Journal of Machine Learning Research 18 (2017) 1-49 Submitted 1/17; Revised 8/17; Published 11/17

Regularized Estimation and Testing for High-Dimensional Multi-Block Vector-Autoregressive Models

Jiahe Lin JIAHELIN@UMICH.EDU
Department of Statistics
University of Michigan
Ann Arbor, MI 48109, USA

George Michailidis* GMICHAIL@UFL.EDU
Department of Statistics and the Informatics Institute
University of Florida
Gainesville, FL 32611, USA

Editor: Jie Peng

Abstract

Dynamical systems comprising of multiple components that can be partitioned into distinct *blocks* originate in many scientific areas. A pertinent example is the interactions between financial assets and selected macroeconomic indicators, which has been studied at *aggregate level*—e.g. a stock index and an employment index—extensively in the macroeconomics literature. A key shortcoming of this approach is that it ignores potential influences from other related components (e.g. Gross Domestic Product) that may impact the system's dynamics and structure and thus produces incorrect results. To mitigate this issue, we consider a multi-block linear dynamical system with Granger-causal ordering between blocks, wherein the blocks' temporal dynamics are described by vector autoregressive processes and are influenced by blocks higher in the system hierarchy. We derive the maximum likelihood estimator for the posited model for Gaussian data in the high-dimensional setting based on appropriate regularization schemes for the parameters of the block components. To optimize the underlying non-convex likelihood function, we develop an iterative algorithm with convergence guarantees. We establish theoretical properties of the maximum likelihood estimates, leveraging the decomposability of the regularizers and a careful analysis of the iterates. Finally, we develop testing procedures for the null hypothesis of whether a block "Granger-causes" another block of variables. The performance of the model and the testing procedures are evaluated on synthetic data, and illustrated on a data set involving log-returns of the US S&P100 component stocks and key macroeconomic variables for the 2001–16 period.

Keywords: Vector-autoregression; Stability; Block-coordinate descent; Consistency; Global testing

1. Introduction.

The study of linear dynamical systems has a long history in control theory (Kumar and Varaiya, 1986) and economics (Hansen and Sargent, 2013) due to their analytical tractability and ease to estimate their parameters. Such systems in their so-called *reduced form* give

∗. Corresponding Author. Post Address: 205 Griffin Floyd Hall, 1 University Ave, Gainesville, FL, 32611.

rise to Vector Autoregressive (VAR) models (Lütkepohl, 2005) that have been widely used in macroeconomic modeling for policy analysis (Sims, 1980, 1992), in financial econometrics (Gourieroux and Jasiak, 2001), and more recently in functional genomics (Shojaie et al., 2012), financial systemic risk analysis (Billio et al., 2012) and neuroscience (Seth, 2013).

In many applications, the components of the system under consideration can be naturally partitioned into *interacting blocks*. For example, Cushman and Zha (1997) studied the impact of monetary policy in a small open economy, where the economy under consideration is modeled as one block, while variables in other (foreign) economies as the other. Both blocks have their own autoregressive structure, and the inter-dependence between blocks is unidirectional: the foreign block *influences* the small open economy, but *not* the other way around, thus effectively introducing a *linear ordering* amongst blocks. Another example comes from the connection between the stock market and employment macroeconomic variables (Fitoussi et al., 2000; Phelps, 1999; Farmer, 2015) that focuses on the impact through a wealth effect mechanism of the former on the latter. Once again, the underlying hypothesis of interest is that the stock market influences employment, but not the other way around. In another application domain, molecular biologists conduct time course experiments on cell lines or animal models and collect data across multiple molecular compartments (transcripotmics, proteomics, metabolomics, lipidomics) in order to delineate mechanisms for disease onset and progression or to study basic biological processes. In this case, the interactions amongst the blocks (molecular compartments) are clearly delineated (transciptomic compartment influencing the proteomic and metabolomic ones), thus leading again to a linear ordering of the blocks.

The proposed model also encompasses the popular in marketing, regional science and growth theory VAR-X model, provided that the temporal evolution of the *exogenous* block of variables "X" exhibits autoregressive dynamics. For example, Nijs et al. (2001) examine the sensitivity of over 500 product prices to various marketing promotion strategies (the exogenous block), while Pauwels and Weiss (2008) examine changes in subscription rates, search engine referrals and marketing efforts of customers when switched from a free account to a fee-based structure, the latter together with customer characteristics representing the exogenous block. Pesaran et al. (2004) examine regional inter-dependencies, building a model where country specific macroeconomic indicators evolve according to a VAR model and they are influenced exogenously by key macroeconomic variables from neighboring countries/regions. Finally, Abeysinghe (2001) studies the impact of the price of oil on Gross Domestic Product growth rates for a number of countries, while controlling for other exogenous variables such as the country's consumption and investment expenditures along with its trade balance.

The proposed model gives rise to a network structure that in its most general form corresponds to a multi-partite graph, depicted in Figure 1 for 3 blocks, that exhibits a *directed acyclic structure* between the constituent blocks, and can also exhibit additional dependence between the nodes in each block. Selected properties of such multi-block structures, known as *chain graphs*, have been studied in the literature. Specifically, their maximum likelihood estimation for *independent and identically distributed* Gaussian data under a high-dimensional *sparse* regime is thoroughly investigated in Lin et al. (2016), where a provably convergent estimation procedure is introduced and its theoretical properties are established.

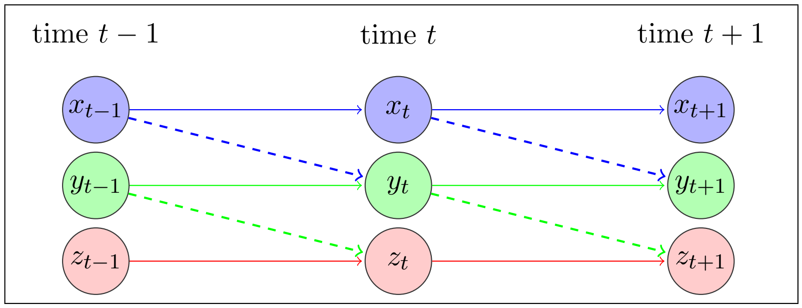


Figure 1: Diagram for a dynamic system with three groups of variables

Given the wide range of applications of multi-block VAR models, which in addition encompass the widely used VAR-X model, the key contributions of the current paper are fourfold: (i) formulating the model as a recursive dynamical system and examining its stability properties; (ii) developing a provably convergent algorithm for obtaining the regularized maximum likelihood estimates (MLE) of the model parameters under high-dimensional scaling; (iii) establishing theoretical properties of the ML estimates; and (iv) devising a testing procedure for the parameters that connect the constituent blocks of the model: if the null hypothesis is not rejected, then one is dealing with a set of independently evolving VAR models, otherwise with the posited multi-block VAR model. Finally, the model, estimation and testing procedures are illustrated on an important problem in macroeconomics, as gleaned by the background of the problem and discussion of the results provided in Section 6.

For the multi-block VAR model, we assume that the time series within each block are generated by a Gaussian VAR process. Further, the transition matrices within and across blocks can be either *sparse* or *low rank*. The posited regularized Gaussian likelihood function is not *jointly convex* in all the model parameters, which poses a number of technical challenges that are compounded by the presence of temporal dependence. These are successfully addressed and resolved in Section 3, where we provide a numerically convergent algorithm and establish the theoretical properties of the resulting ML estimates, that constitutes a key contribution in the study of multi-block VAR models.

The remainder of this paper is organized as follows. In Section 2, we introduce the model setup and the corresponding estimation procedure. In Section 3, we provide consistency properties of the obtained ML estimates under a high-dimensional scaling. In Section 4, we introduce the proposed testing framework, both for low-rank and sparse interaction matrices between the blocks. Section 5 contains selected numerical results that assess the performance of the estimation and testing procedures. Finally, an application to financial and macroeconomic data that was previously discussed as motivation for the model under consideration is presented in Section 6.

Notation. Throughout this paper, we use $|||A|||_1$ and $|||A|||_\infty$ respectively to denote the matrix induced 1-norm and infinity norm of $A \in \mathbb{R}^{m\times n}$, that is, $|||A|||_1 = \max_{1\le j\le n}\sum_{i=1}^m |a_{ij}|$, $|||A|||_\infty = \max_{1\le i\le m}\sum_{j=1}^n |a_{ij}|$, and use $\|A\|_1$ and $\|A\|_\infty$ respectively to denote the element-wise 1-norm and infinity norm: $\|A\|_1 = \sum_{i,j}|a_{ij}|$, $\|A\|_\infty = \max_{i,j}|a_{ij}|$. Moreover, we use $|||A|||_*$, $|||A|||_F$ and $|||A|||_{op}$ to denote the nuclear, Frobenius and operator norms of A, respectively. For two matrices A and B of commensurate dimensions, denote their inner product by $\langle\!\langle A, B \rangle\!\rangle = \mathrm{trace}(A'B)$. Finally, we write $A \gtrsim B$ if there exists some absolute constant c that is independent of the model parameters such that $A \geq cB$.

2. Problem Formulation.

To convey the main ideas and the key technical contributions, we consider a recursive linear dynamical system comprising of two blocks of variables, whose structure is given by:

$$\begin{aligned} X_t &= AX_{t-1} + U_t, \\ Z_t &= BX_{t-1} + CZ_{t-1} + V_t, \end{aligned} \tag{1}$$

where $X_t \in \mathbb{R}^{p_1}, Z_t \in \mathbb{R}^{p_2}$ are the variables in groups 1 and 2, respectively. The temporal intra-block dependence is captured by transition matrices A and C, while the inter-block dependence by B. Noise processes $\{U_t\}$ and $\{V_t\}$, respectively, capture additional contemporaneous intra-block dependence of X_t and Z_t, after conditioning on their respective past values. Further, we assume that U_t and V_t follow mean zero Gaussian distributions with covariance matrices given by Σ_u and Σ_v, i.e.,

$$U_t \sim \mathcal{N}(0, \Sigma_u), \quad \text{and} \quad V_t \sim \mathcal{N}(0, \Sigma_v).$$

With the above model setup, the parameters of interest are transition matrices $A \in \mathbb{R}^{p_1 \times p_1}$, $B \in \mathbb{R}^{p_2 \times p_1}$ and $C \in \mathbb{R}^{p_2 \times p_2}$, as well as the covariances Σ_u, Σ_v. In high-dimensional settings, different combinations of structural assumptions can be imposed on these transition matrices to enable their estimation from limited time series data. In particular, the intra-block transition matrices A and C are sparse, while the inter-block matrix B can be either sparse or low rank. Note that the block of X_t variables acts as an *exogenous* effect to the evolution of the Z_t block (e.g., Cushman and Zha, 1997; Nicolson et al., 2016). Further, we assume $\Omega_u := \Sigma_u^{-1}$ and $\Omega_v := \Sigma_v^{-1}$ are sparse.

Remark 1 *For ease of exposition, we posit a* VAR(1) *modeling structure. Extensions to general multi-block structures akin to the one depicted in Figure 1 and* VAR(d) *specifications are rather straightforward and briefly discussed in Section 7.*

The triangular (recursive) structure of the system enables a certain degree of separability between X_t and Z_t. In the posited model, X_t is a stand-alone VAR(1) process, and the time series in block Z_t is "Granger-caused" by that in block X_t, but not vice versa. The second equation in (1), as mentioned in the introductory section, also corresponds to the so-called "VAR-X" model in the econometrics literature (e.g., Sims, 1980; Bianchi et al., 2010), that extends the standard VAR model to include influences from lagged values of *exogenous* variables. Consider the joint process $W_t = (X_t', Z_t')'$, it corresponds to a VAR(1) model whose transition matrix G has a block triangular form:

$$W_t = GW_{t-1} + \varepsilon_t, \qquad \text{where} \quad G := \begin{bmatrix} A & O \\ B & C \end{bmatrix}, \quad \varepsilon_t = \begin{bmatrix} U_t \\ V_t \end{bmatrix}. \tag{2}$$

The model in (2) can also be viewed from a Structural Equations Modeling viewpoint involving time series data, and also has a Moving Average representation corresponding to a structural VAR representation with Granger causal ordering (Lütkepohl, 2005). As mentioned in the introductory section, the focus of this paper is model parameter estimation under high-dimensional scaling, rather than their cause and effect relationship. For a comprehensive discourse of causality issues for VAR models, we refer the readers to Granger (1969); Lütkepohl (2005), and references therein.

Next, we introduce the notion of stability and spectrum with respect to the system.

System Stability. To ensure that the joint process $\{W_t\}$ is stable (Lütkepohl, 2005), we require the spectral radius, denoted by $\rho(\cdot)$, of the transition matrix G to be smaller than 1, which is guaranteed by requiring that $\rho(A) < 1$ and $\rho(C) < 1$, since

$$|\lambda \mathrm{I}_{p_1 \times p_2} - G| = \begin{vmatrix} \lambda \mathrm{I}_{p_1} - A & O \\ -B & \lambda \mathrm{I}_{p_2} - C \end{vmatrix} = |\lambda \mathrm{I}_{p_1} - A||\lambda \mathrm{I}_{p_2} - C|,$$

implying that the set of eigenvalues of G is the union of the sets of eigenvalues of A and C, hence

$$\rho(A) < 1 \ , \rho(C) < 1, \qquad \Rightarrow \quad \rho(G) = \max\{\rho(A), \rho(C)\} < 1.$$

The latter relation implies that the stability of such a recursive system imposes spectrum constraints only on the diagonal blocks that govern the intra-block evolution, whereas the off-diagonal block that governs the inter-block interaction is left unrestricted.

Spectrum of the joint process. Throughout, we assume that the spectral density of $\{W_t\}$ exists, which then possesses a special structure as a result of the block triangular transition matrix G. Formally, we define the spectral density of $\{W_t\}$ as

$$f_W(\theta) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \Gamma_W(h) e^{-ih\theta}, \qquad \theta \in [-\pi, \pi],$$

where $\Gamma_W(h) := \mathbb{E} W_t W'_{t+h}$. For two (generic) processes $\{X_t\}$ and $\{Z_t\}$, define their cross-covariance as $\Gamma_{X,Z}(h) = \mathbb{E} X_t Z'_{t+h}$ and $\Gamma_{Z,X}(h) = \mathbb{E} Z_t X'_{t+h}$. In general, $\Gamma_{X,Z}(h) \neq \Gamma_{Z,X}(h)$. The cross-spectra are defined as:

$$f_{X,Z}(\theta) := \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \Gamma_{X,Z}(h) e^{-ih\theta}, \quad \text{and} \quad f_{Z,X}(\theta) := \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \Gamma_{Z,X}(h) e^{-ih\theta}, \ \ \theta \in [-\pi, \pi].$$

For the model given in (2), by writing out the dynamics of Z_t, the cross-spectra between X_t and Z_t are given by

$$f_{X,Z}(\theta)(\mathrm{I}_{p_2} - C' e^{-i\theta}) = f_X(\theta) B' e^{-i\theta}, \qquad \text{and} \qquad (\mathrm{I}_{p_2} - C e^{i\theta}) f_{Z,X}(\theta) = B e^{i\theta} f_X(\theta). \quad (3)$$

Similarly, we have

$$(\mathrm{I}_{p_2} - C e^{i\theta}) f_Z(\theta) = B e^{i\theta} f_{X,Z}(\theta) + f_{V,Z}(\theta). \qquad (4)$$

Combining (3) and (4), the spectrum of the joint process W_t is given by

$$f_W(\theta) = \left[H_1(e^{i\theta})\right]^{-1} \Big(\left[H_2(e^{i\theta})\right] \left[\mathbf{1}_{2\times 2} \otimes f_X(\theta)\right] \left[H_2(e^{-i\theta})\right]^\top + \begin{bmatrix} O & O \\ O & \Sigma_v \end{bmatrix} \Big) \left[H_1(e^{-i\theta})\right]^{-\top}, \qquad (5)$$

where $\mathbf{1}_{2\times 2}$ is a 2×2 matrix with all entries being 1, and

$$H_1(x) := \begin{bmatrix} \mathrm{I}_{p_1} & O \\ O & \mathrm{I}_{p_2} - Cx \end{bmatrix} \in \mathbb{R}^{(p_1+p_2)\times(p_1+p_2)}, \quad H_2(x) := \begin{bmatrix} \mathrm{I}_{p_1} & O \\ O & Bx \end{bmatrix} \in \mathbb{R}^{(p_1+p_2)\times(2p_1)}.$$

Equation (5) implies that the spectrum of the joint process $\{W_t\}$ can be decomposed into the sum of two parts: the first, is a function of $f_X(\theta)$, while the second part involves the idiosyncratic error process $\{V_t\}$, which only affects the right-bottom block of the spectrum. Note that since $\{W_t\}$ is a VAR(1) process, its matrix-valued characteristic polynomial is given by

$$\mathcal{G}(\theta) := \mathrm{I}_{(p_1+p_2)} - G\theta,$$

and its spectral density also takes the following form (c.f. Hannan, 1970):

$$f_W(\theta) = \frac{1}{2\pi}\big[\mathcal{G}^{-1}(e^{i\theta})\big]\Sigma_\varepsilon\big[\mathcal{G}^{-1}(e^{i\theta})\big]^*,$$

where

$$\mathcal{G}(x) = \begin{bmatrix} \mathrm{I}_{p_1} - Ax & O \\ -Bx & \mathrm{I}_{p_2} - Cx \end{bmatrix}, \qquad \Sigma_\varepsilon = \begin{bmatrix} \Sigma_u & O \\ O & \Sigma_v \end{bmatrix},$$

and $\mathcal{G}^*$ is the conjugate transpose. One can easily reach the same conclusion as in (5) by multiplying each term, followed by some algebraic manipulations.

2.1 Estimation.

Next, we outline the algorithm for obtaining the ML estimates of the transition matrices A, B and C and inverse covariance matrices Σ_u^{-1} and Σ_v^{-1} from time series data. We allow for a potential high-dimensional setting, where the ambient dimensions p_1 and p_2 of the model exceed the total number of observations T.

Given centered times series data $\{x_0, \cdots, x_T\}$ and $\{z_0, \cdots, z_T\}$, we use $\mathcal{X}^T$ and $\mathcal{Z}^T$ respectively, to denote the "response" matrix from time 1 to T, that is:

$$\mathcal{X}^T = \begin{bmatrix} x_1 & x_2 & \dots & x_T \end{bmatrix}' \quad \text{and} \quad \mathcal{Z}^T = \begin{bmatrix} z_1 & z_2 & \dots & z_T \end{bmatrix}',$$

and use $\mathcal{X}$ and $\mathcal{Z}$ without the superscript to denote the "design" matrix from time 0 to $T-1$:

$$\mathcal{X} = \begin{bmatrix} x_0 & x_1 & \dots & x_{T-1} \end{bmatrix}' \quad \text{and} \quad \mathcal{Z} = \begin{bmatrix} z_0 & z_1 & \dots & z_{T-1} \end{bmatrix}'.$$

We use $\mathcal{U}$ and $\mathcal{V}$ to denote the error matrices. To obtain estimates for the parameters of interest, we formulate optimization problems using a penalized log-likelihood function, with regularization terms corresponding to the imposed structural assumptions on the model parameters–sparsity and/or low-rankness. To solve the optimization problems, we employ block-coordinate descent algorithms, akin to those described in Lin et al. (2016), to obtain the solution.

As previously mentioned, $\{X_t\}$ is not "Granger-caused" by Z_t and hence it is a stand-alone VAR(1) process; this enables us to separately estimate the parameters governing the X_t process (A and Σ_u^{-1}) from those of the Z_t process (B, C, and Σ_v^{-1}).

Estimation of A and Σ_u^{-1}. Conditional on the initial observation x_0, the likelihood of $\{x_t\}_{t=1}^T$ is given by:

$$\begin{aligned} L(x_T, x_{T-1}, \cdots, x_1|x_0) &= L(x_T|x_{T-1}, \cdots, x_0)L(x_{T-1}|x_{T-2}, \cdots, x_0)\cdots L(x_1|x_0) \\ &= L(x_T|x_{T-1})L(x_{T-1}|x_{T-2})\cdots L(x_1|x_0), \end{aligned}$$

where the second equality follows from the Markov property of the process. The log-likelihood function is given by:

$$\ell(A, \Sigma_u^{-1}) = \frac{T}{2}\log\det(\Sigma_u^{-1}) - \frac{1}{2}\sum_{t=1}^{T}(x_t - Ax_{t-1})'\Sigma_u^{-1}(x_t - Ax_{t-1}) + \text{constant}.$$

Letting $\Omega_u := \Sigma_u^{-1}$, then the penalized maximum likelihood estimator can be written as

$$(\widehat{A}, \widehat{\Omega}_u) = \underset{\substack{A\in\mathbb{R}^{p_1\times p_2} \\ \Omega_u\in\mathbb{S}^{++}_{p_1\times p_1}}}{\arg\min} \Big\{\mathrm{tr}\big[\Omega_u(\mathcal{X}^T - \mathcal{X}A')'(\mathcal{X}^T - \mathcal{X}A')/T\big] - \log\det\Omega_u + \lambda_A\|A\|_1 + \rho_u\|\Omega_u\|_{1,\text{off}}\Big\}. \quad (6)$$

Algorithm 1 describes the key steps for obtaining $\widehat{A}$ and $\widehat{\Omega}_u$.

Algorithm 1: Computational procedure for estimating A and Σ_u^{-1}.

Input: Time series data $\{x_t\}_{t=1}^T$, tuning parameter λ_A and ρ_u.

1 **Initialization:** Initialize with $\widehat{\Omega}_u^{(0)} = \mathrm{I}_{p_1}$, then

$$\widehat{A}^{(0)} = \arg\min_A\big\{\tfrac{1}{T}|||\mathcal{X}^T - \mathcal{X}A'|||_F^2 + \lambda_A\|A\|_1\big\};$$

2 **Iterate until convergence:**

– Update $\widehat{\Omega}_u^{(k)}$ by graphical Lasso (Friedman et al., 2008) on the residuals with the plug-in estimate $\widehat{A}^{(k)}$;

– Update $\widehat{A}^{(k)}$ with the plug-in $\widehat{\Omega}_u^{(k-1)}$ and cyclically update each row with a Lasso penalty, which solves

$$\min_A\big\{\tfrac{1}{T}\mathrm{tr}\big[\widehat{\Omega}_u^{(k-1)}(\mathcal{X}^T - \mathcal{X}A)'(\mathcal{X}^T - \mathcal{X}A)/T\big] + \lambda_A\|A\|_1\big\}. \quad (7)$$

Output: Estimated sparse transition matrix $\widehat{A}$ and sparse $\widehat{\Omega}_u$.

Estimation of B, C and Σ_v^{-1}. Similarly, to obtain estimates of B, C and $\Omega_v := \Sigma_v^{-1}$, we formulate the optimization problem as follows:

$$\begin{aligned} (\widehat{B}, \widehat{C}, \widehat{\Omega}_v) = \underset{\substack{B\in\mathbb{R}^{p_2\times p_1}, C\in\mathbb{R}^{p_2\times p_2} \\ \Omega_v\in\mathbb{S}^{++}_{p_2\times p_2}}}{\arg\min} &\Big\{\mathrm{tr}\big[\Omega_v(\mathcal{Z}^T - \mathcal{X}B' - \mathcal{Z}C')'(\mathcal{Z}^T - \mathcal{X}B' - \mathcal{Z}C')/T\big] - \log\det\Omega_v \\ &+ \lambda_B\mathcal{R}(B) + \lambda_C\|C\|_1 + \rho_v\|\Omega_v\|_{1,\text{off}}\Big\}, \end{aligned} \quad (8)$$

where the regularizer $\mathcal{R}(B) = \|B\|_1$ if B is assumed to be sparse, and $\mathcal{R}(B) = |||B|||_*$ if B is assumed to be low rank. Algorithm 2 outlines the procedure for obtaining estimates $\widehat{B}$, $\widehat{C}$

Algorithm 2: Computational procedure for estimating B, C and Σ_v^{-1}.

Input: Time series data $\{x_t\}_{t=1}^T$ and $\{z_t\}_{t=1}^T$, tuning parameters λ_B, λ_C, ρ_v.

1 **Initialization:** Initialize with $\widehat{\Omega}_v^{(0)} = \mathrm{I}_{p_2}$, then

$$(\widehat{B}^{(0)}, \widehat{C}^{(0)}) = \arg\min_{(B,C)} \big\{ \tfrac{1}{T} \big|\big|\big| \mathcal{Z}^T - \mathcal{X}B' - \mathcal{Z}C' \big|\big|\big|_F^2 + \lambda_B \mathcal{R}(B) + \lambda_C \|C\|_1 \big\}.$$

2 **Iterate until convergence:**

– Update $\widehat{\Omega}_v^{(k)}$ by graphical Lasso on the residuals with the plug-in estimates $\widehat{B}^{(k)}$ and $\widehat{C}^{(k)}$;

– For fixed $\widehat{\Omega}_v^{(k)}$, $(\widehat{B}^{(k+1)}, \widehat{C}^{(k+1)})$ solves

$$\min_{B,C} \big\{ \tfrac{1}{T} \mathrm{tr}\big[\widehat{\Omega}_v^{(k)} (\mathcal{Z}^T - \mathcal{X}B' - \mathcal{Z}C')'(\mathcal{Z}^T - \mathcal{X}B' - \mathcal{Z}C') \big] + \lambda_B \mathcal{R}(B) + \lambda_C \|C\|_1 \big\}.$$

· Fix $\widehat{C}^{[s]}$, update $\widehat{B}^{[s+1]}$ by Lasso or singular value thresholding, which solves

$$\min_{B} \big\{ \tfrac{1}{T} \mathrm{tr}\big[\widehat{\Omega}_v^{(k)} (\mathcal{Z}^T - \mathcal{X}B' - \mathcal{Z}\widehat{C}^{[s]'})'(\mathcal{Z}^T - \mathcal{X}B' - \mathcal{Z}\widehat{C}^{[s]'}) \big] + \lambda_B \mathcal{R}(B) \big\};$$

· Fix $\widehat{B}^{[s]}$, update $\widehat{C}^{[s]}$ by Lasso, which solves

$$\min_{C} \big\{ \tfrac{1}{T} \mathrm{tr}\big[\widehat{\Omega}_v^{(k)} (\mathcal{Z}^T - \mathcal{X}\widehat{B}^{[s]'} - \mathcal{Z}C')'(\mathcal{Z}^T - \mathcal{X}\widehat{B}^{[s]'} - \mathcal{Z}C') \big] + \lambda_C \|C\|_1 \big\}.$$

Output: Estimated transition matrices $\widehat{B}$, $\widehat{C}$ and sparse $\widehat{\Omega}_v$.

and $\widehat{\Omega}_v$. Note that $\widehat{B}^{(k)}$ and $\widehat{C}^{(k)}$ need to be treated as a "joint block" in the outer update and convergence of the "joint block" is required before moving on to updating Ω_v.

Note that the objective function in (6) is not *jointly convex* in both parameters, but *biconvex*. Similarly in (8), the objective function is biconvex in $\big[(B, C), \Omega_v\big]$. Consequently, convergence to a stationary point is guaranteed, as long as estimates from all iterations lie within a ball around the true value of the parameters, with the radius of the ball upper bounded by a universal constant that only depends on model dimensions and sample size (Lin et al., 2016, Theorem 4.1). This condition is satisfied upon the establishment of consistency properties of the estimates.

To establish consistency properties of the estimates requires the existence of good initial values for the model parameters (A, Ω_u), and (B, C, Ω_v), respectively, in the sense that they are sufficiently close to the true parameters. For the (A, Ω_u) parameters, the results in Basu and Michailidis (2015) guarantee that for random realizations of $\{X_t, E_t\}$, with sufficiently large sample size, the errors of $\widehat{A}^{(0)}$ and $\widehat{\Omega}_u^{(0)}$ are bounded with high probability, which provides us with good initialization values. Yet, additional handling of the bounds is required to ensure that estimates from subsequent iterations are also uniformly close to the true value (see Section 3.2 Theorem 5). A similar property for $(\widehat{B}^{(0)}, \widehat{C}^{(0)}, \widehat{\Omega}_v^{(0)})$ and subsequent iterations is established in Section 3.2 Theorem 6 (see also Theorem 15 in Appendix A).

3. Theoretical Properties.

In this section, we investigate the theoretical properties of the penalized maximum likelihood estimation procedure proposed in Section 2, with an emphasis on the error bounds for the obtained estimates. We focus on the model specification in which the inter-block transition matrix B is *low rank*, which is of interest in many applied settings. Specifically, we consider the consistency properties of $\widehat{A}$ and $(\widehat{B},\widehat{C})$ that are solutions to the following two optimization problems:

$$(\widehat{A},\widehat{\Omega}_u) = \arg\min_{A,\Omega_u}\Big\{\mathrm{tr}\left[\Omega_u(\mathcal{X}^T-\mathcal{X}A')'(\mathcal{X}^T-\mathcal{X}A')/T\right] - \log\det\Omega_u + \lambda_A\|A\|_1 + \rho_u\|\Omega_u\|_{1,\mathrm{off}}\Big\}, \quad (9)$$

and

$$\begin{aligned}(\widehat{B},\widehat{C},\widehat{\Omega}_v) = \arg\min_{B,C,\Omega_v}\Big\{&\mathrm{tr}\big[\Omega_v(\mathcal{Z}^T-\mathcal{X}B'-\mathcal{Z}C')'(\mathcal{Z}^T-\mathcal{X}B'-\mathcal{Z}C')/T\big] - \log\det\Omega_v \\ &+\lambda_B|||B|||_* + \lambda_C\|C\|_1 + \rho_v\|\Omega_v\|_{1,\mathrm{off}}\Big\}.\end{aligned} \quad (10)$$

The case of a sparse B can be handled similarly to that of A and/or C with minor modifications (details shown in Supplementary Material **??**).

3.1 A road map for establishing the consistency results.

Next, we outline the main steps followed in establishing the theoretical properties for the model parameters. Throughout, we denote with a superscript "$\star$" the true value of the corresponding parameters.

The following key concepts, widely used in high-dimensional regularized estimation problems, are needed in subsequent developments.

Definition 2 (Restricted Strong Convexity (RSC)) *For some generic operator* $\mathfrak{X}:\mathbb{R}^{m_1\times m_2}\mapsto\mathbb{R}^{T\times m_1}$*, it satisfies the RSC condition with respect to norm* Φ *with curvature* $\alpha_{RSC}>0$ *and tolerance* $\tau>0$ *if*

$$\frac{1}{2T}|||\mathfrak{X}(\Delta)|||_F^2 \geq \alpha_{RSC}|||\Delta|||_F^2 - \tau\Phi^2(\Delta), \qquad \text{for some } \Delta\in\mathbb{R}^{m_1\times m_2}.$$

Note that the choice of the norm Φ is context specific. For example, in sparse regression problems, $\Phi(\Delta)=\|\Delta\|_1$ corresponds to the element-wise ℓ_1 norm of the matrix (or the usual vector ℓ_1 norm for the vectorized version). The RSC condition becomes equivalent to the *restricted eigenvalue (RE) condition* (see Loh and Wainwright, 2012; Basu and Michailidis, 2015, and references therein). This is the case for the problem of estimating transition matrix A. For estimating B and C, define $\mathcal{Q}$ to be the weighted regularizer $\mathcal{Q}(B,C):=|||B|||_*+\frac{\lambda_C}{\lambda_B}\|C\|_1$, and the associated norm Φ in this setting is defined as

$$\Phi(\Delta) := \inf_{B_{\mathrm{aug}}+C_{\mathrm{aug}}=\Delta}\mathcal{Q}(B,C).$$

Definition 3 (Diagonal dominance) *A matrix* $\Omega\in\mathbb{R}^{p\times p}$ *is strictly diagonally dominant if*

$$|\Omega_{ii}| > \sum_{j\neq i}|\Omega_{ij}|, \quad \forall\ \ i=1,\cdots,p.$$

Definition 4 (Incoherence condition (Ravikumar et al., 2011)) *A matrix* $\Omega\in\mathbb{R}^{p\times p}$ *satisfies the incoherence condition if:*

$$\max_{e\in(S_\Omega)^c}\|H_{eS_\Omega}(H_{S_\Omega S_\Omega})^{-1}\|_1 \leq 1-\xi, \quad \text{for some } \xi\in(0,1),$$

where $H_{S_\Omega S_\Omega}$ denotes the Hessian of the log-determinant barrier $\log\det\Omega$ *restricted to the true edge set of* Ω *denoted by* S_Ω*, and* H_{eS} *is similarly defined.*

The above two conditions are associated with the inverse covariance matrices Ω_u and Ω_v. Specifically, the diagonal dominance condition is required for $\Omega_u^\star$ and $\Omega_v^\star$ as we build the consistency properties for $\widehat{A}$ and $(\widehat{B},\widehat{C})$ with the penalized maximum likelihood formulation. The incoherence condition is primarily required for establishing the consistency of $\widehat{\Omega}_u$ and $\widehat{\Omega}_v$.

We additionally introduce the upper and lower extremes of the spectrum, defined as

$$\mathcal{M}(f_X) := \operatorname*{esssup}_{\theta\in[-\pi,\pi]} \Lambda_{\max}(f_X(\theta)) \qquad \text{and} \qquad \mathfrak{m}(f_X) := \operatorname*{essinf}_{\theta\in[-\pi,\pi]} \Lambda_{\min}(f_X(\theta)).$$

Analogously, the upper extreme for the cross-spectrum is given by:

$$\mathcal{M}(f_{X,Z}) := \operatorname*{esssup}_{\theta\in[-\pi,\pi]} \sqrt{\Lambda_{\max}(f^*_{X,Z}(\theta) f_{X,Z}(\theta))},$$

with $f^*_{X,Z}(\theta)$ being the conjugate transpose of $f_{X,Z}(\theta)$. With this definition,

$$\mathcal{M}(f_{X,Z}) = \mathcal{M}(f_{Z,X}).$$

Next, consider the solution to (9) that is obtained by the alternate update between A and Ω_u. If Ω_u is held fixed, then A solves (11), and we denote the solution by $\bar{A}$ and its corresponding vectorized version as $\bar{\beta}_A := \operatorname{vec}(\bar{A})$:

$$\bar{\beta}_A := \operatorname*{arg\,min}_{\beta\in\mathbb{R}^{p_1^2}} \big\{ -2\beta'\gamma_X + \beta'\Gamma_X\beta + \lambda_A\|\beta\|_1 \big\}, \tag{11}$$

where

$$\Gamma_X = \Omega_u \otimes \tfrac{\mathcal{X}'\mathcal{X}}{T}, \qquad \gamma_X = \tfrac{1}{T}\left(\Omega_u \otimes \mathcal{X}'\right)\operatorname{vec}(\mathcal{X}^T). \tag{12}$$

Using a similar notation, if A is held fixed, then Ω_u solves (13) with the solution being $\bar{\Omega}_u$:

$$\bar{\Omega}_u := \operatorname*{arg\,min}_{\Theta\in\mathbb{S}^{++}_{p_1\times p_1}} \big\{ \log\det\Omega_u - \operatorname{trace}\left(S_u\Omega_u\right) + \rho_u\|\Omega_u\|_{1,\text{off}} \big\}, \tag{13}$$

where

$$S_u = \tfrac{1}{T}(\mathcal{X}^T - \mathcal{X}A')'(\mathcal{X}^T - \mathcal{X}A'). \tag{14}$$

For *fixed* realizations of $\mathcal{X}$ and $\mathcal{U}$, by Basu and Michailidis (2015), the error bound of $\bar{\beta}_A$ relies on (1) Γ_X satisfying the RSC condition defined above; and (2) the tuning parameter λ_A is chosen in accordance with a deviation bound condition associated with $\|\mathcal{X}'\mathcal{U}\Omega_u/T\|_\infty$. By Ravikumar et al. (2011), the error bound of $\bar{\Omega}_u$ relies on how well S_u concentrates around $\Sigma_u^\star$, that is, $\|S_u - \Sigma_u^\star\|_\infty$. Specifically, for (12) and (14), with $\Omega_u^\star$ and $A^\star$ plugged in respectively, for *random* realizations of $\mathcal{X}$ and $\mathcal{U}$, these conditions hold with high probability. In the actual implementation of the algorithm, however, quantities in (12) and (14) are substituted by estimates so that at iteration k, $\widehat{\beta}_A^{(k)}$ and $\widehat{\Omega}_u^{(k)}$ solve

$$\begin{aligned}
\widehat{\beta}_A^{(k)} &:= \operatorname*{arg\,min}_{\beta\in\mathbb{R}^{p_1^2}} \big\{ -2\beta'\widehat{\gamma}_X^{(k)} + \beta'\widehat{\Gamma}_X^{(k)}\beta + \lambda_A\|\beta\|_1 \big\}, \\
\widehat{\Omega}_u^{(k)} &:= \operatorname*{arg\,min}_{\Omega_u\in\mathbb{S}^{++}_{p_1\times p_1}} \big\{ \log\det\Omega_u - \operatorname{trace}\big(\widehat{S}_u^{(k)}\Omega_u\big) + \rho_u\|\Omega_u\|_{1,\text{off}} \big\},
\end{aligned}$$

where

$$\widehat{\Gamma}_X^{(k)} = \widehat{\Omega}_u^{(k-1)} \otimes \frac{\mathcal{X}'\mathcal{X}}{T}, \quad \widehat{\gamma}_X^{(k)} = \frac{1}{T}\big(\widehat{\Omega}_u^{(k-1)} \otimes \mathcal{X}\big), \quad \widehat{S}_u^{(k)} = \frac{1}{T}\big[\mathcal{X}^T - \mathcal{X}(\widehat{A}^{(k)})'\big]'\big[\mathcal{X}^T - \mathcal{X}(\widehat{A}^{(k)})'\big].$$

As a consequence, to establish the finite-sample bounds of $\widehat{A}$ and $\widehat{\Omega}_u$ given in (9), we need $\widehat{\Gamma}_X^{(k)}$ to satisfy the RSC condition, a bound on $\|\mathcal{X}'\mathcal{U}\widehat{\Omega}_u^{(k-1)}\|_\infty$ and a bound on $\|\widehat{S}_u^{(k)} - \Sigma_u^\star\|_\infty$ for all k. Toward this end, we prove that for random realizations of $\mathcal{X}$ and $\mathcal{U}$, with high probability, the RSC condition for $\widehat{\Gamma}_X^{(k)}$ and the universal bounds for $\|\mathcal{X}'\mathcal{U}\widehat{\Omega}_u^{(k-1)}\|_\infty$ and $\|\widehat{S}_u^{(k)} - \Sigma_u^\star\|_\infty$ hold for *all iterations* k, albeit the quantities of interest rely on estimates from the previous or current iterations. Consistency results of $\widehat{A}$ and $\widehat{\Omega}_u$ then readily follow.

Next, consider the solution to (10) that alternately updates (B, C) and Ω_v. As the regularization term involves both the nuclear norm penalty and the ℓ_1 norm penalty, additional handling of the norms is required which leverages the idea of decomposable regularizers (Agarwal et al., 2012). Specifically, if Ω_v and (B, C) are respectively held fixed, then

$$(\bar{B}, \bar{C}) := \arg\min_{B,C} \Big\{ \tfrac{1}{T}\mathrm{tr}\big[\Omega_v(\mathcal{Z}^T - \mathcal{X}B' - \mathcal{Z}C')'(\mathcal{Z}^T - \mathcal{X}B' - \mathcal{Z}C')\big] + \lambda_B \|\!|\!| B |\!|\!|_* + \lambda_C \|C\|_1 \Big\},$$
$$\bar{\Omega}_v := \arg\min_{\Omega_v} \big\{ \log\det\Omega_v - \mathrm{trace}\big(S_v\Omega_v\big) + \rho_v\|\Omega_v\|_{1,\text{off}} \big\},$$

where $S_v = \frac{1}{T}(Z^T - \mathcal{X}B' - \mathcal{Z}C')'(Z^T - \mathcal{X}B' - \mathcal{Z}C')$. If we let $\mathcal{W} := [\mathcal{X}, \mathcal{Z}] \in \mathbb{R}^{T\times(p_1+p_2)}$, and define the operator $\mathfrak{W}_{\Omega_v} : \mathbb{R}^{p_2\times(p_1+p_2)} \mapsto \mathbb{R}^{T\times p_2}$ induced jointly by $\mathcal{W}$ and Ω_v as

$$\mathfrak{W}_{\Omega_v}(\Delta) := \mathcal{W}\Delta'\Omega_v^{1/2} \quad \text{for } \Delta \in \mathbb{R}^{p_2\times(p_1+p_2)}, \tag{15}$$

then $\bar{B}_{\text{aug}} := [\bar{B}, O_{p_2\times p_2}]$ and $\bar{C}_{\text{aug}} := [O_{p_2\times p_1}, \bar{C}]$ are equivalently given by

$$(\bar{B}_{\text{aug}}, \bar{C}_{\text{aug}}) = \arg\min_{B,C} \Big\{ \tfrac{1}{T} \Big|\!\Big|\!\Big| \mathcal{Z}^T\Omega_v^{1/2} - \mathfrak{W}_{\Omega_v}(B_{\text{aug}} + C_{\text{aug}}) \Big|\!\Big|\!\Big|_F^2 + \lambda_B \|\!|\!| B |\!|\!|_* + \lambda_C\|C\|_1 \Big\}, \tag{16}$$

where $B_{\text{aug}} := [B, O_{p_2\times p_2}], C_{\text{aug}} := [O_{p_2\times p_1}, C] \in \mathbb{R}^{p_2\times(p_1+p_2)}$. Then, for fixed realizations of $\mathcal{Z}$, $\mathcal{X}$ and $\mathcal{V}$, with an extension of Agarwal et al. (2012) the error bound of $(\bar{B}_{\text{aug}}, \bar{C}_{\text{aug}})$ relies on (1) the operator $\mathfrak{W}_{\Omega_v}$ satisfying the RSC condition; and (2) tuning parameters λ_B and λ_C are respectively chosen in accordance with the deviation bound conditions associated with

$$\|\!|\!| \mathcal{W}'\mathcal{V}\Omega_v/T |\!|\!|_{op} \qquad \text{and} \qquad \|\!|\!| \mathcal{W}'\mathcal{V}\Omega_v/T |\!|\!|_\infty. \tag{17}$$

The error bound of $\bar{\Omega}_v$ again relies on $\|S_v - \Sigma_v^\star\|_\infty$. In an analogous way, for the actual alternate update,

$$(\widehat{B}_{\text{aug}}^{(k)}, \widehat{C}_{\text{aug}}^{(k)}) = \arg\min_{B,C} \Big\{ \tfrac{1}{T} \Big|\!\Big|\!\Big| \mathcal{Z}^T\big[\widehat{\Omega}_v^{(k-1)}\big]^{1/2} - \mathfrak{W}_{\widehat{\Omega}_v^{(k-1)}}(B_{\text{aug}} + C_{\text{aug}}) \Big|\!\Big|\!\Big|_F^2 + \lambda_B \|\!|\!| B |\!|\!|_* + \lambda_C\|C\|_1 \Big\},$$
$$\widehat{\Omega}_v^{(k)} := \arg\min_{\Omega_v} \big\{ \log\det\Omega_v - \mathrm{trace}\big(\widehat{S}_v^{(k)}\Omega_v\big) + \rho_v\|\Omega_v\|_{1,\text{off}} \big\},$$

and the error bound of $(\widehat{B}, \widehat{C}, \widehat{\Omega}_v)$ defined in (10) depends on the properties of $\mathfrak{W}_{\widehat{\Omega}_v^{(k)}}$, $\|\!|\!| \mathcal{W}'\mathcal{V}\Omega_v^{(k)}/T |\!|\!|_{op}$ and $\|\!|\!| \mathcal{W}'\mathcal{V}\Omega_v^{(k)}/T |\!|\!|_\infty$ for all k. Specifically, when Ω_v and (B, C) (in (15) and (17), resp.) are substituted by estimated quantities, we prove that the RSC condition and bounds hold with high probability for random realizations of $\mathcal{Z}$, $\mathcal{X}$ and $\mathcal{V}$, for *all iterations* k, which then establishes the consistency properties of $(\widehat{B}, \widehat{C})$ and $\widehat{\Omega}_v$.

3.2 Consistency results for the Maximum Likelihood estimators.

Theorems 5 and 6 below give the error bounds for the estimators in (9) and (10) obtained through Algorithms 1 and 2, using random realizations coming from the stable VAR system defined in (1). As previously mentioned, to establish error bounds for both the transition matrices and the inverse covariance matrix obtained from alternating updates, we need to take into account that the quantities associated with the RSC condition and the deviation bound condition are based on *estimated quantities* obtained from the previous iteration. On the other hand, the sources of randomness contained in the observed data are fixed, hence errors from observed data stop accumulating once all sources of randomness are considered after a few iterations, which govern both the leading term of the error bounds and the probability for the bounds to hold.

Specifically, using the same notation as defined in Section 3.1, we obtain the error bounds of the estimated transition matrices and inverse covariance matrices iteratively, building upon that for all iterations k:

(1) $\widehat{\Gamma}_X^{(k)}$ or the operator $\mathfrak{W}_{\widehat{\Omega}_v^{(k)}}$ satisfies the RSC condition;

(2) deviation bounds hold for $\|\mathcal{X}'\mathcal{U}\widehat{\Omega}_u^{(k)}/T\|_\infty$, $\|\mathcal{W}'\mathcal{V}\widehat{\Omega}_v^{(k)}/T\|_\infty$, and $|\!|\!|\mathcal{W}'\mathcal{V}\widehat{\Omega}_v^{(k)}/T|\!|\!|_{\text{op}}$;

(3) a good concentration given by $\|\widehat{S}_u^{(k)} - \Sigma_u^\star\|_\infty$ and $\|\widehat{S}_v^{(k)} - \Sigma_v^\star\|_\infty$.

We keep track of how the bounds change in each iteration until convergence, by properly controlling the norms and track the rate of the error bound that depends on p_1, p_2 and T, and reach the conclusion that the error bounds *hold uniformly for all iterations*, for the estimates of both the transition matrices A, B and C and the inverse covariance matrices Ω_u and Ω_v.

Theorem 5 *Consider the stable Gaussian VAR process defined in (1) in which $A^\star$ is assumed to be $s_A^\star$-sparse. Further, assume the following:*

C.1 The incoherence condition holds for $\Omega_u^\star$.

C.2 $\Omega_u^\star$ is diagonally dominant.

C.3 The maximum node degree of $\Omega_u^\star$ satisfies $d_{\Omega_u^\star}^{\max} = o(p_1)$.

Then, for random realizations of $\{X_t\}$ and $\{U_t\}$, and the sequence $\{\widehat{A}^{(k)}, \widehat{\Omega}_u^{(k)}\}_k$ returned by Algorithm 1 outlined in Section 2.1, there exist constants $c_1, c_2, \tilde{c}_1, \tilde{c}_2 > 0, \tau > 0$ such that for sample size $T \gtrsim \max\{(d_{\Omega_u^\star}^{\max})^2, s_A^\star\} \log p_1$, with probability at least

$$1 - c_1 \exp(-c_2 T) - \tilde{c}_1 \exp(-\tilde{c}_2 \log p_1) - \exp(-\tau \log p_1),$$

the following hold for all $k \geq 1$ for some $C_0, C_0' > 0$ that are functions of the upper and lower extremes $\mathcal{M}(f_X), \mathfrak{m}(f_X)$ of the spectrum $f_X(\theta)$ and do not depend on p_1, T or k:

(i) $\widehat{\Gamma}_X^{(k)}$ satisfies the RSC condition;

(ii) $\|\mathcal{X}'\mathcal{U}\widehat{\Omega}_u^{(k)}/T\|_\infty \leq C_0\sqrt{\frac{\log p_1}{T}}$;

(iii) $\|\widehat{S}_u^{(k)} - \Sigma_u^\star\|_\infty \leq C_0'\sqrt{\frac{\log p_1}{T}}$.

As a consequence, the following bounds hold uniformly for all iterations $k \geq 1$:

$$\left|\!\left|\!\left|\widehat{A}^{(k)} - A^\star\right|\!\right|\!\right|_F = O\Big(\sqrt{\frac{s_A^\star \log p_1}{T}}\Big), \qquad \left|\!\left|\!\left|\widehat{\Omega}_u^{(k)} - \Omega_u^\star\right|\!\right|\!\right|_F = O\Big(\sqrt{\frac{(s_{\Omega_u}^\star + p_1)\log p_1}{T}}\Big).$$

It should be noted that the above result establishes the *consistency for the ML estimates* of the model presented in Basu and Michailidis (2015).

Theorem 6 *Consider the stable Gaussian VAR system defined in (1) in which $B^\star$ is assumed to be low rank with rank $r_B^\star$ and $C^\star$ is assumed to be $s_C^\star$-sparse. Further, assume the following*

C.1 The incoherence condition holds for $\Omega_v^\star$.

C.2 $\Omega_v^\star$ is diagonally dominant.

C.3 The maximum node degree of $\Omega_v^\star$ satisfies $d_{\Omega_v^\star}^{\max} = o(p_2)$.

Then, for random realizations of $\{X_t\}$, $\{Z_t\}$ and $\{V_t\}$, and the sequence $\{(\widehat{B}^{(k)}, \widehat{C}^{(k)}), \widehat{\Omega}_v^{(k)}\}_k$ returned by Algorithm 2 outlined in Section 2.1, there exist constants $\{c_i, \tilde{c}_i\}, i = (0,1,2)$ and $\tau > 0$ such that for sample size $T \gtrsim (d_{\Omega_v^\star}^{\max})^2(p_1 + 2p_2)$, with probability at least

$$1 - c_0\exp\{-\tilde{c}_0(p_1+p_2)\} - c_1\exp\{-\tilde{c}_1(p_1+2p_2)\} - c_2\exp\{-\tilde{c}_2\log[p_2(p_1+p_2)]\} - \exp\{-\tau\log p_2\},$$

the following hold for all $k \geq 1$ for $C_0, C_0', C_0'' > 0$ that are functions of the upper and lower extremes $\mathcal{M}(f_W), \mathfrak{m}(f_W)$ of the spectrum $f_W(\theta)$ and of the upper extreme $\mathcal{M}(f_{W,V})$ of the cross-spectrum $f_{W,V}(\theta)$ and do not depend on p_1, p_2 or T:

(i) $\widehat{\Gamma}_W^{(k)}$ *satisfies the RSC condition;*

(ii) $\|\mathcal{W}'\mathcal{V}\widehat{\Omega}_v^{(k)}/T\|_\infty \leq C_0\sqrt{\frac{(p_1+p_2)+p_2}{T}}$ *and* $|\!|\!|\mathcal{W}'\mathcal{V}\widehat{\Omega}_v^{(k)}/T|\!|\!|_{op} \leq C_0'\sqrt{\frac{(p_1+p_2)+p_2}{T}}$;

(iii) $\|\widehat{S}_v^{(k)} - \Sigma_v^\star\|_\infty \leq C_0''\sqrt{\frac{(p_1+p_2)+p_2}{T}}$.

As a consequence, the following bounds hold uniformly for all iterations $k \geq 1$:

$$\left|\!\left|\!\left|\widehat{B}^{(k)} - B^\star\right|\!\right|\!\right|_F^2 + \left|\!\left|\!\left|\widehat{C}^{(k)} - C^\star\right|\!\right|\!\right|_F^2 = O\Big(\frac{\max\{r_B^\star, s_C^\star\}(p_1+2p_2)}{T}\Big), \quad \left|\!\left|\!\left|\widehat{\Omega}_v^{(k)} - \Omega_v^\star\right|\!\right|\!\right|_F = O\Big(\sqrt{\frac{(s_{\Omega_v}^\star + p_2)(p_1+2p_2)}{T}}\Big).$$

Remark 7 *It is worth pointing out that the initializers $\widehat{A}^{(0)}$ and $(\widehat{B}^{(0)}, \widehat{C}^{(0)})$ are slightly different from those obtained in successive iterations, as they come from the penalized least square formulation where the inverse covariance matrices are temporarily assumed diagonal. Consistency results for these initializers under deterministic realizations are established in Theorems 14 and 15 (see Appendix A), and the corresponding conditions are later verified for random realizations in Lemmas 16 to 19 (see Appendix B). These theorems and lemmas serve as the stepping stone toward the proofs of Theorems 5 and 6.*

Further, the constants C_0, C_0', C_0'' reflect both the temporal dependence among X_t and Z_t blocks, as well as the cross-sectional dependence within and across the two blocks.

3.3 The effect of temporal and cross-dependence on the established bounds.

We conclude this section with a discussion on the error bounds of the estimators that provides additional insight into the impact of temporal and cross dependence within and between the blocks; specifically, how the exact bounds depend on the underlying processes through their spectra when explicitly taking into consideration the triangular structure of the joint transition matrix.

First, we introduce additional notations needed in subsequent technical developments. The definition of the spectral densities and the cross-spectrum are the same as previously defined in Section 2 and their upper and lower extremes are defined in Section 3.1. For $\{X_t\}$ defined in (1), let $\mathcal{A}(\theta) = \mathrm{I}_{p_1} - A\theta$ denote the characteristic matrix-valued polynomial of $\{X_t\}$ and $\mathcal{A}^*(\theta)$ denote its conjugate. We further define its upper and lower extremes by:

$$\mu_{\max}(\mathcal{A}) = \max_{|\theta|=1} \Lambda_{\max}\left(\mathcal{A}^*(\theta)\mathcal{A}(\theta)\right), \quad \mu_{\min}(\mathcal{A}) = \min_{|\theta|=1} \Lambda_{\min}\left(\mathcal{A}^*(\theta)\mathcal{A}(\theta)\right).$$

The same set of quantities for the joint process $\{W_t = (X_t', Z_t')'\}$ are analogously defined, that is,

$$\mathcal{G}(\theta) = \mathrm{I}_{p_1+p_2} - G\theta, \quad \mu_{\max}(\mathcal{G}) = \max_{|\theta|=1} \Lambda_{\max}\left(\mathcal{G}^*(\theta)\mathcal{G}(\theta)\right), \quad \mu_{\min}(\mathcal{G}) = \min_{|\theta|=1} \Lambda_{\min}\left(\mathcal{G}^*(\theta)\mathcal{G}(\theta)\right).$$

Using the result in Theorem 6 as an example, we show how the error bound depends on the underlying processes $\{(X_t', Z_t')'\}$. Specifically, we note that the bounds for $(\widehat{B}^{(k)}, \widehat{C}^{(k)})$ can be equivalently written as

$$|||\widehat{B}^{(k)} - B^\star|||_F^2 + |||\widehat{C}^{(k)} - C^\star|||_F^2 \leq \bar{C}\Big(\tfrac{\max\{r_B^\star, s_C^\star\}(p_1+2p_2)}{T}\Big),$$

which holds for all k and some constant $\bar{C}$ that does not depend on p_1, p_2 or T. Specifically, by Theorem 15, Lemmas 18 and 19,

$$C_0 \propto \left[\mathcal{M}(f_W) + \tfrac{1}{2\pi}\Lambda_{\max}(\Sigma_v) + \mathcal{M}(f_{W,V})\right] / \mathfrak{m}(f_W).$$

This indicates that the exact error bound depends on $\mathfrak{m}(f_W), \mathcal{M}(f_W)$ and $\mathcal{M}(f_{W,V})$. Next, we provide bounds on these quantities. The joint process W_t as we have noted in (2), is a VAR(1) process with characteristic polynomial $\mathcal{G}(\theta)$ and spectral density $f_W(\theta)$. The bounds for $\mathfrak{m}(f_W)$ and $\mathcal{M}(f_W)$ are given by Basu and Michailidis (2015, Proposition 2.1), that is,

$$\mathfrak{m}(f_W) \geq \frac{\min\{\Lambda_{\min}(\Sigma_u), \Lambda_{\min}(\Sigma_v)\}}{(2\pi)\mu_{\max}(\mathcal{G})} \quad \text{and} \quad \mathcal{M}(f_W) \leq \frac{\max\{\Lambda_{\max}(\Sigma_u), \Lambda_{\max}(\Sigma_v)\}}{(2\pi)\mu_{\min}(\mathcal{G})}. \tag{18}$$

Consider the bound for $\mathcal{M}(f_{W,V})$. First, we note that $\{V_t\}$ is a sub-process of the joint error process $\{\varepsilon_t\}$, where $\varepsilon_t = (U_t', V_t')'$. Then, by Lemma 24,

$$\mathcal{M}(f_{W,V}) \leq \mathcal{M}(f_{W,\varepsilon}) \leq \mathcal{M}(f_W)\mu_{\max}(\mathcal{G}),$$

where the second inequality follows from Basu and Michailidis (2015, Proof of Proposition 2.4).

What are left to be bounded are $\mu_{\min}(\mathcal{G})$ and $\mu_{\max}(\mathcal{G})$. By Proposition 2.2 in Basu and Michailidis (2015), these two quantities are bounded by:

$$\mu_{\max}(\mathcal{G}) \leq \Big[1 + \frac{\vert\vert\vert G\vert\vert\vert_\infty + \vert\vert\vert G\vert\vert\vert_1}{2}\Big]^2 \tag{19}$$

and

$$\mu_{\min}(\mathcal{G}) \geq (1-\rho(G))^2 \cdot \vert\vert\vert P\vert\vert\vert_{op}^{-2} \cdot \Big\vert\Big\vert\Big\vert P^{-1}\Big\vert\Big\vert\Big\vert_{op}^{-2},$$

where $G = P\Lambda_G P^{-1}$ with Λ_G being a diagonal matrix consisting of the eigenvalues of G. Since $\vert\vert\vert P^{-1}\vert\vert\vert_{op} \geq \vert\vert\vert P\vert\vert\vert_{op}^{-1}$, it follows that

$$\vert\vert\vert P\vert\vert\vert_{op}^{-2} \cdot \Big\vert\Big\vert\Big\vert P^{-1}\Big\vert\Big\vert\Big\vert_{op}^{-2} \geq \Big\vert\Big\vert\Big\vert P^{-1}\Big\vert\Big\vert\Big\vert_{op}^{2} \cdot \Big\vert\Big\vert\Big\vert P^{-1}\Big\vert\Big\vert\Big\vert_{op}^{-2} = 1,$$

and therefore

$$\mu_{\min}(\mathcal{G}) \geq \left(1 - \max\{\rho(A), \rho(C)\}\right)^2. \tag{20}$$

Remark 8 The impact of the system's lower-triangular structure on the established bounds. *Consider the bounds in (19) and (20). An upper bound of $\mu_{\max}(\mathcal{G})$ depends on $\vert\vert\vert G\vert\vert\vert_\infty$ and $\vert\vert\vert G\vert\vert\vert_1$, whereas a lower bound of $\mu_{\min}(\mathcal{G})$ involves only the spectral radius of G. Combined with (18), this suggests that the lower extreme of the spectral density is associated with the average of the maximum weighted in-degree and out-degree of the system, whereas the upper extreme is associated with the stability condition: the* less the system is intra- and inter-connected*, the* tighter the bound *for the lower extreme will be; similarly, the* more stable *(exhibits smaller temporal dependence) the system is, the* tighter the bound *for the upper extreme will be. Finally, we note that an upper bound for $(\vert\vert\vert G\vert\vert\vert_\infty + \vert\vert\vert G\vert\vert\vert_1)$ is given by*

$$\max\{\vert\vert\vert A\vert\vert\vert_\infty + \vert\vert\vert B\vert\vert\vert_\infty, \vert\vert\vert C\vert\vert\vert_\infty\} + \max\{\vert\vert\vert A\vert\vert\vert_1, \vert\vert\vert B\vert\vert\vert_1 + \vert\vert\vert C\vert\vert\vert_1\}.$$

The presence of $\vert\vert\vert B\vert\vert\vert_\infty$ and $\vert\vert\vert B\vert\vert\vert_1$ depicts the role of the inter-connectedness between $\{X_t\}$ and $\{Z_t\}$ on the lower extreme of the spectrum, which is associated with the overall curvature of the joint process.

The impact of the system's lower-triangular structure on the system capacity. *With G being a lower-triangular matrix, we only require $\rho(A) < 1$ and $\rho(C) < 1$ to ensure the stability of the system. This enables the system to have "larger capacity" (can accommodate more cross-dependence within each block), since the two sparse components A and C can exhibit larger average weighted in- and out-degrees compared with a system where G does not possess such triangular structure. In the case where G is a complete matrix, one deals with a $(p_1 + p_2)$-dimensional VAR system and $\rho(G) < 1$ is required to ensure its stability. As a consequence, the average weighted in- and out-degree requirements for each time series become more restrictive.*

4. Testing Group Granger-Causality.

In this section, we develop a procedure for testing the hypothesis $H_0 : B = 0$. Note that without the presence of B, the blocks X_t and Z_t in the model become *decoupled* and can be treated as two separate VAR models, whereas with a nonzero B, the group of variables in

Z_t is collectively "Granger-caused" by those in X_t. Moreover, since we are testing whether or not the entire block of B is zero, we do not need to rely on the exact distribution of its individual entries, but rather on the properly measured correlation between the responses and the covariates. To facilitate presentation of the testing procedure, we illustrate the proposed framework via a simpler model setting with $Y_t = \Pi X_t + \epsilon_t$ and testing whether $\Pi = 0$; subsequently, we translate the results to the actual setting of interest, namely, whether or not $B = 0$ in the model $Z_t = BX_{t-1} + CZ_{t-1} + V_t$.

The testing procedure focuses on the following sequence of tests for the rank of B:

$$H_0 : \text{rank}(B) \leq r, \qquad \text{for an arbitrary} \quad r < \min(p_1, p_2). \tag{21}$$

Note that the hypothesis of interest, $B = 0$ corresponds to the special case with $r = 0$. To test for it, we develop a procedure associated with *canonical correlations*, which leverages ideas present in the literature (see Anderson, 2002).

As mentioned above, we consider a simpler setting similar to that in Anderson (2002), given by

$$Y_t = \Pi X_t + \epsilon_t,$$

where $Y_t \in \mathbb{R}^{p_2}$, $X \in \mathbb{R}^{p_1}$ and ϵ_t is independent of X_t. At the population level, let

$$\mathbb{E}Y_tY_t' = \Sigma_Y, \qquad \mathbb{E}X_tX_t' = \Sigma_X, \qquad \mathbb{E}Y_tX_t' = \Sigma_{YX} = \Sigma_{XY}'.$$

The population canonical correlations between Y_t and X_t are the roots of

$$\begin{vmatrix} -\rho\Sigma_Y & \Sigma_{YX} \\ \Sigma_{XY} & -\rho\Sigma_X \end{vmatrix} = 0,$$

i.e., the nonnegative solutions to

$$|\Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY} - \rho^2\Sigma_Y| = 0, \tag{22}$$

with ρ being the unknown. By the results in Anderson (2002), the number of positive solutions to (22) is equal to the rank of Π, and indicates the "degree of dependency" between processes Y_t and X_t. This suggests that if $\text{rank}(\Pi) \leq r < p$, we would expect $\sum_{k=r+1}^{p} \lambda_k$ to be small, where the λ's solve the eigen-equation

$$|S_{YX}S_X^{-1}S_{XY} - \lambda S_Y| = 0, \qquad \text{with} \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p,$$

and S_X, S_{XY} and S_Y are the sample counterparts corresponding to Σ_X, Σ_{XY} and Σ_Y, respectively.

With this background, we switch to our model setting given by

$$Z_t = BX_{t-1} + CZ_{t-1} + V_t, \tag{23}$$

where V_t is assumed to be independent of X_{t-1} and Z_{t-1}, B encodes the canonical correlation between Z_t and X_{t-1}, conditional on Z_{t-1}. We use the same notation as in Section 3; that is, let $\Gamma_X(h) = \mathbb{E}X_tX'_{t+h}$, $\Gamma_Z(h) = \mathbb{E}Z_tZ'_{t+h}$, and $\Gamma_{X,Z}(h) = \mathbb{E}X_tZ'_{t+h}$, with (h) omitted whenever $h = 0$. At the population level, under the Gaussian assumption,

$$\begin{bmatrix} Z_t \\ X_{t-1} \\ Z_{t-1} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Gamma_Z & \Gamma'_{X,Z}(1) & \Gamma_Z(1) \\ \Gamma_{X,Z}(1) & \Gamma_X & \Gamma_{X,Z} \\ \Gamma'_Z(1) & \Gamma'_{X,Z} & \Gamma_Z \end{bmatrix} \right),$$

which suggests that conditionally,

$$Z_t \big| Z_{t-1} \sim \mathcal{N}\left(\Gamma_Z(1)\Gamma_Z^{-1}Z_{t-1}, \Sigma_{00}\right) \qquad \text{and} \qquad X_{t-1}\big|Z_{t-1} \sim \mathcal{N}\left(\Gamma_{X,Z}\Gamma_Z^{-1}Z_{t-1}, \Sigma_{11}\right),$$

where

$$\Sigma_{00} := \Gamma_Z - \Gamma_Z(1)\Gamma_Z^{-1}\Gamma'_Z(1) \qquad \text{and} \qquad \Sigma_{11} := \Gamma_X - \Gamma_{X,Z}\Gamma_Z^{-1}\Gamma'_{X,Z}. \tag{24}$$

Then, we have that jointly

$$\begin{bmatrix} Z_t \\ X_{t-1} \end{bmatrix} \Big| Z_{t-1} \sim \mathcal{N}\left(\begin{bmatrix} \Gamma_Z(1) \\ \Gamma_{X,Z} \end{bmatrix} \Gamma_Z^{-1} Z_{t-1}\ ,\ \begin{bmatrix} \Gamma_Z & \Gamma'_{XZ}(1) \\ \Gamma_{XZ}(1) & \Gamma_Z \end{bmatrix} - \begin{bmatrix} \Gamma_Z(1) \\ \Gamma_{XZ} \end{bmatrix} \Gamma_Z^{-1} \begin{bmatrix} \Gamma'_Z(1) & \Gamma_{ZX} \end{bmatrix} \right),$$

so the partial covariance matrix between Z_t and X_{t-1} conditional on Z_{t-1} is given by

$$\Sigma_{10} = \Sigma'_{01} := \Gamma_{X,Z}(1) - \Gamma_Z(1)\Gamma_Z^{-1}\Gamma_{X,Z}. \tag{25}$$

The population canonical correlations between Z_t and X_{t-1} conditional on Z_{t-1} are the non-negative roots of

$$\left|\Sigma_{01}\Sigma_{11}^{-1}\Sigma_{10} - \rho^2\Sigma_{00}\right| = 0,$$

and the number of positive solutions corresponds to the rank of B; see Anderson (1951) for a discussion in which the author is interested in estimating and testing linear restrictions on regression coefficients. Therefore, to test $\text{rank}(B) \leq r$, it is appropriate to examine the behavior of $\Psi_r := \sum_{k=r+1}^{\min(p_1,p_2)} \phi_k$, where ϕ's are ordered non-increasing solutions to

$$|S_{01}S_{11}^{-1}S_{10} - \phi S_{00}| = 0, \tag{26}$$

and S_{01}, S_{11} and S_{00} are the empirical surrogates for the population quantities Σ_{01}, Σ_{11} and Σ_{00}. For subsequent developments, we make the very mild assumption that $p_1 < T$ and $p_2 < T$ so that $\mathcal{Z}'\mathcal{Z}$ is invertible.

Proposition 9 gives the tail behavior of the eigenvalues and Corollary 11 gives the testing procedure for block "Granger-causality" as a direct consequence.

Proposition 9 *Consider the model setup given in (23), where $B \in \mathbb{R}^{p_2 \times p_1}$. Further, assume all positive eigenvalues μ of the following eigen-equation are of algebraic multiplicity one:*

$$\left|\Sigma_{01}\Sigma_{11}^{-1}\Sigma_{10} - \mu\Sigma_{00}\right| = 0, \tag{27}$$

where Σ_{00}, Σ_{11} and Σ_{01} are given in (24) and (25). The test statistic for testing

$$H_0 : \text{rank}(B) \leq r, \qquad \textit{for an arbitrary} \quad r < \min(p_1, p_2),$$

is given by

$$\Psi_r := \sum_{k=r+1}^{\min(p_1,p_2)} \phi_k,$$

where ϕ_k*'s are ordered decreasing solutions to the eigen-equation* $|S_{01}S_{11}^{-1}S_{10} - \phi S_{00}| = 0$ *where*

$$S_{11} = \tfrac{1}{T}\mathcal{X}'(I-P_z)\mathcal{X}, \quad S_{00} = \tfrac{1}{T}\left(\mathcal{Z}^T\right)'(I-P_z)\left(\mathcal{Z}^T\right), \quad S_{10} = S_{01}' = \tfrac{1}{T}\mathcal{X}'(I-P_z)\left(\mathcal{Z}^T\right),$$

and $P_z = \mathcal{Z}(\mathcal{Z}'\mathcal{Z})^{-1}\mathcal{Z}'$*. Moreover, the limiting behavior of* Ψ_r *is given by*

$$T\Psi_r \sim \chi^2_{(p_1-r)(p_2-r)}.$$

Remark 10 *We provide a short comment on the assumption that the positive solutions to (27) have algebraic multiplicity one in Proposition 9. This assumption is imposed on the eigen-equation associated with population quantities, to exclude the case where a positive root has algebraic multiplicity greater than one and its geometric multiplicity does not match the algebraic one, and hence we would fail to obtain r mutually independent canonical variates and the rank-r structure becomes degenerate. With the imposed assumption which is common in the canonical correlation analysis literature (e.g. Anderson, 2002), such a scenario is automatically excluded. Specifically, this condition is not stringent, as for* ϕ*'s that are solutions to the eigen-equation associated with sample quantities, the distinctiveness amongst roots is satisfied with probability 1 (see Hsu, 1941b, Proof of Lemma 3).*

Corollary 11 (Testing group Granger-causality) *Under the model setup in (23), the test statistic for testing* $B = 0$ *is given by*

$$\Psi_0 := \sum_{k=1}^{\min(p_1,p_2)} \phi_k,$$

with ϕ_k *being the ordered decreasing solutions of*

$$\left| S_{01}\left[\mathrm{diag}(S_{11})\right]^{-1} S_{10} - \phi S_{00} \right| = 0.$$

Asymptotically, $T\Psi_0 \sim \chi^2_{p_1p_2}$*. To conduct a level* α *test, we reject the null hypothesis* $H_0 : B = 0$ *if*

$$\Psi_0 > \frac{1}{T}\chi^2_{p_1p_2}(\alpha),$$

where $\chi^2_d(\alpha)$ *is the upper* α *quantile of the* χ^2 *distribution with* d *degrees of freedom.*

Remark 12 *Corollary 11 is a special case of Proposition 9 with the null hypothesis being* $H_0 : r = 0$*, which corresponds to the Granger-causality test. Under this particular setting, we are able to take the inverse with respect to diag*(S_{11})*, yet maintain the same asymptotic distribution due to the fact that* $S_{01} = S_{10} = 0$ *under the null hypothesis* $B = 0$*. This enables us to perform the test even with* $p_1 > T$*.*

The above testing procedure takes advantage of the fact that when $B = 0$, the canonical correlations among the partial regression residuals after removing the effect of Z_{t-1} are very close to zero. However, the test may not be as powerful under a sparse alternative, i.e., $H_A : B$ is sparse. In Supplementary Material **??**, we present a testing procedure that specifically takes into consideration the fact that the alternative hypothesis is sparse, and the corresponding performance evaluation is shown in Section 5.2 under this setting.

5. Performance Evaluation.

Next, we present the results of numerical studies to evaluate the performance of the developed ML estimates (Section 2.1) of the model parameters, as well as that of the testing procedure (Section 4).

5.1 Simulation results for the estimation procedure.

A number of factors may potentially influence the performance of the estimation procedure; in particular, the model dimension p_1 and p_2, the sample size T, the rank of $B^\star$ and the sparsity level of $A^\star$ and $C^\star$, as well as the spectral radius of $A^\star$ and $C^\star$. Hence, we consider several settings where these parameters vary.

For all settings, the data $\{x_t\}_1^T$ and $\{z_t\}_1^T$ are generated according to the model

$$\begin{aligned} x_t &= A^\star x_{t-1} + u_t, \\ z_t &= B^\star x_{t-1} + C^\star z_{t-1} + v_t. \end{aligned}$$

For the sparse components, each entry in $A^\star$ and $C^\star$ is nonzero with probability $2/p_1$ and $1/p_2$ respectively, and the nonzero entries are generated from $\mathsf{Unif}\left([-2.5, -1.5] \cup [1.5, 2.5]\right)$, then scaled down so that the spectral radii $\rho(A)$ and $\rho(C)$ satisfy the stability condition. For the low rank component, each entry in $B^\star$ is generated from $\mathsf{Unif}(-10, 10)$, followed by singular value thresholding, so that $\text{rank}(B^\star)$ conforms with the model setup. For the contemporaneous dependence encoded by $\Omega_u^\star$ and $\Omega_v^\star$, both matrices are generated according to an Erdös-Rényi random graph, with sparsity being 0.05 and condition number being 3.

Table 1 depicts the values of model parameters under different model settings. Specifically, we consider three major settings in which the size of the system, the rank of the cross-dependence component, and the stability of the system vary. The sample size is fixed at $T = 200$ unless otherwise specified. Additional settings examined (not reported due to space considerations) are consistent with the main conclusions presented next.

We use sensitivity (SEN), specificity (SPC) and relative error in Frobenius norm (Error) as criteria to evaluate the performance of the estimates of transition matrices A, B and C. Tuning parameters are chosen based on BIC. Since the exact contemporaneous dependence is not of primary concern, we omit the numerical results for $\widehat{\Omega}_u$ and $\widehat{\Omega}_v$.

$$\text{SEN} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{SPE} = \frac{\text{TN}}{\text{FP} + \text{TN}}, \quad \text{Error} = \frac{|||\text{Est.} - \text{Truth}|||_F}{|||\text{Truth}|||_F}.$$

Table 2 illustrates the performance for each of the parameters under different simulation settings considered. The results are based on an average of 100 replications and their

Table 1: Model parameters under different model settings

		model parameters				
		p_1	p_2	rank(B^*)	ρ_A	ρ_C
model dimension	A.1	50	20	5	0.5	0.5
	A.2	100	50	5	0.5	0.5
	A.3	200	50	5	0.5	0.5
	A.4	50	100	5	0.5	0.5
rank	B.1	100	50	10	0.5	0.5
	B.2	100	50	20	0.5	0.5
spectral radius	C.1	50	20	5	0.8	0.5
	C.2	50	20	5	0.5	0.8
	C.3	50	20	5	0.8	0.8

Table 2: Performance evaluation of $\widehat{A}$, $\widehat{B}$ and $\widehat{C}$ under different model settings.

	performance of $\widehat{A}$			performance of $\widehat{B}$		performance of $\widehat{C}$		
	SEN	SPC	Error	rank$(\widehat{B})$	Error	SEN	SPC	Error
A.1	0.98(0.014)	0.99(0.004)	0.34(0.032)	5.2(0.42)	0.11(0.008)	1.00(0.000)	0.97(0.008)	0.15(0.074)
A.2	0.97(0.014)	0.99(0.001)	0.38(0.015)	5.2(0.42)	0.31(0.011)	0.97(0.008)	0.97(0.004)	0.28(0.033)
A.3	0.99(0.005)	0.96(0.002)	0.87(0.011)	5.8(0.92)	0.54(0.022)	0.98(0.000)	0.92(0.009)	0.28(0.028)
A.4	0.96(0.0261)	0.99(0.002)	0.36(0.034)	5.2(0.42)	0.32(0.012)	0.95(0.009)	0.98(0.001)	0.37(0.010)
B.1	0.97(0.008)	0.99(0.001)	0.37(0.017)	11.4(1.17)	0.15(0.008)	1.00(0.000)	0.99(0.001)	0.09(0.021)
B.2	0.98(0.008)	0.99(0.001)	0.38(0.016)	21.2(0.91)	0.12(0.006)	1.00(0.000)	0.99(0.001)	0.08(0.018)
C.1	1.00(0.000)	0.97(0.005)	0.25(0.015)	5.6(0.52)	0.23(0.006)	1.00(0.000)	0.92(0.021)	0.11(0.072)
C.2	0.99(0.007)	0.95(0.004)	0.45(0.022)	5.0(0.00)	0.31(0.014)	1.00(0.000)	0.92(0.019)	0.04(0.011)
C.3	1.00(0.000)	0.96(0.004)	0.18(0.013)	6.7(1.16)	0.19(0.011)	1.00(0.000)	0.87(0.029)	0.14(0.067)
C.3'	1.00(0.000)	0.99(0.002)	0.13(0.016)	5.2(0.42)	0.23(0.005)	1.00(0.000)	0.90(0.021)	0.06(0.023)

standard deviations are given in parentheses. Overall, the results are highly satisfactory and all the parameters are estimated with a high degree of accuracy. Further, all estimates were obtained in less than 20 iterations, thus indicating that the estimation procedure is numerically stable. As expected, when the the spectral radii of A and C increase thus leading to less stable $\{X_t\}$ and $\{Z_t\}$ processes, a larger sample size is required for the estimation procedure to match the performance of the setting with same parameters but smaller $\rho(A)$ and $\rho(C)$. This is illustrated in row C.3' of Table 2, where the sample size is increased to $T = 500$, which outperforms the results in row C.3 in which $T = 200$ and broadly matches that of row A.1.

Since in some application settings the data may deviate from the posited Gaussian assumption, we further investigate the robustness of the algorithm in the presence of heavier than Gaussian distributions in the Supplementary Material **??**. Further, a brief comparison of the ML estimates with their two-step (non-iterative) counterparts is given in the Supplementary Material **??** , that illustrates the numerical benefits of the proposed procedure.

Lastly, we examine the performance with respect to one-step-ahead forecasting. Recall that VAR models are widely used for forecasting purposes in many application areas

(Lütkepohl, 2005). The performance metric is given by the relative error as measured by the ℓ_2 norm of the out-of-sample points x_{T+1} and z_{T+1}, where the predicted values are given by $\widehat{x}_{T+1} = \widehat{A}x_T$ and $\widehat{z}_{T+1} = \widehat{B}x_T + \widehat{C}z_T$, respectively. It is worth noting that both $\{X_t\}$ and $\{Z_t\}$ are mean-zero processes. However, since the transition matrix of $\{X_t\}$ is subject to the spectral radius constraints to ensure the stability of the corresponding process, the magnitude of the realized value $\{x_t\}$ is small; whereas for $\{Z_t\}$, since no constraints are imposed on the B coefficient matrix that encodes the inter-dependence, z_t's has the capacity of having relative large values in magnitude. Consequently, the relative error of $\widehat{x}_{T+1}$ is significantly larger than that of $\widehat{z}_{T+1}$, partially due to the small total magnitude of the denominator.

The results show that an increase in the spectral radius (keeping the other structural parameters fixed) leads to a decrease of the relative error, since future observations become more strongly correlated over time. On the other hand, an increase in dimension leads to a deterioration in forecasting, since the available sample size impacts the quality of the parameter estimates. Finally, an increase in the rank of the B matrix is beneficial for forecasting, since it plays a stronger role in the system's temporal evolution.

Table 3: One-step-ahead relative forecasting error.

		$\frac{\lVert\widehat{x}_{T+1}-x_{T+1}\rVert_2}{\lVert x_{T+1}\rVert_2}$	$\frac{\lVert\widehat{z}_{T+1}-z_{T+1}\rVert_2}{\lVert z_{T+1}\rVert_2}$
baseline	A.1	0.89(0.066)	0.23(0.075)
spectral radius	C.1	0.62(0.100)	0.10(0.035)
	C.2	0.93(0.062)	0.17(0.059)
	C.3	0.68(0.096)	0.10(0.045)
rank	B.1	0.92(0.044)	0.14(0.038)
	B.2	0.94(0.042)	0.14(0.025)
dimension	A.2	0.87(0.051)	0.24(0.073)
	A.3	0.96(0.040)	0.44(0.139)
	A.4	0.89(0.059)	0.274(0.068)

5.2 Simulation results for the block Granger-causality test.

Next, we illustrate the empirical performance of the testing procedure introduced in Section 4, together with the one specifically tailored to a sparse alternative (described in detail in the Supplementary Material ??) with the null hypothesis being $B^\star = 0$ and the alternative being $B^\star \neq 0$, either low rank or sparse. Specifically, when the alternative hypothesis is true and has a low-rank structure, we use the general testing procedure proposed in Section 4, whereas when the alternative is true and sparse, we use the testing procedure presented in Supplementary Material ??. We focus on evaluating the type I error (empirical false rejection rate) when $B^\star = 0$, as well as the power of the test when $B^\star$ has nonzero entries.

For both testing procedures, the transition matrix $A^\star$ is generated with each entry being nonzero with probability $2/p_1$, and the nonzeros are generated from $\mathsf{Unif}\,([-2.5, -1.5] \cup [1.5, 2.5])$, then further scaled down so that $\rho(A^\star) = 0.5$. For transition matrix $C^\star$, each entry is nonzero with probability $1/p_2$, and the nonzeros are generated from $\mathsf{Unif}\,([-2.5, -1.5] \cup [1.5, 2.5])$, then further scaled down so that $\rho(C^\star) = 0.5$ or 0.8, depending on the simulation setting. Finally, we only consider the case where v_t and u_t have diagonal covariance matrices.

We use sub-sampling as in Politis et al. (1999) with the number of subsamples set to 3,000; an alternative would have been a block bootstrap procedure (e.g., Hall, 1985). Note that the length of the subsamples varies across simulation settings in order to gain insight on how sample size impacts the type I error or the power of the test.

Low-rank testing. To evaluate the type I error control and the power of the test, we primarily consider the case where $\text{rank}(B^\star) = 0$, with the data alternatively generated based on $\text{rank}(B^\star) = 1$. We test the hypothesis $H_0 : \text{rank}(B) = 0$ and tabulate the empirical proportion of falsely rejecting H_0 when $\text{rank}(B^*) = 0$ (type I error) and the probability that we reject H_0 when $\text{rank}(B^*) = 1$ (power). In addition, we also show how the testing procedure performs when the underlying $B^\star$ has rank $r \geq 0$. In particular, when $\text{rank}(B^\star) = r^\star$, the type I error of the test corresponds to the empirical proportion of rejections of the null hypothesis $H_0 : r \leq r^\star$, while the power of the test to the empirical proportion of rejections of the null hypothesis set to $H_0 : r \leq (r^\star - 1)$. The latter resembles the sequential test in Johansen (1988).

Empirically, we expect that when $B^\star = 0$, the value of the proposed test statistic mostly falls below the cut-off value (upper α quantile), while when $\text{rank}(B^\star) = 1$, the value of the proposed test statistic mostly falls beyond the critical value $\chi^2(\alpha)_{p_1 p_2}/T$ with T being the sample size, hence leading to a detection. Table 4 gives the type I error of the test when setting $\alpha = 0.1, 0.05, 0.1$, and the power of the test using the upper 0.01 quantile of the reference distribution as the cut-off, for different combinations of model dimensions (p_1, p_2) and sample size.

Based on the results shown in Table 4, it can be concluded that the proposed low-rank testing procedure accurately detects the presence of "Granger causality" across the two blocks, when the data have been generated based on a truly multi-layer VAR system. Further, when $B^\star = 0$, the type I error is close to the nominal α level for sufficiently large sample sizes, but deteriorates for increased model dimensions. In particular, relatively large values of p_2 and larger spectral radius $\rho(C^\star)$ negatively impact the empirical false rejection proportion, which deviates from the desired control level of the type I error. In the case where $\text{rank}(B^\star) = r > 0$, the testing procedure provides satisfactory type I error control for larger sample sizes and excellent power.

Sparse testing. Since the rejection rule of the HC-statistic is based on empirical process theory (Shorack and Wellner, 2009) and its dependence on α is not explicit, we focus on illustrating how the empirical proportion of false rejections (type I error) varies with the sample size T, the model dimensions (p_1, p_2) and the spectral radius of $C^\star$. To show the power of the test, each entry in $B^\star$ is nonzero with probability $q \in (0, 1)$ such that $q(p_1 p_2) = (p_1 p_2)^\theta$ with $\theta = 0.6$, to ensure the overall sparsity of $B^\star$ satisfies the sparsity requirement posited in Proposition **??**. The magnitude is set such that the signal-to-noise ratio is 1.2. Note that the actual number of parameters is $p_1 p_2$, while the total number of subsamples used is 3000 with the length of subsamples varying according to different simulation settings to demonstrate the dependence of type I error and power on sample sizes.

Table 4: Empirical type I error and power for low-rank testing

	(p_1, p_2)	sample size	type I error ($B^\star = 0$) $\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	power (rank$(B^\star) = 1$) cut-off $\chi^2(0.01)_{p_1 p_2}/T$
$\rho(C^\star) = 0.5$	$(20, 20)$	$T = 500$	0.028	0.123	0.227	1
		$T = 1000$	0.015	0.073	0.137	1
		$T = 2000$	0.011	0.059	0.118	1
	$(50, 20)$	$T = 500$	0.070	0.228	0.355	1
		$T = 1000$	0.026	0.125	0.226	1
		$T = 2000$	0.013	0.094	0.163	1
	$(20, 50)$	$T = 500$	0.484	0.751	0.857	1
		$T = 1000$	0.089	0.246	0.375	1
		$T = 2000$	0.020	0.088	0.164	1
	$(100, 50)$	$T = 500$	<u>0.997</u>	<u>0.999</u>	<u>1</u>	1
		$T = 1000$	0.608	0.828	0.908	1
		$T = 2000$	0.166	0.374	0.511	1
$\rho(C^\star) = 0.8$	$(20, 50)$	$T = 500$	0.533	0.789	0.880	1
		$T = 1000$	0.130	0.306	0.452	1
		$T = 2000$	0.045	0.145	0.252	1
	$(50, 20)$	$T = 500$	0.083	0.250	0.382	1
		$T = 1000$	0.039	0.133	0.234	1
		$T = 2000$	0.019	0.096	0.174	1
			type I error ($H_0 : r \leq 5$) $\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	power ($H_0 : r \leq 4$) cut-off $\chi^2(0.01)_{(p_1-4)(p_2-4)}/T$
$\rho(C^\star) = 0.5$ rank$(B^\star) = 5$	$(20, 50)$	$T = 500$	0.092	0.274	0.400	1
		$T = 1000$	0.034	0.140	0.236	1
		$T = 2000$	0.022	0.096	0.178	1
	$(50, 20)$	$T = 500$	0.454	0.722	0.829	1
		$T = 1000$	0.126	0.313	0.452	1
		$T = 2000$	0.062	0.184	0.284	1

Table 5: Empirical type I error and power for sparse testing

	(p_1, p_2)	type I error ($B^\star = 0$) 200	500	1000	2000	power (SNR$(B^\star) = 0.8$) 200	500	1000	2000
$\rho(C^\star) = 0.5$	$(20, 20)$	0.244	0.097	0.074	0.055	1	1	1	1
	$(50, 20)$	0.393	0.131	0.108	0.074	1	1	1	1
	$(20, 50)$	<u>0.996</u>	0.351	0.153	0.093	1	1	1	1
	$(100, 50)$	<u>1.000</u>	<u>0.963</u>	0.270	0.115	1	1	1	1
$\rho(C^\star) = 0.8$	$(50, 20)$	0.402	0.158	0.112	0.075	0.829	0.996	1	1
	$(20, 50)$	<u>0.999</u>	0.430	0.166	0.111	1	1	1	1

Based on the results shown in Table 5, when $B^\star = 0$, the proposed testing procedure can effectively detect the absence of block "Granger causality", provided that the sample size is moderately large compared to the total number of parameters being tested. However, if the model dimension is large, whereas the sample size is small, the test procedure becomes problematic and fails to provide legitimate type I error control as desired. When $B^\star$ is nonzero, empirically the test is always able detect its presence, as long as the effective signal-to-noise ratio is beyond the detection threshold.

6. Real Data Analysis Illustration.

We employ the developed framework and associated testing procedures to address one of the motivating applications. Specifically, we analyze the temporal dynamics of the log-returns of stocks with large market capitalization and key macroeconomic variables, as well as their cross-dependence. Specifically, using the notation in (1), we assume that the X_t block consists of the stock log-returns, while the macroeconomic variables form the Z_t block. With this specification, we assume that the macroeconomic variables are "Granger-caused" by the stock market, but not vice versa. Note that our framework allows us to pose and test a more general question than previous work in the economics literature considered. For example, Farmer (2015) building on previous work by Fitoussi et al. (2000); Phelps (1999) tests only the relationship between the employment index and the composite stock index, using a bivariate VAR model. On the other hand, our framework enables us to consider the components of the S&P 100 index and the "medium" list of macroeconomic variables considered in the work of Stock and Watson (2005).

Next, we provide a brief description of the data used. The stock data consist of monthly observations of 71 stocks corresponding to a stable set of historical components comprising the S&P 100 index for the 2001-2016 period. The macroeconomic variables are chosen from the "medium" list in Stock and Watson (2005). The complete lists of stocks and macroeconomic variables, together with the preprocessing to ensure stationarity used in this study are given in Supplementary Material **??**.

We start the analysis by using the VAR model for the stock log-returns to study their evolution over the 2001-2016 period. Analogously to the strategy employed by Billio et al. (2012), we consider 36-month-long rolling-windows for fitting the model $X_t = AX_{t-1} + U_t$, for a total of 143 estimates of the transition matrix A. VAR models involving more than 1 lag were also fitted to the data, but did not indicate temporal dependence beyond lag 1. To obtain the final estimates across all 143 subsamples, we employ *stability selection* (Meinshausen and Bühlmann, 2010), with the threshold set at 0.6 for including an edge in A.[1] Figure 2 depicts the global clustering coefficient (Luce and Perry, 1949) of the skeleton of the estimated A over all 143 rolling windows, with the time stamps on the horizontal axis specifying the starting time of the corresponding window. The results clearly indicate strong connectivity in lead-lag stock relationships during the financial crisis period March 2007-June 2009. Next, we present the analysis based on the VAR-X component of our model, given by $Z_t = BX_{t-1} + CZ_{t-1} + V_t$ with the stock log-returns corresponding to the X_t block and the (stationary) macroeconomic variables to the Z_t block. As before, we fit the data within each rolling window, with the tuning parameters based on a search over a 10×10 lattice (with $(\lambda_B, \lambda_C) \in [0.5, 4] \times [0.2, 2]$, equal-spaced) using the BIC. It should be noted that for the majority of the rolling windows, the rank of B is 1 (data not shown). The sparsity level of the estimated C over the 143 rolling windows is depicted in Figure 3. The connectivity patterns in C show more complex and nuanced patterns than

1. The threshold is set at a relatively low level to compensate for the relative small rolling window size.

Figure 2: Global clustering coefficient of estimated A over different periods

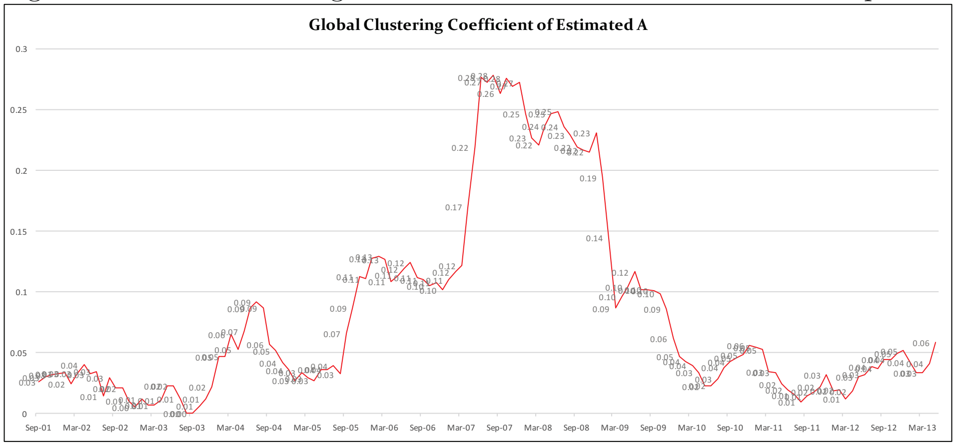
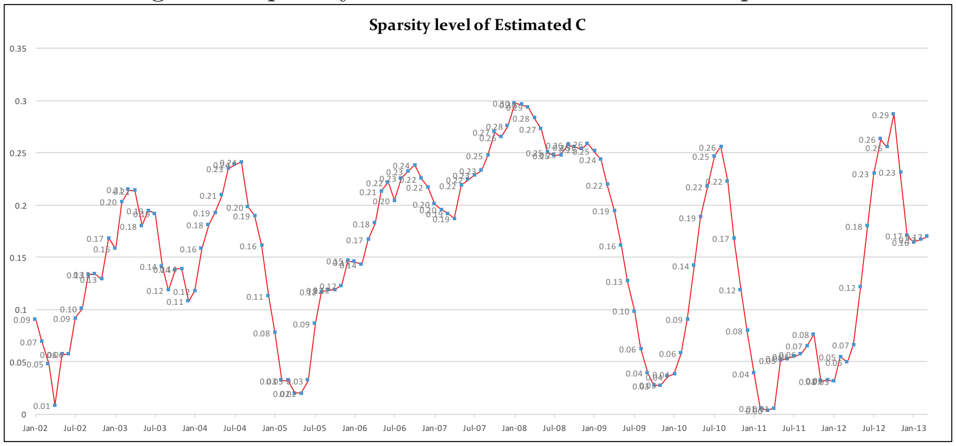


Figure 3: Sparsity of estimated C over different periods



for stocks. A more detailed analysis of the various peaks identified through our model is given in Supplementary Material **??**.

Based on the previous findings, we partition the time frame spanning 2001-2016 into the following periods: pre- (2001/07–2007/03), during- (2007/01–2009/12) and post-crisis (2010/01-2016/06) one. We estimate the model parameters using the data within the entire sub-period(s).

The estimation procedure of the transition matrix A for different periods is identical to that described above using subsamples over rolling-windows. For the pre- and post- crisis periods, since we have 76 and 77 samples respectively, the stability selection threshold is set at 0.75, whereas for the during-crisis period, at 0.6 to compensate for the small sample size (36). Table 6 shows the average R-square for all 71 stocks, as well as its standard deviation, which is calculated based on in-sample fit; i.e.,the proportion of variation explained by using the VAR(1) model to fit the data. The overall sparsity level and the spectral radius of the estimated transition matrices A are also presented. The results are consistent with the previous finding of increased connectivity during the crisis. Further, for all periods the estimate of the spectral radius is fairly large, indicating strong temporal dependence of the log-returns.

Table 6: Summary for estimated A within different periods.

	2001/07–2007/03	2007/01–2009/12	2010/01–2016/06
Averaged R sq	0.31	<u>0.72</u>	0.28
Sd of R sq	0.103	0.105	0.094
Sparsity level of $\widehat{A}$	0.17	0.23	0.19
Spectral radius of $\widehat{A}$	0.67	0.90	0.75

Next, we focus on the key motivation for developing the proposed modeling framework, namely the inter-dependence of stocks and macroeconomic variables over the specified three sub-periods. The p-value for testing the hypothesis of lack of block "Granger causality" $H_0 : B = 0$, together with the spectral radius and the sparsity level for the estimated C transition matrices are listed in Table 7. Specifically, for all three periods, the rank of estimated B is 1, indicating that the stock market as captured by its leading stocks, "Granger-causes" the temporal evolution of the macroeconomic variables. The fact that the rank of B is 1, indicates that the inter-block influence can be captured as a single portfolio acting in unison. To investigate the relative importance of each sector in the portfolio, we group the stocks by sectors. The proportion of each sector (up to normalization) is obtained by summing up the loadings (first right singular vector of the estimated B) of the stocks within this sector, weighted by their market capitalization (results shown in Supplementary Material **??**).

Next, we discuss some key relationships emerging from the model. We start with total employment (ETTL), whose dynamics are only influenced by its own past values as seen by the lack of an incoming arrow in Figure **??**. Further, an examination of the left singular vector (see Table 8) of B strongly indicates the impact of the stock market on total employment. This finding is consistent with the analysis in Farmer (2015), which argues that the crash of the stock market provides a plausible explanation for the great recession. However, the analysis in Farmer (2015) is based on bivariate VAR models involving only employment and the stock index. Therefore, there is a possibility that the stock market is reacting to some other information captured by other macroeconomic variables, such as GDP, capital spending, inflation, interest rates, etc. However, our high-dimensional VAR model simultaneously analyzes a key set of macroeconomic variables and also accounts for the influence of the largest stocks in the market. Hence, it automatically overcomes the criticism leveraged by Sims (1992) about misinterpretations of findings from small scale VAR models due to the omission of important variables, and further echoed in the discussion in Bernanke et al. (2005). Other interesting findings are presented in Supplementary Material **??**.

Table 7: Summary for estimated B and C within different periods.

	2001/07–2007/03	2007/01–2009/12	2010/01–2016/16
p-value for testing $H_0 : B = 0$	0.075	0.009	0.044
Sparsity level of $\widehat{C}$	0.06	0.25	0.06
Spectral radius of $\widehat{C}$	0.35	0.76	0.40

Table 8: Left Singular Vectors of Estimated B for different periods

	Pre-Crisis	During-Crisis	Post-Crisis
FFR	-0.24	-0.26	-0.23
T10yr	-0.09	0.14	0.16
UNEMPL	-0.07	0.01	-0.07
IPI	-0.43	0.34	0.26
ETTL	0.33	0.24	0.13
M1	0.23	-0.12	-0.47
AHES	-0.01	0.30	0.17
CU	-0.49	0.32	0.27
M2	0.10	-0.04	-0.32
HS	0.51	-0.02	-0.02
EX	-0.18	0.41	0.06
PCEQI	-0.07	-0.18	0.41
GDP	0.10	-0.02	0.05
PCEPI	0.00	0.14	-0.01
PPI	-0.15	0.00	0.06
CPI	0.01	0.15	-0.31
SP.IND	-0.06	-0.53	0.38

Remark 13 *We also applied our multi-block model with the first block X_t corresponding to the macro-economic variables and the second block Z_t the stocks variables (results not shown). The key question is whether there is also "Granger causality" from the broader economy to the stock market. The results are inconclusive due to sample size issues that do not allow us to properly test for the key hypothesis whether $B = 0$ or not. Specifically, the length of the sub-periods is short compared to the dimensionality required for the test procedure. A similar issue arises, which is related to the detection boundary for the sparse*

testing procedure during the crisis period. Further, for a sparse B, an examination of its entries shows that Employment Total did not impact the stock market, which is in line with the conclusion reached at the aggregate level by Farmer (2015). On the other hand, GDP negatively impacts stock log-returns, which may act as a leading indicator for suppressed investment and business growth and hence future stock returns.

7. Discussion.

We briefly discuss generalizations of the model to the case of more than two blocks, as mentioned in the introductory section. For the sake of concreteness, consider a triangular recursive linear dynamical system given by:

$$\begin{aligned} X_t^{(1)} &= A_{11}X_{t-1}^{(1)} + \epsilon_t^{(1)}, \\ X_t^{(2)} &= A_{12}X_{t-1}^{(1)} + A_{22}X_{t-1}^{(2)} + \epsilon_t^{(2)}, \\ X_t^{(3)} &= A_{13}X_{t-1}^{(1)} + A_{23}X_{t-1}^{(2)} + A_{33}X_{t-1}^{(3)} + \epsilon_t^{(3)}, \\ &\vdots \end{aligned} \tag{28}$$

where $X^{(j)} \in \mathbb{R}^{p_j}$ denotes the variables in group j, A_{ij} $(i < j)$ encodes the dependency of $X^{(j)}$ on the past values of variables in group i, and A_{jj} encodes the dependency on its own past values. Further, $\{\epsilon_t^{(j)}\}$ is the innovation process that is neither temporally, nor cross-sectionally correlated, i.e.,

$$\mathrm{Cov}(\epsilon_t^{(j)}, \epsilon_s^{(j)}) = 0 \ (s \neq t), \quad \mathrm{Cov}(\epsilon_t^{(i)}, \epsilon_s^{(j)}) = 0 \ (i \neq j, \ \forall \ (s,t)), \quad \mathrm{Cov}(\epsilon_t^{(j)}, \epsilon_t^{(j)}) = \big(\Omega^{(j)}\big)^{-1},$$

with $\Omega^{(j)}$ capturing the conditional contemporaneous dependency of variables within group j. The model in (28) can also be viewed from a multi-layered time-varying network perspective: nodes in each layer are "Granger-caused" by nodes from its previous layers, and are also dependent on its own past values. As previously mentioned, in various real applications, it is of interest to obtain estimates of the transition matrices, and/or test if "Granger-causality" is present between interacting blocks; i.e., to test $A_{ij} = 0$ for some $i \neq j$.

The triangular structure of the system decouples the estimation of the transition matrices from each equation, and hence a straightforward extension of the estimation procedure presented in Section 2.1 becomes applicable. Specifically, to obtain estimates of the transition matrices A_{ij}'s for fixed j and $1 \leq i \leq j$, and the inverse covariance $\Omega^{(j)}$, the optimization problem is formulated as follows:

$$\begin{aligned} (\{\widehat{A}_{ij}\}_{i\leq j}, \widehat{\Omega}^{(j)}) = \underset{A_{ij},\Omega^{(j)}}{\arg\min} \Big\{ -\log\det\Omega^{(j)} + \frac{1}{T}\sum_{t=1}^{T}\big(x_t^{(j)} - \sum_{i=1}^{j} A_{ij}x_{t-1}^{(i)}\big)'\Omega^{(j)}\big(x_t^{(j)} - \sum_{1\leq i\leq j} A_{ij}x_{t-1}^{(i)}\big) \\ + \sum_{i=1}^{j}\mathcal{R}(A_{ij}) + \rho^{(j)}\|\Omega^{(j)}\|_{1,\text{off}}\Big\}, \end{aligned} \tag{29}$$

where the exact expression for the $\mathcal{R}(A_{ij})$ adapts to the structural assumption imposed on the corresponding transition matrix (sparse/low-rank). Solving (29) again requires an iterative algorithm involving the alternate update between transition matrices and the inverse covariance matrices. Further, for updating the values of the transition matrices, a cyclic block-coordinate updating procedure is used.

Consistency results can be established analogously to those provided in Section 3, under the posited conditions of restricted strong convexity (RSC) and a deviation bound. With a larger number of interacting blocks of variables, lower bounds for the lower extremes of the spectra involve all corresponding transition matrices. The error rates that can be obtained are as follows: (i) if equation k only involves sparse transition matrices, then the finite-sample bounds of the transition matrices in this layer in Frobenius norm are of the order $O\big(\sqrt{\frac{\log p_k + \log \sum_{i\leq k} p_k}{T}}\big)$, while (ii) if some of the transition matrices are assumed low rank, then the corresponding finite sample bounds are of the order $O\big(\sqrt{\frac{p_k + \sum_{i\leq k} p_k}{T}}\big)$.

Another generalization that can be handled algorithmically with the same estimation procedure discussed above is the presence of d-lags in the specification of the linear dynamical system. Based on the consistency results developed in this work, together with the theoretical findings for VAR(d) models presented in Basu and Michailidis (2015), we expect all the established theoretical properties of the transition matrices estimates to go through under appropriate RSC and deviation bound conditions.

Acknowledgments

The authors would like to thank Action Editor Jie Peng and two anonymous referees for many constructive comments and suggestions. The work of GM was supported in part by NSF grants CCF-1540093, DMS-1545277 and IIS-1632730 and by NIH grant 5-R01-GM11402902.

Appendix A. Additional Theorems and Proofs for Theorems.

In this section, we introduce two additional theorems that respectively establish the consistency properties for the initializers $\widehat{A}^{(0)}$ and $(\widehat{B}^{(0)}, \widehat{C}^{(0)})$, for *fixed* realizations of the processes $\{X_t\}$ and $\{Z_t\}$. Specifically, $\widehat{A}^{(0)}$ and $(\widehat{B}^{(0)}, \widehat{C}^{(0)})$ are solutions to the following optimization problems:

$$\widehat{A}^{(0)} := \arg\min_{A} \big\{ \tfrac{1}{T} |||\mathcal{X}^T - \mathcal{X}A'|||_F + \lambda_A \|A\|_1 \big\}, \tag{30}$$

$$(\widehat{B}^{(0)}, \widehat{C}^{(0)}) := \arg\min_{B,C} \big\{ \tfrac{1}{T} |||\mathcal{Z}^T - \mathcal{X}B' - \mathcal{Z}C'|||_F + \lambda_B |||B|||_* + \lambda_C \|C\|_1 \big\}. \tag{31}$$

Note that they also correspond to estimators of the setting where there is no contemporaneous dependence among the idiosyncratic error processes. If we additionally introduce operators $\mathfrak{X}_0$ and $\mathfrak{W}_0$ defined as

$$\begin{aligned} \mathfrak{X}_0: \quad & \mathfrak{X}_0(\Delta) = \mathcal{X}'\Delta, \quad \text{for } \Delta \in \mathbb{R}^{p_1 \times p_1}, \\ \mathfrak{W}_0: \quad & \mathfrak{W}_0(\Delta) = \mathcal{W}'\Delta, \quad \text{for } \Delta \in \mathbb{R}^{p_2 \times (p_1+p_2)} \quad \text{where } \mathcal{W} := [\mathcal{X}, \mathcal{Z}], \end{aligned}$$

then (30) and (31) can be equivalently written as

$$\begin{aligned} \widehat{A}^{(0)} &:= \arg\min_{A} \big\{ \tfrac{1}{T} |||\mathcal{X}^T - \mathfrak{X}_0(A)|||_F + \lambda_A \|A\|_1 \big\}, \\ (\widehat{B}^{(0)}, \widehat{C}^{(0)}) &:= \arg\min_{B,C} \big\{ \tfrac{1}{T} |||\mathcal{Z}^T - \mathfrak{W}_0(B_{\text{aug}}, C_{\text{aug}})|||_F + \lambda_B |||B|||_* + \lambda_C \|C\|_1 \big\}, \end{aligned}$$

where $B_{\text{aug}} := [B, O_{p_2 \times p_2}], C_{\text{aug}} := [O_{p_2 \times p_1}, C]$.

Theorem 14 (Error bounds for $\widehat{A}^{(0)}$) *Suppose the operator $\mathfrak{X}_0$ satisfies the RSC condition with norm $\Phi(\Delta) = \|\Delta\|_1$, curvature $\alpha_{RSC} > 0$ and tolerance $\tau > 0$, so that*

$$s_A^\star \tau \leq \alpha_{RSC}/32.$$

Then, with regularization parameter λ_A satisfying $\lambda_A \geq 4\|\mathcal{X}'\mathcal{U}/T\|_\infty$, the solution to (30) satisfies the following bounds:

$$|||\widehat{A}^{(0)} - A^\star|||_F \leq 12\sqrt{s_A^\star}\lambda_A/\alpha_{RSC} \qquad \text{and} \qquad \|\widehat{A} - A^\star\|_1 \leq 48 s_A^\star \lambda_A/\alpha_{RSC}.$$

Theorem 15 (Error bound for $(\widehat{B}^{(0)}, \widehat{C}^{(0)})$) *Let $\mathcal{J}_{C^\star}$ be the support set of $C^\star$ and $s_C^\star$ denote its cardinality. Let $r_B^\star$ be the rank of $B^\star$. Assume that $\mathfrak{W}_0$ satisfies the RSC condition with norm*

$$\Phi(\Delta) := \inf_{B_{aug} + C_{aug} = \Delta} \mathcal{Q}(B, C), \qquad \text{where} \quad \mathcal{Q}(B, C) := |||B|||_* + \tfrac{\lambda_C}{\lambda_B}\|C\|_1,$$

curvature α_{RSC} and tolerance τ such that

$$128\tau r_B^\star < \alpha_{RSC}/4 \qquad \text{and} \qquad 64\tau s_C^\star (\lambda_C/\lambda_B)^2 < \alpha_{RSC}/4.$$

Then, with regularization parameters λ_B and λ_C satisfying

$$\lambda_B \geq 4 \big|\big|\big| \mathcal{W}'\mathcal{V}/T \big|\big|\big|_{op} \quad \text{and} \quad \lambda_C \geq 4 \big\| \mathcal{W}'\mathcal{V}/T \big\|_\infty,$$

the solution to (31) satisfies the following bounds:

$$|||\widehat{B}^{(0)} - B^\star|||_F^2 + |||\widehat{C}^{(0)} - C^\star|||_F^2 \leq 4\left(2r_B^\star\lambda_B^2 + s_C^\star\lambda_C^2\right)/\alpha_{RSC}^2.$$

In the rest of this subsection, we first prove Theorem 14 and 15, then prove Theorem 5 and 6, whose statements are given in Section 3.2.

Proof [**Proof of Theorem 14**]

For the ease of notation, in this proof, we use $\widehat{A}$ to refer to $\widehat{A}^{(0)}$ whenever there is no ambiguity. Let $\beta_A^\star = \mathrm{vec}(A^\star)$ and denote the residual matrix and its vectorized version by $\Delta_A = \widehat{A} - A^\star$ and $\Delta_{\beta_A} = \widehat{\beta}_A - \beta_A^\star$, respectively. By the optimality of $\widehat{A}$ and the feasibility of $A^\star$, the following *basic inequality* holds:

$$\tfrac{1}{T}|||\mathfrak{X}_0(\Delta_A)|||_F^2 \leq \tfrac{2}{T}\langle\!\langle \Delta_A, \mathcal{X}'\mathcal{U}\rangle\!\rangle + \lambda_A\left\{\|A^\star\|_1 - \|A^\star + \Delta_A\|_1\right\},$$

which is equivalent to:

$$\Delta_{\beta_A}'\widehat{\Gamma}_X^{(0)}\Delta_{\beta_A} \leq \tfrac{2}{T}\langle \Delta_{\beta_A}, \mathrm{vec}(\mathcal{X}'\mathcal{U})\rangle + \lambda_A\left\{\|\beta_A^\star\|_1 - \|\beta_A^\star + \Delta_{\beta_A}\|_1\right\}, \tag{32}$$

where $\widehat{\Gamma}_X^{(0)} = \mathrm{I}_{p_1} \otimes \frac{\mathcal{X}'\mathcal{X}}{T}$. By Hölder's inequality and the triangle inequality, an upper bound for the right-hand-side of (32) is given by

$$\tfrac{2}{T}\|\Delta_{\beta_A}\|_1\big\|\mathcal{X}'\mathcal{U}\big\|_\infty + \lambda_A\|\Delta_{\beta_A}\|_1. \tag{33}$$

Now with the specified choice of λ_A, by Lemma **??**, $\|\Delta_{\beta_A|\mathcal{J}_{A^\star}}\|_1 \leq 3\|\Delta_{\beta_A|\mathcal{J}^c_{A^\star}}\|_1$ i.e., $\Delta_{\beta_A} \in \mathcal{C}(\mathcal{J}_{A^\star}, 3)$, hence $\|\Delta_{\beta_A}\|_1 \leq 4\|\Delta_{\beta_A|\mathcal{J}_{A^\star}}\|_1 \leq 4\sqrt{s_A^\star}\|\Delta_{\beta_A}\|$. By choosing $\lambda_A \geq 4\|\mathcal{X}'\mathcal{U}/T\|_\infty$, (33) is further upper bounded by

$$\frac{3}{2}\lambda_A\|\Delta_{\beta_A}\|_1 \leq 6\sqrt{s_A^\star}\lambda_A\|\Delta_{\beta_A}\|. \tag{34}$$

Combined with the RSC condition and the upper bound given in (34), we have

$$\begin{aligned}
\frac{\alpha_{\mathrm{RSC}}}{2}\|\Delta_{\beta_A}\|^2 - \frac{\tau}{2}\|\Delta_{\beta_A}\|_1^2 &\leq \frac{1}{2}\Delta_{\beta_A}'\widehat{\Gamma}_X^{(0)}\Delta_{\beta_A} \leq 3\sqrt{s_A^\star}\lambda_A\|\Delta_{\beta_A}\|,\\
\frac{\alpha_{\mathrm{RSC}}}{4}\|\Delta_{\beta_A}\|^2 &\leq \left(\frac{\alpha_{\mathrm{RSC}}}{2} - \frac{16 s_A^\star\tau}{4}\right)\|\Delta_{\beta_A}\|^2 \leq 3\sqrt{s_A^\star}\lambda_A\|\Delta_{\beta_A}\|,
\end{aligned}$$

which implies

$$\|\Delta_{\beta_A}\| \leq 12\sqrt{s_A^\star}\lambda_A/\alpha_{\mathrm{RSC}} \qquad \text{and} \qquad \|\Delta_{\beta_A}\|_1 \leq 48 s_A^\star\lambda_A/\alpha_{\mathrm{RSC}}.$$

It is easy to see that these bounds also hold for $\|\Delta_A\|_F$ and $\|\Delta_A\|_1$, respectively. ■

Next, to prove Theorem 15, we introduce the following two sets of subspaces $\{\mathcal{S}_\Theta, \mathcal{S}_\Theta^\perp\}$ and $\{\mathcal{R}_\Theta, \mathcal{R}_\Theta^c\}$ associated with some generic matrix $\Theta \in \mathbb{R}^{m_1\times m_2}$, in which the nuclear norm and the ℓ_1-norm are decomposable, respectively (see Negahban et al., 2012). Specifically, let the singular value decomposition of Θ be $\Theta = U\Sigma V'$ with U and V being orthogonal matrices. Let $r = \mathrm{rank}(\Theta)$, and we use U^r and V^r to denote the first r columns of U and V associated with the r singular values of Θ. Further, define

$$\begin{aligned}
\mathcal{S}_\Theta &:= \left\{\Delta \in \mathbb{R}^{m_1\times m_2} | \mathrm{row}(\Delta) \subseteq V^r \quad \text{and} \quad \mathrm{col}(\Delta) \subseteq U^r\right\},\\
\mathcal{S}_\Theta^\perp &:= \left\{\Delta \in \mathbb{R}^{m_1\times m_2} | \mathrm{row}(\Delta) \perp V^r \quad \text{and} \quad \mathrm{col}(\Delta) \perp U^r\right\}.
\end{aligned} \tag{35}$$

Then, for an arbitrary (generic) matrix $M \in \mathbb{R}^{m_1 \times m_2}$, its restriction on the subspace $\mathcal{S}(\Theta)$ and $\mathcal{S}^{\perp}(\Theta)$, denoted by $M_{\mathcal{S}(\Theta)}$ and $M_{\mathcal{S}^{\perp}(\Theta)}$ respectively, are given by:

$$M_{\mathcal{S}_\Theta} = U \begin{bmatrix} \widetilde{M}_{11} & \widetilde{M}_{12} \\ \widetilde{M}_{21} & O \end{bmatrix} V' \qquad \text{and} \qquad M_{\mathcal{S}_\Theta^\perp} = U \begin{bmatrix} O & O \\ O & \widetilde{M}_{22} \end{bmatrix} V',$$

where $\Theta = U\Sigma V'$ and $\widetilde{M}$ is defined and partitioned as

$$\widetilde{M} = U'MV = \begin{bmatrix} \widetilde{M}_{11} & \widetilde{M}_{12} \\ \widetilde{M}_{21} & \widetilde{M}_{22} \end{bmatrix}, \quad \text{where} \ \ \widetilde{M}_{11} \in \mathbb{R}^{r \times r}.$$

Note that by Lemma **??**, $M_{\mathcal{S}_\Theta} + M_{\mathcal{S}_\Theta^\perp} = M$. Moreover, when Θ is restricted to the subspace induced by itself $\Theta_{\mathcal{S}_\Theta}$ (and we write $\Theta_{\mathcal{S}}$ for short for this specific case), the following decomposition for the nuclear norm holds:

$$|||\Theta|||_* = |||\Theta_{\mathcal{S}} + \Theta_{\mathcal{S}^\perp}|||_* = |||\Theta_{\mathcal{S}}|||_* + |||\Theta_{\mathcal{S}^\perp}|||_*.$$

Let $\mathcal{J}(\Theta)$ be the set of indexes in which Θ is nonzero. Analogously, we define

$$\begin{aligned} \mathcal{R}_\Theta &:= \left\{ \Delta \in \mathbb{R}^{m_1 \times m_2} | \Delta_{ij} = 0 \ \ \text{for} \ \ (i,j) \notin \mathcal{J}(\Theta) \right\}, \\ \mathcal{R}_\Theta^c &:= \left\{ \Delta \in \mathbb{R}^{m_1 \times m_2} | \Delta_{ij} = 0 \ \ \text{for} \ \ (i,j) \in \mathcal{J}(\Theta) \right\}. \end{aligned} \tag{36}$$

Then, for an arbitrary matrix M, $M_{\mathcal{J}_\Theta} \in \mathcal{R}_\Theta$ is obtained by setting the entries of M whose indexes are not in $\mathcal{J}(\Theta)$ to 0, and $M_{\mathcal{J}_\Theta^c} \in \mathcal{R}_\Theta^c$ is obtained by setting the entries of M whose indexes are in $\mathcal{J}(\Theta)$ to 0. Then, the following decomposition holds:

$$\left\| M_{\mathcal{J}_\Theta} + M_{\mathcal{J}_\Theta^c} \right\|_1 = \| M_{\mathcal{J}_\Theta} \|_1 + \left\| M_{\mathcal{J}_\Theta^c} \right\|_1.$$

Proof [**Proof of Theorem 15**] Again for the ease of notation, in this proof, we drop the superscript and use $(\widehat{B}^{(0)}, \widehat{C}^{(0)})$ to denote $(\widehat{B}, \widehat{C})$ whenever there is no ambiguity. Define $\mathcal{Q}$ to be the weighted regularizer:

$$\mathcal{Q}(B, C) = |||B|||_* + \tfrac{\lambda_C}{\lambda_B} \|C\|_1.$$

Note that $(B^\star, C^\star)$ is always feasible, and by the optimality of $(\widehat{B}, \widehat{C})$, the following inequality holds:

$$\frac{1}{T} |||\mathcal{Z}^T - \mathfrak{W}_0(\widehat{B}_{\text{aug}} + \widehat{C}_{\text{aug}})|||_F^2 + \lambda_B |||\widehat{B}|||_* + \lambda_C \|\widehat{C}\|_1 \le \frac{1}{T} |||\mathcal{Z}^T - \mathfrak{W}_0(B^\star + C^\star)|||_F^2 + \lambda_B |||B^\star|||_* + \lambda_C \|C^\star\|_1,$$

By defining $\Delta_{\text{aug}}^B = \widehat{B}_{\text{aug}} - B_{\text{aug}}^\star = [\Delta^B, O]$, $\Delta_{\text{aug}}^C = \widehat{C}_{\text{aug}} - C_{\text{aug}}^\star = [O, \Delta^C]$, we obtain the following *basic inequality*:

$$\tfrac{1}{T} \left|\left|\left| \mathfrak{W}_0(\Delta_{\text{aug}}^B + \Delta_{\text{aug}}^C) \right|\right|\right|_F^2 \le \tfrac{2}{T} \langle\!\langle \Delta_{\text{aug}}^B + \Delta_{\text{aug}}^C, \mathcal{W}'\mathcal{V} \rangle\!\rangle + \lambda_B \mathcal{Q}(B^\star, C^\star) - \lambda_B \mathcal{Q}(\widehat{B}, \widehat{C}). \tag{37}$$

By Hölder's inequality and Lemma 21, we have

$$\begin{aligned} \tfrac{1}{T} \left|\left|\left| \mathfrak{W}_0(\Delta_{\text{aug}}^B + \Delta_{\text{aug}}^C) \right|\right|\right|_F^2 \le{} & \tfrac{2}{T} \left(\left|\left|\left| \Delta_{\mathcal{S}_{B^\star}}^B \right|\right|\right|_* + \left|\left|\left| \Delta_{\mathcal{S}_{B^\star}^\perp}^B \right|\right|\right|_* \right) |||\mathcal{W}'\mathcal{V}|||_{op} + \tfrac{2}{T} \left(\left\| \Delta_{\mathcal{J}_{C^\star}^c}^C \right\|_1 + \left\| \Delta_{\mathcal{J}_{C^\star}^c}^C \right\|_1 \right) \|\mathcal{W}'\mathcal{V}\|_\infty \\ & + \lambda_B \mathcal{Q}(\Delta_{\mathcal{S}_{B^\star}}^B, \Delta_{\mathcal{J}_{C^\star}}^C) - \lambda_B \mathcal{Q}(\Delta_{\mathcal{S}_{B^\star}^\perp}^B, \Delta_{\mathcal{J}_{C^\star}^c}^C). \end{aligned} \tag{38}$$

With the specified choice of λ_B and λ_C, after some algebra, (38) is further bounded by

$$\frac{3\lambda_B}{2}\mathcal{Q}(\Delta^B_{\mathcal{S}_{B^\star}}, \Delta^C_{\mathcal{J}_{C^\star}}) - \frac{\lambda_B}{2}\mathcal{Q}(\Delta^B_{\mathcal{S}^\perp_{B^\star}}, \Delta^C_{\mathcal{J}^c_{C^\star}}).$$

By Lemma 22, and using this upper bound, we obtain

$$\frac{\alpha_{\text{RSC}}}{2}(|||\Delta^B|||_F^2 + |||\Delta^C|||_F^2) - \frac{\lambda_B}{2}\mathcal{Q}(\Delta^B, \Delta^C) \le \frac{3\lambda_B}{2}\mathcal{Q}(\Delta^B_{\mathcal{S}_{B^\star}}, \Delta^C_{\mathcal{J}_{C^\star}}) - \frac{\lambda_B}{2}\mathcal{Q}(\Delta^B_{\mathcal{S}^\perp_{B^\star}}, \Delta^C_{\mathcal{J}^c_{C^\star}}).$$

By the triangle inequality, $\mathcal{Q}(\Delta^B, \Delta^C) \le \mathcal{Q}(\Delta^B_{\mathcal{S}_{B^\star}}, \Delta^C_{\mathcal{J}_{C^\star}}) + \mathcal{Q}(\Delta^B_{\mathcal{S}^\perp_{B^\star}}, \Delta^C_{\mathcal{J}^c_{C^\star}})$, rearranging gives

$$\frac{\alpha_{\text{RSC}}}{2}(|||\Delta^B|||_F^2 + |||\Delta^C|||_F^2) \le 2\lambda_B\mathcal{Q}(\Delta^B_{\mathcal{S}_{B^\star}}, \Delta^C_{\mathcal{J}_{C^\star}}). \tag{39}$$

By Lemma **??**, with $N = B^\star$, $M_1 = \Delta^B_{\mathcal{S}_{B^\star}}$, $M_2 = \Delta^B_{\mathcal{S}^\perp_{B^\star}}$, we get

$$\text{rank}(\Delta^B_{\mathcal{S}_{B^\star}}) \le 2r_B^\star \qquad \text{and} \qquad \langle\!\langle \Delta^B_{\mathcal{S}_{B^\star}}, \Delta^B_{\mathcal{S}^\perp_{B^\star}} \rangle\!\rangle = 0,$$

which implies $|||\Delta^B_{\mathcal{S}_{B^\star}}|||_* \le (\sqrt{2r_B^\star})|||\Delta^B_{\mathcal{S}_{B^\star}}|||_F \le (\sqrt{2r_B^\star})|||\Delta^B|||_F$. Since $\Delta^C_{\mathcal{J}_{C^\star}}$ has at most $s_C^\star$ nonzero entries, it follows that $\|\Delta^C_{\mathcal{J}_{C^\star}}\|_1 \le \sqrt{s_C^\star}|||\Delta^C_{\mathcal{J}_{C^\star}}|||_F \le \sqrt{s_C^\star}|||\Delta^C|||_F$. Therefore,

$$\mathcal{Q}(\Delta^B_{\mathcal{S}_{B^\star}}, \Delta^C_{\mathcal{J}_{C^\star}}) = \lambda_B|||\Delta^B_{\mathcal{S}_{B^\star}}|||_* + \lambda_C\|\Delta^C_{\mathcal{J}_{C^\star}}\|_1 \le \lambda_B(\sqrt{2r_B^\star})|||\Delta^B|||_F + \lambda_C(\sqrt{s_C^\star})|||\Delta^C|||_F$$

With an application of the Cauchy-Schwartz inequality, (39) yields:

$$\frac{\alpha_{\text{RSC}}}{2}(|||\Delta^B|||_F^2 + |||\Delta^C|||_F^2) \le \sqrt{2r_B^\star\lambda_B^2 + s_C^\star\lambda_C^2} * \sqrt{|||\Delta^B|||_F^2 + |||\Delta^C|||_F^2}$$

and we obtain the following bound:

$$|||\Delta^B|||_F^2 + |||\Delta^C|||_F^2 \le 4\left(2r_B^\star\lambda_B^2 + s_C^\star\lambda_C^2\right)/\alpha_{\text{RSC}}^2.$$

■

Proof [**Proof of Theorem 5**] At iteration 0, $\widehat{A}^{(0)}$ solves the following optimization problem:

$$\widehat{A}^{(0)} = \underset{A\in\mathbb{R}^{p_1\times p_1}}{\arg\min}\left\{\tfrac{1}{T}|||\mathcal{X}^T - \mathcal{X}A'|||_F^2 + \lambda_A|||A|||_*\right\}.$$

By Theorem 14, its error bound is given by

$$\|\widehat{A}^{(0)} - A^\star\|_1 \le 48 s_A^\star\lambda_A/\alpha_{\text{RSC}},$$

provided that $\widehat{\Gamma}^{(0)}_X = \mathrm{I}_{p_1}\otimes\mathcal{X}'\mathcal{X}/T$ satisfies the RSC condition, and the regularization parameter λ_A satisfies $\lambda_A \ge 4\|\mathcal{X}'\mathcal{U}/T\|_\infty$. For random realizations $\mathcal{X}$ and $\mathcal{U}$, by Lemma 16 and Lemma 17, there exist constants c_i and c_i' such that for sample size $T \gtrsim s_A^\star\log p_1$, with probability at least $1 - c_1\exp(-c_2 T\min\{1, \omega^{-2}\})$, where $\omega = c_3\mu_{\max}(\mathcal{A})/\mu_{\min}(\mathcal{A})$

($\mathbf{E_1}$) $\quad \widehat{\Gamma}^{(0)}_X$ satisfies RSC condition with $\alpha_{\text{RSC}} = \Lambda_{\min}(\Sigma_u^\star)/(2\mu_{\max}(\mathcal{A}))$, and with probability at least $1-c_1'\exp(-c_2'\log p_1)$,

$$(\mathbf{E_2})\qquad \|\mathcal{X}'\mathcal{U}/T\|_\infty \le C_0\sqrt{\frac{\log p_1}{T}},\qquad \text{for some constant } C_0.$$

Hence with probability at least $1-c_1\exp(-c_2T)-c_1'\exp(-c_2'\log p_1)$,

$$\|\widehat{A}^{(0)}-A^\star\|_1 = O\Big(s_A^\star\sqrt{\tfrac{\log p_1}{T}}\Big).$$

Moving onto $\widehat{\Omega}_u^{(0)}$, which is given by

$$\widehat{\Omega}_u^{(0)} = \underset{\Omega_u\in\mathbb{S}_{++}^{p_1\times p_1}}{\arg\min}\left\{\log\det\Omega_u - \mathrm{trace}\big(\widehat{S}_u^{(0)}\Omega_u\big)+\rho_u\|\Omega_u\|_{1,\text{off}}\right\},$$

where $\widehat{S}_u^{(0)} = \frac{1}{T}(\mathcal{X}^T-\mathcal{X}\widehat{A}^{(0)\prime})'(\mathcal{X}^T-\mathcal{X}\widehat{A}^{(0)\prime})$. By Theorem 1 in Ravikumar et al. (2011), the error bound for $\widehat{\Omega}_u^{(0)}$ relies on how well $\widehat{S}_u^{(0)}$ concentrates around $\Sigma_u^\star$, more specifically, $\|\widehat{S}_u^{(0)}-\Sigma_u^\star\|_\infty$. Note that

$$\|\widehat{S}_u^{(0)}-\Sigma_u^\star\|_\infty \le \|S_u-\Sigma_u^\star\|_\infty + \|\widehat{S}_u^{(0)}-S_u\|_\infty,$$

where $S_u=\mathcal{U}'\mathcal{U}/T$ is the sample covariance based on true errors. For the first term, by Ravikumar et al. (2011), there exists constant $\tau_0>2$ such that with probability at least $1-1/p_1^{\tau_0-2}=1-\exp(-\tau\log p_1)$ $(\tau>0)$, the following bound holds:

$$(\mathbf{E_3})\qquad \|S_u-\Sigma_u^\star\|_\infty \le C_1\sqrt{\frac{\log p_1}{T}},\qquad \text{for some constant } C_1.$$

For the second term,

$$\widehat{S}_u^{(0)}-S_u = \tfrac{2}{T}\mathcal{U}'\mathcal{X}(A^\star-\widehat{A}^{(0)})' + (A^\star-\widehat{A}^{(0)})\left(\tfrac{\mathcal{X}'\mathcal{X}}{T}\right)(A^\star-\widehat{A}^{(0)})' := I_1+I_2,$$

For I_1, based on the analysis of $\|A^\star-\widehat{A}^{(0)}\|_1$ and $\|\mathcal{X}'\mathcal{U}/T\|_\infty$,

$$\|I_1\|_\infty \le 2\|\!|A^\star-\widehat{A}^{(0)}|\!\|_\infty\|\tfrac{1}{T}\mathcal{X}'\mathcal{U}\|_\infty \le 2\|A^\star-\widehat{A}^{(0)}\|_1\|\tfrac{1}{T}\mathcal{X}'\mathcal{U}\|_\infty = O\Big(\tfrac{s_A^\star\log p_1}{T}\Big)$$

For I_2,

$$\begin{aligned}\|(A^\star-\widehat{A}^{(0)})\left(\tfrac{\mathcal{X}'\mathcal{X}}{T}\right)(A^\star-\widehat{A}^{(0)})'\|_\infty &\le \|\!|A^\star-\widehat{A}^{(0)}|\!\|_\infty\|\!|A^\star-\widehat{A}^{(0)}|\!\|_1\|\tfrac{\mathcal{X}'\mathcal{X}}{T}\|_\infty\\ &\le \|A^\star-\widehat{A}^{(0)}\|_1^2\|\tfrac{\mathcal{X}'\mathcal{X}}{T}\|_\infty,\end{aligned}$$

where by Proposition 2.4 in Basu and Michailidis (2015) and then taking a union bound, with probability at least $1-c_1''\exp(-c_2''\log p_1)$ $(c_1'',c_2''>0)$,

$$(\mathbf{E_4})\qquad \|\tfrac{\mathcal{X}'\mathcal{X}}{T}\|_\infty \le C_2\sqrt{\frac{\log p_1}{T}}+\Lambda_{\max}(\Gamma_X),\qquad \text{for some constant } C_2.$$

Hence,

$$\|I_2\|_\infty = O\Big((s_A^\star)^2\big(\tfrac{\log p_1}{T}\big)^{3/2}\Big) + O\Big((s_A^\star)^2\tfrac{\log p_1}{T}\Big)$$

Combining all terms, and since we assume that $T^{-1}\log p_1$ is small, $O(\sqrt{T^{-1}\log p_1})$ becomes the leading term, and the following bound holds with probability at least $1-c_1\exp(-c_2T)-c_1'\exp(-c_2'\log p_1)-c_1''\exp(-c_2''\log p_1)-\exp(-\tau\log p_1)$:

$$\|\widehat{S}_u^{(0)}-\Sigma_u^\star\|_\infty = O\Big(\sqrt{\tfrac{\log p_1}{T}}\Big).$$

Consequently,

$$\|\widehat{\Omega}_u^{(0)}-\Omega_u^\star\|_\infty = O\Big(\sqrt{\tfrac{\log p_1}{T}}\Big).$$

At iteration 1, the vectorized $\widehat{A}^{(1)}$ solves

$$\widehat{\beta}_A^{(1)} = \underset{\beta\in\mathbb{R}^{p_1^2}}{\arg\min}\big\{-2\beta'\widehat{\gamma}_X^{(1)}+\beta'\widehat{\Gamma}_X^{(1)}\beta+\lambda_A\|\beta\|_1\big\},$$

where

$$\widehat{\gamma}_X^{(1)} = \tfrac{1}{T}\big(\widehat{\Omega}_u^{(0)}\otimes\mathcal{X}'\big)\mathrm{vec}(\mathcal{X}^T),\qquad \widehat{\Gamma}_X^{(1)} = \widehat{\Omega}_u^{(0)}\otimes\tfrac{\mathcal{X}'\mathcal{X}}{T}.$$

The error bound for $\widehat{\beta}_A^{(1)}$ relies on (1) $\widehat{\Gamma}_X^{(1)}$ satisfying the RSC condition, which holds for sample size $T\gtrsim(d^{\max}_{\Omega_u^\star})^2\log p_1$ upon $\|\widehat{\Omega}_u^{(0)}-\Omega_u^\star\|_\infty = O(\sqrt{T^{-1}\log p_1})$; and (2) a bound for $\|\mathcal{X}'\mathcal{U}\widehat{\Omega}_u^{(0)}/T\|_\infty$. For $\|\mathcal{X}'\mathcal{U}\widehat{\Omega}_u^{(0)}/T\|_\infty$,

$$\tfrac{1}{T}\mathcal{X}'\mathcal{U}\widehat{\Omega}_u^{(0)} = \tfrac{1}{T}\mathcal{X}'\mathcal{U}\Omega_u^\star + \tfrac{1}{T}\mathcal{X}'\mathcal{U}(\widehat{\Omega}_u^{(0)}-\Omega_u^\star) := I_3+I_4.$$

For I_3, by Lemma 3 in Lin et al. (2016) and with the aid of Proposition 2.4 in Basu and Michailidis (2015), again with probability at least $1-c_1'''\exp(-c_2'''\log p_1)$ we get

$$(\mathbf{E_5})\qquad \big\|\tfrac{1}{T}\mathcal{X}'\mathcal{U}\Omega_u^\star\big\|_\infty \le C_3\sqrt{\frac{\log p_1}{T}},\qquad \text{for some constant } C_3.$$

For I_4, by Corollary 3 in Ravikumar et al. (2011), we get

$$\Big\|\tfrac{1}{T}\mathcal{X}'\mathcal{U}(\widehat{\Omega}_u^{(0)}-\Omega_u^\star)\Big\|_\infty \le d^{\max}_{\Omega_u^\star}\|\tfrac{1}{T}\mathcal{X}'\mathcal{U}\|_\infty\|\widehat{\Omega}_u^{(0)}-\Omega_u^\star\|_\infty = O\Big(\tfrac{\log p_1}{T}\Big).$$

Combining all terms and taking the leading one, once again we have

$$\|\widehat{A}^{(1)}-A^\star\|_1 = O\Big(s_A^\star\sqrt{\tfrac{\log p_1}{T}}\Big),$$

which holds with probability at least $1-c_1\exp(-c_2T)-\tilde{c}_1\exp(-\tilde{c}_2\log p_1)-\exp(-\tau\log p_1)$, by letting $\tilde{c}_1=\max\{c_1',c_1'',c_1'''\}$ and $\tilde{c}_1=\min\{c_2',c_2'',c_2'''\}$. It should be noted that up to this step, all sources of randomness from the random realizations have been captured by events from $\mathbf{E_1}$ to $\mathbf{E_5}$; thus, for $\widehat{\Omega}_u^{(1)}$ and iterations thereafter, the probability for which the bounds hold will no longer change, and the same holds for the error bounds for $\widehat{A}^{(k)}$ and $\widehat{\Omega}_u^{(k)}$ in terms of the relative order with respect to the dimension p_1 and sample size T. Therefore, we conclude that with high probability, for all iterations k,

$$\|\mathcal{X}'\mathcal{U}\widehat{\Omega}_u^{(k)}/T\|_\infty = O\Big(\sqrt{\tfrac{\log p_1}{T}}\Big), \qquad \|\widehat{S}_u^{(k)} - \Sigma_u^\star\|_\infty = O\Big(\sqrt{\tfrac{\log p_1}{T}}\Big).$$

With the aid of Theorem 14, it then follows that

$$|||\widehat{A}^{(k)} - A^\star|||_F = O\Big(\sqrt{\tfrac{s_A^\star \log p_1}{T}}\Big), \qquad |||\widehat{\Omega}_u^{(k)} - \Omega_u^\star|||_F = O\Big(\sqrt{\tfrac{(s_{\Omega_u^\star}+p_1)\log p_1}{T}}\Big).$$

■

Proof [**Proof of Theorem 6**] At iteration 0, $(\widehat{B}^{(0)}, \widehat{C}^{(0)})$ solves the following optimization:

$$(\widehat{B}^{(0)}, \widehat{C}^{(0)}) = \underset{(B,C)}{\arg\min} \left\{ \tfrac{1}{T} |||\mathcal{Z}^T - \mathcal{X}B' - \mathcal{Z}C'|||_F^2 + \lambda_B |||B|||_* + \lambda_C \|C\|_1 \right\}.$$

Let $W_t = (X_t', Z_t')' \in \mathbb{R}^{p_1+p_2}$ be the joint process and $\mathcal{W}$ be the realizations, with operators $\mathfrak{W}_0$ identically defined to that in Theorem 15 . By Theorem 15,

$$|||\widehat{B}^{(0)} - B^\star|||_F^2 + |||\widehat{C}^{(0)} - C^\star|||_F^2 \le 4(2r_B^\star \lambda_B^2 + s_C^\star \lambda_C^2)/\alpha_{\text{RSC}}^2,$$

provided that $\mathfrak{W}$ satisfies the RSC condition and λ_B, λ_C respectively satisfy

$$\lambda_B \ge 4\big|\big|\big|\mathcal{W}'\mathcal{V}/T\big|\big|\big|_{op} \quad \text{and} \quad \lambda_C \ge 4\big\|\mathcal{W}'\mathcal{V}/T\big\|_\infty.$$

In particular, by Lemma 18 for random realizations of $\mathcal{X}$, $\mathcal{Z}$ and $\mathcal{V}$, for sample size $T \gtrsim c_0(p_1 + 2p_2)$, with probability at least $1 - c_1\exp\{-c_2(p_1+p_2)\}$,

$$(\mathbf{E_1'}) \qquad \mathfrak{W}_0 \text{ satisfies the RSC condition.}$$

By Lemma 19, for sample size $T \gtrsim (p_1 + 2p_2)$ and some constant $C_1, C_2 > 0$,

$$(\mathbf{E_2'}) \qquad \big|\big|\big|\mathcal{W}'\mathcal{V}/T\big|\big|\big|_{op} \le C_1\sqrt{\frac{p_1+2p_2}{T}} \quad \text{and} \quad \big\|\mathcal{W}'\mathcal{V}/T\big\|_\infty \le C_2\sqrt{\frac{\log(p_1+p_2)+\log p_2}{T}},$$

with probability at least $1 - c_1'\exp\{-c_2'(p_1+2p_2)\}$ and $1 - c_1''\exp\{-c_2''\log[p_2(p_1+p_2)]\}$, respectively. Hence, with probability at least

$$1 - c_1\exp\{-c_2(p_1+p_2)\} - c_1'\exp\{-c_2'(p_1+2p_2)\} - c_1''\exp\{-c_2''\log[p_2(p_1+p_2)]\},$$

the following bound holds for the initializers as long as sample size $T \gtrsim (p_1 + 2p_2)$:

$$|||\widehat{B}^{(0)} - B^\star|||_F^2 + |||\widehat{C}^{(0)} - C^\star|||_F^2 = O\Big(\tfrac{p_1+2p_2}{T}\Big) + O\Big(\tfrac{\log(p_1+p_2)+\log p_2}{T}\Big). \tag{40}$$

Considering the estimation of $\widehat{\Omega}_v^{(0)}$, it solves a graphical Lasso problem:

$$\widehat{\Omega}_v^{(0)} = \underset{\Omega_v \in \mathbb{S}_{++}^{p_2\times p_2}}{\arg\min} \left\{ \log\det\Omega_v - \text{trace}\big(\widehat{S}_u^{(0)}\Omega_v\big) + \rho_v\|\Omega_v\|_{1,\text{off}} \right\},$$

where $\widehat{S}_v^{(0)} = \frac{1}{T}(\mathcal{Z}^T - \mathcal{X}\widehat{B}^{(0)\prime} - \mathcal{Z}\widehat{C}^{(0)\prime})'(\mathcal{Z}^T - \mathcal{X}\widehat{B}^{(0)\prime} - \mathcal{Z}\widehat{C}^{(0)\prime})$. Similar to the proof of Theorem 5, the error bound for $\widehat{\Omega}_v^{(0)}$ depends on $\|\widehat{S}_v^{(0)} - \Sigma_v^\star\|_\infty$, which can be decomposed as

$$\|\widehat{S}_v^{(0)} - \Sigma_v^\star\|_\infty \leq \|S_v - \Sigma_v^\star\|_\infty + \|\widehat{S}_v^{(0)} - S_v\|_\infty,$$

where $S_v = \mathcal{V}'\mathcal{V}/T$ is the sample covariance based on the true errors. For the first term, by Lemma 1 in Ravikumar et al. (2011), there exists constant $\tau_0 > 2$ such that with probability at least $1 - 1/p_2^{\tau_0-2} = 1 - \exp(-\tau\log p_2)$ $(\tau > 0)$, the following bound holds:

$$(\mathbf{E_3'}) \qquad \|S_v - \Sigma_v^\star\|_\infty \leq C_3\sqrt{\frac{\log p_1}{T}}, \qquad \text{for some constant } C_3.$$

For the second term, let $\Pi = [B, C] \in \mathbb{R}^{p_2\times(p_1+p_2)}$, then

$$\widehat{S}_v^{(0)} - S_v = \frac{2}{T}\mathcal{V}'\mathcal{W}(\Pi^\star - \widehat{\Pi}^{(0)})' + (\Pi^\star - \widehat{\Pi}^{(0)})\left(\tfrac{\mathcal{W}'\mathcal{W}}{T}\right)(\Pi^\star - \widehat{\Pi}^{(0)})' := I_1 + I_2,$$

For I_1, we have

$$\|\tfrac{2}{T}\mathcal{V}'\mathcal{W}(\Pi^\star - \widehat{\Pi}^{(0)})'\|_\infty \leq \|\tfrac{2}{T}\mathcal{V}'\mathcal{W}(\Pi^\star - \widehat{\Pi}^{(0)})'\|_F \leq 2|||\tfrac{1}{T}\mathcal{W}'\mathcal{V}|||_{op}|||\Pi^\star - \widehat{\Pi}^{(0)}|||_F.$$

Consider the leading term of $|||\Pi^\star - \widehat{\Pi}^{(0)}|||_F$ as in (40), whose rate is $O(\sqrt{T^{-1}(p_1+2p_2)})$. We therefore obtain

$$\|I_1\|_\infty \leq \|I_1\|_F = O\Big(\tfrac{p_1+2p_2}{T}\Big).$$

Similarly for I_2,

$$\|I_2\|_\infty \leq \|I_2\|_F \leq |||\Pi^\star - \widehat{\Pi}^{(0)}|||_F^2|||\tfrac{\mathcal{W}'\mathcal{W}}{T}|||_{op},$$

where with a similar derivation to that in Lemma 23, for sample size $T \gtrsim (p_1+p_2)$, with probability at least $1 - c_1'''\exp\{-c_2'''(p_1+p_2)\}$, we get

$$(\mathbf{E_4'}) \qquad |||\tfrac{\mathcal{W}'\mathcal{W}}{T}|||_{op} \leq C_4\sqrt{\frac{p_1+2p_2}{T}} + \Lambda_{\max}(\Gamma_X), \qquad \text{for some constant } C_4.$$

Hence,

$$\|I_2\|_\infty \leq \|I_2\|_F \leq= O\Big((\frac{p_1+2p_2}{T})^{3/2}\Big).$$

Combining all terms and then taking the leading one, with probability at least

$$1 - c_1\exp\{-c_2(p_1+p_2)\} - c_1'\exp\{-c_2'(p_1+2p_2)\} - c_1''\exp\{-c_2''\log[p_2(p_1+p_2)]\} \\ -c_1'''\exp\{-c_2'''(p_1+p_2)\} - \exp(-\tau\log p_2),$$

we obtain

$$\|\widehat{S}_v^{(0)} - \Sigma_v^\star\|_\infty = O\Big(\sqrt{\tfrac{p_1+2p_2}{T}}\Big).$$

Note that here with the required sample size, $(p_1+2p_2)/T$ is a small quantity, and therefore

$$O\Big((\tfrac{p_1+2p_2}{T})^{3/2}\Big) \leq O\Big(\tfrac{p_1+2p_2}{T}\Big) \leq O\Big(\sqrt{\tfrac{p_1+2p_2}{T}}\Big).$$

At iteration 1, the bound of $|||\widehat{B}^{(1)} - B^\star|||_F^2 + |||\widehat{C}^{(1)} - C^\star|||_F^2$ relies on the following two quantities:

$$\Big|\Big|\Big|\tfrac{1}{T}\mathcal{W}'\mathcal{V}\widehat{\Omega}_v^{(0)}\Big|\Big|\Big|_{op} \qquad \text{and} \qquad \Big\|\tfrac{1}{T}\mathcal{W}'\mathcal{V}\widehat{\Omega}_v^{(0)}\Big\|_\infty.$$

Using a similar derivation to that in the proof of Theorem 5,

$$\Big\|\tfrac{1}{T}\mathcal{W}'\mathcal{V}\widehat{\Omega}_v^{(0)}\Big\|_\infty \le \Big\|\tfrac{1}{T}\mathcal{W}'\mathcal{V}(\widehat{\Omega}_v^{(0)} - \Omega_v^\star)\Big\|_\infty + \big\|\tfrac{1}{T}\mathcal{W}'\mathcal{V}\Omega_v^\star\big\|_\infty, \tag{41}$$

where by viewing $\mathcal{V}\Omega_v^\star$ as some random realization coming from a certain sub-Gaussian process, with probability at least $1 - \bar{c}_1'' \exp\{-\bar{c}_2'' \log[p_2(p_1+p_2)]\}$, we get

$$(\mathbf{E_5'}) \qquad \big\|\tfrac{1}{T}\mathcal{W}'\mathcal{V}\Omega_v^\star\big\|_\infty \le C_5\sqrt{\frac{\log(p_1+p_2)+\log p_2}{T}}, \qquad \text{for some constant } C_5,$$

and

$$\begin{aligned}\Big\|\tfrac{1}{T}\mathcal{W}'\mathcal{V}(\widehat{\Omega}_v^{(0)} - \Omega_v^\star)\Big\|_\infty &\le d_{\max}^{\Omega_v^\star}\|\tfrac{1}{T}\mathcal{W}'\mathcal{V}\|_\infty \|\widehat{\Omega}_v^{(0)} - \Omega_v^\star\|_\infty \\ &= O\Big(\sqrt{\tfrac{\log(p_1+p_2)+\log p_2}{T}}\Big)\cdot O\Big(\sqrt{\tfrac{p_1+2p_2}{T}}\Big).\end{aligned}$$

For $|||\frac{1}{T}\mathcal{W}'\mathcal{V}\widehat{\Omega}_v^{(0)}|||_{op}$, similarly we have

$$\Big|\Big|\Big|\tfrac{1}{T}\mathcal{W}'\mathcal{V}\widehat{\Omega}_v^{(0)}\Big|\Big|\Big|_{op} \le |||\tfrac{1}{T}\mathcal{W}'\mathcal{V}(\widehat{\Omega}_v^{(0)} - \Omega_v^\star)|||_{\text{op}} + |||\tfrac{1}{T}\mathcal{W}'\mathcal{V}\Omega_v^\star|||_{\text{op}}, \tag{42}$$

where with probability at least $1 - \bar{c}_1' \exp\{-\bar{c}_2'(p_1+p_2)\}$,

$$(\mathbf{E_6'}) \qquad |||\tfrac{1}{T}\mathcal{W}'\mathcal{V}\Omega_v^\star|||_{op} \le C_6\sqrt{\frac{p_1+2p_2}{T}} \qquad \text{for some constant } C_6,$$

and

$$\begin{aligned}\Big|\Big|\Big|\tfrac{1}{T}\mathcal{W}'\mathcal{V}(\widehat{\Omega}_v^{(0)} - \Omega_v^\star)\Big|\Big|\Big|_{op} &\le |||\tfrac{1}{T}\mathcal{W}'\mathcal{V}|||_{\text{op}} |||\widehat{\Omega}_v^{(0)} - \Omega_v^\star|||_{\text{op}} \\ &\le |||\tfrac{1}{T}\mathcal{W}'\mathcal{V}|||_{op}\Big[d_{\max}^{\Omega_v^\star}\|\widehat{\Omega}_v^{(0)} - \Omega_v^\star\|_\infty\Big] = O\Big(\tfrac{p_1+2p_2}{T}\Big),\end{aligned}$$

where the second inequality follows from Corollary 3 of Ravikumar et al. (2011). Combining all terms from (41) and (42), the leading term gives the following bound:

$$|||\widehat{B}^{(1)} - B^\star|||_F^2 + |||\widehat{C}^{(1)} - C^\star|||_F^2 \le C_7\Big(\frac{p_1+2p_2}{T}\Big) \qquad \text{for some constant } C_7,$$

and this error rate coincides with that in the bound of $|||\widehat{B}^{(0)} - B^\star|||_F^2 + |||\widehat{C}^{(0)} - C^\star|||_F^2$. This implies that for $\widehat{\Omega}_v^{(1)}$ and iterations thereafter, the error rate remains unchanged. Moreover, all sources of randomness have been captured up to this step in events $\mathbf{E_1'}$ to $\mathbf{E_6'}$, and therefore the probability for the bounds to hold no longer changes. Consequently, the following bounds hold for all iterations k:

$$\|\mathcal{W}'\mathcal{V}\widehat{\Omega}_v^{(k)}/T\|_\infty = |||\mathcal{W}'\mathcal{V}\widehat{\Omega}_v^{(k)}/T|||_{\text{op}} = O\Big(\sqrt{\tfrac{p_1+2p_2}{T}}\Big)$$

and

$$\|\widehat{S}_v^{(k)} - \Sigma_v^\star\|_\infty = O\Big(\sqrt{\tfrac{p_1+2p_2}{T}}\Big),$$

with probability at least

$$1-c_0\exp\{-\tilde{c}_0(p_1+p_2)\}-c_1\exp\{-\tilde{c}_1(p_1+2p_2)\}-c_2\exp\{-\tilde{c}_2\log[p_2(p_1+p_2)]\}-\exp\{-\tau\log p_2\}.$$

for some new positive constants $c_i, \tilde{c}_i$ $(i=0,1,2)$ and τ.[2] The above bounds directly imply the bound in the statement in Theorem 6, with the aid of Theorem 15. ■

Appendix B. Key Lemmas and Their Proofs.

In this section, we verify the conditions that are required for establishing the consistency results in Theorem 14 and 15, under random realizations of $\mathcal{X}$, $\mathcal{Z}$, $\mathcal{U}$ and $\mathcal{V}$.

The following two lemmas verify the conditions for establishing the consistency properties for $\widehat{A}^{(0)}$. Specifically, Lemma 16 establishes that with high probability, $\mathfrak{X}_0$ satisfies the RSC condition. Further, Lemma 17 gives a high probability upper bound for $\|\mathcal{X}'\mathcal{U}/T\|_\infty$ for random $\mathcal{X}$ and $\mathcal{U}$.

Lemma 16 (Verification of the RSC condition) *For the* VAR(1) *model* $\{X_t\}$ *posited in* (1)*, there exist* $c_i > 0$ $(i = 1,2,3)$ *such that for sample size* $T \gtrsim \max\{\omega^2, 1\} s_A^\star \log p_1$*, with probability at least*

$$1 - c_1\exp\left[-c_2 T\min\{1,\omega^{-2}\}\right], \qquad \omega = c_3\frac{\Lambda_{\max}(\Sigma_u)\mu_{\max}(\mathcal{A})}{\Lambda_{\min}(\Sigma_u)\mu_{\min}(\mathcal{A})},$$

the following inequality holds

$$\frac{1}{2T}\|\!|\!|\mathfrak{X}_0(\Delta)|\!|\!|_F^2 \geq \alpha_{RSC}|\!|\!|\Delta|\!|\!|_F^2 - \tau\|\Delta\|_1^2, \qquad \textit{for } \Delta\in\mathbb{R}^{p_1\times p_1},$$

where $\alpha_{RSC} = \frac{\Lambda_{\min}(\Sigma_u)}{\mu_{\max}(\mathcal{A})}$, $\tau = 4\alpha_{RSC}\max\{\omega^2,1\}\log p_1/T$.

Proof [**Proof of Lemma 16**] For the specific VAR(1) process $\{X_t\}$ given in (1), using Proposition 4.2 in Basu and Michailidis (2015) with $d=1$ directly gives the result. Specifically, we note that by letting $\theta = \mathrm{vec}(\Delta)$,

$$\tfrac{1}{T}|\!|\!|\mathfrak{X}_0(\Delta)|\!|\!|_F^2 = \theta'\widehat{\Gamma}_X^{(0)}\theta,$$

where $\widehat{\Gamma}_X^{(0)} = \mathrm{I}_{p_1}\otimes(\mathcal{X}'\mathcal{X}/T)$, and $\|\theta\|_2^2 = |\!|\!|\Delta|\!|\!|_F^2$, $\|\theta\|_1 = \|\Delta\|_1$. ■

Lemma 17 (Verification of the deviation bound) *For the model in* (1)*, there exist constants* $c_i > 0$*,* $i = 0,1,2$ *such that for* $T \gtrsim 2\log p_1$*, with probability at least* $1 - c_1\exp(-2c_2\log p_1)$*, the following bound holds:*

$$\|\mathcal{X}'\mathcal{U}/T\|_\infty \leq c_0\Lambda_{\max}(\Sigma_u)\left[1 + \frac{1}{\mu_{\min}(\mathcal{A})} + \frac{\mu_{\max}(\mathcal{A})}{\mu_{\min}(\mathcal{A})}\right]\sqrt{\frac{2\log p_1}{T}}. \tag{43}$$

2. Here we slightly abuse the notations and redefine $c_0 := \max\{c_1, c_1'''\}$, $c_1 := \max\{c_1', \bar{c}_1'\}$, $\tilde{c}_1 := \min\{c_2', \bar{c}_2'\}$, $c_2 = \max\{c_1'', \bar{c}_1''\}$, $\tilde{c}_2 := \min\{c_2'', \bar{c}_2''\}$.

Proof [**Proof of Lemma 17**] First, we note that,

$$\|\mathcal{X}'\mathcal{U}/T\|_\infty = \max_{\substack{1\le i\le p_1\\1\le j\le p_1}} \left|e_i'\left(\mathcal{X}'\mathcal{U}/T\right)e_j\right|.$$

Applying Proposition 2.4(b) in Basu and Michailidis (2015) for an arbitrary pair of (i,j) gives:

$$\mathbb{P}\left(\left|e_i'\left(\mathcal{X}'\mathcal{U}/T\right)e_j\right| > \eta\left[\Lambda_{\max}(\Sigma_u)\left(1+\frac{1}{\mu_{\min}(\mathcal{A})}+\frac{\mu_{\max}(\mathcal{A})}{\mu_{\min}(\mathcal{A})}\right)\right]\right) \le 6\exp[-cT\min\{\eta,\eta^2\}].$$

Setting $\eta = c_0\sqrt{2\log p_1/T}$ and taking a union bound over all $1\le i\le p_1, 1\le j\le p_1$, we get that for some $c_1, c_2>0$, with probability at least $1-c_1\exp[-2c_2\log p_1]$,

$$\max_{\substack{1\le i\le p_1\\1\le j\le p_1}} \left|e_i'\left(\mathcal{X}'\mathcal{U}/T\right)e_j\right| \le c_0\Lambda_{\max}(\Sigma_u)\left[1+\frac{1}{\mu_{\min}(\mathcal{A})}+\frac{\mu_{\max}(\mathcal{A})}{\mu_{\min}(\mathcal{A})}\right]\sqrt{\frac{2\log p_1}{T}}.$$

■

In the next two lemmas, Lemma 18 gives an RSC curvature that holds with high probability for $\mathfrak{W}$ induced by a random $\mathcal{W}$, and Lemma 19 gives a high probability upper bound for $|||W'\mathcal{V}/T|||_{\text{op}}$ and $\|W'\mathcal{V}/T\|_\infty$.

Lemma 18 (Verification of the RSC condition) *Consider the covariance stationary process $W_t = (X_t', Z_t')' \in \mathbb{R}^{p_1+p_2}$ whose spectral density exists. Suppose $\mathfrak{m}(f_W)>0$. There exist constants $c_i>0, i=1,2,3$ such that with probability at least $1-2c_1\exp(-c_2(p_1+p_2))$, the RSC condition for $\mathfrak{W}$ induced by a random $\mathcal{W}$ holds for α_{RSC} and tolerance 0, where*

$$\alpha_{RSC} = \pi\mathfrak{m}(f_W)/4,$$

whenever $T \gtrsim c_3(p_1+p_2)$.

Proof [**Proof of Lemma 18**] First ,we note that the following inequality holds, for any $\mathcal{W}$:

$$\frac{1}{2T}|||\mathfrak{W}_0(\Delta)|||_F^2 = \frac{1}{2T}\left|\left|\left|\mathcal{W}'\Delta\right|\right|\right|_F^2 = \frac{1}{2T}\sum_{j=1}^{p_2}\left\|[\mathcal{W}'\Delta]_j\right\|_2^2 \ge \frac{1}{2}\Lambda_{\min}\left(\widehat{\Gamma}_W^{(0)}\right)|||\Delta|||_F^2. \tag{44}$$

where $\widehat{\Gamma}_W^{(0)} = \mathcal{W}'\mathcal{W}/T$. Applying Lemma 4 in Negahban and Wainwright (2011) on $\mathcal{W}$ together with Proposition 2.3 in Basu and Michailidis (2015), the following bound holds with probability at least $1-2c_1\exp[-c_2(p_1+p_2)]$, as long as $T\gtrsim c_3(p_1+p_2)$:

$$\Lambda_{\min}\left(\widehat{\Gamma}_W^{(0)}\right) \ge \frac{\Lambda_{\min}(\Gamma_W(0))}{4} \ge \frac{\pi}{2}\mathfrak{m}(f_W),$$

where $\Gamma_W(0) = \mathbb{E}W_tW_t'$. Combining with (44), the RSC condition holds with $\kappa(\mathfrak{W}) = \pi\mathfrak{m}(f_W)/4$. ■

Lemma 19 (Verification of the deviation bound) *There exist constants $c_i > 0$ and $c_i' > 0, i = 1, 2, 3$ such that the following statements hold:*

(a) With probability at least $1 - c_1 \exp[-c_2(p_1 + 2p_2)]$, as long as $T \gtrsim c_3(p_1 + 2p_2)$,

$$\left|\!\left|\!\left|\mathcal{W}'\mathcal{V}/T\right|\!\right|\!\right|_{op} \leq c_0 \left[\mathcal{M}(f_W) + \tfrac{1}{2\pi}\Lambda_{\max}(\Sigma_v) + \mathcal{M}(f_{W,V})\right] \sqrt{\frac{p_1 + 2p_2}{T}}. \tag{45}$$

(b) With probability at least $1-c_1' \exp(-c_2' \log(p_1+p_2)-c_2' \log p_2)$, as long as $T \gtrsim c_3' \log[(p_1+p_2)p_2]$,

$$\left\|\mathcal{W}'\mathcal{V}/T\right\|_\infty \leq c_0' \left[\mathcal{M}(f_W) + \tfrac{1}{2\pi}\Lambda_{\max}(\Sigma_v) + \mathcal{M}(f_{W,V})\right] \sqrt{\frac{\log(p_1 + p_2) + \log p_2}{T}}. \tag{46}$$

Proof [**Proof of Lemma 19**] (a) is a direct application of Lemma 23 on processes $\{W_t\} \in \mathbb{R}^{(p_1+p_2)}$ and $\{V_t\} \in \mathbb{R}^{p_2}$, and (b) is a direct application of Lemma 17. ■

Appendix C. Auxiliary Lemmas and Their Proofs.

Lemma 20 *Consider two centered stationary Gaussian processes $\{X_t\}$ and $\{Z_t\}$. Further, assume that the spectral density of the joint process $\{(X_t', Z_t')'\}$ exists. Denote their cross-covariance by $\Gamma_{X,Z}(\ell) := \mathrm{Cov}(X_t, Z_{t+\ell})$, and their cross-spectral density is defined as*

$$f_{X,Z}(\theta) := \frac{1}{2\pi}\sum_{\ell=-\infty}^{\infty} \Gamma_{X,Z}(\ell)e^{-i\ell\theta}, \qquad \theta \in [-\pi, \pi],$$

whose upper extreme is given by:

$$\mathcal{M}(f_{X,Z}) = \mathrm{esssup}_{\theta\in[-\pi,\pi]}\sqrt{\Lambda_{\max}\left(f^*_{X,Z}(\theta)f_{X,Z}(\theta)\right)}.$$

Let $\mathcal{X}$ and $\mathcal{Z}$ be data matrices with sample size n. Then, there exists a constant $c > 0$, such that for any $u, v \in \mathbb{R}^p$ with $\|u\| \leq 1$, $\|v\| \leq 1$, we have

$$\mathbb{P}\left[\left|u'\left(\frac{\mathcal{X}'Z}{T} - \mathrm{Cov}(X_t, Z_t)\right)v\right| > 2\pi\left(\mathcal{M}(f_X) + \mathcal{M}(f_Z) + \mathcal{M}(f_{X,Z})\right)\eta\right] \leq 6\exp\left(-cT\min\{\eta, \eta^2\}\right).$$

Proof Let $\xi_t = \langle u, X_t\rangle$, $\eta_t = \langle v, Z_t\rangle$. Let $f_X(\theta), f_Z(\theta)$ denote the spectral density of $\{X_t\}$ and $\{Z_t\}$, respectively. Then, the spectral density of $\{\xi_t\}$ and $\{\eta_t\}$, respectively, is $f_\xi(\theta) = u' f_X(\theta)u$, $f_\eta(\theta) = v' f_Z(\theta)v$. Also, we note that $\mathcal{M}(f_\xi) \leq \mathcal{M}(f_X)$, $\mathcal{M}(f_\eta) \leq \mathcal{M}(f_Z)$. Then,

$$\begin{aligned}\frac{2}{T}\left[\sum_{t=0}^{T}\xi_t\eta_t - \mathrm{Cov}(\xi_t, \eta_t)\right] &= \left[\frac{1}{T}\sum_{t=0}^{T}(\xi_t + \eta_t)^2 - \mathrm{Var}(\xi_t + \eta_t)\right] \\ &\quad - \left[\frac{1}{T}\sum_{t=0}^{T}(\xi_t)^2 - \mathrm{Var}(\xi_t)\right] - \left[\frac{1}{T}\sum_{t=0}^{T}(\eta_t)^2 - \mathrm{Var}(\eta_t)\right].\end{aligned} \tag{47}$$

By Proposition 2.7 in Basu and Michailidis (2015),

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=0}^{T}(\xi_t)^2 - \mathrm{Var}(\xi_t)\right| > 2\pi\mathcal{M}(f_X)\eta\right) \geq 2\exp\left[-cn\min(\eta,\eta^2)\right],$$

and

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=0}^{T}(\eta_t)^2 - \mathrm{Var}(\eta_t)\right| > 2\pi\mathcal{M}(f_Z)\eta\right) \geq 2\exp\left[-cn\min(\eta,\eta^2)\right].$$

What remains to be considered is the first term in (47), whose spectral density is given by

$$f_{\xi+\eta}(\theta) = u' f_X(\theta) u + v' f_Z(\theta) z + u' f_{X,Z}(\theta) v + v' f^*_{X,Z}(\theta) u,$$

and its upper extreme satisfies

$$\mathcal{M}(f_{\xi+\eta}) \leq \mathcal{M}(f_X) + \mathcal{M}(f_Z) + 2\mathcal{M}(f_{X,Z}).$$

Hence, we get:

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=0}^{T}(\xi_t+\eta_t)^2 - \mathrm{Var}(\xi_t+\eta_t)\right| > 2\pi[\mathcal{M}(f_X) + \mathcal{M}(f_Z) + 2\mathcal{M}(f_{X,Z})]\eta\right) \geq 2\exp[-cn\min(\eta,\eta^2)].$$

Combining all three terms yields the desired result. ∎

Lemma 21 *Define the error matrix by* $\Delta^B = \widehat{B} - B^\star$ *and* $\Delta^C = \widehat{C} - C^\star$*, and let the weighted regularizer* $\mathcal{Q}$ *be defined as*

$$\mathcal{Q}(B,C) = |||B|||_* + \frac{\lambda_C}{\lambda_B}\|C\|_1.$$

With the subspaces defined in (35) and (36), the following inequality holds:

$$\mathcal{Q}(B^\star, C^\star) - \mathcal{Q}(\widehat{B}, \widehat{C}) \leq \mathcal{Q}(\Delta^B_{\mathcal{S}_{B^\star}}, \Delta^C_{\mathcal{J}_{C^\star}}) - \mathcal{Q}(\Delta^B_{\mathcal{S}^\perp_{B^\star}}, \Delta^C_{\mathcal{J}^c_{C^\star}}).$$

Proof First, from definitions (35) and (36), we know that $B^\star_{\mathcal{S}^\perp} = 0$ and $C^\star_{\mathcal{J}^c_{C^\star}} = 0$. Using the definition of $\mathcal{Q}$, we obtain

$$\mathcal{Q}(B^\star, C^\star) = \left|\left|\left| B^\star_{\mathcal{S}} + B^\star_{\mathcal{S}^\perp}\right|\right|\right|_* + \frac{\lambda_C}{\lambda_B}\left\|C^\star_{\mathcal{J}^\star_C} + C^\star_{\mathcal{J}^c_{C^\star}}\right\|_1 = |||B^\star_{\mathcal{S}}|||_* + \frac{\lambda_C}{\lambda_B}\left\|C^\star_{\mathcal{J}^\star_C}\right\|_1,$$

and

$$\begin{aligned}
\mathcal{Q}(\widehat{B}, \widehat{C}) &= \mathcal{Q}(B^\star + \Delta^B, C^\star + \Delta^C) \\
&= \left|\left|\left| B^\star_{\mathcal{S}} + \Delta^B_{\mathcal{S}^\perp_{B^\star}} + \Delta^B_{\mathcal{S}_{B^\star}} + B^\star_{\mathcal{S}^\perp}\right|\right|\right|_* + \frac{\lambda_C}{\lambda_B}\left\|C^\star_{\mathcal{J}^\star_C} + \Delta^C_{\mathcal{J}_{C^\star}} + C^\star_{\mathcal{J}^c_{C^\star}} + \Delta^C_{\mathcal{J}^c_{C^\star}}\right\|_1 \\
&\geq \left|\left|\left| B^\star_{\mathcal{S}} + \Delta^B_{\mathcal{S}^\perp_{B^\star}}\right|\right|\right|_* - |||\Delta^B_{\mathcal{S}_{B^\star}}|||_* + \frac{\lambda_C}{\lambda_B}\left(\left\|C^\star_{\mathcal{J}^\star_C} + \Delta^C_{\mathcal{J}_{C^\star}}\right\|_1 + \left\|\Delta^C_{\mathcal{J}^c_{C^\star}}\right\|_1\right) \\
&\geq |||B^\star_{\mathcal{S}}|||_* + \left|\left|\left|\Delta^B_{\mathcal{S}^\perp_{B^\star}}\right|\right|\right|_* - |||\Delta^B_{\mathcal{S}_{B^\star}}|||_* + \frac{\lambda_C}{\lambda_B}\left(\left\|C^\star_{\mathcal{J}^\star_C}\right\|_1 + \|\Delta^C_{\mathcal{J}_{C^\star}}\|_1 - \left\|\Delta^C_{\mathcal{J}^c_{C^\star}}\right\|_1\right).
\end{aligned}$$

The decomposition of the first term comes from the construction of $\Delta^B_{\mathcal{S}^\perp_{B^\star}}$. It then follows that

$$
\begin{aligned}
\mathcal{Q}(B^\star, C^\star) - \mathcal{Q}(\widehat{B}, \widehat{C}) &\leq \frac{\lambda_C}{\lambda_B}\left\|C^\star_{\mathcal{J}^\star_C}\right\|_1 + |||\Delta^B_{\mathcal{S}_{B^\star}}|||_* - \left|\left|\left|\Delta^B_{\mathcal{S}^\perp_{B^\star}}\right|\right|\right|_* + \frac{\lambda_C}{\lambda_B}\left(\|\Delta^C_{\mathcal{J}_{C^\star}}\|_1 - \left\|\Delta^C_{\mathcal{J}^c_{C^\star}}\right\|_1 - \left\|C^\star_{\mathcal{J}^\star_C}\right\|_1\right) \\
&= |||\Delta^B_{\mathcal{S}_{B^\star}}|||_* + \frac{\lambda_C}{\lambda_B}\|\Delta^C_{\mathcal{J}_{C^\star}}\|_1 - \left(\left|\left|\left|\Delta^B_{\mathcal{S}^\perp_{B^\star}}\right|\right|\right|_* + \frac{\lambda_C}{\lambda_B}\left\|\Delta^C_{\mathcal{J}^c_{C^\star}}\right\|_1\right) \\
&= \mathcal{Q}(\Delta^B_{\mathcal{S}_{B^\star}}, \Delta^C_{\mathcal{J}_{C^\star}}) - \mathcal{Q}(\Delta^B_{\mathcal{S}^\perp_{B^\star}}, \Delta^C_{\mathcal{J}^c_{C^\star}}).
\end{aligned}
$$

■

Lemma 22 *Under the conditions of Theorem 15, the following bound holds:*

$$\frac{1}{T}|||\mathfrak{W}_0(\Delta^B_{aug} + \Delta^C_{aug})|||_F^2 \geq \frac{\alpha_{RSC}}{2}(|||\Delta^B|||_F^2 + |||\Delta^C|||_F^2) - \frac{\lambda_B}{2}\mathcal{Q}(\Delta^B, \Delta^C).$$

Proof This lemma directly follows from Lemma 2 in Agarwal et al. (2012), by setting $\Theta^\star = B^\star$, $\Gamma^\star = C^\star$, with the regularizer $\mathcal{R}(\cdot)$ being the element-wise ℓ_1 norm. Note that $\sigma_j(B^\star) = 0$ for $j = r+1, \cdots, \min\{p_1, p_2\}$ since $\mathrm{rank}(B) = r$. For our problem, it suffices to set $\mathbb{M}^\perp$ as $\mathcal{J}^c_{C^\star}$, and therefore $\|C^\star_{\mathcal{J}^c_{C^\star}}\|_1 = 0$. ■

Lemma 23 *Consider the two centered Gaussian processes $\{X_t\} \in \mathbb{R}^{p_1}$ and $\{Z_t\} \in \mathbb{R}^{p_2}$, and denote their cross covariance matrix by $\Gamma_{X,Z}(h) = (X_t, Z_{t+h}) = \mathbb{E}(X_t Z'_{t+h})$. Let $\mathcal{X}$ and $\mathcal{Z}$ denote the data matrix. There exist positive constants $c_i > 0$ such that whenever $T \gtrsim c_3(p_1 + p_2)$, with probability at least*

$$1 - c_1 \exp[-c_2(p_1 + p_2)],$$

the following bound holds:

$$\frac{1}{T}|||\mathcal{X}'\mathcal{Z}|||_{op} \leq \mathbb{Q}_{X,Z}\sqrt{\frac{p_1 + p_2}{T}} + 4|||\Gamma_{X,Z}(0)|||_{op},$$

where

$$\mathbb{Q}_{X,Z} = c_0\left[\mathcal{M}(f_X) + \mathcal{M}(f_Z) + \mathcal{M}(f_{X,Z})\right].$$

Proof The main structure of this proof follows from that of Lemma 3 in Negahban and Wainwright (2011), and here we focus on how to handle the temporal dependency present in our problem. Let $S^p = \{u \in \mathbb{R}^p | \|u\| = 1\}$ denote the p-dimensional unit sphere. The operator norm has the following variational representation form:

$$\frac{1}{T}|||\mathcal{X}'\mathcal{Z}|||_{op} = \frac{1}{n}\sup_{u \in S^{p_1}}\sup_{v \in S^{p_2}} u'\mathcal{X}'\mathcal{Z}v.$$

For positive scalars s_1 and s_2, define

$$\Psi(s_1, s_2) = \sup_{u \in s_1 S^{p_1}}\sup_{v \in s_2 S^{p_2}}\langle \mathcal{X}u, \mathcal{Z}v\rangle,$$

and the goal is to establish an upper bound for $\Psi(1,1)/T$. Let $\mathcal{A} = \{u^1, \cdots, u^A\}$ and $\mathcal{B} = \{v^1, \cdots, v^B\}$ denote the 1/4 coverings of S^{p_1} and S^{p_2}, respectively. Negahban and Wainwright (2011) showed that

$$\Psi(1,1) \leq 4 \max_{u^a \in \mathcal{A}, v^b \in \mathcal{B}} \langle \mathcal{X} u^a, \mathcal{Z} v^b \rangle,$$

and by Anderson et al. (1998) and Anderson (2011), there exists a 1/4 covering of S^{p_1} and S^{p_2} with at most $A \leq 8^{p_1}$ and $B \leq 8^{p_2}$ elements, respectively. Consequently,

$$\mathbb{P}\left[\left|\frac{1}{T}\Psi(1,1)\right| \geq 4\delta\right] \leq 8^{p_1+p_2} \max_{u^a, v^b} \mathbb{P}\left[\frac{|(u^a)'\mathcal{X}\mathcal{Z}(v^b)|}{T} \geq \delta\right].$$

What remains to be bounded is

$$\frac{1}{T} u' \mathcal{X}' \mathcal{Z} v, \qquad \text{for an arbitrary fixed pair of } (u,v) \in S^{p_1} \times S^{p_2}.$$

By Lemma 20, we have

$$\mathbb{P}\left[\left|u'\left(\frac{\mathcal{X}'Z}{T}\right)v\right| > 2\pi\left(\mathcal{M}(f_X) + \mathcal{M}(f_Z) + \mathcal{M}(f_{X,Z})\right)\eta + |||\Gamma_{X,Z}(0)|||_{op}\right] \leq 6\exp\left(-cT\min\{\eta, \eta^2\}\right).$$

Therefore, we have

$$\mathbb{P}\left[\left|\frac{1}{T}\Psi(1,1)\right| \geq 8\pi\left(\mathcal{M}(f_X) + \mathcal{M}(f_Z) + \mathcal{M}(f_{X,Z})\right)\eta + 4|||\Gamma_{X,Z}(0)|||_{op}\right] \leq 6\exp\left[(p_1+p_2)\log 8 - cT\min\{\eta, \eta^2\}\right].$$

With the specified choice of sample size T, the probability vanishes by choosing $\eta = c_0\sqrt{\frac{p_1+p_2}{T}}$, for c_0 large enough, and we yield the conclusion in Lemma 23. ∎

Lemma 24 *Let $\{X_t\}$ and $\{\varepsilon_t\}$ be two generic processes, where $\varepsilon_t = (U_t', V_t')'$. Suppose the spectral density of the joint process (X_t', ε_t') exists. Then, the following inequalities hold*

$$\mathfrak{m}(f_{X,V}) \geq \mathfrak{m}(f_{X,\varepsilon}), \qquad \mathcal{M}(f_{X,V}) \leq \mathcal{M}(f_{X,\varepsilon}).$$

Proof By definition, the spectral density $f_{X,\varepsilon}(\theta)$ can be written as

$$\begin{aligned} f_{X,\varepsilon}(\theta) &= \left(\frac{1}{2\pi}\right)\sum_{\ell=-\infty}^{\infty}\Gamma_{X,\varepsilon}(\ell)e^{-i\ell\theta}, \qquad \theta\in[-\pi,\pi] \\ &= \left(\frac{1}{2\pi}\right)\sum_{\ell=-\infty}^{\infty}\left(\mathbb{E}X_tU_{t+\ell}, \quad \mathbb{E}X_tV_{t+\ell}\right)e^{-i\ell\theta} \\ &= (f_{X,U}(\theta), \quad f_{X,V}(\theta)). \end{aligned}$$

It follows that

$$\mathcal{M}(f_{X,\varepsilon}) = \operatorname*{ess\,sup}_{\theta\in[-\pi,\pi]}\sqrt{\Lambda_{\max}(H(\theta))},$$

where

$$H(\theta) = \begin{bmatrix} f^*_{X,U}(\theta) \\ f^*_{X,V}(\theta) \end{bmatrix} \begin{bmatrix} f_{X,U}(\theta) & f_{X,V}(\theta) \end{bmatrix} = \begin{bmatrix} f^*_{X,U}(\theta) f_{X,U}(\theta) & f^*_{X,U}(\theta) f_{X,V}(\theta) \\ f^*_{X,V}(\theta) f_{X,U}(\theta) & f^*_{X,V}(\theta) f_{X,V}(\theta) \end{bmatrix}.$$

Note that

$$\mathcal{M}(f_{X,V}) = \operatorname*{ess\ sup}_{\theta \in [-\pi,\pi]} \sqrt{\Lambda_{\max}(f^*_{X,V}(\theta) f_{X,V}(\theta))}.$$

By Lemma **??**, $\forall \theta$, $\Lambda_{\min}(f^*_{X,V}(\theta) f_{X,V}(\theta)) \geq \Lambda_{\min}(H(\theta))$ and $\Lambda_{\max}(f^*_{X,V}(\theta) f_{X,V}(\theta)) \leq \Lambda_{\max}(H(\theta))$, hence

$$\mathfrak{m}(f_{X,V}) \geq \mathfrak{m}(f_{X,\varepsilon}), \qquad \mathcal{M}(f_{X,V}) \leq \mathcal{M}(f_{X,\varepsilon}).$$

■

Appendix D. Proof of Proposition(s).

Proof [**Proof of Proposition 9**] The joint process $W_t = \{(X_t', Z_t')'\}$ is a stationary VAR(1) process, and it follows that

$$S_w(h) := \begin{bmatrix} S_x(h) & S_{x,z}(h) \\ S_{z,x}(h) & S_z(h) \end{bmatrix} = \frac{1}{T} \sum_{t=1}^{T} w_t w_{t+h}' \xrightarrow{p} \Gamma_W(h) := \mathbb{E} W_t W_{t+h}', \quad \text{as} \quad T \to \infty,$$

which implies

$$S_x \xrightarrow{p} \Gamma_X, \quad S_z \xrightarrow{p} \Gamma_Z, \quad S_{x,z} \xrightarrow{p} \Gamma_{X,Z}, \quad S_{x,z}(1) \xrightarrow{p} \Gamma_{X,Z}(1).$$

Note that sample partial regression residual covariances can be obtained by

$$S_{00} = S_z - S_z(1) S_z^{-1} S_z'(1), \qquad S_{11} = S_x - S_{x,z} S_z^{-1} S_{x,z}', \qquad S_{10} = S_{x,z}(1) - S_z(1) S_z^{-1} S_{x,z}'.$$

An application of the Continuous Mapping Theorem yields

$$S_{00} \xrightarrow{p} \Sigma_{00}, \qquad S_{10} \xrightarrow{p} \Sigma_{10}, \qquad S_{11} \xrightarrow{p} \Sigma_{11}.$$

By Hsu (1941a,b), the limiting behavior of $T\Psi_r$ is given by

$$T\Psi_r \sim \chi^2_{(p_1 - r)(p_2 - r)}, \qquad \text{as} \quad T \to \infty.$$

Note that since μ is of multiplicity one and the ordered eigenvalues are continuous functions of the matrices, the following holds:

$$\phi_k \xrightarrow{p} \mu_k, \qquad \forall\, k = 1, \ldots, \min(p_1, p_2).$$

■

References

Tilak Abeysinghe. Estimation of direct and indirect impact of oil price on growth. *Economics Letters*, 73(2):147–153, 2001.

Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Noisy matrix decomposition via convex relaxation: optimal rates in high dimensions. *The Annals of Statistics*, 40(2): 1171–1197, 2012.

Charles W Anderson, Erik A Stolz, and Sanyogita Shamsunder. Multivariate autoregressive models for classification of spontaneous electroencephalographic signals during mental tasks. *IEEE Transactions on Biomedical Engineering*, 45(3):277–286, 1998.

Theodore Wilbur Anderson. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics*, 22(3):327–351, 1951.

Theodore Wilbur Anderson. Canonical correlation analysis and reduced rank regression in autoregressive models. *The Annals of Statistics*, 30(4):1134–1154, 2002.

Theodore Wilbur Anderson. *The Statistical Analysis of Time Series*, volume 19. John Wiley & Sons, 2011.

Ery Arias-Castro, Emmanuel J Candès, and Yaniv Plan. Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *The Annals of Statistics*, 39(5):2533–2556, 2011.

Sumanta Basu and George Michailidis. Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567, 2015.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Ben S Bernanke, Jean Boivin, and Piotr Eliasz. Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach. *The Quarterly Journal of Economics*, 120(1):387–422, 2005.

Carluccio Bianchi, Alessandro Carta, Dean Fantazzini, Maria Elena De Giuli, and Mario A Maggi. A copula-VAR-X approach for industrial production modeling and forecasting. *Applied Economics*, 42(25):3267–3277, 2010.

Monica Billio, Mila Getmansky, Andrew W Lo, and Loriana Pelizzon. Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics*, 104(3):535–559, 2012.

Celso Brunetti, Jeffrey H Harris, Shawn Mankad, and George Michailidis. Interconnectedness in the interbank market. *Social Sciences Research Network*, 2015.

David O Cushman and Tao Zha. Identifying monetary policy in a small open economy under flexible exchange rates. *Journal of Monetary Economics*, 39(3):433–448, 1997.

David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004.

Roger EA Farmer. The stock market crash really did cause the great recession. *Oxford Bulletin of Economics and Statistics*, 77(5):617–633, 2015.

US Financial Crisis Inquiry Commission. *The financial crisis inquiry report: Final report of the national commission on the causes of the financial and economic crisis in the United States*. Public Affairs, 2011.

Jean-Paul Fitoussi, David Jestaz, Edmund S Phelps, Gylfi Zoega, Olivier Blanchard, and Christopher A Sims. Roots of the recent recoveries: labor reforms or private sector forces? *Brookings Papers on Economic Activity*, 2000(1):237–311, 2000.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Christian Gourieroux and Joann Jasiak. *Financial Econometrics: Problems, Models, and Methods*. Princeton University Press Princeton, NJ, 2001.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 3(3):424–438, 1969.

Peter Hall. Resampling a coverage pattern. *Stochastic Processes and Their Applications*, 20(2):231–246, 1985.

Edward James Hannan. *Multiple Time Series*. John Wiley & Sons, 1970.

Lars Peter Hansen and Thomas J Sargent. *Recursive Models of Dynamic Linear Economies*. Princeton University Press, 2013.

PL Hsu. On the limiting distribution of roots of a determinantal equation. *Journal of the London Mathematical Society*, 1(3):183–194, 1941a.

PL Hsu. On the problem of rank and the limiting distribution of fisher's test function. *The Annals of Eugenics*, 11(1):39–41, 1941b.

Søren Johansen. Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, 12(2):231–254, 1988.

P.R. Kumar and Pravin Varaiya. *Stochastic Systems: Estimation, Identification and Adaptive Control*. Prentice Hall, 1986.

Jiahe Lin, Sumanta Basu, Moulinath Banerjee, and George Michailidis. Penalized maximum likelihood estimation of multi-layered gaussian graphical models. *Journal of Machine Learning Research*, 17(146):1–51, 2016.

Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664, 2012.

R Duncan Luce and Albert D Perry. A method of matrix analysis of group structure. *Psychometrika*, 14(2):95–116, 1949.

Helmut Lütkepohl. *New Introduction to Multiple Time series Analysis*. Springer Science & Business Media, 2005.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Sahand Negahban and Martin J Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097, 2011.

Sahand Negahban, Bin Yu, Martin J Wainwright, and Pradeep K Ravikumar. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Yurii Nesterov. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

Yurii Nesterov. On an approach to the construction of optimal methods of minimization of smooth convex functions. *Ekonomika i Mateaticheskie Metody*, 24(3):509–517, 1988.

Yurii Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

Yurii Nesterov. Gradient methods for minimizing composite objective function. Technical report, UCL, 2007.

William Nicolson, David Matteson, and Jacob Bien. VARX-L: Structured regularization for large vector autoregressions with exogenous variables. *arXiv preprint 1508.07497*, 2016.

Vincent R Nijs, Marnik G Dekimpe, Jan-Benedict EM Steenkamps, and Dominique M Hanssens. The category-demand effects of price promotions. *Marketing Science*, 20(1): 1–22, 2001.

Koen Pauwels and Allen Weiss. Moving from free to fee: How online firms market to change their business model successfully. *Journal of Marketing*, 72(3):14–31, 2008.

M Hashem Pesaran, Til Schuermann, and Scott M Weiner. Modeling regional interdependencies using a global error-correcting macroeconometric model. *Journal of Business & Economic Statistics*, 22(2):129–162, 2004.

Edmund S Phelps. Behind this structural boom: the role of asset valuations. *The American Economic Review*, 89(2):63–68, 1999.

Dimitris N Politis, Joseph P Romano, and Michael Wolf. *Subsampling*. Springer Series in Statistics, 1999.

Huitong Qiu, Sheng Xu, Fang Han, Han Liu, and Brian Caffo. Robust estimation of transition matrices in high dimensional heavy-tailed vector autoregressive processes. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1843–1851, 2015.

Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Bruno Rémillard, Nicolas Papageorgiou, and Frédéric Soustra. Copula-based semiparametric models for multivariate time series. *Journal of Multivariate Analysis*, 110:30–42, 2012.

Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18(82):1–9, 2013.

Anil K Seth. Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, 17(11):565–573, 2013.

Ali Shojaie, Sumanta Basu, and George Michailidis. Adaptive thresholding for reconstructing regulatory networks from time-course gene expression data. *Statistics in Biosciences*, 4(1):66–83, 2012.

Galen R Shorack and Jon A Wellner. *Empirical Processes with Applications to Statistics*. SIAM, 2009.

Christopher A Sims. Macroeconomics and reality. *Econometrica: Journal of the Econometric Society*, 48(1):1–48, 1980.

Christopher A Sims. Interpreting the macroeconomic time series facts: The effects of monetary policy. *European Economic Review*, 36(5):975–1000, 1992.

James H Stock and Mark W Watson. An empirical comparison of methods for forecasting using many predictors. *Manuscript, Princeton University*, 2005.

Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. Technical report, University of Washington, Seattle, 2008.