HIGH-DIMENSIONAL VARS WITH COMMON FACTORS

By

Ke Miao, Peter C.B. Phillips, and Liangjun Su

COWLES FOUNDATION PAPER NO. 1837

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

2023

http://cowles.yale.edu/

Journal of Econometrics 233 (2023) 155–183

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

High-dimensional VARs with common factors[☆]

Ke Miao [a,b], Peter C.B. Phillips [c,d,e,f], Liangjun Su [g,*]

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

ARTICLE INFO

Article history:
Received 7 July 2020
Received in revised form 11 February 2022
Accepted 15 February 2022
Available online 9 March 2022

JEL classification:
C13
C33
C38
C51

Keywords:
Common factors
Connectedness
Cross-sectional dependence
Investor fear
High-dimensional VAR
Nuclear-norm regularization

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 ℓ_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.

1. Introduction

In a pathbreaking study, Mann and Wald (1943) introduced vector autoregressions (VARs) and developed the limit theory for estimation and inference.[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

[☆] 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.

[*] Correspondence to: School of Economics and Management, Tsinghua University, Beijing, 100084, China.
E-mail address: sulj@sem.tsinghua.edu.cn (L. Su).

[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).

https://doi.org/10.1016/j.jeconom.2022.02.002
0304-4076/© 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), Candes 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 the use of 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 ℓ_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 ℓ_1-nuclear-norm regularized least squares estimation problem that minimizes the sum of squared residuals with an ℓ_1-norm penalty imposed on the transition matrices and a nuclear norm penalty on the low rank matrix Θ representing the common component. Imposing the ℓ_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), Koltchinskii 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 Θ. Applying a singular value thresholding (SVT) procedure on the singular values of the estimate of Θ, 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 $v = (v_1, \dots, v_N)' \in \mathbb{R}^N$ be a matrix and vector. Denote v_I as the subvector of v whose entries are indexed by a set $I \subset [N] \equiv \{1, \dots, N\}$ and denote $A_{I,J}$ as the submatrix of A whose rows and columns are indexed by I and J, respectively. Let $A_{*,J} \equiv A_{[M],J}$ be the submatrix of A whose columns are indexed by J, $A_{I,*} \equiv A_{I,[N]}$ 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 $A_{*,j} \equiv A_{*,\{j\}}$ for $j \in [N]$ and $A_{i,*} \equiv A_{\{i\},*}$ for $i \in [M]$.

Define the ℓ_0, ℓ_q ($q \geq 1$), and ℓ_∞ norms of a vector $v \in \mathbb{R}^N$ as follows

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

where $\mathbf{1}(\cdot)$ is the indicator function. In the special case $q=2$, $|\cdot|_2$ denotes the Euclidean norm of v and can be rewritten as $|v|$ for notational simplicity.

For $1 \le q < \infty$, define the ℓ_q, $\ell_{\max}$, Frobenius (F), and nuclear ($_*$) norms of the matrix A as follows

$$\|A\|_q \equiv \max_{|v|_q=1} \|Av\|_q,\ \ \|A\|_{\max} \equiv \max_{i,j} |a_{ij}|,\ \ \|A\|_{\mathrm{F}} \equiv \left(\sum_{i,j} |a_{ij}|^2\right)^{1/2} \text{ and } \|A\|_* \equiv \sum_{k=1}^{\min(N,M)} \psi_k(A),$$

where $\psi_k(A)$ denotes the kth 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 ℓ_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 ℓ_p-norm as $E(x)$ and $\|x\|_p \equiv [E(|x|_p^p)]^{1/p}$.

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

2. 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

$$Y_t = \sum_{j=1}^{p} A_j^0 Y_{t-j} + \Lambda^0 f_t^0 + u_t, \quad t=1,\ldots,T, \tag{2.1}$$

where $A_1^0,\ldots,A_p^0$ are $N \times N$ transition matrices, $\Lambda^0 = (\lambda_1^0,\ldots,\lambda_N^0)'$ is an $N \times R^0$ factor loading matrix, f_t^0 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_j^0's, Λ^0, and $F^0 \equiv (f_1^0,\ldots,f_T^0)'$. 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

$$\underbrace{\begin{bmatrix} Y_1' \\ \vdots \\ Y_T' \end{bmatrix}}_{\mathbf{Y}} = \underbrace{\begin{bmatrix} Y_0' & \cdots & Y_{1-p}' \\ \vdots & \ddots & \vdots \\ Y_{T-1}' & \cdots & Y_{T-p}' \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} A_1^{0\prime} \\ \vdots \\ A_p^{0\prime} \end{bmatrix}}_{B^0} + \underbrace{\begin{bmatrix} f_1^{0\prime} \\ \vdots \\ f_T^{0\prime} \end{bmatrix}}_{F^0} \underbrace{\begin{bmatrix} \lambda_1^{0\prime} \\ \vdots \\ \lambda_N^{0\prime} \end{bmatrix}'}_{\Lambda^{0\prime}} + \underbrace{\begin{bmatrix} u_1' \\ \vdots \\ u_T' \end{bmatrix}}_{\mathbf{U}}, \tag{2.2}$$

where $\mathbf{Y} \in \mathbb{R}^{T\times N}$, $\mathbf{X} \in \mathbb{R}^{T\times Np}$, $B^0 \in \mathbb{R}^{Np\times N}$, and $\mathbf{U} \in \mathbb{R}^{T\times N}$. Let $\Theta^0 \equiv F^0\Lambda^{0\prime}$ denote the common component. A key observation here is that Θ^0 is a low rank matrix. However, due to the presence of $\mathbf{X}B^0$, the direct use of principal component analysis (PCA) on $\mathbf{Y}$ cannot deliver a consistent estimate of the common factors. Note that under some regularity conditions, $\left\|\Theta^0\right\|_{\text{op}} = O_P(\sqrt{NT})$ and $\|\mathbf{U}\|_{\text{op}} = O_P(\sqrt{N}+\sqrt{T})$.[4] For the pure factor model as in Bai (2003), the separation of Θ^0 from $\mathbf{U}$ hinges on this order difference. The exact probability order of $\left\|\mathbf{X}B^0\right\|_{\text{op}}$ depends on the underlying data generating process but is in general not of smaller order than $O_P(\sqrt{NT})$,[5] which makes it difficult to separate the low rank matrix Θ^0 from $\mathbf{Y}$ 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 Θ^0 are the eigenvalues of $(F^{0\prime}F^0\Lambda^{0\prime}\Lambda^0)^{1/2}$. Assuming $F^{0\prime}F^0/T \xrightarrow{p} \Sigma_F$ and $\Lambda^{0\prime}\Lambda^0/N \xrightarrow{p} \Sigma_\Lambda$ with Σ_F and Σ_Λ 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 $O_P(\sqrt{N}+\sqrt{T})$.

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

$$\|\mathbf{X}B^0\|_{\text{op}} = \left[\lambda_{\max}\left(\mathbf{X}B^0B^{0\prime}\mathbf{X}'\right)\right]^{1/2} \ge c\left[\lambda_{\max}\left(\mathbf{X}\mathbf{X}'\right)\right]^{1/2} = c\,\|\mathbf{X}\|_{\text{op}}\,.$$

By definition of the operator norm, $\|\mathbf{X}\|_{\text{op}} \ge (NT)^{-1/2}\iota_T'\mathbf{X}\iota_N = (NT)^{-1/2}\sum_{i,t} y_{i,t-1} \asymp \sqrt{NT}$ provided $\frac{1}{NT}\sum_{i,t} y_{i,t-1} \xrightarrow{p} c_y \neq 0$ which may occur, say, when the common component $\lambda_i^{0\prime}f_t^0$ does not have mean zero. Here $a_{NT} \asymp b_{NT}$ denotes that a_{NT} and b_{NT} are of the same probability order. Then $\|\mathbf{X}B^0\|_{\text{op}}$ is at least of probability order $\sqrt{NT}$.

2.1. Stationarity analysis

Let $X_t \equiv \mathbf{X}'_{t,*}$. 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.,

$$\underbrace{\begin{bmatrix} Y_t \\ Y_{t-1} \\ \vdots \\ Y_{t-p+1} \end{bmatrix}}_{X_{t+1}} = \underbrace{\begin{bmatrix} A_1^0 & A_2^0 & \cdots & A_{p-1}^0 & A_p^0 \\ I_N & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_N & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & I_N & \mathbf{0} \end{bmatrix}}_{\Phi} \underbrace{\begin{bmatrix} Y_{t-1} \\ Y_{t-2} \\ \vdots \\ Y_{t-p} \end{bmatrix}}_{X_t} + \underbrace{\begin{bmatrix} \Lambda^0 f_t^0 \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}}_{\mathcal{F}_t} + \underbrace{\begin{bmatrix} u_t \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}}_{\mathcal{U}_t}. \tag{2.3}$$

If one treats $\mathcal{F}_t + \mathcal{U}_t$ 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

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

In the low-dimensional framework, the process is stationary if $\mathcal{A}(z)$ has no roots in and on the complex unit circle, or equivalently the largest modulus of the eigenvalues of Φ is less than 1. To achieve identification, we need to study the Gram or signal matrix $S_X \equiv \mathbf{X}'\mathbf{X}/T$ and its population version $\Sigma_X \equiv E(X_t X_t')$. Basu and Michailidis (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(∞)) as

$$X_{t+1} = \sum_{j=0}^{\infty} \Phi^j (\mathcal{F}_{t-j} + \mathcal{U}_{t-j}) \equiv X_{t+1}^{(f)} + X_{t+1}^{(u)}, \tag{2.4}$$

where $X_{t+1}^{(f)} \equiv \sum_{j=0}^{\infty} \Phi^j \mathcal{F}_{t-j}$ and $X_{t+1}^{(u)} \equiv \sum_{j=0}^{\infty} \Phi^j \mathcal{U}_{t-j}$. 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 $\mathcal{F}_t$ is a high-dimensional matrix with rank R^0 and explosive nonzero eigenvalues. Even if the largest modulus of the eigenvalues of Φ 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 ith entry of $X_{t+1}^{(f)}$. Let $e_{j,M}$ be the jth column of I_M. Noting that $y_{it}^{(f)} = (e_{1,p} \otimes e_{i,N})' X_{t+1}^{(f)}$, we can write $y_{it}^{(f)}$ as the MA(∞) process

$$y_{it}^{(f)} = \sum_{j=0}^{\infty} (e_{1,p} \otimes e_{i,N})' \Phi^j (e_{1,p} \otimes \Lambda^0) f_{t-j}^0 \equiv \sum_{j=0}^{\infty} \alpha_{iN}^{(f)}(j) f_{t-j}^0,$$

in which the f_t^0 are allowed to be serially correlated. To ensure $y_{it}^{(f)} = O_P(1)$, the coefficients $\alpha_{iN}^{(f)}(j)$ need to be well-behaved. Note that we generally do not have $\|\Phi\|_{\text{op}} \leq 1$, 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

$$y_{it}^{(u)} = \sum_{j=0}^{\infty} \alpha_{iN}^{(u)}(j) u_{t-j} \text{ and } \alpha_{iN}^{(u)}(j) \equiv (e_{1,p} \otimes e_{i,N})' \Phi^j (e_{1,p} \otimes I_N).$$

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 $\underline{c}$ and $\bar{c}$ denote generic positive constants that may vary across their occurrences. Throughout the paper, we will treat Λ^0 as nonrandom. To ensure the stationarity of $\{Y_t\}$, we impose the following assumption.

Assumption A.1. (i) $u_t = C^{(u)} \epsilon_t^{(u)}$, where $\epsilon_t^{(u)} = (\epsilon_{1,t}^{(u)}, \ldots, \epsilon_{m,t}^{(u)})'$, the $\epsilon_{i,t}^{(u)}$ 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 $C^{(u)} C^{(u)\prime} = \Sigma_u$ and $\underline{c} \leq \psi_{\min}(\Sigma_u) \leq \psi_{\max}(\Sigma_u) \leq \bar{c}$;

(ii) $\left\{f_t^0\right\}$ follows a strictly stationary linear process given by

$$f_t^0 - \mu_f = \sum_{j=0}^{\infty} C_j^{(f)} \epsilon_{t-j}^{(f)},$$

where $\epsilon_t^{(f)} \equiv (\epsilon_{1,t}^{(f)}, \ldots, \epsilon_{R^0,t}^{(f)})'$ are i.i.d. with mean 0 and covariance matrix I_{R^0} across t, $\sup_{m\geq 1}(m+1)^{\alpha} \sum_{j=m}^{\infty} \|C_j^{(f)}\|_{\max} \leq \bar{c} < \infty$ for some constant $\alpha > 1$;

(iii) $\max_{1\leq r\leq R^0} |||\epsilon_{r,t}^{(f)}|||_q < \bar{c}$ and $\max_{1\leq i\leq m} |||\epsilon_{i,t}^{(u)}|||_q < \bar{c}$ for some $q > 4$;

(iv) $\{\epsilon_t^{(u)}\}$ is independent of $\{\epsilon_t^{(f)}\}$;

(v) the largest modulus of the eigenvalues of Φ is bounded uniformly in (N, p) by some constant $\rho \in (0, 1)$;

(vi) $\sup_{N,p} \|(\Phi^j)_{[N],[N]}\|_{\text{op}} \leq \bar{c}\rho^j$ and $\sup_{N,p} |\alpha_{iN}^{(f)}(j)| < \bar{c}\rho^j$;

(vii) $\sup_{N,p} \max_{|z|=1} \psi_{\max}(\mathcal{A}^*(z)\mathcal{A}(z)) \leq \bar{c}$, where $|z|$ denotes the modulus of z in the complex plane, and $\mathcal{A}^*(z)$ denotes the conjugate transpose of $\mathcal{A}(z)$.

Assumption A.1(i) 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 $\psi_{\max}(\Sigma_u)$ 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 $\epsilon_{i,t}^{(u)}$ and $\epsilon_{i,t}^{(f)}$ have finite qth 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 $\{\epsilon_t^{(u)}\}$ and $\{\epsilon_t^{(f)}\}$, which facilitates separate study of $y_{it}^{(f)}$ and $y_{it}^{(u)}$.[6] Assumption A.1(v) is a standard assumption to ensure stationarity. Assumption A.1(vi) is a high level condition to ensure that $E(y_{it}^2)$ is uniformly bounded. Assumption A.1(vii) helps to bound the minimum eigenvalue of Σ_X. From the inequalities

$$\max_{|z|=1} \psi_{\max}(\mathcal{A}^*(z)\mathcal{A}(z)) \leq (\max_{|z|=1} \|\mathcal{A}(z)\|_{\text{op}})^2 \leq 1 + \sum_{k=1}^{p} \|A_j^0\|_{\text{op}},$$

it is evident that requiring all the A_j^0's to have finite operator norms is a sufficient condition for (vii).

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

Proposition 2.1. *Suppose that Assumption A.1 holds. (i) Then Y_t is a stationary process, $\sup_i E(y_{it}^2) < \infty$, and*

$$\psi_{\min}(\Sigma_X) \geq \frac{\psi_{\min}(\Sigma_u)}{\max_{|z|=1} \psi_{\max}(\mathcal{A}^*(z)\mathcal{A}(z))}.$$

(ii) Let $\Sigma_{XF} \equiv E(X_t f_t^{0\prime})$, and $\Sigma \equiv \Sigma_X - \Sigma_{XF}\Sigma_F^{-1}\Sigma_{XF}'$. We also have $\psi_{\min}(\Sigma) \geq \frac{\psi_{\min}(\Sigma_u)}{\max_{|z|=1} \psi_{\max}(\mathcal{A}^*(z)\mathcal{A}(z))}$.

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 Σ_X and Σ : they are bounded away from 0 under Assumption A.1(i) and (vii), but the largest eigenvalues of Σ_X and Σ 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^0 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 Θ^0, the transition matrix B^0 and the factor matrix F^0. This step estimates the low rank matrix Θ^0 together with the transition matrix B^0 via an ℓ_1-nuclear-norm regularization procedure. The ℓ_1-penalty is imposed on the transition matrix B to encourage sparsity and the nuclear-norm penalty is imposed on the common component matrix Θ 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 ℓ_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 ℓ_1-norm and nuclear-norm are convex in their respective matrices. We will show that the resulting estimators $\tilde{B}$ and $\tilde{\Theta}$ are consistent for B^0 and Θ^0, respectively, up to certain scale in terms of the Frobenius norm. Given the preliminary estimate $\tilde{\Theta}$, 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 $\{\epsilon_t^{(u)}\}$ and $\{\epsilon_t^{(f)}\}$ serves as a part of the identification conditions.

thresholding (SVT) procedure and to obtain a consistent estimator $\tilde{F}$ of F^0 in terms of the Frobenius norm up to a certain rotation matrix via singular value decomposition (SVD). Nevertheless, in this step we are unable to establish pointwise consistency for the elements of $\tilde{B}$, $\tilde{\Theta}$ and $\tilde{F}$.

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 ℓ_1-regularized time series regression of $\mathbf{Y}_{*,i}$ on $(\mathbf{X},\tilde{F})$ to obtain an updated estimate $\dot{B}_{*,i}$ of the ith column $(B^0_{*,i})$ of the transition matrix B^0 along with estimates of the factor loadings λ_i^0 for $i = 1, \ldots, N$. Here, the plain ℓ_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 $\dot{B}$, 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 ℓ_1-nuclear-norm regularized estimator to estimate the coefficient matrix B^0 and the low rank matrix Θ^0 simultaneously. We impose a sparsity condition on B^0 and use ℓ_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 Θ^0. The first step estimator is given by the following procedure.

First-step estimator: *Let $\gamma_1 = \gamma_1(N,T) = c_1 T^{-1/2}\log N$ 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 ℓ_1- nuclear-norm regularized regression:*

$$(\tilde{B}, \tilde{\Theta}) = \underset{(B,\Theta)}{\operatorname{argmin}}\, \mathcal{L}(B,\Theta), \text{ where}$$
$$\mathcal{L}(B,\Theta) \equiv \frac{1}{2NT}\|\mathbf{Y} - \mathbf{X}B - \Theta\|_{\mathrm{F}}^2 + \frac{\gamma_1}{N}|\mathrm{vec}(B)|_1 + \frac{\gamma_2}{\sqrt{NT}}\|\Theta\|_*. \tag{3.1}$$

2. *Estimate the number of factors R^0 by singular value thresholding (SVT) as follows:*

$$\hat{R} = \sum_{i=1}^{N\wedge T} \mathbf{1}\{\psi_i(\tilde{\Theta}) \geq (\gamma_2\sqrt{NT}\|\tilde{\Theta}\|_{\mathrm{op}})^{1/2}\},$$

   where $\psi_i(\tilde{\Theta})$ are the singular values of $\tilde{\Theta}$.

3. *Obtain a preliminary estimate of F^0. Let the singular value decomposition (SVD) of $\tilde{\Theta}$ be $\tilde{\Theta} = \tilde{U}\tilde{D}\tilde{V}'$, where $\tilde{D} = \mathrm{diag}(\psi_1(\tilde{\Theta}), \ldots, \psi_{N\wedge T}(\tilde{\Theta}))$. Set $\tilde{F} = \sqrt{T}\tilde{U}_{*,[\hat{R}]}$.*

Remark 3.1. The objective function $\mathcal{L}(B,\Theta)$ is the sum of squared residuals with both the nuclear-norm regularization on Θ and ℓ_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 Θ. The solution can be obtained following the result of Lemma 1 of MW (2019).[7] In the M-step, we fix Θ 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 $\tilde{B}$ and $\tilde{\Theta}$, we establish a non-asymptotic inequality for their estimation errors. For $\hat{R}$, we show that $\hat{R} = R^0$ with a high probability.

To proceed, we introduce a key invertibility condition for the linear operator $(\Delta^{(1)}, \Delta^{(2)}) \mapsto \mathbf{X}\Delta^{(1)} + \Delta^{(2)}$ 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 $A = USV'$, where S =diag$(s_1, \ldots, s_q)$, with q =rank(A). Then $\operatorname{argmin}_\Theta \left(\frac{1}{2}\|A-\Theta\|_{\mathrm{F}}^2 + \gamma\|\Theta\|_*\right)$ is given by $U\cdot\mathrm{diag}((s_1 - \gamma)_+, \ldots, (s_q-\gamma)_+)\cdot V'$, where $(s)_+ = \max(0,s)$.

[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 $B^0_{ji} \neq 0$. Let $J_i^c = [Np] \backslash J_i$. Let the SVD of Θ^0 be $\Theta^0 = U^0 D^0 V^{0\prime}$, where $U^0 \in \mathbb{R}^{T\times R^0}$ and $V^0 \in \mathbb{R}^{N\times R^0}$. For a $T \times N$ matrix $\Delta^{(2)}$, define the operators

$$\mathcal{P}(\Delta^{(2)}) \equiv U^0_{*,[R^0]} U^{0\prime}_{*,[R^0]} \Delta^{(2)} V^0_{*,[R^0]} V^{0\prime}_{*,[R^0]} \text{ and } \mathcal{M}(\Delta^{(2)}) \equiv \Delta^{(2)} - \mathcal{P}(\Delta^{(2)}).$$

Hence, the operator $\mathcal{P}(\cdot)$ projects a matrix onto a 'low-rank' space which contains Θ^0. For some $c > 0$, the 'cone' $\mathcal{C}_{NT}(c) \subset \mathbb{R}^{Np\times N} \times \mathbb{R}^{T\times N}$ is a set of $(\Delta^{(1)}, \Delta^{(2)})$ satisfying the restriction:

$$\frac{\gamma_1 \sum_{i=1}^N \left|\Delta^{(1)}_{J_i^c,i}\right|_1}{N} + \frac{\gamma_2 \left\|\mathcal{M}(\Delta^{(2)})\right\|_*}{\sqrt{NT}} \leq c\frac{\gamma_1 \sum_{i=1}^N \left|\Delta^{(1)}_{J_i,i}\right|_1}{N} + c\frac{\gamma_2 \left\|\mathcal{P}(\Delta^{(2)})\right\|_*}{\sqrt{NT}}.$$

We impose the following condition.

Assumption A.2 (*Restricted Strong Convexity*)**.** Let

$$\Phi_{\gamma_1,\gamma_2}(\Delta^{(1)}, \Delta^{(2)}) \equiv \frac{\gamma_1}{N}\sum_{i=1}^N \left|\Delta^{(1)}_{J_i,i}\right|_1 + \frac{\gamma_2}{\sqrt{NT}}\|\mathcal{P}(\Delta^{(2)})\|_*$$

be a tolerance function. If $(\Delta^{(1)}, \Delta^{(2)}) \in \mathcal{C}_{NT}(3)$, then there exist positive constants κ, $\kappa'$and κ'' such that

$$\left\|\mathbf{X}\Delta^{(1)} + \Delta^{(2)}\right\|_{\mathrm{F}}^2 \geq T\cdot\kappa' \left\|\Delta^{(1)}\right\|_{\mathrm{F}}^2 + \kappa \left\|\Delta^{(2)}\right\|_{\mathrm{F}}^2 - \kappa''\Phi_{\gamma_1,\gamma_2}(\Delta^{(1)}, \Delta^{(2)})$$

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 $k_i = |J_i|$, $K_J \equiv \sup_i k_i$ and $K_a \equiv \frac{1}{N}\sum_{i=1}^N k_i$. The next assumption involves a regularity condition on the errors:

Assumption A.3. $\|\mathbf{U}\|_{\text{ op}} / \sqrt{NT} \leq \gamma_2/2$, where γ_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), MW (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 $\epsilon^{(u)}_{it}$ 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*

$$N^{-1/2}\left\|\tilde{B} - B^0\right\|_F \leq \bar{c}(\gamma_1\sqrt{K_a} \vee \gamma_2) \textit{ and } (NT)^{-1/2}\left\|\tilde{\Theta} - \Theta^0\right\|_F \leq \bar{c}(\gamma_1\sqrt{K_a} \vee \gamma_2),$$

with probability at least $1 - \varepsilon_{NT} - \bar{c}'(pN^2T^{1-q/4}(\log N)^{-q/2} + pN^{2-\underline{c}\log N})$ *for some finite positive constants* $\underline{c}$, $\bar{c}$, *and* $\bar{c}'$.

Theorem 3.1 establishes non-asymptotic inequalities for the estimation errors of $\tilde{B}$ and $\tilde{\Theta}$ in terms of the Frobenius norm. Note that B^0 and Θ^0 are both high-dimensional matrices with N^2p and NT entries, respectively, and the Frobenius norms are normalized correspondingly by $\sqrt{N}$ and $\sqrt{NT}$. Without any sparsity or approximate sparsity assumption, $\|B^0\|_{\mathrm{F}}^2$ 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 $\frac{1}{N}\|B^0\|_{\mathrm{F}}^2 = \frac{1}{N}\sum_{i=1}^N \left|B^0_{*,i}\right|^2 = O(K_a)$ provided the elements in B^0 are uniformly bounded from the above. This motivates the use of $N^{-1/2}$ to normalize $\|\tilde{B} - B^0\|_{\mathrm{F}}$. The first result in Theorem 3.1 ensures that

$$\frac{1}{N}\|\tilde{B} - B^0\|_{\mathrm{F}}^2 = \frac{1}{N}\sum_{i=1}^N \left|\tilde{B}_{*,i} - B^0_{*,i}\right|^2 \leq \bar{c}^2(((\gamma_1\sqrt{K_a}) \vee \gamma_2)^2) \text{ with high probability.}$$

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

The probability in Theorem 3.1 converges to one when ε_{NT}, $pN^2T^{1-q/4}(\log N)^{-q/2}$ and $pN^{2-\underline{c}\log N}$ 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 $T^{-1}\|\mathbf{U}'\mathbf{X}\|_{\max}$. Specifically, we need a sharp probability bound for a partial sum like $T^{-1}\sum_{t=1}^T y_{i,t-k}u_{jt}$. 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 $y_{i,t-k}u_{jt}$

has a nonlinear Wold presentation $y_{i,t-k}u_{jt} = g_{ijk}(\ldots, \epsilon_{t-1}, \epsilon_t)$, where the $\epsilon_t \equiv (\epsilon_t^{(u)\prime}, \epsilon_t^{(f)\prime})'$ are i.i.d. random variables under Assumption A.1. Then one can verify that the *dependence-adjusted norm* (see Wu and Wu (2016)) of $y_{i,t-k}u_{jt}$ 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 B^0:

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

(ii) Let $\Sigma_F = E(f_t^0 f_t^{0\prime})$. There are constants $s_1 > \cdots > s_{R^0} > 0$ so that s_j equals the jth largest eigenvalue of $\Sigma_F^{1/2}\Sigma_\Lambda\Sigma_F^{1/2}$;

(iii) $K_a = o(T\left(N^{-1/2} + T^{-1/2}\right)/(\log N)^2)$.

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 $\Sigma_F^{1/2}\Sigma_\Lambda\Sigma_F^{1/2}$ 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 $T\left(N^{-1/2} + T^{-1/2}\right)/(\log N)^2$ here, which ensures accuracy of $\tilde{\Theta}$. 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.

Assumption A.4(iii) ensures $\gamma_1\sqrt{K_a} = o(N^{-1/4} + T^{-1/4})$. Consequently, Theorem 3.1 implies that both $N^{-1/2}\|\tilde{B} - B^0\|_{\mathrm{F}}$ and $(NT)^{-1/2}\|\tilde{\Theta} - \Theta^0\|_{\mathrm{F}}$ are $o_P(N^{-1/4} + T^{-1/4})$. This rate can be improved to $O_P(N^{-1/2} + T^{-1/2}\log N)$ if we restrict our attention to the case where $K_a = O(1)$. These error bounds help us to establish the following result which establishes the consistency of $\hat{R}$ and the mean-square convergence rate of $\tilde{F}$.

Theorem 3.2. *Suppose Assumptions A.1–A.4 hold. There exist positive constants $\underline{c}$, $\bar{c}$ and $\bar{c}'$ and a random matrix $\tilde{H}$ depending on (F^0, Λ^0) such that (i) $\hat{R} = R^0$ and (ii) $\|\tilde{F} - F^0\tilde{H}\|_F/\sqrt{T} \le \bar{c}(\gamma_1\sqrt{K_a} \vee \gamma_2)$, both with probability larger than $1 - \varepsilon_{NT} - \bar{c}'(pN^2T^{1-q/4}(\log N)^{-q/2} + pN^{2-\underline{c}\log N})$.*

Theorem 3.2(i) establishes the consistency of $\hat{R}$ and the mean-square convergence rate of $\tilde{F}$ in large samples. Intuitively, since $\tilde{\Theta}$ is a consistent estimator of $\Theta^0 \equiv F^0\Lambda^{0\prime}$ with well-controlled estimation errors, we expect the first R^0 singular values of $\tilde{\Theta}$ to be $O_P(\sqrt{NT})$ and the other singular values to be $O_P[\sqrt{NT}(\gamma_1\sqrt{K_a} \vee \gamma_2)]$. Then the hard SVT procedure can distinguish the $\sqrt{NT}$-order singular values from those of smaller order. Alternatively, given the consistency of $\tilde{B}$ established in Theorem 3.1, the 'residual' $\mathbf{Y} - \mathbf{X}\tilde{B}$ is an approximation of $F^0\Lambda^{0\prime} + \mathbf{U}$. 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 $\tilde{B}$, $\tilde{\Theta}$ 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 $\mathbf{Y}_{*,i}$ on $(\mathbf{X}, \tilde{F})$ for each $i \in [N]$ by imposing an ℓ_1-norm penalty on the coefficient of $\mathbf{X}$. 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 $\gamma_3 = c_3(\gamma_1\sqrt{K_a} \vee \gamma_2)$ for some constant c_3. For each $i \in [N]$, solve the minimization problem:*

$$(\dot{B}'_{*,i}, \dot{\lambda}'_i)' = \underset{(v',\lambda')' \in \mathbb{R}^{Np+R^0}}{\operatorname{argmin}} \frac{1}{2T}\|\mathbf{Y}_{*,i} - \mathbf{X}v - \tilde{F}\lambda\|_{\mathrm{F}}^2 + \gamma_3|v|_1, \tag{3.2}$$

where the Lasso penalty is only imposed on the coefficients of $\mathbf{X}$. The second-step estimators of B^0 and Λ^0 are given by $\dot{B} = (\dot{B}_{*,1}, \ldots, \dot{B}_{*,N})$ and $\dot{\Lambda} = (\dot{\lambda}_1, \ldots, \dot{\lambda}_N)'$.

Remark 3.2. Note that the ℓ_1-norm penalty is only imposed on the coefficient of $\mathbf{X}$. In the proof of Theorem 3.3, we show that $\dot{B}_{*,i}$ solves the Lasso problem with dependent variable $\mathbb{M}_{\tilde{F}}\mathbf{Y}_{*,i}$ and regressors $\mathbb{M}_{\tilde{F}}\mathbf{X}$.

[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 qth order moments of u_{it} and allow for serial correlations in the error term. The term $pN^2T^{1-q/4}(\log N)^{-q/2}$ 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 ($\max_{i,j}|\tilde{B}_{ij} - \tilde{B}_{ij}| = o_P(1)$).

3.2.1. Non-asymptotic results for the second step estimator

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

Theorem 3.3. *Suppose that Assumptions A.1–A.4 hold and $32K_J\gamma_3 \leq \psi_{\min}(\Sigma)$. Then*

$$\|\dot{B} - B^0\|_{max} \leq \max_{1\leq i\leq N} |\dot{B}_{*,i} - B^0_{*,i}|_1 \leq \frac{48}{[\psi_{\min}(\Sigma_X)]^2} K_J\gamma_3$$

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

Theorem 3.3 establishes uniform convergence rates for the elements of $\dot{B}$. Compared to Theorems 3.1–3.2, one additional term $pNe^{-\underline{c}T}$ appears in the probability bound. This term decays in the exponential rate of T and is in general dominated by $p^2N^2T^{1-q/4}(\log N)^{-q/2}$ 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 $\tilde{\Sigma} = \mathbf{X}'\mathbb{M}_{\tilde{F}}\mathbf{X}/T$. 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 $Np > T$, $\tilde{\Sigma}$ is always singular, which leads to $\min_{|v|\neq 0} \frac{v'\tilde{\Sigma}v}{|v|^2} = 0$. The minimum has to be replaced by a minimum over a smaller set in order to obtain a nonzero lower bound. Let $J \subset [Np]$ be an index set and $J^c = [Np]\backslash J$. We say the REC is satisfied for some $K \in [Np]$ if

$$\min_{|J|\leq K} \min_{\substack{|v|\neq 0 \\ |v_{J^c}|_1 \leq 3|v_J|_1}} \frac{v'\tilde{\Sigma}v}{|v_J|^2} \equiv \kappa_{\tilde{\Sigma}}(K) > 0, \tag{3.3}$$

where J has cardinality no bigger than K. In (3.3), the minimum is restricted to those vectors for which $|v_{J^c}|_1 \leq 3|v_J|_1$, where J has cardinality no larger than K. In this restricted space, we establish that (3.3) is satisfied with a high probability for $K = K_J$ in Lemma A.4(v). See also the proof of Lemma A.4(v) 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 leads 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 $Np \times N$ matrix with entries*

$$w_{ki} = \begin{cases} 1 & \text{if } |\dot{B}_{ki}| < \alpha\gamma_4 \\ 0 & \text{if } |\dot{B}_{ki}| \geq \alpha\gamma_4 \end{cases}, \tag{3.4}$$

where $k \in [Np]$, $i \in [N]$, 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 Λ using

$$(\hat{B}^{(\ell)\prime}_{*,i}, \hat{\lambda}^{(\ell)\prime}_i)' = \operatorname*{argmin}_{(v',\lambda')'\in\mathbb{R}^{NP+\hat{R}}} \frac{1}{2T}\left\|\mathbf{Y}_{*,i} - \mathbf{X}v - \hat{F}^{(\ell-1)}\lambda\right\|_F^2 + \gamma_4\sum_{k=1}^{pN} w_{ki}|v_k|,$$

where v_k is the kth entry of v, $i \in [N]$. Let $\hat{B}^{(\ell)} \equiv (\hat{B}^{(\ell)}_{*,1}, \ldots, \hat{B}^{(\ell)}_{*,N})$.

3. (Update $\hat{F}^{(\ell)}$) *Obtain the SVD of $\mathbf{Y} - \mathbf{X}\hat{B}^{(\ell)}$ as $\mathbf{Y} - \mathbf{X}\hat{B}^{(\ell)} = \hat{U}^{(\ell)}\hat{D}^{(\ell)}\hat{V}^{(\ell)\prime}$. Obtain an updated estimate of F^0 as $\hat{F}^{(\ell)} = \sqrt{T}\hat{U}^{(\ell)}_{*,[\hat{R}]}$. Set $\ell = \ell + 1$.*

4. *Iterate steps 2–3 for a finite times ℓ^*. 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 $w_{ki} \equiv \frac{\gamma_{\text{prec}}}{|\dot{B}_{ki}| \vee \gamma_{\text{prec}}}$, where $\gamma_{\text{prec}} = \alpha \gamma_4$.

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 $\hat{B}^{(\ell)}$ 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 $\hat{B}^{(\ell)} =_s B^0$, or $\hat{B}^{(\ell)}$ is sign-consistent for B^0, if and only if sgn($\hat{B}^{(\ell)}_{*,i}$) =sgn($B^0_{*,i}$) for all $i \in [N]$, where sgn$(B_{*,i}) \equiv [\text{sgn}(B_{1,i}), \ldots,\text{sgn}(B_{Np,i})]'$, 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 γ_4 where $\gamma_4 = o(\min_{i\in[N]} \min_{k\in J_i} |B^0_{ki}|)$ and $(K_J^{3/2} T^{-1/2} \log N + K_J N^{-1/2}) = o(\gamma_4)$;
(ii) $N^{-1} \sum_{i=1}^N k_i^2 = O(1)$ and $K_J \log N \cdot (N^{-1/2} \vee T^{-1/2}) = o(1)$;
(iii) $N^2 T^{1-q/4} (\log N)^{-q/2} \to 0$ 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 $\min_{i\in[N]} \min_{k\in J_i} |B^0_{ki}|$ has to be larger than γ_4 in order to separate the nonzero entries from zeros. By Assumption A.5(i) and Theorem 3.3, we can show that $\max_{k\in J_i} w_{ki} = 0$ and $\min_{k\in J_i^c} w_{ki} = 1$ 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 $\|\mathbf{X}(\hat{B}^{(\ell)} - B^0)\|_{\text{F}}$ has a desired convergence rate. The first restriction is imposed to simplify the asymptotic analysis and it implies $K_a = O(1)$ so that we can drop K_a 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 $o[(\sqrt{N} \wedge \sqrt{T})/\log N]$ 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.[11] 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 $\hat{B}^{(\ell)}$ and the preliminary convergence rates of $\hat{B}^{(\ell)}$ and $\hat{F}^{(\ell)}$.

Theorem 3.4. *Suppose that Assumptions A.1–A.5 hold. Then for a fixed ℓ, we have*
(i) $P(\hat{B}^{(\ell)} =_s B^0) \to 1$ as $(N, T) \to \infty$;
(ii) $\|\mathbf{X}(\hat{B}^{(\ell)} - B^0)\|_F/\sqrt{NT} = O_P(\gamma_1 + \gamma_2)$;
(iii) $\|\hat{F}^{(\ell)} - F^0\tilde{H}\|_F/\sqrt{T} = O_P(\gamma_1 + \gamma_2)$.

Theorem 3.4(i) shows that $\hat{B}^{(\ell)}$ 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 $B^0_{*,i}$, 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 $\hat{B}$, $\hat{F}$ and $\hat{\Lambda}$. Now, $\hat{F}$ corresponds to the first $\hat{R}$ eigenvectors of $(\mathbf{Y} - \mathbf{X}\hat{B})(\mathbf{Y} - \mathbf{X}\hat{B})'$, rescaled by $\sqrt{T}$, and one can expand the estimation error $\hat{F} - F^0\tilde{H}$ as in Bai and Ng (2002) and Bai (2009). Then, by examining the product of $\hat{F} - F^0\tilde{H}$ with other terms, a sharper bound for some intermediate estimates can be obtained. Finally we can improve the probability order of each element in $\hat{B}_{J_i,i} - B^0_{J_i,i}$ to $O(T^{-1/2})$.

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

Theorem 3.5. *Suppose that Assumptions A.1–A.5 hold. Let S_i denote an $L \times k_i$ selection matrix such that $S_i S_i' = \mathbb{I}_L$ 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 $S_i(\Sigma_{J_i,J_i})^{-1} S_i'$ exists and is given by Ω_i. Conditional on the event $\{\hat{B} =_s B^0\}$, for each $i \in [N]$, we have $\sqrt{T} S_i(\hat{B}_{J_i,i} - B^0_{J_i,i}) \xrightarrow{d} N(\mathbf{0}, \sigma_i^2 \Omega_i)$ where $\sigma_i^2 = E\left(u_{it}^2\right)$.*

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 $\hat{B}_{J_i,i}$ 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 $T^{q/4-1}$ 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 $\hat{B}_{J_i,i}$. In the special case where k_i is fixed, we can take $S_i = I_{|J_i|}$ and obtain the usual joint asymptotic normal distribution for all elements of $\hat{B}_{J_i,i}$.

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: $y_{it} = \rho_i^0 y_{i,t-1} + \lambda_i^{0\prime} f_t^0 + u_{it}$. Clearly, the key strict stationarity condition in Assumption A.1(v) now becomes $\sup_{N\geq 1} \max_{1\leq i\leq N} \left|\rho_i^0\right| \leq \rho \in (0,1)$. Our final estimator of ρ_i^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 ρ^0, one can average our last stage estimates on ρ_i^0 to obtain a $\sqrt{NT}$ -consistent estimate of ρ^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 γ_ℓ, for $\ell = 1, \ldots, 4$. For γ_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 γ_2 is one such that

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

for some $c > 0$ with high probability. Suppose $\mathbf{U}$ is a random matrix with i.i.d. sub-Gaussian entries that have mean zero and variance σ_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 $\gamma_2 = \frac{\hat{\sigma}_y}{C}(\sqrt{N}+\sqrt{T})/\sqrt{NT}$ for some $C > 1$ and $\hat{\sigma}_y$ is the sample standard deviation of Y. After obtaining an estimate $\hat{\sigma}_u$ of σ_u, we can calculate a suitable γ_2 via simulation. Specifically, we can simulate the random matrices $\mathbf{U}$ with i.i.d. $N(0, \hat{\sigma}_u^2)$. Then we let $\gamma_2 = Q(\|\mathbf{U}\|_{\text{op}}, 0.95)/\sqrt{NT}$, where $Q(x, \alpha)$ denotes the α^{th} quantile of x.

For γ_1, γ_3, and γ_4, we propose to use the 5-fold cross validation (CV) process. Let $\gamma = (\gamma_1, \gamma_3, \gamma_4)'$. For the first-step estimation, the procedure goes as follows:

1. Partition the data into 5 separate sets along the time dimension: $\mathbf{T}_1, \ldots, \mathbf{T}_5 \subset [T]$;
2. For $k = 1, \ldots, 5$, fit the model to the training set by excluding the k th fold data. Denote the estimators by $\tilde{B}^{(\gamma,k)}$ and $\tilde{\Lambda}^{(\gamma,k)}$, where $\tilde{\Lambda}^{(\gamma,k)}$ is a $N \times R$ matrix containing the first R right singular vectors of $\tilde{\Theta}$. Calculate the sum of squared prediction errors
$$cv(\gamma,k) = \text{tr}[(\mathbf{Y}_{\mathbf{T}_k,*} - \mathbf{X}_{\mathbf{T}_k,*}\tilde{B}^{(\gamma,k)})\mathbb{M}_{\tilde{\Lambda}^{(\gamma,k)}}(\mathbf{Y}_{\mathbf{T}_k,*} - \mathbf{X}_{\mathbf{T}_k,*}\tilde{B}^{(\gamma,k)})'];$$
3. Compute the CV error for a fixed tuning parameter by $CV(\gamma) = \sum_{k=1}^5 cv(\gamma,k)$.
4. Select $\gamma^* = \text{argmin}_\gamma\, CV(\gamma)$.

Remark 3.4. Once the sample $\mathbf{T}_k$ is excluded, we cannot obtain an estimate of $F_{\mathbf{T}_k,*}$. Hence we cannot obtain the residuals by deducting the estimate of $F_{\mathbf{T}_k,*}\Lambda'$. For this reason, we multiply $\mathbf{Y}_{\mathbf{T}_k,*} - \mathbf{X}_{\mathbf{T}_k,*}\tilde{B}^{(\gamma,k)}$ by $\mathbb{M}_{\tilde{\Lambda}^{(\gamma,k)}}$ to project out $F_{\mathbf{T}_k,*}\Lambda'$ in the above procedure.T

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 $A_k^0 = \mathbf{0}$ 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 $a_k = \|\hat{A}_k\|_{\text{F}}^2 \vee c$ for some small positive constant c and $k \in [p_{\max}]$;
3. The criterion function we consider is given by the ratio
$$GR(p) = \frac{\sum_{k=p}^{p_{\max}} a_k}{\sum_{k=p+1}^{p_{\max}} a_k},\ \ p = 1, \ldots, p_{\max} - 1.$$
The term GR refers to the growth ratio of $\sum_{k=p}^{p_{\max}} a_k$.
4. Obtain the estimator of p^0 as $\hat{p} = \text{argmax}_{1\leq k<p_{\max}}\, GR\,(k)$.

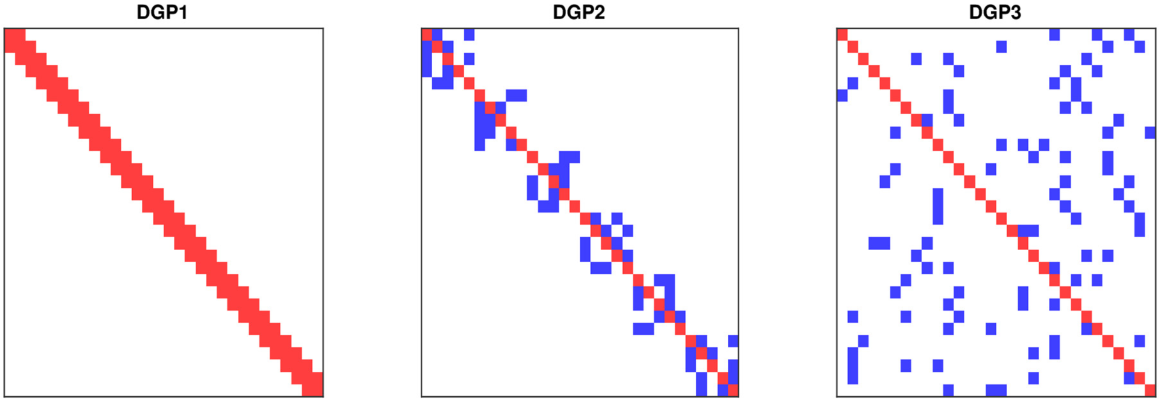

Fig. 1. Structure of the transition matrices in the simulations.

Remark 3.5. Some remarks on this procedure are in order. First, one can also simply run an ℓ_1-nuclear-norm penalized regression with $p_{\max}$, which is the first step of the estimation procedure given in Section 3.1. We only require that $\|\hat{A}_k - A_k^0\|_{\mathrm{F}}$ converge to zero at a certain rate. Second, in practice one may obtain a very small or even zero value for $\|\hat{A}_k\|_{\mathrm{F}}^2$ when $k > p^0$. In this case, if we directly use $a_k = \|\hat{A}_k\|_{\mathrm{F}}^2$, 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 allow some A_k^0 with $k < p^0$ to be a matrix of zeros. If we believe all the A_k^0 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 $p_{\max}$ 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

$$Y_t = A_1^0 Y_{t-1} + \Lambda^0 f_t^0 + u_t, \tag{4.1}$$

where A_1^0 varies across different DGPs, $\Lambda^0 = (\lambda_1^0, \ldots, \lambda_N^0)'$. The factor loadings λ_{ri}^0, for $r = 1, \ldots, R^0$, are independently and identically distributed (i.i.d.) standard normal random variables. The factors f_{tr}^0, for $r = 1, \ldots, R^0$, follow the autoregressive process

$$f_{tr}^0 = \rho_f \cdot f_{t-1,r}^0 + \epsilon_{tr}^{(f)},$$

where $\rho_f = 0.6$ and $\epsilon_{tr}^{(f)}$ 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): $(A_1^0)_{ij} = 0.3 \cdot \mathbf{1}(|i-j| \leq 1)$.

DGP 2 (Block-diagonal transition matrix): We generate a block-diagonal matrix $A_1^0 = \mathrm{bdiag}(S_1, \ldots, S_K)$, where the S_k's are 5×5 random matrices. The diagonal entries of S_k are fixed with $(S_k)_{i,i} = 0.3$. 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_1^0 to be 0.3 (i.e. $(A_1^0)_{ii} = 0.3$). In each row of A_1^0, 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.

DGP	N	T	Number of factors		Step 1		Step 2		Step 3	
			UER	OER	TPR	FPR	TPR	FPR	TPR	FPR
1	30	100	0.0%	0.0%	97.4%	19.3%	98.8%	18.5%	93.7%	8.0%
	30	200	0.0%	0.0%	99.6%	19.1%	99.9%	18.1%	99.4%	5.8%
	30	400	0.0%	0.0%	99.9%	21.8%	100.0%	19.5%	99.9%	4.9%
	60	100	0.0%	0.0%	96.8%	12.7%	98.2%	12.2%	90.5%	5.1%
	60	200	0.0%	0.0%	99.9%	12.2%	100.0%	11.7%	99.1%	2.6%
	60	400	0.0%	0.0%	100.0%	11.9%	100.0%	11.1%	99.9%	1.7%
2	30	100	0.0%	0.0%	86.2%	21.8%	83.9%	18.9%	94.0%	15.7%
	30	200	0.0%	0.0%	95.3%	28.0%	93.7%	24.8%	99.4%	12.8%
	30	400	0.0%	0.0%	99.2%	37.0%	98.7%	33.3%	99.9%	8.2%
	60	100	0.0%	0.0%	76.7%	10.3%	76.5%	9.4%	90.6%	10.7%
	60	200	0.0%	0.0%	88.9%	12.5%	89.7%	12.0%	99.2%	8.9%
	60	400	0.0%	0.0%	96.4%	17.7%	95.8%	16.7%	100.0%	5.5%
3	30	100	0.0%	0.0%	93.2%	24.9%	92.3%	22.0%	96.5%	17.4%
	30	200	0.0%	0.0%	98.1%	31.4%	97.6%	27.6%	99.6%	11.7%
	30	400	0.0%	0.0%	99.5%	38.4%	99.3%	34.4%	99.7%	7.3%
	60	100	0.0%	0.0%	88.1%	12.8%	88.4%	11.8%	95.9%	11.8%
	60	200	0.0%	0.0%	96.1%	15.6%	95.5%	13.9%	99.8%	9.4%
	60	400	0.0%	0.0%	98.9%	19.5%	98.6%	17.9%	100.0%	4.5%

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 squares 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, $Y_t = A_1^0 Y_{t-1} + u_t$, 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 Yılmaz (2014; hereafter DY), Diebold and Yılmaz (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.

DGP	N	T	All entries				Nonzero entries			
			Oracle	Step 1	Step 2	Step 3	Oracle	Step 1	Step 2	Step 3
1	30	100	0.019	0.063	0.059	0.050	0.062	0.145	0.132	0.117
	30	200	0.014	0.055	0.051	0.033	0.044	0.118	0.106	0.066
	30	400	0.010	0.052	0.049	0.029	0.033	0.100	0.092	0.047
	60	100	0.013	0.044	0.041	0.038	0.061	0.150	0.138	0.131
	60	200	0.010	0.035	0.032	0.021	0.043	0.108	0.098	0.066
	60	400	0.007	0.033	0.031	0.016	0.032	0.089	0.080	0.041
2	30	100	0.018	0.065	0.065	0.057	0.056	0.177	0.184	0.154
	30	200	0.012	0.055	0.055	0.038	0.039	0.142	0.150	0.103
	30	400	0.009	0.047	0.047	0.027	0.028	0.110	0.119	0.070
	60	100	0.012	0.050	0.049	0.044	0.054	0.204	0.205	0.179
	60	200	0.008	0.042	0.041	0.028	0.038	0.170	0.168	0.114
	60	400	0.006	0.035	0.035	0.019	0.027	0.138	0.143	0.081
3	30	100	0.019	0.065	0.064	0.055	0.051	0.150	0.155	0.127
	30	200	0.013	0.053	0.053	0.035	0.035	0.117	0.123	0.082
	30	400	0.009	0.047	0.047	0.027	0.025	0.095	0.100	0.058
	60	100	0.013	0.050	0.049	0.042	0.049	0.173	0.173	0.146
	60	200	0.009	0.039	0.040	0.024	0.034	0.135	0.140	0.085
	60	400	0.006	0.033	0.033	0.015	0.024	0.109	0.113	0.056

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.

DGP	N	T	LASSO				Conservative LASSO			
			TPR	FPR	RMSE$_a$	RMSE$_b$	TPR	FPR	RMSE$_a$	RMSE$_b$
1	30	100	78.7%	34.9%	0.115	0.208	78.4%	45.2%	0.178	0.227
	30	200	88.9%	37.7%	0.094	0.178	88.1%	43.3%	0.129	0.173
	30	400	95.3%	45.0%	0.083	0.150	94.5%	43.0%	0.103	0.134
	60	100	71.0%	22.6%	0.086	0.216	72.8%	39.5%	0.161	0.240
	60	200	86.7%	25.7%	0.070	0.179	87.0%	38.9%	0.114	0.175
	60	400	94.9%	30.2%	0.058	0.148	95.3%	37.9%	0.083	0.128
2	30	100	86.2%	59.6%	0.150	0.202	81.9%	54.8%	0.211	0.233
	30	200	95.0%	61.5%	0.107	0.152	91.7%	51.4%	0.139	0.159
	30	400	98.9%	66.3%	0.080	0.113	97.7%	50.5%	0.098	0.110
	60	100	77.0%	46.6%	0.135	0.218	74.1%	48.9%	0.222	0.263
	60	200	91.6%	51.9%	0.100	0.165	86.8%	44.6%	0.143	0.175
	60	400	98.3%	56.1%	0.072	0.120	96.7%	44.4%	0.097	0.116
3	30	100	89.2%	59.2%	0.139	0.186	85.7%	55.9%	0.196	0.215
	30	200	96.2%	61.4%	0.102	0.141	94.0%	54.3%	0.133	0.148
	30	400	99.1%	67.1%	0.079	0.107	98.3%	53.2%	0.096	0.106
	60	100	82.0%	46.1%	0.126	0.203	79.8%	50.6%	0.208	0.247
	60	200	94.0%	51.7%	0.093	0.151	90.5%	46.6%	0.135	0.164
	60	400	98.8%	55.5%	0.068	0.110	97.6%	45.0%	0.091	0.109

Note: We report the true positive rate (TPR), false positive rate (FPR), root mean squared errors of all entries (RMSE$_a$) and nonzero 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 cyclical' 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:

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

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 $\log(\tilde{\sigma}_{it}^2) - \overline{\log(\tilde{\sigma}_{i\cdot}^2)}$.

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 $y_{it} = y_{it}^{(f)} + y_{it}^{(u)}$, where $y_{it}^{(f)}$ is due to the common factors and $y_{it}^{(u)}$ is due to the idiosyncratic errors. Then $\nu_i \equiv \text{var}(y_{it}^{(f)})/\text{var}(y_{it})$ measures the proportion of variance in y_{it} that is due to common factors and $\bar{\nu} \equiv \sum_{i=1}^N \text{var}(y_{it}^{(f)})/\sum_{i=1}^N \text{var}(y_{it})$ measures the corresponding object across the whole panel.

For the idiosyncratic component $y_{it}^{(u)}$ we calculate the measure of connectedness proposed by DY (2014). As discussed in Section 2, we have $y_{it}^{(u)} = \sum_{j=0}^\infty \alpha_{iN}^{(u)}(j) C^{(u)} \epsilon_{t-j}^{(u)}$, where $\alpha_{iN}^{(u)}(j) = (e_{1,p} \otimes e_{i,N})' \Phi^j (e_{1,p} \otimes I_N)$ and $\epsilon_t^{(u)} \sim (0, I_m)$. For simplicity, suppose that $m = N$. Then one can treat $\epsilon_{it}^{(u)}$ as the idiosyncratic shock to individual i. The variance of the H-step ahead prediction error due to $\{\epsilon_{j,t+h}^{(u)}\}_{h=1}^H$ is $s_{ij}^H = \sum_{h=0}^{H-1}([\alpha_{iN}^{(u)}(h) C^{(u)}]_j)^2$. If we can identify both Φ and $C^{(u)}$, we can easily estimate the variance decomposition matrix $\check{D}^H$ with (i,j)th entry $s_{ij}^H / \sum_{k=1}^N s_{ik}^H$. However, $C^{(u)}$ is not identified without further assumption. Although we cannot identify $C^{(u)}$, the matrix $\Sigma_u = C^{(u)} C^{(u)\prime}$ is identified. DY (2014) propose to calculate the H-step generalized variance decomposition matrix[12] $D^H = [d_{ij}^H]_{N\times N}$, where

$$d_{ij}^H = \frac{\sigma_{jj}^{-1} \sum_{h=0}^{H-1} (\alpha_{iN}^{(u)}(h) \Sigma_u e_{j,N})^2}{\sum_{h=0}^{H-1} \alpha_{iN}^{(u)}(h) \Sigma_u \alpha_{iN}^{(u)}(h)'}, \text{ and } e_{j,N} \text{ is the } j\text{th column of } I_N.$$

Unlike $\check{D}^H$, the row sums of D^H are not necessarily unity. We normalize D^H to $\tilde{D}^H$ with (i,j)th entry $\tilde{d}_{ij}^H = d_{ij}^H / \sum_{k=1}^N d_{ik}^H$ so that $\sum_{j=1}^N \tilde{d}_{ij}^H = 1$ and $\sum_{i,j=1}^N \tilde{d}_{ij}^H = N$. Hence, the overall connectedness in the $y_{it}^{(u)}$'s can be measured as $\tilde{d}^H = \sum_{i\neq j} \tilde{d}_{ij}^H / N$. In addition, we let $\tilde{d}_{i\leftarrow}^H \equiv \sum_{j\neq i} \tilde{d}_{ij}^H$. Following DY (2014), we call $\tilde{d}_{i\leftarrow}^H$ the 'FROM' index, as it measures the proportion of generalized variance decomposition that is due to other individuals. Similarly, we let $\tilde{d}_{\leftarrow j}^H \equiv \sum_{i\neq j} \tilde{d}_{ij}^H$ and 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 $p_{\max} = 8$. The result gives $\hat{p} = 4$. When we run the regression with $p = 4$, the number of factors is determined to be one ($\hat{R} = 1$).

Fig. 2 reports the heat map representing the estimates of the $\hat{A}_k$'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 $\hat{A}_k$ 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 $\hat{A}_k$ 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 ν_i, $\tilde{d}_{i\leftarrow}^H$, and $\tilde{d}_{\leftarrow j}^H$. Almost all the ν_i's are above 50%, and the overall variation due to the common factors is $\bar{\nu} = 56.1\%$. These results imply that market level investor fear plays a dominant roll in investor trading behavior. After conditioning on the factors, we consider the idiosyncratic part by looking at $\tilde{d}_{i\leftarrow}^H$, $\tilde{d}_{\leftarrow j}^H$ 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

[12] 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).

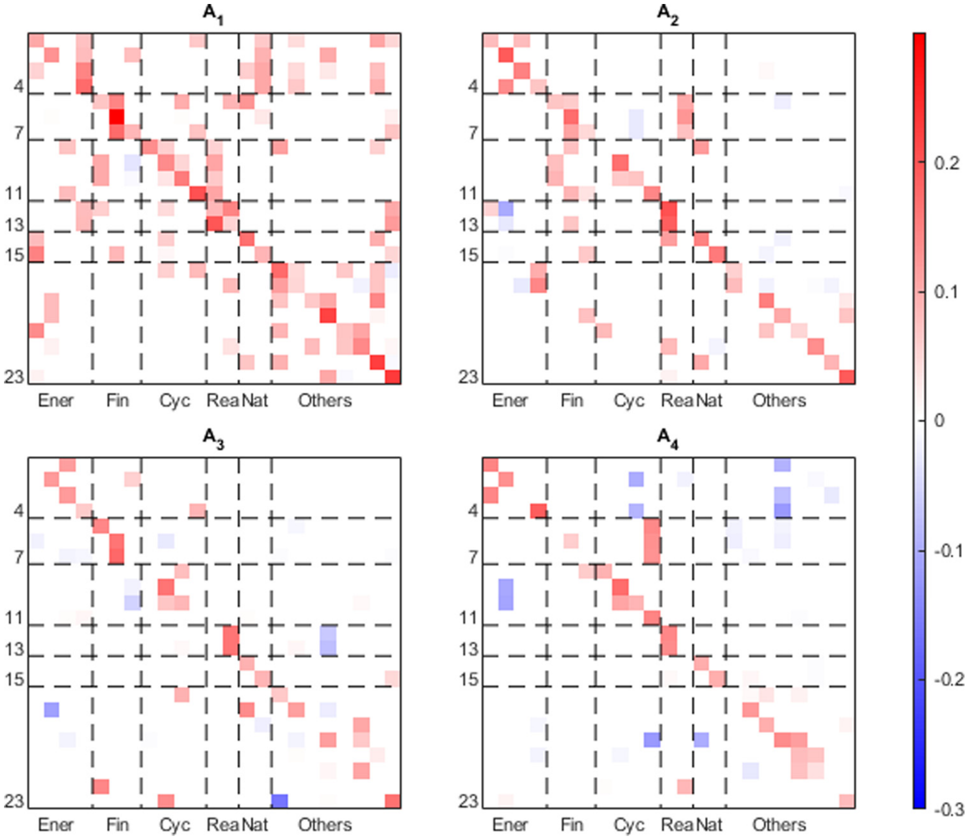

Fig. 2. Heat map of the transition matrices A_k's. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

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 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.

Connectedness measures by estimates of VAR with CFs model								
TICKER	XLE	XOP	IYE	OIH	XLF	KBE	KRE	XLY
ν_i	64.9%	59.1%	65.4%	58.0%	65.2%	56.8%	56.6%	72.0%
FROM	71.4%	65.4%	71.7%	64.3%	61.7%	61.3%	62.3%	51.8%
TO$_i$	106.8%	86.0%	103.9%	71.5%	57.8%	72.6%	51.4%	37.3%
TICKER	XHB	ITB	XRT	IYR	VNQ	XLB	XME	XLK
ν_i	53.6%	49.5%	60.1%	50.7%	49.7%	67.2%	56.9%	70.5%
FROM$_i$	60.5%	58.3%	36.5%	57.9%	58.6%	37.5%	44.1%	39.0%
TO$_i$	56.3%	41.7%	19.0%	79.7%	74.4%	26.3%	37.2%	37.3%
TICKER	SMH	XLV	IBB	XLP	XLU	XLI	GDX	Average
ν_i	54.8%	64.3%	50.7%	61.3%	50.6%	67.7%	31.0%	$\bar{\nu} = 56.1\%$
FROM$_i$	31.9%	38.3%	28.8%	30.7%	29.7%	40.9%	27.7%	$\bar{d}^{12} = 49.8\%$
TO$_i$	23.3%	34.1%	33.0%	21.2%	19.6%	20.7%	19.1%	
Connectedness measures by estimates of pure VAR model								
TICKER	XLE	XOP	IYE	OIH	XLF	KBE	KRE	XLY
FROM$_i$	89.3%	87.1%	89.4%	87.0%	89.6%	86.8%	87.6%	90.9%
TO$_i$	105.0%	79.5%	103.0%	77.7%	112.9%	97.0%	89.1%	110.5%
TICKER	XHB	ITB	XRT	IYR	VNQ	XLB	XME	XLK
FROM$_i$	87.3%	86.3%	88.8%	85.7%	86.2%	90.1%	88.8%	89.8%
TO$_i$	95.8%	80.8%	79.1%	94.0%	89.6%	105.6%	80.1%	103.8%
TICKER	SMH	XLV	IBB	XLP	XLU	XLI	GDX	Average
FROM$_i$	87.6%	88.1%	83.8%	88.4%	85.7%	89.8%	76.5%	$\bar{d}^{12} = 87.40\%$
TO$_i$	74.8%	81.2%	60.8%	80.0%	60.0%	104.3%	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.

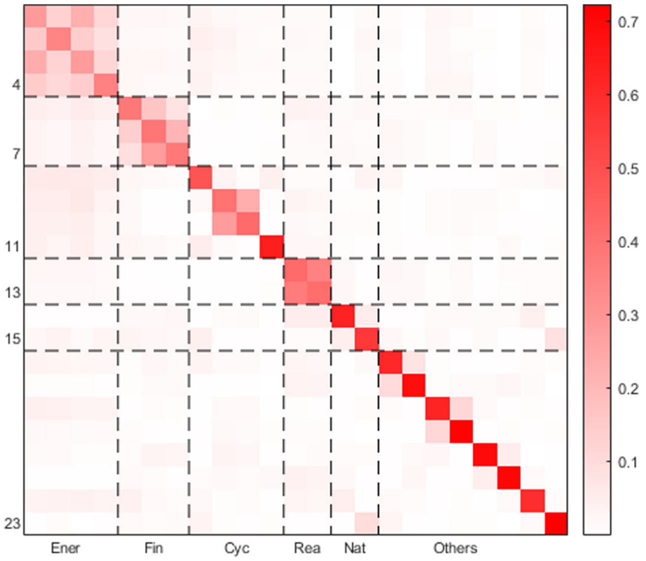

Fig. 3. Heat map of $\tilde{D}^{12}$.

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 ℓ_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 $y_{it}^{(u)}$ and $y_{it}^{(f)}$ are mutually independent. It suffices to study them separately. By Assumption A.1(i), we can write $y_{it}^{(u)}$ as the linear process

$$y_{it}^{(u)}=\sum_{j=0}^{\infty}\alpha_{iN}^{(u)}(j)u_{t-j}=\sum_{j=0}^{\infty}\alpha_{iN}^{(u)}(j)C^{(u)}\epsilon_{t-j}^{(u)}\equiv\sum_{j=0}^{\infty}C_j^{(i,u)}\epsilon_{t-j}^{(u)},$$

where $C_j^{(i,u)}\equiv\alpha_{iN}^{(u)}(j)C^{(u)}$. Under Assumption A.1(vi), one can bound $|(e_{1,p}\otimes e_{i,N})'\Phi^j|$ by $\psi_{\max}([\Phi^j]_{[N],[N]})\le\bar{c}\rho^j$. It follows that $|\alpha_{iN}^{(u)}(j)|\le\bar{c}\rho^j$. Then the MA($\infty$) representation of $y_{it}^{(u)}$ is valid with $E(y_{it}^{(u)})=0$ and $\mathrm{Var}(y_{it}^{(u)})=\sum_{j=0}^{\infty}\alpha_{iN}^{(u)}(j)\Sigma_u\alpha_{iN}^{(u)}(j)'<\infty$.

Under Assumption A.1(vi), we can also show that $E(|y_{it}^{(f)}|)\le\sum_{j=0}^{\infty}|\alpha_{iN}^{(f)}(j)|\left|\mu_f\right|<\infty$. The MA($\infty$) representation of $y_{it}^{(f)}$ is

$$y_{it}^{(f)}=E(y_{it}^{(f)})+\sum_{j=0}^{\infty}\alpha_{iN}^{(f)}(j)(f_{t-j}^0-\mu_f)=E(y_{it})+\sum_{j=0}^{\infty}C_j^{(i,f)}\epsilon_{t-j}^{(f)},$$

where $C_j^{(i,f)}\equiv\sum_{k=0}^{j}\alpha_{iN}^{(f)}(k)C_{j-k}^{(f)}$. Under Assumption A.1(vi), $|C_j^{(i,f)}|\le\sum_{k=0}^{j}|\alpha_{iN}^{(f)}(k)|\cdot\|C_{j-k}^{(f)}\|_{\mathrm{op}}$. In addition, by Assumption A.1(ii),

$$\sum_{j=0}^{\infty}\sum_{k=0}^{j}\rho^k\|C_{j-k}^{(f)}\|_{\max}=\sum_{k=0}^{\infty}\rho^k\sum_{j=k}^{\infty}\|C_{j-k}^{(f)}\|_{\max}\le\bar{c}\sum_{k=0}^{\infty}\rho^k(k+1)^{-\alpha},$$

for some constant $\bar{c}<\infty$. Hence $C_j^{(i,f)}$ is absolutely summable, $\mathrm{Var}(y_{it}^{(f)})=\sum_{j=0}^{\infty}C_j^{(i,f)}C_j^{(i,f)\prime}<\infty$, and the MA($\infty$) representation of $y_{it}^{(f)}$ is valid.

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

$$\psi_{\min}(\Sigma_X^{(u)})\ge\frac{\psi_{\min}(\Sigma_u)}{\max_{|z|=1}\psi_{\max}(\mathcal{A}^*(z)\mathcal{A}(z))}.$$

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

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

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 Θ^0 and Δ, we have*

(i) $\left\|\Theta^0+\mathcal{M}(\Delta)\right\|_*=\left\|\Theta^0\right\|_*+\|\mathcal{M}(\Delta)\|_*$;

(ii) $\|\Delta\|_F^2=\|\mathcal{M}(\Delta)\|_F^2+\|\mathcal{P}(\Delta)\|_F^2$;

(iii) $rank(\mathcal{P}(\Delta))\le 2R^0$;

(iv) $\|\Delta\|_F^2=\sum_j\psi_j(\Delta)^2$ and $\|\Delta\|_*^2\le\|\Delta\|_F^2 rank(\Delta)$;

For any conformable matrices M_1 and M_2, the following statement holds:

(v) $|tr(M_1M_2)|\le\|M_1\|_{max}|vec(M_2)|_1$ and $|tr(M_1M_2)|\le\|M_1\|_{op}\|M_2\|_*$.

Lemma A.2. *Suppose that Assumption A.1 hold. There exist absolute constants $c, \underline{c}, \bar{c} \in (0, \infty)$ such that*
(i) $\left\|\mathbf{U}'\mathbf{X}\right\|_{max}/T \le \gamma_1/2$ *with probability greater than* $1-\bar{c}(pN^2T^{1-q/4}(\log N)^{-q/2}+pN^{2-\underline{c}\log N})$*;*
(ii) $\left\|\mathbf{U}'\mathbb{P}_{F^0}\mathbf{X}\right\|_{max}/T \le c\cdot\gamma_1$ *with probability greater than* $1-\bar{c}[pN(T^{1-q/4}(\log N)^{-q/2} \vee e^{-\underline{c}T})+pN^{1-\underline{c}\log N}]$.

Proof of Theorem 3.1. Let $\tilde{\Delta}^{(1)} = \tilde{B}-B^0$ and $\tilde{\Delta}^{(2)} = \tilde{\Theta}-\Theta^0$. Define the event

$$\mathcal{E}_{NT}^{(1)} = \{\left\|\mathbf{U}'\mathbf{X}\right\|_{\max}/T \le \gamma_1/2,\, \|\mathbf{U}\|_{\text{op}}/\sqrt{NT} \le \gamma_2/2\}.$$

By Lemma A.2(i) and Assumption A.3(i), $\mathcal{E}_{NT}^{(1)}$ holds with probability at least $1-\bar{c}[pN^2T^{1-q/2}(\log N)^{-q/2}\ +pN^{2-\underline{c}\log N}]$. By the definition of $(\tilde{B}, \tilde{\Theta})$, we have

$$\begin{aligned}0 &\ge \mathcal{L}(\tilde{B},\tilde{\Theta})-\mathcal{L}(B^0,\Theta^0)\\ &= \frac{1}{2NT}(\|\mathbf{Y}-\mathbf{X}\tilde{B}-\tilde{\Theta}\|_{\text{F}}^2-\|\mathbf{U}\|_{\text{F}}^2)+\frac{\gamma_1}{N}(|\text{vec}(\tilde{B})|_1-|\text{vec}(B^0)|_1)+\frac{\gamma_2}{\sqrt{NT}}(\|\tilde{\Theta}\|_*-\|\Theta^0\|_*)\\ &\equiv d_1+d_2+d_3. \end{aligned} \tag{A.1}$$

To establish the asymptotic properties of $\tilde{B}$ and $\tilde{\Theta}$, we study d_1, d_2 and d_3 in turn.

First, consider d_1. By the identity $\mathbf{Y} = \mathbf{X}B^0+\Theta^0+\mathbf{U}$, we have

$$\left\|\mathbf{Y}-\mathbf{X}\tilde{B}-\tilde{\Theta}\right\|_{\text{F}}^2-\|\mathbf{U}\|_{\text{F}}^2 = \left\|\mathbf{X}\tilde{\Delta}^{(1)}+\tilde{\Delta}^{(2)}\right\|_{\text{F}}^2-2\,\text{tr}[\mathbf{U}'(\mathbf{X}\tilde{\Delta}^{(1)}+\tilde{\Delta}^{(2)})].$$

For $\text{tr}[\mathbf{U}'(\mathbf{X}\tilde{\Delta}^{(1)}+\tilde{\Delta}^{(2)})]$, conditional on $\mathcal{E}_{NT}^{(1)}$, we apply the triangle inequality and Lemma A.1(v) to obtain

$$\begin{aligned}\frac{1}{NT}|\text{tr}[\mathbf{U}'(\mathbf{X}\tilde{\Delta}^{(1)}+\tilde{\Delta}^{(2)})]| &\le \frac{1}{NT}\|\mathbf{U}'\mathbf{X}\|_{\max}|\text{vec}(\tilde{\Delta}^{(1)})|_1+\frac{1}{NT}\|\mathbf{U}\|_{\text{op}}\|\tilde{\Delta}^{(2)}\|_*\\ &\le \frac{\gamma_1}{2N}|\text{vec}(\tilde{\Delta}^{(1)})|_1+\frac{\gamma_2}{2\sqrt{NT}}\|\tilde{\Delta}^{(2)}\|_*.\end{aligned}$$

It follows that

$$\begin{aligned}d_1 &\ge \frac{1}{2NT}\|\mathbf{X}\tilde{\Delta}^{(1)}+\tilde{\Delta}^{(2)}\|_{\text{F}}^2-\frac{\gamma_1}{2N}|\text{vec}(\tilde{\Delta}^{(1)})|_1-\frac{\gamma_2}{2\sqrt{NT}}\|\tilde{\Delta}^{(2)}\|_*\\ &\ge \frac{1}{2NT}\|\mathbf{X}\tilde{\Delta}^{(1)}+\tilde{\Delta}^{(2)}\|_{\text{F}}^2-\frac{\gamma_1}{2N}\sum_{i=1}^N\left(|\tilde{\Delta}_{J_i,i}^{(1)}|_1+|\tilde{\Delta}_{J_i^c,i}^{(1)}|_1\right)\\ &\quad -\frac{\gamma_2}{2\sqrt{NT}}\left(\|\mathcal{P}(\tilde{\Delta}^{(2)})\|_*+\|\mathcal{M}(\tilde{\Delta}^{(2)})\|_*\right).\end{aligned} \tag{A.2}$$

Next, consider d_2. By the identities $|\tilde{B}_{*,i}|_1 = |\tilde{B}_{J_i,i}|_1+|\tilde{B}_{J_i^c,i}|_1$ and $|B_{*,i}^0|_1 = |B_{J_i,i}^0|_1$, we have

$$d_2 = \frac{\gamma_1}{N}\sum_{i=1}^N(|\tilde{B}_{J_i,i}|_1+|\tilde{B}_{J_i^c,i}|_1-|B_{J_i,i}^0|_1) \ge \frac{\gamma_1}{N}\sum_{i=1}^N(|\tilde{\Delta}_{J_i^c,i}^{(1)}|_1-|\tilde{\Delta}_{J_i,i}^{(1)}|_1), \tag{A.3}$$

where we use the fact that $|\tilde{B}_{J_i,i}|_1+|\tilde{\Delta}_{J_i,i}^{(1)}|_1 \ge |B_{J_i,i}^0|_1$ by the triangle inequality and that $|\tilde{B}_{J_i^c,i}|_1 = |\tilde{\Delta}_{J_i^c,i}^{(1)}|_1$ as $B_{J_i^c,i}^0 = 0$.

Now, consider d_3. By the triangle inequality and Lemma A.1(i), we have

$$\begin{aligned}\|\tilde{\Theta}\|_* &= \|\tilde{\Delta}^{(2)}+\Theta^0\|_* = \|\Theta^0+\mathcal{P}(\tilde{\Delta}^{(2)})+\mathcal{M}(\tilde{\Delta}^{(2)})\|_*\\ &\ge \|\Theta^0+\mathcal{M}(\tilde{\Delta}^{(2)})\|_*-\|\mathcal{P}(\tilde{\Delta}^{(2)})\|_*\\ &= \|\Theta^0\|_*+\|\mathcal{M}(\tilde{\Delta}^{(2)})\|_*-\|\mathcal{P}(\tilde{\Delta}^{(2)})\|_*.\end{aligned}$$

It follows that

$$d_3 \ge \frac{\gamma_2}{\sqrt{NT}}(\|\mathcal{M}(\tilde{\Delta}^{(2)})\|_*-\|\mathcal{P}(\tilde{\Delta}^{(2)})\|_*). \tag{A.4}$$

Combining the results in (A.1)–(A.4), we have

$$\begin{aligned}&\frac{1}{2NT}\|\mathbf{X}\tilde{\Delta}^{(1)}+\tilde{\Delta}^{(2)}\|_{\text{F}}^2+\frac{\gamma_1}{2N}\sum_{i=1}^N\|\tilde{\Delta}_{J_i^c,i}^{(1)}\|_1+\frac{\gamma_2}{2\sqrt{NT}}\|\mathcal{M}(\tilde{\Delta}^{(2)})\|_*\\ &\le \frac{3\gamma_1}{2N}\sum_{i=1}^N\|\tilde{\Delta}_{J_i,i}^{(1)}\|_1+\frac{3\gamma_2}{2\sqrt{NT}}\|\mathcal{P}(\tilde{\Delta}^{(2)})\|_*.\end{aligned} \tag{A.5}$$

The above inequality indicates that $(\tilde{\Delta}^{(1)}, \tilde{\Delta}^{(2)}) \in \mathcal{C}_{NT}(3)$. By Assumption A.2, with probability $1-\varepsilon_{NT}$ we have

$$\frac{1}{N}\|\tilde{\Delta}^{(1)}\|_{\mathrm{F}}^2 + \frac{1}{NT}\|\tilde{\Delta}^{(2)}\|_{\mathrm{F}}^2 - \kappa_2 \Phi_{\gamma_1,\gamma_2}(\tilde{\Delta}^{(1)}, \tilde{\Delta}^{(2)}) \leq \kappa_1 \frac{1}{NT}\|\mathbf{X}\tilde{\Delta}^{(1)} + \tilde{\Delta}^{(2)}\|_{\mathrm{F}}^2, \tag{A.6}$$

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

$$\begin{aligned}\frac{1}{NT}\|\mathbf{X}\tilde{\Delta}^{(1)} + \tilde{\Delta}^{(2)}\|_{\mathrm{F}}^2 &\leq 3\Phi_{\gamma_1,\gamma_2}(\tilde{\Delta}^{(1)}, \tilde{\Delta}^{(2)}) = \frac{3\gamma_1}{N}\sum_{i=1}^{N}\left|\tilde{\Delta}^{(1)}_{J_i,i}\right|_1 + \frac{3\gamma_2}{\sqrt{NT}}\|\mathcal{P}(\tilde{\Delta}^{(2)})\|_* \\ &\leq 3\gamma_1\sqrt{K_a}\frac{\|\tilde{\Delta}^{(1)}\|_{\mathrm{F}}}{\sqrt{N}} + 3\sqrt{2R^0}\gamma_2\frac{\|\tilde{\Delta}^{(2)}\|_{\mathrm{F}}}{\sqrt{NT}} \\ &\leq 3\sqrt{2}(\gamma_1\sqrt{K_a} \vee (\sqrt{2R^0}\gamma_2))\sqrt{\frac{1}{N}\|\tilde{\Delta}^{(1)}\|_{\mathrm{F}}^2 + \frac{1}{NT}\|\tilde{\Delta}^{(2)}\|_{\mathrm{F}}^2},\end{aligned} \tag{A.7}$$

where the second inequality holds by Lemma A.1(ii)–(iv) and the fact that $\sum_{i=1}^N |\tilde{\Delta}^{(1)}_{J_i,i}|_1 \leq \sqrt{NK_a}(\sum_{i=1}^N |\tilde{\Delta}^{(1)}_{J_i,i}|^2)^{1/2} \leq \sqrt{NK_a}\|\tilde{\Delta}^{(1)}\|_{\mathrm{F}}$, where recall that $K_a = N^{-1}\sum_{i=1}^N k_i$ 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 $(1-\varepsilon_{NT})(1-\bar{c}[pN^2T^{1-q/2}(\log N)^{-q/2} + pN^{2-\underline{c}\log N}]) \geq 1-\varepsilon_{NT} - \bar{c}[pN^2T^{1-q/2}(\log N)^{-q/2} + pN^{2-\underline{c}\log N}]$,

$$\frac{1}{N}\|\tilde{\Delta}^{(1)}\|_{\mathrm{F}}^2 + \frac{1}{NT}\|\tilde{\Delta}^{(2)}\|_{\mathrm{F}}^2 \leq (3\kappa_1+\kappa_2)\sqrt{2}[(\gamma_1\sqrt{K_a}) \vee (\sqrt{2R^0}\gamma_2)]\sqrt{\frac{1}{N}\|\tilde{\Delta}^{(1)}\|_{\mathrm{F}}^2 + \frac{1}{NT}\|\tilde{\Delta}^{(2)}\|_{\mathrm{F}}^2},$$

which implies that $\frac{1}{\sqrt{N}}\|\tilde{\Delta}^{(1)}\|_{\mathrm{F}} \leq \bar{c}(\gamma_1\sqrt{K_a} \vee \gamma_2)$ and $\frac{1}{\sqrt{NT}}\|\tilde{\Delta}^{(2)}\|_{\mathrm{F}} \leq \bar{c}(\gamma_1\sqrt{K_a} \vee \gamma_2)$ with $\bar{c} = (3\kappa_1+\kappa_2)\sqrt{2}(1 \vee \sqrt{2R^0}) < \infty$. 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 $S_F \equiv F^{0\prime}F^0/T$. Then for any $x > 0$,*

$$P(T^{1/2}\|S_F - \Sigma_F\|_{max} > x) \leq C_1 x^{-q/2}T^{1-q/4} + C_2\exp\left(-C_3x^2\right)$$

for some absolute constants C_ℓ, $\ell = 1, 2, 3$.

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

$$\mathcal{E}_{NT}^{(2)} = \{\left\|\mathbf{U}'\mathbf{X}\right\|_{\max}/T \leq \gamma_1/2,\ \|\mathbf{U}\|_{\mathrm{op}}/\sqrt{NT} \leq \gamma_2/2 \text{ and } \|S_F - \Sigma_F\|_{\mathrm{op}} \leq c\sqrt{\log NT}^{-1/2}\},$$

where c is a large positive constant. One can verify that for some positive constants $\bar{c}'$ and $\underline{c}$,

$$P(\mathcal{E}_{NT}^{(2)}) \geq 1 - \bar{c}'(pN^2T^{1-q/4}(\mathrm{log}N)^{-q/2} + pN^{2-\underline{c}\mathrm{log}N})$$

by Lemmas A.2–A.3. From Theorem 3.1, we have with probability at least $1-\varepsilon_{NT} - \bar{c}'(pN^2T^{1-q/4}(\log N)^{-q/2} + pN^{2-\underline{c}\log N})$,

$$(NT)^{-1/2}\|\tilde{\Theta} - \Theta^0\|_{\mathrm{op}} \leq (NT)^{-1/2}\|\tilde{\Theta} - \Theta^0\|_{\mathrm{F}} \leq \bar{c}(\gamma_1\sqrt{K_a} \vee \gamma_2).$$

Next, we show that $\mathcal{E}_{NT}^{(2)}$ implies the desired results.

Step 1: Bound the eigenvalues.

Let $S_\Lambda = \Lambda^{0\prime}\Lambda^0/N$ and $S_F = F^{0\prime}F^0/T$. Let $\hat{s}_1 \geq \cdots \geq \hat{s}_{R^0}$ be the R^0 nonzero eigenvalues of $\frac{1}{NT}\Theta^0\Theta^{0\prime} = \frac{1}{T}F^{0\prime}S_\Lambda F^0$. Note that $\hat{s}_1, \ldots, \hat{s}_{R^0}$ are the same as the eigenvalues of $S_F^{1/2}S_\Lambda S_F^{1/2}$. Conditional on the event $\mathcal{E}_{NT}^{(2)}$ and by Assumption A.4(i)–(ii), we have

$$|\hat{s}_j - s_j| \leq \bar{c}(\sqrt{\mathrm{log}N}T^{-1/2} + N^{-1/2}) \text{ for some } \bar{c} < \infty \text{ and} j \in [R^0].$$

This also implies that $\|\Theta^0\|_{\mathrm{op}} = \sqrt{(s_1 + o_P(1))NT}$. For $j > R^0$, simply define $\hat{s}_j = s_j = 0$.

Let $\tilde{s}_1 \geq \cdots \geq \tilde{s}_{N\wedge T}$ be the eigenvalues of $\frac{1}{NT}\tilde{\Theta}\tilde{\Theta}'$. Again by the Weyl's theorem, we have for $j = 1, 2, \ldots$

$$\begin{aligned}|\tilde{s}_j - s_j| &\leq |\tilde{s}_j - \hat{s}_j| + |\hat{s}_j - s_j| \\ &\leq \frac{1}{NT}\|\tilde{\Theta}\tilde{\Theta}' - \Theta^0\Theta^{0\prime}\|_{\mathrm{op}} + |\hat{s}_j - s_j| \\ &\leq \frac{2}{NT}\|\Theta^0\|_{\mathrm{op}}\|\tilde{\Theta} - \Theta^0\|_{\mathrm{op}} + \frac{1}{NT}\|\tilde{\Theta} - \Theta^0\|_{\mathrm{op}}^2 + |\hat{s}_j - s_j|,\end{aligned}$$

implying $|\tilde{s}_j - s_j| \leq \bar{c}(\gamma_1\sqrt{K_a} \vee \gamma_2)$ for $j = 1, 2, \ldots$ Then for $j \in [R^0]$, w.p.a.1,

$$|\hat{s}_{j-1} - \tilde{s}_j| \geq |\hat{s}_{j-1} - \hat{s}_j| - \left|\hat{s}_j - \tilde{s}_j\right| \geq (s_{j-1} - s_j)/2 \text{ and}$$

$$|\tilde{s}_j - \hat{s}_{j+1}| \geq |\hat{s}_j - \hat{s}_{j+1}| - |\tilde{s}_j - \hat{s}_j| \geq (s_j - s_{j+1})/2, \tag{A.8}$$

with $\hat{s}_{R^0+1} = s_{R^0+1} = 0$.

Step 2: Prove the consistency of $\hat{R}$.

Note that $\psi_r(\tilde{\Theta}) = \sqrt{NT\tilde{s}_r}$. By the result in Step 1, we have that $\psi_r(\tilde{\Theta}) \geq \sqrt{[s_{R^0} - o_P(1)]NT}$ for all $r \leq R^0$, and

$$\psi_{R^0+1}(\tilde{\Theta}) \leq \psi_{R^0+1}(\Theta^0) + \left\|\tilde{\Theta} - \Theta^0\right\|_{\text{op}} \leq \left\|\tilde{\Theta} - \Theta^0\right\|_{\text{F}} \leq \sqrt{NT}\bar{c}(\gamma_1\sqrt{K_a} \vee \gamma_2) = \sqrt{NT}o(\gamma_2^{1/2})$$

where we use the condition that $\gamma_1\sqrt{K_a} = o(\gamma_2^{1/2})$ under Assumption A.4(iii). These results, in conjunction with the fact that $(\gamma_2\sqrt{NT}\|\tilde{\Theta}\|_{\text{op}})^{1/2} \asymp \sqrt{NT}\sqrt{\gamma_2}$ with $\gamma_2 = c_2(N^{-1/2} + T^{-1/2})$,[13] imply that

$$\min_{r\leq R^0} \psi_r(\tilde{\Theta}) \geq (\gamma_2\sqrt{NT}\|\tilde{\Theta}\|_{\text{op}})^{1/2} \text{ and } \psi_{R^0+1}(\tilde{\Theta}) < (\gamma_2\sqrt{NT}\|\tilde{\Theta}\|_{\text{op}})^{1/2}$$

with probability at least $1 - \varepsilon_{NT} - \bar{c}'(N^2T^{1-q/4}(\log N)^{-q/2} + N^{2-\underline{c}\log N})$ for sufficiently large (N, T). Then we have $\hat{R} = R^0$ with probability at least $1 - \varepsilon_{NT} - \bar{c}'(N^2T^{1-q/4}(\log N)^{-q/2} + N^{2-\underline{c}\log N})$ for sufficiently large (N, T).

Step 3: Characterize the eigenvectors.

Next, we show that there is an $R^0 \times R^0$ matrix $\tilde{H}$, such that the columns of $\frac{1}{\sqrt{T}}F^0\tilde{H}$ are the first R^0 eigenvectors of $\Theta^0\Theta^{0\prime}$. Let v be the $R^0 \times R^0$ matrix whose columns are the eigenvectors of $S_F^{1/2}S_\Lambda S_F^{1/2}$. Then $D = v'S_F^{1/2}S_\Lambda S_F^{1/2}v$ is a diagonal matrix of the eigenvalues of $S_F^{1/2}S_\Lambda S_F^{1/2}$ that are distinct by Assumption A.4(ii). Let $\tilde{H} = S_F^{-1/2}v$. Then

$$\begin{aligned}
\frac{1}{NT}\Theta^0\Theta^{0\prime}F^0\tilde{H} &= \frac{1}{T}F^0S_\Lambda F^{0\prime}F^0\tilde{H} = F^0S_\Lambda S_F\tilde{H} = F^0S_\Lambda S_F^{1/2}v \\
&= F^0S_F^{1/2}S_F^{-1/2}S_\Lambda S_F^{1/2}v = F^0S_F^{1/2}vv'S_F^{-1/2}S_\Lambda S_F^{1/2}v \\
&= F^0\tilde{H}D.
\end{aligned}$$

In addition, we have $(F^0\tilde{H})'F^0\tilde{H}/T = v'S_F^{-1/2}\frac{F^{0\prime}F^0}{T}S_F^{-1/2}v = v'v = I_{R^0}$. So the columns of $\frac{1}{\sqrt{T}}F^0\tilde{H}$ are the eigenvectors of $\Theta^0\Theta^{0\prime}$, with corresponding eigenvalues in D.

Step 4: Prove the convergence.

We bound $\left\|\tilde{F} - F^0\tilde{H}\right\|_{\text{F}}$ conditional on the event $\hat{R} = R^0$. By the Davis–Kahan $\sin(\Theta)$ theorem (see, e.g., Yu et al. (2014)) and (A.8),

$$\begin{aligned}
\frac{1}{\sqrt{T}}\|\tilde{F} - F^0\tilde{H}\|_{\text{F}} &\leq \frac{\frac{1}{NT}\|\tilde{\Theta}\tilde{\Theta}' - \Theta^0\Theta^{0\prime}\|_{\text{op}}}{\min_{j\leq R^0}\min\{|\hat{s}_{j-1} - \tilde{s}_j|, |\tilde{s}_j - \hat{s}_{j+1}|\}} \\
&\leq \bar{c}\frac{1}{NT}\|\tilde{\Theta}\tilde{\Theta}' - \Theta^0\Theta^{0\prime}\|_{\text{op}} \leq \bar{c}(\gamma_1\sqrt{K_a} \vee \gamma_2).
\end{aligned}$$

Next we have

$$\begin{aligned}
\left\|\mathbb{P}_{\tilde{F}} - \mathbb{P}_{F^0}\right\|_{\text{F}} = \left\|\frac{1}{T}\tilde{F}\tilde{F}' - \mathbb{P}_{F^0}\right\|_{\text{F}} &\leq 2\bar{c}\left\|\frac{1}{\sqrt{T}}\tilde{F} - \frac{1}{\sqrt{T}}F^0\tilde{H}\right\|_{\text{F}} + \left\|\frac{F^0\tilde{H}\tilde{H}'F^{0\prime}}{T} - \mathbb{P}_{F^0}\right\|_{\text{F}} \\
&\leq \bar{c}(\gamma_1\sqrt{K_a} \vee \gamma_2),
\end{aligned}$$

where the second equality is from the fact that $\tilde{H}\tilde{H}' = S_F^{-1/2}vv'S_F^{-1/2} = S_F^{-1}$. 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* $\tilde{\Sigma} \equiv T^{-1}\mathbf{X}'\mathbf{X} - T^{-2}\mathbf{X}'\tilde{F}\tilde{F}'\mathbf{X}$. *Then there exist some constants* $\underline{c}$, $\bar{c}$ *and* $\bar{c}'$ *such that with probability larger than* $1 - \bar{c}'[p^2N^2T^{1-q/4}(\log N)^{-q/2} + pNe^{-\underline{c}T} + p^2N^{2-\underline{c}\log N}]$ *we have*

(i) $\|\tilde{H}\|_{max} \leq \|\tilde{H}\|_\infty \leq \bar{c}$ *and* $\|\tilde{H}^{-1}\|_F \leq \bar{c}$;

(ii) $\max_{1\leq j\leq pN}|\mathbf{X}_{*,j}|/\sqrt{T} < \bar{c}$ *and* $\max_{1\leq j\leq N}|\mathbf{U}_{*,j}|/\sqrt{T} < \bar{c}$;

(iii) $\|F^{0\prime}\mathbf{U}\|_{max}/T \leq \log N \cdot T^{-1/2}/(16\bar{c}^2)$ *and* $\left\|T^{-1}\mathbf{X}'F^0 - \Sigma_{XF}\right\|_{max} \leq \bar{c}T^{-1/2}\log N$;

(iv) $\|\tilde{\Sigma} - \Sigma\|_{max} \leq \gamma_3$;

[13] Write $a \asymp b$ to denote that both a/b and b/a are stochastically bounded.

(v) Suppose $16K_J\gamma_3 \leq \psi_{\min}(\Sigma)/2$. *Then* $\tilde{\Sigma}$ *satisfies the restricted eigenvalue condition for* K_J *in* (3.3) *and* $\kappa_{\tilde{\Sigma}}(K_J) \geq \psi_{\min}(\Sigma)/2$.

Proof of Theorem 3.3. Fix $\bar{c}$ as in Lemma A.4. In this proof, we choose a large enough constant c_3 such that $\gamma_3 = c_3(\gamma_1\sqrt{K_a} \vee \gamma_2)$ with $c_3 \geq 2 \vee (16\bar{c}^2) \vee (16\bar{c}^4)$. Let $\mathcal{E}_{NT}^{(3)}$ be the joint events of

(1) $T^{-1}\left\|\mathbf{U}'\mathbf{X}\right\|_{\max} \leq \gamma_3/4$; (2) $\max_{1\leq j\leq pN}|\mathbf{X}_{*,j}|/\sqrt{T} \leq \bar{c}$;
(3) $\max_{1\leq j\leq N}|\mathbf{U}_{*,j}|/\sqrt{T} \leq \bar{c}$; (4) $\|\tilde{F} - F^0\tilde{H}\|_{\mathrm{F}}/\sqrt{T} \leq \gamma_3/(16\bar{c}^3)$;
(5) $\|F^{0\prime}\mathbf{U}\|_{\max}/T \leq \gamma_3/(16\bar{c}^2)$; (6) $\|\tilde{H}\|_\infty \vee \|\tilde{H}^{-1}\|_{\mathrm{F}} \leq \bar{c}$;
(7) $\hat{R} = R^0$;

and (8) $\tilde{\Sigma}$ satisfies the restricted eigenvalue condition for K_J in (3.3) with $\kappa_{\tilde{\Sigma}}(K_J) \geq \psi_{\min}(\Sigma)/2$. Under Assumptions A.1–A.3, by Lemmas A.2 and A.4, $\mathcal{E}_{NT}^{(3)}$ holds with probability larger than $1 - \varepsilon_{NT} - \bar{c}'[p^2N^2T^{1-q/4}(\log N)^{-q/2} + pNe^{-\underline{c}T} + p^2N^{2-\underline{c}\log N}]$. Conditional on the event $\mathcal{E}_{NT}^{(3)}$, we also have the event that

$$\begin{aligned}(9)\ T^{-1}\|\tilde{F}'\mathbf{U}\|_{\max} &\leq T^{-1}\|(\tilde{F} - F^0\tilde{H})'\mathbf{U}\|_{\max} + T^{-1}\|\tilde{H}'F^{0\prime}\mathbf{U}\|_{\max}\\ &\leq T^{-1}\|\tilde{F} - F^0\tilde{H}\|_{\mathrm{F}} \cdot \max_{1\leq j,N}\|\mathbf{U}_{*,j}\| + \|\tilde{H}'\|_\infty T^{-1}\|F^{0\prime}\mathbf{U}\|_{\max}\\ &\leq \gamma_3/(8\bar{c}),\end{aligned}$$

and that

$$\begin{aligned}(10)\ \max_{1\leq i\leq N} T^{-1/2}|\lambda_i^{0\prime}F^{0\prime}\mathbb{M}_{\tilde{F}}| &\leq \max_{1\leq i\leq N}|\lambda_i^0| \cdot T^{-1/2}\|(F^0 - \tilde{F}\tilde{H}^{-1})'\mathbb{M}_{\tilde{F}}\|_{\mathrm{F}}\\ &\leq \bar{c}T^{-1/2}\|\tilde{F} - F^0\tilde{H}\|_{\mathrm{F}}\left\|\tilde{H}^{-1}\right\|_{\mathrm{F}} \leq \gamma_3/(8\bar{c}).\end{aligned}$$

Conditional on the event $\mathcal{E}_{NT}^{(3)}$, we establish the bound of $|\dot{\Delta}_{*,i}|_1 \equiv |\dot{B}_{*,i} - B^0_{*,i}|_1$ for $i \in [N]$.

Step 1. Concentrate out λ.

The objective function (3.2) is a least squares objective function with respect to λ. Given $\dot{B}_{*,i}$, we have that

$$\dot{\lambda}_i = (\tilde{F}'\tilde{F})^{-1}\tilde{F}'(\mathbf{Y}_{*,i} - \mathbf{X}\dot{B}_{*,i}) = T^{-1}\tilde{F}'(\mathbf{Y}_{*,i} - \mathbf{X}\dot{B}_{*,i}),$$

where the second equality holds by the identity $\tilde{F}'\tilde{F}/T = I_T$. After concentrating out λ_i, the optimization problem becomes

$$\dot{B}_{*,i} = \underset{v\in\mathbb{R}^{Np}}{\operatorname{argmin}}\ \frac{1}{2T}\|\mathbb{M}_{\tilde{F}}(\mathbf{Y}_{*,i} - \mathbf{X}v)\|_F^2 + \gamma_3|v|_1, \tag{A.9}$$

where $\mathbb{M}_{\tilde{F}} = I_T - \tilde{F}\tilde{F}/T$.

Step 2. Compare the objective functions at $\dot{B}_{*,i}$ and $B^0_{*,i}$.

By the identity $\mathbf{Y}_{*,i} = \mathbf{X}B^0_{*,i} + F^0\lambda_i^0 + \mathbf{U}_{*,i}$ and the definition of $\dot{B}_{*,i}$, we have

$$\begin{aligned}0 &\geq \frac{1}{2T}[\|\mathbb{M}_{\tilde{F}}(\mathbf{Y}_{*,i} - \mathbf{X}\dot{B}_{*,i})\|_{\mathrm{F}}^2 - \|\mathbb{M}_{\tilde{F}}(F^0\lambda_i^0 + \mathbf{U}_{*,i})\|_{\mathrm{F}}^2] + \gamma_3(|\dot{B}_{*,i}|_1 - |B^0_{*,i}|_1)\\ &= \frac{1}{2T}\|\mathbb{M}_{\tilde{F}}\mathbf{X}\dot{\Delta}_{*,i}\|_{\mathrm{F}}^2 - \frac{1}{T}\mathrm{tr}[(F^0\lambda_i^0 + \mathbf{U}_{*,i})'\mathbb{M}_{\tilde{F}}\mathbf{X}\dot{\Delta}_{*,i}] + \gamma_3(|\dot{B}_{*,i}|_1 - |B^0_{*,i}|_1),\end{aligned}$$

where $\dot{\Delta} \equiv \dot{B} - B^0$ and $\dot{\Delta}_{*,i}$ denotes the ith column of $\dot{\Delta}$. Then by Lemma A.1(v), we have

$$\begin{aligned}\frac{1}{T}\|(F^0\lambda_i^0 + \mathbf{U}_{*,i})'\mathbb{M}_{\tilde{F}}\mathbf{X}\|_{\max}|\dot{\Delta}_{*,i}|_1 &\geq \frac{1}{T}\mathrm{tr}[(F^0\lambda_i^0 + \mathbf{U}_{*,i})'\mathbb{M}_{\tilde{F}}\mathbf{X}\dot{\Delta}_{*,i}]\\ &\geq \frac{1}{2T}\|\mathbb{M}_{\tilde{F}}\mathbf{X}\dot{\Delta}_{*,i}\|_F^2 + \gamma_3(|\dot{B}_{*,i}|_1 - |B^0_{*,i}|_1)\\ &\geq \frac{1}{2T}\|\mathbb{M}_{\tilde{F}}\mathbf{X}\dot{\Delta}_{*,i}\|_F^2 + \gamma_3|\dot{\Delta}_{J_i^c,i}|_1 - \gamma_3|\dot{\Delta}_{J_i,i}|_1,\end{aligned}$$

where the last inequality follows because

$$\begin{aligned}|\dot{B}_{*,i}|_1 - |B^0_{*,i}|_1 &= |\dot{\Delta}_{*,i} + B^0_{*,i}|_1 - |B^0_{*,i}|_1 = |\dot{\Delta}^0_{J_i^c,i}|_1 + |\dot{\Delta}_{J_i,i} + B^0_{*,i}|_1 - |B^0_{*,i}|_1\\ &\geq |\dot{\Delta}_{J_i^c,i}|_1 - |\dot{\Delta}_{J_i,i}|_1.\end{aligned}$$

Step 3. Bound $T^{-1}\max_i[\|(F^0\lambda_i^0 + U_{*,i})'\mathbb{M}_{\tilde{F}}X\|_{\max}]$, conditional on the event $\mathcal{E}_{NT}^{(3)}$.

By the triangle and Cauchy Schwartz inequalities and the fact that $T^{-1/2}\|\tilde{F}\|_{\mathrm{op}} = 1$, we have

$$\begin{aligned}
& T^{-1}\|(F^0\lambda_i^0 + \mathbf{U}_{*,i})'\mathbb{M}_{\tilde{F}}\mathbf{X}\|_{\max} \\
\leq\ & T^{-1}\|\lambda_i^{0\prime}F^{0\prime}\mathbb{M}_{\tilde{F}}\mathbf{X}\|_{\max} + T^{-1}\|\mathbf{U}_{*,i}'\mathbb{M}_{\tilde{F}}\mathbf{X}\|_{\max} \\
\leq\ & \max_{1\leq j\leq Np} T^{-1/2}|\mathbf{X}_{*,j}| \cdot T^{-1/2}|\lambda_i^{0\prime}F^{0\prime}\mathbb{M}_{\tilde{F}}| + \max_{1\leq j\leq Np} T^{-1}|\mathbf{U}_{*,i}'\mathbf{X}_{*,j}| + T^{-2}\|\mathbf{U}_{*,i}'\tilde{F}\tilde{F}'\mathbf{X}\|_{\max} \\
\leq\ & \max_{1\leq j\leq Np} T^{-1}|\mathbf{U}_{*,i}'\mathbf{X}_{*,j}| + \left\{T^{-1}|\mathbf{U}_{*,i}'\tilde{F}| + T^{-1/2}|\lambda_i^{0\prime}F^{0\prime}\mathbb{M}_{\tilde{F}}|\right\} \max_{1\leq j\leq Np} T^{-1/2}|\mathbf{X}_{*,j}|.
\end{aligned}$$

Combining events (1), (9) and (10), the right hand side of the above inequality is bounded by $\gamma_3/2$ conditional on the event $\mathcal{E}_{NT}^{(3)}$.

Step 4. Obtain the final bound for $|\dot{B}_{*,i} - B_{*,i}^0|_1$.

Combining the results in Steps 2–3 and using the identity $|\dot{\Delta}_{*,i}|_1 = |\dot{\Delta}_{J_i,i}|_1 + |\dot{\Delta}_{J_i^c,i}|_1$, we have that conditional on the event $\mathcal{E}_{NT}^{(3)}$,

$$3\gamma_3|\dot{\Delta}_{J_i,i}|_1 \geq \frac{1}{T}\|\mathbb{M}_{\tilde{F}}\mathbf{X}\dot{\Delta}_{*,i}\|_{\mathrm{F}}^2 + \gamma_3|\dot{\Delta}_{J_i^c,i}|_1.$$

It follows that $|\dot{\Delta}_{J_i^c,i}|_1 \leq 3|\dot{\Delta}_{J_i,i}|_1$ and conditional on $\mathcal{E}_{NT}^{(3)}$,

$$\dot{\Delta}_{*,i}'\tilde{\Sigma}\dot{\Delta}_{*,i} \leq 3\gamma_3|\dot{\Delta}_{J_i,i}|_1 \leq 3\gamma_3\sqrt{k_i}|\dot{\Delta}_{J_i,i}| \leq \frac{6\sqrt{k_i}}{\psi_{\min}(\Sigma)}\gamma_3\sqrt{\dot{\Delta}_{*,i}'\tilde{\Sigma}\dot{\Delta}_{*,i},}$$

where the last inequality holds by event (8) in $\mathcal{E}_{NT}^{(3)}$. It follows that $\sqrt{\dot{\Delta}_{*,i}'\tilde{\Sigma}\dot{\Delta}_{*,i},} \leq \frac{6\sqrt{k_i}}{\psi_{\min}(\Sigma)}\gamma_3$ and $|\dot{\Delta}_{J_i,i}|_1 \leq \frac{2\sqrt{k_i}}{\psi_{\min}(\Sigma)}\sqrt{\dot{\Delta}_{*,i}\tilde{\Sigma}\dot{\Delta}_{*,i},} \leq \frac{12k_i}{(\psi_{\min}(\Sigma))^2}\gamma_3$. Consequently, we have established that

$$|\dot{\Delta}_{*,i}|_1 = |\dot{\Delta}_{J_i,i}|_1 + |\dot{\Delta}_{J_i^c,i}|_1 \leq 4|\dot{\Delta}_{J_i,i}|_1 \leq \frac{48}{(\psi_{\min}(\Sigma))^2}k_i\gamma_3.$$

Then the conclusion in Theorem 3.3 follows. ■

A.3. Theoretical analysis of the third-step estimators

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

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

(i) $\psi_{\min}(\tilde{\Sigma}_{J_i J_i}) \geq \underline{c}$;

(ii) $\|\tilde{\Sigma}_{J_i^c,J_i}\|_{max} \leq \bar{c}$ and $\psi_{\max}(\tilde{\Sigma}_{J_i^c,J_i}) \leq \bar{c}k_i$,

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

Proof of Theorem 3.4. For any n-dimensional vector $v = (v_1, \ldots, v_n)'$, denote $\mathrm{abs}(v) = (|v_1|, \ldots, |v_n|)'$. We say that $v < \tilde{v}$ if and only if $v_i < \tilde{v}_i'$ for all $i \in [n]$. Let $W^{(i)} = \mathrm{diag}(w_{1i}, \ldots, w_{Np,i})$, $W^{(1,i)} = W_{J_i J_i}^{(i)}$ and $W^{(0,i)} = W_{J_i^c J_i^c}^{(i)}$. The following proof is by induction. Based on the error bounds for $\hat{F}^{(\ell)}$'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 $\|\hat{F}^{(0)} - F^0\tilde{H}\|_{\mathrm{F}}/\sqrt{T} = O_P(\gamma_1\sqrt{K_a} + \gamma_2)$.

For notational simplicity, let $\tilde{\Sigma}^{(\ell)}$ denote $T^{-1}\mathbf{X}'\mathbb{M}_{\hat{F}^{(\ell)}}\mathbf{X}$ for $\ell = 0, 1, 2, \ldots$

(i) For all (k, i)'s such that $B_{ki}^0 = 0$, $\sup_{(k,i):B_{ki}^0=0}|\dot{B}_{ki}| \leq \|\dot{B} - B^0\|_{\max} \leq O_P(K_J\gamma_3) = o_P(\gamma_4)$. It follows that $W^{(0,i)} = I_{|J_i^c|}$ with w.p.a.1. For all (k, i)'s such that $B_{ki}^0 \neq 0$,

$$\min_{k,i:B_{ki}^0\neq 0}|\dot{B}_{ki}| \geq \min_{i\in[N]}\min_{k\in J_i}|B_{ki}^0| - \|\dot{B} - B^0\|_{\max} = \min_{i\in[N]}\min_{k\in J_i}|B_{ki}^0| - o_P(\gamma_4) \geq \alpha\gamma_4 \text{ w.p.a.1}$$

by Assumption A.5(i). It follows that $W^{(1,i)} = \mathbf{0}$ w.p.a.1. For each $i \in [N]$, the estimator $\hat{B}_{*,i}^{(\ell)}$ can be written as

$$\hat{B}_{*,i}^{(\ell)} = \operatorname*{argmin}_{v\in\mathbb{R}^{NP}} \mathcal{L}^{(i)}(v, \hat{F}^{(\ell-1)}),$$

where $\mathcal{L}^{(i)}(v, F) \equiv \frac{1}{2T}(\mathbf{Y}_{*,i} - \mathbf{X}v)'\mathbb{M}_{\hat{F}^{(\ell-1)}}(\mathbf{Y}_{*,i} - \mathbf{X}v) + \gamma_4\sum_{k=1}^{pN} w_{ki}|v_k|$. Following the proof of Proposition 1 of Zhao and Yu (2006), $\mathrm{sgn}(\hat{B}_{*,i}^{(l)})$ $=\mathrm{sgn}(B_{*,i}^0)$ is implied by event $\mathcal{E}_{i,1} \cap \mathcal{E}_{i,2}$, where

$$\mathcal{E}_{i,1} \equiv \left\{\mathrm{abs}[T^{-1/2}\tilde{\Sigma}_{J_i,J_i}^{-1}\mathbf{X}_{*,J_i}'\mathbb{M}_{\hat{F}^{(\ell-1)}}(\mathbf{U}_{*,i} + F^0\lambda_i^0)] < T^{1/2}\mathrm{abs}(B_{J_i,i}^0) - T^{1/2}\gamma_4\mathrm{abs}[\tilde{\Sigma}_{J_i,J_i}^{-1}W^{(1,i)}\mathrm{sgn}(B_{J_i,i}^0)]\right\}$$

and

$$\begin{aligned}\mathcal{E}_{i,2} &\equiv \{\mathrm{abs}[T^{-1/2}(-\tilde{\Sigma}_{J_i^c,J_i}\tilde{\Sigma}_{J_i,J_i}^{-1}\cdot\mathbf{X}_{*,J_i}'+\mathbf{X}_{*,J_i^c}')\mathbb{M}_{\hat{F}^{(\ell-1)}}(\mathbf{U}_{*,i}+F^0\lambda_i^0)]\\&\le T^{1/2}\gamma_4 W^{(0,i)}\cdot\iota_{|J_i^c|}-T^{1/2}\gamma_4\mathrm{abs}[\tilde{\Sigma}_{J_i^c,J_i}\tilde{\Sigma}_{J_i,J_i}^{-1}W^{(1,i)}\mathrm{sgn}(B_{J_i,i}^0)]\}.\end{aligned}$$

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

First, we consider $\mathcal{E}_{i,1}$. It suffices to show that each entry of $T^{-1/2}\mathrm{abs}[\tilde{\Sigma}_{J_i,J_i}^{-1}\mathbf{X}_{*,J_i}'\mathbb{M}_{\hat{F}^{(\ell-1)}}(\mathbf{U}_{*,i}+F^0\lambda_i^0)]$ is o_P $(\sqrt{T}\min_i\min_{k\in J_i}|B_{ki}^0|)$. Applying the triangle inequality, one has

$$\begin{aligned}&T^{-1/2}\mathrm{abs}[\tilde{\Sigma}_{J_i,J_i}^{-1}\mathbf{X}_{*,J_i}'\mathbb{M}_{\hat{F}^{(\ell-1)}}(\mathbf{U}_{*,i}+F^0\lambda_i^0)]\\&\le T^{-1/2}\mathrm{abs}(\tilde{\Sigma}_{J_i,J_i}^{-1}\mathbf{X}_{*,J_i}'\mathbb{M}_{\hat{F}^{(\ell-1)}}\mathbf{U}_{*,i})+T^{-1/2}\mathrm{abs}(\tilde{\Sigma}_{J_i,J_i}^{-1}\mathbf{X}_{*,J_i}'\mathbb{M}_{\hat{F}^{(\ell-1)}}F^0\lambda_i^0)\\&\le T^{-1/2}\mathrm{abs}(\tilde{\Sigma}_{J_i,J_i}^{-1}\mathbf{X}_{*,J_i}'\mathbb{M}_{F^0}\mathbf{U}_{*,i})+T^{-1/2}\mathrm{abs}[\tilde{\Sigma}_{J_i,J_i}^{-1}\mathbf{X}_{*,J_i}'(\mathbb{P}_{F^0}-\mathbb{P}_{\hat{F}^{(\ell-1)}})\mathbf{U}_{*,i}]\\&\quad+T^{-1/2}\mathrm{abs}[\tilde{\Sigma}_{J_i,J_i}^{-1}\mathbf{X}_{*,J_i}'\mathbb{M}_{\hat{F}^{(\ell-1)}}(\hat{F}^{(\ell-1)}-F^0\tilde{H})\tilde{H}^{-1}\lambda_i^0].\end{aligned}\tag{A.10}$$

Note that $\max_i\|\tilde{\Sigma}_{J_i,J_i}^{-1}\|_{\mathrm{op}}\le\bar{c}$ w.p.a.1 by Lemma A.5(i). This, in conjunction with Lemma A.2(i)–(ii), implies that the first term on the right hand side of (RHS) of (A.10) is uniformly $O_P(\log N)$. With $\|\hat{F}^{(\ell-1)}-F^0\tilde{H}\|_{\mathrm{F}}/\sqrt{T}=O_P(\gamma_1\sqrt{K_a}+\gamma_2)=O_P((\log N)T^{-1/2}\sqrt{K_a}+N^{-1/2})$,[14] we have $\|\mathbb{P}_{F^0}-\mathbb{P}_{\hat{F}^{(\ell-1)}}\|_{\mathrm{op}}=O_P((\log N)T^{-1/2}\sqrt{K_a}+N^{-1/2})$. Note that Lemma A.4(ii) ensures $\max_{1\le j\le pN}\|\mathbf{X}_{*,j}\|/\sqrt{T}$ and $\max_{1\le j\le N}\|\mathbf{U}_{*,j}\|/\sqrt{T}$ are both bounded by an absolute constant. It follows that each entry of the second term on the RHS of (A.10) is $O_P(\log N\cdot\sqrt{K_a}+\sqrt{T/N})$. Similarly, each entry of the third term on the RHS is $O_P(\log N\cdot\sqrt{K_a}+\sqrt{T/N})$. These results, along with the fact that $\log N\cdot T^{-1/2}\sqrt{K_a}=o(\min_i\min_{k\in J_i}|B_{ki}^0|)$ and $N^{-1/2}=o(\min_i\min_{k\in J_i}|B_{ki}^0|)$ in Assumption A.5 imply that $P(\mathcal{E}_{i,1})\to 1$.

Next, we consider $\mathcal{E}_{i,2}$. Similar to the analysis for $\mathcal{E}_{i,1}$, we can use Lemma A.5(ii) to show that each entry of $T^{-1/2}(-\tilde{\Sigma}_{J_i^c,J_i}\tilde{\Sigma}_{J_i,J_i}^{-1}\cdot\mathbf{X}_{*,J_i}'+\mathbf{X}_{*,J_i^c}')\mathbb{M}_{\hat{F}^{(\ell-1)}}(\mathbf{U}_{*,i}+F^0\lambda_i^0)$ is $O_P(K_J\log N\cdot\sqrt{K_a}+K_J\sqrt{T/N})=o(\sqrt{T}\gamma_3)$. By the fact that $\gamma_3=o(\gamma_4)$, we have $P(\mathcal{E}_{i,2})\to 1$, as $(N,T)\to\infty$.

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

$$\tilde{\Sigma}_{J_iJ_i}(\hat{B}_{J_i,i}^{(\ell)}-B_{J_i,i}^0)=\frac{1}{T}\mathbf{X}_{*,J_i}'\mathbb{M}_{\hat{F}^{(\ell-1)}}(F^0\lambda_i^0+\mathbf{U}_{*,i}),$$

for $i\in[N]$. Then

$$\begin{aligned}\left|\hat{B}_{J_i,i}^{(\ell)}-B_{J_i,i}^0\right|&=\left|\tilde{\Sigma}_{J_iJ_i}^{-1}\frac{1}{T}\mathbf{X}_{*,J_i}'\mathbb{M}_{\hat{F}^{(\ell-1)}}(F^0\lambda_i^0+\mathbf{U}_{*,i})\right|\\&\le\underline{c}^{-1}\left|\frac{1}{T}\mathbf{X}_{*,J_i}'\mathbb{M}_{\hat{F}^{(\ell-1)}}F^0\lambda_i^0\right|+\underline{c}^{-1}\left|\frac{1}{T}\mathbf{X}_{*,J_i}'\mathbb{M}_{\hat{F}^{(\ell-1)}}\mathbf{U}_{*,i}\right|\equiv\underline{c}^{-1}(A_{1i}+A_{2i}),\end{aligned}$$

where we use the fact that $\max_i\left\|\tilde{\Sigma}_{J_iJ_i}^{-1}\right\|_{\mathrm{op}}\le\underline{c}^{-1}$ w.p.a.1 by Lemma A.5(i). Note that uniformly in $i\in[N]$,

$$\begin{aligned}A_{1i}^2&=\frac{1}{T^2}\left|\mathbf{X}_{*,J_i}'\mathbb{M}_{\hat{F}^{(\ell-1)}}(\hat{F}^{(\ell-1)}\tilde{H}^{-1}-F^0)\lambda_i^0\right|^2\\&=\frac{1}{T^2}\mathrm{tr}\left(\lambda_i^{0\prime}(\hat{F}^{(\ell-1)}\tilde{H}^{-1}-F^0)'\mathbb{M}_{\hat{F}^{(\ell-1)}}\mathbf{X}_{*J_i}\mathbf{X}_{*J_i}'\mathbb{M}_{\hat{F}^{(\ell-1)}}(\hat{F}^{(\ell-1)}\tilde{H}^{-1}-F^0)\lambda_i^0\right)\\&\le\psi_{\max}\left(\frac{1}{T}\mathbb{M}_{\hat{F}^{(\ell-1)}}\mathbf{X}_{*,J_i}\mathbf{X}_{*,J_i}'\mathbb{M}_{\hat{F}^{(\ell-1)}}\right)\frac{1}{T}\left\|\hat{F}^{(\ell-1)}\tilde{H}^{-1}-F^0\right\|^2\left\|\lambda_i^0\right\|^2\\&=\psi_{\max}\left(\frac{1}{T}\mathbf{X}_{*,J_i}'\mathbb{M}_{\hat{F}^{(\ell-1)}}\mathbf{X}_{*,J_i}\right)\frac{1}{T}\left\|\hat{F}^{(\ell-1)}\tilde{H}^{-1}-F^0\right\|^2\left\|\lambda_i^0\right\|^2\\&\le\bar{c}k_i\frac{1}{T}\left\|\hat{F}^{(\ell-1)}\tilde{H}^{-1}-F^0\right\|^2=k_i\cdot O_P[(\gamma_1+\gamma_2)^2],\end{aligned}$$

and

$$A_{2i}^2=\left|\frac{1}{T}\mathbf{X}_{*,J_i}'\mathbb{M}_{\hat{F}^{(\ell-1)}}\mathbf{U}_{*,i}\right|^2\le 2\left|\frac{1}{T}\mathbf{X}_{*,J_i}'\mathbf{U}_{*,i}\right|^2+2\left|\frac{1}{T}\mathbf{X}_{*,J_i}'\hat{F}^{(\ell-1)}\frac{1}{T}\hat{F}^{(\ell-1)\prime}\mathbf{U}_{*,i}\right|^2.$$

[14] This claim holds for $\ell=1$ by Theorem 3.2. Given this claim, we can show that $\|\hat{F}^{(\ell)}-F^0\tilde{H}\|_{\mathrm{F}}/\sqrt{T}=O_P((\mathrm{log}N)T^{-1/2}\sqrt{K_a}+N^{-1/2})$ for each ℓ using the results below.

It is standard to show that $\left|\frac{1}{T}\mathbf{X}'_{*,J_i}\mathbf{U}_{*,i}\right| \leq k_i^{1/2} O_P(T^{-1/2}\log N)$ uniformly in i. In addition,

$$
\begin{aligned}
\left|\frac{1}{T}\mathbf{X}'_{*,J_i}\hat{F}^{(\ell-1)}\frac{1}{T}\hat{F}^{(\ell-1)\prime}\mathbf{U}_{*,i}\right|^2 &= \operatorname{tr}\left(\frac{1}{T^2}\hat{F}^{(\ell-1)\prime}\mathbf{X}_{*J_i}\mathbf{X}'_{*,J_i}\hat{F}^{(\ell-1)}\frac{1}{T^2}\hat{F}^{(\ell-1)\prime}\mathbf{U}_{*,i}\mathbf{U}'_{*,i}\hat{F}^{(\ell-1)}\right) \\
&\leq \psi_{\max}\left(\frac{1}{T^2}\hat{F}^{(\ell-1)\prime}\mathbf{X}_{*J_i}\mathbf{X}'_{*,J_i}\hat{F}^{(\ell-1)}\right)\frac{1}{T^2}\left|\hat{F}^{(\ell-1)\prime}\mathbf{U}_{*,i}\right|^2 \\
&\leq \psi_{\max}\left(\frac{1}{T}\mathbf{X}_{*,J_i}\mathbf{X}'_{*,J_i}\right)\frac{1}{T^2}\left|\hat{F}^{(\ell-1)\prime}\mathbf{U}_{*,i}\right|^2 \\
&= k_i \cdot O_{\mathrm{P}}[(\gamma_1+\gamma_2)^2] \text{ uniformly in } i,
\end{aligned}
$$

where the last equality follows from the fact $\psi_{\max}(\frac{1}{T}\mathbf{X}_{*,J_i}\mathbf{X}'_{*,J_i}) \leq \bar{c}$ w.p.a.1 and $\max_i \frac{1}{T}|\hat{F}^{(\ell-1)\prime}\mathbf{U}_{*,i}| = O_P(\gamma_1+\gamma_2)$ by similar arguments as used to obtain event (9) in the proof of Theorem 3.3. Then uniformly in $i \in [N]$, we have $A_{2i}^2 \leq k_i \cdot O_{\mathrm{P}}[(\gamma_1+\gamma_2)^2]$ and

$$
\left|\hat{B}^{(\ell)}_{J_i,i} - B^0_{J_i,i}\right|^2 \leq k_i \cdot O_{\mathrm{P}}[(\gamma_1+\gamma_2)^2].
$$

It follows that

$$
\begin{aligned}
\frac{\|\mathbf{X}(\hat{B}^{(\ell)}-B^0)\|_{\mathrm{F}}^2}{NT} &= \frac{1}{N}\sum_{i=1}^N \frac{\left|\mathbf{X}(\hat{B}^{(\ell)}_{*,i}-B^0_{*,i})\right|^2}{T} = \frac{1}{N}\sum_{i=1}^N \frac{1}{T}(\hat{B}^{(\ell)}_{J_i,i}-B^0_{J_i,i})'\mathbf{X}'_{*,J_i}\mathbf{X}_{*J_i}(\hat{B}^{(\ell)}_{J_i,i}-B^0_{J_i,i})' \\
&\leq \frac{1}{N}\sum_{i=1}^N \left|\hat{B}^{(\ell)}_{J_i,i}-B^0_{J_i,i}\right|^2 \left\|\frac{1}{T}\mathbf{X}'_{*,J_i}\mathbf{X}_{*J_i}\right\|_{\mathrm{op}} \\
&= \frac{1}{N}\sum_{i=1}^N k_i^2 \cdot O_{\mathrm{P}}[(\gamma_1+\gamma_2)^2].
\end{aligned}
$$

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

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

$$
\begin{aligned}
\frac{1}{\sqrt{NT}}\left\|\mathbf{U}-\mathbf{X}(\hat{B}^{(\ell)}-B^0)\right\|_{\mathrm{op}} &\leq \frac{1}{\sqrt{NT}}\|\mathbf{U}\|_{\mathrm{op}} + \frac{1}{\sqrt{NT}}\left\|\mathbf{X}(\hat{B}^{(\ell)}-B^0)\right\|_{\mathrm{op}} \\
&\leq O_P(\gamma_2) + O_P(\gamma_1+\gamma_2) = O_P(\gamma_1+\gamma_2).
\end{aligned}
$$

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

Proof of Theorem 3.5. Let $\hat{\Sigma} = \mathbf{X}'\mathbb{M}_{\hat{F}}\mathbf{X}/T$. From the proof of Theorem 3.4, we have

$$
\hat{\Sigma}_{J_iJ_i}(\hat{B}_{J_i,i} - B^0_{J_i,i}) = \frac{1}{T}\mathbf{X}'_{*,J_i}\mathbb{M}_{\hat{F}}F^0\lambda_i^0 + \frac{1}{T}\mathbf{X}'_{*,J_i}\mathbb{M}_{\hat{F}}\mathbf{U}_{*,i} - \gamma_4 W^{(1,i)}\operatorname{sgn}(B^0_{J_i,i}). \tag{A.11}
$$

By Theorem 3.3 and Assumption A.5(i), $\max_{k\in J_i} w_{ki} = 0$ w.p.a.1, which implies that $\gamma_4 W^{(1,i)}\operatorname{sgn}(B^0_{J_i,i}) = o_p(T^{-1/2})$.

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

$$
\begin{aligned}
\hat{F}V_{NT} &= \frac{1}{NT}\left(\mathbf{Y}-\mathbf{X}\hat{B}^{(\ell^*-1)}\right)\left(\mathbf{Y}-\mathbf{X}\hat{B}^{(\ell^*-1)}\right)'\hat{F} \\
&= \frac{1}{NT}\sum_{i=1}^N \left(\mathbf{Y}_{*,i}-\mathbf{X}_{*J_i}\hat{B}^{(\ell^*-1)}_{J_i,i}\right)\left(\mathbf{Y}_{*,i}-\mathbf{X}_{*J_i}\hat{B}^{(\ell^*-1)}_{J_i,i}\right)'\hat{F},
\end{aligned}
$$

where V_{NT} is a diagonal matrix that consists of the $\hat{R}$ largest eigenvalues of the matrix $T\times T$ matrix $(NT)^{-1}(\mathbf{Y}-\mathbf{X}\hat{B}^{(\ell^*-1)})(\mathbf{Y}-\mathbf{X}\hat{B}^{(\ell^*-1)})'$, arranged in descending order along its diagonal line. One can use a similar expansion as in the Proof of Proposition S1.1, to study $\mathbb{M}_{\hat{F}}$. The estimation error of $\hat{B}^{(\ell)}$ depends on ℓ and the error of $\hat{B}^{(1)}$, but part that is due to $\hat{B}^{(1)}$ will decay fast. After finite steps, we have that

$$
S_i\hat{\Sigma}_{J_iJ_i}(\hat{B}_{J_i,i} - B^0_{J_i,i}) = \frac{1}{T}S_i\mathbf{X}'_{*,J_i}\mathbb{M}_{F^0}\mathbf{U}_{*,i} + o_P(T^{-1/2}).
$$

By arguments as used in the proof of Lemma A.2, we can readily show that $\left\|\frac{1}{T}\mathbf{X}'_{*,J_i}\mathbb{M}_{F^0}\mathbf{X}_{*,J_i} - \Sigma_{J_iJ_i}\right\|_{\mathrm{F}} = O_P\left(K_J T^{-1/2}\log N\right)$ and $|\frac{1}{\sqrt{T}}\mathbf{X}'_{*,J_i}\mathbb{P}_{F^0}\mathbf{U}_{*,i} - \frac{1}{\sqrt{T}}(F^0\Sigma_F^{-1}[\Sigma_{FX}]_{J_i,*})'\mathbf{U}_{*,i}| = O_P(K_J^{1/2}T^{-1/2}\log N)$, where $[\Sigma_{FX}]_{J_i,*} = \frac{1}{T}E\left[F^{0\prime}\mathbf{X}_{*,J_i}\right]$ is a $R^0 \times k_i$ matrix.

It follows that

$$\sqrt{T}S_i(\hat{B}_{J_i,i} - B^0_{J_i,i}) = \frac{1}{\sqrt{T}} S_i(\Sigma_{J_iJ_i})^{-1}(\mathbf{X}_{*,J_i} - F^0\Sigma_F^{-1}[\Sigma_{FX}]_{J_i,*})'\mathbf{U}_{*,i} + o_P(1)$$

$$\equiv T^{-1/2}\sum_{t=1}^{T} z^*_{it}u_{it} + o_P(1),$$

where $z^*_{it} = S_i(\Sigma_{J_iJ_i})^{-1}z^0_{it}$ and z^0_{it} denotes the tth column of the $k_i \times T$ matrix $(\mathbf{X}_{*,J_i} - F^0\Sigma_F^{-1}[\Sigma_{FX}]_{J_i,*})'$. Under Assumption A.1, $\{z^*_{it}u_{it}, t \geq 1\}$ is a martingale difference sequence (m.d.s.) and we can readily verify the conditions of the martingale central limit theorem by straightforward moment calculations and obtain $\sqrt{T}S_i(\check{B}_{J_i,i} - B^0_{J_i,i}) \xrightarrow{d} N(0, \sigma_i^2\Omega_i)$, where $\sigma_i^2 = E\left(u^2_{it}\right)$. ■

Appendix B. Nagaev inequality for time series

In various places, we need to a sharp probability bound for partial sums like $T^{-1}\sum_{t=1}^{T} y_{i,t-k}u_{jt}$, which is nonlinear in shocks $\{u_{jt}\}$ 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 $S_n = \sum_{i=1}^{n} a_ie_i$, where the $a_i \in \mathbb{R}$ are nonrandom, the scalar process $\{e_i\}$ has the form $e_i = g(..., \varepsilon_{i-1}, \varepsilon_i)$, where the ε_i are independently and identically distributed (i.i.d.) random variables, and $g(\cdot)$ is a measurable function. Letting $\mathcal{F}_i \equiv (..., \varepsilon_{i-1}, \varepsilon_i)$, we write $e_i = g(\mathcal{F}_i)$. Then a coupled process e^*_i can be defined as $e^*_i = g(\mathcal{F}^*_i)$, where $\mathcal{F}^*_i = (..., \varepsilon_{-1}, \varepsilon^*_0, \varepsilon_1, \ldots, \varepsilon_{i-1}, \varepsilon_i)$ and ε^*_0 is an independent copy of ε_0. Recall that $\|\cdot\|_q \equiv (E|\cdot|^q_q)^{1/q} < \infty$. Assuming that $\|e_i\|_q < \infty$ for some $q \geq 1$, we define the functional dependence measure

$$\delta_{i,q}(e.) \equiv \|e_i - e^*_i\|_q = \|g(\mathcal{F}_i) - g(\mathcal{F}^*_i)\|_q,$$

where $e^*_i = g(\mathcal{F}^*_i)$. The measure $\delta_{i,q}(e.)$ reflects the effect of shock ε_0 on e_i. Accordingly, we assume the cumulative effect of ε_0 on $\{e_i\}_{i\geq m}$ to be summable and given by

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

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

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

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 $a = (a_1, \ldots, a_n)'$ 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 $\|e.\|_{q,\alpha} < \infty$ for some $q > 2$ and $\alpha > 1$. Then for all $x > 0$,*

$$P(|S_n| > x) \leq C_1\frac{|a|^q_q\|e.\|^q_{q,\alpha}}{x^q} + C_2 exp\left(-\frac{C_3x^2}{n\|e.\|^2_{2,\alpha}}\right),$$

where C_1, C_2, C_3 are constants that only depend of q and α.

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

Ahn, S.C., Horenstein, A.R., 2013. Eigenvalue ratio test for the number of factors. Econometrica 81, 1203–1227.
Alizadeh, S., Brandt, M.W., Diebold, F.X., 2002. Range-based estimation of stochastic volatility models. J. Finance 57, 1047–1091.
Bai, J., 2003. Inferential theory for factor models of large dimensions. Econometrica 71, 135–171.
Bai, J., 2009. Panel data models with interactive fixed effects. Econometrica 77, 1229–1279.
Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. Econometrica 70, 191–221.
Bai, J., Ng, S., 2019. Rank regularized estimation of approximate factor models. J. Econometrics 212, 78–96.
Bai, Z., Saranadasa, H., 1996. Effect of high dimension: by an example of a two sample problem. Statist. Sinica 6, 311–329.
Barigozzi, M., Brownlees, C., 2019. Nets: Network estimation for time series. J. Appl. Econometrics 34, 347–364.
Barigozzi, M., Hallin, M., 2017. A network analysis of the volatility of high dimensional financial series. J. R. Stat. Soc. Ser. C. Appl. Stat. 66, 581–605.
Basu, S., Michailidis, G., 2015. Regularized estimation in sparse high-dimensional time series models. Ann. Statist. 43, 1535–1567.

Belloni, A., Chen, D., Chernozhukov, V., Hansen, C., 2012. Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica 80, 2369–2429.

Belloni, A., Chen, M., Padilla, O.H.M., 2019. High dimensional latent panel quantile regression with an application to asset pricing. arXiv preprint arXiv:1912.02151.

Bernanke, B.S., Boivin, J., Eliasz, P., 2005. Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach. Q. J. Econ. 120, 387–422.

Bickel, P.J., Ritov, Y., Tsybakov, A.B., 2009. Simultaneous analysis of Lasso and Dantzig selector. Ann. Statist. 37, 1705–1732.

Billio, M., Getmansky, M., Lo, A.W., Pelizzon, L., 2012. Econometric measures of connectedness and systemic risk in the finance and insurance sectors. J. Financ. Econ. 104, 535–559.

Candes, E., Tao, T., 2007. The Dantzig selector: Statistical estimation when p is much larger than n. Ann. Statist. 35, 2313–2351.

Caner, M., Kock, A.B., 2018. Asymptotically honest confidence regions for high dimensional parameters by the desparsified conservative LASSO. J. Econometrics 203, 143–168.

Chen, S., Qin, Y., 2010. A two-sample test for high-dimensional data with applications to gene-set testing. Ann. Statist. 38, 808–835.

Cheng, X., Hansen, B.E., 2015. Forecasting with factor-augmented regression: A frequentist model averaging approach. J. Econometrics 186, 280–293.

Chernozhukov, V., Hansen, C., Liao, Y., Zhu, Y., 2019. Inference for Heterogeneous Effects using Low-Rank Estimations. CEMMAP working paper.

Chudik, A., Pesaran, M.H., 2011. Infinite-dimensional VARs and factor models. J. Econometrics 163, 4–22.

Chudik, A., Pesaran, M.H., Tosetti, E., 2011. Weak and strong cross-section dependence and estimation of large panels. Econom. J. 14, C45–C90.

Demirer, M., Diebold, F.X., Liu, L., Yılmaz, K., 2018. Estimating global bank network connectedness. J. Appl. Econometrics 33, 1–15.

Diebold, F.X., Yılmaz, K., 2014. On the network topology of variance decompositions: Measuring the connectedness of financial firms. J. Econometrics 182, 119–134.

Diebold, F.X., Yılmaz, K., 2015. Financial and Macroeconomic Connectedness: A Network Approach to Measurement and Monitoring. Oxford University Press.

Fama, E.F., French, K.R., 1993. Common risk factors in the returns on stocks and bonds. J. Financ. Econ. 33, 3–56.

Fan, J., Gong, W., Zhu, Z.Z., 2019. Generalized high-dimensional trace regression via nuclear norm regularization. J. Econometrics 212, 177–202.

Fan, Y., Han, F., Park, H., 2020. Estimation and Inference on Granger Causality in a Latent High-Dimensional Gaussian Vector Autoregressive Model. Working Paper, Dept. of Economics, University of Washington.

Fan, J., Peng, H., 2004. Nonconcave penalized likelihood with a diverging number of parameters. Ann. Statist. 32, 928–961.

Feng, J., 2019. Regularized quantile regression with interactive fixed effects. arXiv preprint arXiv:1911.00166.

Forni, M., Hallin, M., Lippi, M., Reichlin, L., 2000. The generalized dynamic-factor model: Identification and estimation. Rev. Econ. Stat. 82, 540–554.

Garman, M.B., Klass, M.J., 1980. On the estimation of security price volatilities from historical data. J. Bus. 53, 67–78.

Geweke, J., 1977. The dynamic factor analysis of economic time series. In: Latent Variables in Socio-Economic Models.

Giannone, D., Reichlin, L., Sala, L., 2004. Monetary policy in real time. NBER Macroecon. Annu. 19, 161–200.

Guibert, Q., Lopez, O., Piette, P., 2019. Forecasting mortality rate improvements with a high-dimensional VAR. Insurance Math. Econom. 88, 255–272.

Guo, S., Wang, Y., Yao, Q., 2016. High-dimensional and banded vector autoregressions. Biometrika 103, 889–903.

Hallin, M., Liška, R., 2007. Determining the number of factors in the general dynamic factor model. J. Amer. Statist. Assoc. 102, 603–617.

Han, F., Lu, H., Liu, H., 2015. A direct estimation of high dimensional stationary vector autoregressions. J. Mach. Learn. Res. 16, 3115–3150.

Haufe, S., Müller, K.R., Nolte, G., Krämer, N., 2010. Sparse causal discovery in multivariate time series. In: Causality: Objectives and Assessment. pp. 97–106.

Hautsch, N., Schaumburg, J., Schienle, M., 2014. Financial network systemic risk contributions. Rev. Finance 19, 685–738.

Hecq, A., Margaritella, L., Smeekes, S., 2020. Granger causality testing in high-dimensional VARs: a post-double-selection procedure. arXiv. org.

Huang, J., Ma, S., Zhang, C.H., 2008. Adaptive lasso for sparse high-dimensional regression models. Statist. Sinica 1603–1618.

Hurn, S., Martin, V.L., Phillips, P.C.B., Yu, J., 2020. Financial Econometric Modelling. Oxford University Press.

Kock, A.B., Callot, L., 2015. Oracle inequalities for high dimensional vector autoregressions. J. Econometrics 186, 325–344.

Koltchinskii, V., Lounici, K., Tsybakov, A.B., 2011. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. Ann. Statist. 39, 2302–2329.

Koop, G., Pesaran, M.H., Potter, S.M., 1996. Impulse response analysis in nonlinear multivariate models. J. Econometrics 74, 119–147.

Lam, C., Fan, J., 2008. Profile-kernel likelihood inference with diverging number of parameters. Ann. Statist. 36, 2232–2260.

Leeper, E.M., Sims, C.A., Zha, T., Hall, R.E., Bernanke, B.S., 1996. What Does Monetary Policy Do? Brookings Papers on Economic Activity, 1996, pp. 1–78.

Lu, X., Su, L., 2016. Shrinkage estimation of dynamic panel data models with interactive fixed effects. J. Econometrics 190, 148–175.

Ludvigson, S., Ng, S., 2007. The empirical risk-return relation: a factor analysis approach. J. Financ. Econ. 83, 171–222.

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

Ma, S., Lan, W., Su, L., Tsai, C.-L., 2020. Testing alphas in conditional time-varying factor models with high dimensional assets. J. Bus. Econom. Statist. 38, 214–227.

Ma, S., Su, L., Zhang, Y., 2021. Detecting Latent Communities in Network Formation Model. Working Paper, Singapore Management University.

Mann, H.B., Wald, A., 1943. On the statistical treatment of linear stochastic difference equations. Econometrica 173–220.

Moon, H.R., Weidner, M., 2017. Dynamic linear panel regression models with interactive fixed effects. Econom. Theory 33, 158–195.

Moon, H.R., Weidner, M., 2019. Nuclear norm regularized estimation of panel regression models. arXiv:1810.10987.

Negahban, S.N., Ravikumar, P., Wainwright, M.J., Yu, B., 2012. A unified framework for high dimensional analysis of M-estimators with decomposable regularizers. Statist. Sci. 27, 538–557.

Negahban, S., Wainwright, M.J., 2011. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. Ann. Statist. 39, 1069–1097.

Negahban, S., Yu, B., Wainwright, M.J., Ravikumar, P., 2009. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. Adv. Neural Inf. Process. Syst. 22, 1348–1356.

Onatski, A., 2010. Determining the number of factors from empirical distribution of eigenvalues. Rev. Econ. Stat. 92, 1004–1016.

Pesaran, H.H., Shin, Y., 1998. Generalized impulse response analysis in linear multivariate models. Econom. Lett. 58, 17–29.

Qian, J., Su, L., 2016. Shrinkage estimation of regression models with multiple structural changes. Econom. Theory 32, 1376–1433.

Rohde, A., Tsybakov, A.B., 2011. Estimation of high-dimensional low-rank matrices. Ann. Statist. 39, 887–930.

Sims, C.A., 1980. Macroeconomics and reality. Econometrica 48, 1–48.

Sims, C.A., 1992. Interpreting the macroeconomic time series facts: The effects of monetary policy. Eur. Econ. Rev. 36, 975–1000.

Sims, C.A., 1993. A nine-variable probabilistic macroeconomic forecasting model. In: Business Cycles, Indicators and Forecasting. University of Chicago Press, pp. 179–212.

Stock, J.H., Watson, M.W., 1999. Forecasting inflation. J. Monetary Econ. 44, 293–335.

Stock, J.H., Watson, M.W., 2002. Macroeconomic forecasting using diffusion indexes. J. Bus. Econom. Statist. 20, 147–162.

Stock, J.H., Watson, M.W., 2005. Understanding changes in international business cycle dynamics. J. Eur. Econom. Assoc. 3, 968–1006.

Su, L., Wang, X., 2017. On time-varying factor models: Estimation and testing. J. Econometrics 198, 84–101.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58, 267–288.

Vershynin, R., 2018. High-Dimensional Probability: An Introduction with Applications in Data Science. Cambridge University Press.

Wu, W.B., 2005. Nonlinear system theory: Another look at dependence. Proc. Natl. Acad. Sci. 102, 14150–14154.

Wu, W.B., Wu, Y.N., 2016. Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. Electron. J. Stat. 10, 352–379.

Yu, Y., Wang, T., Samworth, R.J., 2014. A useful variant of the Davis–Kahan theorem for statisticians. Biometrika 102, 315–323.

Zhao, P., Yu, B., 2006. On model selection consistency of lasso. J. Mach. Learn. Res. 7, 2541–2563.

Zou, H., 2006. The adaptive lasso and its oracle properties. J. Amer. Statist. Assoc. 101, 1418–1429.