

# Regularized Estimation of High-dimensional Factor-Augmented Autoregressive (FAVAR) Models

Jiahe Lin\*    George Michailidis†

## 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 very 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 successfully 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.

Key words: 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<sup>1</sup> 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.

---

\*Ph.D. Candidate, Department of Statistics, University of Michigan.

†Corresponding Author. Professor, Department of Statistics and the Informatics Institute, University of Florida.

<sup>1</sup>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

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 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. (2016) 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), 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) came up with a compromise model, called 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 include those in a standard VAR. 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

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

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  $F_t$ :

$$Y_t = \Lambda F_t + \Gamma X_t + e_t. \quad (2)$$

Hence, 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).

Note that there is very limited theoretical work (e.g. 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 ( $p_1 + p_2$ ) assumption is rather restrictive in many applications as discussed next. The model has been extensively used in empirical work in economics and finance (e.g. Eickmeier et al., 2014; Caggiano et al., 2014), where customarily a very small size block  $X_t$  is considered. For example, in the paper that introduced the FAVAR model Bernanke et al. (2005)  $X_t$  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 model considered is augmented by one factor summarizing the macroeconomic indicators and its dependence over time involves 7 lags, thus increasing the sample size requirement for its estimation. In a recent application, Stock and Watson (2016) applies the FAVAR model to macroeconomics effects of oil supply shocks, the VAR model comprises 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 a growing need for large scale FAVAR models and this papers aims to examine their estimation in high-dimensions, leveraging sparsity constraints on key model parameters.

The key contributions of this paper are the investigation of the theoretical properties of estimates of the FAVAR model parameters under high-dimensional scaling, together with the introduction of an identifiability constraint compatible with the high-dimensional nature of the model. At the technical level there are two sets of challenges that are successfully resolved: (i) the calibration equation involves both an observed set of predictor variables and a set of latent factors, and their interactions require careful handling to enable accurate estimation of the factors, which is crucial for estimating the transition matrix since they constitute part of the input to the VAR system; and (ii) the presence of a block of variables in the VAR system that are subject to error due to it being an estimated quantity introduces a number of technical challenges, which are compounded by the presence of temporal dependence.

**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  $\|A\|$  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  $\|A\|_F$  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<sup>2</sup> 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 sample size. Further, we assume that the transition matrices  $\{A^{(i)}\}_{i=1}^d$  and the regression coefficient matrix  $\Gamma$  are *sparse*. Finally, the factor loading matrix  $\Lambda$  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: for any invertible matrix  $Q_1 \in \mathbb{R}^{p_1 \times p_1}$  and  $Q_2 \in \mathbb{R}^{p_1 \times p_2}$ ,

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

where

$$\tilde{\Lambda} := \Lambda Q_1, \quad \tilde{F}_t := Q_1^{-1} F_t - Q_1^{-1} Q_2 X_t, \quad \tilde{\Gamma} := \Gamma + \Lambda Q_2. \quad (3)$$

Hence, the key model parameters  $(\Lambda, \Gamma)$  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 indexed 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  $\Lambda$ ,  $\Gamma$  and  $F_t$ .

Various schemes have been proposed in the literature to address this issue. Specifically, [Bernanke et al. \(2005\)](#) imposes the necessary restrictions through the coefficient matrices of the calibration equation, requiring  $\Lambda = \begin{bmatrix} I_{p_1} \\ * \end{bmatrix}$  and  $\Gamma_{[1:p_1], \cdot} = 0$ ; that is, the upper  $p_1 \times p_1$  block of  $\Lambda$  is set to the identity matrix and the first  $p_1$  rows of  $\Gamma$  to zero. [Bai et al. \(2016\)](#) considers different sets of restrictions (respectively labeled as IRa, IRb and IRc), all involving parameters from both the calibration and the VAR equations; in particular,  $p_1 p_2$  of the total restrictions required are imposed through  $\text{Cov}(w_t^X, w_t^F) = O$  and the remaining  $p_1^2$  ones are imposed in an analogous fashion to those in classical factor analysis.

In the low-dimensional setting ( $p_2$  fixed), one can proceed to estimate the parameters subject to these restrictions. For example, [Bernanke et al. \(2005\)](#) uses a single-step Bayesian likelihood

<sup>2</sup>We do not impose the restriction that  $p_2$  is smaller than the available sample size.

approach that fully incorporates their proposed identifiability restrictions, yet is computationally intensive. The procedure in Bai et al. (2016) requires the projection onto the orthogonal space spanned by samples of  $X_t$  as the very first step and the inverse matrix associated with of the sample covariance of  $w_X^t$  for further rotation. However, in high-dimensional settings, the growing dimension  $p_2$  of the observed block  $X_t$  will further exacerbate the computational inefficiency of the aforementioned Bayesian approach. Further, neither the projection step nor the matrix inversion one are possible, which automatically renders the estimation procedure proposed in Bai et al. (2016) infeasible<sup>3</sup>.

Next, we introduce an alternative identification scheme (**IR+**) that is compatible with the model specification and can also be seamlessly incorporated in the estimation procedure. First, we require:

(**IR**)  $\Lambda = \begin{bmatrix} I_{p_1} \\ * \end{bmatrix}$ : the upper  $p_1 \times p_1$  block of  $\Lambda$  is an identity matrix, while the bottom block is left unconstrained.

Note that (IR) only involves  $p_1^2$  constraints and yields uniquely identifiable  $\Lambda$  and  $F$ , for any given product  $\Lambda F_t$ . Note that the latent factor under (IR) remains completely unrestricted which is desirable given its use in the VAR system. The (IR) constraint corresponds to a commonly employed identifiability scheme in classical factor analysis (e.g. Bai and Ng, 2013). Specifically, with (IR), 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

$$\check{F}_t = F_t - Q_2 X_t \quad \check{\Gamma} := \Gamma + \Lambda Q_2. \quad (4)$$

All such models encoded by  $(\check{F}_t, \check{\Gamma})$ , form an equivalence class indexed by  $Q_2$  that specifies the transformation. We denote this equivalence class by  $\mathcal{C}(Q_2)$  and the magnitude of  $Q_2$  can be interpreted as a rough measure of discordance between the true data-generating model encoded by  $(F_t, \Gamma)$  and those encoded by  $(\check{F}_t, \check{\Gamma})$ . In particular, such discordance becomes zero when  $Q_2 = O$  and  $\mathcal{C}(Q_2)$  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. For example, as previously mentioned, Bernanke et al. (2005) impose the restrictions through  $\Gamma$  by constraining its first  $p_1$  rows to be equal to zero. Nevertheless, from a model perspective it translates to expressing the first  $p_1$  coordinates of  $Y_t$  as noisy versions of  $F_t$ , which in turn makes it difficult to appropriately choose those coordinates in applications. Bai et al. (2016) requires  $\text{Cov}(w_t^X, w_t^F) = O$  which resolves the identifiability issue at the population level, but this constraint can not be operationalized in the high-dimensional setting as explained above.

An applicable constraint to high-dimensional settings is given by  $\text{Cov}(F_t, X_t) = O$  which yields the necessary  $p_1 p_2$  restrictions. Yet, it 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, we first 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

<sup>3</sup>For a full account of the estimation procedure in Bai et al. (2016) and why it fails to go through in high dimensions, see Appendix D

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

$$\mathcal{S}(\check{\Theta}) := \{\check{\Theta} := \check{\mathbf{F}}\Lambda^\top \mid \check{\mathbf{F}} \text{ are samples of } \check{F}_t \text{ defined through (4)}\}.$$

Further, let  $\Theta$  (without the check) 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  $\Theta \in \mathcal{S}(\check{\Theta})$  and  $\check{\Theta}$  coincides with  $\Theta$  when  $Q_2 = 0$ . In addition, we require that all elements in  $\mathcal{S}(\check{\Theta})$  to satisfy the following constraint:

**(Compactness)**  $\|\check{\Theta}/\sqrt{n}\|_{\text{op}} \leq \phi$ , that is, the largest singular value of  $(\check{\Theta}/\sqrt{n})$  does not exceed a pre-specified value  $\phi$ .

(Compactness) limits the spikiness of all possible  $\check{\Theta}$ 's by imposing a *box constraint* on their eigen-spectra, and restricts the factor hyperplane set induced by  $\mathcal{C}(Q_2)$  to its  $\phi$ -radius subset  $\mathcal{S}_\phi(\check{\Theta})$ , where

$$\mathcal{S}_\phi(\check{\Theta}) := \{\|\check{\Theta}/\sqrt{n}\|_{\text{op}} \leq \phi \mid \check{\Theta} \in \mathcal{S}(\check{\Theta})\}.$$

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. Since  $\Theta \in \mathcal{S}(\check{\Theta})$ ,  $\phi \geq \phi_0 := \|\Theta/\sqrt{n}\|_{\text{op}}$ . The  $\sqrt{n}$  factor is introduced to reflect proper scaling with respect to the available number of samples. Note that this constraint also indirectly limits the magnitude of  $Q_2$ , since by singular value inequalities<sup>4</sup> and (4), we get

$$\|\mathbf{X}Q_2^\top \Lambda^\top / \sqrt{n}\|_{\text{op}} - \|\Theta/\sqrt{n}\|_{\text{op}} \leq \|\check{\Theta}/\sqrt{n}\|_{\text{op}} \leq \phi.$$

The above gives that

$$\|Q_2\|_{\text{op}} \leq \frac{\phi + \phi_0}{\sigma_{\min}(\mathbf{X}/\sqrt{n})\sigma_{\min}(\Lambda)}, \quad (5)$$

where  $\sigma_{\min}$  denotes the smallest nonzero singular value that comes from the reduced SVD of the corresponding matrix. Even though the bound in (5) may not be the tightest, it nevertheless imposes an effective constraint on  $Q_2$ , since it no longer allows  $Q_2$  to take arbitrary values in the set of  $p_1 \times p_2$  matrices. Consequently, the size of the equivalence class  $\mathcal{C}(Q_2)$  is also limited, which implies that although the models encoded by  $(F_t, \Gamma)$  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.

In summary, our proposed identification scheme (IR+) entails 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. Hence, (IR+) can be viewed as an *approximate identification* scheme of the true data generating model.

Thus, for estimation purposes henceforth, it becomes adequate to focus on this restricted equivalence class, rather than its individual elements. The (IR+) constraint 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).

<sup>4</sup>For two generic matrices  $A$  and  $B$  of commensurate dimensions, let  $\sigma_1 \geq \sigma_2 \geq \dots$  denote their singular values in decreasing order, then the following inequality holds:  $\sigma_i(A+B) \geq \sigma_i(A) - \sigma_1(B)$ . This can be derived from Theorem 3.4.1 in Horn and Johnson (1990).

*Remark 1.* It is worth pointing out that the sparsity requirement on  $\Gamma$  further limits the size of the equivalence class  $\mathcal{C}(Q_2)$ . To see this, note that for an arbitrary element in  $\mathcal{C}(Q_2)$ ,  $\check{\Gamma}$  satisfies  $\check{\Gamma} = \Gamma + \Lambda Q_2$ . In order for  $\Gamma$  and  $\check{\Gamma}$  to have the same support,  $Q_2$  needs to be further restricted. However, since the support set of  $\Gamma$  is unknown, this further implicit restriction on structural equivalence can not be enforced or verified, and the effective equivalence class can not be characterized through the support set either.

## 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  $Z_t = AZ_{t-1} + W_t$ :

$$\begin{bmatrix} F_t \\ X_t \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} F_{t-1} \\ X_{t-1} \end{bmatrix} + \begin{bmatrix} w_t^F \\ w_t^X \end{bmatrix}. \quad (6)$$

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  $\mathcal{A}_d(L) := I - A^{(1)}L - \dots - A^{(d)}L^d$ , it can always be written in the form of a VAR(1) model for some  $dp$ -dimensional process  $\tilde{Z}_t$  (see Lütkepohl, 2005, for details).

Based on the introduced model identification scheme (IR+), we propose the following procedure to estimate the FAVAR model with a sparse coefficient matrix  $\Gamma$  and a dense loading matrix  $\Lambda$ , together with 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  $\mathbf{X}_{n-1} := [x_1, \dots, x_{n-1}]^\top$  denote the predictor matrix and  $\mathbf{X}_n := [x_2, \dots, x_n]^\top$  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 (6) 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  $\Lambda, \Gamma$  and subsequently the transition matrices  $\{A_{ij}\}_{i,j=1,2}$ .

- Stage I: estimation of the calibration equation under (IR+). We formulate the following *constrained optimization* problem using a least squares loss function and incorporating the sparsity-induced  $\ell_1$  regularization of the sparse block  $\Gamma$ , the rank constraint on the hyperplane  $\Theta$ , and (Compactness):

$$\begin{aligned} (\hat{\Theta}, \hat{\Gamma}) &:= \arg \min_{\Theta \in \mathbb{R}^{n \times q}, \Gamma \in \mathbb{R}^{q \times p_2}} \left\{ \frac{1}{2n} \|\mathbf{Y} - \Theta - \mathbf{X}\Gamma^\top\|_{\text{F}}^2 + \lambda_\Gamma \|\Gamma\|_1 \right\}, \\ &\text{subject to} \quad \text{rank}(\Theta) \leq r, \quad \|\Theta/\sqrt{n}\|_{\text{op}} \leq \phi. \end{aligned} \quad (7)$$

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

$$\hat{\mathbf{F}} = \hat{\mathbf{F}}^{\text{PC}} (\hat{\Lambda}_1^{\text{PC}})^\top, \quad \hat{\Lambda} = \hat{\Lambda}^{\text{PC}} (\hat{\Lambda}_1^{\text{PC}})^{-1}, \quad (8)$$

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

- 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

$$\widehat{A} := \arg \min_{A \in \mathbb{R}^{(p_1+p_2) \times (p_1+p_2)}} \left\{ \frac{1}{2n} \|\widehat{\mathbf{Z}}_n - \widehat{\mathbf{Z}}_{n-1} A^\top\|_F^2 + \lambda_A \|A\|_1 \right\}, \quad (9)$$

where  $\widehat{\mathbf{Z}}_n := [\widehat{\mathbf{F}}_n, \mathbf{X}_n]$  and  $\widehat{\mathbf{Z}}_{n-1}$  is analogously defined. The  $\ell_1$ -norm penalty induces sparsity on  $A$  according to the model assumption.

In principle, there may be additional contemporaneous dependence amongst the coordinates of the error processes  $e_t, w_t^X, w_t^F$ , respectively. In that case, one has to make additional assumptions on the structure of the inverses of covariance matrices  $\Sigma_e, \Sigma_w^X$  and  $\Sigma_w^F$  (e.g. sparsity) and modify the loss function accordingly. The complete estimation procedure for a VAR system whose error process exhibits contemporaneous dependence is discussed in detail in [Lin and Michailidis \(2017\)](#) and an analogous strategy can be adopted for this model. 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 (9) 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  $\Theta$  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+) 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 of an arbitrary  $\check{\Theta}$  ( $\check{\Theta} \in \mathcal{C}(Q_2)$ ) is close to the true data generating  $\Theta$ . 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 the 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. \(2012\)](#) studies 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. \(2012\)](#) studies 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 compactness 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+) 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+) 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 the formulation (7) and (9), whose global optima correspond to  $(\hat{\Theta}, \hat{\Gamma})$  and  $\hat{A}$ , respectively.

Since (9) relies not only on prime observable quantities (namely  $X_t$ ), but also on estimated quantities from Stage I (namely  $\hat{F}_t$ ), the analysis requires a careful examination of how the estimation error in the factor propagates to that for  $A$ . 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  $\hat{\Gamma}$ ,  $\hat{\Theta}$ <sup>5</sup> and  $\hat{A}$  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  $A$ , the error of the plug-in estimate  $\hat{\mathbf{F}}$  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.

Throughout, we use superscript  $\star$  to denote the true value of the parameters of interest, and  $\Delta$  for errors of the estimators; e.g.,  $\Delta_A = \hat{A} - A^\star$ . All proofs are deferred to the relevant Appendices.

**A road map for establishing the 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 error; 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 (9) 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, provide the high-probability error bounds for  $\hat{\Theta}$  and  $\hat{\Gamma}$ . The high-probability error bound of the estimated factors is subsequently established, 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,

---

<sup>5</sup>Consequently, the error bounds of  $\hat{\mathbf{F}}$  and  $\hat{\Lambda}$  under IR are also obtained.

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  $\Phi$  with curvature  $\alpha_{\text{RSC}} > 0$  and tolerance  $\tau_n \geq 0$ , if

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

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

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

$$\hat{A} := \min_A \left\{ \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}A^\top\|_{\text{F}}^2 + \lambda_A \|A\|_1 \right\},$$

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  $\lambda_A$  to be selected in accordance with the deviation condition, if

$$\lambda_A \geq c_0 \|\mathcal{H}_A - \mathcal{G}_A(A^*)^\top\|_\infty.$$

**Definition 3** (Spectrum and its extremes). For a  $p$ -dimensional stationary process  $X_t$ , its spectral density  $f_X(\omega)$  is defined as  $f_X(\omega) := \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \Sigma_X(h) e^{i\omega h}$ , where  $\Sigma_X(h) := \mathbb{E}(X_t X_{t+h}^\top)$ . Its upper and lower extremes are defined as

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

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

$$\mathcal{M}(f_{X,Y}) := \text{ess sup}_{\omega \in [-\pi, \pi]} \sqrt{\Lambda_{\max}(f_{X,Y}^*(\omega) f_{X,Y}(\omega))}, \quad \text{where } * \text{ denotes the conjugate transpose.}$$

Additionally, we let  $S_{\mathbf{X}} := \frac{1}{n} \mathbf{X}^\top \mathbf{X}$  denote the sample covariance matrix, and similar quantities (e.g.,  $S_{\mathbf{E}}$ ) are analogously defined. Denote the density level of  $\Gamma^*$  by  $s_{\Gamma^*} := \|\Gamma^*\|_0$ , and that of  $A^*$  by  $s_{A^*}$ .

We start by providing error bounds for  $\hat{\Gamma}$  and  $\hat{\Theta}$ , as well as those of the corresponding  $\hat{\mathbf{F}}$  and  $\hat{\Lambda}$  extracted under IR. For the optimization problem given in (7), we assume that  $r \geq p_1$  and  $\phi$  is always compatible with the true data generating mechanism, so that  $\Theta^*$  is always feasible.

The error bounds of  $\widehat{\Theta}$  and  $\widehat{\Gamma}$  for deterministic realizations rely on: (i)  $\mathbf{X}$  satisfying the RSC condition with curvature  $\alpha_{\text{RSC}}^{\mathbf{X}}$ ; and (ii) the tuning parameter  $\lambda_{\Gamma}$  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

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

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 (9), 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

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

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 strength. Moreover, the deviation bound is of the form  $h_2(\Delta_{\mathbf{F}})$ , which can be viewed as another function of the error<sup>6</sup>. 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 (7), 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  $\Delta_{\Theta}$  and  $\Delta_{\Gamma}$  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  $\alpha_{\text{RSC}}^{\mathbf{X}} > 0$  and a tolerance  $\tau_{\mathbf{X}}$  for which*

$$\tau_{\mathbf{X}} \cdot (p_1 + r + 4s_{\Gamma^*}) < \min\{\alpha_{\text{RSC}}^{\mathbf{X}}, 1\}/16.$$

*Then, for any matrix pair  $(\Theta^*, \Gamma^*)$  satisfying  $\|\Theta^*/\sqrt{n}\|_{op} \leq \phi$  that generates  $Y_t$ , for estimators  $(\widehat{\Theta}, \widehat{\Gamma})$  obtained by solving (7) with regularization parameters  $\lambda_{\Gamma}$  satisfying*

$$\lambda_{\Gamma} \geq \max\{2\|\mathbf{X}^{\top} \mathbf{E}/n\|_{\infty}, \Lambda_{\max}^{1/2}(S_{\mathbf{E}}), (p_1 + r)\phi\Lambda_{\max}^{1/2}(S_{\mathbf{X}})\},$$

*the following bound holds:*

$$\|\Delta_{\Gamma}\|_{\mathbf{F}}^2 + \|\Delta_{\Theta}/\sqrt{n}\|_{\mathbf{F}}^2 \leq \frac{16\lambda_{\Gamma}^2 (p_1 + r + (2\sqrt{s_{\Gamma^*}} + 1)^2)}{\min\{\alpha_{\text{RSC}}^{\mathbf{X}}, 1\}^2}. \quad (10)$$

---

<sup>6</sup>note the deviation bound in principle also depends on other population quantities such as  $\mathfrak{m}(f_Z)$ ,  $\mathcal{M}(f_Z)$ ,  $\Lambda_{\max}(S_w)$  etc.

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  $\Delta_{\Theta}$ , the above bound automatically holds for  $\Delta_{\mathbf{F}}$ . Further, the following lemma provides the relative error of the estimated  $\Lambda$  under IR and the condition  $\Lambda_{\max}^{1/2}(S_{\mathbf{F}})$ , which translates to the requirement that the leading signal of  $\mathbf{F}$  overrules the averaged row error of  $\Delta_{\Theta}$ .

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

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

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 (9), based on observed  $\mathbf{X}$  and estimated  $\widehat{\mathbf{F}}$ , and assuming  $\Delta_{\mathbf{F}}$  is fixed.

**Proposition 2** (Bound for  $\Delta_A$  under fixed realization and a non-random  $\Delta_{\mathbf{F}}$ ). *Consider the estimator  $\widehat{A}$  obtained by solving (9). 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  $s_{A^*}\tau_{\mathbf{Z}} < \alpha_{RSC}^{\widehat{\mathbf{Z}}}/64$ ;
- A2.  $\|\widehat{\mathbf{Z}}_{n-1}^{\top}(\widehat{\mathbf{Z}}_n - \widehat{\mathbf{Z}}_{n-1}(A^*)^{\top})/n\|_{\infty} \leq C(n, p_1, p_2)$  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  $\widehat{A}$ :

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

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 (7) and (9) 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

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

with  $\mathcal{A}_X(z) := \mathbf{I} - A_{22}z$ . 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 5 respectively verify the RSC condition associated with  $\mathbf{X}$  and establish the high probability bounds for  $\|\mathbf{X}^\top \mathbf{E}/n\|_\infty$ ,  $\Lambda_{\max}(S_{\mathbf{X}})$  and  $\Lambda_{\max}(S_{\mathbf{E}})$ .

**Lemma 2** (Verification of the RSC condition for  $\mathbf{X}$ ). *Consider  $\mathbf{X} \in \mathbb{R}^{n \times p_2}$  whose rows correspond to a random realization  $\{x_1, \dots, x_n\}$  of the stable Gaussian  $\{X_t\}$  process, and its dynamics is governed by (6). Then, there exist positive constants  $c_i$  ( $i = 1, 2$ ) 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*

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

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  $1 - c_1 \exp(-c_2 \log(p_2 q))$ , the following bound holds:*

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

**Lemma 4** (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  $\{x_1, \dots, x_n\}$  of the stable Gaussian  $\{X_t\}$  process, and its dynamics is governed by (6). Then, there exist positive constants  $c_i$  ( $i = 0, 1, 2$ ) such that for sample size  $n \gtrsim p_2$ , with probability at least  $1 - c_1 \exp(-c_2 p_2)$ , the following bound holds for the eigen-spectrum of  $S_{\mathbf{X}}$ :*

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

**Lemma 5** (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  $\Sigma_e$ . Then, for sample size  $n \gtrsim q$ , with probability at least  $1 - \exp(-n/2)$ , the following bound holds:*

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

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  $\|\widehat{\mathbf{Z}}_{n-1}^\top (\widehat{\mathbf{Z}}_n - \widehat{\mathbf{Z}}_{n-1} (A^*)^\top) / n\|_\infty$ , 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*

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

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 (6). Suppose the lower and upper extremes of its spectral density  $f_Z(\omega)$  satisfy

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

for some constant  $c_0 \geq 6\sqrt{165\pi}$ . Then, with probability at least  $1 - c_1 \exp(-c_2 n)$ ,  $\widehat{\mathbf{Z}}_{n-1}$  satisfies the RSC condition with curvature

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

and tolerance

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

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

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

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

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

*Remark 2.* 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 (“Original 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  $\mathbf{m}(f_X)/\mathcal{M}^{1/2}(f_X)$  and the latter by  $\Lambda_{\max}^{1/2}(S_{\Delta_{\mathbf{F}_{n-1}}})$ . In particular, the RSC curvature is pushed toward zero compared with that in the Original 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 Original 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  $\|\Delta_{\mathbf{F}}/\sqrt{n}\|_{\mathbf{F}} \asymp \mathcal{O}(1)$ , the dominating term of the three components is

the one attributed to the non-random error<sup>7</sup> 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 (7) and (9) 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 5 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  $\{x_1, \dots, x_n\}$  and  $\{y_1, \dots, y_n\}$  obtained from the stable Gaussian processes  $X_t$  and  $Y_t$ , whose dynamics are described in (6) and (2). Suppose the following conditions hold for some  $(C_{X,l}, C_{X,u})$  and  $(C_{e,l}, C_{e,u})$ :*

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

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

Then, there exist universal constants  $\{C_i\}$  and  $\{c_i\}$  such that for sample size  $n \gtrsim q$ , by solving (7) with regularization parameter

$$\lambda_\Gamma = \max \left\{ C_1(2\pi\mathcal{M}(f_X) + \Lambda_{\max}(\Sigma_e))\sqrt{\frac{\log(p_2q)}{n}}, C_2(p_1 + r)\phi\mathcal{M}^{1/2}(f_X), C_3\Lambda_{\max}^{1/2}(\Sigma_e) \right\}, \quad (15)$$

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

$$\|\Delta_\Theta / \sqrt{n}\|_F^2 + \|\Delta_\Gamma\|_F^2 \lesssim C(\mathbf{m}(f_X), \mathcal{M}(f_X), \Lambda_{\max}(\Sigma_e)) \cdot \kappa(s_{\Gamma^*}, p_1^3, r^3, \phi) =: K_1, \quad (16)$$

for some function  $C(\mathbf{m}(f_X), \mathcal{M}(f_X), \Lambda_{\max}(\Sigma_e))$  that does not depend on  $n, p_2, q$ , and  $\kappa(\cdot)$  that depends linearly on  $s_{\Gamma^*}, p_1^3, r^3$  and the box constraint  $\phi$ .

Note that the above bound also holds if we replace  $\Delta_\Theta$  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  $Z_t$ :*

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

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 (9) with  $\lambda_A$  satisfying

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

---

<sup>7</sup>with the implicit assumption that  $\log(p_1 + p_2)/n \asymp o(1)$  which is indeed the case for this study.

with probability at least

$$\left(1 - c_1 \exp\{-c_2 \log(p_2 q)\}\right) \left(1 - c'_1 \exp\{-c'_2 \log(p_1 + p_2)\}\right), \quad (17)$$

the following bound holds for  $\Delta_A$ :

$$\|\Delta_A\|_F^2 \leq \check{C}(K_1, \mathbf{m}(f_Z), \mathcal{M}(f_Z)) \cdot \check{\kappa}(s_{A^*}),$$

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

*Remark 3.* Note that to establish the high probability finite-sample error bound of the transition matrix estimate  $\hat{A}$ , the sample size requirement  $n \gtrsim q$  for the proposed estimation procedure is more stringent compared to that for the Original Problem, with the latter given by  $n \gtrsim \sqrt{\log(p_1 + p_2)}$ . The root of this discrepancy is due to the estimated factor, whose accurate recovery from the calibration equation requires the concentration of  $\Lambda_{\max}(S_E)$  that provides adequate control over  $\Delta_{\mathbf{F}}$ , which in turn places the tightest condition on the sample size.

*Remark 4.* 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  $\tilde{Z}_t^\top := (Z_t, Z_{t-1}, \dots, Z_{t-d+1})$  that satisfies

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

In particular, with probability at least  $(1 - c_1 \exp\{-c_2 \log(p_2 q)\})(1 - c'_1 \exp\{-c'_2 \log(d(p_1 + p_2))\})$ , the following bound holds for the estimate of  $\tilde{A}$ :

$$\|\Delta_{\tilde{A}}\|_F^2 \leq \check{C}(K_1, \mathbf{m}(f_{\tilde{Z}}), \mathcal{M}(f_{\tilde{Z}})) \cdot \check{\kappa}(s_{\tilde{A}^*}).$$

However, note that although the error bound is still of the same form, the stronger temporal dependence yields a larger  $\check{C}(K_1, \mathbf{m}(f_{\tilde{Z}}), \mathcal{M}(f_{\tilde{Z}}))$  through the RSC curvature parameter; specifically, a smaller value of  $\mathbf{m}(f_{\tilde{Z}})$ . 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  $\lambda_A$ .

## 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  $\phi$  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 C certain relative extreme settings where the proposed relaxation fails to provide accurate estimates of the model parameters.

**Implementation.** The following relaxation of (7) is used in practice:

$$\min_{\Theta, \Gamma} f(\Theta, \Gamma) := \left\{ \frac{1}{2n} \|\mathbf{Y} - \Theta - \mathbf{X}\Gamma^\top\|_F^2 + \lambda_\Gamma \|\Gamma\|_1 \right\}, \quad \text{subject to} \quad \text{rank}(\Theta) \leq r, \quad (18)$$

---

**Algorithm 1:** Computational procedure for estimating  $A$ ,  $\Gamma$  and  $\Lambda$ .

---

**Input:** Time series data  $\{x_i\}_{i=1}^n$  and  $\{y_i\}_{i=1}^n$ ,  $(\lambda_\Gamma, r)$ , and  $\lambda_A$ .

- 1 **Stage I:** recover the latent factors by solving (18), through iterating between (1.1) and (1.2) until  $|f(\Theta^{(m)}, \Gamma^{(m)}) - f(\Theta^{(m-1)}, \Gamma^{(m-1)})| < \text{tolerance}$ :
- 2 (1.1) Update  $\hat{\Theta}^{(m)}$  by singular value thresholding (SVT): do SVD on the lagged value-adjusted hyperplane, i.e.,  $\mathbf{Y} - \mathbf{X}(\hat{\Gamma}^{(m-1)})^\top = UDV$ , where  $D := \text{diag}(d_1, \dots, d_{\min(n,q)})$ , and construct  $\hat{\Theta}^{(m)}$  by
 
$$\hat{\Theta}^{(m)} = UD_rV, \quad \text{where } D_r := \text{diag}(d_1, \dots, d_r, 0, \dots, 0).$$
- 3 (1.2) Update  $\hat{\Gamma}^{(m)}$  with the plug-in  $\hat{\Theta}^{(m)}$  so that each row  $j$  is obtained with Lasso regression (in parallel) and solves

$$\min_{\beta} \left\{ \frac{1}{2n} \|(\mathbf{Y} - \hat{\Theta}^{(m)})_{\cdot j} - \mathbf{X}\beta\|^2 + \lambda_A \|\beta\|_1 \right\}.$$

- 4 **Stage I output:**  $\hat{\Theta}$  and  $\hat{\Gamma}$ ; the estimated factor  $\hat{\mathbf{F}}$  and  $\hat{\Lambda}$  via (8) under (IR).
- 5 **Stage II:** estimate the transition matrix by solving (9): update each row of  $A$  (in parallel) by solving the Lasso problem:

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

- 6 **Stage II output:**  $\hat{A}$ .

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

---

which leads to Algorithm 1. The implementation of Stage I requires the pair of tuning parameters  $(\lambda_\Gamma, r)$  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 (2015), which searches for  $(\lambda_\Gamma, r)$  over a lattice that minimizes

$$\text{PIC}(\lambda_\Gamma, r) := \frac{1}{nq} \left\| \mathbf{Y} - \hat{\Theta} - \mathbf{X}\hat{\Gamma}^\top \right\|_F^2 + \hat{\sigma}^2 \left[ \frac{\log n}{n} \|\hat{\Gamma}\|_0 + r \left( \frac{n+p}{nq} \right) \log(nq) \right],$$

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

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

where  $\text{RSS}_i := \|(\mathbf{X}_n)_{\cdot i} - \mathbf{X}_{n-1}\hat{A}_i^\top\|^2$  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  $\Sigma_w^X$ ,  $\Sigma_X^F$  and  $\Sigma_e$  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). In particular, (IR) is imposed on the true value of the parameter, hence  $\Lambda^*$  that is used for generating  $Y_t$  always satisfies the restriction  $\Lambda = \begin{bmatrix} I_{p_1} \\ * \end{bmatrix}$ .

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  $\Gamma$  and that of entries of  $\Lambda$  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 sparsity 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  $A$ ,  $\rho(A)$ , 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  $\tilde{A}$  is smaller than 1<sup>8</sup>, where we let  $p = p_1 + p_2$  and

$$\tilde{A} := \begin{bmatrix} A^{(1)} & A^{(2)} & \dots & A^{(d)} \\ I_p & O & & O \\ \vdots & \ddots & \ddots & \vdots \\ O & O & I_p & O \end{bmatrix}.$$

Table 1 lists the simulation settings and their parameter setup.

	$q$	$p_1$	$p_2$	$s_{A_{11}}$	$s_{A_{12}}$	$s_{A_{21}}$	$s_{A_{22}}$	SNR( $Y_t$ )
A1	100	5	50	$s_A = 3/(p_1 + p_2)$				1.5
A2	200	10	100	$s_A = 3/(p_1 + p_2)$				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	$s_{A^{(1)}} = 3/(p_1 + p_2)$ $s_{A^{(2)}} = 2/(p_1 + p_2)$				2
B2 ( $d = 4$ )	200	5	100	0.5	$3/p_2$	0.5	$3/p_2$	2
				0.2	$2/p_2$	0.25	$2/p_2$	
				$s_{A^{(3)}} = 2/(p_1 + p_2)$ $s_{A^{(4)}} = 2/(p_1 + p_2)$				
B3 ( $d = 4$ )	100	5	25	0.5	$2/p_2$	0.5	$2/p_2$	2
				0.2	$1.5/p_2$	0.1	$1.5/p_2$	
				$s_{A^{(3)}} = 1/(p_1 + p_2)$ $s_{A^{(4)}} = 0.8/(p_1 + p_2)$				

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

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 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

<sup>8</sup>In practice, this can be achieved by first generating  $A^{(1)}, \dots, A^{(d)}$ , align them in  $\tilde{A}_{\text{initial}}$  and obtain the scale factor  $\zeta := \rho_{\text{target}}/\rho(\tilde{A}_{\text{initial}})$ , then scale  $A^{(i)}$  by  $\zeta^i$ . The validity of this procedure follows from simple algebraic manipulations.

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 sparsity in the component that governs the  $F_{t-i} \rightarrow X_t$  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.

**Performance evaluation.** We consider both the estimation and 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  $\Gamma$ ), defined as

$$\text{SEN} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{SPC} = \frac{\text{TN}}{\text{FP} + \text{TN}}, \quad \text{Err} = \|\Delta_M\|_{\text{F}} / \|M^*\|_{\text{F}} \text{ (for some generic matrix } M\text{)}.$$

We also track the estimated size of the latent component (i.e., the rank constraint in (7), jointly with  $\lambda_{\Gamma}$  is selected by PIC), as well as the relative errors of  $\hat{\Theta}$ ,  $\hat{\mathbf{F}}$  and  $\hat{\Lambda}$ . For forecasting, we focus on evaluating the  $h$ -step-ahead predictions for the  $X_t$  block. Specifically, for settings A1–A4, we consider  $h = 1$ ; for settings B1–B3, 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 for any generic time series  $X_t \in \mathbb{R}^p$ , each of its coordinates  $j = 1, \dots, p$  follows a centered random walk:

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

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

$$\text{RE} = \|\hat{x}_{T+h} - x_{T+h}\|_2^2 / \|x_{T+h}\|_2^2, \quad \text{RER} = \frac{\frac{1}{p_2} \sum_{j=1}^{p_2} \left| \frac{\hat{x}_{T+h,j} - x_{T+h,j}}{x_{T+h,j}} \right|}{\frac{1}{p_2} \sum_{j=1}^{p_2} \left| \frac{\tilde{x}_{T+h,j} - x_{T+h,j}}{x_{T+h,j}} \right|},$$

where RE measures the  $\ell_2$  norm of the relative error of the forecast to the true value; whereas for RER, 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  $\hat{x}_{T+h}$  and that of the benchmark  $\tilde{x}_{T+h}$  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. Table 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  $\hat{\Theta}$  and  $\hat{\Gamma}$  are well estimated, while the rank slightly underestimated. Further, the SEN and SPC measures of  $\hat{\Gamma}$  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  $\hat{\Theta}$  further decreases (results omitted for space considerations)<sup>9</sup>. Under the given IR, we decompose the estimated factor hyperplane into the factor block and its loadings. The results show

<sup>9</sup>This also comes up when comparing the relative error of  $\hat{\Theta}$  in the A1-A4 settings to that in the B1-B2 ones, where the latter two have a higher SNR.

	PIC-selected $r$	Err( $\hat{\Theta}$ )	Err( $\hat{\mathbf{F}}$ )	Err( $\hat{\mathbf{A}}$ )	SEN( $\hat{\Gamma}$ )	SPC( $\hat{\Gamma}$ )	Err( $\hat{\Gamma}$ )
A1	4.8(.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)

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

	lag	SEN( $\hat{A}$ )	SPC( $\hat{A}$ )	Err( $\hat{A}$ )	SEN( $\hat{A}_{22}$ )	SPC( $\hat{A}_{22}$ )	Err( $\hat{A}_{22}$ )
A1		0.99(.003)	0.95(.012)	0.35(.019)	0.99(.001)	0.96(.013)	0.31(.022)
A2		0.98(.008)	0.97(.004)	0.46(.018)	0.99(.001)	0.98(.003)	0.39(.017)
A3		0.86(.050)	0.98(.006)	0.73(.029)	0.93(.032)	0.98(.005)	0.65(.034)
A4		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)

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

		A1	A2	A3	A4	B1	B2	B3
$h = 1$	RE	0.53(.117)	0.60(.075)	0.80(.075)	0.56(.109)	0.62(.060)	0.89(.091)	0.81(.094)
	RER	0.38(.065)	0.38(.046)	0.45(.064)	0.40(.055)	0.35(.171)	0.42(.217)	0.32(.129)
$h = 2$	RE					0.66(.127)	0.94(.173)	0.90(.402)
	RER					0.24(.071)	0.29(.118)	0.26(.174)

Table 4: Evaluation of forecasting performance.

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% RE (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 RER measure). Of note, the 2-step-ahead forecasting value for settings with more lags outperforms the benchmark by an even wider margin with the RER ratio decreasing to less than 0.3.

## 5 Application to Commodity Price Interlinkages.

Interlinkages between commodity prices represent an active research area in economics, together with 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, 2008, 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.

We apply the same estimation procedure for each of the above three sub-periods. Starting with the calibration equation, we estimate the factor hyperplane  $\Theta$  and the sparse regression coefficient matrix  $\Gamma$ , then extract the factors based on the estimated factor hyperplane 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  $\Gamma X_t$ . Based on the composition of the factors, we note that the

	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		−				*		*		

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

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, 2008, 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  $(\hat{F}_t^\top, X_t^\top)^\top$  system. The estimated transition matrices are depicted in Figures 1 to 3 as networks. 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.

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+) 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 (IR+) constraint is not particularly stringent, especially in settings where the coefficient matrix  $\Gamma$  of the observed predictor variables in the calibration equation exhibits sufficient level of sparsity.

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. This introduces a problem of independent interest, namely what statistical guarantees can be established for the estimates of the transition matrix of a VAR system under high-dimensional scaling when one block (or even all) of the variables are subject to error. Similar 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. \(1996\)](#) examines parameter estimation of a univariate autoregressive process with error-in-variables and in more recent work [Komunjer and Ng \(2014\)](#) investigates parameter identification of VAR-X and dynamic panel VAR models subject to measurement errors.

The results obtained in this paper provide some initial insights, based on the roadmap used to establish them. Building on the discussion in Remark 2, consider the following setting where one is interested in estimating a VAR system with one block of variables contaminated by some *non-random*  $\mathbf{Z}$ , so that the transition matrix is obtained by solving

$$\min_A \left\{ \frac{1}{2n} \left\| \begin{bmatrix} \mathbf{x}_n^{(1)} \\ \mathbf{x}_n^{(2)} + \mathbf{z}_n \end{bmatrix} - \begin{bmatrix} \mathbf{x}_{n-1}^{(1)} \\ \mathbf{x}_{n-1}^{(2)} + \mathbf{z}_{n-1} \end{bmatrix} A^\top \right\|_F^2 + \lambda_A \|A\|_1 \right\},$$

whereas the true data generating mechanism is that  $\begin{pmatrix} X_t^{(1)} \\ X_t^{(2)} \end{pmatrix}$  jointly follow a VAR(1) model. Then, based on Lemma 6 and 7, as long as the RSC condition on the corresponding quantity is satisfied with high probability and the tuning parameter is chosen in accordance with the deviation bound condition, the error of the estimated transition matrix is still well-bounded. In particular, if the magnitude of  $\mathbf{Z}$  satisfies  $\|\mathbf{Z}/\sqrt{n}\|_F \asymp o(1)$ , then the error of the estimated transition matrix would still be  $\mathcal{O}(\sqrt{\log(p_1 + p_2)/n})$ , which is identical to that of a VAR model without error-in-variables, despite the fact that the estimation is based on contaminated quantities rather than uncontaminated samples. In addition, the presence of the contaminating  $\mathbf{Z}$  does not affect the sample size requirement with the latter remaining at  $n \gtrsim \sqrt{\log(p_1 + p_2)}$ , although it does affect the exact error bound through both the deviation bound and the curvature in the RSC condition. Thus, it is of interest to investigate the conditions required on a *random*  $\mathbf{Z}$ , so that the VAR estimates exhibit similar rates to those without contamination and this constitutes a topic of future research.

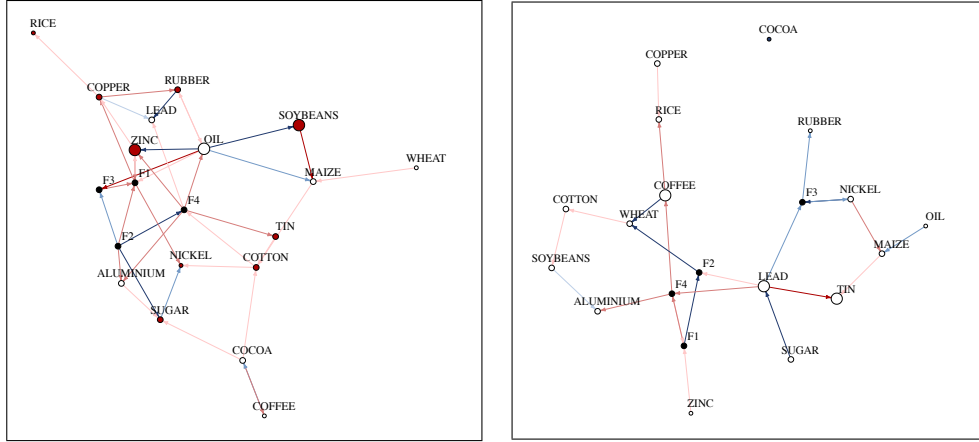


Figure 1: Estimated transition matrices for Pre-crisis period.

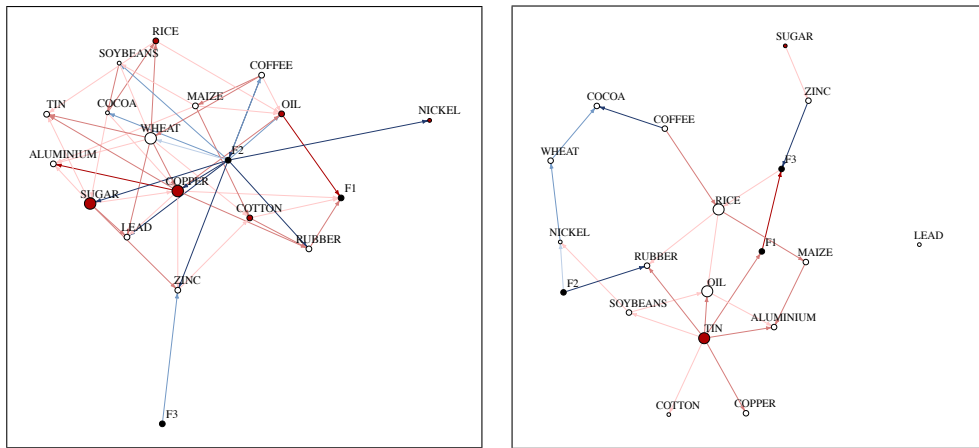


Figure 2: Estimated transition matrices for the Crisis period.

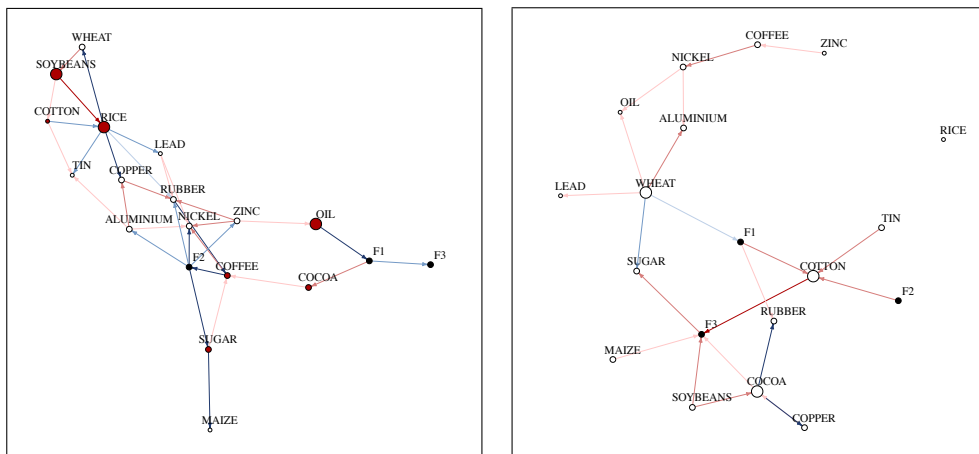


Figure 3: Estimated transition matrices for Post-crisis period

Left panel:  $\widehat{A}^{(1)}$ ; right panel:  $\widehat{A}^{(2)}$ . 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.

## 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.,  $\hat{\Theta}$  and  $\hat{\Gamma}$ . In the second part, we give proofs for the statements related to Stage II estimates, namely  $\hat{A}$ , with an emphasis on how to obtain the final high probability error bound through properly conditioning on related events.

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

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

$$\frac{1}{2n} \|\mathbf{X}\Delta_\Gamma^\top + \Delta_\Theta\|_F^2 \leq \frac{1}{n} \left( \langle \Delta_\Gamma^\top, \mathbf{X}^\top \mathbf{E} \rangle + \langle \Delta_\Theta, \mathbf{E} \rangle \right) + \lambda_\Gamma \left( \|\Gamma^*\|_1 - \|\hat{\Gamma}\|_1 \right), \quad (20)$$

which after rearranging terms gives

$$\frac{1}{2n} \|\mathbf{X}\Delta_\Gamma^\top\|_F^2 + \frac{1}{2} \|\Delta_\Theta / \sqrt{n}\|_F^2 \leq \frac{1}{n} \langle \mathbf{X}\Delta_\Gamma^\top, \hat{\Theta} - \Theta^* \rangle + \frac{1}{n} \left( \langle \Delta_\Gamma^\top, \mathbf{X}^\top \mathbf{E} \rangle + \langle \Delta_\Theta, \mathbf{E} \rangle \right) + \lambda_\Gamma \left( \|\Gamma^*\|_1 - \|\hat{\Gamma}\|_1 \right). \quad (21)$$

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  $\lambda_\Gamma$  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  $\Gamma$  and its complement:

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

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

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

and note that they satisfy

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

and

$$\|\Delta_{\mathbb{S}}\|_1 \leq \sqrt{s} \|\Delta_{\mathbb{S}}\|_F \leq \sqrt{s_{\Gamma^*}} \|\Delta\|_F. \quad (22)$$

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

$$\frac{\alpha_{\text{RSC}}^{\mathbf{X}}}{2} \|\Delta_\Gamma\|_F^2 - \tau_{\mathbf{X}} \|\Delta_\Gamma\|_1^2. \quad (23)$$

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:

$$\langle \Delta_\Gamma^\top, \mathbf{X}^\top \mathbf{E} \rangle \leq \|\Delta_\Gamma\|_1 \|\mathbf{X}^\top \mathbf{E}\|_\infty, \quad \langle \Delta_\Theta, \mathbf{E} \rangle \leq \|\Delta_\Theta\|_* \|\mathbf{E}\|_{op} = n \|\Delta_\Theta\|_* \Lambda_{\max}^{1/2}(S_{\mathbf{E}}); \quad (24)$$

for the last term, since

$$\|\widehat{\Gamma}\|_1 = \|\Gamma_{\mathcal{S}}^* + \Gamma_{\mathcal{S}^c}^* + \Delta_{\Gamma|\mathcal{S}} + \Delta_{\Gamma|\mathcal{S}^c}\|_1 = \|\Gamma_{\mathcal{S}}^* + \Delta_{\Gamma|\mathcal{S}}\|_1 + \|\Delta_{\mathcal{S}|\mathcal{S}^c}\|_1 \geq \|\Gamma_{\mathcal{S}}^*\|_1 - \|\Delta_{\Gamma|\mathcal{S}}\|_1 + \|\Delta_{\Gamma|\mathcal{S}^c}\|_1,$$

the following inequality holds:

$$\|\Gamma^*\|_1 - \|\widehat{\Gamma}\|_1 \leq \|\Delta_{\Gamma|\mathcal{S}}\|_1 - \|\Delta_{\Gamma|\mathcal{S}^c}\|_1. \quad (25)$$

Using the non-negativity of the RHS in (20), by choosing  $\lambda_{\Gamma} \geq \max\left\{2\|\mathbf{X}^{\top}\mathbf{E}/n\|_{\infty}, \Lambda_{\max}^{1/2}(S_{\mathbf{E}})\right\}$ , the following inequality holds:

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

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

$$\begin{aligned} \frac{\lambda_{\Gamma}}{2}\|\Delta_{\Gamma|\mathcal{S}^c}\|_1 &\leq \lambda_{\Gamma}\sqrt{p_1 + r}\|\Delta_{\Theta}/\sqrt{n}\|_{\text{F}} + \frac{3\lambda_{\Gamma}}{2}\|\Delta_{\Gamma|\mathcal{S}}\|_1, \\ \frac{\lambda_{\Gamma}}{2}\|\Delta_{\Gamma|\mathcal{S}}\|_1 + \frac{\lambda_{\Gamma}}{2}\|\Delta_{\Gamma|\mathcal{S}^c}\|_1 &\leq \lambda_{\Gamma}\sqrt{p_1 + r}\|\Delta_{\Theta}/\sqrt{n}\|_{\text{F}} + \frac{3\lambda_{\Gamma}}{2}\|\Delta_{\Gamma|\mathcal{S}}\|_1 + \frac{\lambda_{\Gamma}}{2}\|\Delta_{\Gamma|\mathcal{S}}\|_1, \\ \|\Delta_{\Gamma}\|_1 &\leq \sqrt{4(p_1 + r)}\|\Delta_{\Theta}/\sqrt{n}\|_{\text{F}} + 4\|\Delta_{\Gamma|\mathcal{S}}\|_1 \leq \sqrt{4(p_1 + r)}\|\Delta_{\Theta}/\sqrt{n}\|_{\text{F}} + 4\sqrt{s}\|\Delta_{\Gamma}\|_{\text{F}}, \end{aligned}$$

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

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

that is,

$$\|\Delta_{\Gamma}\|_1^2 \leq 4(p_1 + r + 4s) \left[ \|\Delta_{\Gamma}\|_{\text{F}}^2 + \|\Delta_{\Theta}/\sqrt{n}\|_{\text{F}}^2 \right]. \quad (26)$$

Combine (23) and (26), a lower bound for the LHS of (21) is given by

$$\left( \frac{\alpha_{\text{RSC}}^{\mathbf{X}}}{2} - 4\tau_{\mathbf{X}}(p_1 + r + 4s) \right) \|\Delta_{\Gamma}\|_{\text{F}}^2 + \left( \frac{1}{2} - 4\tau_{\mathbf{X}}(p_1 + r + 4s) \right) \|\Delta_{\Theta}/\sqrt{n}\|_{\text{F}}^2. \quad (27)$$

Step (ii). For the first term in the RHS of (21), using the duality of the nuclear-operator norm pair, the following inequality holds:

$$\frac{1}{n} |\langle \mathbf{X}\Delta_{\Gamma}^{\top}, \widehat{\Theta} - \Theta^* \rangle| \leq \frac{1}{n} |\langle \mathbf{X}\Delta_{\Gamma}^{\top}, \widehat{\Theta} \rangle| + \frac{1}{n} |\langle \mathbf{X}\Delta_{\Gamma}^{\top}, \Theta^* \rangle| \quad (28)$$

$$\leq \|\mathbf{X}\Delta_{\Gamma}^{\top}/\sqrt{n}\|_{\text{op}} \|\widehat{\Theta}/\sqrt{n}\|_* + \|\mathbf{X}\Delta_{\Gamma}^{\top}/\sqrt{n}\|_{\text{op}} \|\Theta^*/\sqrt{n}\|_*. \quad (29)$$

For  $\|\mathbf{X}\Delta_{\Gamma}^{\top}/\sqrt{n}\|_{\text{op}}$ , we have

$$\|\mathbf{X}\Delta_{\Gamma}^{\top}/\sqrt{n}\|_{\text{op}} \leq \|\mathbf{X}/\sqrt{n}\|_{\text{op}} \|\Delta_{\Gamma}^{\top}\|_{\text{op}} \leq \|\mathbf{X}/\sqrt{n}\|_{\text{op}} \|\Delta_{\Gamma}^{\top}\|_{\text{F}} = \Lambda_{\max}^{1/2}(S_{\mathbf{X}}) \|\Delta_{\Gamma}\|_{\text{F}}, \quad (30)$$

where the first inequality comes from the sub-multiplicativity of the nuclear norm. Combining with  $(\mathbb{R}^+)$  box constraint on the eigen-spectrum and the feasibility of  $\Theta^*$ , we obtain  $\|\Theta^*/\sqrt{n}\|_* \leq p_1\phi$  and  $\|\widehat{\Theta}/\sqrt{n}\|_* \leq r\phi$ , thus (28) is upper bounded by

$$(p_1 + r)\phi\Lambda_{\max}^{1/2}(S_{\mathbf{X}}) \|\Delta_{\Gamma}\|_{\text{F}}.$$

Further, combining with (24) and (25), as long as  $\lambda_\Gamma \geq \{\|\mathbf{X}^\top \mathbf{E}/n\|_\infty, \Lambda^{1/2}(S_{\mathbf{E}}), (p_1+r)\phi\Lambda_{\max}^{1/2}(S_{\mathbf{X}})\}$ , the following upper bound holds for the RHS of (21):

$$\begin{aligned} & \lambda_\Gamma \|\Delta_\Gamma\|_{\mathbf{F}} + \lambda_\Gamma \|\Delta_\Gamma\|_1 + \lambda_\Gamma \sqrt{p_1+r} \|\Delta_\Theta/\sqrt{n}\|_{\mathbf{F}} + \lambda_\Gamma (\|\Delta_\Gamma\|_{\mathbb{S}} - \|\Delta_\Gamma\|_{\mathbb{S}^c}) \\ & \leq \lambda_\Gamma \left( (2\sqrt{s_{\Gamma^*}} + 1) \|\Delta_\Gamma\|_{\mathbf{F}} + \sqrt{p_1+r} \|\Delta_\Theta/\sqrt{n}\|_{\mathbf{F}} \right) \\ & \leq \lambda_\Gamma \sqrt{(2\sqrt{s_{\Gamma^*}} + 1)^2 + (p_1+r)^2} \sqrt{\|\Delta_\Gamma\|_{\mathbf{F}}^2 + \|\Delta_\Theta/\sqrt{n}\|_{\mathbf{F}}^2}. \end{aligned} \quad (31)$$

Step (iii). Combine (27) and (31), by rearranging terms and requiring  $\tau_{\mathbf{X}}$  to satisfy  $\tau_{\mathbf{X}}(p_1+r+4s_{\Gamma^*}) < \min\{\alpha_{\text{RSC}}^{\mathbf{X}}, 1\}/16$ , the following inequality holds:

$$\frac{\min\{\alpha_{\text{RSC}}^{\mathbf{X}}, 1\}}{4} \left( \|\Delta_\Gamma\|_{\mathbf{F}}^2 + \|\Delta_\Theta/\sqrt{n}\|_{\mathbf{F}}^2 \right) \leq \lambda_\Gamma \sqrt{(2\sqrt{s_{\Gamma^*}} + 1)^2 + (p_1+r)^2} \sqrt{\|\Delta_\Gamma\|_{\mathbf{F}}^2 + \|\Delta_\Theta/\sqrt{n}\|_{\mathbf{F}}^2},$$

which gives

$$\|\Delta_\Gamma\|_{\mathbf{F}}^2 + \|\Delta_\Theta/\sqrt{n}\|_{\mathbf{F}}^2 \leq \frac{16\lambda_\Gamma^2 \left( (p_1+r) + (2\sqrt{s_{\Gamma^*}} + 1)^2 \right)}{\min\{\alpha_{\text{RSC}}^{\mathbf{X}}, 1\}^2}.$$

□

*Proof sketch for Theorem 1.* First we note that the requirement on the tuning parameter  $\lambda_\Gamma$  determines the leading term in the ultimate high probability error bound. By Lemma 4 and 5, to have adequate concentration for the leading eigenvalue  $\Lambda_{\max}(\cdot)$  of the sample covariance matrices, the requirement imposed on the sample size makes  $\sqrt{\log(p_2q)/n}$  a lower order term relative to  $\mathcal{M}^{1/2}(f_X)$  and  $\Lambda_{\max}^{1/2}(\Sigma_e)$ , with the latter two being  $\mathcal{O}(1)$  terms. Consequently, the choice of the tuning parameter effectively becomes

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

and by conditions C1 and C2, there exists some constant  $C$  such that  $\lambda_\Gamma^2 \leq C$ . 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  $P_1 := 1 - c_1 \exp(-c_2 \log(p_2q))$ :

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

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

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

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

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

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  $P_1 P_{2,\text{RSC}}$  where we define  $P_{2,\text{RSC}} := 1 - c'_1 \exp(-c'_2 n)$ , 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  $P_1 P_{2,\text{DB}}$  where we define  $P_{2,\text{DB}} := 1 - c'_1 \exp(-c'_2 \log(p_1 + p_2))$ , the following bound holds for the deviation bound  $C(n, p_1, p_2)$  *unconditionally*:<sup>10</sup>

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

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  $\min\{P_1 P_{2,\text{RSC}}, P_1 P_{2,\text{DB}}\}$ , the bound in Theorem 2 holds.  $\square$

## B Proof for Lemmas.

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

*Proof of Lemma 1.* Note that

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

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

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

<sup>10</sup>Note that it can be shown that  $\|\varepsilon_n\|_{\mathbf{F}}^2 = O(\|\Delta_{\mathbf{F}}\|_{\mathbf{F}}^2)$

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

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

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

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

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

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

One such bound is given by

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

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

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

□

*Proof of Lemma 2.* First, suppose we have

$$\frac{1}{2}v' S_{\mathbf{X}} v = \frac{1}{2}v' \left( \frac{\mathbf{X}'\mathbf{X}}{n} \right) v \geq \frac{\alpha_{\text{RSC}}}{2} \|v\|_2^2 - \tau_n \|v\|_1^2, \quad \forall v \in \mathbb{R}^p; \quad (32)$$

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

$$\frac{1}{2n} \|\mathbf{X}\Delta\|_{\mathbf{F}}^2 = \frac{1}{2} \sum_{j=1}^q \Delta_j' \left( \frac{\mathbf{X}'\mathbf{X}}{n} \right) \Delta_j \geq \frac{\alpha_{\text{RSC}}}{2} \sum_{j=1}^q \|\Delta_j\|_2^2 - \tau_n \sum_{j=1}^q \|\Delta_j\|_1^2 \geq \frac{\alpha_{\text{RSC}}}{2} \|\Delta\|_{\mathbf{F}}^2 - \tau_n \|\Delta\|_1^2.$$

Therefore, it suffices to verify that (32) 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  $\eta > 0$ ,

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

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:

$$\mathbb{P} \left[ \sup_{v \in \mathbb{K}(2k)} |v'(S_{\mathbf{X}} - \Gamma_X(h))v| > 2\pi \mathcal{M}(g_X)\eta \right] \leq 2 \exp \left( -cn \min\{\eta, \eta^2\} + 2k \min\{\log p, \log(21ep/2k)\} \right).$$

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  $\Gamma = S_{\mathbf{X}} - \Gamma_X(0)$  and  $\delta = \pi\mathfrak{m}(g_X)/27$ , so that the following holds

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

with probability at least  $1 - 2\exp(-cn \min\{\gamma^{-2}, 1\} + 2k \log p)$  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 [\(32\)](#) holds with

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

and so does also the RSC condition.  $\square$

*Proof of Lemma 3.* We note that

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

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:

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

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  $1 \leq i \leq p_2, 1 \leq j \leq q$ , and the following bound holds:

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

Set  $\eta = c'\sqrt{\log p/n}$  for  $c' > (1/c)$  and with the choice of  $n \gtrsim \log(p_2 q)$ ,  $\min\{\eta^2, \eta\} = \eta^2$ , then with probability at least  $1 - c_1 \exp(-c_2 \log p_2 q)$ , there exists some  $c_0$  such that the following bound holds:

$$\frac{1}{n}\|\mathbf{X}^\top \mathbf{E}\|_\infty \leq c_0(2\pi\mathcal{M}(g_X) + \Lambda_{\max}(\Sigma_e))\sqrt{\frac{\log(p_2 q)}{n}}.$$

$\square$

Before proving [Lemma 4](#), we first state [Lemma B.1](#) which provides a concentration inequality in the operator norm.

**Lemma B.1.** *Consider the stationary centered Gaussian process  $\{X_t\} \in \mathbb{R}^p$ , whose spectral density function  $g_X(\omega)$  exists and the maximum eigenvalue is bounded a.e. on  $[-\pi, \pi]$ . Then, for  $\mathbf{X}$  whose rows are random realizations  $\{x_0, \dots, x_{n-1}\}$  of  $\{X_t\}$ , the following bound holds for  $S_{\mathbf{X}} = \mathbf{X}^\top \mathbf{X}/n$ , for some  $c > 0$ :*

$$\mathbb{P}\left[\|S_{\mathbf{X}} - \Gamma_X(0)\|_{op} > 4\pi\mathcal{M}(g_X)\eta\right] \leq 2 \exp(-cn \min\{\eta, \eta^2\} + p \log 8).$$

*Proof of Lemma B.1.* First, we note that by Basu and Michailidis (2015, Proposition 2.4), the following inequality holds for any fixed  $v \in S^p$ , where  $S^p := \{v \in \mathbb{R}^p : \|v\| = 1\}$  is the  $p$ -dimensional unit sphere:

$$\mathbb{P}\left[|v'(S_{\mathbf{X}} - \Gamma_X(0))v| > 2\pi\mathcal{M}(g_X)\eta\right] \leq 2\exp(-cn \min\{\eta, \eta^2\}). \quad (33)$$

Additionally, by Vershynin (2010, Lemma 5.4),

$$\|S_{\mathbf{X}} - \Gamma_X(0)\|_{op} = \sup_{v \in S^p} |v'(S_{\mathbf{X}} - \Gamma_X(0))v| \leq (1 - 2\delta)^{-1} \sup_{v \in \mathcal{N}_\delta} v'[S_{\mathbf{X}} - \Gamma_X(0)]v,$$

where  $\mathcal{N}_\delta$  is a  $\delta$ -net of  $S^p$  for some  $\delta \in [0, 1)$ , which guarantees that the sphere can essentially be replaced by its  $\delta$ -net whose cardinality is finite. Towards this end, based upon (33), take the union bound over all vectors  $v$  in the  $\frac{1}{4}$ -net of  $S^p$ , whose cardinality is at most  $8^p$  (e.g. Anderson, 2011), we have

$$\begin{aligned} \mathbb{P}\left[\| \frac{1}{n}X'X - \Gamma_X(0) \|_{op} > 4\pi\mathcal{M}(g_X)\eta\right] &\leq \mathbb{P}\left[\sup_{v \in \mathcal{N}_\delta} |v'(S - \Gamma_X(0))v| > 4\pi\mathcal{M}(g_X)\eta\right] \\ &\leq 8^p \cdot 2\exp\left(-cn \min\{\eta, \eta^2\}\right). \end{aligned}$$

□

*Proof of Lemma 4.* The result follows in a straightforward manner based on Lemma B.1. Specifically, by letting  $\eta = c'\sqrt{p_2/n}$  for  $c' > (\log 8/c)$  and with  $n \gtrsim p$  so that  $\min\{\eta^2, \eta\} = \eta^2$ , then if we relax  $\Lambda_{\max}(\Gamma_X(0))$  by its upper bound  $2\pi\mathcal{M}(g_X)$  (Basu and Michailidis, 2015, Proposition 2.3), with probability at least  $1 - c_1 \exp(-c_2 p_2)$ , the following bound holds for some  $c_0$ :

$$\Lambda_{\max}(S_{\mathbf{X}}) \leq c_0\mathcal{M}(g_X).$$

□

*Proof of Lemma 5.* 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:

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

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

$$\|S_{\mathbf{E}}\|_{op} \leq \|\Sigma_e\|_{op} + \|S_{\mathbf{E}} - \Sigma_e\|_{op} \leq \Lambda_{\max}(\Sigma_e) + \Lambda_{\max}(\Sigma_e)\delta(n, q, \eta).$$

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

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

□

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

$$\frac{1}{2}\theta^\top \hat{\Gamma}_{\mathbf{Z}}\theta \geq \frac{\alpha_{\text{RSC}}^{\hat{\mathbf{Z}}}}{2}\|\theta\|^2 - \tau_{\mathbf{Z}}\|\theta\|_1^2, \quad \forall \theta \in \mathbb{R}^p.$$

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

$$\widehat{\Gamma}_{\mathbf{Z}} = \Gamma_{\mathbf{Z}} + \left( \frac{1}{n} \mathbf{Z}_{n-1}^\top \Delta_{\mathbf{Z}_{n-1}} + \frac{1}{n} \Delta_{\mathbf{Z}_{n-1}}^\top \mathbf{Z}_{n-1} \right) + \left( \frac{1}{n} \Delta_{\mathbf{Z}_{n-1}}^\top \Delta_{\mathbf{Z}_{n-1}} \right), \quad (34)$$

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

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

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  $\theta \in \mathbb{K}(2s)$ <sup>11</sup>, the following bound holds

$$|\theta^\top \Gamma \theta| \leq 27\delta \left( \|\theta\|_2^2 + \frac{1}{s} \|\theta\|_1^2 \right), \quad \forall \theta \in \mathbb{R}^p. \quad (35)$$

Then, based on (35), consider  $\Gamma = \widehat{\Gamma} - \Sigma$  then rearrange terms, so that  $\theta^\top \widehat{\Gamma} \theta \geq \theta^\top \Sigma \theta - \frac{27\delta}{2} \left( \|\theta\|_2^2 + \frac{1}{2} \|\theta\|_1^2 \right)$ . The RE condition follows by setting  $\delta$  to be some quantity related to  $\Lambda_{\min}(\Sigma)$ .

In light of this, for the first two terms in (34), let

$$\Psi := \Gamma_{\mathbf{Z}} + \left( \frac{1}{n} \mathbf{Z}_{n-1}^\top \Delta_{\mathbf{Z}_{n-1}} + \frac{1}{n} \Delta_{\mathbf{Z}_{n-1}}^\top \mathbf{Z}_{n-1} \right),$$

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

$$\left| \theta^\top [\Psi - \Gamma_{\mathbf{Z}}(0)] \theta \right| \leq \left| \theta^\top (\Gamma_{\mathbf{Z}} - \Gamma_{\mathbf{Z}}(0)) \theta \right| + \left| \theta^\top \left( \frac{1}{n} \mathbf{Z}'_{n-1} \Delta_{\mathbf{Z}_{n-1}} + \frac{1}{n} \Delta'_{\mathbf{Z}_{n-1}} \mathbf{Z}_{n-1} \right) \theta \right|,$$

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

$$(i) \quad \left| \theta^\top (\Gamma_{\mathbf{Z}} - \Gamma_{\mathbf{Z}}(0)) \theta \right|, \quad (ii) \quad \left| \theta^\top \left( \frac{1}{n} \mathbf{Z}'_{n-1} \Delta_{\mathbf{Z}_{n-1}} + \frac{1}{n} \Delta'_{\mathbf{Z}_{n-1}} \mathbf{Z}_{n-1} \right) \theta \right|.$$

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

$$\mathbb{P} \left[ \left| \theta^\top (\Gamma_{\mathbf{Z}} - \Gamma_{\mathbf{Z}}(0)) \theta \right| > 2\pi \mathcal{M}(f_{\mathbf{Z}}) \eta \right] \leq 2 \exp \left[ -cn \min\{\eta^2, \eta\} \right],$$

and further with probability at least  $1 - 2 \exp \left( -cn \min\{\eta^2, \eta\} + 2s \min\{\log p, \log(21ep/2s)\} \right)$ , the following bound holds:

$$\sup_{\theta \in \mathbb{K}(2s)} \left| \theta^\top (\Gamma_{\mathbf{Z}} - \Gamma_{\mathbf{Z}}(0)) \theta \right| < 2\pi \mathcal{M}(f_{\mathbf{Z}}) \eta. \quad (36)$$

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:

$$\sup_{\theta \in \mathbb{K}(2s)} \left| \frac{1}{n} \langle \mathbf{Z}_{n-1} \theta, \Delta_{\mathbf{Z}_{n-1}} \theta \rangle \right| \leq \left( \sup_{\theta \in \mathbb{K}(2s)} \left\| \frac{\mathbf{Z}_{n-1} \theta}{\sqrt{n}} \right\| \right) \left( \sup_{\|\theta\| \leq 1} \left\| \frac{\Delta_{\mathbf{Z}_{n-1}} \theta}{\sqrt{n}} \right\| \right). \quad (37)$$

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

For the first term in (37), since rows of  $\mathbf{Z}_{n-1}$  are time series realizations from (6), 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  $Q_{st} = \theta' \Gamma_Z(t-s)\theta$ . To get its upper bound, we bound its square, and use again (36), that is,

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

For the second term  $\|\Delta_{\mathbf{Z}_{n-1}}\theta/\sqrt{n}\|$ , this is non-random, and for all  $\|\theta\| \leq 1$ ,  $\|\Delta_{\mathbf{Z}_{n-1}}\theta/\sqrt{n}\| \leq \Lambda_{\max}^{1/2}(S_{\Delta_{\mathbf{Z}_{n-1}}}) = \Lambda_{\max}^{1/2}(S_{\Delta_{\mathbf{F}_{n-1}}})$ . Therefore, the following bound holds for (37):

$$\sup_{\theta \in \mathbb{K}(2s)} \left| \frac{1}{n} \langle \mathbf{Z}_{n-1}\theta, \Delta_{\mathbf{Z}_{n-1}}\theta \rangle \right| \leq \Lambda_{\max}^{1/2}(S_{\Delta_{\mathbf{F}_{n-1}}}) \sqrt{2\pi \mathcal{M}(f_Z) + 2\pi \mathcal{M}(f_Z) \eta}. \quad (38)$$

Combine (36) and (38) that are respectively the bounds for (i) and (ii), and the following bound holds with probability at least  $1 - 2 \exp(-cn \min\{\eta^2, \eta\} + 2s \min\{\log p, \log(21ep/2s)\})$ :

$$\sup_{\theta \in \mathbb{K}(2s)} \left| \theta^\top \left( \Psi - \Gamma_Z(0) \right) \theta \right| \leq 2\pi \mathcal{M}(f_Z) \eta + 2\Lambda_{\max}^{1/2}(S_{\Delta_{\mathbf{F}_{n-1}}}) \sqrt{2\pi \mathcal{M}(f_Z) + 2\pi \mathcal{M}(f_Z) \eta}. \quad (39)$$

Now applying Loh and Wainwright (2012, Lemma 12) to  $\Gamma = \Psi - \Gamma_Z(0)$ , and  $\delta$  being the RHS of (39), then the following bound holds:

$$\theta^\top \widehat{\Gamma} \theta \geq 2\pi \mathfrak{m}(f_Z) \|\theta\|_2^2 - 27\delta (\|\theta\|_2^2 + \frac{1}{s} \|\theta\|_1^2) = (2\pi \mathfrak{m}(f_Z) - 27\delta) \|\theta\|^2 - \frac{27\delta}{s} \|\theta\|_1^2.$$

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

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

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

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

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

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

$$\begin{aligned} \frac{1}{n} \|\widehat{\mathbf{Z}}_{n-1}^\top (\widehat{\mathbf{Z}}_n - \widehat{\mathbf{Z}}_{n-1}(A^*)^\top)\|_\infty &= \frac{1}{n} \left\| \left( \mathbf{Z}_{n-1} + \Delta_{\mathbf{Z}_{n-1}} \right)^\top \left( \mathbf{W} + \Delta_{\mathbf{Z}_n} - \Delta_{\mathbf{Z}_{n-1}}(A^*)^\top \right) \right\|_\infty \\ &\leq \left\| \frac{1}{n} \mathbf{Z}_{n-1}^\top \mathbf{W} \right\|_\infty + \left\| \frac{1}{n} \Delta_{\mathbf{Z}_{n-1}}^\top \mathbf{W} \right\|_\infty + \left\| \frac{1}{n} \mathbf{Z}_{n-1}^\top \left( \Delta_{\mathbf{Z}_n} - \Delta_{\mathbf{Z}_{n-1}}(A^*)^\top \right) \right\|_\infty \\ &\quad + \left\| \frac{1}{n} \Delta_{\mathbf{Z}_{n-1}}^\top \left( \Delta_{\mathbf{Z}_n} - \Delta_{\mathbf{Z}_{n-1}}(A^*)^\top \right) \right\|_\infty \\ &:= T_1 + T_2 + T_3 + T_4. \end{aligned} \quad (40)$$

We provide bounds on each term 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

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

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

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

is Gaussian, and the following tail bound holds:

$$\mathbb{P}[|\kappa_{ij}| \geq t] \leq e \cdot \exp\left(-\frac{cnt^2}{\Lambda_{\max}(\Sigma_w) \max_{i \in \{1, \dots, p_1+p_2\}} \|\Delta_{\mathbf{Z}_{n-1}, i}^\top / \sqrt{n}\|_2^2}\right) = e \cdot \exp\left(-\frac{cnt^2}{\Lambda_{\max}(\Sigma_w) \max_{i \in \{1, \dots, p_1\}} \|\Delta_{\mathbf{F}_{n-1}, i}^\top / \sqrt{n}\|_2^2}\right).$$

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

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

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

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

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

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

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  $(Q^i)_{st} = [\Gamma_Z(t - s)]_{ii}$  satisfying  $\Lambda_{\max}(Q^i) \leq \Lambda_{\max}(\Gamma_Z(0)) \leq 2\pi \mathcal{M}(f_Z)$ , so for any  $(i, j)$ ,  $(\frac{1}{n} \mathbf{Z}_{n-1}^\top \varepsilon_n)_{ij}$  satisfies

$$\mathbb{P}\left[\left| \left( \frac{1}{n} \mathbf{Z}_{n-1}^\top \varepsilon_n \right)_{ij} \right| > t\right] \leq \exp\left(1 - \frac{cnt^2}{\Lambda_{\max}(\Gamma_Z(0)) \max_{j \in \{1, \dots, p_1\}} \|\varepsilon_{n, j} / \sqrt{n}\|_2^2}\right).$$

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

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

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

$$\frac{1}{n} \|\mathbf{Z}_{n-1}^\top (\Delta_{\mathbf{Z}_n} - \Delta_{\mathbf{Z}_{n-1}}(A^*)^\top)\|_\infty \leq c_0 (2\pi \mathcal{M}(f_Z))^{1/2} \max_{j \in \{1, \dots, (p_1+p_2)\}} \|\varepsilon_{n, j} / \sqrt{n}\| \sqrt{\frac{\log(p_1 + p_2)}{n}}.$$

For  $T_4$ , it is deterministic, and satisfies

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

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  $1 - c_1 \exp(-c_2 \log(p_1 + p_2))$ , the bound in (14) holds.  $\square$

### C 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  $\begin{pmatrix} X_t \\ F_t \end{pmatrix} \sim \mathcal{N}(0, \Sigma)$  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. Note that the simulation settings are characterized by adequate sparsity in  $\Gamma$ , which in turn limits the size of the equivalence class  $\mathcal{C}(Q_2)$  as mentioned in Remark 1 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  $\Gamma$ . Specifically, for both  $\Lambda$  and  $\Gamma$ , their entries are generated from  $\text{Unif}\{(-1.5, -1.2) \cup (1.2, 1.5)\}$ . Additionally, we specify one 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  $\Theta$  based on different initializers:

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

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

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  $\Theta^* + 0.1 * \mathbf{Z}_{n \times q}$ , where each entry  $\Theta^*$  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  $\phi_0$ . 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  $\Gamma$ . Specifically, using the same generating mechanism for  $\Lambda$  and  $\Gamma$  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  $\Gamma$  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  $\Gamma$ .

Finally, we consider a case that lies between the above two settings, that is, there is a structured sparsity pattern in  $\Gamma$ . Specifically, we set the last 5 columns of  $\Gamma$  to be dense while the remaining ones are sparse. The overall sparsity level of  $\Gamma$  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  $\Gamma$  with 10% uniformly distributed non-zeros entries, due to the presence of the five dense columns. As the

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

Table 7: Performance evaluation for  $\hat{\Theta}$  obtained from different initializers under a structured-sparse setting.

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  $\Gamma$ , 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  $\Gamma$  is unknown.

## D An Outline of the Estimation Procedure in Low-dimensional Settings.

For the sake of completeness, we outline the estimation procedure proposed in Bai et al. (2016) and elaborate on the reasons that it is not applicable in high-dimensional settings. Note that the restriction  $\text{Cov}(w_t^X, W_t^F) = O$  is universal for all sets of identifications considered. Given a sample version corresponding to the calibration equation

$$\mathbf{Y} = \mathbf{F}\Lambda^\top + \mathbf{X}\Gamma^\top + \mathbf{E},$$

and that to the VAR equation

$$\mathbf{Z}_n = \mathbf{Z}_{n-1}A^\top + \mathbf{W},$$

the estimation procedure is based on the following steps.

1. Project and estimate a factor model. Specifically, by left multiplying  $\mathbb{P}_{\mathbf{X}^\perp} := \mathbf{I}_{p_1} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ , the model to estimate becomes

$$\mathbb{P}_{\mathbf{X}^\perp} \mathbf{Y} = \mathbb{P}_{\mathbf{X}^\perp} \mathbf{F}\Lambda^\top + \mathbb{P}_{\mathbf{X}^\perp} \mathbf{E}.$$

Proceed by doing factor analysis on  $\mathbb{P}_{\mathbf{X}^\perp} \mathbf{Y}$  through a quasi-Maximum Likelihood procedure as detailed in Bai and Li (2012), and obtain *intermediate estimates* denoted by  $\tilde{\Lambda}$ ,  $\tilde{\mathbf{F}}$ ,  $\tilde{\Gamma}$  and  $\tilde{\Sigma}_{ee}$ .

2. Estimate a VAR model based on  $(\tilde{\mathbf{F}}, \mathbf{X})$ , and denote the intermediate estimate of the transition matrix by  $\tilde{A}$  and the residual by  $\tilde{\mathbf{W}}$ . Calculate the sample covariance matrix of  $\tilde{\mathbf{W}}$ , partitioned as  $[\tilde{\Sigma}_w^{ff}, \tilde{\Sigma}_w^{fx}; \tilde{\Sigma}_w^{xf}, \tilde{\Sigma}_w^{xx}]$ .

3. Calculate a rotation matrix depending on the identification restrictions (either IRa, IRb or IRc) so that the one under consideration is satisfied. Specifically, all such rotation matrices involve

$$\tilde{\Sigma}_w^{ff \cdot x} := \tilde{\Sigma}_w^{ff} - \tilde{\Sigma}_w^{fx} (\tilde{\Sigma}_w^{xx})^{-1} \tilde{\Sigma}_w^{xf}.$$

Apply the rotation matrix (or their related transformations) to all previous intermediate estimates to obtain the final estimates.

To initiate the procedure,  $\mathbb{P}_{\mathbf{x}^\perp}$  is required for the first step; yet, this quantity is not readily available in high-dimensional settings where  $p_2 \geq n$ . Moreover, subsequent calculations of the rotation matrix are based on  $\tilde{\Sigma}_w^{ff \cdot x}$ , with the latter relying on  $(\tilde{\Sigma}_w^{xx})^{-1}$ , which again is not properly defined under high-dimensional scaling.

## E List of Commodities and Macroeconomic Variables.

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

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

Name	Description	tCode	Category	Region
IPL.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	Commercial 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
USDL.US	Trade weighted U.S.dollar index	5	Interest & Exchange Rates	US
CPLUS	CPI: all items	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 against 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 items	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:  $\Delta X_t$ ; 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

- Agarwal, A., S. Negahban, and M. J. Wainwright (2012). Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics*, 1171–1197.
- Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*, Volume 2. Wiley New York.
- Anderson, T. W. (2011). *The Statistical Analysis of Time Series*, Volume 19. John Wiley & Sons.
- Ando, T. and J. Bai (2015). Selecting the regularization parameters in high-dimensional panel data models: Consistency and efficiency. *Econometric Review*, 1–29.
- Bai, J. and K. Li (2012). Statistical analysis of factor models of high dimension. *The Annals of Statistics* 40(1), 436–465.
- Bai, J., K. Li, and L. Lu (2016). Estimation and inference of factor models. *Journal of Business & Economic Statistics* 34(4), 620–641.
- Bai, J. and S. Ng (2008). Large dimensional factor analysis. *Foundations and Trends® in Econometrics* 3(2), 89–163.
- Bai, J. and S. Ng (2013). Principal components estimation and identification of static factors. *Journal of Econometrics* 176(1), 18–29.
- Bañbura, M., D. Giannone, and L. Reichlin (2010). Large bayesian vector auto regressions. *Journal of Applied Econometrics* 25(1), 71–92.
- Basu, S. and G. Michailidis (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics* 43(4), 1535–1567.
- Bernanke, B. S., J. Boivin, and P. Elias (2005). Measuring the effects of monetary policy: a factor-augmented vector autoregressive (favar) approach. *The Quarterly Journal of Economics* 120(1), 387–422.
- Caggiano, G., E. Castelnuovo, and N. Groshenny (2014). Uncertainty shocks and unemployment dynamics in us recessions. *Journal of Monetary Economics* 67, 78–92.
- Chanda, K. C. et al. (1996). Asymptotic properties of estimators for autoregressive models with errors in variables. *The Annals of Statistics* 24(1), 423–430.
- Chandrasekaran, V., P. A. Parrilo, and A. S. Willsky (2012). Latent variable graphical model selection via convex optimization. *The Annals of Statistics* 40(4), 1935–1967.
- Eickmeier, S., L. Gambacorta, and B. Hofmann (2014). Understanding global liquidity. *European Economic Review* 68, 1–18.
- Frankel, J. A. (2008). The effect of monetary policy on real commodity prices. *Asset Prices and Monetary Policy*, 291.
- Frankel, J. A. (2014). Effects of speculation and interest rates in a carry trade model of commodity prices. *Journal of International Money and Finance* 42, 88–112.

- Hall, E. C., G. Raskutti, and R. Willett (2016). Inference of high-dimensional autoregressive generalized linear models. *arXiv preprint arXiv:1605.02693*.
- Horn, R. A. and C. R. Johnson (1990). *Matrix analysis*. Cambridge university press.
- Komunjer, I. and S. Ng (2014). Measurement errors in dynamic models. *Econometric Theory* 30(1), 150–175.
- Lin, J. and G. Michailidis (2017). Regularized estimation and testing for high-dimensional multi-block vector-autoregressive models. *Journal of Machine Learning Research* 18(117), 1–49.
- Litterman, R. B. (1986). Forecasting with bayesian vector autoregressions five years of experience. *Journal of Business & Economic Statistics* 4(1), 25–38.
- Loh, P.-L. and M. J. Wainwright (2012). High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity. *The Annals of Statistics* 40(3), 1637–1664.
- Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.
- Lütkepohl, H. (2014). Structural vector autoregressive analysis in a data rich environment. Technical report, Deutsches Institut für Wirtschaftsforschung.
- Melnyk, I. and A. Banerjee (2016). Estimating structured vector autoregressive models. In *International Conference on Machine Learning*, pp. 830–839.
- Negahban, S., B. Yu, M. J. Wainwright, and P. K. Ravikumar (2012). A unified framework for high-dimensional analysis of  $M$ -estimators with decomposable regularizers. *Statistical Science* 27(4), 538–557.
- Nicholson, W. B., D. S. Matteson, and J. Bien (2017). Varx-l: Structured regularization for large vector autoregressions with exogenous variables. *International Journal of Forecasting* 33(3), 627–651.
- Seth, A. K., A. B. Barrett, and L. Barnett (2015). Granger causality analysis in neuroscience and neuroimaging. *Journal of Neuroscience* 35(8), 3293–3297.
- Shojaie, A. and G. Michailidis (2010). Discovering graphical granger causality using the truncating lasso penalty. *Bioinformatics* 26(18), i517–i523.
- Sims, C. A. (1980). Macroeconomics and reality. *Econometrica: Journal of the Econometric Society*, 1–48.
- Stock, J. H. and M. W. Watson (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* 97(460), 1167–1179.
- Stock, J. H. and M. W. Watson (2016). Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. In J. B. Taylor and H. Uhlig (Eds.), *Handbook of Macroeconomics*, Volume 2A, Chapter 8, pp. 415–525. Elsevier.
- Stock, J. H. and M. W. Watson (2017). Twenty years of time series econometrics in ten pictures. *Journal of Economic Perspectives* 31(2), 59–86.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv:1011.3027*.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using  $\ell_1$ -constrained quadratic programming (lasso). *IEEE Transactions on Information Theory* 55(5), 2183–2202.