Low Rank and Structured Modeling of High-Dimensional Vector Autoregressions

Sumanta Basu, Xianqi Li, and George Michailidis, *Member, IEEE*

***Abstract*—Network modeling of high-dimensional time series data is a key learning task due to its widespread use in a number of application areas, including macroeconomics, finance, and neuroscience. While the problem of *sparse* modeling based on vector autoregressive models (VAR) has been investigated in depth in the literature, more complex network structures that involve low rank and group sparse components have received considerably less attention, despite their presence in data. Failure to account for low-rank structures results in spurious connectivity among the observed time series, which may lead practitioners to draw incorrect conclusions about pertinent scientific or policy questions. In order to accurately estimate a network of Granger causal interactions after accounting for latent effects, we introduce a novel approach for estimating *low-rank and structured sparse* high-dimensional VAR models. We introduce a regularized framework involving a combination of nuclear norm and lasso (or group lasso) penalties. Subsequently, we establish nonasymptotic probabilistic upper bounds on the estimation error rates of the low-rank and the structured sparse components. We also introduce a fast estimation algorithm and finally demonstrate the performance of theproposed modeling framework over standard sparse VAR estimates through numerical experiments on synthetic and real data.**

***Index Terms*—Lasso, group lasso, nuclear norm, low rank, vector autoregression, probabilistic bounds, identifiability, fast algorithm.**

I. Introduction

THE problem of learning the network structure among a large set of time series arises in many signal processing, economic, finance and biomedical applications. Examples include processing signals obtained from radars [1], [2], macroeconomic policy making and forecasting [3], assessing connectivity among financial firms [4], reconstructing gene regulatory interactions from time-course genomic data [5] and understanding connectivity between brain regions from fMRI measurements [6]. Vector Autoregressive (VAR) models provide a principled framework for these tasks.

Formally, a VAR model for p-dimensional time series X_t is defined in its simplest possible form involving a single time-lag as

$$X^t = B'X^{t-1} + \epsilon^t, \quad t = 1, \ldots, T, \tag{1}$$

where B is a $p \times p$ transition matrix specifying the lead-lag cross dependencies among the p time series and $\{\epsilon^t\}$ is a zero mean error process. VAR models for small number of time series (low-dimensional) have been thoroughly studied in the literature [7]. However, the above mentioned applications, where dozens to hundreds of time series are involved, created the need for the study of VAR models under high dimensional scaling and the assumption that their interactions are *sparse* to compensate for the possible lack of adequate number of time points (samples; see [8] and references therein). There has been a growing body of literature on sparse estimation of large scale VAR models, including alternative penalties beyond the popular ℓ_1 penalty (lasso), such as the Berhu regularization introduced in [9], group lasso type penalties employed in [21], [22], as well as non-convex penalties akin to a square-root lasso investigated in [11]. Further, [10] examines estimation of the transition matrix and the inverse covariance matrix of the error process through a joint sparse penalty. Note that the problem of sparse estimation of these two model parameters separately from a least squares and maximum likelihood viewpoints is addressed in [3], [8], respectively, where in addition probabilistic finite sample error bounds for the obtained estimates are obtained.

Nevertheless, there are occasions where the sparsity assumption may not be sufficient. For example, during financial crisis periods, returns on assets move together in a more concerted manner [4], [12], while transcription factors regulate a large number of genes that may lead to hub-node network structures [13]. Similarly, in brain connectivity networks, particular tasks activate a number of regions that cross talk in a collaborative manner [14]. Hence, it is of interest to study VAR models under high dimensional scaling where the transition matrix governing the temporal dynamics exhibits a more complex structure; e.g. it is *low rank* and/or *(group) sparse*.

In a low-dimensional regime, where the number of time points scales to infinity, but the number of time series under study remains fixed, [15] examined asymptotic properties of VAR models, where the parameters exhibit reduced rank structure and

Manuscript received April 1, 2018; revised October 11, 2018; accepted November 20, 2018. Date of publication January 3, 2019; date of current version January 14, 2019. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xavier Mestre. The work of S. Basu was supported by the National Science Foundation under Grant DMS-1812128. The work of X. Li was supported by the University of Florida Informatics Institute Fellowship. The work of G. Michailidis was supported by the National Science Foundation under Grants IIS-1632730, CNS-1422078, CCF-1540093, DMS-1545277. *(Corresponding author: George Michailidis.)*

S. Basu is with the Department of Statistical Science, Cornell University, Ithaca, NY 14850 USA (e-mail: sumbose@cornell.edu).

X. Li is with the Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02115 USA (e-mail: xianqili@ufl.edu).

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

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

Digital Object Identifier 10.1109/TSP.2018.2887401

also discussed connections with canonical correlation analysis of such models presented in [16]. Specifically, the transition matrix B in (2) can be written as the product of two rank-k matrices Φ, Ψ, i.e. $B = \Phi\Psi'$, so that in the resulting model specification of the original p time series is expressed as linear combinations $Z^t = \Psi X^t$ of the original ones, and Φ specifies the dependence between X^t and Z^t; namely $X^t = \Phi' Z^{t-1} + \epsilon^t$. Hence, to obtain Φ and Ψ, [15] suggests to estimate the parameters of the original model in (1) under the constraint that $B = \Phi\Psi$ and that rank$(B) = k$. Other works include low rank approximations of Hankel matrices that represent the input-output structure of a linear time invariant systems and were studied in [17], [18]. Finally, a brief mention to the possibility that the VAR transition matrix may exhibit such a structure appeared as a motivating example in [19].

On the other hand, there is a mature literature on imposing low rank plus sparse, or pure group sparse structure for many learning tasks for independent and identically distributed (i.i.d.) data. Examples include group sparsity in regression and graphical modeling [20], low rank and sparse matrix approximations for dimension reduction [18], etc. However, as shown in [8], the presence of temporal dependence across observations induces intricate dependencies between both rows and columns of the design matrix of the corresponding least squares estimation problem, as well as between the design matrix and the error term, that require careful handling to establish consistency properties for the model parameters under sparsity and high dimensional scaling. These issues are further compounded when more complex regularizing norms are involved, as discussed in [21]. In this paper, the authors model grouping structures within each column of B, but do not consider a low-rank component. In contrast, we focus on groups potentially spanning across different columns and allow a low-rank component in B.

More recently, [37] and [38] extended the framework of [18] beyond decomposing an observable matrix, to the problem of Gaussian process identification, by assuming a low-rank plus sparse structure on the inverse spectral density and the transfer function of a general VAR(d) system, respectively. Our work is complementary to these recent developments. We directly model the transition matrix of a VAR(1) process that enables us to identify *a directed network of* (group) sparse Granger causal relationships that are of interest in a number of applications; e.g. in financial economics where firms with higher out-degree are of particular interest in measuring systemic risk [4], [12]. Further, our approach explicitly addresses identifiability issues for extracting the respective low-rank and sparse components, which in turn are leveraged to obtain probabilistic error bounds that characterize the quality of their estimates. The latter provide insights to the practitioner on sample size requirements and tuning parameter selection for real data applications. Finally, note that our approach to the issue of identifiability builds on [19], wherein we characterize the degree of unidentifiability which guides in an explicit manner the selection of the tuning parameters used in the proposed optimization algorithm.

Further, to estimate the posited model in (1) with B being both low-rank and structured sparse (henceforth indicating that it could be either *pure sparse or group sparse or both*), we also introduce a fast accelerated proximal gradient algorithm, inspired by [30], [31], for the corresponding optimization problems. The key idea is that instead of searching for the local Lipschitz constant of the gradient of the smooth component of the objective function, the proposed algorithm utilizes a safeguarded Barzilai-Borwein (BB) initial stepsize [25] and employs relaxed line search conditions to achieve better performance in practice. The latter enables the selection of more "aggressive" stepsizes, while preserving the accelerated convergence rate of $\mathcal{O}(\frac{1}{k^2})$, where k denotes the number of iterations required until convergence. Finally, the performance of the model parameters under different structures together with the associated estimation procedure based on the accelerated proximal gradient algorithm are calibrated on synthetic data, and illustrated on three data sets examining realized volatilities of stock prices of 75 large financial firms before, during and after the 2007–09 US financial crisis.

Notation: Throughout the paper, we employ the following notation: $\|.\|$, $\|.\|_2$ and $\|.\|_F$ denote the ℓ_2-norm of a vector, the spectral norm and the Frobenius norm of a matrix, respectively. For a $p \times p$ matrix B, the symbol $\|B\|_*$ is used to denote the nuclear norm, i.e. $\sum_{j=1}^{p} \sigma_j(B)$, the sum of the singular values of a matrix, while $B^\dagger$ denotes the conjugate transpose of a matrix B. For any matrix B, we use $\|B\|_0$ to denote $card(vec(B))$, $\|B\|_1$ for $\|vec(B)\|_1$ and $\|B\|_{\max}$ to denote $\|vec(B)\|_\infty$. Further, if $\{G_1, G_2, \ldots, G_K\}$ denote a partition of $\{1, 2, \ldots, p^2\}$ into K non-overlapping groups, then we use $\|B\|_{2,1}$ to denote $\sum_{k=1}^{K} \|(B)_{G_k}\|_F$, $\|B\|_{2,\max}$ for $\max_{k=1,2,\ldots K} \|(B)_{G_k}\|_F$, while $\|B\|_{2,0}$ denotes the number of nonzero groups in B. Here, with a little abuse of notation, we use B_{G_k} to denote $vec(B)_{G_k}$. In addition, $\Lambda_{\max}(.)$, $\Lambda_{\min}(.)$ denote the maximum and minimum eigenvalues of a symmetric or Hermitian matrix. For any integer $p \geq 1$, we use $\mathbb{S}^{p-1}$ to denote the unit ball $\{v \in \mathbb{R}^p : \|v\| = 1\}$. We also use $\{e_1, e_2, \ldots\}$ generically to denote unit vectors in $\mathbb{R}^p$, when p is clear from the context. Finally, for positive real numbers A, B, we write $B \gtrsim A$ if there exists an absolute positive constant c, independent of the model parameters, such that $B \geq cA$.

II. Model Formulation and Estimation Procedure

Consider a VAR(1) model where the transition matrix B is low-rank plus structured sparse given by

$$X^t = B' X^{t-1} + \epsilon^t, \quad \epsilon^t \overset{i.i.d.}{\sim} N(0, \Sigma_\epsilon), \tag{2}$$

$$B = L^* + R^*, \; rank(L^*) = r, \tag{3}$$

where L^* corresponds to the low rank component and R^* represents either a sparse S^*, or group-sparse component G^*. It is further assumed that the number of non-zero elements in the sparse case is $\|S^*\|_0 = s$, while in the group sparse case the number of non-zero groups is $\|G^*\|_{2,0} = g$, with $r \ll p, s \ll p^2$ and $g \ll p^2$. The matrix L^* captures persistence structure across *all* p time series and enables the model to be applicable in settings where there are strong cross-autocorrelations, a feature that the standard sparse VAR model is not equipped to handle. The sparse or group sparse component captures additional cross-sectional autocorrelation structure among the time series.

Finally, it is assumed that the error terms are serially uncorrelated. Our objective is to estimate L^* and R^* accurately based on a relatively small number of samples $N \ll p^2$.

Stability: In order to ensure consistent estimation, we assume that the posited VAR model in (2) is stable; i.e. its characteristic polynomial $\mathcal{B}(z) := I_p - B^{'}z$ satisfies $\det(\mathcal{B}(z)) \neq 0$ on the unit circle of the complex plane $\{z \in \mathcal{C} : |z| = 1\}$. This is a common assumption in the literature of multivariate time series [7], required for consistency and asymptotic normality of low-dimensional VAR models. This assumption also ensures that the spectral density of the VAR model

$$f_X(\theta) = \frac{1}{2\pi}\left(\mathcal{B}^{-1}\left(e^{i\theta}\right)\right)\Sigma_\epsilon\left(\mathcal{B}^{-1}\left(e^{i\theta}\right)\right)^\dagger, \ \ \theta \in [-\pi, \pi], \quad (4)$$

is bounded above in spectral norm.

It was shown in [8] that this condition is sufficient to establish consistency of regularized VAR estimates of a sparse transition matrix. Further, the following quantities play a central role in the error bounds of the regularized estimates:

$$\begin{aligned}
\mathcal{M}(f_X) &= \sup_{\theta\in[-\pi,\pi]} \Lambda_{\max}(f_X(\theta)),\\
\mathfrak{m}(f_X) &= \sup_{\theta\in[-\pi,\pi]} \Lambda_{\min}(f_X(\theta)),\\
\mu_{\max}(\mathcal{B}) &= \max_{|z|=1} \Lambda_{\max}\left(\mathcal{B}^\dagger(z)\mathcal{B}(z)\right),\\
\mu_{\min}(\mathcal{B}) &= \min_{|z|=1} \Lambda_{\min}\left(\mathcal{B}^\dagger(z)\mathcal{B}(z)\right). \qquad (5)
\end{aligned}$$

As shown in [8], $\mathcal{M}(f_X)$ and $\mathfrak{m}(f_X)$ together capture the narrowness of the spectral density of a time series. Processes with stronger temporal and cross-sectional dependence have narrower spectra that in turn lead to slower convergence rates for the regularized estimates. For VAR models, $\mathcal{M}(f_X)$ and $\mathfrak{m}(f_X)$ are related to $\mu_{\max}(\mathcal{B})$ and $\mu_{\min}(\mathcal{B})$ as follows:

$$\mathfrak{m}(f_X) \geq \frac{1}{2\pi}\frac{\Lambda_{\min}(\Sigma_\epsilon)}{\mu_{\max}(\mathcal{B})}, \ \ \mathcal{M}(f_X) \leq \frac{1}{2\pi}\frac{\Lambda_{\max}(\Sigma_\epsilon)}{\mu_{\min}(\mathcal{B})}. \quad (6)$$

Proposition 2.2 in [8] provides a lower bound on $\mu_{\min}(\mathcal{B})$. For the special structure of the models considered here, we can get an improved upper bound on $\mu_{\max}(\mathcal{B})$, as shown in the following lemma:

Lemma II.1: For a stable VAR(1) model of the class (2), we have

$$\mu_{\max}(\mathcal{B}) \leq \left[1 + l + (v_{in} + v_{out})/2\right]^2 \quad (7)$$

where l is the largest singular value of L^*, $v_{in} = \max_{1\leq j\leq p} \sum_{i=1}^p |R^*_{ij}|$ and $v_{out} = \max_{1\leq i\leq p} \sum_{j=1}^p |R^*_{ij}|$.

Proof: $\|\mathcal{B}(z)\| = \|I - (L^* + R^*)z\| \leq \|I\| + \|L^*\| + \|R^*\|$ for any $z \in \mathbb{C}$ with $|z| = 1$. The result follows from the fact that $\mu_{\max}(\mathcal{B}) = \max_{|z|=1} \|\mathcal{B}(z)\|^2$. ■

A. Estimation Procedure

The estimation of VAR model parameters is based on the following regression formulation (see [7]). Given $T + 1$ consecutive observations $\{X^0, \ldots, X^T\}$ from the VAR model, we work with the autoregressive design as follows:

$$\underbrace{\begin{bmatrix}(X^T)'\\ \vdots \\ (X^1)'\end{bmatrix}}_{\mathcal{Y}} = \underbrace{\begin{bmatrix}(X^{T-1})'\\ \vdots \\ (X^0)\end{bmatrix}}_{\mathcal{X}} B + \underbrace{\begin{bmatrix}(\epsilon^T)'\\ \vdots \\ (\epsilon^1)'\end{bmatrix}}_{E}. \quad (8)$$

This is a standard regression problem with $N \equiv T$ samples and $q = p^2$ variables. Our goal is to estimate L^*, R^* with high accuracy when $N \ll p^2$.

There is an *inherent identifiability* issue in the estimation of the components L^* and R^*. Suppose the low-rank component L^* itself is s-sparse or g-group sparse and the sparse or group-sparse component R^* is of rank r. In that scenario, we can not hope for any method to estimate L^* and R^* separately without imposing any further constraints. So, a minimal condition for low-rank and sparse (or group-sparse) recovery is that the low rank component should not be too sparse and the sparse (group-sparse) component should not be low-rank.

This issue has been rigorously addressed in the literature (e.g. [18]) for independent and identically distributed data and resolved by imposing an *incoherence* condition. Such a condition is sufficient for *exact* recovery of the low rank and the sparse or group-sparse component by solving a convex program. In a recent paper, [19] showed that in a noisy setting where exact recovery of the two components is impossible, it is still possible to achieve good estimation error under a comparatively mild assumption. In particular, they formulated a general measure for the *radius of non-identifiability* of the problem under consideration and established a non-asymptotic upper bound on the estimation error $\|\hat{L} - L^*\|_F^2 + \|\hat{R} - R^*\|_F^2$, which depends on this radius. The key idea is to allow for simultaneously sparse (or group-sparse) and low-rank matrices in the model, and control for the error introduced. We refer the readers to the above paper for a more detailed discussion on this notion of non-identifiability. In this work, the low-rank plus sparse (group-sparse) decomposition problem under restrictions on the radius of non-identifiability takes the form

$$\begin{aligned}
(\hat{L}, \hat{R}) &= \underset{\substack{L,R\in\mathbb{R}^{p\times p}\\ L\in\Omega}}{\operatorname{argmin}} \ l(L, R),\\
l(L, R) &:= \frac{1}{2}\|\mathcal{Y} - \mathcal{X}(L + R)\|_F^2 + \lambda_N\|L\|_* + \mu_N\|R\|_\diamond. \qquad (9)
\end{aligned}$$

Here $\Omega = \{L \in \mathbb{R}^{p\times p} : \|L\|_{\max} \leq \frac{\alpha}{p}\}$ (for sparse) or $\{L \in \mathbb{R}^{p\times p} : \|L\|_{2,\max} \leq \frac{\beta}{\sqrt{K}}\}$ (for group sparse), $\|\cdot\|_\diamond$ represents $\|\cdot\|_1$ or $\|\cdot\|_{2,1}$ depending on sparsity or group sparsity of R, and λ_N and μ_N are non-negative tuning parameters controlling the regularizations of low-rank and sparse/group-sparse parts. The parameters α and β control the degree of non-identifiability of the matrices allowed in the model class. For instance, larger values of α provide sparser estimates of S and allow simultaneously sparse and low-rank components to be absorbed in $\hat{L}$. A smaller value of α, on the other hand, tends to produce a matrix L with smaller rank and pushes the simultaneously low-rank and sparse components to be absorbed in $\hat{S}$. In practice,

we recommend choosing α and β in the range $[1, p]$ and $[1, K]$, respectively. The issue of selecting them robustly in practice is discussed in Section V.

Remark: On certain occasions, it may be useful to have both sparse *and* group-sparse components in the model, in addition to the low rank structure. We then have $R^* = S^* + G^*$ in (9) with $\|R\|_\diamond = \|S\|_1 + \frac{\nu_N}{\mu_N}\|G\|_{2,1}$. However, to guarantee the *simultaneous identifiability* of the sparse and group-sparse components from the low-rank component, stronger conditions need to be imposed on L; namely, $\Omega = \{L \in \mathbb{R}^{p\times p} : \|L\|_{\max} \leq \frac{\alpha}{p}\ \&\ \|L\|_{2,\max} \leq \frac{\beta}{\sqrt{K}}\}$.

Remark: Note that the estimated VAR model is not guaranteed to be stable, although the error bound analysis in Section III ensures its stability with high probability, as long as the sample size is large enough and the true generative model is stable. For network reconstruction and visualization purposes, stability of the estimated VAR is not strictly required. However, enforcing stability is essential for forecasting purposes. When there is a small deviation of the estimated model from stability (e.g the spectral radius of the estimated $\hat{B}$ is a little over 1), stability can be enforced through a post-processing step of shrinking the moduli of eigenvalues of $\hat{B}$ below 1, while keeping its eigenvectors unchanged. This type of projection argument is common in covariance and correlation matrix estimation with missing data for ensuring positive definiteness of the estimates [41]. However, in case of moderate to large deviation from stability, a closer look at the individual time series is recommended to re-assess the validity of the VAR formulation. For example, in macroeconomics, it is customary to use suitable transformations of the component time series to ensure that each of the individual time series and the resulting VAR model is stable, as opposed to modeling the individual and the joint time series (without transformation) as unit root and co-integrating processes. For instance, see [39], [40] for specific recommendations on useful transformations for macroeconomic time series.

III. Theoretical Properties of the Estimates

Next, we derive non-asymptotic probabilistic upper bounds on the estimation errors of the low-rank plus structured sparse components of the transition matrix B. The main result shows that consistent estimation is possible with a sample size of the order $N \asymp p\,\mathcal{M}^2(f_X)/\mathfrak{m}^2(f_X)$, as long as the process $\{X^t\}$ is stable and the radius of nondentifiability, as measured by $\|L^*\|_{\max}$ and/or $\|L^*\|_{2,\max}$, is small in an appropriate sense detailed next.

To establish the results, we first consider fixed realizations of $\mathcal{X}$ and E and impose the following assumptions:

1) *Restricted Strong Convexity (RSC):* There exist $\zeta > 0$ and $\tau_N > 0$ such that

$$\frac{1}{2}\|\mathcal{X}\Delta\|_{\mathrm{F}}^2 \geq \frac{\zeta}{2}\|\Delta\|_{\mathrm{F}}^2 - \tau_N\Phi^2(\Delta),\ \text{ for all }\ \Delta \in \Re^{p\times p}$$

where $\Phi(\Delta) = \inf_{L+R=\Delta}\{\lambda_N\|L\|_* + \mu_N\|R\|_\diamond\}$, and

2) *Deviation Conditions:* There exists a deterministic function $\phi(B, \Sigma_\epsilon)$ of the model parameters B and Σ_ϵ such that

$$\|\mathcal{X}'E/N\|_2 \leq \phi(B,\Sigma_\epsilon)\sqrt{p/N},\ \text{ and}$$

$$\|\mathcal{X}'E/N\|_{\max} \leq \phi(B,\Sigma_\epsilon)\sqrt{\frac{2\log p}{N}},\ \text{ and}$$

$$\|\mathcal{X}'E/N\|_{2,\max} \leq \phi(B,\Sigma_\epsilon)\frac{\sqrt{m\log K}}{\sqrt{N}},$$

where m is the size of the largest group $\max_{1\leq k\leq K} card(G_k)$. Later on, we show that assumptions 1) and 2) are indeed satisfied with high probability when the data are generated from model (2).

Next, we present the non-asymptotic upper bounds on the estimation errors of the low-rank plus structured sparse components, respectively.

Proposition 1: (a) Suppose that the matrix L^* has rank at most r, while the matrix S^* has at most s nonzero entries. Then, for any $\lambda_N \geq 4\|\mathcal{X}'E\|_2$ and $\mu_N \geq 4\|\mathcal{X}'E\|_{\max} + \frac{4\zeta\alpha}{p}$, any solution $(\hat{L}, \hat{S})$ of (9) satisfies

$$\|\hat{L} - L^*\|_F^2 + \|\hat{S} - S^*\|_F^2 \leq \frac{4}{\zeta^2}\left(\frac{9}{2}\lambda_N^2 r + 4\mu_N^2 s\right).$$

(b) Suppose that the matrix L^* has rank at most r, while the matrix G^* has at most g non-zero groups. Then, for any $\lambda_N \geq 4\|\mathcal{X}'E\|_2$ and $\mu_N \geq 4\|\mathcal{X}'E\|_{2,\max} + \frac{4\zeta\beta}{\sqrt{K}}$, any solution $(\hat{L}, \hat{G})$ of (9) satisfies

$$\|\hat{L} - L^*\|_F^2 + \|\hat{G} - G^*\|_F^2 \leq \frac{4}{\zeta^2}\left(\frac{9}{2}\lambda_N^2 r + 4\mu_N^2 g\right)$$

Remark: It should be noted that if each group in G^* has only one element, then we have $K = p^2$ and g non-zero entries. For such cases, part (b) of Proposition 1 becomes identical to part (a).

As a byproduct, we also give the estimation error bound of the transition matrix which can be characterized by the sparse plus group-sparse and the low-rank plus sparse and group-sparse components, respectively, under the assumption that the strength of the connections in the group-sparse component G is weak; i.e. $G \in \Psi$ with $\Psi = \{G \in \mathbb{R}^{p\times p} : \|G\|_{\max} \leq \frac{\gamma}{p}\}$, where $\gamma \in [1, p]$.

Proposition 2: (a) Suppose that the matrix S^* has at most s nonzero entries, while the matrix G^* has at most g non-zero groups. Then, for any $\mu_N \geq 4\|\mathcal{X}'E\|_{\max} + \frac{4\zeta\gamma}{p}$ and $\nu_N \geq 4\|\mathcal{X}'E\|_{2,\max}$, any solution $(\hat{S}, \hat{G})$ of (9) satisfies

$$\|\hat{S} + \hat{G} - S^* - G^*\|_F^2 \leq \frac{4}{\zeta^2}(8\mu_N^2 s + 9\nu_N^2 g).$$

(b) Suppose that the matrix L has rank at most r, while the matrix S has at most s nonzero entries and the matrix G has at most g non-zero groups. Then, for any $\lambda_N \geq 4\|\mathcal{X}'E\|_2$, $\mu_N \geq 4\|\mathcal{X}'E\|_{\max} + \frac{4\zeta\alpha}{p} + \frac{4\zeta\gamma}{p}$, and $\nu_N \geq 4\|\mathcal{X}'E\|_{2,\max} + \frac{4\zeta\beta}{\sqrt{K}}$, any

solution $(\hat{L}, \hat{S}, \hat{G})$ of (9) satisfies

$$\|\hat{L} - L^*\|_F^2 + \|\hat{S} + \hat{G} - S^* - G^*\|_F^2 \leq \frac{4}{\zeta^2}\left(9\lambda_N^2 r + \frac{25}{2}\mu_N^2 s + 8\nu_N^2 g\right).$$

See Appendix A for detailed proofs of Propositions 1 and 2.

Note that the objective in Proposition 2 is not the accurate recovery of the S^* and G^* components *separately*. The latter can be in principle achieved, if one sets γ to a very small value.

In order to obtain meaningful results in the context of our problem, we need upper bounds on $\|\mathcal{X}'E\|_2$, $\|\mathcal{X}'E\|_{\max}$ and $\|\mathcal{X}'E\|_{2,\max}$ and a lower bound on $\Lambda_{\min}(\mathcal{X}'\mathcal{X})$ that hold with high probability. For the case of independent and identically distributed data, such high-probability deviation bounds are established in [19]. However, for time series data all entries of the $\mathcal{X}$ matrix are dependent on each other, and hence it is a nontrivial technical task to establish such deviation bounds. A key technical contribution of this work is to derive these deviation bounds, which lead to meaningful analysis for VAR models. The results rely on the measure of stability defined in (5) and an analysis of the joint spectrum of $\{X^{t-1}\}$ and $\{\epsilon^t\}$ undertaken next.

Proposition 3: Consider a random realization of $\{X^0, \ldots, X^T\}$ generated according to a stable VAR(1) process (2) and form the autoregressive design (8). Define

$$\phi(B, \Sigma_\epsilon) = \Lambda_{\max}(\Sigma_\epsilon)\left[1 + \frac{1 + \mu_{\max}(\mathcal{B})}{\mu_{\min}(\mathcal{B})}\right]$$

Then, there exist universal positive constants $c_i > 0$ such that

1) for $N \gtrsim p$,

$$\mathbb{P}\left[\left\|\frac{\mathcal{X}'E}{N}\right\|_2 > c_0\phi(B, \Sigma_\epsilon)\sqrt{\frac{p}{N}}\right] \leq c_1 \exp\left[-c_2 \log p\right]$$

and for any $N \gtrsim \log p$,

$$\mathbb{P}\left[\left\|\frac{\mathcal{X}'E}{N}\right\|_{\max} > c_0\phi(B, \Sigma_\epsilon)\sqrt{\frac{\log p}{N}}\right] \leq c_1 \exp\left[-c_2 \log p\right]$$

and for $N \gtrsim m \log p$,

$$\mathbb{P}\left[\left\|\frac{\mathcal{X}'E}{N}\right\|_{2,\max} > c_0\phi(B, \Sigma_\epsilon)\frac{\sqrt{m \log p}}{\sqrt{N}}\right] \leq c_1 \exp\left[-c_2 \log p\right]$$

2) for $N \gtrsim p\mathcal{M}^2(f_X)/\mathfrak{m}^2(f_X)$,

$$\mathbb{P}\left[\Lambda_{\min}\left(\frac{\mathcal{X}'\mathcal{X}}{N}\right) > \frac{\Lambda_{\min}(\Sigma_\epsilon)}{2\mu_{\max}(\mathcal{B})}\right] \leq c_1 \exp\left[-c_2 \log p\right]$$

See Appendix B for the detailed proof of Proposition 3.

Applying the above deviation bounds to the non-asymptotic errors of Propositions 1, we obtain the final result for approximate recovery of the low-rank and the structured sparse components using nuclear and $\ell_1/\ell_{2,1}$ norm relaxations, as we show next.

Proposition 4: Consider the setup of Proposition 3. There exist universal positive constants $c_i > 0$ such that for $N \gtrsim p\mathcal{M}^2(f_X)/\mathfrak{m}^2(f_X)$, and $\|L^*\|_{\max} \leq \alpha/p$, any solution $(\hat{L}, \hat{S})$ of the program (9) satisfies, with probability at least $1 - c_1 \exp[-c_2 \log p]$,

$$\|\hat{S} - S^*\|_F^2 + \|\hat{L} - L^*\|_F^2 \leq \frac{c_0\phi^2(B, \Sigma_\epsilon)\mu_{\max}^2(\mathcal{B})}{\Lambda_{\min}^2(\Sigma_\epsilon)} \frac{(rp + s\log p)}{N} + \frac{32\Lambda_{\min}^2(\Sigma_\epsilon)}{\mu_{\max}^2(\mathcal{B})}\frac{s\alpha^2}{p^2}. \tag{10}$$

Remark: The error bound presented in the above proposition consists of two key terms. The first term is the error of estimation emanating from randomness in the data and limited sample capacity. For a given model, this error goes to zero as the sample size increases. The second term represents the error due to the unidentifiability of the problem. This is more fundamental to the structure of the true low-rank and structured sparse components, and depends only on the model parameters and does not vanish, even with infinite sample size.

Further, the estimation error is a product of two terms - the second term $(rp + s\log p)/N$ involves the dimensionality parameters and matches the parametric convergence rate for independent observations. The effect of dependence in the data is captured through the first part of the term: $\frac{c_0\phi^2(B,\Sigma_\epsilon)\mu_{\max}^2(\mathcal{B})}{\Lambda_{\min}^2(\Sigma_\epsilon)}$. As discussed in [8], this term is larger when the spectral density is more spiky, indicating a stronger temporal and cross-sectional dependence in the data.

Proposition 5: Consider the setup of Proposition 3. There exist universal positive constants $c_i > 0$ such that for $N \gtrsim p\mathcal{M}^2(f_X)/\mathfrak{m}^2(f_X)$, for any G^0 with $\|L^*\|_{2,\max} \leq \beta/\sqrt{K}$, any solution $(\hat{L}, \hat{G})$ of the program (9) satisfies, with probability at least $1 - c_1 \exp[-c_2 \log p]$,

$$\|\hat{G} - G^*\|_F^2 + \|\hat{L} - L^*\|_F^2 \leq \frac{c_0\phi^2(B, \Sigma_\epsilon)\mu_{\max}^2(\mathcal{B})}{\Lambda_{\min}^2(\Sigma_\epsilon)} \frac{(rp + g(m\log p))}{N} + \frac{32\Lambda_{\min}^2(\Sigma_\epsilon)}{\mu_{\max}^2(\mathcal{B})}\frac{g\beta^2}{K}. \tag{11}$$

See Appendix B for detailed proofs of Propositions 4 and 5.

Remark: Based on Proposition 5, analogous conclusions can be obtained to those for the low rank plus sparse case.

IV. Computational Algorithm and Its Convergence Properties

Next, we introduce a fast algorithm for estimating the transition matrix B from data. For ease of presentation and to convey the key ideas clearly, we first present the algorithm for B representing a single structure (e.g. only low rank, or only group sparse, or only sparse), and in addition establish its convergence properties. Subsequently, we modify the algorithm to handle the composite structures considered in this paper and also establish its convergence.

The fast network structure learning (FNSL) Algorithm 1 is described next. A safeguarded BB initial value is selected, as

Algorithm 1: Fast Network Structure Learning (FNSL) method.

Choose $C \geq 0, \sigma > 1, \eta_{0,1} \geq \eta_{\min}$. Set $\alpha_1 = 1, B_1^{ag} = B_1$, and $Q_1 = 0$.

For $i = 1, 2, \ldots, k$, 1.

//Backtracking

1) Set $\eta_i = \alpha_i \eta_{0,i}$, where $\eta_{0,i}$ is from (12). Solve α_i from $\frac{1}{\alpha_{i-1}\eta_{i-1}} = \frac{1-\alpha_i}{\alpha_i \eta_i}$ for $i > 1$. Compute

$$B_i^{md} = (1-\alpha_i)B_i^{ag} + \alpha_i B_i,$$

$$B_{i+1} = \arg\min_B \Big\{ \langle \nabla l(B_i^{md}), B \rangle + \frac{\eta_i}{2}\|B - B_i\|_F^2 + P_B(B, \lambda) \Big\},$$

$$\Gamma_i = \|B_{i+1} - B_i\|^2 - \frac{\alpha_i}{\eta_i}\|\mathcal{X}(B_{i+1} - B_i)\|_F^2,$$

$$Q_{i+1} = \beta_i Q_i + \Gamma_i, \text{ where } 0 \leq \beta_i \leq \left(1 - \frac{1}{i}\right)^2.$$

2) If $Q_{i+1} < -C/i^2$, then replace $\eta_{0,i}$ by $\sigma\eta_{0,i}$ and return to step 1.

//Updating iterates

3) Compute

$$B_{i+1}^{ag} = (1-\alpha_i)B_i^{ag} + \alpha_i B_{i+1}.$$

EndFor

Output B_{k+1}^{ag}.

the initial choice of the nominal step η_i, i.e.

$$\eta_{0,i} = \max\left\{\eta_{\min}, \frac{\|\mathcal{X}(B_i - B_{i-1})\|_F^2}{\|B_i - B_{i-1}\|_F^2}\right\} \text{ for } i > 1. \quad (12)$$

For notational convenience, the penalty term for estimating the transition matrix B is denoted by $P_B(B, \lambda)$, where $\lambda > 0$ denotes the tuning parameter. The specific B_{i+1} update depends on the employed penalty term; for an ℓ_1 penalty inducing sparsity, it corresponds to soft-thresholding [27], for a group sparse penalty to group soft-thresholding [20], while for a nuclear norm penalty to singular value thresholding [28].

It can also been seen from Algorithm 1, that for $\alpha_i \equiv 1$ for $\forall\ i \geq 1$, then $B_i^{md} = B_i$ and $B_{i+1}^{ag} = B_{i+1}$, which leads to the traditional gradient descent algorithm. Indeed, Algorithm 1 is obtained by incorporating an efficient backtracking strategy into the accelerated multi-step scheme by [29], [30]. It provides a different way to look for a larger stepsize by employing a relaxed line search condition, instead of searching for the gradient Lipschitz constant of the data fidelity term. Steps 1 and 2 constitute the backtracking ones. Both B_i^{md} and B_i^{ag} are linear combinations of all past iterations of B_i, but based on different weights as can be seen from their updates, i.e. $B_i^{ag} = (1-\alpha_{i-1})B_{i-1}^{ag} + \alpha_{i-1}B_i$ and $B_i^{md} = (1-\alpha_i)B_i^{ag} + \alpha_i B_i$. Here 'ag' simply denotes 'aggregate'. The data fidelity term in the cost function is linearized at B_i^{md}. Since it is used after we have obtained B_i^{ag} and before we obtain B_{i+1}, we use 'md' to denote a 'middle' update. Further, Γ_i and Q_i are the parameters most closely related to our line search conditions. Intuitively, if we set $\Gamma_i \geq 0$, we are certainly able to guarantee the accelerated rate of convergence. However, this will render the stepsize smaller. Fortunately, our convergence analysis enables us to relax this condition by utilizing the summing up procedure, with Q_i corresponding to the part of the sum of Γ_i. Thus, we can impose a relaxed termination condition on Q_i (see step 2 of Algorithm 1) without impacting the rate of convergence while being able to obtain a more aggressive stepsize. In fact, parameter C in Step 2 plays an important role, i.e. the number of the trial steps can be reduced significantly when a relatively larger C is selected. However, the value of C can not be too large either, since it might impair the convergence rate in terms of the objective function value.

The convergence rate of Algorithm 1 is established next.

Proposition 6: Let $\{B_{k+1}^{ag}\}$ be generated by Algorithm 1. Then, for any $k \geq 1$

$$l(B_{k+1}^{ag}) - l(\hat{B}) \leq \frac{2\sigma\|\mathcal{X}\|_2^2\|B_0 - \hat{B}\|_F^2 + \tilde{C}}{(k+1)^2} \quad (13)$$

where $\tilde{C}$ is a finite positive number independent of k.

See Appendix C for its detailed proof. It should be noted that the convergence rate of $\{B_{k+1}\}$ is still an open problem, which needs to be addressed further in future, considering its good performance for some cases.

Next, we enhance the algorithm for solving (9) in the general case. The accelerated convergence rate can be obtained by following the proof of Proposition 6. Similarly to the case in Algorithm 1, the initial trial step of η_i is a safeguarded BB choice

$$\eta_{0,i} = \max\left\{\eta_{\min}, \frac{\|\mathcal{X}(L_i + R_i - L_{i-1} - R_{i-1})\|_F^2}{\|L_i + R_i - L_{i-1} - R_{i-1}\|_F^2}\right\} \quad (14)$$

The update of the L component is based on singular value thresholding, while that of the R component on (group) soft-thresholding. Note that the most expensive computational operation corresponds to the singular value decomposition (SVD) when updating L_{i+1}. As mentioned earlier, the proposed algorithm is able to look for larger magnitude step sizes by conducting fewer number of line searches, due to employing more relaxed line search conditions. Actually, this is an important improvement considering the computational cost of SVD. Indeed, the efficiency of the proposed algorithm can be enhanced further if we employ the truncated SVD [32] instead of the full SVD.

The convergence rate of the proposed Algorithm 2 (given in supplement due to limited space) is established next.

Proposition 7: Let $(L_{k+1}^{ag}, R_{k+1}^{ag})$ be a sequence of updates generated by Algorithm 2. Then, for any $k \geq 1$

$$l(L_{k+1}^{ag}, R_{k+1}^{ag}) - l(\hat{L}, \hat{R}) \leq \frac{2\sigma\|\mathcal{X}\|_2^2\big(\|L_0 - \hat{L}\|_F^2 + \|R_0 - \hat{R}\|_F^2\big) + \tilde{C}}{(k+1)^2} \quad (15)$$

where $\tilde{C}$ is a finite positive number independent of k.

TABLE I
PARAMETER SETTINGS IN THE PROPOSED ALGORITHMS FOR ALL THE EXPERIMENTS

Parameters	σ	$\eta_{\min}$	C	β_k
Values	2	$\|\mathcal{X}'\mathcal{X}\|_2/10$	100	1/k

Proposition 7 is a direct extension of Proposition 6 and can be obtained by following the roadmap of the proof for the former result.

V. PERFORMANCE EVALUATION

Next, we present experimental results on both synthetic and real data. Specifically, the first two experiments focus on large-scale network learning with single penalty term to show the efficiency and effectiveness of the proposed algorithms, while the remaining ones assess the accuracy of recovering low rank plus structured sparse transition matrices B.

A. Performance Metrics and Experimental Settings

We introduce the performance metrices used in the numerical work. For network estimation, we use the true positive rate (TPR) and false alarm rate (FAR) defined as:

- TPR := $\frac{\sharp\{\hat{b}_{ij}\neq 0 \text{ and } b_{ij}\neq 0\}}{\sharp\{b_{ij}\neq 0\}}$
- FAR := $\frac{\sharp\{b_{ij}=0 \text{ and } \hat{b}_{ij}\neq 0\}}{\sharp\{b_{ij}=0\}}$

where b_{ij} and $\hat{b}_{ij}$ are the correspnding elemnets in B and $\hat{B}$, respectively. The estimation error (EE) and out-of-sample prediction error (PE) are defined as

- EE := $\frac{\|\hat{B}-B\|_F}{\|B\|_F}$
- PE := $\|\hat{\mathcal{Y}}-\mathcal{Y}\|_F^2/\|\mathcal{Y}\|_F^2$

To select the optimal value of the tuning parameters, we combine the three (or two or one, respectively) -dimensional grid search method with the AIC/BIC/forward cross-validation criteria. We will specify the criterion on a case by case basis for the following experiments. In examples B and C, the tuning parameter λ is selected by the AIC criterion. A grid of 100 values in the interval $[0, \|\mathcal{X}'\mathcal{Y}\|_{\max}]$ is used for λ. In examples D, E and F, we utilize a two/three-dimensional grid search to select the optimal values of λ_N, μ_N and/or ν_N as that for λ. For the experiments employing synthetic data, the tuning parameters are selected by assuming the rank of the true low-rank transition matrix and/or the non-zero group-sparse components of the true group-sparse transition matrix are known. We will specify the forward cross-validation procedure for the real data case in example G.

For all the experiments, the parameters used in the proposed algorithms are depicted in Table I. Also we set $\Sigma_\epsilon = \epsilon^2 I$ and $\epsilon^2 = 1$. We rescale the entries of B to ensure stability of the process (the spectral radius $\rho \in (0.45, 0.95)$. All the results are based on 50 replications. Finally, all algorithms are run in the MATLAB R2015a environment on a PC equipped with 12GB memory.

TABLE II
PERFORMANCE COMPARISON OF FNSL WITH A VARIANT OF FISTA ON A LARGE-SCALE SPARSE NETWORK STRUCTURE LEARNING PROBLEM

p	N	method	(TPR, FAR)(%)	EE	T
800	1000	FISTA	(88.5, 14.4)	0.22	11.9
		FNSL	(88.6, 13.9)	0.22	**6.1**
	1500	FISTA	(90.8, 12.3)	0.17	14.7
		FNSL	(90.7, 12.0)	0.17	**8.2**
	2000	FISTA	(91.3, 11.0)	0.14	17.6
		FNSL	(91.3, 11.3)	0.14	**10.8**
900	1000	FISTA	(87.8, 15.0)	0.24	13.1
		FNSL	(87.8, 14.7)	0.24	**7.2**
	1500	FISTA	(89.1, 10.7)	0.19	16.3
		FNSL	(89.1, 10.5)	0.19	**8.7**
	2000	FISTA	(90.9, 12.0)	0.16	20.7
		FNSL	(90.9, 11.8)	0.16	**13.5**
1000	1000	FISTA	(88.5, 15.2)	0.25	18.6
		FNSL	(88.4, 14.8)	0.25	**11.2**
	1500	FISTA	(90.3, 14.1)	0.20	21.5
		FNSL	(90.2, 13.8)	0.20	**12.8**
	2000	FISTA	(91.2, 12.1)	0.16	24.7
		FNSL	(91.2, 12.4)	0.16	**15.1**

B. Large-Scale Sparse Network Learning

We start by comparing the performance of the proposed algorithm 1 with FISTA with line search [33] to solve problem (9) with a sparse transition matrix.

We consider three different VAR(1) models with $p = 800$, 900 and 1000 variables. For each of these models, we generate $N = 1000, 1500$, and 2000 observations from a Gaussian VAR(1) process (2). The $p \times p$ transition matrix B with sparsity is generated in the following way. First, the topology is generated from a directed random graph $G(p, \xi)$, where the edge from one node to another node occurs independently with probability $\xi = 10/p$. Then, the strength of the edges is generated independently from a Gaussian distribution. This process is repeated until we obtain a transition matrix B with a desired spectral radius ρ. We compare TPR, FAR, EE, and computational time (denoted by T).

Table II shows the experimental results for sparse network structure with different network size p and sample size N. It can be seen that the proposed algorithm performs similarly to FISTA in terms of TPR, FAR, and estimation error. To show the efficiency of the proposed algorithm, we also compare the computational time in seconds in terms of the convergence of the objective function value. Clearly, the proposed algorithm outperforms FISTA in efficiency, especially when the network and sample size become larger. This is mainly due to the relaxed line search scheme, as previously discussed. To further support our claim, we also show the graphs of the decreasing objective function value vs. CPU, see Fig. 1 when $p = 1000$ and $N = 2000$.

Table III shows comparisons between a variant of FISTA and FNSL in terms of objective function values, CPU time in seconds and the number of matrix products for a network of size

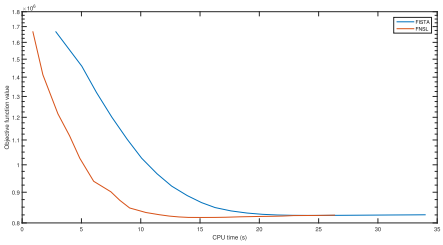

Fig. 1. Performance comparison of a variant of FISTA and the proposed algorithm: plotting the objective function values vs. CPU time for sparse network learning problem with $p = 1000$ and $N = 2000$.

TABLE III
COMPARISON OF OBJECTIVE FUNCTION VALUE, CPU TIME IN SECONDS, AND THE NUMBER OF MATRIX PRODUCTS (AX) FOR A VARIANT OF FISTA AND FNSL ON A SPARSE NETWORK STRUCTURE FOR DIFFERENT SAMPLE SIZES

Algorithms	Objective value	CPU	AX
$p = 1000, N = 1000$			
FISTA	4.140e+5	18.6	66
FNSL	4.089e+5	11.2	44
$p = 1000, N = 1500$			
FISTA	6.137e+5	21.5	62
FNSL	6.071e+5	12.8	38
$p = 1000, N = 2000$			
FISTA	8.239e+5	24.7	66
FNSL	8.169e+5	15.1	41

$p = 1000$ with sample size 1000, 1500 and 2000, respectively. For each data set, FISTA needs around 30 iterations to reach convergence and the total number of line searches is 3 for all iterations, while FNSL needs no more than 20 iterations and the total number of line searches is no more than 4 for all iterations. This illustrates the computational savings of the proposed algorithm.

C. Network Learning With a Low-Rank Transition Matrix

To further show the efficiency of the proposed algorithms, we compare the performance of a variant of FISTA [34] and FNSL on estimating low-rank transition matrices.

We consider three different VAR(1) models with $p = 200$, 300 and 400 variables. For each of these models, we generate $N = 400, 1200$, and 2000 observations from a Gaussian VAR(1) process (2). The $p \times p$ low-rank transition matrix B is generated with rank $\lfloor p/25 \rfloor + 1$. Subsequently, we rescale the entries of B to ensure the spectral radius ρ lies in $(0.45, 0.95)$. We compare the rank of the estimated transition matrix, denoted by $\hat{r}$, EE, and computational time T.

Table IV shows the experimental results for low-rank network structure with different network and sample size. Both FISTA and the proposed algorithm achieve good recovery of the transition matrix B with the correct rank and they have similar performance in terms of estimation error. Clearly, the proposed algorithm outperforms FISTA in efficiency for this case as well.

TABLE IV
PERFORMANCE COMPARISON OF A VARIANT OF FISTA AND FNSL ON ESTIMATING LOW-RANK TRANSITION MATRICES PROBLEMS

p	N	method	$\hat{r}$	EE	T
200	400	FISTA	8	0.80	4.9
		FNSL	8	0.80	**3.4**
	1200	FISTA	8	0.63	9.7
		FNSL	8	0.63	**6.9**
	2000	FISTA	8	0.57	14.2
		FNSL	8	0.57	**10.5**
300	400	FISTA	12	0.84	7.8
		FNSL	12	0.84	**5.7**
	1200	FISTA	12	0.72	16.1
		FNSL	12	0.72	**11.7**
	2000	FISTA	12	0.68	18.2
		FNSL	12	0.68	**13.8**
400	400	FISTA	16	0.87	8.5
		FNSL	16	0.87	**5.9**
	1200	FISTA	16	0.82	20.5
		FNSL	16	0.82	**17.4**
	2000	FISTA	16	0.75	31.4
		FNSL	16	0.75	**25.1**

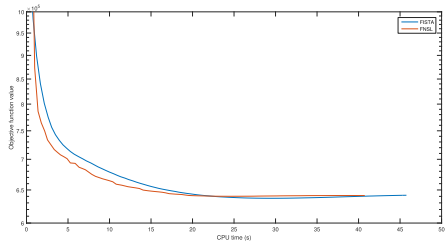

Fig. 2. Performance comparison of a variant of FISTA and the proposed algorithm: plotting the objective function values vs. CPU time for a low-rank transition matrix estimation problem with with $p = 400$ and $N = 2000$.

The graphs of the decreasing objective function value vs. CPU are depicted in Fig. 2.

The first two experiments focused on computational efficiency of the proposed algorithms, while retaining good network estimation properties. Next, we demonstrate their accuracy for learning structured sparse networks.

D. Sparse Plus Low-Rank Network Learning

Next, we investigate estimation of sparse plus low-rank transition matrices and compare it to ordinary least square (OLS) and lasso estimates.

We consider three different VAR(1) models with $p = 50$, 75 and 100 variables. For each of these models, we generate $N = 100$ and 200 observations from the model defined in (2) where B can be decomposed into a low-rank matrix L of rank $\lfloor p/25 \rfloor + 1$ and a sparse matrix S with 2–4% non-zero entries. We rescale the entries of B to ensure stability of the process (the spectral radius is set to $\rho(B) = 0.7$). We compare the estimation and out-of-sample prediction errors. The number of out of samples is set to 10.

TABLE V
TRUE POSITIVE RATE AND FALSE ALARM RATE OF THE L+S MODEL ON IDENTIFYING THE SPARSE COMPONENT S WITH DIFFERENT α

α	(TPR, FAR)(%)
p/8	(84.5, 17.4)
p/4	(82.4, 17.2)
p/2	(80.4, 17.5)
p	(73.2, 17.1)
2p	(54.7, 17.0)
4p	(40.2, 17.8)
8p	(22.7, 17.4)

TABLE VI
PERFORMANCE COMPARISON OF L+S WITH OLS AND LASSO

p	N	model	(TPR, FAR)(%)	EE	PE
50	100	OLS	(-, -)	0.84	0.72
		Lasso	(73.2, 30.0)	0.69	0.53
		L+S	**(76.3, 18.9)**	**0.48**	**0.47**
	200	OLS	(-, -)	0.52	0.41
		Lasso	(77.3, 35.0)	0.57	0.45
		L+S	**(80.4, 17.5)**	**0.31**	**0.36**
75	100	OLS	(-, -)	0.75	0.37
		Lasso	(71.0, 24.7)	0.75	0.37
		L+S	**(79.0, 18.0)**	**0.51**	**0.29**
	200	OLS	(-, -)	0.53	0.18
		Lasso	(77.0, 28.6)	0.67	0.22
		L+S	**(83.8, 18.3)**	**0.36**	**0.16**
100	100	OLS	(-, -)	3.7	4.0
		Lasso	(57.3, 29.0)	1.06	1.05
		L+S	**(52.3, 20.1)**	**0.92**	**1.0**
	200	OLS	(-, -)	2.07	1.73
		Lasso	(59.4, 25.5)	0.86	0.95
		L+S	**(60.4, 20.5)**	**0.72**	**0.90**

First, we study the influence of α in (9) on this learning problem with $p = 50$ and $N = 200$. From Table V, that a smaller α parameter leads to markedly improved identification of all the true nonzero entries in the sparse component, which consequently leads to better separating the sparse component S from the low-rank component L.

The corresponding estimation errors are reported in Table VI. In all three settings, we find that the low-rank plus sparse VAR estimates outperform the estimates using ordinary least-squares (OLS) and lasso, as expected. We observe that as the ratio of N/p increases, OLS may produce lower estimation error than lasso, even though the OLS model is not interpretable for this case. Further, we note that the estimation errors of all three methods decrease with increasing sample sizes as expected and predicted by theory. In addition, we illustrate how the squared Frobenius norm error in Proposition 4 of the VAR model with low rank plus sparse transition matrix scales with the sample size N and dimension p, when the rank r of B is fixed. The network size p is set to $50, 100, 150$ and 200, respectively, while the rank r is fixed to be 2 for all p. Sparsity s and ε are defined similarly as above, while the sample size is set to $N \in (150, 5500)$. The squared Frobenius norm error of

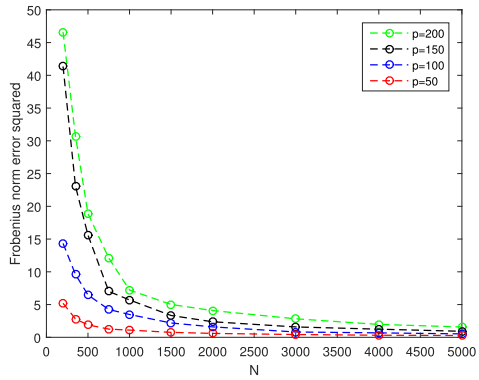

Fig. 3. Estimation error of low rank plus sparse structure $\|S - \hat{S}\|_{\mathrm{F}}^2 + \|L - \hat{L}\|_{\mathrm{F}}^2$ with different network size p and sample size N.

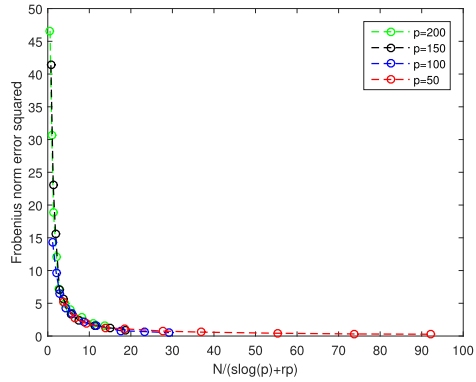

Fig. 4. Estimation error of low rank plus sparse structure $\|S - \hat{S}\|_{\mathrm{F}}^2 + \|L - \hat{L}\|_{\mathrm{F}}^2$ with rescaled sample size $N/(slog(p) + rp)$.

estimation given by $\|S - \hat{S}\|_{\mathrm{F}}^2 + \|L - \hat{L}\|_{\mathrm{F}}^2$ is depicted in Fig. 3, which displays the errors for different values of p, plotted against the sample size N. As predicted by our theoretical result, the error is larger for larger p. Fig. 4 displays the error against the rescaled sample size $N/(s\,log(p) + rp)$. it can be seen that the corresponding error curves for different values of p align well, which is consistent with the estimation error obtained in Proposition 4.

In addition to its improved estimation and prediction performance, the low-rank plus sparse modeling strategy aids in recovering the underlying Granger causal network after accounting for the latent structure. In Figs. 5, we demonstrate this using a VAR(1) model with $p = 50$ and $n = 200$. The top panel of the Figs. 5 displays the true transition matrix B, its low-rank component L and the structure of its sparse component S. The bottom panel of the Figs. 5 displays the structure of the Granger causal networks estimated by the method of Lasso and the low-rank plus sparse modeling strategy. As predicted by the theory, it can be that the lasso estimate of the Granger causal network selects many false positives due to its failure to account for the latent structure. On the other hand, the sparse component S provides an estimate exhibiting significantly fewer false positives entries.

It is interesting to note that the estimation performance of the regularized estimates in low-rank plus sparse VAR models

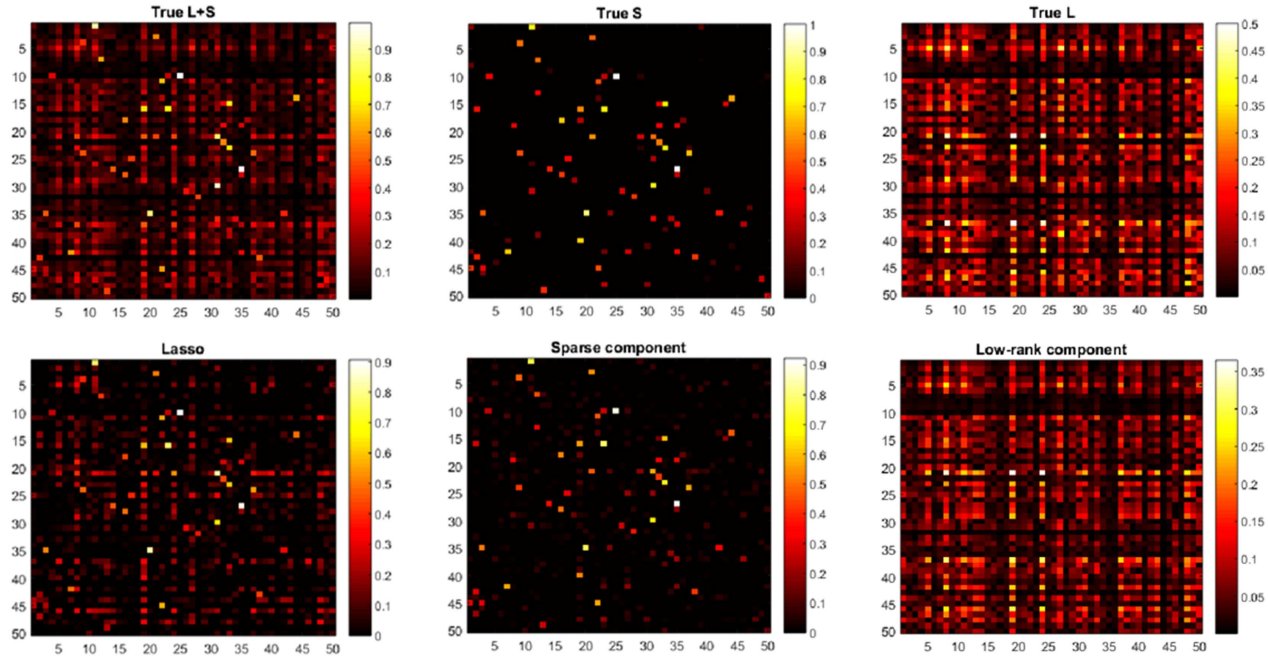

Fig. 5. Estimated Granger causal networks using lasso and low-rank plus sparse VAR estimates. The top panel displays the true transition matrix B, the structure of its sparse component S and its low-rank component L. The bottom panel displays the structure of the Granger causal networks estimated by lasso ($\hat{B}_{Lasso}$), the low-rank plus sparse modeling strategy ($\hat{S}$) and the estimated low-rank component ($\hat{L}$).

is worse than the performance of lasso in sparse VAR models of similar dimension [8], even for the same sample sizes. This is in line with the error bounds presented in Proposition 4. The estimation error in low-rank plus sparse models is of the order of $O(rp + s\log p)/N$, while the error of lasso in sparse VAR models scales at a faster rate of $O(s\log p/N)$. Further note that a s-sparse VAR requires estimating s parameters in S, while the presence of r factors introduces an additional rp parameters in the loading matrix Λ.

E. Sparse Plus Group-Sparse Plus Low-Rank Network Learning

Finally, we conduct numerical experiments to assess the performance of low-rank plus sparse plus group-sparse modeling in VAR analysis and compare it to the performance of sparse plus group-sparse and low-rank plus sparse estimates.

We consider three different VAR(1) models with $p = 50$, 100 and 150 variables. For each of these models, we generate $N = 200$ and 300 observations from the Gaussian VAR(1) process defined in (2), where B can be decomposed into a low-rank matrix L of rank $\lfloor p/25 \rfloor + 1$, a sparse matrix S with 2–4% non-zero entries, and a group-sparse matrix G with each column corresponding to a different group for a total of p groups. We rescale the entries of B to ensure stability of the process (the spectral radius is set to $\rho(B) = 0.7$) and compare the estimation and out-of-sample prediction errors, with the number of out-samples set to 10.

The corresponding estimation errors are reported in Table VII. In those three settings, we find that the low-rank plus sparse plus group-sparse VAR estimates performs only slightly better than low-rank plus sparse VAR estimates. One of the reasons lie in that the ability of the identification will degrade as more structures are involved. The other one is that multiple-times shrinkage for the multiple structures lead to severe bias estimation. Even though the group structures can be recovered completely, some non-zero elements in the sparse component vanished. An ad-hoc way to improve the performance for this case is to combine these two methods together. However, both methods outperform the estimates using sparse plus group-sparse VAR. We also

TABLE VII
PERFORMANCE COMPARISON OF L+S+G WITH S+G AND L+S

p	N	method	(TPR, FAR)(%)	EE	PE
50	200	S+G	(85.5, 34.1)	0.46	0.53
		L+S	(82.6, 26.4)	0.41	0.51
		L+S+G	**(83.3, 26.9)**	**0.41**	**0.51**
	300	S+G	(91.7, 47.0)	0.37	0.58
		L+S	(88.6, 24.4)	0.31	0.56
		L+S+G	**(90.6, 24.9)**	**0.30**	**0.56**
100	200	S+G	(92.3, 49.9)	0.48	0.73
		L+S	(84.3, 28.4)	0.44	0.72
		L+S+G	**(85.3, 27.3)**	**0.44**	**0.72**
	300	S+G	(94.8, 49.0)	0.44	0.72
		L+S	(89.6, 25.4)	0.37	0.70
		L+S+G	**(90.0, 25.1)**	**0.36**	**0.70**
150	200	S+G	(92.0, 50.5)	0.64	0.73
		L+S	(83.3, 28.8)	0.57	0.71
		L+S+G	**(84.0, 28.1)**	**0.55**	**0.70**
	300	S+G	(93.6, 50.2)	0.55	0.71
		L+S	(85.6, 27.4)	0.46	0.68
		L+S+G	**(86.4, 27.6)**	**0.46**	**0.68**

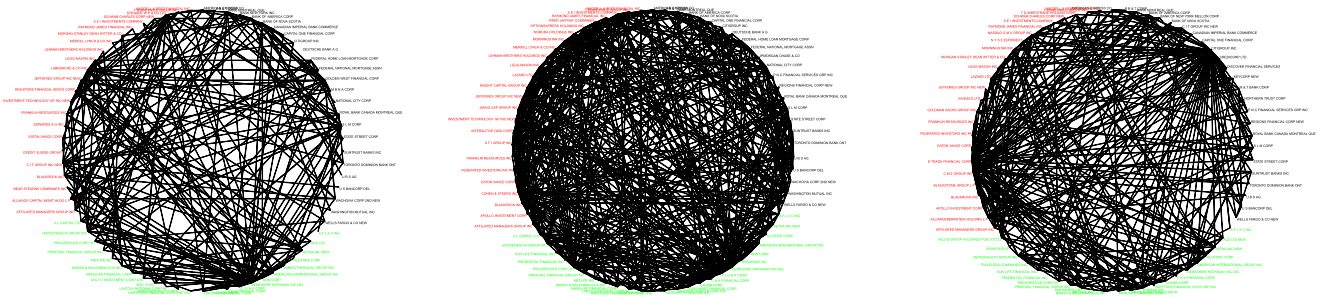

Fig. 6. Estimated Granger causal networks based on L+S estimates. The left panel depicts the structure of the Granger causal network of the estimated sparse component $\hat{S}$ in the pre-crisis period and has 298 edges, the middle during the crisis and has 592 edges, while the right panel in the post-crisis periods and has 345 edges.

observe that, as the ratio of N/p increases, the estimation errors of all three methods decrease with increasing sample sizes, as expected and predicted by theory.

F. Structured Network Learning of Asset Pricing Data

Finally, we employ the proposed framework to learn Granger causal networks of asset pricing data obtained from the University of Chicago's Center for Research in Security Prices (CRSP) and retrieved from the Wharton Research Data Service (WRDS). Specifically, we examine the network structure of realized volatilities of financial institutions representing banks (BA), primary broker/dealers (PB) and insurance companies (INS). The analysis was performed across the following time periods: September 2002–August 2005, September 2006–August 2008 and September 2010–August 2012 that correspond to instances before the financial crisis of 2008 (pre-crisis period), the build-up and apex of the crisis and the post-crisis period, respectively. For each period, we collected data on 75 firms with 25 companies in each of the three categories - BA, PB and INS based on the size of the average market capitalization of each firm, but dropped a few due to duplicate/missing observations. The final form of the variables used are based on the log transformation of the difference between the highest and lowest stock price during a day that acts as a proxy for realized volatility and subsequently detrending it.

To select the tuning parameters, we employ the following forward cross-validation procedure: (1) We use a time window of length W, the available number of time points in the data. Then, starting from time t, we use the most recent W observations to estimate B and denote the transition matrix estimate by $\hat{B}_t$. We use the next W' observations (right after W observations) to validate $\hat{B}_t$. (2) We select the optimal tuning parameters $(\lambda_N^{'}, \mu_N^{'})$ so that

$$(\lambda_N^{'}, \mu_N^{'}) = \arg\min \left\{ \frac{1}{\lfloor \frac{m-500}{25} \rfloor} \sum_{t=500+25*i}^{m} Err(\hat{B}_t) \right\}$$

where $Err(\hat{B}_t) = \|\mathcal{Y}_{W'} - \mathcal{X}_{W'}\hat{B}_t\|_F^2$ and $i = 0, ...$

In our analysis, we set $W = 500$ and $W' = 50$. Further, to separate the sparse component from the low-rank component as much as possible, we set $\alpha = p/10$. The learned Granger causal network structures (the sparse component $\hat{S}$) estimated by a sparse plus low-rank model are depicted in Fig. 6. It can be seen that even in the presence of a low-rank component, the sparse component exhibits a certain density (about 5% in the pre-crisis period, rising to 10% during the crisis and dropping down to about 6% in the post-crisis period). This increased connectivity during the crisis period has been observed in Granger causal networks for log-returns as well [4]. In the Supplement, we also provide the Granger causal network structures estimated by only assuming sparsity of the transition matrix B (see Supplement Fig. 5). A similar increased connectivity pattern is observed during the crisis period. Also note that after accounting for the low-rank component, the estimated sparse component is significantly more sparse than that estimated by a lasso approach, thus enabling us to better examine specific firms that are key drivers in the volatility network.

VI. Discussion

Our modeling and technical developments were based on a VAR(1) model. However, it can be generalized to VAR(d) models in different ways, depending on the context of the application. One possible formulation where the low rank component stays the same across lags can be expressed as

$$X_t = \sum_{\ell=1}^{d} (L + S_\ell) X_{t-\ell} + v_t.$$

This model can be posed as a 1-order (L+S) model on the concatenated process $\tilde{X}_t = [X_t^\top, X_{t-1}^\top, \ldots, X_{t-d+1}^\top]^\top$ using the standard transformation [7]:

$$\underbrace{\begin{bmatrix} X_t \\ X_{t-1} \\ \vdots \\ X_{t-d} \end{bmatrix}}_{\tilde{X}_t} = \underbrace{\begin{bmatrix} L & L & \ldots & L \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 \end{bmatrix}}_{\tilde{L}} \underbrace{\begin{bmatrix} X_{t-1} \\ X_{t-2} \\ \vdots \\ X_{t-d-1} \end{bmatrix}}_{\tilde{X}_{t-1}}$$

$$+ \underbrace{\begin{bmatrix} S_1 & S_2 & \ldots & S_d \\ I & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & I & 0 \end{bmatrix}}_{\tilde{S}} \underbrace{\begin{bmatrix} X_{t-1} \\ X_{t-2} \\ \vdots \\ X_{t-d-1} \end{bmatrix}}_{\tilde{X}_{t-1}} + \underbrace{\begin{bmatrix} v_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_{\tilde{v}_t}.$$

Since the matrices $\tilde{L}$ and $\tilde{S}$ are respectively low-rank and sparse, our proposed algorithm can be used and the error bounds will be applicable. Further, to maintain the special structure of these new matrices, additional constraints can be imposed to the posited objective function or the final estimates can be projected to this space.

Finally, motivated by [37], [38], it would be of interest to consider analogous developments in the frequency domain and address identifiability issues, as well as establish finite sample error bounds.

APPENDIX A

The proof of Proposition 1 and part (a) of Proposition 2 can be easily obtained by the following proof for part (b) of Proposition 2.

Proof of part (b) of Proposition 2: By the optimality of $(\hat{L}, \hat{S}, \hat{G})$ and the feasibility of (L^*, S^*, G^*), we have

$$\begin{aligned}\frac{1}{2}\|\mathcal{Y}-\mathcal{X}(\hat{L}+\hat{S}+\hat{G})\|_F^2+\lambda_N\|\hat{L}\|_*+\mu_N\|\hat{S}\|_1\\ +\nu_N\|\hat{G}\|_{2,1}\leq\frac{1}{2}\|\mathcal{Y}-\mathcal{X}(L^*+S^*+G^*)\|_F^2\\ +\lambda_1\|L^*\|_*+\mu_N\|S^*\|_1+\nu_N\|G^*\|_{2,1}\end{aligned}\tag{16}$$

By setting $\hat{\Delta}^L=\hat{L}-L^*$, $\hat{\Delta}^S=\hat{S}-S^*$, and $\hat{\Delta}^G=\hat{G}-G^*$ and combining with $\mathcal{Y}=\mathcal{X}(L^*+S^*+G^*)+E$, we have

$$\begin{aligned}\frac{1}{2}\|\mathcal{X}(\hat{\Delta}^L+\hat{\Delta}^S+\hat{\Delta}^G)\|_F^2\leq\langle\hat{\Delta}^L+\hat{\Delta}^S+\hat{\Delta}^G,\mathcal{X}'E\rangle\\ +\lambda_N\|L^*\|_*+\mu_N\|S^*\|_1+\nu_N\|G^*\|_{2,1}-\lambda_N\|L+\hat{\Delta}^L\|_*\\ -\mu_N\|S^*+\hat{\Delta}^S\|_1-\nu_N\|G^*+\hat{\Delta}^G\|_{2,1}\end{aligned}$$

By Lemma 1 in [19] and lemma 2.3 in [35], we obtain

$$\begin{aligned}\frac{1}{2}\|\mathcal{X}(\hat{\Delta}^L+\hat{\Delta}^S+\hat{\Delta}^G)\|_F^2\leq\langle\hat{\Delta}^L+\hat{\Delta}^S+\hat{\Delta}^G,\\ \mathcal{X}'E\rangle+\lambda_N(\|\hat{\Delta}_A^L\|_*-\|\hat{\Delta}_B^L\|_*)\\ +2\lambda_N\sum_{j=r+1}^{d}\sigma_j(L^*)+\mu_N(\|\hat{\Delta}_M^S\|_1-\|\hat{\Delta}_{M^\perp}^S\|_1)\\ +2\mu_N\|S_{M^\perp}^*\|_1+\nu_N(\|\hat{\Delta}_N^G\|_{2,1}-\|\hat{\Delta}_{N^\perp}^G\|_{2,1})\\ +2\nu_N\|G_{N^\perp}^*\|_{2,1}\end{aligned}\tag{17}$$

where the matrices $(A,B)\in\{(\bar{A},\bar{B}):\bar{A}\bar{B}'=0\ \&\ \bar{A}'\bar{B}=0\}$, $(M,M^\perp)$ and $(N,N^\perp)$ denote an arbitrary subspace pair for which $\|S\|_1$ and $\|G\|_{2,1}$ are decomposable, respectively. Since

$$\begin{aligned}\langle\hat{\Delta}^L+\hat{\Delta}^S+\hat{\Delta}^G,\mathcal{X}'E\rangle\leq\|\hat{\Delta}^L\|_*\|\mathcal{X}'E\|_2\\ +\|\hat{\Delta}^S\|_1\|\mathcal{X}'E\|_{\max}+\|\hat{\Delta}^G\|_{2,1}\|\mathcal{X}'E\|_{2,\max}\\ \leq(\|\hat{\Delta}_A^L\|_*+\|\hat{\Delta}_B^L\|_*)\|\mathcal{X}'E\|_2+(\|\hat{\Delta}_M^S\|_1+\|\hat{\Delta}_{M^\perp}^S\|_1)\\ \|\mathcal{X}'E\|_{\max}+(\|\hat{\Delta}_N^G\|_{2,1}+\|\hat{\Delta}_{N^\perp}^G\|_{2,1})\|\mathcal{X}'E\|_{2,\max}\end{aligned}\tag{18}$$

Substituting (18) into (17) and recalling conditions for λ_N, μ_N and ν_N, we have

$$\begin{aligned}\frac{1}{2}\|\mathcal{X}(\hat{\Delta}^L+\hat{\Delta}^S+\hat{\Delta}^G)\|_F^2\leq\frac{3}{2}\lambda_N\|\hat{\Delta}_A^L\|_*\\ +\frac{3}{2}\mu_N\|\hat{\Delta}_M^S\|_1+\frac{3}{2}\nu_N\|\hat{\Delta}_N^G\|_{2,1}+2\lambda_N\sum_{j=r+1}^{d}\\ \sigma_j(L^*)+2\mu_N\|S_{M^\perp}^*\|_1+2\nu_N\|G_{N^\perp}^*\|_{2,1}\end{aligned}\tag{19}$$

By the RSC condition, the constraints on L and G in (9), and the definition of μ_N and ν_N, we have

$$\begin{aligned}\frac{1}{2}\|\mathcal{X}(\hat{\Delta}^L+\hat{\Delta}^S+\hat{\Delta}^G)\|_F^2\geq\frac{\zeta}{2}\|\hat{\Delta}^L+\hat{\Delta}^S+\hat{\Delta}^G\|_F^2\\ \geq\frac{\zeta}{2}\|\hat{\Delta}^L\|_F^2+\frac{\zeta}{2}\|\hat{\Delta}^S\|_F^2+\frac{\zeta}{2}\|\hat{\Delta}^G\|_F^2-\zeta|\langle\hat{\Delta}^L,\hat{\Delta}^S\rangle|\\ -\zeta|\langle\hat{\Delta}^L,\hat{\Delta}^G\rangle|-\zeta|\langle\hat{\Delta}^G,\hat{\Delta}^S\rangle|\\ \geq\frac{\zeta}{2}\|\hat{\Delta}^L\|_F^2+\frac{\zeta}{2}\|\hat{\Delta}^S\|_F^2+\frac{\zeta}{2}\|\hat{\Delta}^G\|_F^2-\zeta\|\hat{\Delta}^L\|_{\max}\|\hat{\Delta}^S\|_1\\ -\zeta\|\hat{\Delta}^L\|_{2,\max}\|\hat{\Delta}^G\|_{2,1}\quad-\zeta\|\hat{\Delta}^G\|_{\max}\|\hat{\Delta}^S\|_1\\ \geq\frac{\zeta}{2}\|\hat{\Delta}^L\|_F^2+\frac{\zeta}{2}\|\hat{\Delta}^S\|_F^2+\frac{\zeta}{2}\|\hat{\Delta}^G\|_F^2-\frac{\mu_N}{2}\|\hat{\Delta}^S\|_1\\ -\frac{\nu_N}{2}\|\hat{\Delta}^G\|_{2,1}-\frac{\mu_N}{2}\|\hat{\Delta}^S\|_1\end{aligned}$$

Inserting the above inequality into (19), we have

$$\begin{aligned}\frac{\zeta}{2}(\|\hat{\Delta}^L\|_F^2+\|\hat{\Delta}^S\|_F^2+\|\hat{\Delta}^G\|_F^2)\leq\frac{3}{2}\lambda_N\|\hat{\Delta}_A^L\|_*\\ +\frac{3}{2}\mu_N\|\hat{\Delta}_M^S\|_1+\frac{3}{2}\nu_N\|\hat{\Delta}_N^G\|_{2,1}+\mu_N\|\hat{\Delta}^S\|_1+\frac{\nu_N}{2}\|\hat{\Delta}^G\|_{2,1}\\ +2\lambda_N\sum_{j=r+1}^{d}\sigma_j(L^*)+2\lambda_N\|S_{M^\perp}^*\|_1+2\mu_N\|G_{N^\perp}^*\|_{2,1}\end{aligned}$$

By the compatibility constant in [19], we have

$$\begin{aligned}\frac{\zeta}{2}(\|\hat{\Delta}^L\|_{\mathrm{F}}^2+\|\hat{\Delta}^S\|_{\mathrm{F}}^2+\|\hat{\Delta}^G\|_{\mathrm{F}}^2)\leq\left(\frac{3}{2}\lambda_N\sqrt{2r}\right)\|\hat{\Delta}^L\|_{\mathrm{F}}\\ +\left(\frac{5}{2}\mu_N\right)\sqrt{s}\|\hat{\Delta}^S\|_{\mathrm{F}}+2\nu_N\sqrt{g}\|\hat{\Delta}^G\|_{\mathrm{F}}+2\lambda_N\sum_{j=r+1}^{d}\sigma_j(L^*)\\ +2\mu_N\|S_{M^\perp}^*\|_1+2\nu_N\|G_{N^\perp}^*\|_{2,1}\end{aligned}$$

By our assumptions, we have

$$\begin{aligned}\frac{\zeta}{4}(\|\hat{\Delta}^L\|_F^2+\|\hat{\Delta}^S\|_F^2+\|\hat{\Delta}^G\|_F^2)\\ \leq\sqrt{\left(\frac{3}{2}\lambda_1\sqrt{2r}\right)^2+\left(\frac{5}{2}\lambda_2\sqrt{s}\right)^2+(2\lambda_3\sqrt{g})^2}\\ \sqrt{\|\hat{\Delta}^L\|_F^2+\|\hat{\Delta}^S\|_F^2+\|\hat{\Delta}^G\|_F^2}\end{aligned}$$

Combining with the inequality $\|\hat{\Delta}^L\|_F^2+\|\hat{\Delta}^S+\hat{\Delta}^G\|_F^2\leq 2(\|\hat{\Delta}^L\|_F^2+\|\hat{\Delta}^S\|_F^2+\|\hat{\Delta}^G\|_F^2)$, we conclude part (b) of Corollary 2. ■

APPENDIX B

Proof of Proposition 3:

1) We seek to find upper bounds on $\|\mathcal{X}'E/N\|_{\max}$, $\|\mathcal{X}'E/N\|$ and $\|\mathcal{X}'E/N\|_{2,\max}$ that hold with high probability. Note that such an upper bound for $\|\mathcal{X}'E/N\|_{\max}$ has already been derived in [8]. Here we adopt a different technique that takes a unified approach to provide upper bounds on both quantities. To this end, note that the two norms have the following representations

$$\frac{1}{N}\|\mathcal{X}'E\| = \sup_{u,v\in\mathbb{S}^{p-1}} \frac{1}{N}u'\mathcal{X}'Ev,$$

$$\frac{1}{N}\|\mathcal{X}'E\|_{\max} = \sup_{u,v\in\{e_1,\ldots,e_p\}} \frac{1}{N}u'\mathcal{X}'Ev$$

For any given $u, v \in \mathbb{S}^{p-1}$, we first provide a bound on $u'(\mathcal{X}'E/N)v$.
Using Proposition 2.4 of [8], we obtain

$$\mathbb{P}\left[|u'(\mathcal{X}'E/N)v| > 2\pi\eta\phi(B,\Sigma_\epsilon)\right] \le 6\exp\left[-cN\min\{\eta,\eta^2\}\right] \quad (20)$$

for any $u, v \in \mathbb{S}^{p-1}$ and any $\eta > 0$.
To derive the deviation bound on $\|\mathcal{X}'E/N\|_{\max}$, we simply take a union bound over the p^2 possible choices of $u, v \in \{e_1, e_2, \ldots, e_p\}$. This leads to

$$\mathbb{P}\left[\|\mathcal{X}'E/N\|_{\max} > 2\pi\eta\phi(B,\Sigma_\epsilon)\right] \le 6\exp\left[-cN\min\{\eta,\eta^2\} + 2\log p\right]$$

Since $N \gtrsim p$, we can set $\eta = \sqrt{(2+c_1)\log p/cN}$ so that $\eta < 1$ (i.e., $\eta^2 < \eta$) will be satisfied for large enough N. This implies that

$$\mathbb{P}\left[\|\mathcal{X}'E/N\|_{\max} > c_0\phi(B,\Sigma_\epsilon)\right] \le c_1\exp\left[-c_2\log p\right]$$

for some universal constants $c_i > 0$.
Similarly, for any group G_i of size m_i, we have

$$\mathbb{P}\left[\|vec(\mathcal{X}_r'E_s/N,\ (r,s)\in G_i)\| > 2\pi\sqrt{m_i}\eta\phi(B,\Sigma_\epsilon)\right] \le 6\exp\left[-cN\min\{\eta,\eta^2\} + \log m_i\right]. \quad (21)$$

Taking a union bound over K non-overlapping groups G_i leads to

$$\mathbb{P}\left[\|\mathcal{X}'E/N\|_{2,max} > 2\pi\sqrt{m}\eta\phi(B,\Sigma_\epsilon)\right] \le 6\exp\left[-cN\min\{\eta,\eta^2\} + \log p\right], \quad (22)$$

where $m = \max_{i=1,\ldots,K} m_i$. As before, setting $\eta = \sqrt{\log p/N}$ implies

$$\mathbb{P}\left[\|\mathcal{X}'E/N\|_{2,max} > 2\pi\sqrt{m\log p/N}\phi(B,\Sigma_\epsilon)\right] \le c_1\exp\left[-c_2\log p\right] \quad (23)$$

for some $c_i > 0$.
To derive the deviation bound on the spectral norm, we discretize the unit ball S^{p-1} using an ϵ-net $\mathcal{N}$ of cardinality at most $(1+2/\epsilon)^p$. An argument along the line of Supplementary Lemma F.2 in [8] then shows that for a small enough $\epsilon > 0$,

$$\sup_{u,v\in\mathbb{S}^{p-1}} |u'(\mathcal{X}'E/N)v| \le \kappa \sup_{u,v\in\mathcal{N}} |u'(\mathcal{X}'E/N)v|$$

for some constant $\kappa > 1$, possibly dependent on ϵ. As before, taking a union bound over the $(1+2/\epsilon)^{2p}$ choices of u and v, we get

$$\mathbb{P}\left[\|\mathcal{X}'E/N\| > 2\pi\kappa\eta\phi(B,\Sigma_\epsilon)\right] \le 6\exp\left[-cN\min\{\eta,\eta^2\} + 2p\log(1+2/\epsilon)\right]$$

Since $N \gtrsim p$, choosing $\eta = \sqrt{(c_1+2\log(1+2/\epsilon))p/cN}$ ensures $\eta < 1$ for large enough N. Setting η as above concludes the proof.

2) We want to obtain a lower bound on the minimum eigenvalue of $\mathcal{X}'\mathcal{X}/N$ that holds with high probability.
Since $\Lambda_{\min}(\mathcal{X}'\mathcal{X}/N) = \inf_{v\in\mathbb{S}^{p-1}} v'(\mathcal{X}'\mathcal{X}/N)v$, we start with the single deviation bound of Proposition 2.3 in [8],

$$\mathbb{P}\left[\left|v'\left(\mathcal{X}'\mathcal{X}/N - \Gamma_X(0)\right)v\right| > 2\pi\eta\mathcal{M}(f_X)\right] \le 2\exp\left[-cN\min\{\eta,\eta^2\}\right]$$

for any $v \in \mathbb{S}^{p-1}$ and $\eta > 0$.
The next step is to extend this single deviation bound uniformly on the set $\mathbb{S}^{p-1}$. As in the proof of part 1, we construct a ϵ-net of cardinality at most $(1+2/\epsilon)^p$ and approximate the quadratic form using its values on the net. This yields the following deviation bound

$$\mathbb{P}\left[\sup_{v\in\mathbb{S}^{p-1}}\left|v'\left(\frac{\mathcal{X}'\mathcal{X}}{N} - \Gamma_X(0)\right)v\right| > 2\kappa\pi\eta\mathcal{M}(f_X)\right] \le 2\exp\left[-cN\min\{\eta,\eta^2\} + p\log\left(1+\frac{2}{\epsilon}\right)\right]$$

for some constant $\kappa > 1$. Seting $\eta = \mathfrak{m}(f_X)/4\kappa\pi\mathcal{M}(f_X) < 1$ and noting that $N \gtrsim \mathcal{M}^2(f_X)/\mathfrak{m}^2(f_X)p$, we conclude

$$\mathbb{P}\left[\sup_{v\in\mathbb{S}^{p-1}}\left|v'\left(\mathcal{X}'\mathcal{X}/N - \Gamma_X(0)\right)v\right| > \mathfrak{m}(f_X)/2\right] \le c_0\exp\left[-c_1\log p\right]$$

The result follows from the lower bound on $\mathfrak{m}(f_X)$ presented in (6) and the fact that $v'\Gamma_X(0)v \ge \mathfrak{m}(f_X)$ for all $v \in \mathbb{S}^{p-1}$. ■

Proof of Proposition 4 and 5: Clearly, setting ζ to the lower bound on $\mathfrak{m}(f_X)$ as in (6) satisfies the RSC. Combining the estimates of Proposition 1 and 3 leads to Proposition 4 and 5 after simple algebraic computation. ■

APPENDIX C

In the following proof, we denote $\frac{1}{2}\|\mathcal{Y} - \mathcal{X}B\|_F^2$ and the regularization term by $H(B)$ and $P_B(B,\lambda)$, respectively.

Proof of Proposition 6: By the differentiability of H, we have

$$H(B_{i+1}^{ag}) = H(B_i^{md}) + \int_0^1 < \nabla H(B_i^{md} + \tau(B_{i+1}^{ag} - B_i^{md})), B_{i+1}^{ag} - B_i^{md} > d\tau \tag{24}$$

Then, by the definition of $H(B)$ and B_{i+1}^{ag}, and the relationship $B_{i+1}^{ag} - B_i^{md} = \alpha_i(B_{i+1} - B_i)$, we have

$$\begin{aligned}
l(B_{i+1}^{ag}) &= H(B_{i+1}^{ag}) + P_B(B_{i+1}^{ag}, \lambda) \\
&= H(B_i^{md}) + \int_0^1 \langle \mathcal{X}^T \mathcal{X}(B_i^{md} + \tau(B_{i+1}^{ag} - B_i^{md})) \\
&\quad - \mathcal{X}^T \mathcal{Y}, B_{i+1}^{ag} - B_i^{md} \rangle d\tau + P_B(B_{i+1}^{ag}, \lambda) \\
&= H(B_i^{md}) + \int_0^1 \langle \mathcal{X}^T(\mathcal{X} B_i^{md} - \mathcal{Y}), B_{i+1}^{ag} - B_i^{md} \rangle d\tau \\
&\quad + \int_0^1 \tau \|\mathcal{X}(B_{i+1}^{ag} - B_i^{md})\|^2 d\tau + P_B(B_{i+1}^{ag}, \lambda) \\
&= H(B_i^{md}) + \langle \nabla H(B_i^{md}), B_{i+1}^{ag} - B_i^{md} \rangle \\
&\quad + \frac{1}{2}\|\mathcal{X}(B_{i+1}^{ag} - B_i^{md})\|^2 + P_B(B_{i+1}^{ag}, \lambda) \\
&= H(B_i^{md}) + (1-\alpha_i)\langle \nabla H(B_i^{md}), B_i^{ag} - B_i^{md} \rangle \\
&\quad + \alpha_i \langle \nabla H(B_i^{md}), B_{i+1} - B_i^{md} \rangle \\
&\quad + \frac{\alpha_i^2}{2}\|\mathcal{X}(B_{i+1} - B_i)\|^2 \\
&\quad + (1-\alpha_i) P_B(B_i^{ag}, \lambda) + \alpha_i P_B(B_{i+1}, \lambda) \\
&= (1-\alpha_i)(H(B_i^{md}) + \langle \nabla H(B_i^{md}), B_i^{ag} - B_i^{md} \rangle \\
&\quad + P_B(B_i^{ag}, \lambda)) + \alpha_i (H(B_i^{md}) \\
&\quad + \langle \nabla H(B_i^{md}), B_{i+1} - B_i^{md} \rangle) \\
&\quad + \frac{\alpha_i^2}{2}\|\mathcal{X}(B_{i+1} - B_i)\|^2 + \alpha_i P_B(B_{i+1}, \lambda).
\end{aligned} \tag{25}$$

By the convexity of $H(B)$ and (25), we have

$$\begin{aligned}
l(B_{i+1}^{ag}) &= (1-\alpha_i)(H(B_i^{md}) + \langle \nabla H(B_i^{md}), \\
&\quad B_i^{ag} - B_i^{md} \rangle + P_B(B_i^{ag})) \\
&\quad + \alpha_i (H(B_i^{md}) + \langle \nabla H(B_i^{md}), B - B_i^{md} \rangle) \\
&\quad + \alpha_i \langle \nabla H(B_i^{md}), B_{i+1} - B \rangle \\
&\quad + \frac{\alpha_i^2}{2}\|\mathcal{X}(B_{i+1} - B_i)\|^2 + \alpha_i P_B(B_{i+1}, \lambda) \\
&\le (1-\alpha_i) L(B_i^{ag}) + \alpha_i L(B) \\
&\quad + \alpha_i \langle \nabla H(B_i^{md}), B_{i+1} - B \rangle \\
&\quad + \frac{\alpha_i^2}{2}\|\mathcal{X}(B_{i+1} - B_i)\|^2 \\
&\quad + \alpha_i P_B(B_{i+1}, \lambda) - \alpha_i P_B(B, \lambda)
\end{aligned} \tag{26}$$

Subtracting $l(B)$ from both sides of (26) and rearranging some terms, we have

$$\begin{aligned}
&[l(B_{i+1}^{ag}) - l(B)] - (1-\alpha_i)[l(B_i^{ag}) - l(B)] \\
&\le \alpha_i \langle \nabla H(B_i^{md}), B_{i+1} - B \rangle + \frac{\alpha_i^2}{2}\|\mathcal{X}(B_{i+1} - B_i)\|^2 \\
&\quad + \alpha_i \langle \xi, B_{i+1} - B \rangle
\end{aligned} \tag{27}$$

where $\xi \in \partial P_B(B_{i+1}, \lambda)$. On the other hand, by the first-order optimality conditions for the sequence B_{i+1} in Algorithm 1, we have

$$\begin{aligned}
&\langle \nabla H(B_i^{md}), B_{i+1}^e \rangle + \eta_i \langle B_{i+1} - B_i, B_{i+1}^e \rangle \\
&\quad + \langle \partial P_B(B_{i+1}, \lambda), B_{i+1} - B \rangle \le 0
\end{aligned} \tag{28}$$

Combining (27) and (28), we obtain

$$\begin{aligned}
&[l(B_{i+1}^{ag}) - l(B)] - (1-\alpha_i)[l(B_i^{ag}) - l(B)] \\
&\le \alpha_i \left\{ \eta_i \langle B_i - B_{i+1}, B_{i+1}^e \rangle + \frac{\alpha_i}{2}\|\mathcal{X}(B_{i+1} - B_i)\|^2 \right\} \\
&\le \alpha_i \left\{ \frac{\eta_i}{2}(\|B_i^e\|^2 - \|B_{i+1}^e\|^2 - \|B_{i+1} - B_i\|^2) \right. \\
&\quad \left. + \frac{\alpha_i}{2}\|\mathcal{X}(B_{i+1} - B_i)\|^2 \right\}
\end{aligned} \tag{29}$$

where we used the relationship $2\langle a-b, a-c \rangle = -\|b-c\|^2 + \|a-c\|^2 + \|a-b\|^2$ and the definition of B_{i+1}^e.

Dividing both sides of (29) by $\alpha_i \eta_i$, we have

$$\begin{aligned}
&\frac{1}{\alpha_i \eta_i}[l(B_{i+1}^{ag}) - l(B)] - \frac{(1-\alpha_i)}{\alpha_i \eta_i}[L(B_i^{ag}) - L(B)] \\
&\le \frac{1}{2}(\|B_i^e\|^2 - \|B_{i+1}^e\|^2) - \frac{1}{2}\Big(\|B_{i+1} - B_i\|^2 \\
&\quad - \frac{\alpha_i}{\eta_i}\|\mathcal{X}(B_{i+1} - B_i)\|^2\Big) \\
&\le \frac{1}{2}(\|B_i^e\|^2 - \|B_{i+1}^e\|^2) - \frac{1}{2}\Gamma_i
\end{aligned} \tag{30}$$

Adding $\frac{(\beta_i Q_i + \Gamma_i)}{2}$ to both sides of (30), we have

$$\begin{aligned}
&\frac{1}{\alpha_i \eta_i}[l(B_{i+1}^{ag}) - l(B)] - \frac{(1-\alpha_i)}{\alpha_i \eta_i}[l(B_i^{ag}) - l(B)] \\
&\quad + \frac{(\beta_i Q_i + \Gamma_i)}{2} \\
&\le \frac{1}{2}(\|B_i^e\|^2 - \|B_{i+1}^e\|^2) + \frac{(\beta_i - 1)Q_i}{2} + \frac{Q_i}{2}
\end{aligned} \tag{31}$$

Since $Q_{i+1} = \beta_i Q_i + \Gamma_i$, $0 \le \beta_i \le (1 - \frac{1}{i})^2$, and $Q_i \ge -\frac{C}{(i-1)^2}$, we obtain

$$\begin{aligned}
&\frac{1}{\alpha_i \eta_i}[l(B_{i+1}^{ag}) - l(B)] - \frac{(1-\alpha_i)}{\alpha_i \eta_i} l(B_i^{ag}) - l(B)] \\
&\quad + \frac{Q_{i+1}}{2} \\
&\le \frac{1}{2}(\|B_i^e\|^2 - \|B_{i+1}^e\|^2) + \frac{(1-\beta_i)C}{2(i-1)^2} + \frac{Q_i}{2}
\end{aligned} \tag{32}$$

Setting $B = \hat{B}$, by the relationship $\frac{1}{\alpha_i \eta_i} = \frac{1-\alpha_{i+1}}{\alpha_{i+1}\eta_{i+1}}$, and $\alpha_1 = 1$, we obtain

$$\begin{aligned} &\frac{1}{\alpha_i \eta_i}[l(B_{i+1}^{ag}) - l(B)] \\ &\quad \leq \frac{1}{2}\|B_0 - \hat{B}\|^2 + \sum_{i=2}^{k} \frac{(1-\beta_i)C}{(i-1)^2} + \frac{C}{k^2} \end{aligned} \tag{33}$$

after summing (32) from $i = 1$ to k.

Next we show the upper bound of $\alpha_k \eta_k$. Since $\eta_{\min} \leq \eta_{0,1}$, we have $\eta_{\min} \leq \|\mathcal{X}^T \mathcal{X}\|_2$. Then, by definition of $\eta_{0,i}$, we get

$$\eta_{\min} \leq \eta_{0,i} \leq ||\mathcal{X}^T \mathcal{X}||_2. \tag{34}$$

Denote $\sigma^l \eta_{0,i}$ by $\eta_i^{'}$, where l is the number of line search in Step 3 of Algorithm 1. By $\frac{1}{\alpha_i \eta_i} = \frac{1-\alpha_{i+1}}{\alpha_{i+1}\eta_{i+1}}$ and the definition of η_i, we have

$$\begin{aligned} \frac{1}{\alpha_i \sqrt{\eta_i^{'}}} &= \frac{\sqrt{1-\alpha_{i+1}}}{\alpha_{i+1}\sqrt{\eta_{i+1}^{'}}} \\ &\leq \frac{1}{\alpha_{i+1}\sqrt{\eta_{i+1}^{'}}} - \frac{1}{2\eta_{i+1}^{'}} \quad \text{for } i \geq 1 \end{aligned} \tag{35}$$

Then, by induction we can get, with $\alpha_1 = 1$,

$$\left(\frac{1}{\sqrt{\eta_1^{'}}} + \frac{1}{2}\sum_{i=2}^{k} \frac{1}{\sqrt{\eta_k^{'}}}\right)^2 \leq \frac{1}{\alpha_k^2 \eta_k^{'}}$$

which implies

$$\begin{aligned} \alpha_k \eta_k &\leq \frac{1}{\left(\frac{1}{\sqrt{\eta_1^{'}}} + \frac{1}{2}\sum_{i=2}^{k} \frac{1}{\sqrt{\eta_k^{'}}}\right)^2} \\ &\leq \frac{4\sigma||\mathcal{X}^T \mathcal{X}||_2}{(k+1)^2} \quad \text{for } k \geq 1 \end{aligned} \tag{36}$$

where we used (35) and the definition of $\eta_i^{'}$.

Combining (33) and (36), we obtain (13). ■

ACKNOWLEDGMENT

The authors would like to thank the Associate Editor and three anonymous referees for many constructive suggestions and comments.

REFERENCES

[1] A. L. Swindlehurst and P. Stoica, "Maximum likelihood methods in radar array signal processing," *Proc. IEEE*, vol. 86, no. 2, pp. 421–441, Feb. 1998.

[2] J. R. Roman, M. Rangaswamy, D. W. Davis, Q. Zhang, B. Himed, and J. Michels, "Parametric adaptive matched filter for airborne radar applications," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 36, no. 2, pp. 677–692, Apr. 2000.

[3] J. Lin and G. Michailidis, "Regularized estimation and testing for high-dimensional multi-block vector-autoregressive models," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 4188–4236, 2017.

[4] S. Basu, S. Das, G. Michailidis, and A. K. Purnanandam, *A System-Wide Approach to Measure Connectivity in the Financial Sector* (2017). [Online]. Available: http://dx.doi.org/10.2139/ssrn.2816137

[5] G. Michailidis and F. dAlché-Buc, "Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues," *Math. Biosciences*, vol. 246, no. 2, pp. 326–334, 2013.

[6] M. P. Van Den Heuvel and H. E. H. Pol, "Exploring the brain network: A review on resting-state fmri functional connectivity," *Eur. Neuropsychopharmacology*, vol. 20, no. 8, pp. 519–534, 2010.

[7] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*. Berlin, Germany: Springer-Verlag, 2005.

[8] S. Basu and G. Michailidis, "Regularized estimation in sparse high-dimensional time series models," *Ann. Statist.*, vol. 43, no. 4, pp. 1535–1567, 2015.

[9] Y. He, Y. She, and D. Wu, "Stationary-sparse causality network learning," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 3073–3104, 2013.

[10] Y. She, Y. He, S. Li, and D. Wu, "Joint association graph screening and decomposition for large-scale linear dynamical systems," *IEEE Trans. Signal Process.*, vol. 63, no. 2, pp. 389–401, Jan. 2015.

[11] J. He, "Sparse estimation based on square root nonconvex optimization in high-dimensional data," *Neurocomputing*, vol. 282, pp. 122–135, 2018.

[12] 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, 2012.

[13] K. M. Tan, P. London, K. Mohan, S.-I. Lee, M. Fazel, and D. M. Witten, "Learning graphical models with hub," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3297–3331, 2014.

[14] M. G. Sharaev, V. V. Zavyalova, V. L. Ushakov, S. I. Kartashov, and B. M. Velichkovsky, "Effective connectivity within the default mode network: Dynamic causal modeling of resting-state FMRI data," *Frontiers Human Neuroscience*, vol. 10, 2016.

[15] R. P. Velu, G. C. Reinsel, and D. W. Wichern, "Reduced rank models for multiple time series," *Biometrika*, vol. 73, no. 1, pp. 105–118, 1986.

[16] G. E. Box and G. C. Tiao, "A canonical analysis of multiple time series," *Biometrika*, vol. 64, no. 2, pp. 355–365, 1977.

[17] M. Fazel, H. Hindi, and S. P. Boyd, "Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices," in *Proc. Amer. Control Conf.*, 2003, vol. 3, pp. 2156–2162.

[18] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, "Rank-sparsity incoherence for matrix decomposition," *SIAM J. Optim.*, vol. 21, no. 2, pp. 572–596, 2011.

[19] 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, 2012. [Online]. Available: http://dx.doi.org/10.1214/12-AOS1000

[20] E. Yang and A. C. Lozano, "Sparse+ group-sparse dirty models: Statistical guarantees without unreasonable conditions and a case for non-convexity," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3911–3920.

[21] I. Melnyk and A. Banerjee, "Estimating structured vector autoregressive models," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 830–839.

[22] S. Basu, A. Shojaie, and G. Michailidis, "Network Granger causality with inherent grouping structure," *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 417–453, 2015.

[23] M. Fazel, "Matrix rank minimization with applications," Ph.D. dissertation, Stanford Univ., Stanford, CA, USA, 2002.

[24] X. Yuan and J. Yang, "Sparse and low-rank matrix decomposition via alternating direction methods," *Pacific J. Optim.*, vol. 9, pp. 167–180, 2009.

[25] J. Barzilai and J.M. Borwein, "Two-point step size gradient methods," *IMA J. Numer. Anal.*, vol. 8, no. 1, pp. 141–148, 1988.

[26] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky, "Latent variable graphical model selection via convex optimization," *Ann. Statist.*, vol. 40, no. 4, pp. 1935–1967, 2012.

[27] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Commun. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413–1457, 2004.

[28] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010

[29] Y. E. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*. Norwell, MA, USA: Kluwer, 2004.

[30] P. Tseng, "On accelerated proximal gradient methods for convex-concave optimization," submitted for publication.

[31] Y. Chen, X. Li, Y. Ouyang, and E. Pasiliao, "Accelerated Bregman operator splitting with backtracking," *Inverse Problems Imag.*, vol. 11, no. 6, pp. 1047–1070, 2017.

[32] Z. Lin, "Some software packages for partial svd computation," 2011, arXiv:1108.1548.

[33] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sciences*, vol. 2, no. 1, pp. 183–202, 2009.
[34] S. Ji and J. Ye, "An accelerated gradient method for trace norm minimization," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 457–464.
[35] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, 2010.
[36] S. Negahban, B. Yu, M. J. Wainwright, and P. K. Ravikumar, "A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers," in *Advances Neural Inf. Process. Syst.*, vol. 22, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Eds. Cambridge, MA, USA: MIT Press, 2009, pp. 1348–1356.
[37] M. Zorzi, and R. Sepulchre, "AR identification of latent-variable graphical models," *IEEE Trans. Autom. Control*, vol. 61, no. 9, pp. 2327–2340, Sep. 2016.
[38] M. Zorzi and A. Chiuso, "Sparse plus low rank network identification: A nonparametric approach," *Automatica*, 76, pp. 355–366, 2017.
[39] D. Giannone, M. Lenza, and G. E. Primiceri, "Prior selection for vector autoregressions," *Rev. Econ. Statist.*, vol. 97, no. 2, pp. 436–451, 2015.
[40] M. Babura, D. Giannone, and L. Reichlin, "Large Bayesian vector auto regressions," *J. Appl. Econometrics*, vol. 25, no. 1, pp. 71–92, 2010.
[41] N. J. Higham, "Computing the nearest correlation matrix—A problem from finance," *IMA J. Numer. Anal.*, vol. 22, no. 3, pp. 329–343, 2002.

Sumanta Basu received the B. Stat. and M. Stat. degrees from the Indian Statistical Institute, Kolkata, India, in 2006 and 2008, respectively, and the Ph.D. degree in statistics from the University of Michigan, Ann Arbor, MI, USA, in 2014. From 2014 to 2016, he was a Postdoctoral Scholar with the University of California, Berkeley, CA, USA, and the Lawrence Berkeley National Laboratory, Berkeley, CA, USA. Since 2016, he has been an Assistant Professor with the Department of Biological Statistics and Computational Biology and the Department of Statistical Science, Cornell University, Ithaca, NY, USA. His research interest is in developing statistical machine learning methods for structure learning and prediction of complex, high-dimensional systems arising in biological and social sciences.

Xianqi Li received the B.Sc. degree in information and computational mathematics from Liaoning University, Shenyang, China, in 2007, the M.Sc. degree in mathematics from the University of Texas-Pan American (now named as The University of Rio Grande Valley), Edinburg, TX, USA, in August 2009, and the M.Sc. degree in electrical and computer engineering and the Ph.D. degree in mathematics from the University of Florida, Gainesville, FL, USA, in May 2015 and May 2018, respectively. Since July 2018, he has been a Postdoctoral Researcher with the Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. His research interests include optimization, algorithms, and their applications to image analysis and network structure learning problems.

George Michalidis (M'05) received the Ph.D. degree in mathematics from the University of California, Los Angeles, CA, USA, in 1996. From 1996 to 1998, he was a Postdoctoral Fellow with the Department of Operations Research, Stanford University, Stanford, CA, USA. In 1998, he joined the University of Michigan, Ann Arbor, MI, USA, where he spent 17 years as a faculty member in statistics, electrical engineering, and computer science. He is currently with the University of Florida, Gainesville, FL, USA, where he is the Director of the Informatics Institute, and a faculty with the Departments of Statistics and Computer and Information Science and Engineering. His current research interests include analysis of high-dimensional data with network structure, stochastic network modeling, and performance evaluation, queuing analysis and congestion control, statistical modeling and analysis of Internet traffic and network tomography. Dr. Michailidis served as the Editor-in-Chief for the *Electronic Journal of Statistics* from 2013 to 2015 and serves on a number of editorial boards of other leading statistics journals. He served as Symposium Chair for the Support for Storage, Renewable Sources, and Microgrids Symposium for SmartGridComm 2012 and also as the the Secretary of the IEEE Subcommittee on SmartGrid Communications from 2009 to 2017. He is a Fellow of the Institute of Mathematical Statistics, the American Statistical Association, and the International Statistical Institute.