# System Identification of High-Dimensional Linear Dynamical Systems With Serially Correlated Output Noise Components

Jiahe Lin and George Michailidis, *Member, IEEE*

***Abstract*—We consider identification of linear dynamical systems comprising of high-dimensional signals, where the output noise components exhibit strong serial, and cross-sectional correlations. Although such settings occur in many modern applications, such dependency structure has not been fully incorporated in existing approaches in the literature. In this paper, we explicitly incorporate the dependency structure present in the output noise through lagged values of the observed multivariate signals. We formulate a constrained optimization problem to identify the space spanned by the latent states, and the transition matrices of the lagged values *simultaneously*, wherein the constraints reflect the low rank nature of the state information, and the sparsity of the transition matrices. We establish theoretical properties of the estimators, and introduce an easy-to-implement computational procedure for empirical applications. The performance of the proposed approach, and the implementation procedure is evaluated on synthetic data, and compared with competing approaches, and further illustrated on a data set involving weekly stock returns of 75 US large financial institutions for the 2001–2017 period.**

***Index Terms*—Alternating minimization, convergence, convex optimization, high-probability error bounds.**

## I. Introduction

DISCRETE time linear dynamical systems (LDS) are widely used in a number of scientific fields including modeling the behavior of engineering control systems [54], as well as data in economics [28], [29] and medical studies [31], [49]. A linear time-invariant system, also known as a linear state-space model, assumes that the system consists of a multivariate *latent* state $f_t$ and a multivariate *observed* signal $X_t \in \mathbb{R}^p$, whose dynamics are governed by the following equations:

$$X_t = \Lambda f_t + u_t, \qquad \text{(observation equation)}$$
$$f_t = \Phi f_{t-1} + v_t, \qquad \text{(state equation)}$$

wherein both $u_t$ and $\eta_t$ are mean-zero Gaussian random vectors. One approach to identifying the parameters of such a system is the expectation-maximization (EM) algorithm [26], [46] built upon the Markov property of the system, where the E-step computes the expected log-likelihood using the Kalman Filter (KF; [32], [33]) and a forward-backward recursion, and the M-step calculates the parameter estimates by taking partial derivatives with respect to the Gaussian log-likelihood. Other popular approaches include subspace-based identification methods that relate the state space representation to more traditional input-output forms, e.g., N4SID [51], MOESP [52]; see also [43] and references therein.

A KF-based method is generally applicable to systems with states and output signals of moderate dimensions, but has its limitations when these dimensions become large. Past literature on the subject has considered various extensions of the classical KF to larger systems with the aid of regularization, and the corresponding modeling frameworks can be categorized as follows:

- **C1** Systems comprising of a large number of states. In this setting, despite the fact that the system has a large number of states, the effective or the intrinsic number of states is assumed to be small; system identification is accomplished either by imposing a sparsity constraint on the state estimates [5], [19] or a low-rank constraint on the state transition matrix [16]; other variants are discussed in [36], [45] and references therein.
- **C2** Systems comprising of a large panel of observed signals. In this setting, the number of states remains small, whereas that of the observed signals is large. For system identification, [20] proposed an EM-type algorithm with regularization on the state loading matrix, under the rather stringent assumption that the covariance matrix of the observation noise $u_t$ is diagonal, i.e., conditional on the states $f_t$, the observed signals are mutually independent. [34] considered distributing the calculations for the KF, assuming that the covariance matrix of $u_t$ is block diagonal, with each block having small dimension. [14] considered sketching the signals, which reduces their dimension and enables using standard KF-based algorithms.

Recent applications of LDS in electric circuits [39], neuroscience [49] and finance [23] involve large panels of signals whose sizes often exceed the number of available observations.

Manuscript received January 22, 2020; revised July 18, 2020; accepted August 23, 2020. Date of publication August 31, 2020; date of current version October 6, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Athanasios A. Rontogiannis. The work of George Michailidis was supported by NSF under Grants IIS 1632730, DMS 1821220, and DMS 1830175. *(Corresponding author: George Michailidis.)*

Jiahe Lin is with the Department of Statistics, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: jiahelin@umich.edu).

George Michailidis is with the Departments of Statistics, and Computer Science, and the Informatics Institute, University of Florida, Gainesville, FL 32611 USA (e-mail: gmichail@umich.edu).

This article has supplementary downloadable material available at https://ieeexplore.ieee.org, provided by the authors.

Code for implementing the proposed methodology is available online at https://github.com/jhlinplus/Linear_Dynamical_System.

Digital Object Identifier 10.1109/TSP.2020.3020397

Further, as noted in [20] such signals may exhibit additional temporal and cross-sectional dependency through their output noise, even after accounting for their common structure as captured by latent states. Note that such non-diagonal correlation structure for the output noise would potentially render the classical EM-algorithms infeasible due to the singularity of intermediate quantities involved in the E-step. Without further constraints on the structure of the model parameters (e.g., [20]), an alternative approach is to use principal component analysis, e.g., PCA-ID [21], which shares similarities with statistical factor analysis (e.g., [18], [48]). The latent states are estimated based on the Singular Value Decomposition of the observed signals $X_t$. However, for the number of latent states to be correctly identified and the parameter to be consistently estimated, the correlation amongst the coordinates of the observation noise $u_t$ needs to be sufficiently weak, as discussed in [8], [9]. In particular, the presence of strongly correlated coordinates in $u_t$ can lead to overestimation of the number of states [27] and also has detrimental effects for prediction [3].

In this paper, we consider identification of large scale LDS (setting C2) under a significantly more relaxed assumption on the correlation structure of $u_t$, which allows for both cross-sectional and temporal dependence amongst its coordinates. Further, we consider high-dimensional scaling of the LDS for our theoretical analysis, wherein the dimension of the signal grows with the number of observations (time points). Hence, the main contributions are: (a) the generalization of the linear state-space model to accommodate both high-dimensional observed signals and strongly serially and cross-correlated coordinates of the output noise; (b) the identification of system parameters and prediction of future signal values, through the formulation of a *penalized least squares* problem that is solved by a block-coordinate descent algorithm; and (c) establishing finite sample high-probability error bounds for the convergent solution estimates of the aforementioned problem.

The remainder of this paper is organized as follows. In Section II, we introduce our model setup, an identification procedure for the model parameters and discuss how to predict future values of the signal vector. Theoretical properties of the obtained estimates of the model parameters are established in Section III. In Section IV, we introduce an empirical implementation procedure and present the performance evaluation of the estimates based on synthetic data. In Section V, an application of our model to weekly stock return data of large US financial institutions for the period from 2001 to 2017 is considered. Finally, Section VI concludes the paper.

*Notation:* Throughout this paper, for some generic matrix $A$ of dimension $m \times n$, we use $|||\cdot|||$to denote its matrix norms, including the operator norm $|||A|||_{\text{op}}$, the Frobenius norm $|||A|||_{\text{F}}$, the nuclear norm $|||A|||_*$, $|||A|||_1 = \max_{1\le j\le n}\sum_{i=1}^m |a_{ij}|$, and $|||A|||_\infty = \max_{1\le i\le m}\sum_{j=1}^n |a_{ij}|$. We use $\|A\|_1 = \sum_{i,j}|a_{ij}|$ and $\|A\|_\infty = \max_{i,j}|a_{ij}|$ to denote the element-wise 1-norm and infinity norm. Additionally, we use $\varrho(A)$ to denote its spectral radius $(\max|\lambda(A)|)$. For two matrices $A$ and $B$ of commensurate dimensions, denote their inner product by $\langle\langle A, B\rangle\rangle = \text{trace}(A^\top B)$. Finally, we write $A \gtrsim B$ if there exists some absolute constant $c$ that is independent of the model parameters such that $A \ge cB$; $A \lesssim B$ is analogously defined.

## II. Problem Formulation, Identification and Prediction

We start by introducing the model in question assuming that the output noise $u_t$ in the (observation equation) follows a sparse $\text{VAR}(d)$ model that simultaneously incorporates the cross-sectional and serial structure amongst its coordinates; that is, $u_t = \mathcal{B}_d(L)\epsilon_t$ where $L$ is the lag operator, and $\mathcal{B}_d(L) := \text{I}_p - B_1 L - B_2 L^2 - \cdots B_d L^d$ is the lagged matrix polynomial for some *weakly sparse* $B_i$'s. To convey the main arguments, we assume without loss of generality that $d = 1$ and let $\mathcal{B}(L) \equiv \mathcal{B}_1(L) = \text{I}_p - BL$ for ease of exposition, and present the extension to the general lag case in Supplement-I.

To this end, consider an LDS comprising of the latent state variable $f_t$ and the observed signal $X_t \in \mathbb{R}^p$; $f_t$ follows some VAR model with lagged polynomial $\Phi(L) := \text{I} - \sum_{h=1}^w \Phi_h L^h$ and the dynamics of $X_t$ are governed by the latent factor $f_t$; that is,

$$\begin{aligned} f_t &= \Phi_1 f_{t-1} + \cdots \Phi_w f_{t-w} + v_t, \\ X_t &= \tilde{\Lambda} f_t + u_t; \end{aligned} \tag{1}$$

$u_t$ is the output noise whose dynamics satisfy $\mathcal{B}(L)u_t = \epsilon_t$, as previously mentioned. Multiplying both sides of (1) by $\mathcal{B}(L)$ leads to the following lag-adjusted representation for the observation equation:

$$X_t = \Lambda F_t + BX_{t-1} + \epsilon_t, \tag{2}$$

where $F_t \in \mathbb{R}^K (K \ll p)$ collects the lags of $f_t$, so that it only impacts the dynamics of $X_t$ *contemporaneously* through $\Lambda$, i.e.,

$$F_t = \begin{bmatrix} f_t \\ f_{t-1} \end{bmatrix} \quad \text{and} \quad \Lambda = \left[\tilde{\Lambda},\, -B\tilde{\Lambda}\right].$$

Further, $\epsilon_t$ is a mean zero white noise process and is *strictly exogenous*, satisfying $\text{Cov}(X_{t-1}, \epsilon_{t+h}) = 0$ and $\text{Cov}(F_t, \epsilon_{t+h}) = 0, \forall\, h \ge 0$. The primary focus of this paper is on the observation equation in (2), including the identification of its components (transition matrix $B$, the loading matrix $\Lambda$ and the state $F_t$) and the prediction of the observed signal $X_t$, under high-dimensional scaling, where both the number of observations $T$ and the size of the signal $p$ grow. Note that we only require $B$ to be *weakly sparse*, the notion of which can be formalized through the definition of an $\ell_q$ ball constraint of radius $R_q$ for some fixed $q \in [0, 1]$ (c.f., [41]):

$$\mathbb{B}_q(R_q) := \left\{ B \in \mathbb{R}^{p\times p} : \sum_{i,j}^p |B_{ij}|^q \le R_q \right\}. \tag{3}$$

The case of exact sparsity corresponds to $q = 0$, where $B \in \mathbb{B}_q(R_0)$ has at most $R_0$ nonzero entries; whereas for $q \in (0, 1]$, the $R_q$ ball imposes constraints on the decay rate of the elements $|B_{ij}|$'s.

To ensure that the observed signal process $X_t$ is covariance stationary, we require that the spectral radius of $B$ satisfies $\varrho(B) < 1$ without further restricting $\Lambda$. Additionally, under the

assumption that the spectral density of $X_t$ exists, the spectral density of the filtered process $Z_t := \mathcal{B}(L)X_t = \Lambda F_t + \epsilon$ satisfies

$$g_Z(\omega) = \Lambda g_F(\omega)\Lambda^\top + g_\epsilon(\omega) + g_{\epsilon,F}(\omega)\Lambda^\top + \Lambda g_{F,\epsilon}(\omega);$$

correspondingly, the spectral density of $X_t$ is given by

$$g_X(\omega) = \left[\mathcal{B}^{-1}(e^{i\omega})\right] g_Z(\omega) \left[\mathcal{B}^{-1}(e^{i\omega})\right]^*.$$

Note that $g_X(\omega)$ and $g_{X,Y}(\omega)$ denote the spectrum and cross-spectrum of some generic processes $\{X_t\}$ and $\{Y_t\}$:

$$g_X(\omega) := \frac{1}{2\pi}\sum_{h=-\infty}^{\infty} \Sigma_X(h)e^{-i\omega h} \quad \text{and}$$

$$g_{X,Y}(\omega) := \frac{1}{2\pi}\sum_{h=-\infty}^{\infty} \Sigma_{X,Y}(h)e^{-i\omega h}, \tag{4}$$

with $\Sigma_X(h) := \mathbb{E}(X_t X_{t-h}^\top)$ and $\Sigma_{X,Y}(h) := \mathbb{E}(X_t Y_{t-h}^\top)$.

### A. Identification Through a Convex Program

First note that the parameters of the model posited in (2) are not uniquely identifiable in the absence of additional constraints. In this study, we tackle the parameter identification problem through the following two steps: in Step (a), we formulate a convex program and develop an algorithm that can identify the state hyperplane (defined in the sequel) and the transition matrix $B$; subsequently, in Step (b), we reconstruct the latent states from the estimated state hyperplane, subject to the imposed identifiability constraint.

*a) Identification of the state hyperplane and the transition matrix:* Given a snapshot of the $p$-dimensional observable $X_t$ process $\{x_0, x_1, \ldots, x_T\}$ of length $(T+1)$, let $\mathsf{X}_T := [x_1, \ldots, x_T]^\top$ and $\mathsf{X}_{T-1} := [x_0, \ldots, x_{T-1}]^\top$ respectively denote the response matrix and the lagged regressor matrix of size $T \times p$; $\mathsf{F} \in \mathbb{R}^{T\times K}$ and $\mathsf{E} \in \mathbb{R}^{T\times p}$ are analogously defined with $F_t$'s and $\epsilon_t$'s $(t = 1, \ldots, T)$ stacked in their rows, respectively. We additionally define the *state hyperplane* associated with $\mathsf{F}$ as $\Theta := \mathsf{F}\Lambda^\top \in \mathbb{R}^{T\times p}$, and note that $\Theta$ is a low-rank component with rank at most $K$. With the above notation, the sample version for the observation equation in (2) can be written as

$$\mathsf{X}_T = \Theta + \mathsf{X}_{T-1}B^\top + \mathsf{E}.$$

Under the assumption that the transition matrix $B \in \mathbb{R}^{p\times p}$ is sparse and the state hyperplane $\Theta = \mathsf{F}\Lambda^\top \in \mathbb{R}^{T\times p}$ is low rank, we formulate the following constrained optimization problem:

$$\min_{B,\Theta} \left\{\tfrac{1}{2T}\|\mathsf{X}_T - \Theta - \mathsf{X}_{T-1}B^\top\|_{\mathrm{F}}^2\right\},$$

$$\text{subject to } \operatorname{rank}(\Theta) \le r,\ \|B\|_1 \le \xi, \text{ for some } r \text{ and } \xi, \tag{5}$$

whose feasible region is determined through a rank constraint imposed on $\Theta$ and a sparsity-inducing norm constraint imposed on $B$.

The rank constraint in (5) renders the feasible region *non-convex* and makes it particularly hard to characterize the obtained solution analytically, which depends on the initial values provided to the algorithm. Thus, as commonly undertaken in the literature (e.g., [2]), we consider a tight convex relaxation of the rank constraint, and the solution to the convexified program has convergence guarantees independent of the initializer. Formally, we consider obtaining the estimator through the convex program in (6), which can be derived from (5) by alternatively considering the nuclear norm constraint for the state hyperplane and the $\ell_1$ norm constraint for the sparse transition matrix $B$ in Lagrangian form:

$$(\widehat{B}, \widehat{\Theta}) = \arg\min_{B,\Theta} \mathcal{L}(B, \Theta; \mathsf{X}_T, \mathsf{X}_{T-1}, \lambda_B, \lambda_\Theta),$$

$$\mathcal{L} := \frac{1}{2T}\|\mathsf{X}_T - \Theta - \mathsf{X}_{T-1}B^\top\|_{\mathrm{F}}^2 + \lambda_B\|B\|_1 + \lambda_\Theta\left\|\tfrac{\Theta}{\sqrt{T}}\right\|_*, \tag{6}$$

where $\lambda_B$ and $\lambda_\Theta$ are tuning parameters. The solution $(\widehat{B}, \widehat{\Theta})$ can be obtained by a block-coordinate descent algorithm that alternately minimizes with respect to $B$ and $\Theta$, as outlined in Algorithm 1.

*b) Reconstruction of the states:* The solution to (8) provides an estimate of the state hyperplane, based on which realizations of the $K$-dimensional latent state process $\{F_t\}$ can be reconstructed under certain identifiability restrictions. Note that for any invertible matrix $Q \in \mathbb{R}^{K\times K}$, the following equality holds:

$$\Theta = \mathsf{F}\Lambda^\top = \left[\mathsf{F}Q^\top\right]\left[\Lambda Q^{-1}\right]^\top := \check{\mathsf{F}}\check{\Lambda}^\top;$$

hence given $\Theta$, to fully identify the states and the corresponding loading matrix $(\mathsf{F}, \Lambda)$ from their observationally equivalent counterpart $(\check{\mathsf{F}}, \check{\Lambda})$, a total number of $K^2$ restrictions is required

---

**Algorithm 1:** Estimating $B$ and $\Theta$ by Solving (6).

**Input:** Observed signals $\{x_t\}_{t=0}^T$, tuning parameters $\lambda_B$, $\lambda_\Theta$

1 **Initialization: set** $B^{(0)} = O$;

2 **Iterate until convergence**:

3 For fixed $\widehat{B}^{(m-1)}$, update $\widehat{\Theta}^{(m)}$ as:

$$\widehat{\Theta}^{(m)} = \arg\min_\Theta \mathcal{L}(\Theta; \widehat{B}^{(m-1)}, \mathsf{X}_T, \mathsf{X}_{T-1}, \lambda_B, \lambda_\Theta),$$

where the minimum can be obtained by a proximal gradient descent algorithm involving a soft singular value thresholding (SVT) step; each inner iteration indexed by $t$ solves the following minimization with some stepsize $\zeta$:

$$\widehat{\Theta}^{(m,[t+1])} = \arg\min_\Theta \Big\{\langle\!\langle \Theta, \nabla\mathcal{G}^{(m)}(\widehat{\Theta}^{(m,[t])})\rangle\!\rangle + (\zeta/2)\|\Theta - \Theta^{(m,[t])}\|_{\mathrm{F}}^2 + \lambda_\Theta\left\|\tfrac{\Theta}{\sqrt{T}}\right\|_*\Big\},$$

$$\mathcal{G}^{(m)}(\Theta) := \tfrac{1}{2T}\|\mathsf{X}_T - \mathsf{X}_{T-1}(\widehat{B}^{(m-1)})^\top - \Theta\|_{\mathrm{F}}^2.$$

4 For fixed $\widehat{\Theta}^{(m)}$, update $\widehat{B}^{(m)}$ as

$$\widehat{B}^{(m)} = \arg\min_B \mathcal{L}(\widehat{B}^{(m-1)}; \widehat{\Theta}^{(m)}, \mathsf{X}_T, \mathsf{X}_{T-1}, \lambda_B, \lambda_\Theta),$$

where each row $j$ solves a Lasso regression (in parallel):

$$\widehat{B}_{j\cdot}^{(m)} = \arg\min_{\beta\in\mathbb{R}^p}\left\{\tfrac{1}{2T}\left\|\left[\mathsf{X}_T - \widehat{\Theta}^{(m)}\right]_{\cdot j} - \mathsf{X}_{T-1}\beta\right\|^2 + \lambda_B\|\beta\|_1\right\}.$$

**Output:** Estimated sparse transition matrix $\widehat{B}$ and the low rank hyperplane $\widehat{\Theta}$.

---

to address such indeterminacy. Specifically, let the singular value decomposition of $\widehat{\Theta}$ be $\widehat{\Theta} = \widehat{U}\widehat{D}\widehat{V}^\top$. Depending on the application of interest, the following identification restrictions lead to state estimates with different properties (see also [10]):

- R1 Orthogonal states: the states are assumed orthogonal, i.e., $\frac{1}{T}\mathsf{F}^\top\mathsf{F} = I_K$, with $\Lambda^\top\Lambda$ being diagonal. In this case, the states' estimate is given by $\widehat{\mathsf{F}} = \sqrt{T}\widehat{U}$.
- R2 Orthogonal loadings: the loading are assumed orthogonal, i.e., $\frac{1}{p}\Lambda'\Lambda = I_K$, with $\mathsf{F}^\top\mathsf{F}$ being diagonal. In this case, the state' estimate is given by $\widehat{\mathsf{F}} = \frac{1}{\sqrt{p}}\widehat{U}\widehat{D}$.
- R3 Unrestricted states: the $K \times K$ upper sub-matrix of $\Lambda$ is assumed an identity matrix and the states $\mathsf{F}$ are left unrestricted, i.e., $\Lambda = [\begin{smallmatrix} I_K \\ * \end{smallmatrix}]$. Consequently, the states' estimate $\widehat{\mathsf{F}}$ is given by the first $K$ columns of $\widehat{\Theta}$, i.e., $\widehat{\mathsf{F}} = \widehat{\Theta}_{,1:K}$.

It is worth noting that irrespective of the identification restrictions, the space spanned by the estimated states is invariant once $\widehat{\Theta}$ is obtained; moreover, predicting future values of $X_t$ does not require an exact recovery of $F_t$, as discussed next.

### B. Signal/Output Prediction

Given estimates $\widehat{B}$ of the transition matrix and $\widehat{\Theta}$ of the hyperplane, we consider the following procedure that first obtains prediction of the filtered process $Z_t := X_t - BX_{t-1}$ through projection onto the space spanned by the states, followed by a lag adjustment to obtain those of the $X_t$ signal.

To this end, according to the model in (2), the filtered process $Z_t$ can be represented as $Z_t = \Lambda F_t + \epsilon_t$, whose $h$-step-ahead best linear predictor based on $F_{T-k}, k \geq 0$ is given by the projection $\text{Proj}(Z_{T+h} \,|\, \text{Span}(\mathsf{F}, T))$, where $\text{Span}(\mathsf{F}, T)$ denotes the linear space spanned by $\{F_t\}_{t=1}^T$ [47]. In particular, based on estimate $\widehat{B}$, the filtered process $Z_t$ can be estimated through $\widehat{z}_t := x_t - \widehat{B}x_{t-1}$, whose common space estimate corresponds to $\widehat{\Theta}$. Using the surrogate process $\{\widehat{z}_t\}$, let the sample covariance be $\widehat{\Sigma}_Z(h) := \frac{1}{T-h}\sum_{t=h+1}^T \widehat{z}_t\widehat{z}_{t-h}^\top$; the $h$-step-ahead prediction of $\{z_t\}$ is then given by

$$\widehat{z}_{T+h} = \left\{\widehat{\Sigma}_Z(h)\widehat{V}\left[\widehat{V}^\top\widehat{\Sigma}_Z(0)\widehat{V}\right]^{-1}\right\}(\widehat{V}^\top\widehat{z}_T), \qquad (7)$$

where columns of $\widehat{V}$ are the right singular vectors of $\widehat{\Theta}$ corresponding to the nonzero singular values, and they are effectively a set of orthonormal bases for the space spanned by the states. In the case where $h = 1$, $\widehat{x}_{T+1} = \widehat{B}x_T + \widehat{z}_{T+1|T}$; in the case where $h > 1$, $\widehat{x}_{T+h}$ can be obtained inductively by sequentially estimating $x_{T+i}$, for all $0 < i \leq h$. Algorithm 2 outlines the prediction procedure.

The empirical performance of the parameter estimation and the prediction procedure is considered in Section IV under various data generating mechanisms.

**Algorithm 2:** Obtaining an $h$-Step-Ahead Prediction of the Signals.

**Input:** Time series data $\{x_t\}_{t=0}^T$, estimates $\widehat{B}$ and $\widehat{\Theta}$

1 Obtain the filtered process $\widehat{z}_t := \widehat{x}_t - \widehat{B}x_{t-1}$ and its sample cross-covariance estimates $\widehat{\Sigma}_z(i)$, for all $0 \leq i \leq h$;

2 For $i = 1, 2, \ldots, h$:

3 Obtain $\widehat{z}_{T+i}$ through

$$\widehat{z}_{T+i} := \left\{\widehat{\Sigma}_Z(i)\widehat{V}\left[\widehat{V}^\top\widehat{\Sigma}_Z(0)\widehat{V}\right]^{-1}\right\}(\widehat{V}^\top\widehat{z}_T)$$

where columns of $\widehat{V}$ are the right singular vectors of $\widehat{\Theta}$ corresponding to the nonzero singular values;

4 Obtain $\widehat{x}_{T+i}$ through $\widehat{x}_{T+i} := \widehat{z}_{T+i} + \widehat{B}\widehat{x}_{T+i-1}$.

**Output:** Predicted values $\widehat{x}_{T+i}, 0 < i \leq h$.

## III. Theoretical Properties

To establish statistical properties of the estimators, a ball constraint on the feasible region of $\Theta$ is required to incur additional compactness on the low rank component that limits the spikiness of its entries, and this enables identification of the sparse component $B$. To this end, throughout this section, we consider estimators that are solutions to the following convex program:

$$(\widehat{B}, \widehat{\Theta}) = \arg\min{}_{B,\Theta}\mathcal{L}, \text{ subject to } \Theta \in \mathbb{B}_\infty(\phi, \mathsf{X}_{T-1})$$
$$\mathcal{L} := \tfrac{1}{2T}|||\mathsf{X}_T - \Theta - \mathsf{X}_{T-1}B^\top|||_{\mathrm{F}}^2 + \lambda_B\|B\|_1 + \lambda_\Theta|||\tfrac{\Theta}{\sqrt{T}}|||_*, \qquad (8)$$

where $\mathbb{B}_\infty(\phi, \mathsf{X}_{T-1})$ is a box constraint given by

$$\mathbb{B}_\infty(\phi, \mathsf{X}_{T-1}) := \left\{\Theta : \|\Theta\|_\infty \leq \frac{\phi}{\sqrt{Tp}\cdot|||\mathsf{X}_{T-1}/\sqrt{T}|||_{\text{op}}}\right\}.$$

$\phi$ is chosen such that the true value of the parameters $\Theta^\star$ is always feasible. We will provide further illustration on the interpretation of such a box constraint in Section III-A; see also Remark 4 in Supplement-III. $(\widehat{B}, \widehat{\Theta})$ falls into the class of *regularized $M$-estimators*, whose theoretical properties have been extensively studied in the statistical literature for diverse settings (e.g., [1], [38]).

A road map to establish properties of the estimators for $(\Theta, B)$ is given next: first in Section III-A we derive non-asymptotic statistical error bounds for $\widehat{\Theta}$ and $\widehat{B}$ under certain regularity conditions, when the proposed estimation procedure is based on a *deterministic realization* of the observable process $\{X_t\}$. In particular, the required regularity conditions primarily entail the *restricted strong convexity* (RSC) condition [2] and that the choice of $\lambda_B$ and $\lambda_\Theta$ being in accordance with some *deviation condition* [38]. Subsequently, in Section III-B, we establish that the required conditions are satisfied with high probability, and provide probabilistic analogues of key model parameters' error bounds for *random realizations* drawn from the underlying observable Gaussian process $\{X_t\}$ and the latent process $\{F_t\}$. We also briefly discuss how the model identifiability issue is tackled through the constrained formulation adopted in (8). Finally in Section III-C, from a numerical perspective, we establish the convergence of the proposed iterative algorithm to a stationary point. All proofs are deferred to Appendices A, B and the Supplement. Throughout our exposition, we use superscript $\star$ to

denote the true value of the parameters of interest, and denote the errors of the estimators by $\Delta_\Theta := \widehat{\Theta} - \Theta^\star$ and $\Delta_B := \widehat{B} - B^\star$, respectively. Additionally, the underlying processes are assumed Gaussian, although empirically the proposed estimator exhibits decent performance even in the presence of heavy tails (see Section IV).

### A. Statistical Error Bounds with Deterministic Realizations

We start by introducing additional notation required for the ensuing technical developments. Let $\ell_T(B, \Theta; \mathsf{X})$ denote the loss function, given by

$$\ell_T(B, \Theta; \mathsf{X}) := \frac{1}{2\,T} |||\mathsf{X}_T - \Theta - \mathsf{X}_{T-1} B^\top|||_{\mathrm{F}}^2.$$

The dimension of the latent state $F_t$ is given by $K$ and thus $\mathrm{rank}(\Theta^\star) = K$. Further, given some thresholded level $\eta > 0$, let $\mathcal{S}_\eta^\star$ denote the *strong* support set of $B^\star$, and we use $s_\eta$ to denote its cardinality, that is,

$$\mathcal{S}_\eta^\star := \{(i,j) : |B_{ij}^\star| > \eta\} \quad \text{and} \quad s_\eta := \|\mathcal{S}_\eta^\star\|_0. \tag{9}$$

Finally, let $\widehat{\Sigma}_{\mathsf{E}} := \frac{1}{T}\mathsf{E}^\top\mathsf{E}$ denote the sample covariance matrix of the noise process and let $\Lambda_{\max}(\widehat{\Sigma}_{\mathsf{E}})$ be its maximum eigenvalue. Formally, the *RSC condition* (c.f., [2], [41]) is defined as follows.

*Definition 1. (Restricted Strong Convexity (RSC)):* For some generic data matrix $\mathsf{X} \in \mathbb{R}^{T\times p}$, it satisfies the RSC condition with respect to norm $\Phi$ with curvature $\alpha_{\mathrm{RSC}} > 0$ and tolerance $\tau_{\mathrm{tol}} \geq 0$ if

$$\frac{1}{2\,T}|||\mathsf{X}\Delta|||_{\mathrm{F}}^2 \geq \frac{\alpha_{\mathrm{RSC}}}{2}|||\Delta|||_{\mathrm{F}}^2 - \tau_{\mathrm{tol}}\Phi^2(\Delta), \qquad \forall\, \Delta \in \mathbb{R}^{p\times p}.$$

In our context, we consider the $\ell_1$ norm regularizer and thus $\Phi(\Delta) = \|\Delta\|_1$.

Further, for high dimensional sparse VAR models (which corresponds to $\Theta = 0$ in the proposed formulation in (2)), the tuning parameter $\lambda_B$ needs to satisfy a *deviation condition* [13], [38], namely,

$$\lambda_B \geq c_0 \|\nabla_B \ell_T - \nabla_B^2 \ell_T (B^\star)^\top\|_\infty, \text{ for some constant } c_0 > 0,$$

which can be simplified to $\lambda_B \geq \|\mathsf{X}_{T-1}^\top \mathsf{E}/T\|_\infty$. Under the current model setup, however, the deviation condition is significantly more involved and requires proper modifications to incorporate quantities associated with the state hyperplane, as seen in Theorem 1.

Before stating the main results, we provide a brief discussion on the box constraint on $\Theta$, which aims to "limit" the spikiness of the low rank component, and hence the interaction between the spaces respectively spanned by the latent states and the observable lag $X_{t-1}$ — in particular, for $\Theta$ and $B$ to be properly recovered, such interaction can not be too large. Due to the basis vectors of the space spanned by the states being latent, a direct restriction on the interaction is impractical and conceptually unsatisfying, whereas the box constraint adopted effectively restricts the product of the signals from the two spaces and serves our objective, as shown in the proof of Theorem 1 and Remark 3. Note that this constraint is in similar spirit to the ones in the literature (e.g., [2], [40]), and the presence of the norm of $\mathsf{X}$ in the box constraint is necessary due to the two spaces having distinct bases.

*Theorem 1 (Error bound for $(\widehat{B}, \widehat{\Theta})$ under fixed realizations):* Suppose fixed realizations $\mathsf{X}_{T-1} \in \mathbb{R}^{T\times p}$ of process $X_t \in \mathbb{R}^p$ satisfy the RSC condition with curvature $\alpha_{\mathrm{RSC}} > 0$ and a tolerance $\tau_{\mathrm{tol}}$ such that

$$128\tau_{\mathrm{tol}}\left(s_\eta + (2\,K)\left(\frac{\lambda_\Theta}{\lambda_B}\right)^2\right) < \min\{\alpha_{\mathrm{RSC}}, 1\}. \tag{10}$$

Then, for any matrix pair $(B^\star, \Theta^\star)$ that generates the evolution of the $X_t$ process, for estimators $(\widehat{B}, \widehat{\Theta})$ obtained by solving the optimization (8) with regularization parameters $\lambda_B$ and $\lambda_\Theta$ satisfying

$$\lambda_B \geq 2\|\mathsf{X}_{T-1}^\top \mathsf{E}/T\|_\infty + 4\phi/\sqrt{Tp} \text{ and } \lambda_\Theta \geq \Lambda_{\max}^{1/2}(\widehat{\Sigma}_{\mathsf{E}}), \tag{11}$$

the following error bound holds for some positive constants $C_1$, $C_2$ and $C_3$:

$$|||\Delta_B|||_{\mathrm{F}}^2 + |||\Delta_\Theta/\sqrt{T}|||_{\mathrm{F}}^2 \leq C_1 \cdot \mathcal{E}_B + C_2 \cdot \mathcal{E}_\Theta + C_3 \cdot \mathcal{E}_{\tau_{\mathrm{tol}}}, \tag{12}$$

where $\alpha' := \min\{\alpha_{\mathrm{RSC}}, 1\}$,

$$\mathcal{E}_B := \left(\frac{\lambda_B}{\alpha'}\right)^2 \left\{ s_\eta + \frac{\alpha'}{\lambda_B} \sum_{(i,j)\notin\mathcal{S}_\eta^\star} |B_{ij}^\star| \right\},$$

$$\mathcal{E}_\Theta := \left(\frac{\lambda_\Theta}{\alpha'}\right)^2 K,$$

$$\mathcal{E}_{\tau_{\mathrm{tol}}} := \left(\frac{\tau_{\mathrm{tol}}}{\alpha'}\right) \left( \sum_{(i,j)\notin\mathcal{S}_\eta^\star} |B_{ij}^\star| \right)^2.$$

Next, we comment on the error bound in (12) and the required conditions in (10). The error bound encompasses three terms that are respectively associated with the transition matrix $B$, the low rank component $\Theta$, and the tolerance $\tau_{\mathrm{tol}}$ that measures the extent to which the log-likelihood function deviates from strong convexity (see Definition 1). Both $\mathcal{E}_B$ and $\mathcal{E}_\Theta$ depend on three components: (1) the overall curvature of the objective function as captured by $\alpha_{\mathrm{RSC}}$, (2) the interaction structure between various components of the underlying processes, as captured by the tuning parameters $\lambda_B$ and $\lambda_\Theta$, and (3) the inherent structure of the parameters as captured by $s_\eta$, $\sum_{(i,j)\notin\mathcal{S}_\eta^\star} |B_{ij}|$ and $K$ — in particular, due to the approximately sparse structure of $B^\star$, both the density level $s_\eta$ of its strong support set and the magnitude of its "weak" entries play a role, with the two respectively reflecting the *estimation error* and the *approximation error* (c.f., [2]) after proper scaling. The curvature as measured by $\alpha_{\mathrm{RSC}}$ dictates the constraint to which the tolerance $\tau_{\mathrm{tol}}$ needs to conform (see Equation (10)), and such a constraint is also interrelated to $s_\eta$ and $K$ — for (10) to be satisfied, neither $K$ nor $s_\eta$ can be too large. Regarding the tuning parameters, $\lambda_B$ can be sub-divided into two terms: the cross-product term $\|\mathsf{X}_{T-1}^\top \mathsf{E}/T\|_\infty$ measures the maximum interaction between the design matrix $\mathsf{X}_{T-1}$ and the noise $\mathsf{E}$, which according to the model assumption (population level) should center around 0; the term $\phi/\sqrt{Tp}$ corresponds to

an upper bound on the interaction between the latent and the observed spaces, respectively spanned by $F_t$ and $X_{t-1}$. For $\lambda_\Theta$, we require that it dominates the maximum signal coming from the white noise process in the form of $\Lambda_{\max}^{1/2}(\widehat{\Sigma}_{\mathsf{E}})$. Thus, a smaller $\lambda_B$ is needed when interactions between associated terms are weaker and similarly a smaller $\lambda_\Theta$ is needed if the magnitude of the noise is weaker, thus leading to a tighter error bound for the estimates. Finally, it is worth noting that $\mathcal{E}_{\tau_{\text{tol}}}$ is a result of the weakly sparse structure of $B$; in the special case where $B$ is exactly sparse, this term would be 0.

Corollary 1 gives the bound of $\Delta_B$ and $\Delta_\Theta$ with specific choice of the thresholded level $\eta$, which determines the strong support set (see Equation (9)) of the true value $B^\star$ lying in the $\ell_q$ ball of radius $R_q$ (see definition in Equation (3)).

*Corollary 1:* Under the same set of conditions as in Theorem 1, with $B^\star \in \mathbb{B}_q(R_q)$, by choosing the thresholded level according to $\eta = \lambda_B/\alpha'$ where $\alpha' := \min\{\alpha_{\text{RSC}}, 1\}$, the following error bound holds for some positive constants $C_1, C_2$ and $C_3$:

$$|||\Delta_B|||_{\text{F}}^2 + |||\Delta_\Theta/\sqrt{T}|||_{\text{F}}^2 \leq C_1 \cdot \left(\frac{\lambda_B}{\alpha'}\right)^{2-q} R_q + C_2 \cdot \left(\frac{\lambda_\Theta}{\alpha'}\right)^2 K + C_3 \cdot \frac{\tau_{\text{tol}}}{\alpha'} \left(\frac{\lambda_B}{\alpha'}\right)^{2-2q} R_q^2.$$

### B. High Probability Bounds Under Random Realizations

Next, we provide high probability bounds/concentrations for key quantities associated with the derived error bound in Section III-A, for random realizations of the underlying Gaussian processes. Specifically, this involves the verification of the RSC condition, as well as the examination of quantities associated with the deviation condition to which the choice of $(\lambda_B, \lambda_\Theta)$ needs to conform, as shown in (11).

We introduce additional notation for subsequent developments. For some generic process $\{X_t\}$, in addition to the autocovariance function $\Sigma_X(h)$ and its spectral density $g_X(\omega)$, we define its maximum and minimum eigenvalue associated with the spectral density $g_X(\omega)$ introduced in Section II as follows (see also [13]):

$$\mathcal{M}(g_X) := \operatorname*{ess\,sup}_{\omega\in[-\pi,\pi]} \Lambda_{\max}(g_X(\omega)),$$

$$\mathfrak{m}(g_X) := \operatorname*{ess\,inf}_{\omega\in[-\pi,\pi]} \Lambda_{\min}(g_X(\omega)).$$

For two generic centered processes $\{X_t\}$ and $\{Y_t\}$ that are assumed jointly covariance stationary, whose cross-spectral density is given by $g_{X,Y}(\omega)$ (see (4)), the upper extreme for $g_{X,Y}(\omega)$ is analogously defined as

$$\mathcal{M}(g_{X,Y}) := \operatorname*{ess\,sup}_{\omega\in[-\pi,\pi]} \sqrt{\Lambda_{\max}\left(g_{X,Y}^*(\omega) g_{X,Y}(\omega)\right)}.$$

In general $g_{X,Y}(\omega) \neq g_{Y,X}(\omega)$, but $\mathcal{M}(g_{X,Y}) = \mathcal{M}(g_{Y,X})$.

For the processes involved in our proposed model, recall that $\{X_t\}$, $\{\epsilon_t\}$ and $\{F_t\}$ are mean zero Gaussian processes. In particular, $\{\epsilon_t\}$ is a noise process that does not exhibit temporal nor cross-sectional dependence, hence it is effectively a Gaussian random vector with covariance $\Sigma_\epsilon = \sigma_\epsilon^2 \mathrm{I}_p$, and its spectral density simplifies to $g_\epsilon(\omega) = \frac{\Sigma_\epsilon}{2\pi}$. Further, we define the shifted process $\{\widetilde{\epsilon}_t := \epsilon_{t+1}\}$ for notation convenience.

The following lemma verifies that with high probability, for random realizations of the process $\{X_t\}$, the RSC condition is satisfied provided that the sample size is sufficiently large.

*Lemma 1 (verification of the RSC condition):* Consider $\mathsf{X} \in \mathbb{R}^{T\times p}$ whose rows are some random realization $\{x_0, \ldots, x_{T-1}\}$ of the stable $\{X_t\}$ process with dynamic given in (2). Then there exist positive constants $c_i\,(i = 1, 2)$ such that with probability at least $1 - c_1 \exp(-c_2 T)$, the RSC condition holds for $\mathsf{X}$ with curvature $\alpha_{\text{RSC}}$ and tolerance $\tau_{\text{tol}}$ satisfying

$$\alpha_{\text{RSC}} = \pi \mathfrak{m}(g_X), \quad \text{and} \quad \tau_{\text{tol}} = \gamma^2 \left(\frac{\alpha_{\text{RSC}}}{2}\right)\left(\frac{\log p}{T}\right),$$

where $\gamma := 54\mathcal{M}(g_X)/\mathfrak{m}(g_X)$, provided that $T \gtrsim s_\eta \log p$.

The next lemma establishes a high probability bound for the interaction term $(\mathsf{X}_{T-1}^\top \mathsf{E}/T)$ that influences the choice of $\lambda_B$ through its elementwise $\ell_\infty$ norm.

*Lemma 2 (High probability bound for $\|\mathsf{X}_{T-1}^\top \mathsf{E}/T\|_\infty$):* Consider $\mathsf{X} \in \mathbb{R}^{T\times p}$ and $\mathsf{E} \in \mathbb{R}^{T\times p}$ whose rows are random realizations $\{x_0, \ldots, x_{T-1}\}$ and $\{\epsilon_1, \ldots, \epsilon_T\}$ drawn from the processes in (2). There exist positive constants $c_i\,(i = 0, 1, 2)$ such that for sample size $T \gtrsim \log p$, with probability at least $1 - c_1 \exp(-c_2 \log p)$, the following bound holds:

$$\|\mathsf{X}^\top \mathsf{E}/T\|_\infty \leq c_0 \left(\mathcal{M}(g_X) + \mathcal{M}(g_\epsilon) + \mathcal{M}(g_{X,\widetilde{\epsilon}})\right) \sqrt{\frac{\log p}{T}}. \tag{13}$$

Note that with the definition of the shifted processes $\{\widetilde{\epsilon}_t\}$, we have $g_{X,\widetilde{\epsilon}}(\omega) = e^{-ih\omega} g_{X,\epsilon}(\omega)$, which implies $\mathcal{M}(g_{X,\widetilde{\epsilon}}) = \mathcal{M}(g_{X,\epsilon})$. Hence, the term that measures the upper extreme of the cross-spectrum between $X_t$ and the shifted process in (13) can be replaced by its unshifted counterpart. Moreover, since $g_\epsilon(\omega) = \frac{\sigma_\epsilon}{2\pi}$, its upper extreme is given by $\mathcal{M}(g_\epsilon) = \Lambda_{\max}(\Sigma_\epsilon)/(2\pi) = \sigma_\epsilon^2/(2\pi)$.

The next lemma provides upper bounds for the maximum eigenvalue of the sample covariance matrix $\widehat{\Sigma}_{\mathsf{E}}$.

*Lemma 3 (High probability concentration for $\Lambda_{\max}(\widehat{\Sigma}_{\mathsf{E}})$):* Consider $\mathsf{E} \in \mathbb{R}^{T\times p}$ whose rows are independent realizations of the mean zero Gaussian random vector $\epsilon_t$ with covariance $\Sigma_\epsilon$. Then, for sample size $T \gtrsim p$, with probability at least $1 - \exp(-T/2)$, the following bound holds:

$$\Lambda_{\max}(\widehat{\Sigma}_{\mathsf{E}}) \leq 9\Lambda_{\max}(\Sigma_\epsilon).$$

Proofs for Lemmas 1 to 3 can be found in Appendix B.

Up to this stage, we have verified the RSC condition and obtained the high probability bounds for quantities that are associated with the choice of $(\lambda_B, \lambda_\Theta)$, for random realizations from the underlying processes. Theorem 2 combines the results in Corollary 1 and Lemmas 1 to 3, and provides a high probability error bound of the estimates when the data are random realizations from the underlying processes, as stated next.

*Theorem 2 (High probability error bound for random realizations):* Consider data of length $(T + 1)$ $\{x_0, \ldots, x_T\}$ from the $p$-dimensional observable process $\{X_t\}$, whose dynamics are

described in (2) with $B^\star \in \mathbb{B}_q(R_q)$. Then, there exist universal positive constants $c_i$ $(i = 1, 2, 3)$ and $c_i'$ $(i = 1, 2)$, such that for sample size $T \gtrsim p$ and regularization parameters

$$\lambda_B = c_1 \left(\mathcal{M}(g_X) + \mathcal{M}(g_\epsilon) + \mathcal{M}(g_{X,\epsilon})\right) \sqrt{\tfrac{\log p}{T}} + c_2 \tfrac{\phi}{\sqrt{Tp}},$$

$$\lambda_\Theta = c_3 \Lambda_{\max}^{1/2}(\Sigma_\epsilon),$$

with probability at least $1 - c_1' \exp(-c_2' \log p)$, the solution $(\widehat{B}, \widehat{\Theta})$ of the conex problem (8) has the following estimation error bound for some thresholded level $\eta > 0$:

$$|||\Delta_B|||_{\mathrm{F}}^2 + |||\Delta_\Theta/\sqrt{T}|||_{\mathrm{F}}^2 \leq C_1 \cdot \mathcal{E}_B' + C_2 \cdot \mathcal{E}_\Theta' + C_3 \cdot \mathcal{E}_{\tau_{\mathrm{tol}}}', \tag{14}$$

where

$$\mathcal{E}_B' := \lambda_B^2 \left(s_\eta + \|B^\star_{\mathcal{S}_\eta^c}\|_1\right) = \left(s_\eta + \|B^\star_{\mathcal{S}_\eta^c}\|_1\right) \cdot \mathcal{O}\left(\left[\frac{\log p}{T}\right]\right),$$

$$\mathcal{E}_\Theta' := \lambda_\Theta^2 K = \mathcal{O}(K),$$

$$\mathcal{E}_{\tau_{\mathrm{tol}}}' := \tau_{\mathrm{tol}} \left(\|B^\star_{\mathcal{S}_\eta^c}\|_1^2\right) = \left(\|B^\star_{\mathcal{S}_\eta^c}\|_1^2\right) \cdot \mathcal{O}\left(\left[\frac{\log p}{T}\right]\right).$$

$C_i$ $(i = 1, 2, 3)$ are positive constants that depend on the upper and lower extremes of the underlying processes, but are independent of $T$ and $p$.

Note that Theorem 2 requires that $T \gtrsim p$ for the corresponding quantities to properly concentrate, which in turn leads to the estimation errors for $\Delta_B$ and $\Delta_\Theta$ being jointly bounded. This sample size requirement is of the same order as in classical factor analysis[1] literature (e.g., [10]) for independent and identically distributed (iid) data, and also shows up for the problem of recovering a low-rank component based on noisy data in high-dimensional statistics (e.g., [2]).

Corollary 2 provides the order of each term in the error bound as a function of $T, p, K$ and $R_q$, when the level of thresholding is set at some prespecified level.

*Corollary 2:* Under the same set of conditions as in Theorem 2, by choosing the level of thresholding as $\kappa\lambda_B$ with $\kappa := \max\{\mathfrak{m}^{-1}(g_X), \pi\}$, the following holds for $\mathcal{E}_B'$, $\mathcal{E}_\Theta'$ and $\mathcal{E}_{\tau_{\mathrm{tol}}}'$:

$$\mathcal{E}_B' = \mathcal{O}\left(\left[\frac{\log p}{T}\right]^{1-q/2} R_q\right),$$

$$\mathcal{E}_\Theta' = \mathcal{O}(K) = \mathcal{O}(1),$$

$$\mathcal{E}_{\tau_{\mathrm{tol}}}' = \mathcal{O}\left(\left[\frac{\log p}{T}\right]^{2-q} R_q^2\right).$$

As a consequence, the following error bound holds for $\widehat{B}$ and $\widehat{\Theta}$ with probability at least $1 - c_1' \exp(-c_2' \log p)$:

$$|||\Delta_B|||_{\mathrm{F}}^2 + |||\Delta_\Theta/\sqrt{T}|||_{\mathrm{F}}^2 \lesssim \mathcal{O}(1). \tag{15}$$

We further discuss the bounds obtained in (14) and (15) next.

[1]In classical factor analysis, for both the factors and its loadings to be consistently estimated, both $\sqrt{p}/T \to 0$ and $\sqrt{T}/p \to 0$ are required to hold simultaneously.

*Remark 3:* The bound in (14) has a similar form to the one in Section 3.4 of [2] for the case where an $\ell_1$ norm regularizer is considered; in their setting, the coefficient matrix of the linear model for iid sub-Gaussian data can be decomposed into a sparse and a low rank component. When the two components are exactly sparse/low-rank, under certain regularity conditions, the error bound in [2] takes the form $\lambda^2\, s + \mu^2\, K$ with appropriately chosen tuning parameters $\lambda$ and $\mu$. In particular, for random realizations, one gets

$$\lambda \gtrsim |||\mathsf{X}'\mathsf{E}/T|||_\infty + \frac{\phi}{\sqrt{p \cdot p}} = \mathcal{O}\left(\sqrt{\log p/T}\right),$$

$$\mu \gtrsim |||\mathsf{X}'\mathsf{E}/T|||_{op} = \mathcal{O}\left(\sqrt{p/T}\right); \tag{16}$$

the second term in (16) is due to the box constraint imposed on the low rank component (with $\phi$ being the compatibility constant) to encourage its identifiability w.r.t. the sparse one. Consequently, the error bound becomes $s \cdot \mathcal{O}(\log p/T) + K \cdot \mathcal{O}(p/T)$ and vanishes for $p = o(T)$. However, it is worth highlighting that in the setting of [2], the two components share the same *observed* bases (i.e., coordinates of the $X_t$'s) that span the space of the predictors; moreover, the bases are assumed uncorrelated with the noise $\epsilon_t$ at the population level, which leads to a vanishing bound, as long as $(\mathsf{X}'\mathsf{E}/T)$ properly concentrates. In contrast, the nature of the non-vanishing upper bound provided in (15) is a consequence of the fact that for an LDS, $F_t$ is latent, and therefore there is no observed basis for the low rank component in the formulated optimization problem, which renders the latent state hyperplane $\Theta$ and the error term being not distinguishable. Mathematically, this is manifested through the choice of $\lambda_\Theta$ that needs to dominate $\Lambda_{\max}^{1/2}(\widehat{\Sigma}_E)$ (see (11)); for random realizations, this quantity *concentrates*, but does not vanish. Note that the structure of the underlying optimization problem is similar to that of the *noisy matrix completion* problem, for which the same phenomenon of a non-vanishing error bound occurs [17], due to the violation of the Restricted Isometry Property. Further details on model identifiability are given in Supplement-III.

### C. Convergence Analysis of Algorithm 1

The convergence property of Algorithm 1 can be established using familiar arguments and exploiting its convex nature. Specifically, the objective function is given by

$$f(B, \Theta) := \ell_T(\mathsf{X}; B, \Theta) + \lambda_B \|B\|_1 + \lambda_\Theta |||\Theta/\sqrt{T}|||_*$$

and is *jointly convex* in $(B, \Theta)$, with a convex feasible region $\mathbb{B}_\infty(\phi, \mathsf{X}_{T-1})$. Thus, it directly follows from [50] that the alternating minimization that generates the sequence $\{(\widehat{B}^{(k)}, \widehat{\Theta}^{(k)})\}$ converges to a stationary point which is also a global optimum, though the global optimum is not necessarily unique.

To conclude this section, we remark that the theoretical formulation in (8) can be solved in an analogous way to Algorithm 1. Specifically, the update of $\Theta$ requires modification to satisfy the constraint on the feasible region of $\Theta$, and the partial minimization can be solved by employing the composite gradient descent algorithm of [42] that involves singular value thresholding steps.

**Algorithm 3:** Empirical Implementation for Obtaining $\widehat{B}_{\text{emp}}$ and $\widehat{\Theta}_{\text{emp}}$ Through Alternate Minimization.

**Input:** Time series data $\{x_t\}_{t=0}^{T}$, tuning parameter $\lambda_B$, rank constraint $r$.

1 **Initialization:** Initialize with $\bar{B}^{(0)} = 0$ and $\bar{\Theta}^{(0)} = \text{SVT}(\mathsf{X}_T)$;

2 **Iterate until convergence:**

3 Update $\bar{B}^{(m)}$ with the plug-in $\bar{\Theta}^{(m-1)}$ so that each row $j$ is obtained with Lasso regression (in parallel) and solves

$$\bar{B}_{j\cdot}^{(m)} = \arg\min_{\beta}\left\{\frac{1}{2T}\|\left[\mathsf{X}_T - \bar{\Theta}^{(m-1)}\right]_{\cdot j} - \mathsf{X}_{T-1}\beta\|^2 + \lambda_B\|\beta\|_1\right\}.$$

4 Update $\bar{\Theta}^{(m)}$ by singular value thresholding (SVT): do SVD on the lagged value-adjusted hyperplane, i.e., let

$$UDV^\top := \mathsf{X}_T - \mathsf{X}_{T-1}\bar{B}^{(m)},$$

where $D := \text{diag}(d_1, \ldots, d_{\min(T,p)})$, and construct $\bar{\Theta}^{(m)}$ by

$$\bar{\Theta}^{(m)} = UD_rV, \quad \text{where } D_r := \text{diag}\,(d_1, \ldots, d_r, 0, \ldots, 0)\,.$$

**Output:** Estimated sparse transition matrix $\widehat{B}_{\text{emp}} = \bar{B}^{(\infty)}$ and the low rank hyperplane $\widehat{\Theta}_{\text{emp}} = \bar{\Theta}^{(\infty)}$.

Nevertheless, the modified algorithm is also convergent, as the one in Algorithm 1.

## IV. Implementation and Performance Evaluation

Next, we present results for simulation studies under various settings to demonstrate the performance of the proposed formulation of the LDS model.

*An empirical algorithmic relaxation.* The actual implementation of Algorithm 1 requires $\lambda_B$, $\lambda_\Theta$ as inputs, which in practice are challenging to select. On the other hand, the computation procedure designed for solving the convex program in (6) suggests that to obtain the estimates boils down to alternating between the following two steps: (1) a regularized regression (lasso) update on the rows of $B$; and (2) an SVT update on $\Theta$. This naturally motivates the following steps in the implemented version of the algorithm, outlined next in Algorithm 3.

Algorithm 3 outlines the algorithmic relaxation to obtaining $(\widehat{B}, \widehat{\Theta})$ in (6), and it can be viewed as an alternating minimization algorithm that solves

$$\min_{B,\Theta}\left\{\tfrac{1}{2T}\|\mathsf{X}_T - \Theta - \mathsf{X}_{T-1}B^\top\|_{\text{F}}^2 + \lambda_B\|B\|_1\right\},$$
$$\text{subject to } \text{rank}(\Theta) \le r. \tag{17}$$

For each update, the partial minimization step with respect to $\Theta$ or $B$ ensures that the value of the objective function is always non-ascending, which together with the fact that the objective function is bounded below guarantees convergence of the objective function iterates. In practice, the algorithm is terminated when the descent magnitude of the objective function between successive iterations is smaller than some pre-specified tolerance level. This algorithm does not provide guarantees of convergence to a stationary point of the sequence of $(\bar{\Theta}^{(k)}, \bar{B}^{(k)})$ iterates, which requires stronger assumptions — either the convexity of the objective function and the constraint region, or the uniform compactness of the generated sequence of iterates.

*Choice of the tuning parameter $\lambda_B$ and the rank constraint $r$.* The implementation of Algorithm 3 requires a specific pair of $(\lambda_B, r)$ as input. We consider choosing the optimal pair of $(\lambda_B, r)$ based on the information criterion proposed in [4], called the Panel Information Criterion (PIC) and defined as:

$$\text{PIC}(\lambda_B, r) := \tfrac{1}{Tp}\|\mathsf{X}_T - \widehat{\Theta}_{\text{emp}} - \mathsf{X}_{T-1}\widehat{B}_{\text{emp}}^\top\|_{\text{F}}^2 + \widehat{\sigma}^2\left[\tfrac{\log T}{T}\|\widehat{B}_{\text{emp}}\|_0 + r(\tfrac{T+p}{Tp})\log(Tp)\right], \tag{18}$$

where $\widehat{\sigma}^2 = \frac{1}{Tp}\|\mathsf{X}_T - \widehat{\Theta}_{\text{emp}} - \mathsf{X}_{T-1}\widehat{B}_{\text{emp}}^\top\|_{\text{F}}^2$ and $(\widehat{B}_{\text{emp}}, \widehat{\Theta}_{\text{emp}})$ are solutions to (17) with the plug-in $(\lambda_B, r)$ pair. The optimal pair $(\lambda_B, r)$ is then selected in two steps: in step 1, we obtain $(\lambda_B^0, r^0)$ that gives the smallest PIC over a lattice $\mathcal{G}_{\lambda_B} \times \mathcal{G}_r := \{\lambda_B^{(1)}, \ldots, \lambda_B^{(j_1)}\} \times \{r^{(1)}, \ldots, r^{(j_2)}\}$; in step 2, we fix $r$ at $(d+1) \times r^0$ where $d$ is the number of lags corresponding to the sparse VAR($d$) model, and seek for $\lambda_B^{\text{opt}}$ over a grid that minimizes $\text{PIC}(\lambda_B, (d+1)r^0)$. The optimal pair of tuning parameters is then given by $(\lambda_B^{\text{opt}}, r^{\text{opt}}) := (\lambda_B^{\text{opt}}, (d+1) \times r^0)$, with $r^{\text{opt}}$ being the effective dimensionality of the states.

*Data generating mechanism.* Synthetic data are generated according to the model representation in 2. Starting from the standard state-space model representation $X_t = \widetilde{\Lambda} f_t + u_t$, $u_t$ is serially correlated and follows a weakly sparse VAR($d$) model;[2] at each timestamp $t$, the $K$-dimensional state $f_t$ is generated according to a VAR($w$) model $f_t = \Phi_1 f_{t-1} + \cdots + \Phi_w f_{t-w} + v_t$ where $v_t \sim \mathcal{N}(0, \sigma_v^2 \text{I})$; decorrelating $u_t$ leads to the following dynamic of $X_t$:

$$X_t = \widetilde{\Lambda} f_t - B\widetilde{\Lambda} f_{t-1} + BX_{t-1} + \epsilon_t =: \Lambda F_t + BX_{t-1} + \epsilon_t,$$

where $\Lambda = [\widetilde{\Lambda}, B\widetilde{\Lambda}]$ and $F_t = (f_t^\top, f_{t-1}^\top)^\top \in \mathbb{R}^{2K}$.

We consider several simulation settings as listed in Table I to test various facets of the model, primarily encompassing the dimensionality of the system $p$ and the states, as well as the sparsity structure of $B$ and its spectral radius that captures the level of autocorrelation. In addition to settings S0 to S4 where $\epsilon_t$ is Gaussian, to test the robustness of the proposed model in the presence of heavy tails, we consider also cases where $\epsilon_t$ follows some multivariate $t$ distribution (S5, S6). Throughout all numerical experiments presented in this section, the sample size $T$ is fixed at 200 and the spectral radius of the VAR($w$) system is set randomly from $\mathsf{Unif}[0.6, 0.8]$.

[2]Throughout this section, we assume $d = 1$; additional consideration for $d > 1$ have been deferred to Supplement-I.

TABLE I
SIMULATION SETTINGS FOR DATA GENERATED ACCORDING TO A LAG-ADJUSTED STATE-SPACE REPRESENTATION

| | signal dim $p$ | sparsity struct. of $B$ | spec. radius $\varrho(B)$ | dim of $F_t$ $K$ | VAR lag # for $f_t$ | structure of $\Sigma_\epsilon$ | state/lag relative strength |
|---|---|---|---|---|---|---|---|
| S0 | 100 | $2/p$, exact | 0.7 | 2 | 1 | diagonal - $\mathcal{N}$ | strong state $\approx 3:2$ |
| S1 | 100 | $5/p$, weak | 0.7 | 2 | 1 | Toeplitz(0.2) - $\mathcal{N}$ | strong state $\approx 2:1$ |
| S2 | 300 | $2/p$, weak | 0.7 | 5 | 1 | diagonal - $\mathcal{N}$ | strong state $\approx 2:1$ |
| S3 | 200 | $2/p$, exact | 0.9 | 5 | 2 | diagonal - $\mathcal{N}$ | strong lag $\approx 2:3$ |
| S4 | 200 | $2/p$, weak | 0.7 | 5 | 4 | Toeplitz(0.2) - $\mathcal{N}$ | strong state $\approx 3:2$ |
| S5 | 100 | $2/p$, exact | 0.7 | 5 | 1 | diagonal - $t_4$ | strong state $\approx 3:2$ |
| S6 | 200 | $2/p$, weak | 0.7 | 5 | 1 | Toeplitz(0.2) - $t_8$ | strong state $\approx 1:1$ |

To generate the sparse transition matrix $B$, for each row that corresponds to the coefficients of each single time series regression, its (strong) support set is randomly generated to meet the specified density level requirement (i.e., $2/p$ or $5/p$); nonzero entries are then generated from $\pm\mathsf{Unif}[m_B - 0.1, m_B + 0.1]$ for some $m_B > 0$ that dictates the magnitude. In the case of a weakly sparse $B$, entries in the weak support set are generated from $\mathsf{Unif}([-10\%m_B, 10\%m_B])$ and thus having smaller magnitude compared with those in the strong one. Finally, all entries are scaled to the specified $\varrho(B)$ level is satisfied to ensure that the system is stationary. For the dense loading matrix $\Lambda$, its entries are generated from $\pm\mathsf{Unif}[m_\Lambda - 0.1, m_\Lambda + 0.1]$ for some $m_\Lambda > 0$. It is worth noting that the values of $m_B$ and $m_\Lambda$ are set so that the state/lag relative strength is satisfied, measured by the empirical relative signal-to-noise ratio for the $\Lambda F_t$ and the $BX_{t-1}$ component.

For comparison purposes, we also present the performance evaluation of the common space recovery (to be defined later) and the prediction accuracy of the signals in the case where components are identified via standard Principal Components analysis (e.g., [48]) with the lag-adjusting term ignored. Note that for most settings, as the dimensionality of the signal exceeds the number of available data points, identification through a Kalman-filter based EM algorithm is not necessarily feasible and thus its performance is not considered in the comparisons.

*Performance evaluation.* To measure the accuracy of the parameter estimates and signal prediction, we focus on the following four components of the model:

- For the sparse transition matrix $B$ we use sensitivity $\mathrm{SEN} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$, specificity $\mathrm{SPC} = \frac{\mathrm{TN}}{\mathrm{FP+TN}}$ and relative error in Frobenius norm ($\mathrm{RErr}_B$) as evaluation criteria; in the case where $B$ is weakly sparse, despite the fact that entries in the weak support set are not exactly zero, they are effectively deemed as zeros for comparison purpose.
- For the state hyperplane $\Theta$, we measure its relative error in Frobenius norm ($\mathrm{RErr}_\Theta$) and its relative *projection error*, defined as $\mathrm{ProjErr}_\Theta := |||\Pi_{\widehat{\Theta}} - \Pi_{\Theta^\star}|||_{\mathrm{F}} / |||\Pi_{\Theta^\star}|||_{\mathrm{F}}$, where $\Pi_{\Theta^\star} := Q_{\Theta^\star} Q_{\Theta^\star}^\top$ with $Q_{\Theta^\star}$ being the orthonormal basis of $\Theta^\star$; $\Pi_{\widehat{\Theta}}$ can be analogously defined. Note that the following correspondence between $\sin\theta$ distance and the projection error holds: $|||\sin\theta(\widehat{\Theta}, \Theta^\star)|||_{\mathrm{F}}^2 = \frac{1}{2}|||\Pi_{\widehat{\Theta}} - \Pi_{\Theta^\star}|||_{\mathrm{F}}$; moreover, this metric is not applicable in high-dimensional regimes ($p \geq T$) where it would remain at zero.
- For the common space, in the case where it is estimated with the proposed lag-adjusted state-space representation, at the population level it is captured by $BX_{t-1} + \Lambda F_t$ and hence its estimate is given by $\widehat{\Theta} + \mathsf{X}_{T-1}\widehat{B}^\top$; whereas in the case where the model is estimated based on classical PC analysis, the estimated state hyperplane coincides with that of the common space. For this quantity, we present the relative error in Frobenius norm of the estimates.
- For the one-step-ahead prediction, we measure its squared $\ell_2$ norm w.r.t. the oracle $x_{T+1}^\star$, that is, $\|\widehat{x}_{T+1} - x_T^\star\|^2 / \|x_T^\star\|^2$, where the oracle is given by $x_{T+1}^\star = Bx_T + \Lambda F_{T+1}$ and can be viewed as the "denoised" version of $x_{T+1}$.

As Table II demonstrates, for all four components, estimates obtained from Algorithm 3 exhibit good performance. In particular, (i) the proposed method is robust to the sparsity structure of $B$, as both exactly-sparse and weakly-sparse settings yield very satisfactory strong support recovery (see S1, S2 and S4). (ii) A larger panel size $p$ leads to improved state hyperplane recovery, as manifested in the form of smaller relative error in its magnitude estimation, although it requires the sparsity of the transition matrix to decrease accordingly (recall that it is set to $2/p$); however, the performance deteriorates as the dynamics of $f_t$ become more complex (e.g., S4). (iii) A strong signal in the lag-space leads to improved recovery of $B$, despite the presence of stronger temporal dependence which empirically incurs the algorithm to take more iterations to converge (e.g., S3). For all settings, PIC correctly determines the dimension of the states, which translates into the correct identification of the rank constraint.

Additionally, the proposed model is relative robust to the presence of heavy tails, although the performance deteriorates compared to the Gaussian case. Specifically, when the distribution shows significant deviation from Gaussian (e.g., S5), the degradation manifests itself through less satisfactory recovery in the support of $B$ and larger error of the estimated factor space; whereas the forecasting performance isn't affected. On the other hand, with lighter tails (e.g., S6), the performance becomes comparable to the Gaussian case.

Finally, in most settings, the proposed method outperforms the standard PC analysis in both the common space recovery and the one-step-ahead forecasting.

## V. APPLICATION TO RETURNS OF US FINANCIAL ASSETS

State-space models have been widely used in financial applications (e.g., [24]). In particular, they have been employed in analyzing the dynamics of asset returns, either for the purpose of identifying risk factors (i.e., the latent states), or for estimating the covariance structure amongst assets for better portfolio diversification and asset allocation (e.g., [25]). We

TABLE II
PERFORMANCE EVALUATION FOR VARIOUS SIMULATION SETTINGS, MEDIAN ACROSS 100 REPLICATIONS

| | $B$ recovery (lag-adj state-space) | | | $\Theta$ recovery (lag-adj state-space) | | | common space recovery | | one-step-ahead prediction | |
|---|---|---|---|---|---|---|---|---|---|---|
| | SEN | SPC | $\mathrm{RErr}_B$ | $\widehat{K}$ | $\mathrm{ProjErr}_\Theta$ | $\mathrm{RErr}_\Theta$ | lag-adj state-space | PC analysis | lag-adj state-space | PC analysis |
| S0 | 0.99 | 0.98 | 0.28 | 2 | 0.15 | 0.20 | 0.13 | 0.32 | 0.51 | 0.60 |
| S1 | 0.97 | 0.92 | 0.51 | 2 | 0.16 | 0.47 | 0.27 | 0.47 | 0.56 | 0.91 |
| S2 | 0.99 | 0.95 | 0.74 | 5 | – | 0.58 | 0.35 | 0.45 | 0.60 | 0.73 |
| S3 | 0.99 | 0.98 | 0.19 | 5 | – | 0.26 | 0.22 | 0.72 | 0.36 | 0.92 |
| S4 | 0.98 | 0.97 | 0.58 | 5 | – | 0.51 | 0.32 | 0.44 | 0.47 | 0.58 |
| S5 | 0.92 | 0.92 | 0.61 | 5 | 0.31 | 0.48 | 0.10 | 0.13 | 0.43 | 0.42 |
| S6 | 0.98 | 0.93 | 0.47 | 5 | – | 0.53 | 0.35 | 0.63 | 0.55 | 0.90 |

apply the proposed modeling framework to a set of stock returns corresponding to 75 large US financial institutions, spanning the period of 2001–2017. This time period contains a number of significant events for the financial industry, including the growth of mortgage bank securities [6] in the early 2000 s, rapid changes in monetary policy in 2004–2006, the great financial crisis [22] in 2008–2009 and the European debt crisis in 2011–2012. Our analysis identifies a number of interesting patterns, especially around the period 2007–2009 encompassing the beginning, height and immediate aftermath of the US financial crisis, both through changes in the structure of the latent states and that of partial autocorrelations governed by the VAR model transition matrix $B$ of the log-returns of these financial assets.

*Data.* The data consist of weekly stock risk-free returns[3] corresponding to 75 large financial institutions in terms of market capitalization, for the period of January 2001 to December 2017 and were obtained from the Center for Research in Security Prices (CRSP) database. The 75 companies are categorized into three sectors: banks (SIC code 6000–6199), broker/dealers (SIC code 6200–6299) and insurance companies (SIC code 6300–6499), with 25 in each sector (see also [15]). As we require that the data be available for the entire time span under consideration, 56 firms are kept for further analysis, since the remaining ones either went bankrupt or were forced to merge with financially healthier companies (e.g. Lehman Brothers and Merill Lynch in 2008, respectively). Additionally, based on previous analysis (c.f., [15], [35]) the time period is broken into the following sub-periods: 2001–2006 (pre-crisis), 2007–2009 (crisis), 2010–2017 (post-crisis).

The analysis is based on 104-week-long rolling windows to avoid issues with possible non-stationarity of the data, a commonly used strategy in the literature ([15], [35]), that allows monitoring changes in the dimension of the state over time, as well as the sparsity level of $B$ which measures the connectivity of the partial autocorrelation network across these financial institutions. We fit the proposed lag-adjusted LDS in each time window, with tuning parameters selected according to the PIC criterion (see Section IV). As Fig. 1 shows, sharp changes are observed in the temporal dependence structure of stock returns—two change points respectively correspond to the beginning of the 2007 sub-prime mortgage crisis and the ending of the 2008–2009 global financial crisis. Specifically, for the pre- and post-crisis periods, the density of the transition

[3]The risk-free return of Stock $i$ at time $t$ is calculated as $\widetilde{r}_{i,t} = r_{i,t} - r_{\mathrm{rf},t} = \frac{p_{i,t}-p_{i,(t-1)}}{p_{i,(t-1)}} - r_{\mathrm{rf},t}$, where $p_{i,t}$ is its stock price at time $t$ and $r_{\mathrm{rf},t}$ is the risk-free rate.

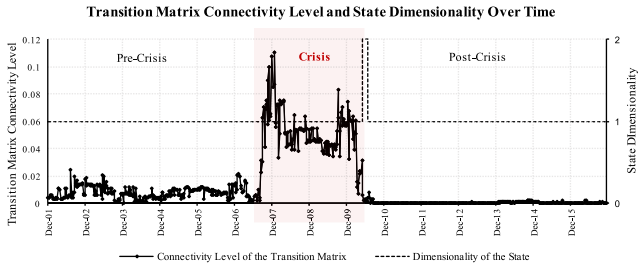


Fig. 1. Results after fitting the model to the real data based on 104-week-long rolling windows over time. Left axis: the connectivity level of $\widehat{B}$; right axis: intrinsic dimensionality of the states. Time stamps on the horizontal axis correspond to the mid-point of the window in question.

matrix $B$ stays at a level close to zero, suggesting that little serial correlation exists in the observation noise component after the effect of the latent state (that captures the overall market direction [7]) is accounted for. During the crisis period, however, the connectivity level of $\widehat{B}$ witnesses a sharp increase, with the maximum corresponds to the sampling window from Dec 2006 to Dec 2008, during which multiple major events of the financial crisis occurred. Note that with data at the weekly frequency and rolling samples of size 104, the dimensionality of the latent state is identified as one for almost all times, corresponding to a "market factor" with all financial institutions contributing positively to it. Of note, for a short period centered at June-July 2010, the dimension of the estimated latent state becomes two. The key contributor to the second state variable is AIG, with some minor contribution stemming from other insurance companies (e.g., United Health Group, Manulife Financial Corp). Given the timing of the emergence of the second latent state and its primary contributors, a possible explanation relates to the active market activities of AIG. Specifically, the recapitalization deal among AIG, the Federal Reserve and the Treasury was launched in Sept 2010 and closed in Jan 2011; a number of transactions have taken place pre/during/post this time frame from 2010 to 2011, in which AIG sold its subsidiaries or conducted an IPO in the overseas market to raise capital. Additionally, in May 2011, the Treasury initiated its first AIG stock sale to reduce its stake.

To further investigate the composition of the latent state and the temporal dependence structure amongst the observed signals during the financial crisis period, we further focus on the 2008 data. Specifically, we consider *daily* data from January 2008 to December 2008 that cover 253 consecutive trading days and fit the proposed lag-adjusted LDS. Note that for this part of the analysis, the sample consists of 72 stocks. Using PIC, a two-dimensional latent state is identified, with the first component

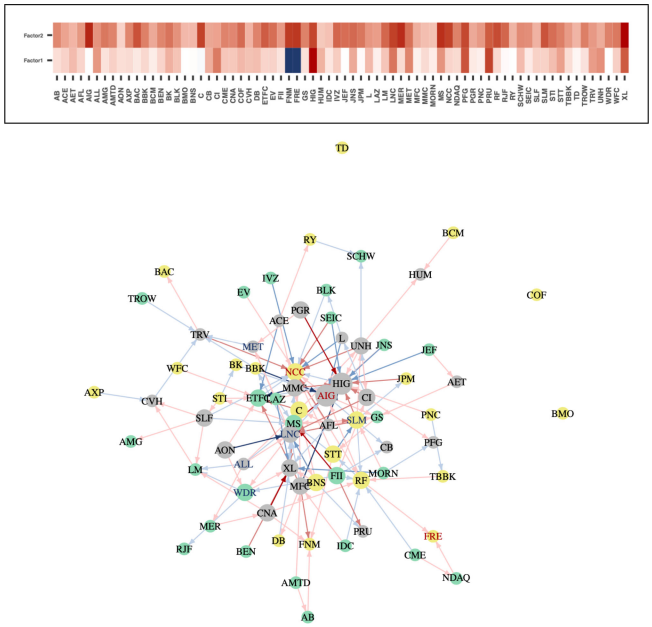


Fig. 2. Top panel: factor composition. Bottom panel: partial autocorrelation network during the crisis, after proper thresholding of entries with small magnitudes. Top 5 emitters: MFC, MMC, MS, AON, PGR. Top receivers (in white bars): HIG, NCC, LNC, XL, AIG. Node colors indicate their respective sector (gray–INS, yellow-BA, green–PB).

capturing 55% of the R-squared statistic, followed by 11% for the second component. For reconstruction purposes, we assume they are orthogonal so that their respective composition can be directly retrieved from the singular vectors of $\widehat{\Theta}$, as shown in Section II-A.

As depicted in the left panel of Fig. 2, all financial institutions contribute positively to the first component, with dominating contributors across all three groups (banks, insurance companies and brokers/dealers). The composition of the second component exhibits an interesting pattern: two negative contributors are FRE (Freddie Mac) and FNM (Fannie Mae), while the positive ones are primarily from the insurance group. However, AIG—unlike its peers—shows almost zero contribution to the second component, albeit its strong contribution to the first one. The latter is consistent with other findings in the literature that AIG played a prominent role during the crisis; see [30] and references therein..[4] In the lower panel of Fig. 2, we plot the partial autocorrelation network of the firms during the crisis after properly thresholding the entries that have small magnitudes, with red edges denoting positive links and with blue negative ones. Nodes that belong to the same group are colored identically. A careful examination of the node weighted in/out-degrees shows that the top emitters are relatively uniform, in the sense that their weighted out-degrees do not differ by much; whereas the top receivers are dominating, since the weighted in-degrees for top receivers are significantly higher compared with the rest. Further, an examination of the individual names show that top emitters heavily concentrate in the insurance sector, and some of the top receivers are also major contributors to the states' composition, e.g., AIG to the 1st component, HIG to the 2nd, etc. This finding partially aligns with the role that many insurance companies played in magnifying the impact of the crisis on the overall stability of the financial system, due to their large insurance underwriting of Credit Default Swaps and subsequent exposure to accentuated risks [22]. However, this analysis points to the importance of insurance companies based on publicly available data and before their role in the crisis was fully revealed and understood. It is worth noting that with the same set of data, classical factor analysis using the information criterion proposed in [9] only identifies 1 factor, which further substantiates the aforementioned point that classical factor analysis may lead to skewed inference when strong correlation amongst the coordinates of the observation noise is present.

[4]According to an estimate as of January 2010, AIG accounted for 38% of the total losses incurred by insurance companies ($98.2 bn out of $261 bn) since 2007. Source: Bloomberg, see also [44]

## VI. Discussion

In this paper, we introduced a novel modeling framework that generalizes the classical state-space model on LDS, to accommodate large scale panels of observed signals/time series in the presence of strong cross-correlations across components of the observation noise process. This is accomplished by including lags of the observed signals and further assuming that the autoregressive structure is sparse for identification purposes. Note that the transition matrix of the autoregressive structure provides useful and interpretable information, as shown in our application study and also noted in [23], [37]. The LDS parameters are estimated through penalized least squares. Specifically, based on the proposed algorithm, the estimators possess finite-sample high probability error bounds that can be expressed in terms of the key structural parameters ($T, p, K$ and sparsity), and they exhibit superior empirical performance in synthetic data.

As mentioned in the introductory section, the LDS modeling framework assumes that the dynamics of a large panel of signals are driven by a low-dimensional state variable; given the latent nature of the state variables, they are estimated through the "compression" of the observed ones which is a low-rank component relative to the system. Another class of models for multivariate signals assumes that the temporal evolution of an observed signal is influenced by others within the system (and possibly external sources), where the dependency amongst the coordinates can be modeled explicitly and thus the contemporaneous relationship among them can be examined directly. For example, the posited model in [12] enables one to identify the few "neighbors" that impact the value of a signal, as well as the magnitude of such influence. Further, the online version allows for the data to be collected in a sequential manner, and the coefficient matrix that encodes the inter-relationship across signals can slowly vary. Note that at the modeling class level, this is similar to a VAR-X model (e.g., see [35] and references therein) in that both models aim to express explicitly the dependency of one signal on the others within the system, and also allow for impact from external sources, although the former focuses on the contemporaneous impact, whereas the VAR-X model focuses on lead-lag relationships.

## APPENDIX A
## PROOFS FOR STATISTICAL ERROR BOUNDS

Before presenting the proof of Theorem 1, we first define key quantities associated with the regularizers. Given some $\eta$ (to be specified later), let $\mathcal{S}_\eta^\star$ denote the thresholded support set of $B^\star$, i.e., $\mathcal{S}_\eta^\star := \{(i,j) : |B_{ij}^\star| > \eta\}$, and let the SVD of $\Theta^\star$ be $\Theta^\star = (U^\star)D^\star(V^\star)^\top$, with $U_K^\star$ and $V_K^\star$ respectively denoting the first $K$ columns of $U^\star$ and $V^\star$. Let $\mathbb{S}$, $\mathbb{M}$ and their complements respectively be defined as follows:

$$\mathbb{S} := \left\{\Delta \in \mathbb{R}^{p\times p} \,|\, \Delta_{ij} = 0 \text{ for } (i,j) \notin \mathcal{S}_\eta^\star\right\},$$
$$\mathbb{S}^c := \left\{\Delta \in \mathbb{R}^{p\times p} \,|\, \Delta_{ij} = 0 \text{ for } (i,j) \in \mathcal{S}_\eta^\star\right\},$$

and

$$\mathbb{M} := \left\{\Delta \in \mathbb{R}^{T\times p} \,|\, \text{row}(\Delta) \subseteq V_K^\star \text{ and } \text{col}(\Delta) \subseteq U_K^\star\right\},$$
$$\mathbb{M}^\perp := \left\{\Delta \in \mathbb{R}^{T\times p} \,|\, \text{row}(\Delta) \perp V_K^\star \text{ and } \text{col}(\Delta) \perp U_K^\star\right\}.$$

Further, for some generic matrix $\Delta_1 \in \mathbb{R}^{p\times p}$, we define its projection on $\mathbb{S}$ and $\mathbb{S}^c$ (denoted by $\Delta_{1|\mathbb{S}}$ and $\Delta_{1|\mathbb{S}^c}$, resp.) as

$$\Delta_{1|\mathbb{S},ij} := \mathbf{1}\left\{(i,j) \in \mathcal{S}_\eta^\star\right\}\Delta_{1,ij} \text{ and}$$
$$\Delta_{1|\mathbb{S}^c,ij} := \mathbf{1}\left\{(i,j) \notin \mathcal{S}_\eta^\star\right\}\Delta_{1,ij}. \tag{19}$$

With the above definitions and projections, $\forall\, \Delta_1 \in \mathbb{R}^{p\times p}$, we can write

$$\Delta_1 = \Delta_{1|\mathbb{S}} + \Delta_{1|\mathbb{S}^c},\ \ \|\Delta_1\|_1 = \|\Delta_{1|\mathbb{S}}\|_1 + \|\Delta_{1|\mathbb{S}^c}\|_1, \tag{20}$$

and note that the following inequality holds:

$$\|\Delta_{1|\mathbb{S}}\|_1 \le \sqrt{s}\|\Delta_{1|\mathbb{S}}\|_{\text{F}} \le \sqrt{s}\|\Delta_1\|_{\text{F}}, \tag{21}$$

as $\Delta_{1|\mathbb{S}}$ has at most $s$ nonzero entries where $s := |\mathcal{S}_\eta^\star|$. In an analogous way, for some generic matrix $\Delta_2 \in \mathbb{R}^{T\times p}$, its projections on $\mathbb{M}$ and $\mathbb{M}^\perp$ (denoted by $\Delta_{2|\mathbb{M}}$ and $\Delta_{2|\mathbb{M}^\perp}$, resp.) are defined as

$$\Delta_{2|\mathbb{M}} := U^\star \begin{bmatrix} \widetilde{\Delta}_{2,11} & \widetilde{\Delta}_{2,12} \\ \widetilde{\Delta}_{2,21} & O \end{bmatrix} (V^\star)^\top \quad \text{and}$$
$$\Delta_{2|\mathbb{M}^\perp} := U^\star \begin{bmatrix} O & O \\ O & \widetilde{\Delta}_{2,22} \end{bmatrix} (V^\star)^\top, \tag{22}$$

where $\widetilde{\Delta}_2$ is given as below and partitioned as:

$$\widetilde{\Delta}_2 = (U^\star)^\top\Delta_2(V^\star) = \begin{bmatrix} \widetilde{\Delta}_{2,11} & \widetilde{\Delta}_{2,12} \\ \widetilde{\Delta}_{2,21} & \widetilde{\Delta}_{2,22} \end{bmatrix}, \text{ with } \widetilde{\Delta}_{2,11} \in \mathbb{R}^{K\times K}.$$

Note that the following relationship holds $\forall\, \Delta_2 \in \mathbb{R}^{T\times p}$:

$$\Delta_2 = \Delta_{2|\mathbb{M}} + \Delta_{2|\mathbb{M}^\perp},$$
$$\|\Delta_2\|_* = \|\Delta_{2|\mathbb{M}} + \Delta_{2|\mathbb{M}^\perp}\|_* = \|\Delta_{2|\mathbb{M}}\|_* + \|\Delta_{2|\mathbb{M}^\perp}\|_*. \tag{23}$$

Next, we introduce concepts and lemmas regarding *decomposable regularizers* [41]. Define the weighted regularizer as

$$\mathcal{R}(B,\Theta) := \|B\|_1 + \frac{\lambda_\Theta}{\lambda_B}\|\Theta/\sqrt{T}\|_*,$$

and let $\Delta_B := \widehat{B} - B^\star$ and $\Delta_\Theta := \widehat{\Theta} - \Theta^\star$.

*Lemma 4:* With the definitions of projections in (19) and (22), the following inequality holds:

$$\mathcal{R}(B^\star,\Theta^\star) - \mathcal{R}(\widehat{B},\widehat{\Theta}) \le \mathcal{R}(\Delta_{B|\mathbb{S}},\Delta_{\Theta|\mathbb{M}}) - \mathcal{R}(\Delta_{B|\mathbb{S}^c},\Delta_{\Theta|\mathbb{M}^\perp}) + 2\mathcal{R}(B_{\mathbb{S}^c}^\star,\Theta_{\mathbb{M}^\perp}^\star).$$

*Lemma 5:* With the definition of (22), the following holds for some generic $\Delta \in \mathbb{R}^{T\times p}$:

$$\text{rank}(\Delta_{\mathbb{M}}) \le 2\cdot\text{rank}(\Theta^\star).$$

The proofs of these two lemmas are deferred to Supplement-II. Based on the above preparatory steps, we present next the proof of Theorem 1.

*Proof of Theorem 1:* We prove the bound for $\Delta_B := \widehat{B} - B^\star$ and $\Delta_\Theta := \widehat{\Theta} - \Theta^\star$ under the imposed regularity conditions, where $(\widehat{B},\widehat{\Theta})$ is the solution to the optimization problem (8). Using the optimality of $(\widehat{B},\widehat{\Theta})$ and the feasibility of $(B^\star,\Theta^\star)$, the following *basic inequality* holds:

$$\tfrac{1}{2T}\|\mathsf{X}_{T-1}\Delta_B^\top + \Delta_\Theta\|_{\text{F}}^2 \le \tfrac{1}{T}\left(\langle\langle\Delta_B^\top, \mathsf{X}_{T-1}^\top\mathsf{E}\rangle\rangle + \langle\langle\Delta_\Theta,\mathsf{E}\rangle\rangle\right) + \lambda_B\left(||B^\star||_1 - ||\widehat{B}||_1\right) + \lambda_\Theta\left(\|\Theta^\star/\sqrt{T}\|_* - \|\widehat{\Theta}/\sqrt{T}\|_*\right). \tag{24}$$

The LHS can be equivalently written as

$$\tfrac{1}{2T}\|\mathsf{X}_{T-1}\Delta_B^\top + \Delta_\Theta\|_{\text{F}}^2 = \tfrac{1}{2T}\left(\|\mathsf{X}_{T-1}\Delta_B^\top\|_{\text{F}}^2 + \|\Delta_\Theta\|_{\text{F}}^2 + 2\langle\langle\mathsf{X}_{T-1}\Delta_B^\top, \widehat{\Theta} - \Theta^\star\rangle\rangle\right),$$

and by rearranging, (24) becomes

$$\tfrac{1}{2T}\|\mathsf{X}_{T-1}\Delta_B^\top\|_{\text{F}}^2 + \tfrac{1}{2}\|\Delta_\Theta/\sqrt{T}\|_{\text{F}}^2 \le \tfrac{1}{T}\langle\langle\mathsf{X}_{T-1}\Delta_B^\top, \widehat{\Theta} - \Theta^\star\rangle\rangle + \tfrac{1}{T}\langle\langle\Delta_B^\top, \mathsf{X}_{T-1}^\top\mathsf{E}\rangle\rangle + \tfrac{1}{T}\langle\langle\Delta_\Theta,\mathsf{E}\rangle\rangle + \lambda_B\left(||B^\star||_1 - ||\widehat{B}||_1\right) + \lambda_\Theta\left(\|\Theta^\star/\sqrt{T}\|_* - \|\widehat{\Theta}/\sqrt{T}\|_*\right). \tag{25}$$

Based on (25), the rest of the proof is divided into three parts: in part (i), we provide a lower bound for the LHS primarily using the RSC condition; in part (ii), we provide an upper bound for the RHS with the designated choice of $\lambda_B$ and $\lambda_\Theta$; in part (iii), we align the two sides and obtain the error bound after some rearrangement.

Part i. In this part, we obtain a lower bound for the LHS of (25). Using the RSC condition for $\mathsf{X}_{n-1}$, the following lower bound holds for the LHS of (25):

$$\begin{aligned}&\tfrac{1}{2T}\|\mathsf{X}_{T-1}\Delta_B^\top\|_{\text{F}}^2 + \tfrac{1}{2}\|\Delta_\Theta/\sqrt{T}\|_{\text{F}}^2 \\ &\ge \tfrac{\alpha_{\text{RSC}}}{2}\|\Delta_B\|_{\text{F}}^2 + \tfrac{1}{2}\|\Delta_\Theta/\sqrt{T}\|_{\text{F}}^2 - \tau_{\text{tol}}\|\Delta_B\|_1^2.\end{aligned} \tag{26}$$

To further lower-bound (26), consider an upper bound for $||\Delta_B||_1$ with the aid of (24). By Hölder's inequality, the following inequalities hold for the inner products:

$$\frac{1}{T}\langle\langle\Delta_B^\top, \mathsf{X}_{T-1}^\top\mathsf{E}\rangle\rangle \le \|\Delta_B\|_1\|\mathsf{X}_{T-1}^\top\mathsf{E}/T\|_\infty, \tag{27}$$

and

$$\tfrac{1}{T}\langle\langle\Delta_\Theta,\mathsf{E}\rangle\rangle \le \|\tfrac{\Delta_\Theta}{\sqrt{T}}\|_*\|\tfrac{\mathsf{E}}{\sqrt{T}}\|_{op} = \|\Delta_\Theta/\sqrt{T}\|_*\Lambda_{\max}^{1/2}(\widehat{\Sigma}_{\mathsf{E}}). \tag{28}$$

By choosing $\lambda_B \geq 2\|\mathsf{X}_{T-1}^\top \mathsf{E}/T\|_\infty$ and $\lambda_\Theta \geq \Lambda_{\max}^{1/2}(\widehat{\Sigma}_{\mathsf{E}})$, the following inequality can be derived from the non-negativity of the RHS in (24):

$$\begin{aligned}
0 &\leq \tfrac{\lambda_B}{2}\|\Delta_B\|_1 + \lambda_\Theta \|\!|\Delta_\Theta/\sqrt{T}|\!\|_* + \lambda_B \mathcal{R}(B^\star, \Theta^\star) \\
&\quad - \lambda_B \mathcal{R}(\widehat{B}, \widehat{\Theta}) \\
&\overset{(1)}{\leq} \tfrac{\lambda_B}{2}\|\Delta_{B|\mathbb{S}}\|_1 + \tfrac{\lambda_B}{2}\|\Delta_{B|\mathbb{S}^c}\|_1 + \lambda_\Theta \|\!|\tfrac{\Delta_{\Theta|\mathbb{M}}}{\sqrt{T}}|\!\|_* + \lambda_\Theta \|\!|\tfrac{\Delta_{\Theta|\mathbb{M}^\perp}}{\sqrt{T}}|\!\|_* \\
&\quad + \lambda_B \left(\mathcal{R}(\Delta_{B|\mathbb{S}}, \Delta_{\Theta|\mathbb{M}}) - \mathcal{R}(\Delta_{B|\mathbb{S}^c}, \Delta_{\Theta|\mathbb{M}^\perp}) + 2\mathcal{R}(B^\star_{\mathbb{S}^c}, \Theta^\star_{\mathbb{M}^\perp})\right),
\end{aligned}$$

where the first two terms in (1) come from (20), the next two terms come from (23) and the last three terms use Lemma 4. After writing out $\mathcal{R}(\cdot,\cdot)$ and rearranging, we obtain

$$\begin{aligned}
\tfrac{\lambda_B}{2}\|\Delta_{B|\mathbb{S}^c}\|_1 &\leq \tfrac{3\lambda_B}{2}\|\Delta_{B|\mathbb{S}}\|_1 + 2\lambda_\Theta \|\!|\tfrac{\Delta_{\Theta|\mathbb{M}}}{\sqrt{T}}|\!\|_* \\
&\quad + 2\mathcal{R}(B^\star_{|\mathbb{S}^c}, \Theta^\star_{|\mathbb{M}^\perp});
\end{aligned}$$

adding $\frac{\lambda_B}{2}\|\Delta_{B|\mathbb{S}}\|_1$ to both sides gives

$$\|\Delta_B\|_1 \leq 4\mathcal{R}(\Delta_{B|\mathbb{S}}, \Delta_{\Theta|\mathbb{M}}) + 4\mathcal{R}(B^\star_{|\mathbb{S}^c}, \Theta^\star_{|\mathbb{M}^\perp}). \tag{29}$$

Note that for $\mathcal{R}(\Delta_{B|\mathbb{S}}, \Delta_{\Theta|\mathbb{M}})$, using (21) and Lemma 5,

$$\begin{aligned}
\mathcal{R}(\Delta_{B|\mathbb{S}}, \Delta_{\Theta|\mathbb{M}}) &= \|\Delta_{B|\mathbb{S}}\|_1 + \frac{\lambda_\Theta}{\lambda_B}\|\!|\tfrac{\Delta_{\Theta|\mathbb{M}}}{\sqrt{T}}|\!\|_* \\
&\leq \sqrt{s}\|\!|\Delta_{B|\mathbb{S}}|\!\|_{\mathrm{F}} + \frac{\lambda_\Theta}{\lambda_B}(\sqrt{2\,K})\|\!|\tfrac{\Delta_{\Theta|\mathbb{M}}}{\sqrt{T}}|\!\|_{\mathrm{F}} \\
&\leq \sqrt{s}\|\!|\Delta_B|\!\|_{\mathrm{F}} + \frac{\lambda_\Theta}{\lambda_B}(\sqrt{2\,K})\|\!|\Delta_\Theta/\sqrt{T}|\!\|_{\mathrm{F}}.
\end{aligned} \tag{30}$$

Plug (30) into (29), and by the Cauchy-Schwartz inequality, we have

$$\begin{aligned}
&||\Delta_B||_1^2 \\
&\leq 32\left(s + (2\,K)(\tfrac{\lambda_\Theta}{\lambda_B})^2\right)\left(\|\!|\Delta_B|\!\|_{\mathrm{F}}^2 + \|\!|\tfrac{\Delta_\Theta}{\sqrt{T}}|\!\|_{\mathrm{F}}^2\right) + 32\|B^\star_{\mathbb{S}^c}\|_1^2.
\end{aligned} \tag{31}$$

Combine (26) and (31), a lower bound for the LHS of (25) is given by

$$\begin{aligned}
&\left[\frac{\alpha_{\mathrm{RSC}}}{2} - 32\tau_{\mathrm{tol}}\left(s + (2\,K)(\tfrac{\lambda_\Theta}{\lambda_B})^2\right)\right]\|\!|\Delta_B|\!\|_{\mathrm{F}}^2 \\
&+ \left[\frac{1}{2} - 32\tau_{\mathrm{tol}}\left(s + (2\,K)(\tfrac{\lambda_\Theta}{\lambda_B})^2\right)\right]\|\!|\Delta_\Theta/\sqrt{T}|\!\|_{\mathrm{F}}^2 \\
&- 32\tau_{\mathrm{tol}}\|B^\star_{\mathbb{S}^c}\|_1^2.
\end{aligned}$$

With the designated choice of $\tau_{\mathrm{tol}}$ satisfying $32\tau_{\mathrm{tol}}(s + (2\,K)(\frac{\lambda_\Theta}{\lambda_B})^2) \leq \min\{\alpha_{\mathrm{RSC}}, 1\}/4$, the above bound can be further lower bounded by

$$\frac{\min\{\alpha_{\mathrm{RSC}}, 1\}}{4}\left(\|\!|\Delta_B|\!\|_{\mathrm{F}}^2 + \|\!|\Delta_\Theta/\sqrt{T}|\!\|_{\mathrm{F}}^2\right) - 32\tau_{\mathrm{tol}}\|B^\star_{\mathbb{S}^c}\|_1^2. \tag{32}$$

Part ii. Next, we obtain an upper bound for the RHS of (25). Using the triangle inequality and Hölder's inequality, the first term satisfies

$$\begin{aligned}
&\tfrac{1}{T}|\langle\!\langle \mathsf{X}_{T-1}\Delta_B^\top, \widehat{\Theta} - \Theta^\star\rangle\!\rangle| \\
&\quad \leq \tfrac{1}{T}|\langle\!\langle \Delta_B^\top, \mathsf{X}_{T-1}^\top\widehat{\Theta}\rangle\!\rangle| + \tfrac{1}{T}|\langle\!\langle \Delta_B^\top, \mathsf{X}_{T-1}^\top\Theta^\star\rangle\!\rangle| \\
&\quad \leq \|\Delta_B\|_1\|\mathsf{X}_{T-1}^\top\widehat{\Theta}/T\|_\infty + \|\Delta_B\|_1\|\mathsf{X}_{T-1}^\top\Theta^\star/T\|_\infty \\
&\quad \leq \|\Delta_B\|_1\|\!|\tfrac{\mathsf{X}_{T-1}}{T}|\!\|_1\|\widehat{\Theta}\|_\infty + \|\Delta_B\|_1\|\!|\tfrac{\mathsf{X}_{T-1}}{T}|\!\|_1\|\Theta^\star\|_\infty.
\end{aligned} \tag{33}$$

Using the fact that both $\Theta^\star$ and $\widehat{\Theta}$ are feasible and satisfy the box constraint $\Theta \in \mathbb{B}_\infty(\phi, \mathsf{X}_{T-1})$, the RHS of (33) is upper bounded by $\frac{2\phi}{\sqrt{Tp}} \cdot \|\Delta_B\|_1$; thus, by choosing $\lambda_B \geq 4\phi/\sqrt{Tp}$, we have

$$\frac{1}{T}|\langle\!\langle \mathsf{X}_{T-1}\Delta_B^\top, \widehat{\Theta} - \Theta^\star\rangle\!\rangle| \leq \frac{\lambda_B}{2}||\Delta_B||_1.$$

With (27) and (28), by choosing $\lambda_B \geq 2\|\mathsf{X}_{T-1}^\top\mathsf{E}/T\|_\infty + 4\phi/\sqrt{Tp}$ and $\lambda_\Theta \geq \Lambda_{\max}^{1/2}(\widehat{\Sigma}_{\mathsf{E}})$, the following upper bound holds for the RHS of (25):

$$\begin{aligned}
&\lambda_B\|\Delta_B\|_1 + \lambda_\Theta\|\Delta_\Theta\|_* \\
&+ \lambda_B\left(\mathcal{R}(\Delta_{B|\mathbb{S}}, \Delta_{\Theta|\mathbb{M}}) + 2\mathcal{R}(B^\star_{|\mathbb{S}^c}, \Theta^\star_{|\mathbb{M}^\perp})\right) - \mathcal{R}(\Delta_{B|\mathbb{S}^c}, \Delta_{\Theta|\mathbb{M}^\perp}) \\
&\leq \lambda_B\left(\|\Delta_{B|\mathbb{S}}\|_1 + \|\Delta_{B|\mathbb{S}^c}\|_1\right) + \lambda_\Theta\left(\|\!|\tfrac{\Delta_{\Theta|\mathbb{M}}}{\sqrt{T}}|\!\|_* + \|\!|\tfrac{\Delta_{\Theta|\mathbb{M}^\perp}}{\sqrt{T}}|\!\|_*\right) \\
&+ \lambda_B\left(\mathcal{R}(\Delta_{B|\mathbb{S}}, \Delta_{\Theta|\mathbb{M}}) - \mathcal{R}(\Delta_{B|\mathbb{S}^c}, \Delta_{\Theta|\mathbb{M}^\perp})\right) \\
&+ 2\lambda_B\mathcal{R}(B^\star_{\mathbb{S}^c}, \Theta^\star_{\mathbb{M}^\perp})
\end{aligned}$$

where the inequality uses (20) and (23). By writing out $\mathcal{R}(\cdot,\cdot)$ and canceling terms, the right hand side is further upper bounded by $2\lambda_B\|\Delta_{B|\mathbb{S}}\|_1 + 2\lambda_\Theta\|\!|\Delta_{\Theta|\mathbb{M}}/\sqrt{T}|\!\|_* + 2\lambda_B\mathcal{R}(B^\star_{\mathbb{S}^c}, \Theta^\star_{\mathbb{M}^\perp})$, that is,

$$2\lambda_B\mathcal{R}(\Delta_{B|\mathbb{S}}, \Delta_{\Theta|\mathbb{M}}) + 2\lambda_B\mathcal{R}(B^\star_{\mathbb{S}^c}, \Theta^\star_{\mathbb{M}^\perp}).$$

Using (30), a final upper bound for the RHS of (25) can be written as

$$(2\lambda_B)\sqrt{s}\|\!|\Delta_B|\!\|_{\mathrm{F}} + (2\lambda_\Theta)\sqrt{2\,K}\|\!|\Delta_\Theta/\sqrt{T}|\!\|_{\mathrm{F}} + (2\lambda_B)\|B^\star_{\mathbb{S}^c}\|_1. \tag{34}$$

Part iii. Aligning (32) and (34) then rearranging terms associated with $\Delta_B$ and $\Delta_\Theta$ gives the claimed bound in (12). ■

*Proof of Corollary 1:* First we note that by the definition of $\mathbf{B}_q(R_q)$, the following holds for the strong support set

$$R_q \geq \sum_{i,j}|B_{ij}|^q \geq \sum_{(i,j)\in\mathcal{S}_\eta^\star}|B_{ij}|^q \geq \eta^q s_\eta, \tag{35}$$

which then gives $\eta^{-q}R_q$. Further, the following inequality holds for the weak support set:

$$\sum_{(i,j)\notin\mathcal{S}_\eta^\star}|B_{ij}| = \sum_{(i,j)\notin\mathcal{S}_\eta^\star}(|B_{ij}|^q|B_{ij}|^{1-q}) \leq R_q\eta^{1-q}. \tag{36}$$

Setting $\eta = \lambda_B/\alpha'$ and plugging (35) and (36) into (12) yields the desired result. ■

Theorem 2 and Corollary 2 can be readily obtained by plugging Lemmas 1 to 3 into to Theorem 1 and Corollary 1 respectively, and thus their proofs are omitted.

## APPENDIX B
## PROOFS FOR LEMMAS

*Proof of Lemma 1:* First, suppose we have that $\forall\, v \in \mathbb{R}^p$,

$$\frac{1}{2} v' \widehat{\Sigma}_{\mathsf{X}} v = \frac{1}{2} v' \left( \frac{\mathsf{X}'\mathsf{X}}{T} \right) v \geq \frac{\alpha_{\text{RSC}}}{2} \|v\|_2^2 - \tau_{\text{tol}} \|v\|_1^2; \quad (37)$$

then, for all $\Delta \in \mathbb{R}^{p \times p}$, and letting $\Delta_j$ denote its $j$th column, the RSC condition automatically holds since

$$\begin{aligned} \tfrac{1}{2T} |||\mathsf{X}\Delta|||_{\text{F}}^2 &= \frac{1}{2} \sum_{j=1}^{q} \Delta_j' \left( \tfrac{\mathsf{X}'\mathsf{X}}{T} \right) \Delta_j \\ &\geq \frac{\alpha_{\text{RSC}}}{2} \sum_{j=1}^{q} \|\Delta_j\|_2^2 - \tau_{\text{tol}} \sum_{j=1}^{q} \|\Delta_j\|_1^2 \\ &\geq \frac{\alpha_{\text{RSC}}}{2} |||\Delta|||_{\text{F}}^2 - \tau_{\text{tol}} \|\Delta\|_1^2. \end{aligned}$$

Therefore, it suffices to verify that (37) holds. In [13, Proposition 4.2], the authors prove a similar result under the assumption that $X_t$ is a VAR$(d)$ process. Here, we adopt the same proof strategy and state the result for a *more general process* $X_t$.

Specifically, by [13, Proposition 2.4(a)], $\forall v \in \mathbb{R}^p, \|v\| \leq 1$ and $\eta > 0$,

$$\begin{aligned} &\mathbb{P}\left[ \left| v' \left( \widehat{\Sigma}_{\mathsf{X}} - \Sigma_X(h) \right) v \right| > 2\pi \mathcal{M}(g_X) \eta \right] \\ &\leq 2\eta \exp\left( -cT \min\{\eta^2, \eta\} \right). \end{aligned}$$

Applying the discretization in [13, Lemma F.2] and taking the union bound, define $\mathbb{K}(2\,s) := \{v \in \mathbb{R}^p, \|v\| \leq 1, \|v\|_0 \leq 2k\}$, and the following inequality holds:

$$\begin{aligned} &\mathbb{P}\left[ \sup_{v \in \mathbb{K}(2\,k)} \left| v' \left( \widehat{\Sigma}_{\mathsf{X}} - \Sigma_X(h) \right) v \right| > 2\pi \mathcal{M}(g_X) \eta \right] \\ &\leq 2 \exp\left( -cT \min\{\eta, \eta^2\} + 2\,k \min\{\log p, \log(21ep/2\,k)\} \right). \end{aligned}$$

With the specified $\gamma = 54\mathcal{M}(g_X)/\mathfrak{m}(g_X)$, set $\eta = \gamma^{-1}$, then apply results from [38, Lemma 12] with $\Gamma = \widehat{\Sigma}_{\mathsf{X}} - \Sigma_X(0)$ and $\delta = \pi \mathfrak{m}(g_X)/27$, so that the following holds

$$\frac{1}{2} v' \widehat{\Sigma}_{\mathsf{X}} v \geq \frac{\alpha_{\text{RSC}}}{2} \|v\|^2 - \frac{\alpha_{\text{RSC}}}{2\,k} \|v\|_1^2,$$

with probability at least $1 - 2\exp(-cT \min\{\gamma^{-2}, 1\} + 2k \log p)$ and note that $\min\{\gamma^{-2}, 1\} = \gamma^{-2}$ since $\gamma > 1$. Finally, let $k = \min\{cT\gamma^{-2}/(c' \log p), 1\}$ for some $c' > 2$, and conclude that with probability at least $1 - c_1 \exp(-c_2\, T)$, the inequality in (37) holds with

$$\alpha_{\text{RSC}} = \pi \mathfrak{m}(g_X), \qquad \tau_{\text{tol}} = \alpha_{\text{RSC}} \gamma^2 \frac{\log p}{2\,T},$$

and so does also the RSC condition. ∎

*Proof of Lemma 2:* We note that

$$\frac{1}{T} ||\mathsf{X}^\top \mathsf{E}||_\infty = \max_{1 \leq i,j \leq p} \left| e_i^\top \left( \mathsf{X}^\top \mathsf{E}/T \right) e_j \right|,$$

where $e_i$ is the $p$-dimensional standard basis with the $i$th entry being 1. Applying [13, Proposition 2.4(b)], for an arbitrary pair of $(i, j)$, the following inequality holds:

$$\begin{aligned} &\mathbb{P}\left[ \left| e_i^\top \left( \mathsf{X}^\top \mathsf{E}/T \right) e_j \right| > 2\pi \left( \mathcal{M}(g_X) + \mathcal{M}(g_\epsilon) + \mathcal{M}(g_{X,\widetilde{\epsilon}}) \right) \eta \right] \\ &\leq 6 \exp\left( -cT \min\{\eta^2, \eta\} \right). \end{aligned}$$

Taking a union bound over all $1 \leq i, j \leq p$, and the following bound holds:

$$\begin{aligned} &\mathbb{P}\Big[ \max_{1 \leq i,j \leq p} \left| e_i^\top \left( \mathsf{X}^\top \mathsf{E}/T \right) e_j \right| > 2\pi \left( \mathcal{M}(g_X) + \mathcal{M}(g_\epsilon) \right. \\ &\quad \left. + \mathcal{M}(g_{X,\widetilde{\epsilon}}) \right) \eta \Big] \\ &\leq 6 \exp\left( -cn \min\{\eta^2, \eta\} + 2 \log p \right). \end{aligned}$$

Set $\eta = c' \sqrt{\log p / T}$ for $c' > (2/c)$ and with the choice of $T \gtrsim \log p$, $\min\{\eta^2, \eta\} = \eta^2$, then with probability at least $1 - c_1 \exp(-c_2 \log p)$, the following bound holds:

$$\frac{1}{T} \|\mathsf{X}^\top \mathsf{E}\|_\infty \leq c_0 \left( \mathcal{M}(g_X) + \mathcal{M}(g_\epsilon) + \mathcal{M}(g_{X,\widetilde{\epsilon}}) \right) \sqrt{\frac{\log p}{T}}.$$

∎

*Proof of Lemma 3:* For $\mathsf{E}$ whose rows are iid realizations of a Gaussian random vector $\epsilon_t$, by [53, Lemma 9], the following bound holds:

$$\mathbb{P}\left[ \left| \left| \left| \widehat{\Sigma}_{\mathsf{E}} - \Sigma_\epsilon \right| \right| \right|_{op} \geq \Lambda_{\max}(\Sigma_\epsilon) \delta(T, p, \eta) \right] \leq 2 \exp(-T\eta^2/2),$$

where $\delta(T, p, \eta) := 2(\sqrt{\frac{p}{T}} + \eta) + (\sqrt{\frac{p}{T}} + \eta)^2$. In particular, by the triangle inequality, with probability at least $1 - 2\exp(-T\eta^2/2)$,

$$\begin{aligned} |||\widehat{\Sigma}_{\mathsf{E}}|||_{op} &\leq |||\Sigma_\epsilon|||_{op} + |||\widehat{\Sigma}_{\mathsf{E}} - \Sigma_\epsilon|||_{op} \leq \Lambda_{\max}(\Sigma_\epsilon) \\ &\quad + \Lambda_{\max}(\Sigma_\epsilon) \delta(T, p, t). \end{aligned}$$

So for $T \geq p$, by setting $\eta = 1$, which yields $\delta(T, p, \eta) \leq 8$ so that with probability at least $1 - 2\exp(-T/2)$, the following bound holds:

$$\Lambda_{\max}(\widehat{\Sigma}_{\mathsf{E}}) \leq 9 \Lambda_{\max}(\Sigma_\epsilon).$$

∎

## ACKNOWLEDGMENT

The authors would like to thank the AE Dr. Rontogiannis and two anonymous referees for many constructive comments and suggestions.

## REFERENCES

[1] A. Agarwal, S. Negahban, and M. J. Wainwright, "Fast global convergence of gradient methods for high-dimensional statistical recovery," *Ann. Statist.*, vol. 40, no. 5, pp. 2452–2482, Oct. 2012.

[2] A. Agarwal, S. Negahban, and M. J. Wainwright, "Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions," *Ann. Statist.*, vol. 40, no. 2, pp. 1171–1197, Apr. 2012.

[3] H. M. Anderson and F. Vahid, "Forecasting the volatility of australian stock returns," *J. Bus. Economic Statist.*, vol. 25, no. 1, pp. 76–90, Jan. 2007.

[4] T. Ando and J. Bai, "Selecting the regularization parameters in high-dimensional panel data models: Consistency and efficiency," *Econometric Reviews*, vol. 37, no. 3, pp. 183–211, Apr. 2016.

[5] D. Angelosante, S. I. Roumeliotis, and G. B. Giannakis, "Lasso-Kalman smoother for tracking sparse signals," in *Proc. Conf. Rec. 43rd Asilomar Conf. Signals, Syst. Computers*., 2009, pp. 181–185.

[6] A. B. Ashcraft, P. Goldsmith-Pinkham, and J. I. Vickery, "MBS ratings and the mortgage credit boom." Elsevier BV, 2010. [Online]. Available: https://www.newyorkfed.org/medialibrary/media/research/staffreports/sr449.pdf

[7] M. Avellaneda and J.-H. Lee, "Statistical arbitrage in the US equities market," *Quantitative Finance*, vol. 10, no. 7, pp. 761–782, Aug. 2010.

[8] J. Bai, "Inferential theory for factor models of large dimensions," *Econometrica*, vol. 71, no. 1, pp. 135–171, Jan. 2003.
[9] J. Bai and S. Ng, "Determining the number of factors in approximate factor models," *Econometrica*, vol. 70, no. 1, pp. 191–221, Jan. 2002.
[10] J. Bai and S. Ng, "Large dimensional factor analysis," *Found. Trends in Econometrics*, vol. 3, no. 2, pp. 89–163, 2008.
[11] J. Bai, K. Li, and L. Lu, "Estimation and inference of FAVAR models," *J. Bus. Economic Statist.*, vol. 34, no. 4, pp. 620–641, Sep. 2016.
[12] B. Baingana, G. Mateos, and G. B. Giannakis, "Dynamic structural equation models for tracking topologies of social networks," in *Proc. IEEE 5th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 2013, pp. 292–295.
[13] S. Basu and G. Michailidis, "Regularized estimation in sparse high-dimensional time series models," *Ann. Statist.*, vol. 43, no. 4, pp. 1535–1567, Aug. 2015.
[14] D. Berberidis and G. B. Giannakis, "Data sketching for large-scale Kalman filtering," *IEEE Trans. Signal Process.*, vol. 65, no. 14, pp. 3688–3701, Jul. 2017.
[15] M. Billio, M. Getmansky, A. W. Lo, and L. Pelizzon, "Econometric measures of connectedness and systemic risk in the finance and insurance sectors," *J. Financial Econ.*, vol. 104, no. 3, pp. 535–559, Jun. 2012.
[16] B. Boots, G. J. Gordon, and S. M. Siddiqi, "A constraint generation approach to learning stable linear dynamical systems," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 1329–1336.
[17] E. J. Candes and Y. Plan, "Matrix completion with noise," *Proc. IEEE*, vol. 98, no. 6, pp. 925–936, Jun. 2010.
[18] G. Chamberlain and M. Rothschild, "Arbitrage, factor structure, and mean-variance analysis on large asset markets," *Econometrica: J. of the Econometric Society*, vol. 51, no. 5, pp. 1281–1304, Sep. 1983.
[19] A. Charles, M. S. Asif, J. Romberg, and C. Rozell, "Sparsity penalties in dynamical system estimation," in *Proc. 45th Annu. Conf. Inf. Sci. Syst.*, Mar. 2011, pp. 1–6.
[20] S. Chen *et al.*, "An M-estimator for reduced-rank system identification," *Pattern Recognition Letters*, vol. 86, pp. 76–81, Jan. 2017.
[21] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto, "Dynamic textures," *Int. J. Comput. Vis.*, vol. 51, no. 2, pp. 91–109, 2003.
[22] B. Eichengreen and K. H. Orourke, *From Commodity to Fiat and Now to Crypto: What Does Hist. Tell Us?* National Bureau of Economic Research, Jan. 2019.
[23] B. Eichengreen, A. Mody, M. Nedeljkovic, and L. Sarno, "How the subprime crisis went global: Evidence from bank credit default swap spreads," *J. Int. Money Fin.*, vol. 31, no. 5, pp. 1299–1318, Apr. 2009.
[24] R. J. Elliott, J. van der Hoek, and W. P. Malcolm, "Pairs trading," *Quantitative Finance*, vol. 5, no. 3, pp. 271–276, Jun. 2005.
[25] J. Fan, Y. Li, and K. Yu, "Vast volatility matrix estimation using high-frequency data for portfolio selection," *J. Amer. Stat. Assoc.*, vol. 107, no. 497, pp. 412–428, Mar. 2012.
[26] Z. Ghahramani and G. E. Hinton, "Technical Progress Report. Part I. Report no. 342," Tech. Rep.CRG-TR-96-2, Jan. 1969.
[27] R. Greenaway-Mcgrevy, C. Han, and D. Sul, "Estimating the number of common factors in serially dependent approximate factor models," *Econ. Letters*, vol. 116, no. 3, pp. 531–534, Sep. 2012.
[28] A. C. Harvey and R. G. Pierse, "Estimating missing observations in economic time series," *J. Amer. Stat. Assoc.*, vol. 79, no. 385, pp. 125–131, Mar. 1984.
[29] A. C. Harvey and P. H. J. Todd, "Forecasting economic time series with structural and Box-Jenkins models: A case study," *J. Bus. Economic Statist.*, vol. 1, no. 4, pp. 299–307, Oct. 1983.
[30] S. Hu, Y. Lucotte, and S. Tokpavi, "Measuring network systemic risk contributions: A leave-one-out approach," *J. Economic Dyn. Control*, vol. 100, pp. 86–114, Mar. 2019.
[31] R. H. Jones, *Longitudinal Data with Serial Correlation: A State-space Approach*. London, U.K.: Chapman and Hall/CRC, May 2018.
[32] R. E. Kalman, "A new approach to linear filtering and prediction problems," *J. Basic Eng.*, vol. 82, no. 1, pp. 35–45, Mar. 1960.
[33] R. E. Kalman and R. S. Bucy, "New results in linear filtering and prediction theory," *J. Basic Eng.*, vol. 83, no. 1, pp. 95–108, Mar. 1961.
[34] U. A. Khan and J. M. F. Moura, "Distributing the Kalman filter for large-scale systems," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 4919–4935, Oct. 2008.
[35] J. Lin and G. Michailidis, "Regularized estimation of high-dimensional factor-augmented vector autoregressive (FAVAR) models," *SSRN Electron. J.*, vol. 18, no. 117, pp. 1–49, 2018.
[36] Z. Liu and M. Hauskrecht, "Clinical time series prediction: Toward a hierarchical dynamical system framework," *Artif. Intell. Med.*, vol. 65, no. 1, pp. 5–18, Sep. 2015.
[37] Z. Liu and P. Spencer, "Modelling sovereign credit spreads with international macro-factors: The case of brazil 1998–2009," *J. Banking Finance*, vol. 37, no. 2, pp. 241–256, Feb. 2013.
[38] P.-L. Loh and M. J. Wainwright, "High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity," *Ann. Statist.*, vol. 40, no. 3, pp. 1637–1664, Jun. 2012.
[39] T. Martinez-Marin, "State-space formulation for circuit analysis," *IEEE Trans. Educ.*, vol. 53, no. 3, pp. 497–503, Aug. 2010.
[40] S. Negahban and M. J. Wainwright, "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise," *J. Mach. Learn. Res.*, vol. 13, pp. 1665–1697, May 2012.
[41] S. N. Negahban, P. K. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of $m$-estimators with decomposable regularizers," *Stat. Sci.*, vol. 27, no. 4, pp. 538–557, Nov. 2012.
[42] Y. Nesterov and L. Scrimali, "Solving strongly monotone variational and quasi-variational inequalities," Tech. Rep. 4, 2011.
[43] S. J. Qin, "An overview of subspace identification," *Comput. Chem. Eng.*, vol. 30, no. 10–12, pp. 1502–1513, Sep. 2006.
[44] S. Schich, "Insurance companies and the financial crisis," *OECD J.: Financial Market Trends*, vol. 2009, no. 2, pp. 123–151, Mar. 2010.
[45] Q. She, Y. Gao, K. Xu, and R. H. Chan, "Reduced-rank linear dynamical systems," in *Proc. 32nd AAAI Conf. on Artificial Intelligence (AAAI-18). AAAI press*, pp. 4050–4057, 2018.
[46] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the EM algorithm," *J. Time Ser. Anal.*, vol. 3, no. 4, pp. 253–264, Jul. 1982.
[47] J. H. Stock and M. W. Watson, "Forecasting using principal components from a large number of predictors," *J. Amer. Stat. Assoc.*, vol. 97, no. 460, pp. 1167–1179, Dec. 2002.
[48] J. H. Stock and M. W. Watson, "Implications of dynamic factor models for VAR analysis," *National Bureau of Economic Research*, Tech. Rep. 11467, Jul. 2005.
[49] J. Taghia *et al.*, "Uncovering hidden brain state dynamics that regulate performance and decision-making during cognition," *Nature Commun.*, vol. 9, no. 1, pp. 1–19, Jun. 2018.
[50] P. Tseng, "Convergence of a block coordinate descent method for non-differentiable minimization," *J. Optim. Theory Appl.*, vol. 109, no. 3, pp. 475–494, Jun. 2001.
[51] P. V. Overschee and B. D. Moor, "N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems," *Automatica*, vol. 30, no. 1, pp. 75–93, Jan. 1994.
[52] M. Verhaegen, "Identification of the deterministic part of MIMO state space models given in innovations form from input-output data," *Automatica*, vol. 30, no. 1, pp. 61–74, Jan. 1994.
[53] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using $\ell_1$-constrained quadratic programming (Lasso)," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2183–2202, May 2009.
[54] R. L. Williams and D. A. Lawrence, *Linear State-Space Control Systems*. Hoboken, New Jersey: Wiley, Feb. 2007.

**Jiahe Lin** received the B.S. degree in mathematics and statistics from the University of Illinois at Urbana Champaign, Champaign, IL, USA, in 2013, and the Ph.D. degree in statistics from the University of Michigan-Ann Arbor, Ann Arbor, MI, USA, in 2018.

**George Michailidis** (Member, IEEE) received the B.S. degree in economics from the University of Athens, Athens, Greece, in 1987, the M.A. degrees in both economics and mathematics from the University of California, Los Angeles (UCLA), Los Angeles, CA, USA, and the Ph.D. degree in mathematics from UCLA. After a Postdoc in operations research with Stanford University, he joined in 1998 the Department of Statistics, University of Michigan in 1998, where he became a Full Professor. In 2015, he joined the University of Florida as the Founding Director of the Informatics Institute. He is a fellow of the American Statistical Association, the Institute of Mathematical Statistics, and the International Statistical Institute. His research interests include high-dimensional statistics, machine learning, network analysis, stochastic control, optimization and queueing theory.