High-dimensional VARs with common factors

Ke Miao\textsuperscript{a,b}, Peter C.B. Phillips\textsuperscript{c,d,e,f}, Liangjun Su\textsuperscript{g}\textsuperscript{*}

\textsuperscript{a} Institute of World Economy, School of Economics, Fudan University, China
\textsuperscript{b} Shanghai Institute of International Finance and Economics, China
\textsuperscript{c} Yale University, United States of America
\textsuperscript{d} University of Auckland, New Zealand
\textsuperscript{e} University of Southampton, UK
\textsuperscript{f} Singapore Management University, Singapore
\textsuperscript{g} School of Economics and Management, Tsinghua University, China

\textbf{Abstract}

This paper studies high-dimensional vector autoregressions (VARs) augmented with common factors that allow for strong cross-sectional dependence. Models of this type provide a convenient mechanism for accommodating the interconnectedness and temporal co-variability that are often present in large dimensional systems. We propose an $\ell_1$-nuclear-norm regularized estimator and derive the non-asymptotic upper bounds for the estimation errors as well as large sample asymptotics for the estimates. A singular value thresholding procedure is used to determine the correct number of factors with probability approaching one. Both the LASSO estimator and the conservative LASSO estimator are employed to improve estimation precision. The conservative LASSO estimates of the non-zero coefficients are shown to be asymptotically equivalent to the oracle least squares estimates. Simulations demonstrate that our estimators perform reasonably well in finite samples given the complex high-dimensional nature of the model. In an empirical illustration we apply the methodology to explore dynamic connectedness in the volatilities of financial asset prices and the transmission of 'investor fear'. The findings reveal that a large proportion of connectedness is due to the common factors. Conditional on the presence of these common factors, the results still document remarkable connectedness due to the interactions between the individual variables, thereby supporting a common factor augmented VAR specification.

© 2022 Elsevier B.V. All rights reserved.

1. Introduction

In a pathbreaking study, Mann and Wald (1943) introduced vector autoregressions (VARs) and developed the limit theory for estimation and inference.\textsuperscript{1} The VAR approach was further developed and promoted for empirical macroeconomic research in an influential paper by Sims (1980). Since then, the methodology has become one of the most heavily

\textsuperscript{*} The authors thank Serena Ng, an associate editor and three anonymous referees for their constructive comments. Miao and Su thank the National Natural Science Foundation of China for financial support under the Grant Numbers 72103046 and 72133002, respectively. Phillips acknowledges support from the National Science Foundation under Grant No. SES 18-50860, and a Kelly Fellowship from the University of Auckland, New Zealand.

\textsuperscript{1} The extension to the structural VAR (SVAR) case was developed in the final section of Mann and Wald (1943); but this contribution seems to have passed unnoticed in the vast literature on SVAR modeling. For further discussion, see Hurn et al. (2020).

\textsuperscript{10}161jj. 3034-0476/© 2022 Elsevier B.V. All rights reserved.
used tools in applied finance and macroeconomics. It offers a simple and useful method to capture rich dynamics and interconnectedness in multiple time series. Unrestricted VARs can be efficiently estimated by least squares regression, which makes them particularly attractive in applied research. But low dimensional VARs often suffer from the notorious omitted variable bias problem, which makes the approach vulnerable to misleading inferences on both coefficients and impulse responses. In a series of articles (e.g., Sims (1992, 1993), and Leeper et al. (1996)) Sims and his coauthors have explored options to include more variables in VARs to improve forecasting performance.

Over the last decade, high-dimensional VARs have been frequently employed to conduct large dimensional time series investigations in economics, finance, and other social sciences. Inspired by the influential works of Tibshirani (1996), Zhao and Yu (2006), Zou (2006), Candès and Tao (2007) and Huang et al. (2008), researchers in this area have frequently utilized Lasso-type regularized estimation to address the difficulties of over-parameterization in large dimensions. For example, Haufe et al. (2010) propose to use high-dimensional VARs to estimate causal interactions in multivariate time series via group-Lasso; Guibert et al. (2019) propose to improve the forecast of mortality rates by using the elastic-net to estimate a high-dimensional VAR; BB (2019), Barigozzi and Hallin (2017), and Demirer et al. (2018) apply Lasso, adaptive Lasso or elastic-net methods to high-dimensional VARs or generalized dynamic factor models to estimate networks and construct measures of financial sector connectedness. All these papers focus on empirical applications rather than theory development. In theoretical work, BM (2015) study deviation bounds for Gaussian processes and investigate the $\ell_1$-regularized estimation of transition matrices in sparse VAR models; KC (2015) establish oracle inequalities for high-dimensional VAR models; Han et al. (2015) propose a generalized Dantzig selector in high-dimensional VARs; Guo et al. (2016) study a class of VAR models with banded coefficient matrices. These studies have opened up new avenues for handling high-dimensional VAR models in practical work.

All the aforementioned studies assume that the VAR errors exhibit at most weak cross-sectional dependence (c.f., Chudik et al. (2011)). However, as the number of cross section units becomes large relative to the number of time periods, the cross-sectional dependence in the error terms is often strong.\(^2\) It is well known that ignoring strong cross-sectional dependence in the error terms typically leads to inaccurate estimation and misleading inferences. In response to this limitation, the present paper proposes a new high-dimensional VAR model in which some common factors (CFs) feature in the determination of each time series besides the idiosyncratic errors and lagged values of the time series themselves. This high-dimensional VAR model with CFs allows for serial correlation among the CFs, which in turn leads to correlations between the CFs and the lagged time series. To properly control for the presence of CFs in this model it is necessary to estimate the factor component and the transition matrices simultaneously. Practical implementation also requires the determination of the number of factors and lag length.

A mentioned above, we choose to model strong cross-sectional dependence through a latent factor structure. In principle our analysis is closely related to certain dynamic factor models, especially the generalized dynamic factor model (GDFM) of Forni et al. (2000) that generalizes the dynamic factor model proposed by Geweke (1977). The proposed model has a GDFM representation with certain restrictions on the coefficients.\(^3\) In recent decades, the approximate static and dynamic factor models have been extensively studied. Examples of theoretical work include Forni et al. (2000), Bai and Ng (2002), Bai (2003), and Hallin and Liška (2007), among others. Applied finance and macroeconomic examples include Fama and French (1993), Stock and Watson (1999, 2002), Giannone et al. (2004), Bernanke et al. (2005), Ludvigson and Ng (2007), and Cheng and Hansen (2015). The success of factor models in these empirical analyses arguably establishes that strong cross-sectional dependence is pervasive in real financial and macroeconomic data. In these applications, dynamic factor model methods are utilized to summarize information from a large panel data. Specifically, the estimated dynamic factors serve as predictors or regressors to study univariate or fixed-dimensional time series. In contrast, our model is a generalization of the pure VAR model that seeks to study the complicated time series dynamics and cross-section interactions in high-dimensional time series. The latent factor structure is employed to control for strong cross-sectional dependence and the factors themselves are regarded as systematic shocks. Chudik and Pesaran (2011) also consider a factor-augmented infinite dimensional VAR model. For simplicity, they construct a model in which the factor-induced strong cross-sectional dependence is explicitly separated from other sources of cross-sectional dependence. They mention the possibility of using high-dimensional VARs with CFs but do not explicitly analyze the model. In an earlier work, Stock and Watson (2005) proposed a factor-structural VAR (FSVAR) model that appears similar to ours except that it is a fixed dimensional system and requires factors to be serially uncorrelated over time. In the panel data literature, Bai (2009) proposes to use a latent factor structure to capture unobserved heterogeneity and strong cross sectional dependence. Lu and Su (2016) and Moon and Weidner (2017) consider dynamic panel regressions with interactive fixed effects (IFEs). Our high-dimensional VAR model with IFEs includes both homogeneous and heterogeneous pure dynamic panels with IFEs as special cases.

\(^2\) For example, one can follow Forni et al. (2000) and look at the largest eigenvalues of the spectral density matrices of the N-dimensional error term, or study the eigenvalues of their covariance matrix. In many empirical datasets, it is commonly found that these diverge to infinity at rate N, which is highly suggestive of strong cross-sectional dependence as defined in Chudik et al. (2011).

\(^3\) Since our model is proposed to capture the dynamic mechanism of high-dimensional time series through VAR modeling, it assists in both network and spillover effect analyses. In contrast, the GDFM is proposed to distill information from high-dimensional time series with the estimated factors often assisting in studying dynamics in univariate time series or low dimensional time series.
To estimate a high-dimensional VAR model with CFs, we propose a three-step procedure. In the first step, we consider an $\ell_1$-nuclear-norm regularized least squares estimation problem that minimizes the sum of squared residuals with an $\ell_1$-norm penalty imposed on the transition matrices and a nuclear norm penalty on the low rank matrix $\Theta$ representing the common component. Imposing the $\ell_1$-norm penalty helps to estimate the sparse transition matrices, and the nuclear-norm penalty helps to estimate the low rank matrix arising from the CFs and factor loadings. The nuclear-norm regularization has recently become popular in the estimation of low rank matrices in statistics and econometrics; see, Negahban and Wainwright (2011), Rohde and Tsybakov (2011), Negahban et al. (2009, 2012), Bai and Ng (2019), Belloni et al. (2019), Fan et al. (2019), Feng (2019), Kolchinski et al. (2011), Moon and Weidner (2019), Chernozhukov et al. (2019), and Ma et al. (2021), among others. All these previous works focus on the error bounds (in Frobenius norm) for the nuclear-norm regularized estimates, except Moon and Weidner, 2019, Chernozhukov et al. (2019) and Ma et al. (2021) who study inference problems in linear or nonlinear panel data models with a low-rank structure. Like the latter authors, we simply use the nuclear-norm regularization to obtain consistent initial estimates. Under some regularity conditions, we establish the non-asymptotic bounds for the estimation error of the transition matrices and the low rank matrix $\Theta$. Applying a singular value thresholding (SVT) procedure on the singular values of the estimate of the $\Theta$, we obtain a consistent estimate of the number of factors. Then, given the estimated factor number, preliminary estimates of the CFs can be obtained.

In the second step, we include the estimated CFs as regressors and consider a generalized Lasso estimator to obtain an updated estimate of the transition matrices. We show that the estimation errors can be uniformly controlled, which facilitates the construction of weights for subsequent estimation by conservative Lasso in the third step. Under some regularity conditions, we show that this third step conservative Lasso estimator of the transition matrices achieves sign consistency (see, e.g., Zhao and Yu (2006)). Besides, the third step estimator of the transition matrices, factors and factor loadings are asymptotically equivalent to the corresponding oracle least squares estimators that are obtained by using detailed information about the form of the true regression model. We also study the asymptotic distributions of the oracle efficient estimators of the transition matrices.

The usefulness of our methodology is demonstrated in a real-data example. The illustration revisits the financial connectedness measures proposed by DY (2014) and the results document strong evidence for the existence of CFs in the volatilities of 23 sector exchange traded funds (ETFs). The findings show that CFs account for a large proportion of the variation in these volatilities; and, conditional on the CFs, a high level of connectedness remains present among the idiosyncratic components. This empirical application demonstrates the particularly useful features of the high-dimensional VAR model with CFs that enable this model to capture the dynamic evolution of time series with strong cross-sectional dependence while distinguishing variations that originate from different sources.

The present paper contributes to the fields of both high-dimensional time series analysis and regularized estimation. First, a new high-dimensional VAR model with CFs is proposed for which there are four main advantages: (i) it provides a convenient tool to study rich dynamics in high-dimensional time series while controlling for the presence of strong cross-sectional dependence; (ii) taking into account the influence of unobserved common factors helps to alleviate potential endogeneity issues due to serial correlation in the unobserved common factors; (iii) the common factor structure can be consistently estimated and used to identify systematic shocks, which are of interest in empirical work; and (iv) the model framework follows the lead of Demirer et al. (2018) in studying spillover effects via constructing a measure of connectedness in a VAR-based network. Second, our analysis of regularized estimation is new in three directions: (i) we relax the Gaussianity assumptions that are commonly assumed in the existing literature (see, e.g., BM (2015), and KC (2015)); (ii) we establish sharp probability bounds for processes with serial dependence, utilizing techniques developed by Wu (2005) and Wu and Wu (2016); and (iii) our methodology utilizes a combination of different types of regularization in the estimation procedure and establishes non-asymptotic error bounds.

The remainder of the paper is organized as follows. Section 2 introduces the model and provides conditions for stationarity in the analysis of the high-dimensional system. Section 3 develops the three-step estimation procedure and examines its theoretical properties. In Section 4, we conduct Monte Carlo experiments to evaluate the finite sample performance of our estimators. The model and methods are applied to study financial connectedness in Section 5. Section 6 concludes. Proofs of the main results in the paper are given in Appendix A. Further technical details are provided in the online Supplementary Material.

**Notation**

To proceed, we introduce some notation. Let $A = (a_{ij}) \in \mathbb{R}^{M \times N}$ and $v = (v_1, \ldots, v_N)' \in \mathbb{R}^N$ be a matrix and vector. Denote $v_i$ as the subvector of $v$ whose entries are indexed by a set $I \subset [N] \equiv \{1, \ldots, N\}$ and denote $A_{ij}$ as the submatrix of $A$ whose rows and columns are indexed by $i$ and $j$, respectively. Let $A_{\ast j} \equiv A_{\{j\}, \ast}$ be the submatrix of $A$ whose columns are indexed by $j$, $A_{i \ast} \equiv A_{\ast, \{i\}}$ be the submatrix of $A$ whose rows are indexed by $i$. For notational simplicity, we also write the individual columns and rows of $A$ respectively as $A_{\ast j} \equiv A_{\ast, \{j\}}$ for $j \in [N]$ and $A_{i \ast} \equiv A_{\{i\}, \ast}$ for $i \in [M]$.

Define the $\ell_0$, $\ell_q$ ($1 \leq q \leq 1$), and $\ell_\infty$ norms of a vector $v \in \mathbb{R}^N$ as follows

$$
|v|_0 \equiv \sum_{i=1}^N 1(v_i \neq 0), \quad |v|_q \equiv \left( \sum_{i=1}^N |v_i|^q \right)^{1/q}, \quad \text{and} \quad |v|_\infty \equiv \max_{1 \leq i \leq N} |v_i|,
$$

157
component analysis (PCA) on $\|A\|_q = \max_{|i|=1} \|Av_i\|_q$, $\|A\|_{\max} = \max_{i,j} |a_{ij}|$, $\|A\|_F = \left(\sum_{i,j} |a_{ij}|^2\right)^{1/2}$ and $\|A\|_\psi = \sum_{k=1}^{\min(N,M)} \psi_k(A)$, where $\psi_k(A)$ denotes the $k$th largest singular value of $A$ for $k = 1, \ldots, \min(N,M)$. Denote the largest and smallest singular values of $A$ as $\psi_{\max}(A)$ and $\psi_{\min}(A)$. In the special case $q = 2$, the $\ell_2$ matrix norm is also denoted as the operator norm: $\|A\|_{\text{op}} \equiv \|A\|_2 = \psi_{\max}(A)$. For a random variable or vector $x$, we denote its expectation and $\ell_p$-norm as $E(x)$ and $\|x\|_p \equiv [E(|x|^p)]^{1/p}$.

For a $T \times R$ full rank matrix $F$ with $T > R$, we denote the corresponding orthogonal projection matrices as $P_F = F(F'F)^{-1}F'$ and $M_F = I_T - P_F$, where $I_T$ denotes the $T \times T$ identity matrix. Let $\text{vec}(\cdot)$ denote the (columnwise) vectorization operator, and $\otimes$ be the (right hand) Kronecker operator. Let $\lor$ and $\land$ denote the max and min operators, viz., $a \lor b = \max(a, b)$ and $a \land b = \min(a, b)$.

\section{Model}

For an $N$-dimensional vector-valued time series $\{Y_t\} = \{(y_{1t}, \ldots, y_{Nt})'\}$, the high-dimensional VAR model of order $p$ with CFs is given by

$$Y_t = \sum_{j=1}^{p} A^0_{-j} Y_{t-j} + A^0_0 + u_t, \quad t = 1, \ldots, T,$$

where $A^0_1, \ldots, A^0_p$ are $N \times N$ transition matrices, $\Lambda^0 = (\lambda^0_{1}, \ldots, \lambda^0_{N})'$ is an $N \times R^0$ factor loading matrix, $f^0_t$ is an $R^0$-dimensional vector of common factors, and $u_t \equiv (u_{1t}, \ldots, u_{Nt})'$ is an $N$-dimensional vector of unobserved idiosyncratic errors. Throughout this paper we use the superscript 0 to denote true values. The coefficients of interest are the $A^0_0$, $\Lambda^0$, and $f^0 \equiv (f^0_1, \ldots, f^0_T)$. In practice, we need to determine the number of factors and the VAR order $p$. We propose a method to consistently determine $p$ in Section 3.5. The number of factors can be determined in the first step of our estimation procedure introduced in Section 3.1. The analytic framework allows for both the number of cross-sectional units $N$ and the number of time periods $T$ to pass to infinity. The lag length is also allowed to (slowly) grow to infinity with $(N, T)$. Estimation is then a natural high-dimensional problem with the number of parameters, $N^2p + R^0N + R^0T$, growing linearly with $T$ and quadratically with $N$.

It is convenient to reformulate model (2.1) in multivariate regression form as

$$Y_t = \begin{pmatrix}
Y_t' \\
Y_{t-1}' \\
\vdots
\end{pmatrix} = \begin{pmatrix}
Y_0' \\
Y_{t-1}' \\
\vdots
\end{pmatrix} + \begin{pmatrix}
A^0_0 \\
A^0_1 \\
\vdots
\end{pmatrix} \Psi^0 + \begin{pmatrix}
u_t'
\end{pmatrix},$$

where $Y \in \mathbb{R}^{T \times N}$, $X \in \mathbb{R}^{T \times Np}$, $B^0 \in \mathbb{R}^{Np \times N}$, and $U \in \mathbb{R}^{T \times N}$. Let $\Theta^0 = F^0A^0$ denote the common component. A key observation here is that $\Theta^0$ is a low rank matrix. However, due to the presence of $XB^0$, the direct use of principal component analysis (PCA) on $Y$ cannot deliver a consistent estimate of the common factors. That note under some regularity conditions, $\|\Theta^0\|_{\text{op}} = O_p(\sqrt{NT})$ and $\|U\|_{\text{op}} = O_p(\sqrt{NT})$. For the pure factor model as in Bai (2003), the separation of $\Theta^0$ from $U$ hinges on this order difference. The exact probability order of $\|B^0\|_{\text{op}}$ depends on the underlying data generating process but is in general not of smaller order than $O_p(\sqrt{NT})$, which makes it difficult to separate the low rank matrix $\Theta^0$ from $Y$ without information about $B^0$. Besides, when the common factors are themselves serially correlated, pure VAR($p$) estimation generally suffers from endogeneity bias issues.

4 The nonzero singular values of $\Theta^0$ are the eigenvalues of $(F^0F^0A^0A^0)^{1/2}$. Assuming $F^0F^0/\Sigma_{\Delta} \xrightarrow{p} \Sigma_{\Delta}$ and $A^0A^0/\Sigma_{\Delta} \xrightarrow{p} \Sigma_{\Delta}$ with $\Sigma_{\Delta}$ and $\Sigma_{\Delta}$ both nonsingular ensures the first part of the stated claim. Theorem 4.4.5 of Vershynin (2018) shows that the operator norm of a $T \times N$ random matrix with independent, mean zero, and sub-gaussian entries is $O_p(\sqrt{NT})$.

5 Let $\lambda_{\min}(B^0B^0)$ be $N$- and $T$- vectors of ones. Suppose that $\lambda_{\min}(B^0B^0) \geq c > 0$ and $p = 1$. Then

$$\|B^0\|_{\text{op}} \leq \left[\lambda_{\min}(X^0B^0X^0)^{1/2} \geq c \left[\lambda_{\min}(X^0X^0)^{1/2} \right] \right]^{1/2} \leq c \|X^0\|_{\text{op}}.$$
2.1. Stationarity analysis

Let \( X_t \equiv X_{t+j} \). The \( N \)-dimensional VAR(\( p \)) process \( \{Y_t\} \) can be rewritten in a companion form as an \( Np \)-dimensional VAR(1) process with CFs, viz.,

\[
\begin{bmatrix}
Y_t \\
Y_{t-1} \\
\vdots \\
Y_{t-p+1}
\end{bmatrix} = \begin{bmatrix}
A_1^0 & A_2^0 & \cdots & A_{p-1}^0 & A_p^0 \\
I_N & 0 & \cdots & 0 & 0 \\
0 & I_N & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & I_N & 0
\end{bmatrix} \begin{bmatrix}
Y_{t-1} \\
Y_{t-2} \\
\vdots \\
Y_{t-p}
\end{bmatrix} + \begin{bmatrix}
A_0^0 \phi \\
0 \\
0 \\
\vdots \\
0
\end{bmatrix} + \begin{bmatrix}
u_t \\
u_{t-1} \\
\vdots \\
u_{t-p+1}
\end{bmatrix}.
\tag{2.3}
\]

If one treats \( F_t + U_t \) as an impulse at period \( t \), the process \( \{X_{t+1}\} \) in (2.3) can be regarded as a high-dimensional VAR(1) process. We can write the reverse characteristic polynomial (see, e.g., p.16 of Lütkepohl (2005)) of \( Y_t \) as

\[
A(z) \equiv I_N - \sum_{j=1}^p A_j^0 z^j.
\]

In the low-dimensional framework, the process is stationary if \( A(z) \) has no roots in and on the complex unit circle, or equivalently the largest modulus of the eigenvalues of \( \Phi \) is less than 1. To achieve identification, we need to study the Gram or signal matrix \( S_X \equiv X'X/T \) and its population version \( \Sigma_X \equiv E(X_iX_i') \). Basu and Mikhailidis (2015; hereafter BM) study the deviation bounds for the Gram matrix, using a Gaussianity assumption and boundedness of the spectral density function. Following this approach we impose some conditions that ensure \( S_X \) is well behaved.

To proceed, write \( X_{t+1} \) as a moving average process of infinite order (MA(\( \infty \))) as

\[
X_{t+1} = \sum_{j=0}^\infty \Phi^j(F_{t-j} + U_{t-j}) \equiv X_{t+1}^{(f)} + X_{t+1}^{(u)},
\tag{2.4}
\]

where \( X_{t+1}^{(f)} \equiv \sum_{j=0}^\infty \Phi^jF_{t-j} \) and \( X_{t+1}^{(u)} \equiv \sum_{j=0}^\infty \Phi^jU_{t-j} \). Then, the stationarity of \( Y_t \) can be studied by considering \( X_{t+1}^{(f)} \) and \( X_{t+1}^{(u)} \). First, consider \( X_{t+1}^{(f)} \), the component due to the common factors. Note that the covariance matrix of \( F_t \) is a high-dimensional matrix with rank \( R_d \) and explosive nonzero eigenvalues. Even if the largest modulus of the eigenvalues of \( \Phi \) is smaller than 1, the variances of the entries of \( X_{t+1}^{(f)} \) are not assured to be uniformly bounded. Specifically, we consider \( y_{it}^{(f)} \), which is the \( i \)-th entry of \( X_{t+1}^{(f)} \). Let \( e_{j,M} \) be the \( j \)-th column of \( I_M \). Noting that \( y_{it}^{(f)} = (e_{1,p} \otimes e_{1,N})X_{t+1}^{(f)} \), we can write \( y_{it}^{(f)} \) as the MA(\( \infty \)) process

\[
y_{it}^{(f)} = \sum_{j=0}^\infty (e_{1,p} \otimes e_{1,N}) \Phi^j(e_{1,p} \otimes A^0)^j f_{t-j}^0 \equiv \sum_{j=0}^\infty \alpha_{it}^{(f)}(j)f_{t-j}^0,
\]

in which the \( f_0^0 \) are allowed to be serially correlated. To ensure \( y_{it}^{(f)} = O_p(1) \), the coefficients \( \alpha_{it}^{(f)}(j) \) need to be well-behaved. Note that we generally do not have \( \|\Phi\|_{op} \leq 1 \), as explained in the supplement of BM (2015). In Assumption A.1, we impose sufficient conditions that ensure the \( y_{it}^{(f)} \) are well-behaved. The online supplementary material provides a discussion of these conditions.

For the process \( \{X_{t+1}^{(u)}\} \), stationarity is assured if we assume the covariance matrix of \( U_t \) is well-behaved and \( u_t \) is serially uncorrelated as in BM (2015) and Kock and Callot (2015; hereafter KC). Similarly to \( y_{it}^{(f)} \), we define \( y_{it}^{(u)} \) such that

\[
y_{it} \equiv y_{it}^{(f)} + y_{it}^{(u)},
\tag{2.5}
\]

with explicit form

\[
y_{it}^{(u)} = \sum_{j=0}^\infty \alpha_{it}^{(u)}(j)u_{t-j} \text{ and } \alpha_{it}^{(u)}(j) \equiv (e_{1,p} \otimes e_{1,N}) \Phi^j(e_{1,p} \otimes I_M).
\]

Again, imposing zero serial correlation and weak cross-sectional correlation across the \( u_{it} \) is insufficient to ensure that \( y_{it}^{(u)} = O_p(1) \) uniformly.

Let \( \zeta \) and \( \hat{\zeta} \) denote generic positive constants that may vary across their occurrences. Throughout the paper, we will treat \( A^0 \) as nonrandom. To ensure the stationarity of \( \{Y_t\} \), we impose the following assumption.

Assumption A.1. (i) \( u_t = c^{(u)}e^{(u)}_t \), where \( e^{(u)}_t = (e^{(u)}_{1,t}, \ldots, e^{(u)}_{m,t})' \), the \( e^{(u)}_{i,t} \) are i.i.d. random variables across \((i, t)\) with mean zero and variance 1, and \( C^{(u)} \) is an \( N \times m \) matrix such that \( C^{(u)}C^{(u)'} = \Sigma_u \) and \( \zeta \leq \psi_{\max}(\Sigma_u) \leq \psi_{\max}(\Sigma_u) \leq \hat{\zeta} \);
(ii) \( \{f_t^0\} \) follows a strictly stationary linear process given by
\[
f_t^0 - \mu_f = \sum_{j=0}^{\infty} c(j)^{-1} e_t^{(j)},
\]
where \( e_t^{(j)} \equiv (e_{it}^{(j)}, \ldots, e_{it}^{(j)}) \)'s are i.i.d. with mean 0 and covariance matrix \( I_{p0} \) across \( t \), \( \sup_{m>1}(m+1)^\alpha \sum_{j=m}^{\infty} \|c(j)^{-1}\|_{\text{max}} \leq \tilde{c} < \infty \) for some constant \( \alpha > 1 \);
(iii) \( \max_{1 \leq i \leq p} \|e_{i0}^{(j)}\|_q < \tilde{c} \) and \( \max_{1 \leq i \leq m} \|e_{i1}^{(j)}\|_q < \tilde{c} \) for some \( q > 4 \);
(iv) \( \{e_t^{(j)}\} \) is independent of \( \{t\} \);
(v) the largest modulus of the eigenvalues of \( \Phi \) is bounded uniformly in \( (N, p) \) by some constant \( \rho \in (0, 1) \);
(vi) \( \sup_{N,p} \|\Phi(\cdot)e_{i0}^{(j)}\|_{\text{op}} \leq \tilde{c}_\rho \) and \( \sup_{N,p} \|\Phi(\cdot)e_{i1}^{(j)}\|_{\text{op}} \leq \tilde{c}_\rho \);
(vii) \( \sup_{N,p} \max_{|z|=1} \max_{1 \leq i \leq p} \|A_i^*(z)A(z)\| \leq \tilde{c} \), where \( |z| \) denotes the modulus of \( z \) in the complex plane, and \( A^*(z) \) denotes the conjugate transpose of \( A(z) \).

**Assumption A.1** is frequently made in high-dimensional time series analysis; see, e.g., Bai and Saranadasa (1996), Chen and Qin (2010) and Ma et al. (2020). At the cost of more complicated notations, one can allow \( \psi_{\min}(\Sigma_u) \) to converge to zero and \( \psi_{\min}(\Sigma_u) \) to diverge to infinity, both at a slow rate. Assumption A.1(ii) assumes the common factors to be stationary and allows for weak serial correlation. The factors can have nonzero mean so that the \( y_{it} \) can also have nonzero mean. Assumption A.1(iii) requires that both \( e_{i0}^{(j)} \) and \( e_{i1}^{(j)} \) have finite 4th order moments, which is a weak assumption compared to the Gaussian distribution assumption of BM (2015) and KC (2015). Assumption A.1(iv) requires independence between \( \{e_{i0}^{(j)}\} \) and \( \{e_{i1}^{(j)}\} \), which facilitates separate study of \( y_{it}^0 \) and \( y_{it}^1 \). Assumption A.1(v) is a standard assumption to ensure stationarity. Assumption A.1(vi) is a high level condition to ensure that \( E(y_{it}^0) \) is uniformly bounded. Assumption A.1(vii) helps to bound the minimum eigenvalue of \( \Sigma_X \). From the inequalities
\[
\max_{|z|=1} \max_{1 \leq i \leq p} \|A_i^*(z)A(z)\| \leq (\max_{|z|=1} \|A(z)\|_{\text{op}})^2 \leq 1 + \sum_{k=1}^{p} \|A_k^0\|_{\text{op}},
\]
it is evident that requiring all the \( A_k^0 \)'s to have finite operator norms is a sufficient condition for (vii).

The online Supplementary Material provides further discussion on Assumption A.1(i)--(vii). The following proposition ensures the stationarity of the process \( Y_t \) and establishes a lower bound for \( \psi_{\min}(\Sigma_u) \).

**Proposition 2.1.** Suppose that Assumption A.1 holds. (i) Then \( Y_t \) is a stationary process, sup \( E(y_{it}^0) < \infty \), and
\[
\psi_{\min}(\Sigma_u) \geq \frac{\psi_{\min}(\Sigma_u)}{\psi_{\min}(\max_{1 \leq i \leq p} \|A_i^*(z)A(z)\|)},
\]
(ii) Let \( \Sigma_{XF} \equiv E(X_t f_t^0) \), and \( \Sigma \equiv \Sigma_X - \Sigma_{XF} \Sigma_{XF}^{-1} \Sigma_{XF} \). We also have \( \psi_{\min}(\Sigma) \geq \psi_{\min}(\Sigma_u) \frac{\psi_{\min}(\max_{1 \leq i \leq p} \|A_i^*(z)A(z)\|)}{\max_{1 \leq i \leq p} \|A_i^*(z)A(z)\|}. \)

**Proposition 2.1(ii)** is a direct consequence of Proposition 2.3 and equation (2.6) of BM (2015). With the presence of common factors, we only have the well-behaved lower bounds for the eigenvalues of \( \Sigma_X \) and \( \Sigma \): they are bounded away from 0 under Assumption A.1(i) and (vii), but the largest eigenvalues of \( \Sigma_X \) and \( \Sigma \) still diverge to infinity at rate \( N \).

3. Estimation method and theory

This section develops a three-step estimation procedure for the model and establishes its non-asymptotic and asymptotic properties. For the moment, we assume that the VAR order \( p \) is known but the true number of factors \( R \) is unknown. In practice, we can determine \( p \) via a data-driven method as introduced in Section 3.5.

The three-step estimation procedure can be summarized as follows:

**Step 1: Initial estimates of the low rank matrix \( \Theta^0 \), the transition matrix \( B^0 \) and the factor matrix \( F^0 \).** This step estimates the low rank matrix \( \Theta^0 \) together with the transition matrix \( B^0 \) via an \( \ell_1 \)-nuclear-norm regularization procedure. The \( \ell_1 \)-penalty is imposed on the transition matrix \( B \) to encourage sparsity and the nuclear-norm penalty is imposed on the common component matrix \( \Theta \) for its low rank structure. Since the nuclear norm of a matrix is given by the summation of its singular values, the nuclear-norm can be regarded intuitively as a matrix version of the usual \( \ell_1 \)-norm imposed on the singular values and thereby assists in achieving a low rank estimate. The approach has two advantages: one is that it does not require the specification of the number of factors a priori, and the other is that the minimization problem is a convex problem due to the fact both the \( \ell_1 \)-norm and nuclear-norm are convex in their respective matrices. We will show that the resulting estimators \( \hat{B} \) and \( \hat{\Theta} \) are consistent for \( B^0 \) and \( \Theta^0 \), respectively, up to certain scale in terms of the Frobenius norm. Given the preliminary estimate \( \hat{\Theta} \), it is possible to estimate \( R^0 \) consistently via a hard singular value

---

6 As discussed in Section E of the online supplement, the process \( X_t \) has a generalized dynamic factor representation. The orthogonality between \( \{\epsilon_t^{(0)}\} \) and \( \{\epsilon_t^{(1)}\} \) serves as a part of the identification conditions.
Step 2: Initial estimates of the elements of the factor loadings and transition matrix. This step applies plain Lasso to estimate the elements of the factor loadings and transition matrix. Specifically, we run the \( \ell_1 \)-regularized time series regression of \( Y_t \) on \( (X_t, \hat{F}) \) to obtain an updated estimate \( \hat{B}_{n,i} \) of the \( i \)th column \( (B^n_{0,i}) \) of the transition matrix \( B^0 \) along with estimates of the factor loadings \( \hat{\lambda}^0_i \) for \( i = 1, \ldots, N \). Here, the plain \( \ell_1 \)-penalty is imposed on the transition matrix only, and we cannot apply the adaptive Lasso here because we do not have the element-wise rates yet. We will establish the uniform consistency for the elements of \( \hat{B} \), which is required for the construction of the weights to be used for the conservative Lasso in the third step. As is well known, despite the fact the Lasso used in this step encourages sparsity in the estimate, it does not deliver selection/sign consistency (see Zhao and Yu (2006)) or oracle-efficient estimation.

Step 3: Final estimates of the transition matrix, factors and factor loadings. With the uniform elementwise rates for the loadings and transition matrix estimates, we apply iterative conservative Lasso to obtain updated estimates of the transition matrix, factors and factor loadings. Like adaptive Lasso, conservative Lasso can yield sign consistency and oracle efficient estimates.

3.1. First-step estimator

In the first step, we propose an \( \ell_1 \)-nuclear-norm regularized estimator to estimate the coefficient matrix \( B^0 \) and the low rank matrix \( \Theta^0 \) simultaneously. We impose a sparsity condition on \( B^0 \) and use \( \ell_1 \)-norm regularization to achieve the selection of regressors. We adopt nuclear norm regularized estimation to obtain the initial consistent estimate of the low rank matrix \( \Theta^0 \). The first step estimator is given by the following procedure.

First-step estimator: Let \( \gamma_1 = \gamma_1(N, T) = c_1T^{-1/2} \log N \) and \( \gamma_2 = \gamma_2(N, T) = c_2(N^{-1/2} + T^{-1/2}) \) for some positive constants \( c_1 \) and \( c_2 \).

1. Estimate the coefficient matrix \( B^0 \) and the low rank matrix \( \Theta^0 \) by running the following \( \ell_1 \)-nuclear-norm regularized regression:
   \[
   (\hat{B}, \hat{\Theta}) = \arg\min_{(B, \Theta)} \mathcal{L}(B, \Theta),
   \]
   \[
   \mathcal{L}(B, \Theta) = \frac{1}{2NT} \| Y - XB - \Theta \|_F^2 + \frac{\gamma_1}{N} \| \text{vec}(B) \|_1 + \frac{\gamma_2}{\sqrt{NT}} \| \Theta \|_*.
   \]

2. Estimate the number of factors \( R^0 \) by singular value thresholding (SVT) as follows:
   \[
   \hat{R} = \sum_{i=1}^{N \times T} 1\{\psi_i(\hat{\Theta}) \geq (\gamma_2 \sqrt{NT} \| \hat{\Theta} \|_{op})^{1/2}\},
   \]
   where \( \psi_i(\hat{\Theta}) \) are the singular values of \( \hat{\Theta} \).

3. Obtain a preliminary estimate of \( F^0 \). Let the singular value decomposition (SVD) of \( \hat{\Theta} \) be \( \hat{\Theta} = \tilde{U} \tilde{D} \tilde{V}' \), where \( \tilde{D} = \text{diag}(\psi_1(\hat{\Theta}), \ldots, \psi_{N \times T}(\hat{\Theta})) \). Set \( \tilde{F} = \sqrt{T} \tilde{U} \cdot |\hat{\Theta}| \).

Remark 3.1. The objective function \( \mathcal{L}(B, \Theta) \) is the sum of squared residuals with both the nuclear-norm regularization on \( \Theta \) and \( \ell_1 \)-norm regularization on \( B \). To obtain the numerical solution, we can apply an EM type algorithm. In the E-step, we fix \( B \) and update the estimate of \( \Theta \). The solution can be obtained following the result of Lemma 1 of MW (2019).\(^7\) In the M-step, we fix \( \Theta \) and update \( B \). The optimization problem can be decomposed to \( N \) Lasso-type linear regression problems.

3.1.1. Non-asymptotic results for the first-step estimator

In this subsection we establish the non-asymptotic properties of the first step estimator. In particular, for \( \hat{B} \) and \( \hat{\Theta} \), we establish a non-asymptotic inequality for their estimation errors. For \( \hat{R} \), we show that \( \hat{R} = \hat{R}^0 \) with a high probability.

To proceed, we introduce a key invertibility condition for the linear operator \( (\Delta^{(1)}, \Delta^{(2)}) \mapsto X\Delta^{(1)} + \Delta^{(2)} \) when \( (\Delta^{(1)}, \Delta^{(2)}) \) is restricted to lie in a ‘cone’. We follow the lead of Negahban et al. (2012) and refer to the condition as ‘restricted strong convexity’.\(^8\) Our condition imposed here takes a form related to that of MW (2019) and Chernozhukov

\(^7\) Let the SVD of \( A \) be \( A = USV' \), with \( S = \text{diag}(s_1, \ldots, s_q) \), with \( q = \text{rank}(A) \). Then \( \arg\min_{\Theta} \left\{ \frac{1}{2} \| A - \Theta \|_F^2 + \gamma \| \Theta \|_* \right\} \) is given by \( U \cdot \text{diag}(s_1 - \gamma, \ldots, s_q - \gamma) \cdot V' \), where \( (s_i) = \max(0, s_i) \).

\(^8\) As remarked by Negahban et al. (2012), the loss function is often not strongly convex for high-dimensional regressions. This failure leads to a difficulty in showing the desired convergence rate for the estimators. In this context, a suitable choice of the regularization parameter helps ensure that the estimate lies in a restricted set in the parameter space. Consequently, it suffices to assume that the function is strongly convex over this restricted set.
et al. (2019). To define the ‘cone’, let $J_i \subset \{Np\}$ be an index set such that $j \in J_i$ if and only if $B_{ij}^{0} \neq 0$. Let $J_i = \{Np\} \setminus J_i$. Let the SVD of $\Theta^0$ be $\Theta^0 = U^0 D \bar{V}^0$, where $U^0 \in \mathbb{R}^{T \times R}$ and $V^0 \in \mathbb{R}^{T \times R}$. For a $T \times N$ matrix $\Delta^{(2)}$, define the operators $P(\Delta^{(2)}) \equiv U_{\{i\} \setminus \{\} \cup \{1\}}^{(2)} V_{\{i\} \setminus \{\} \cup \{1\}}^{(2)}$ and $M(\Delta^{(2)}) \equiv (\Delta^{(2)} - P(\Delta^{(2)}))$. Hence, the operator $P(\cdot)$ projects a matrix onto a ‘low-rank’ space which contains $\Theta^0$. For some $c > 0$, the ‘cone’ $C_{NT}(c) \subset \mathbb{R}^{Np \times N} \times \mathbb{R}^{T \times N}$ is a set of $(\Delta^{(1)}, \Delta^{(2)})$ satisfying the restriction: $\gamma_1 \sum_{i=1}^{N} | \Delta_{i,j}^{(1)} | 1 = \gamma_2 \| M(\Delta^{(2)}) \|_* \leq c \gamma_1 \sum_{i=1}^{N} | \Delta_{i,j}^{(1)} | 1 + c \gamma_2 \| P(\Delta^{(2)}) \|_*$. We impose the following condition.

**Assumption A.2** (Restricted Strong Convexity). Let $\Phi_{\gamma_1, \gamma_2}(\Delta^{(1)}, \Delta^{(2)}) \equiv \gamma_1 \sum_{i=1}^{N} | \Delta_{i,j}^{(1)} | 1 + \gamma_2 \sqrt{NT} \| P(\Delta^{(2)}) \|_*$ be a tolerance function. If $(\Delta^{(1)}, \Delta^{(2)}) \in C_{NT}(3)$, then there exist positive constants $\kappa$, $\kappa'$ and $\kappa''$ such that $\| X \Delta^{(1)} + \Delta^{(2)} \|_F^2 \geq \gamma_1 \kappa' \| \Delta^{(1)} \|_F^2 + \kappa \| \Delta^{(2)} \|_F^2 - \kappa'' \Phi_{\gamma_1, \gamma_2}(\Delta^{(1)}, \Delta^{(2)})$ with probability $1 - \varepsilon_{NT}$ and $\varepsilon_{NT} \to 0$ as $(N, T) \to \infty$.

Assumption A.2 is a high level condition whose verification is challenging without imposing further conditions on the parameter space. In Section F of the online supplement, we provide some discussion of Assumption A.2. In particular, we give two conditions that are sufficient for Assumption A.2. The second condition can be relatively easier to verify and the first condition can be argued on intuitive grounds. See also the discussion in MW (2019) following their Assumption 1. Let $k_i = |J_i|$, $K = \sup_i k_i$, and $K_0 \equiv \frac{1}{N} \sum_{i=1}^{N} k_i$. The next assumption involves a regularity condition on the errors:

**Assumption A.3.** $\| U \|_{op} / \sqrt{NT} \leq \gamma_2 / 2$, where $\gamma_2$ is the tuning parameter for the nuclear norm regularization.

Recall that $\gamma_2 = C_2(N^{-1/2} + T^{-1/2})$. Assumption A.3 requires that the error matrix have an operator norm of order $O(p \sqrt{N} + \sqrt{T})$. This condition is standard in the literature; see, e.g., Lu and Su (2016), Moon and Weidner (2017), Su and Wang (2017), Moon (2019), and Chernozhukov et al. (2019). Moon and Weidner (2017) provide examples to ensure the above assumption holds. In particular, it is satisfied when the $\epsilon_{it}^{(0)}$ are i.i.d. sub-Gaussian (see, e.g., Vershynin (2018)).

**Theorem 3.1.** Suppose that Assumptions A.1–A.3 hold. Then we have $N^{-1/2} \| \bar{B} - B^0 \|_F \leq \tilde{c}(\gamma_1 \sqrt{K_0} \vee \gamma_2)$ and $(NT)^{-1/2} \| \bar{\Theta} - \Theta^0 \|_F \leq \tilde{c}(\gamma_1 \sqrt{K_0} \vee \gamma_2)$, with probability at least $1 - \varepsilon_{NT} - \tilde{c}(pN^2 \gamma_1 \sqrt{T} \log N \gamma_1 / 2 + pN^2 \log \log N)$ for some finite positive constants $\tilde{c}$, $\check{c}$, and $\tilde{c}$.

Theorem 3.1 establishes non-asymptotic inequalities for the estimation errors of $\bar{B}$ and $\bar{\Theta}$ in terms of the Frobenius norm. Note that $B^0$ and $\Theta^0$ are both high-dimensional matrices with $Np$ and $NT$ entries, respectively, and the Frobenius norms are normalized correspondingly by $\sqrt{N}$ and $\sqrt{NT}$. Without any sparsity or approximate sparsity assumption, $\| B^0 \|_F^2$ can be as large as $O(N^2)$. Assumption A.4(iii) below specifies an average control on the sparsity of the columns of $B^0$, which ensures that $\frac{1}{N} \| B^0 \|_F^2 = \frac{1}{N} \sum_{i=1}^{N} | B_{i,i}^0 |^2 = O(K_0)$ provided the elements in $B^0$ are uniformly bounded from the above. This motivates the use of $N^{-1/2}$ to normalize $\| \bar{B} - B^0 \|_F$. The first result in Theorem 3.1 ensures that $\frac{1}{N} \| \bar{B} - B^0 \|_F = \frac{1}{N} \sum_{i=1}^{N} \| B_{i,i} - B_{i,i}^0 \|_2^2 \leq \tilde{c}^2((\gamma_1 \sqrt{K_0} \vee \gamma_2)^2)$ with high probability.

That is, on average, the Euclidean distance between the columns of $\bar{B}$ and $B^0$ is bounded by a small term $\tilde{c}((\gamma_1 \sqrt{K_0}) \vee \gamma_2)$. Similarly, $\Theta^0$ has Frobenius norm of order $\sqrt{NT}$, which motivates the use of $(NT)^{-1/2}$ to normalize $\| \bar{\Theta} - \Theta^0 \|_F$. The second result in the theorem ensures that the large dimensional matrix estimate $\bar{\Theta}$ is sufficiently close to the truth $\Theta^0$ in terms of Frobenius norm: the entries of $\bar{\Theta}$ converge to those of $\Theta^0$ at rate $(\gamma_1 \sqrt{K_0}) \vee \gamma_2$ on average.

The probability in Theorem 3.1 converges to one when $\varepsilon_{NT}, pN^2 \gamma_1 \sqrt{T} \log N \gamma_1 / 2$ and $pN^{1/2} \log \log N$ all converge to zero. In general, the second term dominates the third one for finite $q$ and divergent $(N, T)$. If the error terms are sub-exponential, we can allow $q$ to diverge to infinity in which case the third term could dominate the second one. To prove the above theorem, we need to establish a bound for $T^{-1} \| U \|_{max}$ specifically, we need a sharp probability bound for a partial sum like $T^{-1} \sum_{t=1}^{T} e_{it} - y_{it} u_{ij}$. To achieve such a bound we resort to a Nagaev-type inequality, as introduced by Wu (2005) and Wu and Wu (2016), allowing for both dependence among summands and non-Gaussianity. The summand $y_{i,t-k} u_{ij}$
has a nonlinear Wold presentation \( y_{t−k} = g_{k}(e_{t−1}, ..., e_{t−k−1}) \), where the \( e_{t} \equiv \left( e^{(x)}_{t}, e^{(y)}_{t} \right) \) are i.i.d. random variables under Assumption A.1. Then one can verify that the \textit{dependence-adjusted norm} (see Wu and Wu (2016)) of \( y_{t−k} \) is well bounded so that one can obtain a sharp probability bound using the Nagaev-type inequality for nonlinear processes.\(^9\)

Next, we impose an assumption on the common factor and the factor loadings and a sparsity condition on \( \Sigma_{A}^{0} \):

**Assumption A.4.** (i) There exists an \( \bar{N} \) such that for all \( N > \bar{N} \), \( \|A^{0}A^{0}/N - \Sigma_{A}^{0}\|_{max} \leq \bar{c}N^{-1/2} \) for an \( R^{0} \times R^{0} \) matrix \( \Sigma_{A} \geq 0 \) and \( \|A^{0}\|_{max} \leq \bar{c} \);

(ii) Let \( \Sigma_{F} = \text{E}(f^{0}_{i} f^{0}_{j}) \). There are constants \( s_{1} > ... > s_{q} > 0 \) so that \( s_{j} \) equals the \( j \)th largest eigenvalue of \( \Sigma_{F}^{1/2} \Sigma_{A} \Sigma_{F}^{1/2} \);

(iii) \( K_{a} = o(T(N^{-1/2} + T^{-1/2})/|\log N)^{2} |) \).

**Assumption A.4(i)–(ii)** requires that the factors and the factor loadings are strong/pervasive with well-behaved sample second moments. **Assumption A.4(ii)** requires distinct eigenvalues of \( \Sigma_{F}^{1/2} \Sigma_{A} \Sigma_{F}^{1/2} \) in order to identify the corresponding eigenvectors. **Assumption A.4(iii)** imposes a sparsity condition on the transition matrix. We allow \( K_{a} \) (and thus \( K_{j} \)) to diverge to infinity at a rate slower than \( T(N^{-1/2} + T^{-1/2})/(\log N)^{2} \) here, which ensures accuracy of \( \hat{\Theta} \). Such a strict sparsity condition can be relaxed to an approximate sparsity condition as in Belloni et al. (2012) but that extension is not pursued here.

**Theorem 3.2.** Suppose Assumptions 1–4 hold. There exist positive constants \( \bar{c}, \bar{c} \) and \( \bar{c}' \) and a random matrix \( \tilde{F} \) depending on \( (F^{0}, A^{0}) \) such that (i) \( \tilde{F} = F^{0} \tilde{H} / \sqrt{T} \) and (ii) \( \|F^{0} F^{0} / ||F^{0} F^{0}/\sqrt{T} \leq \bar{c} (\sqrt{N} / \sqrt{K_{a}} + \gamma_{2}) \), both with probability larger than \( 1 - \bar{c}_{NT} = \bar{c} (pN^{2}/T^{1/4} (|\log N)^{-1/2} + pN^{2}^{-1/2} \log N) \).

**Theorem 3.2(i)** establishes the consistency of \( \tilde{F} \) and the mean-square convergence rate of \( \tilde{F} \) in large samples. Intuitively, since \( \hat{\Theta} \) is a consistent estimator of \( \Theta^{0} \equiv F^{0} A^{0} \) with well-controlled estimation errors, we expect the first \( R^{0} \) singular values of \( \hat{\Theta} \) to be \( O_{P}(\sqrt{NT}) \) and the other singular values to be \( O_{P}(\sqrt{N} / \sqrt{K_{a}} + \gamma_{2}) \). Then the hard SVT procedure can distinguish the \( \sqrt{NT} \) order singular values from those of smaller order. Alternatively, given the consistency of \( \hat{B} \) established in Theorem 3.1, the ‘residual’ \( Y - \hat{X} \tilde{B} \) is an approximation of \( F^{0} A^{0} + U \). One can also apply the methods of Bai and Ng (2002), Onatski (2010) and Ahn and Horenstein (2013) to determine the number of factors. **Theorem 3.2(ii)** establishes the convergence rate of \( \tilde{F} \). The \( R^{0} \times R^{0} \) transformation matrix \( \tilde{H} \) is similar to the transform matrix \( H \) in Bai (2003) but only depends on the true values whereas \( H \) in Bai (2003) also depends on the estimator.

Despite the fact that we can establish weak consistency of \( \tilde{B}, \hat{\Theta} \) and \( \tilde{F} \) in terms of the Frobenius norm in Theorems 3.1–3.2, we cannot obtain pointwise consistency or asymptotic distributions for the elements of these estimators. The major role for the first-step procedure is to obtain an initial estimator that can be used subsequently to enhance estimation properties.

**3.2. Second-step estimator**

In the second-step of the procedure we run a time series regression of \( Y_{n,i} \) on \( (X, \tilde{F}) \) for each \( i \in [N] \) by imposing an \( \ell_{1} \)-norm penalty on the coefficient of \( X \). The goal is to obtain an estimator of \( B^{0} \) whose elements uniformly converge to the true values.\(^10\) Given the uniform convergence property, the second-step estimator indicates how likely the corresponding true parameter value is to zero or not. The estimator can then be employed to construct adaptive- or conservative-Lasso weights in a third step with further enhanced properties.

**Second-step estimator:** Let \( \gamma_{3} = c_{3} (\sqrt{N} / \sqrt{K_{a}} + \gamma_{2}) \) for some constant \( c_{3} \). For each \( i \in [N] \), solve the minimization problem:

\[
(\hat{B}_{n,i}, \hat{\lambda}_{i}) = \arg \min_{(v', \lambda') \in \mathbb{R}^{Np + d0}} \frac{1}{2T} \|Y_{n,i} - Xv - \tilde{F}\lambda\|_{F}^{2} + \gamma_{3} |v|_{1},
\]

where the Lasso penalty is only imposed on the coefficients of \( X \). The second-step estimators of \( B^{0} \) and \( A^{0} \) are given by \( B = (\hat{B}_{n,1}, ..., \hat{B}_{n,N}) \) and \( A = (\hat{\lambda}_{1}, ..., \hat{\lambda}_{N}) \).

**Remark 3.2.** Note that the \( \ell_{1} \)-norm penalty is only imposed on the coefficient of \( X \). In the proof of Theorem 3.3, we show that \( \hat{B}_{n,i} \) solves the Lasso problem with dependent variable \( \tilde{M}_{F} Y_{n,i} \) and regressors \( \tilde{M}_{F} X \).

\(^{9}\) Both KC (2015) and BM (2015) impose i.i.d. and Gaussianity assumptions on the error terms and derive exponential probability bounds for the partial sums. In contrast, we only assume the existence of finite \( q \) th order moments of \( u_{t} \), and allow for serial correlations in the error term. The term \( pN^{2}/T^{1/4} (|\log N)^{-1/2} \) in the probability bound reflects the price of relaxing the Gaussianity assumption.

\(^{10}\) By contrast the first-step estimator \( \tilde{B} \) converges to \( B^{0} \) in Frobenius norm after normalization; but this convergence does not ensure either the pointwise convergence or uniform convergence (\( \max_{i} |\tilde{B}_{i} - B_{i}| = o_{P}(1) \)).
3.2.1. Non-asymptotic results for the second step estimator

The following theorem establishes non-asymptotic properties for \( \hat{B} \) by delivering an \( \ell_{\max} \)-norm bound for the estimation error of \( \hat{B} \).

**Theorem 3.3.** Suppose that Assumptions A1–A4 hold and \( 32K_f \gamma_3 \leq \psi_{\min}(\Sigma) \). Then

\[
\| \hat{B} - B^0 \|_{\max} \leq \max_{1 \leq i \leq N} | |\hat{B}_{\cdot i} - B^0_{\cdot i}| | \leq \frac{48}{(\psi_{\min}(\Sigma x))^2} K_f \gamma_3
\]

with probability larger than \( 1 - \varepsilon_{NT} - c[p^2 N^2 T^{-1/4}(\log N)^{-q/2} + pN^2 \varepsilon^T + p^2 N^2 \xi^2 \log N] \) for some finite positive constants \( \xi \) and \( c \).

Theorem 3.3 establishes uniform convergence rates for the elements of \( \hat{B} \). Compared to Theorems 3.1–3.2, one additional term \( pN^2 \varepsilon^T \) appears in the probability bound. This term decays in the exponential rate of \( T \) and is in general dominated by \( p^2 N^2 T^{-1/4}(\log N)^{-q/2} \) when \( T \to \infty \).

A key step in the proof of Theorem 3.3 is to establish a restricted eigenvalue condition (REC) as in Bickel et al. (2009) and KC (2015). Due to the presence of common factors in the model, one needs to establish the REC on \( \tilde{\Sigma} = X'MX/T \). Recall that \( \psi_{\min}(\Sigma) \) is bounded away from 0 by Proposition 2.1, but the sample analog \( \tilde{\Sigma} \) does not have such a property. In fact, if \( N_p > T, \tilde{\Sigma} \) is always singular, which leads to \( \min_{|v|_F \neq 0} \frac{v' \tilde{\Sigma} v}{|v|^2} = 0 \). The minimum has to be replaced by a minimum over a smaller set in order to obtain a nonzero lower bound. Let \( J \subset [N_p] \) be an index set and \( J^c = [N_p]\backslash J \). We say the REC is satisfied for some \( K \in [N_p] \) if

\[
\min_{|J| \leq K} \min_{|v|_F \neq 0, |v|_0 \leq 3|v|_1} \frac{v' \tilde{\Sigma} v}{|v|^2} \leq \kappa_2(K) > 0,
\]

where \( J \) has cardinality no bigger than \( K \). In (3.3), the minimum is restricted to those vectors for which \( |v_{J^c}| \leq 3|v|_1 \), where \( J \) has cardinality no larger than \( K \). In this restricted space, we establish that (3.3) is satisfied with a high probability for \( K = K_f \) in Lemma A.4(\( \psi \)). See also the proof of Lemma A.4(\( \psi \)) in the Online Supplement.

3.3. Third-step estimator

In the first and second steps, we impose penalties on the elements in the coefficient matrix \( B \). These penalties introduce asymptotic bias into the estimator of the transition matrix. Zou (2006) proposed an adaptive Lasso technique in a linear regression framework that penalizes the true zero parameters more than the non-zero ones. Zou shows that the adaptive Lasso estimator is asymptotically equivalent to the oracle least-squares estimator that is obtained using the true information concerning the relevant regressors in the regression model. KC (2015) explored the use of the adaptive Lasso method in a high-dimensional VAR framework.

In practice, the regressors with zero coefficient estimates in the preliminary stage, which are usually plain Lasso estimates, are excluded in the adaptive Lasso. Hence, any incorrect regressor exclusion by the preliminary stage estimates directly lead to invalid regressor selection in adaptive Lasso. To solve this problem, the conservative Lasso, which gives the regressors that are excluded by the initial estimator a second chance, is introduced (e.g., Caner and Kock (2018)). In this subsection, we extend the conservative Lasso estimator to the framework of high-dimensional VAR with CFs. To ensure stationarity in the high-dimensional VAR, most nonzero entries of the transition matrices have to be bounded above by one in absolute value. Some of them may even shrink to zero as \( N \) goes to infinity. In finite samples, the first and second step estimates may be wrongly estimated to be zero, leading to poor finite sample performance. This is the reason that we recommend the use of conservative Lasso in this step. In the Monte Carlo simulations reported later we find that the conservative Lasso tends to outperform the adaptive Lasso.

**Third-step estimator (Conservative Lasso): Implement the following procedure:**

1. (Set weights) Let \( \gamma_4 = \gamma_4(N, T) \). Let \( W \) be a \( N_p \times N \) matrix with entries

\[
w_{ki} = 1 \quad \text{if} \quad |B_{ki}| < \alpha \gamma_4,
0 \quad \text{if} \quad |B_{ki}| \geq \alpha \gamma_4,
\]

where \( k \in [N_p], i \in [N], \) and \( \alpha > 0 \). Set \( \hat{F}^{(0)} = \tilde{F} \).

2. (Update \( \hat{B}^{(\ell)} \)) For integer \( \ell \geq 1 \), update the estimates of \( B \) and \( \Lambda \) using

\[
(\hat{B}_{\cdot i}^{(\ell)}, \hat{\Lambda}_{\cdot i}^{(\ell)}) = \arg \min_{(\hat{B}_{\cdot i}^{(\ell)}, \hat{\Lambda}_{\cdot i}^{(\ell)}) \in R^p \times R} \frac{1}{2T} \| Y_{\cdot i} - X\hat{B}^{(\ell)} - \hat{F}^{(\ell-1)} \|_F^2 + \gamma_4 \sum_{k=1}^{K_p} u_{ki} |v_{ki}|,
\]

where \( u_{ki} \) is the \( k \)th entry of \( v, i \in [N] \). Let \( \hat{B}_{\cdot i} = (\hat{B}_{1i}, \ldots, \hat{B}_{N_i}) \).

3. (Update \( \hat{F}^{(\ell)} \)) Obtain the SVD of \( Y - X\hat{B}^{(\ell)} \) as \( Y - X\hat{B}^{(\ell)} = U^{(\ell)} \hat{F}^{(\ell)} \hat{V}^{(\ell)} \). Obtain an updated estimate of \( F \) as \( \hat{F}^{(\ell)} = \sqrt{U^{(\ell)} \hat{F}^{(\ell-1)} U^{(\ell)} } \). Set \( \ell = \ell + 1 \).

4. Iterate steps 2–3 for a finite times \( \ell^* \). Denote the final estimators by \( \hat{B} = \hat{B}^{(\ell^*)}, \hat{F} = \hat{F}^{(\ell^*-1)} \) and \( \hat{\Lambda} = \hat{\Lambda}^{(\ell^*)} \).
Remark 3.3. Note that the weights do not change with iterations in the above procedure. It is worth mentioning that the weights $w_{ki}$ can take other forms. For example, Caner and Kock (2018) also consider $w_{ki} \equiv \frac{\gamma_{perc}}{|B_{ki}| \sqrt{\gamma_{perc}}}$, where $\gamma_{perc} = \alpha \gamma_4$.

Recall that $\hat{F}(0) = \tilde{F}$, which is estimated in the first step and has slower convergence rate. Iterations help to improve the factor estimates. In our simulations, the iterations often numerically converged in less than ten steps.

3.3.1. Asymptotic properties of the third-step estimator

We establish two results: (i) the conservative Lasso estimator $\hat{B}(\cdot)$ achieves variable-selection or sign consistency; and (ii) $\hat{B}$ is asymptotically equivalent to the oracle least squares estimator. Following Zhao and Yu (2006) and Huang et al. (2008), we say that $\hat{B}(\cdot) = B^0$, or $\hat{B}(\cdot)$ is sign-consistent for $B^0$, if and only if $\text{sgn}(\hat{B}_{k,i}) = \text{sgn}(B^0_{k,i})$ for all $i \in [N]$, where $\text{sgn}(B_{k,i}) \equiv \{\text{sgn}(B_{1,i}), \ldots, \text{sgn}(B_{N,i})\}$. and

$$\text{sgn}(B_{k,i}) \equiv \begin{cases} 1 & \text{if } B_{k,i} > 0 \\ 0 & \text{if } B_{k,i} = 0 \\ -1 & \text{if } B_{k,i} < 0. \end{cases}$$

Assumption A.5. (i) As $(N, T) \to \infty$, the magnitude of nonzero coefficients are of larger asymptotic order than $\gamma_4$ where $\gamma_4 = o(\min_{|j| \in [N]} |\text{min}_{k \neq j} |B^0_{k,i}|)$ and $(k_j^{3/2}T^{-1/2} \log N + K_jN^{-1/2}) = o(\gamma_4)$;

(ii) $N^{-1} \sum_{i=1}^{N} k_i^2 \equiv O(1)$ and $K_j \log N \cdot (N^{-1/2} \vee T^{-1/2}) = o(1)$;

(iii) $N^{-2}I^{-4/3}(\log N)^{-4/3} \to 0$ and $T/N^2 \to 0$ as $(N, T) \to \infty$.

Assumption A.5(i) assumes the nonzero entries of $B^0$ are not too small, a standard condition in the adaptive Lasso literature. The lower bound $\min_{|j| \in [N]} |\text{min}_{k \neq j} |B^0_{k,i}|$ has to be larger than $\gamma_4$ in order to separate the nonzero entries from zeros. By Assumption A.5(i) and Theorem 3.3, we can show that $\max_{|j| \in [N]} w_{ki} = 0$ and $\min_{|j| \in [N]} w_{ki} = 1$ with probability approaching one (w.p.a.1). In this case, we only place a penalty on the true zero entries asymptotically. Assumption A.5(ii) imposes some conditions on $K_j$ and the $k_i$ to ensure that $\|X(\hat{B}(\cdot) - B^0)\|_F$ has a desired convergence rate. The first restriction is imposed to simplify the asymptotic analysis and it implies $K_j = O(1)$ so that we can drop $K_j$ in subsequent asymptotic orders. Assumption A.5(ii) can be satisfied if most columns in $B^0$ have a finite number of nonzero coefficients while some columns in $B^0$ can have $0(\sqrt{N} \wedge \sqrt{T} / \log N)$ nonzero coefficients. Assumption A.5(iii) imposes conditions on the relative rates at which $N$ and $T$ pass to infinity and these depend on the number $(q)$ of moments for the innovation processes in the errors and factors. In the special case where $N$ and $T$ pass to infinity at the same rate, this condition requires $q \geq 12$.

The following theorem establishes the variable selection consistency of $\hat{B}(\cdot)$ and the preliminary convergence rates of $\hat{B}(\cdot)$ and $\tilde{F}(\cdot)$.

Theorem 3.4. Suppose that Assumptions A.1–A.5 hold. Then for a fixed $i$, we have

(i) $P(\hat{B}(\cdot) \equiv B^0) \to 1$ as $(N, T) \to \infty$;

(ii) $\|X(\hat{B}(\cdot) - B^0)\|_F / \sqrt{NT} = O_P(\gamma_1 + \gamma_2)$;

(iii) $\|\hat{F}(\cdot) - F^0\|_F / \sqrt{T} = O_P(\gamma_1 + \gamma_2)$.

Theorem 3.4(i) shows that $\hat{B}(\cdot)$ has the oracle property in that it selects the correct variables w.p.a.1. Due to the presence of common factors and the possibly divergent number $(k_i)$ of nonzero coefficients in $B^0_{k,i}$, we can only obtain a preliminary rate $O_P(\gamma_1 + \gamma_2)$ in Theorem 3.4(ii)-(iii).

To improve the rate of convergence, we study the final estimators $\hat{B}, \tilde{F}$ and $\hat{A}$. Now, $\tilde{F}$ corresponds to the first $\tilde{K}$ eigenvectors of $(Y - X\hat{B})(Y - X\hat{B})'$, rescaled by $\sqrt{T}$, and one can expand the estimation error $\tilde{F} - F^0H$ as in Bai and Ng (2002) and Bai (2009). Then, by examining the product of $\tilde{F} - F^0H$ with other terms, a sharper bound for some intermediate estimates can be obtained. Finally we can improve the probability order of each element in $\hat{B}_{k,i} - B^0_{k,i}$ to $O(T^{-1/2})$.

The following theorem reports the asymptotic distribution of $\hat{B}_{k,i}$.

Theorem 3.5. Suppose that Assumptions A.1–A.5 hold. Let $S_i$ denote an $L \times k_i$ selection matrix such that $S_iS_i' = I_k$ and $L$ is a fixed integer. Suppose that as $(N, T) \to \infty$ or $(N, T, p) \to \infty$ in the case $p \to \infty$, the limit of $S_i(\Sigma_{k,i}^{-1})S_i$ exists and is given by $\Sigma_i$. Conditioned on the event $\{\hat{B} = B^0\}$, for each $i \in [N]$, we have $\sqrt{T}S_i(\hat{B}_{k,i} - B^0_{k,i}) \overset{d}{\to} N(0, \sigma^2_i \Omega)$, where $\sigma^2_i = E(u_{ki}^2)$.

Note that a selection matrix $S_i$ is used in Theorem 3.5 that is not needed if $k_i$ is fixed. Intuitively, since $k_i$ is allowed to diverge to infinity as $(N, T) \to \infty$, asymptotic normality of $\hat{B}_{k,i}$ can be obtained directly when $k_i \to \infty$. Instead, we follow standard practice for estimation and inference with a divergent number of parameters (see, e.g., Fan and Peng (2004), Lam and Fan (2008), and Qian and Su (2016)) and prove asymptotic normality for an arbitrary but finite number

---

11 The first requirement assumes $T^{3/4 - 1}$ dominates $N^2$. For a VAR system, $T$ is the number of observations and $N$ is the dimension of the system. Large $T$ compared to $N$ is desirable for good regression results. The second requirement implies that $N$ cannot be too small compared to $T$. Because $N$ affects the estimation accuracy of the factors, only when the estimation errors for the factors are well controlled can the asymptotic oracle properties in Theorems 3.4–3.5 be established.
of linear combinations of the elements of $\hat{B}_{i,j}$. In the special case where $k_i$ is fixed, we can take $S_i = I_{i,j}$ and obtain the usual joint asymptotic normal distribution for all elements of $\hat{B}_{i,j}$.

As mentioned in the Introduction section, our model includes the pure dynamic panels with IFEs as special cases. For clarity, consider the high dimensional VAR(1) model with IFEs. If one finds that VAR(1) coefficient matrix is diagonal, then the model can be written as a heterogeneous dynamic panel of the form: $y_{it} = \rho^0 y_{i,t-1} + \Lambda^0_i f^0_t + u_{it}$. Clearly, the key strict stationarity condition in Assumption A.1(v) now becomes $\sup_{N\geq 1} \max_{1 \leq i \leq N} |\rho^0| \leq \rho \in (0, 1)$. Our final estimator of $\rho^0$ enjoys the same first order asymptotic property as the usual PCA estimator based on such a pure dynamic panel model specification. Furthermore, if one has prior knowledge that these AR(1) coefficients are common across all cross sectional units and given by $\rho^0$, one can average our last stage estimates on $\rho^0$ to obtain a $\sqrt{NT}$-consistent estimate of $\rho^0$, after proper bias correction. But due to the space limitation, we do not conduct formal asymptotic analyses here.

3.4. Tuning parameter selection

In practice, we need to select the tuning parameters $\gamma_\ell$, for $\ell = 1, \ldots, 4$. For $\gamma_2$, which is the tuning parameter for the nuclear norm penalty, we adopt a simple plug-in approach similar to that introduced in Chernozhukov et al. (2019). An ideal tuning parameter for $\gamma_2$ is one such that

$$\|U\|_\text{op}/\sqrt{NT} \leq (1 - c)\gamma_2$$

for some $c > 0$ with high probability. Suppose $U$ is a random matrix with i.i.d. sub-Gaussian entries that have mean zero and variance $\sigma_u^2$, its operator norm is bounded by $C \sigma_u(\sqrt{N} + \sqrt{T})$ for some $C > 0$ with high probability (see Vershynin (2018)). One can first use $\gamma_2 = \frac{\sigma_u^2}{\sqrt{N} + \sqrt{T}}/\sqrt{NT}$ for some $C > 1$ and $\hat{\gamma}_2$ is the sample standard deviation of $Y$. After obtaining an estimate $\hat{\sigma}_u$ of $\sigma_u$, we can calculate a suitable $\gamma_2$ via simulation. Specifically, we can simulate the random matrices $U$ with i.i.d. $N(0, \hat{\sigma}_u^2)$, and then let $\gamma_2 = Q(\|U\|_\text{op}, 0.95)/\sqrt{NT}$, where $Q(x, a)$ denotes the $a$th quantile of $x$.

For $\gamma_1$, $\gamma_3$, and $\gamma_4$, we propose to use the 5-fold cross validation (CV) process. Let $\gamma = (\gamma_1, \gamma_3, \gamma_4)^T$. For the first-step estimation, the procedure goes as follows:

1. Partition the data into 5 separate sets along the time dimension: $T_1, \ldots, T_5 \subset [T]$;
2. For $k = 1, \ldots, 5$, fit the model to the training set by excluding the $k$th fold data. Denote the estimators by $\tilde{B}^{(r,k)}$ and $\tilde{A}^{(r,k)}$, where $A^{(r,k)}$ is a $N \times R$ matrix containing the first $R$ right singular vectors of $\tilde{\Theta}$. Calculate the sum of squared prediction errors

$$c(u, \gamma, k) = \text{tr}[(Y_{T_k,*} - X_{T_k,*} \tilde{B}^{(r,k)}) M_{A^{(r,k)}(Y_{T_k,*} - X_{T_k,*} \tilde{B}^{(r,k)})}]$$

3. Compute the CV error for a fixed tuning parameter by $\text{CV}(\gamma) = \sum_{k=1}^5 c(u, \gamma, k)$.
4. Select $\gamma^* = \arg\min_{\gamma'} \text{CV}(\gamma')$.

Remark 3.4. Once the sample $T_k$ is excluded, we cannot obtain an estimate of $F_{T_k,*}$. Hence we cannot obtain the residuals by deducting the estimate of $F_{T_k,*}A'$. For this reason, we multiply $Y_{T_k,*} - X_{T_k,*} \tilde{B}^{(r,k)}$ by $M_{A^{(r,k)}}$ to project out $F_{T_k,*}A'$ in the above procedure.

For the second and third step estimators, the CV procedure can be constructed in a similar way.

3.5. Lag length selection

In the above estimation procedure, we have so far assumed that the lag length $p$ is known. In practice, the lag length $p$ is usually unknown and requires estimation. To address this uncertainty we propose a procedure to determine the lag length $p$. Suppose we estimate the model with a lag setting $p_{\max} \geq p^0$, where we use the superscript ‘0’ to denote the true parameter. The model with $p_{\max}$ lags continues to be a correctly specified model except that $A_0^0 = 0$ for $k > p^0$. Due to Lasso regularization, the elements of the estimator $\hat{A}_p$ for $p > p^0$ should converge to zero. For this reason, we propose to determine the lag length by the following procedure:

1. Given $p_{\max}$, obtain the estimates $\hat{A}_k$ for $k \in [p_{\max}]$;
2. Calculate $\alpha_k = \|\hat{A}_k\|^2_F / \sigma$ for some small positive constant $\sigma$ and $k \in [p_{\max}]$;
3. The criterion function we consider is given by the ratio

$$GR(p) = \frac{\sum_{k=p}^{p_{\max}} \alpha_k}{\sum_{k=p+1}^{p_{\max}} \alpha_k}, \quad p = 1, \ldots, p_{\max} - 1.$$

The term $GR$ refers to the growth ratio of $\sum_{k=p}^{p_{\max}} \alpha_k$.
4. Obtain the estimator of $p^0$ as $\hat{p} = \arg\max_{1 \leq k < p_{\max}} GR(k)$.

166
Remark 3.5. Some remarks on this procedure are in order. First, one can also simply run an $\ell_1$-nuclear-norm penalized regression with $p_{\text{max}}$, which is the first step of the estimation procedure given in Section 3.1. We only require that $\|\hat{A}_k - A^0_k\|_F$ converge to zero at a certain rate. Second, in practice one may obtain a very small or even zero value for $\|\hat{A}_k\|_F^2$ when $k > p^0$. In this case, if we directly use $a_k = \|\hat{A}_k\|_F^2$, the growth ratio may possibly choose a larger $p$ than $p^0$. To solve this problem, we bound $a_k$ below by some constant $c > 0$. Third, the GR($p$) criterion function is constructed to allowsome $A^0_k$ with $k < p^0$ to be a matrix of zeros. If we believe all the $A^0_k$ are nonzero matrices for $k \in [p^0]$, one can also consider the criterion function $FR(p) = a_p/a_{p+1}$, where the term $FR$ refers to the Frobenius norm ratio. Fourth, in order to allow $p^0$ to be divergent, one should allow $p_{\text{max}}$ to go to infinity.

4. Monte Carlo simulations

This section reports the results of a set of Monte Carlo experiments designed to evaluate the finite sample performance of the estimation procedures given above.

4.1. Data generating processes

We consider three cases with $p = 1$. For each data generating process (DGP), we generate data from the following high-dimensional VAR(1) system with CFs

$$Y_t = A^0_1 Y_{t-1} + A^0 f^0_t + u_t,$$

where $A^0_1$ varies across different DGPs, $A^0 = (\lambda^0_1, \ldots, \lambda^0_N)'$. The factor loadings $\lambda^0_r$ for $r = 1, \ldots, R^0$, are independently and identically distributed (i.i.d.) standard normal random variables. The factors $f^0_r$, for $r = 1, \ldots, R^0$, follow the autoregressive process

$$f^0_r = \rho_f \cdot f^0_{r-1} + \epsilon_r^{(f)},$$

where $\rho_f = 0.6$ and $\epsilon_r^{(f)}$ are i.i.d. $N(0, 1)$. The idiosyncratic errors are generated as $u_{it} = s \cdot \epsilon_{it}^{(u)}$, where $s$ controls the signal-to-noise ratio, and the $\epsilon_{it}^{(u)}$ are i.i.d. $N(0, 1)$.

**DGP 1** (Tridiagonal transition matrix): $(A^0_k)_{ij} = 0.3 \cdot 1(|i - j| \leq 1)$.

**DGP 2** (Block-diagonal transition matrix): We generate a block-diagonal matrix $A^0_k = \text{bdia}(S_1, \ldots, S_K)$, where the $S_k$’s are $5 \times 5$ random matrices. The diagonal entries of $S_k$ are fixed with $(S_k)_{ii} = 0.3$. In each column of $S_k$, we randomly choose 2 out of 4 off-diagonal entries and set them to be $-0.3$.

**DGP 3** (Random transition matrix): We fix the diagonal entries of $A^0_1$ to be 0.3 (i.e. $(A^0_1)_{ii} = 0.3$). In each row of $A^0_1$, we randomly choose 3 out of $N - 1$ entries and set them to be $-0.3$.

Fig. 1 illustrates the structure of the random transition matrices used in our simulation. For each DGP, we consider $N = 30, 60,$ and $T = 100, 200, 400$, leading to six combinations of cross-sectional and time series dimensions. The number of replications is set to be 500.

4.2. Implementation and estimation results

For each DGP, we consider the feasible estimator proposed in this paper and the oracle least squares estimator. The oracle estimators are obtained by using information regarding the true number of factors and the true regressors.
Table 1
Model selection accuracy.

<table>
<thead>
<tr>
<th>DGP</th>
<th>N</th>
<th>T</th>
<th>Number of factors</th>
<th>Step 1</th>
<th>Step 2</th>
<th>Step 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>UER</td>
<td>TPR</td>
<td>FPR</td>
</tr>
<tr>
<td>1</td>
<td>30</td>
<td>100</td>
<td>0.0%</td>
<td>0.0%</td>
<td>97.4%</td>
<td>19.3%</td>
</tr>
<tr>
<td>30</td>
<td>200</td>
<td>0.0%</td>
<td>0.0%</td>
<td>99.6%</td>
<td>19.1%</td>
<td>99.9%</td>
</tr>
<tr>
<td>30</td>
<td>400</td>
<td>0.0%</td>
<td>0.0%</td>
<td>99.9%</td>
<td>21.8%</td>
<td>100.0%</td>
</tr>
<tr>
<td>60</td>
<td>100</td>
<td>0.0%</td>
<td>0.0%</td>
<td>96.8%</td>
<td>12.7%</td>
<td>98.2%</td>
</tr>
<tr>
<td>60</td>
<td>200</td>
<td>0.0%</td>
<td>0.0%</td>
<td>99.9%</td>
<td>12.2%</td>
<td>100.0%</td>
</tr>
<tr>
<td>60</td>
<td>400</td>
<td>0.0%</td>
<td>0.0%</td>
<td>100.0%</td>
<td>11.9%</td>
<td>100.0%</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>100</td>
<td>0.0%</td>
<td>0.0%</td>
<td>86.2%</td>
<td>21.8%</td>
</tr>
<tr>
<td>30</td>
<td>200</td>
<td>0.0%</td>
<td>0.0%</td>
<td>95.3%</td>
<td>28.0%</td>
<td>93.7%</td>
</tr>
<tr>
<td>30</td>
<td>400</td>
<td>0.0%</td>
<td>0.0%</td>
<td>99.2%</td>
<td>37.0%</td>
<td>98.7%</td>
</tr>
<tr>
<td>60</td>
<td>100</td>
<td>0.0%</td>
<td>0.0%</td>
<td>76.7%</td>
<td>10.3%</td>
<td>76.5%</td>
</tr>
<tr>
<td>60</td>
<td>200</td>
<td>0.0%</td>
<td>0.0%</td>
<td>88.9%</td>
<td>12.5%</td>
<td>89.7%</td>
</tr>
<tr>
<td>60</td>
<td>400</td>
<td>0.0%</td>
<td>0.0%</td>
<td>96.4%</td>
<td>17.7%</td>
<td>95.8%</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>100</td>
<td>0.0%</td>
<td>0.0%</td>
<td>93.2%</td>
<td>24.9%</td>
</tr>
<tr>
<td>30</td>
<td>200</td>
<td>0.0%</td>
<td>0.0%</td>
<td>98.1%</td>
<td>31.4%</td>
<td>97.6%</td>
</tr>
<tr>
<td>30</td>
<td>400</td>
<td>0.0%</td>
<td>0.0%</td>
<td>99.5%</td>
<td>38.4%</td>
<td>99.3%</td>
</tr>
<tr>
<td>60</td>
<td>100</td>
<td>0.0%</td>
<td>0.0%</td>
<td>88.1%</td>
<td>12.8%</td>
<td>88.4%</td>
</tr>
<tr>
<td>60</td>
<td>200</td>
<td>0.0%</td>
<td>0.0%</td>
<td>96.1%</td>
<td>15.6%</td>
<td>95.5%</td>
</tr>
<tr>
<td>60</td>
<td>400</td>
<td>0.0%</td>
<td>0.0%</td>
<td>98.9%</td>
<td>19.5%</td>
<td>98.6%</td>
</tr>
</tbody>
</table>

Note: We report the under/over-estimation rate (UER and OER) of the number of factors in the UER and OER columns, respectively. The TPR (true positive rate) columns report the average shares of relevant variables included. The FPR (false positive rate) columns report the average shares of irrelevant variables included.

Table 1 reports the model selection accuracy. For each combination of $N$ and $T$ in each DGP, the fourth and fifth columns report the under- and over-estimation rate of $\hat{R}$, respectively. The TPR (true positive rate) columns report the average shares of relevant variables included. The FPR (false positive rate) columns report the average shares of irrelevant variables included. We summarize some important findings from Table 1. First, the proposed hard singular value thresholding (SVT) procedure can correctly determine the number of factors for each case. Second, with $N$ fixed, the TPR increases with $T$ in all cases as expected. All three-step estimators can include almost all the true regressors when $T = 400$. Third, among the three estimators, the third-step conservative Lasso estimator includes the least irrelevant regressors in almost all settings. In addition, only the conservative Lasso estimators tend to exclude more irrelevant regressors as $T$ increases, while the FPRs of the first and second step estimators increase as $T$ grows.

Table 2 reports the estimation errors of both the feasible estimators and the oracle least squares estimators. We report the root mean squared errors (RMSEs) for all entries and nonzero entries, respectively. We summarize some important findings from Table 2. First, as expected, the oracle least squares estimator uniformly outperforms the feasible estimators. This is mainly due to the fact that the FPRs of the feasible estimators were never zero. Second, the RMSE of the oracle estimator for nonzero entries decreases with $T$ at the $\sqrt{T}$-rate and alters with $N$ slightly. This is consistent with our theoretical prediction that the oracle least square estimator converges to the true values at the $\sqrt{T}$-rate. Third, the conservative Lasso outperforms the other two feasible estimators in terms of RMSEs in all cases.

For all DGPs, we also consider estimation of a misspecified VAR(1) model, $Y_t = A_0 Y_{t-1} + u_t$, where the common factors are ignored. We first estimate the model with a Lasso penalty as in KC (2015). Then we construct the weights as in (3.4) and use conservative Lasso to estimate the misspecified model. Table 3 reports the performance of these two estimators. We summarize some findings from Table 3. First, the FPRs for both estimators are quite high. This indicates that the misspecification may lead to non-sparse estimates of the transition matrices when the presence of strong cross-sectional dependence is not properly accounted for. Second, the estimators for the misspecified model also have higher RMSEs. Third, in many cases, the conservative Lasso estimator performs even worse than the Lasso estimator in terms of RMSEs. These findings show that it is important to take into account the presence of a factor structure in the estimation of a VAR with CFs.

5. Empirical application

5.1. Evaluating a network of financial assets volatilities

In recent years, financial asset connectedness has been an active topic in financial econometrics. Examples of contributions to this literature include Barigozzi and Brownlees (2019; hereafter BB), Barigozzi and Hallin (2017), Billio et al. (2012), Diebold and Yilmaz (2014; hereafter DY), Diebold and Yilmaz (2015), and Hautsch et al. (2014). Some of these authors directly model the large panels of time series they are studying as a VAR process without the potential presence of common factors. In this work a Lasso-type method has been employed to estimate the transition matrices. However,
Table 2
Root mean squared errors of the feasible and oracle transition matrix estimators.

<table>
<thead>
<tr>
<th>DGP</th>
<th>N</th>
<th>T</th>
<th>All entries</th>
<th>Oracle</th>
<th>Step 1</th>
<th>Step 2</th>
<th>Step 3</th>
<th>Nonzero entries</th>
<th>Oracle</th>
<th>Step 1</th>
<th>Step 2</th>
<th>Step 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30</td>
<td>100</td>
<td>0.019</td>
<td>0.063</td>
<td>0.059</td>
<td>0.050</td>
<td>0.062</td>
<td>0.145</td>
<td>0.132</td>
<td>0.120</td>
<td>0.117</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>100</td>
<td>0.012</td>
<td>0.047</td>
<td>0.049</td>
<td>0.027</td>
<td>0.039</td>
<td>0.142</td>
<td>0.150</td>
<td>0.103</td>
<td>0.070</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>100</td>
<td>0.013</td>
<td>0.050</td>
<td>0.049</td>
<td>0.044</td>
<td>0.054</td>
<td>0.204</td>
<td>0.205</td>
<td>0.196</td>
<td>0.179</td>
<td></td>
</tr>
</tbody>
</table>

Note: We report the root mean squared errors (RMSEs) of the feasible and oracle transition matrix estimators. Columns 4–7 report the RMSEs of all entries, and Columns 8–11 report the RMSEs of non-zero entries.

Table 3
Results of misspecified estimates.

<table>
<thead>
<tr>
<th>DGP</th>
<th>N</th>
<th>T</th>
<th>LASSO</th>
<th>Conservative LASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TPR</td>
<td>FPR</td>
<td>RMSE_a</td>
<td>RMSE_b</td>
</tr>
<tr>
<td>1</td>
<td>30</td>
<td>100</td>
<td>78.7%</td>
<td>34.9%</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>100</td>
<td>86.2%</td>
<td>59.6%</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>100</td>
<td>89.2%</td>
<td>59.2%</td>
</tr>
</tbody>
</table>

Note: We report the true positive rate (TPR), false positive rate (FPR), root mean squared errors of all entries (RMSE_a) and non-zero entries (RMSE_b) of misspecified estimates. We consider the LASSO estimator as in Kock and Callot (2015) and a conservative LASSO estimator. The LASSO estimator was used to construct weights for the conservative LASSO.

Barigozzi and Hallin (2017) and Barigozzi and Brownlees document evidence for the existence of a factor structure in volatility. Barigozzi and Hallin (2017) consider controlling for the presence of common factors by means of a dynamic factor model. BB (2019) use the regression residuals of individual volatilities on observed factors (e.g., market volatility or sector-specific volatility) to represent the idiosyncratic components of the volatilities. Neither of these papers provides a theoretical justification for the procedures employed.

In this empirical application, we extend the measure of connectedness of DY (2014) and study the connectedness of financial assets. More specifically, we explore connectedness in a panel of volatility measures. As remarked in DY (2014), the volatilities of financial assets can be interpreted as a form of 'investor fear'. Volatility connectedness may then be interpreted as representing 'fear connectedness' across assets. In this context it is natural to take into account common factors, which reflect confidence in the market. Spillover effects across assets is another reason for connectedness. We use the econometric methodology derived in the present work to analyze a panel of return volatilities of 23 sector ETF funds. The findings show that common factors account for 56.1% of the overall variability. Conditioning on these factors, the interdependence across individuals still captures a relatively high proportion of the variation.
5.1.1. Data description and empirical framework

We collect the weekly ‘open price’, ‘close price’, ‘high price’ and ‘low price’ of a series of sector ETF funds from Yahoo finance. The fund names and tickers are listed in Table S1 in the online supplement. The funds fall into 11 categories. The ‘Energy’, ‘Financial’ and ‘Consumer cyclicals’ are three large categories, each of which contains three to four funds. Each of the other categories contain at most two funds. The sample spans July 2007 to August 2019, which corresponds to 688 weeks. As volatility is unobserved, we use observed price data to estimate it. Specifically, we follow Garman and Klass (1980) and Alizadeh et al. (2002) to measure asset volatility as follows:

$$\hat{\sigma}_{it}^2 = 0.511(H_{it} - L_{it})^2 - 0.019[ (C_{it} - O_{it})(H_{it} + L_{it} - 2O_{it}) - 2(H_{it} - O_{it})(L_{it} - O_{it}) ] - 0.383(C_{it} - O_{it})^2,$$

where \(O_{it}, C_{it}, H_{it}\) and \(L_{it}\) are natural logarithms of weekly ‘open price’, ‘close price’, ‘high price’ and ‘low price’, respectively. Some descriptive statistics of the volatilities are presented in Table S2 in the online supplement. As in DY (2014) we normalize the data by taking natural logarithms and then center each time series. That is, our panel data variable \(y_{it}\) is given by \(\log(\hat{\sigma}_{it}^2) \rightarrow \log(\hat{\rho}_{it}^2)\).

Given the panel of volatilities, we fit the data to our VAR model with CFs in (2.1). From the decomposition (2.5), we have \(y_{it} = y_{i1}^{(f)} + y_{i1}^{(u)}\), where \(y_{i1}^{(f)}\) is due to the common factors and \(y_{i1}^{(u)}\) is due to the idiosyncratic errors. Then \(\nu_{it} \equiv \text{var}(y_{i1}^{(f)})/\text{var}(y_{it})\) measures the proportion of variance in \(y_{it}\) that is due to common factors and \(\nu \equiv \sum_{i=1}^{N}\text{var}(y_{i1}^{(f)})/\sum_{i=1}^{N}\text{var}(y_{it})\) measures the corresponding object across the whole panel.

For the idiosyncratic component \(y_{i1}^{(u)}\) we calculate the measure of connectedness proposed by DY (2014). As discussed in Section 2, we have \(y_{i1}^{(u)} = \sum_{j=0}^{\infty}\epsilon_{ij}(j)C^{(u)}e_{j,N}\), where \(\epsilon_{ij}(j) = (e_1 \otimes \epsilon_{i,j,N})\Phi(e_1 \otimes I_N)\) and \(\epsilon_{i,j,N} \sim (0, I_N)\). For simplicity, suppose that \(m = N\). Then one can treat \(\epsilon_{i,j,N}\) as the idiosyncratic shock to individual \(i\). The variance of the H-step ahead prediction error due to \(\epsilon_{ij}(j)\) is \(s_{ij}^H = \sum_{h=0}^{H-1}[(C^{(u)}(h)C^{(u)})_j]^2\). If we can identify both \(\Phi\) and \(C^{(u)}\), we can easily estimate the variance decomposition matrix \(\tilde{D}^H\) with \((i,j)\)th entry \(s_{ij}^H/\sum_{j=1}^{N}s_{ij}^H\). However, \(C^{(u)}\) is not identified without further assumption. Although we cannot identify \(C^{(u)}\), the matrix \(\Sigma_{i,j} = C^{(u)}C^{(u)\prime}\) is identified. DY (2014) propose to calculate the H-step generalized variance decomposition matrix by \(\tilde{D}^H = [\tilde{d}_{ij}N \times N]\), where

$$d_{ij}^H = \frac{\alpha_{ij}(0)\sum_{h=0}^{H-1}\alpha_{ij}(h)\Sigma_{ij}e_{j,N}^2}{\sum_{h=0}^{H-1}\alpha_{ij}(h)\Sigma_{ij}e_{j,N}^2},$$

Unlike \(\tilde{D}^H\), the row sums of \(\tilde{D}^H\) are not necessarily unity. We normalize \(D^H\) to \(\tilde{D}^H\) with \((i,j)\)th entry \(\tilde{d}_{ij}^H = d_{ij}^H/\sum_{j=1}^{N}d_{ij}^H\) so that \(\sum_{j=1}^{N}\tilde{d}_{ij}^H = 1\) and \(\sum_{i=1}^{N}\tilde{d}_{ij}^H = N\). Hence, the overall connectedness in the \(y_{i1}^{(u)}\)’s can be measured as \(\tilde{\nu}^H = \sum_{i,j}\tilde{d}_{ij}^H/N\). In addition, we let \(\tilde{d}_{i,j}^{\infty} \equiv \sum_{j \neq i}\tilde{d}_{ij}^H\). Following DY (2014), we call \(\tilde{d}_{i,j}^{\infty}\) the ‘FROM’ index, as it measures the proportion of generalized variance decomposition that is due to other individuals. Similarly, we let \(\tilde{d}_{i,j}^{\infty}\) call this the ‘TO’ index.

5.1.2. Estimation results

We use the procedure proposed in Section 3.4 to determine the lag length with \(p_{\text{max}} = 8\). The result gives \(\hat{p} = 4\). When we run the regression with \(p = 4\), the number of factors is determined to be one \((R = 1)\).

Fig. 2 reports the heat map representing the estimates of the \(\tilde{A}_k\)’s. The element value is represented by color intensity on the scale shown in the figure. In total, 330 out of 2116 entries are nonzero. There are three interesting findings. First, most of the nonzero entries are estimated to be positive. The positive coefficients represent propagation of investor fear across assets. Second, the diagonal elements of the \(\tilde{A}_k\)s are mostly nonzero. The magnitude of the diagonal elements is larger than that of the off diagonal elements on average. Third, the number of nonzero coefficients in \(\tilde{A}_k\) decreases as \(k\) increases and the average magnitude of the entries also decreases. These results suggest that more recent investor fear is more influential in raising present investor fear.

Next, we calculate the statistics introduced in the last subsection. The upper panel of Table 4 provides the estimates of \(\nu_{it}, \tilde{d}_{i,j}^\infty,\) and \(\tilde{d}_{i,j}^\infty\). Almost all the \(\nu_{it}\)’s are above 50%, and the overall variation due to the common factors is \(\nu = 56.1\%.\). These results imply that market level investor fear plays a dominant role in investor trading behavior. After conditioning on the factors, we consider the idiosyncratic part by looking at \(\tilde{d}_{i,j}^\infty, \tilde{d}_{i,j}^\infty,\) and the H-step generalized variance decomposition matrix \(\tilde{D}^H\). The ‘FROM’ index ranges between 27.7% and 71.7%. Interestingly, the ‘energy’ and ‘finance’ funds have higher ‘FROM’ index compared to other funds. A similar observation applies for the ‘TO’ index. Specifically, the ‘TO’ index of XLE and IYE are close to 100% and both are ‘energy’ funds. The energy industry is therefore instrumental in transmitting considerable investor fear to the entire market. This finding has a strong intuitive basis as oil prices have been extremely volatile in recent years and energy prices affect all industries. The fund GDX (VanEck Vectors Gold Miners ETF) has the

\[\text{12}\text{ The generalized variance decomposition (GVD) framework was introduced by Koop et al. (1996) and Pesaran and Shin (1998). It provides an order-invariant framework to estimate the variance decomposition. For more details see Section 2.3 of DY (2014).}\]
Fig. 2. Heatmap of the transition matrices $A_k$’s. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

least connectedness. It receives only 27.7% connectedness from other assets and transmits only 19.1% connectedness to others. The overall connectedness measure is 49.8%. Conditioning on the factors, there is still substantive transmission of investor fear across individual assets. Fig. 3 reports the heat map of the H-step generalized variance decomposition matrix $\tilde{D}_H$ at $H = 12$. We observe that the interconnections within the same category are high, whereas connectedness across categories is relatively low.

The lower panel of Table 4 provides measures of connectedness based on pure VAR model estimation as in Demirer et al. (2018). Without controlling for common factors, the ‘FROM’ and ‘TO’ index of each fund becomes much larger. Importantly, little heterogeneity is now observed across categories. These results indicate that all the connectedness due to common factors is now absorbed and interpreted as individual level connectedness, leading to potentially misleading inferences.

In sum, our framework extends traditional VAR analyses of financial asset connectedness to control for the presence of possible common factors in the determination of volatility. This extension leads to new interpretations of the data that give a prominent role to the presence of a single common factor in volatility connectedness. Our results show that this common factor accounts for more than a half of the variation in the data, thereby contributing substantially to observed connectedness. But even allowing for the influence of this common factors there is still a remarkable degree of connectedness arising from spillover channels that operate among the assets themselves.

6. Conclusion

This paper proposes a methodology to study the properties of regularized estimates of high-dimensional VARs with unobserved common factors. The presence of common factors introduces strong cross sectional dependence into the process. Incorporating such dependence is particularly important in high-dimensional disaggregated data where connectedness between the variables may arise through different channels. Dependence and connectedness are found to be especially relevant in studying the transmission of investor fear across financial assets in our study of asset price volatility.
### Table 4
Connectedness measures across funds.

<table>
<thead>
<tr>
<th>TICKER</th>
<th>XLE</th>
<th>XOP</th>
<th>IYE</th>
<th>OIH</th>
<th>XLF</th>
<th>KBE</th>
<th>KRE</th>
<th>XLY</th>
</tr>
</thead>
<tbody>
<tr>
<td>FROM</td>
<td>64.9%</td>
<td>59.1%</td>
<td>65.4%</td>
<td>58.0%</td>
<td>65.2%</td>
<td>56.8%</td>
<td>56.6%</td>
<td>72.0%</td>
</tr>
<tr>
<td>TO</td>
<td>106.8%</td>
<td>86.0%</td>
<td>103.9%</td>
<td>71.5%</td>
<td>57.8%</td>
<td>72.6%</td>
<td>51.4%</td>
<td>37.3%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TICKER</th>
<th>XHB</th>
<th>ITB</th>
<th>XRT</th>
<th>IYR</th>
<th>VNQ</th>
<th>XLB</th>
<th>XME</th>
<th>XLK</th>
</tr>
</thead>
<tbody>
<tr>
<td>FROM</td>
<td>53.6%</td>
<td>49.5%</td>
<td>60.1%</td>
<td>50.7%</td>
<td>49.7%</td>
<td>67.2%</td>
<td>56.9%</td>
<td>70.5%</td>
</tr>
<tr>
<td>TO</td>
<td>56.3%</td>
<td>41.7%</td>
<td>19.0%</td>
<td>79.7%</td>
<td>74.4%</td>
<td>26.3%</td>
<td>37.2%</td>
<td>37.3%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TICKER</th>
<th>SMH</th>
<th>XLV</th>
<th>IBB</th>
<th>XLP</th>
<th>XLU</th>
<th>XLI</th>
<th>GDX</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>FROM</td>
<td>54.8%</td>
<td>64.3%</td>
<td>50.7%</td>
<td>61.3%</td>
<td>50.6%</td>
<td>67.7%</td>
<td>31.0%</td>
<td>56.1%</td>
</tr>
<tr>
<td>TO</td>
<td>60.5%</td>
<td>58.3%</td>
<td>36.5%</td>
<td>57.9%</td>
<td>58.6%</td>
<td>37.5%</td>
<td>44.1%</td>
<td>39.0%</td>
</tr>
</tbody>
</table>

### Connectedness measures by estimates of pure VAR model

<table>
<thead>
<tr>
<th>TICKER</th>
<th>XLE</th>
<th>XOP</th>
<th>IYE</th>
<th>OIH</th>
<th>XLF</th>
<th>KBE</th>
<th>KRE</th>
<th>XLY</th>
</tr>
</thead>
<tbody>
<tr>
<td>FROM</td>
<td>89.3%</td>
<td>87.1%</td>
<td>89.4%</td>
<td>87.0%</td>
<td>89.6%</td>
<td>86.8%</td>
<td>87.6%</td>
<td>90.9%</td>
</tr>
<tr>
<td>TO</td>
<td>105.0%</td>
<td>79.5%</td>
<td>103.0%</td>
<td>77.7%</td>
<td>112.9%</td>
<td>97.0%</td>
<td>89.1%</td>
<td>110.5%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TICKER</th>
<th>SMH</th>
<th>XLV</th>
<th>IBB</th>
<th>XLP</th>
<th>XLU</th>
<th>XLI</th>
<th>GDX</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>FROM</td>
<td>87.3%</td>
<td>86.3%</td>
<td>88.8%</td>
<td>85.7%</td>
<td>86.2%</td>
<td>90.1%</td>
<td>88.8%</td>
<td>89.8%</td>
</tr>
<tr>
<td>TO</td>
<td>95.8%</td>
<td>80.8%</td>
<td>79.1%</td>
<td>94.0%</td>
<td>89.6%</td>
<td>105.6%</td>
<td>80.1%</td>
<td>103.8%</td>
</tr>
</tbody>
</table>

| FROM | 87.6% | 88.1% | 83.8% | 88.4% | 85.7% | 89.8% | 76.5% | 76.4% |
| TO   | 74.8% | 81.2% | 60.8% | 80.0% | 60.0% | 104.3% | 45.8% | 45.8% |

Note: Cyc, Rea, Natu, Tech, Heal, Def, Util, Indu and EMP stand for consumer cyclical, real estate, natural resource, technology, health care, consumer defensive, utilities, industrials and equity precious metals, respectively.

**Fig. 3.** Heat map of $\tilde{D}^{12}$.

In practice, our procedure is implemented as follows. First, given the order $p$ of the VAR process, which can be estimated via a growth ratio criterion, we obtain preliminary estimates of the transition matrices and common component via $\ell_1$-nuclear norm regularizations, with which one can estimate the number of factors consistently and obtain a preliminary consistent estimate of the common factors. Second, we estimate the model using a generalized Lasso procedure by including the preliminary estimate of the common factors as regressors. In the third stage conservative Lasso is used to obtain the final estimates, which are shown to be asymptotically equivalent to the oracle least squares estimates.

The methods and results in this paper open up multiple avenues for further research. First, following BB (2019) it may be useful in practice to impose some sparsity assumptions on the large dimensional error variance matrix and develop estimation methods to achieve this. Second, the model studied here does not allow for structural change in
the transition matrices or the factor loadings (c.f., Su and Wang (2017)). It will also be interesting and challenging to study high-dimensional VAR models with common factors that may involve time-varying transition matrices and factor loadings, which can help capture empirical evolution in institutional and regulatory frameworks. Third, the framework and methods provide the facility to implement Granger-causality testing in the presence of common factors. Existing Granger-causality tests mainly focus on low-dimensional VAR models, most often bivariate or trivariate VAR models. Exceptions include Hecq et al. (2020) and Fan et al. (2020) who consider Granger causality tests in high-dimensional VARs based on post-double-selection and debiased estimators, respectively, but do not allow for strong cross-sectional dependence. These avenues of future work provide many options for further technical and applied research on high-dimensional VAR systems.

Appendix A. Proofs of the main results

Proof of Proposition 2.1. (i) By Assumption A.1(iv), the $y^{(u)}_t$ and $y^{(f)}_t$ are mutually independent. It suffices to study them separately. By Assumption A.1(i), we can write $y^{(u)}_t$ as the linear process

$$y^{(u)}_t = \sum_{j=0}^{\infty} c^{(u)}_j \epsilon_{t-j} = \sum_{j=0}^{\infty} \alpha^{(u)}_j(c^{(u)}_j | \epsilon_{t-j} = \sum_{j=0}^{\infty} c^{(j)}_j \epsilon_{t-j},$$

where $c^{(i)}_j \equiv \alpha^{(u)}_j(c^{(u)}_j).$ Under Assumption A.1(vii), one can bound $\| (e_{1,p} \otimes e_{1,n})^\prime \| \psi_{\text{max}}(\| \Phi \|_{1,1}) \leq \bar{c} \rho^{j}.$ It follows that $\| e^{(u)}_j \| \leq \bar{c} \rho^{j}.$ Then the MA($\infty$) representation of $y^{(u)}_t$ is valid with $E(y^{(u)}_t) = 0$ and $\text{Var}(y^{(u)}_t) = \sum_{j=0}^{\infty} \alpha^{(u)}_j(c^{(u)}_j) < \infty.$

Under Assumption A.1(vii), we can also show that $E(|y^{(f)}_t|) \leq \sum_{j=0}^{\infty} | \alpha^{(f)}_j | \mu_f \leq \infty.$ The MA($\infty$) representation of $y^{(f)}_t$ is

$$y^{(f)}_t = E(y^{(f)}_t) + \sum_{j=0}^{\infty} \alpha^{(f)}_j(c^{(f)}_j) \epsilon_{t-j} = E(y^{(f)}_t) + \sum_{j=0}^{\infty} c^{(f)}_j \epsilon_{t-j},$$

where $c^{(f)}_j \equiv \sum_{k=0}^{\infty} \alpha^{(f)}_j(k) c^{(f)}_{j-k}.$ Under Assumption A.1(vi), $\| c^{(f)}_j \| \leq \sum_{k=0}^{\infty} \| c^{(f)}_j(k) \| \| c^{(f)}_{j-k} \|$ and $\text{Var}(y^{(f)}_t) \leq \sum_{j=0}^{\infty} \| \alpha^{(f)}_j \| \| c^{(f)}_{j-k} \| \| c^{(f)}_{j-k} \| \| \epsilon_{t-j} \| \leq \bar{c} \sum_{k=0}^{\infty} \rho^{k} (k+1)^{-\beta},$

for some constant $\bar{c} < \infty.$ Hence $c^{(f)}_j$ is absolutely summable, $\text{Var}(y^{(f)}_t) = \sum_{j=0}^{\infty} c^{(f)}_j c^{(f)}_j \leq \infty,$ and the MA($\infty$) representation of $y^{(f)}_t$ is valid.

Similar to the decomposition (2.5), we can write $X_t = X^{(u)}_t + X^{(f)}_t.$ For $X_t$, due to the independence between $X^{(u)}_t$ and $X^{(f)}_t$, we can also write $\Sigma_X = \Sigma^{(u)}_X + \Sigma^{(f)}_X,$ where $\Sigma^{(u)}_X \equiv E(X^{(u)}_t X^{(u)}_t)$ and $\Sigma^{(f)}_X \equiv E(X^{(f)}_t X^{(f)}_t).$ Since $\Sigma^{(f)}_X$ is positive semi-definite, we have $\psi_{\text{min}}(\Sigma_X) \geq \psi_{\text{min}}(\Sigma^{(u)}_X).$ It suffices to show $\psi(\Sigma^{(f)}_X)$ is bounded below. By Proposition 2.3 of BM (2015), we have

$$\psi_{\text{min}}(\Sigma^{(u)}_X) \geq \psi_{\text{min}}(\Sigma_A) \max_{x_{1}=1} \psi_{\text{max}}(A(z),A(z)).$$

Given Assumption A.1(vii), we have that $\psi_{\text{min}}(\Sigma^{(u)}_X)$ is bounded below by some constant.

(ii) By virtue of the independence between $X^{(u)}_t$ and $X^{(f)}_t$, it can also be shown that $\psi_{\text{min}}(\Sigma_X)$. ■

A.1. Analysis of the first-step estimators

To prove Theorem 3.1, we need the following two lemmas whose proofs can be found in the online supplement.

**Lemma A.1.** For the $T \times N$ matrices $\Theta^0$ and $\Delta,$ we have

(i) $\| \Theta^0 + M(\Delta) \|_* \leq \| \Theta^0 \|_* + \| M(\Delta) \|_*;
(ii) \| \Delta \|_F^2 = \| M(\Delta) \|_F^2 + \| P(\Delta) \|_F^2;
(iii) rank(P(\Delta)) \leq 2R^2;
(iv) \| \Delta \|_F^2 = \sum_{j=1}^p \psi_{\text{min}}(\lambda_j) \lambda_j^2$ and $\| \Delta \|_F^2 \leq \| \Delta \|_F^2 \text{rank}(\Delta);
For any conformable matrices $M_1$ and $M_2,$ the following statement holds:
(v) $| \text{tr}(M_1 M_2) | \leq \| M_1 \|_\infty \| M_2 \|_1 \text{ and } | \text{tr}(M_1 M_2) | \leq \| M_1 \|_F \| M_2 \|_F \leq \| M_1 \|_F \| M_2 \|_\infty.$
Lemma A.2. Suppose that Assumption A.1 hold. There exist absolute constants $c, c_1, c_2 \in (0, \infty)$ such that
(i) $\| \mathbf{UX} \|_{\text{max}} / T \leq \gamma_1 / 2$ with probability greater than $1 - c_1 \tilde{c}(pN^2 T^{1-\eta/4} (\log N)^{-\eta/2} + pN^{-\epsilon} \log N)$;
(ii) $\| \mathbf{Y}^{\gamma_2} \mathbf{X} \|_{\text{max}} / T \leq c \cdot \gamma_1$ with probability greater than $1 - \tilde{c}_1[pN(T^{1-\eta/4} (\log N)^{-\eta/2} \vee e^{2T}) + pN^{1-\epsilon} \log N)]$.

Proof of Theorem 3.1. Let $\tilde{\Delta}^{(1)} = \tilde{\mathbf{B}} - \mathbf{B}^0$ and $\tilde{\Delta}^{(2)} = \tilde{\Theta} - \Theta^0$. Define the event
$$\mathcal{E}^{(1)}_{NT} = \{ \| \mathbf{UX} \|_{\text{max}} / T \leq \gamma_1 / 2, \| \mathbf{U} \|_{\text{op}} / \sqrt{NT} \leq \gamma_2 / 2 \}.$$ 
By Lemma A.2(i) and Assumption A.3(i), $\mathcal{E}^{(1)}_{NT}$ holds with probability at least $1 - \tilde{c}(pN^2 T^{1-\eta/4} (\log N)^{-\eta/2} + pN^{-\epsilon} \log N)$. By the definition of $(\tilde{\mathbf{B}}, \tilde{\Theta})$, we have
$$0 \geq \mathcal{L}(\tilde{\mathbf{B}}, \tilde{\Theta}) - \mathcal{L}(\mathbf{B}^0, \Theta^0) = \frac{1}{2NT} \| \mathbf{Y} - \mathbf{XB} - \tilde{\Theta} \|_F^2 - \| \mathbf{U} \|_F^2 + \frac{\gamma_1}{N} \| \text{vec}(\tilde{\mathbf{B}}) \|_1 - \| \text{vec}(\mathbf{B}^0) \|_1 + \frac{\gamma_2}{\sqrt{NT}} (\| \tilde{\Theta} \|_* - \| \Theta^0 \|_*)$$
$$= d_1 + d_2 + d_3.$$ (A.1)

To establish the asymptotic properties of $\tilde{\mathbf{B}}$ and $\tilde{\Theta}$, we study $d_1, d_2$ and $d_3$ in turn. First, consider $d_1$. By the identity $\mathbf{Y} = \mathbf{XB}^0 + \Theta^0 + \mathbf{U}$, we have
$$\| \mathbf{Y} - \mathbf{XB} - \tilde{\Theta} \|_F^2 = \| \mathbf{X} \tilde{\Delta}^{(1)} + \tilde{\Delta}^{(2)} \|_F^2 - 2 \text{tr}[\mathbf{U}(\mathbf{X} \tilde{\Delta}^{(1)} + \tilde{\Delta}^{(2)})].$$
For $\text{tr}[\mathbf{U}(\mathbf{X} \tilde{\Delta}^{(1)} + \tilde{\Delta}^{(2)})]$, conditional on $\mathcal{E}^{(1)}_{NT}$, we apply the triangle inequality and Lemma A.1(v) to obtain
$$\frac{1}{NT} \text{tr}[\mathbf{U}(\mathbf{X} \tilde{\Delta}^{(1)} + \tilde{\Delta}^{(2)})] \leq \frac{1}{NT} \| \mathbf{UX} \|_{\text{max}} \| \text{vec}(\tilde{\Delta}^{(1)}) \|_1 + \frac{1}{NT} \| \mathbf{U} \|_{\text{op}} \| \tilde{\Delta}^{(2)} \|_*$$
$$\leq \frac{\gamma_1}{2N} \| \text{vec}(\tilde{\Delta}^{(1)}) \|_1 + \frac{\gamma_2}{2N} \| \tilde{\Delta}^{(2)} \|_*.$$ It follows that
$$d_1 \geq \frac{1}{2NT} \| \mathbf{X} \tilde{\Delta}^{(1)} + \tilde{\Delta}^{(2)} \|_F^2 - \frac{\gamma_1}{2N} \| \text{vec}(\tilde{\Delta}^{(1)}) \|_1 - \frac{\gamma_2}{2N} \| \tilde{\Delta}^{(2)} \|_*$$
$$\geq \frac{1}{2NT} \| \mathbf{X} \tilde{\Delta}^{(1)} + \tilde{\Delta}^{(2)} \|_F^2 - \frac{\gamma_1}{2N} \sum_{i=1}^N \| \tilde{\Delta}^{(1)}_{i\cdot} \|_1 + \tilde{\Delta}^{(1)}_{i\cdot} \|_1$$
$$- \frac{\gamma_2}{2N} \sum_{i=1}^N \| \text{vec}(\tilde{\Delta}^{(1)}_{i\cdot}) \|_* + \| \tilde{\Delta}^{(2)} \|_*.$$ (A.2)

Next, consider $d_2$. By the identities $\tilde{\mathbf{B}}_{\cdot i1} = \tilde{B}_{\cdot i1} + [\tilde{\mathbf{B}}_{i\cdot}]_{11}$ and $|B^0_{\cdot i1}|_1 = |B^0_{i\cdot}]_{11}$, we have
$$d_2 = \frac{\gamma_1}{N} \sum_{i=1}^N \| \tilde{\mathbf{B}}_{\cdot i1} + [\tilde{\mathbf{B}}_{i\cdot}]_{11} - |\mathbf{B}^0_{\cdot i1}|_1 \|_1 \geq \frac{\gamma_1}{N} \sum_{i=1}^N \| \tilde{\mathbf{B}}_{\cdot i1} + [\tilde{\mathbf{B}}_{i\cdot}]_{11} - |\tilde{\mathbf{B}}_{i\cdot}]_{11} \|_1,$$ (A.3)
where we use the fact that $|\tilde{\mathbf{B}}_{\cdot i1} + [\tilde{\mathbf{B}}_{i\cdot}]_{11}| \geq |\mathbf{B}^0_{\cdot i1}|$ by the triangle inequality and that $|\tilde{\mathbf{B}}_{i\cdot}]_{11} = |\tilde{\mathbf{B}}_{i\cdot}]_{11}$ as $B^0_{\cdot i1} = 0$.

Now, consider $d_3$. By the triangle inequality and Lemma A.1(i), we have
$$\| \tilde{\Theta} \|_* = \| \tilde{\Delta}^{(2)} + \Theta^0 \|_* = \| \Theta^0 + \mathcal{P}(\tilde{\Delta}^{(2)}) + \mathcal{M}(\tilde{\Delta}^{(2)}) \|_*$$
$$\geq \| \Theta^0 \|_* + \mathcal{M}(\tilde{\Delta}^{(2)}) \|_* - \| \mathcal{P}(\tilde{\Delta}^{(2)}) \|_*$$
$$= \| \Theta^0 \|_* + \| \mathcal{M}(\tilde{\Delta}^{(2)}) \|_* - \| \mathcal{P}(\tilde{\Delta}^{(2)}) \|_*.$$ It follows that
$$d_3 \geq \frac{\gamma_2}{\sqrt{NT}} (\| \mathcal{M}(\tilde{\Delta}^{(2)}) \|_* - \| \mathcal{P}(\tilde{\Delta}^{(2)}) \|_*).$$ (A.4)

Combining the results in (A.1)–(A.4), we have
$$\frac{1}{2NT} \| \mathbf{X} \tilde{\Delta}^{(1)} + \tilde{\Delta}^{(2)} \|_F^2 + \frac{\gamma_1}{2N} \sum_{i=1}^N \| \tilde{\Delta}^{(1)}_{i\cdot} \|_1 + \frac{\gamma_2}{2N} \| \mathcal{M}(\tilde{\Delta}^{(2)}) \|_*$$
$$\leq \frac{3\gamma_1}{2N} \sum_{i=1}^N \| \tilde{\Delta}^{(1)}_{i\cdot} \|_1 + \frac{3\gamma_2}{2N} \| \mathcal{P}(\tilde{\Delta}^{(2)}) \|_*.$$ (A.5)
The above inequality indicates that \((\tilde{\Lambda}^{(1)}, \tilde{\Lambda}^{(2)}) \in C_{NT}(3)\). By Assumption A.2, with probability \(1 - \epsilon_{NT}\) we have

\[
\frac{1}{N} \| \tilde{\Lambda}^{(1)} \|_F^2 + \frac{1}{NT} \| \tilde{\Lambda}^{(2)} \|_F^2 - \kappa_2 \Phi_{\gamma_1, \gamma_2}(\tilde{\Lambda}^{(1)}, \tilde{\Lambda}^{(2)}) \leq \kappa_1 \frac{1}{NT} \| X \tilde{\Lambda}^{(1)} + \tilde{\Lambda}^{(2)} \|_F^2,
\]

(A.6)

where \(\kappa_1 = (\kappa \wedge \kappa')^{-1}\) and \(\kappa_2 = \kappa_1 \kappa''\). By the inequality (A.5), we have

\[
\frac{1}{NT} \| X \tilde{\Lambda}^{(1)} + \tilde{\Lambda}^{(2)} \|_F^2 \leq 3 \Phi_{\gamma_1, \gamma_2}(\tilde{\Lambda}^{(1)}, \tilde{\Lambda}^{(2)}) = \frac{3Y_1}{N} \sum_{i=1}^N |\tilde{\Lambda}_{j_i, i}|_1 + \frac{3Y_2}{\sqrt{NT}} \| \mathcal{P}(\tilde{\Lambda}^{(2)}) \|_\infty
\]

\[
\leq 3 \sqrt{Y_1 \kappa_0} \| \tilde{\Lambda}^{(1)} \|_F + 3 \sqrt{Y_2} \| \tilde{\Lambda}^{(2)} \|_F \leq \frac{3 \sqrt{2} (\sqrt{Y_1 \kappa_0} \vee \sqrt{Y_2}) \left( \frac{1}{N} \| \tilde{\Lambda}^{(1)} \|_F^2 + \frac{1}{NT} \| \tilde{\Lambda}^{(2)} \|_F^2 \right)}{\sqrt{2} \sqrt{Y_1 \kappa_0} \vee \sqrt{Y_2}},
\]

(A.7)

where the second inequality holds by Lemma A.1(ii)–(iv) and the fact that \(\sum_{i=1}^N |\tilde{\Lambda}_{j_i, i}|_1 \leq \sqrt{N \kappa_0} \left( \sum_{i=1}^N |\tilde{\Lambda}_{j_i, i}|^2 \right)^{1/2} \leq \sqrt{N \kappa_0} \| \tilde{\Lambda}^{(1)} \|_F\), where recall that \(K_0 = N^{-1} \sum_{i=1}^N k_i\) and \(k_i \equiv |J_i|\) denotes the cardinality of the set \(J_i\). Combining (A.6)–(A.7), we have with probability at least \((1 - \epsilon_{NT}) (1 - c[pN^2 T^{1-\eta/4}(\log N)^{-\eta/4} + pN^{2-\eta} \log N]) \geq 1 - \epsilon_{NT} - c[pN^2 T^{1-\eta/4}(\log N)^{-\eta/2} + pN^{2-\eta} \log N]]\),

\[
\frac{1}{N} \| \tilde{\Lambda}^{(1)} \|_F^2 + \frac{1}{NT} \| \tilde{\Lambda}^{(2)} \|_F^2 \leq 3(k_1 + k_2) \sqrt{2} (\sqrt{Y_1 \kappa_0} \vee \sqrt{Y_2}) \left( \frac{1}{N} \| \tilde{\Lambda}^{(1)} \|_F^2 + \frac{1}{NT} \| \tilde{\Lambda}^{(2)} \|_F^2 \right),
\]

which implies that \(\frac{1}{N^2} \| \tilde{\Lambda}^{(1)} \|_F \leq \tilde{c}(\sqrt{Y_1 \kappa_0} \vee Y_2)\) and \(\frac{1}{N^2} \| \tilde{\Lambda}^{(2)} \|_F \leq \tilde{c}(\gamma_1 \sqrt{K_0} \vee Y_2)\) with \(\tilde{c} = (3k_1 + k_2) \sqrt{2} (1 + \sqrt{Y_2}) < \infty\). This completes the proof.

To prove Theorem 3.2, we need the following lemma which is proved in the online supplement.

**Lemma A.3.** Suppose that Assumptions A.1 and A.3 hold. Let \(S_F \equiv F^0 F^0 / T\). Then for any \(x > 0\),

\[
P(T^{1/2} \| S_F - \Sigma_F \|_{max} > x) \leq C_1 x^{-q/2} T^{1-n/4} + C_2 \exp \left( -C_3 x^2 \right)
\]

for some absolute constants \(C_i, \ell = 1, 2, 3\).

**Proof of Theorem 3.2.** We operate conditional on the event that

\[
c_{NT}^{(2)} = \left\| U^X \right\|_{max} / T \leq \gamma_1 / 2, \left\| U \right\|_{op} / \sqrt{NT} \leq \gamma_2 / 2 \text{ and } \left\| S_F - \Sigma_F \right\|_{op} \leq c \sqrt{\log NT^{-1/2}},
\]

where \(c\) is a large positive constant. One can verify that for some positive constants \(\tilde{c}'\) and \(c\),

\[
P(c_{NT}^{(2)}) \geq 1 - c' (pN^2 T^{1-n/4}(\log N)^{-n/4} + pN^{2-\eta} \log N\]

by Lemmas A.2–A.3. From Theorem 3.1, we have with probability at least \((1 - \epsilon_{NT} - \tilde{c}'(pN^2 T^{1-n/4}(\log N)^{-n/4} + pN^{2-\eta} \log N\]

\[
(NT)^{-1/2} \| \tilde{\Theta} - \Theta^0 \|_{op} \leq (NT)^{-1/2} \| \tilde{\Theta} - \Theta^0 \|_{F} \leq \tilde{c}(\gamma_1 \sqrt{K_0} \vee Y_2).\]

Next, we show that \(c_{NT}^{(2)}\) implies the desired results.

**Step 1: Bound the eigenvalues.**

Let \(S_A = A^0 A^0 / N\) and \(S_F = F^0 F^0 / T\). Let \(s_1 \geq \cdots \geq s_{p^0}\) be the \(R_0\) nonzero eigenvalues of \(1/N \Theta^0 \Theta^0 = F^0 S_A F^0\). Note that \(s_1, \ldots, s_{p^0}\) are the same as the eigenvalues of \(C^1_{F^0} S_A C^1_{F^0}\). Conditional on the event \(c_{NT}^{(2)}\) and by Assumption A.4(i)–(ii), we have

\[
|s_j - s_i| \leq \tilde{c}(\sqrt{\log NT^{-1/2} + N^{-1/2}}) \text{ for some } \tilde{c} < \infty \text{ and } d_j \in [R_0].
\]

This also implies that \(\| \Theta^0 \|_{op} = \sqrt{(s_1 + o_F(1)) NT}\). For \(j > R_0\), simply define \(s_j = s_j = 0\).

Let \(s_1 \geq \cdots \geq s_{p^0}\) be the eigenvalues of \(1/N \tilde{\Theta} \tilde{\Theta}^\dagger\). Again by the Weyl’s theorem, we have for \(j = 1, 2, \ldots\)

\[
|s_j - s_k| \leq \frac{1}{N} \| \tilde{\Theta} \tilde{\Theta}^\dagger - \Theta^0 \Theta^0 \|_{op} + |s_j - s_k|
\]

\[
\leq \frac{2}{NT} \| \tilde{\Theta} \tilde{\Theta}^\dagger - \Theta^0 \|_{op}^2 + \frac{1}{NT} \| \tilde{\Theta} - \Theta^0 \|_{op}^2 + |s_j - s_k|,
\]

implying \(|s_j - s_k| \leq \tilde{c}(\gamma_1 \sqrt{K_0} \vee Y_2)\) for \(j = 1, 2, \ldots\). Then for \(j \in [R_0]\), w.p.a.1,

\[
|s_j - s_k| \geq |s_{j-1} - s_{j-1}| - |s_j - s_k| \geq (s_{j-1} - s_{j-1}) / 2 \text{ and}
\]

\[
\tilde{c}(\gamma_1 \sqrt{K_0} \vee Y_2).
\]
\[ |\tilde{s}_j - \tilde{s}_{j+1}| \geq |\bar{s}_j - \bar{s}_{j+1}| - |\bar{s}_j - \bar{s}_i| \geq (\bar{s}_j - \bar{s}_{j+1})/2, \]  
(A.8)

with \( \tilde{s}_{\rho+1} = s_{\rho+1} = 0 \).

**Step 2: Prove the consistency of \( \hat{\theta} \).**

Note that \( \psi_1(\tilde{\theta}) = \sqrt{NT}\tilde{s}_0 \). By the result in Step 1, we have that \( \psi_1(\tilde{\theta}) \geq \sqrt{|s_{\rho} - \sigma(T)|}NT \) for all \( r \leq R^0 \), and

\[
\psi_{\rho+1}(\hat{\theta}) \leq \psi_{\rho+1}(\theta^0) + \| \hat{\theta} - \theta^0 \|_p \leq \| \hat{\theta} - \theta^0 \|_F \leq \sqrt{NT}(\gamma_1\sqrt{K_0} \vee \gamma_2) = \sqrt{NT}o(\gamma_2^{1/2})
\]

where we use the condition that \( \gamma_1\sqrt{K_0} = o(\gamma_2^{1/2}) \) under Assumption A.4(iii). These results, in conjunction with the fact that \( (\gamma_2\sqrt{NT}|\hat{\theta}||_p^{1/2}) > \sqrt{NT}\sqrt{\gamma_2} \) with \( \gamma_2 = c_2(N^{-1/2} + T^{-1/2}) \), imply that

\[
\min_{\rho \leq \rho_0} \psi_1(\tilde{\theta}) \geq (\gamma_2\sqrt{NT}|\hat{\theta}||_p^{1/2}) \text{ and } \psi_{\rho+1}(\hat{\theta}) < (\gamma_2\sqrt{NT}|\hat{\theta}||_p^{1/2})^{1/2}
\]

with probability at least \( 1 - \epsilon_{NT} - \tilde{c}(N^2T^{-1/4}|\log N|^{-q/4} + N^2|\log N|^{-q/4} + N^2|\log N|) \) for sufficiently large \( (N, T) \). Then we have \( \hat{\theta} = \rho^0 \) with probability at least \( 1 - \epsilon_{NT} - \tilde{c}(N^2T^{-1/4}|\log N|^{-q/4} + N^2|\log N|) \) for sufficiently large \( (N, T) \).

**Step 3: Characterize the eigenvectors.**

Next, we show that there is an \( \rho^0 \times \rho^0 \) matrix \( \tilde{H} \), such that the columns of \( \frac{1}{\sqrt{T}}F^0\tilde{H} \) are the first \( \rho^0 \) eigenvectors of \( \theta^0\theta^0 \). Let \( v \) be the \( \rho^0 \times \rho^0 \) matrix whose columns are the eigenvectors of \( \rho^{1/2}\rho\rho^{1/2} \). Then \( D = v\rho^{1/2}\rho\rho^{1/2}v \) is a diagonal matrix of the eigenvalues of \( \rho^{1/2}\rho\rho^{1/2} \) that are distinct by Assumption A.4(ii). Let \( \tilde{H} = S^{-1/2}v \). Then

\[
\frac{1}{NT}\theta^0\theta^0F^0\tilde{H} = \frac{1}{T}F^0\rho\rho^0F^0\tilde{H} = F^0\rho\rho\tilde{H} = F^0\rho\rho^{1/2}v
\]

\[
= F^{1/2}S^{-1/2}\rho^{1/2}S^{1/2}v = F^{1/2}S^{-1/2}vS^{1/2}v = F^{1/2}\tilde{H}D.
\]

In addition, we have \( (F^0\tilde{H})F^0\tilde{H}/T = vS^{1/2}S^{1/2}v = v = I_{\rho^0} \). So the columns of \( \frac{1}{\sqrt{T}}F^0\tilde{H} \) are the eigenvectors of \( \theta^0\theta^0 \), with corresponding eigenvalues in \( D \).

**Step 4: Prove the convergence.**

We bound \( \| F - F^0\tilde{H} \|_F \) conditional on the event \( \hat{\theta} = \rho^0 \). By the Davis–Kahan sin(\( \theta \)) theorem (see, e.g., Yu et al. (2014)) and (A.8),

\[
\frac{1}{\sqrt{T}}\| F - F^0\tilde{H} \|_F \leq \frac{1}{\sqrt{T}}\| \hat{\theta} - \hat{\theta}' \|_p \| \theta^0\theta^0 \|_p \leq \tilde{c}(\gamma_1\sqrt{K_0} \vee \gamma_2).
\]

Next we have

\[
\| F - F^0\tilde{H} \|_F = \| F^0\tilde{H} - F^0\tilde{H} \|_F \leq 2\tilde{c} \| F^0\tilde{H} - F^0\tilde{H} \|_F + \| F^0\tilde{H}F^0\tilde{H}^T - F^0\tilde{H}F^0\tilde{H} \|_F
\]

\[
\leq \tilde{c}(\gamma_1\sqrt{K_0} \vee \gamma_2),
\]

where the second equality is from the fact that \( \hat{\theta} = \tilde{H} = S^{-1/2}v'vS^{-1/2}v = S^{-1} \). This proves the second result in the theorem.

**A.2. Theoretical analysis of the second-step estimators**

To prove Theorem 3.3, we need to add a further lemma.

**Lemma A.4.** Suppose that Assumptions A.1–A.3 hold. Let \( \tilde{\Sigma} = T^{-1}\tilde{X}'\tilde{X} - T^{-2}\tilde{X}'\tilde{X} \). Then there exist some constants \( \bar{c}, \tilde{c} \) such that with probability larger than \( 1 - \tilde{c}(p^2N^2T^{-1/4}(\log N)^{-q/2} + pN^2\tilde{c}T + p^2N^2\tilde{c}(\log N) \) we have

(i) \( \| H \|_{\text{max}} \leq \| H \|_{\infty} \leq \bar{c} \) and \( \| H^{-1} \|_{\text{F}} \leq \bar{c} \);

(ii) \( \max_{1 \leq i \leq p} |X_{i,i}|/\sqrt{T} < \tilde{c} \) and \( \max_{1 \leq i \leq n} |X_{i,j}|/\sqrt{T} < \tilde{c} \);

(iii) \( \frac{p^0U}{\text{max}} / T \leq \log N \cdot T^{-1/2}/(16\tilde{c}^2) \) and \( \| T^{-1}\tilde{X}F^0 - \Sigma_{\text{X}} \|_{\text{max}} \leq \tilde{c}T^{-1/2}\log N \);

(iv) \( \| \tilde{\Sigma} - \Sigma \|_{\text{max}} \leq \gamma_1 \).

13 Write \( a \approx b \) to denote that both \( a/b \) and \( b/a \) are stochastically bounded.

(1) $T^{-1}\|U\mathbf{x}\|_{\text{max}} \leq \gamma_3/4$; (2) $\max_{1 \leq j \leq n}\|X_{s,j}\|/\sqrt{T} \leq \tilde{c}$;

(3) $\max_{1 \leq j \leq n}|U_{s,j}|/\sqrt{T} \leq \tilde{c}$; (4) $\|\hat{F} - F^0\hat{H}\|_F/\sqrt{T} \leq \gamma_3/(16\tilde{c})$;

(5) $\|F^0\mathbf{u}\|_{\text{max}}/T \leq \gamma_3/(16\tilde{c})$; (6) $\|\hat{H}\|_{\infty} \vee \|\hat{H}^{-1}\|_F \leq \tilde{c}$;

(7) $\hat{R} = R^0$.

and (8) $\hat{\Sigma}$ satisfies the restricted eigenvalue condition for $K_\ell$ in (3.3) with $\kappa_\varepsilon(K_\ell) \geq \psi_{\text{min}}(\Sigma)/2$. Under Assumptions A.1–A.3, by Lemmas A.2 and A.4, $\varepsilon_N^{(3)}$ holds with probability larger than $1 - \varepsilon_{NT} - \tilde{c}[p^2N^2T^{-1/4}(\log N)^{-q/2} + pN\varepsilon - CT + p^2N^{2-\varepsilon}\log N]$. Conditional on the event $\varepsilon_N^{(3)}$, we also have the event that

(9) $T^{-1}\|\hat{F}\|_{\max} \leq T^{-1}\|\hat{F} - F^0\hat{H}\|_F + T^{-1}\|\hat{H}^0\|_F$;

and that

(10) $\max_{1 \leq i \leq n} |\hat{\Delta}_{s,i}| \leq \max_{1 \leq i \leq n} |\Delta_{s,i}| - 1/2$.

Conditioned on the event $\varepsilon_N^{(3)}$, we establish the bound of $|\hat{\Delta}_{s,i}|_1 \equiv |\hat{b}_{s,i} - B_{s,i}^0|$ for $i \in [N]$.

Step 1. Concentrate out $\lambda$.

The objective function (3.2) is a least squares objective function with respect to $\lambda$. Given $\hat{b}_{s,i}$, we have that

$\hat{\lambda}_i = (T\hat{F})^{-1}\hat{F}(Y_{s,i} - \hat{B}_{s,i}) = T^{-1}\hat{F}(Y_{s,i} - \hat{B}_{s,i})$,\n
where the second equality holds by the identity $T\hat{F}/T = I_T$. After concentrating out $\lambda_i$, the optimization problem becomes

$\hat{b}_{s,i} = \arg\min_{v \in \mathbb{R}^p} \frac{1}{2T}\|M_F(Y_{s,i} - \mathbf{v})\|_F^2 + \gamma_3|v|_1$, \hfill (A.9)

where $M_F = I_T - T\hat{F}/T$.

Step 2. Compare the objective functions at $\hat{b}_{s,i}$ and $B_{s,i}^0$.

By the identity $Y_{s,i} = \mathbf{X}B_{s,i}^0 + F^0\lambda_i^0 + U_{s,i}$ and the definition of $\hat{b}_{s,i}$, we have

$0 \geq \frac{1}{2T}[\|M_F(Y_{s,i} - \mathbf{X}\hat{b}_{s,i})\|_F^2 - \|M_F(F^0\lambda_i^0 + U_{s,i})\|_F^2] + \gamma_3(|\hat{b}_{s,i}|_1 - |B_{s,i}^0|_1)$

$= \frac{1}{2T}\|M_F\Delta_{s,i}\|_F^2 - \frac{1}{T}\text{tr}(F^0\lambda_i^0 + U_{s,i})^\prime M_F \Delta_{s,i} + \gamma_3(|\hat{b}_{s,i}|_1 - |B_{s,i}^0|_1)$,

where $\Delta \equiv \hat{B} - B^0$ and $\Delta_{s,i}$ denotes the $i$th column of $\Delta$. Then by Lemma A.1(v), we have

$\frac{1}{T}\|F^0\lambda_i^0 + U_{s,i})^\prime M_F \Delta_{s,i}\|_{\text{max}} \geq \frac{1}{T}\text{tr}(F^0\lambda_i^0 + U_{s,i})^\prime M_F \Delta_{s,i}$

$\geq \frac{1}{2T}\|M_F\Delta_{s,i}\|_F^2 + \gamma_3(|\hat{b}_{s,i}|_1 - |B_{s,i}^0|_1)$

$\geq \frac{1}{2T}\|M_F\Delta_{s,i}\|_F^2 + \gamma_3(|\hat{b}_{s,i}|_1 - |B_{s,i}^0|_1) - \gamma_3|\hat{b}_{s,i}|_1$,

where the last inequality follows because

$|\hat{b}_{s,i}|_1 - |B_{s,i}^0|_1 = |\hat{\Delta}_{s,i} + B_{s,i}^0|_1 - |B_{s,i}^0|_1 = |\hat{\Delta}_{s,i}^0|_1 + |\hat{\Delta}_{s,i}^0 + B_{s,i}^0|_1 - |B_{s,i}^0|_1$

$\geq |\hat{\Delta}_{s,i}^0|_1 - |\hat{\Delta}_{s,i}^0|_1$. \hfill (A.9)

Step 3. Bound $T^{-1}\max_1[\|F^0\lambda_i^0 + U_{s,i})^\prime M_F \Delta_{s,i}\|_{\text{max}}]$, conditional on the event $\varepsilon_N^{(3)}$. \hfill (177)
By the triangle and Cauchy Schwartz inequalities and the fact that $T^{-1/2} \| \hat{F} \|_{op} = 1$, we have
\[
T^{-1} \| (F_0^0 + U_{i,j}) M_{\ell,i} X \|_{max} \\
= T^{-1} \| F_0^0 M_{\ell,i} X \|_{max} + T^{-1} \| U_{i,j} M_{\ell,i} X \|_{max} \\
\leq T^{-1/2} |X_{i,j}| \cdot T^{-1/2} \| \lambda_0^0 F_0^0 M_{\ell,i} \|_{max} + \max_{1 \leq j \leq N_p} T^{-1} \| U_{i,j} X_{i,j} \| + T^{-2} \| U_{i,j} F_0^0 M_{\ell,i} X \|_{max} \\
\leq \max_{1 \leq j \leq N_p} T^{-1} |U_{i,j} X_{i,j}| + \left( T^{-1} |U_{i,j} F_0^0 M_{\ell,i} X \|_{max} \right) \max_{1 \leq j \leq N_p} T^{-1/2} |X_{i,j}|.
\]
Combining events (1), (9) and (10), the right hand side of the above inequality is bounded by $\gamma_3/2$ conditional on the event $\mathcal{E}_N^3$.

**Step 4. Obtain the final bound for $|\hat{B}_{i,j} - B_{i,j}^0|$.**

Combining the results in Steps 2–3 and using the identity $|\hat{\Delta}_{s,i}|_1 = |\hat{\Delta}_{j,i}|_1 + |\hat{\Delta}_{j',i}|_1$, we have that conditional on the event $\mathcal{E}_N^3$,
\[
3\gamma_3 |\hat{\Delta}_{j,i}|_1 \geq \frac{1}{T} \| M_{\ell,i} X \Delta_{s,i} \|_F^2 + \gamma_3 |\hat{\Delta}_{j',i}|_1.
\]
It follows that $|\hat{\Delta}_{j',i}|_1 \leq 3 |\hat{\Delta}_{j,i}|_1$ and conditional on $\mathcal{E}_N^3$,
\[
\hat{\Delta}_{s,i} \hat{\Sigma}_{s,i} \leq 3\gamma_3 |\hat{\Delta}_{j,i}|_1 \leq 3\gamma_3 \sqrt{k_i} |\hat{\Delta}_{j,i}|_1 \leq \frac{6k_i}{\psi_{\min}(\Sigma)} \gamma_3 \sqrt{\hat{\Delta}_{s,i} \hat{\Sigma}_{s,i}},
\]
where the last inequality holds by event (8) in $\mathcal{E}_N^3$. It follows that $\sqrt{\hat{\Delta}_{s,i} \hat{\Sigma}_{s,i}} \leq \frac{6k_i}{\psi_{\min}(\Sigma)} \gamma_3$ and $|\hat{\Delta}_{j,i}|_1 \leq \frac{2\sqrt{k_i}}{\psi_{\min}(\Sigma)} \gamma_3$. Consequently, we have established that
\[
|\hat{\Delta}_{s,i}|_1 = |\hat{\Delta}_{j,i}|_1 + |\hat{\Delta}_{j',i}|_1 \leq 4 |\hat{\Delta}_{j,i}|_1 \leq \frac{48}{(\psi_{\min}(\Sigma))^2} k_i \gamma_3.
\]
Then the conclusion in Theorem 3.3 follows.

**A.3. Theoretical analysis of the three-step estimators**

To prove Theorems 3.4 and 3.5, we need the following lemma.

**Lemma A.5.** Suppose that Assumptions A1–A5 hold. Then uniformly over $i = 1, \ldots, N$, the following results hold w.p.a.1:

(i) $\psi_{\min}(\hat{\Sigma}_{j,i}) \geq \xi$;

(ii) $\| \hat{\Sigma}_{j,i} \|_{max} \leq \tilde{c}$ and $\psi_{\max}(\hat{\Sigma}_{j,i}) \leq \tilde{c} k_i$,

for some finite constant $\tilde{c}$.

**Proof of Theorem 3.4.** For any $n$-dimensional vector $v = (v_1, \ldots, v_n)'$, denote $abs(v) = (|v_1|, \ldots, |v_n|)'$. We say that $v < \tilde{v}$ if and only if $v_i < \tilde{v}_i$ for all $i \in [n]$. Let $W^{(0)} = \text{diag}(w_{11}, \ldots, w_{NN})$, $W^{(1,j)} = W_{j,j}^{(1)}$ and $W^{(0,j)} = W_{j,j}^{(0)}$. The following proof is by induction. Based on the error bounds for $\hat{F}^{(i)}$’s, we show that results (i)–(iii) hold for the $(\ell + 1)$-th step estimators. Then the results follow as we have already established that $\| \hat{F}^{(0)} - F^0 \|_{F}/\sqrt{T} = O_p(\gamma_1 \sqrt{K_d} + \gamma_2)$.

For notational simplicity, let $\hat{\Sigma}^{(i)}$ denote $T^{-1} X M_{\ell,i} X$ for $\ell = 0, 1, 2, \ldots$.

(i) For all $(k, i)$’s such that $B_{0,i}^{(k)} = 0$, $\min_{k \neq j} \frac{1}{k_d} |\hat{B}_{i,k} - B_{j,k}^0| \leq \| \hat{B} - B^0 \|_{max} \leq O_p(K_1 \gamma_3) = O_p(\gamma_4)$. It follows that $W^{(0,j)} = I_{j,j}^f$ with w.p.a.1. For all $(k, i)$’s such that $B_{0,i}^{(k)} \neq 0$,
\[
\min_{k,i} \frac{1}{k_d} \frac{1}{k_d} |\hat{B}_{i,k} - B_{j,k}^0| = \min_{i \in [N]} \frac{1}{k_d} |\hat{B}_{i,k} - B_{j,k}^0| = \min_{i \in [N]} \frac{1}{k_d} |\hat{B}_{i,k} - B_{j,k}^0| = o_p(\gamma_4) \leq \gamma_4 \text{ w.p.a.1}
\]
by Assumption A5(i). It follows that $W^{(1,i)} = 0$ w.p.a.1. For each $i \in [N]$, the estimator $\hat{B}_{s,i}^{(\ell)}$ can be written as
\[
\hat{B}_{s,i}^{(\ell)} = \arg \min_{v \in \mathbb{R}^N} \mathcal{L}^{(i)}(v, \hat{F}^{(\ell-1)}),
\]
where $\mathcal{L}^{(i)}(v, F) \equiv \frac{1}{2T} (Y_s - Xv)' M_{\ell,i} X_{s,i} (Y_s - Xv) + \gamma_4 \sum_{k=1}^{N_p} w_{kl} |v_k|$. Following the proof of Proposition 1 of Zhao and Yu (2006), $\text{sgn}(\hat{B}_{s,i}^{(0)}) = \text{sgn}(B_{0,i})$ is implied by event $\mathcal{E}_{i,1} \cap \mathcal{E}_{i,2}$, where
\[
\mathcal{E}_{i,1} \equiv \left\{ \begin{array}{l}
\left[ T^{-1/2} \hat{\Sigma}_{j,i}^{-1} X_{s,i} M_{\ell,i}^{-1} (U_{s,i} + F^0 \hat{\lambda}_i^{(0)}) \right] < T^{1/2} \text{abs}(B_{0,i}) - T^{1/2} \gamma_4 \text{abs} \left[ \hat{\Sigma}_{j,i}^{-1} W^{(1,i)} \text{sgn}(B_{0,i}) \right] \end{array} \right\}
\]
178
and

\[ e_{i,2} \equiv \| \text{abs}[T^{-1/2}(-\tilde{\Sigma}_{i,h}^{-1} \cdot X_{i,h}^{*} + X_{i,h}^{*} \cdot \tilde{M}_{\tilde{F}(-1)}(U_{i,i} + F^0_{\lambda_1}^0)] \| \leq T^{-1/2} \gamma_4 W(0,0), \quad \{ e_{i,1} \} - T^{-1/2} \gamma_4 \text{abs}[\tilde{\Sigma}_{i,h}^{-1} W^{(1,0)} \text{sgn}(B_{0,1})] \}. \]

We prove (i) by showing that \( e_{i,1} \) and \( e_{i,2} \) hold w.p.a.1.

First, we consider \( e_{i,1} \). It suffices to show that each entry of \( T^{-1/2} \text{abs}[\tilde{\Sigma}_{i,h}^{-1} X_{i,h}^{*} \cdot \tilde{M}_{\tilde{F}(-1)}(U_{i,i} + F^0_{\lambda_1}^0)] \) is \( o_p \left( \sqrt{T} \min \{ \min_{k \in \{1\}} |B_{0,1}\} \right) \). Applying the triangle inequality, one has

\[
T^{-1/2} \text{abs}[\tilde{\Sigma}_{i,h}^{-1} X_{i,h}^{*} \cdot \tilde{M}_{\tilde{F}(-1)}(U_{i,i} + F^0_{\lambda_1}^0)] \\
\leq T^{-1/2} \text{abs}[\tilde{\Sigma}_{i,h}^{-1} X_{i,h}^{*} \cdot \tilde{M}_{\tilde{F}(-1)} U_{i,i}] + T^{-1/2} \text{abs}[\tilde{\Sigma}_{i,h}^{-1} X_{i,h}^{*} \cdot \tilde{M}_{\tilde{F}(-1)} F^0_{\lambda_1}^0] \\
\leq T^{-1/2} \text{abs}[\tilde{\Sigma}_{i,h}^{-1} X_{i,h}^{*} \cdot \tilde{M}_{\tilde{F}(-1)} \hat{H} - F^0_{\lambda_1}^0] \\
+ T^{-1/2} \text{abs}[\tilde{\Sigma}_{i,h}^{-1} X_{i,h}^{*} \cdot \tilde{M}_{\tilde{F}(-1)} (\hat{F}(-1) - \tilde{H}) - F^0_{\lambda_1}^0].
\]

(A.10)

Note that max \( \| \tilde{\Sigma}_{i,h}^{-1} \| \| \| \) \( \leq \tilde{c} \) w.p.a.1 by Lemma A.5(i). This, in conjunction with Assumption A.2(ii)–(ii), implies that the first term on the right hand side of (RHS) of (A.10) is uniformly \( O_p(\log N) \). With \( \| \hat{F}(-1) - F^0 H \| / \sqrt{T} = O_p(\sqrt{\log N} N^{-1/2}) \), we have \( \| F_{\lambda_1}^0 - \hat{F}(-1) \| \| \| \leq O_p(\log N) N^{-1/2} \). Note that Lemma A.4(ii) ensures max \( \| X_{i,i} \| / \sqrt{T} \) and max \( \| X_{i,i} \| / \sqrt{T} \) are both bounded by an absolute constant. It follows that each entry of the second term on the RHS of (A.10) is uniformly \( O_p(\log N \cdot \sqrt{\log N}) \). Similarly, each entry of the third term on the RHS is \( O_p(\log N \cdot \sqrt{\log N} + \sqrt{T\log N}) \). This results, along with the fact that \( \log N \cdot T^{-1/2} \sqrt{\log N} = o(\min(\min_{k \in \{1\}} |B_{0,1}\|, N^{-1/2}) \) and \( N^{-1/2} = o(\min(\min_{k \in \{1\}} |B_{0,1}\|, N^{-1/2}) \), we have \( P(e_{i,1}) \rightarrow 1 \), as \( (N, T) \rightarrow \infty \).

Next, we consider \( e_{i,2} \). Similar to the analysis for \( e_{i,1} \), we can use Lemma A.5(ii) to show that each entry of \( T^{-1/2} \text{abs}[\tilde{\Sigma}_{i,h}^{-1} X_{i,h}^{*} \cdot \tilde{M}_{\tilde{F}(-1)} (U_{i,i} + F^0_{\lambda_1}^0)] \) is \( O_p(\lambda_1^{(1)} \log N \cdot \sqrt{\log N}) \). By the fact that \( \gamma_3 = o(\gamma_4) \), we have \( P(e_{i,2}) \rightarrow 1 \), as \( (N, T) \rightarrow \infty \).

(ii) Conditional on the event \( \{ \hat{B}(-1) = B^0 \} \), we can follow the proof of Lemma 1 in Zhao and Yu (2006) to establish the first order condition that

\[
\tilde{\Sigma}_{i,h} = \hat{B}(-1) - B^0_{i,i},
\]

for \( i \in [N] \). Then

\[
\left\| \tilde{\Sigma}_{i,h} \right\| \leq \tilde{c} \text{ w.p.a.1 by Lemma A.5(i). Note that uniformly in } i \in [N],
\]

\[
A_1^2 = \frac{1}{T^2} \left\| \tilde{\Sigma}_{i,h} \cdot \tilde{M}_{\tilde{F}(-1)}(\hat{F}(-1) \tilde{H}^{-1} - F^0_{\lambda_1}^0) \right\| \leq \frac{1}{T^2} \text{tr} \left( \left\| F^0_{\lambda_1}^0 \cdot \tilde{M}_{\tilde{F}(-1)} \hat{H}^{-1} - F^0_{\lambda_1}^0 \right\|^2 \right)
\]

\[
\leq \psi_{\max} \left( \frac{1}{T} \tilde{M}_{\tilde{F}(-1)} X_{i,h} X_{i,h}^{*} \tilde{M}_{\tilde{F}(-1)} \right) \left\| \hat{F}(-1) \tilde{H}^{-1} - F^0_{\lambda_1}^0 \right\|^2 \right\|^2 + \psi_{\max} \left( \frac{1}{T} X_{i,h}^* X_{i,h} \tilde{M}_{\tilde{F}(-1)} \hat{H}^{-1} - F^0_{\lambda_1}^0 \right) \left\| \hat{F}(-1) \tilde{H}^{-1} - F^0_{\lambda_1}^0 \right\|^2
\]

\[ \leq \tilde{c} \tilde{k} \frac{1}{T} \left\| \hat{F}(-1) \tilde{H}^{-1} - F^0_{\lambda_1}^0 \right\|^2 \leq k_1 \cdot o_p(\gamma_1 + \gamma_2^2), \]

and

\[
A_2^2 = \left\| \frac{1}{T} X_{i,h}^* \tilde{M}_{\tilde{F}(-1)} U_{i,i} \right\|^2 \leq 2 \left( \frac{1}{T} X_{i,h}^* U_{i,i} \right)^2 + 2 \left( \frac{1}{T} X_{i,h}^* \hat{F}(-1) \right) \left\| \hat{F}(-1) \right\|^2.
\]

This claim holds for \( \epsilon = 1 \) by Theorem 3.2. Given this claim, we can show that \( \| \hat{F}(-1) - F^0 H \| / \sqrt{T} = O_p(\log N) N^{-1/2} \sqrt{\log N} + N^{-1/2} \) for each \( \ell \) using the results below.
It is standard to show that \( \| \frac{1}{T} X_{s,j} U_{s,i} \| \leq k_i^{1/2} \sigma_T (T^{-1/2} \log N) \) uniformly in \( i \). In addition,

\[
\| \frac{1}{T} X_{s,j} \hat{\epsilon}^{(e-1)} \| \leq \psi \max \left( \left\| \frac{1}{T} \hat{\epsilon}^{(e-1)} X_{s,j} X_{s,j} \right\| \right) \frac{1}{T} \| \hat{\epsilon}^{(e-1)} U_{s,i} \| \leq \psi \max \left( \left\| \frac{1}{T} \hat{\epsilon}^{(e-1)} U_{s,i} \| \right\| \right) = k_i \cdot \sigma_T (\gamma_1 + \gamma_2)^2 \text{ uniformly in } i,
\]

where the last equality follows from the fact \( \psi \max (\frac{1}{T} X_{s,j} X_{s,j}) \leq \bar{c} \text{ w.p.a.1} \) and \( \max_i \left\| \frac{1}{T} \hat{\epsilon}^{(e-1)} U_{s,i} \right\| = \sigma_T (\gamma_1 + \gamma_2) \) by similar arguments as used to obtain event (9) in the proof of Theorem 3.3. Then uniformly in \( i \in [N] \), we have

\[
\| \hat{B}^{(e)}_{s,j} - B^0_{s,j} \|^2 \leq k_i \cdot \sigma_T (\gamma_1 + \gamma_2)^2.
\]

It follows that

\[
\| X(\hat{B}^{(e)} - B^0) \|^2 \leq \frac{1}{N} \sum_{i=1}^N \left( \| X(\hat{B}^{(e)}_{s,j} - B^0_{s,j}) \|^2 \right) = \frac{1}{N} \sum_{i=1}^N \frac{1}{T} \| (\hat{B}^{(e)}_{s,j} - B^0_{s,j}) X_{s,j} X_{s,j} (\hat{B}^{(e)}_{s,j} - B^0_{s,j}) \|_{op}
\leq \frac{1}{N} \sum_{i=1}^N \left( \| \hat{B}^{(e)}_{s,j} - B^0_{s,j} \| \right) \| X_{s,j} X_{s,j} \|_{op}
= \frac{1}{N} \sum_{i=1}^N k_i \cdot \sigma_T (\gamma_1 + \gamma_2)^2.
\]

Then the result in (ii) follows under Assumption A.5(iii).

(iii) Note that \( Y - X\hat{B}^{(e)} - F^0 \Lambda^0 = U - X(\hat{B}^{(e)} - B^0) \). By the result in (ii) and Assumption A.3(i),

\[
\frac{1}{\sqrt{NT}} \left\| U - X(\hat{B}^{(e)} - B^0) \right\|_{op} \leq \frac{1}{\sqrt{NT}} \left\| U \right\|_{op} + \frac{1}{\sqrt{NT}} \left\| X(\hat{B}^{(e)} - B^0) \right\|_{op} \leq \sigma_T (\gamma_2) + \sigma_T (\gamma_1 + \gamma_2) = \sigma_T (\gamma_1 + \gamma_2).
\]

One can apply analyses similar to proof of Theorem 3.2 to obtain the desired result. □

Proof of Theorem 3.5. Let \( \hat{\Sigma} = X M_{ij} X / T \). From the proof of Theorem 3.4, we have

\[
\hat{\Sigma}_{s,j}(\hat{B}_{s,j} - B^0_{s,j}) = \frac{1}{T} X_{s,j} M_{ij} F^0 \Lambda^0 + \frac{1}{T} X_{s,j} M_{ij} U_{s,i} - \gamma_4 W^{(1,0)} \text{sgn}(B^0_{s,j}).
\]

By Theorem 3.3 and Assumption A.5(i), \( \max_w w_{kj} = 0 \) w.p.a.1, which implies that \( \gamma_4 W^{(1,0)} \text{sgn}(B^0_{s,j}) = O_T (T^{-1/2}) \).

Noting that the columns of \( \hat{F}/\sqrt{T} \) are the first \( \hat{K} \) eigenvectors of \( \frac{1}{NT}(Y - X\hat{B}^{(e-1)})(Y - X\hat{B}^{(e-1)})' \), we have

\[
\hat{F}_{VNT} = \frac{1}{NT} \left( Y - X\hat{B}^{(e-1)} \right) \left( Y - X\hat{B}^{(e-1)} \right)' \hat{F}
= \frac{1}{NT} \sum_{i=1}^N \left( Y_{s,i} - X_{s,j} \hat{B}_{s,j}^{(e-1)} \right) \left( Y_{s,i} - X_{s,j} \hat{B}_{s,j}^{(e-1)} \right)' \hat{F}.
\]

where \( V_{NT} \) is a diagonal matrix that consists of the \( \hat{K} \) largest eigenvalues of the matrix \( T \times T \) matrix \( (NT)^{-1}(Y - X\hat{B}^{(e-1)})(Y - X\hat{B}^{(e-1)})' \), arranged in descending order along its diagonal line. One can use a similar expansion as in the Proof of Proposition S.1.1, to study \( M_{ij} \). The estimation error of \( \hat{B}^{(e)} \) depends on \( \ell \) and the error of \( B^{(1)} \), but part that is due to \( \hat{B}^{(e)} \) will decay fast. After finite steps, we have that

\[
\frac{1}{T} X_{s,j} M_{ij} \sigma_{s,j} \hat{\Sigma}_{s,j} \hat{\Sigma}_{s,j} = \frac{1}{T} S_{s,j} \sigma_{s,j} \sigma_{s,j} + O_T (K^1 T^{-1/2} \log N).
\]

By arguments as used in the proof of Lemma A.2, we can readily show that \( \frac{1}{\sqrt{T}} X_{s,j} M_{ij} \sigma_{s,j} \hat{\Sigma}_{s,j} = \frac{1}{\sqrt{T}} \sigma_{s,j} \sigma_{s,j} + O_T (K^1 T^{-1/2} \log N) \), and \( \frac{1}{\sqrt{T}} X_{s,j} \sigma_{s,j} \sigma_{s,j} = O_T (K^1 T^{-1/2} \log N) \), where \( [\Sigma_{s,j}]_{s,s} = \frac{1}{T} F^{(0)} X_{s,j} \) is a \( \sigma_{s,j} \times k_i \) matrix.
Appendix B. Nagaev inequality for time series

In various places, we need to a sharp probability bound for partial sums like $T^{-1} \sum_{t=1}^{T} X_t$, which is nonlinear in shocks $\{u_t\}$ and non-Gaussian. Wu (2005) provides a simple functional measure to quantify the degree of dependence in nonlinear systems. With the dependence measure, Wu and Wu (2016) establish a Nagaev-type inequality for nonlinear processes, under mild conditions.

In Theorem B.1, we aim to bound a partial sum of the form $S_n = \sum_{i=1}^{n} a_i e_i$, where the $a_i \in \mathbb{R}$ are nonrandom, the scalar process $\{e_i\}$ has the form $e_i = g(\ldots, e_{i-1}, e_i)$, where the $e_i$ are independently and identically distributed (i.i.d.) random variables, and $g(\cdot)$ is a measurable function. Letting $F_i \equiv (\ldots, e_{i-1}, e_i)$, we write $e_i = g(F_i)$. Then a coupled process $e^* = g(F^*_i)$. The measure $\delta_i(q)(\cdot)$ reflects the effect of shock $e_0$ on $e_i$. Accordingly, we assume the cumulative effect of $e_0$ on $\{e_i\}_{i=m}$ to be summable and given by

$$\Delta_{m,q}(e) \equiv \sum_{i=m}^{\infty} \delta_i(q)(\cdot) < \infty.$$

We can then define the dependence-adjusted norm (DAN):

$$\|e\|_{q,a} \equiv \sup_{m \geq 0} (m + 1)^{\alpha} \Delta_{m,q}(e).$$

With these definitions we present the following Nagaev inequality for time series as a simplified version of Theorem 2 of Wu and Wu (2016).

**Theorem B.1.** Let $a = (a_1, \ldots, a_n)^T$ and $|a|_q = (\sum_{i=1}^{n} |a_i|^q)^{1/q}$. Suppose that $\sum_{i=1}^{n} a_i^2 = n$, $E(e_i) = 0$, and $\|e\|_{q,a} < \infty$ for some $q > 2$ and $\alpha > 1$. Then for all $x > 0$,

$$P(\{|S_n| > x\}) \leq C_1 \frac{|a|_q \|e\|_{q,a}^q}{x^q} + C_2 \exp\left(-\frac{C_3 x^2}{n \|e\|_{L_q}^2}\right),$$

where $C_1, C_2, C_3$ are constants that only depend on $q$ and $\alpha$. The above lemma is used repeatedly in proving some technical lemmas that are needed in the proof of our main results.

Appendix C. Online supplement for “high dimensional var with common factors”

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2022.02.002.

References


Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. Econometrica 70, 191–221.


Lewin, F., 2016. Shrinkage estimation of dynamic panel data models with interactive fixed effects. J. Econometrics 190, 148–175.


