsilon$, we need a much larger $\mathcal{O}(\frac{n\log n}{\rho^2\epsilon^2(1-\sigma_{\max})^4})$ number of samples.

Substituting $\rho = 1$, we have the full observation case. In [11], the error in estimating Σ^0 when $Q_w = I$ was found to be

$$\|\hat{\Sigma}^0 - \Sigma^0\|_{\max} \le \sqrt{\frac{32\log(2n^2/\delta)}{T(1-\sigma_{\max})^3}},$$

which is the same scaling in ρ, n, and T to the bound we obtain. Our bound has a slightly weaker factor of $(1-\sigma_{\max})^{-2}$ as opposed to $(1-\sigma_{\max})^{-1.5}$ as we have loosened the analysis to obtain bounds on the operator norm. It is possible to tighten it further to $(1-\sigma_{\max}^2)^{-1.5}$. The second difference that arises is different constants as we have used a concentration result for multiplicative noise which is not needed in the noiseless fully observed case of $\rho = 1$. In the scalar case $n = 1$, the results from [5, Eq. 3.11] implies that with high probability,

$$\|\hat{\Sigma}^k - \Sigma^k\|_{\max} \le \sqrt{\frac{C_1}{T} + \frac{C_2}{\rho^2 T}}$$

for some constants $C_1, C_2 > 0$, which agrees with Corollary 1.

Next, we consider estimation of the transition matrix A using the relation (4). We have the following result:

Theorem 2. *Let $\sigma_{\min}$ be the minimal singular value of Σ^0. For $T = \Omega(\frac{\log n}{\theta(0)_*(1-\sigma_{\max})^4})$, with probability at least $1-\delta$ we have*

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

Theorem 2 indicates that the error is proportional to the square root of the dimension which is much larger than the log factor in (6) and (7). This unfortunate fact could lead to very loose bounds in the high dimensional setting where n is on the same order as the number of time samples.

One way to get around this penalizing $\sqrt{n}$ factor is to impose structural constraints such as sparsity. As in [12] we further assume $A \in \mathcal{A}(q, s, A_1)$ where

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

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.

The following theorem provides a guide for selecting appropriate λ in (5) along with performance bounds which are logarithmic in dimension.

Theorem 3. *Let $A \in \mathcal{A}(q, s, A_1)$ and let*

$$\lambda = (1 + A_1)\gamma(\delta/2),$$

with $\gamma(\cdot)$ defined in Theorem 1. Then, with probability greater than $1 - \delta$,

$$\|\hat{A} - \hat{A}\|_1 \le 4s(2\lambda\|\Sigma^{0\dagger}\|_1)^{1-q}$$
$$\|\hat{A} - \hat{A}\|_{\max} \le 2\|\Sigma^{0\dagger}\|_1\lambda.$$

These theorems imply that we can bound $\|\hat{A} - A\|$ to within $\mathcal{O}(\frac{1}{\rho}\sqrt{\frac{\log n}{T}})$ with high probability.

When $s = n$, we obtain $\|\hat{A} - \hat{A}\|_1 \lesssim \Theta(n\gamma(\delta/2))$ with high probability which is also the upper bound on $\|\hat{A} - A\|_1$ we get from Theorem 2. Alternatively, if we need $\|\hat{A} - A\|_2 \le \epsilon$, we need a fraction $\frac{s^2}{n}$ of the time samples when we consider $A, A^\intercal \in \mathcal{A}(q, s, A_1)$ compared to the estimator (3).

IV. Proof Sketches of Theorems 1-2

In this section, we provide a sketch of the proofs for Theorems 1 and 2. A detailed version of these proofs as well as the proof for Theorem 3 is in the supplementary document [21].

We first outline the steps in proving the bound for $\|\hat{\Sigma}^0 - \Sigma^0\|_{\max}$ in Theorem 1. We focus on the convergence of each element of the sample covariance matrix $\hat{\Sigma}_{ij}^0$. In order to do this, our analysis hinges on expressing the various errors in terms of quadratic and linear products of Gaussian vectors as well as a concentration result to determine limits on the multiplicative noise.

We assume that the process begins $T_p \ge 0$ time units before observations take place. In other words, $x_{-T_p} = x_S$. We create stacked vectors of noise $W = [w_{-T_p+1}|\ldots|w_0|w_1|\ldots|w_T]$

and of initial conditions of the same dimension $X_S = [x_S|\mathbf{0}]$. Let the stacked vector of observations of position i be the T-dimensional vector $Z_i = [z_{1,i}|z_{2,i}|\dots|z_{T,i}]$. Similarly, the stacked vector of observation noise for position i is $V_i = [v_{1,i}|v_{2,i}|\dots|v_{T,i}]$. We recall that $P_{t,i}$ is 1 if the i^{th} position of noisy observation of x_t is observed in the sampling case or is the multiplicative noise otherwise. Finally, create the T-diagonal matrix $P_i = \mathrm{diag}([P_{1,i}|\dots|P_{T,i}])$.

Consider the matrix $\Phi_i \in \mathbf{R}^{T\times n(T+T_p)}$, defined as

$$\Phi_i = \begin{bmatrix} A_i^{T_p} & \dots & A_i & \mathbf{I}_i & \dots & 0 \\ A_i^{T_p+1} & \dots & A_i^2 & A_i & \dots & 0 \\ \vdots & & & \ddots & & \vdots \\ A_i^{T_p+T-1} & \dots & A_i^T & A_i^{T-1} & \dots & \mathbf{I}_i \end{bmatrix} \tag{8}$$

where L_i for a matrix L is its i^{th} row.

We now have

$$\begin{aligned} Z_i &= P_i\left(\Phi_i(W+X_S)+V_i\right) \\ \hat{\Sigma}^0_{i,j} &= \frac{1}{(T)\theta(0)_{i,j}} Z_i^\intercal Z_j - (Q_v)_{i,j}. \end{aligned} \tag{9}$$

We further define $P(0)_{i,j} = P_iP_j$. This is a diagonal matrix with the terms arising from a bounded distribution. In the sampling case, the marginal distribution is $\mathcal{B}(\theta(0)_{i,j})$ for the elements. We can expand the terms of (9) as

$$\begin{aligned} T_1 &= \frac{(W^\intercal\Phi_i^\intercal + V_i^\intercal)P(0)_{i,j}(\Phi_jW+V_j)}{T\theta(0)_{i,j}} - (Q_v)_{i,j} \\ T_2 &= \frac{X_S^\intercal}{T\theta(0)_{i,j}}\Big[\big(\Phi_i^\intercal P(0)_{i,j}\Phi_j + \Phi_j^\intercal P(0)_{i,j}\Phi_i\big)W \\ &\quad + \Phi_i^\intercal P(0)_{i,j}V_j + \Phi_j^\intercal P(0)_{i,j}V_i\Big] \\ T_3 &= \frac{X_S^\intercal\Phi_i^\intercal P(0)_{i,j}\Phi_jX_S}{T\theta(0)_{i,j}} \\ \hat{\Sigma}^0_{i,j} &= T_1+T_2+T_3 \end{aligned}$$

We now have to bound the deviation of individual terms T_i from their mean. The term T_1 indicates the deviation if the system started from zero state. The effect of the initial state is through terms T_2 and T_3.

It can be seen that T_1 can be written as $z^\intercal L_1 z/(T\theta(0)_{i,j})$ where $z \sim \mathcal{N}(0,I)$. We show that $\|L_1\|_F^2 = \mathcal{O}(\mathrm{Tr}(P(0)_{i,j}^2))$. We need to bound what the average multiplicative noise is for the ij^{th} term of the matrix. As the multiplicative noise is bounded, it converges rapidly to its mean. In other words $\mathrm{Tr}(P(0)_{i,j}^2)/(T\theta(0)_{i,j}) \le 1 + \mathcal{O}(T^{-1/2})$. The proof follows by expressing $\mathrm{Tr}(P(0)_{i,j}^2)$ as a collection of sums of independent terms and then applying the Hoeffding bound as the terms are bounded. In the independent sampling case with probability ρ, this implies that the number of effective samples for the ij^{th} element of Σ^0 is ρ^2T when $i \neq j$ and ρT otherwise. We see that term T_1 is a sub-exponential or χ^2 random variable and deviation from its mean falls as $\mathcal{O}(T^{-1/2})$.

Similarly, T_2 can be written as $l_2^\intercal z/(T\theta(0)_{i,j})$ where $\|l_2\|_2 = \mathcal{O}((1-\sigma_{\max})^{-2})$. Term T_2 is a Gaussian random variable and converges to its mean zero as $\mathcal{O}(T^{-1})$. Finally, we see that the term T_3 decays as $\mathcal{O}(T^{-1})$. Thus we see that the convergence of $\hat{\Sigma}^0$ to its mean is dominated by the convergence of the T_1, implying that the initial state does not matter much as long as $x_S = o(T^{-1/2})$. We can union bound the maximum deviation of T_1, T_2 for all elements of $\hat{\Sigma}^k$ and this gives us an additional factor of $\log n$.

We have characterized how rapidly $\hat{\Sigma}^0$ converges to its mean, but also need to quantify how far $\mathbb{E}[\hat{\Sigma}^0]$ is from the true covariance matrix Σ^k. It is shown that $\|\mathbb{E}[\hat{\Sigma}^0] - \Sigma^0\|_2 = \mathcal{O}(\sigma_{\max}^{T_p}/T)$. If $T_p \gg 1$, then the system has been running for a long time and is stationary. Otherwise, the system has a transient period where it transitions from non-stationarity to stationarity. Thus the estimate is asymptotically unbiased. The bias does not play a significant role as the sample covariance matrix converges to its mean at a much slower rate.

Theorem 1 is now proved by observing that the deviation can be split as

$$\begin{aligned} \|\Sigma^0 - \hat{\Sigma}^0\|_{\max} &\le \|\hat{\Sigma}^0 - \mathbb{E}[\hat{\Sigma}^0]\|_{\max} + \|\mathbb{E}[\hat{\Sigma}^0] - \Sigma^0\|_{\max} \\ &\le \max_{i,j}|T_1 - \mathbb{E}[T_1]| + \max_{i,j}|T_2| + \|\mathbb{E}[\hat{\Sigma}^0] - \Sigma^0\|_{\max}. \end{aligned}$$

The bounds on $\|\hat{\Sigma}^0 - \Sigma^0\|_2 = \max_{\|u\|_2,\|v\|_2\le 1} u^\intercal\Sigma^0 v$ follow from a covering argument.

We now prove Theorem 2. In the case of identifying dense A, we further assume that the condition number of Σ^0 is finite. This can occur when A is a full rank matrix. Let the minimum singular value of Σ^0 be greater than $\sigma_{\min}$. From Theorem 1, we have with high probability that the deviation from the mean of estimators of Σ^1 and Σ^0 is less than $\gamma(\delta/2)$, a quantity that falls as $\mathcal{O}((T-k)^{-1/2})$. In other words defining $\Delta\Sigma^0 = \hat{\Sigma}^0 - \Sigma^0$ and $\Delta\Sigma^1 = \hat{\Sigma}^1 - \Sigma^1$, we have $\|\Delta\Sigma^0\|_2, \|\Delta\Sigma^1\|_2 \le 4\sqrt{n}\gamma(\delta/2)$. We now use a standard linear algebra result from [22] that relates the error in estimating $\Sigma^{0\dagger}$. The result states that $\|\Delta\Sigma^{0\dagger}\|_2 \le \|\Sigma^{0\dagger}\|_2^2\|\Delta\Sigma^0\|_2$. Theorem 2 follows from a decomposition of the error as

$$\begin{aligned} \|\hat{A} - A\|_2 &\le \|\hat{\Sigma}^{1\intercal}\hat{\Sigma}^{0\dagger} - \Sigma^{1\intercal}\hat{\Sigma}^{0\dagger} + \Sigma^{1\intercal}\hat{\Sigma}^{0\dagger} - \Sigma^{1\intercal}\Sigma^{0\dagger}\|_2 \\ &\le (\|\Delta\Sigma^{0\dagger}\|_2 + \|\Sigma^{0\dagger}\|_2)\|\Delta\Sigma^1\|_2 + \|\Sigma^1\|_2\|\Delta\Sigma^{0\dagger}\|_2 \end{aligned}$$

V. Examples

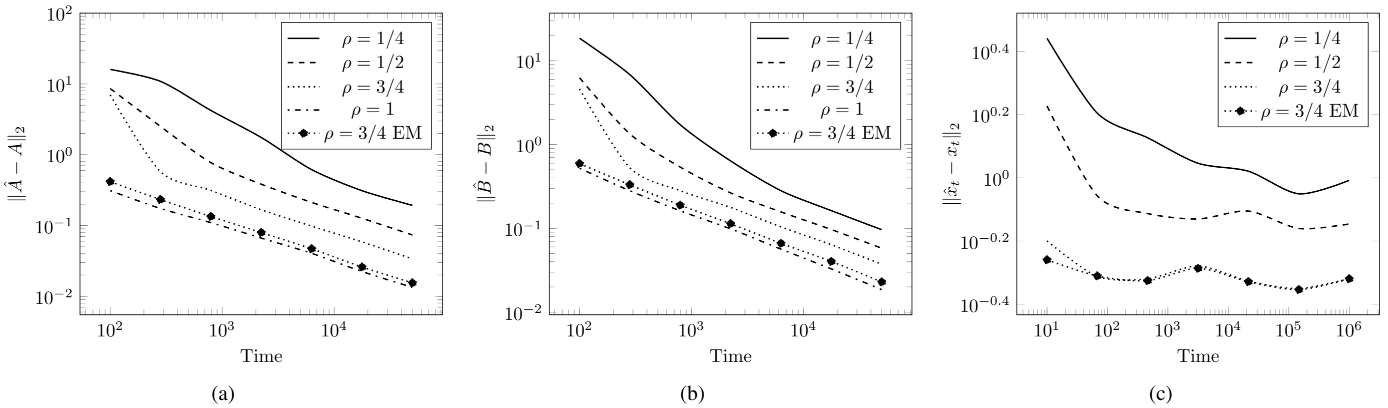
A. Joint System and State Identification

In this subsection, we consider an application to the joint system and state identification of a stable linear dynamical system with inputs that are chosen randomly and known. In the case where inputs are not randomly chosen but are known, classical system identification techniques can be applied to jointly infer the parameters and state [23]. However, there are no non-asymptotic guarantees in the literature.

Specifically, we consider the dynamical system

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t + w_t \\ z_t &= P_tx_t, \end{aligned}$$

where A and B are unknown. We start from unknown initial state x_0. At each point, we input $u_t \overset{\text{iid}}{\sim} \mathcal{N}(0,I)$ whose value is known. Noise $w_t \overset{\text{iid}}{\sim} \mathcal{N}(0,I)$ is added. $P_t = \mathrm{diag}(p_t)$ and



(a)

(b)

(c)

Fig. 1. Results for joint state and system identification with partially observed samples (a) Error in estimating the transition matrix with different sampling rates. (b) Error in estimating the input matrix B (c) Error in estimating the final state through Kalman filtering.

$p_t \overset{\text{iid}}{\sim} \mathcal{B}(\rho)$ or observations are seen with probability ρ and with no multiplicative noise. We can equivalently recast the problem with state and noise matrices

$$\hat{x}_t = \begin{bmatrix} x_t \\ u_t \end{bmatrix}, \hat{w}_t = \begin{bmatrix} w_t \\ u_{t+1} \end{bmatrix}$$

as,

$$\hat{x}_{t+1} = \begin{bmatrix} A & B \\ 0 & 0 \end{bmatrix} \hat{x}_t + \hat{w}_t$$
$$\hat{z}_t = \begin{bmatrix} P_t & 0 \\ 0 & I \end{bmatrix} \hat{x}_t.$$

This new transition matrix is stable if $\|A\|_2 \leq \sigma_{\max} < 1$ as it is an upper triangular matrix.

The matrix $\hat{P}_t$ is the effective multiplicative noise at each time instant. It is a diagonal matrix with entries $\hat{p}_t = [P_{t,1}, \ldots, P_{t,n}, \mathbf{1}(n)]$ where $P_{t,i} \overset{\text{iid}}{\sim} \mathcal{B}(\rho)$. We have $\theta(k) = \mathbb{E}[\hat{p}_t \hat{p}_t^\intercal]$. Thus, we run the algorithm with the following $\theta(k)$

$$\theta(k) = \begin{bmatrix} \rho^2 \mathbf{1}\mathbf{1}^\intercal + (\rho - \rho^2) I \mathbf{1}(k=0) & \rho \mathbf{1}\mathbf{1}^\intercal \\ \rho \mathbf{1}\mathbf{1}^\intercal & \mathbf{1}\mathbf{1}^\intercal \end{bmatrix}.$$

Since our matrix A has this structure, we estimate $\hat{A}_{1:n,:} = (\hat{\Sigma}^1_{:,1:n})^\intercal \hat{\Sigma}^0$. Given an estimate of the matrices $\hat{A}$ and $\hat{B}$, we can perform Kalman filtering and Rauch-Tung-Striebel (RTS) smoothing to estimate the state we are in [24].

We compare the results of estimator (3) to the EM estimate [25]. In the EM estimate, we solve for the joint ML estimate of the unobserved state sequence as well as system matrices A and B. This is a non-convex problem and is solved iteratively. In the E step, we compute an approximation to the log-likelihood function while keeping A and B constant - this results in Kalman filtering and RTS smoothing to obtain the hidden state sequence x_t. In the M step, we solve for A and B by maximizing the approximation to the log-likelihood while keeping the state-sequence constant. The method is computationally intensive with memory requirements $\mathcal{O}(T)$ larger than in estimator (3) as well as requiring $\mathcal{O}(T)$ more inversions of matrices of size $n \times n$. This method is not computationally feasible for high-dimensional systems. The estimate of A can in fact serve to initialize the EM procedure.

In this example, dimension $n = 7$ and dense transition matrices with $\sigma_{\max} = 0.8, \sigma_{\min} = 0.2$ are chosen. Results average 16 runs. Fig. 1 presents the error in estimating the transition matrices as well as the error in estimating the final state. As can be seen from the plots, the error in identifying the system has slope $-1/2$ which validates the theory that the error is proportional to $T^{-1/2}$. We also verify that as the sampling rate doubles, the error reduces by a factor of 4. The performance of the expensive EM procedure (with 10 iterations) is seen for $\rho = 3/4$ and it is seen to perform better than estimator (3).

B. *Sparse Transition Matrices*

In this example, we consider the system

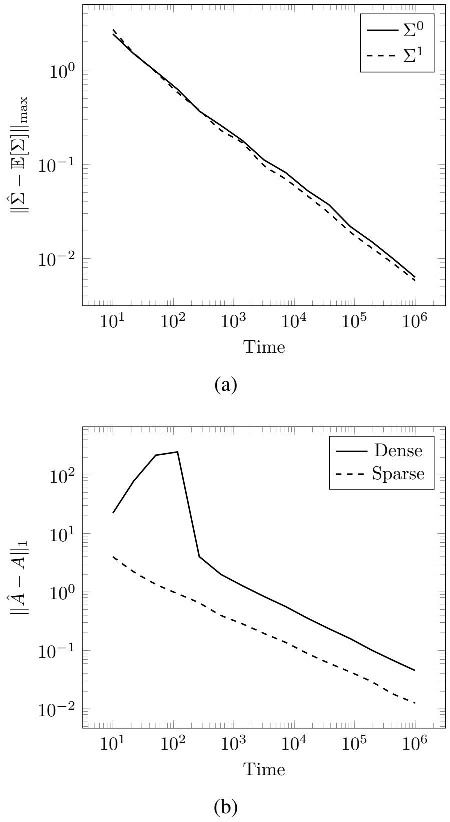
$$x_{t+1} = Ax_t + w_t$$
$$z_t = P_t(x_t + v_t),$$

where $w_t \overset{\text{iid}}{\sim} \mathcal{N}(0, I), v_t \overset{\text{iid}}{\sim} \mathcal{N}(0, I)$ and we have a sparse unknown A. The matrix P_t is diagonal with entries $p_t = [P_{t,1}, \ldots, P_{t,n}]$ and $P_{t,i} \overset{\text{iid}}{\sim} \mathcal{U}([0,1])$. Thus, we see all the observations corrupted by multiplicative noise and zero-mean additive noise. In this case, we cannot use the EM based estimator of the previous subsection as we do not know P_t, the multiplicative noise at each time instant.

We consider 10 instances of a sparse stable 30 dimensional system where there are an average of 20 non-zero elements in each. Results comparing the performance of the sparse and dense estimator are shown in Fig. 2. As expected the sparse identification method outperforms the dense estimator of the transition matrix. The scaling of the error in the covariance matrix is seen to be around $T^{-1/2}$ as predicted.

VI. Conclusion

We have considered the problem of system identification of vector autoregressive processes with partial observations corrupted by multiplicative and additive noise. This problem



(a)

(b)

Fig. 2. Results for sparse system identification with observations having multiplicative and additive noise (a) Maximum entrywise deviation of sample covariance matrices from the actual value. (b) Error in transition matrix estimation assuming it is dense and sparse.

is motivated by the difficulty in storing, processing or communicating a large number of observations in high-dimensional VAR processes.

An estimator of the covariance matrices of the process which can be arbitrarily initialized is first described. This is used to obtain the transition matrix in both the general case and one with structural constraint of sparsity which is relevant in the high-dimensional regime. The error in estimating the transition matrix scaled as $\mathcal{O}(\frac{1}{\rho\sqrt{T}})$, which implies that the number of time samples required to estimate the matrix to a certain accuracy increases quadratically as the sampling ratio decreases.

The estimators were validated through simulation and applied to the problem of jointly estimating the state and system of a stable linear dynamical system with random inputs.

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] B. Porat, *Digital processing of random signals: theory and methods*. Prentice-Hall, Inc., 1994.

[4] T. Kailath, "A view of three decades of linear filtering theory," *IEEE Transactions on Information Theory*, vol. 20, no. 2, pp. 146–181, 1974.

[5] 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.

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

[7] 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.

[8] 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.

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

[10] 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

[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] 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.

[13] 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.

[14] L. Balzano, R. Nowak, and B. Recht, "Online identification and tracking of subspaces from highly incomplete information," in *Proc. of the 48th Annual Allerton Conference on Communication, Control, and Computing*. IEEE, 2010, pp. 704–711.

[15] Y. Chi, Y. C. Eldar, and R. Calderbank, "Petrels: Parallel subspace estimation and tracking by recursive least squares from partial observations," *Signal Processing, IEEE Transactions on*, vol. 61, no. 23, pp. 5947–5959, 2013.

[16] A. Gonen, D. Rosenbaum, Y. C. Eldar, and S. Shalev-Shwartz, "Subspace learning with partial information," *Journal of Machine Learning Research*, vol. 17, no. 52, pp. 1–21, 2016.

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

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

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

[20] 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.

[21] M. Rao, A. Kipnis, T. Javidi, Y. C. Eldar, and A. Goldsmith, "System identification from partial samples: Proofs," http://stanford.edu/~milind/reports/system_id_cdc_proof.pdf, accessed: 2016-03-15.

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

[23] J.-N. Juang, M. Phan, L. G. Horta, and R. W. Longman, "Identification of observer/kalman filter markov parameters-theory and experiments," *Journal of Guidance, Control, and Dynamics*, vol. 16, no. 2, pp. 320–329, 1993.

[24] H. E. Rauch, C. Striebel, and F. Tung, "Maximum likelihood estimates of linear dynamic systems," *AIAA journal*, vol. 3, no. 8, pp. 1445–1450, 1965.

[25] S. Gibson and B. Ninness, "Robust maximum-likelihood estimation of multivariable dynamic systems," *Automatica*, vol. 41, no. 10, pp. 1667–1682, 2005.