On Factor Models with Random Missing: EM Estimation, Inference, and Cross Validation *

Sainan Jin[a], Ke Miao[b], and Liangjun Su[c]
[a]School of Economics, Singapore Management University, Singapore.
[b]School of Economics, Fudan University, Shanghai, China
[c]School of Economics and Management, Tsinghua University, Beijing, China

August 16, 2020

Abstract

We consider the estimation and inference in approximate factor models with random missing values. We show that with the low rank structure of the common component, we can estimate the factors and factor loadings consistently with the missing values replaced by zeros. We establish the asymptotic distributions of the resulting estimators and those based on the EM algorithm. We also propose a cross-validation-based method to determine the number of factors in factor models with or without missing values and justify its consistency. Simulations demonstrate that our cross validation method is robust to fat tails in the error distribution and significantly outperforms some existing popular methods in terms of correct percentage in determining the number of factors. An application to the factor-augmented regression models shows that a proper treatment of the missing values can improve the out-of-sample forecast of some macroeconomic variables.

JEL Classification: C23, C33, C38; C55

Key Words: Cross-validation; Expectation-Maximization (EM) algorithm; Factor models; Matrix completion; Missing at random; Principal component analysis; Singular value decomposition

1 Introduction

Since the seminal work of Geweke (1977), Sargent and Sims (1977), Chamberlain and Rothschild (1983), factor models have been widely used in economics and finance. Some important theoretical contributions include Stock and Watson (1998), Forni et al. (2000), Bai and Ng (2002), Bai (2003), Hallin and Liśka (2007), Onatski (2009, 2010, 2012), and Ahn and Horenstein (2013), among others. Nevertheless, all these authors assume a balanced panel in their asymptotic analyses.

*The authors sincerely thank Serena Ng, an associate editor and an anonymous referee for many constructive comments on the paper. Su acknowledges the funding support provided by Tsinghua University. Address Correspondence to: Liangjun Su, School of Economics and Management, Tsinghua University, Beijing, 100084, China; E-mail: sulj@sem.tsinghua.edu.cn, Phone: +86 10 62789506.

Empirical data typically contain a variety of irregularities, including occasionally missing observations, unbalanced panel, and mixed frequency (e.g., monthly and quarterly) data. One simple way to handle missing data is to omit the cross-sectional units with missing values; see, e.g., Ludvigson and Ng (2007). But this will result in efficiency loss that can be substantial in some applications. To handle the missing data problem in factor models effectively, two methods have been proposed: the expectation-maximization (EM) algorithm and the Kalman filter (KF). These two methods have been widely used to handle missing data for principal component (PC) estimation with missing data and state space estimation with missing data. The details on how missing data are handled differ a lot in PC and state space applications. For the PC estimation with missing data, Stock and Watson (2002) propose an iterative method based on the EM algorithm that has proved to be easy and effective. Schumacher and Breitung (2008) apply Stock and Watson's methodology to nowcast German gross domestic product (GDP).

The state space framework has been adapted to missing data by either allowing the measurement equation to vary depending on what data are available at a given time point or keeping the dimension of the measurement equation to be the same over time by including a proxy value for the missing observation while adjusting the model parameters so that the Kalman filter places no weights on the missing observation. See Giannone et al. (2008), Mariano and Murasawa (2010), Doz et al. (2011), Jungbacker et al. (2011), Pinheiro et al. (2013), Bańbura and Modugno (2014), Bai and Li (2016) and Marcellino and Sivec (2016) for variations on this latter approach. In particular, Giannone et al. (2008) propose a two-step procedure that is able to solve the "ragged edge" problem in an approximate factor model when data are observed at different frequencies. They estimate the model by PC analysis with truncated balanced panel in the first step and update the estimates of factors by the KF with unbalanced panel data in the second step. Doz et al. (2011) show the consistency of the two-step estimators, and Bai and Li (2016) derive the asymptotic distribution of the Kalman filter estimator. Jungbacker et al. (2011) propose a new state space formulation of the factor model and apply the KF to estimate the underlying parameters with computational efficiency when the observations are missing at random. In view of the fact that it is not straightforward to apply Giannone et al.'s (2008) methodology to mixed frequency datasets with series of different lengths or, in general, to any pattern of missing data, Bańbura and Modugno (2014) propose a modified EM algorithm to allow for an arbitrary pattern of missing data where the KF is incorporated to estimate the factors in the maximization-step. A drawback of their approach is that for large cross-sections, the dimension of the augmented state vector becomes very large, which leads to computational inefficiency. Pinheiro et al. (2013) also propose an EM algorithm to estimate a dynamic factor model for panel data sets with jagged edge without significantly increasing the computation time relative to the balanced panel case. In addition, Foroni and Marcellino (2013) survey methods for handling mixed-frequency data, including dynamic factor models and alternative approaches; Stock and Watson (2016) summarize the advantage and disadvantage of the state space estimation for factor models with missing observations. During the revision, we found that Bai and Ng (2019b) also consider factor analysis with missing data and show that in spite of missing values in the data, every entry of the common component matrix can be consistently estimated using their tall-wide (TW)

algorithm that involves two applications of principal components.

Despite the popularity of the EM algorithm in empirical researches, the asymptotic properties of the resulting estimators have been rarely studied. To the best of our knowledge, there is no formal study of the asymptotic properties for the EM estimators of the factors and factor loadings for the PC estimation with missing observations. As Bai and Ng (2019b) remarked, "While convergence of the algorithm can be established, the asymptotic properties of the converged estimates are not well understood."

In this paper we consider the EM estimation of approximate factor models with missing observations. For simplicity, we focus on the case where the missing occurs at random and remark in the end on the other forms of missing. As Stock and Watson (2016) remark, all the procedures in common use adopt the assumption that the data are missing at random, that is, whether a datum is missing is independent of the latent variables, and the missing-at-random assumption arguably is a reasonable assumption for the main sources of missing data in dynamic factor models in most macroeconomic applications to date. In the case of random missing, we draw support from the literature on matrix completion in the computer science. It is well known that the low rank matrix such as the common component matrix in factor models can be recovered in the presence of missing observations when the noise matrix exhibit certain sparsity feature; see Cai et al. (2010), Candès and Plan (2010) and Candès and Li (2011). We show that similar phenomenon occurs when the noise matrix does not have any sparsity feature but has a lower order spectral norm than the common component matrix. In computation, we can simply replace the missing observations by zeros and conduct the usual PC analysis for a scaled version of the data matrix where the scale is determined by the percentage of observed values in the data. We show that the resulting estimators of factors, factor loadings, and common components are consistent but not asymptotically normal in general. Following the EM algorithm, we replace the missing observations by such initial estimators of the common components and obtain updated PC estimators. This procedure can be iterated until convergence. We show that the final estimators of the factors, factor loadings, and PCs are asymptotically more efficient than the initial estimators. We also characterize the efficiency loss for such EM estimators relative to the PC estimators without missing observations.[1]

In some sense, the pure approximate factor model possesses the "self-fulfilling" property in that one does not need to observe all values in the data matrix in order to estimate the factors, factor loadings and common components and the missing values can be well recovered from the observed data. Such a self-fulfilling property motivates us to propose a novel method to determine the number of pervasive factors in approximate factor models no matter whether the original data contains missing observations or not. Our key insight is that we can draw each observation at random with probability p to construct the pseudo-data matrix with missing values. The original data are then divided into two sets, with one set containing the training observations used for the PC estimation for any prescribed number of factors (say, R) and the other set containing the held-out entries used for the out-of-sample evaluation. Then we can construct a cross-validation (CV) objective function that

[1] Recently, Athey et al. (2018) have developed new methods for estimating causal effects in panel data models with missing values based on the matrix completion methods. But they do not provide any distribution or inference theory.

is indexed by R and choose R to minimize it. We show that this procedure consistently estimates the number of true factors. The finite sample performance of this procedure can be improved via iterations and some design for stability selection (e.g., Meinshausen and Bühlmann (2010)). Monte Carlo simulations indicate that our new estimator of the number of factors significantly outperforms some existing popular estimators including those based on either information criterion (Bai and Ng (2002)), or eigenvalue distribution function (Onatski (2010)), or eigenvalue/growth ratio (Ahn and Horenstein (2013)). Moreover, our simulations also demonstrate that our new estimators are robust to fat tails in the error terms.

The paper is organized as follows. Section 2 introduces the EM estimators of factor models with random missing and their asymptotic properties. Section 3 proposes a novel method to determine the number of factors in approximate factor models. In Section 4, we report the Monte Carlo simulation results for our EM estimators of the factors, factor loadings and common components, and compare our method of determining the number of factors with the methods of Bai and Ng (2002), Onatski (2010), and Ahn and Horenstein (2013). In Section 5, we apply our method to an empirical application and show that it helps the out-of-sample forecasts based on factor-augmented regressions. Final remarks are contained in Section 6. The proofs of the results in Sections 2 and 3 are relegated to Appendices A and B, respectively. The proofs of the technical lemmas and theorems in Appendices A and B along with some additional simulation results can be found in the online supplement.

NOTATION. For an $m \times n$ real matrix A, we denote its transpose as A', its entrywise L_∞ norm as $\|A\|_\infty (\equiv \max_{i,t} |A_{it}|)$, its Frobenius norm as $\|A\|$ $(\equiv [\text{tr}(AA')]^{1/2})$, its spectral norm as $\|A\|_{\text{sp}}$ $(\equiv \sqrt{\mu_1 (A'A)})$ and its Moore-Penrose generalized inverse as A^+, where $\equiv$ means "is defined as" and $\mu_s(\cdot)$ denotes the sth largest eigenvalue of a real symmetric matrix by counting eigenvalues of multiplicity multiple times. Note that the two norms are equal when A is a vector. We will frequently use the submultiplicative property of these norms and the fact that $\|A\|_{\text{sp}} \leq \|A\| \leq \|A\|_{\text{sp}} \text{rank}(A)^{1/2}$. We also use $\mu_{\max}(B)$ and $\mu_{\min}(B)$ to denote the largest and smallest eigenvalues of a symmetric matrix B, respectively. We use $B > 0$ to denote that B is positive definite. Let $P_A \equiv A(A'A)^+ A'$ and $M_A \equiv I_m - P_A$, where I_m denotes an $m \times m$ identity matrix. The operator $\xrightarrow{P}$ denotes convergence in probability, $\xrightarrow{d}$ convergence in distribution, and plim probability limit. Let $\vee$ and $\wedge$ denote the max and min operators, respectively. E.g., $N \vee T = \max(N, T)$. Let $[N] = \{1, 2, ..., N\}$ and $[T] = \{1, 2, ..., T\}$. We use $(N, T) \to \infty$ to denote that N and T pass to infinity jointly. We let $\delta_{NT} = \sqrt{N} \wedge \sqrt{T}$.

2 Large Dimensional Factor Models with Random Missing

In this section, we consider the PCA estimation of large dimensional factor models with observations that are missing at random by assuming the true number of factors is known. We will propose a novel cross validation method to determine the number of factors in the next section.

For simplicity and clarity, we shall work on the approximate factor model of Stock and Watson (2002), Bai and Ng (2002) and Bai (2003) when missing at random observations are present. In this

case, Stock and Watson (2002) propose an iterative method based on the EM algorithm that has proved to be easy and effective. Despite the popularity of the EM algorithm in empirical researches, there is no formal study of the asymptotic properties of the resulting estimators of the factors and factor loadings. Below we propose to obtain the initial estimators in the EM algorithm by replacing the missing values in the data matrix by zeros and show that one can derive the asymptotic distributions of such initial estimators and the resultant iterative estimators. As the iterative estimators converge to the EM estimators, the asymptotic properties of the EM estimators can be derived.

2.1 EM Estimation

We consider the following factor model

$$X_{it} = \lambda_i' F_t + \varepsilon_{it}, \tag{2.1}$$

where $i = 1, ..., N$, $t = 1, \ldots, T$, F_t and λ_i are $R \times 1$ vectors of factors and factor loadings, respectively, and ε_{it} is the idiosyncratic error term. Following the lead of Stock and Watson (2002) and Bai et al. (2015), we study the estimation of the factors and factor loadings when some of the observations, X_{it}, are missing at random. Let $X = (X_1, ..., X_N)$ and $\varepsilon = (\varepsilon_1, ..., \varepsilon_N)$, where $X_i \equiv (X_{i1}, \ldots, X_{iT})'$ and $\varepsilon_i \equiv (\varepsilon_{i1}, \ldots, \varepsilon_{iT})'$ for $i = 1, ..., N$. We can write (2.1) in matrix form:

$$X = F\Lambda' + \varepsilon \tag{2.2}$$

where $F = (F_1, ..., F_T)'$ and $\Lambda = (\lambda_1, ..., \lambda_N)'$. We will use $F^0 = \left(F_1^0, ..., F_T^0\right)'$ and $\Lambda^0 = (\lambda_1^0, ..., \lambda_N^0)'$ to denote the true values of F and Λ, respectively. Let $\Omega \subset [N] \times [T]$ be the index set of the observations that are observed. That is,

$$\Omega = \{(i,t) \in [N] \times [T] : X_{it} \text{ is observed}\}.$$

Let G denote a $T \times N$ matrix with (t,i)th element given by $g_{it} = \mathbf{1}\{(i,t) \in \Omega\}$. Under the random missing mechanism, g_{it}'s are independently and identically distributed as Bernoulli(q) with $q \in (0,1]$ and independent of X, F^0, Λ^0 and ε. So the population missing probability is given by $1-q \in [0,1)$. Let $|\Omega|$ denote the cardinality of the set Ω. It is easy to see that $\tilde{q} \equiv |\Omega| / (NT)$ is a $\sqrt{NT}$-consistent estimator of q.

2.1.1 The initial estimates

Let $\tilde{X} = X \circ G$ and $\tilde{X}_{it} = X_{it} g_{it}$, where $\circ$ denotes the Hadamard product. Our key observation is that the common component

$$C^0 \equiv F^0 \Lambda^{0\prime}$$

is a low rank matrix and ε is the noise component. In this case, it is possible to recover C^0 even when a large proportion of elements in the data matrix X are missing at random.

Let $E\left(\frac{1}{q}\tilde{X}|F^0, \Lambda^0\right)$ denote the $T \times N$ matrix with a typical element given by $E\left(\frac{1}{q}\tilde{X}_{it}|F_t^0, \lambda_i^0\right)$. Under the standard condition that $E\left(\varepsilon_{it}|F_t^0, \lambda_i^0\right) = 0$, we can readily verify that $E\left(\frac{1}{q}\tilde{X}|F^0, \Lambda^0\right) =$

$F^0\Lambda^{0\prime}$.[2] This motivates us to estimate F^0 and Λ^0 by minimizing the following least squares objective function

$$\mathcal{L}_{NT}^0(F,\Lambda) \equiv \frac{1}{NT}\text{tr}\left[\left(\frac{1}{\tilde{q}}\tilde{X} - F\Lambda'\right)\left(\frac{1}{\tilde{q}}\tilde{X} - F\Lambda'\right)'\right] \tag{2.3}$$

under the identification restrictions: $F'F/T = I_R$ and $\Lambda'\Lambda$ is a diagonal matrix. By concentrating out Λ and using the normalization that $F'F/T = I_R$, the above minimization problem is identical to maximizing $\frac{1}{\tilde{q}^2}\text{tr}\{F'\tilde{X}\tilde{X}'F\}$.[3] The estimated factor matrix, denoted by $\hat{F}^{(0)}$ is $\sqrt{T}$ times the eigenvectors corresponding to the R largest eigenvalues of the $T \times T$ matrix $\frac{1}{NT\tilde{q}^2}\tilde{X}\tilde{X}'$:

$$\frac{1}{NT\tilde{q}^2}\tilde{X}\tilde{X}'\hat{F}^{(0)} = \hat{F}^{(0)}\hat{D}^{(0)}, \tag{2.4}$$

where $\hat{D}^{(0)}$ is an $R \times R$ diagonal matrix consisting of the R largest eigenvalues of $\left(NT\tilde{q}^2\right)^{-1}\tilde{X}\tilde{X}'$, arranged in descending order along its diagonal line. Then the estimator of Λ' is given by

$$\hat{\Lambda}^{(0)\prime} = \frac{1}{\tilde{q}}\left(\hat{F}^{(0)\prime}\hat{F}^{(0)}\right)^{-1}\hat{F}^{(0)\prime}\tilde{X} = \frac{1}{T\tilde{q}}\hat{F}^{(0)\prime}\tilde{X}. \tag{2.5}$$

Let $\hat{F}_t^{(0)}$ denote the tth column of $\hat{F}^{(0)\prime}$ and $\hat{\lambda}_i^{(0)}$ the ith column of $\hat{\Lambda}^{(0)\prime}$. We can obtain an initial estimate of the (t,i)th element, C_{it}^0, of C^0 by $\hat{C}_{it}^{(0)} = \hat{\lambda}_i^{(0)\prime}\hat{F}_t^{(0)}$. We will show that the initial estimators $\hat{F}_t^{(0)}, \hat{\lambda}_i^{(0)}$ and $\hat{C}_{it}^{(0)}$ are consistent and follow mixture normal distributions under some standard conditions.

2.1.2 The iterated estimates

Despite the consistency of the initial estimators, they are not asymptotically efficient. To improve the efficiency, we consider iterative estimators. Let $\ell \geq 1$ be an integer. Suppose that we have obtained the estimates $\hat{F}_t^{(\ell-1)}, \hat{\lambda}_i^{(\ell-1)}$ and $\hat{C}_{it}^{(\ell-1)}$. In step ℓ, we can replace the missing values (X_{it}) in the matrix X with the estimated common components $\hat{C}_{it}^{(\ell-1)}$. Define the $T \times N$ matrix $\hat{X}^{(\ell)}$ with its (t,i)th element given by

$$\hat{X}_{it}^{(\ell)} = \begin{cases} X_{it} & \text{if } (i,t) \in \Omega \\ \hat{C}_{it}^{(\ell-1)} & \text{if } (i,t) \in \Omega_\perp \end{cases}, \ \ell \geq 1,$$

where $\Omega_\perp = \{(i,t) \in [N] \times [T] : (i,t) \notin \Omega\}$. Then we can conduct the PC analysis based on $\hat{X}^{(\ell)}$ under the identification restrictions that $F'F/T = I_R$ and $\Lambda'\Lambda$ is a diagonal matrix. The estimated

[2] The idea of scaling the matrix $\tilde{X}/q$ can be traced back to the machine learning literature; see, e.g., Negahban and Wainwright (2012). Their interest focuses on matrix completion in the matrix norms under noisy sampling and for both exact and near low-rank matrices, while our work focuses on the estimation of the factors and factor loadings.

[3] Following the lead of Bai and Ng (2002), one can alternatively consider concentrating out F under the identification restrictions that $\Lambda'\Lambda/N = I_R$ and $F'F$ is a diagonal matrix, which is computationally more efficient if $T \gg N$. The estimates of F and Λ will be different, but the estimate of $C = F\Lambda'$ would be the same. The theoretical analyses below can be derived analogously with apparent modifications.

factor matrix, denoted by $\hat{F}^{(\ell)}$, is $\sqrt{T}$ time the eigenvectors corresponding to the R largest eigenvalues of the $T \times T$ matrix $\frac{1}{NT}\hat{X}^{(\ell)}\hat{X}^{(\ell)\prime}$:

$$\frac{1}{NT}\hat{X}^{(\ell)}\hat{X}^{(\ell)\prime}\hat{F}^{(\ell)} = \hat{F}^{(\ell)}\hat{D}^{(\ell)},$$

where $\hat{D}^{(\ell)}$ is a diagonal matrix consisting of the R largest eigenvalues of $\frac{1}{NT}\hat{X}^{(\ell)}\hat{X}^{(\ell)\prime}$ arranged in descending order along its diagonal line. Then the estimator of Λ' is given by

$$\hat{\Lambda}^{(\ell)\prime} = \left(\hat{F}^{(\ell)\prime}\hat{F}^{(\ell)}\right)^{-1}\hat{F}^{(\ell)\prime}\hat{X}^{(\ell)} = \frac{1}{T}\hat{F}^{(\ell)\prime}\hat{X}^{(\ell)}.$$

Let $\hat{F}_t^{(\ell)}$ denote the tth column of $\hat{F}^{(\ell)\prime}$ and $\hat{\lambda}_i^{(\ell)}$ the ith column of $\hat{\Lambda}^{(\ell)\prime}$. We obtain the updated estimate of C_{it}^0 by $\hat{C}_{it}^{(\ell)} = \hat{\lambda}_i^{(\ell)\prime}\hat{F}_t^{(\ell)}$. We will study the asymptotic properties of $\hat{F}_t^{(\ell)}, \hat{\lambda}_i^{(\ell)}$ and $\hat{C}_{it}^{(\ell)}$, $\ell = 1, 2, ...,$ below.

Remark 1 (Connection with Stock and Watson's (2002) EM estimation) Stock and Watson (2002, SW hereafter) propose an EM algorithm to conduct the PC analysis for panel data with missing values. The least squares objective function they consider is given by

$$V(F,\Lambda) = \frac{1}{NT}\text{tr}\left[\left[\left(X - F\Lambda'\right)\circ G\right]\left[\left(X - F\Lambda'\right)\circ G\right]'\right] = \frac{1}{NT}\sum_{i=1}^{N}\sum_{i=1}^{N}\left(X_{it} - \lambda_i' F_t\right)^2 g_{it}.$$

Minimization of $V(F,\Lambda)$ requires iterative methods. SW motivate the EM algorithm by assuming that ε_{it}'s are independently and identically distributed (i.i.d.) according to $N\left(0,\sigma^2\right)$. They suggest various ways to obtain the initial estimates. For example, when the full dataset contains a subset constituting a balanced panel, they suggest using estimates of the factors from the balanced subset as the starting value $\hat{F}_t^{(0)}$. Given the estimates $\hat{C}_{it}^{(\ell-1)}$ at stage $\ell-1$, our construction of the expectation object $\hat{X}_{it}^{(\ell)}$ is the same as SW's and so is our ℓth stage estimator. But SW do not provide any theoretical justification for their EM estimates. They study neither numerical convergence of the EM algorithm nor the asymptotic properties of the EM estimators. In contrast, with our well-chosen initial estimators, we will show that our proposed procedure attains numerical convergence and formally establish the asymptotic properties of the resulting EM estimators.

2.2 Asymptotic properties of the initial estimators $\hat{F}_t^{(0)}, \hat{\lambda}_i^{(0)}$ and $\hat{C}_{it}^{(0)}$

Let M denote a generic finite positive constant that may vary across lines. We make the following assumptions.

Assumption A.1 (i) $\max_t E\left\|F_t^0\right\|^{4/\gamma_1} \leq M$ for some $\gamma_1 \in (0,1)$ and $T^{-1}F^{0\prime}F^0 \xrightarrow{P} \Sigma_{F^0} > 0$ for some $R \times R$ matrix Σ_{F^0} as $T \to \infty$.

(ii) $\max_i E\left\|\lambda_i^0\right\|^{4/\gamma_2} \leq M$ for some $\gamma_2 \in (0,1)$ and $N^{-1}\Lambda^{0\prime}\Lambda^0 \xrightarrow{P} \Sigma_{\Lambda^0} > 0$ for some $R \times R$ matrix Σ_{Λ^0} as $N \to \infty$.

(iii) $\max_{i,t} E[\left(\lambda_i^{0\prime}F_t^0\right)^4] \leq M$.

(iv) The eigenvalues of $\Sigma_{\Lambda^0}\Sigma_{F^0}$ are distinct from each other.

(v) $N^{-1}\Lambda^{0\prime}\Lambda^0 - \Sigma_{\Lambda_0} = O_P(N^{-1/2})$ and $T^{-1}F^{0\prime}F^0 - \Sigma_{F_0} = O_P(T^{-1/2})$.

Assumption A.2 (i) $E\left(\varepsilon_{it}|\lambda_i^0, F_t^0\right) = 0$, $E\left(\varepsilon_{it}^4\right) \leq M$, and $\|\varepsilon\|_{\text{sp}} = O_P(\max(\sqrt{N}, \sqrt{T}))$.
(ii) $\max_s \sum_{t=1}^T |\gamma_N(s,t)| \leq M$, where $\gamma_N(s,t) = N^{-1}\sum_{i=1}^N |E(\varepsilon_{is}\varepsilon_{it})|$.
(iii) $\max_{t,s} E\left|N^{-1/2}\sum_{i=1}^N [\varepsilon_{it}\varepsilon_{is} - E(\varepsilon_{it}\varepsilon_{is})]\right|^2 \leq M$.

Assumption A.1 parallels Assumptions A-B in Bai (2003) and Assumption A.2 is analogous to Assumption C in Bai (2003). The major difference is that we require both the factors and factor loadings have finite moments higher than the usual fourth order. Bai and Ng (2002) and Bai (2003) assume finite fourth moments for F_t^0 but require that λ_i^0 be uniformly bounded. Assumption A.1(v) imposes the standard convergence rates for $N^{-1}\Lambda^{0\prime}\Lambda^0$ and $T^{-1}F^{0\prime}F^0$. It implies that $\mu_r(\frac{1}{NT}F^0\Lambda^{0\prime}\Lambda^0F^{0\prime}) - \sigma_r^2 = O_P\left(\delta_{NT}^{-1}\right)$ for $r = 1, ..., R$, where $\sigma_r^2 = \mu_r\left(\Sigma_{\Lambda^0}\Sigma_{F^0}\right)$. Assumption A.2(i) is also assumed in Su and Chen (2013), Lu and Su (2016), and Moon and Weidner (2017). In particular, Moon and Weidner (2017) demonstrate that this condition can be satisfied for various error processes.

The following theorem establishes the mean squared convergence of $\hat{F}_t^{(0)}$. Define

$$\hat{H}^{(0)} = \left(N^{-1}\Lambda^{0\prime}\Lambda^0\right)T^{-1}F^{0\prime}\hat{F}^{(0)}(\hat{D}^{(0)})^{-1},$$

where $\hat{D}^{(0)}$ is asymptotically nonsingular by Lemma A.1.

Theorem 2.1 *Suppose Assumptions A.1 and A.2 hold. Then* $\frac{1}{T}\left\|\hat{F}^{(0)} - F^0\hat{H}^{(0)}\right\|^2 = O_P\left(\delta_{NT}^{-2}\right)$ *where* $\delta_{NT} = \sqrt{N} \wedge \sqrt{T}$.

Theorem 2.1 reports the mean squared (MS) convergence rate of $\hat{F}_t^{(0)}$. It implies that we can estimate the space spanned by the columns of F^0 consistently.

To proceed, we assume the following limiting objects exist and are finite:

$$\Gamma_{1g,t}(q) = \lim_{N\to\infty}\text{Var}\left(\frac{1}{\sqrt{N}q}\sum_{i=1}^N \lambda_i^0\varepsilon_{it}g_{it}\right),\ \Gamma_{2g,t}(q) = \text{plim}_{N\to\infty}\frac{1-q}{qN}\sum_{i=1}^N \lambda_i^0\lambda_i^{0\prime}\left(\lambda_i^{0\prime}F_t^0\right)^2,$$

$$\Phi_{1g,i}(q) = \lim_{T\to\infty}\text{Var}\left(\frac{1}{\sqrt{T}q}\sum_{t=1}^T F_t^0\varepsilon_{it}g_{it}\right),\ \Phi_{2g,i}(q) = \text{plim}_{T\to\infty}\frac{1-q}{qT}\sum_{t=1}^T F_t^0F_t^{0\prime}\left(\lambda_i^{0\prime}F_t^0\right)^2.$$

Let

$$\Gamma_{g,t}(q) = \Gamma_{1g,t}(q) + \Gamma_{2g,t}(q) \text{ and } \Phi_{g,i}(q) = \Phi_{1g,i}(q) + \Phi_{2g,i}(q).$$

Note that $\Gamma_{2g,t}$ and $\Phi_{2g,i}$ and therefore $\Gamma_{g,t}$ and $\Phi_{g,i}$ are generally random objects under our assumptions that allow for random factors and random factor loadings. To study the asymptotic distributions of $\hat{F}_t^{(0)}, \hat{\lambda}_i^{(0)}$ and $\hat{C}_{it}^{(0)}$, we add the following assumptions.

Assumption A.3 (i) Either $\max_{t,s} E\left\|\frac{1}{\sqrt{N}}\sum_{i=1}^N \chi_{i,st}\right\|^4 \leq M$ or $E\left\|\frac{1}{\sqrt{NT}}\sum_{s=1}^T\sum_{i=1}^N F_s^0\chi_{i,st}\right\|^2 \leq M$, where $\chi_{i,st} = \varepsilon_{it}\varepsilon_{is} - E(\varepsilon_{it}\varepsilon_{is})$.
(ii) $E\left\|\frac{1}{\sqrt{NT}}\sum_{s=1}^T\sum_{i=1}^N F_s^0\lambda_i^{0\prime}\varepsilon_{is}\right\|^2 \leq M$.

(iii) Let $\sigma_{ij,ts} = E(\varepsilon_{it}\varepsilon_{js})$. $\max_t N^{-1}\sum_{i=1}^N \sigma_{ii,tt} \leq M$, $\max_{1\leq t\leq T} N^{-1}\sum_{i=1}^N\sum_{j=1}^N |\sigma_{ij,tt}| \leq M$, $\max_{1\leq i\leq N} T^{-1}\sum_{t=1}^T\sum_{s=1}^T |\sigma_{ii,ts}| \leq M$, and $(NT)^{-1}\sum_{i=1}^N\sum_{j=1}^N\sum_{t=1}^T\sum_{s=1}^T |\sigma_{ij,ts}| \leq M$.

Assumption A.4 (i) $\frac{1}{\sqrt{N}q}\sum_{i=1}^N \lambda_i^0\varepsilon_{it}g_{it} \xrightarrow{d} N(0,\Gamma_{1g,t})$,

(ii) $\frac{1}{\sqrt{T}q}\sum_{t=1}^T F_t^0\varepsilon_{it}g_{it} \xrightarrow{d} N(0,\Phi_{1g,i})$.

The first part of Assumption A.3(i) strengthens Assumption A.2(iii) and is also assumed in Bai and Ng (2002, Assumption C.5) and Bai (2003, Assumptions C.5 and F.1). Bai (2003) assumes that the second part of A.3(i) holds simultaneously with the first part, which we do not need. In the special case where $E(F_s^0\chi_{i,st}) \neq 0$ for enough (s,t) pairs (e.g., when $E(F_s^0) = 0$ but $E(F_s^0\varepsilon_{it}\varepsilon_{is}) \neq 0$ for all $s > t$), the second part of A.3(i) is not satisfied. Assumption A.3(ii) is the same as Assumption F.2 in Bai (2003), and Assumption A.3(iii) is comparable with and slightly stronger than Assumption C.2-C.4 in Bai and Ng (2002) and Assumptions C.2-C.4 and E.1-E.2 in Bai (2003). Assumption A.4(i)-(ii) is parallel to Assumption F.3-F.4 in Bai (2003) and reduces to the latter in the case when $g_{it} = 1$ for all (i,t).

Let $\mathcal{G}_{Ni}^t = \sigma(\{g_{jt}, j \leq i), \Lambda^0, F_t^0)$, the minimal sigma-field generated from $\{g_{jt},\ j \leq i\}$ and (Λ^0, F_t^0). Let $\mathcal{G}^t = \sigma\left(\cup_{N=1}^\infty \mathcal{G}_{NN}^t\right)$. Analogously, let $\mathcal{G}_{Tt}^i = \sigma\left(\{g_{is}, s \leq t), \lambda_i^0, F^0\right)$, $\mathcal{G}^i = \sigma\left(\cup_{T=1}^\infty \mathcal{G}_{TT}^i\right)$, and $\mathcal{G}^{it} = \sigma\left(\mathcal{G}^i \cup \mathcal{G}^t\right)$.

The following theorem presents the asymptotic distributions of $\hat{F}_t^{(0)}, \hat{\lambda}_i^{(0)}$ and $\hat{C}_{it}^{(0)}$ based on the notation of stable convergence.

Theorem 2.2 *Suppose Assumptions A.1-A.4 hold. Suppose that* $(T^{1/2} + N^{1/2})\delta_{NT}^{-2} = o(1)$. *Let* $\hat{\Pi}_{tN}^{(0)} = \sqrt{N}(\hat{F}_t^{(0)} - \hat{H}^{(0)\prime}F_t^0)$ *and* $\hat{\Pi}_{iT}^{(0)} = \sqrt{T}(\hat{\lambda}_i^{(0)} - (\hat{H}^{(0)})^{-1}\lambda_i^0)$. *Then as* $(N,T) \to \infty$

(i) $\hat{\Pi}_{tN}^{(0)} = (\hat{D}^{(0)})^{-1}\frac{1}{T}\hat{F}^{(0)\prime}F^0\frac{1}{\sqrt{N}q}\sum_{i=1}^N \lambda_i^0\xi_{it} + O_P\left(N^{1/2}\delta_{NT}^{-2}\right) \to N\left(0, D^{-1}Q\Gamma_{g,t}(q)Q'D^{-1}\right)$ $\mathcal{G}^t$*-stably,*

(ii) $\hat{\Pi}_{iT}^{(0)} = \hat{H}^{(0)\prime}\frac{1}{\sqrt{T}q}\sum_{t=1}^T F_t^0\xi_{it} + O_P\left(T^{1/2}\delta_{NT}^{-2}\right) \to N\left(0, (Q')^{-1}\Phi_{g,i}(q)Q^{-1}\right)$ $\mathcal{G}^i$*-stably,*

(iii) $\left(\frac{1}{N}\Sigma_{F,it}^{(0)}(q) + \frac{1}{T}\Sigma_{\Lambda,it}^{(0)}(q)\right)^{-1/2}\left(\hat{C}_{it}^{(0)} - C_{it}^0\right) \xrightarrow{d} N(0,1)$,

where $\xi_{it} = \varepsilon_{it}g_{it} + \lambda_i^{0\prime}F_t^0(g_{it} - q)$, $\Sigma_{F,it}^{(0)}(q) = \lambda_i^{0\prime}\Sigma_{\Lambda^0}^{-1}\Gamma_{g,t}(q)\Sigma_{\Lambda^0}^{-1}\lambda_i^0$ *and* $\Sigma_{\Lambda,it}^{(0)}(q) = F_t^{0\prime}\Sigma_{F^0}^{-1}\Phi_{g,i}(q)\Sigma_{F^0}^{-1}F_t^0$ *signify the contributions of the factor and factor loading estimators to the asymptotic variance of* $\hat{C}_{it}^{(0)}$, *respectively, and* D *denotes the diagonal matrix consisting of the eigenvalues of* $\Sigma_{\Lambda^0}^{1/2}\Sigma_{F^0}\Sigma_{\Lambda^0}^{1/2}$ *in descending order with the corresponding eigenvector matrix denoted as* Υ *such that* $\Upsilon'\Upsilon = I_R$ *and* $Q = D^{1/2}\Upsilon'\Sigma_{\Lambda^0}^{-1/2}$.

Theorem 2.2 parallels Theorems 1-3 in Bai (2003). Bai (2003) obtains the asymptotic normal distributions for his estimators of factors and factor loadings. In contrast, we show that the sequence$\left\{\hat{\Pi}_{tN}^{(0)},\ N \geq 1\right\}$ converges $\mathcal{G}^t$-stably as $(N,T) \to \infty$ to a mixture normal whose asymptotic variance is random but measurable with respect to certain limit sigma-field, and similarly, the sequence $\left\{\hat{\Pi}_{iT}^{(0)},\ T \geq 1\right\}$ converges $\mathcal{G}^i$-stably as $(N,T) \to \infty$ to a mixture normal whose asymptotic variance is random but measurable with respect to certain limit sigma-field. We refer the reader directly to the Häusler and Luschgy (2015) for stable convergence in general and the stable

martingale central limit theorem in particular. To understand the limiting distribution of $\hat{\Pi}_{tN}^{(0)}$ in Theorem 2.2(i), we notice that its influence function depends on ξ_{it} through two terms, $\varepsilon_{it}g_{it}$ and $\lambda_i^{0\prime}F_t^0\left(g_{it}-q\right)$. The first term also appears in the influence function for the factor estimators in the absence of random missing at time t (i.e., $g_{it}=1\ \forall i$) while the second term is introduced by the random missing mechanism. Due to the presence of common factor F_t^0 for all cross-sectional units, $\frac{1}{\sqrt{N}q}\sum_{i=1}^{N}\lambda_i^0\lambda_i^{0\prime}F_t^0\left(g_{it}-q\right)$ does not have a limiting normal distribution. Instead, it converges to $N\left(0,\Gamma_{2g,t}\right)$ $\mathcal{G}^t$-stably as $N\to\infty$, where $N\left(0,\Gamma_{2g,t}\right)$ can be regarded as a normal random vector with random variance given by $\Gamma_{2g,t}$. In the special case where F_t^0 is nonrandom, the limiting distribution reduces to the usual normal distribution. Similar remarks apply for $\hat{\Pi}_{iT}^{(0)}$ in Theorem 2.2(ii). Theorem 2.2(iii) only reports the limiting distribution for the normalized common component estimator. One can also follow the analyses of parts (i)-(ii) in the theorem and report the stable limiting distribution of $\delta_{NT}(\hat{C}_{it}^{(0)}-C_{it}^0)$ as $(N,T)\to\infty$.

By Corollary 6.3 in Häusler and Luschgy (2015) and the Cramér-Wold device, we can show that

$$\begin{aligned}\left[(D^{-1}Q\Gamma_{g,t}Q'D^{-1}\right]^{-1/2}\hat{\Pi}_{tN}^{(0)}&\xrightarrow{d}N\left(0,I_R\right)\ \text{as}\ (N,T)\ \to\ \infty,\ \text{and}\\ \left[(Q')^{-1}\Phi_{g,i}Q^{-1}\right]^{-1/2}\hat{\Pi}_{iT}^{(0)}&\xrightarrow{d}N\left(0,I_R\right)\ \text{as}\ (N,T)\ \to\ \infty.\end{aligned}$$

With these results and the result in Theorem 2.2(iii), we could make inference on the factors, factor loadings, and common component. Since these estimates are not the final estimates, we will study the asymptotic properties of the iterated estimators of these objects later on.

2.3 Asymptotic properties of the iterated estimators of the factors and factor loadings

Let $\hat{H}^{(\ell)}=\left(N^{-1}\Lambda^{0\prime}\Lambda^0\right)T^{-1}F^{0\prime}\hat{F}^{(\ell)}(\hat{D}^{(\ell)})^{-1}$. To study the asymptotic properties of $\hat{F}_t^{(\ell)},\hat{\lambda}_i^{(\ell)}$ and $\hat{C}_{it}^{(\ell)}$, we add the following assumption.

Assumption A.5 (i) $\max_t\left\|\frac{1}{N}\sum_{i=1}^{N}\zeta_{1,it}\right\|=O_P\left((N/\ln N)^{-1/2}\right)$ and $\max_{t,s}\left\|\frac{1}{N}\sum_{i=1}^{N}\lambda_i^0\varepsilon_{it}g_{it}g_{is}\right\|=O_P\left((N/\ln N)^{-1/2}\right)$, where $\zeta_{1,it}=\lambda_i^0\varepsilon_{it}g_{it}$ and $\lambda_i^0\lambda_i^{0\prime}F_t^0\left(g_{it}-q\right)$.

(ii) $\max_i\left\|\frac{1}{T}\sum_{t=1}^{T}\zeta_{2,it}\right\|=O_P\left((T/\ln T)^{-1/2}\right)$, where $\zeta_{2,it}=F_t^0\varepsilon_{it}g_{it}$ and $F_t^0\lambda_i^{0\prime}F_t^0\left(g_{it}-q\right)$.

(iii) $\max_t\left\|\frac{1}{NT}\sum_{i,s}\zeta_{3,its}\right\|=O_P\left(\delta_{NT}^{-2}\ln N\right)$ and $\max_t\left\|\frac{1}{NT}\sum_i\sum_{s=1,s\neq t}^{T}F_s^0F_s^{0\prime}\lambda_i^0\lambda_i^{0\prime}(g_{is}-q)(g_{it}-q)\right\|$ $=O_P\left(\delta_{NT}^{-2}\ln N\right)$, where $\zeta_{3,its}=F_s^0\left[\varepsilon_{it}\varepsilon_{is}-E(\varepsilon_{it}\varepsilon_{is})\right]g_{it}g_{is}$, $F_s^0F_s^{0\prime}\lambda_i^0\varepsilon_{it}g_{it}(g_{is}-q)$ and $\lambda_i^0F_s^{0\prime}\varepsilon_{is}g_{is}(g_{it}-q)$.

Assumption A.5 imposes some high level conditions that are similar to those imposed in Su et al. (2015) and Su and Wang (2017). Following these authors, one can verify Assumption A.5 under some primitive conditions on $\left\{\lambda_i^0,F_t^0,\varepsilon_{it}\right\}$. The conditions in Assumption A.5 are needed for the establishment of the uniform convergence results in Theorem 2.4 below. They still allow weak cross-section or serial dependence in the error terms or weak serial dependence in the factors but do rule out unit-root type nonstationary behavior or long-memory behavior along the time dimension in the error terms and factors. When unit root or long memory is of concern in factor models with random

missing, we admit that our assumptions (similar those of BN and Bai (2003)) will be violated and one needs to specify a different set of assumptions. But this goes beyond the scope of the current paper and is left for future research.

The following theorem establishes the mean squared convergence of $\hat{F}_t^{(\ell)}$.

Theorem 2.3 *Suppose Assumptions A.1-A.5 hold. Then* $\frac{1}{T}\left\|\hat{F}^{(\ell)} - F^0\hat{H}^{(\ell)}\right\|^2 = O_P\left(\delta_{NT}^{-2}\right)$ *for each* ℓ.

The following theorem reports the asymptotic distributions of $\hat{F}_t^{(\ell)}$, $\hat{\lambda}_i^{(\ell)}$ and $\hat{C}_{it}^{(\ell)}$.

Theorem 2.4 *Suppose Assumptions A.1-A.5 hold. Suppose that* $\sqrt{N}(T^{\gamma_1/4}\delta_{NT}^{-2}\ln T + T^{-1+3\gamma_1/4}) = o(1)$ *and* $\sqrt{T}(N^{\gamma_2/4}\delta_{NT}^{-2}\ln N + N^{-1+3\gamma_2/4}) = o(1)$. *Let* $\hat{\Pi}_{tN}^{(\ell)} = \sqrt{N}(\hat{F}_t^{(\ell)} - \hat{H}^{(\ell)\prime}F_t^0)$ *and* $\hat{\Pi}_{iT}^{(\ell)} = \sqrt{T}(\hat{\lambda}_i^{(\ell)} - \hat{H}^{(\ell)-1}\lambda_i^0)$. *Then*

(i) $\hat{\Pi}_{tN}^{(\ell)} = D^{-1}Q\frac{1}{\sqrt{N}}\sum_{i=1}^N \lambda_i^0\varepsilon_{it}g_{it} + (1-q)\,\hat{\Pi}_{tN}^{(\ell-1)} + o_P(1)$ *uniformly in* t *and*

$$\hat{\Pi}_{tN}^{(\ell)} \xrightarrow{d} N(0, D^{-1}Q\Gamma_{1g,t}(q)\,Q'D^{-1}) \text{ as } (\ell, N, T) \to \infty,$$

(ii) $\hat{\Pi}_{iT}^{(\ell)} = (Q')^{-1}\frac{1}{\sqrt{T}}\sum_{t=1}^T F_t^0\varepsilon_{it}g_{it} + (1-q)\,\hat{\Pi}_{iT}^{(\ell-1)} + o_P(1)$ *uniformly in* i *and*

$$\hat{\Pi}_{iT} \xrightarrow{d} N(0, (Q')^{-1}\Phi_{1g,i}(q)\,Q^{-1}) \text{ as } (\ell, N, T) \to \infty,$$

(iii) $(\frac{1}{N}\Sigma_{1F,it} + \frac{1}{T}\Sigma_{1\Lambda,it})^{-1/2}(\hat{C}_{it}^{(\ell)} - C_{it}^0) \xrightarrow{d} N(0,1)$ *as* $(\ell, N, T) \to \infty$,

where $\Gamma_{1g,t}, \Phi_{1g,i}$, D *and* Q *are as defined in the last subsection, and* $\Sigma_{1F,it} = \lambda_i^{0\prime}\Sigma_{\Lambda^0}^{-1}\Gamma_{1g,t}(q)\,\Sigma_{\Lambda^0}^{-1}\lambda_i^0$, *and* $\Sigma_{1\Lambda,it} = F_t^{0\prime}\Sigma_{F^0}^{-1}\Phi_{1g,i}(q)\,\Sigma_{F^0}^{-1}F_t^0$ *signify the contribution of the factor and factor loading estimators to the asymptotic variance of* $\hat{C}_{it}^{(\ell)}$ *for large* ℓ, *respectively.*

Remark 2 Note that $\Gamma_{g,t}(q) = \Gamma_{1g,t}(q) + \Gamma_{2g,t}(q)$ and $\Phi_{g,i}(q) = \Phi_{1g,i}(q) + \Phi_{2g,i}(q)$. A comparison of Theorem 2.4 with Theorem 2.2 indicates that $\hat{F}_t^{(\ell)}$, $\hat{\lambda}_i^{(\ell)}$ and $\hat{C}_{it}^{(\ell)}$ are asymptotically more efficient than $\hat{F}_t^{(0)}$, $\hat{\lambda}_i^{(0)}$ and $\hat{C}_{it}^{(0)}$, respectively. In theory, the distributional results in Theorem 2.4 require $\ell \to \infty$. In practice, ℓ can diverge to infinity at an arbitrarily slow rate. To see this point, we take a close look at the iterative relationship between $\hat{\Pi}_{tN}^{(\ell)}$ and $\hat{\Pi}_{tN}^{(\ell-1)}$. Let $\beta_{F,t} = \frac{1}{N}\sum_{i=1}^N \lambda_i^0\varepsilon_{it}g_{it}$. Note that the result in Theorem 2.4(i) implies

$$\hat{\Pi}_{tN}^{(\ell)} = D^{-1}Q\sqrt{N}\beta_{F,t}\sum_{s=0}^{\ell-1}(1-q)^s + (1-q)^\ell\,\hat{\Pi}_{tN}^{(0)} + o_P(1),$$

where the first term is the dominant term and the second term can be made arbitrarily small for sufficiently large ℓ. In practice, we find it is not necessary to iterate too many times so that we can stop the iteration when $(1-q)^\ell$ is small enough. For example, we can iterate ℓ^* times such that $(1-q)^{\ell^*} \asymp \epsilon_{NT}$ for some small positive number ϵ_{NT}. Simulations suggest that $\ell^* = \lfloor \ln(\epsilon_{NT})/\ln(1-q) \rfloor$ with $\epsilon_{NT} = 0.001$ works very well for all data generating processes under our investigation. Note that $\ell^* = 3$, 4, and 5 for $q = 0.9$, 0.8, and 0.7, respectively. This suggests a small number of iterations is sufficient.

Remark 3 (Comparison with the oracle estimators) We can also compare the asymptotic variances of our EM estimators with those of the oracle estimators that are obtained in the absence of missing values (viz., $q = 1$). For example, we consider the factor estimation and use $\hat{F}_t^{\text{oracle}}$ to denote the oracle estimator of F_t^0 with the corresponding rotational matrix $\hat{H}^{\text{oracle}}$. It is well known that the asymptotic variance-covariance (Avar) of $\sqrt{N}(\hat{F}_t^{\text{oracle}} - \hat{H}^{\text{oracle}\prime} F_t^0)$ is given by $D^{-1}Q\Gamma_t^{\text{oracle}}Q'D^{-1}$, where

$$\Gamma_t^{\text{oracle}} = \lim_{N\to\infty} \text{Var}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\lambda_i^0\varepsilon_{it}\right).$$

In contrast, by the law of iterated expectations

$$\begin{aligned}
\Gamma_{1g,t}(q) &= \lim_{N\to\infty}\left\{\text{Var}\left[E\left(\frac{1}{\sqrt{N}q}\sum_{i=1}^{N}\lambda_i^0\varepsilon_{it}g_{it}|\Lambda^0,\varepsilon\right)\right] + E\left[\text{Var}\left(\frac{1}{\sqrt{N}q}\sum_{i=1}^{N}\lambda_i^0\varepsilon_{it}g_{it}|\Lambda^0,\varepsilon\right)\right]\right\} \\
&= \lim_{N\to\infty}\left\{\text{Var}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\lambda_i^0\varepsilon_{it}\right) + \frac{1-q}{q}E\left(\frac{1}{N}\sum_{i=1}^{N}\lambda_i^0\lambda_i^{0\prime}\varepsilon_{it}^2\right)\right\} \\
&\geq \Gamma_t^{\text{oracle}}.
\end{aligned}$$

The difference, $\Gamma_{1g,t}(q) - \Gamma_t^{\text{oracle}}$, given by $\lim_{N\to\infty}\frac{1-q}{q}E\left(\frac{1}{N}\sum_{i=1}^{N}\lambda_i^0\lambda_i^{0\prime}\varepsilon_{it}^2\right)$, reflects the cost of missing $(1-q)$ proportion of observations. The larger proportion of missing observations, the larger value $\Gamma_{1g,t}(q)$ is. In the absence of cross-sectional correlation among $\left\{\lambda_i^0\varepsilon_{it}\right\}$, it is easy to verify that

$$\Gamma_{1g,t}(q) = \frac{1}{q}\lim_{N\to\infty}E\left(\frac{1}{N}\sum_{i=1}^{N}\lambda_i^0\lambda_i^{0\prime}\varepsilon_{it}^2\right) = \frac{1}{q}\Gamma_t^{\text{oracle}}.$$

So q reflects the relative asymptotic efficiency of the EM estimator compared to the oracle estimator. Analogous remarks hold for our EM estimators of the factor loadings.

With the results in Theorem 2.4, we can make inference on the factors, factor loadings, and common component. Below we focus on the inference on the factors due to the widespread use of estimated factors, say, in various factor-augmented regression or forecasting models.

2.4 Inference on the factors

Let $\hat{F}_t$, $\hat{\lambda}_i$, and $\hat{C}_{it}$ denote $\hat{F}_t^{(\ell)}$, $\hat{\lambda}_i^{(\ell)}$, and $\hat{C}_{it}^{(\ell)}$ respectively, when $\ell \to \infty$. To make inference on the factors, we need to estimate the asymptotic variance $V_F \equiv D^{-1}Q\Gamma_{1g,t}(q)Q'D^{-1}$ consistently. By Lemma A.1 in the appendix, we can consistently estimate D by the diagonal matrix $\hat{D} = \hat{D}^{(\infty)}$, that contains the R largest eigenvalues of $(NT)^{-1}\hat{X}^{(\infty)}\hat{X}^{(\infty)\prime}$, arranged in descending order. So the key is to estimate $Q\Gamma_{1g,t}Q'$ consistently.

To estimate $Q\Gamma_{1g,t}(q)Q'$, we consider two cases: (1) $\left\{\lambda_i^0\varepsilon_{it}g_{it}\right\}$ are cross-sectionally uncorrelated; (2) $\left\{\lambda_i^0\varepsilon_{it}g_{it}\right\}$ are cross-sectionally correlated. In Case (1), we have a simplified expression for $\Gamma_{1g,t}(q)$:

$$\Gamma_{1g,t}(q) = \lim_{N\to\infty}\frac{1}{Nq^2}\sum_{i=1}^{N}\text{Var}\left(\lambda_i^0\varepsilon_{it}g_{it}\right) = \lim_{N\to\infty}\frac{1}{Nq^2}\sum_{i=1}^{N}E\left[\lambda_i^0\lambda_i^{0\prime}\left(\varepsilon_{it}^g\right)^2\right],$$

where $\varepsilon_{it}^g = \varepsilon_{it} g_{it}$. Noting that with $\tilde{H} \equiv \hat{H}^{(\infty)}$, $\tilde{H}^{-1} \xrightarrow{p} Q$ by Lemma A.4(iii) in the appendix, it is easy to show that a consistent estimator of $Q\Gamma_{1g,t}(q) Q'$ is given by

$$\hat{\Gamma}_{1g,t}^{(1)} = \frac{1}{N\tilde{q}^2} \sum_{i=1}^{N} \hat{\lambda}_i \hat{\lambda}_i' (\hat{\varepsilon}_{it}^g)^2,$$

where $\hat{\varepsilon}_{it}^g = (X_{it} - \hat{C}_{it}) g_{it}$.

In Case (2), for simplicity we consider the case where the factor loadings are nonrandom and the process $\{\varepsilon_{it},\ t \geq 1\}$ is covariance stationary. Let $\varepsilon_{\cdot t}^g = \left(\varepsilon_{1t}^g, \varepsilon_{2t}^g, ..., \varepsilon_{Nt}^g\right)'$. Let $\Sigma^g \equiv E(\varepsilon_{\cdot t}^g \varepsilon_{\cdot t}^{g\prime}) = \{\sigma_{ij}^g\}$, which is an $N \times N$ matrix. Then $\Gamma_{1g,t}(q) = \lim_{N\to\infty} \frac{1}{Nq^2} \mathrm{Var}\left(\Lambda^{0\prime} \varepsilon_{\cdot t}^g\right) = \lim_{N\to\infty} \frac{1}{Nq^2} \Lambda^{0\prime} \Sigma^g \Lambda^0$. Suppose that $\tilde{\Sigma}^g$ is a consistent estimator of Σ^g in the sense $\left\|\hat{\Sigma}^g - \Sigma^g\right\|_{\mathrm{sp}} = o_P(1)$. Then we can readily show that a consistent estimator of $Q\Gamma_{1g,t}Q'$ is given by

$$\hat{\Gamma}_{1g,t}^{(2)} \equiv \frac{1}{N\tilde{q}^2} \hat{\Lambda}' \hat{\Sigma}^g \hat{\Lambda}.$$

Fortunately, a feasible consistent estimator of Σ^g is available as ε_{it}^g can be estimated by $\hat{\varepsilon}_{it}^g$ and there is no need to estimate the error terms corresponding to those missing observations. To see this, define

$$\hat{\sigma}_{ij}^g = \frac{1}{T} \sum_{t=1}^{T} \hat{\varepsilon}_{it}^g \hat{\varepsilon}_{jt}^g \text{ and } \hat{\theta}_{ij} = \frac{1}{T} \sum_{t=1}^{T} \left(\hat{\varepsilon}_{it}^g \hat{\varepsilon}_{jt}^g - \hat{\sigma}_{ij}^g\right)^2.$$

We follow the lead of Fan, Liao and Mincheva (2013, FLM hereafter) and propose to estimate Σ^g by $\hat{\Sigma}^g = \left\{\hat{\sigma}_{ij}^{g,\mathcal{T}}\right\}$, where

$$\hat{\sigma}_{ij}^{g,\mathcal{T}} = \begin{cases} \hat{\sigma}_{ij}^g & \text{if } i = j \\ s_{ij}(\hat{\sigma}_{ij}^g) & \text{if } i \neq j \end{cases},$$

where $s_{ij}(\cdot)$ is the soft thresholding function: $s_{ij}(z) \equiv sgn(z)(|z| - \tau_{ij})_+$, $\tau_{ij} = c_0 \omega_{NT} (\hat{\theta}_{ij})^{1/2}$, $\omega_{NT} = [\max(N^{-1+\gamma_2/2}, T^{-1} \ln T)]^{1/2}$, and c_0 is a positive constant.[4] We will show that $\left\|\hat{\Sigma}^g - \Sigma^g\right\|_{\mathrm{sp}} = o_P(1)$ under some additional conditions.

When Λ^0 is random, the above procedure also works under the additional restriction that $\mathrm{Var}\left(\varepsilon_{\cdot t}^g | \Lambda^0\right) = \mathrm{Var}(\varepsilon_{\cdot t}^g) = \Sigma^g$. To see this, we notice that by the variance decomposition formula, we have

$$\begin{aligned} \Gamma_{1g,t}(q) &= \lim_{N\to\infty} \frac{1}{Nq^2} E\left[\mathrm{Var}\left(\Lambda^{0\prime} \varepsilon_{\cdot t}^g | \Lambda^0\right)\right] + \lim_{N\to\infty} \frac{1}{Nq^2} \mathrm{Var}\left(E\left(\Lambda^{0\prime} \varepsilon_{\cdot t}^g | \Lambda^0\right)\right) \\ &= \lim_{N\to\infty} \frac{1}{Nq^2} E\left[\Lambda^{0\prime} \mathrm{Var}\left(\varepsilon_{\cdot t}^g | \Lambda^0\right) \Lambda^0\right] + 0 = \lim_{N\to\infty} \frac{1}{Nq^2} E\left[\Lambda^{0\prime} \Sigma^g \Lambda^0\right]. \end{aligned}$$

$\frac{1}{Nq^2} E\left[\Lambda^{0\prime} \Sigma^g \Lambda^0\right]$ can be estimated in the same procedure as outlined above.

To allow for possible cross-sectional dependence, we recommend using $\hat{\Gamma}_{1g,t}^{(2)}$ and will justify the consistency of this estimator below. To proceed, we add the following assumption.

[4] In our simulations and applications, we let $c_0 = 1$. In most situations, when $c_0 = 1$, $\tilde{\Sigma}^g$ is positive definite. Otherwise, we choose c_0 to be the smallest value such that $\tilde{\Sigma}^g$ is positive definite. For details, see FLM's Section 4.

Assumption A.6 (i) The process $\{\varepsilon_{\cdot t}^g,\ t \geq 1\}$ is covariance-stationary with covariance matrix $\Sigma^g = E(\varepsilon_{\cdot t}^g \varepsilon_{\cdot t}^{g\prime}) = \left\{\sigma_{ij}^g\right\}$.

(ii) There exists $\gamma_3 \in [0,1)$ such that $\max_i \sum_j \left|\sigma_{ij}^g\right|^{\gamma_3} \leq M$.

(iii) Let $\omega_{NT} = [\max(N^{-1+\gamma_2/2}, T^{-1}\ln T)]^{1/2}$. $T^{-1/2+\gamma_1/4}(N^{\gamma_2/4} + T^{\gamma_1/4})(\ln T)^{1/2} \to 0$ and $T^{-1+\gamma_1/4}\omega_{NT}^{1-\gamma_3} N^{1/2} \to 0$ as $(N,T) \to \infty$.

Assumption A.6(i) is typically assumed in the literature when there is no missing value. Assumption A.6(ii) strengthens the standard weak cross-sectional dependence condition $\max_i \sum_j \left|\sigma_{ij}^g\right| = O(1)$; see, e.g., FLM. It is satisfied if $\varepsilon_{\cdot t}^g$'s satisfy certain m-dependence condition cross-sectionally or the correlation between ε_{it}^g and ε_{jt}^g vanishes sufficiently fast as the "distance" between i and j increases, perhaps after reordering of the data along the cross-sectional dimension. Assumption A.6(iii) imposes further restrictions on the relative magnitude of N and T.

The following theorem reports the consistency of $\hat{D}^{-1}\hat{\Gamma}_{1g,t}\hat{D}^{-1}$.

Theorem 2.5 *Suppose that Assumptions A.1-A.6 hold. Then* $\hat{D}^{-1}\hat{\Gamma}_{1g,t}\hat{D}^{-1} \xrightarrow{p} D^{-1}Q\Gamma_{1g,t}(q)\,Q'D^{-1}$, *where* $\hat{\Gamma}_{1g,t} = \hat{\Gamma}_{1g,t}^{(2)}$.

Given the above result, we can make inference on the global factors. The procedure is standard and omitted for brevity.

3 Determining the Number of Factors via Cross Validation

In this section, we propose a novel method to determine the number of factors via cross-validation (CV). In comparison with existing methods, our method has the following features. First, our method is inspired by the results in Section 2. With the theories in Section 2, it seems natural to consider a CV method to determine the number of factors. The key insight for such a method to work is that we can consistently estimate the common component for the factor models with random missing. Therefore we can randomly hold some observations for the out-of-sample evaluation and use the remaining observations to estimate the common component. Second, our method can be used no matter whether there are random missing observations in the original data matrix or not. In contrast, all popular existing methods for the determination of the number of factors do not allow for missing values without suitable modifications. Third, as our simulations show, our method works well in a variety of scenarios in comparison with existing methods adjusted to incorporate missing values. Fourth, like many existing methods, our CV method is easy to implement and computationally efficient.

For notational simplicity, we first focus on the CV method when the original dataset does not have missing value problems and then study the case with missing values.

3.1 The cross validation method

Let R denote the generic number of factors with the true value given by R_0. The key insight for our CV method is that one can consistently estimate the common component for the factor models with random missing. Given the $T \times N$ matrix of observations X, we propose to randomly sample elements in X with a fixed probability $p \in (0,1)$ and leave the rest $(1-p)$-proportion of observations as held-out entries for the out-of-sample evaluation.

As before, let $\Omega^* \subset [N] \times [T]$ be the index set of the training entries and $\Omega_\perp^*$ the index set of the held-out entries. Define the operator $P_{\Omega^*} : \mathbb{R}^{T\times N} \to \mathbb{R}^{T\times N}$ by

$$(P_{\Omega^*} X)_{ti} = X_{it} g_{it}^* = X_{it} \mathbf{1}\{(i,t) \in \Omega^*\},$$

where $g_{it}^* = \mathbf{1}\{(i,t) \in \Omega^*\}$. Let G^* denote a $T \times N$ matrix with (t,i)th element given by g_{it}^*. Now we can regard $P_{\Omega^*} X$ as the $T \times N$ data matrix with missing values replaced by zeros. Given $P_{\Omega^*} X$, we apply the proposed EM algorithm to recover the data via estimating the common component matrix C for any given number of factors.

To proceed, we consider the full singular value decomposition (SVD) for $\frac{1}{p} P_{\Omega^*} X$:

$$\frac{1}{p} P_{\Omega^*} X = \tilde{U} \tilde{\Sigma} \tilde{V}' = \sum_{r=1}^{T \wedge N} \tilde{u}_r \tilde{v}_r' \tilde{\sigma}_r,$$

where $\tilde{U} = (\tilde{u}_1, ..., \tilde{u}_T)$ and $\tilde{V} = (\tilde{v}_1, ..., \tilde{v}_N)$ are respectively the $T \times T$ matrix of left singular vectors and $N \times N$ matrix of right singular vectors of $\frac{1}{p} P_{\Omega^*} X$, and $\tilde{\Sigma}$ is the $T \times N$ 'diagonal' matrix that contains the singular values, $\tilde{\sigma}_1, \tilde{\sigma}_2, ..., \tilde{\sigma}_{T\wedge N}$, arranged in descending order along the main diagonal line. Given any $R \leq T \wedge N$ and the training entries in $P_{\Omega^*} X$, we can estimate the common component C by the singular value thresholding procedure:

$$\tilde{C}_R = S_H\left(\frac{1}{p} P_{\Omega^*} X, R\right) = \tilde{U}_R \tilde{\Sigma}_R \tilde{V}_R' = \sum_{r=1}^{R} \tilde{u}_r \tilde{v}_r' \tilde{\sigma}_r, \tag{3.1}$$

where $S_H(\cdot, R)$ is the rank-R truncated SVD of $\cdot$, the subscript H stands for hard thresholding, $\tilde{U}_R = (\tilde{u}_1, ..., \tilde{u}_R)$, $\tilde{V}_R = (\tilde{v}_1, ..., \tilde{v}_R)$, and $\tilde{\Sigma}_R = \text{diag}(\tilde{\sigma}_1, ..., \tilde{\sigma}_R)$. We can regard $\tilde{C}_R$ as a matrix-completion version of $P_{\Omega^*} X$. Let $\tilde{C}_{R,it}$ denote the (t,i)th element of $\tilde{C}_R$ for $R \geq 1$. Let $\tilde{C}_{0,it} = 0$ for all (i,t). We propose to choose R to minimize the following CV criterion function

$$\widetilde{CV}(R) = \sum_{(i,t)\in\Omega_\perp^*} \left[X_{it} - \tilde{C}_{R,it}\right]^2. \tag{3.2}$$

Let $\tilde{R} = \arg\min_{0\leq R\leq R_{\max}} \widetilde{CV}(R)$ where $R_{\max}$ is a fixed integer that is no less than R_0. We will show the consistency of $\tilde{R}$ under some regularity conditions.

Note that the CV function in (3.2) is based on the initial estimator $\tilde{C}_R$ of the common component matrix C^0. As demonstrated in the last subsection, one can update the estimator of C^0 via the EM algorithm and obtain a more efficient estimator of C. It is expected that using such a more efficient

estimator would yield better finite sample performance for the choice of the correct number of factors. As before, let $\hat{C}_{R,it}^{(0)} = \tilde{C}_{R,it}$ and $\ell \geq 1$ be an integer. Suppose that we have obtained the estimates $\hat{C}_{R,it}^{(\ell-1)}$. In step ℓ, we can replace the zero elements in $X^* \equiv P_{\Omega^*} X$ with the estimated common components $\hat{C}_{R_{\max},it}^{(\ell-1)}$.[5] Define the $T \times N$ matrix $\hat{X}^{*(\ell)}$ with its (t,i)th element given by

$$\hat{X}_{it}^{*(\ell)} = \begin{cases} X_{it} & \text{if } (i,t) \in \Omega^* \\ \hat{C}_{R_{\max},it}^{(\ell-1)} & \text{if } (i,t) \in \Omega_\perp^* \end{cases}, \ \ell \geq 1, \tag{3.3}$$

where $\Omega_\perp^* = \{(i,t) \in [N] \times [T] : (i,t) \notin \Omega^*\}$. Then we can conduct the singular value thresholding procedure:

$$\hat{C}_R^{(\ell)} = S_H\left(\hat{X}^{*(\ell)}, R\right) = \hat{U}_R^{(\ell)} \hat{\Sigma}_R^{(\ell)} \hat{V}_R^{(\ell)\prime}, \tag{3.4}$$

where $\hat{U}_R^{(\ell)\prime} \hat{U}_R^{(\ell)} = I_R$, $\hat{V}_R^{(\ell)\prime} \hat{V}_R^{(\ell)} = I_R$, and $\hat{\Sigma}_R^{(\ell)}$ is a diagonal matrix that contains the R largest singular values of $\hat{X}^{*(\ell)}$ arranged in descending order along its diagonal line. Following Remark 2, we recommend repeating the above procedure for $\ell = 1, ..., \ell^* \equiv \lfloor \ln(\epsilon_{NT}) / \log(p) \rfloor$ where, e.g., $\epsilon_{NT} = 0.001$. Let $\hat{C}_R = \hat{C}_R^{(\ell^*)}$ and $\hat{R} = \arg\min_{0 \leq R \leq R_{\max}} \widehat{CV}(R)$, where

$$\widehat{CV}(R) = \sum_{(i,t) \in \Omega_\perp^*} \left[X_{it} - \hat{C}_{R,it}\right]^2. \tag{3.5}$$

We will show the consistency of $\hat{R}$ under some regularity conditions.

Remark 4. Recently, Zeng, Xia, and Zhang (2019, ZXZ hereafter) proposed a double CV method to determine the number of factors consistently. Our approach differs from theirs in two aspects. First, ZXZ's method applies CV twice, first along the directions of observations (i.e., along t in our notation by K-fold CV) and then variables (i.e., along i in our notation by leave-one-out CV). In contrast, our CV method applies CV only once over random splitting of the NT observations and tends to be relatively more straightforward to implement. Second, ZXZ only consider balanced panels and their method does not apply for factor models with missing values. We show below that our method also works for factor models with missing values.

3.2 The consistency of the CV method

Let $\tilde{u}_r$ and $\tilde{v}_r$ denote the rth left and right singular vectors of X^*/p, respectively, associated with its rth largest singular value. We add one assumption.

Assumption A.7. (i) For $r = R_0+1, ..., R_{\max}$, $P(\|\tilde{u}_r\|_\infty \|\tilde{v}_r\|_\infty \leq 1/(c_0\sqrt{(N+T)\log(N+T)})) \to 1$ for some fixed $c_0 < \infty$ as $(N,T) \to \infty$, $\|\tilde{u}_r\|_\infty = o_P(1)$, and $\|\tilde{v}_r\|_\infty = o_P(1)$;

[5] We conjecture that one can replace $\hat{C}_{R_{\max},it}^{(\ell-1)}$ by $\hat{C}_{R,it}^{(\ell-1)}$ in which case $\hat{X}_{it}^{*(\ell)}$ becomes

$$\hat{X}_{R,it}^{*(\ell)} = \begin{cases} X_{it} & \text{if } (i,t) \in \Omega^* \\ \hat{C}_{R,it}^{(\ell-1)} & \text{if } (i,t) \in \Omega_\perp^* \end{cases}, \ell \geq 1.$$

But the justification for this method is far more complicated than the proof of Theorem 3.2 below because of the dependence of $\hat{X}_{R,it}^{*(\ell)}$ on R and the inconsistency of $\hat{C}_{R,it}^{(\ell-1)}$ for $R < R_0$.

(ii) $\max_{(i,t)\in\Omega_{\perp}^{*}} \sum_{(j,s)\in\Omega_{\perp}^{*}} |E\left[\varepsilon_{it}\varepsilon_{js}|P_{\Omega^{*}}X,\Omega^{*}\right]| = o_P(\delta_{NT}^{2})$.

Assumption A.7(i) is a high level condition that restricts the spikeness of singular vectors of X. A similar condition is also imposed in Negahban and Wainwright (2012). Since $\|\tilde{u}_r\|_2 = \|\tilde{v}_r\|_2 = 1$, on average each entry of $\tilde{u}_r\tilde{v}_r'$ is of the order $(NT)^{-1/2}$. We require the maximum entry is bounded by the order $((N+T)\log(N+T))^{-1/2}$. We can show that $\tilde{u}_r$ and $\tilde{v}_r$ are asymptotically equal to the $(r-R_0)$th singular vector of $\varsigma^{*} \equiv \varepsilon \circ G^{*} + F^{0}\Lambda^{0\prime} \circ [G^{*} - E(G^{*})]/p$, where each entry has zero mean. As we do not have the explicit form of $\tilde{u}_r$ and $\tilde{v}_r$, it is difficult to show its spikeness. It is well known that for an i.i.d. Gaussian random matrix, the elements of its right and left eigenvectors are uniformly distributed on the unit spheres S^{N-1} and S^{T-1}, respectively. Then Assumption A.7(i) is satisfied in this case. It is expected that the singular vectors of a general random matrix behave similarly. Assumption A.7 (ii) is a higher level condition that requires low degree of correlations among $\{\varepsilon_{it}\}$, conditional on kept-in information. It is satisfied when ε_{it} is i.i.d. and the factors and factor loadings are nonrandom. When we have $|E\left[\varepsilon_{it}\varepsilon_{js}|P_{\Omega^{*}}X,\Omega^{*}\right]| \leq M\rho^{|t-s|+|j-i|}$ for some $M < \infty$ and $\rho < 1$ perhaps after reordering the data along the cross-sectional direction, the condition is also satisfied.

The next two theorems establish the selection consistency of our CV method based on $\widetilde{CV}(R)$ and $\widehat{CV}(R)$.

Theorem 3.1 *Suppose Assumptions A.1-A.3 hold, and Assumptions A.4-A.5 hold with $g_{it} \equiv 1$. Then $P\left(\tilde{R} < R_0\right) \to 0$ as $(N,T) \to \infty$. If Assumption A.7 also holds, then $P\left(\tilde{R} > R_0\right) \to 0$ as $(N,T) \to \infty$.*

Theorem 3.2 *Suppose Assumptions A.1-A.3 hold, and Assumptions A.4-A.5 hold with $g_{it} \equiv 1$. Then $P\left(\hat{R} < R_0\right) \to 0$ as $(N,T) \to \infty$. If Assumption A.7 also holds, then $P\left(\hat{R} > R_0\right) \to 0$ as $(N,T) \to \infty$.*

Theorems 3.1 and 3.2 indicate that the CV estimators $\tilde{R}$ and $\hat{R}$ consistently estimate the true number of factors R_0 in large samples when Assumptions A.1-A.5 and A.7 hold. As we show in the proof of Theorem 3.1, the consistency of $\tilde{R}$ is established by demonstrating that

$$\begin{aligned} \widetilde{CV}(R) - \widetilde{CV}(R_0) &= (1-p)\sum_{r=R+1}^{R_0}\sigma_r^2 + O_P\left(\delta_{NT}^{-1}\right) \text{ when } R < R_0, \text{ and} \\ \text{plim}_{(N,T)\to\infty}\delta_{NT}^{2}\left[\widetilde{CV}(R) - \widetilde{CV}(R_0)\right] &\geq \frac{1-p}{256}(R-R_0)\,c_\sigma > 0 \text{ when } R > R_0, \end{aligned}$$

where c_σ is the lower probability bound of $\delta_{NT}^{2}(NT)^{-1}\tilde{\sigma}_r^{2}$ for $r \in \{R_0+1,...,R_{\max}\}$. Note that $\tilde{\sigma}_r^{2}$ diverges to infinity in probability at the rate NT for $r \in \{1,...,R_0\}$ and $(NT)^{-1}\tilde{\sigma}_r^{2}$ converges to zero in probability at the rate δ_{NT}^{-2} when $r \in \{R_0+1,...,R_{\max}\}$. Similar remarks hold true for $\widehat{CV}(R) - \widehat{CV}(R_0)$.

3.3 CV in the presence of random missing

From the proof of Theorem 3.1 we can see that the same result holds with some modifications when the original data matrix X contains random missing values. To see the modifications, we continue to

use $\Omega \subset [N]\times[T]$ to denote the index set of the observations that are observed. Let $g_{it} = \mathbf{1}\{(i,t) \in \Omega\}$ and $\tilde{q} \equiv |\Omega|/(NT)$. As before, $P(g_{it} = 1) = q \in (0,1]$ and g_{it} is independent of X, F^0, Λ^0 and ε. In this case, we consider the SVD for $\frac{1}{p\tilde{q}} P_{\Omega^*} P_{\Omega} X$:

$$\frac{1}{p\tilde{q}} P_{\Omega^*} P_{\Omega} X = \tilde{U}\tilde{\Sigma}\tilde{V}',$$

where $\tilde{U}$ is now the $T \times T$ matrix of left singular vectors of $\frac{1}{p\tilde{q}} P_{\Omega^*} P_{\Omega} X$, $\tilde{V}$ is the $N \times N$ matrix of right singular vector of $\frac{1}{p\tilde{q}} P_{\Omega^*} P_{\Omega} X$, and $\tilde{\Sigma}_R$ contains the singular values of $\frac{1}{p\tilde{q}} P_{\Omega^*} P_{\Omega} X$ arranged in descending order along its diagonal line. Then we estimate the common component C by the singular value thresholding procedure:

$$\tilde{C}_R = S_H\left(\frac{1}{p\tilde{q}} P_{\Omega^*} P_{\Omega} X, R\right) = \tilde{U}_R \tilde{\Sigma}_R \tilde{V}_R', \tag{3.6}$$

where $\tilde{U}_R$, $\tilde{V}_R$, and $\tilde{\Sigma}_R$ are defined as before. Let $\tilde{R} \in \{0, 1, 2, ..., R_{\max}\}$ minimize the following CV function

$$\widetilde{CV}(R) = \sum_{(i,t)\in\Omega^*_{\perp}\cap\Omega} \left[X_{it} - \tilde{C}_{R,it}\right]^2, \tag{3.7}$$

where $\tilde{C}_{R,it}$ denote the (t,i)th element of $\tilde{C}_R$. Following the proof of Theorem 3.1, we can also show that $P(\tilde{R} = R_0) \to 1$ as $(N,T) \to \infty$ in this case.

As in the last subsection, we can consider iterative estimates of C. Let $\hat{C}^{(0)}_{R,it} = \tilde{C}_{R,it}$. Suppose that we have obtained the estimates $\hat{C}^{(\ell-1)}_{R,it}$. In step ℓ, we can replace the zero elements in $P_{\Omega^*} P_{\Omega} X$ with the estimated common components $\hat{C}^{(\ell-1)}_{R_{\max},it}$.[6] Define the $T \times N$ matrix $\hat{X}^{*(\ell)}$ with its (t,i)th element given by

$$\hat{X}^{*(\ell)}_{it} = \begin{cases} X_{it} & \text{if } (i,t) \in \Omega \cap \Omega^* \\ \hat{C}^{(\ell-1)}_{R_{\max},it} & \text{if } (i,t) \in (\Omega \cap \Omega^*)_{\perp} \end{cases}, \ \ell \geq 1. \tag{3.8}$$

Then we can conduct the singular value thresholding procedure:

$$\hat{C}^{(\ell)}(R) = S_H\left(\hat{X}^{*(\ell)}, R\right) = \hat{U}^{(\ell)}_R \hat{\Sigma}^{(\ell)}_R \hat{V}^{(\ell)\prime}_R, \tag{3.9}$$

where $\hat{U}^{(\ell)\prime}_R \hat{U}^{(\ell)}_R = I_R$, $\hat{V}^{(\ell)\prime}_R \hat{V}^{(\ell)}_R = I_R$, and $\hat{\Sigma}^{(\ell)}_R$ is a diagonal matrix that contains the R largest singular values of $\hat{X}^{*(\ell)}$ arranged in descending order along its diagonal line. Following Remark 2, let $\hat{C}_R = \hat{C}^{(\ell^*)}_R$ and $\hat{R} = \arg\min_{0\leq R\leq R_{\max}} \widehat{CV}(R)$, where

$$\widehat{CV}(R) = \sum_{(i,t)\in\Omega^*_{\perp}\cap\Omega} \left[X_{it} - \hat{C}_{R,it}\right]^2. \tag{3.10}$$

Following the proof of Theorem 3.2, we can also show that $P(\hat{R} = R_0) \to 1$ as $(N,T) \to \infty$ in this case.

[6] We conjecture that one can replace $\hat{C}^{(\ell-1)}_{R_{\max},it}$ by $\hat{C}^{(\ell-1)}_{R,it}$ in which case $\hat{X}^{*(\ell)}_{it}$ becomes

$$\hat{X}^{*(\ell)}_{R,it} = \begin{cases} X_{it} & \text{if } (i,t) \in \Omega \cap \Omega^* \\ \hat{C}^{(\ell-1)}_{R,it} & \text{if } (i,t) \in (\Omega \cap \Omega^*)_{\perp} \end{cases}, \ell \geq 1.$$

3.4 Averaging CV and stability selection

The CV method in Sections 3.1 and 3.3 is based on a single random draw for the training set of observations. The resulting performance of the CV method can be affected by the quality of such a draw. In practice, we can always average $\widetilde{CV}(R)$ or $\widehat{CV}(R)$ over a large number (say, J) of draws.

Recognizing the notorious difficulty in the estimation of discrete structures, such as in variable selection and cluster analysis, Meinshausen and Bühlmann (2010) introduce stability selection based on subsampling in combination with some selection algorithms. The procedure serves as a general method to reduce noise by repeating the model selection many times over random splits of the data. Our CV procedure can benefit from the stability selection since it relies on random data splits. An additional benefit of stability selection in our context is that it is more robust to the choices of p and J. The algorithm is given below.

Algorithm 1 (The CV procedure)

1. For $(j,k) \in [J] \times [K]$

   (a) Randomly choose a subset of training observations $\Omega \subset [N] \times [T]$ where each observation in X can be chosen with probability p.

   (b) Apply the thresholding SVD in (3.1) or (3.6) to obtain $\tilde{C}_R$ or that in (3.4) or (3.9) to obtain $\hat{C}_R$ for $R = 0, 1, ..., R_{\max}$, respectively. Here $\tilde{C}_0$ and $\hat{C}_0$ are $T \times N$ matrices of zeros.

   (c) For each $R \in \{0, 1, ..., R_{\max}\}$, calculate the CV value via (3.2) or (3.7) and denote it as $\widetilde{CV}^{(j,k)}(R)$ or that via (3.5) or (3.10) and denote it as $\widehat{CV}^{(j,k)}(R)$.

2. Let $\widetilde{CV}_k(R) = \frac{1}{J}\sum_{j=1}^{J} \widetilde{CV}^{(j,k)}(R)$ and $\widehat{CV}_k(R) = \frac{1}{J}\sum_{j=1}^{J} \widehat{CV}^{(j,k)}(R)$ for $k = 1, ..., K$. Let

$$\tilde{R}_k = \arg \min_{0 \le R \le R_{\max}} \widetilde{CV}_k(R) \text{ and } \hat{R}_k = \arg \min_{0 \le R \le R_{\max}} \widehat{CV}_k(R) \text{ for } k = 1, ..., K.$$

   Let $\tilde{R}$ and $\hat{R}$ denote the modes in $\{\tilde{R}_1, ..., \tilde{R}_K\}$ and $\{\hat{R}_1, ..., \hat{R}_K\}$, respectively. $\tilde{R}$ and $\hat{R}$ serve as the estimator of the true number of factors without and with iterations.

We will evaluate the finite sample performance of $\tilde{R}$ and $\hat{R}$ through simulations by setting $K = 10$ and $J = 5$.

3.5 Modification of existing methods

There are many methods developed for determining the number of factors in econometrics. Popular examples are Bai and Ng's (2002) PC and IC method, Onatski's (2010) ED method, and Ahn and Horenstein's (2013) GR and ER methods. However, these methods are not directly applicable when the panel is unbalanced. Inspired by the results in Section 2, we propose some modifications to the existing methods in this subsection. Specifically, we can modify these methods in the following two ways:

- With missing observations, the balanced panel X is not directly observed. We cannot directly calculate the eigenvalues of $X'X/(NT)$, on which ED, GR and ER depend. However, our analysis in Section 3 suggests that eigenvalues of $\tilde{X}'\tilde{X}/(\tilde{q}^2NT)$ should have similar asymptotic properties to that of $X'X/(NT)$, where $\tilde{X} = X \circ G$ is observed. Then one potential valid modification of these methods is to work on the eigenvalues of $\tilde{X}'\tilde{X}/(\tilde{q}^2NT)$ directly. The idea can also be applied to modify PC and IC methods of Bai and Ng (2002). It is well known that the IC and PC methods are also closely related to the eigenvalues of $X'X/(NT)$. The objective function $V(k, \hat{F}^k)$ given by equation (7) of Bai and Ng (2002) has the property that

$$V(k, \hat{F}^k) = \sum_{r=k+1}^{N\wedge T} \mu_r[X'X/(NT)].$$

  Accordingly, we can also modify PC and IC to work on $\tilde{X}'\tilde{X}/(\tilde{q}^2NT)$.

  Although directly working on $\tilde{X}/\tilde{q}$ is asymptotically valid, this approach is not efficient and its finite sample performance can be very poor when the proportion of missing observations is high.

- To reduce the information loss, we consider the matrix completion with iterative estimates. However, we have to estimate factors and factor loadings with $R = R_{\max}$ in this case, since the number of factors is unknown. Then we work on $\check{X}$, which is completed by our proposed iterative estimates with $R_{\max}$ factors. According to the analysis in the proof of Theorem 3.2, the first R_0 singular values of $\check{X}/\sqrt{NT}$ should be bounded below by some constants and its (R_0+1)th, ..., $R_{\max}$th singular values are $O_P(\delta_{NT}^{-1})$.

  In this case, it is easy to justify that the PC and IC methods continue to work. However, it is difficult to see whether the ED, GR and ER methods can continue to work on the completed matrix because they have more requirements on the (R_0+1)th, ..., $R_{\max}$th singular values.

4 Monte Carlo Simulations

In this section, we conduct Monte Carlo simulations to evaluate the finite sample performance of our proposed EM estimators and CV method.

4.1 Data generating processes

In this subsection, we introduce the data generating processes (DGPs). Although we are considering a static factor model, the factors can be a dynamic vector process in practice (Ludvigson and Ng, 2007). In our Monte Carlo experiments, we generate the factors according to

$$F_t - \mu_f\iota_R = \rho_f(F_{t-1} - \mu_f\iota_R) + (1-\rho_f^2)^{1/2}v_t,\ t = 1, ..., T,$$

where ι_R is an $R \times 1$ vector of ones, μ_f is a scalar, v_t is independent and identically distributed (i.i.d.) from $N(0, (1-\rho_f^2)I_R)$, and $\rho_f \in (0, 1)$. To avoid the start-up effect, we throw away the first

1000 observations of $\{F_t\}$ and use the next T observations for the estimation below. For the factor loadings, we let λ_{ir}, $i = 1, ..., N$ and $r = 1, ..., R$, be i.i.d. draws from $c_s \cdot N(1, 1)$, where c_s is a constant controlling the signal strength.

Next, we introduce the generation of the idiosyncratic error terms ε_{it} in DGPs 1–5:

DGP 1. (Errors without serial or cross-sectional correlation) We let

$$\varepsilon_{it} = \left[0.9 + 0.1(\lambda_i' F_t)^2 / E(\lambda_i' F_t)^2\right] u_{it},$$

where u_{it} is i.i.d. from $t(5)$, the Student's t-distribution with 5 degrees of freedom. Apparently, we allow for conditional heteroskedasticity but no serial or cross-sectional correlation among ε_{it}'s.

DGP 2. (AR(1) errors across time) We generate ε_{it} via an AR(1) process: $\varepsilon_{it} = \rho\varepsilon_{i,t-1} + u_{it}$, where u_{it} is i.i.d. $N\left(0, 1-\rho^2\right)$ and $\rho = 0.5$. To eliminate the start-up effect of the initial value, we throw away the first 100 observations.

DGP 3. (MA(1) errors across individuals) We generate $\varepsilon_{it} = u_{it} + u_{i-1.t}$, where u_{it} is i.i.d. $N(0, 1/\sqrt{2})$.

DGP 4. (Errors with both serial and cross-sectional dependence) We generate $\varepsilon_{it} = u_{it} + bu_{i,t-1} + bu_{i-1,t} + b^2 u_{i-1,t-1}$, where u_{it} is i.i.d. $N(0, 1)$ and $b = 0.3$.

DGPs 1-4 satisfy all assumptions in the paper. We employ them to evaluate the finite sample performance of the proposed estimators and CV methods under various scenarios. With the presence of cross-sectional dependence in the error terms in DGPs 3-4, $\left\{\lambda_i^0 \varepsilon_{it} g_{it}\right\}$ are also cross-sectionally correlated. The covariance estimator $\hat{\Gamma}_{1g,t}^{(1)}$ introduced in Section 2.4 is invalid. We can use these two DGPs to check the performance of robust inference procedure using the estimator $\hat{\Gamma}_{1g,t}^{(2)}$.

DGP 5. (Errors with a fat-tailed distribution) The setting for ε_{it} is the same as DGP 1 except that u_{it} is i.i.d. from Student's t-distribution with 3 degrees of freedom.

In this case, the error term ε_{it} does not have a finite fourth moment, which violates Assumption A.2(i). DGP 5 can be used to check the robustness of the proposed estimators and the CV method in the fat-tailed error scenario.

In all experiments, we fix the number of factors to be 3. We set $\mu_f = 0.6$, $\rho_f = 0.3$ and choose c_s such that the signal to noise ratio (SNR) equals 4 for each DGP. Specifically, we define SNR as $\text{var}(\lambda_i' F_t)/\text{var}(\varepsilon_{it})$. For each DGP, we consider $N = 50, 100$ and $T = 50, 100$. The number of replications is 1000 in all cases.

Table 1: Under/Over-estimation rate (%) with complete data

DGP	N	T	$\widetilde{CV}$	ED	GR	ER	PC	IC
1	50	50	2.9/0.0	0.0/4.9	27.8/0.0	78.6/0.0	0.0/9.9	0.0/3.2
	50	100	0.1/0.0	0.0/3.4	6.9/0.2	63.1/0.0	0.0/3.5	0.0/2.4
	100	50	0.1/0.0	0.0/3.1	8.7/0.0	59.1/0.0	0.0/2.8	0.0/1.7
	100	100	0.0/0.0	0.0/2.2	0.2/0.0	24.2/0.0	0.0/1.0	0.0/0.6
2	50	50	0.8/0.0	0.0/2.8	41.4/0.0	86.7/0.0	0.0/85.0	0.0/17.7
	50	100	0.1/0.0	0.0/0.1	10.2/0.0	67.6/0.0	0.0/2.1	0.0/0.1
	100	50	0.0/0.0	0.0/0.5	21.6/0.0	80.0/0.0	0.0/74.0	0.0/10.4
	100	100	0.0/0.0	0.0/0.0	0.5/0.0	40.1/0.0	0.0/0.3	0.0/0.0
3	50	50	0.8/0.0	0.0/0.3	3.0/0.0	47.2/0.0	0.0/28.1	0.0/0.5
	50	100	0.0/0.0	0.0/0.0	0.0/0.0	21.3/0.0	0.0/1.1	0.0/0.0
	100	50	0.0/0.0	0.0/0.0	0.0/0.0	17.2/0.0	0.0/0.0	0.0/0.0
	100	100	0.0/0.0	0.0/0.0	0.0/0.0	2.1/0.0	0.0/0.0	0.0/0.0
4	50	50	1.4/0.0	0.0/0.3	26.7/0.0	77.4/0.0	0.0/9.2	0.0/0.1
	50	100	0.0/0.0	0.0/0.1	7.1/0.0	59.1/0.0	0.0/0.1	0.0/0.0
	100	50	0.1/0.0	0.0/0.0	6.9/0.0	59.3/0.0	0.0/0.0	0.0/0.0
	100	100	0.0/0.0	0.0/0.0	0.2/0.0	23.5/0.0	0.0/0.0	0.0/0.0
5	50	50	4.7/3.0	0.0/35.1	44.2/3.2	84.4/0.6	0.0/59.6	0.0/38.6
	50	100	0.4/3.2	0.0/33.4	27.7/4.5	76.9/1.0	0.0/43.4	0.0/32.6
	100	50	0.4/3.5	0.0/33.4	27.1/4.8	73.5/1.5	0.0/45.0	0.0/33.0
	100	100	0.1/1.9	0.0/32.6	13.6/3.7	52.5/1.3	0.0/36.6	0.0/29.8

Note: We report the under/over-estimation rate with complete data. We consider $\widetilde{CV}$ with leave-out probability $p = 0.9$, ED of Onatski (2010), GR and ER of Ahn and Horenstein (2013), and PC and IC of Bai and Ng (2002). The number of replications is 1000.

4.2 Simulation results

In this subsection, we present our simulation results in two parts. In the first part, we examine the accuracy of the CV method proposed in section 3, measured by the empirical rate of correct determination of the number of factors. In the second part, we estimate the model with the estimated number of factors and report the finite sample performance of the proposed estimators in Section 2.

4.2.1 Determining the number of factors

In this subsection, we use the CV method to determine the number of factors for data with or without random missing observations. For both cases, we let $R_{\max} = 5$ and use the leave-out probability $p = 0.9$. To implement the averaging CV and stability selection method in Section 3.4, we set $K = 10$ and $J = 5$. For the case of incomplete data, we consider two random missing probabilities: $q = 0.7, 0.9$.

When the original data forms a balanced panel, there are existing methods including the growth ratio (GR) and eigenvalue ratio (ER) of Ahn and Horenstein (2013), the edge distribution (ED) of Onatski (2010) and the PC and IC methods of Bai and Ng (2002), among others. We also report the performance of these methods for comparison. Table 1 presents the under/over-estimation rate with complete data. We summarize some important findings from Table 1. First, for DGP 1 where the error terms have neither serial nor cross-sectional dependence, all the methods under investigation

show a pattern of convergence as both N and T increases, and the CV method with $p = 0.9$ obviously outperforms all the other methods. Second, for DGPs 2-4 with serial or cross-sectional dependence in errors, the performance of various methods are similar to that for DGP 1. Among all the methods under study, ER, PC and IC tend to be outperformed by the CV and ED methods. Third, for DGP 5, with fat-tail distributed errors, our CV method tends to outperform all existing methods by a big margin. Specifically, ED, PC and IC over-estimated more than 29.8% times for all four combinations of N and T, and GR and ER tend to under-estimate the number of factors. From the performance of these five existing methods, we observe slow rate of convergence. The result for DGP 5 indicates the CV method is somewhat robust to error terms with fat tails.

When the original data has random missing observations, we follow the discussion in Section 3.5 to modify the existing methods and compare them with our CV method. In our simulation studies, we have explored both approaches introduced in Section 3.5. To save space, we only report the results of the best performed approach. Specifically, we modify these methods as follows:

M-ED, **M-GR**, and **M-ER**: We directly use ED, GR and ER algorithms to work on the eigenvalues of $\tilde{X}'\tilde{X}/(\tilde{q}^2NT)$.

M-PC and **M-IC**: We adopt an approach which is slightly different from what is introduced in Section 3.5. The algorithms are given by:

1. Fix $R = 1, ..., R_{\max}$ and conduct the following steps:
   (a) Obtain the iterated estimates $\hat{C}_{R,it}^{(\ell^*)}$ as in Section 2.2;
   (b) Calculate the modified objective function: $\tilde{V}(R) = \frac{1}{|\Omega|}\sum_{(i,t)\in\Omega}(X_{it} - \hat{C}_{R,it}^{(\ell^*)})^2$;
   (c) The criteria functions are of the form $\tilde{V}(R) + R \cdot g(N,T)$, where $g(N,T)$ is the penalty function for PC and IC in Bai and Ng (2002).
2. Choose R that minimizes the criterion function.

Table 2 presents the under/over-estimation rate with incomplete data over 1000 Monte Carlo replications for $q = 0.7$. The case for $q = 0.9$ is reported in Table A1 in the additional online supplement. We consider the two CV methods discussed in Section 3.4, namely, $\widetilde{CV}(R)$ and $\widehat{CV}(R)$ with $\hat{C}_{R_{\max},it}^{(\ell-1)}$ used in the ℓth iteration. As in Remark 2, we stop the iterations when $\ell = \ell^*$. The results are reported in the '$\widetilde{CV}$' and '$\widehat{CV}$' columns of Table 2.

We summarize some important findings from Table 2. First, both CV methods yield decreasing percentage of under/over-estimation rate as either N or T increases, and $\widehat{CV}$ has better finite sample performance than $\widetilde{CV}$. Therefore, using iterations to complete missing observations helps improve the finite sample performance of the CV method. Second, for the other competing methods, only M-ED performs well for all DGPs with large N and T, but its finite sample performance is not as good as $\widehat{CV}$. Under the random missing case and the assumptions in Onatski (2000), it is possible to justify M-ED theoretically. Third, M-GR and M-ER severely under-estimate the number of factors for all DGPs whereas M-PC and M-IC slightly over-estimate the number of factors in small samples

Table 2: Under/Over-estimation rate (%) with missing data ($q = 0.7$)

DGP	N	T	$\widetilde{CV}$	$\widehat{CV}$	M-ED	M-GR	M-ER	M-PC	M-IC
1	50	50	63.8/0.0	9.2/0.0	49.5/9.5	99.4/0.0	99.8/0.0	0.0/67.8	0.0/31.4
	50	100	26.4/0.1	0.1/0.0	26.8/4.9	99.4/0.0	100.0/0.0	0.0/30.7	0.0/11.9
	100	50	33.2/0.1	0.4/0.0	30.0/6.7	98.9/0.0	99.7/0.0	0.0/27.9	0.0/10.5
	100	100	1.8/0.0	0.0/0.0	4.1/1.3	98.7/0.0	99.8/0.0	0.0/9.2	0.0/4.7
2	50	50	55.3/0.0	3.4/0.0	43.7/6.5	98.9/0.0	99.7/0.0	0.0/74.7	0.0/44.1
	50	100	17.3/0.1	0.1/0.0	20.2/1.9	99.2/0.0	99.8/0.0	0.0/30.9	0.0/5.9
	100	50	24.3/0.0	0.2/0.0	24.1/3.3	98.1/0.0	99.8/0.0	0.0/63.0	0.0/34.4
	100	100	0.9/0.0	0.0/0.0	1.8/0.6	98.4/0.0	99.9/0.0	0.0/19.2	0.0/2.6
3	50	50	58.8/0.0	3.3/0.0	46.3/6.7	99.4/0.0	100.0/0.0	0.0/59.2	0.0/23.6
	50	100	20.1/0.0	0.2/0.0	21.0/1.7	98.3/0.0	99.6/0.0	0.0/37.0	0.0/6.7
	100	50	29.0/0.0	0.2/0.0	26.9/5.4	98.8/0.0	99.8/0.0	0.0/12.2	0.0/0.7
	100	100	0.9/0.0	0.0/0.0	1.7/0.9	99.1/0.0	99.8/0.0	0.0/1.0	0.0/0.0
4	50	50	57.5/0.1	4.3/0.0	43.3/7.4	98.6/0.0	99.7/0.0	0.0/45.7	0.0/12.3
	50	100	20.3/0.0	0.0/0.0	20.5/2.9	99.3/0.0	99.9/0.0	0.0/10.6	0.0/0.8
	100	50	26.4/0.0	0.4/0.0	23.8/4.6	98.6/0.0	99.9/0.0	0.0/10.3	0.0/0.9
	100	100	0.9/0.0	0.0/0.0	2.1/1.1	98.4/0.0	99.9/0.0	0.0/0.0	0.0/0.0
5	50	50	63.1/1.1	12.9/1.8	47.6/13.8	99.0/0.1	99.6/0.1	0.0/87.2	0.0/71.2
	50	100	27.2/2.7	0.8/2.1	27.6/10.2	99.7/0.0	100.0/0.0	0.0/79.2	0.0/63.0
	100	50	33.2/2.7	1.6/2.7	29.9/11.7	99.1/0.1	99.8/0.0	0.0/73.9	0.0/60.7
	100	100	2.6/4.4	0.0/1.3	5.1/8.0	98.7/0.2	99.9/0.0	0.0/69.0	0.0/59.5

Note: We report the under/over-estimation rate with missing data, where each entry is observed with probability $q = 0.7$. We consider $\widetilde{CV}$ and $\widehat{CV}$ with leave-out probability $p = 0.9$. For comparison, we also consider the M-ED,M-ER, M-PC, and M-IC, which are modified from ED of Onatski (2010), GR and ER of Ahn and Horenstein (2013), and PC and IC of Bai and Ng (2002), respectively. The number of replications is 1000.

in DGPs 1-4 and severely over-estimate the number of factors in DGP 5. Fourth, when the proportion of missing observations is small ($q = 0.9$ in Table A1 in the online supplement), the performance of all methods improved remarkably but our $\widehat{CV}$ method still outperforms all other methods in all cases. To sum up, our CV methods have great finite sample performance compared to the modified existing methods.

4.2.2 Estimation of Λ and F

In this subsection, we work on the scenario with random missing observations where $q = 0.7$ and 0.9. We estimate the factors and factor loadings using the method introduced in Section 2 and make inferences on the factors. Specifically, we consider the initial estimates ($\ell = 0$) and the ℓ^*th step estimates. For comparison, we also report the oracle estimate that is obtained with the knowledge of missing observations and the correct number of factors.

Tables 3 reports the feasible estimation results for $q = 0.7$, where the number of factors is determined by $\widehat{CV}$. As Table 2 indicates, the $\widehat{CV}$-based estimate, $\hat{R}$, is quite accurate. But we still have both nonnegligible percentages of under- and over-estimation when $(N, T) = (50, 50)$.

To measure the consistency of our estimates, we calculate the mean squared error (MSE) of $\hat{C}_{it}$'s. The results are reported in the *'MSE'* columns of Table 3. We observe that the MSE of the

Table 3: Mean squared error and coverage probability of confidence intervals with missing data ($q = 0.7$)

			MSE of $\hat{C}_{it}$			CP of Standard CI			CP of Robust CI		
DGP	N	T	$\hat{C}_{it}^{(oracle)}$	$\hat{C}_{it}^{(0)}$	$\hat{C}_{it}^{(\ell^*)}$	$\hat{F}_t^{(oracle)}$	$\hat{F}_t^{(0)}$	$\hat{F}_t^{(\ell^*)}$	$\hat{F}_t^{(oracle)}$	$\hat{F}_t^{(0)}$	$\hat{F}_t^{(\ell^*)}$
1	50	50	0.250	1.445	0.437	91.1%	96.2%	86.9%	93.4%	98.3%	91.2%
	50	100	0.186	0.993	0.290	91.3%	94.3%	89.6%	93.3%	96.4%	92.5%
	100	50	0.187	1.039	0.294	91.6%	95.9%	91.2%	92.8%	97.7%	92.2%
	100	100	0.124	0.613	0.185	94.0%	95.5%	91.8%	95.2%	96.6%	94.1%
2	50	50	0.284	1.390	0.420	87.6%	94.8%	84.2%	90.6%	97.8%	88.7%
	50	100	0.197	0.932	0.282	89.4%	95.7%	87.7%	92.3%	97.0%	91.0%
	100	50	0.229	0.996	0.312	88.3%	96.0%	87.8%	88.6%	98.1%	91.0%
	100	100	0.142	0.604	0.193	92.3%	96.9%	92.0%	92.9%	98.1%	93.4%
3	50	50	0.248	1.369	0.392	82.5%	93.4%	80.5%	89.1%	97.2%	85.9%
	50	100	0.192	0.926	0.278	81.8%	92.8%	82.3%	91.1%	96.3%	88.3%
	100	50	0.176	0.962	0.263	80.2%	94.3%	80.1%	86.5%	96.8%	84.4%
	100	100	0.121	0.588	0.173	83.5%	93.1%	86.4%	90.7%	96.1%	90.4%
4	50	50	0.255	1.367	0.398	82.3%	93.7%	81.8%	86.5%	97.9%	86.6%
	50	100	0.187	0.934	0.273	84.1%	94.4%	85.1%	87.9%	97.6%	89.2%
	100	50	0.193	0.975	0.280	86.4%	95.8%	87.9%	89.0%	97.4%	90.3%
	100	100	0.127	0.591	0.179	86.4%	93.4%	87.0%	89.8%	95.4%	90.1%
5	50	50	0.319	1.547	0.524	91.9%	96.9%	87.1%	93.0%	98.5%	91.4%
	50	100	0.255	1.123	0.375	92.1%	94.4%	90.2%	93.3%	96.2%	93.1%
	100	50	0.258	1.160	0.375	92.4%	96.1%	91.2%	93.7%	98.2%	92.1%
	100	100	0.156	0.689	0.230	94.6%	95.7%	92.4%	95.3%	96.9%	93.8%

Note: We report the mean squared errors (MSE) of $\hat{C}_{it}$ and the coverage probabilities (CP) of the 95% confidence intervals (CIs) for F_t^0's. Each entry is observed with probability $q = 0.7$. We consider the feasible estimates with $\ell = 0$ and $\ell = \ell^*$, and the oracle estimate that is obtained using the information of R^0 and missing observations. The standard CIs and the robust CIs are constructed using $\hat{\Gamma}_{1g,t}^{(1)}$ and $\hat{\Gamma}_{1g,t}^{(2)}$ in Section 2.4, respectively.

ℓ^*th-step estimator is smaller than that of the initial estimator. However, as expected, the feasible estimators that use random missing data are less efficient than the oracle estimator. This is consistent with Remark 3 in Section 2.3. In fact, despite the presence of serial dependence, or cross-sectional dependence, or both in DGPs 2-4, the MSE of the ℓ^*th step estimator is approximately equal to that of the oracle estimator multiplied by $1/q$ in DGPs 1-4. DGP 5 is an exception because of the violation of the moment conditions on the error terms.

Applied researchers are often interested in obtaining the estimates of the first eigenvector of $F^0\Lambda^{0\prime}\Lambda^0F^0$, which asymptotically equals to the first column of F^0H, when the associated eigenvalues are ordered in descending order. In our simulations, we conduct inference on $(F^0H)_{t^*1}$, which is the t^*th entry of the first column of F^0H. In each replication, we randomly choose a t^* and construct confidence intervals of $(F^0H)_{t^*1}$. Following the results of Theorem 2.2 and Theorem 2.4, the 95% confidence interval (CI) is given by

$$[\hat{F}^{(\ell)}_{t^*1} - 1.96([\hat{\Sigma}^{(\ell)}_{F_t}]_{11})^{1/2}/\sqrt{N}, \hat{F}^{(\ell)}_{t^*1} + 1.96([\hat{\Sigma}^{(\ell)}_{F_t}]_{11})^{1/2}/\sqrt{N}].$$

To estimate the covariance matrix, we consider both the standard covariance matrix estimate based on $\hat{\Gamma}^{(1)}_{1g,t}$ and the robust one based on $\hat{\Gamma}^{(2)}_{1g,t}$ introduced in Section 2.4.

We refer to the 95% confidence intervals constructed using $\hat{\Gamma}^{(1)}_{1g,t}$ and $\hat{\Gamma}^{(2)}_{1g,t}$ as standard CI and robust CI, respectively. In Table 3, we report the coverage probability (CP) of the standard CIs and robust CIs for both the oracle estimate and our feasible estimates for $q = 0.7$. We summarize the main findings. First, for DGPs 1–2 and 5 where there is no cross-sectional dependence in the error terms, both standard and robust covariance estimators provide asymptotically valid inferences. The CP approaches the nominal CP as the sample size increases. Second, for DGPs 3-4 where there is cross-sectional dependence in the error terms, the standard CIs are invalid, and the standard CIs based on both the oracle estimates and the ℓ^*th step estimates tend to under-cover the true parameters. In contrast, the robust CI method provides asymptotically valid inference and yields remarkably better performance. This suggests that ignoring the cross-sectional dependence may lead to the incorrect estimation of the standard errors of the factor estimator and one should incorporate the cross-sectional dependence in the estimation of the standard errors. Third, the CP for the CIs based on the initial estimates tend to over-cover the true values. However, we should refrain from making an inference based on the initial estimates in large samples. This is because the initial estimates tend to have a larger variance than the ℓ^*th step estimates, leading to much wider CIs compared to those based on the ℓ^*th step estimates.

In Table A2 of the online supplement, we also report the simulation results for the feasible estimates that are associated with the case $q = 0.9$, i.e., only 10% observations are missing at random. In addition, we report the estimation and inference results of the infeasible estimates, which are obtained by using the correct number of factors, in Tables A3 and A4 of the online supplement.

5 Empirical Application: Forecasting Macroeconomic Variables

In this section, we show the usefulness of the proposed method by considering factor-augmented regressions to forecast macroeconomic variables. The procedure starts from estimating a set of latent factors using panel data. In practice, some variables have missing observations due to a short collection history or lagged publications. A simple and frequently used method to deal with this problem is to delete those units/variables with missing observations to obtain a balanced panel and the PC estimators of latent factors (PC-F). Inevitably, we lose some useful information by doing so. To fully exploit the information of predictors with missing observations, we can use the EM estimators to estimate the latent factors (EM-F). In our application, we use EM-F or PC-F to forecast macroeconomic variables, respectively. Then we show that EM-F outperforms PC-F in terms of predictive MSE.

Below we consider the forecasts of the U.S. real gross domestic product (RGDP), gross domestic product (GDP), industrial production (IP) and real disposal personal income (RDPI) at 1, 2 and 4 quarter horizons. These four time series are collected from the Federal Reserve Bank website.

5.1 Implementation

We use a panel dataset FRED-QD, which is an unbalanced panel at the quarterly frequency. FRED-QD is a quarterly frequency companion of FRED-MD that is introduced by McCracken and Ng (2016, MN hereafter). The dataset consists of 248 quarterly U.S. indicators from 1959Q1 to 2018Q2. We use 125 time series that are used in SW to estimate the latent factors. Despite the fact that the missing is not exactly random in this application, we demonstrate that the appropriate use of the EM algorithm can outperform the simple approach by deleting those individuals with missing observations in terms of out-of-sample prediction.

We take 1960Q1 as the start of the sample. Then we lose two periods of observations due to data transformations as in MN and obtain an unbalanced panel with $(T, N) = (236, 125)$. There are 37 variables containing 1594 missing observations in total. Following the lead of MN, we check for outliers in each variable where an outlier is defined as an observation that deviates from the observed sample median by more than 10 times interquantile range. The outliers are removed and treated as missing observations. As a result, the total number of missing observations becomes 1602 ($\hat{q} = 0.946$). All columns of the data matrix X are standardized to have zero mean and unit standard deviation before estimating EM-F. To estimate PC-F, we drop 37 variables with missing observations to obtain a balanced panel with $(T, N) = (236, 88)$. We also standardize the balanced panel before estimating PC-F. We estimate the first factor by PC and EM and use them to do the out-of-sample forecasting.

Next, we consider the forecast based on the following factor-augmented autoregression (FA-AR) models:

$$y_{t+h}^h = \phi_h^{(1)} + \phi_h^{(2)}(L)\hat{F}_t + \phi_h^{(3)}(L)y_t + \varepsilon_{t+h}^h, \ h = 1, 2, \ 4, \tag{5.1}$$

where y_t is one of the four macro-variables (i.e., RGDP, GDP, IP, and RDPI), $\hat{F}_t$ is the estimated

Table 4: Comparison of forecast results

		Real GDP			GDP			IP			RDPI		
		MSE	ratio		MSE	ratio		MSE	ratio		MSE	ratio	
period	horizon		AR	PC-F		AR	PC-F		AR	PC-F		AR	PC-F
1987	h=1	4.571	0.923	0.985	6.665	0.921	1.004	11.488	0.911	0.929	11.896	0.958	0.988
|	h=2	2.986	0.853	0.968	5.349	0.874	1.003	13.091	0.896	0.922	4.505	0.888	0.985
2016	h=4	2.683	0.948	0.927	5.727	0.940	0.996	13.489	0.969	0.994	2.565	0.841	0.989
1997	h=1	4.734	0.870	1.009	6.745	0.892	1.000	12.131	0.853	0.896	14.982	0.957	0.987
|	h=2	3.246	0.813	0.957	5.531	0.851	0.998	15.583	0.875	0.918	5.085	0.856	0.995
2016	h=4	3.020	0.924	0.955	5.924	0.916	0.997	16.964	0.948	0.983	2.832	0.809	0.983
2007	h=1	5.049	0.746	0.982	8.170	0.794	0.984	16.818	0.805	0.862	20.446	0.941	0.982
|	h=2	4.247	0.749	0.922	7.167	0.801	1.004	23.777	0.851	0.886	6.565	0.785	0.985
2016	h=4	4.445	0.901	0.950	8.145	0.923	1.011	26.810	0.904	0.936	4.047	0.777	0.973

Note: We report the MSE of forecasts using EM-F and its ratios to the MSEs associated with the AR or FA-AR using PC-F.

vector of factors, $\phi_h^{(1)}$ is the intercept term, L is the lag operator, and $\phi_h^{(2)}(L)$ and $\phi_h^{(3)}(L)$ are finite-order polynomials of the lag operators. For all four variables to be forecasted, we treat them as $I(1)$ series and define the dependent variable as average annualized quarterly growth rate. As an example, for IP, we define

$$y_{t+h}^h = (400/h)\ln(IP_{t+h}/IP_t) \text{ and } y_t = 400\ln(IP_t/IP_{t-1}).$$

All the models are estimated recursively by ordinary least squares (OLS). We use BIC to select the number of autoregressive lags (from 0 to 6) and lags of the first factor (from 1 to 6) in EM-F and PC-F, respectively.

5.2 Forecast results

We consider three out-of-sample periods, namely, 1987Q1-2016Q4, 1997Q1-2016Q4 and 2007Q1-2016Q4. Table 4 reports the MSE of forecasts using EM-F and its ratio to the MSE associated with the autoregression (AR) or FA-AR using PC-F, where the AR model is used with $\hat{F}_t$ absent in (5.1) and the number of lags are also determined by the Bayesian information criterion (BIC). Ratios smaller than one are in favor of the method using EM-F. For all the four macroeconomic variables under investigation, the forecasts using EM-F outperforms those only using autoregression. Therefore, we can conclude that the estimated latent factors contain some predictive power. For Real GDP, IP and RDPI, the forecast using EM-F provides smaller MSE for almost all horizons and periods compared to that using PC-F. For GDP, we can see that the forecasts using EM-F and PC-F have comparable performance. In short, the EM estimation of the factors generally help the out-of-sample forecast of some major macroeconomic variables.

6 Conclusion

In this paper we study the asymptotic properties of the EM estimators of factors and factor loadings in an approximate factor model with random missing. Based on the asymptotic results, we also propose a novel cross-validation method to determine the number of factors in factor models with or

without random missing observations. Simulations demonstrate the good finite sample performance of the proposed method and empirical applications suggest the usefulness of our method.

The paper can be extended in various directions. First, we only consider random missing and it is possible to extend our method to allow for heterogenous missing or missing with certain patterns. Second, we focus on a pure approximate factor model and one may consider the extension to the panel data models with multi-factor error structure and random missing values (see, Bai et al. (2015) and Athey et al. (2018)) or factor-augmented vector-autoregressive (FAVAR) models with missing values. We are exploring some of these topics in ongoing works.

APPENDIX

A Proofs of the main results in Section 2

In this appendix, we prove the main results in Section 2 by calling upon some technical lemmas whose proofs can be found in the online supplement. For notational simplicity, we use $\tilde{F}$, $\tilde{\Lambda}$, $\tilde{C}$, $\tilde{D}$, $\tilde{H}$, $\tilde{F}_t, \tilde{\lambda}_i$ and $\tilde{C}_{it}$ to denote $\hat{F}^{(0)}$, $\hat{\Lambda}^{(0)}$, $\hat{C}^{(0)}$, $\hat{D}^{(0)}$, $\hat{H}^{(0)}$, $\hat{F}_t^{(0)}, \hat{\lambda}_i^{(0)}$ and $\hat{C}_{it}^{(0)}$, respectively.

To prove Theorem 2.1, we need the following lemma.

Lemma A.1 *Suppose that Assumptions A.1-A.2 hold. Then* $T^{-1}\tilde{F}'\left(NT\tilde{q}^2\right)^{-1}\tilde{X}\tilde{X}'\tilde{F} = \tilde{D} = D + \delta_{NT}^{-(1-\gamma/2)}$, *where* $\gamma = \gamma_1 \vee \gamma_2$, $\tilde{D}$ *is an* $R \times R$ *diagonal matrix consisting of the* R *largest eigenvalues of* $\left(NT\tilde{q}^2\right)^{-1}\tilde{X}\tilde{X}'$, *and* D *is an* $R \times R$ *matrix consisting of the* R *eigenvalues of* $\Sigma_{\Lambda^0}\Sigma_{F^0}$, *arranged in descending order along the diagonal line.*

Proof of Theorem 2.1. From the principal component analysis, we have the identity $\left(NT\tilde{q}^2\right)^{-1}\tilde{X}\tilde{X}'\tilde{F} = \tilde{F}\tilde{D}$. By Lemma A.1 and Assumption A.1, $\tilde{D}$ is asymptotically nonsingular so that we can post-multiply both sides by $\tilde{D}^{-1}$ to obtain $\tilde{F} = \left(NT\tilde{q}^2\right)^{-1}\tilde{X}\tilde{X}'\tilde{F}\tilde{D}^{-1}$. Recall that $\tilde{H} = \left(N^{-1}\Lambda^{0\prime}\Lambda^0\right)^{-1}T^{-1}F^{0\prime}\tilde{F}\tilde{D}^{-1}$. Noting that the (t,i)th element of $\tilde{X}$ is given by $\tilde{X}_{it} = \left(\lambda_i^{0\prime}F_t^0 + \varepsilon_{it}\right)g_{it} = \lambda_i^{0\prime}F_t^0 q + \varepsilon_{it}g_{it} + \lambda_i^{0\prime}F_t^0\left(g_{it} - q\right)$, we have

$$\begin{aligned}
\tilde{F}_t - \tilde{H}'F_t^0 &= \frac{1}{NT\tilde{q}^2}\tilde{D}^{-1}\sum_{s=1}^{T}\tilde{F}_s\sum_{i=1}^{N}\{E\left(\varepsilon_{is}\varepsilon_{it}\right)g_{is}g_{it} + \left[\varepsilon_{is}\varepsilon_{it} - E\left(\varepsilon_{is}\varepsilon_{it}\right)\right]g_{is}g_{it} \\
&\quad + F_s^{0\prime}\lambda_i^0\varepsilon_{it}g_{is}g_{it} + F_t^{0\prime}\lambda_i^0\varepsilon_{is}g_{is}g_{it} + F_s^{0\prime}\lambda_i^0\lambda_i^{0\prime}F_t^0\left(g_{is} - q\right)q \\
&\quad + F_s^{0\prime}\lambda_i^0\lambda_i^{0\prime}F_t^0\left(g_{it} - q\right)q + F_s^{0\prime}\lambda_i^0\lambda_i^{0\prime}F_t^0\left(g_{is} - q\right)\left(g_{it} - q\right)\} + O_p((NT)^{-1/2}) \\
&\equiv a_{1t} + a_{2t} + ... + a_{7t} + O_p((NT)^{-1/2}), \qquad \text{(A.1)}
\end{aligned}$$

where, e.g., $a_{1t} = \frac{1}{NT\tilde{q}^2}\tilde{D}^{-1}\sum_{s=1}^{T}\tilde{F}_s\sum_{i=1}^{N}E\left(\varepsilon_{is}\varepsilon_{it}\right)g_{is}g_{it}$ and the first equality used the fact $\tilde{q} - q = O_p((NT)^{-1/2})$. It follows that $T^{-1}\sum_{t=1}^{T}\left\|\tilde{F}_t - \tilde{H}'F_t^0\right\|^2 \leq 7\sum_{l=1}^{7}T^{-1}\sum_{t=1}^{T}\|a_{lt}\|^2 + O_p((NT)^{-1/2})$ by the Cauchy-Schwarz (CS) inequality. For a_{1t}, we have

$$\begin{aligned}
T^{-1}\sum_{t=1}^{T}\|a_{1t}\|^2 &\leq \left\|\tilde{D}^{-1}\right\|^2 T^{-1}\sum_{t=1}^{T}\left\|\frac{1}{T\tilde{q}^2}\sum_{s=1}^{T}\tilde{F}_s\frac{1}{N}\sum_{i=1}^{N}E\left(\varepsilon_{is}\varepsilon_{it}\right)g_{is}g_{it}\right\|^2 \\
&\leq \frac{1}{T\tilde{q}^4}\left\|\tilde{D}^{-1}\right\|^2\frac{1}{T}\sum_{s=1}^{T}\left\|\tilde{F}_s\right\|^2\frac{1}{T}\sum_{s=1}^{T}\sum_{t=1}^{T}\left|\frac{1}{N}\sum_{i=1}^{N}E\left(\varepsilon_{is}\varepsilon_{it}\right)g_{is}g_{it}\right|^2 \\
&\leq \frac{R}{T\tilde{q}^4}\left\|\tilde{D}^{-1}\right\|^2\frac{1}{T}\sum_{s=1}^{T}\sum_{t=1}^{T}\left|\gamma_N\left(s,t\right)\right|^2 = O_P\left(T^{-1}\right),
\end{aligned}$$

where the second inequality follows from the CS inequality and the third inequality follows from the fact that $\frac{1}{T}\sum_{s=1}^{T}\left\|\tilde{F}_s\right\|^2 = \frac{1}{T}\text{tr}(\tilde{F}'\tilde{F}) = \text{tr}(I_R) = R$ and that $|g_{it}| \leq 1$, and the last equality holds by Assumption A.2. Similarly, for a_{2t}, we have $T^{-1}\sum_{t=1}^{T}\|a_{2t}\|^2 \leq \frac{R}{T\tilde{q}^4}\left\|\tilde{D}^{-1}\right\|^2\frac{1}{T}\sum_{s=1}^{T}\sum_{t=1}^{T}\zeta_{1g,st}^2$,

where $\zeta_{1g,st} = \frac{1}{N}\sum_{i=1}^{N}[\varepsilon_{is}\varepsilon_{it} - E(\varepsilon_{is}\varepsilon_{it})]g_{is}g_{it}$. Noting that $\zeta_{1g,st} = \frac{1}{N}\sum_{i=1}^{N}[\varepsilon_{is}\varepsilon_{it} - E(\varepsilon_{is}\varepsilon_{it})]$ $\{q^2 + (g_{is}-q)q + (g_{it}-q)q + (g_{is}-q)(g_{it}-q)\} \equiv \sum_{l=1}^{4}\zeta_{1g,stl}$, where, e.g., $\zeta_{1g,st1} = \frac{1}{N}\sum_{i=1}^{N}[\varepsilon_{is}\varepsilon_{it} - E(\varepsilon_{is}\varepsilon_{it})]q^2$, we have $\zeta_{1g,st}^2 \leq 4\sum_{l=1}^{4}\zeta_{1g,stl}^2$. By Assumption A.2,

$$\begin{aligned}
\frac{1}{T}\sum_{s=1}^{T}\sum_{t=1}^{T}E(\zeta_{1g,st,1}^2) &= \frac{q^4}{TN}\sum_{s=1}^{T}\sum_{t=1}^{T}E\left[\frac{1}{N^{1/2}}\sum_{i=1}^{N}[\varepsilon_{is}\varepsilon_{it} - E(\varepsilon_{is}\varepsilon_{it})]\right]^2 = O(T/N), \\
\frac{1}{T}\sum_{s=1}^{T}\sum_{t=1}^{T}E(\zeta_{1g,st,2}^2) &= \frac{q^2}{T}\sum_{s=1}^{T}\sum_{t=1}^{T}E\left[\frac{1}{N}\sum_{i=1}^{N}[\varepsilon_{is}\varepsilon_{it} - E(\varepsilon_{is}\varepsilon_{it})](g_{is}-q)\right]^2 \\
&= \frac{q^2}{TN^2}\sum_{s=1}^{T}\sum_{t=1}^{T}\sum_{i=1}^{N}E[\varepsilon_{is}\varepsilon_{it} - E(\varepsilon_{is}\varepsilon_{it})]^2 = O(T/N).
\end{aligned}$$

Similarly, we can show that $\frac{1}{T}\sum_{s=1}^{T}\sum_{t=1}^{T}E(\zeta_{1g,st,l}^2) = O(T/N)$ for $l = 3, 4$. Then $T^{-1}\sum_{t=1}^{T}\|a_{2t}\|^2 = O_P(N^{-1})$ by Markov inequality. For a_{3t}, we have $T^{-1}\sum_{t=1}^{T}\|a_{3t}\|^2 \leq \frac{R}{\tilde{q}^4}\left\|\tilde{D}^{-1}\right\|^2\frac{1}{T^2}\sum_{s=1}^{T}\sum_{t=1}^{T}\zeta_{2g,st}^2$, where $\zeta_{2g,st} = \frac{1}{N}\sum_{i=1}^{N}F_s^{0\prime}\lambda_i^0\varepsilon_{it}g_{is}g_{it}$. Using $g_{is} = q + (g_{is}-q)$, we have

$$\begin{aligned}
\frac{1}{T^2}\sum_{s=1}^{T}\sum_{t=1}^{T}\zeta_{2g,st}^2 &\leq \frac{2}{T^2}\sum_{s=1}^{T}\sum_{t=1}^{T}\left[\frac{1}{N}\sum_{i=1}^{N}F_s^{0\prime}\lambda_i^0\varepsilon_{it}g_{it}q\right]^2 + \frac{2}{T^2}\sum_{s=1}^{T}\sum_{t=1}^{T}\left[\frac{1}{N}\sum_{i=1}^{N}F_s^{0\prime}\lambda_i^0\varepsilon_{it}g_{it}(g_{is}-q)\right]^2 \\
&\equiv 2A_1 + 2A_2, \text{ say.}
\end{aligned}$$

Noting that $\frac{1}{T}\sum_{t=1}^{T}E\left\|\frac{1}{N}\sum_{i=1}^{N}\lambda_i^0\varepsilon_{it}g_{it}\right\|^2 = \frac{1}{N^2T}\sum_{t=1}^{T}\sum_{i=1}^{N}E[\|\lambda_i^0\|^2\varepsilon_{it}^2]E(g_{it}^2) = O(N^{-1})$ under Assumptions A.1(ii) and A.2(i), we have $A_1 \leq \frac{1}{T}\sum_{s=1}^{T}\|F_s^0\|^2\frac{1}{T}\sum_{t=1}^{T}\left\|\frac{1}{N}\sum_{i=1}^{N}\lambda_i^0\varepsilon_{it}g_{it}\right\|^2 = O_P(N^{-1})$. Similarly, $A_2 = O_P(T^{-1})$ by Markov inequality. It follows that $T^{-1}\sum_{t=1}^{T}\|a_{3t}\|^2 = O_P(N^{-1}+T^{-1})$. Analogously, we can show that $T^{-1}\sum_{t=1}^{T}\|a_{4t}\|^2 = O_P(N^{-1}+T^{-1})$. Next, $T^{-1}\sum_{t=1}^{T}\|a_{5t}\|^2 \leq \frac{Rq^2}{\tilde{q}^4}\left\|\tilde{D}^{-1}\right\|^2\frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\zeta_{3g,st}^2 = O_P(N^{-1})$, where $\zeta_{3g,st} = \frac{1}{N}\sum_{i=1}^{N}F_s^{0\prime}\lambda_i^0\lambda_i^{0\prime}F_t^0(g_{is}-q)$ and the last equality follows from the Markov inequality and the fact that $\frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}E(\zeta_{3g,st}^2) = \frac{q(1-q)}{N^2T^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}E[(F_s^{0\prime}\lambda_i^0\lambda_i^{0\prime}F_t^0)^2] = O(N^{-1})$. Similarly, we can show that $T^{-1}\sum_{t=1}^{T}\|a_{6t}\|^2 = O_P(N^{-1})$ and $T^{-1}\sum_{t=1}^{T}\|a_{7t}\|^2 = O_P(N^{-1})$.

In sum, we have shown that $T^{-1}\sum_{t=1}^{T}\left\|\tilde{F}_t - \tilde{H}'F_t^0\right\|^2 = O_P(N^{-1}+T^{-1})$. ■

To prove Theorem 2.2, we need the following lemma.

Lemma A.2 *Suppose that Assumptions A.1-A.3 hold. Then*
(i) $T^{-1}\tilde{F}'F^0 = Q + O_P(\delta_{NT}^{-(1-\gamma/2)})$,
(ii) $\tilde{H} = Q^{-1} + O_P(\delta_{NT}^{-(1-\gamma/2)})$,
(iii) $\frac{1}{T}\sum_{t=1}^{T}(\tilde{F}_t - \tilde{H}F_t^0)\varepsilon_{it}g_{it} = O_P(\delta_{NT}^{-2})$,
(iv) $\frac{1}{T}\sum_{t=1}^{T}(\tilde{F}_t - \tilde{H}'F_t^0)(\tilde{F}_t - \tilde{H}'F_t^0)'g_{it} = O_P(\delta_{NT}^{-2})$,
(v) $\frac{1}{T}\sum_{t=1}^{T}\tilde{F}_t(\tilde{F}_t - \tilde{H}'F_t^0)'g_{it} = O_P(\delta_{NT}^{-2})$,
(vi) $\frac{1}{T}\sum_{t=1}^{T}(\tilde{F}_t - \tilde{H}'F_t^0)F_t^{0\prime}\tilde{H}(g_{it}-q) = O_P(\delta_{NT}^{-2})$,
(vii) $\frac{1}{T}\sum_{t=1}^{T}\tilde{F}_t\tilde{F}_t'(g_{it}-q) = \tilde{H}'\frac{1}{T}\sum_{t=1}^{T}F_t^0F_t^{0\prime}\tilde{H}(g_{it}-q) + O_P(\delta_{NT}^{-2})$,
(viii) $\tilde{H}\tilde{H}' = (\frac{1}{T}F^{0\prime}F^0)^{-1} + O_P(\delta_{NT}^{-2})$.

Proof of Theorem 2.2. (i) By the decomposition in (A.1) and Lemma A.1, it suffices to show that $A_{lt} = \tilde{D}a_{lt} = o_P\left(N^{-1/2}\right)$ for $l = 1, 2, 4, 5, 7$ and $\sqrt{N}\tilde{D}\left(a_{3t} + a_{6t}\right) \xrightarrow{d} N\left(0, Q\Gamma_{g,t}Q'\right)$. For A_{1t}, we make the following decomposition $A_{1t} = \frac{1}{NT\tilde{q}^2}\sum_{s=1}^{T}\left(\tilde{F}_s - \tilde{H}'F_s^0\right)\sum_{i=1}^{N}E\left(\varepsilon_{is}\varepsilon_{it}\right)g_{is}g_{it}$ $+\frac{\tilde{H}'}{\tilde{q}^2}\frac{1}{NT}\sum_{s=1}^{T}F_s^0\sum_{i=1}^{N}E\left(\varepsilon_{is}\varepsilon_{it}\right)g_{is}g_{it} \equiv A_{1t,1} + \frac{\tilde{H}'}{\tilde{q}^2}A_{1t,2}$. By the CS inequality and Theorem 2.1,

$$\|A_{1t,1}\| \leq \frac{1}{\tilde{q}^2}\left\{\frac{1}{T}\sum_{s=1}^{T}\left\|\tilde{F}_s - \tilde{H}'F_s^0\right\|^2\right\}^{1/2}\left\{\frac{1}{T}\sum_{s=1}^{T}\left\|\frac{1}{N}\sum_{i=1}^{N}E\left(\varepsilon_{is}\varepsilon_{it}\right)g_{is}g_{it}\right\|^2\right\}^{1/2} = O_P\left(\delta_{NT}^{-1}\right)O_P(T^{-1/2}),$$

where we use the fact that $\frac{1}{T}\sum_{s=1}^{T}\left\|\frac{1}{N}\sum_{i=1}^{N}E\left(\varepsilon_{is}\varepsilon_{it}\right)g_{is}g_{it}\right\|^2 \leq \frac{1}{T}\max_t\sum_{s=1}^{T}\left|\gamma_N\left(s,t\right)\right|^2 = O\left(T^{-1}\right)$. For $A_{1t,2}$, we have $E\left\|A_{1t,2}\right\| \leq \frac{\max_s E\left\|F_s^0\right\|}{T}\sum_{s=1}^{T}\left|\gamma_N\left(s,t\right)\right| = O\left(T^{-1}\right)$. It follows that $A_{1t,2} = O_P\left(T^{-1}\right)$ and $A_{1t} = O_P\left(\delta_{NT}^{-1}T^{-1/2}\right)$. For A_{2t}, we have

$$A_{2t} = \frac{1}{NT\tilde{q}^2}\sum_{s=1}^{T}\left(\tilde{F}_s - \tilde{H}'F_s^0\right)\sum_{i=1}^{N}\chi_{i,st}g_{is}g_{it} + \frac{\tilde{H}'}{\tilde{q}^2}\frac{1}{NT}\sum_{s=1}^{T}F_s^0\sum_{i=1}^{N}\chi_{i,st}g_{is}g_{it} \equiv A_{2t,1} + \frac{\tilde{H}'}{\tilde{q}^2}A_{2t,2},$$

where $\chi_{i,st} = \varepsilon_{is}\varepsilon_{it} - E\left(\varepsilon_{is}\varepsilon_{it}\right)$. As in the analysis of $A_{1t,1}$, we can show that $\|A_{2t,1}\| = O_P\left(\delta_{NT}^{-1}N^{-1/2}\right)$ by the CS inequality and Theorem 2.1. Next, $A_{2t,2} = \frac{1}{NT}\sum_{s=1}^{T}F_s^0\sum_{i=1}^{N}\chi_{i,st}[q^2 + \left(g_{is} - q\right)q + \left(g_{it} - q\right)q + \left(g_{is} - q\right)\left(g_{it} - q\right)] \equiv \sum_{l=1}^{4}A_{2t,2l}$. By straightforward moment calculations, we can show that $E\left\|A_{2t,2l}\right\|^2 = O((NT)^{-1})$ under Assumptions A.3(i) and A.1(i) for $l = 1, 2, 3, 4$. Then $A_{2t,2} = O_P((NT)^{-1/2})$. For A_{3t}, we use $g_{is} = q + \left(g_{is} - q\right)$ and $\tilde{F}_s = \left(\tilde{F}_s - \tilde{H}'F_s^0\right) + \tilde{H}'F_s^0$ to make the following decomposition

$$\begin{aligned}
A_{3t} &= \frac{1}{T\tilde{q}^2}\sum_{s=1}^{T}\tilde{F}_sF_s^{0\prime}\frac{1}{N}\sum_{i=1}^{N}\lambda_i^0\varepsilon_{it}g_{is}g_{it} \\
&= \frac{q}{T\tilde{q}^2}\sum_{s=1}^{T}\tilde{F}_sF_s^{0\prime}\frac{1}{N}\sum_{i=1}^{N}\lambda_i^0\varepsilon_{it}g_{it} + \frac{1}{T\tilde{q}^2}\sum_{s=1}^{T}(\tilde{F}_s - \tilde{H}'F_s^0)F_s^{0\prime}\frac{1}{N}\sum_{i=1}^{N}\lambda_i^0\varepsilon_{it}g_{it}(g_{is} - q) \\
&\quad + \frac{\tilde{H}'}{\tilde{q}^2}\left[\frac{1}{T}\sum_{s=1}^{T}F_s^0F_s^{0\prime}\frac{1}{N}\sum_{i=1}^{N}\lambda_i^0\varepsilon_{it}g_{it}(g_{is} - q)\right] \equiv A_{3t,1} + A_{3t,2} + \frac{\tilde{H}'}{\tilde{q}^2}A_{3t,3}.
\end{aligned}$$

By the CS inequality and Theorem 2.1,

$$\|A_{3t,2}\| \leq \frac{1}{\tilde{q}^2}\left\{\frac{1}{T}\sum_{s=1}^{T}\left\|\tilde{F}_s - \tilde{H}'F_s^0\right\|^2\right\}^{1/2}\left\{\frac{1}{T}\sum_{s=1}^{T}\left\|F_s^{0\prime}\frac{1}{N}\sum_{i=1}^{N}\lambda_i^0\varepsilon_{it}g_{it}(g_{is} - q)\right\|^2\right\}^{1/2} = O_P(\delta_{NT}^{-1})O_P(N^{-1/2}),$$

where we use the fact that $\frac{1}{T}\sum_{s=1}^{T}E\left\|F_s^{0\prime}\frac{1}{N}\sum_{i=1}^{N}\lambda_i^0\varepsilon_{it}g_{it}(g_{is} - q)\right\|^2 = O\left(N^{-1}\right)$. For $A_{3t,3}$, it is easy to verify that $E\left(A_{3t,3}\right) = O\left(T^{-1}\right)$ and $E\left\|A_{3t,3}\right\|^2 = O((NT)^{-1} + T^{-2})$. Then $A_{3t,3} = O_P(\delta_{NT}^{-1}T^{-1/2})$ and $A_{3t} = \frac{1}{T}\tilde{F}'F^0\frac{1}{Nq}\sum_{i=1}^{N}\lambda_i^0\varepsilon_{it}g_{it} + O_P\left(\delta_{NT}^{-2}\right)$, where we use the fact that $\tilde{q} = q + O_P((NT)^{-1/2})$. For

A_{4t}, we apply $g_{it} = q + (g_{it} - q)$ and $\tilde{F}_s = (\tilde{F}_s - \tilde{H}'F_s^0) + \tilde{H}'F_s^0$ to make the following decomposition

$$\begin{aligned} A_{4t} &= \frac{q}{NT\tilde{q}^2}\sum_{s=1}^{T}\left(\tilde{F}_s - \tilde{H}'F_s^0\right)\sum_{i=1}^{N}\lambda_i^{0\prime}\varepsilon_{is}g_{is}F_t^0 + \frac{1}{NT\tilde{q}^2}\sum_{s=1}^{T}\left(\tilde{F}_s - \tilde{H}'F_s^0\right)\sum_{i=1}^{N}\lambda_i^{0\prime}\varepsilon_{is}g_{is}\left(g_{it}-q\right)F_t^0 \\ &\quad + \frac{q\tilde{H}'}{\tilde{q}^2}\left[\frac{1}{NT}\sum_{s=1}^{T}F_s^0\sum_{i=1}^{N}\lambda_i^{0\prime}\varepsilon_{is}g_{is}\right]F_t^0 + \frac{\tilde{H}'}{\tilde{q}^2}\left[\frac{1}{NT}\sum_{s=1}^{T}F_s^0\sum_{i=1}^{N}\lambda_i^{0\prime}\varepsilon_{is}g_{is}\left(g_{it}-p\right)\right]F_t^0 \\ &\equiv A_{4t,1}F_t^0 + A_{4t,2}F_t^0 + \frac{q\tilde{H}'}{\tilde{q}^2}A_{4t,3}F_t^{0\prime} + \frac{\tilde{H}'}{\tilde{q}^2}A_{4t,4}F_t^0. \end{aligned}$$

For $A_{4t,1}$ and $A_{4t,2}$, we can readily use the CS inequality and Theorem 2.1 to show that $A_{4t,1} = O_P\left(\delta_{NT}^{-1}N^{-1/2}\right)$ and $A_{4t,2} = O_P\left(\delta_{NT}^{-2}\right)$. For $A_{4t,3}$ we apply $g_{it} = q + (g_{it} - q)$, the CS inequality, and Assumption A.3(ii) to obtain $E\left\|A_{4t,3}\right\|^2 \leq \frac{2}{N^2T^2}E||\sum_{s=1}^{T}\sum_{i=1}^{N}F_s^0\lambda_i^{0\prime}\varepsilon_{is}q||^2 + \frac{2}{N^2T^2}E||\sum_{s=1}^{T}\sum_{i=1}^{N}F_s^0\lambda_i^{0\prime}\varepsilon_{is} \times (g_{is}-q)||^2 = O\left((NT)^{-1}\right) + O\left((NT)^{-1}\right) = O\left((NT)^{-1}\right)$. It follows that $A_{4t,3} = O_P((NT)^{-1/2})$. Similarly, $A_{4t,4} = O_P((NT)^{-1/2})$. Then $A_{4t} = O_P(\delta_{NT}^{-2})$.

For A_{5t}, we use $\tilde{F}_s = (\tilde{F}_s - \tilde{H}'F_s^0) + \tilde{H}'F_s^0$ to obtain $A_{5t} = \frac{q}{\tilde{q}^2}(A_{5t,1} + \tilde{H}'A_{5t,2})F_t^0$, where $A_{5t,1} = \frac{1}{T}\sum_{s=1}^{T}(\tilde{F}_s - \tilde{H}'F_s^0)F_s^{0\prime}\frac{1}{N}\sum_{i=1}^{N}\lambda_i^0\lambda_i^{0\prime}\left(g_{is}-q\right)$ and $A_{5t,2} = \frac{1}{NT}\sum_{s=1}^{T}F_s^0F_s^{0\prime}\sum_{i=1}^{N}\lambda_i^0\lambda_i^{0\prime}\left(g_{is}-q\right)$. By the CS inequality and Theorem 2.1,

$$\|A_{5t,1}\| = \left\{\frac{1}{T}\sum_{s=1}^{T}\left\|\tilde{F}_s - \tilde{H}'F_s^0\right\|^2\right\}^{1/2}\left\{\frac{1}{T}\sum_{s=1}^{T}\left\|F_s^{0\prime}\frac{1}{N}\sum_{i=1}^{N}\lambda_i^0\lambda_i^{0\prime}\left(g_{is}-q\right)\right\|^2\right\}^{1/2} = O_P(\delta_{NT}^{-1}N^{-1/2}),$$

where we use the fact that $\frac{1}{T}\sum_{s=1}^{T}E\left\|F_s^{0\prime}\frac{1}{N}\sum_{i=1}^{N}\lambda_i^0\lambda_i^{0\prime}\left(g_{is}-q\right)\right\|^2 = O\left(N^{-1}\right)$. Similarly, $E\left\|A_{5t,2}\right\|^2 = \frac{q(1-q)}{(NT)^2}\sum_{s=1}^{T}\sum_{i=1}^{N}E\left\|F_s^0F_s^{0\prime}\lambda_i^0\lambda_i^{0\prime}\right\|^2 = O((NT)^{-1})$ under Assumption A.1(iii). Then $A_{5t,2} = O_P((NT)^{-1/2})$ and $A_{5t} = O_P\left(\delta_{NT}^{-2}\right)$. For A_{6t}, we apply the fact that $\tilde{q} = q + O_P((NT)^{-1/2})$ to obtain

$$A_{6t} = \frac{1}{T}\tilde{F}'F^0\frac{q}{\sqrt{N}\tilde{q}^2}\sum_{i=1}^{N}\lambda_i^0\lambda_i^{0\prime}F_t^0\left(g_{it}-q\right) = \frac{1}{T}\tilde{F}'F^0\frac{1}{\sqrt{N}q}\sum_{i=1}^{N}\lambda_i^0\lambda_i^{0\prime}F_t^0\left(g_{it}-q\right) + O_P((NT)^{-1/2}).$$

For A_{7t}, we have $A_{7t} = [\frac{1}{T\tilde{q}^2}\sum_{s=1}^{T}\left(\tilde{F}_s - \tilde{H}'F_s^0\right)F_s^{0\prime}\frac{1}{N}\sum_{i=1}^{N}\alpha_{i,st}]F_t^0 + \frac{\tilde{H}'}{\tilde{q}^2}[\frac{1}{T}\sum_{s=1}^{T}F_sF_s^{0\prime}\frac{1}{N}\sum_{i=1}^{N}\alpha_{i,st}]F_t^0$ $\equiv A_{7t,1}F_t^0 + \frac{\tilde{H}'}{\tilde{q}^2}A_{7t,2}F_t^{0\prime}$, where $\alpha_{i,st} = \lambda_i^0\lambda_i^{0\prime}\left(g_{is}-q\right)\left(g_{it}-q\right)$. As in the analysis of A_{5t}, we can readily show that $A_{7t,1} = O_P\left(\delta_{NT}^{-2}\right)$ and $A_{7t,2} = O_P((NT)^{-1/2})$. Then $A_{7t,1} = O_P(\delta_{NT}^{-2})$.

In sum, we have

$$\sqrt{N}(\tilde{F}_t - \tilde{H}'F_t^0) = \tilde{D}^{-1}\frac{1}{T}\tilde{F}'F^0\frac{1}{\sqrt{N}q}\sum_{i=1}^{N}\lambda_i^0\left[\varepsilon_{it}g_{it} + \lambda_i^{0\prime}F_t^0\left(g_{it}-q\right)\right] + O_P(N^{1/2}\delta_{NT}^{-2}). \tag{A.2}$$

By Assumption A.4(i), $\frac{1}{\sqrt{N}q}\sum_{i=1}^{N}\lambda_i^0\varepsilon_{it}g_{it} \xrightarrow{d} N\left(0,\Gamma_{1g,t}\right)$. Let $\omega \in \mathbb{R}^R$ be a nonrandom vector with $\|\omega\| = 1$. Let $\varphi_{it} = \frac{1}{\sqrt{N}q}\omega'\lambda_i^0\lambda_i^{0\prime}F_t^0\left(g_{it}-q\right)$ and $\mathcal{G}_{Ni}^t = \sigma\left(\{g_{jt}, j \leq i), \Lambda^0, F_t^0\right)$, the sigma-field generated from $\{g_{jt},\ j \leq i\}$ and (Λ^0, F^0). Let $\mathcal{G}^t = \sigma\left(\cup_{N=1}^{\infty}\mathcal{G}_{NN}^t\right)$. By the independence of g_{it} along the i-dimension, we have $E\left(\varphi_{it}|\mathcal{F}_{Nt,i-1,t}\right) = 0$ and $\sum_{i=1}^{N}E\left(\varphi_{it}^2|\mathcal{G}_{N,i-1}^t\right) = \frac{1-q}{Nq}\sum_{i=1}^{N}\left(\omega'\lambda_i^0\lambda_i^{0\prime}F_t^0\right)^2 = \omega'\frac{1-q}{Nq}\sum_{i=1}^{N}\lambda_i^0\lambda_i^{0\prime}\left(F_t^{0\prime}\lambda_i^0\right)^2\omega \xrightarrow{p} \omega'\Gamma_{2g,t}\omega$. Let $\epsilon = \frac{4}{\lambda_2} - 4$. Then by Assumption A.1(ii), $\sum_{i=1}^{N}E(|\varphi_{it}|^{2+\epsilon/2}$

$|\mathcal{G}_{N,i-1}^t) \leq \frac{\left\|F_t^0\right\|^{2+\epsilon/2}}{N^{\delta/2}} \frac{1}{N} \sum_{i=1}^N \left\|\lambda_i^0\right\|^{4+\epsilon} \xrightarrow{p} 0$, which is sufficient for the conditional Lindeberg condition in Häusler and Luschgy (2015) to hold. Then by the stable martingale central limit theorem (e.g., Theorem 6.1 in Häusler and Luschgy (2015)), we have

$$\frac{1}{\sqrt{N}q} \sum_{i=1}^N \lambda_i^0 \lambda_i^{0\prime} F_t^0 \left(g_{it} - q\right) = \sum_{i=1}^N \varphi_{it} \to N\left(0, \Gamma_{2g,t}\right) \quad \mathcal{G}^t\text{-stably as } N \to \infty,$$

where $\Gamma_{2g,t}$ is $\mathcal{G}_\infty^t$ measurable. Noting that $\text{Cov}(\frac{1}{\sqrt{N}q} \sum_{i=1}^N \lambda_i^0 \varepsilon_{it} g_{it}, \frac{1}{\sqrt{N}q} \sum_{i=1}^N \lambda_i^0 \lambda_i^{0\prime} F_t^0 \left(g_{it} - q\right)) = \frac{1}{Nq^2} \sum_{i=1}^N \sum_{j=1}^N E(\lambda_i^0 \lambda_j^{0\prime} \varepsilon_{it} \lambda_j^{0\prime} F_t^0) E\left[g_{it}\left(g_{jt} - q\right)\right] = \frac{1-q}{Nq} \sum_{i=1}^N E\left[\lambda_i^0 \lambda_i^{0\prime} \varepsilon_{it} \lambda_i^{0\prime} F_t^0\right] = 0$ by the i.i.d. of g_{it}, the independence between $\{g_{it}\}$ and $\left\{\Lambda^0, F^0, \varepsilon\right\}$, and Assumption A.2(i), we also have

$$\frac{1}{\sqrt{N}q} \sum_{i=1}^N \lambda_i^0 \left[\varepsilon_{it} g_{it} + \lambda_i^{0\prime} F_t^0 \left(g_{it} - q\right)\right] \to N\left(0, \Gamma_{1g,t} + \Gamma_{2g,t}\right) \ \mathcal{G}^t\text{-stably as } N \to \infty.$$

Then by Lemmas A.1(i) and A.2(i) and Corollary 6.3 in Häusler and Luschgy (2015), we have

$$\begin{aligned}
\sqrt{N}(\tilde{F}_t - \tilde{H}' F_t^0) &= \tilde{D}^{-1} \frac{1}{T} \tilde{F}' F^0 \frac{1}{\sqrt{N}q} \sum_{i=1}^N \lambda_i^0 \left[\varepsilon_{it} g_{it} + \lambda_i^{0\prime} F_t^0 \left(g_{it} - q\right)\right] + O_P\left(N^{1/2} \delta_{NT}^{-2}\right) \\
&\to N\left(0, D^{-1} Q \left(\Gamma_{1g,t} + \Gamma_{2g,t}\right) Q' D^{-1}\right) \ \mathcal{G}^t\text{-stably as } (N,T) \to \infty.
\end{aligned}$$

This completes the proof of (i).

(ii) Noting that $\tilde{\Lambda}' = \frac{1}{T\tilde{q}} \tilde{F}' \tilde{X}$, $\tilde{X} = (F^0 \Lambda^{0\prime} + \varepsilon) \circ G$, and $\frac{1}{T} \sum_{t=1}^T \tilde{F}_t \tilde{F}_t' = I_R$, we have

$$\begin{aligned}
\tilde{\lambda}_i - \tilde{H}^{-1} \lambda_i^0 &= \frac{1}{T\tilde{q}} \sum_{t=1}^T \tilde{F}_t \left(\varepsilon_{it} + F_t^{0\prime} \lambda_i^0\right) g_{it} - \tilde{H}^{-1} \lambda_i^0 \\
&= \frac{1}{T\tilde{q}} \sum_{t=1}^T \tilde{F}_t \left\{\varepsilon_{it} + \left[\tilde{F}_t' \tilde{H}^{-1} + \left(F_t^{0\prime} - \tilde{F}_t' \tilde{H}^{-1}\right)\right] \lambda_i^0\right\} g_{it} - \tilde{H}^{-1} \lambda_i^0 \\
&= \frac{\tilde{H}'}{T\tilde{q}} \sum_{t=1}^T F_t^0 \varepsilon_{it} g_{it} + \frac{1}{T\tilde{q}} \sum_{t=1}^T \left(\tilde{F}_t - \tilde{H}' F_t^0\right) \varepsilon_{it} g_{it} + \frac{1}{T\tilde{q}} \sum_{t=1}^T \tilde{F}_t \left(\tilde{H}' F_t^0 - \tilde{F}_t\right)' \tilde{H}^{-1} \lambda_i^0 g_{it} \\
&\quad + \frac{1}{T\tilde{q}} \sum_{t=1}^T \tilde{F}_t \tilde{F}_t' \tilde{H}^{-1} \lambda_i^0 (g_{it} - q) + \frac{q - \tilde{q}}{\tilde{q}} \tilde{H}^{-1} \lambda_i^0 \equiv \sum_{l=1}^5 B_{li}.
\end{aligned}$$

By Lemma A.2(ii)-(v) and (vii), $\sqrt{T} B_{1i} = \tilde{H}' \frac{1}{\sqrt{T}q} \sum_{t=1}^T F_t^0 \varepsilon_{it} g_{it} + o_P(1)$ and $\sqrt{T} B_{li} = O_P\left(T^{1/2} \delta_{NT}^{-2}\right) = o_P(1)$ for $l = 2, 3$. By Lemma A.2(ii) and (vii), $\sqrt{T} B_{4i} = \tilde{H}' \frac{1}{\sqrt{T}q} \sum_{t=1}^T F_t^0 F_t^{0\prime} \lambda_i^0 (g_{it} - q) + O_P\left(T^{1/2} \delta_{NT}^{-2}\right)$. Noting that $\tilde{q} - q = O_P((NT)^{-1/2})$, we have $\sqrt{T} B_{5i} = O_P(N^{-1/2})$. Therefore we have shown that

$$\sqrt{T}\left(\tilde{\lambda}_i - \tilde{H}^{-1} \lambda_i^0\right) = \tilde{H}' \frac{1}{\sqrt{T}q} \sum_{t=1}^T F_t^0 \left[\varepsilon_{it} g_{it} + F_t^{0\prime} \lambda_i^0 (g_{it} - q)\right] + O_P(T^{1/2} \delta_{NT}^{-2}). \tag{A.3}$$

Recall that $\mathcal{G}_{Tt}^i = \sigma\left(\{g_{is}, s \leq t), \lambda_i^0, F^0\right)$ denotes the sigma-field generated from $\{\{g_{is}, s \leq t)\}$ and (λ_i^0, F^0) and $\mathcal{G}^i = \sigma\left(\cup_{T=1}^\infty \mathcal{G}_{TT}^i\right)$. Following the analysis at the end of the proof of part (i), we can

show that $\sqrt{T}\left(\tilde{\lambda}_i - \tilde{H}^{-1}\lambda_i^0\right) \to N\left(0, (Q')^{-1}\left(\Phi_{1g,t} + \Phi_{2g,t}\right)(Q)^{-1}\right)$ $\mathcal{G}^i$-stably as $N \to \infty$, where we use Lemma A.2(ii) and the fact $\text{Cov}(\frac{1}{\sqrt{T}q}\sum_{t=1}^T F_t^0\varepsilon_{it}g_{it}, \frac{1}{\sqrt{T}q}\sum_{t=1}^T F_t^0 F_t^{0\prime}\lambda_i^0(g_{it}-q)) = 0$.

(iii) Let $\varsigma_{it} = \varepsilon_{it}g_{it} + \lambda_i^{0\prime}F_t^0\left(g_{it}-q\right)$. By the proofs of (i) and (ii),

$$\begin{aligned}
\tilde{C}_{it} - C_{it}^0 &= \lambda_i^{0\prime}(\tilde{H}')^{-1}(\tilde{F}_t - \tilde{H}'F_t^0) + \tilde{F}_t'(\tilde{\lambda}_i - \tilde{H}^{-1}\lambda_i^0) \\
&= \lambda_i^{0\prime}(\tilde{H}')^{-1}(\tilde{F}_t - \tilde{H}'F_t^0) + F_t^{0\prime}\tilde{H}(\tilde{\lambda}_i - \tilde{H}^{-1}\lambda_i^0) + O_P((NT)^{-1/2}) \\
&= \lambda_i^{0\prime}(\tilde{H}')^{-1}\tilde{D}^{-1}(\frac{1}{T}\tilde{F}'F^0)\frac{1}{Nq}\sum_{i=1}^N \lambda_i^0\varsigma_{it} + F_t^{0\prime}\tilde{H}\tilde{H}'\frac{1}{Tq}\sum_{t=1}^T F_t^0\varsigma_{it} + O_P\left(\delta_{NT}^{-2}\right) \\
&= \lambda_i^{0\prime}(\frac{1}{N}\Lambda^{0\prime}\Lambda^0)^{-1}\frac{1}{Nq}\sum_{i=1}^N \lambda_i^0\varsigma_{it} + F_t^{0\prime}(\frac{1}{T}F^{0\prime}F^0)^{-1}\frac{1}{Tq}\sum_{t=1}^T F_t^0\varsigma_{it} + O_P\left(\delta_{NT}^{-2}\right),
\end{aligned}$$

where the second equality follows from the fact that $\tilde{F}_t - \tilde{H}'F_t^0 = O_P\left(N^{-1/2}\right)$ and $\tilde{\lambda}_i - \tilde{H}^{-1}\lambda_i^0 = O_P\left(T^{-1/2}\right)$, the third equality holds by the results in (i) and (ii), and fourth equality holds because $(\tilde{H}')^{-1}\tilde{D}^{-1}\frac{1}{T}\tilde{F}'F^0 = (\frac{1}{N}\Lambda^{0\prime}\Lambda^0)^{-1}$ by the definition of $\tilde{H}$ and $\tilde{H}\tilde{H}' = (\frac{1}{T}F^{0\prime}F^0)^{-1} + O_P\left(\delta_{NT}^{-2}\right)$ by Lemma A.2(viii). Following the proof of Theorem 3 in Bai (2003), we can readily show that $(\frac{1}{N}\Sigma_{1it} + \frac{1}{T}\Sigma_{2it})^{-1/2}\left(\tilde{C}_{it} - C_{it}^0\right) \xrightarrow{d} N(0,1)$, where $\Sigma_{1it} = \lambda_i^{0\prime}\Sigma_{\Lambda^0}^{-1}\Gamma_{g,t}\Sigma_{\Lambda^0}^{-1}\lambda_i^0$ and $\Sigma_{2it} = F_t^{0\prime}\Sigma_{F^0}^{-1}\Phi_{g,i}\Sigma_{F^0}^{-1}F_t^0$. ■

To prove Theorems 2.3-2.4, we introduce some notations. Recall that $\hat{H}^{(\ell)} = \left(N^{-1}\Lambda^{0\prime}\Lambda^0\right)^{-1} \times T^{-1}F^{0\prime}\hat{F}^{(\ell)}\hat{D}^{(\ell)-1}$. Define

$$\begin{aligned}
\hat{\phi}_{F,t}^{(0)} &= \hat{D}^{(0)-1}\frac{1}{T}\hat{F}^{(0)\prime}F^0\frac{1}{Nq}\sum_{i=1}^N \lambda_i^0\left[\varepsilon_{it}g_{it} + \lambda_i^{0\prime}F_t^0\left(g_{it}-q\right)\right], \\
\hat{\phi}_{\Lambda,i}^{(0)} &= \hat{H}^{(0)\prime}\frac{1}{Tq}\sum_{t=1}^T F_t^0\left[\varepsilon_{it}g_{it} + F_t^{0\prime}\lambda_i^0(g_{it}-q)\right], \\
\hat{\phi}_{F,t}^{(\ell)} &= \hat{D}^{(\ell)-1}\frac{1}{T}\hat{F}^{(\ell)\prime}F^0\frac{1}{N}\sum_{i=1}^N \lambda_i^0\varepsilon_{it}^{(\ell)} \text{ for } \ell \geq 1, \text{ and} \\
\hat{\phi}_{\Lambda,i}^{(\ell)} &= \hat{H}^{(\ell)\prime}\frac{1}{T}\sum_{t=1}^T F_t^0\varepsilon_{it}^{(\ell)} \text{ for } \ell \geq 1,
\end{aligned}$$

where $\varepsilon_{it}^{(\ell)}$ is defined sequentially in (A.6) below, and $\hat{\phi}_{F,t}^{(\ell)}$ and $\hat{\phi}_{\Lambda,i}^{(\ell)}$ denote the leading influence functions of $\hat{F}_t^{(\ell)} - \hat{H}^{(\ell)\prime}F_t^0$ and $\hat{\lambda}_i^{(\ell)} - (\hat{H}^{(\ell)})^{-1}\lambda_i^0$, respectively. Let $\hat{r}_{F,t}^{(\ell)} = \hat{F}_t^{(\ell)} - \hat{H}^{(\ell)\prime}F_t^0 - \hat{\phi}_{F,t}^{(\ell)}$ and $\hat{r}_{\Lambda,i}^{(\ell)} = \hat{\lambda}_i^{(\ell)} - (\hat{H}^{(\ell)})^{-1}\lambda_i^0 - \hat{\phi}_{\Lambda,i}^{(\ell)}$ where $\ell \geq 0$. Then

$$\hat{\lambda}_i^{(\ell)\prime}\hat{F}_t^{(\ell)} = \left[(\hat{H}^{(\ell)})^{-1}\lambda_i^0 + \hat{\phi}_{\Lambda,i}^{(\ell)} + \hat{r}_{\Lambda,i}^{(\ell)}\right]'\left[\hat{H}^{(\ell)\prime}F_t^0 + \hat{\phi}_{F,t}^{(\ell)} + \hat{r}_{F,t}^{(\ell)}\right] = \lambda_i^{0\prime}F_t^0 + \eta_{it}^{(\ell)}, \tag{A.4}$$

where $\eta_{it}^{(\ell)} = \eta_{1,it}^{(\ell)} + \eta_{2,it}^{(\ell)}$,

$$\begin{aligned}
\eta_{1,it}^{(\ell)} &= F_t^{0\prime}\hat{H}^{(\ell)}\hat{\phi}_{\Lambda,i}^{(\ell)} + \lambda_i^{0\prime}(\hat{H}^{(\ell)\prime})^{-1}\hat{\phi}_{F,t}^{(\ell)} + \lambda_i^{0\prime}(\hat{H}^{(\ell)\prime})^{-1}\hat{r}_{F,t}^{(\ell)} + F_t^{0\prime}\hat{H}^{(\ell)\prime}\hat{r}_{\Lambda,i}^{(\ell)}, \text{ and} \\
\eta_{2,it}^{(\ell)} &= \hat{\phi}_{\Lambda,i}^{(\ell)\prime}\hat{\phi}_{F,t}^{(\ell)} + \hat{\phi}_{\Lambda,i}^{(\ell)\prime}\hat{r}_{F,t}^{(\ell)} + \hat{\phi}_{F,t}^{(\ell)\prime}\hat{r}_{\Lambda,i}^{(\ell)} + \hat{r}_{\Lambda,i}^{(\ell)\prime}\hat{r}_{F,t}^{(\ell)}.
\end{aligned} \tag{A.5}$$

Let $\bar{g}_{it} = 1 - g_{it}$ and

$$\varepsilon_{it}^{(\ell)} = \varepsilon_{it} g_{it} + \eta_{it}^{(\ell-1)} \bar{g}_{it}, \ \ell \geq 1. \tag{A.6}$$

By (A.4) and (A.6), we have

$$\hat{X}_{it}^{(\ell)} = \left(\lambda_i^{0\prime} F_t^0 + \varepsilon_{it}\right) g_{it} + \hat{\lambda}_i^{(\ell-1)\prime} \hat{F}_t^{(\ell-1)} \bar{g}_{it} = \left(\lambda_i^{0\prime} F_t^0 + \varepsilon_{it}\right) g_{it} + \left(\lambda_i^{0\prime} F_t^0 + \eta_{it}\right) \bar{g}_{it} = \lambda_i^{0\prime} F_t^0 + \varepsilon_{it}^{(\ell)}. \tag{A.7}$$

This expression will be used repeatedly in the following derivation.

The following three lemmas are used in the proofs of Theorems 2.3 and 2.4. When Lemmas A.3-A.5 hold for $\ell = 1$, Theorems 2.3 and 2.4 also hold for $\ell = 1$. With the results in Lemmas A.3-A.5 and Theorems 2.3 and 2.4 for $\ell = 1$, we can show that they also hold for $\ell = 2$. This procedure is repeated until convergence which requires ℓ to be at most of order $\ln N$.

Lemma A.3 *Suppose that Assumptions A.1-A.5 hold. Then for any $\ell \geq 1$ we have*

(i) $\max_t \left\| \hat{\phi}_{F,t}^{(\ell-1)} \right\| = O_P((N/\ln N)^{-1/2})$ *and* $\max_i \left\| \hat{\phi}_{\Lambda,i}^{(\ell-1)} \right\| = O_P((T/\ln T)^{-1/2})$,

(ii) $\max_t \left\| \hat{r}_{F,t}^{(\ell-1)} \right\| = O_P(T^{\gamma_1/4}\delta_{NT}^{-2} \ln T + T^{-1+3\gamma_1/4})$ *and* $\max_i \left\| \hat{r}_{\Lambda,i}^{(\ell-1)} \right\| = O_P(N^{\gamma_2/4}\delta_{NT}^{-2} \ln N)$,

(iii) $\max_{i,t} \left\| \eta_{1,it}^{(\ell-1)} \right\| = O_P(\delta_{NT}^{-1+\gamma/2} \ln N)$ *and* $\max_{i,t} \left\| \eta_{2,it}^{(\ell-1)} \right\| = O_P\left(\delta_{NT}^{-2} \ln N\right)$,

(iv) $\max_t \left\| \frac{1}{N} \sum_{i=1}^N \hat{\phi}_{\Lambda,i}^{(\ell-1)} \varepsilon_{it} g_{it} \right\| = O_P\left(T^{-1+\gamma_1/4} + \delta_{NT}^{-2} \ln N\right)$, $\left\| \frac{1}{N} \sum_{i=1}^N \hat{\phi}_{\Lambda,i}^{(\ell-1)} \lambda_i^{0\prime} \bar{g}_{it} \right\| = O_P(T^{-1+\gamma_1/4} + N^{\gamma_2/4}\delta_{NT}^{-2} \ln N)$, *and* $\max_t \left\| \frac{1}{N} \sum_{i=1}^N \hat{r}_{\Lambda,i}^{(\ell-1)} \lambda_i^{0\prime} \bar{g}_{it} \right\| = O_P\left(\delta_{NT}^{-2} \ln N\right)$,

(v) $\max_i \left\| \frac{1}{T} \sum_{t=1}^T \hat{\phi}_{F,t}^{(\ell-1)} F_t^{0\prime} \bar{g}_{it} \right\| = O_P\left(\delta_{NT}^{-2} \ln N + N^{-1+\gamma_2/2}\right)$ *and* $\max_i \left\| \frac{1}{T} \sum_{t=1}^T \hat{r}_{F,t}^{(\ell-1)} F_t^{0\prime} \bar{g}_{it} \right\| = O_P\left(\delta_{NT}^{-2} \ln N\right)$,

(vi) $\max_t \frac{1}{N} \sum_{i=1}^N \left\| \eta_{it}^{(\ell-1)} \right\|^2 = O_P(T^{-1+\gamma_1/2} + N^{-1} \ln N)$ *and* $\max_i \frac{1}{T} \sum_{t=1}^T \left\| \eta_{it}^{(\ell-1)} \right\|^2 = O_P(N^{-1+\gamma_2/2} + T^{-1} \ln N)$,

(vii) $\frac{1}{NT} \sum_{t=1}^T \sum_{i=1}^N (1 + \left\| F_t^0 \right\|^2)(\eta_{it}^{(\ell-1)})^2 = O_P\left(\delta_{NT}^{-2}\right)$,

(viii) $\frac{1}{NT} \sum_{s=1}^T F_s^0 \sum_{i=1}^N \lambda_i^{0\prime} \eta_{is}^{(\ell-1)} \bar{g}_{is} = O_P\left(\delta_{NT}^{-2} \ln N\right)$,

(ix) $\max_t \left\| \frac{1}{NT} \sum_{s=1}^T F_s^0 \sum_{i=1}^N \eta_{it}^{(\ell-1)} \bar{g}_{it} \varepsilon_{is} g_{is} \right\| = O_P(T^{-1+\gamma_1/4} + (NT/\ln N)^{-1/2})$,

(x) $\max_t \left\| \frac{1}{NT} \sum_{s=1}^T F_s^0 \sum_{i=1}^N \varepsilon_{it} g_{it} \eta_{is}^{(\ell-1)} \bar{g}_{is} \right\| = O_P\left(T^{-1+\gamma_1/4} + \delta_{NT}^{-2} \ln N\right)$.

Lemma A.4 *Suppose that Assumptions A.1-A.5 hold. Then for any $\ell \geq 1$ we have*

(i) $T^{-1} \hat{F}^{(\ell)\prime} (NT)^{-1} \hat{X}^{(\ell)} \hat{X}^{(\ell)\prime} \hat{F}^{(\ell)} = \hat{D}^{(\ell)} = D + O_P\left(\delta_{NT}^{-1} \ln N\right)$,

(ii) $T^{-1} \hat{F}^{(\ell)\prime} F^0 = Q + O_P\left(\delta_{NT}^{-1} \ln N\right)$,

(iii) $\hat{H}^{(\ell)} = Q^{-1} + O_P\left(\delta_{NT}^{-1} \ln N\right)$,

(iv) $\frac{1}{T} \sum_{t=1}^T (\hat{F}_t^{(\ell)} - \hat{H}^{(\ell)\prime} F_t^0) F_t^{0\prime} = O_P\left(\delta_{NT}^{-2}\right)$,

(v) $\max_i \left\| \frac{1}{T} \sum_{t=1}^T (\hat{F}_t^{(\ell)} - \hat{H}^{(\ell)\prime} F_t^0) \varepsilon_{it}^{(\ell)} \right\| = O_P\left(N^{-1/2+\gamma_2/4}\delta_{NT}^{-1} + \delta_{NT}^{-2} \ln N\right)$.

Lemma A.5 *Suppose that Assumptions A.1-A.5 hold. Then*

(i) $\hat{\phi}_{F,t}^{(\ell)} = D^{-1} Q \beta_{F,t} + (1-q) \hat{\phi}_{F,t}^{(\ell-1)} + O_P\left(T^{\gamma_1/4}\delta_{NT}^{-2} \ln T + T^{-1+\gamma_1/4}\right)$,

(ii) $\hat{\phi}_{\Lambda,i}^{(\ell)} = (Q')^{-1} \beta_{\Lambda,i} + (1-q) \hat{\phi}_{\Lambda,i}^{(\ell-1)} + O_P\left(N^{\gamma_2/4}\delta_{NT}^{-2} \ln N + N^{-1+3\gamma_2/4}\right)$,

where $\beta_{F,t} = \frac{1}{N} \sum_{i=1}^N \lambda_i^0 \varepsilon_{it} g_{it}$, *and* $\beta_{\Lambda,i} = \frac{1}{T} \sum_{t=1}^T F_t^0 \varepsilon_{it} g_{it}$.

The proof of Theorem 2.4 below suggests that $\hat{\phi}_{F,t}^{(\ell)}$ and $\hat{\phi}_{\Lambda,i}^{(\ell)}$ are associated with the leading influence functions of $\hat{F}_t^{(\ell)} - \hat{H}^{(\ell)\prime} F_t^0$ and $\hat{\lambda}_i^{(\ell)} - (\hat{H}^{(\ell)})^{-1}\lambda_i^0$, respectively.

Proof of Theorem 2.3. The proof follows closely from that of Theorem 2.1 and we only outline the main differences. From the identity $\hat{F}^{(\ell)} = (NT)^{-1}\hat{X}^{(\ell)}\hat{X}^{(\ell)\prime}\hat{F}^{(\ell)}\hat{D}^{(\ell)-1}$ where $\hat{D}^{(\ell)}$ is asymptotically nonsingular by Lemma A.4(i), we have by (A.7),

$$\hat{F}_t^{(\ell)} - \hat{H}^{(\ell)\prime} F_t^0 = \frac{1}{NT}\hat{D}^{(\ell)-1}\sum_{s=1}^{T}\hat{F}_s^{(\ell)}\sum_{i=1}^{N}\left\{\varepsilon_{it}^{(\ell)}\varepsilon_{is}^{(\ell)} + F_s^{0\prime}\lambda_i^0\varepsilon_{it}^{(\ell)} + F_t^{0\prime}\lambda_i^0\varepsilon_{is}^{(\ell)}\right\} \equiv \hat{a}_{1t}^{(\ell)} + \hat{a}_{2t}^{(\ell)} + \hat{a}_{3t}^{(\ell)}. \quad \text{(A.8)}$$

Then $T^{-1}\sum_{t=1}^{T}\left\|\hat{F}_t^{(\ell)} - \hat{H}^{(\ell)\prime}F_t^0\right\|^2 \leq 3\sum_{l=1}^{3}T^{-1}\sum_{t=1}^{T}(\hat{a}_{lt}^{(\ell)})^2$ by the CS inequality. For $\hat{a}_{1t}^{(\ell)}$, using $\varepsilon_{it}^{(\ell)} = \varepsilon_{it}g_{it} + \eta_{it}^{(\ell-1)}\bar{g}_{it}$ and the CS inequality, we have

$$\begin{aligned}
T^{-1}\sum_{t=1}^{T}\left\|\hat{D}^{(\ell)}\hat{a}_{1t}^{(\ell)}\right\|^2 &\leq 4T^{-1}\sum_{t=1}^{T}\left\{\left\|\frac{1}{T}\sum_{s=1}^{T}\hat{F}_s^{(\ell)}\frac{1}{N}\sum_{i=1}^{N}\varepsilon_{it}g_{it}\varepsilon_{is}g_{is}\right\|^2 + \left\|\frac{1}{T}\sum_{s=1}^{T}\hat{F}_s^{(\ell)}\frac{1}{N}\sum_{i=1}^{N}\eta_{it}^{(\ell-1)}\bar{g}_{it}\eta_{is}^{(\ell-1)}\bar{g}_{is}\right\|^2\right. \\
&\quad \left. + \left\|\frac{1}{T}\sum_{s=1}^{T}\hat{F}_s^{(\ell)}\frac{1}{N}\sum_{i=1}^{N}\varepsilon_{it}g_{it}\eta_{is}^{(\ell-1)}\bar{g}_{is}\right\|^2 + \left\|\frac{1}{T}\sum_{s=1}^{T}\hat{F}_s^{(\ell)}\frac{1}{N}\sum_{i=1}^{N}\eta_{it}^{(\ell-1)}\bar{g}_{it}\varepsilon_{is}g_{is}\right\|^2\right\} \\
&\equiv 4(\hat{A}_{1,1} + \hat{A}_{1,2} + \hat{A}_{1,3} + \hat{A}_{1,4}),
\end{aligned}$$

where we suppress the dependence of $\hat{A}_1$'s on ℓ. Following the analyses of $T^{-1}\sum_{t=1}^{T}\|a_{1t}\|^2$ and $T^{-1}\sum_{t=1}^{T}\|a_{2t}\|^2$ in the proof of Theorem 2.1, we can readily show that $\hat{A}_{1,1} = O_P\left(\delta_{NT}^{-2}\right)$. For $\hat{A}_{1,2}$ and $\hat{A}_{1,3}$, we can apply the fact $\hat{F}^{(\ell)\prime}\hat{F}^{(\ell)}/T = I_R$, the CS inequality, and Lemma A.3(vii) to obtain

$$\begin{aligned}
\hat{A}_{1,2} &\leq \frac{R}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\left(\frac{1}{N}\sum_{i=1}^{N}\eta_{it}^{(\ell-1)}\bar{g}_{it}\eta_{is}^{(\ell-1)}\bar{g}_{is}\right)^2 \leq R\left\{\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}(\eta_{it}^{(\ell-1)})^2\right\}^2 = O_P\left(\delta_{NT}^{-4}\right), \text{ and} \\
\hat{A}_{1,3} &\leq \frac{R}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\left(\frac{1}{N}\sum_{i=1}^{N}\varepsilon_{it}g_{it}\eta_{is}^{(\ell-1)}\bar{g}_{is}\right)^2 \leq \frac{R}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}|\varepsilon_{it}g_{it}|^2\frac{1}{NT}\sum_{i=1}^{N}\sum_{s=1}^{T}(\eta_{is}^{(\ell-1)})^2 = O_P\left(\delta_{NT}^{-2}\right).
\end{aligned}$$

Analogously, $\hat{A}_{1,4} = O_P\left(\delta_{NT}^{-2}\right)$. It follows that $\hat{A}_1 = O_P\left(\delta_{NT}^{-2}\right)$. For $\hat{a}_{2t}^{(\ell)}$, we have

$$\begin{aligned}
T^{-1}\sum_{t=1}^{T}\left\|\hat{D}^{(\ell)}\hat{a}_{2t}^{(\ell)}\right\|^2 &= T^{-1}\sum_{t=1}^{T}\left\|\frac{1}{T}\sum_{s=1}^{T}\hat{F}_s^{(\ell)}\frac{1}{N}\sum_{i=1}^{N}F_s^{0\prime}\lambda_i^0\varepsilon_{it}^{(\ell)}\right\|^2 \leq \frac{R}{T^2}\sum_{s=1}^{T}\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{i=1}^{N}F_s^{0\prime}\lambda_i^0\varepsilon_{it}^{(\ell)}\right)^2 \\
&\leq \frac{2R}{T^2}\sum_{s=1}^{T}\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{i=1}^{N}F_s^{0\prime}\lambda_i^0\varepsilon_{it}g_{it}\right)^2 + \frac{2R}{T^2}\sum_{s=1}^{T}\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{i=1}^{N}F_s^{0\prime}\lambda_i^0\eta_{it}^{(\ell-1)}\bar{g}_{it}\right)^2.
\end{aligned}$$

By the analysis of $T^{-1}\sum_{t=1}^{T}\|a_{3t}\|^2$ in the proof of Theorem 2.1, the first term is $O_P(\delta_{NT}^{-2})$. For the second term, by the CS inequality and Lemma A.3(vii) it is bounded above by $\frac{2R}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|F_s^{0\prime}\lambda_i^0\right\|^2 \times\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\eta_{it}^{(\ell-1)})^2 = O_P(\delta_{NT}^{-2})$. Then $T^{-1}\sum_{t=1}^{T}\left\|\hat{a}_{2t}^{(\ell)}\right\|^2 = O_P(\delta_{NT}^{-2})$. Analogously, we can show that $T^{-1}\sum_{t=1}^{T}\left\|\hat{D}^{(\ell)}\hat{a}_{3t}^{(\ell)}\right\|^2 = O_P(\delta_{NT}^{-2})$. In sum, we have shown that $T^{-1}\sum_{t=1}^{T}\left\|\hat{F}_t^{(\ell)} - \hat{H}^{(\ell)\prime}F_t^0\right\|^2 = O_P(\delta_{NT}^{-2})$. ■

Proof of Theorem 2.4. (i) Let $\hat{A}_{lt}^{(\ell)} = \hat{D}^{(\ell)}\hat{a}_{lt}^{(\ell)}$ for $l = 1, 2, 3$. By the decomposition in (A.8) and Lemma A.4(i), we will bound $\hat{A}_{lt}^{(\ell)}$ for $l = 1, 3$ and find the leading influence function for $\hat{A}_{2t}^{(\ell)}$. For $\hat{A}_{1t}^{(\ell)}$, we use $\hat{F}_s^{(\ell)} = (\hat{F}_s^{(\ell)} - \hat{H}^{(\ell)\prime}F_s^0) + \hat{H}^{(\ell)\prime}F_s^0$ to make the decomposition

$$\hat{A}_{1t}^{(\ell)} = \frac{1}{NT}\sum_{s=1}^{T}(\hat{F}_s^{(\ell)} - \hat{H}^{(\ell)\prime}F_s^0)\sum_{i=1}^{N}\varepsilon_{it}^{(\ell)}\varepsilon_{is}^{(\ell)} + \hat{H}^{(\ell)\prime}\frac{1}{NT}\sum_{s=1}^{T}F_s^0\sum_{i=1}^{N}\varepsilon_{it}^{(\ell)}\varepsilon_{is}^{(\ell)} \equiv \hat{A}_{1t,1}^{(\ell)} + \hat{H}^{(\ell)\prime}\hat{A}_{1t,2}^{(\ell)}.$$

It is easy to show that $\hat{A}_{1t,1}^{(\ell)}$ is of smaller order than $\hat{A}_{1t,2}^{(\ell)}$. We focus on the study of $\hat{A}_{1t,2}^{(\ell)}$. By (A.6), we have $\hat{A}_{1t,2}^{(\ell)} = \frac{1}{NT}\sum_{s=1}^{T}F_s^0\sum_{i=1}^{N}(\varepsilon_{it}\varepsilon_{is}g_{it}g_{is} + \eta_{it}^{(\ell-1)}\eta_{is}^{(\ell-1)}\bar{g}_{it}\bar{g}_{is} + \varepsilon_{it}g_{it}\eta_{is}^{(\ell-1)}\bar{g}_{is} + \eta_{it}^{(\ell-1)}\bar{g}_{it}\varepsilon_{is}g_{is}) \equiv \sum_{l=1}^{4}\hat{A}_{1t,2l}^{(\ell)}$. By the analysis of $A_{1t,2}$ and $A_{2t,2}$ in the proof of Theorem 2.2(i), $\max_t\left\|\hat{A}_{1t,21}^{(\ell)}\right\| = O_P\left(T^{-1+\gamma_1/4} + \delta_{NT}^{-2}\ln N\right)$. By Lemma A.3(vi)-(vii) and the CS inequality,

$$\begin{aligned}\max_t\left\|\hat{A}_{1t,22}^{(\ell)}\right\| &\leq \left\{\max_t\frac{1}{N}\sum_{i=1}^{N}(\eta_{it}^{(\ell-1)})^2\right\}^{1/2}\left\{\frac{1}{NT}\sum_{s=1}^{T}\sum_{i=1}^{N}\left\|F_s^0\right\|^2(\eta_{is}^{(\ell-1)})^2\bar{g}_{is}\right\}^{1/2}\\ &= O_P\left(\delta_{NT}^{-1}(T^{-1/2+\gamma_1/4} + (N/\ln N)^{-1/2})\right).\end{aligned}$$

By Lemma A.3(ix)-(x), $\hat{A}_{1t,23}^{(\ell)} + \hat{A}_{1t,24}^{(\ell)} = O_P\left(T^{-1+\gamma_1/4} + \delta_{NT}^{-2}\ln N\right)$. Thus $\hat{A}_{1t,2}^{(\ell)} = O_P\left(T^{-1+\gamma/4} + \delta_{NT}^{-2}\ln N\right)$ and $\hat{A}_{1t}^{(\ell)} = O_P\left(T^{-1+\gamma_1/4} + \delta_{NT}^{-2}\ln N\right)$.

For $\hat{A}_{3t}^{(\ell)}$, we apply (A.6) and $\hat{F}_s^{(\ell)} = (\hat{F}_s^{(\ell)} - \hat{H}^{(\ell)\prime}F_s^0) + \hat{H}^{(\ell)\prime}F_s^0$ to make the decomposition

$$\begin{aligned}\hat{A}_{3t}^{(\ell)} &= \frac{1}{NT}\sum_{s=1}^{T}\hat{F}_s^{(\ell)}\sum_{i=1}^{N}\lambda_i^0\varepsilon_{is}'g_{is}F_t^{0\prime} + \frac{1}{NT}\sum_{s=1}^{T}(\hat{F}_s^{(\ell)} - \hat{H}^{(\ell)\prime}F_s^0)\sum_{i=1}^{N}\lambda_i^{0\prime}\eta_{is}^{(\ell-1)}\bar{g}_{is}F_t^0\\ &\quad + \hat{H}^{(\ell)\prime}[\frac{1}{NT}\sum_{s=1}^{T}F_s^0\sum_{i=1}^{N}\lambda_i^{0\prime}\eta_{is}^{(\ell-1)}\bar{g}_{is}]F_t^0 \equiv \left(\hat{A}_{3,1}^{(\ell)} + \hat{A}_{3,2}^{(\ell)} + \hat{H}^{(\ell)\prime}\hat{A}_{3,3}^{(\ell)}\right)F_t^0.\end{aligned}$$

Following the analysis of $A_{4t,1}$ and $A_{4t,2}$ in the proof of Theorem 2.2(i), we can show that $\hat{A}_{3,1}^{(\ell)} = O_P\left(\delta_{NT}^{-2}\right)$. For $\hat{A}_{3,2}^{(\ell)}$, we have by the CS inequality, Theorem 2.3 and Lemma A.3(vii)-(vii)

$$\begin{aligned}\left\|\hat{A}_{3,2}^{(\ell)}\right\| &\leq \frac{1}{T^{1/2}}\left\|\hat{F}^{(\ell)} - F^0\hat{H}^{(\ell)}\right\|\left\{\frac{1}{T}\sum_{s=1}^{T}\left\|\frac{1}{N}\sum_{i=1}^{N}\lambda_i^{0\prime}\eta_{is}^{(\ell-1)}\bar{g}_{is}\right\|^2\right\}^{1/2}\\ &\leq O_P\left(\delta_{NT}^{-1}\right)\left\{\frac{1}{N}\sum_{i=1}^{N}\left\|\lambda_i^0\right\|^2\frac{1}{TN}\sum_{s=1}^{T}\sum_{i=1}^{N}(\eta_{is}^{(\ell-1)})^2\bar{g}_{is}\right\}^{1/2} = O_P\left(\delta_{NT}^{-2}\right),\end{aligned}$$

and $\left\|\hat{A}_{3,3}^{(\ell)}\right\| \leq \left\|\frac{1}{NT}\sum_{s=1}^{T}\sum_{i=1}^{N}F_s^0\lambda_i^{0\prime}\eta_{is}^{(\ell-1)}\bar{g}_{is}\right\| = O_P\left(\delta_{NT}^{-2}\right)$. It follows that $\max_t\left\|\hat{A}_{3t}^{(\ell)}\right\| = \max_t\left\|F_t^0\right\|$ $\times O_P(\delta_{NT}^{-2}) = O_P\left(T^{\gamma_1/4}\delta_{NT}^{-2}\right)$.

It follows that

$$\begin{aligned}\hat{\Pi}_{tN}^{(\ell)} &\equiv \sqrt{N}(\hat{F}_t^{(\ell)} - \hat{H}^{(\ell)\prime}F_t^0) = \sqrt{N}\hat{\phi}_{F,t}^{(\ell)} + O_P(\sqrt{N}(T^{\gamma_1/4}\delta_{NT}^{-2}\ln T + T^{-1+3\gamma_1/4}))\\ &= \sqrt{N}[\hat{D}^{(\ell-1)}]^{-1}\frac{1}{T}\hat{F}^{(\ell-1)\prime}F^0\frac{1}{N}\sum_{i=1}^{N}\lambda_i^0\varepsilon_{it}^{(\ell)}\bar{g}_{it} + O_P(\sqrt{N}(T^{\gamma_1/4}\delta_{NT}^{-2}\ln T + T^{-1+3\gamma_1/4}))\\ &= D^{-1}Q\sqrt{N}\beta_{F,t} + (1-q)\sqrt{N}\hat{\phi}_{F,t}^{(\ell-1)} + O_P(\sqrt{N}(T^{\gamma_1/4}\delta_{NT}^{-2}\ln T + T^{-1+3\gamma_1/4})),\end{aligned}$$

where the remainder term $O_P(\sqrt{N}(T^{\gamma_1/4}\delta_{NT}^{-2}\ln T + T^{-1+3\gamma_1/4}))$ holds uniformly in t. This, in conjunction with Lemma A.5(i), implies that

$$\begin{aligned}\hat{\Pi}_{tN}^{(\ell)} &= D^{-1}Q\sqrt{N}\beta_{F,t} + (1-q)\hat{\Pi}_{tN}^{(\ell-1)} + O_P(\sqrt{N}(T^{\gamma_1/4}\delta_{NT}^{-2}\ln T + T^{-1+3\gamma_1/4})) \\ &= D^{-1}Q\sqrt{N}\beta_{F,t}\sum_{s=0}^{\ell-1}(1-q)^s + (1-q)^\ell\hat{\Pi}_{tN}^{(0)} + o_P(1) \\ &\xrightarrow{d} N\left(0, D^{-1}Q\Gamma_{1g,t}(q)Q'D^{-1}\right) \text{ as } (N,T,\ell)\to\infty.\end{aligned}$$

(ii) Let $\varepsilon^{(\ell)}$ be the $T\times N$ matrix with (t,i)th element given by $\varepsilon_{it}^{(\ell)}$. Noting that $\hat{\Lambda}^{(\ell)\prime} = \frac{1}{T}\hat{F}^{(\ell)\prime}\hat{X}^{(\ell)}$, $\hat{X}^{(\ell)} = F^0\Lambda^{0\prime} + \varepsilon^{(\ell)}$, and $\frac{1}{T}\sum_{t=1}^T \hat{F}_t^{(\ell)}\hat{F}_t^{(\ell)\prime} = I_R$, we have

$$\begin{aligned}\hat{\lambda}_i^{(\ell)} - \hat{H}^{(\ell)-1}\lambda_i^0 &= \frac{1}{T\tilde{q}}\sum_{t=1}^T \hat{F}_t^{(\ell)}\left\{\varepsilon_{it}^{(\ell)} + \left[\hat{F}_t^{(\ell)\prime}\hat{H}^{(\ell)-1} + (F_t^{0\prime} - \hat{F}_t^{(\ell)\prime}\hat{H}^{(\ell)-1})\right]\lambda_i^0\right\} - \hat{H}^{(\ell)-1}\lambda_i^0 \\ &= \frac{\hat{H}^{(\ell)\prime}}{T}\sum_{t=1}^T F_t^0\varepsilon_{it}^{(\ell)} + \frac{1}{T}\sum_{t=1}^T(F_t^{0\prime} - \hat{F}_t^{(\ell)\prime}\hat{H}^{(\ell)-1})\varepsilon_{it}^{(\ell)} + \frac{1}{T}\sum_{t=1}^T \hat{F}_t^{(\ell)}(\hat{H}^{(\ell)\prime}F_t^0 - \hat{F}_t^{(\ell)})'\hat{H}^{(\ell)-1}\lambda_i^0 \\ &\equiv \hat{B}_{1i}^{(\ell)} + \hat{B}_{2i}^{(\ell)} + \hat{B}_{3i}^{(\ell)}.\end{aligned}$$

By Lemma A.4(iv)-(v), we have $\max_i \left\|\hat{B}_{2i}^{(\ell)}\right\| = O_P\left(N^{-1/2+\gamma_2/4}\delta_{NT}^{-1} + \delta_{NT}^{-2}\ln N\right)$ and $\max_i \left\|\hat{B}_{3i}^{(\ell)}\right\| = \max_i\left\|\lambda_i^0\right\| O_P\left(\delta_{NT}^{-2}\right) = O_P\left(N^{\gamma_2/4}\delta_{NT}^{-2}\right)$. It follows that

$$\begin{aligned}\hat{\Pi}_{iT}^{(\ell)} &\equiv \sqrt{T}\left(\hat{\lambda}_i^{(\ell)} - \hat{H}^{(\ell)-1}\lambda_i^0\right) = \sqrt{T}\hat{B}_{1i}^{(\ell)} + O_P(\sqrt{T}(N^{\gamma_2/4}\delta_{NT}^{-2}\ln N + N^{-1+3\gamma_2/4})) \\ &= (Q')^{-1}\sqrt{T}\beta_{\Lambda,i} + (1-q)\sqrt{T}\hat{\phi}_{\Lambda,i}^{(\ell-1)} + o_P(1),\end{aligned}$$

where the remainder term $O_P(\sqrt{T}(N^{\gamma_2/4}\delta_{NT}^{-2}\ln N + N^{-1+3\gamma_2/4}))$ holds uniformly in i. This, in conjunction with Lemma A.5(ii), implies that

$$\begin{aligned}\hat{\Pi}_{iT}^{(\ell)} &= (Q')^{-1}\sqrt{T}\beta_{\Lambda,i} + (1-q)\hat{\Pi}_{iT}^{(\ell-1)} + O_P(\sqrt{T}N^{\gamma_2/4}\delta_{NT}^{-2}\ln N) \\ &= (Q')^{-1}\sqrt{T}\beta_{\Lambda,i}\sum_{s=0}^{\ell-1}(1-q)^s + (1-q)^\ell\hat{\Pi}_{iT}^{(0)} + O_P(\sqrt{T}(N^{\gamma_2/4}\delta_{NT}^{-2}\ln N + N^{-1+3\gamma_2/4})) \\ &\xrightarrow{d} N\left(0, (Q')^{-1}\Phi_{1g,i}(q)Q^{-1}\right) \text{ as } (N,T,\ell)\to\infty.\end{aligned}$$

(iii) By the proof of (i) and (ii) and as in the proof of Theorem 2.2(iii), we have

$$\begin{aligned}\hat{C}_{it}^{(\ell)} - C_{it}^0 &= \hat{\lambda}_i^{(\ell)\prime}\hat{F}_t^{(\ell)} - \lambda_i^{0\prime}F_t^0 = \lambda_i^{0\prime}(\hat{H}^{(\ell)\prime})^{-1}(\hat{F}_t^{(\ell)} - \hat{H}^{(\ell)\prime}F_t^0) + \hat{F}_t^{(\ell)\prime}(\hat{\lambda}_i^{(\ell)} - \hat{H}^{(\ell)-1}\lambda_i^0) \\ &= \frac{1}{\sqrt{N}}\lambda_i^{0\prime}(\hat{H}^{(\ell)\prime})^{-1}\sqrt{N}(\hat{F}_t^{(\ell)} - \hat{H}^{(\ell)\prime}F_t^0) + \frac{1}{\sqrt{T}}F_t^{0\prime}\hat{H}^{(\ell)}\sqrt{T}(\hat{\lambda}_i^{(\ell)} - \hat{H}^{(\ell)-1}\lambda_i^0) \\ &\quad + O_P((N^{\gamma_2/4} + T^{\gamma_1/4})(NT)^{-1/2}) \\ &= \frac{1}{\sqrt{N}}\lambda_i^{0\prime}(\hat{H}^{(\ell)\prime})^{-1}\hat{\Pi}_{tN}^{(\ell)} + \frac{1}{\sqrt{T}}F_t^{0\prime}\hat{H}^{(\ell)}\hat{\Pi}_{iT}^{(\ell)} + o_P(1).\end{aligned}$$

Then we have $(\frac{1}{N}\Sigma_{1F,it} + \frac{1}{T}\Sigma_{1\Lambda,it})^{-1/2}(\hat{C}_{it}^{(\ell)} - C_{it}^0) \xrightarrow{d} N(0,1)$ as $(N,T,\ell)\to\infty$, where $\Sigma_{1F,it} = \lambda_i^{0\prime}\Sigma_{\Lambda^0}^{-1}\Gamma_{1g,t}(q)\Sigma_{\Lambda^0}^{-1}\lambda_i^0$ and $\Sigma_{1\Lambda,it} = F_t^{0\prime}\Sigma_{F^0}^{-1}\Phi_{1g,i}(q)\Sigma_{F^0}^{-1}F_t^0$. ■

To prove Theorem 2.5, we need the following lemma.

Lemma A.6 *Suppose that Assumptions A.1-A.6 hold. Then*
(i) $\max_i \frac{1}{T}\sum_{t=1}^{T}|\hat{\varepsilon}_{it}-\varepsilon_{it}|^2 = O_P(N^{-1+\gamma_2/2}+T^{-1}\ln T)$,
(ii) $\max_{i,t}|\hat{\varepsilon}_{it}-\varepsilon_{it}| = O_P\left((T^{-1/2+\gamma_1/4}+N^{-1/2+\gamma_2/4})(\ln T)^{1/2}\right) = o_P(1)$,
(iii) $\left\|\hat{\Sigma}^g-\Sigma^g\right\|_{sp} = o_P(1)$.

Proof of Theorem 2.5. To show $\hat{D}^{-1}\hat{\Gamma}^{(2)}_{1g,t}\hat{D}^{-1}\xrightarrow{p} D^{-1}Q\Gamma_{1g,t}(q)Q'D^{-1}$, it suffices to show that (i) $\hat{D}^{-1}\xrightarrow{p}D^{-1}$ and (ii) $\hat{\Gamma}^{(2)}_{1g,t}\xrightarrow{p}Q\Gamma_{1g,t}Q'$. (i) holds by Lemma A.4(i) and positive definiteness of D. To show (ii), we recall that $\hat{\Gamma}^{(2)}_{1g,t}=\frac{1}{N\tilde{q}^2}\hat{\Lambda}'\hat{\Sigma}^g\hat{\Lambda}$ and $\Gamma_{1g,t}(q)=\lim_{N\to\infty}\Gamma_{1g,tN}(q)$, where $\Gamma_{1g,tN}(q)=\frac{1}{Nq^2}\Lambda^{0\prime}\Sigma^g\Lambda^0$. Then by the triangle inequality, we have

$$\begin{aligned}\left\|\hat{\Gamma}^{(2)}_{1g,t}-Q\Gamma_{1g,t}Q'\right\|_{\text{sp}} &\leq \frac{1}{N\tilde{q}^2}\left\|\hat{\Lambda}'\hat{\Sigma}^g\hat{\Lambda}-Q\Lambda^{0\prime}\Sigma^g\Lambda^0Q'\right\|_{\text{sp}}+\frac{1}{N}\left\|Q\Lambda^{0\prime}\Sigma^g\Lambda^0Q'\right\|_{\text{sp}}\left(\frac{1}{q^2}-\frac{1}{\tilde{q}^2}\right)\\ &\quad+\left\|Q\left(\Gamma_{1g,tN}(q)-\Gamma_{1g,t}(q)\right)Q'\right\|_{\text{sp}}.\end{aligned}$$

The last term on the right hand side (rhs) of the last expression is $o_P(1)$ and the second term is $O_P((NT)^{-1/2})$ by noting that $\tilde{q}-q=O_P((NT)^{-1/2})$. For the first term, we have

$$\left\|\hat{\Lambda}'\hat{\Sigma}^g\hat{\Lambda}-Q\Lambda^{0\prime}\Sigma^g\Lambda^0Q'\right\|_{\text{sp}}\leq\left\|[\hat{\Lambda}-\Lambda^0Q']'\hat{\Sigma}^g\hat{\Lambda}\right\|_{\text{sp}}+\left\|Q\Lambda^{0\prime}(\hat{\Sigma}^g-\Sigma^g)\hat{\Lambda}\right\|_{\text{sp}}+\left\|Q\Lambda^{0\prime}\Sigma^g[\hat{\Lambda}-\Lambda^0Q']\right\|_{\text{sp}}.$$

It is standard to show $\frac{1}{N}\left\|\hat{\Lambda}-\Lambda^0Q\right\|^2\leq\frac{1}{N}\left\|\hat{\Lambda}-\Lambda^0\hat{H}^{(\ell)-1}\right\|^2+\frac{1}{N}\left\|\Lambda^0(\hat{H}^{(\ell)-1}-Q)\right\|^2=o_P(1)$ by using the expression of $\hat{\lambda}_i-\hat{H}^{(\ell)-1}\lambda_i^0$ in the proof of Theorem 2.4(ii) and Lemma A.4(iii). In addition, $\left\|\hat{\Sigma}^g\right\|_{\text{sp}}\leq\|\Sigma^g\|_{\text{sp}}+\left\|\hat{\Sigma}^g-\Sigma^g\right\|_{\text{sp}}=O(1)+o_P(1)=O_P(1)$ by Lemma A.6. It follows that

$$\begin{aligned}\frac{1}{N}\left\|[\hat{\Lambda}-\Lambda^0Q']'\hat{\Sigma}^g\hat{\Lambda}\right\|_{\text{sp}} &\leq \left\|\hat{\Sigma}^g\right\|_{\text{sp}}\frac{1}{N^{1/2}}\left\|\hat{\Lambda}\right\|_{\text{sp}}\frac{1}{N^{1/2}}\left\|\hat{\Lambda}-\Lambda^0Q\right\|_{\text{sp}}\\ &\leq O_P(1)\frac{1}{N^{1/2}}\left\{\left\|\hat{\Lambda}-\Lambda^0(\hat{H}^{(\ell)\prime})^{-1}\right\|_{\text{sp}}+\left\|\Lambda^0\left[(\hat{H}^{(\ell)\prime})^{-1}-Q'\right]\right\|_{\text{sp}}\right\}=o_P(1).\end{aligned}$$

Similarly, by Lemma A.6, we have $\frac{1}{N}||Q\Lambda^{0\prime}(\hat{\Sigma}^g-\Sigma^g)\hat{\Lambda}||_{\text{sp}}\leq\|Q\|_{\text{sp}}\frac{1}{N^{1/2}}\left\|\Lambda^0\right\|_{\text{sp}}\frac{1}{N^{1/2}}||\hat{\Lambda}||_{\text{sp}}||\hat{\Sigma}^g-\Sigma^g||_{\text{sp}}=o_P(1)$. By the same token, we have $\frac{1}{N}||Q\Lambda^{0\prime}\Sigma^g[\hat{\Lambda}-\Lambda^0Q']||_{\text{sp}}=o_p(1)$. It follows that $\frac{1}{N}||\hat{\Lambda}'\hat{\Sigma}^g\hat{\Lambda}-Q\Lambda^{0\prime}\Sigma^g\Lambda^0Q'||_{\text{sp}}=o_P(1)$. ■

B Proofs of the main results in Section 3

In this appendix, we prove the main results in Section 3 by calling upon some technical lemmas and theorems whose proofs can be found in the online supplement.

To proceed, we introduce some notations. Note that the true number of factors is assumed to be R_0 but the working model is given by $X=F(R)\Lambda(R)'+\varepsilon(R)$, where we make the dependence of F and Λ on the assumed number of factors (R) explicit and $\varepsilon(R)\equiv X-F(R)\Lambda(R)'$. As in Bai and Ng (2019a), we want to establish the connection between the usual principal component (PC) estimators of the factors and factor loadings and the SVD estimators.

Let $X^*=P_{\Omega^*}X$. Note that $\tilde{C}_R=S_H(\frac{1}{p}P_{\Omega^*}X,R)=\tilde{U}_R\tilde{\Sigma}_R\tilde{V}_R'$, $\tilde{U}_R$ and $\tilde{V}_R$ are respectively the eigenvector matrices of $\frac{1}{p^2}X^*X^{*\prime}$ and $\frac{1}{p^2}X^{*\prime}X^*$ associated with their R largest eigenvalues, and

the diagonal elements of $\tilde{\Sigma}_R^2$ are the R largest eigenvalues of $\frac{1}{p^2}X^*X^{*\prime}$. Let $\tilde{F}^R$ and $\tilde{\Lambda}^R$ denote the conventional principal component (PC) estimators of $F(R)$ and $\Lambda(R)$ under the normalization restrictions that $T^{-1}F(R)'F(R) = I_R$ and $\Lambda(R)'\Lambda(R)$ =diagonal matrix. It is well known that $\tilde{F}^R$ is given by $\sqrt{T}$ times the normalized eigenvector matrix of $\frac{1}{p^2}X^*X^{*\prime}$ associated with its R largest eigenvalues and $\tilde{\Lambda}^{R\prime} = (\tilde{F}^{R\prime}\tilde{F}^R)^{-1}\tilde{F}^{R\prime}\frac{1}{p}X^* = \tilde{F}^{R\prime}\frac{1}{Tp}X^*$. This indicates that

$$\tilde{F}^R = \sqrt{T}\tilde{U}_R. \tag{B.1}$$

In addition, we consider the full SVD of $\frac{1}{p}X^*$: $\frac{1}{p}X^* = \tilde{U}\tilde{\Sigma}\tilde{V}' = \sum_{r=1}^{T\wedge N}\tilde{u}_r\tilde{v}_r'\tilde{\sigma}_r$. Then $\frac{1}{p}X^{*\prime}\tilde{U} = \tilde{V}\tilde{\Sigma}'\tilde{U}'\tilde{U} = \tilde{V}\tilde{\Sigma}'$. This implies that

$$\tilde{V}_R\tilde{\Sigma}_R = \frac{1}{p}X^{*\prime}\tilde{U}_R = \frac{\sqrt{T}}{Tp}X^{*\prime}\tilde{F}^R = \sqrt{T}\tilde{\Lambda}^R. \tag{B.2}$$

(B.1) says that $\tilde{U}_R$ is a scaled version of $\tilde{F}^R$ and (B.2) says that each column of $\tilde{V}_R$ is a scaled version of the corresponding column of $\tilde{\Lambda}^R$. It is easy to see that

$$\tilde{U}_R\tilde{\Sigma}_R\tilde{V}_R' = \tilde{F}^R\tilde{\Lambda}^{R\prime}. \tag{B.3}$$

That is, both the SVD and the PCA yield the same estimates of the common component once R is given. Following the lead of Bai and Ng (2002), we consider a rotational version of $\tilde{F}^R$: $\breve{F}^R = \left(NTp^2\right)^{-1}X^*X^{*\prime}\tilde{F}^R$. Let $\breve{H}_{1R} = (N^{-1}\Lambda^{0\prime}\Lambda^0)(T^{-1}F^{0\prime}\tilde{F}^R)$. The properties of $\breve{F}^R$ can be established along the lines of proofs in Bai and Ng (2002) and those in the proof of Theorem 2.1 in the presence of random missing values.

Alternatively, we can consider the PC estimation under the normalization restrictions that $N^{-1}\Lambda(R)'\Lambda(R) = I_R$ and $F(R)'F(R)$ =diagonal matrix. Let $\bar{F}^R$ and $\bar{\Lambda}^R$ denote the conventional PC estimators of $F(R)$ and $\Lambda(R)$ in this case. Then following the above arguments, we can show that

$$\bar{\Lambda}^R = \sqrt{N}\tilde{V}_R,\ \tilde{U}_R\tilde{\Sigma}_R = \sqrt{N}\bar{F}^R,\ \text{and}\ \tilde{U}_R\tilde{\Sigma}_R\tilde{V}_R' = \bar{F}^R\bar{\Lambda}^{R\prime}. \tag{B.4}$$

Following the lead of Bai and Ng (2002), we consider a rotational version of $\bar{\Lambda}_R$: $\breve{\Lambda}^R = \left(NTp^2\right)^{-1}X^{*\prime}X^*\bar{\Lambda}^R$. Let $\breve{H}_{2R} = (T^{-1}F^{0\prime}F^0)(N^{-1}\Lambda^{0\prime}\bar{\Lambda}^R)$.

Finally, let $\tilde{D}_R$ denote the $R\times R$ diagonal matrix that contains the R largest eigenvalues of $(NTp^2)^{-1}X^*X^{*\prime}$ arranged in descending order along its diagonal line. Note that $\tilde{D}_R = (NT)^{-1}\tilde{\Sigma}_R^2$. Recall that $\bar{g}_{it}^* = \mathbf{1}\{(i,t)\in\Omega_\perp^*\}$ and $g_{it}^* = \mathbf{1}\{(i,t)\in\Omega^*\}$. Let $\bar{G}^*$ be the $T\times N$ matrix with (t,i)th element given by $\bar{g}_{it}^*$. Define G^* analogously. Let e_{rR} denote the rth column of the $R\times R$ identity matrix I_R. Similarly, e_{rN} and e_{rT} denote the rth column of I_N and I_T, respectively. Note that $\tilde{u}_r \equiv \tilde{U}_Re_{rR}$ and $\tilde{v}_r \equiv \tilde{V}_Re_{rR}$, $r = 1,...,R$, denote the rth column of $\tilde{U}_R$ and $\tilde{V}_R$, respectively. In addition, $\tilde{C}_R = \sum_{r=1}^R\tilde{u}_r\tilde{v}_r'\tilde{\sigma}_r$.

The proof of Theorem 3.1 needs the following three lemmas.

Lemma B.1 *Suppose that all the conditions but Assumption A.7 in Theorem 3.1 hold. Then*
(i) $\frac{1}{T}\left\|\sqrt{T}\tilde{U}_R\tilde{D}_R - F^0\breve{H}_{1R}\right\|^2 = O_P\left(\delta_{NT}^{-2}\right)$,
(ii) $\frac{1}{N}\left\|\sqrt{N}\tilde{V}_R\tilde{D}_R - \Lambda^0\breve{H}_{2R}\right\|^2 = O_P\left(\delta_{NT}^{-2}\right)$.

Lemma B.2 *Let $\breve{\sigma}_r = (NT)^{-1/2}\tilde{\sigma}_r$. Let σ_r^2 denote the rth largest eigenvalue of $\Sigma_{F^0}\Sigma_{\Lambda^0}$ for $r = 1, ..., R_0$. Suppose that all the conditions but Assumption A.7 in Theorem 3.1 hold. Then*
(i) $\breve{\sigma}_r^2 = \sigma_r^2 + O_P\left(\delta_{NT}^{-1}\right)$ for $r = 1, ..., R_0$,
(ii) $\breve{\sigma}_{R_0+r}^2 = O_P\left(\delta_{NT}^{-2}\right)$ for $r \geq 1$,
(iii) $\delta_{NT}^2\breve{\sigma}_{R_0+r}^2 \geq c_\sigma + o_P(1)$ for some positive constant c_σ and any $r \geq 1$ with $R_0 + r \leq R$.

Lemma B.3 *Let $\tilde{u}_r$ and $\tilde{v}_r$ be the rth left and right singular vector of $\frac{1}{p}X^*$. Suppose that all the conditions but Assumption A.7 in Theorem 3.1 hold. Then for $r = R_0 + 1, ..., R_{\max}$, we have $\tilde{u}_r'F^0 = O_P(\delta_{NT}^{-1})$ and $\tilde{v}_r'\Lambda^0 = O_P(\delta_{NT}^{-1})$.*

To proceed, we define some notations. For a real matrix Γ, recall that $\|\Gamma\|$ and $\|\Gamma\|_\infty$ denote its Frobenius norm and entrywise L_∞ norm, respectively. We use $\|\Gamma\|_*$ to denote the nuclear norm of Γ, which is defined as the summation of the singular values of Γ. For a nonzero matrix $\Gamma \in \mathbb{R}^{T\times N}$, we define two measures to control its spikeness and rank. First, we define the spikeness ratio as $\alpha_{sp}(\Gamma) \equiv \frac{\sqrt{NT}\|\Gamma\|_\infty}{\|\Gamma\|}$, which satisfies $1 \leq \alpha_{sp}(\Gamma) \leq \sqrt{NT}$. The lower bound can be reached when all the entries of Γ are the same, and the upper bound can be reached when there is only one nonzero entry in Γ. Next, we define a tractable measure of how close Γ is to a low-rank matrix via the ratio $\beta_{ra}(\Gamma) \equiv \frac{\|\Gamma\|_*}{\|\Gamma\|}$. Note that $1 \leq \beta_{ra}(\Gamma) \leq \delta_{NT} \equiv \sqrt{N} \wedge \sqrt{T}$. Let $d = (N + T)/2$. Define the constraint set

$$\mathcal{C}_{NT}(c_0) \equiv \left\{\Gamma \in \mathbb{R}^{N\times T}, \Gamma \neq 0 \mid \alpha_{sp}(\Gamma)\beta_{ra}(\Gamma) \leq \frac{1}{c_0}\sqrt{\frac{NT}{d\log d}}\right\}, \tag{B.5}$$

where c_0 is a universal constant. For a low rank matrix $\Gamma \in \mathcal{C}_{NT}(c_0)$, the constraint requires it to be not very spiky.

The following two theorems are needed to show that the probability of overselecting the number of factors is approaching zero.

Theorem B.4 *Let G be a $T \times N$ random matrix with all entries i.i.d. from the Bernoulli distribution with parameter $p \in (0, 1)$. There are universal constants c_0, c_1, c_2, and c_3 such that*

$$\left\|\frac{1}{\sqrt{p}}\Gamma \circ G\right\| \geq \frac{1}{8}\|\Gamma\|\left\{1 - \frac{c_3\alpha_{sp}(\Gamma)}{\sqrt{NT}}\right\} \qquad \text{for all } \Gamma \in \mathcal{C}_{NT}(c_0)$$

with probability greater than $1 - c_1\exp(-c_2 NT\log d/d)$.

Theorem B.5 *Let G be a $T \times N$ random matrix with all entries i.i.d. from the Bernoulli distribution with parameter $p \in (0, 1)$. Then*

$$\sup_{\Gamma\in\mathcal{C}_{1NT}}\|\Gamma \circ [G - E(G)]\|_{sp} = O_P\left(c_{1NT} + c_{2NT} + c_{3NT}\sqrt{(N + T)\log\log(N + T)} + 1/\log(N + T)\right),$$

where $\mathcal{C}_{1NT} \equiv \mathcal{C}_{1NT}(c_{1NT}, c_{2NT}, c_{3NT}) \equiv \{\Gamma \in \mathbb{R}^{N\times T}, \mid \Gamma = UV', U \in \mathbb{R}^T$ and $V \in \mathbb{R}^N$ are vectors such that $\|U\| = \|V\| = 1$, $\|U\|_\infty \leq c_{1NT}$, $\|V\|_\infty \leq c_{2NT}$, $\|U\|_\infty\|\|V\|_\infty\| \leq c_{3NT}\}$.

Proof of Theorem 3.1. Noting that $X = C^0 + \varepsilon$, we have $\widetilde{CV}(R) = \frac{1}{NT}\left\|(X - \tilde{C}_R) \circ \bar{G}^*\right\|^2$ $= \frac{1}{NT}\left\|(C^0 - \tilde{C}_R) \circ \bar{G}^*\right\|^2 + \frac{1}{NT}\left\|\varepsilon \circ \bar{G}^*\right\|^2 + \frac{2}{NT}\mathrm{tr}\Big\{\Big[(C^0 - \tilde{C}_R) \circ \bar{G}^*\Big]\left(\varepsilon \circ \bar{G}^*\right)'\Big\} \equiv \widetilde{CV}_1(R) + \widetilde{CV}_2 + 2\widetilde{CV}_3(R)$, where $\widetilde{CV}_2$ does not depend on R. Then

$$\widetilde{CV}(R) - \widetilde{CV}(R_0) = \Big[\widetilde{CV}_1(R) - \widetilde{CV}_1(R_0)\Big] + 2\Big[\widetilde{CV}_3(R) - \widetilde{CV}_3(R_0)\Big]. \tag{B.6}$$

It is sufficient to study the asymptotic properties of $\widetilde{CV}_1(R) - \widetilde{CV}_1(R_0)$ and $\widetilde{CV}_3(R) - \widetilde{CV}_3(R_0)$ under the under-fitted and over-fitted cases, respectively.

We first study the under-fitted case where $R < R_0$. Noting that $\|A\|^2 - \|B\|^2 = \mathrm{tr}(A'A - B'B)$ $= \mathrm{tr}\{(A-B)'(A-B)\} + 2\mathrm{tr}\big((A-B)'B\big)$, we have

$$\begin{aligned}\widetilde{CV}_1(R) - \widetilde{CV}_1(R_0) &= \frac{1}{NT}\left\|(\tilde{C}_R - \tilde{C}_{R_0}) \circ \bar{G}^*\right\|^2 + \frac{2}{NT}\mathrm{tr}\left\{\Big[(\tilde{C}_R - \tilde{C}_{R_0}) \circ \bar{G}^*\Big]'\Big[(\tilde{C}_{R_0} - C^0) \circ \bar{G}^*\Big]\right\} \\ &\equiv \widetilde{CV}_{11}(R) + 2CV_{12}(R). \end{aligned} \tag{B.7}$$

Noting that $\tilde{C}_{R_0} - \tilde{C}_R = \sum_{r=R+1}^{R_0} \tilde{u}_r \tilde{v}_r' \tilde{\sigma}_r$, $\tilde{u}_r = \tilde{U}_{R_0} e_{rR_0}$, $\tilde{v}_r = \tilde{V}_{R_0} e_{rR_0}$, and $\breve{\sigma}_r = (NT)^{-1/2}\tilde{\sigma}_r$, we have

$$\begin{aligned}\widetilde{CV}_{11}(R) &= \frac{1}{NT}\left\|\left(\sum_{r=R+1}^{R_0} \tilde{u}_r \tilde{v}_r' \tilde{\sigma}_r\right) \circ \bar{G}^*\right\|^2 = \frac{1}{NT}\left\|\left(\sum_{r=R+1}^{R_0} \tilde{U}_{R_0} e_{rR_0} e_{rR_0}' \tilde{V}_{R_0}' \tilde{\sigma}_r\right) \circ \bar{G}^*\right\|^2 \\ &= \frac{1}{NT}\left\|\left(\sum_{r=R+1}^{R_0} \left(\sqrt{N}\tilde{U}_{R_0}\tilde{D}_{R_0}\right)\tilde{D}_{R_0}^{-1} e_{rR_0} e_{rR_0}' \tilde{D}_{R_0}^{-1}(\sqrt{T}\tilde{V}_{R_0}\tilde{D}_{R_0})' \breve{\sigma}_r\right) \circ \bar{G}^*\right\|^2. \end{aligned} \tag{B.8}$$

Let $\varsigma_{1R} = \sqrt{N}\tilde{U}_R\tilde{D}_R - F^0\breve{H}_{1R}$ and $\varsigma_{2R} = \sqrt{N}\tilde{V}_R\tilde{D}_R - \Lambda^0\breve{H}_{2R}$. Then $\sqrt{N}\tilde{U}_{R_0}\tilde{D}_{R_0} = F^0\breve{H}_{1R_0} + \varsigma_{1R_0}$ and $\sqrt{N}\tilde{V}_{R_0}\tilde{D}_{R_0} = \Lambda^0\breve{H}_{2R_0} + \varsigma_{2R_0}$. It is easy to apply Lemma B.1 to show that

$$\begin{aligned}&\widetilde{CV}_{11}(R) \\ =\ & \frac{1}{NT}\left\|\left(\sum_{r=R+1}^{R_0}\left(F^0\breve{H}_{1R_0} + \varsigma_{1R_0}\right)\tilde{A}_{rR_0}(\Lambda^0\breve{H}_{2R_0} + \varsigma_{2R_0})'\breve{\sigma}_r\right) \circ \bar{G}^*\right\|^2 \\ =\ & \frac{1}{NT}\left\|\left(\sum_{r=R+1}^{R_0} F^0\breve{H}_{1R_0}\tilde{A}_{rR_0}\breve{H}_{2R_0}'\Lambda^{0\prime}\breve{\sigma}_r\right) \circ \bar{G}^*\right\|^2 + O_P\left(\delta_{NT}^{-1}\right) \\ =\ & \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\sum_{r=R+1}^{R_0} e_{iT}'F^0\breve{H}_{1R_0}\tilde{A}_{rR_0}\breve{H}_{2R_0}'\Lambda^{0\prime}e_{iN}\breve{\sigma}_{R_0}\right)^2 \bar{g}_{it}^* + O_P\left(\delta_{NT}^{-1}\right) \\ =\ & \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{r=R+1}^{R_0}\sum_{l=R+1}^{R_0}\mathrm{tr}\left\{\breve{H}_{1R_0}\tilde{A}_{rR_0}\breve{H}_{2R_0}'\Lambda^{0\prime}e_{iN}e_{iN}'\Lambda^0\breve{H}_{2R_0}\tilde{A}_{lR_0}'\breve{H}_{1R_0}'F^0e_{tT}e_{tT}'F^{0\prime}\right\}\breve{\sigma}_r\breve{\sigma}_l\bar{g}_{it}^* \\ &+O_P\left(\delta_{NT}^{-1}\right) \\ =\ & \sum_{r=R+1}^{R_0}\sum_{l=R+1}^{R_0}[\mathrm{vec}(\breve{H}_{1R_0}\tilde{A}_{rR_0}\breve{H}_{2R_0}')]'\left\{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}[(\Lambda^{0\prime}e_{iN}e_{iN}'\Lambda^0)\otimes(F^{0\prime}e_{tT}e_{tT}'F^{0\prime})]\bar{g}_{it}^*\right\}\breve{\sigma}_r\breve{\sigma}_l \\ &\times\mathrm{vec}(\breve{H}_{1R_0}\tilde{A}_{lR_0}\breve{H}_{2R_0}') + O_P\left(\delta_{NT}^{-1}\right) \end{aligned} \tag{B.9}$$

where $\tilde{A}_{rR} = \tilde{D}_R^{-1} e_{rR} e_{rR}' \tilde{D}_R^{-1}$, $\bar{g}_{it}^* = \mathbf{1}\{(i,t) \in \Omega_\perp^*\}$, and the last equality follows from the fact that $\text{tr}(A_1A_2A_3A_4) = [\text{vec}(A_1)]'(A_2 \otimes A_4')\text{vec}(A_3')$ and the Fubini theorem. Now using $\bar{g}_{it}^* = (1-p) + [\bar{g}_{it}^* - (1-p)]$ and the fact that $\bar{g}_{it}^*$ are i.i.d. and independent of $\left(\Lambda^{0\prime}, F^{0\prime}\right)$, we can readily show that

$$\begin{aligned}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}[(\Lambda^0 e_{iN}' e_{iN}' \Lambda^{0\prime}) \otimes (F^0 e_{tT} e_{tT}' F^{0\prime})]\bar{g}_{it}^* &= \frac{1-p}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}[(\Lambda^0 e_{iN}' e_{iN}' \Lambda^{0\prime}) \otimes (F^0 e_{tT} e_{tT}' F^{0\prime})] \\ &\quad + O_P((NT)^{-1/2}).\end{aligned}$$

It follows that

$$\begin{aligned}\widetilde{CV}_{11}(R) &= (1-p)\sum_{r=R+1}^{R_0}\sum_{l=R+1}^{R_0}[\text{vec}(\breve{H}_{1R_0}\tilde{A}_{rR_0}\breve{H}_{2R_0}')]'\left\{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}[(\Lambda^{0\prime} e_{iN} e_{iN}' \Lambda^{0\prime}) \otimes (F^{0\prime} e_{tT} e_{tT}' F^{0\prime})]\right\} \\ &\quad \times \text{vec}(\breve{H}_{1R_0}\tilde{A}_{lR_0}\breve{H}_{2R_0}')\breve{\sigma}_r\breve{\sigma}_l + O_P\left(\delta_{NT}^{-1}\right) \\ &= \frac{1-p}{NT}\left\|\sum_{r=R+1}^{R_0}\tilde{u}_r\tilde{v}_r'\tilde{\sigma}_r\right\|^2 + O_P\left(\delta_{NT}^{-1}\right) = (1-p)\sum_{r=R+1}^{R_0}(NT)^{-1}\tilde{\sigma}_r^2 + O_P\left(\delta_{NT}^{-1}\right) \\ &= (1-p)\sum_{r=R+1}^{R_0}\sigma_r^2 + O_P\left(\delta_{NT}^{-1}\right),\end{aligned}$$

where the second equality is obtained by reversing the operations in (B.9) and (B.8), the third equality holds by the fact that $\tilde{U}_R'\tilde{U}_R = I_R$ and $\breve{V}_R'\breve{V}_R = I_R$, and the fourth equality follows because $(NT)^{-1}\tilde{\sigma}_r^2 = \sigma_r^2 + O_P(\delta_{NT}^{-1})$ for $r \le R_0$ by Lemma B.2(i).

Following the proof of Theorem 2.4, we can show that $\frac{1}{NT}\left\|(C^0 - \tilde{C}_{R_0}) \circ \bar{G}^*\right\|^2 \le \frac{1}{NT}\left\|C^0 - \tilde{C}_{R_0}\right\|^2 = O_P\left(\delta_{NT}^{-2}\right)$. Then $\left|\widetilde{CV}_{12}(R)\right| \le \left\{\frac{1}{NT}\left\|\left(\tilde{C}_R - \tilde{C}_{R_0}\right) \circ \bar{G}^*\right\|^2\right\}^{1/2}\left\{\frac{1}{NT}\left\|\left(C^0 - \tilde{C}_{R_0}\right) \circ \bar{G}^*\right\|^2\right\}^{1/2} = O_P(1)\,O_P\left(\delta_{NT}^{-1}\right) = O_P\left(\delta_{NT}^{-1}\right)$. It follows that $\widetilde{CV}_1(R) - \widetilde{CV}_1(R_0) = (1-p)\sum_{r=R+1}^{R_0}\sigma_r^2 + O_P\left(\delta_{NT}^{-1}\right)$.

Next, $\widetilde{CV}_3(R) - \widetilde{CV}_3(R_0) = \frac{1}{NT}\text{tr}\{[(\tilde{C}_{R_0} - \tilde{C}_R) \circ \bar{G}^*]\left(\varepsilon \circ \bar{G}^*\right)'\}$. Noting that $\frac{1}{NT}\left\|\varepsilon \circ \bar{G}^*\right\|^2 \le \frac{1}{NT}\|\varepsilon\|^2 = O_P(1)$, we can apply Lemma B.1 and follow the analysis of $\widetilde{CV}_{11}(R)$ to show that

$$\begin{aligned}&\widetilde{CV}_3(R) - \widetilde{CV}_3(R_0) \\ &= \frac{1}{NT}\text{tr}\left\{\left(\sum_{r=R+1}^{R_0}\left[\left(F^0\breve{H}_{1R_0} + \varsigma_{1R_0}\right)\tilde{A}_{rR_0}(\Lambda^0\breve{H}_{2R_0} + \varsigma_{2R_0})'\breve{\sigma}_r\right] \circ \bar{G}^*\right)\left(\varepsilon \circ \bar{G}^*\right)'\right\} \\ &= \frac{1}{NT}\text{tr}\left\{\left(\sum_{r=R+1}^{R_0}\left[F^0\breve{H}_{1R_0}\tilde{A}_{rR_0}\breve{H}_{2R_0}'\Lambda^{0\prime}\breve{\sigma}_r\right] \circ \bar{G}^*\right)\left(\varepsilon \circ \bar{G}^*\right)'\right\} + O_P\left(\delta_{NT}^{-1}\right) \\ &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{r=R+1}^{R_0}\text{tr}\left(e_{iT}'F^0\breve{H}_{1R_0}\tilde{A}_{rR_0}\breve{H}_{2R_0}'\Lambda^{0\prime}e_{iN}\right)\breve{\sigma}_r\varepsilon_{it}\bar{g}_{it}^* + O_P\left(\delta_{NT}^{-1}\right) \\ &= \sum_{r=R+1}^{R_0}\text{tr}\left(\breve{H}_{1R_0}\tilde{A}_{rR_0}\breve{H}_{2R_0}'\ \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Lambda^{0\prime}e_{iN}e_{iT}'F^0\varepsilon_{it}\bar{g}_{it}^*\right)\breve{\sigma}_r + O_P\left(\delta_{NT}^{-1}\right) \\ &= (1-p)\sum_{r=R+1}^{R_0}\text{tr}\left(\breve{H}_{1R_0}\tilde{A}_{rR_0}\breve{H}_{2R_0}'\ \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Lambda^{0\prime}e_{iN}e_{iT}'F^0\varepsilon_{it}\right)\breve{\sigma}_r + O_P\left(\delta_{NT}^{-1}\right) = O_P\left(\delta_{NT}^{-1}\right),\end{aligned}$$

where the last line follows from the fact that $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Lambda^{0\prime}e_{iN}e_{iT}^{\prime}F^{0}\varepsilon_{it}=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\lambda_{i}^{0}F_{t}^{0\prime}\varepsilon_{it}=O_{P}((NT)^{-1/2})$.

In sum, we have shown that when $R<R_{0}$, $\widetilde{CV}(R)-\widetilde{CV}(R_{0})=(1-p)\sum_{r=R+1}^{R_{0}}\sigma_{r}^{2}+O_{P}\left(\delta_{NT}^{-1}\right)$. This implies that $P(\tilde{R}<R_{0})\rightarrow 0$ as $(N,T)\rightarrow\infty$.

Now, we study the overfitted case where $R>R_{0}$. We continue to use the decompositions in (B.6) and (B.7). We first study $\widetilde{CV}_{11}(R)$. When $R>R_{0}$, $\tilde{D}_{R}^{-1}\neq O_{P}(1)$ and thus $\tilde{A}_{rR}\neq O_{P}(1)$. This implies that we cannot use similar arguments as used in the case where $R<R_{0}$. In addition, $\tilde{C}_{R}-\tilde{C}_{R_{0}}$ is not independent of $\bar{G}^{*}$, which further complicates the analysis. To tackle the problem, we call upon Assumption A.7. Let $\tilde{\Gamma}_{R}\equiv\tilde{C}_{R}-\tilde{C}_{R_{0}}$. By Assumption A.7(i), we have $\left\|\tilde{\Gamma}_{R}\right\|_{\infty}\leq\sum_{r=R_{0}+1}^{R}\tilde{\sigma}_{r}/(c_{0}\sqrt{(N+T)\log(N+T)})$ with probability approaching 1 (w.p.a.1). In addition, by the definitions of Frobenius and nuclear norms, $\left\|\tilde{\Gamma}_{R}\right\|=(\sum_{r=R_{0}+1}^{R}\tilde{\sigma}_{r}^{2})^{1/2}$ and $\left\|\tilde{\Gamma}_{R}\right\|_{*}=\sum_{r=R_{0}+1}^{R}\tilde{\sigma}_{r}$. By the Jensen inequality and the fact that $R\leq R_{\max}$,

$$\frac{\sqrt{NT}\left\|\tilde{\Gamma}_{R}\right\|_{\infty}\left\|\tilde{\Gamma}_{R}\right\|_{*}}{\left\|\tilde{\Gamma}_{R}\right\|^{2}}\leq\frac{R_{\max}-R_{0}}{c_{0}}\sqrt{\frac{NT}{(N+T)\log(N+T)}}\leq\frac{1}{\tilde{c}_{0}}\sqrt{\frac{NT}{d_{NT}\log d_{NT}}},$$

where $d_{NT}=\frac{1}{2}(N+T)$ and $\tilde{c}_{0}=\sqrt{2}c_{0}/(R_{\max}-R_{0})$. Therefore, $\tilde{\Gamma}_{R}\in\mathcal{C}_{NT}(\tilde{c}_{0})$ w.p.a.1. Then we can apply Theorem B.4 and the fact that $\left\|\tilde{\Gamma}_{R}\right\|_{\infty}/\left\|\tilde{\Gamma}_{R}\right\|=o_{P}(1)$ to obtain that $\left\|\frac{1}{\sqrt{1-p}}\tilde{\Gamma}_{R}\circ\bar{G}^{*}\right\|\geq\frac{1}{16}\left\|\tilde{\Gamma}_{R}\right\|$ w.p.a.1. It follows that $\widetilde{CV}_{11}(R)=\frac{1}{NT}\left\|\tilde{\Gamma}_{R}\circ\bar{G}^{*}\right\|^{2}\geq\frac{1-p}{256}\frac{1}{NT}\left\|\tilde{\Gamma}_{R}\right\|^{2}=\frac{1-p}{256}\sum_{r=R_{0}+1}^{R}\breve{\sigma}_{r}^{2}$ w.p.a.1, where $\breve{\sigma}_{r}^{2}=O_{P}\left(\delta_{NT}^{-2}\right)$ for $r=R_{0}+1,...,R_{\max}$ by Lemma B.2(ii). Then by Lemma B.2(iii) we have $\operatorname{plim}_{(N,T)\rightarrow\infty}\delta_{NT}^{2}\widetilde{CV}_{11}(R)\geq(R-R_{0})\frac{1-p}{256}c_{\sigma}>0$.

Next, we study $\widetilde{CV}_{12}(R)$. Noting that $\tilde{\Gamma}_{R}=\tilde{C}_{R}-\tilde{C}_{R_{0}}=\sum_{r=R_{0}+1}^{R}\tilde{u}_{r}\tilde{v}_{r}^{\prime}\tilde{\sigma}_{r}$, we have $\widetilde{CV}_{12}(R)=\frac{1}{NT}\mathrm{tr}\left\{(\tilde{\Gamma}_{R}\circ\bar{G}^{*})\left[(\tilde{C}_{R_{0}}-C^{0})\circ\bar{G}^{*}\right]^{\prime}\right\}=\sum_{r=R_{0}+1}^{R}\frac{\tilde{\sigma}_{r}}{NT}\mathrm{tr}\left\{(\tilde{u}_{r}\tilde{v}_{r}^{\prime})^{\prime}\left[(\tilde{C}_{R_{0}}-C^{0})\circ\bar{G}^{*}\right]\right\}\equiv\sum_{r=R_{0}+1}^{R}CV_{12r}$. In addition,

$$\frac{1}{\sqrt{NT}}\mathrm{tr}\left\{(\tilde{u}_{r}\tilde{v}_{r}^{\prime})^{\prime}(\tilde{C}_{R_{0}}-C^{0})\right\}=\frac{-1}{\sqrt{NT}}\mathrm{tr}\left\{(\tilde{u}_{r}\tilde{v}_{r}^{\prime})^{\prime}C^{0}\right\}=\frac{-1}{\sqrt{NT}}\mathrm{tr}\left\{\tilde{u}_{r}^{\prime}F^{0}\Lambda^{0\prime}\tilde{v}_{r}\right\}=O_{P}(\delta_{NT}^{-4}),$$

where the first equality holds by the orthogonality between $\tilde{u}_{r}$ and $\tilde{C}_{R_{0}}$ for $r>R_{0}$ and the third equality holds by Lemma B.3. It follows that $CV_{12r}=\frac{\tilde{\sigma}_{r}}{NT}\mathrm{tr}\left\{(\tilde{u}_{r}\tilde{v}_{r}^{\prime})^{\prime}\left[(\tilde{C}_{R_{0}}-C^{0})\circ\bar{G}^{*}\right]\right\}=\overline{CV}_{12r}+O_{P}(\delta_{NT}^{-4})$, where $\overline{CV}_{12r}=-\frac{\tilde{\sigma}_{r}}{NT}\mathrm{tr}\left\{(\tilde{u}_{r}\tilde{v}_{r}^{\prime})^{\prime}\ \left[(\tilde{C}_{R_{0}}-C^{0})\circ(G^{*}-E(G^{*}))\right]\right\}$. Note that

$$\begin{aligned}\left|\overline{CV}_{12r}\right| &= \frac{\tilde{\sigma}_{r}}{\sqrt{NT}}\frac{1}{\sqrt{NT}}\left|\mathrm{tr}\left\{(\tilde{C}_{R_{0}}-C^{0})\left[(\tilde{u}_{r}\tilde{v}_{r}^{\prime})\circ(G^{*}-E(G^{*}))\right]\right\}\right| \\ &\leq O_{P}\left(\delta_{NT}^{-1}\right)\frac{1}{\sqrt{NT}}\left\|\tilde{C}_{R_{0}}-C^{0}\right\|_{*}\left\|(\tilde{u}_{r}\tilde{v}_{r}^{\prime})\circ(G^{*}-E(G^{*}))\right\|_{\mathrm{sp}} \\ &\leq O_{P}\left(\delta_{NT}^{-2}\right)\sup_{\Gamma\in\mathcal{C}_{1NT}(c_{1NT},c_{2NT},c_{3NT})}\left\|\Gamma\circ(G^{*}-E(G^{*}))\right\|_{\mathrm{sp}}=o_{P}\left(\delta_{NT}^{-2}\right),\end{aligned}$$

where first inequality follows the fact that $\frac{\tilde{\sigma}_{r}}{\sqrt{NT}}=O_{P}\left(\delta_{NT}^{-1}\right)$ and $\left|\mathrm{tr}\left(AB\right)\right|\leq\left\|A\right\|_{*}\left\|B\right\|_{\mathrm{sp}}$, the second inequality follows because $\frac{1}{\sqrt{NT}}\left\|\tilde{C}_{R_{0}}-C^{0}\right\|_{*}\leq\frac{\sqrt{2R_{0}}}{\sqrt{NT}}\left\|\tilde{C}_{R_{0}}-C^{0}\right\|=O_{P}\left(\delta_{NT}^{-1}\right)$ and the last equality

holds by Theorem B.5 with $c_{1NT} = o(1)$, $c_{2NT} = o(1)$ and $c_{3NT} = 1/\sqrt{(N+T)\log(N+T)}$. Then we have $\widetilde{CV}_{12}(R) = o_P\left(\delta_{NT}^{-2}\right)$.

Now, we study $\widetilde{CV}_3(R) - \widetilde{CV}_3(R_0)$. Note that

$$\begin{aligned}
\widetilde{CV}_3(R) - \widetilde{CV}_3(R_0) &= \frac{1}{NT}\mathrm{tr}\left\{\left[\left(\tilde{C}_{R_0} - \tilde{C}_R\right)\circ \bar{G}^*\right]\left(\varepsilon \circ \bar{G}^*\right)'\right\} \\
&= -\sum_{r=R_0+1}^{R}\frac{\tilde{\sigma}_r}{NT}\mathrm{tr}\left\{\left(\tilde{u}_r\tilde{v}_r'\right)'\left(\varepsilon\circ\bar{G}^*\right)\right\} = -\sum_{r=R_0+1}^{R}\frac{\tilde{\sigma}_r}{NT}\tilde{u}_r'\left(\varepsilon\circ\bar{G}^*\right)\tilde{v}_r \\
&= -\sum_{r=R_0+1}^{R}\frac{\tilde{\sigma}_r}{NT}\sum_{i,t}\tilde{u}_{tr}\tilde{v}_{ir}\varepsilon_{it}(1-g_{it}^*) \equiv -\sum_{r=R_0+1}^{R} CV_{3r},
\end{aligned}$$

where $\tilde{u}_{tr}$ and $\tilde{v}_{ir}$ denote the tth and ith entries of $\tilde{u}_r$ and $\tilde{v}_r$, respectively. Noting that $\tilde{\sigma}_r^2/(NT) = O_P(\delta_{NT}^{-2})$, we have

$$\begin{aligned}
E_{\mathcal{D}_{NT}}\left[CV_{3r}^2\right] &= \frac{\tilde{\sigma}_r^2}{NT}\frac{1}{NT}\sum_{(i,t)\in\Omega_\perp^*}\sum_{(j,s)\in\Omega_\perp^*}\tilde{u}_{tr}\tilde{v}_{ir}\tilde{u}_{sr}\tilde{v}_{jr}E_{\mathcal{D}_{NT}}(\varepsilon_{it}\varepsilon_{js}) \\
&\le O_P(\delta_{NT}^{-2})\frac{1}{2NT}\sum_{(i,t)\in\Omega_\perp^*}\sum_{(j,s)\in\Omega_\perp^*}\left(\tilde{u}_{tr}^2\tilde{v}_{ir}^2 + \tilde{u}_{sr}^2\tilde{v}_{jr}^2\right)\left|E_{\mathcal{D}_{NT}}(\varepsilon_{it}\varepsilon_{js})\right| \\
&= O_P(\delta_{NT}^{-2})\frac{1}{NT}\sum_{(i,t)\in\Omega_\perp^*}\tilde{u}_{tr}^2\tilde{v}_{ir}^2\sum_{(j,s)\in\Omega_\perp^*}\left|E_{\mathcal{D}_{NT}}(\varepsilon_{it}\varepsilon_{js})\right| \\
&\le O_P(\delta_{NT}^{-2})\frac{1}{NT}\max_{(i,t)\in\Omega_\perp^*}\sum_{(j,s)\in\Omega_\perp^*}\left|E_{\mathcal{D}_{NT}}(\varepsilon_{it}\varepsilon_{js})\right| = o_P(\delta_{NT}^{-4}),
\end{aligned}$$

where $E_{\mathcal{D}_{NT}}(\cdot) = E(\cdot|P_{\Omega^*}X, \Omega^*)$, the first inequality holds by the CS inequality, the second inequality holds by the fact that $\sum_{(i,t)\in\Omega_\perp^*}\tilde{u}_{tr}^2\tilde{v}_{ir}^2 \le \|\tilde{u}_r\|^2\|\tilde{v}_r\|^2 = 1$, and the last equality holds by Assumption 7(ii). Hence, $CV_{3r} = o_p(\delta_{NT}^{-2})$ for each $r \in (R_0, R]$ and $\widetilde{CV}_3(R) - \widetilde{CV}_3(R_0) = o_p(\delta_{NT}^{-2})$. It follows that

$$\mathrm{plim}_{(N,T)\to\infty}\delta_{NT}^2\left[\widetilde{CV}(R) - \widetilde{CV}(R_0)\right] \ge \frac{(R-R_0)(1-p)}{256}c_\sigma > 0 \text{ for any } R > R_0.$$

This implies that $P(\tilde{R} > R_0) \to 0$ as $(N,T) \to \infty$. This completes the proof of the theorem. ■

Proof of Theorem 3.2. The proof is essentially the same as that of Theorem 3.1 given the results in Theorem 2.4. Here, we only outline the major differences. Let $\hat{X}^* = \hat{X}^{*(\ell^*)}$. Noting that $\hat{C}_R = S_H(\hat{X}^*, R) = \hat{U}_R\hat{\Sigma}_R\hat{V}_R'$, $\hat{U}_R$ and $\hat{V}_R$ are respectively the eigenvector matrices of $\hat{X}^*\hat{X}^{*\prime}$ and $\hat{X}^{*\prime}\hat{X}^*$ associated with their R largest eigenvalues, and the diagonal elements of $\hat{\Sigma}_R^2$ are the R largest eigenvalues of $\hat{X}^*\hat{X}^{*\prime}$. Let $\hat{F}^R$ and $\hat{\Lambda}^R$ denote the conventional principal component (PC) estimators of $F(R)$ and $\Lambda(R)$ based on $\hat{X}^*$ under the normalization restrictions that $T^{-1}F(R)'F(R) = I_R$ and $\Lambda(R)'\Lambda(R)$ =diagonal matrix. Let $\ddot{F}^R$ and $\ddot{\Lambda}^R$ denote the conventional PC estimators of $F(R)$ and $\Lambda(R)$ based on $\hat{X}^*$ under the normalization restrictions that $N^{-1}\Lambda(R)'\Lambda(R) = I_R$ and $F(R)'F(R)$ =diagonal matrix. Let $\ddot{H}_{1R} = (N^{-1}\Lambda^{0\prime}\Lambda^0)(T^{-1}\Lambda^{0\prime}\hat{F}^R)$ and $\ddot{H}_{2R} = (T^{-1}F^{0\prime}F^0)(N^{-1}F^{0\prime}\ddot{\Lambda}^R)$. Let $\hat{D}_R$

denote the $R \times R$ diagonal matrix that contains the R largest eigenvalues of $(NT)^{-1}\hat{X}^*\hat{X}^{*\prime}$ arranged in descending order along its diagonal line. Note that $\hat{D}_R = (NT)^{-1}\hat{\Sigma}_R^2$.

Following the proof of Theorem 2.4, we can show that $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|\hat{C}_{R_{\max},it}^{(\ell^*-1)} - C_{it}^0\right\|^2 = O_P\left(\delta_{NT}^{-2}\right)$. With this result, we can show that the results analogous to those in Lemmas B.1-B.2 hold: (i) $\frac{1}{T}\left\|\sqrt{T}\hat{U}_R\hat{D}_R - F^0\ddot{H}_{1R}\right\|^2 = O_P\left(\delta_{NT}^{-2}\right)$, (ii) $\frac{1}{N}\left\|\sqrt{N}\hat{V}_R\hat{D}_R - \Lambda^0\ddot{H}_{2R}\right\|^2 = O_P\left(\delta_{NT}^{-2}\right)$, (iii) $\ddot{\sigma}_r^2 = \sigma_r^2 + O_P\left(\delta_{NT}^{-1}\right)$ for $r = 1, ..., R_0$, (iv) $\ddot{\sigma}_{R_0+r}^2 = O_P\left(\delta_{NT}^{-2}\right)$ for $r \geq 1$, and (v) $\delta_{NT}^2\ddot{\sigma}_{R_0+r}^2 \geq c_\sigma + o_P(1)$ for some positive constant c_σ and any $r \geq 1$ with $R_0 + r \leq R$, where $\ddot{\sigma}_r = (NT)^{-1/2}\hat{\sigma}_r$. Noting that $X = C^0 + \varepsilon$, we make the following decomposition

$$\begin{aligned}
\widehat{CV}(R) &= \frac{1}{NT}\left\|(X - \hat{C}_R)\circ\bar{G}^*\right\|^2 \\
&= \frac{1}{NT}\left\|(C^0 - \hat{C}_R)\circ\bar{G}^*\right\|^2 + \frac{1}{NT}\left\|\varepsilon\circ\bar{G}^*\right\|^2 + \frac{2}{NT}\mathrm{tr}\left\{\left[(C^0 - \hat{C}_R)\circ\bar{G}^*\right]\left(\varepsilon\circ\bar{G}^*\right)'\right\} \\
&\equiv \widehat{CV}_1(R) + \widehat{CV}_2 + 2\widehat{CV}_3(R).
\end{aligned}$$

Then we have $\widehat{CV}(R) - \widehat{CV}(R_0) = [\widehat{CV}_1(R) - \widehat{CV}_1(R_0)] + 2[\widehat{CV}_3(R) - \widehat{CV}_3(R_0)]$. When $R < R_0$, we can follow the proof of Theorem 3.1 and apply the above results in (i)-(iii) to show that

$$\widehat{CV}_1(R) - \widehat{CV}_1(R_0) = (1-p)\sum_{r=R+1}^{R_0}\sigma_r^2 + O_P\left(\delta_{NT}^{-1}\right) \text{ and } \widehat{CV}_3(R) - \widehat{CV}_3(R_0) = O_P\left(\delta_{NT}^{-1}\right).$$

Then $\widehat{CV}(R) - \widehat{CV}(R_0) = (1-p)\sum_{r=R+1}^{R_0}\sigma_r^2 + O_P\left(\delta_{NT}^{-1}\right)$ and $P(\hat{R} < R_0) \to 0$ as $(N,T) \to \infty$.

Similarly, when $R > R_0$, we can follow the proof of Theorem 3.1 and apply the above results in (i)-(ii) and (iv)-(v) and analogous results to those in Theorems B.4-B.5 to show that

$$\widehat{CV}_1(R) - \widehat{CV}_1(R_0) \geq \frac{(1-p)}{256}\sum_{r=R_0+1}^{R}\ddot{\sigma}_r^2 + O_P\left(\delta_{NT}^{-3}\right) \text{ and } \widehat{CV}_3(R) - \widehat{CV}_3(R_0) = o_P\left(\delta_{NT}^{-2}\right).$$

Then $\mathrm{plim}_{(N,T)\to\infty}\delta_{NT}^2[\widehat{CV}(R) - \widehat{CV}(R_0)] \geq \frac{(R-R_0)(1-p)}{256}c_\sigma > 0$ and $P(\hat{R} > R_0) \to 0$ as $(N,T) \to \infty$. This completes the proof of the theorem. ■

REFERENCES

Ahn, S., Horenstein, A., 2013. Eigenvalue ratio test for the number of factors. Econometrica 81, 1203-1227.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., Khosravi, K., 2018. Matrix completion methods for causal panel data models. Working Paper, Graduate School of Business, Stanford University.

Bai, J., 2003. Inferential theory for factors models of large dimensions. Econometrica 71, 135-173.

Bai, J. and Li, K., 2016. Maximum likelihood estimation and inference for approximate factor models of high dimension. Review of Economics and Statistics, 98, pp.298-309.

Bai, J., Liao, Y., Yang, J., 2015. Unbalanced panel data models with interactive effects. In B.H. Baltagi (eds), The Oxford Handbook of Panel Data, pp., 149-170.

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

Bai, J., Ng, S., 2019a. Rank regularized estimation of approximate factor models. Journal of Econometrics 212, 78-96.

Bai, J., Ng, S., 2019b. Matrix completion, counterfactuals, and factor analysis of missing data. Working paper, Department of Economics, Columbia University.

Bańbura, M., Modugno, M., 2014. Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. Journal of Applied Econometrics 29, 133-160.

Cai, J-F., Candès, E.J., Shen, Z., 2010. A singular value thresholding algorithm for matrix completion. SIAM Journal on Optimization 20, 1956-1982.

Candès, E.J., Li, X., 2011. Robust principal component analysis. Journal of the ACM 58(3), 11:1-37.

Candès, E.J., Plan, Y., 2010. Matrix completion with noise. Proceedings of the IEEE 98(6), 925-936.

Chamberlain, G., Rothschild, M., 1983. Arbitrage, factor structure, and mean-variance analysis on large asset markets. Econometrica 51(5), 1281-1304.

Doz, C., Giannone, D., Reichlin, L., 2011. A two-step estimator for large approximate dynamic factor models based on Kalman filtering. Journal of Econometrics 164, 188-205.

Fan, J., Liao, Y., Mincheva, M., 2013. Large covariance estimation by thresholding principal orthogonal complements. Journal of the Royal Statistical Society: Series B 75, 603-680.

Foroni, C., Marcellino, M.G., 2013. A Survey of Econometric Methods for Mixed-Frequency Data. SSRN Electronic Journal.

Foroni, M., Hallin, M., Lippi, M., Reichlin, L., 2000. The generalized dynamic factor model: identification and estimation. Review of Economics and Statistics 82, 540-554.

Geweke, J.F., 1977. The dynamic factor analysis of economic time series models. In D. Aigner and A. Goldberger (eds.), Latent Variables in Socioeconomic Models, pp. 365-383, North-Holland, Amsterdam.

Giannone, D., Reichlin, L., Small, D., 2008. Nowcasting: The real-time informational content of macroeconomic data. Journal of Monetary Economics 55, 665-676.

Hallin, M., Liśka, R., 2007. Determining the number of factors in the general dynamic factor model. Journal of the American Statistical Association 102, 603-617.

Häusler, E., Luschgy, H., 2015. Stable Convergence and Stable Limit Theorems. Springer, New York.

Jungbacker, B., Koopman, S., Wel, M.V.D., 2011. Maximum likelihood estimation for dynamic factor models with missing data. Journal of Economic Dynamics and Control 35, 1358-1368.

Kapetanios, G., 2010. A testing procedure for determining the number of factors in approximate factor models with large datasets. Journal of Business & Economic Statistics 28, 397-409.

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

Ludvigson, S., Ng, S., 2007. The empirical risk-return relation: a factor analysis approach. Journal of Financial Economics 83,171-222.

Ludvigson, S., Ng, S., 2009. A factor analysis of bond risk premia. Review of Financial Studies 22, 5027-5067.

Marcellino, M., Sivec, V., 2016. Monetary, fiscal, and oil shocks: evidence based on mixed frequency structural FAVARs. Journal of Econometrics 193, 335-348.

Mariano, R.S., Murasawa, Y., 2010. A coincident index, common Factors, and monthly real GDP. Oxford Bulletin of Economics and Statistics 72, 27-46.

McCracken, M. W., Ng, S., 2016. FRED-MD: A monthly database for macroeconomic research. Journal of Business & Economic Statistics 34(4), 574-589.

Meinshausen, N., Bühlmann, 2010. Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72, 417-473.

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

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

Onatski, A., 2009. Testing hypotheses about the number of factors in large factor models. Econometrica 77, 1447-1479.

Onatski, A., 2010. Determining the number of factors from empirical distribution of eigenvalues. Review of Economics and Statistics 92, 1004-1016.

Onatski, A., 2012. Asymptotics of the principal components estimator of large factor models with weakly influential factors. Journal of Econometrics 168, 244-258.

Pinheiro, M., Rua, A., Dias, F., 2013. Dynamic factor models with jagged edge panel data: Taking on board the dynamics of the idiosyncratic components. Oxford Bulletin of Economics and Statistics 75, 80-102.

Sargent, T.J., Sims, C., 1977. Business cycle modelling without pretending to have too much a-priori economic theory. In C. Sims (ed.) New Methods in Business Cycle Research, pp. 45-109. Federal Reserve Bank of Minneapolis, Minneapolis.

Schumacher, C., Breitung, J., 2008. Real-time forecasting of German GDP based on a large factor model with monthly and quarterly data. International Journal of Forecasting 24, 386-398.

Stock, J.H., Watson, M.W., 1998. Diffusion indexes. Working paper 6702, National Bureau of Economic Research.

Stock, J.H., Watson, M.W., 2002. Macroeconomic forecasting using diffusion indexes. Journal of Business & Economic Statistics 20, 147-162.

Stock, J., Watson, M., 2016. Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. Handbook of Macroeconomics 415–525.

Su, L.,Chen, Q., 2013. Testing homogeneity in panel data models with interactive fixed effects. Econometric Theory 29, 1079-1135.

Su, L., Wang, X., 2017. On time-varying factor models: estimation and testing. Journal of Econometrics 198, 84-101.

Su, L., Jin, S., Zhang, Y., 2015. Specification test for panel data models with interactive fixed effects. Journal of Econometrics 186, 222-244.

Zeng, X., Xia, Y. and Zhang, L., 2019. Double Cross Validation for the Number of Factors in Approximate Factor Models. arXiv preprint arXiv:1907.01670.