

# Large Network Autoregressions with Unknown Adjacency Matrix\*

Kenwin Maung

Rutgers University

*Abstract:* Many network econometric models rely on known adjacency matrices. This becomes a problem for investigations when the network structure is not readily accessed or constructed such as those typically observed in macroeconomics and finance. Furthermore, direct estimation may be cumbersome or infeasible if the number of units in the network is large. To deal with this, we propose a Structural Vector Autoregression (SVAR) data-driven approach to recover the network structure via matrix regression under a large  $N$  and  $T$  asymptotic framework. The high-dimensionality of the problem is dealt with by focusing on low-rank representations of the network - hub and authority centralities. We show, both theoretically and through simulations, that the reduced-form estimator is consistent and asymptotically normal, and suggest an identification strategy for the SVAR as implied by its network structure. In our empirical study, we extract volatility connectedness between major US financial institutions and find a greater degree of interconnectedness compared to Diebold and Yilmaz (2014, 2015). We further demonstrate the utility of the estimated network for systemic risk analysis by identifying key propagators of volatility spillovers in the financial sector.

*JEL Classifications:* C13, C33, C51, C55

*Key words:* Large  $N$  and  $T$ , Structural VARs, Spatial models, Network centralities, Composite errors, Matrix-valued time series, Bilinear regression.

---

\*I would like to thank Bin Chen, Nese Yildiz, Michael Gofman, George Alessandria, Elynn Chen and participants from 2024 Dynamic Econometrics Conference, AMES China, SETA 2022, Singapore Economic Review Conference 2022, Society of Economic Measurement Conference 2022, 2022 Rochester Conference in Econometrics, 24th Federal Forecasters Conference, 2022 Economics Graduate Student Conference at WUSTL, and Midwest Econometrics Group 2022 Conference. Any remaining errors are solely mine.

# 1. Introduction

The modeling of networks and cross-sectional relationships has become a major research area in econometrics (Chudik and Pesaran, 2011; Billio et al., 2012; Diebold and Yilmaz, 2009; Zhu et al., 2017; Guðmundsson and Brownlees, 2021). Large networks are particularly important because they manifest where agents interact and thus encode useful information about economic behavior. This scale and type of interaction is only poised to grow with the expansion of social media, global financial integration, trade liberalization, to name a few. Indeed, large networks have been used in studying asset pricing (Ahern, 2013), exchange rates (Richmond, 2019), firm growth (Allen et al., 2019), corporate finance (Gofman and Wu, 2022), among other applications.

To that end, vector autoregressive (VAR) models have proven to be invaluable tools in studying the dynamic interactions between economic units. Under this framework, each entry of the endogenous vector in the VAR represents a cross-sectional unit or a network node. A recent approach is that of the Network Vector Autoregression of Zhu et al. (2017):

$$y_t = \beta_1 W y_{t-1} + \dots + \beta_P W y_{t-P} + v_t \quad (1)$$

where  $y_t$  is a  $N \times 1$  vector of variables,  $W$  is a *known* network adjacency matrix,  $v_t$  is the error term, and  $\beta_1, \dots, \beta_P$  are scalars that measure dynamic network effects. For a general  $W$ , we can see that each  $y_{it}$  ( $i = 1, \dots, N$ ) is influenced by all other units  $y_{jt}$  in the network. The use of a *known adjacency matrix* here is key because it helps to aggregate this cross-sectional relationship by inducing a weighted sum so that instead of estimating  $NP$  coefficients for a given  $i$  (or  $N^2P$  coefficients for the whole system), we only have to estimate a few scalar parameters. As such, many authors require a known adjacency matrix for effective dimension reduction. In the time series context, the Global VAR of Pesaran et al. (2004) and the Infinite-dimensional VAR of Chudik and Pesaran (2011) also employ a similar technique with a prespecified  $W$ . Chen et al. (2020) expand on (1) to allow for community-specific effects. Guðmundsson and Brownlees (2021) propose the Stochastic Block VAR

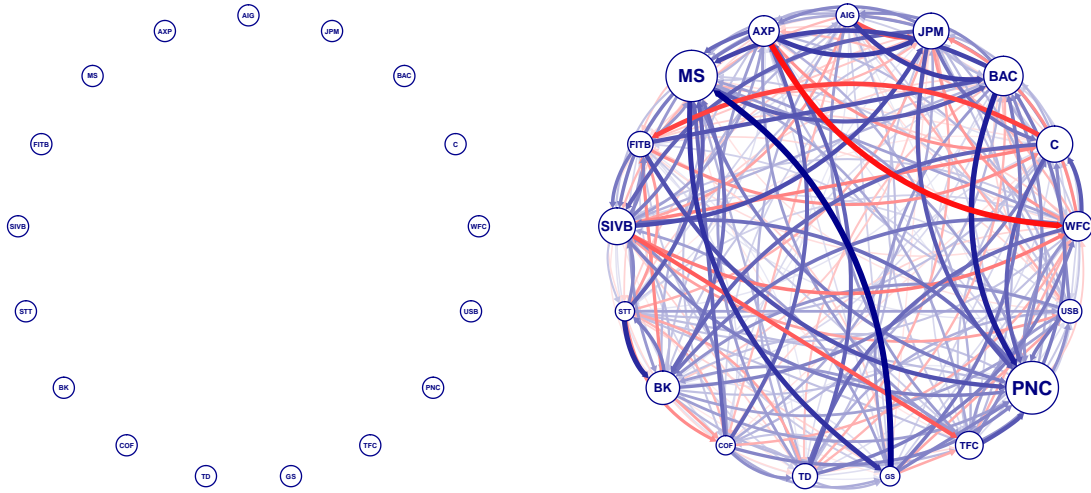
model for when the data is known to be generated from an underlying graph and interpret the autoregressive matrices as functions of a known adjacency matrix. Furthermore, the use of a known adjacency matrix is particularly popular in the spatial econometrics literature (see [Elhorst et al., 2021](#), for a survey).

This approach may be convenient for applications where natural proxies of a network or spatial interactions exist. For example, networks involving geographic units often use some measure of geographic closeness, such as the distance, to proxy exposure to neighboring regions. International trade is another such example where bilateral trade volume can measure connectivity between trading partners. However, these constructions of  $W$  may not readily apply to other applications, such as in studying the financial linkages between assets or monetary policy spillovers between the US and emerging markets. A concrete example of this is that of a volatility network between major financial institutions (FIs). Documenting and analyzing the connectedness between FIs is essential in having a better understanding of financial crises and contagion, and to develop policy tools to safeguard against systemic stress. However, such networks are not observed and it is hard to imagine reasonable proxies that can capture the complex connectivity between FIs as demonstrated in [Figure 1](#).

Even if obvious choices of  $W$  are available, whether by virtue of the data type or economic theory, there remains potentially many degrees of freedom in determining the exact structure of  $W$ . For example, one could construct weights from geographic distances using either a negative exponential decay or a power function ([Corrado and Fingleton, 2012](#)). This then begs the question of which specification to use if indeed the subsequent results are sufficiently different. [Kilian and Lütkepohl \(2017\)](#) argue that it is because of this inherent ad-hocness of weights selection that such models are rarely used in practice.

The objective of this paper is thus to propose a novel way to estimate a generalized version of [\(1\)](#) while treating  $W$  as unknown. Unfortunately, direct estimation of  $W$  without restrictions may be prohibitive when there are many units as it can be of a large dimension ( $N \times N$ ). This is a significant problem as the number of coefficients to be estimated in the VAR diverges at a quadratic rate in  $N$ , leading to parameter proliferation and consistent

Figure 1: Estimated network structure between major US FIs.



Notes: Left figure demonstrates an unobserved network. Right figure presents the estimated network using the method proposed in this paper. Blue indicate a positive link in the adjacency matrix and red indicates negative.

estimation of the large system may no longer be attainable if the sample size  $T$  remains small. Shrinkage methods such as Lasso can be used to uncover the underlying network structure by assuming a sparse network (Manresa, 2013; De Paula et al., 2019; Barigozzi and Brownlees, 2019; Chernozhukov et al., 2021; Miao et al., 2022). These approaches often impose the assumption that only a few entries of  $W$  are non-zero and thus the true model is actually low-dimensional. Under this framework, the key interest would be in recovering the sparse structure of  $W$  using some penalized estimation. Although sparsity might make sense in some applications, it might not be reasonable in others. For example, international trade networks have been observed to be dense (De Benedictis and Tajoli, 2011) and transactions are constantly being made in interbank payment networks (Chapman et al., 2019). As such, we consider here an alternative approach without relying on sparsity. Often, a full estimation of an adjacency matrix is not useful for interpretation because of the numerous linkages involved. Instead, researchers may want to focus on low-rank representations of the network

that reflect the position and influence of some key nodes in the network. Here, we propose to capture a generalized form of eigenvector centralities called "hub and authority centralities" following [Cai et al. \(2021\)](#) and [Allen et al. \(2019\)](#). In our context, this turns out to be an application of rank-reduction to  $W$  so that instead of estimating  $N^2$  adjacency parameters, we require only  $2Nr$  where  $r$  is the reduced-rank of  $W$  (see [Section 2.1](#) for details).

Hence, we consider (1) with an unknown adjacency matrix in a structural VAR (SVAR) under an asymptotic framework where  $N, T, P \rightarrow \infty$ . Specifically, we model

$$y_t = \beta_0 W_0 y_t + \beta_1 W y_{t-1} + \dots + \beta_P W y_{t-P} + \tilde{u}_t \quad (2)$$

where we have  $W$  as an *unknown* low-rank adjacency matrix to be estimated,  $W_0$  is  $W$  with its diagonals set at 0,  $\beta_0$  measures the contemporaneous network, and  $\tilde{u}_t = \tilde{\Lambda} f_t + \tilde{\varepsilon}_t$ . This specification of the error accommodates the dichotomy between latent network-wide common shocks  $f_t$  and the conventional mutually orthogonal idiosyncratic shocks  $\tilde{\varepsilon}_t$ . (2) extends the original Network Vector Autoregression of [Zhu et al. \(2017\)](#) (and [Chen et al. \(2020\)](#)) along these dimensions: (i) most importantly, we drop the requirement for a known adjacency matrix and instead propose to estimate it jointly due to the reasons mentioned above; (ii) we explicitly separate the common and idiosyncratic shocks so that we can study them in isolation when considering model implications such as impulse response analyses; (iii) we generalize the permissible distributions for the error terms to heavy-tailed distributions which is a meaningful departure from the commonly assumed Gaussianity (iv) following much of the spatial econometrics literature, we have incorporated a contemporaneous term to allow for contemporaneous network effects; (v) we allow for the possibility of infinitely many lags. On that note, the latter extension is significant because it improves the flexibility of the model where dynamic spillovers are of particular interest (for e.g. in studying financial contagion)<sup>1</sup>.

---

<sup>1</sup>Furthermore, it is well-known that impulse responses, and subsequently substantive policy implications, can differ greatly depending on the lag order ([Hamilton and Herrera, 2004](#); [Kilian, 2001](#)).

Our contribution is five-fold. Firstly, unlike much of the literature in network and spatial econometrics, we propose a data-driven estimation of the adjacency matrix while allowing for a sufficiently large dimension without relying on sparsity or *a priori* knowledge of the network. The proposed low-rank adjacency matrix approach is closely related to the traditional reduced-rank time series regression of [Velu and Reinsel \(2013\)](#), and our paper extends their theoretical results to the "big data" framework of  $N, T, P \rightarrow \infty$  with parameter matrices of diverging dimensions. Secondly, we extend a nuclear-norm regularization approach ([Chen et al., 2013](#)) for selecting the rank of the adjacency matrix to our context and provide statistical guarantees for its performance. Next, we relax the Gaussian or sub-Gaussian assumption typically required in high-dimensional VAR studies ([Zhu et al., 2017](#); [Chen et al., 2020](#); [Li and Xiao, 2021](#); [Miao et al., 2022](#)) to allow for heavy-tailed distributions along the lines of [Vershynin \(2018\)](#) thus expanding the applicability of our methodology. To the best of our knowledge, the most recent attempt to do so considers sub-Weibull distributions ([Wong et al., 2020](#)) which may still be somewhat restrictive as all moments are required to exist. Additionally, we propose a new identification strategy which involves over-identified estimation of the SVAR to recover the contemporaneous network effects which are often absent in reduced-form treatments such as [Zhu et al. \(2017\)](#) and [Chen et al. \(2020\)](#). Finally, in an application on estimating volatility connectedness between major US FIs, we find a greater degree of interconnectedness compared to the popular network estimation strategy of [Diebold and Yilmaz \(2014\)](#). In addition, we show that our approach dominates many other commonly used procedures in volatility forecasting.

The rest of this paper is organized as follows. [Section 2](#) describes the structural model in greater detail while [Section 3](#) provides an iterated least squares algorithm for estimation. [Section 4](#) contains the main asymptotic results. Monte Carlo experiments are presented in [Section 5](#) to assess the finite sample performance of the proposed estimation, while [Section 6](#) applies it to the problem of estimating volatility connectedness. Finally, [Section 7](#) concludes. All proofs are collected in the appendix.

## 2. Model

Let  $y_t$  be a  $N \times 1$  vector of cross-sectional observations at time  $t$ . We rewrite (2) as:

$$\begin{aligned} (I - \beta_0 W_0)y_t &= \beta_1 W y_{t-1} + \dots + \beta_P W y_{t-P} + \tilde{u}_t \\ &= W \mathcal{Y}_{t-P}^{t-1} \beta + \tilde{u}_t \end{aligned} \quad (3)$$

where

$$\tilde{u}_t = \tilde{\Lambda} f_t + \tilde{\varepsilon}_t. \quad (4)$$

Here,  $\mathcal{Y}_{t-P}^{t-1} = [y_{t-1}, \dots, y_{t-P}]$  is a  $N \times P$  matrix covariate,  $W$  is the unknown  $N \times N$  directed weighted adjacency matrix<sup>2</sup>, and  $\beta = (\beta_1, \dots, \beta_P)^\top$  is  $P \times 1$  and captures the dynamic impact of the network. On the left,  $W_0$  is equivalent to  $W$  but with zeroes imposed on the diagonal and  $\beta_0$  is an unknown scalar parameter representing the contemporaneous network effect. Expressing (3) as a bilinear regression is convenient because it allows us to use estimation tools in the matrix (tensor) regression literature (Zhou et al., 2013; Chen et al., 2021).

The structural error is composed of  $r$  latent common white noise shocks  $f_t = (f_{1t}, \dots, f_{rt})^\top$  while  $\tilde{\varepsilon}_t$  contain  $N$  orthogonal idiosyncratic shocks with variance-covariance matrix normalized to be an identity matrix.  $\tilde{\Lambda}$  is a  $N \times r$  loading matrix. We consider the inclusion of shocks common to all units in addition to the usual orthogonal structural shocks because of the spatio-temporal context of our SVAR. Cesa-Bianchi et al. (2020) consider a similar composite error structure in their panel VAR estimation of multicountry output growth and volatility.

### 2.1. Centralities

One key limitation in the direct estimation of (3) is the large dimensionality of  $W$ . At the same time, it is often not desirable to extract the adjacency matrix in its entirety because

---

<sup>2</sup>The diagonal entries here are non-zero and represent the dynamic effect of a unit on itself.

of difficulties in interpretation especially when the number of nodes,  $N$ , is large. Instead, researchers are often interested in summarizing the network with a few centrality statistics. These statistics highlight the position and importance of specific agents in a network which may be more informative than simply observing an entire network. For example, identifying central units is important because adversarial shocks to them may propagate and amplify more easily throughout the network as opposed to shocks to an isolated node. Given that our SVAR implies a directed network with givers and recipients, we follow [Cai et al. \(2021\)](#) in considering the hub and authority centralities, of which eigenvector centralities are special cases of. In a nutshell, the hub centrality of a given unit  $i$  measures the total authority level of all other units that it links to. On the other hand, the authority centrality of a unit  $j$  is the sum of the hub score of all units it receives links from<sup>3</sup>. Typically, the centralities are estimated as the left and right singular vectors of  $W$  ([Kleinberg, 1999](#)). [Cai et al. \(2021\)](#) propose estimating the centralities from an observed adjacency matrix using a rank one decomposition:

$$\hat{W} = W + E = \tilde{a}\tilde{b}^\top + E$$

where  $\hat{W}$  is the observed matrix,  $E$  is a matrix of errors, and  $W$  is the true weighted adjacency matrix.  $\tilde{a}$  and  $\tilde{b}$  are  $N \times 1$  vectors that contain the hub and authority centralities respectively. Instead of assuming a rank one decomposition as in [Cai et al. \(2021\)](#), we consider a general rank  $r$  decomposition where  $r < N$  so that

$$W = \tilde{a}\tilde{b}^\top \tag{5}$$

and  $\tilde{a}, \tilde{b} \in \mathbb{R}^{N \times r}$ . This implies that our estimation of (3) entails a reduced-rank regression since  $W$  is restricted to rank  $r$ . In the following sections on estimation and asymptotic theory, we assume  $r$  is known. We relax this assumption in section 4.1 and propose a data driven procedure via nuclear norm regularization in selecting  $r$  prior to estimation.

---

<sup>3</sup>See [Cai et al. \(2021\)](#) for a detailed explanation of the centralities and examples

### 3. Estimation

In this section we present an iterated least squares algorithm similar to [Chen et al. \(2021\)](#) to estimate (3) with the rank-reduced hub and authority centralities. We consider also the estimation of the common shocks and loading matrix via principal components analysis à la [Bai \(2009\)](#).

First, rewrite (3) in its reduced form:

$$\begin{aligned} y_t &= (I - \beta_0 W_0)^{-1} W \mathcal{Y}_{t-P}^{t-1} \beta + (I - \beta_0 W_0)^{-1} \tilde{\Lambda} f_t + (I - \beta_0 W_0)^{-1} \tilde{\varepsilon}_t \\ &\equiv A \mathcal{Y}_{t-P}^{t-1} \beta + \Lambda f_t + \varepsilon_t \end{aligned} \quad (6)$$

where  $A = (I - \beta_0 W_0)^{-1} W$ ,  $\Lambda = (I - \beta_0 W_0)^{-1} \tilde{\Lambda}$ , and  $\varepsilon_t = (I - \beta_0 W_0)^{-1} \tilde{\varepsilon}_t$ . Here, note that  $A \mathcal{Y}_{t-P}^{t-1} \beta = (\beta^\top \otimes A) \mathcal{Y}_{t-P}^{t-1}$ . Written this way, it is clear that we have an identification issue as  $(\beta^\top \otimes A) = (c \beta^\top \otimes A/c)$  for any non-zero scalar  $c$ . Hence, for the purpose of estimation, we normalize  $\|A\|_F = 1$ , where  $\|\cdot\|_F$  is the frobenius norm. Given the reduced rank assumption on  $W$ , it is clear that  $A$  is also of the same rank since  $(I - \beta_0 W_0)$  has full rank<sup>4</sup>. Hence, this implies that there exists  $N \times r$  vectors  $a, b$  such that

$$A = ab^\top. \quad (7)$$

Similar to [Chen et al. \(2020\)](#), we treat  $f_t$  as white noise, and hence we can devise a two-step estimation algorithm to solve out problem. In the first step, we ignore the factor structure in  $u_t$  and conduct iterative reduced-rank least squares to estimate  $a, b$  and  $\beta$ . Once we have obtained these estimates, we extract the common factors and factor loading in a second step by conducting principal components analysis on  $y_t - \hat{a} \hat{b}^\top \mathcal{Y}_{t-P}^{t-1} \hat{\beta}$  where the parameters with hats denote estimates from the first step.

---

<sup>4</sup>[De Paula et al. \(2019\)](#) show that for this matrix to be invertible, we require  $|\beta_0| < 1$  and  $\|W_0\| < \infty$ . Like most papers on SVAR modeling, we consider this to be an implicit assumption here.

Before stating the algorithm, we present an optimization result relating to the reduced-rank estimation. This result is a direct application of Theorem 2.2 in [Velu and Reinsel \(2013\)](#) and hence we omit its proof.

**Lemma 1.** *If  $\beta$ , is known, and  $A = ab^\top$  with rank  $r$  known, the solution to*

$$\min_{a,b} \frac{1}{T-P} \sum_{t=P+1}^T \text{tr}\{(y_t - ab^\top \mathcal{Y}_{t-P}^{t-1} \beta)(y_t - ab^\top \mathcal{Y}_{t-P}^{t-1} \beta)^\top\}$$

is given by

$$a^* = V \text{ and } b^{*\top} = V^\top S_{Y^*Y} S_{YY}^{-1}$$

where  $S_{YY} = \frac{1}{T-P} \sum_{t=P+1}^T \mathcal{Y}_{t-P}^{t-1} \beta \beta^\top \mathcal{Y}_{t-P}^{t-1\top}$ ,  $S_{Y^*Y} = \frac{1}{T-P} \sum_{t=P+1}^T y_t \beta^\top \mathcal{Y}_{t-P}^{t-1\top}$ , and  $V$  is a matrix of  $k$  eigenvectors that corresponds to the largest  $k$  eigenvalues of  $S_{Y^*Y} S_{YY}^{-1} S_{Y^*Y}^\top$ .

We will use this result in the first step of our algorithm given initial values of  $\beta^{(0)}$ :

---

**Algorithm 1** Iterated least squares to solve (6)

---

**Require:**  $\beta^{(0)}$ , error tolerance  $e$

- 1: **while**  $\|(\beta^{(m)\top} \otimes a^{(m)} b^{(m)\top}) - (\beta^{(m-1)\top} \otimes a^{(m-1)} b^{(m-1)\top})\|_F > e$  **do**
  - 2:      $a^{(m+1)} \leftarrow V^{(m)}$ ;
  - 3:      $b^{(m+1)\top} \leftarrow V^{(m)\top} S_{Y^*Y}^{(m)} S_{YY}^{-1(m)}$ ;
  - 4:      $\beta^{(m+1)} \leftarrow (\sum_t \mathcal{Y}_{t-P}^{t-1\top} b^{(m+1)} a^{(m+1)\top} a^{(m+1)} b^{(m+1)\top} \mathcal{Y}_{t-P}^{t-1})^{-1} \sum_t \mathcal{Y}_{t-P}^{t-1\top} b^{(m+1)} a^{(m+1)\top} y_t^{(m)}$ ;
  - 5: **end while**
  - 6: Define  $\hat{u}_t = y_t - \hat{a} \hat{b}^\top \mathcal{Y}_{t-P}^{t-1} \hat{\beta}$ , and  $\hat{u}$  as the stacked version of  $\hat{u}_t$ , then the estimator of  $F = (f_{P+1}, \dots, f_T)^\top$  is given by the  $k$  eigenvectors of  $\frac{\hat{u} \hat{u}^\top}{T-P}$  corresponding to the  $k$  largest eigenvalues.
  - 7: Estimate  $\hat{\Lambda} = \frac{\hat{u}^\top \hat{F}}{T-P}$ .
- 

Note that steps 1 to 5 is the first-stage iterative estimation for the model coefficients while  $F$  and  $\Lambda$  are estimated by principal components analysis in steps 6 to 7. From the simulations in Section 5, convergence is attained relatively quickly, typically in less than 10 iterations or so.

### 3.1. Structural identification

In most cases however, the key object of interest is the weighted adjacency matrix which is captured by  $W$  in (3) and the contemporaneous network effect captured by  $\beta_0$ . Given our structure in (3) and its relationship to (6), we propose a straightforward method of extracting these quantities of interest.

Firstly, since  $E(\tilde{\varepsilon}_t \tilde{\varepsilon}_t^\top) = I$  by construction, we have that

$$E(\varepsilon_t \varepsilon_t^\top) = (I - \beta_0 W_0)^{-1} (I - \beta_0 W_0)^{-1\top}.$$

Hence, as usual, we can use the sample variance-covariance matrix of the estimated residuals here. However, this alone would not be enough because the degrees of freedom in  $E(\varepsilon_t \varepsilon_t^\top)$  is  $N(N+1)/2$  while there are  $N^2$  in  $W$  and 1 from  $\beta_0$ . Furthermore, we are unable to estimate the diagonals in  $W$  since only  $W_0$  is used here. We can augment this with the following relationship:

$$A = (I - \beta_0 W_0)^{-1} W.$$

Using  $A$  from the reduced-form model introduces  $N^2$  more degrees of freedom, and we now have a system of overidentified nonlinear equations.

Consider the case where  $W$  can be decomposed into its centralities as  $W = \tilde{a} \tilde{b}^\top$ . The scenario for unrestricted estimation of  $W$  can be dealt with in a very similar way by removing the estimation constraints. Given the reduced-form estimates using Algorithm 1 and the sample variance-covariance matrix of  $\hat{\varepsilon}_t, \hat{\Sigma}$ , we can solve the following constrained system using a non-linear solver<sup>5</sup> to obtain  $\hat{W}$  and  $\hat{\beta}_0$ :

$$\hat{\Sigma} = (I - \beta_0 W_0)^{-1} (I - \beta_0 W_0)^{-1\top}, \tag{8}$$

$$\hat{A} = (I - \beta_0 W_0)^{-1} W, \tag{9}$$

---

<sup>5</sup>For example `Rsolnp` in R.

subject to

$$W = \tilde{a}\tilde{b}^\top,$$

$$\|W\| = 1.$$

Once we have an estimate of  $W$  and  $\beta_0$ , typical statistics of interest such as the impulse response functions and variance decompositions can be constructed using the orthogonal shocks in  $\tilde{\varepsilon}_t$  or the common system-wide shocks in  $f_t$ .

## 4. Asymptotic theory

To provide the statistical guarantees for our proposed estimation of the reduced form model in (6), we require the following regularity conditions.

**Assumption 1 (Composite errors)** (i)  $f_t$  is a zero mean stationary white noise process; (ii)  $\varepsilon_t$  are *i.i.d* with  $\|\varepsilon_t\| = O(\sqrt{N})$  almost surely and  $\sup_{it} E(|\varepsilon_{it}|^{4+\nu}) < \infty$  for some  $\nu > 0$ ; (iii)  $f_t \perp \varepsilon_t$ .

*Remarks.* Assumption 1(ii) is a key assumption here that generalizes the error distribution. It is typical in the high-dimensional VAR literature to impose a specific distribution restriction, such as Gaussian (e.g. [Zhu et al., 2017](#); [Chen et al., 2020](#)) or sub-Gaussian (e.g. [Miao et al., 2022](#)) requirements, on the error term. This is understandable as these restrictions can greatly simplify proofs as many known results, such as concentration and tail inequalities and on covariance estimation, are readily available. Nonetheless, this may leave out some heavy-tailed distributions and could be too restrictive for empirical applications that may require more generality, such as those commonly encountered in finance. [Vershynin \(2018\)](#) show that under the  $\sqrt{N}$  and fourth moment boundedness assumptions, we may recover many of the known results for sub-Gaussian distributions under independence (albeit up to a log rate penalty). We adopt similar conditions here and show that a similar extension can be made for high-dimensional covariance estimation under dependence in [Lemma 3](#). We note that the  $\sqrt{N}$  bound is not particularly restrictive since many commonly used distributions,

such as the Gaussian distribution, satisfy it with high probability. One departure from the suggested assumptions is the strengthening of the moment bound to apply uniformly. This is required for autocovariance estimation of the errors and is not an uncommon assumption in the spatial literature (Lin and Lee, 2010).

**Assumption 2 (Factor structure)** Normalize  $F^\top F/T = I_r$  where  $F = [f_1, \dots, f_T]^\top$  and let  $\Lambda^\top \Lambda$  be a diagonal matrix. Assume

- (i)  $E\|f_t\|^4 \leq C_f < \infty$ ,  $\frac{1}{T-P} \sum_{t=P+1}^T f_t f_t^\top \rightarrow^p \Sigma_f > 0$  as  $T \rightarrow \infty$  for some  $r \times r$  matrix  $\Sigma_f$ ,
- (ii)  $\|\lambda_i\| \leq C_\lambda < \infty$  where  $\lambda_i$  is the  $i^{\text{th}}$  row of  $\Lambda$ ,  $\frac{1}{N} \Lambda^\top \Lambda \rightarrow \Sigma_\lambda > 0$  as  $N \rightarrow \infty$  for some  $r \times r$  matrix  $\Sigma_\lambda$ .

*Remarks.* Assumption 2 is entirely standard and follows that of Bai and Ng (2002). 2(ii) allows for pervasive factors which introduces "strong" cross-sectional dependence into the model (Chudik et al., 2011).

**Assumption 3 (Stationarity)** Normalize  $\|A\|_F = 1$  and define  $\beta_j$  to be the  $j$ -th element of  $\beta$ . Assume the following stability condition holds for  $P \geq 1$ : The roots of the  $P$ -th order polynomial equation

$$z^P - |\beta_1|z^{P-1} - |\beta_2|z^{P-2} - \dots - |\beta_{P-1}|z^1 - |\beta_P| = 0 \quad (10)$$

lie inside the unit circle.

*Remarks.* Note that our stationarity condition is slightly different from that of usual VARs. There are two reasons for this: (i) the normalization on  $A$  suggests that we can direct our attention to disciplining  $\beta$  alone, and (ii) usual sufficient conditions for a stable VAR(P) do not necessarily hold when  $N \rightarrow \infty$ . On the latter point, Chudik and Pesaran (2011) show this using a counterexample that the condition requiring the stacked coefficient matrix, resulting from re-expressing the VAR(P) as a VAR(1), to have all of its eigenvalues to lie in the unit circle is not sufficient to guarantee that  $Var(y_{Nt})$  would be finite as  $N \rightarrow \infty$ . Instead, they recommend binding the spectral norm of the coefficient matrix in their VAR(1) to be strictly less than 1. Unfortunately, this condition does not readily extend to the case

with  $P \geq 2$ . Hence, we rely on Assumption 3 and show in [Proposition 1](#) below that it is sufficient to guarantee stationarity.

**Assumption 4 (Rates)** Allow  $N, T, P \rightarrow \infty$  and let

$$\frac{NP \log(NP)(\log^3((T - P)NP))}{T - P} \rightarrow 0.$$

*Remarks.* Assumption 4 disciplines how fast  $N$  and  $P$  can diverge to infinity relative to the sample size  $T - P$ . This rate ensures successful covariance estimation under a diverging framework with dependence structures and a general error distribution. This rate can be relaxed further if one is willing to accept distributional assumptions. For example, the required rate for Gaussian errors under  $\tau$ -dependence for the data matrices can be relaxed ([Han and Li, 2020](#)) to be

$$\frac{NP}{T - P} \rightarrow 0. \tag{11}$$

**Proposition 1.** *Under Assumptions 1-3, the model in (3) is stable.*

[Proposition 1](#) states that the given assumptions are sufficient for the model given in (6) to be stationary and causal. In other words, for any  $N, T$  and  $P$ , we know that the following infinite-order moving average representation absolutely summable:

$$y_t = \sum_{i=0}^{\infty} \Phi_i u_{t-i}$$

where  $\{\Phi_i\}_{i=0}^{\infty}$  is defined in the appendix. We remark that this result does not depend on any rank condition on  $A$ , but rather only on the convention that  $\|A\|_F = 1$ . This implies that the stability of the system still holds as long as the stated assumptions in [Proposition 1](#) are satisfied.

Next, we consider the consistency of our estimator. Let  $\hat{\beta}$ ,  $\hat{a}$ , and  $\hat{b}$  be the estimates obtained from Algorithm 1.

**Theorem 1.** *Under Assumptions 1-4, we have:*

$$\left\| \hat{\beta}^\top \otimes \hat{a}\hat{b}^\top - \beta^\top \otimes ab^\top \right\| = O_p\left(\sqrt{\frac{NP}{T-P}}\right) = o_p(1).$$

Note that the required rate as stated in Assumption 4 is sufficient for  $\sqrt{NP/T - P}$  to vanish asymptotically, implying consistency in the spectral norm. If the number of lags,  $P$ , remains fixed, the rate reduces to  $\sqrt{N/T}$  which is identical to that of [Li and Xiao \(2021\)](#).

**Theorem 2.** *Let  $L_{N,P}$  be a  $q \times (2Nr + P)$  non-stochastic matrix such that  $L_{N,P}L_{N,P}^\top \rightarrow L$  where  $L$  is some  $q \times q$  non-negative symmetric matrix, for a finite  $q$ . Define  $\hat{\theta} = (\text{vec}(\hat{a}), \text{vec}(\hat{b}), \hat{\beta}^\top)^\top$ . Given Assumptions 1-4, we have:*

$$L_{N,P}\Omega_{Z_t}^{-1/2}\sqrt{T-P}(\hat{\theta} - \theta) \rightarrow^d N(0, L), \quad (12)$$

where

$$\Omega_{Z_t} = \lim_{N,P \rightarrow \infty} \left[ (NP)^{-1}\Sigma_{Z_t} \right] \left[ (NP)^{-2}\Sigma_{Z_t} \otimes \Sigma_u \right] \left[ (NP)^{-1}\Sigma_{Z_t} \right]^\top,$$

for  $\Sigma_{Z_t} = E(Z_t Z_t^\top)$ ,

$$Z_t = \begin{bmatrix} \mathcal{Y}_t^{\beta,b} \otimes I \\ \mathcal{Y}_t^\beta \otimes a^\top \\ I \otimes \mathcal{Y}_t^{a,b} \end{bmatrix}$$

$\mathcal{Y}_t^{\beta,b} = b^\top \mathcal{Y}_{t-P}^{t-1} \beta$ ,  $\mathcal{Y}_t^\beta = \mathcal{Y}_{t-P}^{t-1} \beta$ ,  $\mathcal{Y}_t^{a,b} = \mathcal{Y}_{t-P}^{t-1} b a^\top$ , and  $\Sigma_u = E(u_t u_t^\top)$ .

A few remarks are in order. Firstly, the asymptotic normality result for the reduced rank estimation in part (i) can be viewed as a high-dimensional generalization of the normality result for iterated least squares of reduced rank regressions in [Velu and Reinsel \(2013\)](#). Secondly, we note that the pre-multiplication by  $L_{N,P}$  is simply for ease of interpretation as the dimension of the parameter vector can diverge. Hence, the limiting distributions are presented for arbitrary linear combinations of the underlying parameters.

Our next result shows that the principal component estimator in the second step has limiting mean squared error that converges to 0 in probability.

**Proposition 2.** *Let  $H$  be an invertible matrix and  $\hat{f}_t$  be the estimator of  $f_t$  from the second step. Given Assumptions 1-4, we have*

$$\frac{1}{T-P} \sum_{t=P+1}^T \|\hat{f}_t - H^\top f_t\|^2 = O_p\left(\frac{NP^2}{T-P}\right). \quad (13)$$

We make two remarks on this result. First, note that we require  $f_t$  to be premultiplied by  $H$ . This is necessary because  $\Lambda$  and  $f_t$  are not separately identifiable as  $\Lambda f_t = \Lambda H^{-1\top} H^\top f_t \equiv \bar{\Lambda} \bar{f}_t$ . Next, note that the convergence rate that we have obtained is different from the  $N^{-1} + (T-P)^{-1}$  rate shown in [Fan et al. \(2013\)](#). This stems from the first-stage estimation of  $\hat{\beta} \otimes \hat{a}\hat{b}^\top$ . If the parameters were known prior to principal components analysis, our rate would also be  $N^{-1} + (T-P)^{-1}$ .

#### 4.1. Rank selection

The discussion thus far has been premised on the assumption that the rank of  $W$  (or  $A$ ),  $r$ , is known. In practice, this may not be known a priori. Recently, rank selection with nuclear norm regularization has become increasingly popular in statistics and econometrics ([Yuan et al., 2007](#); [Chen et al., 2013](#)). For example, [Miao et al. \(2022\)](#) utilize the norm to select the number of factors in their factor-augmented VAR, noting that the product of the factors and the loadings correspond to a reduced-rank matrix. We adapt the regularization strategy as suggested in [Chen et al. \(2013\)](#) to our context here while allowing for general errors as opposed to the Gaussian restriction in the original paper.

First, consider the multivariate representation of [\(3\)](#):

$$Y = X \underbrace{(\beta \otimes W)}_{\substack{\equiv Q \\ (NP \times N)}} + U \quad (14)$$

where

$$\underbrace{Y}_{T \times N} = \begin{bmatrix} y_{P+1}^\top \\ \vdots \\ y_T^\top \end{bmatrix}, \quad \underbrace{X}_{T \times NP} = \begin{bmatrix} y_P^\top & \cdots & y_1^\top \\ \vdots & \ddots & \vdots \\ y_{T-1}^\top & \cdots & y_{T-P}^\top \end{bmatrix} \quad \text{and} \quad \underbrace{U}_{T \times N} = \begin{bmatrix} u_{P+1}^\top \\ \vdots \\ u_T^\top \end{bmatrix}.$$

In order to estimate  $\text{rank}(\beta \otimes W) = \text{rank}(\beta)\text{rank}(W) = \text{rank}(W)$ , we optimize the suggested objective function:

$$\frac{1}{2} \|Y - XQ\|_F^2 + \lambda \|XQ\|_* \quad (15)$$

where

$$\| \underbrace{S}_{p \times q} \|_* = \sum_{i=1}^{\min(p,q)} d_i(S)$$

and  $d_i$  refers to the ordered (from largest to smallest) singular values of the matrix  $S$ . The nuclear norm penalty is similar to that of the Lasso penalty in that it aims to shrink singular values down to 0. Since having a matrix  $S$  be of rank  $r$  is equivalent to requiring that the first  $r$  singular values are strictly non-zero while the remaining values are set exactly at zero, the nuclear norm penalty directly recovers the rank of the matrix by identifying the number of non-zero singular values.

Nonetheless, the suggested penalty used in (15) (Chen et al., 2013) is different from the usual nuclear norm penalty of  $\|Q\|_*$  (Yuan et al., 2007). The replacement with  $\|XQ\|_*$  is attractive because as the authors show, will allow for an explicit solution given by:

$$\hat{r} = \max\{i : d_i(X\hat{Q}_{LS}) > \lambda\} \quad (16)$$

where  $\hat{Q}_{LS}$  is the least squares solution to (15). Replacing  $Q$  with  $XQ$  in the norm is justified because  $X$  has full column rank and thus  $\text{rank}(XQ) = \text{rank}(Q)$ . Our next result provides the statistical guarantees for (16) under an additional assumption on the behavior of  $\lambda$ .

**Assumption 5 (Nuclear norm)** Let  $\lambda = O(\sqrt{T} + \sqrt{\log K_N} \sqrt{N \log N})$  where  $N = o(K_N)$  for some sequence of numbers  $K_N$ .

Given this rate condition, the nuclear norm procedure delivers consistent estimates of the

true rank.

**Theorem 3.** *Under Assumptions 1-5, we have  $P(\hat{r} = r) \rightarrow 1$ .*

This result provides the justification of using the nuclear norm approach to select the rank of  $W$  prior to estimation. Hence, one would replace  $r$  in Algorithm 1 with  $\hat{r}$ . Optimization of (15) can easily be done in  $\mathbf{R}$  with the `rrr` package.

## 5. Monte Carlo Simulation

This section studies the finite sample properties of the proposed iterative least-squares algorithms in estimating (6). We consider three experiments.

In the first experiment (**EXP1**), we let  $A = ab^\top$  where  $a = b = (1/\sqrt{N}, \dots, 1/\sqrt{N})$ , which implies that  $A$  is a rank-one matrix with  $1/N$  in every entry. This is equivalent to giving equal weights to every individual in the adjacency matrix. [Kelejian and Prucha \(2002\)](#) and [Lee \(2002\)](#) consider such a matrix in the spatial context in which every individual within a neighborhood is connected to one another and when there is no other intuitive or observable measure of distance.

We relax this characterization of the adjacency matrix by introducing a community structure in the second experiment (**EXP2**). Specifically, we consider the following

$$A = \begin{bmatrix} a_1 b_1^\top & \mathbf{O} \\ \mathbf{O} & a_2 b_2^\top \end{bmatrix}$$

where  $a_1, b_1, a_2, b_2$  are  $N/2 \times 1$  vectors that are drawn from a standard normal distribution, and  $\mathbf{O}$  is a  $N/2 \times N/2$  matrix of zeroes. Clearly, the rank of  $A$  is 2 as both  $a_1 b_1^\top$  and  $a_2 b_2^\top$  are rank 1 matrices. This adjacency matrix limits the network interaction to occur within two groups or two communities. For example, the first unit in the network will only be affected by its own lag and that of units 2 to  $N/2$ , while it is unaffected by units  $N/2 + 1$  to  $N$ .

For both exercises, we consider  $T \in \{300, 600, 1200\}$  and  $N \in \{12, 14, 18\}$ . Since we are considering  $T, N \rightarrow \infty$  (with fixed  $P$  here), we compute the rate stated in Assumption 4

for the  $(T, N)$  pairs  $(300, 12)$ ,  $(600, 14)$  and  $(1200, 18)$ , and obtain the decreasing sequence 54.6, 45.4 and 43.1. Hence, we can study the finite sample performance of our estimator under a double asymptotic framework that is consistent with the required rate in our experiments here. We note that the choice of  $N$  and  $T$  here is motivated by the setup of our empirical application on estimating a volatility network of 17 major FIs from daily data. However, there may be applications where a larger cross-sectional setup with fewer time points is required such as when studying dynamic spillovers between US states with quarterly data. One may thus be interested in the properties of our estimator under a larger  $N$  and smaller  $T$  setup. As such, we consider this framework in a third numerical experiment below.

We fix the number of lags to be 4 and let  $\beta^\top = (0.8, -0.4, 0.2, 0.1)$ . The number of common shocks in  $f_t$  is set at 1 and is drawn from a standard normal distribution and the first  $N/2$  entries of the  $N \times 1$  vector  $\Lambda$  is set at 0.25 while the remaining is -0.25. For the idiosyncratic error term,  $\varepsilon_t$ , we consider two specifications. In the first case, we consider the usual normal errors drawn from  $N(0, \Sigma)$  with the  $(i,j)$ th element of  $\Sigma$  given by  $0.25^{|i-j|}$ . Next, we also consider a "pathological" case with respect to the stated assumptions in [Section 4](#) by drawing the errors from a Student-t distribution with 2.1 degrees of freedom. Although this meant that the distribution has a heavier tail than the normal distribution, which is permissible under Assumption 1(ii), it violates the uniform moment restriction because none of its moments exist beyond that of the variance.

Table 1 reports the mean and standard deviation of  $\|\hat{\beta}^\top \otimes \hat{a}\hat{b}^\top - \beta^\top \otimes ab^\top\|_F$  for the reduced-rank estimation. Looking at both experiments, it appears that the estimation error is declining as  $T$  increases, which is to be expected. More interestingly, we observe the same result when both  $N$  and  $T$  increases (see along the diagonals in Table 1) for both error distributions, which gives us some evidence that our proposed estimator is able to consistently estimate the reduced-rank structure under the double asymptotic framework. This result is surprising particularly for case with t-distributed errors due to the violation of the uniform moment condition. As mentioned in the remarks for Assumption 1, this requirement was needed for consistent estimation of the error covariance matrix under the

Table 1: Simulation results from experiment 1 and 2

<b>EXP1</b>							
Normal	N			Student t (2.1)	N		
	12	14	18		12	14	18
<b>T</b>							
300	0.82 (0.21)	0.93 (0.21)	1.21 (0.25)		1.81 (1.11)	2.10 (0.94)	2.78 (0.88)
600	0.56 (0.13)	0.66 (0.14)	0.83 (0.16)		1.07 (0.29)	1.32 (0.79)	1.66 (0.38)
1200	0.39 (0.09)	0.44 (0.09)	0.58 (0.11)		0.68 (0.18)	0.84 (0.21)	1.09 (0.24)
<b>EXP2</b>							
Normal	N			Student t (2.1)	N		
	12	14	18		12	14	18
<b>T</b>							
300	1.28 (0.27)	1.21 (0.34)	1.50 (0.37)		1.87 (1.19)	3.03 (1.20)	3.74 (1.15)
600	0.52 (0.17)	0.57 (0.15)	1.21 (0.20)		1.17 (0.62)	1.51 (0.93)	2.05 (0.59)
1200	0.32 (0.06)	0.41 (0.09)	0.43 (0.12)		0.78 (0.34)	0.63 (0.34)	1.14 (0.59)

Notes: Results from 500 iterations showing the mean of  $\|\hat{\beta}^\top \otimes \hat{a}\hat{b}^\top - \beta^\top \otimes ab^\top\|_F$  for both error distributions: normal or Student-t with 2.1 degrees of freedom. Values in parentheses report the standard deviations. Detailed description of EXP1 (equally-weighted adjacency matrix) and EXP2 (two communities model) can be found in [Section 5](#).

asymptotic framework that  $N, T \rightarrow \infty$  such that  $N/T \rightarrow 0$  (Han and Li, 2020). Since  $N$  increases slowly relative to  $T$  in our experiments here, we hypothesize that departures from the uniform moment bound can be entertained for some heavy-tailed distributions provided that the variance exists as suggested by these results.

## 5.1. Rank selection

The preceding experiments were conducted with the assumption of a known rank,  $r$  (1 in EXP1 and 2 in EXP2). We now test for the ability of the suggested nuclear-norm regularization approach introduced in Section 4.1 to select the rank of  $A$  in EXP2.

Table 2: Share of iterations with correctly selected rank

Rank selection	N			Student t (2.1)	N		
	12	14	18		12	14	18
<b>T</b>							
300	0.93	0.68	0.30		0.19	0.20	0.17
600	0.87	0.95	0.97		0.61	0.59	0.35
1200	0.99	0.96	0.95		0.78	0.68	0.50

Notes: Share of correctly selected rank out of 500 iterations for EXP2. Details of the nuclear norm regularization can be found in Section 4.1 details of EXP2 (two communities model) can be found in Section 5.

As we can see in Table 2, the nuclear norm selection procedure performs favorably for normally distributed errors particularly when the number of time points is large. However, when we look at the case with Student-t errors, the procedure appears to struggle. Unlike our discussion earlier, this result is not related to the uniform moment bound condition. In fact, strictly speaking, only the existence of second moments is required for the consistency of the selected rank in Theorem 3. We do indeed see some evidence of that when  $T$  is large. The emergent question then is this procedure valid when  $T$  is relatively small. To answer this, we look at the situations when an incorrect rank was selected. Focusing on the Student t case when  $T = 300$ , we report that the percentage of selecting a rank greater than 2 for

$N = 12, 14$  and  $18$  is  $81\%$ ,  $80\%$ , and  $83\%$  respectively. This means that a larger rank was suggested for the majority of the cases with a wrongly selected rank. This is akin to ignoring the lower-rank constraint and is not too concerning of an issue given that fewer restrictions would be imposed when estimating the model. The trade-off then would be the requirement for a larger sample size to ensure computationally feasible estimation.

## 5.2. Larger $N$ , smaller $T$

This section considers an experiment (**EXP3**) with a larger number of cross-sectional units  $N$  and fewer time points  $T$ , which brings our model closer to that of traditional panel data analysis. We consider a setup with a rank 1 adjacency matrix  $A = a_1 b_1^\top$  where  $a_1, b_1$  are standard normal vectors, with 2 lags and  $\beta = (0.2, -0.2)^\top$ . The errors are drawn from a normal distribution identical to that in the previous section. Our experiment involves the following sample sizes:  $T \in \{100, 150, 200\}$  and  $N \in \{40, 50, 60\}$ .

We remark here that as  $N, T \rightarrow \infty$ , we are unable to demonstrate the performance of the estimator under the required rate in Assumption 4. Nonetheless, as mentioned earlier, we can circumvent this issue by imposing some structure on the error terms. Specifically, the design of normal errors allows us to relax the rate to that of (11), which is satisfied by our present choices of  $N$  and  $T$ . This discussion thus demonstrates the trade-off between the available sample sizes and the desired generality of the error process.

To situate the performance of the reduced-rank estimator in this high-dimensional framework, we consider the Lasso as a competitor within the structure of (3). The estimation is similar to the iterative estimation of Algorithm 1, but instead of estimating the reduced-rank matrices  $a, b$  in steps 2-3, replace it with an equation-by-equation Lasso step (as one might do in a linear VAR setting) to recover  $A$ .

Table 3 reports the mean squared error of both approaches. We first remark that the estimation error appears to be higher than that of EXP1 and EXP2 but this is to be expected given the larger number of parameters involved. Nonetheless, compared to the Lasso, the reduced rank method appears to approximately halve the estimation error. One possible

Table 3: Simulation results from experiment 3

RR	N			Lasso	N		
	40	50	60		40	50	60
<b>T</b>							
100	5.54 (0.92)	-	-	10.70 (0.67)	-	-	
150	3.93 (0.53)	5.18 (0.69)	-	7.29 (0.39)	9.69 (0.50)	-	
200	3.16 (0.35)	4.11 (0.43)	5.15 (0.60)	5.72 (0.28)	7.48 (0.32)	9.34 (0.41)	

Notes: Results from 500 iterations showing the mean of  $\|\hat{\beta}^\top \otimes \hat{a}\hat{b}^\top - \beta^\top \otimes ab^\top\|_F$  for EXP3 for both the proposed reduced rank estimator (RR) and the Lasso. Values in parentheses report the standard deviations.

reason for this result is that the matrix  $A$  is not sparse and Lasso might incorrectly penalize some coefficients to zero, thereby resulting in a worse performance.

## 6. Empirical application: Volatility connectedness

In this section, we apply the proposed methodology to study the volatility connectedness between FIs and its implications on systemic risk, in the spirit of [Diebold and Yilmaz \(2014\)](#), [Barigozzi and Brownlees \(2019\)](#), and [Miao et al. \(2022\)](#). On a fundamental level, we expect large FIs to be well-connected for at least three reasons ([Diebold and Yilmaz, 2014](#)): counterparty linkages through asset positions, financial services offered between institutions, and through various deals such as share or asset deals. Our experience with the financial crisis of 2007-2008 greatly highlights the importance of documenting and analyzing the connectedness between FIs in order to have a better understanding of financial crises and contagion, and to inform and develop policy tools to safeguard against similar catastrophic events.

We focus on volatility here for several reasons. Firstly, as we are concerned with systemic risk, volatility is a pertinent metric because it is particularly sensitive to crises. Specifically, volatility is often reflective of investor fear as exemplified by how the VIX is commonly

referred to as the fear index or gauge. By studying the connectedness patterns in volatility, we can map out how this sentiment spillovers from asset to asset. Furthermore, volatility is particularly suited for our methodology here because it tends to exhibit strong serial correlation. This is readily accommodated in our model because we allow for a large number of lags.

Unfortunately, volatility is not directly observable. Hence, we need to proxy for it. Following [Engle et al. \(2012\)](#) and [Barigozzi and Brownlees \(2019\)](#), we use the high-low daily range ([Parkinson, 1980](#)) given below as our volatility measure:

$$0.361 \left( \log(p_{it}^{high}) - \log(p_{it}^{low}) \right)^2 \tag{17}$$

where  $p_{it}^{high}$  and  $p_{it}^{low}$  represent the maximum and minimum stock price of institution  $i$  on day  $t$ . [Diebold and Yilmaz \(2014\)](#) argue that volatility tends to be asymmetrically distributed with a right skew, and suggest using the log transformation, as in (17), to attain approximate normality. Note that since our methodology imposes neither a Gaussianity (or sub-Gaussianity) nor symmetric assumption, our results would be robust to situations when the log transformation fails to induce normality. Nonetheless, we use this measure in our application here as it is parsimonious and easily constructed. Furthermore, [Engle and Gallo \(2006\)](#) show that the range has good explanatory power in predicting future values, while [Brownlees and Gallo \(2006\)](#) demonstrate that it yields excellent forecasting performance relative to high-frequency measures. We do have to note here that our methodology is independent of the choice of volatility proxy, and the researcher may choose alternative proxies if they wish to do so.

We construct the high-low range for 17 major FIs in the US using daily stock return data from 01/02/2009 to 08/31/2022, obtainable from Yahoo! Finance. Our sample of FIs include 1 insurance firm, 1 credit card company, 2 investment banks, and commercial banks for the remaining institutions. A full description of our sample is provided in [Table 4](#), along with summary statistics<sup>6</sup>.

---

<sup>6</sup>We do not report the mean here because it is effectively 0 for every institution. We report the median

Table 4: Summary statistics of the high-low range of 17 major US FIs

<b>Institution</b>	<b>Median</b>	<b>Std. Dev.</b>	<b>Min</b>	<b>Max</b>
American International Group (AIG)	-0.13	1.07	-2.84	5.02
JPMorgan Chase & Co (JPM)	-0.14	1.23	-3.42	4.62
Bank of America Corp (BAC)	-0.08	1.26	-3.82	3.89
Citigroup Inc (C)	-0.09	1.07	-3.21	4.93
Wells Fargo & Co (WFC)	-0.05	1.00	-2.79	3.93
US Bancorp (USB)	-0.07	1.05	-2.88	4.20
PNC Financial Services Group Inc (PNC)	-0.10	1.23	-3.30	4.28
Truist Financial Corp (TFC)	-0.09	1.10	-2.92	4.31
Goldman Sachs Group Inc (GS)	-0.12	1.08	-3.00	4.28
TD Bank (TD)	-0.04	0.98	-3.43	4.32
Capital One Financial Corp (COF)	-0.15	1.16	-3.52	4.20
Bank of New York Mellon Corp (BK)	-0.07	1.05	-2.70	3.99
State Street Corp (STT)	-0.06	1.07	-3.73	3.90
SVB Financial Group (SIVB)	-0.23	2.00	-4.18	5.74
Fifth Third Bancorp (FITB)	-0.11	1.12	-2.89	3.96
Morgan Stanley (MS)	-0.17	1.20	-3.44	3.81
American Express Company (AXP)	-0.22	1.28	-3.61	5.20

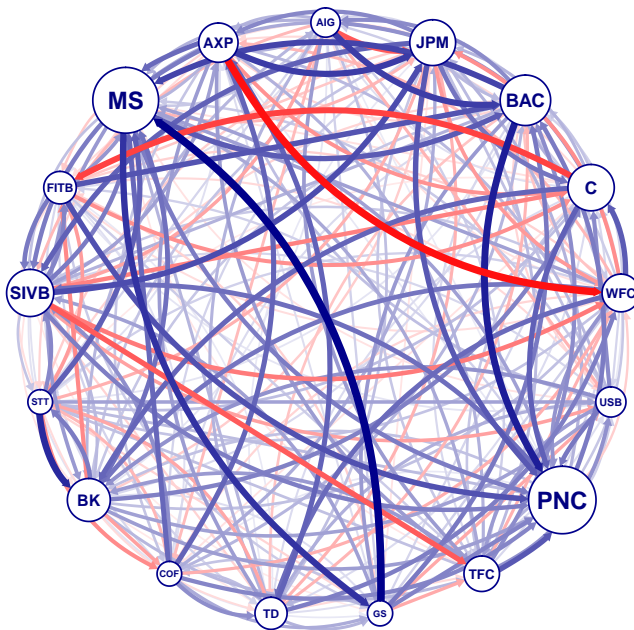
Notes: Daily high-low range constructed following (17).

Before estimating the reduced-rank adjacency matrix, we need to know the rank. We thus apply the nuclear-norm regularization introduced in [Section 4.1](#) as a data-driven approach to select the rank.

## 6.1. Full sample analysis

Figure 2 plots the estimated adjacency matrix,  $W$ . The color of the directed links indicate the sign of the coefficient in the estimated matrix while the width of the links are representative of the magnitude of the weights. We use the "TO" scores (to be defined below) to construct the relative differences in node sizes with the goal of visually representing major propagators of volatility spillovers. Intuitively speaking, the larger the node size, the greater its degree of influence on other institutions.

Figure 2: Estimated adjacency matrix.



Notes: Blue links indicate positive weights, while red links indicate negative weights. Width of arrows is proportional to the magnitude of weights.

---

instead as a measure of central tendency.

Immediately, we can see that the volatility measures of FIs are highly interrelated and lends support to the idea that idiosyncratic shocks to a single institution can easily propagate throughout the system. Nonetheless, the figure is a powerful analytical and exploratory tool on two levels. Firstly, it visualizes the major propagators of volatility spillovers so that institutions of systemic importance are immediately identifiable. In this context, both MS and PNC appears to be the largest originators of spillovers. Secondly, we are also able to identify the institutions that are most exposed to these systemically important actors which is useful in teasing out the vulnerable institutions in the network.

To see this more precisely, we construct a spillover table as in [Diebold and Yilmaz \(2014\)](#). To do so, we first row-normalize the estimated coefficient matrix to sum to 100<sup>7</sup>. The resulting matrix is presented in Table 5. To interpret this table, suppose we fix an institution  $i$  and look across the columns (left to right). Excluding the diagonal term, these values reveal the impact of other institutions on institution  $i$  and we call them the "From" scores. Next, fixing an institution  $i$  again, we look across the rows (up to down), and excluding the diagonal term, it shows the influence of institution  $i$  on the other institutions in the network, and are termed the "To" scores. For each institution, we can construct the aggregate of "To" (and "From") scores by summing them up across institutions except its own. These aggregate scores are reported in the lower panel of Table 5. The "Net" score is obtained as "To" - "From".

To contextualize our results, we compare them with the network estimation strategy of [Diebold and Yilmaz \(2014\)](#) (Henceforth termed "DY"). In a nutshell, the proposed algorithm there is to estimate a linear reduced-form VAR with the purpose of constructing a generalized forecast error variance decomposition (FEVD). The authors then interpret the resulting FEVD matrix as the adjacency matrix. We report the "From" and "To" scores using their method in Table 5 in parentheses.

We notice a stark difference when comparing with their results. Under our framework, the

---

<sup>7</sup>Here we are interested in the magnitudes of the spillovers, hence, we take the absolute values of the coefficients before the normalization procedure.

Table 5: Spillover table

	From	AIG	JPM	BAC	C	WFC	USB	PNC	TFC	GS	TD	COF	BK	STT	SIVB	FITB	MS	AXP
To	AIG	<b>34.3</b> ( <b>29.8</b> )	10.2 (4.5)	10.6 (7.4)	1.5 (5.8)	3.8 (3.4)	5.6 (3.9)	1.1 (4.6)	5.1 (5.5)	0.6 (4.1)	1.6 (2.5)	5.2 (4.1)	4.7 (5.4)	0.1 (3.5)	5.5 (2.3)	0.6 (3.8)	4.8 (5.7)	4.7 (3.6)
	JPM	2.3 (2.6)	<b>8.9</b> ( <b>14.9</b> )	6.7 (8.3)	6.1 (7.5)	6.2 (5.6)	1.6 (5.7)	12.5 (7.3)	0.5 (5.8)	3.8 (6.3)	2.2 (2.6)	3.6 (4.5)	5.6 (5.1)	3.1 (4.0)	15.8 (4.4)	3.5 (4.3)	10.1 (7.1)	7.6 (3.1)
	BAC	7.4 (2.9)	3.8 (8.5)	<b>12.6</b> ( <b>16.2</b> )	4.0 (8.0)	5.4 (5.5)	1.8 (5.1)	14.5 (7.5)	2.5 (5.6)	7.1 (5.8)	0.8 (2.0)	1.7 (4.5)	6.9 (5.3)	1.8 (4.0)	12.3 (3.8)	0.8 (4.9)	12.2 (7.3)	4.5 (2.6)
	C	5.6 (3.4)	6.7 (8.4)	3.0 (8.8)	<b>24.1</b> ( <b>18.7</b> )	1.6 (5.1)	1.3 (5.2)	8.9 (6.9)	2.9 (5.7)	1.6 (5.9)	1.1 (2.2)	1.0 (4.3)	11.7 (5.2)	2.3 (4.0)	5.7 (2.9)	18.6 (3.9)	1.2 (7.0)	2.8 (2.6)
	WFC	1.8 (2.6)	12.4 (7.3)	2.7 (6.9)	7.8 (6.4)	<b>16.6</b> ( <b>17.3</b> )	1.2 (6.8)	4.1 (7.4)	5.7 (5.5)	6.9 (5.0)	13.8 (2.5)	2.9 (4.6)	6.5 (5.9)	1.3 (3.9)	0.1 (4.0)	3.0 (4.9)	6.8 (6.0)	6.4 (3.0)
	USB	3.7 (2.2)	7.9 (7.4)	3.2 (6.5)	6.4 (6.2)	2.0 (6.5)	<b>5.9</b> ( <b>15.3</b> )	12.3 (7.8)	5.7 (7.2)	0.8 (5.0)	7.6 (2.7)	1.5 (4.9)	7.2 (5.6)	1.8 (3.9)	13.9 (4.0)	3.7 (5.1)	9.6 (6.0)	6.9 (3.6)
	PNC	1.4 (2.3)	6.1 (7.6)	10.7 (7.5)	4.0 (6.1)	3.9 (6.0)	0.9 (6.6)	<b>16.0</b> ( <b>15.4</b> )	2.7 (7.3)	5.3 (5.2)	2.3 (2.2)	2.1 (4.6)	6.3 (5.2)	2.0 (3.8)	17.9 (4.8)	2.2 (5.8)	11.9 (6.2)	4.2 (3.2)
	TFC	4.0 (2.6)	6.8 (7.1)	11.9 (7.1)	3.5 (6.2)	8.1 (5.1)	7.2 (7.1)	14.0 (8.5)	<b>13.8</b> ( <b>15.4</b> )	2.5 (5.0)	1.7 (2.1)	1.0 (4.7)	6.7 (5.4)	3.4 (3.8)	0.8 (4.2)	2.2 (6.0)	7.8 (6.2)	4.6 (3.5)
	GS	4.7 (2.7)	3.4 (8.1)	6.5 (7.2)	4.8 (6.7)	3.3 (5.1)	5.7 (4.5)	7.7 (6.1)	6.2 (4.4)	<b>11.5</b> ( <b>17.9</b> )	7.4 (2.4)	2.5 (4.8)	6.6 (4.6)	1.3 (4.2)	5.9 (3.9)	4.9 (3.6)	13.1 (10.1)	4.5 (3.6)
	TD	0.1 (2.5)	7.8 (5.1)	1.5 (4.7)	2.2 (4.4)	10.8 (4.3)	0.5 (4.7)	0.8 (4.7)	4.2 (4.2)	7.1 (4.2)	<b>28.6</b> ( <b>30.2</b> )	5.7 (4.8)	3.3 (3.9)	8.5 (3.0)	3.2 (3.5)	3.4 (3.7)	5.3 (6.0)	7.1 (5.7)
	COF	1.5 (2.7)	7.6 (6.7)	6.3 (6.1)	0.8 (5.4)	7.3 (4.6)	0.1 (5.3)	10.4 (6.2)	1.9 (5.6)	5.4 (5.5)	7.5 (2.7)	<b>5.8</b> ( <b>17.9</b> )	2.0 (4.8)	5.5 (3.8)	13.6 (4.8)	1.8 (5.1)	10.8 (6.9)	11.8 (6.0)
	BK	1.5 (2.5)	4.6 (6.9)	1.5 (6.5)	5.6 (5.8)	3.9 (5.8)	4.0 (6.0)	5.3 (6.7)	2.8 (5.9)	3.6 (5.0)	7.6 (1.9)	7.7 (4.0)	<b>13.8</b> ( <b>17.5</b> )	11.1 (7.1)	3.8 (4.3)	10.1 (4.1)	10.8 (6.9)	2.3 (3.2)
	STT	0.8 (2.4)	6.2 (6.4)	2.4 (6.1)	1.5 (5.9)	2.2 (4.9)	3.5 (5.2)	4.6 (6.0)	2.4 (5.2)	7.1 (5.6)	8.0 (1.9)	7.1 (3.7)	14.2 (8.5)	<b>11.9</b> ( <b>19.5</b> )	3.3 (4.5)	9.3 (3.9)	14.4 (7.2)	0.9 (3.2)
	SIVB	5.4 (1.3)	8.7 (6.1)	0.8 (4.8)	5.3 (3.2)	6.2 (4.2)	3.6 (4.3)	6.8 (6.4)	6.8 (5.1)	0.2 (4.6)	2.0 (1.9)	2.6 (5.1)	2.1 (4.3)	1.4 (5.1)	<b>31.5</b> ( <b>27.9</b> )	5.7 (6.1)	5.9 (5.3)	4.9 (4.1)
	FITB	2.3 (2.2)	4.2 (6.2)	10.8 (7.4)	9.2 (5.4)	0.8 (5.6)	5.6 (6.1)	9.7 (8.2)	6.3 (7.6)	4.8 (4.8)	8.1 (2.2)	0.2 (5.2)	0.9 (4.8)	2.1 (3.8)	9.7 (5.0)	<b>16.6</b> ( <b>16.7</b> )	7.6 (5.6)	1.1 (3.4)
	MS	0.3 (2.6)	5.9 (7.4)	4.8 (7.6)	2.7 (6.7)	1.0 (4.9)	0.0 (4.8)	6.2 (6.3)	1.8 (5.0)	15.8 (8.7)	4.7 (2.2)	0.7 (4.8)	10.7 (5.5)	4.2 (5.0)	3.4 (3.8)	6.5 (3.6)	<b>20.7</b> ( <b>17.3</b> )	10.4 (3.9)
	AXP	7.1 (2.6)	7.4 (6.9)	2.1 (5.1)	8.6 (4.0)	14.3 (3.4)	2.6 (4.9)	0.3 (5.6)	6.1 (5.3)	0.4 (5.3)	5.4 (3.2)	3.5 (6.9)	0.2 (4.7)	2.0 (3.9)	8.5 (4.5)	1.5 (4.3)	8.7 (6.6)	<b>21.2</b> ( <b>22.3</b> )
To		49.9 (40.2)	109.8 (110.6)	85.6 (108.1)	73.9 (93.5)	80.7 (79.9)	45.4 (86.3)	119.0 (106.2)	63.6 (91.7)	73.1 (86.2)	81.8 (37.0)	49.1 (75.4)	95.4 (84.3)	51.9 (66.8)	123.3 (64.5)	77.6 (73.1)	141.0 (106.1)	84.8 (59.8)
From		65.7 (70.2)	91.1 (85.1)	87.4 (83.8)	75.9 (81.3)	83.4 (82.7)	94.1 (84.7)	84.0 (84.6)	86.2 (84.6)	88.5 (82.1)	71.4 (69.8)	94.2 (82.1)	86.2 (82.5)	88.1 (80.5)	68.5 (72.1)	83.4 (83.3)	79.3 (82.7)	78.8 (77.7)
Net		-15.7 (-30.0)	18.7 (25.6)	-1.8 (24.2)	-2.0 (12.2)	-2.7 (-2.8)	-48.7 (1.6)	35.1 (21.7)	-22.5 (7.1)	-15.4 (4.1)	10.4 (-32.8)	-45.0 (-6.7)	9.3 (1.7)	-36.2 (-13.7)	54.8 (-7.6)	-5.8 (-10.1)	61.7 (23.3)	6.0 (-17.8)

Notes: Upper panel presents the row-normalization (sum to 100) of the estimated adjacency matrix,  $W$ . The construction and interpretation of the "To" and "From" scores in the lower panel are described in the text. Numbers in parentheses indicate estimates obtained from the FEVD-based adjacency matrix of [Diebold and Yilmaz \(2014\)](#).

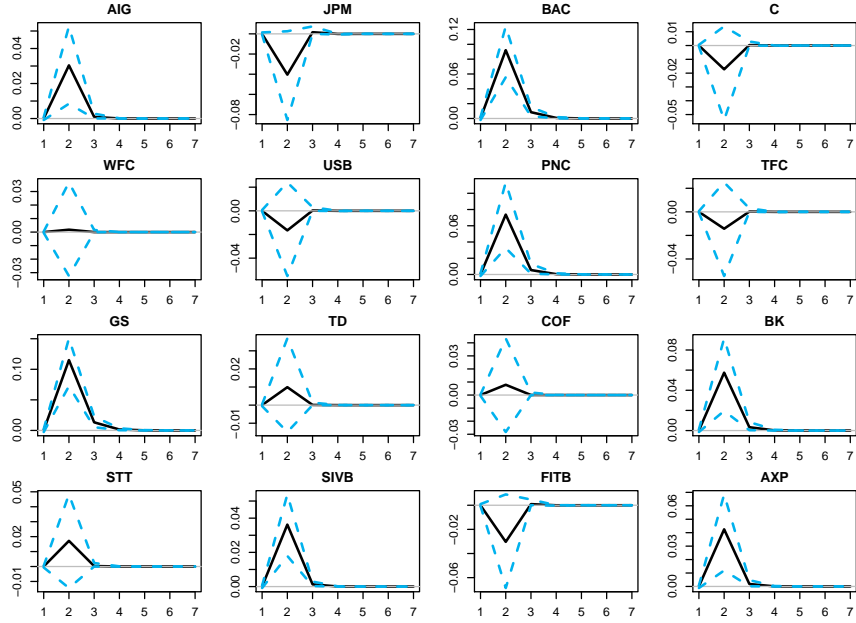
diagonal terms of the adjacency matrix, which represents the dynamic effect of an institution's past on itself, are mostly (12 out of 17) smaller than the estimates obtained from DY. For example, if we look at COF, our estimate yields a value of 5.8 while it is three times greater under DY's framework. The key implication of this is that for a majority of institutions in our sample, their asset volatilities are, relative to the results in DY, more affected by the volatilities of other institutions. By symmetry, it means that there are some institutions that have an oversized impact on others.

To see this heterogeneity in greater detail, we focus on the bottom panel of Table 5. If we take the standard deviation of all the "From" scores obtained from our model we get a value of 8.2, while we get a value of 4.9 under DY's framework. We observe a similar result for the "To" scores, with a standard deviation of 27 under our framework compared to a value of 21.4 from DY. Essentially, this means that, relative to DY, the importance of institutions in propagating and receiving volatility spillovers under our framework is more heterogeneous. We can observe a visual representation of this in Figure 2 as the nodes sizes which indicates the importance of an institution in propagating volatility spillovers (as they were constructed using the "To" scores) vary significantly across institutions. The key implication is that our results yield a landscape with a few important systemic actors as compared to the results obtained using DY where the importance of institutions becomes relatively more homogeneous.

## 6.2. Impulse response analysis

Recall that we have decomposed the structural error into an idiosyncratic and common component. This allows us to study in isolation the impact of a shock to an institution and to observe its propagation through the estimated network. In the context of our application, this tool is particularly useful for determining the risks posed by major propagators of volatility spillovers.

Figure 3: Impulse response to a 1 s.d. positive shock to MS.



Notes: Bold line indicates estimated impulse response while blue lines indicate a bootstrapped 95% confidence interval.

Hence, we study the impacts of a positive idiosyncratic shock<sup>8</sup> (an increase in volatility) to MS. Figure 3 plots the propagation of the positive shock throughout the network over a 7 day horizon.

Here, we see that the institutions most exposed to a positive shock to MS’s volatility are GS and BAC. This result is not particularly surprising given that both are large banks. The former is also an investment bank, hence we should expect a great degree of interconnectivity with MS. Some regional banks, such as PNC and SIVB, are also significantly impacted. Although some institutions react negatively (e.g. C and TFC) in response to the positive shock, none of these appear to be statistically different from 0.

---

<sup>8</sup>One standard deviation shock.

### 6.3. Rolling sample analysis

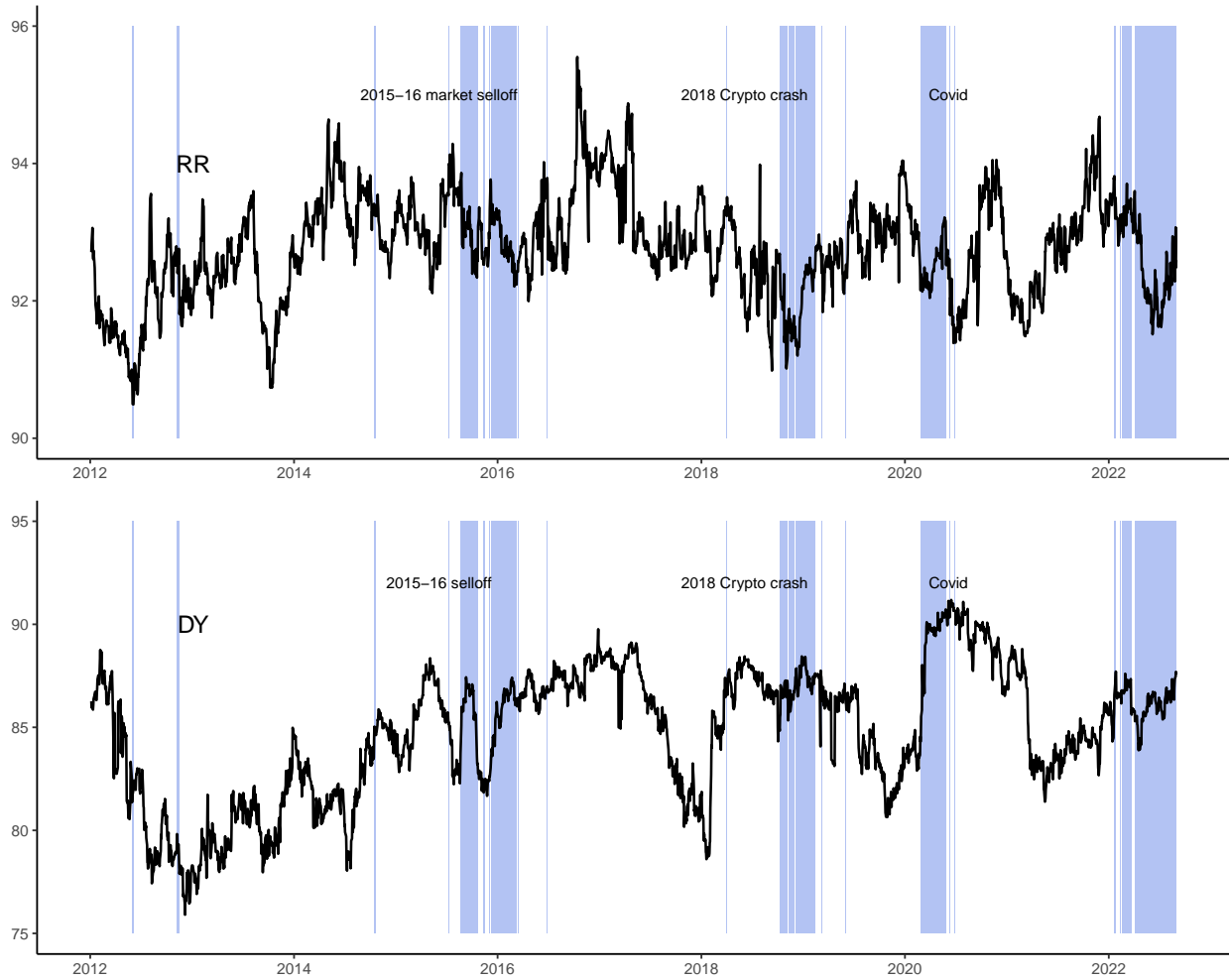
Following [Diebold and Yilmaz \(2014\)](#), we can also apply our model for dynamic analysis with a rolling setup and to compare it with that of DY. In particular, we consider a similar rolling window period of 200 days. Here, we emphasize the merits of our proposed methodology in dimension reduction. Given the relatively small sample size and the large number of institutions, we are unable to reduce the rolling window to be any smaller (e.g. 100 days) because DY starts struggling with the estimation as their model is essentially an unrestricted reduced-form VAR. On the other hand, the reduced-rank estimator had no issues with the smaller sample size. To remain fair, we consider a window of 200 days.

Here, we study the performance of the spillover index over time, which is defined to be the average of either the "From" or "To" scores<sup>9</sup>. [Figure 4](#) displays the evolution of the spillover index estimated from our reduced-rank model (RR) and under DY. The shaded regions are data-driven estimation of bear markets obtained by comparing the daily SP500 index with its trailing 200-day moving average. We observe two stark differences. Firstly, RR produces a spillover index that is on average higher than that of DY throughout the whole sample period. This coheres with our finding in [Section 6.1](#) that our model suggests a greater degree of interconnectedness compared to DY. Furthermore, when looking at the evolution of the spillover index during periods of financial stress, RR responds in a relatively more muted fashion while the response of DY is significantly more pronounced. For example, we observe a spike in the spillover index of DY during the acute months of Covid-19 during early 2020, while the index derived from RR does not appear to react as much. This is likely attributable to the explicit inclusion of common factor shocks in our model. Hence, the common factor would capture the "Covid" shock. On the other hand, DY does not explicitly model for this and it is thus unclear whether their spillover index is reflective of volatility responses to changes in the volatility of other institutions or if the change is in response to a market wide shock.

---

<sup>9</sup>The averages are equivalent.

Figure 4: Spillover index estimated from a 200-day rolling window analysis.



Notes: Bold line indicates the spillover index (average of either "From" or "To" scores) for RR (proposed model) and DY (Diebold and Yilmaz (2014)). Shaded areas indicate bear markets as identified by comparing the daily SP500 index with its trailing 200-day moving average.

## 6.4. Forecasting

At the end of the day, the proposed SVAR model is a time series model that can be used for forming forecasts. Hence, we consider the one-period ahead forecasting of the volatility measure for the 17 financial institutions. A key difference from that of traditional univariate approaches is that we explicitly consider network interactions here, which allows for direct modeling of the spillover effects.

We compare the performance of our model to 4 other approaches. The first approach is the underlying model of DY which is an unrestricted VAR. The second method is that of a VAR(1) with Lasso regularization. This model is related to the NETS algorithm of [Barigozzi and Brownlees \(2019\)](#) in which the coefficient matrix of the VAR(1) is interpreted as the adjacency matrix. A variant of the aforementioned approach is a VAR(1) with Ridge penalties. Lastly, we consider the parsimonious univariate AR(1) as a benchmark.

We use the first 3 years in our sample (01/02/2009 to 12/31/2011) as the training set to estimate the models. Given the estimated model, we conduct our out-of-sample forecasting exercise for the period of 01/02/2012 to the end of the sample. Note that the out-of-sample period includes several bear markets including the volatile period of the acute Covid months. Our accuracy metric is given by the Mean Squared Forecast Error (MSFE) which is the squared discrepancy between the forecast and the realized value, averaged over the entire out-of-sample period.

Table 6 reports the MSFE for the different approaches. We can see that the forecasting performance of RR appears to dominate other forecasting models for a majority of assets (13 out of 17). This forecasting superiority compared to the second-best forecast is statistically significant, as indicated by the Diebold-Mariano test ([Diebold and Mariano, 2002](#)), for most of the results. The unrestricted VAR(P) and AR(1) are both close runner-ups. We remark that the superior forecasting performance of our model here may be attributable to the shorter sample size, which allows it to dominate the imprecisely estimated unrestricted VAR. Hence, in scenarios with a large cross-section and relatively small time sample, it is likely that our method can produce significant forecasting gains.

Table 6: MSFE of different forecasting approaches.

<b>MSFE</b>	<b>RR</b>	<b>VAR(P)</b>	<b>VAR(1)-Lasso</b>	<b>VAR(1)-Ridge</b>	<b>AR(1)</b>	<b>Best</b>	<b>DM pval</b>
<b>AIG</b>	0.63	0.65	0.65	0.68	0.65	RR	0.00
<b>JPM</b>	0.67	0.68	0.70	0.70	0.77	RR	0.00
<b>BAC</b>	0.78	0.80	0.79	0.81	0.80	RR	0.28
<b>C</b>	0.71	0.71	0.70	0.70	0.72	AR(1)	0.10
<b>WFC</b>	0.63	0.65	0.65	0.65	0.68	RR	0.00
<b>USB</b>	0.58	0.61	0.61	0.60	0.68	RR	0.00
<b>PNC</b>	0.66	0.65	0.70	0.70	0.75	VAR(P)	0.00
<b>TFC</b>	0.61	0.61	0.65	0.65	0.71	RR	0.09
<b>GS</b>	0.64	0.64	0.68	0.67	0.73	RR	0.04
<b>TD</b>	0.71	0.72	0.75	0.76	0.73	RR	0.06
<b>COF</b>	0.65	0.64	0.69	0.68	0.79	VAR(P)	0.09
<b>BK</b>	0.62	0.64	0.64	0.65	0.70	RR	0.00
<b>STT</b>	0.68	0.70	0.70	0.70	0.77	RR	0.00
<b>SIVB</b>	0.72	0.74	0.86	0.86	0.88	RR	0.00
<b>FITB</b>	0.60	0.64	0.64	0.64	0.73	RR	0.00
<b>MS</b>	0.75	0.75	0.77	0.78	0.72	AR(1)	0.09
<b>AXP</b>	0.74	0.76	0.80	0.80	0.85	RR	0.00

Notes: RR: Our reduced-rank model; VAR(P): unrestricted VAR(P) of DY; VAR(1)-NETS: VAR(1) with Lasso penalty; VAR(1)-Ridge: VAR(1) with Ridge penalty; AR(1): benchmark univariate autoregressive model; DM pval: p-value from the Diebold-Mariano test comparing superior predictive accuracy of the best model against the second best.

## 7. Concluding remarks

In this paper, we have proposed a novel SVAR methodology to estimate large unknown adjacency matrices from cross-sectional and time series data. Motivated by the centralities and low-rank representations of adjacency matrices, we propose a reduced-rank estimation to deal with the high-dimensionality of the problem. Theoretically, we have shown that reduced-form estimation is consistent and asymptotically normal. Results from numerical experiments show that the proposed iterated least squares algorithms can handle large network autoregressions well and the finite sample performance of the estimators provides support for our theoretical results. The empirical investigation on estimating volatility connectedness between major US FIs highlights the flexibility of the proposed estimation in estimating networks from the returns data of many stocks with limited sample size, a feat that many traditional methodologies may struggle with. More importantly, the ability of the SVAR to isolate idiosyncratic impacts from common shocks provides a tool for studying shock propagation *ceteris paribus* throughout a large network. Notably, this suggests that our framework can be generalized to other applications where a data-driven approach to network modeling is desirable particularly in fields where traditional network proxies do not easily apply.

## References

- Ahern, K.R., 2013. Network centrality and the cross section of stock returns. Available at SSRN 2197370 .
- Allen, F., Cai, J., Gu, X., Qian, J., Zhao, L., Zhu, W., 2019. Ownership network and firm growth: What do five million companies tell about chinese economy. Unpublished working paper .
- Bai, J., 2003. Inferential theory for factor models of large dimensions. *Econometrica* 71, 135–171.
- Bai, J., 2009. Panel data models with interactive fixed effects. *Econometrica* 77, 1229–1279.
- Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. *Econometrica* 70, 191–221.
- Barigozzi, M., Brownlees, C., 2019. Nets: Network estimation for time series. *Journal of Applied Econometrics* 34, 347–364.
- Billio, M., Getmansky, M., Lo, A.W., Pelizzon, L., 2012. Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of financial economics* 104, 535–559.
- Brownlees, C.T., Gallo, G.M., 2006. Financial econometric analysis at ultra-high frequency: Data handling concerns. *Computational statistics & data analysis* 51, 2232–2245.
- Cai, J., Yang, D., Zhu, W., Shen, H., Zhao, L., 2021. Network regression and supervised centrality estimation. Available at SSRN 3963523 .
- Cesa-Bianchi, A., Pesaran, M.H., Rebucci, A., 2020. Uncertainty and economic activity: A multicountry perspective. *The Review of Financial Studies* 33, 3393–3445.
- Chapman, J., Gofman, M., Jafri, S., 2019. High-frequency analysis of financial stability. Technical Report. Working paper, Bank of Canada and University of Rochester.

- Chen, E.Y., Fan, J., Zhu, X., 2020. Community network auto-regression for high-dimensional time series. arXiv preprint arXiv:2007.05521 .
- Chen, K., Dong, H., Chan, K.S., 2013. Reduced rank regression via adaptive nuclear norm penalization. *Biometrika* 100, 901–920.
- Chen, R., Xiao, H., Yang, D., 2021. Autoregressive models for matrix-valued time series. *Journal of Econometrics* 222, 539–560.
- Chernozhukov, V., Härdle, W.K., Huang, C., Wang, W., 2021. Lasso-driven inference in time and space. *The Annals of Statistics* 49, 1702–1735.
- Chudik, A., Pesaran, M.H., 2011. Infinite-dimensional vars and factor models. *Journal of Econometrics* 163, 4–22.
- Chudik, A., Pesaran, M.H., Tosetti, E., 2011. Weak and strong cross-section dependence and estimation of large panels.
- Corrado, L., Fingleton, B., 2012. Where is the economics in spatial econometrics? *Journal of Regional Science* 52, 210–239.
- Davidson, J., 1994. *Stochastic limit theory: An introduction for econometricians*. OUP Oxford.
- De Benedictis, L., Tajoli, L., 2011. The world trade network. *The World Economy* 34, 1417–1454.
- De Paula, Á., Rasul, I., Souza, P., 2019. Identifying network ties from panel data: Theory and an application to tax competition. arXiv preprint arXiv:1910.07452 .
- Diebold, F.X., Mariano, R.S., 2002. Comparing predictive accuracy. *Journal of Business & economic statistics* 20, 134–144.
- Diebold, F.X., Yilmaz, K., 2009. Measuring financial asset return and volatility spillovers, with application to global equity markets. *The Economic Journal* 119, 158–171.

- Diebold, F.X., Yilmaz, K., 2014. On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of econometrics* 182, 119–134.
- Elhorst, J.P., Gross, M., Tereanu, E., 2021. Cross-sectional dependence and spillovers in space and time: Where spatial econometrics and global var models meet. *Journal of Economic Surveys* 35, 192–226.
- Engle, R.F., Gallo, G.M., 2006. A multiple indicators model for volatility using intra-daily data. *Journal of econometrics* 131, 3–27.
- Engle, R.F., Gallo, G.M., Velucchi, M., 2012. Volatility spillovers in east asian financial markets: a mem-based approach. *Review of Economics and Statistics* 94, 222–223.
- Fan, J., Liao, Y., Mincheva, M., 2013. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75, 603–680.
- Gofman, M., Wu, Y., 2022. Trade credit and profitability in production networks. *Journal of Financial Economics* 143, 593–618.
- Guðmundsson, G.S., Brownlees, C., 2021. Detecting groups in large vector autoregressions. *Journal of Econometrics* 225, 2–26.
- Hamilton, J.D., Herrera, A.M., 2004. Comment: oil shocks and aggregate macroeconomic behavior: the role of monetary policy. *Journal of Money, credit and Banking* , 265–286.
- Han, F., Li, Y., 2020. Moment bounds for large autocovariance matrices under dependence. *Journal of Theoretical Probability* 33, 1445–1492.
- Kelejian, H.H., Prucha, I.R., 2002. 2sls and ols in a spatial autoregressive model with equal spatial weights. *Regional Science and Urban Economics* 32, 691–707.
- Kilian, L., 2001. Impulse response analysis in vector autoregressions with unknown lag order. *Journal of Forecasting* 20, 161–179.

- Kilian, L., Lütkepohl, H., 2017. Structural vector autoregressive analysis. Cambridge University Press.
- Kleinberg, J.M., 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)* 46, 604–632.
- Lee, L.F., 2002. Consistency and efficiency of least squares estimation for mixed regressive, spatial autoregressive models. *Econometric theory* 18, 252–277.
- Li, Z., Xiao, H., 2021. Multi-linear tensor autoregressive models. arXiv preprint arXiv:2110.00928 .
- Lin, X., Lee, L.f., 2010. Gmm estimation of spatial autoregressive models with unknown heteroskedasticity. *Journal of Econometrics* 157, 34–52.
- Manresa, E., 2013. Estimating the structure of social interactions using panel data. Unpublished Manuscript. CEMFI, Madrid .
- Miao, K., Phillips, P.C., Su, L., 2022. High-dimensional vars with common factors. *Journal of Econometrics* .
- Parkinson, M., 1980. The extreme value method for estimating the variance of the rate of return. *Journal of business* , 61–65.
- Pesaran, M.H., Schuermann, T., Weiner, S.M., 2004. Modeling regional interdependencies using a global error-correcting macroeconomic model. *Journal of Business & Economic Statistics* 22, 129–162.
- Richmond, R.J., 2019. Trade network centrality and currency risk premia. *The Journal of Finance* 74, 1315–1361.
- Velu, R., Reinsel, G.C., 2013. Multivariate reduced-rank regression: theory and applications. volume 136. Springer Science & Business Media.

- Vershynin, R., 2010. Introduction to the non-asymptotic analysis of random matrices. arXiv preprint arXiv:1011.3027 .
- Vershynin, R., 2018. High-dimensional probability: An introduction with applications in data science. volume 47. Cambridge university press.
- Wong, K.C., Li, Z., Tewari, A., 2020. Lasso guarantees for  $\beta$ -mixing heavy-tailed time series. The Annals of Statistics 48, 1124–1142.
- Yuan, M., Ekici, A., Lu, Z., Monteiro, R., 2007. Dimension reduction and coefficient estimation in multivariate linear regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69, 329–346.
- Zhou, H., Li, L., Zhu, H., 2013. Tensor regression with applications in neuroimaging data analysis. Journal of the American Statistical Association 108, 540–552.
- Zhu, X., Pan, R., Li, G., Liu, Y., Wang, H., 2017. Network vector autoregression. The Annals of Statistics 45, 1096–1123.

## Appendix A. Proofs for Section 4

### Appendix A.1. Proof of Proposition 1

First note that, for any  $N, T$  and  $P$ , (6) has a VAR(P) form:

$$y_t = A\beta_1 y_{t-1} + \dots + A\beta_p y_{t-p} + u_t \quad (\text{A.1})$$

where  $u_t = \Lambda f_t + \varepsilon_t$  is the composite error. Consequently, it can be expressed as the following moving average process:

$$y_t = \sum_{i=0}^{\infty} \Phi_i u_{t-i} \quad (\text{A.2})$$

where  $\Phi_0 = I$ ,  $\Phi_1 = A\beta_1$ ,  $\Phi_2 = A^2\beta_1^2 + \beta_2$  and hence  $\Phi_j = \sum_{l=1}^j \Phi_{j-l} A\beta_l$ . Note that  $\|\Phi_1\| \leq \|A\|_F |\beta_1| = |\beta_1| \equiv |\psi_1|$ ,  $\|\Phi_2\| \leq \|A\|_F^2 |\beta_1|^2 + |\beta_2| = |\beta_1|^2 + |\beta_2| \equiv |\psi_2|$  and so on. We can show that  $\|\Phi_j\| \leq |\psi_j|$  where  $|\psi_j| \equiv \sum_{l=1}^j |\psi_{j-l}| |\beta_l|$ . So

$$y_t \leq \sum_{i=0}^{\infty} \|\Phi_i\| \|u_{t-i}\| \leq \sum_{i=0}^{\infty} |\psi_i| \|u_{t-i}\|.$$

By Assumption 3, this infinite sum exists in mean-square and since this holds for any  $N, T$  and  $P$ ,  $\{y_t\}$  is a well-defined process almost surely.

### Appendix A.2. Proof of Theorem 1

Our proof strategy follows that of [Chen et al. \(2021\)](#) and [Li and Xiao \(2021\)](#). For notational convenience, write  $\text{vec}(\mathcal{Y}_{t-P}^{t-1})$  as  $\bar{y}_{P,t}$ ,  $E(\bar{y}_{P,t} \bar{y}_{P,t}^\top) \equiv \Sigma_Y$  and  $(\beta^\top \otimes ab^\top) \equiv \Phi$ . Then by [Lemma 3](#) and Assumption 4, we have

$$\left\| \frac{1}{T-P} \sum_{t=P+1}^T \bar{y}_{P,t} \bar{y}_{P,t}^\top - E(\bar{y}_{P,t} \bar{y}_{P,t}^\top) \right\| \rightarrow^p 0.$$

We can then find a subsequence of  $T_k, N_k, P_k$  such that the convergence holds almost surely, and for any positive constant  $C$  we have

$$\begin{aligned} & \sup_{\sqrt{\frac{T-P}{NP}} \|\tilde{\Phi} - \Phi\| \leq C} \frac{1}{N_k P_k (T_k - P_k)} \left| \sum_{t=P_k+1}^{T_k} \text{tr} [(\tilde{\Phi} - \Phi) \bar{y}_{P,t} \bar{y}_{P,t}^\top (\tilde{\Phi} - \Phi)^\top] - \right. \\ & \left. (T_k - P_k) \text{tr} [(\tilde{\Phi} - \Phi) \Sigma_Y (\tilde{\Phi} - \Phi)^\top] \right| \rightarrow 0 \text{ a.s.} \end{aligned} \quad (\text{A.3})$$

Let  $C_{NPT}$  be a sequence such that  $C_{NPT} \rightarrow \infty$  as  $N, P, T \rightarrow \infty$ . (A.3) implies that there exists  $C_{N_k, P_k, T_k} < C_{NPT}$  with  $C_{N_k, P_k, T_k} \rightarrow \infty$  and satisfies

$$\begin{aligned} & \sup_{\sqrt{\frac{T_k - P_k}{N_k P_k}} \|\tilde{\Phi} - \Phi\| \leq C_{N_k, P_k, T_k}} \frac{1}{N_k P_k (T_k - P_k)} \left| \sum_{t=P_k+1}^{T_k} \text{tr} [(\tilde{\Phi} - \Phi) \bar{y}_{P,t} \bar{y}_{P,t}^\top (\tilde{\Phi} - \Phi)^\top] \right. \\ & \left. - (T_k - P_k) \text{tr} [(\tilde{\Phi} - \Phi) \Sigma_Y (\tilde{\Phi} - \Phi)^\top] \right| \rightarrow^p 0. \end{aligned} \quad (\text{A.4})$$

Next, we consider

$$\begin{aligned} & \sum_{t=P+1}^T \|y_t - \tilde{\Phi} \bar{y}_{P,t}\|_F^2 - \sum_{t=P+1}^T \|u_t\|^2 \\ & = -2 \sum_{t=P+1}^T \text{tr} [(\tilde{\Phi} - \Phi) \bar{y}_{P,t} u_t^\top] + \sum_{t=P+1}^T \text{tr} [(\tilde{\Phi} - \Phi) \bar{y}_{P,t} \bar{y}_{P,t}^\top (\tilde{\Phi} - \Phi)^\top]. \end{aligned} \quad (\text{A.5})$$

On the boundary where  $\sqrt{\frac{T_k - P_k}{N_k P_k}} \|\tilde{\Phi} - \Phi\| = C_{N_k, P_k, T_k}$ ,

$$\sum_{t=P+1}^T \text{tr} [(\tilde{\Phi} - \Phi) \bar{y}_{P,t} u_t^\top] \leq \|\tilde{\Phi} - \Phi\|_F \left\| \sum_{t=P+1}^T \bar{y}_{P,t} u_t^\top \right\| = O\left((N_k P_k)^2 C_{N_k, P_k, T_k}\right), \quad (\text{A.6})$$

and

$$(T - P) \text{tr} [(\tilde{\Phi} - \Phi) \Sigma_Y (\tilde{\Phi} - \Phi)^\top] \geq (T - P) \lambda_{\min}(\Sigma_Y) \|\tilde{\Phi} - \Phi\|^2 \geq O\left((N_k P_k)^2 C_{N_k, P_k, T_k}^2\right), \quad (\text{A.7})$$

where  $\lambda_{\min}(\Sigma_Y)$  refers to the smallest eigenvalue of  $\Sigma_Y$ . Hence as  $C_{N_k, P_k, T_k} \rightarrow \infty$  and combining the results from (A.4) to (A.7), we conclude that

$$P \left[ \inf_{\sqrt{\frac{T_k - P_k}{N_k P_k}} \|\tilde{\Phi} - \Phi\| = C_{N_k, P_k, T_k}} \sum_{t=P+1}^T \|y_t - \tilde{\Phi} \bar{y}_{P,t}\|_F^2 < \sum_{t=P+1}^T \|u_t\|^2 \right] \rightarrow 0,$$

which by extension implies that

$$P \left[ \inf_{\sqrt{\frac{T-P}{NP}} \|\tilde{\Phi} - \Phi\| \geq C_{NPT}} \sum_{t=P+1}^T \|y_t - \tilde{\Phi} \bar{y}_{P,t}\|_F^2 < \sum_{t=P+1}^T \|u_t\|^2 \right] \rightarrow 0$$

since  $\|y_t - \tilde{\Phi} \bar{y}_{P,t}\|_F^2$  is convex in  $\tilde{\Phi}$ , thus concluding the proof.  $\square$

### Appendix A.3. Proof of Theorem 2

For convenience, let  $\sum_t$  be shorthand for  $\sum_{t=P+1}^T$ . We start from the first order conditions from the iterated least squares problem for  $\hat{a}, \hat{b}, \hat{\beta}$  respectively:

$$\begin{aligned} \sum_t \hat{a} \hat{b}^\top \mathcal{Y}_{t-P}^{t-1} \hat{\beta} \hat{\beta}^\top \mathcal{Y}_{t-P}^{t-1\top} \hat{b} - \sum_t y_t \hat{\beta}^\top \mathcal{Y}_{t-P}^{t-1\top} \hat{b} &= 0 \\ \sum_t \hat{b}^\top \mathcal{Y}_{t-P}^{t-1} \hat{\beta} \hat{\beta}^\top \mathcal{Y}_{t-P}^{t-1\top} - \sum_t \hat{a}^\top y_t \hat{\beta}^\top \mathcal{Y}_{t-P}^{t-1\top} &= 0 \\ \sum_t \mathcal{Y}_{t-P}^{t-1\top} \hat{b} \hat{a}^\top \hat{b}^\top \mathcal{Y}_{t-P}^{t-1} \hat{\beta} - \sum_t \mathcal{Y}_{t-P}^{t-1\top} \hat{b} \hat{a}^\top y_t &= 0. \end{aligned}$$

Substitute in the data generating process for  $y_t = ab^\top \mathcal{Y}_{t-P}^{t-1} \beta + u_t$  into the equations above and we get from the first equation:

$$\begin{aligned} \sum_t (\hat{a} - a) \mathcal{Y}_t^{\beta, b} \mathcal{Y}_t^{\beta, b\top} + \sum_t a (\hat{b} - b)^\top \mathcal{Y}_t^\beta \mathcal{Y}_t^{\beta, b\top} + \sum_t \mathcal{Y}_t^{a, b\top} (\hat{\beta} - \beta) \mathcal{Y}_t^{\beta, b\top} \\ = \sum_t u_t \mathcal{Y}_t^{\beta, b\top} + o_p \left( \sqrt{\frac{NP}{T-P}} \right) \end{aligned}$$

where  $\mathcal{Y}_t^{\beta,b} = b^\top \mathcal{Y}_{t-P}^{t-1} \beta$ ,  $\mathcal{Y}_t^\beta = \mathcal{Y}_{t-P}^{t-1} \beta$ , and  $\mathcal{Y}_t^{a,b} = \mathcal{Y}_{t-P}^{t-1} b a^\top$ ; For the second equation we get,

$$\begin{aligned} & \sum_t a^\top (\hat{a} - a) \mathcal{Y}_t^{\beta,b} \mathcal{Y}_t^{\beta\top} + \sum_t a^\top a (\hat{b} - b)^\top \mathcal{Y}_t^\beta \mathcal{Y}_t^{\beta\top} + \sum_t a^\top \mathcal{Y}_t^{a,b\top} (\hat{\beta} - \beta) \mathcal{Y}_t^{\beta\top} \\ &= \sum_t a^\top u_t \mathcal{Y}_t^{\beta\top} + o_p \left( \sqrt{\frac{NP}{T-P}} \right); \end{aligned}$$

Lastly, for the third equation we have,

$$\begin{aligned} & \sum_t \mathcal{Y}_t^{a,b} (\hat{a} - a) b^\top \mathcal{Y}_t^\beta + \sum_t \mathcal{Y}_t^{a,b} a (\hat{b} - b)^\top \mathcal{Y}_t^\beta + \sum_t \mathcal{Y}_t^{a,b} \mathcal{Y}_t^{a,b\top} (\hat{\beta} - \beta) \\ &= \sum_t \mathcal{Y}_t^{a,b} u_t + o_p \left( \sqrt{\frac{NP}{T-P}} \right); \end{aligned}$$

Taking vectorization on both sides of the equations, we get

$$\sum_t \begin{bmatrix} \mathcal{Y}_t^{\beta,b} \mathcal{Y}_t^{\beta,b\top} \otimes I & \mathcal{Y}_t^{\beta,b} \mathcal{Y}_t^{\beta\top} \otimes a & \mathcal{Y}_t^{\beta,b} \otimes \mathcal{Y}_t^{a,b\top} \\ \mathcal{Y}_t^\beta \mathcal{Y}_t^{\beta\top} \otimes a^\top & \mathcal{Y}_t^\beta \mathcal{Y}_t^{\beta\top} \otimes a^\top a & \mathcal{Y}_t^\beta \otimes a^\top \mathcal{Y}_t^{a,b\top} \\ \mathcal{Y}_t^{\beta,b\top} \otimes \mathcal{Y}_t^{a,b} & \mathcal{Y}_t^{\beta\top} \otimes \mathcal{Y}_t^{a,b} a & I \otimes \mathcal{Y}_t^{a,b} \mathcal{Y}_t^{a,b\top} \end{bmatrix} \begin{bmatrix} \text{vec}(\hat{a} - a) \\ \text{vec}(\hat{b}^\top - b^\top) \\ \hat{\beta} - \beta \end{bmatrix} = \sum_t \begin{bmatrix} \mathcal{Y}_t^{\beta,b} \otimes I \\ \mathcal{Y}_t^\beta \otimes a^\top \\ I \otimes \mathcal{Y}_t^{a,b} \end{bmatrix} u_t$$

which can be written as

$$\sum_t Z_t Z_t^\top \begin{bmatrix} \text{vec}(\hat{a} - a) \\ \text{vec}(\hat{b}^\top - b^\top) \\ \hat{\beta} - \beta \end{bmatrix} = \sum_t Z_t u_t.$$

Define  $\Sigma_{Z_t} = E(Z_t Z_t^\top)$  and

$$\Omega_{Z_t} = \lim_{N,P \rightarrow \infty} \left[ (NP)^{-1} \Sigma_{Z_t} \right] \left[ (NP)^{-2} \Sigma_{Z_t} \otimes \Sigma_u \right] \left[ (NP)^{-1} \Sigma_{Z_t} \right]^\top,$$

where  $\Sigma_u = E(u_t u_t^\top)$ . Then, we show

$$L_{N,P} \Omega_{Z_t}^{-1/2} \frac{1}{\sqrt{T-P}} \sum_t Z_t u_t \rightarrow^D N(0, L),$$

where  $\eta \in \mathbb{R}^q$  is a constant and  $L_{N,P}$  is defined as in the theorem but for convenience, normalize  $L$  such that  $\eta^\top L_{N,P} \eta = 1$ . Define

$$Q_{T(N),t} = \frac{1}{\sqrt{NP_{(T(N))}}} \eta^\top L_{N,P} \Omega_{Z_t}^{-1/2} \frac{1}{\sqrt{T(N) - P_{(T(N))}}} Z_t u_t$$

where  $T(N)$  is an integer-valued function of  $N$  and  $T(N) \rightarrow \infty$  as  $N \rightarrow \infty$ . Define  $\mathcal{F}_{Nt}$  to be an array of  $\sigma$ -algebra increasing in  $t$  for each  $N$  and let  $Q_{T(N)}$  be measurable with respect to  $\mathcal{F}_{Nt}$ .

Since  $Z_t Z_t^\top$  is a function of products of  $y_t$  and the lags, [Lemma 3](#) implies that

$$\frac{1}{T(N) - P_{T(N)}} \sum_t \eta^\top L_{N,P} \tilde{\Sigma}_{Z_t}^{-1/2} \frac{Z_t Z_t^\top}{NP_{(T(N))}} \tilde{\Sigma}_{Z_t}^{-1/2} L_{N,P}^\top \eta \xrightarrow{p} 1$$

where  $\tilde{\Sigma}_{Z_t} \equiv E(\lim_{N,P \rightarrow \infty} (NP)^{-1} Z_t Z_t^\top)$ . Next, given that  $u_t$  is uncorrelated over time, we have that  $Z_t$  and  $u_t$  are independent. Furthermore, the fourth moments of  $u_t$  are finite by Assumption 1. Together, we get

$$\sum_t Q_{T(N),t}^2 \rightarrow 1,$$

and this satisfies the first of two conditions for the martingale central limit theorem (see Theorem 24.3 of [Davidson \(1994\)](#)). Next we verify the following

$$E \left[ \left( \frac{1}{NP} \eta^\top L_{N,P} \Omega_{Z_t}^{-1/2} Z_t u_t u_t^\top Z_t^\top \Omega_{Z_t}^{-1/2} L_{N,P}^\top \eta \right)^2 \right] = O(1)$$

which is sufficient for Lindeberg's condition. Note that

$$\begin{aligned} E \left[ \left( \frac{1}{NP} \eta^\top L_{N,P} \Omega_{Z_t}^{-1/2} Z_t u_t u_t^\top Z_t^\top \Omega_{Z_t}^{-1/2} L_{N,P}^\top \eta \right)^2 \right] \\ \leq C \frac{1}{(NP)^2} \lambda_{\max}(L_N L_N^\top) \lambda_{\max}(\Sigma_{Z_t}) E(\|Z_t u_t u_t^\top Z_t^\top\|^2) \end{aligned}$$

where  $\lambda_{\max}$  refers to the largest singular value of the matrix, and  $C$  is some constant. Firstly,  $L_{N,P}$  is non-random, so  $\lambda_{\max}(L_N L_N^\top)$  is  $O(1)$ . Following [Chen et al. \(2020\)](#), we can show

that  $\lambda_{max}(\Sigma_{Z_t}) = O(NP)$  and likewise, the final term in the product is also  $O(NP)$  following independence. Hence, this verifies the claim. Under stationarity, this condition yields the Lindeberg condition which implies the second requirement for the martingale difference central limit theorem is satisfied. Hence, by Theorem 24.3 of [Davidson \(1994\)](#), and the delta method, we obtain our result. □

## Appendix A.4. Proof of [Proposition 2](#)

First, recall that  $\hat{u}_t = y_t - \hat{a}\hat{b}^\top \mathcal{Y}_{t-P}^{t-1} \hat{\beta}$ , then we have

$$\begin{aligned} \hat{u}_t &= y_t - (\hat{\beta}^\top \otimes \hat{A}) \bar{y}_{P,t} \\ &= u_t - [(\beta^\top \otimes A) - (\hat{\beta}^\top \otimes \hat{A})] \bar{y}_{P,t} \\ &\equiv u_t - \epsilon \bar{y}_{P,t} \\ &\equiv u_t + \delta_{NP,t} \end{aligned}$$

where  $\bar{y}_{P,t} = \text{vec}(\mathcal{Y}_{t-P}^{t-1})$ ,  $\epsilon = (\beta^\top \otimes A) - (\hat{\beta}^\top \otimes \hat{A})$ , and  $\delta_{NP,t} = \epsilon \bar{y}_{P,t}$ . Next, stack the vectors over time:  $\hat{u} = (\hat{u}_{P+1}, \dots, \hat{u}_T)^\top$ ,  $F = (f_{P+1}, \dots, f_T)^\top$ ,  $\mathcal{E} = (\epsilon_{P+1}, \dots, \epsilon_T)$ , and  $\Delta = (\delta_{P+1}, \dots, \delta_T)^\top$  where we have suppressed the  $NP$  subscript of  $\delta$  for convenience. Then we can write

$$\hat{u} = u + \Delta = F\Lambda^\top + \mathcal{E} + \Delta.$$

Let  $V_{NPT}$  be a  $k \times k$  diagonal matrix with the diagonal elements being the largest  $k$  eigenvalues, arranged in descending order, of  $\frac{1}{N(T-P)} \hat{u}\hat{u}^\top$  and  $\hat{F} = (\hat{f}_{P+1}, \dots, \hat{f}_T)^\top$  be the  $k$  eigenvectors, multiplied by  $\sqrt{T-P}$ , of the corresponding eigenvalues.

Using the definition of eigenvectors, we get

$$\begin{aligned} \frac{1}{N(T-P)} \hat{u}\hat{u}^\top \hat{F} &= \hat{F} V_{NPT} \\ \frac{1}{N(T-P)} \hat{u}\hat{u}^\top \hat{F} V_{NPT}^{-1} &= \hat{F}. \end{aligned}$$

Next, define the  $k \times k$  matrix  $H = (\Lambda^\top \Lambda / N)(F^\top \hat{F} / T)V_{NPT}^{-1}$ . Then, we can derive a similar identity as in equation A.1 of [Bai \(2003\)](#) for the discrepancy between the estimated factors and the truth:

$$\begin{aligned} \hat{f}_t - Hf_t &= V_{NPT}^{-1} \left( \frac{1}{T-P} \sum_{s=P+1}^T \hat{f}_s \frac{\varepsilon_s^\top \varepsilon_t}{N} + \frac{1}{T-P} \sum_{s=P+1}^T \hat{f}_s \frac{f_s^\top \Lambda^\top \varepsilon_t}{N} + \frac{1}{T-P} \sum_{s=P+1}^T \hat{f}_s \frac{f_t^\top \Lambda^\top \varepsilon_s}{N} \right. \\ &\quad \left. + \frac{1}{T-P} \sum_{s=P+1}^T \hat{f}_s \frac{\delta_s^\top \delta_t}{N} + \frac{1}{T-P} \sum_{s=P+1}^T \hat{f}_s \frac{\delta_s^\top u_t}{N} + \frac{1}{T-P} \sum_{s=P+1}^T \hat{f}_s \frac{\delta_t^\top u_s}{N} \right). \end{aligned}$$

This equation differs from A.1 for two reasons. When considering the idiosyncratic shocks,  $\varepsilon_t$ , we have assumed that they were *i.i.d* and hence  $E(\varepsilon_s^\top \varepsilon_t) = 0$  for  $t \neq s$ . More importantly, we have additional terms representing the estimation errors from the first-stage estimation, and they are captured in  $\delta$ .

Let  $(\cdot)_i$  indicate the  $i$ -th element of a vector in the parenthesis. Then by Cauchy-Schwarz and for some constant  $M$ ,

$$\begin{aligned} &\max_i \frac{1}{T-P} \sum_{s=P+1}^T (\hat{f}_t - Hf_t)_i^2 \\ &\leq M \left[ \max_i \frac{1}{T-P} \sum_{s=P+1}^T \left( \frac{1}{T-P} \sum_{s=P+1}^T \hat{f}_s \frac{\varepsilon_s^\top \varepsilon_t}{N} \right)^2 + \max_i \frac{1}{T-P} \sum_{s=P+1}^T \left( \frac{1}{T-P} \sum_{s=P+1}^T \hat{f}_s \frac{f_s^\top \Lambda^\top \varepsilon_t}{N} \right)^2 \right. \\ &\quad \left. + \max_i \frac{1}{T-P} \sum_{s=P+1}^T \left( \frac{1}{T-P} \sum_{s=P+1}^T \hat{f}_s \frac{f_t^\top \Lambda^\top \varepsilon_s}{N} \right)^2 + \max_i \frac{1}{T-P} \sum_{s=P+1}^T \left( \frac{1}{T-P} \sum_{s=P+1}^T \hat{f}_s \frac{\delta_s^\top \delta_t}{N} \right)^2 \right. \\ &\quad \left. + \max_i \frac{1}{T-P} \sum_{s=P+1}^T \left( \frac{1}{T-P} \sum_{s=P+1}^T \hat{f}_s \frac{\delta_s^\top u_t}{N} \right)^2 + \max_i \frac{1}{T-P} \sum_{s=P+1}^T \left( \frac{1}{T-P} \sum_{s=P+1}^T \hat{f}_s \frac{\delta_t^\top u_s}{N} \right)^2 \right]. \end{aligned}$$

The first 3 terms are identical to that in Lemma C.9 of [Fan et al. \(2013\)](#), and are shown to be  $O_P((T-P)^{-1} + N^{-1}) = O(N^{-1})$ , where  $N$  goes to infinity slower than  $T-P$ . Since the difference in our case is due to an additional first stage, we have to check for the last three terms. Hence, by [Lemma 4](#) we can bind the additional terms by an order of  $NP^2/(T-P)$ .

To arrive at the desired result, note that

$$\frac{1}{T-P} \sum_{t=P+1}^T \|\hat{f}_t - H^\top f_t\|^2 \leq k \max_{i \leq k} \frac{1}{T-P} \sum_{s=P+1}^T (\hat{f}_t - H f_t)_i^2.$$

## Appendix A.5. Proof of Theorem 3

For clarity, let the true rank be denoted  $r^*$ . Before proceeding, we restate Lemma 1 in [Chen et al. \(2013\)](#) here:

**Lemma 2** ([Chen et al. \(2013\)](#)). *Suppose there exists  $s \leq r^*$  such that  $d_s(XQ) > (1 + \delta)\lambda$  and  $d_{s+1}(XQ) \leq (1 - \delta)\lambda$  for a  $\delta \in (0, 1]$ . Then  $P(\hat{r} = s) \geq 1 - P(d_1(PU) \geq \delta\lambda)$  where  $P$  is the projection matrix from (14).*

Next note that when  $d_{r^*}(XQ) > 2\lambda$ , we have  $d_{r^*}(XQ) > 2\lambda \geq (1 + \delta)\lambda$ . By definition we also have  $d_{r^*+1}(XQ) = 0 \leq (1 - \delta)\lambda$  for some  $\delta \in (0, 1]$ . Hence, applying [Lemma 2](#), we get

$$\begin{aligned} P(\hat{r} = r^*) &\geq 1 - P(d_1(PU) \geq \delta\lambda) \\ &= 1 - P\left(d_1(PU) \geq \sqrt{T} + \sqrt{\log K_N} \sqrt{N \log N}\right) \\ &\geq 1 - 2 \exp(-c) \frac{N}{K_N} \rightarrow 1 \end{aligned}$$

where the second inequality is obtained by Assumption 5 and  $\lambda = \delta^{-1}(\sqrt{T} + \sqrt{\log K_N} \sqrt{N \log N})$ , and the final inequality is obtained from [Lemma 5](#) in the technical appendix.

□

## Appendix B. Technical lemmas

**Lemma 3.** *Under the assumptions of [Theorem 1](#) and for any  $j$ , we have*

$$\begin{aligned} & E \left\| \frac{1}{T-P} \sum_{t=P+1}^T \text{vec}(\mathcal{Y}_{t-P}^{t-1}) \text{vec}(\mathcal{Y}_{t-P-j}^{t-j})^\top - E(\text{vec}(\mathcal{Y}_{t-P}^{t-1}) \text{vec}(\mathcal{Y}_{t-P-j}^{t-j})^\top) \right\| \\ &= O \left( \sqrt{\frac{NP \log(NP)}{T-P}} + \frac{NP \log(NP) (\log^3((T-P)NP))}{T-P} \right). \end{aligned} \quad (\text{B.1})$$

*Proof.* The proof relies on Theorem 2.1 of [Han and Li \(2020\)](#) (termed as "HL" hereafter), however we are unable to directly apply their result to our case because it relies on a sub-Gaussian assumption of the underlying process. To generalize their result to the distributional assumptions as stated in Assumption 1, it is sufficient to check that Proposition 4.1 of HL continues to hold in our case.

To do so, first note that unlike the aforementioned paper, we do not need to truncate our process,  $\text{vec}(\mathcal{Y}_{t-P}^{t-1}) \text{vec}(\mathcal{Y}_{t-P-j}^{t-j})^\top$ , to bind it because  $\|y_t\| = O(\sqrt{N})$  almost surely for all  $t$ . To see this, recall the moving average form of  $y_t$  in [\(A.2\)](#):

$$\|y_t\| \leq \sum_{i=0}^{\infty} \|\Phi_i\| \|u_{t-i}\| = \sum_{i=0}^{\infty} \|\Phi_i\| \|\Lambda f_t + \varepsilon_t\|.$$

Note that  $\|\Lambda f_t\| \leq \|\Lambda\|_F \|f_t\|$ , and we have  $\|f_t\| = 1$  by normalization and  $\|\Lambda\|_F = O(\sqrt{N})$  by Assumption 2(ii). By Assumption 1,  $\|\varepsilon_t\| = O(\sqrt{N})$  almost surely. Together with stationarity, we conclude that  $\|y_t\| = O(\sqrt{N})$  almost surely.

Hence, we can immediately apply a version of the matrix Bernstein inequality for  $\tau$ -mixing processes (Theorem 4.3 in HL). To conserve on notation, write  $\text{vec}(\mathcal{Y}_{t-P}^{t-1})$  as  $\bar{y}_{P,t}$ . Let  $M$  be such that  $\|\bar{y}_{P,t} \bar{y}_{P,t}^\top - E(\bar{y}_{P,t} \bar{y}_{P,t}^\top)\| \leq M$  a.s. For a  $0 < \delta \leq 1$ , we have:

$$P \left( \frac{1}{T-P} \left\| \sum_{t=P+1}^T (\bar{y}_{P,t} \bar{y}_{P,t}^\top - E(\bar{y}_{P,t} \bar{y}_{P,t}^\top)) \right\| \geq z + \sqrt{\frac{\delta}{T-P}} \right)$$

$$\leq NP \exp \left\{ \frac{-(z + \sqrt{\frac{\delta}{T-P}})^2}{8(15^2(T-P)\nu^2 + 60^2M^2/\psi_2) + 2(z + \sqrt{\frac{\delta}{T-P}})M\tilde{\psi}(\tilde{\psi}_1, \psi_2, T-P, NP)} \right\},$$

where

$$\begin{aligned} \nu^2 &= \sup_{K \subseteq \{1, \dots, T-P\}} \frac{1}{\text{card}(K)} \left\| \sum_{i \in K} E(\bar{y}_{P,t} \bar{y}_{P,t}^\top - E(\bar{y}_{P,t} \bar{y}_{P,t}^\top)) \right\|^2, \\ \tilde{\psi}(\tilde{\psi}_1, \psi_2, T-P, NP) &= \frac{\log(T-P)}{\log 2} \max \left\{ 1, \frac{8 \log(\tilde{\psi}_1 (T-P)^6 NP)}{\psi_2} \right\}, \\ \tilde{\psi}_1 &= \max(NP)^{-1}, \psi_1 \end{aligned}$$

and  $\psi_1, \psi_2 > 0$  are constants. Following [Vershynin \(2018\)](#), we bind  $\nu^2$  and  $M$ . For  $\nu^2$ , note that  $\nu^2 = E(\bar{y}_{P,t} \bar{y}_{P,t}^{\top 2}) - E(\bar{y}_{P,t} \bar{y}_{P,t}^\top)^2 \preceq E(\bar{y}_{P,t} \bar{y}_{P,t}^{\top 2})$  where  $X \preceq Y \Rightarrow 0 \preceq Y - X$  implies that  $Y - X$  is a positive-semidefinite matrix. Then  $(\bar{y}_{P,t} \bar{y}_{P,t}^\top)^2 = \|\bar{y}_{P,t}\|^2 \bar{y}_{P,t} \bar{y}_{P,t}^\top \preceq C^2 \sqrt{NP} \bar{y}_{P,t} \bar{y}_{P,t}^\top$ , for some constant  $C$ . Taking expectation on both sides, we get  $\|E(\bar{y}_{P,t} \bar{y}_{P,t}^\top - E(\bar{y}_{P,t} \bar{y}_{P,t}^\top))\|^2 \leq C^2 \sqrt{NP} \|E(\bar{y}_{P,t} \bar{y}_{P,t}^\top)\|$  where  $\|E(\bar{y}_{P,t} \bar{y}_{P,t}^\top)\|$  is also  $O(\sqrt{NP})$ , which implies that  $\nu^2$  is of the same rate as  $A_1$ , which contains an analogous variance term, in proposition 4.1 of HL. Next, we bind  $M$ . Recall that  $M$  is such that  $\|\bar{y}_{P,t} \bar{y}_{P,t}^\top - E[\bar{y}_{P,t} \bar{y}_{P,t}^\top]\| \leq M$ . So,

$$\begin{aligned} \|\bar{y}_{P,t} \bar{y}_{P,t}^\top - E[\bar{y}_{P,t} \bar{y}_{P,t}^\top]\| &\leq \|\bar{y}_{P,t}\|^2 + \|E[\bar{y}_{P,t} \bar{y}_{P,t}^\top]\| \\ &\leq C^2 \sqrt{NP} + \|E[\bar{y}_{P,t} \bar{y}_{P,t}^\top]\| \\ &\leq 2C^2 \sqrt{NP}. \end{aligned}$$

We can set this upper bound to be  $M$  which is tighter than the bound,  $M_\delta$ , proposed in HL. These results together imply that the conclusion of proposition 4.1 will continue to hold even if we replace the sub-Gaussian assumption in HL with Assumption 1.  $\square$

**Lemma 4.** For all  $i \leq k$ , and let  $\hat{f}_{is}$  be the  $i$ -th element of  $\hat{f}_s$ , then we have

$$(i) \quad \frac{1}{T-P} \sum_{t=P+1}^T \left( \frac{1}{T-P} \sum_{s=P+1}^T \hat{f}_{is} \frac{\delta_s^\top \delta_t}{N} \right)^2 = O_p \left( \frac{N^2 P^4}{(T-P)^2} \right);$$

$$(ii) \frac{1}{T-P} \sum_{t=P+1}^T \left( \frac{1}{T-P} \sum_{s=P+1}^T \hat{f}_{is} \frac{\delta_t^\top u_s}{N} \right)^2 = O_p \left( \frac{NP^2}{T-P} \right);$$

$$(iii) \frac{1}{T-P} \sum_{t=P+1}^T \left( \frac{1}{T-P} \sum_{s=P+1}^T \hat{f}_{is} \frac{\delta_s^\top u_t}{N} \right)^2 = O_p \left( \frac{NP^2}{T-P} \right).$$

*Proof.* Firstly,

$$\begin{aligned} \frac{1}{T-P} \sum_{t=P+1}^T \left( \frac{1}{T-P} \sum_{s=P+1}^T \hat{f}_{is} \frac{\delta_s^\top \delta_t}{N} \right)^2 &\leq \frac{1}{T-P} \sum_{t=P+1}^T \left( \frac{1}{T-P} \sum_{s=P+1}^T \hat{f}_{is}^2 \right) \frac{1}{T-P} \sum_{s=P+1}^T \left( \frac{\delta_s^\top \delta_t}{N} \right)^2 \\ &= \frac{1}{T-P} \sum_{t=P+1}^T \frac{1}{T-P} \sum_{s=P+1}^T \left( \frac{\bar{y}_{P,s}^\top \epsilon^\top \bar{y}_{P,t}}{N} \right)^2 \\ &\leq \frac{1}{T-P} \sum_{t=P+1}^T \frac{1}{T-P} \sum_{s=P+1}^T \left( \frac{\|\epsilon\|^2 \|\bar{y}_{P,s}\| \|\bar{y}_{P,t}\|}{N} \right)^2 \\ &= O_p \left( \frac{N^2 P^4}{(T-P)^2} \right), \end{aligned}$$

since  $\bar{y}_{P,s}^\top \epsilon^\top \bar{y}_{P,t} \leq \lambda_1(\epsilon^\top \epsilon) \|\bar{y}_{P,s}\| \|\bar{y}_{P,t}\|$ , where  $\lambda_1(\epsilon^\top \epsilon)$  refers to the largest eigenvalue of  $\epsilon^\top \epsilon$ . However, note that it is equivalent to the square of the spectral norm of  $\epsilon$ ,  $\|\epsilon\|^2$ . The last line follows from [Theorem 1](#). Next, we have for (ii):

$$\begin{aligned} \frac{1}{T-P} \sum_{t=P+1}^T \left( \frac{1}{T-P} \sum_{s=P+1}^T \hat{f}_{is} \frac{\delta_t^\top u_s}{N} \right)^2 &\leq \frac{1}{T-P} \sum_{t=P+1}^T \left( \frac{1}{T-P} \sum_{s=P+1}^T \hat{f}_{is}^2 \right) \frac{1}{T-P} \sum_{s=P+1}^T \left( \frac{\delta_t^\top u_s}{N} \right)^2 \\ &= \frac{1}{T-P} \sum_{t=P+1}^T \frac{1}{T-P} \sum_{s=P+1}^T \left( \frac{\bar{y}_{P,t}^\top \epsilon^\top u_s}{N} \right)^2 \\ &\leq \frac{1}{T-P} \sum_{t=P+1}^T \frac{1}{T-P} \sum_{s=P+1}^T \left( \frac{\|\epsilon\| \|u_s\| \|\bar{y}_{P,t}\|}{N} \right)^2 \\ &= O_p \left( \frac{NP^2}{T-P} \right). \end{aligned}$$

The proof for (iii) is identical to that of (ii). □

**Lemma 5.** *Under Assumptions 1-5, and for some  $K_N$  such that  $N/K_N \rightarrow 0$ , we have for*

some constant  $c$ ,

$$P\left(d_1(PU) \geq \sqrt{T} + \sqrt{\log K_N} \sqrt{N \log N}\right) \leq 2\exp(-c) \frac{N}{K_N} \rightarrow 0.$$

*Proof.*

$$\begin{aligned} & P\left(d_1(PU) \geq \sqrt{T} + \sqrt{\log K_N} \sqrt{N \log N}\right) \\ & \leq P\left(\|U\| \geq \sqrt{T} + \sqrt{\log K_N} \sqrt{N \log N}\right) \\ & \leq 2\exp(-c) \frac{N}{K_N} \end{aligned}$$

where we have used  $\|PU\| \leq \|P\|\|U\| = \|U\|$  and that  $\|P\| = 1$  since the projection is symmetric and idempotent. For the second inequality, since each row of  $U$  is independent and  $O(\sqrt{N})$  almost surely by Assumption 1, we can apply Theorem 5.41 of [Vershynin \(2010\)](#) for matrices with heavy-tailed rows to bound the probability. The conclusion is achieved by construction of  $K_N$ .  $\square$