Journal of Machine Learning Research 21 (2020) 1-51 Submitted 10/19; Revised 3/20; Published 6/20

Regularized Estimation of High-dimensional Factor-Augmented Vector Autoregressive (FAVAR) 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: Xiaotong Shen

Abstract

A factor-augmented vector autoregressive (FAVAR) model is defined by a VAR equation that captures lead-lag correlations amongst a set of observed variables X and latent factors F, and a calibration equation that relates another set of observed variables Y with F and X. The latter equation is used to estimate the factors that are subsequently used in estimating the parameters of the VAR system. The FAVAR model has become popular in applied economic research, since it can summarize a large number of variables of interest as a few factors through the calibration equation and subsequently examine their influence on core variables of primary interest through the VAR equation. However, there is increasing need for examining lead-lag relationships between a large number of time series, while incorporating information from another high-dimensional set of variables. Hence, in this paper we investigate the FAVAR model under high-dimensional scaling. We introduce an appropriate identification constraint for the model parameters, which when incorporated into the formulated optimization problem yields estimates with good statistical properties. Further, we address a number of technical challenges introduced by the fact that estimates of the VAR system model parameters are based on estimated rather than directly observed quantities. The performance of the proposed estimators is evaluated on synthetic data. Further, the model is applied to commodity prices and reveals interesting and interpretable relationships between the prices and the factors extracted from a set of global macroeconomic indicators.

Keywords: Model Identifiability; Compactness; Low-rank plus Sparse Decomposition; Finite-Sample Bounds

1. Introduction

There is a growing need in employing a large set of time series (variables) for modeling social or physical systems. For example, economic policy makers have concluded based on extensive empirical evidence (e.g. Sims, 1980; Bernanke et al., 2005; Bańbura et al., 2010) that large scale models of economic indicators provide improved forecasts, together with better estimates of how current economic shocks propagate into the future, which produces better guidance for policy actions. Another reason for considering large number of time series in

social sciences is that key variables implied by theoretical models for policy decisions[1] are not directly observable, but related to a large number of other variables that collectively act as a good proxy of the unobservable key variables. In other domains such as genomics and neuroscience, advent of high throughput technologies have enabled researchers to obtain measurements on hundreds of genes from functional pathways of interest (Shojaie and Michailidis, 2010) or brain regions (Seth et al., 2015), thus allowing a more comprehensive modeling to gain insights into biological mechanisms of interest. There are two popular modeling paradigms for such large panel of time series, with the first being the Vector Autoregressive (VAR) model (Lütkepohl, 2005) and the second being the Dynamic Factor Model (DFM) (Stock and Watson, 2002; Lütkepohl, 2014).

The VAR model has been the subject of extensive theoretical and empirical work primarily in econometrics, due to its relevance in macroeconomic and financial modeling. However, the number of model parameters increases quadratically with the number of time series included for each lag period considered, and this feature has limited its applicability since in many applications it is hard to obtain adequate number of time points for accurate estimation. Nevertheless, there is a recent body of technical work that leveraging *structured sparsity* and the corresponding regularized estimation framework has established results for consistent estimation of the VAR parameters under high dimensional scaling. Basu and Michailidis (2015) examined Lasso penalized Gaussian VAR models and proved consistency results, while at the same time providing technical tools useful for analysis of sparse models involving temporally dependent data. Melnyk and Banerjee (2016) extended the results to other regularizers, Lin and Michailidis (2017) to the inclusion of exogenous variables (the so-called VAR-X model in the econometrics literature), Hall et al. (2019) to models for count data and Nicholson et al. (2017) to the simultaneous estimation of time lags and model parameters. However, a key requirement for the theoretical developments is a spectral radius constraint that ensures the *stability* of the underlying VAR process (see Basu and Michailidis, 2015; Lin and Michailidis, 2017, for details). For large VAR models, this constraint implies a smaller magnitude on average for all model parameters, which makes their estimation more challenging, unless one compensates with a higher level of sparsity. Nevertheless, very sparse VAR models may not be adequately informative, while their estimation requires larger penalties that in turn induce higher bias due to shrinkage, when the sample size stays fixed.

The DFM model aims to decompose a large number of time series into a few common latent factors and idiosyncratic components. The premise is that these common factors are the key drivers of the observed data, which themselves can exhibit temporal dynamics. They have been extensively used for forecasting purposes in economics (Stock and Watson, 2002), while their statistical properties have been studied in depth (see Bai and Ng, 2008, and references therein). Despite their ability to handle very large number of time series, theoretically appealing properties and extensive use in empirical work in economics, DFMs aggregate the underlying time series and hence are not suitable for examining their individual cross-dependencies. Since in many applications researchers are primarily interested in understanding the interactions between key variables (Sims, 1980; Stock and Watson, 2016),

1. such as the concept of output gap for monetary policy, the latter defined as the difference between the actual output of an economy and its potential output

while accounting for the influence of many others so as to avoid model misspecification that leads to biased results, DFMs may not be the most appropriate model.

To that end, Bernanke et al. (2005) proposed a "fusion" model, namely the Factor Augmented VAR, that aims to summarize the information contained in a large set of time series by a small number of factors and includes those in a standard VAR model. Specifically, let $\{F_t\} \in \mathbb{R}^{p_1}$ be the latent factor and $\{X_t\} \in \mathbb{R}^{p_2}$ the observed sets of variables, they jointly form a VAR system given by

$$\begin{bmatrix} F_t \\ X_t \end{bmatrix} = A^{(1)} \begin{bmatrix} F_{t-1} \\ X_{t-1} \end{bmatrix} + \cdots + A^{(d)} \begin{bmatrix} F_{t-d} \\ X_{t-d} \end{bmatrix} + \begin{bmatrix} w_t^F \\ w_t^X \end{bmatrix}. \tag{1}$$

In addition, there is a large panel of observed time series $Y_t \in \mathbb{R}^q$, whose current values are influenced by both X_t and F_t; i.e., the calibration equation:

$$Y_t = \Lambda F_t + \Gamma X_t + e_t. \tag{2}$$

The primary variables of interest X_t together with the unobserved factors F_t—both are assumed to have small and fixed dimensions—drives the dynamics of the system, and the factors are inferred from (2).

Even in the low-dimensional setting (p_2 fixed), there is very limited theoretical work (Bai et al., 2016) on the FAVAR model and some work on identification restrictions for the model parameters (e.g. Bernanke et al., 2005). However, the fixed dimensionality assumption is rather restrictive in many applications; in particular, the model has been extensively used in empirical work in economics and finance (e.g. Eickmeier et al., 2014; Caggiano et al., 2014), yet customarily a very small size block X_t is considered. For example, in Bernanke et al. (2005) that introduces the FAVAR model, X_t comprises of three "core" economic indicators (industrial production, consumer price index and the federal funds rate) and Y_t of 120 other economic indicators. The VAR system is augmented by one factor summarizing the macroeconomic indicators, and the augmented system shows 7-lag time dependence that significantly increases the sample size requirement for estimation purposes. In a recent application, Stock and Watson (2016) apply the FAVAR model to macroeconomics effects of oil supply shocks; the augmented VAR system consists of 8 times series (observed and latent), but due to the limitation in sample size to avoid non-stationarities ($T = 120$) the lag of the model is fixed to 1. Hence, as argued in Stock and Watson (2016), there is growing need for large scale FAVAR models and this paper aims to examine their estimation and theoretical properties in high-dimensions, leveraging sparsity constraints on key model parameters.

The key contributions of this paper are twofold: (1) the introduction of an identifiability constraint compatible with the high-dimensional nature of the model, under sparsity assumptions on model parameters Γ and $\{A^{(k)}\}$, and (2) the ensuing formulation of the optimization problem that leads to their estimators based on observational data and estimators' high-probability error bounds. At the technical level there are two sets of challenges that are successfully resolved: (i) the calibration equation involves both an observed set of covariates and a set of latent factors, and their interactions require careful handling to enable accurate estimation of the factors that constitute part of the input to the augmented VAR system and are crucial for estimating the transition matrix; and (ii) with the presence

of a block of variables in the VAR system that are subject to error due to being estimated rather than directly observed, a number of new technical challenges emerge and they are compounded by the presence of temporal dependence. Note that for ease of presentation, the main technical developments are shown for Gaussian data (all noise processes in (1) and (2) are assumed to be Gaussian), but the key theoretical results are also established for sub-Gaussian and sub-exponential error processes; see Appendix C for a result of independent theoretical interest, even for the standard sparse VAR model.

Outline of the paper. The remainder of the paper is organized as follows. In Section 2, the model identifiability constraint is introduced, followed by formulation of the objective function to be optimized that obtains estimates of the model parameters. Theoretical properties of the proposed estimators, specifically, their high probability finite-sample error bounds, are investigated in Section 3. Subsequently in Section 4, we introduce an empirical implementation procedure for obtaining the estimates and present its performance evaluation based on synthetic data. An application of the model on interlinkages of commodity prices and the influence of world macroeconomic indicators on them is presented in Section 5, while Section 6 provides some concluding remarks. All proofs and other supplementary materials are deferred to Appendices.

Notations. Throughout this paper, we use $|||A|||_\cdot$ to denote matrix norms for some generic matrix $A \in \mathbb{R}^{m\times n}$. For example, $|||A|||_1$ and $|||A|||_\infty$ respectively denote the matrix induced 1-norm and infinity norm, $|||A|||_{\text{op}}$ the matrix operator norm and $|||A|||_{\text{F}}$ the Frobenius norm. Moreover, We use $\|A\|_1$ and $\|A\|_\infty$ respectively to denote the element-wise 1-norm and infinity norm. For two matrices A and B of commensurate dimensions, denote their inner product by $\langle\!\langle A, B\rangle\!\rangle = \text{tr}(A^\top B)$. Finally, we write $A \gtrsim B$ if there exists some absolute constant c that is independent of the model parameters such that $A \geq cB$; and $A \asymp B$ if $A \gtrsim B$ and $B \gtrsim A$ hold simultaneously.

2. Model Identification and Problem Formulation

The FAVAR model proposed in Bernanke et al. (2005) has the following two components, as seen in Section 1: a system given in (1) that describes the dynamics of the latent block $F_t \in \mathbb{R}^{p_1}$ and the observed block $X_t \in \mathbb{R}^{p_2}$ that jointly follow a stationary VAR(d) model (the "VAR equation"); and the model in (2) that characterizes the contemporaneous dependence of the large observed informational series $Y_t \in \mathbb{R}^q$ as a linear function of X_t and F_t (the "calibration equation"). Further, w_t^F, w_t^X and e_t are all noise terms that are independent of the predictors, and we assume they are serially uncorrelated mean-zero Gaussian random vectors: $w_t^F \sim \mathcal{N}(0, \Sigma_w^F)$, $w_t^X \sim \mathcal{N}(0, \Sigma_w^X)$ and $e_t \sim \mathcal{N}(0, \Sigma_e)$. In this study we consider a potentially large VAR system that has many coordinates, hence in contrast to Bernanke et al. (2005) and Bai et al. (2016) where both p_1 and p_2 are fixed and small, we allow the size of the observed block, p_2, to be large[2] and to grow with the sample size; yet the size of the latent block, p_1, can not be too large and is still assumed fixed. Moreover, the size of the informational series, q, can also be large and grow with the

2. We do not impose the restriction that p_2 is smaller than the available sample size.

sample size. Further, we assume that the transition matrices $\{A^{(i)}\}_{i=1}^{d}$ and the regression coefficient matrix Γ are *sparse*. Finally, the factor loading matrix Λ is assumed to be dense.

2.1. Model identification considerations

The latent nature of F_t leads to the following observational equivalence across the following two models encoded by (Λ, Γ) and $(\widetilde{\Lambda}, \widetilde{\Gamma})$, respectively: for any invertible matrix $Q_1 \in \mathbb{R}^{p_1 \times p_1}$ and $Q_2 \in \mathbb{R}^{p_1 \times p_2}$,

$$Y_t = \Lambda F_t + \Gamma X_t + e_t \;\equiv\; \widetilde{\Lambda}\widetilde{F}_t + \widetilde{\Gamma} X_t + e_t, \qquad (Y_t \in \mathbb{R}^q, F_t \in \mathbb{R}^{p_1}, X_t \in \mathbb{R}^{p_2})$$

where

$$\widetilde{\Lambda} := \Lambda Q_1, \qquad \widetilde{F}_t := Q_1^{-1} F_t - Q_1^{-1} Q_2 X_t, \qquad \widetilde{\Gamma} := \Gamma + \Lambda Q_2. \tag{3}$$

In other words, the key model parameters (Λ, Γ) and the latent factors F_t are *not uniquely* identified, a known problem even in classical factor analysis (Anderson, 1958). Thus, additional restrictions are required to overcome this indeterminacy, since there is an equivalence class parametrized by (Q_1, Q_2) within which individual models are not mutually distinguishable based on observational data. For the FAVAR model, a total number of $p_1^2 + p_1 p_2$ restrictions are needed for unique identification of Λ, Γ and F_t.

Various schemes have been proposed in the literature to address this issue. Specifically, Bernanke et al. (2005) impose the necessary restrictions through the coefficient matrices of the calibration equation, requiring $\Lambda = \left[\begin{smallmatrix} \mathrm{I}_{p_1} \\ * \end{smallmatrix}\right]$ and $\Gamma_{[1:p_1],\cdot} = 0$; that is, the upper $p_1 \times p_1$ block of Λ is set to the identity matrix and the first p_1 rows of Γ to zero. Bai et al. (2016) consider different sets of restrictions that involve combinations of coefficients from the calibration equation and the noise term from the VAR equation. In the low-dimensional setting (p_2 fixed), one can proceed to estimate the parameters subject to these restrictions, by adopting either a single-step Bayesian likelihood approach (Bernanke et al., 2005) or an orthogonal projection-based approach by profiling out X_t (Bai et al., 2016). However, neither approach is applicable in high-dimensional settings, due to the growing dimension p_2 which would render a projection-based approach infeasible or add to the computational demands of a Bayesian procedure.

To overcome these issues in high-dimensional settings, we introduce an alternative identification scheme "IR+Compactness" that is compatible with the model specification and can also be seamlessly incorporated in the estimation procedure, leveraging sparsity of the regression coefficient Γ. Specifically, we first impose constraint (IR):

(IR) $\Lambda = \left[\begin{smallmatrix} \mathrm{I}_{p_1} \\ * \end{smallmatrix}\right]$: the upper $p_1 \times p_1$ block of Λ is an identity matrix, while the bottom block is left unconstrained.

Note that (IR) imposes p_1^2 constraints but crucially not on the latent factors, given their subsequent utilization in the VAR system. Further, it yields uniquely identifiable Λ and F_t, for any given product ΛF_t, and the indeterminacy incurred by $Q_1 \in \mathbb{R}^{p_1 \times p_1}$ in (3) vanishes.

However, the issue is not fully resolved, since for any $Q_2 \in \mathbb{R}^{p_1 \times p_2}$, the following relationship holds:

$$Y_t = \Lambda F_t + \Gamma X_t + e_t \equiv \Lambda \check{F}_t + \check{\Gamma} X_t + e_t,$$

where

$$\check{F}_t = F_t - Q_2 X_t \qquad \check{\Gamma} := \Gamma + \Lambda Q_2. \tag{4}$$

All such models encoded by $(\check{F}_t, \check{\Gamma})$, form an equivalence class parametrized by Q_2 that specifies the transformation. We denote this equivalence class by $\mathcal{C}(Q_2)$. If $Q_2 = O$, then $\mathcal{C}(Q_2)$ degenerates to a singleton that contains only the true data-generating model, which requires the imposition of $p_1 p_2$ restrictions on primary model quantities. One applicable constraint out of theoretical consideration is to impose orthogonality on X_t and F_t — it yields the necessary $p_1 p_2$ restrictions; yet is excessively stringent and limits the appeal of the FAVAR model, while also being challenging to operationalize. Therefore as a good working alternative, we address the identifiability issue through a weaker constraint that effectively limits sufficiently the size of the $\mathcal{C}(Q_2)$.

To this end, let $\mathbf{X} \in \mathbb{R}^{n \times p_2}$, $\mathbf{Y} \in \mathbb{R}^{n \times q}$ and $\mathbf{F} \in \mathbb{R}^{n \times p_1}$ be centered data matrices whose rows are samples of X_t, Y_t and the latent process F_t respectively, and $\check{\mathbf{F}}$ is analogously defined. The characterization of $\mathcal{C}(Q_2)$ is through the sample versions of the underlying processes. Specifically, define the set of *factor hyperplanes* induced by $\mathcal{C}(Q_2)$ by

$$\mathcal{S}(\check{\Theta}) := \{\check{\Theta} := \check{\mathbf{F}}\Lambda^\top \in \mathbb{R}^{n \times q} \mid \check{\mathbf{F}} \text{ are samples of } \check{F}_t \text{ defined through (4)}\},$$

and we let $\Theta^\star$ denote the factor hyperplane associated with the true data-generating model, to distinguish it from some generic element in $\mathcal{S}(\check{\Theta})$ that is denoted by $\check{\Theta}$. Note that $\Theta^\star \in \mathcal{S}(\check{\Theta})$ and $\check{\Theta}$ coincides with $\Theta^\star$ when $Q_2 = 0$. Moreover, all elements in $\mathcal{S}(\check{\Theta})$ are at most of rank p_1, hence a low-rank component relative to their size $n \times q$. Next, in a similar spirit to Negahban and Wainwright (2012), we define the following constrained set:

$$\mathcal{S}_\phi(\check{\Theta}) := \big\{\varphi_{\mathcal{R}}(\check{\Theta}) \le \phi(n,q) \;\big|\; \check{\Theta} \in \mathcal{S}(\check{\Theta})\big\},$$

where $\varphi_{\mathcal{R}}(\Theta)$ is defined according to

$$\varphi_{\mathcal{R}}(\Theta) := \kappa(\mathcal{R}^*)\mathcal{R}^*(\Theta)\|\mathbf{X}/\sqrt{n}\|_{\text{op}},$$

and $\kappa(\mathcal{R}^*) := \sup_{\Theta \neq 0} \big(\|\Theta\|_{\text{F}}/\mathcal{R}^*(\Theta)\big)$ with $\mathcal{R}^*$ being the dual norm of some regularizer $\mathcal{R}$. Base on the above definition, $\varphi_{\mathcal{R}}(\Theta)$ captures the interaction between the factor space and the observed $\mathbf{X}$-space; the product $\kappa(\mathcal{R}^*)\mathcal{R}^*(\Theta)$ measures the spikiness of Θ w.r.t. $\mathcal{R}$, and in the case where $\mathcal{R}$ corresponds to the sparsity-induced ℓ_1-norm which would be the setup of interest in this paper (see Section 2.2), $\mathcal{R}^*(\Theta) = \|\Theta\|_\infty$ and $\kappa(\mathcal{R}^*) = \sqrt{nq}$. With the definition of $\mathcal{S}_\phi(\check{\Theta})$, we impose the following compactness constraint on $\check{\Theta}$ to further encourage identifiability:

(Compactness) $\check{\Theta} \in \mathcal{S}_\phi(\check{\Theta})$ for some $\phi(n,q)$ satisfying $\phi(n,q) \ge \phi^\star := \varphi_{\mathcal{R}}(\Theta^\star)$.

(Compactness) effectively limits the spikiness of all possible $\check{\Theta}$'s by imposing a *box constraint* through the dual norm corresponding to the sparsity regularizer, and for an arbitrary set of fixed realizations, it restricts the factor hyperplane set induced by $\mathcal{C}(Q_2)$ to its ϕ-radius subset $\mathcal{S}_\phi(\check{\Theta})$. This in turn limits the size of the equivalence class $\mathcal{C}(Q_2)$ under consideration, since there is a one-to-one correspondence at the set level between $\mathcal{C}(Q_2)$ and the factor hyperplane set induced by it. This further implies that although the models encoded by (F_t, Γ) and $(\check{F}_t, \check{\Gamma})$ may not be perfectly distinguishable based on observational data, at the

population level the discordance between the two models can not be too large. It is worth pointing out that the bound $\phi(n, q)$ is allowed to grow, but at a much slower rate than the size of $\breve{\Theta}$; specifically, we require $\phi(n, q) = o(\kappa(\mathcal{R}^*))$. For ease of presentation, we use ϕ to denote this bound henceforth and further note that it is in fact a constant in any finite sample setting.

In summary, our proposed identification scheme comprises of two parts: (IR) and (Compactness). The former provides exact identification within the factor hyperplane and narrows the scope of observationally equivalent models to $\mathcal{C}(Q_2)$, while the latter limits its size; and they jointly incur *approximate identification* of the true data generating model; and thus, for estimation purposes henceforth, it becomes adequate to focus on this restricted equivalence class, rather than its individual elements. The proposed scheme is suitable for the high-dimensional nature of the problem and can easily be incorporated in the formulation of the optimization problem for parameter estimation (see Section 2.2), which in turn yields estimates with tight error bounds (see Section 3).

2.2. Proposed formulation

Without loss of generality, we focus on the case where $d = 1$ in subsequent technical developments, so that $Z_t := (F_t^\top, X_t^\top)^\top$ follows a VAR(1) model $Z_t = AZ_{t-1} + W_t$:

$$\begin{bmatrix} F_t \\ X_t \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} F_{t-1} \\ X_{t-1} \end{bmatrix} + \begin{bmatrix} w_t^F \\ w_t^X \end{bmatrix}. \tag{5}$$

The generalization to the VAR(d) ($d > 1$) case is straightforward since for any generic VAR(d) process satisfying $\mathcal{A}_d(L)Z_t = w_t$ where $\mathcal{A}_d(L) := \mathrm{I} - A^{(1)}L - \cdots - A^{(d)}L^d$, it can always be written in the form of a VAR(1) model for some dp-dimensional process $\widetilde{Z}_t$ (see Lütkepohl, 2005, for details).

Based on the introduced model identification scheme (IR+Compactness), we propose the following procedure to estimate the FAVAR model, whose parameters include a sparse coefficient matrix Γ, a dense loading matrix Λ, and a sparse transition matrix A. Observed data matrices $\mathbf{X}$ and $\mathbf{Y}$ are identical to what have been previously defined, and to distinguish the responses from their lagged predictors when considering the VAR system, we let $\mathbf{X}_{n-1} := [x_1, \ldots, x_{n-1}]^\top$ denote the predictor matrix and $\mathbf{X}_n := [x_2, \ldots, x_n]^\top$ the response one; $\mathbf{F}_n, \mathbf{F}_{n-1}, \mathbf{Z}_n, \mathbf{Z}_{n-1}$ are analogously defined. Based on these notations, the sample versions of the VAR system and the calibration equation in (5) and (2) can be written as

$$\mathbf{Z}_n = \mathbf{Z}_{n-1}A^\top + \mathbf{W}, \quad \text{and} \quad \mathbf{Y} = \mathbf{F}\Lambda^\top + \mathbf{X}\Gamma^\top + \mathbf{E} =: \Theta + \mathbf{X}\Gamma^\top + \mathbf{E}.$$

We propose the following estimators obtained from a two-stage procedure for the coefficient matrices Λ, Γ and subsequently the transition matrices $\{A_{ij}\}_{i,j=1,2}$.

- Stage I: estimation of the calibration equation under (IR+Compactness). We formulate the following *constrained optimization* problem using a least squares loss function and incorporating the sparsity-induced ℓ_1 regularization of the sparse block Γ, the

rank constraint on the hyperplane Θ, and (Compactness):

$$\begin{aligned}(\widehat{\Theta},\widehat{\Gamma}) \;:=\; & \underset{\Theta\in\mathbb{R}^{n\times q},\Gamma\in\mathbb{R}^{q\times p_2}}{\arg\min} \big\{\frac{1}{2n}|||\mathbf{Y}-\Theta-\mathbf{X}\Gamma^\top|||_{\mathrm{F}}^2+\lambda_\Gamma\|\Gamma\|_1\big\},\\ & \text{subject to}\quad \mathrm{rank}(\Theta)\le r,\quad \|\Theta\|_\infty\le\frac{\phi}{\sqrt{nq}\cdot|||\mathbf{X}/\sqrt{n}|||_{\mathrm{op}}}.\end{aligned} \tag{6}$$

Once $\widehat{\Theta}$ is obtained, under (IR), the estimated factors $\widehat{\mathbf{F}}$ and the corresponding loading matrix $\widehat{\Lambda}$ are extracted as follows:

$$\widehat{\mathbf{F}}=\widehat{\mathbf{F}}^{\mathrm{PC}}(\widehat{\Lambda}_1^{\mathrm{PC}})^\top,\qquad \widehat{\Lambda}=\widehat{\Lambda}^{\mathrm{PC}}(\widehat{\Lambda}_1^{\mathrm{PC}})^{-1}, \tag{7}$$

where $\widehat{\Lambda}_1^{\mathrm{PC}}$ is the upper p_1 sub-block of $\widehat{\Lambda}^{\mathrm{PC}}$, with $\widehat{\mathbf{F}}^{\mathrm{PC}}$ and $\widehat{\Lambda}^{\mathrm{PC}}$ being the PC estimators (Stock and Watson, 2002) given by $\widehat{\mathbf{F}}^{\mathrm{PC}}:=\sqrt{n}\widehat{U}$ and $\widehat{\Lambda}^{\mathrm{PC}}:=\widehat{V}\widehat{D}/\sqrt{n}$. The estimates $\widehat{U},\widehat{D}$ and $\widehat{V}$ are obtained from the SVD of $\widehat{\Theta}=\widehat{U}\widehat{\Theta}\widehat{V}^\top$. Note that after these algebra, $\widehat{\mathbf{F}}$ corresponds to the first p_1 columns of $\widehat{\Theta}$.

Of note, $|||\mathbf{X}/\sqrt{n}|||_{\mathrm{op}}^2=\Lambda_{\max}(\mathbf{X}^\top\mathbf{X}/n)$ and it can be shown that for any random realizations $\mathbf{X}$, the latter can be bounded with high probability (see Lemma 5).

- Stage II: estimation of the VAR equation based on $\mathbf{X}$ and $\widehat{\mathbf{F}}$. With the estimated factor $\widehat{\mathbf{F}}$ as the surrogate for the true latent factor $\mathbf{F}$, the transition matrix A can be estimated by solving

$$\widehat{A}\;:=\;\underset{A\in\mathbb{R}^{(p_1+p_2)\times(p_1+p_2)}}{\arg\min}\big\{\frac{1}{2n}|||\widehat{\mathbf{Z}}_n-\widehat{\mathbf{Z}}_{n-1}A^\top|||_{\mathrm{F}}^2+\lambda_A\|A\|_1\big\}, \tag{8}$$

where $\widehat{\mathbf{Z}}_n:=[\widehat{\mathbf{F}}_n,\mathbf{X}_n]$ and $\widehat{\mathbf{Z}}_{n-1}$ is analogously defined. The ℓ_1-norm penalty induces sparsity on A according to the model assumption.

In the presence of additional contemporaneous dependence amongst the coordinates for the error processes w_t, one may consider a maximum likelihood-based loss function, but the full estimation would require additional structural assumptions of Σ_w (or its inverse) given the high dimensionality; we do not further elaborate in this study, since our prime interest is estimating the coefficient/transition matrices of the FAVAR model.

The formulation in (8) based on the least squares loss function and the surrogate $\widehat{\mathbf{F}}$ is straightforward. However, the formulation for the calibration equation merits additional discussion. First, note that the factor hyperplane Θ has at most rank p_1 and therefore has low rank structure relative to its size $n\times q$. We impose a rank constraint in the estimation procedure to enforce such structure. Together with the (IR+Compactness) constraint introduced above, the objective then becomes to estimate accurately the parameters of a model within the equivalence class $\mathcal{C}(Q_2)$, in the sense that the estimate obtained by solving (6) effectively corresponds to recovering an arbitrary $\check{\Theta},\check{\Theta}\in\mathcal{C}(Q_2)$; such an estimate, however, will be close to the true data generating $\Theta^\star$. Once this goal is achieved, this would enable accurate estimation of the transition matrix of the VAR system.

From an optimization perspective, the objective function admits a low-rank-plus-sparse decomposition and compactification is necessary for establishing statistical properties of

the global optima in the absence of explicitly specifying the interaction structure between the low rank and the sparse blocks (or the spaces they live in). Note that the form of the compactness constraint is dictated by the statistical problem under consideration. For example, Agarwal et al. (2012) study a multivariate regression problem, where the coefficient is decomposed to a sparse and a low rank block. In that setting, a compactness constraint is imposed through the entry-wise infinity norm bound of the low rank block. Chandrasekaran et al. (2012) study a graphical model with latent variables where the conditional concentration matrix is the parameter of interest. The marginal concentration matrix is decomposed to a sparse and a low rank block via the alignment of the Schur complement, and the compactness constraint is imposed on both blocks and manifests through the corresponding regularization terms in the resulting optimization problem. Hence, the compactness constraint takes different forms but ultimately serves the same goal, namely, to introduce an upper bound on the magnitude of the low rank–sparse block interaction, with the latter being an important component in analyzing the estimation errors. The compacteness constraint adopted for the FAVAR model serves a similar purpose, although the presence of temporal dependence introduces a number of additional technical challenges compared to the two aforementioned settings that consider independent and identically distributed data.

Finally, we remark that the model identification scheme (IR+Compactness) incorporated in the optimization problem as a constraint, enables us to establish high-probability error bounds (relative to the true data generating parameters/factors) for the proposed estimators, as shown next in Section 3. Therefore, although (IR+Compactness) does not encompass the full $p_1^2 + p_1 p_2$ restrictions, it provides sufficient identifiability for estimation purposes.

3. Theoretical Properties

In this section, we investigate the theoretical properties of the estimators proposed in Section 2.2. We focus on formulations (6) and (8), whose global optima correspond to $(\widehat{\Theta}, \widehat{\Gamma})$ and $\widehat{A}$, respectively.

Since (8) relies not only on prime observable quantities (namely X_t), but also on estimated quantities from Stage I (namely $\widehat{\mathbf{F}}$), the analysis requires a careful examination of how the estimation error in the factor propagates to that of $\widehat{A}$. We start by outlining a road map of our proof strategy together with a number of regularity conditions needed in subsequent developments. Section 3.1 establishes error bounds for $\widehat{\Gamma}$, $\widehat{\Theta}$ [3]and $\widehat{A}$ under certain regularity conditions and employing suitable choices of the tuning parameters, for *deterministic realizations* from the underlying observable processes. Specifically when considering the error bound of $\widehat{A}$, the error of the plug-in estimate $\widehat{\mathbf{F}}$ is assumed non-random and given. Subsequently, Section 3.2 examines the probability of the events in which the regularity conditions are satisfied for *random realizations*, and further establishes high-probability upper bounds for quantities to which the tuning parameters need to conform. Finally, the high-probability finite sample error bounds for the estimates obtained based on random realizations of the data generating processes readily follow after properly aligning the conditioning arguments, and the results are presented in Section 3.3. All proofs are deferred to Appendices A and B.

3. Consequently, the error bounds of $\widehat{\mathbf{F}}$ and $\widehat{\Lambda}$ under (IR) are also obtained.

Additional notations. Throughout, we use superscript $\star$ to denote the true value of the parameters of interest, and Δ for errors of the estimators; e.g., $\Delta_A = \widehat{A} - A^\star$. For sample quantities (e.g., $\mathbf{X}$ and $\mathbf{F}$) and their corresponding error (e.g., $\Delta_{\mathbf{F}}$), we use subscript $(n-1)$ to denote their first $n-1$ rows. We let $S_{\mathbf{E}} := \frac{1}{n}\mathbf{E}^\top\mathbf{E}$ denote the sample covariance matrix of $\mathbf{E}$ and the sample covariance of other quantities are analogously defined. Additionally, denote the density level of $\Gamma^\star$ by $s_{\Gamma^\star} := \|\Gamma^\star\|_0$, and that of $A^\star$ by $s_{A^\star}$.

A road map for establishing consistency results. As previously mentioned, the key steps are:

- Part 1: analyses based on deterministic realizations using the optimality of the estimators, assuming the parameters of the objective function (e.g., the Hessian and the penalty parameter) satisfy certain regularity conditions;
- Part 2: analyses based on random realizations that the probability of the regularity conditions being satisfied, primarily involving the utilization of concentration inequalities.

In Part 1, note that the first-stage estimators obtained from the calibration equation are based on observed data and thus the regularity conditions needed are imposed on (functions of) the observed samples. On the other hand, the second-stage estimator relies on the plugged-in first-stage estimates that have bounded errors; therefore, the analysis is carried out in an analogous manner to problems involving error-in-variables. Specifically, the required regularity conditions on quantities appearing in the optimization (8) involve the error of the first stage estimates, with the latter assumed fixed. In Part 2, the focus shifts to the probability of the regularity conditions being satisfied under random realizations, again starting from the first stage estimates, with the aid of Gaussian concentration inequalities and proper accounting for temporal dependence. Once the required regularity conditions are shown to hold with high probability, combining the results established in Part 1 for deterministic realizations, the high-probability error bounds for $\widehat{\Theta}$ and $\widehat{\Gamma}$ are established. The high-probability error bound of the estimated factors readily follows, which ensures that the variables which Stage II estimates rely upon are sufficiently accurate with high probability. Based on the latter result, the regularity conditions required for the Stage II estimates are then verified to hold with high probability at a certain rate. In the FAVAR model, since the estimation of the VAR equation is based on quantities among which one block is subject to error, to obtain an accurate estimate of the transition matrix requires more stringent conditions on population quantities (e.g., extremes of the spectrum), so that the regularity conditions hold with high probability. In essence, the joint process Z_t need to be adequately "regular" in order to get good estimates of the transition matrix , vis-a-vis the case of the standard VAR model where all variables are directly observed.

Next, we introduce the following key concepts that are widely used in establishing theoretical properties of high-dimensional regularized M-estimators (e.g. Negahban et al., 2012; Loh and Wainwright, 2012), as well as quantities that are related to processes exhibiting temporal dependence (see also Basu and Michailidis, 2015).

Definition 1 (Restricted strong convexity (RSC)) *A matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ satisfies the RSC condition with respect to norm Φ with curvature $\alpha_{RSC} > 0$ and tolerance $\tau_n \geq 0$, if*

$$\frac{1}{2n}|||\mathbf{X}\Delta|||_F^2 \geq \frac{\alpha_{RSC}}{2}|||\Delta|||_F^2 - \tau_n\Phi^2(\Delta), \qquad \forall\ \Delta \in \mathbb{R}^{p\times p}.$$

In our setting, we consider the norm $\Phi(\Delta) = \|\Delta\|_1$.

Definition 2 (Deviation condition) *For a regularized M-estimator given in the generic form of*

$$\widehat{A} := \min_A \big\{\frac{1}{2n}|||\mathbf{Y} - \mathbf{X}A^\top|||_F^2 + \lambda_A\|A\|_1\big\},$$

with $\mathcal{H}_A := \frac{1}{n}\mathbf{X}^\top\mathbf{X}$ denoting the Hessian and $\mathcal{G}_A := \frac{1}{n}\mathbf{Y}^\top\mathbf{X}$ denoting the gradient, we define the tuning parameter λ_A to be selected in accordance with the deviation condition, if

$$\lambda_A \geq c_0\|\mathcal{H}_A - \mathcal{G}_A(A^\star)^\top\|_\infty, \quad \textit{for some } c_0.$$

Under the current model setup, however, the exact form of the deviation bound becomes more involved and requires proper modifications to incorporate quantities associated with the factor hyperplane, as seen in Proposition 1.

Definition 3 (Spectrum and its extremes) *For a p-dimensional stationary process X_t, its spectral density $f_X(\omega)$ is defined as*

$$f_X(\omega) := \tfrac{1}{2\pi}\sum_{h=-\infty}^{\infty}\Sigma_X(h)e^{i\omega h}, \tag{9}$$

where $\Sigma_X(h) := \mathbb{E}(X_tX_{t+h}^\top)$. Its upper and lower extremes are defined as

$$\mathcal{M}(f_X) := \operatorname*{ess\ sup}_{\omega\in[-\pi,\pi]}\Lambda_{\max}(f_X(\omega)), \quad \textit{and} \quad \mathfrak{m}(f_X) := \operatorname*{ess\ inf}_{\omega\in[-\pi,\pi]}\Lambda_{\min}(f_X(\omega)).$$

The cross-spectrum for two generic stationary processes X_t and Y_t is defined as

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

where $\Sigma_{X,Y}(h) := \mathbb{E}(X_tY_{t+h}^\top)$, and its upper extreme is defined as

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

where $*$ denotes the conjugate transpose.

We start by providing error bounds for $\widehat{\Gamma}$ and $\widehat{\Theta}$, as well as those of the corresponding $\widehat{\mathbf{F}}$ and $\widehat{\Lambda}$ extracted under (IR). For the optimization problem given in (6), we assume that $r \geq p_1$ and ϕ is always compatible with the true data generating mechanism, so that $\Theta^\star$ is always feasible. To this end, the error bounds of $\widehat{\Theta}$ and $\widehat{\Gamma}$ for deterministic realizations

crucially rely on two components: (i) $\mathbf{X}$ satisfying the RSC condition with curvature $\alpha^{\mathbf{X}}_{\text{RSC}}$; and (ii) the tuning parameter λ_Γ being chosen in accordance with the deviation bound condition that is associated with the interaction between $\mathbf{X}$ and $\mathbf{E}$, the strength of the noise, and the interaction between the space spanned by the factor hyperplane and the observed $\mathbf{X}$. Upon the satisfaction of these conditions, the error bounds of $\widehat{\Theta}$ and $\widehat{\Gamma}$ are given by

$$|||\Delta_\Gamma|||_{\text{F}}^2 + |||\Delta_\Theta/\sqrt{n}|||_{\text{F}}^2 \leq C_1\lambda_\Gamma^2\big(p_1 + r + 4s_{\Gamma^\star}\big)/\min\{\alpha^{\mathbf{X}}_{\text{RSC}}, 1\}^2,$$

and these conditions hold with high probability for random realizations of X_t and Y_t. Since $\widehat{\mathbf{F}}$ is the first p_1 columns of $\widehat{\Theta}$, it possesses an error bound of the similar form.

Next, we briefly sketch the error bounds of $\widehat{A}$. For the optimization in (8), for deterministic realizations, the results in Basu and Michailidis (2015) can be applied with the corresponding RSC condition and deviation condition imposed on quantities associated with $\widehat{\mathbf{Z}}_n$ and $\widehat{\mathbf{Z}}_{n-1}$, and the error for $\widehat{A}$ is in the form of

$$|||\Delta_A|||_{\text{F}}^2 \leq C_2 s_{A^\star}\lambda_A^2/(\alpha^{\widehat{\mathbf{Z}}}_{RSC})^2.$$

Then, for random realizations, assuming $\Delta_{\mathbf{F}}$ known and non-random, to satisfy the corresponding regularity conditions, we additionally require that the following functional involving the spectral density of the underlying joint process Z_t exhibits adequate curvature, that is, $\mathfrak{m}(f_Z)/\sqrt{\mathcal{M}(f_Z)} > c_0 h_1(\Delta_{\mathbf{F}_{n-1}})$ for constant c_0 and some function h_1 of the error $\Delta_{\mathbf{F}_{n-1}}$ that captures its magnitude. Moreover, the deviation bound is of the form $h_2(\Delta_{\mathbf{F}})$, which can be viewed as another function of the error[4]. Further, since $\Delta_{\mathbf{F}}$ is bounded with high probability from the analysis in Stage I, it will be established that $h_1(\Delta_{\mathbf{F}})$ and $h_2(\Delta_{\mathbf{F}})$ are both upper bounded at a certain rate, thus ensuring that the RSC condition and the deviation conditions can both be satisfied unconditionally, by properly choosing the required constants.

3.1. Statistical error bounds with deterministic realizations

Proposition 1 below gives the error bounds for the estimators in (6), assuming certain regularity conditions hold for deterministic realizations of the processes X_t and Y_t, upon suitable choice of the regularization parameters.

Proposition 1 (Bound for Δ_Θ and Δ_Γ under fixed realizations) *Suppose the fixed realizations $\mathbf{X} \in \mathbb{R}^{n\times p_2}$ of process $\{X_t \in \mathbb{R}^{p_2}\}$ satisfies the RSC condition with curvature $\alpha^{\mathbf{X}}_{RSC} > 0$ and a tolerance $\tau_{\mathbf{X}}$ for which*

$$\tau_{\mathbf{X}} \cdot \big(p_1 + r + 4s_{\Gamma^\star}\big) < \min\{\alpha^{\mathbf{X}}_{RSC}, 1\}/16.$$

Then, for any matrix pair $(\Theta^\star, \Gamma^\star)$ satisfying the constraint $\varphi_{\mathcal{R}}(\Theta^\star) \leq \phi$ that generates $\mathbf{Y}$, for estimators $(\widehat{\Theta}, \widehat{\Gamma})$ obtained by solving (6) with regularization parameters λ_Γ satisfying

$$\lambda_\Gamma \geq \max\big\{2\|\mathbf{X}^\top\mathbf{E}/n\|_\infty,\ 4\phi/\sqrt{nq},\ \Lambda^{1/2}_{\max}(S_{\mathbf{E}})\big\},$$

4. note the deviation bound in principle also depends on other population quantities such as $\mathfrak{m}(f_Z)$, $\mathcal{M}(f_Z)$, $\Lambda_{\max}(\Sigma_w)$ etc.

the following bound holds:

$$\|\Delta_\Gamma\|_F^2 + \|\Delta_\Theta/\sqrt{n}\|_F^2 \leq \frac{16\lambda_\Gamma^2\big(p_1 + r + 4s_{\Gamma^\star}\big)}{\min\{\alpha_{RSC}^{\mathbf{X}}, 1\}^2}. \tag{10}$$

Based on Proposition 1, under fixed realizations of X_t and Y_t, the error bounds of $\widehat{\Gamma}$ and $\widehat{\Theta}$ are established. Using these Stage I estimates and the IR condition, estimates of the factors and their loadings can be calculated. In particular, since $\Delta_{\mathbf{F}}$ corresponds to the first p_1 columns of Δ_Θ, the above bound automatically holds for $\Delta_{\mathbf{F}}$. Further, the following lemma provides the relative error of the estimated Λ under (IR) and the condition on $\Lambda_{\max}^{1/2}(S_{\mathbf{F}})$, with the latter translating to the requirement that the leading signal of $\mathbf{F}$ overrules the averaged row error of Δ_Θ.

Lemma 1 (Bound of Δ_Λ) *The following error bound holds for $\widehat{\Lambda}$, provided that $\Lambda_{\max}^{1/2}(S_{\mathbf{F}}) > \|\Delta_\Theta/\sqrt{n}\|_F$:*

$$\frac{\|\Delta_\Lambda\|_F}{\|\Lambda^\star\|_F} \leq \frac{\sqrt{p_1}\cdot\|\Delta_\Theta/\sqrt{n}\|_F}{\Lambda_{\max}^{1/2}(S_{\mathbf{F}}) - \|\Delta_\Theta/\sqrt{n}\|_F}\Big(1 + 1/\|\Lambda^\star\|_F\Big). \tag{11}$$

Up to this point, error bounds have been obtained for all the parameters in the calibration equation. The following proposition establishes the error bound for the estimator obtained from solving (8), based on observed $\mathbf{X}$ and estimated $\widehat{\mathbf{F}}$, and assuming $\Delta_{\mathbf{F}}$ is fixed.

Proposition 2 (Bound for Δ_A under fixed realization and a non-random $\Delta_{\mathbf{F}}$) *Consider the estimator $\widehat{A}$ obtained by solving* (8)*. Suppose the following conditions hold:*

A1. $\widehat{\mathbf{Z}}_{n-1} := [\widehat{\mathbf{F}}_{n-1}, \mathbf{X}_{n-1}]$ *satisfies the RSC condition with curvature $\alpha_{RSC}^{\widehat{\mathbf{Z}}}$ and tolerance $\tau_{\mathbf{Z}}$ for which $s_{A^\star}\tau_{\mathbf{Z}} < \alpha_{RSC}^{\widehat{\mathbf{Z}}}/64$;*

A2. $\|\widehat{\mathbf{Z}}_{n-1}^\top\big(\widehat{\mathbf{Z}}_n - \widehat{\mathbf{Z}}_{n-1}(A^\star)^\top\big)/n\|_\infty \leq C(n, p_1, p_2)$ *where $C(n, p_1, p_2)$ is some function that depends on n, p_1 and p_2.*

Then, for any $\lambda_A \geq 4C(n, p_1, p_2)$, the following error bound holds for $\widehat{A}$:

$$\|\Delta_A\|_F \leq 16\sqrt{s_{A^\star}}\lambda_A/\alpha_{RSC}^{\widehat{\mathbf{Z}}}.$$

Note that Proposition 2 applies the results in Basu and Michailidis (2015, Proposition 4.1) to the setting in this study, where Stage II estimation of the transition matrix is based on $\widehat{\mathbf{Z}}_n$ and $\widehat{\mathbf{Z}}_{n-1}$; consequently, the regularity conditions should be imposed on corresponding quantities associated with $\widehat{\mathbf{Z}}_n$ and $\widehat{\mathbf{Z}}_{n-1}$.

Propositions 1 and 2 give finite sample error bounds for the estimators of the parameters obtained by solving optimization problems (6) and (8) based on fixed realizations of the observable processes X_t and Y_t, and the regularity conditions outlined. Next, we examine and verify these conditions for random realizations of the processes, to establish high probability error bounds for these estimators.

3.2. High probability bounds under random realizations

We provide high probability bounds or concentrations for the quantities associated with the required regularity conditions, for random realizations of X_t and Y_t. Specifically, we note that when X_t is considered separately from the joint system, it follows a high-dimensional VAR-X model (Lin and Michailidis, 2017)

$$X_t = A_{22}X_{t-1} + A_{21}F_{t-1} + w_t^X,$$

whose spectrum $f_X(\omega)$ satisfies

$$f_X(\omega) = \big[\mathcal{A}_X^{-1}(e^{-i\omega})\big]\big(A_{21}f_F(\omega)A_{21}^\top + f_{w^X}(\omega) + f_{w^X,F}(\omega)A_{21}^\top + A_{21}f_{F,w^X}(\omega)\big)\big[\mathcal{A}_X^{-1}(e^{-i\omega})\big]^*,$$

where $\mathcal{A}_X(L) := \mathrm{I} - A_{22}L$. Similar properties hold for F_t. Throughout, we assume $\{X_t\}, \{F_t\}$ and $\{Y_t\}$ are all mean-zero stable Gaussian processes.

Lemmas 2 to 4 respectively verify the RSC condition associated with $\mathbf{X}$ and establish the high probability bounds for $\|\mathbf{X}^\top\mathbf{E}/n\|_\infty$, $\Lambda_{\max}(S_{\mathbf{E}})$ and $\Lambda_{\max}(S_{\mathbf{X}})$.

Lemma 2 (Verification of the RSC condition for X) *Consider $\mathbf{X} \in \mathbb{R}^{n\times p_2}$ whose rows correspond to a random realization $\{x_1, \ldots, x_n\}$ of the stable Gaussian $\{X_t\}$ process, and its dynamics are governed by (5). Then, there exist positive constants $c_i > 0, i = 1, 2$, such that with probability at least $1 - c_1\exp(-c_2 n\min\{\gamma^{-2}, 1\})$ where $\gamma := 54\mathcal{M}(g_X)/\mathfrak{m}(g_X)$, the RSC condition holds for $\mathbf{X}$ with curvature $\alpha_{RSC}^{\mathbf{X}}$ and tolerance $\tau_{\mathbf{X}}$ satisfying*

$$\alpha_{RSC}^{\mathbf{X}} = \pi\mathfrak{m}(f_X), \qquad \tau_{\mathbf{X}} = \alpha_{RSC}\gamma^2\Big(\frac{\log p_2}{n}\Big)/2\ ,$$

provided that $n \gtrsim \log p_2$.

Lemma 3 (High probability bound for $\|\mathbf{X}^\top\mathbf{E}/n\|_\infty$) *There exist positive constants c_i ($i = 0, 1, 2$) such that for sample size $n \gtrsim \log(p_2 q)$, with probability at least $1-c_1\exp(-c_2\log(p_2q))$, the following bound holds:*

$$\|\mathbf{X}^\top\mathbf{E}/n\|_\infty \le c_0\Big(2\pi\mathcal{M}(f_X) + \Lambda_{\max}(\Sigma_e)\Big)\sqrt{\frac{\log p_2 + \log q}{n}}. \tag{12}$$

Lemma 4 (High probability bound for $\Lambda_{\max}(S_{\mathbf{E}})$) *Consider $\mathbf{E} \in \mathbb{R}^{n\times q}$ whose rows are independent realizations of the mean zero Gaussian random vector e_t with covariance Σ_e. Then, for sample size $n \gtrsim q$, with probability at least $1 - \exp(-n/2)$, the following bound holds:*

$$\Lambda_{\max}(S_{\mathbf{E}}) \le 9\Lambda_{\max}(\Sigma_e).$$

Lemma 5 (High probability bound for $\Lambda_{\max}(S_{\mathbf{X}})$) *Consider $\mathbf{X} \in \mathbb{R}^{n\times p_2}$ whose rows correspond to a random realization $\{x_1, \ldots, x_n\}$ of the stable Gaussian $\{X_t\}$ process, and its dynamics are governed by (5). There exist positive constants $c_i > 0, i = 0, 1, 2$, such that for sample size $n \gtrsim p_2$, with probability at least $1 - c_1\exp(-c_2 n)$, the following bound holds:*

$$\Lambda_{\max}(S_{\mathbf{X}}) \le c_0\mathcal{M}(f_X).$$

In the next two lemmas, we verify the RSC condition for random realizations of $\widehat{\mathbf{Z}}_{n-1}$ and obtain the high probability bound $C(n, p_1, p_2)$ for $\|\widehat{\mathbf{Z}}_{n-1}^\top\big(\widehat{\mathbf{Z}}_n - \widehat{\mathbf{Z}}_{n-1}(A^\star)^\top\big)/n\|_\infty$, with the underlying truth $\mathbf{F}$ being random but the error $\Delta_{\mathbf{F}}$ non-random. Note that this can be equivalently viewed as a *conditional* RSC condition and deviation bound, when conditioning on some fixed $\Delta_{\mathbf{F}}$.

Lemma 6 (Verification of RSC for $\widehat{\mathbf{Z}}_{n-1}$) *Consider $\widehat{\mathbf{Z}}_{n-1}$ given by*

$$\widehat{\mathbf{Z}}_{n-1} = \mathbf{Z}_{n-1} + \Delta_{\mathbf{Z}_{n-1}} = [\mathbf{F}_{n-1}, \mathbf{X}_{n-1}] + [\mathbf{\Delta}_{\mathbf{F}_{n-1}}, O],$$

with rows of $[\mathbf{F}_{n-1}, \mathbf{X}_{n-1}]$ being a random realization drawn from process $\{Z_t\}$ whose dynamics are given by (5). Suppose the lower and upper extremes of its spectral density $f_Z(\omega)$ satisfy

$$\mathfrak{m}(f_Z)/\mathcal{M}^{1/2}(f_Z) > c_0 \cdot \Lambda_{\max}^{1/2}\big(S_{\Delta_{\mathbf{F}_{n-1}}}\big), \quad \text{where} \;\; S_{\Delta_{\mathbf{F}_{n-1}}} := \Delta_{\mathbf{F}_{n-1}}^\top \Delta_{\mathbf{F}_{n-1}}/n,$$

for some constant $c_0 \geq 6$. Then, with probability at least $1 - c_1\exp(-c_2 n)$, $\widehat{\mathbf{Z}}_{n-1}$ satisfies the RSC condition with curvature

$$\alpha_{RSC}^{\widehat{\mathbf{Z}}} = \pi\mathfrak{m}(f_Z) - 54\Lambda_{\max}^{1/2}\big(S_{\Delta_{\mathbf{F}_{n-1}}}\big)\sqrt{2\pi\mathcal{M}(f_Z) + \pi\mathfrak{m}(f_Z)/27}, \tag{13}$$

and tolerance

$$\tau_n = \Big(\frac{\pi}{2}\mathfrak{m}(f_Z) + 27\Lambda_{\max}^{1/2}\big(S_{\Delta_{\mathbf{F}_{n-1}}}\big)\sqrt{2\pi\mathcal{M}(f_Z) + \pi\mathfrak{m}(f_Z)/27}\Big)\omega^2\sqrt{\frac{\log(p_1+p_2)}{n}},$$

where $\omega = 54\frac{\mathcal{M}(f_Z)}{\mathfrak{m}(f_Z)}$, provided that the sample size $n \gtrsim \log(p_1+p_2)$.

Lemma 7 (Deviation bound for $\|\widehat{\mathbf{Z}}_{n-1}^\top\big(\widehat{\mathbf{Z}}_n - \widehat{\mathbf{Z}}_{n-1}(A^\star)^\top\big)/n\|_\infty$) *There exist positive constants c_i $(i = 1, 2)$ and C_i $(i = 1, 2, 3)$ such that with probability at least $1 - c_1\exp\big(-c_2\log(p_1+p_2)\big)$ we have*

$$\begin{aligned} C(n, p_1, p_2) \leq\; & C_1\big[\mathcal{M}(f_Z) + \frac{\Lambda_{\max}(\Sigma_w)}{2\pi} + \mathcal{M}(f_{Z,W^+})\big]\sqrt{\frac{\log(p_1+p_2)}{n}} \\ & + \; C_2\big[\mathcal{M}^{1/2}(f_Z)\max_{j\in\{1,\ldots,p_1\}}\|\Delta_{\mathbf{F}_{n,\cdot j}}/\sqrt{n}\|\big]\sqrt{\frac{\log p_1 + \log(p_1+p_2)}{n}} \\ & + \; C_3\big[\Lambda_{\max}^{1/2}(\Sigma_w)\max_{j\in\{1,\ldots,(p_1+p_2)\}}\|\varepsilon_{n,\cdot j}/\sqrt{n}\|\big]\sqrt{\frac{\log(p_1+p_2)}{n}} \\ & + \; \frac{1}{n}\|\Delta_{\mathbf{F}_{n-1}}^\top\Delta_{\mathbf{F}_n}\|_\infty + \frac{1}{n}\|\Delta_{\mathbf{F}_{n-1}}^\top\Delta_{\mathbf{F}_{n-1}}(A_{11}^\star)^\top\|_\infty, \end{aligned} \tag{14}$$

where $\varepsilon_n := \Delta_{\mathbf{Z}_n} - \Delta_{\mathbf{Z}_{n-1}}(A^\star)^\top = [\Delta_{\mathbf{F}_n} - \Delta_{\mathbf{F}_{n-1}}(A_{11}^\star)^\top, -\Delta_{\mathbf{F}_{n-1}}(A_{21}^\star)^\top]$, and $\{W_t^+\} := \{W_{t+1}\}$ is the shifted W_t process.

Remark 1 Before moving to the high probability error bounds of the estimates, we discuss the conditions and the various quantities appearing in Lemmas 6 and 7 that determine the error bound of the estimated transition matrix and underlie the differences between

the original VAR estimation problem based on primal observed quantities (the "vanilla VAR problem" henceforth), and the present one in which one block of the variables enters the VAR system with errors. Note that the statements in the two lemmas are under the assumption that the error in the F_t block is pre-determined and non-random.

As previously mentioned, due to the presence of the error of the latent factor block, the corresponding regularity conditions need to be imposed and verified on quantities with the error incorporated, namely, $\widehat{\mathbf{Z}}$, instead of the original true random realizations $\mathbf{Z}$. Lemma 6 shows that with high probability, the random design matrix although exhibits error-in-variables, will still satisfy the RSC condition with some positive curvature as long as the spectrum of the process Z_t has sufficient regularity relative to the magnitude of the error, with the former determined by $\mathfrak{m}(f_X)/\mathcal{M}^{1/2}(f_X)$ and the latter by $\Lambda^{1/2}_{\max}(S_{\Delta_{\mathbf{F}_{n-1}}})$. In particular, the RSC curvature is pushed toward zero compared with that in the vanilla VAR problem, due to the presence of the second term in (13) that would be 0 if $\Delta_{\mathbf{F}_{n-1}} = 0$, i.e., there were no estimation errors. This curvature affects the constant scalar part of the ultimate high probability error bound obtained for the transition matrix.

Lemma 7 gives the deviation bound associated with the Hessian and the gradient (both random), which comprises of three components attributed to the random samples observed, the non-random error, and their interactions, respectively. Further, it is the relative order of these components that determines the error rate (as a function of model dimensions and the sample size). In particular, for the vanilla VAR problem, only the first term in (14) exists and yields an error rate of $\mathcal{O}(\sqrt{\log(p_1+p_2)/n})$ (see also Basu and Michailidis, 2015). For the current setting, as it is later shown in Theorem 1, since $|||\Delta_{\mathbf{F}}/\sqrt{n}|||_{\mathrm{F}} \asymp \mathcal{O}(1)$, the dominating term of the three components is the one attributed to the non-random error[5] and it ultimately determines the error rate of $\widehat{A}$, which will also be $\mathcal{O}(1)$.

3.3. High probability error bounds for the estimators

Given the results in Sections 3.1 and 3.2, we provide next high probability error bounds for the estimates, obtained by solving the optimization problems in (6) and (8) based on random snapshots from the underlying processes X_t and Y_t.

Theorem 1 combines the results in Proposition 1 and Lemmas 2 to 4 and provides the high probability error bound of the estimates, when $\widehat{\Theta}$ and $\widehat{\Gamma}$ are estimated based on random realizations from the observable processes X_t and Y_t, with the latter driven by both X_t and the latent F_t.

Theorem 1 (High probability error bounds for $\widehat{\Theta}$ and $\widehat{\Gamma}$) *Suppose we are given some randomly observed snapshots $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_n\}$ obtained from the stable Gaussian processes X_t and Y_t, whose dynamics are described in* (5) *and* (2)*. Suppose the following conditions hold for some $(C_{X,l}, C_{X,u})$ and $(C_{e,l}, C_{e,u})$:*

C1. $C_{X,l} \leq \mathfrak{m}(f_X) \leq \mathcal{M}(f_X) \leq C_{X,u}$;

C2. $C_{e,l} \leq \Lambda_{\min}(\Sigma_e) \leq \Lambda_{\max}(\Sigma_e) \leq C_{e,u}$.

5. with the implicit assumption that $\log(p_1+p_2)/n = o(1)$ which is satisfied for this study.

Then, there exist universal constants $\{C_i\}$ *and* $\{c_i\}$ *such that for sample size* $n \gtrsim q$*, by solving* (6) *with regularization parameter*

$$\lambda_\Gamma = \max\Big\{C_1(2\pi\mathcal{M}(f_X) + \Lambda_{\max}(\Sigma_e))\sqrt{\frac{\log(p_2 q)}{n}},\ C_2\phi/\sqrt{nq},\ C_3\Lambda^{1/2}_{\max}(\Sigma_e)\Big\}, \tag{15}$$

the solution $(\widehat{\Theta}, \widehat{\Gamma})$ *has the following bound with probability at least* $1 - c_1\exp(-c_2\log(p_2 q))$*:*

$$|||\Delta_\Theta/\sqrt{n}|||_F^2 + |||\Delta_\Gamma|||_F^2 \lesssim \frac{\lambda_\Gamma^2}{\mathfrak{m}(f_X)}\psi(s_{\Gamma^\star}, p_1, r) =: K_1, \tag{16}$$

for some function $\psi(\cdot)$ *that depends linearly on* $s_{\Gamma^\star}, p_1$ *and* r*.*

Note that the above bound also holds if we replace Δ_Θ by $\Delta_{\mathbf{F}}$ under (IR). Next, using the results in Proposition 2, Lemmas 6 and 7 and combine the bound in Theorem 1, we establish a high probability error bound for the estimated $\widehat{A}$ in Theorem 2.

Theorem 2 (High probability error bound for $\widehat{A}$) *Under the settings and with the procedures in Theorem 1, we additionally assume the following condition holds for the spectrum of the joint process* Z_t*:*

C3. $\mathfrak{m}(f_Z)/\mathcal{M}^{1/2}(f_Z) > C_Z$ *for some constant* C_Z*.*

Then there exists universal constants $\{c_i\}$*,* $\{c_i'\}$ *and* $\{C_i\}$ *such that for sample size* $n \gtrsim q$*, such that the estimator* $\widehat{A}$ *obtained by solving for* (8) *with* λ_A *satisfying*

$$\begin{aligned}\lambda_A = C_1\big(\mathcal{M}(f_Z) + \frac{\Sigma_w}{2\pi} + \mathcal{M}(f_{Z,W^+})\big)\sqrt{\frac{\log(p_1+p_2)}{n}} \\ + C_2\mathcal{M}^{1/2}(f_Z)\sqrt{\frac{\log(p_1+p_2)+\log p_1}{n}} + C_3\Lambda^{1/2}_{\max}(\Sigma_w)\sqrt{\frac{\log(p_1+p_2)}{n}} + C_4,\end{aligned}$$

with probability at least

$$\Big(1 - c_1\exp\{-c_2\log(p_2 q)\}\Big)\Big(1 - c_1'\exp\{-c_2'\log(p_1+p_2)\}\Big), \tag{17}$$

the following bound holds for Δ_A*:*

$$|||\Delta_A|||_F^2 \le \check{C}(K_1, \mathfrak{m}(f_Z), \mathcal{M}(f_Z))\cdot\check{\psi}(s_{A^\star}),$$

for some function $\check{C}(K_1, \mathfrak{m}(f_Z), \mathcal{M}(f_Z))$ *that does not depend on* n, p_2, q *and* $\check{\psi}(\cdot)$ *that depends linearly on* $s_{A^\star}$*. Here* K_1 *denotes the upper bound of the first stage error shown in* (16)*.*

Remark 2 (Rate of convergence) It is worth pointing out similarities in the formulation of the calibration equation and a matrix completion problem. Note that the factor hyperplane corresponds to the low-rank component one seeks to recover in the latter problem in a noisy setting. Hence, the resulting similarity in the rate obtained in our setting to that established for the matrix completion problem (Candes and Plan, 2010), is a consequence of absence of the restricted isometry property (RIP) (see also Gunasekar et al., 2015).

Remark 3 (Sample size requirement) To establish the finite-sample high probability error bound for the estimated transition matrices $\widehat{A}$, the proposed estimation procedure requires the sample size to satisfy $n \gtrsim q$; this condition is more stringent compared to the standard VAR estimation problem under sparsity, given by $n \gtrsim \sqrt{\log(p_1+p_2)}$. However, this is due to the fact that in the FAVAR formulation the F_t block is latent and needs to be estimated from the data and hence comes with "measurement error". The more restrictive sample size requirement reflects the latter fact and is embedded in the factor recovery step in the calibration equation – specifically, the concentration of $\Lambda_{\max}(S_{\mathbf{E}})$ that is necessary for providing adequate control over $\mathbf{\Delta_F}$.

Remark 4 (Generalization to VAR(d)) As a straightforward generalization, for a VAR$(d), d > 1$ system $Z_t = (F_t^\top, X_t^\top)^\top$, a similar error bound holds by considering the augmented process $\widetilde{Z}_t^\top := (Z_t, Z_{t-1}, \ldots, Z_{t-d+1})$ that satisfies

$$\widetilde{Z}_t = \widetilde{A}\widetilde{Z}_{t-1} + \widetilde{W}_t, \qquad \text{where} \qquad \widetilde{A} := \begin{bmatrix} A^{(1)} & A^{(2)} & \ldots & A^{(d)} \\ \mathrm{I}_p & O & O & O \\ \vdots & \ddots & \vdots & \vdots \\ O & O & \mathrm{I}_p & O \end{bmatrix}, \quad \widetilde{W}_t = \begin{bmatrix} W_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

In particular, with probability at least

$$\big(1 - c_1 \exp\{-c_2 \log(p_2 q)\}\big)\big(1 - c_1' \exp\{-c_2' \log(d(p_1+p_2))\}\big),$$

the following bound holds for the estimate of $\widetilde{A}$:

$$\|\Delta_{\widehat{\widetilde{A}}}\|_{\mathrm{F}}^2 \le \widetilde{C}(K_1, \mathfrak{m}(f_{\widetilde{Z}}), \mathcal{M}(f_{\widetilde{Z}})) \cdot \widetilde{\kappa}(s_{\widetilde{A}^\star}).$$

However, note that although the error bound is still of the same form, the stronger temporal dependence yields a larger $\widetilde{C}(K_1, \mathfrak{m}(f_{\widetilde{Z}}), \mathcal{M}(f_{\widetilde{Z}})$ through the RSC curvature parameter; specifically, a smaller value of $\mathfrak{m}(f_{\widetilde{Z}})$. Its impact on the deviation bound will not manifest itself in terms of the order of the error, since it only affects the constants in front of lower order terms in the expression of choosing λ_A.

4. Implementation and Performance Evaluation

We first discuss implementation issues of the proposed problem formulation for the high-dimensional FAVAR model. Specifically, the formulation requires imposing the compactness constraint for identifiability purposes and for obtaining the necessary statistical guarantees for the estimates of the model parameters. However, the value ϕ in the compactness constraint is hard to calibrate in any real data set. Hence, in the implementation we relax this constraint and assess the performance of the algorithm. Due to its importance in constraining the size of the equivalence class $\mathcal{C}(Q_2)$, we examine in Appendix D certain relatively extreme settings where the proposed relaxation fails to provide accurate estimates of the model parameters.

Implementation. The following relaxation of (6) is used in practice:

$$\begin{aligned} \min_{\Theta,\Gamma} f(\Theta,\Gamma) := \big\{\frac{1}{2n}\|\mathbf{Y} - \Theta - \mathbf{X}\Gamma^\top\|_{\mathrm{F}}^2 + \lambda_\Gamma \|\Gamma\|_1\big\} \\ \text{subject to} \quad \mathrm{rank}(\Theta) \le r, \end{aligned} \tag{18}$$

Algorithm 1: Computational procedure for estimating A, Γ and Λ.

Input: Time series data $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$, (λ_Γ, r), and λ_A.

Stage I: recover the latent factors by solving (18), through iterating between (1.1) and (1.2) until $|f(\Theta^{(m)}, \Gamma^{(m)}) - f(\Theta^{(m-1)}, \Gamma^{(m-1)})| <$ tolerance:

– (1.1) Update $\widehat{\Theta}^{(m)}$ by singular value thresholding (SVT): do SVD on the residual hyperplane, i.e., $\mathbf{Y} - \mathbf{X}(\widehat{\Gamma}^{(m-1)})^\top = UDV$, where $D := \operatorname{diag}\big(d_1, \ldots, d_{\min(n,q)}\big)$, and construct $\widehat{\Theta}^{(m)}$ by

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

– (1.2) Update $\widehat{\Gamma}^{(m)}$ with the plug-in $\widehat{\Theta}^{(m)}$ so that each row j is obtained with Lasso regression (in parallel) and solves

$$\min_\beta \big\{ \frac{1}{2n} \|(\mathbf{Y} - \widehat{\Theta}^{(m)})_{\cdot j} - \mathbf{X}\beta\|^2 + \lambda_A \|\beta_1\|_1 \big\}$$

Stage I output: $\widehat{\Theta}$ and $\widehat{\Gamma}$; the estimated factor $\widehat{\mathbf{F}}$ and $\widehat{\Lambda}$ via (7) under (IR);

Stage II: estimate the transition matrix by solving (8): update each row of A (in parallel) by solving the Lasso problem:

$$\min_\beta \big\{ \frac{1}{2n} \|(\widehat{\mathbf{Z}}_n)_{\cdot j} - \widehat{\mathbf{Z}}_{n-1}\beta\|^2 + \lambda_A \|\beta\|_1 \big\}.$$

Stage II output: $\widehat{A}$.

Output: Estimates $\widehat{\Gamma}$, $\widehat{\Lambda}$, $\widehat{A}$ and the latent factor $\widehat{\mathbf{F}}$.

which leads to Algorithm 1. The implementation of Stage I requires the pair of tuning parameters (λ_Γ, r) as input, and the choice of r is particularly critical since it determines the effective size of the latent block. In our implementation, we select the optimal pair based on the Panel Information Criterion (PIC) proposed in Ando and Bai (2018), which searches for (λ_Γ, r) over a lattice that minimizes

$$\mathrm{PIC}(\lambda_\Gamma, r) := \frac{1}{nq} \Big\| \mathbf{Y} - \widehat{\Theta} - \mathbf{X}\widehat{\Gamma}^\top \Big\|_{\mathrm{F}}^2 + \widehat{\sigma}^2 \Big[\frac{\log n}{n} \|\widehat{\Gamma}\|_0 + r\big(\frac{n+q}{nq}\big) \log(nq) \Big],$$

where $\widehat{\sigma}^2 = \frac{1}{nq} \|\mathbf{Y} - \widehat{\Theta} - \mathbf{X}\widehat{\Gamma}^\top\|_{\mathrm{F}}^2$. Analogously, the implementation of Stage II requires λ_A as input, and we select λ_A over a grid of values that minimizes the Bayesian Information Criterion (BIC):

$$\mathrm{BIC}(\lambda_A) = \sum_{i=1}^{q} \log \mathrm{RSS}_i + \frac{\log n}{n} \|\widehat{A}\|_0,$$

where $\mathrm{RSS}_i := \|(\mathbf{X}_n)_{\cdot i} - \mathbf{X}_{n-1}\widehat{A}_{i\cdot}^\top\|^2$ is the residual sum of square of the i^{th} regression. Extensive numerical work shows that these two criteria select very satisfactory values for the tuning parameters, which in turn yield highly accurate estimates of the model parameters.

Simulation setup. Throughout, we assume Σ_w^X, Σ_X^F and Σ_e are all diagonal matrices, and the sample size is fixed at 200, unless otherwise specified. We first generate samples of $F_t \in \mathbb{R}^{p_1}$ and $X_t \in \mathbb{R}^{p_2}$ recursively according to the VAR(d) model in (1), and then the samples of $Y_t \in \mathbb{R}^q$ are generated according to the linear model given in (2). Specifically, (IR) is imposed on the true value of the parameter, hence $\Lambda^\star$ that is used for generating

Y_t always satisfies the restriction $\Lambda = \left[\begin{smallmatrix} \mathrm{I}_{p_1} \\ * \end{smallmatrix}\right]$. Unless otherwise specified, all error terms are generated according to some mean-zero Gaussian distribution.

For the calibration equation, the density level of the sparse coefficient matrix $\Gamma \in \mathbb{R}^{q \times p_2}$ is fixed at $5/p_2$ for each regression; thus, each Y_t coordinate is affected by 5 series (coordinates) from the X_t block on average. The bottom $(q - p_1) \times p_1$ block of the loading matrix $\Lambda \in \mathbb{R}^{q \times p_1}$ is dense. The magnitude of nonzero entries of Γ and that of entries of Λ may vary to capture different levels of signal contributions to Y_t, and we adjust the standard deviation of e_t to maintain the desired level of the signal-to-noise ratio for Y_t (averaged across all coordinates).

For the transition matrix A of the VAR equation, the density for each of its component block $\{A_{ij}\}_{i,j=1,2}$ varies across settings, so as to capture different levels of the influence from the lagged value of the latent block F_t on the observed X_t. Note that to ensure stability of the VAR system, the spectral radius of A, $\varrho(A)$, needs to be smaller than 1. In particular, when a VAR(d) ($d > 1$) system is considered, we need to ensure that the spectral radius of $\widetilde{A}$ is smaller than 1[6], where we let $p = p_1 + p_2$ and

$$\widetilde{A} := \begin{bmatrix} A^{(1)} & A^{(2)} & \cdots & A^{(d)} \\ \mathrm{I}_p & O & O & O \\ \vdots & \ddots & \vdots & \vdots \\ O & O & \mathrm{I}_p & O \end{bmatrix}.$$

Table 1 lists the simulation settings and their parameter setup.

Table 1: Parameter setup for different simulation settings for the VAR equation.

	q	p_1	p_2		$s_{A_{11}}$	$s_{A_{12}}$	$s_{A_{21}}$	$s_{A_{22}}$	SNR(Y_t)
A1	100	5	50			$3/(p_1+p_2)$			1.5
A2	200	10	100			$3/(p_1+p_2)$			1.5
A3	200	5	100		$3/p_1$	$2/p_2$	$2/p_1$	$2/p_2$	1.5
A4	300	5	500		$3/p_1$	$2/p_2$	0.8	$2/p_2$	1.5
B1 ($d=2$)	200	5	100	$A^{(1)}$:		$3/(p_1+p_2)$			2
				$A^{(2)}$:		$2/(p_1+p_2)$			
B2 ($d=4$)	200	5	100	$A^{(1)}$:	0.5	$3/p_2$	0.5	$3/p_2$	2
				$A^{(2)}$:	0.2	$2/p_2$	0.25	$2/p_2$	
				$A^{(3)}$:		$2/(p_1+p_2)$			
				$A^{(4)}$:		$2/(p_1+p_2)$			
B3 ($d=4$)	100	5	25	$A^{(1)}$:	0.5	$2/p_2$	0.5	$2/p_2$	2
				$A^{(2)}$:	0.2	$1.5/p_2$	0.1	$1.5/p_2$	
				$A^{(3)}$:		$1/(p_1+p_2)$			
				$A^{(4)}$:		$0.8/(p_1+p_2)$			
C1	same as setting A1 with t_4 noise for the VAR system								
C2	same as setting B1 with t_8 noise for the VAR system								
C3	same as setting B2 with sub-exponential noise for the VAR system								
C4	same as setting B2 with sub-exponential noise for the VAR system and 500 observations								

Specifically, in settings A1 – A4, $(F_t^\top, X_t^\top)^\top$ jointly follows a VAR(1) model. The (average) signal-to-noise ratio for each regression of Y_t is 1.5. For settings A1 and A2, the

6. In practice, this can be achieved by first generating $A^{(1)}, \ldots, A^{(d)}$, align them in $\widetilde{A}_{\text{initial}}$ and obtain the scale factor $\zeta := \varrho_{\text{target}}/\varrho(\widetilde{A}_{\text{initial}})$, then scale $A^{(i)}$ by ζ^i. The validity of this procedure follows from simple algebraic manipulations.

transition matrix A is uniformly sparse, with A2 corresponding to a larger system; for settings A3 and A4, we increase the density level (the proportion of nonzero entries) for the transition matrices that govern the effect of F_{t-1} on F_t and X_t. In particular, for setting A4, we consider a large system with 500 coordinates in X_t, and the factor effect is almost pervasive on these coordinates (through the lags), as the density level of A_{21} is set at 0.8. Settings B1, B2 and B3 consider settings with more lags ($d = 2$ and $d = 4$, respectively), and to compensate for the higher level of correlation between F_t and X_t, we elevate the signal-to-noise for each regression of Y_t to 2. For B1, the transition matrices for both lags ($A^{(1)}$ and $A^{(2)}$) have uniform sparsity patterns, with $A^{(2)}$ being slightly more sparse compared to $A^{(1)}$; for B2, the transition matrices for the first two lags have higher density in the component that governs the $F_{t-i} \to X_t$ cross effect, and those for the last two lags have uniform sparsity. B3 has approximately the same scale as observed in real data, and due to a small p_2, the system exhibits a higher sparsity level in general. In settings C1 – C4, the error terms of the VAR system are generated from distributions with tails heavier than a Gaussian (e.g. t-distributions, squares of Gaussian which have sub-exponential tails), and the joint process $(F_t', X_t)'$ will be heavy-tailed as a result of the recursive data generating mechanism.

Performance evaluation. We consider both the estimation and the forecasting performance of the proposed estimation procedure. The performance metrics used for estimation are sensitivity (SEN), specificity (SPC) and the relative error in Frobenius norm (Err) for the sparse components (transition matrices A and the coefficient matrix Γ), defined as

$$\text{SEN} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{SPC} = \frac{\text{TN}}{\text{FP} + \text{TN}}, \quad \text{Err} = |||\Delta_M|||_{\text{F}} / |||M^\star|||_{\text{F}}.$$

We also track the estimated size of the latent component (i.e., the rank constraint in (6), jointly with λ_Γ is selected by PIC), as well as the relative errors of $\widehat{\Theta}$, $\widehat{\mathbf{F}}$ and $\widehat{\Lambda}$. For forecasting, we focus on evaluating the h-step-ahead predictions for the X_t block. Specifically, for settings where the VAR system is 1-lag dependent (A1–A4, C1), we consider $h = 1$; for settings where the VAR system has more lag dependencies (B1–B3, C2–C4), we consider $h = 1, 2$. We use the same benchmark model as in Bańbura et al. (2010) which is based on a special case of the Minnesota prior distribution (Litterman, 1986), so that the for any generic time series $X_t \in \mathbb{R}^p$, each of its coordinates $j = 1, \ldots, p$ follows a centered random walk:

$$X_{t,j} = X_{t-1,j} + u_{t,j}, \qquad u_{t,j} \sim \mathcal{N}(0, \sigma_u^2). \tag{19}$$

For each forecast $\widehat{x}_{T+h}$, its performance is evaluated based on the following two measures:

$$\text{rel-err} = \|\widehat{x}_{T+h} - x_{T+h}\|_2^2 / \|x_{T+h}\|_2^2, \qquad \text{rel-err-ratio} = \frac{\frac{1}{p_2}\sum_{j=1}^{p_2} \left|\frac{\widehat{x}_{T+h,j} - x_{T+h,j}}{x_{T+h,j}}\right|}{\frac{1}{p_2}\sum_{j=1}^{p_2} \left|\frac{\widetilde{x}_{T+h,j} - x_{T+h,j}}{x_{T+h,j}}\right|},$$

where rel-err measures the ℓ_2 norm of the relative error of the forecast to the true value; whereas for rel-err-ratio, it measures the ratio between the relative error of the forecast and the above described benchmark. In particular, its numerator and denominator respectively capture the averaged relative error of all coordinates of the forecast $\widehat{x}_{T+h}$ and that of the benchmark $\widetilde{x}_{T+h}$ that evolves according to (19), while the ratio measures how much the

forecast based on the proposed FAVAR model outperforms (< 1) or under-performs (> 1) compared to the benchmark.

All tabulated results are based on the average of 50 replications. Tables 2, 3 and 4, respectively, depict the performance of the estimates of the parameters in the calibration and the VAR equations, as well as the forecasting performance under the settings considered. Based on the results listed in Tables 2 and 3, we notice that in all settings, the parameters in the calibration equation $\widehat{\Theta}$ and $\widehat{\Gamma}$ are well estimated, while the rank slightly underestimated. Further, the SEN and SPC measures of $\widehat{\Gamma}$ show excellent performance regarding support recovery. It is worth pointing out that the estimation accuracy of the parameters in the calibration equation strongly depends on the signal-to-noise ratio of Y_t. In particular, if the signal-to-noise ratio in A1-A4 is increased to 1.8, the rank is always correctly selected by PIC, and the estimation relative error of $\widehat{\Theta}$ further decreases(results omitted for space considerations)[7]. Under the given IR, we decompose the estimated factor hyperplane into the factor block and its loadings. The results show that both quantities exhibit a higher relative error compared to that of the factor hyperplane. Of note, the loadings estimates exhibit a lot of variability as indicated by the high standard deviation in the Table.

Regarding the estimates in the VAR equation, for settings A1, A2 and B1 that are characterized by an adequate degree of sparsity, the recovery of the skeleton of the transition matrices is very good. However, performance deteriorates if the latent factor becomes "more pervasive" (settings A3 and A4), which translates to the A_{21} block having lower sparsity. On the other hand, this does not have much impact on the recovery of the A_{22} sub-block, as for these two settings, SEN and SPC of A_{22} still remain at a high level. For settings with more lags, performance deteriorates (as expected) although SEN and SPC remain fairly satisfactory. On the other hand, the relative error of the transition matrices increases markedly. Nevertheless, the estimates of the first lag transition matrix is better than the remaining ones. Further, the results indicate that smaller size VAR systems (B3) exhibit better performance than larger ones. Finally, in terms of forecasting (results depicted in Table 4), the one-step-ahead forecasting value yields approximately 50% to 90% rel-err (compared to the truth), depending on the specific setting and the actual SNR, while it outperforms the forecast of the benchmark by around 40% (based on the rel-err-ratio measure). Of note, the 2-step-ahead forecasting value for settings with more lags outperforms the benchmark by an even wider margin with the rel-err-ratio decreasing to less than 0.3.

Finally, the proposed methodology is robust in the presence of heavier than Gaussian tails in the VAR processes. Further, note that in setting C3 wherein the temporal dependence is strong and the error terms are generated according to a sub-exponential distribution, the performance of the estimated transition matrices deteriorates significantly, as expected from the theoretical results outlined in Appendix C. Nevertheless, with proper compensation in terms of sample size (setting C4), the performance improves markedly.

7. This also comes up when comparing the relative error of $\widehat{\Theta}$ in the A1-A4 settings to that in the B1-B2 ones, where the latter two have a higher SNR.

Table 2: Performance evaluation of estimated parameters in the calibration equation.

	PIC-selected r	Err($\widehat{\Theta}$)	Err($\widehat{\mathbf{F}}$)	Err($\widehat{\Lambda}$)	SEN($\widehat{\Gamma}$)	SPC($\widehat{\Gamma}$)	Err($\widehat{\Gamma}$)
A1	4.80(.40)	0.32(.010)	0.56(.074)	0.67(.345)	0.99(.007)	0.98(.003)	0.45(.013)
A2	9.96(.19)	0.32(.008)	0.90(.065)	2.54(1.30)	0.99(.005)	0.98(.001)	0.52(.010)
A3	4.78(.54)	0.33(.048)	0.73(.103)	2.59(1.59)	0.99(.003)	0.99(.001)	0.57(.009)
A4	4.42(.49)	0.38(.040)	0.84(.100)	2.66(2.14)	0.97(.009)	0.99(.001)	0.59(.015)
B1	5(0)	0.23(.004)	0.41(.043)	0.54(.020)	1.00(.000)	0.97(.011)	0.27(.014)
B2	5(0)	0.26(.007)	0.38(.047)	0.42(.087)	1.00(.000)	0.99(.002)	0.37(.007)
B3	5(0)	0.25(.007)	0.34(.031)	0.34(.080)	1.00(.000)	0.99(.001)	0.32(.012)
C1	4.96(.20)	0.32(.019)	0.58(.075)	0.86(.564)	0.99 (.001)	0.96(.009)	0.47(.017)
C2	5(0)	0.23(.005)	0.43(.042)	0.54(.155)	1.00 (.000)	0.96(.008)	0.27(.010)
C3	5(0)	0.21(.006)	0.39(.040)	0.41(.123)	1.00 (.000)	0.97(.003)	0.27(.052)
C4	5(0)	0.20(.007)	0.27(.028)	0.25(.041)	1.00 (.000)	0.97(.011)	0.18(.012)

Table 3: Performance evaluation of the estimated transition matrices in the VAR equation.

	coef	SEN($\widehat{A}$)	SPC($\widehat{A}$)	Err($\widehat{A}$)	SEN($\widehat{A}_{22}$)	SPC($\widehat{A}_{22}$)	Err($\widehat{A}_{22}$)
A1	A	0.99(.003)	0.95(.012)	0.35(.019)	0.99(.001)	0.96(.013)	0.31(.022)
A2	A	0.98(.008)	0.97(.004)	0.46(.018)	0.99(.001)	0.98(.003)	0.39(.017)
A3	A	0.86(.050)	0.98(.006)	0.73(.029)	0.93(.032)	0.98(.005)	0.65(.034)
A4	A	0.75(.046)	0.92(.002)	0.71(0.024)	0.99(.001)	0.92(.002)	0.60(.018)
B1	$A^{(1)}$	0.99(.003)	0.98(.002)	0.47(.017)	0.99(.002)	0.98(.002)	0.46(.017)
	$A^{(2)}$	0.97(.010)	0.98(.002)	0.55(.017)	0.98(.011)	0.98(.003)	0.55(.018)
B2	$A^{(1)}$	0.89(.017)	0.88(.003)	0.71(.014)	0.90(.017)	0.99(.003)	0.70(.014)
	$A^{(2)}$	0.75(.028)	0.88(.003)	0.89(.020)	0.77(0.032)	0.88(.003)	0.90(.021)
	$A^{(3)}$	0.84(.025)	0.88(.003)	0.85(.015)	0.85(.027)	0.88(.004)	0.84(.018)
	$A^{(4)}$	0.72(.022)	0.88(.003)	0.99(.017)	0.73(.025)	0.88(.003)	0.98(.017)
B3	$A^{(1)}$	0.93(.034)	0.96(.010)	0.61(.043)	0.94(.035)	0.97(.009)	0.60(.045)
	$A^{(2)}$	0.77(.078)	0.96(.010)	0.74(.044)	0.78(.084)	0.97(.010)	0.74(.046)
	$A^{(3)}$	0.80(.098)	0.96(.012)	0.75(.052)	0.81(.102)	0.97(.010)	0.74(.056)
	$A^{(4)}$	0.74(.122)	0.97(.011)	0.78(.059)	0.72(.134)	0.97(.009)	0.79(.065)
C1	A	0.99(.007)	0.95(.012)	0.42(.024)	0.99(.002)	0.96(.011)	0.38(.024)
C2	$A^{(1)}$	0.99(.004)	0.98(.002)	0.46(.013)	0.99(.003)	0.98(.002)	0.45(.015)
	$A^{(2)}$	0.98(.008)	0.97(.003)	0.54(.018)	0.98(.009)	0.98(.003)	0.54(.019)
C3	$A^{(1)}$	0.93(.013)	0.42(.005)	1.54(.024)	0.93(.013)	0.42(.006)	1.61(.027)
	$A^{(2)}$	0.86(.019)	0.44(.006)	2.11(.029)	0.86(.023)	0.44(.006)	2.30(.032)
	$A^{(3)}$	0.88(.023)	0.44(.006)	2.06(.028)	0.89(.023)	0.44(.005)	2.07(.028)
	$A^{(4)}$	0.82(.023)	0.44(.006)	2.51(.043)	0.83(.025)	0.44(.006)	2.51(.041)
C4	$A^{(1)}$	0.89(.016)	0.96(.002)	0.67(.013)	0.89(.016)	0.96(.002)	0.65(.014)
	$A^{(2)}$	0.73(.025)	0.96(.006)	0.78(.029)	0.74(.026)	0.96(.002)	0.79(.011)
	$A^{(3)}$	0.82(.027)	0.96(.002)	0.74(.015)	0.82(.028)	0.96(.002)	0.74(.017)
	$A^{(4)}$	0.60(.031)	0.96(.002)	0.87(.014)	0.60(.033)	0.96(.002)	0.87(.015)

Table 4: Performance evaluation of forecasting.

	$h=1$		$h=2$	
	rel-err	rel-err-ratio	rel-err	rel-err-ratio
A1	0.53(.117)	0.38(.065)	-	-
A2	0.60(.075)	0.38(.046)	-	-
A3	0.80(.075)	0.45(.064)	-	-
A4	0.56(.109)	0.40(.055)	-	-
B1	0.62(.060)	0.35(.171)	0.66(.127)	0.24(.071)
B2	0.89(.091)	0.42(.217)	0.94(.173)	0.29(.118)
B3	0.81(.094)	0.32(.129)	0.90(.402)	0.26(.174)
C1	0.59(.176)	0.50(.118)	-	-
C2	0.59(.121)	0.41(.350)	0.61(.270)	0.26(.089)
C3	1.25(.305)	0.19(.081)	1.26(.396)	0.15(.059)
C4	0.52(.073)	0.12(.071)	0.51(.168)	0.07(.034)

5. Application to Commodity Price Interlinkages

Interlinkages between commodity prices represent an active research area in economics and have been a source of concern for policymakers. Commodity prices, unlike stocks and bonds, are determined more strongly by global demand and supply considerations. Nevertheless, other factors are also at play as outlined next. The key ones are: (i) the state of the global macro-economy and the state of the business cycle that manifest themselves as direct demand for commodities; (ii) monetary policy, specifically, interest rates that impact the opportunity cost for holding inventories, as well as having an impact on investment and hence production capacity that subsequently contribute to changes in supply and demand in the market; and (iii) the relative performance of other asset classes through portfolio allocation (see Frankel, 2006, 2014, and references therein). We employ the FAVAR model and the proposed estimation method to investigate interlinkages amongst major commodity prices. The X_t block corresponds to the set of commodity prices of interest, while the Y_t block contains representative indicators for the global economic environment. We extract the factors F_t based on the calibration equation and then consider the augmented VAR system of (F_t, X_t), so that the estimated interlinkages amongst commodity prices are based on a larger information set that takes into account broader economic activities.

Data. The commodity price data (X_t) are retrieved from the International Monetary Fund, comprising of 16 commodity prices in the following categories: Metal, Energy (oil) and Agricultural. The set of economic indicators (Y_t) contain core macroeconomic variables and stock market composite indices from major economic entities including China, EU, Japan, UK and US, with a total number of 54 indicators. Specifically, the macroeconomic variables primarily account for: Output & Income (e.g. industrial production index), Labor Market (unemployment), Money & Credit (e.g. M2), Interest & Exchange Rate (e.g. Fed Funds Rate and the effective exchange rate), and Price Index (e.g. CPI). For variables that reflect interest rates, we use both the short-term interest rate such as 6-month LIBOR, and the 10-year T-bond yields from the secondary market. Further, to ensure stationarity of the time series, we take the difference of the logarithm for X_t; for Y_t, we apply the same transformation as proposed in Stock and Watson (2002). A complete list of the commodity prices and economic indicators used in this study is provided in Appendix E. For all time series considered, we use monthly data spanning the January 2001 to December 2016 period. Further, based on previous empirical findings in the literature related to the global financial crisis of 2008 (Stock and Watson, 2017), we break the analysis into the following three sub-periods (Stock and Watson, 2017): pre-crisis (2001–2006), crisis (2007–2010) and post-crisis (2011–2016), each having sample size (available time points) 72, 48, and 72, respectively[8].

We apply the same estimation procedure for each of the above three sub-periods. Starting with the calibration equation, we estimate the factor hyperplane Θ and the sparse regression coefficient matrix Γ, then extract the factors based on the estimated factor hy-

8. For each individual time series, we test for its normality using data spanning the pre-crisis, crisis, and post-crisis periods, respectively. Based on the Shapiro-Wilk test, the null hypothesis of normality is not rejected for selected time series (e.g., ALUMINUM) and rejected for others (e.g., OIL). However, when testing for multivariate normality of the joint distribution of all time series resp. across the three periods, we fail to reject the null hypothesis. The latter result may be due to inadequate power of the test given the relatively small sample size.

perplane under the (IR) condition. For each of the three sub-periods, 4, 3, and 3 factors are respectively identified based on the PIC criterion, with the key variable loadings (collapsed into categories) on each extracted factor listed in Table 5, after adjusting for ΓX_t. Based

Table 5: Composition of the factors identified for three sub-periods. $+$, $-$ and $*$ respectively stand for positive (all economic indicators in that category have a positive sign in Λ), negative and mixed (sign) contribution.

	pre-crisis				crisis			post-crisis		
	F1	F2	F3	F4	F1	F2	F3	F1	F2	F3
bond return	$-$		$+$	$+$	$-$	$+$				$-$
economic output	$+$						$+$		$+$	
equity return	$+$				$-$	$-$		$-$		$+$
interest/exchange rate			$*$					$*$		
labor		$+$			$-$		$-$			
money & credit			$+$		$+$				$+$	
price index		$+$					$+$			$-$
trade		$-$				$*$		$*$		

on the composition of the factors, we note that the factors summarize both the macroeconomic environment and also capture information from the secondary market (bond & equity return), as suggested by economic analysis of potential contributors to commodity price movements (Frankel, 2006, 2014). Hence, the obtained factors summarize the necessary information to include in the VAR system that examines commodity price interlinkages over time. Further, across all three periods considered, Economic Output and Money & Credit indicators contribute positively to the factor composition. In particular, the positive contribution from the M2 measure of money supply for the US during the crisis period and that from the Fed Funds Rate post crisis are pronounced; hence, the estimated factors strongly reflect the effect of the Quantitative Easing policy adopted by the US central bank. The contribution of the other categories are mixed, with that from bond returns being noteworthy due to their role as a proxy for long-term interest rates, which impact both the cost of investment in increasing production capacity and on holding inventories, as well as on the composition of asset portfolios across a range of investment possibilities (stocks, bonds, commodities, etc.).

Next, using these estimated factors, we fit a sparse VAR(2) model to the augmented $(\widehat{F}_t^\top, X_t^\top)^\top$ system. The estimated transition matrices are depicted in Figures 1 to 3 as networks[9]. It is apparent that the factors play an important role, both as emitters and receivers. The effects from the first lag are generally stronger to that from the second one. In particular, focusing on the first lag, the dominant nodes in the system have shifted over time from (OIL, SOYBEANS, ZINC) pre crisis to (SUGAR, WHEAT, COPPER) during the crisis, then to (OIL, SOYBEANS, RICE) post crisis. Based on node weighted degree, the role of OIL is dominant in both pre- and post-crisis periods, but is much weaker during the crisis.

9. In all three figures, the left panel corresponds to $\widehat{A}^{(1)}$ and the right panel corresponds to $\widehat{A}^{(2)}$. Node sizes are proportional to node weighted degrees. Positive edges are in red and negative edges are in blue. Edges with higher saturation have larger magnitudes.

Figure 1: Estimated transition matrices for Pre-crisis period.

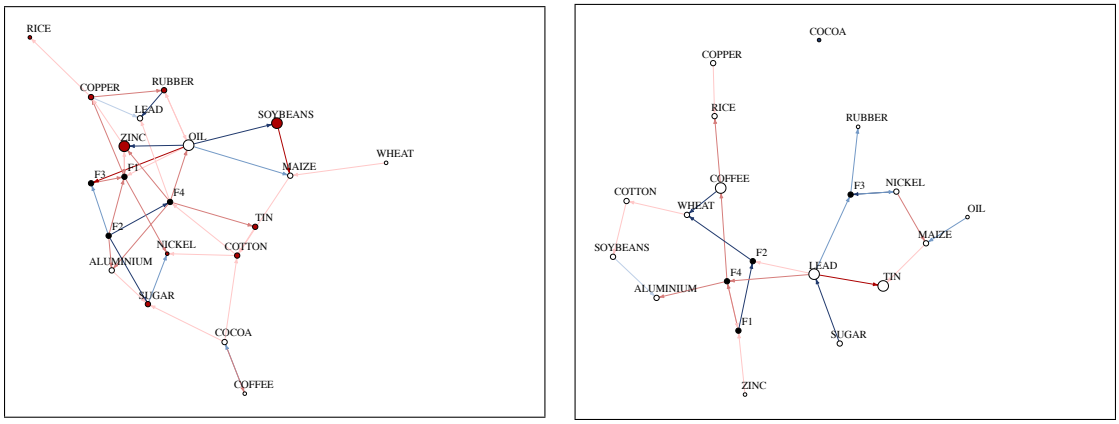


Figure 2: Estimated transition matrices for the Crisis period.

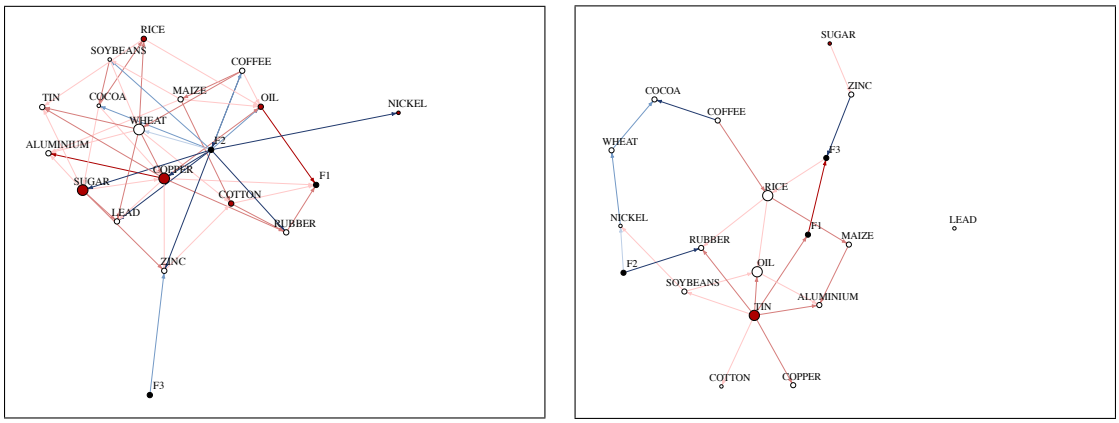
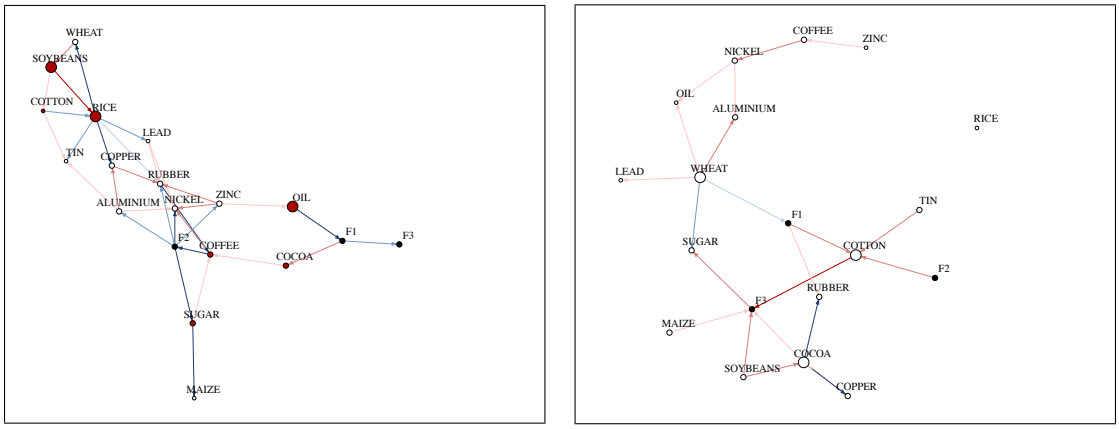


Figure 3: Estimated transition matrices for Post-crisis period



Another key feature of the interlinkage networks is their increased connectivity during the crisis period, vis-a-vis the pre- and post-crisis periods. The same empirical finding has been noted for stock returns (see Lin and Michailidis, 2017, and references therein). Before the global financial crisis of 2008, commodity prices were fast rising primarily due

to increased demand from China. Specifically, as Chinese industrial production quadrupled between 2001 and 2011, its consumption of industrial metals (Copper, Zinc, Aluminum, Lead) increased by 330%, while its oil consumption by 98%. This strong demand shock led to a sharp rise in these commodity prices, particularly accentuated beginning in 2006 (the onset of the crisis period considered in our analysis), briefly disrupted with a quick plunge of commodity prices in 2008 and their subsequent recovery in the ensuing period until late 2010, when demand from China subsided, which coupled with weak demand from the EU, Japan and the US in the aftermath of the crisis created an oversupply that put downward pressure on prices. These events induce strong inter-temporal and cross-temporal correlations amongst commodity prices, and hence are reflected in their estimated interlinkage network.

6. Discussion

This paper considered the estimation of FAVAR model under the high-dimensional scaling. It introduced an identifiability constraint (IR+Compactness) that is suitable for high-dimensional settings, and when such a constraint is incorporated in the optimization problem based upon the calibration equation, the global optimizer corresponds to model parameter estimates with bounded statistical errors. This development also allows for accurate estimation of the transition matrices of the VAR system, despite the plug-in factor block contains error due to the fact that it is an estimated quantity. Extensive numerical work illustrates the overall good performance of the proposed empirical implementation procedure, but also illustrates that the imposed constraint is not particularly stringent, especially in settings where the coefficient matrix Γ of the observed predictor variables in the calibration equation exhibits sufficient level of sparsity.

The key advantage of the FAVAR model is that it can leverage information from a large number of variables, while modeling the cross-temporal dependencies of a smaller number of them that are of primary interest to the analyst.

Recall that the nature of the FAVAR model results in estimating the transition matrix of a VAR system with one block of the observations (factors) being an estimated quantity, rather than conducting the estimation based on observed samples. Similar in flavor problems have been examined in the high-dimensional iid setting (e.g. Loh and Wainwright, 2012), as well as low dimensional time series settings; for example, Chanda et al. (1996) examine parameter estimation of a univariate autoregressive process with error-in-variables and in more recent work Komunjer and Ng (2014) investigate parameter identification of VAR-X and dynamic panel VAR models subject to measurement errors.

Acknowledgments

The authors would like to thank two anonymous referees for constructive comments and suggestions. The work of GM was supported in part by NSF grants IIS 1632730, DMS 1821220 and DMS 1830175.

Appendix A. Proofs for Theorems and Propositions

This section is divided into two parts. In the first part, we provide proofs for the proposition and theorem related to Stage I estimates, i.e., $\widehat{\Theta}$ and $\widehat{\Gamma}$. In the second part, we give proofs for the statements related to Stage II estimates, namely $\widehat{A}$, with an emphasis on how to obtain the final high probability error bound through properly conditioning on related events.

Part 1. Proofs for the $\widehat{\Theta}$ and $\widehat{\Gamma}$ estimates.

Proof of Proposition 1. Using the optimality of $(\widehat{\Gamma}, \widehat{\Theta})$ and the feasibility of $(\Gamma^\star, \Theta^\star)$, the following *basic inequality* holds:

$$\frac{1}{2n}|||\mathbf{X}\Delta_\Gamma^\top + \Delta_\Theta|||_{\mathrm{F}}^2 \leq \frac{1}{n}\Big(\langle\!\langle \Delta_\Gamma^\top, \mathbf{X}^\top\mathbf{E}\rangle\!\rangle + \langle\!\langle \Delta_\Theta, \mathbf{E}\rangle\!\rangle\Big) + \lambda_\Gamma\Big(\|\Gamma^\star\|_1 - \|\widehat{\Gamma}\|_1\Big), \tag{20}$$

which after rearranging terms gives

$$\begin{aligned}\frac{1}{2n}|||\mathbf{X}\Delta_\Gamma^\top|||_{\mathrm{F}}^2 + \frac{1}{2}|||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}^2 \leq \frac{1}{n}\langle\!\langle \mathbf{X}\Delta_\Gamma^\top, \widehat{\Theta} - \Theta^\star\rangle\!\rangle \\ + \frac{1}{n}\Big(\langle\!\langle \Delta_\Gamma^\top, \mathbf{X}^\top\mathbf{E}\rangle\!\rangle + \langle\!\langle \Delta_\Theta, \mathbf{E}\rangle\!\rangle\Big) + \lambda_\Gamma\Big(\|\Gamma^\star\|_1 - \|\widehat{\Gamma}\|_1\Big).\end{aligned} \tag{21}$$

The remainder of the proof proceeds in three steps: in Step (i), we obtain a lower bound for the left-hand-side (LHS) leveraging the RSC condition; in Step (ii), an upper bound for the right-hand-side (RHS) based on the designated choice of λ_Γ is derived; in Step (iii), the two sides are aligned to yield the desired error bound after rearranging terms.

To complete the proof, we first define a few quantities that are associated with the support set of Γ and its complement:

$$\begin{aligned}\mathbb{S} &:= \big\{\Delta \in \mathbb{R}^{q\times p_2} | \Delta_{ij} = 0 \text{ for } (i,j) \notin S_{\Gamma^\star}\big\}, \\ \mathbb{S}^c &:= \big\{\Delta \in \mathbb{R}^{q\times p_2} | \Delta_{ij} = 0 \text{ for } (i,j) \in S_{\Gamma^\star}\big\},\end{aligned}$$

where $S_{\Gamma^\star}$ is the support of $\Gamma^\star$. Further, define $\Delta_{\mathbb{S}}$ and $\Delta_{\mathbb{S}^c}$ as

$$\Delta_{\mathbb{S},ij} = 1\{(i,j) \in S_{\Gamma^\star}\}\Delta_{ij}, \qquad \Delta_{\mathbb{S}^c,ij} = 1\{(i,j) \in S_{\Gamma^\star}^c\}\Delta_{ij},$$

and note that they satisfy

$$\Delta = \Delta_{\mathbb{S}} + \Delta_{\mathbb{S}^c}, \qquad \|\Delta\|_1 = \|\Delta_{\mathbb{S}}\|_1 + \|\Delta_{\mathbb{S}^c}\|_1,$$

and

$$\|\Delta_{\mathbb{S}}\|_1 \leq \sqrt{s}|||\Delta_{\mathbb{S}}|||_{\mathrm{F}} \leq \sqrt{s_{\Gamma^\star}}|||\Delta|||_{\mathrm{F}}. \tag{22}$$

Step (i). Since $\mathbf{X}$ satisfies the RSC condition, the first term on the LHS of (21) is lower bounded by

$$\frac{\alpha_{\mathrm{RSC}}^{\mathbf{X}}}{2}|||\Delta_\Gamma|||_{\mathrm{F}}^2 - \tau_{\mathbf{X}}\|\Delta_\Gamma\|_1^2. \tag{23}$$

To get a lower bound for (23), consider an upper bound for $\|\Delta_\Gamma\|_1$ with the aid of (20). Specifically, for the first two terms in the RHS of (20), by Hölder's inequality, the following inequalities hold for the inner products:

$$\begin{aligned} \langle\!\langle \Delta_\Gamma^\top, \mathbf{X}^\top\mathbf{E} \rangle\!\rangle &\le \|\Delta_\Gamma\|_1 \|\mathbf{X}^\top\mathbf{E}\|_\infty \\ \langle\!\langle \Delta_\Theta, \mathbf{E} \rangle\!\rangle &\le |||\Delta_\Theta|||_* |||\mathbf{E}|||_{op} = n|||\Delta_\Theta|||_* \Lambda_{\max}^{1/2}(S_\mathbf{E}); \end{aligned} \tag{24}$$

for the last term, since

$$\begin{aligned} \|\widehat{\Gamma}\|_1 = \|\Gamma^\star_\mathbb{S} + \Gamma^\star_{\mathbb{S}^c} + \Delta_{\Gamma|\mathbb{S}} + \Delta_{\Gamma|\mathbb{S}^c}\|_1 &= \|\Gamma^\star_\mathbb{S} + \Delta_{\Gamma|\mathbb{S}}\|_1 + \|\Delta_{\mathbb{S}|\mathbb{S}^c}\|_1 \\ &\ge \|\Gamma^\star_\mathbb{S}\|_1 - \|\Delta_{\Gamma|\mathbb{S}}\|_1 + \|\Delta_{\Gamma|\mathbb{S}^c}\|_1, \end{aligned}$$

the following inequality holds:

$$\|\Gamma^\star\|_1 - \|\widehat{\Gamma}\|_1 \le \|\Delta_{\Gamma|\mathbb{S}}\|_1 - \|\Delta_{\Gamma|\mathbb{S}^c}\|_1. \tag{25}$$

Using the non-negativity of the RHS in (20), by choosing

$$\lambda_\Gamma \ge \max\Big\{2\|\mathbf{X}^\top\mathbf{E}/n\|_\infty,\ \Lambda_{\max}^{1/2}(S_\mathbf{E})\Big\}, \tag{26}$$

the following inequality holds:

$$\begin{aligned} 0 \le \frac{\lambda_\Gamma}{2}\|\Delta_\Gamma\|_1 + \lambda_\Gamma|||\Delta_\Theta/\sqrt{n}|||_* + \lambda_\Gamma(\|\Delta_{\Gamma|\mathbb{S}}\|_1 - \|\Delta_{\Gamma|\mathbb{S}^c}\|_1) \\ = \frac{3\lambda_\Gamma}{2}\|\Delta_{\Gamma|\mathbb{S}}\|_1 - \frac{\lambda_\Gamma}{2}\|\Delta_{\Gamma|\mathbb{S}^c}\|_1 + \lambda_\Gamma|||\Delta_\Theta/\sqrt{n}|||_*. \end{aligned}$$

Since $\Delta_\Theta = \widehat{\Theta} - \Theta^\star$ has rank at most $p_1 + r$, $|||\Delta_\Theta/\sqrt{n}|||_* \le \sqrt{p_1+r}|||\Delta_\Theta/\sqrt{n}|||_\mathrm{F}$. It follows that

$$\begin{aligned} \frac{\lambda_\Gamma}{2}\|\Delta_{\Gamma|\mathbb{S}^c}\|_1 &\le \lambda_\Gamma\sqrt{p_1+r}|||\Delta_\Theta/\sqrt{n}|||_\mathrm{F} + \frac{3\lambda_\Gamma}{2}\|\Delta_{\Gamma|\mathbb{S}}\|_1, \\ \frac{\lambda_\Gamma}{2}\|\Delta_{\Gamma|\mathbb{S}}\|_1 + \frac{\lambda_\Gamma}{2}\|\Delta_{\Gamma|\mathbb{S}^c}\|_1 &\le \lambda_\Gamma\sqrt{p_1+r}|||\Delta_\Theta/\sqrt{n}|||_\mathrm{F} + \frac{3\lambda_\Gamma}{2}\|\Delta_{\Gamma|\mathbb{S}}\|_1 + \frac{\lambda_\Gamma}{2}\|\Delta_{\Gamma|\mathbb{S}}\|_1, \\ \|\Delta_\Gamma\|_1 &\le \sqrt{4(p_1+r)}|||\Delta_\Theta/\sqrt{n}|||_\mathrm{F} + 4\|\Delta_{\Gamma|\mathbb{S}}\|_1 \\ &\le \sqrt{4(p_1+r)}|||\Delta_\Theta/\sqrt{n}|||_\mathrm{F} + 4\sqrt{s_{\Gamma^\star}}|||\Delta_\Gamma|||_\mathrm{F}, \end{aligned}$$

where the second line is obtained by adding $\frac{\lambda_\Gamma}{2}\|\Delta_{\Gamma|\mathbb{S}}\|_1$ on both sides, and the last inequality uses (22). Further, by the Cauchy-Schwartz inequality, we have

$$\|\Delta_\Gamma\|_1 \le \sqrt{(\sqrt{4(p_1+r)})^2 + (4\sqrt{s})^2}\sqrt{|||\Delta_\Gamma|||_\mathrm{F}^2 + |||\Delta_\Theta/\sqrt{n}|||_\mathrm{F}^2},$$

that is,

$$\|\Delta_\Gamma\|_1^2 \le 4\big(p_1 + r + 4s\big)\Big[|||\Delta_\Gamma|||_\mathrm{F}^2 + |||\Delta_\Theta/\sqrt{n}|||_\mathrm{F}^2\Big]. \tag{27}$$

Combine (23) and (27), a lower bound for the LHS of (21) is given by

$$\Big(\frac{\alpha^\mathbf{X}_\mathrm{RSC}}{2} - 4\tau_\mathbf{X}\big(p_1 + r + 4s\big)\Big)|||\Delta_\Gamma|||_\mathrm{F}^2 + \Big(\frac{1}{2} - 4\tau_\mathbf{X}\big(p_1 + r + 4s\big)\Big)|||\Delta_\Theta/\sqrt{n}|||_\mathrm{F}^2. \tag{28}$$

Step (ii). For the first term in the RHS of (21), using the duality of ℓ_1-ℓ_∞ dual norm pair, the following inequality holds:

$$\begin{aligned}\frac{1}{n}|\langle\!\langle \mathbf{X}\Delta_\Gamma^\top, \widehat{\Theta}-\Theta^\star\rangle\!\rangle| &\le \frac{1}{n}|\langle\!\langle \Delta_\Gamma^\top, \mathbf{X}^\top\widehat{\Theta}\rangle\!\rangle| + \frac{1}{n}|\langle\!\langle \Delta_\Gamma^\top, \mathbf{X}^\top\Theta^\star\rangle\!\rangle| \\ &\le \|\Delta_\Gamma\|_1\|\mathbf{X}^\top\widehat{\Theta}/n\|_\infty + \|\Delta_\Gamma\|_1\|\mathbf{X}^\top\Theta^\star/n\|_\infty \\ &\le \|\Delta_\Gamma\|_1\cdot\|\mathbf{X}/n\|_1\cdot\|\widehat{\Theta}\|_\infty + \|\Delta_\Gamma\|_1\cdot\|\mathbf{X}/n\|_1\cdot\|\Theta^\star\|_\infty.\end{aligned} \tag{29}$$

Using the fact that both $\Theta^\star$ and $\widehat{\Theta}$ are feasible and satisfy the box constraint $\|\Theta\|_\infty \le \frac{\phi}{\kappa(\mathcal{R}^*)|||\mathbf{X}/\sqrt{n}|||_{\mathrm{op}}}$, it follows that

$$|||\mathbf{X}/n|||_1\|\widehat{\Theta}\|_\infty \le \frac{\phi}{\kappa(\mathcal{R}^*)} \quad \text{and} \quad |||\mathbf{X}/n|||_1\|\Theta^\star\|_\infty \le \frac{\phi}{\kappa(\mathcal{R}^*)},$$

Consequently, (29) is upper bounded by $\frac{2\phi}{\kappa(\mathcal{R}^*)}\cdot\|\Delta_\Gamma\|_1$. By additionally requiring λ_Γ to satisfy

$$\lambda_\Gamma \ge 4\phi/\kappa(\mathcal{R}^*),$$

and combining (24), (25) and (26), the following upper bound holds for the RHS of (21):

$$\begin{aligned}&\tfrac{\lambda_\Gamma}{2}\|\Delta_\Gamma\|_1 + \tfrac{\lambda_\Gamma}{2}\|\Delta_\Gamma\|_1 + \lambda_\Gamma\sqrt{p_1+r}|||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}} + \lambda_\Gamma(\|\Delta_{\Gamma|\mathbb{S}}\|_1 - \|\Delta_{\Gamma|\mathbb{S}^c}\|_1) \\ &\qquad\le \lambda_\Gamma\Big(2\sqrt{s_{\Gamma^\star}}|||\Delta_\Gamma|||_{\mathrm{F}} + \sqrt{p_1+r}|||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}\Big) \\ &\qquad\le \lambda_\Gamma\sqrt{4s_{\Gamma^\star}+p_1+r}\sqrt{|||\Delta_\Gamma|||_{\mathrm{F}}^2 + |||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}^2}.\end{aligned} \tag{30}$$

Step (iii). Combine (28) and (30), by rearranging terms and requiring $\tau_{\mathbf{X}}$ to satisfy $\tau_{\mathbf{X}}(p_1 + r + 4s_{\Gamma^\star}) < \min\{\alpha^{\mathbf{X}}_{\mathrm{RSC}}, 1\}/16$, the following inequality holds:

$$\begin{aligned}&\frac{\min\{\alpha^{\mathbf{X}}_{\mathrm{RSC}},1\}}{4}\Big(|||\Delta_\Gamma|||_{\mathrm{F}}^2 + |||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}^2\Big) \\ &\qquad\le \lambda_\Gamma\sqrt{4s_{\Gamma^\star}+p_1+r}\sqrt{|||\Delta_\Gamma|||_{\mathrm{F}}^2 + |||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}^2},\end{aligned}$$

which gives

$$|||\Delta_\Gamma|||_{\mathrm{F}}^2 + |||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}^2 \le \frac{16\lambda_\Gamma^2\Big(p_1+r+4s_{\Gamma^\star}\Big)}{\min\{\alpha^{\mathbf{X}}_{\mathrm{RSC}},1\}^2}.$$

■

Proof sketch for Theorem 1. First we note that the requirement on the tuning parameter λ_Γ determines the leading term in the ultimate high probability error bound. By Lemma 4, to have adequate concentration for the leading eigenvalue $\Lambda_{\max}(S_{\mathbf{E}})$ of the sample covariance matrices, the requirement imposed on the sample size makes $\sqrt{\log(p_2 q)/n}$ a lower order term relative to $\Lambda^{1/2}_{\max}(\Sigma_e)$, with the latter being an $\mathcal{O}(1)$ term. Consequently, the choice of the tuning parameter effectively becomes

$$\lambda_\Gamma \asymp \mathcal{O}(1),$$

The conclusion readily follows as a result of Proposition 1. ■

Part 2. This part contains the proofs for the results related to $\widehat{A}$.

Proof sketch for Proposition 2. The result follows along the lines of Basu and Michailidis (2015, Proposition 4.1). In particular, in Basu and Michailidis (2015), the authors consider estimation of A based on the directly observed samples of the X_t process, with the restricted eigenvalue (RE) condition imposed on the corresponding Hessian matrix and the tuning parameter selected in accordance to the deviation bound defined in Definition 2. On the other hand, in the current setting, estimation of the transition matrix is based on quantities that are surrogates for the true sample quantities. Consequently, as long as the required conditions are imposed on their counterparts associated with these surrogate quantities, the conclusion directly follows.

Finally, we would like to remark that the RSC condition used is in essence identical to the RE condition required in Basu and Michailidis (2015) in the setting under consideration. ■

Proof of Theorem 2. First, we note that under (IR), by Theorem 1, there exists some constant K_1 that is independent of n, p_1, p_2 and q such that the following event holds with probability at least $\mathsf{P}_1 := 1 - c_1 \exp(-c_2 \log(p_2 q))$:

$$\mathcal{E}_1 := \Big\{ |||\Delta_{\mathbf{F}}/\sqrt{n}|||_{\mathrm{F}} \leq K_1 \Big\}.$$

Conditional on $\mathcal{E}_1$, by Proposition 2, Lemmas 6 and 7, with high probability, the following event holds:

$$\mathcal{E}_2 := \Big\{ |||\Delta_A|||_{\mathrm{F}} \leq \varphi(n, p_1, p_2, K_1) \Big\},$$

for some function $\varphi(\cdot)$ that not only depends on sample size and dimensions, but also on K_1, provided that the "conditional" RSC condition is satisfied. What are left to be examined are: (i) what does $\mathcal{E}_1$ imply in terms of the RSC condition being satisfied *unconditionally*; and (ii) what does $\mathcal{E}_1$ imply in terms of the bound in $\mathcal{E}_2$,

Towards this end, for (i), we note that since

$$\Lambda_{\max}^{1/2}(S_{\Delta_{\mathbf{F}_{n-1}}}) = |||\Delta_{\mathbf{F}}/\sqrt{n}|||_{op} \leq |||\Delta_{\mathbf{F}}/\sqrt{n}|||_{\mathrm{F}} \leq K_1,$$

then as long as C_Z in condition C3 satisfies $C_Z \geq c_0 K_1$ with the specified $c_0 \geq 6\sqrt{165\pi}$, with probability at least $P_1 P_{2,\mathrm{RSC}}$ where we define $\mathsf{P}_{2,\mathrm{RSC}} := 1 - c_1' \exp(-c_2' n)$, by Lemma 6 the required RSC condition is guaranteed to be satisfied with a positive curvature. For (ii), with the aid of Lemma 7, with probability at least $\mathsf{P}_1 \mathsf{P}_{2,\mathrm{DB}}$ where we define $\mathsf{P}_{2,\mathrm{DB}} := 1 - c_1' \exp(-c_2' \log(p_1 + p_2))$, the following bound holds for the deviation bound $C(n, p_1, p_2)$ *unconditionally*:[10]

$$\begin{aligned} C(n, p_1, p_2) \leq\ & C_1\big(\mathcal{M}(f_Z) + \frac{\Sigma_w}{2\pi} + \mathcal{M}(f_{Z,W})\big)\sqrt{\frac{\log(p_1 + p_2)}{n}} \\ & + C_2 \mathcal{M}^{1/2}(f_Z)\sqrt{\frac{\log(p_1 + p_2) + \log p_1}{n}} + C_3 \Lambda_{\max}^{1/2}(\Sigma_w)\sqrt{\frac{\log(p_1 + p_2)}{n}} + C_4, \end{aligned}$$

10. Note that it can be shown that $|||\varepsilon_n|||_{\mathrm{F}}^2 = O(|||\Delta_{\mathbf{F}}|||_{\mathrm{F}}^2)$

where the constants $\{C_i\}$ have already absorbed the upper error bound K_1 of the Stage I estimates, compared with the original expression in Proposition 2. With the required sample size, the constant becomes the leading term, so that there exists some constant K_2 such that *unconditionally*:

$$C(n, p_1, p_2) \leq K_2 \asymp \mathcal{O}(1).$$

Combine (i) and (ii), and with probability at least $\min\{\mathsf{P}_1\mathsf{P}_{2,\mathrm{RSC}}, \mathsf{P}_1\mathsf{P}_{2,\mathrm{DB}}\}$, the bound in Theorem 2 holds. ■

Appendix B. Proof for Lemmas

In this section, we provide proofs for the lemmas in Section 3.2.

Proof of Lemma 1. Note that

$$\begin{aligned}\widehat{\Theta} = \Theta^\star + \Delta_\Theta &= (\mathbf{F} + \Delta_\mathrm{F})(\Lambda^\star + \Delta_\Lambda)^\top \\ \Delta_\Theta &= \Delta_\mathrm{F}(\Lambda^\star)^\top + \widehat{\mathbf{F}}\Delta_\Lambda^\top.\end{aligned}$$

Multiply the left inverse of $\widehat{\mathbf{F}}$ which gives

$$\Delta_\Lambda^\top = \big(\widehat{\mathbf{F}}^\top\widehat{\mathbf{F}}\big)^{-1}\widehat{\mathbf{F}}^\top\Delta_\Theta + \big(\widehat{\mathbf{F}}^\top\widehat{\mathbf{F}}\big)^{-1}\widehat{\mathbf{F}}^\top\Delta_{\mathbf{F}}(\Lambda^\star)^\top.$$

Since for some generic matrix M, we have $|||M^{-1}|||_\mathrm{F} \geq (|||M|||_\mathrm{F})^{-1}$, an application of the triangle inequality gives

$$\begin{aligned}|||\Delta_\Lambda|||_\mathrm{F} &\leq \frac{|||\widehat{\mathbf{F}}|||_\mathrm{F}}{|||\widehat{\mathbf{F}}^\top\widehat{\mathbf{F}}|||_\mathrm{F}}\Big(|||\Delta_\Theta|||_\mathrm{F} + |||\Delta_{\mathbf{F}}(\Lambda^\star)^\top|||_\mathrm{F}\Big) \\ &= \frac{|||\widehat{\mathbf{F}}/\sqrt{n}|||_\mathrm{F}}{|||\frac{1}{n}\widehat{\mathbf{F}}^\top\widehat{\mathbf{F}}|||_\mathrm{F}}(\frac{1}{\sqrt{n}})\Big(|||\Delta_\Theta|||_\mathrm{F} + |||\Delta_{\mathbf{F}}(\Lambda^\star)^\top|||_\mathrm{F}\Big) \\ &\leq \sqrt{p_1}\Lambda_{\max}^{-1/2}(S_{\widehat{\mathbf{F}}})|||\Delta_\Theta/\sqrt{n}|||_\mathrm{F}\Big(1 + |||\Lambda^\star|||_\mathrm{F}\Big),\end{aligned}$$

where $S_{\widehat{\mathbf{F}}} := \frac{1}{n}\widehat{\mathbf{F}}^\top\widehat{\mathbf{F}}$, and the numerator and the denominator of $\frac{|||\widehat{\mathbf{F}}|||_\mathrm{F}}{|||\widehat{\mathbf{F}}^\top\widehat{\mathbf{F}}|||_\mathrm{F}}$ are respectively by

$$|||\widehat{\mathbf{F}}|||_\mathrm{F} \leq \sqrt{p_1}|||\widehat{\mathbf{F}}|||_{\mathrm{op}}, \qquad |||\widehat{\mathbf{F}}^\top\widehat{\mathbf{F}}|||_\mathrm{F} \geq |||\widehat{\mathbf{F}}^\top\widehat{\mathbf{F}}|||_{\mathrm{op}}.$$

Further, note that $|||\widehat{\mathbf{F}}/\sqrt{n}|||_{\mathrm{op}}^2 = \Lambda_{\max}(S_{\widehat{\mathbf{F}}}) = |||S_{\widehat{\mathbf{F}}}|||_{\mathrm{op}}$. What remains is to obtain a lower bound for

$$\Lambda_{\max}^{1/2}\big(S_{\widehat{\mathbf{F}}}\big) = |||(\mathbf{F} + \Delta_{\mathbf{F}})/\sqrt{n}|||_{\mathrm{op}}.$$

One such bound is given by

$$\begin{aligned}|||(\mathbf{F} + \Delta_{\mathbf{F}})/\sqrt{n}|||_{\mathrm{op}} \geq |||\mathbf{F}/\sqrt{n}|||_{\mathrm{op}} - |||\Delta_{\mathbf{F}}/\sqrt{n}|||_{\mathrm{op}} &\geq |||\mathbf{F}/\sqrt{n}|||_{\mathrm{op}} - |||\Delta_{\mathbf{F}}/\sqrt{n}|||_\mathrm{F} \\ &\geq \Lambda_{\max}^{1/2}(S_{\mathbf{F}}) - |||\Delta_\Theta/\sqrt{n}|||_\mathrm{F},\end{aligned}$$

which leads to the following bound for $|||\Delta_\Lambda|||_{\mathrm{F}}$, provided that the RHS is positive:

$$\frac{|||\Delta_\Lambda|||_{\mathrm{F}}}{|||\Lambda^\star|||_{\mathrm{F}}} \leq \sqrt{p_1}\frac{|||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}}{\Lambda_{\max}^{1/2}(S_{\mathbf{F}}) - |||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}}\Big(1 + 1/|||\Lambda^\star|||_{\mathrm{F}}\Big).$$

■

Proof of Lemma 2. First, suppose we have

$$\frac{1}{2}v'S_{\mathbf{X}}v = \frac{1}{2}v'\Big(\frac{\mathbf{X}'\mathbf{X}}{n}\Big)v \geq \frac{\alpha_{\mathrm{RSC}}}{2}\|v\|_2^2 - \tau_n\|v\|_1^2, \quad \forall\ v \in \mathbb{R}^p; \tag{31}$$

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

$$\begin{aligned}\frac{1}{2n}|||\mathbf{X}\Delta|||_{\mathrm{F}}^2 = \frac{1}{2}\sum_{j=1}^q \Delta_j'\big(\tfrac{\mathbf{X}'\mathbf{X}}{n}\big)\Delta_j &\geq \frac{\alpha_{\mathrm{RSC}}}{2}\sum_{j=1}^q\|\Delta_j\|_2^2 - \tau_n\sum_{j=1}^q\|\Delta_j\|_1^2 \\ &\geq \frac{\alpha_{\mathrm{RSC}}}{2}|||\Delta|||_{\mathrm{F}}^2 - \tau_n\|\Delta\|_1^2.\end{aligned}$$

Therefore, it suffices to verify that (31) holds. In Basu and Michailidis (2015, Proposition 4.2), the authors prove a similar result under the assumption that X_t is a VAR(d) process. Here, we adopt the same proof strategy and state the result for a *more general process* X_t.

Specifically, by Basu and Michailidis (2015, Proposition 2.4(a)), $\forall v \in \mathbb{R}^p, \|v\| \leq 1$ and $\eta > 0$,

$$\mathbb{P}\Big[\big|v'\big(S_{\mathbf{X}} - \Sigma_X(h)\big)v\big| > 2\pi\mathcal{M}(g_X)\eta\Big] \leq 2\eta\exp\Big(-cn\min\{\eta^2,\eta\}\Big).$$

Applying the discretization in Basu and Michailidis (2015, Lemma F.2) and taking the union bound, define $\mathbb{K}(2s) := \{v \in \mathbb{R}^p, \|v\| \leq 1, \|v\|_0 \leq 2k\}$, and the following inequality holds:

$$\begin{aligned}&\mathbb{P}\Big[\sup_{v\in\mathbb{K}(2k)}\big|v'\big(S_{\mathbf{X}} - \Sigma_X(h)\big)v\big| > 2\pi\mathcal{M}(g_X)\eta\Big] \\ &\qquad\leq 2\exp\Big(-cn\min\{\eta,\eta^2\} + 2k\min\{\log p, \log(21ep/2k)\}\Big).\end{aligned}$$

With the specified $\gamma = 54\mathcal{M}(g_X)/\mathfrak{m}(g_X)$, set $\eta = \gamma^{-1}$, then apply results from Loh and Wainwright (2012, Lemma 12) with $\Gamma = S_{\mathbf{X}} - \Sigma_X(0)$ and $\delta = \pi\mathfrak{m}(g_X)/27$, so that the following holds

$$\frac{1}{2}v'S_{\mathbf{X}}v \geq \frac{\alpha_{\mathrm{RSC}}}{2}\|v\|^2 - \frac{\alpha_{\mathrm{RSC}}}{2k}\|v\|_1^2,$$

with probability at least $1-2\exp\big(-cn\min\{\gamma^{-2},1\}+2k\log p\big)$ and note that $\min\{\gamma^{-2},1\} = \gamma^{-2}$ since $\gamma > 1$. Finally, let $k = \min\{cn\gamma^{-2}/(c'\log p), 1\}$ for some $c' > 2$, and conclude that with probability at least $1 - c_1\exp(-c_2 n)$, the inequality in (31) holds with

$$\alpha_{\mathrm{RSC}} = \pi\mathfrak{m}(g_X), \qquad \tau_n = \alpha_{\mathrm{RSC}}\gamma^2\frac{\log p}{2n},$$

and so does the RSC condition. ■

Proof of Lemma 3. We note that

$$\frac{1}{n}\Big\|\mathbf{X}^\top\mathbf{E}\Big\|_\infty = \max_{1\le i,j\le p}\big|e_i^\top\big(\mathbf{X}^\top\mathbf{E}/n\big)e_j\big|,$$

where e_i is the p-dimensional standard basis with its i-th entry being 1. Applying Basu and Michailidis (2015, Proposition 2.4(b)), for an arbitrary pair of (i,j), the following inequality holds:

$$\mathbb{P}\Big[\big|e_i^\top\big(\mathbf{X}^\top\mathbf{E}/n\big)e_j\big| > 2\pi\big(\mathcal{M}(g_X) + \frac{\Lambda_{\max}(\Sigma_e)}{2\pi}\big)\eta\Big] \le 6\exp\Big(-cn\min\{\eta^2,\eta\}\Big),$$

and note that e_t is a pure noise term that is assumed to be independent of X_t; hence, there is no cross-dependence term to consider. Take the union bound over all $1\le i\le p_2, 1\le j\le q$, and the following bound holds:

$$\begin{aligned}\mathbb{P}\Big[\max_{1\le i\le p_2,1\le j\le q}\big|e_i^\top\big(\mathbf{X}^\top\mathbf{E}/n\big)e_j\big| > 2\pi\big(\mathcal{M}(g_X) + \frac{\Lambda_{\max}(\Sigma_e)}{2\pi}\big)\eta\Big]& \\ \le 6\exp\Big(-cn\min\{\eta^2,\eta\} + \log(p_2q)\Big).&\end{aligned}$$

Set $\eta = c'\sqrt{\log p/n}$ for $c' > (1/c)$ and with the choice of $n \gtrsim \log(p_2q)$, $\min\{\eta^2,\eta\} = \eta^2$, then with probability at least $1 - c_1\exp(-c_2\log p_2q)$, there exists some c_0 such that the following bound holds:

$$\frac{1}{n}\Big\|\mathbf{X}^\top\mathbf{E}\Big\|_\infty \le c_0\big(2\pi\mathcal{M}(g_X) + \Lambda_{\max}(\Sigma_e)\big)\sqrt{\frac{\log(p_2q)}{n}}.$$

■

Proof of Lemma 4. For $\mathbf{E}$ whose rows are iid realizations of a sub-Gaussian random vector e_t, by Wainwright (2009, Lemma 9), the following bound holds:

$$\mathbb{P}\Big[|||S_\mathbf{E} - \Sigma_e|||_{op} \ge \Lambda_{\max}(\Sigma_e)\delta(n,q,\eta)\Big] \le 2\exp(-n\eta^2/2),$$

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

$$|||S_\mathbf{E}|||_{op} \le |||\Sigma_\epsilon|||_{op} + |||S_\mathbf{E} - \Sigma_\epsilon|||_{op} \le \Lambda_{\max}(\Sigma_\epsilon) + \Lambda_{\max}(\Sigma_\epsilon)\delta(n,q,t).$$

So for $n \gtrsim q$, by setting $\eta = 1$, which yields $\delta(n,q,\eta) \le 8$ so that with probability at least $1 - 2\exp(-n/2)$, the following bound holds:

$$\Lambda_{\max}(S_\mathbf{E}) \le 9\Lambda_{\max}(\Sigma_\epsilon).$$

■

Proof of Lemma 5. To prove this lemma, we use a similar strategy as in the proof of Negahban and Wainwright (2011, Lemma 3) while taking into consideration the temporal dependence present in the rows of $\mathbf{X}$. In the remainder of the proof, we use p (instead of p_2) to denote generically the dimension of the process.

Let $S^p = \{u \in \mathbb{R}^p | \|u\| = 1\}$ denote the p-dimensional unit sphere. Then, $\Lambda_{\max}(S_{\mathbf{X}})$ is the operator norm of $S_{\mathbf{X}}$, which has the following variational representation form:

$$\Lambda_{\max}(S_{\mathbf{X}}) = \frac{1}{n}\left|\left|\left|\mathbf{X}'\mathbf{X}\right|\right|\right|_{\text{op}} = \frac{1}{n}\sup_{u\in S^p} u'\mathbf{X}'\mathbf{X}u.$$

For a positive scalar s, define

$$\Psi(s) := \sup_{u\in sS^{p_1}} \langle \mathbf{X}u, \mathbf{X}u\rangle;$$

the goal is to establish an upper bound for $\Psi(1)/n$. Let $\mathcal{A} = \{u^1, \cdots, u^A\}$ denote the $1/4$ covering of S^p. Negahban and Wainwright (2011) established that

$$\Psi(1) \leq 4\max_{u^a\in\mathcal{A}} \langle \mathbf{X}u^a, \mathbf{X}u^a\rangle;$$

further, according to Anderson (2011), there exists a $1/4$ covering of S^p with at most $|\mathcal{A}| \leq 8^p$ elements. Consequently,

$$\mathbb{P}\Big[\big|\frac{1}{n}\Psi(1)\big| \geq 4\delta\Big] \leq 8^p \max_{u^a} \mathbb{P}\Big[\frac{|(u^a)'\mathbf{X}\mathbf{X}(u^a)|}{n} \geq \delta\Big].$$

What remains to be bounded is $\frac{1}{n}u'\mathbf{X}'\mathbf{X}u$, for an arbitrary $u \in S^p$. By Basu and Michailidis (2015, Proposition 2.4(b)), we have

$$\mathbb{P}\Big[\Big|u'\Big(\big(\frac{\mathbf{X}'\mathbf{X}}{n}\big) - \Sigma_X(0)\Big)u\Big| > 2\pi\mathcal{M}(f_X)\eta\Big] \leq 2\exp\left(-cn\min\{\eta, \eta^2\}\right),$$

and thus

$$\mathbb{P}\Big[u'\big(\frac{\mathbf{X}'\mathbf{X}}{n}\big)u > 2\pi\mathcal{M}(f_X)\eta + \left|\left|\left|\Sigma_X(0)\right|\right|\right|_{\text{op}}\Big] \leq 2\exp\left(-cn\min\{\eta, \eta^2\}\right).$$

Therefore, it follows that

$$\mathbb{P}\Big[\big|\frac{1}{n}\Psi(1)\big| \geq 8\pi\mathcal{M}(f_X)\eta + 4\left|\left|\left|\Sigma_X(0)\right|\right|\right|_{\text{op}}\Big] \leq 2\exp\left(p\log 8 - cn\min\{\eta, \eta^2\}\right).$$

With the specified choice of sample size n, the probability vanishes by choosing $\eta = c_0'$ for constant c_0' sufficiently large. Finally, by Proposition 2.3 in Basu and Michailidis (2015), $\left|\left|\left|\Sigma_X(0)\right|\right|\right|_{\text{op}} \leq 2\pi\mathcal{M}(f_X)$, and thus the conclusion in Lemma 5 holds. ■

Proof of Lemma 6. It suffices to show that the following inequality holds with high probability for some curvature $\alpha^{\widehat{\mathbf{Z}}}_{\mathrm{RSC}} > 0$ and tolerance $\tau_{\mathbf{Z}}$, where we define $\widehat{S}_{\mathbf{Z}} := \frac{1}{n}\widehat{\mathbf{Z}}^{\top}_{n-1}\widehat{\mathbf{Z}}_{n-1}$:

$$\frac{1}{2}\theta^{\top}\widehat{S}_{\mathbf{Z}}\theta \geq \frac{\alpha^{\widehat{\mathbf{Z}}}_{\mathrm{RSC}}}{2}\|\theta\|^2 - \tau_{\mathbf{Z}}\|\theta\|_1^2, \qquad \forall\ \theta \in \mathbb{R}^p.$$

Define $S_{\mathbf{Z}} := \frac{1}{n}\mathbf{Z}^{\top}_{n-1}\mathbf{Z}_{n-1}$, then $\widehat{S}_{\mathbf{Z}}$ can be written as

$$\widehat{S}_{\mathbf{Z}} = S_{\mathbf{Z}} + \Big(\frac{1}{n}\mathbf{Z}^{\top}_{n-1}\Delta_{\mathbf{Z}_{n-1}} + \frac{1}{n}\Delta^{\top}_{\mathbf{Z}_{n-1}}\mathbf{Z}_{n-1}\Big) + \Big(\frac{1}{n}\Delta^{\top}_{\mathbf{Z}_{n-1}}\Delta_{\mathbf{Z}_{n-1}}\Big), \tag{32}$$

First, notice that the last term satisfies the following natural lower bound *deterministically*, since $\Delta_{\mathbf{F}}$ is assumed non-random and $\Delta_{\mathbf{Z}} = [\Delta_{\mathbf{F}}, O]$:

$$\theta^{\top}\Big(\frac{1}{n}\Delta^{\top}_{\mathbf{Z}_{n-1}}\Delta_{\mathbf{Z}_{n-1}}\Big)\theta \geq 0 \qquad \forall\ \theta \in \mathbb{R}^p,$$

which however, does not contribute to the "positive" part of curvature. For the first two terms, we adopt the following strategy, using Lemma 12 in Loh and Wainwright (2012) as an intermediate step. Specifically, Loh and Wainwright (2012, Lemma 12) proves that for any fixed generic matrix $\Gamma \in \mathbb{R}^{p\times p}$ that satisfies $|\theta^{\top}\Gamma\theta| \leq \delta$ for any $\theta \in \mathbb{K}(2s)$[11], the following bound holds

$$|\theta^{\top}\Gamma\theta| \leq 27\delta\big(\|\theta\|_2^2 + \frac{1}{s}\|\theta\|_1^2\big), \qquad \forall\ \theta \in \mathbb{R}^p. \tag{33}$$

Then, based on (33), consider $\Gamma = \widehat{\Gamma} - \Sigma$ then rearrange terms, so that $\theta^{\top}\widehat{\Gamma}\theta \geq \theta^{\top}\Sigma\theta - \frac{27\delta}{2}\big(\|\theta\|_2^2 + \frac{1}{2}\|\theta\|_1^2\big)$. The RE condition follows by setting δ to be some quantity related to $\Lambda_{\min}(\Sigma)$.

In light of this, for the first two terms in (32), let

$$\Psi := S_{\mathbf{Z}} + \Big(\frac{1}{n}\mathbf{Z}^{\top}_{n-1}\Delta_{\mathbf{Z}_{n-1}} + \frac{1}{n}\Delta^{\top}_{\mathbf{Z}_{n-1}}\mathbf{Z}_{n-1}\Big),$$

denote their sum, in order to obtain an upper bound for $\big|\theta^{\top}\big(\Psi - \Sigma_Z(0)\big)\theta\big|$, so that Lemma 12 in Loh and Wainwright (2012) can be applied. To this end, since

$$\Big|\theta^{\top}\big[\Psi - \Sigma_Z(0)\big]\theta\Big| \leq \Big|\theta'(S_{\mathbf{Z}} - \Sigma_Z(0))\theta\Big| + \Big|\theta'\Big(\frac{1}{n}\mathbf{Z}'_{n-1}\Delta_{\mathbf{Z}_{n-1}} + \frac{1}{n}\Delta'_{\mathbf{Z}_{n-1}}\mathbf{Z}_{n-1}\Big)\theta\Big|,$$

we consider getting upper bounds for each of the two terms:

$$\text{(i)}\ \ \Big|\theta'(S_{\mathbf{Z}} - \Sigma_Z(0))\theta\Big|, \qquad \text{(ii)}\ \ \Big|\theta'\Big(\frac{1}{n}\mathbf{Z}'_{n-1}\Delta_{\mathbf{Z}_{n-1}} + \frac{1}{n}\Delta'_{\mathbf{Z}_{n-1}}\mathbf{Z}_{n-1}\Big)\theta\Big|.$$

For (i), we follow the derivation in Basu and Michailidis (2015, Proposition 2.4(a)), that is, for all $\|\theta\| \leq 1$,

$$\mathbb{P}\Big[\Big|\theta'\big(S_{\mathbf{Z}} - \Sigma_Z(0)\big)\theta\Big| > 2\pi\mathcal{M}(f_Z)\eta\Big] \leq 2\exp\big[-cn\min\{\eta^2, \eta\}\big],$$

11. $\mathbb{K}(2s) := \{\theta : \|\theta\|_0 = 2s\}$ is the set of $2s$-sparse vectors.

and further with probability at least

$$1 - 2\exp\big(- cn\min\{\eta^2, \eta\} + 2s\min\{\log p, \log(21ep/2s)\}\big),$$

the following bound holds:

$$\sup_{\theta\in\mathbb{K}(2s)} \Big|\theta^\top \big(S_{\mathbf{Z}} - \Sigma_Z(0)\big)\theta\Big| < 2\pi\mathcal{M}(f_Z)\eta. \tag{34}$$

For (ii), the two terms are identical, with either one given by

$$\frac{1}{n}(\mathbf{Z}_{n-1}\theta)^\top(\Delta_{\mathbf{Z}_{n-1}}\theta).$$

To obtain its upper bound, consider the following inequality, based on which we bound the two terms in the product separately:

$$\sup_{\theta\in\mathbb{K}(2s)} \Big|\frac{1}{n}\langle \mathbf{Z}_{n-1}\theta, \Delta_{\mathbf{Z}_{n-1}}\theta\rangle\Big| \le \Big(\sup_{\theta\in\mathbb{K}(2s)} \|\frac{\mathbf{Z}_{n-1}\theta}{\sqrt{n}}\|\Big)\Big(\sup_{\|\theta\|\le 1} \|\frac{\Delta_{\mathbf{Z}_{n-1}}\theta}{\sqrt{n}}\|\Big). \tag{35}$$

For the first term in (35), since rows of $\mathbf{Z}_{n-1}$ are time series realizations from (5), then if we let $\xi := \mathbf{Z}_{n-1}\theta$, $\xi \sim \mathcal{N}(0_{n\times 1}, Q_{n\times n})$ is Gaussian with $Q_{st} = \theta'\Sigma_Z(t-s)\theta$. To get its upper bound, we bound its square, and use again (34), that is,

$$\sup_{\theta\in\mathbb{K}(2s)} \Big|\theta^\top\Big(\frac{1}{n}\mathbf{Z}_{n-1}^\top\mathbf{Z}_{n-1}\Big)\theta\Big| \le \sup_{\theta\in\mathbb{K}(2s)} \theta'\Sigma_Z(0)\theta + 2\pi\mathcal{M}(f_Z) \le 2\pi\mathcal{M}(f_Z) + 2\pi\mathcal{M}(f_Z)\eta.$$

For the second term $\|\Delta_{\mathbf{Z}_{n-1}}\theta/\sqrt{n}\|$, this is non-random, and for all $\|\theta\| \le 1$, $\|\Delta_{\mathbf{Z}_{n-1}}\theta/\sqrt{n}\| \le \Lambda_{\max}^{1/2}\big(S_{\Delta_{\mathbf{Z}_{n-1}}}\big) = \Lambda_{\max}^{1/2}\big(S_{\Delta_{\mathbf{F}_{n-1}}}\big)$. Therefore, the following bound holds for (35):

$$\sup_{\theta\in\mathbb{K}(2s)} \Big|\frac{1}{n}\langle \mathbf{Z}_{n-1}\theta, \Delta_{\mathbf{Z}_{n-1}}\theta\rangle\Big| \le \Lambda_{\max}^{1/2}\big(S_{\Delta_{\mathbf{F}_{n-1}}}\big)\sqrt{2\pi\mathcal{M}(f_Z) + 2\pi\mathcal{M}(f_Z)\eta}. \tag{36}$$

Combine (34) and (36) that are respectively the bounds for (i) and (ii), and the following bound holds with probability at least $1-2\exp\big(-cn\min\{\eta^2,\eta\}+2s\min\{\log p, \log(21ep/2s)\}\big)$:

$$\sup_{\theta\in\mathbb{K}(2s)} \Big|\theta^\top\Big(\Psi - \Sigma_Z(0)\Big)\theta\Big| \le 2\pi\mathcal{M}(f_Z)\eta + 2\Lambda_{\max}^{1/2}\big(S_{\Delta_{\mathbf{F}_{n-1}}}\big)\sqrt{2\pi\mathcal{M}(f_Z) + 2\pi\mathcal{M}(f_Z)\eta}. \tag{37}$$

Now applying Loh and Wainwright (2012, Lemma 12) to $\Gamma = \Psi - \Sigma_Z(0)$, and δ being the RHS of (37), then the following bound holds:

$$\theta^\top\widehat{S}_{\mathbf{Z}}\theta \ge 2\pi\mathfrak{m}(f_Z)\|\theta\|_2^2 - 27\delta(\|\theta\|_2^2 + \frac{1}{s}\|\theta\|_1^2) = \big(2\pi\mathfrak{m}(f_Z) - 27\delta\big)\|\theta\|^2 - \frac{27\delta}{s}\|\theta\|_1^2.$$

By setting $\eta = \omega^{-1} := \frac{\mathfrak{m}(f_Z)}{54\mathcal{M}(f_Z)}$,

$$\begin{aligned}\delta &= \frac{\pi}{27}\mathfrak{m}(f_Z) + 2\Lambda_{\max}^{1/2}\big(S_{\Delta_{\mathbf{F}_{n-1}}}\big)\sqrt{2\pi\mathcal{M}(f_Z) + \pi\mathfrak{m}(f_Z)/27} \\ &\le \frac{\pi}{27}\mathfrak{m}(f_Z) + 2\Lambda_{\max}^{1/2}\big(S_{\Delta_{\mathbf{F}_{n-1}}}\big)\sqrt{\frac{55\pi}{27}\mathcal{M}(f_Z)}.\end{aligned}$$

Since we have required that $\mathfrak{m}(f_Z)/\mathcal{M}^{1/2}(f_Z) > c_0 \cdot \Lambda_{\max}^{1/2}(S_{\Delta_{\mathbf{F}_{n-1}}})$ with $c_0 \geq 6\sqrt{165\pi}$, $2\pi\mathfrak{m}(f_Z) - 27\delta > 0$. Therefore, the RSC condition is satisfied with curvature

$$\alpha_{\mathrm{RSC}}^{\widehat{\mathbf{Z}}} = 2\pi\mathfrak{m}(f_Z) - 27\delta = \pi\mathfrak{m}(f_Z) - 54\Lambda_{\max}^{1/2}\big(S_{\Delta_{\mathbf{F}_{n-1}}}\big)\sqrt{2\pi\mathcal{M}(f_Z) + \pi\mathfrak{m}(f_Z)/27} > 0,$$

and tolerance $27\delta/(2s)$, with probability at least $1 - 2\exp\Big(-cn\omega^{-2} + 2s\log p\Big)$. Finally, set $s = \lceil cn\omega^{-1}/4\log p\rceil$, we get the desired conclusion. ■

Proof of Lemma 7. First, we note that the quantity of interest can be upper bounded by the following four terms:

$$\begin{aligned}
&\frac{1}{n}\|\widehat{\mathbf{Z}}_{n-1}^{\top}\Big(\widehat{\mathbf{Z}}_n - \widehat{\mathbf{Z}}_{n-1}(A^\star)^{\top}\Big)\|_\infty \\
&\qquad = \frac{1}{n}\|\Big(\mathbf{Z}_{n-1} + \Delta_{\mathbf{Z}_{n-1}}\Big)^{\top}\Big(\mathbf{W} + \Delta_{\mathbf{Z}_n} - \Delta_{\mathbf{Z}_{n-1}}(A^\star)^{\top}\Big)\|_\infty \\
&\qquad \leq \|\frac{1}{n}\mathbf{Z}_{n-1}^{\top}\mathbf{W}\|_\infty + \|\frac{1}{n}\Delta_{\mathbf{Z}_{n-1}}^{\top}\mathbf{W}\|_\infty + \|\frac{1}{n}\mathbf{Z}_{n-1}^{\top}\Big(\Delta_{\mathbf{Z}_n} - \Delta_{\mathbf{Z}_{n-1}}(A^\star)^{\top}\Big)\|_\infty \\
&\qquad\quad + \|\frac{1}{n}\Delta_{\mathbf{Z}_{n-1}}^{\top}\Big(\Delta_{\mathbf{Z}_n} - \Delta_{\mathbf{Z}_{n-1}}(A^\star)^{\top}\Big)\|_\infty \\
&\qquad := T_1 + T_2 + T_3 + T_4. \qquad (38)
\end{aligned}$$

We provide bounds on each term in (38) sequentially. T_1 is the standard Deviation Bound, which according to previous derivations (e.g., Basu and Michailidis (2015) for the expression specifically derived for VAR(1)) satisfies

$$\frac{1}{n}\|\mathbf{Z}_{n-1}^{\top}\mathbf{W}\|_\infty \leq c_0\big[\mathcal{M}(f_Z) + \mathcal{M}(f_W) + \mathcal{M}(f_{Z,W^+})\big]\sqrt{\frac{\log(p_1+p_2)}{n}}$$

with probability at least $1 - c_1\exp(-c_2\log(p_1+p_2))$ for some $\{c_i\}$. For T_2, since rows of $\mathbf{W}$ are iid realizations from $\mathcal{N}(0,\Sigma_w)$, then for $\Delta_{\mathbf{Z}_{n-1}}^{\top}\mathbf{W} \in \mathbb{R}^{(p_1+p_2)\times(p_1+p_2)}$ which has at most $p_1 \times (p_1+p_2)$ nonzero entries, each entry (i,j) given by

$$\kappa_{ij} := \big(\frac{1}{n}\Delta_{\mathbf{Z}_{n-1}}^{\top}\mathbf{W}\big)_{ij} = \frac{1}{n}\Delta_{\mathbf{Z}_{n-1},\cdot i}^{\top}\mathbf{W}_{\cdot j}$$

is Gaussian, and the following tail bound holds:

$$\begin{aligned}
\mathbb{P}\big[|\kappa_{ij}| \geq t\big] &\leq e\cdot\exp\Big(-\frac{cnt^2}{\Lambda_{\max}(\Sigma_w)\max\limits_{i\in\{1,\ldots,p_1+p_2\}}\|\Delta_{\mathbf{Z}\cdot i}/\sqrt{n}\|_2^2}\Big) \\
&= e\cdot\exp\Big(-\frac{cnt^2}{\Lambda_{\max}(\Sigma_w)\max\limits_{i\in\{1,\ldots,p_1\}}\|\Delta_{\mathbf{F}\cdot i}/\sqrt{n}\|_2^2}\Big).
\end{aligned}$$

Taking the union bound over all $p_1 \times (p_1+p_2)$ nonzero entries, the following bound holds:

$$\mathbb{P}\Big[\frac{1}{n}\|\Delta_{\mathbf{Z}_{n-1}}^{\top}\mathbf{W}\|_\infty \geq t\Big] \leq \exp\Big(-\frac{cnt^2}{\Lambda_{\max}(\Sigma_w)\max\limits_{i\in\{1,\ldots,p_1\}}\|\Delta_{\mathbf{F}\cdot i}/\sqrt{n}\|_2^2} + \log\big(ep_1(p_1+p_2)\big)\Big).$$

Choose $t = c_0\big(\Lambda_{\max}^{1/2}(\Sigma_w) \max_{i=1,\ldots,p_1} \|\Delta_{\mathbf{F}\cdot i}/\sqrt{n}\|\big)\sqrt{\frac{\log(p_1(p_1+p_2))}{n}}$, the following bound holds with probability at least $1 - \exp\Big(-c_1 \log\big(p_1(p_1+p_2)\big)\Big)$:

$$\frac{1}{n}\|\Delta_{\mathbf{Z}_{n-1}}^\top \mathbf{W}\|_\infty \le c_0 \Lambda_{\max}^{1/2}(\Sigma_w) \max_{i=1,\ldots,p_1} \|\Delta_{\mathbf{F}\cdot i}/\sqrt{n}\| \sqrt{\frac{\log p_1 + \log(p_1+p_2)}{n}}.$$

For T_3, let $\varepsilon_n := \Delta_{\mathbf{Z}_n} - \Delta_{\mathbf{Z}_{n-1}}(A^\star)^\top = [\Delta_{\mathbf{F}_n} - \Delta_{\mathbf{F}_{n-1}}(A_{11}^\star)^\top, -\Delta_{\mathbf{F}_{n-1}}(A_{21}^\star)^\top]$, then each entry of $\frac{1}{n}\mathbf{Z}_{n-1}^\top \varepsilon_n$ is given by

$$\Big(\tfrac{1}{n}\mathbf{Z}_{n-1}^\top \varepsilon_n\Big)_{ij} = \frac{1}{n}\mathbf{Z}_{n-1,\cdot i}^\top \varepsilon_{n,\cdot j},$$

and it has $(p_1+p_2)\times(p_1+p_2)$ entries. Next, note that column i of $\mathbf{Z}_{n-1} \in \mathbb{R}^n$ can be viewed as a mean-zero Gaussian random vector with covariance matrix Q^i where $(Q^i)_{st} = [\Sigma_Z(t-s)]_{ii}$ satisfying $\Lambda_{\max}(Q^i) \le \Lambda_{\max}(\Sigma_Z(0)) \le 2\pi\mathcal{M}(f_Z)$, so for any (i,j), $\big(\frac{1}{n}\mathbf{Z}_{n-1}^\top \varepsilon_n\big)_{ij}$ satisfies

$$\mathbb{P}\Big[\big|\big(\tfrac{1}{n}\mathbf{Z}_{n-1}^\top \varepsilon_n\big)_{ij}\big| > t\Big] \le \exp\Big(1 - \frac{cnt^2}{\Lambda_{\max}(\Sigma_Z(0)) \max_{j\in\{1,\ldots,p_1\}} \|\varepsilon_{n,\cdot j}/\sqrt{n}\|^2}\Big).$$

Again by taking the union bound over all $(p_1+p_2)^2$ entries, and let

$$t = c_0\big(2\pi\mathcal{M}(f_Z)\big)^{1/2} \max_{j\in\{1,\ldots,p_1\}} \|\varepsilon_{n,\cdot j}/\sqrt{n}\| \sqrt{\frac{\log p_1 + \log(p_1+p_2)}{n}},$$

the following bound holds w.p. at least $1 - \exp(-c_1 \log(p_1+p_2))$:

$$\begin{aligned}
&\frac{1}{n}\|\mathbf{Z}_{n-1}^\top \Big(\Delta_{\mathbf{Z}_n} - \Delta_{\mathbf{Z}_{n-1}}(A^\star)^\top\Big)\|_\infty \\
&\qquad \le c_0\big(2\pi\mathcal{M}(f_Z)\big)^{1/2} \max_{j\in\{1,\ldots,(p_1+p_2)\}} \|\varepsilon_{n,\cdot j}/\sqrt{n}\| \sqrt{\frac{\log(p_1+p_2)}{n}}.
\end{aligned}$$

For T_4, it is deterministic, and satisfies

$$\begin{aligned}
\frac{1}{n}\|\Delta_{\mathbf{Z}_{n-1}}^\top \Big(\Delta_{\mathbf{Z}_n} - \Delta_{\mathbf{Z}_{n-1}}(A^\star)^\top\Big)\|_\infty &\le \|\frac{1}{n}\Delta_{\mathbf{Z}_{n-1}}^\top \Delta_{\mathbf{Z}_n}\|_\infty + \|\frac{1}{n}\Delta_{\mathbf{Z}_{n-1}}^\top \Delta_{\mathbf{Z}_{n-1}}(A^\star)^\top\|_\infty \\
&= \|\frac{1}{n}\Delta_{\mathbf{F}_{n-1}}^\top \Delta_{\mathbf{F}_n}\|_\infty + \|\frac{1}{n}\Delta_{\mathbf{F}_{n-1}}^\top \Delta_{\mathbf{F}_{n-1}}(A_{11}^\star)^\top\|_\infty
\end{aligned}$$

Combine all terms, and there exist some constant C_1, C_2, C_3 and c_1, c_2 such that with probability at least $1 - c_1\exp\big(-c_2\log(p_1+p_2)\big)$, the bound in (14) holds. ■

Appendix C. Generalization of the Main Results to Sub-exponential Tailed Error Processes: a Sketch

In this section, we provide the counterpart of Theorem 1 for the case where the underlying processes are linear with generalized sub-exponential tails. Specifically, the stable joint

VAR process $Z_t = (F_t', X_t)'$ has the following moving average representation with absolutely summable coefficients B_ℓ's (c.f. Rosenblatt (2012)):

$$Z_t = \sum_{\ell=0}^{\infty} B_\ell w_{t-\ell}.$$

In the case where the process is Gaussian, the w_t's correspond to Gaussian white noise processes. Throughout this section, we relax the Gaussian assumption and assume w_t is a white noise process whose coordinates have the following α-sub-exponential tail decay, that is, there exist two constants a, b such that the following holds:

$$\mathbb{P}\Big(|w_{tj}| \geq \xi\Big) \leq a \exp(-b\xi^\alpha), \quad \forall\ \xi > 0. \tag{39}$$

Specifically, the case of sub-Gaussian tails corresponds to $\alpha = 2$, whereas for $\alpha \in (0, 1]$ it leads to distributions with heavier tails, such as the sub-exponential distribution ($\alpha = 1$) or the Weibull distribution; see also Erdős et al. (2012); Götze et al. (2019). As a consequence, X_t and F_t deviate from being Gaussian due to the recursive data generating mechanism. Additionally, we assume the noise term of the calibration equation e_t comes from the same α-sub-exponential family.

Proposition 3 (High probability error bounds for $\widehat{\Theta}$ and $\widehat{\Gamma}$) *Suppose we are given some randomly observed snapshots $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_n\}$ obtained from the stable processes X_t and Y_t, whose dynamics are described in* (5) *and* (2)*. Assume that the same conditions as in Theorem 1 hold. Then, there exist universal positive constants $\{C_i\}$ and $\{c_i\}$ such that by solving* (6) *with regularization parameter*

$$\lambda_\Gamma = \max\Big\{C_1(2\pi\mathcal{M}(f_X) + \Lambda_{\max}(\Sigma_e))\frac{(\log p_2 + \log q)^{1/\alpha}}{\sqrt{n}},\ C_2\phi/\sqrt{nq},\ C_3\Lambda^{1/2}_{\max}(\Sigma_e)\Big\}, \tag{40}$$

the solution $(\widehat{\Theta}, \widehat{\Gamma})$ has the following bound with probability at least $1 - c_1\exp\{-c_2\big(\log(p_2q)\big)^{2/\alpha}\}$:

$$|||\Delta_\Theta/\sqrt{n}|||_F^2 + |||\Delta_\Gamma|||_F^2 \lesssim \frac{\lambda_\Gamma^2}{\mathfrak{m}(f_X)}\psi(s_{\Gamma^\star}, p_1, r), \tag{41}$$

for a sufficiently large sample size and some function $\psi(\cdot)$ that depends linearly on $s_{\Gamma^\star}, p_1$ and r.

Note that the bounds for each individual probabilistic event (e.g., RSC condition, deviation bound) differ from those in the Gaussian case, although their expressions in (41) do not exhibit marked differences compared to the Gaussian case; specifically, the bound for $|||\Delta_\Theta/\sqrt{n}|||_{\mathrm{F}}^2 + |||\Delta_\Gamma|||_{\mathrm{F}}^2$ is governed by the more stringent sample size requirement amongst its building components (i.e., concentration in the operator norm) and the slowest term in terms of probability decay.

In the rest of this section, we sketch the statements and proofs for key lemmas that underlie the high probability statements, assuming α-sub-exponential tail decay where $\alpha \in (0, 1] \cup \{2\}$. In particular, one can verify that the rates obtained below would coincide with

the Gaussian case, if $\alpha = 2$. Similar arguments can be applied to the Stage II estimate to arrive at the counterpart of Theorem 2, which are omitted.

Lemmas C.1 generalizes Hanson-Wright type concentration inequality to samples of X_t.

Lemma C.1 *Consider some generic p-dimensional linear process given in the form of $X_t := \sum_{\ell=0}^{\infty} \Phi_\ell u_{t-\ell}$, where u_t is i.i.d coming from the α-sub-exponential family defined in (39). Denote its realization by $\mathbf{X} \in \mathbb{R}^{n\times p}$ with n consecutive observations stacked in its rows. Then for a deterministic $np \times np$ matrix A, there exists some constant C such that the following bound holds:*

$$\mathbb{P}\Big(\Big|vec(\mathbf{X}^\top)^\top A\, vec(\mathbf{X}^\top) - \mathbb{E}\big[vec(\mathbf{X}^\top)^\top A\, vec(\mathbf{X}^\top)\big]\Big| > 2\pi\eta\mathcal{M}(f_X)\Big) \le \mathcal{T}(\eta, \alpha, A), \quad (42)$$

where

$$\mathcal{T}(\eta, \alpha, A) := 2\exp\Big[-C\min\Big\{\frac{\eta^2}{rk(A)|||A|||_{op}^2}, \Big(\frac{\eta}{|||A|||_{op}}\Big)^{\frac{\alpha}{2}}\Big\}\Big]. \quad (43)$$

Proof of Lemma C.1. Let $\mathrm{vec}(\mathbf{X}^\top) \overset{d}{=} \Omega^{1/2}Z$ where Ω is the covariance matrix of the np-dimensional random vector $\mathrm{vec}(\mathbf{X}^\top)$ and Z satisfies $\mathbb{E}Z = 0, \mathbb{E}(ZZ^\top) = \mathrm{I}_{np}$. Applying Götze et al. (2019, Proposition 1.1) gives

$$\begin{aligned}
&\mathbb{P}\Big(\Big|\mathrm{vec}(\mathbf{X}^\top)^\top A\,\mathrm{vec}(\mathbf{X}^\top) - \mathbb{E}\big[\mathrm{vec}(\mathbf{X}^\top)^\top A\,\mathrm{vec}(\mathbf{X}^\top)\big]\Big| > 2\pi\eta\mathcal{M}(f_X)\Big) \\
&\qquad = \mathbb{P}\Big(\Big|Z^\top\Omega^{1/2}A\Omega^{1/2}Z - \mathbb{E}\big[Z^\top\Omega^{1/2}A\Omega^{1/2}Z\big]\Big| > 2\pi\eta\mathcal{M}(f_X)\Big) \quad (44)\\
&\qquad \le 2\exp\Big\{-c_0\cdot\nu\Big(\Omega^{1/2}A\Omega^{1/2}, \alpha, 2\pi\eta\mathcal{M}(f_X)\Big)\Big\}
\end{aligned}$$

where

$$\nu(A, \alpha, t) := \min\Big\{\frac{t^2}{M^4|||A|||_{\mathrm{F}}^2}, \Big(\frac{t}{M^2|||A|||_{\mathrm{op}}}\Big)^{\alpha/2}\Big\}; \quad (45)$$

both c_0 and M are constants that depend on a, b. Next, we consider the bounds for various norms of $\Omega^{1/2}A\Omega^{1/2}$:

- $|||\Omega^{1/2}A\Omega^{1/2}|||_{\mathrm{op}} \le |||\Omega|||_{\mathrm{op}}|||A|||_{\mathrm{op}} \le 2\pi\mathcal{M}(f_X)|||A|||_{\mathrm{op}}$ where the last inequality follows from Basu and Michailidis (2015, Proposition 2.3) which applies to general linear processes;
- $|||\Omega^{1/2}A\Omega^{1/2}|||_{\mathrm{F}} \le \sqrt{\mathrm{rk}(\Omega^{1/2}A\Omega^{1/2})}|||\Omega^{1/2}A\Omega^{1/2}|||_{\mathrm{op}} \le 2\pi\sqrt{\mathrm{rk}(A)}|||A|||_{\mathrm{op}}\mathcal{M}(f_X)$;

Therefore, the last expression in (44) can be upper bounded by (43) and the claim in (42) follows. ∎

Lemma C.2 is a generalization of Proposition 2.4 in Basu and Michailidis (2015) to the case where the underlying processes come from the α-sub-exponential family.

Lemma C.2 *Consider some generic linear processes given in the form of $X_t := \sum_{\ell=0}^{\infty} \Phi_\ell u_{t-\ell}$, where u_t comes from the α-sub-exponential family. Let $\Sigma_X(0) := Cov(X_t, X_t)$. Denote its realization by $\mathbf{X} \in \mathbb{R}^{n\times p}$ and sample covariance by $S := \frac{1}{n}\mathbf{X}^\top\mathbf{X}$,respectively.*

(i) For unit vectors v_1 and v_2 satisfying $\|v_1\| \leq 1, \|v_2\| \leq 1$, the following bound holds:

$$\mathbb{P}\Big(|v_1'(S - \Sigma_X(0))v_1| > 2\pi\eta\mathcal{M}(f_X)\Big) \leq \mathcal{T}'\big(\eta, \alpha, n\big),$$

and

$$\mathbb{P}\Big(|v_1'(S - \Sigma_X(0))v_2| > 6\pi\eta\mathcal{M}(f_X)\Big) \leq 2\mathcal{T}'\big(\eta, \alpha, n\big).$$

(ii) Consider the linear process $Z_t := \sum_{\ell=0}^{\infty} \Psi_\ell w_{t-\ell} \in \mathbb{R}^q$ with w_t coming from the same family of distributions as u_t and satisfies $Cov(X_t, Z_t) = 0$; $\mathbf{Z}$ is similarly defined. Then, the following bound holds:

$$\mathbb{P}\Big(|v_1'(\mathbf{X}^\top\mathbf{Z})v_2| > 2\pi\eta(\mathcal{M}(f_Z) + \mathcal{M}(f_Z) + \mathcal{M}(f_{X,Z}))\Big) \leq 3\,\mathcal{T}'\big(\eta, \alpha, n\big),$$

where $\mathcal{M}(f_{X,Z})$ is identically defined to the quantity in Section 3.

$\mathcal{T}'$ has the following functional form:

$$\mathcal{T}'(\eta, \alpha, n) = c_1 \exp\big[-c_2 \min\{n\eta^2, (n\eta)^{\alpha/2}\}\big], \quad \textit{for some constants } c_1, c_2.$$

Proof of Lemma C.2. First we note that with $A = \mathrm{I}_n$ and the definition of $\mathcal{T}(\eta, \alpha, n)$, the following holds for some constant $C > 0$:

$$\mathcal{T}(n\eta, \alpha, A) = 2\exp\Big[-C\min\{n\eta^2, (n\eta)^{\alpha/2}\}\Big].$$

Let $y_t := v_1^\top X_t$ and $\mathbf{Y} = \mathbf{X}v_1 \in \mathbb{R}^n$ be n consecutive observations of the scalar process $\{y_t\}$, then

$$v_1'Sv_1 \overset{d}{=} \frac{1}{n}\mathbf{Y}^\top\mathbf{Y} \qquad \text{and} \qquad v_1'\Sigma_X(0)v_1 = \mathbb{E}\big[\mathbf{Y}^\top\mathbf{Y}/n\big].$$

Apply Lemma C.1 to process $\{Y_t\}$ with $A = \mathrm{I}_n$ (since moment properties are preserved under linear transformations), to obtain

$$\mathbb{P}\Big(|v_1'(S-\Sigma_X(0))v_1| > 2\pi\eta\mathcal{M}(f_Y)\Big) = \mathbb{P}\Big(\big|\mathbf{Y}^\top\mathbf{Y} - \mathbb{E}\mathbf{Y}^\top\mathbf{Y}\big| > 2\pi(n\eta)\mathcal{M}(f_Y)\Big) \leq \mathcal{T}'\big(\eta, \alpha, n\big).$$

Further, by Lemma C.6 in Sun et al. (2018), it follows that $\mathcal{M}(f_Y) \leq \|v_1\|^2\mathcal{M}(f_X) = \mathcal{M}(f_X)$; hence, the following bound holds:

$$\mathbb{P}\Big(|v_1'(S - \Sigma_X(0))v_1| > 2\pi\eta\mathcal{M}(f_X)\Big) \leq \mathcal{T}'\big(\eta, \alpha, n\big).$$

This proves the first part in (i). The rest of the proof follows along similar lines to the derivation of Proposition 2.4 in Basu and Michailidis (2015), and we give an outline without getting into too many details. For $|v_1'(S - \Sigma_X(0))v_2|$, one considers the decomposition

$$2|v_1'(S-\Sigma_X(0))v_2| \leq |v_1'(S-\Sigma_X(0))v_1| + |v_2'(S-\Sigma_X(0))v_2| + |(v_1+v_2)'(S-\Sigma_X(0))(v_1+v_2)|$$

with $\|v_1 + v_2\| \leq 2$. Repeating the steps above to each of the three terms yields the desired result.

For $|v_1'(\mathbf{X}^\top\mathbf{Z})v_2|$, let $\widetilde{y}_t = v_2^\top Z_t$; then $v_1'(\mathbf{X}^\top\mathbf{Z})v_2 = \frac{1}{n}\sum_{t=1}^n y_t\widetilde{y}_t$ and it satisfies the following decomposition

$$\begin{aligned}\frac{2}{n}\sum_{t=1}^n y_t\widetilde{y}_t &= \Big[\frac{1}{n}\sum_{t=1}^n(y_t+\widetilde{y}_t)^2 - \mathrm{Var}(y_t+\widetilde{y}_t)\Big] - \Big[\frac{1}{n}\sum_{t=1}^n y_t^2 - \mathrm{Var}(y_t)\Big] - \Big[\frac{1}{n}\sum_{t=1}^n \widetilde{y}_t^2 - \mathrm{Var}(\widetilde{y}_t)\Big] \\ &=: \big[\mathbf{G}^\top\mathbf{G} - \mathbb{E}\mathbf{G}^\top\mathbf{G}\big] - \big[\mathbf{Y}^\top\mathbf{Y} - \mathbb{E}\mathbf{Y}^\top\mathbf{Y}\big] - \big[\widetilde{\mathbf{Y}}^\top\widetilde{\mathbf{Y}} - \mathbb{E}\widetilde{\mathbf{Y}}^\top\widetilde{\mathbf{Y}}\big],\end{aligned}$$

where $\{g_t := y_t + \widetilde{y}_t\}$ is the summation process; $\mathbf{G}$ and $\widetilde{\mathbf{Y}}$ are analogously defined to $\mathbf{Y}$. Repeating the above steps to each term. Note that

$$\mathcal{M}(f_g) \le \mathcal{M}(f_Z) + \mathcal{M}(f_X) + \mathcal{M}(f_{X,Z}),$$

and this completes the proof. ■

The following lemma considers the deviation bound. Of note, to ensure the deviation bound vanishes, the sample size requirement would be $n \gtrsim (\log p + \log q)^{\frac{2}{\alpha}}$.

Lemma C.3 (high probability deviation bound) *There exist positive constants C and $c_i > 0$ such that the following deviation bound holds*

$$\|\mathbf{X}^\top\mathbf{E}/n\|_\infty \le C\cdot(\log p + \log q)^{\frac{1}{\alpha}}/\sqrt{n}$$

with probability at least

$$1 - c_1\exp\big\{-c_2\big(\log(pq)\big)^{2/\alpha}\big\},$$

for any random realizations $\mathbf{X}\in\mathbb{R}^{n\times p}$ and $\mathbf{E}\in\mathbb{R}^{n\times q}$, drawn from the linear processes $\{X_t\in\mathbb{R}^p\}$ and $\{\varepsilon_t\in\mathbb{R}^q\}$ that are constructed as linear filters of the white noise processes coming from some α-sub-exponential family.

Proof of Lemma C.3. Apply Lemma C.2, so that for any standard basis vector e_k and e_j, the following holds:

$$\mathbb{P}\Big(|e_k'(\mathbf{X}^\top\mathbf{E})e_j| > 2\pi\eta(\mathcal{M}(f_X)+\mathcal{M}(f_\varepsilon)+\mathcal{M}(f_{X,\varepsilon}))\Big) \le 3\,\mathcal{T}'\big(\eta,\alpha,n\big).$$

Taking the union bound across all pq elements, with probability at least $1-3(pq)\mathcal{T}'\big(\eta,\alpha,n\big) = 1 - 3c_1\exp\{-c_2\min\{n\eta^2,(n\eta)^{\alpha/2}\}+\log(pq)\}$, the following bound holds:

$$\|\mathbf{X}^\top\mathbf{E}/n\|_\infty \le 2\pi(\mathcal{M}(f_X)+\mathcal{M}(f_\varepsilon)+\mathcal{M}(f_{X,\varepsilon}))\cdot\eta.$$

Set $\eta := c_0(\log p+\log q)^{\frac{1}{\alpha}}/\sqrt{n}$, the desired result holds for some sufficiently large c_0 provided that $n^{\alpha/4} \gtrsim \log(pq)^{(2/\alpha-1/2)}$ (which ensures that $\min\{n\eta^2,(n\eta)^{\alpha/2}\}$). Specifically, in the context of this problem, the most stringent sample size requirement is dictated by the concentration for the operator norm (see Lemma C.5), and therefore this sample size requirement is automatically fulfilled. ■

The following lemma verifies the RSC condition.

Lemma C.4 (Verification of RSC) *Consider a snapshot of random realizations* $\mathbf{X} \in \mathbb{R}^{n\times p}$ *drawn from the linear process* $X_t := \sum_{\ell=0}^{\infty} \Phi_\ell u_{t-\ell}$ *with* u_t *coming from the* α*-sub-exponential family. Then RSC holds for* $\mathbf{X}$ *with parameter* $\alpha_{RSC} = \pi\mathfrak{m}(f_X)$ *and tolerance* $\tau := c_0\alpha_{RSC}\log p/(n^{\alpha/2})$*, with probability at least* $1 - c_1\exp\{-c_2 n^{\alpha/2}\}$.

Proof of Lemma C.4. Let $S = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$. First, suppose we have

$$\frac{1}{2}v'Sv = \frac{1}{2}v'\Big(\frac{\mathbf{X}'\mathbf{X}}{n}\Big)v \geq \frac{\alpha_{\text{RSC}}}{2}\|v\|_2^2 - \tau\|v\|_1^2, \quad \forall\ v \in \mathbb{R}^p; \tag{46}$$

then, for all $\Delta \in \mathbb{R}^{pz\times pz}$, and letting Δ_j denote its jth column, the RSC condition automatically holds since

$$\frac{1}{2T}|||\mathbf{X}\Delta|||_{\text{F}}^2 = \frac{1}{2}\sum_{j=1}^{p}\Delta_j'\big(\tfrac{\mathbf{X}'\mathbf{X}}{n}\big)\Delta_j \geq \frac{\alpha_{\text{RSC}}}{2}\sum_{j=1}^{p}\|\Delta_j\|_2^2 - \tau\sum_{j=1}^{pz}\|\Delta_j\|_1^2 \geq \frac{\alpha_{\text{RSC}}}{2}|||\Delta|||_{\text{F}}^2 - \tau\|\Delta\|_1^2.$$

Therefore, it suffices to verify that (46) holds. By Lemma C.2, $\forall v \in \mathbb{R}^p, \|v\| \leq 1$ and $\eta > 0$,

$$\mathbb{P}\Big[\big|v'\big(S - \Sigma_X(0)\big)v\big| > 2\pi\mathcal{M}(f_X)\eta\Big] \leq 2\mathcal{T}'(\eta,\alpha,n).$$

Applying the discretization argument in Basu and Michailidis (2015, Lemma F.2 & Lemma F.3), define $\mathbb{K}(2s) := \{v \in \mathbb{R}^p, \|v\| \leq 1, \|v\|_0 \leq 2s\}$, and taking the union bound in this $2s$-sparse cone gives the following inequality:

$$\begin{aligned}\mathbb{P}\Big[\sup_{v\in\mathbb{K}(2s)}\big|v'\big(S - \Sigma_X(0)\big)v\big| > 2\pi\mathcal{M}(f_X)\eta\Big] &\leq 2\cdot\min\{p^s, (21e\cdot p/s)^s\}\cdot\mathcal{T}'(\eta,\alpha,n)\\ &= 2c_1\exp\Big[-c_2\min\{n\eta^2, (n\eta)^{\alpha/2}\} + s\min\{\log p, \log(21ep/s)\}\Big].\end{aligned} \tag{47}$$

Let $\eta = \mathfrak{m}(f_X)/[54\mathcal{M}(f_X)]$, then apply results from Loh and Wainwright (2012, Lemma 12) with $\Gamma = S - \Sigma_X(0)$ and $\delta = \pi\mathfrak{m}(f_X)/27$, so that the following holds

$$\frac{1}{2}v'Sv \geq \frac{\alpha_{\text{RSC}}}{2}\|v\|^2 - \frac{\alpha_{\text{RSC}}}{2s}\|v\|_1^2, \qquad \text{where } \ \alpha_{\text{RSC}} = \pi\mathfrak{m}(f_X),$$

with probability at least $1 - 2\min\{p^s, (21e\cdot p/s)^s\}\mathcal{T}'(\eta,\alpha,n)$. By letting $s := c_0' n^{\alpha/2}/\log p$ for some small constant c_0, then τ can be expressed as $\tau = c_0\alpha_{\text{RSC}}\log p/(n^{\alpha/2})$ and the bound holds with probability at least $1 - c_1\exp\{-c_2 n^{\alpha/2}\}$. ∎

Lemma C.5 (High probability bound for $\Lambda_{\max}(S_{\mathbf{E}})$) *Consider* $\mathbf{E} \in \mathbb{R}^{n\times q}$ *whose rows are independent realizations drawn from some mean-zero* α*-sub-exponential distribution with covariance* Σ_e*. Then, the following holds for some constants* $c_i > 0$ *provided that the sample size satisfies* $n^{\alpha/2} \gtrsim q$*:*

$$\Lambda_{\max}(S_{\mathbf{E}}) \leq c_0\Lambda_{\max}(\Sigma_e),$$

with probability at least $1 - c_1\exp(-c_2 n^{\alpha/2})$.

Proof of Lemma C.5. The main arguments of the proof follow closely along the lines of those in the proof of Lemma 5, while ignoring the temporal dependence. Specifically, using similar covering arguments, with the tail decay as in Lemma C.2, there exists some constant $c_i > 0$ such that

$$\mathbb{P}\Big[\Lambda_{\max}(S_{\mathbf{E}}) \geq c_0\eta|||\Sigma_e|||_{\text{op}} + |||\Sigma_e|||_{\text{op}}\Big] \leq c_1 \exp\{-c_2 \min\{n\eta^2, (n\eta)^{\alpha/2}\} + q\log 8\}.$$

By choosing η to be a sufficiently large constant, with $n^{\alpha/2} \gtrsim q$, the statement in the lemma holds.

Remark 5 To ensure concentration of the operator norm, with the specified choice of η, the sample size requirement in (C.5) is more stringent than that of the Gaussian case. In particular, for the case of sub-exponential tails with $\alpha = 1$, this would imply a sample size requirement $\sqrt{n} \gtrsim q$. If however, the elements of the random noise vector e_t's are bounded, that is, $\|e_t\|_2 \leq \sqrt{C}$ almost surely for some $C > 0$, one can directly apply the matrix Bernstein inequality to obtain the following bound (Wainwright, 2019, Corollary 6.20):

$$\mathbb{P}\Big[\Big|\Lambda_{\max}(S_{\mathbf{E}}) - |||\Sigma_e|||_{\text{op}}\Big| \geq \eta\Big] \leq 2q \exp\Big\{\frac{-n\eta^2}{2C(|||\Sigma_e|||_{\text{op}} + \eta)}\Big\}.$$

Depending on how C grows with q, the sample size requirement could potentially be more relaxed to attain concentration. ■

Appendix D. Additional Numerical Studies

In this section, we investigate selected scenarios where the relaxed implementation on estimating the calibration equation may fail to produce good estimates, due to the absence of the compactness constraint. For illustration purposes, it suffices to consider the setting where X_t and F_t jointly follow a multivariate Gaussian distribution and are independent and identically distributed across samples. Throughout, we set $n = 200, p_1 = 5, p_2 = 50, q = 100$, and $\binom{X_t}{F_t} \sim \mathcal{N}(0, \Sigma)$ with $\Sigma_{ij} = 0.25$ $(i \neq j)$ and $\Sigma_{ii} = 1$. The noise level is fixed at $\sigma_e = 1$.

First, we note that based on the performance evaluation shown in Section 4, the estimates demonstrate good performance even without the compactness constraint. The simulation settings are characterized by adequate sparsity in Γ, which in turn limits the size of the equivalence class $\mathcal{C}(Q_2)$ as mentioned in Section 2.1. Therefore, we focus on the following two issues: (i) whether sparsity encourages additional "approximate identification"; and (ii) whether a good initializer helps constrain estimates from subsequent iterations to a ball around the true value.

We start by considering a non-sparse Γ. Specifically, for both Λ and Γ, their entries are generated from $\mathsf{Unif}\{(-1.5, -1.2) \cup (1.2, 1.5)\}$. Additionally, we specify one alternative model in $\mathcal{C}(Q_2)$ by setting $Q_2 = \mathbf{5}_{p_1 \times p_2}$, which will generate the corresponding $\check{\mathbf{F}}$, $\check{\Theta}$ and $\check{\Gamma}$. Table 6 depicts the performance of the estimated Θ based on different initializers:

The results in Table 6 show that the algorithm converges (if at all) to different local optima whose values may deviate markedly for the true ones. Specifically, initializer $\Theta^\star +$

Table 6: Performance evaluation of $\widehat{\Theta}$ obtained from different initializers under a non-sparse setting.

initializer $\widehat{\Theta}^{(0)}$	$\Theta^\star$	$\mathbf{0}_{n\times q}$	$\Theta^\star + 0.1 * \mathbf{Z}_{n\times q}$	$\check{\Theta}$
Rel.Err	0.09	0.63	fail to converge within 5000 iterations	1.82 (0.02, relative to $\check{\Theta}$)

$0.1 * \mathbf{Z}_{n\times q}$, where each entry $\Theta^\star$ is perturbed by an iid standard Gaussian random variable scaled by 0.1, fails to converge. Note that the perturbation is small, but the operator norm of the initializer far exceeds ϕ_0. Initializer $\check{\Theta}$ yields an estimate that is far from the true data-generating factor hyperplane, yet close to its observationally equivalent one. This suggests that in non-sparse settings, without imposing the compactness constraint on the equivalence class, a good initializer is required for the actual relaxed implementation to produce a fairly good estimate of the true data generating parameters.

However, this is not the case if there is sufficient sparsity in Γ. Specifically, using the same generating mechanism for Λ and Γ as in Section 4, we found that even with different initializers, the algorithm always produces estimates that are close to each other and also exhibit good performance. This finding strongly suggests that sparsity in Γ effectively shrinks the size of the equivalence class and the algorithm after a few iterations produces updates that are close to each other, irrespective of the initializer employed. Hence, the effective equivalence class is constrained to the one whose elements are encoded by $\check{\Gamma}$ that have similar characteristics in terms of the location of the non-zero parameters to Γ.

Finally, we consider a case that lies between the above two settings, that is, there is a structured sparsity pattern in Γ. Specifically, we set the last 5 columns of Γ to be dense while the remaining ones are sparse. The overall density level of Γ is fixed at 10%. Note that in this case, the size of the corresponding equivalence class is much larger to the one corresponding to a Γ with 10% uniformly distributed non-zeros entries, due to the presence of the five dense columns.

Table 7: Performance evaluation for $\widehat{\Theta}$ with different initializers under structured sparsity.

initializer $\widehat{\Theta}^{(0)}$	$\Theta^\star$	$\mathbf{0}_{n\times q}$	$\Theta^\star + 0.1 * \mathbf{Z}_{n\times q}$	$\mathbf{20}_{n\times q}$
Rel.Err	0.65	0.65	0.65	0.68

As the results in Table 7 indicate, when the initializer starts to deviate from the true value, there exist initializers that would yield inferior estimates.

In summary, in a non-sparse setting without compactification of the equivalence class, different initializers yield drastically different estimates that are not close enough to the true data-generating model, as expected by the approximate (IR+) condition employed. The problem is largely mitigated for sufficiently sparse Γ, which leads to shrinking the equivalence class. However, an exact characterization of the equivalence class is hard to obtain in practice, since the location of the non-zero entries in Γ is unknown.

Appendix E. List of Commodities and Macroeconomic Variables

Table 8: List of commodities considered in this study. Data source: International Monetary Fund.

Commodity	Key	Description
ALUMINUM	PALUM	Aluminum, 99.5% minimum purity, LME spot price
COCOA	PCOCO	Cocoa beans, International Cocoa Organization cash price
COFFEE	PCOFFOTM	Coffee, Other Mild Arabicas, International Coffee Organization New York cash price
COPPER	PCOPP	Copper, grade A cathode, LME spot price
COTTON	PCOTTIND	Cotton, Cotton Outlook 'A Index', Middling 1-3/32 inch staple
LEAD	PLEAD	Lead, 99.97% pure, LME spot price
MAIZE	PMAIZMT	Maize (corn), U.S. No.2 Yellow, FOB Gulf of Mexico, U.S. price
NICKEL	PNICK	Nickel, melting grade, LME spot price
OIL	POILAPSP	Crude Oil (petroleum), simple average of three spot prices
RICE	PRICENPQ	Rice, 5 percent broken milled white rice, Thailand nominal price quote
RUBBER	PRUBB	Rubber, Singapore Commodity Exchange, No. 3 Rubber Smoked Sheets, 1st contract
SOYBEANS	PSOYB	Soybeans, U.S. soybeans, Chicago Soybean futures contract (first contract forward)
SUGAR	PSUGAUSA	Sugar, U.S. import price, contract no.14 nearest futures position
TIN	PTIN	Tin, standard grade, LME spot price
WHEAT	PWHEAMT	Wheat, No.1 Hard Red Winter, ordinary protein
ZINC	PZINC	Zinc, high grade 98% pure

Name	Description	tCode	Category	Region
IPI_US	IP Index: total	5	Output & Income	US
CUM_US	Capacity Utilization: manufacturing	2	Output & Income	US
UNEMP_US	Civilian unemployment rate: all	2	Labor Market	US
HOUST_US	Housing Starts: ttl new privately owned	4	Housing	US
ISR_US	Total Business: inventories to sales ratio	2	Consumption	US
M2_US	M2 Money Stock	6	Money & Credit	US
BUSLN_US	Commericial and industrial loans	6	Money & Credit	US
REALN_US	Real estate loans at all commercial banks	6	Money & Credit	US
FFR_US	Effective federal funds rate	2	Interest & Exchange Rates	US
TB10Y_US	10-year treasury rate	2	Interest & Exchange Rates	US
BAA_US	Moody's Baa corporate bond yield	2	Interest & Exchange Rates	US
USDI_US	Trade weighted U.S.dollar index	5	Interest & Exchange Rates	US
CPI_US	CPI: all iterms	5	Prices	US
PCEPI_US	Personal Consumption Expenditure: chain index	5	Prices	US
SP500_US	S&P's Common Stock Price Index: composite	5	Stock Market	US
CPI_EU	Consumer Price Indices, percent change	2	Prices	EU
IPI_EU	Industrial Production Index: total industry (excluding construction)	5	Output & Income	EU
IPICP_EU	Industrial Production Index: construction	5	Output & Income	EU
M3_EU	Monetary aggregate M3	6	Money & Credit	EU
LOANRES_EU	Credit to resident sectors, non-MFI excluding gov	6	Money & Credit	EU
LOANGOV_EU	Credit to general government sector	6	Money & Credit	EU
PPI_EU	Producer Price Index: total industry (excluding construction)	6	Prices	EU
UNEMP_EU	Unemployment rate: total	2	Labor Market	EU
IMPORT_EU	Total trade: import value	6	Trade	EU
EXPORT_EU	Total trade: export value	6	Trade	EU
EB1Y_EU	Euribor 1 year	2	Interest & Exchange Rates	EU
TB10Y_EU	10-year government benchmark bond yield	2	Interest & Exchange Rates	EU
EFFEXR_EU	ECB nominal effective exchange rate againt group of trading partners	2	Interest & Exchange Rates	EU
EUROSTOXX50_EU	Euro STOXX composite index	5	Stock Market	EU
IOP_UK	Index of Production	5	Output & Income	UK
CPI_UK	CPI Index	5	Prices	UK
PPI_UK	Output of manufactured products	5	Prices	UK
UNEMP_UK	Unemployment rate: aged 16 and over	2	Labor Market	UK
EFFEXR_UK	Effective exchange rate index, Sterling	2	Interest & Exchange Rates	UK
TB10Y_UK	10-year British government stock, nominal par yield	2	Interest & Exchange Rates	UK
LIBOR6M_UK	6 month interbank lending rate, month end	2	Interest & Exchange Rates	UK
M3_UK	Monetary aggregate M3	6	Money & Credit	UK
CPI_CN	CPI: all iterms	5	Prices	CN

PPI_CN	Producer price index for industrial products (same month last year = 100)	2	Prices	CN
M2_CN	Monetary aggregate M2	6	Money & Credit	CN
EFFEXR_CN	Real broad effective exchange rate	2	Interest & Exchange Rates	CN
EXPORT_CN	Value goods	6	Trade	CN
IMPORT_CN	Value goods	6	Trade	CN
INDGR_CN	Growth rate of industrial value added (last year = 100)	2	Output & Income	CN
SHANGHAI_CN	Shanghai Composite Index	5	Stock Market	CN
TB10Y_JP	10-year government benchmark bond yield	2	Interest & Exchange Rates	JP
EFFEXR_JP	Real broad effective exchange rate	2	Interest & Exchange Rates	JP
CPI_JP	CPI Index: all items	5	Prices	JP
M2_JP	Monetary aggregate M2	6	Money & Credit	JP
UNEMP_JP	Unemployment rate: aged 15-64	2	Labor Market	JP
IPI_JP	Production of Total Industry	5	Output & Income	JP
IMPORT_JP	Import price index: all commodities	6	Trade	JP
EXPORT_JP	Value goods	6	Trade	JP
NIKKEI225_JP	NIKKEI 225 composite index	5	Stock Market	JP

Table 9: List of macroeconomic variables in this study.

Data source: Fred St.Louis, ECB Statistical Data Warehouse, UK Office for National Statistics, Bank of England, National Bureau of Statistics of China, YAHOO!. tCode: 1: none; 2: ΔX_t; 3: $\Delta^2 X_t$; 4: $\log X_t$; 5: $\Delta \log X_t$; 6: $\Delta^2 \log X_t$; 7: $\Delta(X_t/X_{t-1} - 1)$.

References

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

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

Theodore Wilbur Anderson. *An Introduction to Multivariate Statistical Analysis*, volume 2. Wiley New York, 1958.

Tomohiro Ando and Jushan Bai. Selecting the regularization parameters in high-dimensional panel data models: Consistency and efficiency. *Econometric Reviews*, 37 (3):183–211, 2018.

Jushan Bai and Serena Ng. Large dimensional factor analysis. *Foundations and Trends® in Econometrics*, 3(2):89–163, 2008.

Jushan Bai, Kunpeng Li, and Lina Lu. Estimation and inference of favar models. *Journal of Business & Economic Statistics*, 34(4):620–641, 2016.

Marta Bańbura, Domenico Giannone, and Lucrezia Reichlin. Large Bayesian vector auto regressions. *Journal of Applied Econometrics*, 25(1):71–92, 2010.

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

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.

Giovanni Caggiano, Efrem Castelnuovo, and Nicolas Groshenny. Uncertainty shocks and unemployment dynamics in us recessions. *Journal of Monetary Economics*, 67:78–92, 2014.

Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

Kamal C Chanda et al. Asymptotic properties of estimators for autoregressive models with errors in variables. *The Annals of Statistics*, 24(1):423–430, 1996.

Venkat Chandrasekaran, Pablo A Parrilo, and Alan S Willsky. Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1935–1967, 2012.

Sandra Eickmeier, Leonardo Gambacorta, and Boris Hofmann. Understanding global liquidity. *European Economic Review*, 68:1–18, 2014.

László Erdős, Horng-Tzer Yau, and Jun Yin. Bulk universality for generalized Wigner matrices. *Probability Theory and Related Fields*, 154:341–407, 2012.

Jeffrey A Frankel. The effect of monetary policy on real commodity prices. Technical report, National Bureau of Economic Research, 2006.

Jeffrey A Frankel. Effects of speculation and interest rates in a carry trade model of commodity prices. *Journal of International Money and Finance*, 42:88–112, 2014.

Friedrich Götze, Holger Sambale, and Arthur Sinulis. Concentration inequalities for polynomials in α-sub-exponential random variables. *arXiv preprint arXiv:1903.05964*, 2019.

Suriya Gunasekar, Arindam Banerjee, and Joydeep Ghosh. Unified view of matrix completion under general structural constraints. In *Advances in Neural Information Processing Systems*, pages 1180–1188, 2015.

Eric C Hall, Garvesh Raskutti, and Rebecca Willett. Learning high-dimensional generalized linear autoregressive models. *IEEE Transactions on Information Theory*, 65(4):2401–2422, 2019.

Ivana Komunjer and Serena Ng. Measurement errors in dynamic models. *Econometric Theory*, 30(1):150–175, 2014.

Jiahe Lin and George Michailidis. Regularized estimation and testing for high-dimensional multi-block vector-autoregressive models. *Journal of Machine Learning Research*, 18 (117):1–49, 2017.

Robert B Litterman. Forecasting with Bayesian vector autoregressions—five years of experience. *Journal of Business & Economic Statistics*, 4(1):25–38, 1986.

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.

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

Helmut Lütkepohl. Structural vector autoregressive analysis in a data rich environment. Technical report, Deutsches Institut für Wirtschaftsforschung, 2014.

Igor Melnyk and Arindam Banerjee. Estimating structured vector autoregressive models. In *International Conference on Machine Learning*, pages 830–839, 2016.

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 and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: optimal bounds with noise. *Journal of Machine Learning Research*, 13(53):1665–1697, 2012.

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.

William B Nicholson, David S Matteson, and Jacob Bien. VARX-L: Structured regularization for large vector autoregressions with exogenous variables. *International Journal of Forecasting*, 33(3):627–651, 2017.

Murray Rosenblatt. *Stationary Sequences and Random Fields*. Springer Science & Business Media, 2012.

Anil K Seth, Adam B Barrett, and Lionel Barnett. Granger causality analysis in neuroscience and neuroimaging. *Journal of Neuroscience*, 35(8):3293–3297, 2015.

Ali Shojaie and George Michailidis. Discovering graphical granger causality using the truncating lasso penalty. *Bioinformatics*, 26(18):i517–i523, 2010.

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

James H Stock and Mark W Watson. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179, 2002.

James H Stock and Mark W Watson. Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. In John B Taylor and Harald Uhlig, editors, *Handbook of Macroeconomics*, volume 2A, chapter 8, pages 415–525. Elsevier, 2016.

James H Stock and Mark W Watson. Twenty years of time series econometrics in ten pictures. *Journal of Economic Perspectives*, 31(2):59–86, 2017.

Yiming Sun, Yige Li, Amy Kuceyeski, and Sumanta Basu. Large spectral density matrix estimation by thresholding. *arXiv preprint arXiv:1812.00532*, 2018.

Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.

Martin J Wainwright. *High-dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.