Efficient Estimation of Approximate Factor Models

Bai, Jushan and Liao, Yuan

Columbia University, University of Maryland - College Park

September 2012

Efficient Estimation of Approximate Factor Models

Jushan Bai

Yuan Liao

Columbia University

University of Maryland

Department of Economics

Department of Mathematics

September 26, 2012

Abstract

We study the estimation of a high dimensional approximate factor model in the presence of both cross sectional dependence and heteroskedasticity. The classical method of principal components analysis (PCA) does not efficiently estimate the factor loadings or common factors because it essentially treats the idiosyncratic error to be homoskedastic and cross sectionally uncorrelated. For the efficient estimation, it is essential to estimate a large error covariance matrix. We assume the model to be conditionally sparse, and propose two approaches to estimating the common factors and factor loadings; both are based on maximizing a Gaussian quasi-likelihood and involve regularizing a large covariance sparse matrix. In the first approach the factor loadings and the error covariance are estimated separately while in the second approach they are estimated jointly. Extensive asymptotic analysis has been carried out. In particular, we develop the inferential theory for the two-step estimation. Because the proposed approaches take into account the large error covariance matrix, they produce more efficient estimators than the classical PCA methods or methods based on a strict factor model.

Keywords: High dimensionality, unknown factors, principal components, sparse matrix, conditional sparse, thresholding, cross-sectional correlation, penalized maximum likelihood, adaptive lasso, heteroskedasticity

1 Introduction

In many applications of economics, finance, and other scientific fields, researchers often face a large panel data set in which there are multiple observations for each individual; here individuals can be families, firms, countries, etc. Modern applications usually involve data-rich environments in which both the number of observations for each individual and the number of individuals are large. One useful method for summarizing information in a large dataset is the factor model:

$$y_{it} = \alpha_i + \lambda'_{0i} f_t + u_{it}, \quad i \leq N, t \leq T, \tag{1.1}$$

where α_i is an individual effect, λ_{0i} is an $r \times 1$ vector of factor loadings and f_t is an $r \times 1$ vector of common factors; u_{it} denotes the idiosyncratic component of the model. Note that y_{it} is the only observable random variable in this model. If we write $y_t = (y_{1t}, ..., y_{Nt})'$, $\Lambda_0 = (\lambda_{01}, ..., \lambda_{0N})'$, $\alpha = (\alpha_1, ..., \alpha_N)'$ and $u_t = (u_{1t}, ..., u_{Nt})'$, then model (1.1) can be equivalently written as

$$y_t = \alpha + \Lambda_0 f_t + u_t.$$

Because y_{it} is the only observable in the model, both factors and loadings are treated as parameters to estimate. As was shown by Chamberlain and Rothschild (1983), in many applications of factor analysis, it is desirable to allow dependence among the error terms $\{u_{it}\}_{i \leq N, t \leq T}$ not only serially but also cross-sectionally. This gives rise to the *approximate factor model*, in which the $N \times N$ covariance matrix $\Sigma_{u0} = \text{cov}(u_t)$ is not diagonal. In addition, the diagonal entries may vary in a large range. As a result, efficiently estimating the factor model under both large N and large T is difficult because to take into account both cross-sectional heteroskedasticity and dependence of $\{u_{it}\}_{i \leq N, t \leq T}$, it is essential to estimate the large covariance Σ_{u0}. The latter has been known as a challenging problem when N is larger than T.

In this paper, we assume the model to be *conditionally sparse*, in the sense that Σ_{u0} is a sparse matrix with bounded eigenvalues. This assumption effectively reduces the number of parameters to be estimated in the model, and allows a consistent estimation of Σ_{u0}. The latter is needed to efficiently estimate the factor loadings. In addition, it enables the model to identify the common components $\alpha_i + \lambda'_{0i} f_t$ asymptotically as $N \rightarrow \infty$. We propose two alternative methods, both are likelihood-based. The first one is a two-step procedure. In step one, we apply the *principal orthogonal complement thresholding* (POET) estimator of Fan et al. (2012) to estimate Σ_{u0} using the adaptive thresholding as in Cai and Liu (2011); in step two, we estimate the factor loadings by maximizing a Gaussian-quasi likelihood func-

tion, which depends on the covariance estimator in the first step. These two steps can be carried out iteratively. We also propose an alternative method for jointly estimating the factor loadings and the error covariance matrix by maximizing a weighted l_1 penalized likelihood function. The likelihood penalizes the estimation of the off-diagonal entries of the error covariance and automatically produces a sparse covariance estimator. We present asymptotic analysis for both methods. In particular, we derive the uniform rate of convergence and limiting distribution of the estimators for the two-step procedure. The analysis of the joint-estimation is more difficult as it involves penalizing a large covariance with diverging eigenvalues. We establish the consistency for this method.

Moreover, we achieve the "sparsistency" for the estimated error covariance matrix in factor analysis (see Section 3 for detailed explanations). The estimated covariance is consistent for both approaches under the normalized Frobenius norm even when N is much larger than T. This is important in the applications of approximate factor models.

There has been a large literature on estimating the approximate factor model. Stock and Watson (1998, 2002) and Bai (2003) considered the principal components analysis (PCA), and they developed large-sample inferential theory. However, the PCA essentially treats u_{it} to have the same variance across i, hence is inefficient when cross-sectional heteroskedasticity is present. Choi (2012) proposed a generalized PCA that requires $N < T$ to invert the error sample covariance matrix. More recently, Bai and Li (2012) estimated the factor loadings by maximizing the Gaussian-quasi likelihood, which addresses the heteroskedasticity under large N, but they consider the strict factor model in which $(u_{1t}, ..., u_{Nt})$ are uncorrelated. Additional literature on factor analysis includes, e.g., Bai and Ng (2002), Wang (2009), Dias, Pinherio and Rua (2008), Breitung and Tenhofen (2011), Han (2012), etc; most of these studies are based on the PCA method. In contrast, our methods are maximum-likelihood-based. Maximum likelihood methods have been one of the fundamental tools for statistical estimation and inference.

Our approach is closely related to the large covariance estimation literature, which has been rapidly growing in recent years. There are in general two ways to estimate a sparse covariance in the literature: thresholding and penalized maximum likelihood. For our two-step procedure, we apply the POET estimator recently proposed by Fan et al. (2012), corresponding to the thresholding approach of Bickel and Levina (2008a), Rothman et al. (2009) and Cai and Liu (2011). For the joint estimation procedure, we use the penalized likelihood, corresponding to that of Lam and Fan (2009), Bien and Tibshirani (2011), etc. In either way, we need to show that the impact of estimating the large covariances is asymptotically negligible for an efficient estimation, which is not easy in our context since the likelihood function is highly nonlinear, and $\Lambda_0\Lambda_0'$ contains a few eigenvalues that grow very fast. It was

recently shown by Fan et al. (2012) that estimating a covariance matrix with fast diverging eigenvalues is a challenging problem. Other works on large covariance estimation include Cai and Zhou (2012), Fan et al. (2008), Jung and Marron (2009), Witten, Tibshirani and Hastie (2009), Deng and Tsui (2010), Yuan (2010), Ledoit and Wolf (2012), El Karoui (2008), Pati et al. (2012), Rohde and Tsybakov (2011), Zhou et al. (2011), Ravikumar et al. (2011) etc.

This paper focuses on high-dimensional *static factor models* although the factors and errors can be serially correlated. The model considered is different from the generalized *dynamic factor models* as in Forni, Hallin, Lippi and Reichlin (2000), Forni and Lippi (2001), Hallin and Liška (2007), and other references therein. Both static and dynamic factor models are receiving increasing attention in applications of many fields.

The paper is organized as follows. Section 2 introduces the conditional sparsity assumption and the likelihood function. Section 3 proposes the two-step estimation procedure. In particular, we present asymptotic inferential theory of the estimators. Both uniform rate of convergence and limiting distributions are derived. Section 4 gives the joint estimation as an alternative procedure, where we demonstrate the estimation consistency. Section 5 illustrates some numerical examples which compare the proposed methods with the existing ones in the literature. Finally, Section 6 concludes with further discussions. All proofs are given in the appendix.

Notation

Let $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote the maximum and minimum eigenvalues of a matrix A respectively. Also Let $\|A\|_1$, $\|A\|$ and $\|A\|_F$ denote the l_1, spectral and Frobenius norms of A, respectively. They are defined as $\|A\|_1 = \max_i \sum_j |A_{ij}|$, $\|A\| = \sqrt{\lambda_{\max}(A'A)}$, $\|A\|_F = \sqrt{\text{tr}(A'A)}$. Note that $\|A\|$ is also the Euclidean norm when A is a vector. For two sequences a_T and b_T, we write $a_T \ll b_T$, and equivalently $b_T \gg a_T$, if $a_T = o(b_T)$ as $T \to \infty$.

2 Approximate Factor Models

2.1 The model

The approximate factor model (1.1) implies the following covariance decomposition:

$$\Sigma_{y0} = \Lambda_0 \, \text{cov}(f_t) \, \Lambda_0' + \Sigma_{u0}, \tag{2.1}$$

assuming f_t to be uncorrelated with u_t, where Σ_{y0} and Σ_{u0} denote the $N \times N$ covariance matrices of y_t and u_t; $\text{cov}(f_t)$ denotes the $r \times r$ covariance of f_t, all assumed to be time-invariant. The approximate factor model typically requires the idiosyncratic covariance Σ_{u0} have bounded eigenvalues and $\Lambda_0' \Lambda_0$ have eigenvalues diverging at rate $O(N)$. One of the

4

key concepts of approximate factor models is that it allows Σ_{u0} to be non-diagonal.

Stock and Watson (1998) and Bai (2003) derived the rates of convergence as well as the inferential theory of the method of principal component analysis (PCA) for estimating the factors and loadings. Let $Y = (y_1, ..., y_T)'$ be the $T \times N$ data matrix. Then PCA estimates the $T \times r$ factor matrix F by maximizing $\mathrm{tr}(F'(YY')F)$ subject to normalization restrictions for F. The PCA method essentially restricts to have cross-sectional homoskedasticity and independence. Thus it is known to be inefficient when the idiosyncratic errors are either cross sectionally heteroskedastic or correlated.

This paper aims at the efficient estimation of the approximate factor model, and assumes the number of factors r to be known. In practice, r can be estimated from the data, and there has been a large literature addressing its consistent estimation, e.g., Bai and Ng (2002), Kapetanios (2010), Onatski (2010), Alessi et al. (2010), Hallin and Liška (2007), Lam and Yao (2012), among others.

2.2 Conditional sparsity

An efficient estimation of the factor loadings and factors should take into account both cross-sectional dependence and heteroskedasticity, which will then involve estimating $\Sigma_{u0} = \mathrm{cov}(u_t)$, or more precisely, the precision matrix Σ_{u0}^{-1}. In a data-rich environment, N can be either comparable with or much larger than T. Then estimating Σ_{u0} is a challenging problem even when the idiosyncratics $\{u_{it}\}_{i \leq N, t \leq T}$ are observable, because the sample covariance is nonsingular when $N > T$, whose spectrum is inconsistent (Johnstone and Ma 2009).

Under the regular approximate factor model considered by Chamberlain and Rothschild (1983) and Stock and Watson (2002), it is difficult to estimate Σ_{u0} without further structural assumptions. A natural assumption to go one-step further is that of sparsity, which assumes that many off-diagonal elements of Σ_{u0} be either zero or vanishing as the dimensionality increases. In an approximate factor model, it is more appropriate to assume Σ_{u0} be a sparse matrix instead of Σ_{y0}. Due to the presence of common factors, we call such a special structure of the factor model to be *conditionally sparse.*

Therefore, the model studied in the current paper is the approximate factor model with conditional sparsity (sparsity structure on Σ_{u0}), which is sightly more restrictive than that of Chamberlain and Rothschild (1983). The conditional sparsity is required to regularize a large idiosyncratic covariance, which allows us to take both cross sectional correlation and heteroskedasticity into account, and is needed for an efficient estimation. However, such an assumption is still quite general and covers most of the applications of factor models in economics, finance, genomics, and many important applied areas.

2.3 Maximum likelihood

Compared to PCA, a more efficient estimation for model (2.1) of high dimension is based on a Gaussian quasi-likelihood approach. Let $\bar{f} = T^{-1}\sum_{t=1}^{T} f_t$. Because of the existence of α, the model $y_t = \Lambda_0 f_t + \alpha + u_t$ is observationally equivalent to $y_t = \Lambda_0 f_t^* + \alpha^* + u_t$, where $f_t^* = f_t - \bar{f}$ and $\alpha^* = \alpha + \Lambda_0 \bar{f}$. Therefore without loss of generality, we assume $\bar{f} = 0$. The Guassian quasi-likelihood for Σ_y is given by

$$-N^{-1}\log|\det(\Sigma_y)| - N^{-1}\mathrm{tr}(S_y\Sigma_y^{-1})$$

where $S_y = T^{-1}\sum_{t=1}^{T}(y_t - \bar{y})(y_t - \bar{y})'$ is the sample covariance matrix, with $\bar{y} = T^{-1}\sum_{t=1}^{T} y_t$. Plugging in (2.1), using the notation $S_f = \frac{1}{T}\sum_{t=1}^{T} f_t f_t'$, we obtain the quasi-likelihood function for the factors and loadings:

$$-\frac{1}{N}\log|\det(\Lambda S_f \Lambda' + \Sigma_u)| - \frac{1}{N}\mathrm{tr}\left(S_y(\Lambda S_f \Lambda' + \Sigma_u)^{-1}\right), \qquad (2.2)$$

where $\Lambda = (\lambda_1, ..., \lambda_N)'$ is an $N \times r$ matrix of factor loadings.

It has been well known that the factors and loadings are not separably identified without further restrictions. Note that the factors and loadings enter the likelihood through $\Lambda S_f \Lambda'$. Hence for any invertible $r \times r$ matrix $\bar{H}$, if we define $\Lambda^* = \Lambda\bar{H}^{-1}$, $f_t^* = \bar{H}f_t$ and $S_{f^*} = \frac{1}{T}\sum_{t=1}^{T} f_t^* f_t^{*'}$, then $\Lambda^* S_{f^*}\Lambda^{*'} = \Lambda S_f \Lambda'$, and they produce observationally equivalent models. In this paper, we focus on a usual restriction for MLE of factor analysis (see e.g., Lawley and Maxwell 1971) as follows:

$$S_f = I_r, \text{ and } \Lambda'\Sigma_u^{-1}\Lambda \text{ is diagonal,} \qquad (2.3)$$

and the diagonal entries of $\Lambda'\Sigma_u^{-1}\Lambda$ are distinct and are arranged in a decreasing order. Restriction (2.3) guarantees a unique solution to the maximization of the log-likelihood function up to a column sign change for Λ. Therefore we assume the estimator $\widehat{\Lambda}$ and Λ_0 have the same column signs, as part of the identification conditions.

The negative log-likelihood function (2.2) simplifies to

$$-L(\Lambda, \Sigma_u) = \frac{1}{N}\log|\det(\Lambda\Lambda' + \Sigma_u)| + \frac{1}{N}\mathrm{tr}\left(S_y(\Lambda\Lambda' + \Sigma_u)^{-1}\right). \qquad (2.4)$$

In the presence of cross sectional dependence, Σ_{u0} is not necessarily diagonal. Therefore there can be up to $O(N^2)$ free parameters in the likelihood function (2.4). There are in general two main regularization approaches to estimating a large sparse covariance: (adaptive) thresholding (Bickel and Levina 2008a, Rothman et al. 2009, Cai and Liu 2011, etc.)

and penalized maximum likelihood (Lam and Fan 2009, Bien and Tibshirani 2011). Correspondingly in this paper, we propose two methods for regularizing the likelihood function to efficiently estimate the factor loadings as well as the unknown factors. One estimates Σ_{u0} and Λ_0 in two steps and the other estimates them jointly.

3 Two-Step Estimation

The two-step estimation estimates (Λ_0, Σ_{u0}) separately. In the first step, we estimate Σ_{u0} by the principal orthogonal complement thresholding (POET), proposed by Fan et al. (2012), and in the second step we estimate Λ_0 only, using the quasi-maximum likelihood, replacing Σ_u by the covariance estimator obtained in step one.

3.1 Step one: covariance estimation by thresholding

The POET is based on a spectrum expansion of the sample covariance matrix and adaptive thresholding. Let $(\nu_j, \xi_j)_{j=1}^N$ be the eigenvalues-vectors of the sample covariance S_y of y_t, in a decreasing order such that $\nu_1 \geq \nu_2 \geq ... \geq \nu_N$. Then S_y has the following spectrum decomposition:

$$S_y = \sum_{i=1}^r \nu_i \xi_i \xi_i' + R$$

where $R = \sum_{i=r+1}^N \nu_i \xi_i \xi_i'$ is the *orthogonal complement component*. Define a general thresholding function $s_{ij}(z) : \mathbb{R} \to \mathbb{R}$ as in Rothman et al. (2009) and Cai and Liu (2011) with an entry-dependent threshold τ_{ij} such that:

(i) $s_{ij}(z) = 0$ if $|z| < \tau_{ij}$;

(ii) $|s_{ij}(z) - z| \leq \tau_{ij}$.

(iii) There are constants $a > 0$ and $b > 1$ such that $|s_{ij}(z) - z| \leq a\tau_{ij}^2$ if $|z| > b\tau_{ij}$.

Examples of $s_{ij}(z)$ include the hard-thresholding: $s_{ij}(z) = zI_{(|z|>\tau_{ij})}$; SCAD (Fan and Li 2001), MPC (Zhang 2010) etc. Then we obtain the step-one consistent estimator for Σ_{u0}:

$$\widehat{\Sigma}_u^{(1)} = (s_{ij}(R_{ij}))_{N \times N}, \text{ where } R = (R_{ij})_{N \times N}.$$

We can choose the threshold as $\tau_{ij} = C\sqrt{R_{ii}R_{jj}}(\sqrt{(\log N)/T} + 1/\sqrt{N})$ for some universal constant $C > 0$, which corresponds to applying the threshold $C(\sqrt{(\log N)/T} + 1/\sqrt{N})$ to the correlation matrix of R [defined to be $\text{diag}(R)^{-1/2}R\,\text{diag}(R)^{-1/2}$]. The POET estimator also has an equivalent expression using PCA. Let $\{\widehat{u}_{it}^{\text{PCA}}\}_{i \leq N, t \leq T}$ denote the PCA estimators of $\{u_{it}\}_{i \leq N, t \leq T}$ (Bai 2003). Then $\widehat{\Sigma}_{u,ij}^{(1)} = s_{ij}(T^{-1}\sum_{t=1}^T \widehat{u}_{it}^{\text{PCA}}\widehat{u}_{jt}^{\text{PCA}})$.

It was shown by Fan et al. (2012) that under some regularity conditions $\|\widehat{\Sigma}_u^{(1)} - \Sigma_{u0}\| = O_p(N^{-1/2} + T^{-1/2}(\log N)^{1/2})$, which guarantees the positive definiteness asymptotically, given that $\lambda_{\min}(\Sigma_{u0}) > 0$ is bounded away from zero.

3.2   Step two: estimating factor loadings and factors

Replacing Σ_u in (2.4) by $\widehat{\Sigma}_u^{(1)}$, we obtain the objective function for Λ. Under the identification condition (2.3), in this step, we estimate the loadings as:

$$
\begin{aligned}
\widehat{\Lambda}^{(1)} &= \arg\min_{\Lambda \in \Theta_\lambda} L_1(\Lambda) \\
&= \arg\min_{\Lambda \in \Theta_\lambda} \frac{1}{N} \log|\det(\Lambda\Lambda' + \widehat{\Sigma}_u^{(1)})| + \frac{1}{N}\mathrm{tr}(S_y(\Lambda\Lambda' + \widehat{\Sigma}_u^{(1)})^{-1})
\end{aligned} \tag{3.1}
$$

where Θ_λ is a parameter space for the loading matrix, to be defined later. Suppose that $y_t \sim N(0, \Lambda_0\Lambda_0' + \Sigma_{u0})$, the negative log-likelihood is then the same (up to a constant) as (3.1) except that $\widehat{\Sigma}_u^{(1)}$ should be replaced by Σ_{u0}. Consequently, (3.1) can be treated as a Gaussian quasi-likelihood of Λ, which will give an efficient estimation of Λ_0 since it takes into account the cross sectional heteroskedasticity and dependence in Σ_{u0} through its consistent estimator.

After obtaining $\widehat{\Lambda}^{(1)}$, we estimate f_t via the generalized least squares (GLS) as suggested by Bai and Li (2012):

$$
\widehat{f}_t^{(1)} = (\widehat{\Lambda}^{(1)\prime}(\widehat{\Sigma}_u^{(1)})^{-1}\widehat{\Lambda}^{(1)})^{-1}\widehat{\Lambda}^{(1)\prime}(\widehat{\Sigma}_u^{(1)})^{-1}(y_t - \bar{y}).
$$

The proposed two-step procedure can be carried out iteratively.   After obtaining $(\widehat{\Lambda}^{(1)}, \widehat{f}_t^{(1)})$, we update

$$
\widehat{u}_t = y_t - \widehat{\Lambda}^{(1)}\widehat{f}_t^{(1)}, \quad \widehat{\Sigma}_u^{(1)} = (s_{ij}(T^{-1}\sum_{t=1}^{T}\widehat{u}_{it}\widehat{u}_{jt}))_{N \times N}.
$$

Then $\widehat{\Sigma}_u^{(1)}$ in the objective function (3.1) is updated, which gives updated $\widehat{\Lambda}^{(1)}$ and $\widehat{f}_t^{(1)}$ respectively. This procedure can be continued until convergence.

3.3   Positive definiteness

The objective function (3.1) requires $\Lambda\Lambda' + \widehat{\Sigma}_u^{(1)}$ be positive definite for any given finite sample. A sufficient condition is the finite-sample positive definiteness of $\widehat{\Sigma}_u^{(1)}$, which also

depends on the choice of the adaptive threshold value τ_{ij}. We specify

$$\tau_{ij} = C\alpha_{ij}\left(\sqrt{\frac{\log N}{T}} + \frac{1}{\sqrt{N}}\right)$$

where α_{ij} is an entry-dependent value that captures the variability of individual variables such as $\sqrt{R_{ii}R_{jj}}$; $C > 0$ is a pre-determined universal constant. More concretely, the finite sample positive definiteness depends on the choice of C. If we write $\widehat{\Sigma}_u^{(1)} = \widehat{\Sigma}_u^{(1)}(C)$ in step one to indicate its dependence on the threshold, then C should be chosen in the interval $(C_{\min}, C_{\max}]$, where

$$C_{\min} = \inf\{M : \lambda_{\min}(\widehat{\Sigma}_u^{(1)}(C)) > 0, \forall C > M\},$$

and $C_{\max}$ is a large constant that thresholds all the off-diagonal elements of $\widehat{\Sigma}_u^{(1)}$ to zero. Then by construction, $\widehat{\Sigma}_u^{(1)}(C)$ is finite-sample positive definite for any $C > C_{\min}$ (see Figure 1).

Figure 1: Minimum eigenvalue of $\lambda_{\min}(\widehat{\Sigma}_u^{(1)}(C))$
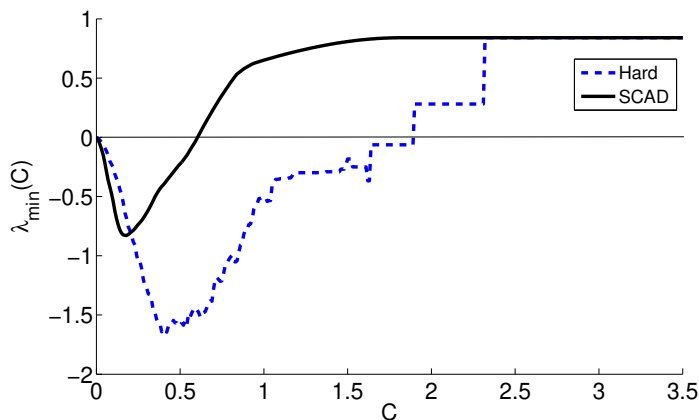
Data are simulated from the setting of Section 5 with $T = 100, N = 150$. Both hard and SCAD with adaptive thresholds (Cai and Liu 2011) are plotted.

3.4 Asymptotic analysis

We now present the asymptotic analysis of the proposed two-step estimator. We first list a set of regularity conditions and then present the consistency. A more refined set of assumptions are needed to achieve the optimal rate of convergence as well as the limiting distributions.

3.4.1 Consistency

Assumption 3.1. *Let $\Sigma_{u0,ij}$ denote the (i,j)th entry of Σ_{u0}. There is $q \in [0,1)$ such that*

$$m_N \equiv \max_{i \leq N} \sum_{j=1}^{N} |\Sigma_{u0,ij}|^q = o(\min(\sqrt{N}, \sqrt{T/\log N})).$$

In particular, when $q = 0$, we define $m_N = \max_{i \leq N} \sum_{j=1}^{N} I_{(\Sigma_{u0,ij} \neq 0)}$, which corresponds to the "exactly sparse" case.

The first assumption sets a condition on the sparsity of Σ_{u0}, under which Fan et al. (2012) showed that the POET estimator $\widehat{\Sigma}_u^{(1)}$ is consistent under the operator norm. The sparsity is in terms of the maximum row sum, considered by Bickel and Levina (2008a).

The following assumption provides the regularity conditions on the data generating process. We introduce the strong mixing condition. Let $\mathcal{F}_{-\infty}^0$ and $\mathcal{F}_T^\infty$ denote the σ-algebras generated by $\{(f_t, u_t) : -\infty \leq t \leq 0\}$ and $\{(f_t, u_t) : T \leq t \leq \infty\}$ respectively. In addition, define the mixing coefficient

$$\alpha(T) = \sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_T^\infty} |P(A)P(B) - P(AB)|. \tag{3.2}$$

Assumption 3.2. *(i) $\{u_t, f_t\}_{t \geq 1}$ is strictly stationary. In addition, $Eu_{it} = Eu_{it}f_{jt} = 0$ for all $i \leq p, j \leq r$ and $t \leq T$.*
(ii) There exist constants $c_1, c_2 > 0$ such that $c_2 < \lambda_{\min}(\Sigma_{u0}) \leq \lambda_{\max}(\Sigma_{u0}) < c_1$, and $\max_{j \leq N} \|\lambda_{0j}\| < c_1$.
(iii) Exponential tail: There exist $r_1, r_2 > 0$ and $b_1, b_2 > 0$, such that for any $s > 0$, $i \leq p$ and $j \leq r$,

$$P(|u_{it}| > s) \leq \exp(-(s/b_1)^{r_1}), \quad P(|f_{jt}| > s) \leq \exp(-(s/b_2)^{r_2}).$$

(iv) Strong mixing: There exists $r_3 > 0$ such that $3r_1^{-1} + 1.5r_2^{-1} + r_3^{-1} > 1$, and $C > 0$ satisfying: for all $T \in \mathbb{Z}^+$,
$$\alpha(T) \leq \exp(-CT^{r_3}).$$

The following assumptions are standard in the approximate factor models, see e.g., Stock and Watson (1998, 2002) and Bai (2003). In particular, Assumption 3.3 implies that the first r eigenvalues of $\Lambda_0 \Lambda_0'$ are growing rapidly at $O(N)$. Intuitively, it requires the factors be pervasive in the sense that they impact a non-vanishing proportion of time series $\{y_{1t}\}_{t \leq T}, ..., \{y_{Nt}\}_{t \leq T}$.

Assumption 3.3. *There is a $\delta > 0$ such that for all large N,*

$$\delta^{-1} < \lambda_{\min}(N^{-1}\Lambda_0'\Lambda_0) \leq \lambda_{\max}(N^{-1}\Lambda_0'\Lambda_0) < \delta.$$

Therefore all the eigenvalues of $N^{-1}\Lambda_0'\Lambda_0$ are bounded away from both zero and infinity as $N \to \infty$.

Assumption 3.4. *There exists $M > 0$ such that for all $t \leq T$ and $s \leq T$,*
(i) $E[N^{-1/2}(u_s'u_t - Eu_s'u_t)]^4 < M$,
(ii) $E\|N^{-1/2}\sum_{j=1}^N \lambda_{0j}u_{jt}\|^4 < M$.

The following assumption defines the threshold τ_{ij} on the (i,j)th entry of R_{ij} for the step-one POET estimator.

Assumption 3.5. *The threshold $\tau_{ij} = C\alpha_{ij}(\sqrt{(\log N)/T} + 1/\sqrt{T})$ where $\alpha_{ij} > 0$ is entry-dependent, either stochastic or deterministic, such that $\forall \epsilon > 0$, there are positive C_1 and C_2 so that*

$$P(C_1 < \min_{i,j \leq N} \alpha_{ij} \leq \max_{i,j \leq N} \alpha_{i,j} < C_2) > 1 - \epsilon \tag{3.3}$$

for all large N and T. Here $C > 0$ is a deterministic constant.

Condition (3.3) requires the rate $\tau_{ij} \asymp (\sqrt{(\log N)/T} + 1/\sqrt{T})$ uniformly in (i,j). This condition is satisfied by the universal threshold $\alpha_{ij} = \alpha$ for all (i,j), the correlation threshold $\alpha_{ij} = \sqrt{R_{ii}R_{jj}}$ as discussed before, and the adaptive threshold in Cai and Liu (2011).

For identification, we require the objective function be minimized subject to the diagonality of $\Lambda'(\widehat{\Sigma}_u^{(1)})^{-1}\Lambda$. In addition, since Assumption 3.3 is essential in asymptotically identifying the covariance decomposition $\Sigma_{y0} = \Lambda_0\Lambda_0' + \Sigma_{u0}$, we need to take it into account when minimizing the objective function. Therefore we assume δ in Assumption 3.3 is sufficiently large, which leads to the following parameter space:

$$\Theta_\lambda = \{\Lambda : \quad \delta^{-1} < \lambda_{\min}(N^{-1}\Lambda'\Lambda) \leq \lambda_{\max}(N^{-1}\Lambda'\Lambda) < \delta,$$
$$\Lambda'(\widehat{\Sigma}_u^{(1)})^{-1}\Lambda \text{ is diagonal.}\} \tag{3.4}$$

Write $\gamma^{-1} = 3r_1^{-1} + 1.5r_2^{-1} + r_3^{-1} + 1$ and $\widehat{\Lambda}^{(1)} = (\widehat{\lambda}_1^{(1)}, ..., \widehat{\lambda}_N^{(1)})'$. We have the following theorem.

Theorem 3.1. *Suppose $(\log N)^{6/\gamma} = o(T)$, $T = o(N^2)$. Under Assumptions 3.1-3.5,*

$$\frac{1}{\sqrt{N}}\|\widehat{\Lambda}^{(1)} - \Lambda_0\|_F = o_p(1), \quad \max_{j \leq N}\|\widehat{\lambda}_j^{(1)} - \lambda_{0j}\| = o_p(1).$$

By a more careful large-sample analysis, we can improve the above result and derive the rate of convergence. Throughout the paper, we will frequently use the notation:

$$\omega_T = \frac{1}{\sqrt{N}} + \sqrt{\frac{\log N}{T}}.$$

Theorem 3.2. *Under the Assumptions of Theorem 3.1,*

$$\frac{1}{\sqrt{N}}\|\widehat{\Lambda}^{(1)} - \Lambda_0\|_F = O_p(m_N\omega_T^{1-q}), \quad \max_{j \leq N}\|\widehat{\lambda}_j^{(1)} - \lambda_{0j}\| = O_p(m_N\omega_T^{1-q}),$$

where m_N and q are defined in Assumption 3.1.

Remark 3.1. In the above theorem m_N does not need to be bounded. But in order to achieve the $\sqrt{T}$-consistency for each $\widehat{\lambda}_j$, the uniform rate of convergence above would require it be bounded (which is a strong assumption on the sparsity of Σ_{u0}). Later in Section 3.4.3 we will enhance this convergence rate so that the boundedness of m_N is not necessary and $\sqrt{T}$-consistency can still be achieved. This will require additional regularity conditions.

3.4.2 Covariance estimation and sparsistency

In order to obtain the limiting distribution for each individual $\widehat{\lambda}_j^{(1)}$, we also need to achieve the *sparsistency* for estimating Σ_{u0}. By sparsistency, we mean the property that all small entries of Σ_{u0} are estimated as exactly zeros with a probability arbitrarily close to one. Besides being important for deriving the limiting distribution of $\widehat{\lambda}_j^{(1)}$, the sparsistency itself is of independent interest for large covariance estimation, and has been studied by many authors, for instance, Lam and Fan (2009) and Rothman et al. (2009). To our best knowledge, this is the first place where the sparsistency for an estimated idiosyncratic Σ_{u0} is achieved in a high dimensional approximate factor model.

Let S_L and S_U denote two disjoint sets and respectively include the indices of small and large elements of Σ_{u0} in absolute value, and

$$\{(i,j) : i \leq N, j \leq N\} = S_L \cup S_U.$$

Because the diagonal elements represent the individual variances of the idiosyncratic components, we assume $(i,i) \in S_U$ for all $i \leq N$. The sparsity assumes that most of the indices (i,j) belong to S_L when $i \neq j$. A special case arises when Σ_{u0} is strictly sparse, in the sense that its elements in small magnitudes (S_L) are exactly zero. For the banded matrix as an example,

$$\Sigma_{u0,ij} \neq 0 \text{ if } |i-j| \leq k; \quad \Sigma_{u0,ij} = 0 \text{ if } |i-j| > k$$

for some fixed k. Then $S_L = \{(i,j) : |i - j| > k\}$ and $S_U = \{(i,j) : |i - j| \leq k\}$.

The following assumption quantifies the "small" and "big" entries of Σ_{u0}. By "small" entries we mean those of smaller order than $\omega_T = N^{-1/2} + T^{-1/2}(\log N)^{1/2}$. The partition $\{(i,j) : i \leq N, j \leq N\} = S_L \cup S_U$ may not be unique. Our analysis suffices as long as such a partition exists.

Assumption 3.6. *There is a partition $\{(i,j) : i \leq N, j \leq N\} = S_L \cup S_U$ such that $(i,i) \in S_U$ for all $i \leq N$ and S_L is nonempty. In addition,*

$$\max_{(i,j) \in S_L} |\Sigma_{u0,ij}| \ll \omega_T \ll \min_{(i,j) \in S_U} |\Sigma_{u0,ij}|.$$

The conditional sparsity assumption requires most off-diagonal entries of Σ_{u0} be inside S_L, hence it is reasonable to have $S_L \neq \emptyset$ in the condition. It is likely that S_U only contains the diagonal elements. It then essentially corresponds to the strict factor model where Σ_{u0} is almost a diagonal matrix and error terms are only weakly cross-sectionally correlated. That is also a special case of Assumption 3.6.

Theorem 3.3. *Under Assumption 3.6 and those of Theorem 3.2, for any $\epsilon > 0$ and $M > 0$, there is an integer $N_0 > 0$ such that as long as T and $N > N_0$,*

$$P(\widehat{\Sigma}_{u,ij}^{(1)} = 0, \forall (i,j) \in S_L) > 1 - \epsilon,$$
$$P(|\widehat{\Sigma}_{u,ij}^{(1)}| > M\omega_T, \forall (i,j) \in S_U) > 1 - \epsilon.$$

It was shown by Fan et al. (2012) that $\|(\widehat{\Sigma}_u^{(1)})^{-1} - \Sigma_{u0}^{-1}\| = O_p(m_N \omega_T^{1-q})$. Theorem 3.4 below demonstrates a strengthened convergence rate for the averaged estimation error.

Assumption 3.7. *There is $c > 0$ such that $\|\Sigma_{u0}^{-1}\|_1 < c$.*

In addition to Assumptions 3.1 and 3.6, we require the following condition on the sparsity of Σ_{u0}, which further characterizes S_L and S_U:

Assumption 3.8. *The index sets S_L and S_U satisfy: $\sum_{i \neq j, (i,j) \in S_U} 1 = O(N)$ and $\sum_{(i,j) \in S_L} |\Sigma_{u0,ij}| = O(1)$.*

Assumption 3.8 requires that the number of off-diagonal large entries of Σ_{u0} be of order $O(N)$, and that the absolute sum of the small entries is bounded. This assumption is satisfied, for example, if $\{u_{it}\}_{i \leq N}$ follows an heteroskedastic MA(p) process with a fixed p, where $\sum_{i \neq j, (i,j) \in S_U} 1 = O(N)$ and $\sum_{(i,j) \in S_L} |\Sigma_{u0,ij}| = 0$. It is also satisfied by banded

13

matrices (Bickel and Levina 2008b, Cai and Yuan 2012) and block-diagonal matrices with fixed block size.

Define an $r \times N$ matrix $\Xi = \Lambda_0' \Sigma_{u0}^{-1} = (\xi_1, ..., \xi_N)$. Then $\|\Sigma_{u0}^{-1}\|_1 < c$ implies

$$\max_{j \leq N} \|\xi_j\| = \max_{j \leq N} \|\sum_{i=1}^{N} \lambda_{0i}(\Sigma_{u0}^{-1})_{ij}\| \leq \|\Sigma_{u0}^{-1}\|_1 \max_{j \leq N} \|\lambda_{0j}\| < \infty.$$

The following assumption corresponds to those of PCA in Bai (2003), and also extends to the non-diagonal Σ_{u0}.

Assumption 3.9. *(i)* $E\|\frac{1}{\sqrt{TN}}\sum_{s=1}^{T} f_s(u_s' u_t - E u_s' u_t)\|^2 = O(1)$
(ii) For each element $d_{i,kl}$ of $\xi_i \xi_i'$ ($k, l \leq r$),
$\frac{1}{N\sqrt{NT}}\sum_{j=1}^{N}\sum_{i=1}^{N}\sum_{t=1}^{T}(u_{it}u_{jt} - Eu_{it}u_{jt})\lambda_{0i}\lambda_{0j}'d_{i,kl} = O_p(1),$
$\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}(u_{it}^2 - Eu_{it}^2)\xi_i\xi_i' = O_p(1).$
(iii) For each element $d_{ij,kl}$ of $\xi_i \xi_j'$,
$\frac{1}{N\sqrt{NT}}\sum_{i\neq j,(i,j)\in S_U}\sum_{t=1}^{T}\sum_{v=1}^{N}(u_{it}u_{vt} - Eu_{it}u_{vt})\lambda_{0j}\lambda_{0v}'d_{ij,kl} = O_p(1),$
$\frac{1}{\sqrt{NT}}\sum_{i\neq j,(i,j)\in S_U}\sum_{t=1}^{T}(u_{it}u_{jt} - Eu_{it}u_{jt})\xi_i\xi_j' = O_p(1).$

Under Assumption 3.9, we can achieve the following improved rate of convergence for the averaged estimation error $\widehat{\Sigma}_u^{(1)} - \Sigma_{u0}$:

Theorem 3.4. *Under the assumptions of Theorem 3.3 and Assumption 3.9,*

$$\frac{1}{N}\|\Lambda_0'[(\widehat{\Sigma}_u^{(1)})^{-1} - \Sigma_{u0}^{-1}]\Lambda_0\|_F = O_p(m_N^2 \omega_T^{2-2q}).$$

Remark 3.2. 1. A simple application of
$\|(\widehat{\Sigma}_u^{(1)})^{-1} - \Sigma_{u0}^{-1}\| = O_p(m_N \omega_T^{1-q})$ by Fan et al. (2012) yields
$\frac{1}{N}\|\Lambda_0'[(\widehat{\Sigma}_u^{(1)})^{-1} - \Sigma_{u0}^{-1}]\Lambda_0\|_F = O_p(m_N \omega_T^{1-q})$. In contrast, the rate we present in Theorem 3.4 requires more refined asymptotic analysis. It shows that after weighted by the factor loadings, the averaged convergence rate is faster.

2. The condition on the large-entry-set S_U in Assumption 3.8 can be relaxed a bit to $\sum_{i\neq j,(i,j)\in S_U} 1 = O(N^{1+\epsilon})$ for an arbitrarily small $\epsilon > 0$, which will allow less sparse covariances. For example, Suppose $\{u_{it}\}_{i\leq N}$ follows a cross sectional AR(1) process such that

$$u_{it} = \rho u_{i-1,t} + e_{it}$$

for $|\rho| < 1$ and $\{e_{it}\}_{i\leq N, t\leq T}$ being independent across both i and t. We can then find a partition $S_L \cup S_U$ such that $\sum_{(i,j)\in S_L} |\Sigma_{u0,ij}| = O(1)$ and $\sum_{i\neq j,(i,j)\in S_U} 1 = O(N^{1+\epsilon})$ for any $\epsilon > 0$. Theorems 3.4 and 3.5 below still hold. But conditions in Assumption

3.9 need to be adjusted accordingly. For example, in condition (iii) the normalizing constant $\frac{1}{N\sqrt{N}}$ in the first equation should be changed to $\frac{1}{N^{\epsilon+3/2}}$, and $\frac{1}{\sqrt{N}}$ in the second equation should be changed to $\frac{1}{N^{(1+\epsilon)/2}}$. The current Assumption 3.9, on the other hand, keeps our presentation simple.

3.4.3 Limiting distribution

As a result of Theorem 3.4, the impact of estimating Σ_{u0} at step one is asymptotically negligible. This enables us to achieve the $\sqrt{T}$-consistency and the limiting distribution of $\widehat{\lambda}_j^{(1)}$ for each j. We impose further assumptions.

Assumption 3.10. (i) $\frac{1}{N\sqrt{NT}} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T} (u_{it}u_{jt} - Eu_{it}u_{jt})\xi_i\xi_j' = O_p(1)$.
For each $j \leq N, \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} (u_{it}u_{jt} - Eu_{it}u_{jt})\xi_i = O_p(1)$.
(iii) $\frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} \xi_i u_{it} f_t' = O_p(1)$.

Theorem 3.5. *Suppose $0 \leq q < 1/2$, and $T = o(N^{2-2q})$. In addition, $m_N^2\omega_T^{2-2q} = o(T^{-1/2})$. Then under the assumptions of Theorem 3.4 and Assumption 3.10, for each $j \leq N$,*

$$\sqrt{T}(\widehat{\lambda}_j^{(1)} - \lambda_{0j}) \to^d N_r\left(0, E(u_{jt}f_tf_t')\right).$$

We make some technical remarks regarding Theorem 3.5.

Remark 3.3. 1. The condition $m_N^2\omega_T^{2-2q} = o(T^{-1/2})$ (roughly speaking, this is $m_N = o(T^{1/4})$ when N is very large and $q = 0$) strengthens the sparsity condition of Assumption 3.1. The required upper bound for m_N is tight. Roughly speaking, the estimation error of $\widehat{\Sigma}_u^{(1)}$ plays a role in the asymptotic expansion of $\sqrt{T}(\widehat{\lambda}_j^{(1)} - \lambda_{0j})$ only through an averaged term as in Theorem 3.4. Condition $m_N^2\omega_T^{2-2q} = o(T^{-1/2})$ is required for that term to be asymptotically negligible.

2. The asymptotic normality also holds jointly for finitely many estimators. For any finite and fixed k, we have,

$$\sqrt{T}(\widehat{\lambda}_1^{(1)'} - \lambda_{01}', \cdots, \widehat{\lambda}_k^{(1)'} - \lambda_{0k}')' \to^d N_{kr}(0, E[\text{cov}(u_t^k|f_t) \otimes f_tf_t']).$$

where $\text{cov}(u_t^k|f_t) = \text{cov}(u_{1t}, ..., u_{kt}|f_t)$.

3. If Assumption 3.10(i) is replaced by a uniform convergence, by assuming $\max_{j \leq N} \|\frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} (u_{it}u_{jt} - Eu_{it}u_{jt})\xi_i\| = O_p(\sqrt{N \log N})$, we can then improve

15

the uniform rate of convergence in Theorem 3.2 and obtain

$$\max_{j \leq N} \|\widehat{\lambda}_j^{(1)} - \lambda_{0j}\| = O_p(\sqrt{\frac{\log N}{T}}).$$

3.4.4 Estimation of common factors

For the limiting distribution of $\widehat{f}_t^{(1)}$, we make the following additional assumption:

Assumption 3.11. *There is a positive definite matrix Q such that for each $t \leq T$,*

$$\frac{1}{N}\Lambda_0'\Sigma_{u0}^{-1}\Lambda_0 \to Q, \quad \frac{1}{\sqrt{N}}\Lambda_0'\Sigma_{u0}^{-1}u_t = \frac{1}{\sqrt{N}}\sum_{j=1}^{N}\xi_j u_{jt} \to N_r(0, Q).$$

For the next assumption, we define $\beta_t = \Sigma_{u0}^{-1}u_t$. Then β_t has mean zero and covariance matrix Σ_{u0}^{-1}.

Assumption 3.12. *For any fixed $t \leq T$,*
(i) $\frac{1}{\sqrt{NT}}\sum_{s=1}^{T}\sum_{i=1}^{N}f_s u_{is}\beta_{it} = O_p(1)$,
$\frac{1}{NT\sqrt{N}}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{s=1}^{T}\xi_i(u_{is}u_{js} - Eu_{is}u_{js})\beta_{jt} = o_p(1)$
$\frac{1}{T\sqrt{N}}\sum_{i=1}^{N}\sum_{s=1}^{T}(u_{is}^2 - Eu_{is}^2)\xi_i\beta_{it} = o_p(1)$
$\frac{1}{T\sqrt{N}}\sum_{i\neq j,(i,j)\in S_U}\sum_{s=1}^{T}(u_{is}u_{js} - Eu_{is}u_{js})\xi_i\beta_{jt} = o_p(1)$.
(ii) For each $k \leq r$,
$\frac{1}{NT\sqrt{N}}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{s=1}^{T}(u_{is}u_{js} - Eu_{is}u_{js})\lambda_{0i}\lambda_{0j}'\xi_{ik}\beta_{it} = o_p(1)$,
$\frac{1}{NT\sqrt{N}}\sum_{i\neq j,(i,j)\in S_U}\sum_{s=1}^{T}\sum_{l=1}^{N}(u_{is}u_{ls} - Eu_{is}u_{ls})\lambda_{0j}\lambda_{0l}'\xi_{ik}\beta_{jt} = o_p(1)$.

Theorem 3.6. *Under the assumptions of Theorem 3.5, we have for each fixed $t \leq T$,*

$$\|\widehat{f}_t^{(1)} - f_t\| = O_p(m_N\omega_T^{1-q}(\log T)^{1/r_1+1/r_2}).$$

where $r_1, r_2 > 0$ are defined in Assumption 3.2.
If in addition Assumptions 3.11, 3.12 are satisfied and $\sqrt{N}m_N^2\omega_T^{2-2q} = o(1)$. Then when $T^{1/(2-2q)} \ll N \ll T^{2-2q}$,
$$\sqrt{N}(\widehat{f}_t^{(1)} - f_t) \to^d N(0, Q^{-1}).$$

Remark 3.4. 1. It follows from Theorem 3.6 that for each fixed t, $\widehat{f}_t^{(1)}$ is a root-N consistent estimator of f_t. Root-N consistency for the estimated common factors also holds for the principal components estimator as in Bai (2003). In addition, the above limiting distribution holds only when $N = o(T^2)$.

2. If we strengthen the assumption to $\max_{t \leq T} \|\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \xi_i u_{it}\| = O_p(\log T)$, then the uniform rate of convergence can be achieved:

$$\max_{t \leq T} \|\widehat{f}_t^{(1)} - f_t\| = O_p(m_N \omega_T^{1-q}(\log T)^{1/r_1 + 1/r_2 + 1}).$$

To compare this rate with that of the PCA estimator, we consider for simplicity, the strictly sparse case $q = 0$. Then when $N = o(T^{3/2})$ and m_N is either bounded or growing slowly $(m_N^2 \ll \min\{\sqrt{T}, T^{3/2}/N\})$, the above rate is faster than that of the PCA estimator. (The above rate is $O_p((\log T)^{1/r_1 + 1/r_2}/\sqrt{N})$ when $N = O(T)$, whereas the uniform convergence rate for PCA estimator is $O_p(T^{1/4}/\sqrt{N})$.)

4 Joint Estimation

4.1 l_1- penalized maximum likelihood

One can also jointly estimate (Λ_0, Σ_{u0}) to take into account the cross-sectional dependence and heteroskedasticity simultaneously. As in the sparse covariance estimation literature (e.g., Lam and Fan 2009, Bien and Tibshirani 2011), we penalize the off-diagonal elements of the error covariance estimator, and minimize the following weighted-l_1 penalized objective function, motivated by a penalized Gaussian likelihood function:

$$
\begin{aligned}
(\widehat{\Lambda}^{(2)}, \widehat{\Sigma}_u^{(2)}) &= \arg \min_{(\Lambda, \Sigma_u) \in \Theta_\lambda \times \Gamma} L_2(\Lambda, \Sigma_u) \\
&= \arg \min_{\Lambda \in \Theta_\lambda \times \Gamma} \frac{1}{N} \log |\det(\Lambda \Lambda' + \Sigma_u)| + \frac{1}{N} \mathrm{tr}(S_y(\Lambda \Lambda' + \Sigma_u)^{-1}) \\
&\quad + \frac{1}{N} \sum_{i \neq j} \mu_T w_{ij} |\Sigma_{u,ij}|,
\end{aligned}
\tag{4.1}
$$

where Γ is the parameter space for Σ_u, to be defined later. We introduce the weighted l_1-penalty $N^{-1}\mu_T \sum_{i \neq j} w_{ij} |\Sigma_{u,ij}|$ with $w_{ij} \geq 0$ to penalize the inclusion of many off-diagonal elements of $\Sigma_{u,ij}$ in small magnitudes, which therefore produces a sparse estimator $\widehat{\Sigma}_u^{(2)}$. Here μ_T is a tuning parameter that converges to zero at a not-too-fast rate; w_{ij} is an entry-dependent weight parameter, which can be either deterministic or stochastic. Popular choices of w_{ij} in the literature include:

Lasso The choice $w_{ij} = 1$ for all $i \neq j$ gives the well-known Lasso penalty $N^{-1}\mu_T \sum_{i \neq j} |\Sigma_{u,ij}|$ studied by Tibshirani (1996). The Lasso penalty puts an equal weight to each element of the idiosyncratic covariance matrix.

Adaptive-Lasso Let $\widehat{\Sigma}^*_{u,ij}$ be a preliminary consistent estimator of $\Sigma_{u0,ij}$. Let $w_{ij} = |\widehat{\Sigma}^*_{u,ij}|^{-\gamma}$ for some $\gamma > 0$, then

$$\frac{\mu_T}{N} \sum_{i \neq j} w_{ij} |\Sigma_{u,ij}| = \frac{\mu_T}{N} \sum_{i \neq j} |\widehat{\Sigma}^*_{u,ij}|^{-\gamma} |\Sigma_{u,ij}|$$

corresponds to the adaptive-lasso penalty proposed by Zou (2006). Note that the adaptive-lasso puts an entry-adaptive weight on each off-diagonal element of Σ_u, whose reciprocal is proportional to the preliminary estimate. If the true element $\Sigma_{u0,ij} \in S_L$, the weight $|\widehat{\Sigma}^*_{u,ij}|^{-\gamma}$ should be quite large, and results in a heavy penalty on that entry. The preliminary estimator $\widehat{\Sigma}^*_{u,ij}$ can be taken, for example, as the PCA estimator $\widehat{\Sigma}^{PCA}_{u,ij} = T^{-1} \sum_{t=1}^T \widehat{u}^{PCA}_{it} \widehat{u}^{PCA'}_{jt}$. It was shown by Bai (2003) that under mild conditions, $\widehat{\Sigma}^{PCA}_{u,ij} - \Sigma_{u0,ij} = O_p(N^{-1/2} + T^{-1/2})$.

SCAD: Fan and Li (2001) proposed to use, for some $a > 2$ (e.g, $a = 3.7$)

$$w_{ij} = I_{(|\widehat{\Sigma}^*_{u,ij}| \leq \mu_T)} + \frac{(a - |\widehat{\Sigma}^*_{u,ij}|/\mu_T)_+}{a-1} I_{(|\widehat{\Sigma}^*_{u,ij}| > \mu_T)}.$$

The notation z_+ stands for the positive part of z; z_+ is z if $z > 0$, zero otherwise. Here $\widehat{\Sigma}^*_{u,ij}$ is still a preliminary consistent estimator, which can be taken as the PCA estimator.

4.2 Consistency of the joint estimation

We assume the parameter space for Σ_{u0} to be, for some known sufficiently large $M > 0$,

$$\Gamma = \{\Sigma_u : \|\Sigma_u\|_1 < M, \|\Sigma_u^{-1}\|_1 < M\}.$$

Then $\Sigma_{u0} \in \Gamma$ implies that all the eigenvalues of Σ_{u0} are bounded away from both zero and infinity. There are many examples where both the covariance and its inverse have bounded row sums. For example, for each t, when $\{u_{it}\}_{i=1}^N$ follows a cross sectional autoregressive process $AR(p)$ for some fixed p, then the maximum row sum of Σ_{u0} is bounded. The inverse of Σ_{u0} is a banded matrix, whose maximum row sum is also bounded.

As before we assume $T^{-1} \sum_{t=1}^T f_t f_t' = I_r$ and $\Lambda_0' \Sigma_{u0}^{-1} \Lambda$ be diagonal for identification. In addition, Assumptions 3.2 and 3.3 for the two-step estimation are still needed. Those conditions such as strong mixing, weakly dependence and bounded eigenvalues of $N^{-1} \Lambda_0' \Lambda_0$ regulate the data generating process, and asymptotically identify the covariance decomposition (2.1).

The conditions for the partition $\{(i,j) : i, j \leq N\} = S_L \cup S_U$ of Σ_{u0} are replaced by the following, which are weaker than those of two-step estimation in Assumption 3.8. Define the number of off-diagonal large entries:

$$D = \sum_{i \neq j, (i,j) \in S_U} 1. \tag{4.2}$$

Assumption 4.1. *There exists a partition $\{(i,j) : i \leq N, j \leq N\} = S_L \cup S_U$ where S_U and S_L are disjoint, which satisfies:*
(i) $\Sigma_{u0,ii} \in S_U$ for all $i \leq N$,
(ii) $D = o(\min\{N\sqrt{T/\log N}, N^2/\log N\})$,
(iii) $\sum_{(i,j) \in S_L} |\Sigma_{u0,ij}| = o(N)$.

The following assumption is imposed on the penalty parameters. Define the weights ratios

$$\alpha_T = \frac{\max_{i \neq j, (i,j) \in S_U} w_{ij}}{\min_{(i,j) \in S_L} w_{ij}}, \quad \beta_T = \frac{\max_{(i,j) \in S_L} w_{ij}}{\min_{(i,j) \in S_L} w_{ij}}.$$

Assumption 4.2. *The tuning parameter μ_T and the weights $\{w_{ij}\}_{i \leq N, j \leq N}$ satisfy:*
(i)

$$\alpha_T = o_p\left[\min\left\{\sqrt{\frac{T}{\log N}}\frac{N}{D}, \left(\frac{T}{\log N}\right)^{1/4}\sqrt{\frac{N}{D}}, \frac{N}{\sqrt{D\log N}}\right\}\right],$$

$$\beta_T \sum_{(i,j) \in S_L} |\Sigma_{u0,ij}| = o_p(N),$$

(ii) $\mu_T \max_{(i,j) \in S_L} w_{ij} \sum_{(i,j) \in S_L} |\Sigma_{u0,ij}| = o(\min\{N, N^2/D, N^2/(D\alpha_T^2)\})$,
$\mu_T \max_{i \neq j, (i,j) \in S_U} w_{ij} = o(\min\{N/D, \sqrt{N/D}, N/(D\alpha_T)\})$,
$\mu_T \min_{(i,j) \in S_L} w_{ij} \gg \sqrt{\log N/T} + (\log N)/N$.

The above assumption is not as complicated as it looks, and is satisfied by many examples. For instance, the Lasso penalty sets $w_{ij} = 1$ for all $i, j \leq N$. Hence $\alpha_T = \beta_T = 1$. Then condition (i) of Assumption 4.2 follows from Assumption 4.1(ii), which is also satisfied if $D = O(N)$. Condition (ii) is also straightforward to verify. This immediately implies the following lemma.

Lemma 4.1 (Lasso). *Choose $w_{ij} = 1$ for all $i, j \leq N, i \neq j$. Suppose in addition $D = O(N)$ and $\log N = o(T)$. Then Assumption 4.2 is satisfied if the tuning parameter $\mu_T = o(1)$ is such that*

$$\sqrt{\frac{\log N}{T}} + \frac{\log N}{N} = o(\mu_T).$$

One of the attractive features of this lemma is that the condition on μ_T does not depend on the unknown Σ_{u0}. We will present the adaptive lasso and SCAD as another two examples of the weighted-l_1 penalty in Section 4.3 below, both satisfy the above assumption.

Our main theorem is stated as follows.

Theorem 4.1. *Suppose* $\log N = o(T)$. *Under Assumptions 3.2, 3.3, 3.7, 4.1, and 4.2, the penalized ML estimator satisfies: as T and $N \to \infty$,*

$$\frac{1}{N}\|\widehat{\Sigma}_u^{(2)} - \Sigma_{u0}\|_F^2 \to^p 0, \quad \frac{1}{N}\|\widehat{\Lambda}^{(2)} - \Lambda_0\|_F^2 \to^p 0.$$

For each $t \leq T$,

$$\|\widehat{f}_t^{(2)} - f_t\| = o_p(1).$$

Remark 4.1. 1. The consistency for $\widehat{f}_t^{(2)}$ can be made uniformly in $t \leq T$ if the condition is strengthened to $\max_{t \leq T} \|N^{-1/2} \sum_{i=1}^{N} \xi_i u_{it}\| = o_p(\sqrt{N})$.

2. To establish the consistency in the high dimensional literature, one usually constructs a neighborhood of the true parameters $(\Lambda_0, \Sigma_{u0}) \in U$ (e.g., Rothman et al. 2008, Lam and Fan 2009), and show that with probability approaching one, $L_2(\Lambda_0, \Sigma_{u0}) > \sup_{(\Lambda, \Sigma_u) \notin U} L_2(\Lambda, \Sigma_u)$. This strategy however, does not work here due to the technical difficulty in dealing with the term $(\Lambda\Lambda' + \Sigma_u)$ in the likelihood function, because its largest r eigenvalues are unbounded and grow at rate $O(N)$ uniformly in the parameter space. One of the contributions of Theorem 4.1 is to achieve consistency using a new strategy to deal with the penalized likelihood function, which involves diverging eigenvalues.

In this paper we only present the consistency for the joint estimation, which is already technically difficult as one needs to deal with an equilibrium of the first order conditions for both $(\widehat{\Lambda}^{(2)}, \widehat{\Sigma}_u^{(2)})$ simultaneously. Deriving the limiting distributions for the joint estimators is difficult, and we leave this as a future topic.

4.3 Two examples

We present two popular choices for the weights as examples: one is adaptive lasso, proposed by Zou (2006), and the other is SCAD by Fan and Li (2001). Both weights depend on a preliminary consistent estimate of each element of Σ_{u0}. In the high dimensional approximate factor model, a simple consistent estimate for each element can be obtained by the principal component analysis (Stock and Watson 1998 and Bai 2003).

To simplify the presentation, we will assume that $D = O(N)$, which controls the number of off-diagonal large entries of Σ_{u0}. Moreover, we retain Assumption 3.6:

$$\max\{|\Sigma_{u0,ij}| : (i,j) \in S_L\} \ll \omega_T \ll \min\{|\Sigma_{u0,ij}| : \Sigma_{u0,ij} \in S_U\},$$

and recall that $\omega_T = \sqrt{\frac{\log N}{T}} + \frac{1}{\sqrt{N}}$.

Let the initial estimate $\widehat{\Sigma}^*_{u,ij} = R_{ij}$, where R_{ij} is the PCA estimator of $\Sigma_{u0,ij}$ as in Bai (2003). The adaptive lasso chooses the weights to be, for some constant $\gamma \in (0, 1]$,

$$\text{(Adaptive Lasso)}: \quad w_{ij} = (|\widehat{\Sigma}^*_{u,ij}| + \delta_T)^{-\gamma}, \tag{4.3}$$

where $\delta_T = o(1)$ is a pre-determined nonnegative sequence. The additive δ_T was not included in the original definition of adaptive lasso in Zou (2006), but has often been seen in recent literature, e.g., Xue and Zou (2012). We include it here in the weights to prevent w_{ij} getting too large if $|\widehat{\Sigma}^*_{u,ij}|$ is very close to zero. The adaptive lasso has been used extensively in the high dimensional literature, see for example, Huang, Ma and Zhang (2006), van de Geer, Bühlmann and Zhou (2011), Caner and Fan (2011), etc.

Another important example is SCAD, defined as: for some $a > 2$,

$$\text{(SCAD)}: \quad w_{ij} = I_{(|\widehat{\Sigma}^*_{u,ij}| \le \mu_T)} + \frac{(a - |\widehat{\Sigma}^*_{u,ij}|/\mu_T)_+}{a - 1} I_{(|\widehat{\Sigma}^*_{u,ij}| > \mu_T)}. \tag{4.4}$$

We have the following theorem.

Theorem 4.2. *Suppose either the Adaptive Lasso or SCAD is used for the weighted-l_1 penalized objective function. Also, suppose $\log N = o(T)$, $D = O(N)$, $\sum_{(i,j)\in S_L} |\Sigma_{u0,ij}| = o(N)$ and Assumptions 3.2, 3.3, 3.7, 4.1 hold. In addition, assume the tuning parameters are such that:*
(i) for Adaptive Lasso,

$$\omega_T \left(\frac{\sum_{(i,j)\in S_L} |\Sigma_{u0,ij}|}{N} \right)^{1/\gamma} \ll \delta_T \ll \omega_T, \tag{4.5}$$

$$\omega_T^{1+\gamma} \ll \mu_T \ll \omega_T^{\gamma}; \tag{4.6}$$

(ii) for SCAD:

$$\left(\frac{\log N}{T} \right)^{1/4} + \left(\frac{\log N}{N} \right)^{1/2} \ll \mu_T \ll \min_{i\neq j, (i,j)\in S_U} |\Sigma_{u0,ij}|. \tag{4.7}$$

21

Then Assumption 4.2 is satisfied, and

$$\frac{1}{N}\|\widehat{\Sigma}_u^{(2)} - \Sigma_{u0}\|_F^2 = o_p(1), \quad \frac{1}{N}\|\widehat{\Lambda}^{(2)} - \Lambda_0\|_F^2 = o_p(1).$$

$$\|\widehat{f}_t^{(2)} - f_t\| = o_p(1).$$

As in the case of Lemma 4.1, an attractive feature of this theorem is that, if both the upper bound of $\sum_{(i,j)\in S_L} |\Sigma_{u0,ij}|$ and the lower bound of $\min_{i\neq j,(i,j)\in S_U} |\Sigma_{u0,ij}|$ are known, [e.g., in the strictly sparse model, $\sum_{(i,j)\in S_L} |\Sigma_{u0,ij}| = 0$, and assume $\min_{i\neq j,(i,j)\in S_U} |\Sigma_{u0,ij}|$ is bounded away from zero as in MA(1)] then Conditions (4.5) - (4.7) do not depend on any other unknown feature of Σ_{u0}.

5 Numerical Examples

We propose a novel algorithm to numerically minimize the objective function $L_2(\Lambda, \Sigma_u)$ (4.1) for joint estimation, which combines the EM algorithm with the majorize-minimize method recently proposed by Bien and Tibshirani (2011). The algorithm uses the PCA as initial values, and updates the estimator iteratively. At each iteration, an EM-algorithm is carried out to estimate Λ and the empirical residual covariance $\frac{1}{T}\sum_{t=1}^{T}\widehat{u}_t\widehat{u}_t'$. Then a majorize-minimize method (Bien and Tibshirani 2011) is used to obtain a positive definite estimate of the covariance Σ_u based on $\frac{1}{T}\sum_{t=1}^{T}\widehat{u}_t\widehat{u}_t'$ and soft-thresholding. The algorithm is summarized as follows (see Bai and Li (2012) and Bien and Tibshirani (2011) for detailed descriptions of the algorithm).

1. Initialize $\widehat{\Lambda}$ and $\widehat{u}$ as the PCA estimators. Initialize $\widehat{\Sigma}_u$ as a diagonal matrix of the sample covariance based on the PCA residuals.

2. At step k+1, $\widehat{\Lambda}_{k+1} = AM^{-1}$, where
$M = \widehat{\Lambda}_k'\widehat{\Sigma}_{y,k}^{-1}S_y\widehat{\Sigma}_{y,k}^{-1}\widehat{\Lambda}_k + I_r - \widehat{\Lambda}_k'\widehat{\Sigma}_{y,k}^{-1}\widehat{\Lambda}_k,$

$$A = S_y\widehat{\Sigma}_{y,k}^{-1}\widehat{\Lambda}_k, \quad \widehat{\Sigma}_{y,k} = \widehat{\Lambda}_k\widehat{\Lambda}_k' + \widehat{\Sigma}_{u,k}.$$

Let $S_{u,k} = S_y - A\widehat{\Lambda}_{k+1}' - \widehat{\Lambda}_{k+1}A' + \widehat{\Lambda}_{k+1}M\widehat{\Lambda}_{k+1}'$.

3. Still at step $k+1$, For some small value $t > 0$, let $B = \widehat{\Sigma}_{u,k} - t(\widehat{\Sigma}_{u,k}^{-1} - \widehat{\Sigma}_{u,k}^{-1}S_{u,k}\widehat{\Sigma}_{u,k}^{-1})$. Let

$$\widehat{\Sigma}_{u,k+1} = S(B, \lambda t K)$$

where $S(A, B)_{ij} = sign(A_{ij})(A_{ij} - B_{ij})$ and K is a matrix whose off-diagonal K_{ij} is $|(S_{u,k})_{ij}|^{-\gamma}$ and diagonal elements are zero.

4. Repeat 2-3 until converge.

We present a numerical experiment to illustrate the performance of the proposed method. The data was generated as following: $\{e_{it}\}_{i \leq N, t \leq T}$ are both serially and cross-sectionally independent as $N(0, 1)$. Let

$$u_{1t} = e_{1t}, \quad u_{2t} = e_{2t} + a_1 e_{1t}, \quad u_{3t} = e_{3t} + a_2 e_{2t} + b_1 e_{1t},$$

$$u_{i+1,t} = e_{i+1,t} + a_i e_{it} + b_{i-1} e_{i-1,t} + c_{i-2} e_{i-2,t},$$

where $\{a_i, b_i, c_i\}_{i=1}^N$ are i.i.d. $N(0, 0.7^2)$. Let the two factors $\{f_{1t}, f_{2t}\}$ be i.i.d. $N(0, 1)$, and $\{\lambda_{i,1}, \lambda_{i,2}\}_{i \leq N}$ be uniform on $[0, 1]$. Then Σ_{u0} is a banded matrix.

We apply the adaptive lasso penalty for our joint estimation, with various choices of the tuning parameters γ and μ_T. The result is compared with the PCA estimator and the regular maximum likelihood restricted to diagonal $\widehat{\Sigma}_u$ (DML, Bai and Li 2012). More specifically, DML estimates (Λ_0, Σ_{u0}) by:

$$\min_{\Sigma_{u,ij}=0 \text{ for } i \neq j} \min_{\Lambda} \frac{1}{N} \log |\Lambda\Lambda' + \Sigma_u| + \frac{1}{N} \text{tr}(S_y(\Lambda\Lambda' + \Sigma_u)^{-1}). \tag{5.1}$$

Therefore DML forces the covariance estimator to be diagonal even though the true Σ_{u0} is not. Hence it does not take the idiosyncratic cross-sectional dependence into account.

For each estimator, the smallest canonical correlation (the higher the better) between the estimator and the parameter has been used as a measurement to assess the accuracy of each estimator. Tables 1 and 2 list the results of the estimated factor loadings and common factors from joint-estimation.

We have also computed the canonical correlations between the estimators and the true parameters using the regularized two-step method (Section 3) with iterations. For computational simplicity, the threshold value in the first step has been fixed to be the adaptive threshold of Fan et al. (2012) with a universal constant $C = 1$, which we find to maintain the finite-sample positive definiteness well. The results demonstrate that both two-step and joint estimations have higher canonical correlations, and thus outperform the PCA and DML.

Our EM plus majorize-minimize algorithm maximizes an approximate penalized likelihood function. Developing an efficient algorithm for maximizing the original likelihood function will be a future research direction.

Table 1: Canonical correlations between $\widehat{\Lambda}^{(2)}$ and Λ_0

| | | | | Penalized ML | | | |
| | | PCA | DML | $\gamma = 1$ | | $\gamma = 5$ | |
T	N			$\mu_T = 0.08$	$\mu_T = 0.3$	$\mu_T = 0.08$	$\mu_T = 0.3$
50	50	0.205	0.199	0.212	0.222	0.230	0.234
50	100	0.429	0.558	0.591	0.613	0.627	0.631
50	150	0.328	0.470	0.494	0.495	0.515	0.507
100	50	0.496	0.519	0.560	0.537	0.558	0.537
100	100	0.394	0.574	0.621	0.648	0.648	0.658
100	150	0.774	0.819	0.837	0.829	0.840	0.836

Canonical correlations are presented. DML is defined in (5.1). Penalized ML uses the one-step adaptive Lasso estimation.

Table 2: Canonical correlations between $\widehat{F}^{(2)}$ and F

| | | | | Penalized ML | | | |
| | | PCA | DML | $\gamma = 1$ | | $\gamma = 5$ | |
T	N			$\mu_T = 0.08$	$\mu_T = 0.3$	$\mu_T = 0.08$	$\mu_T = 0.3$
50	50	0.232	0.234	0.251	0.267	0.279	0.283
50	100	0.477	0.640	0.671	0.732	0.748	0.749
50	150	0.411	0.599	0.623	0.638	0.666	0.650
100	50	0.430	0.446	0.503	0.473	0.508	0.474
100	100	0.371	0.579	0.647	0.688	0.687	0.697
100	150	0.820	0.867	0.880	0.892	0.912	0.903

Canonical correlations are presented. Penalized ML uses the one-step adaptive Lasso estimation.

Table 3: Canonical correlations between the regularized two-step ML estimators (Section 3) and the true parameters

T	N	Factor loadings			Factors		
		PCA	DML	Two-step ML	PCA	DML	Two-step ML
50	50	0.205	0.199	0.241	0.232	0.234	0.277
50	100	0.429	0.558	0.643	0.477	0.640	0.752
50	150	0.328	0.470	0.565	0.411	0.599	0.731
100	50	0.496	0.519	0.548	0.430	0.446	0.469
100	100	0.394	0.574	0.717	0.371	0.579	0.758
100	150	0.774	0.819	0.846	0.820	0.867	0.927

The $SCAD(\tau_{ij})$ threshold has been used for the covariance estimation, where $\tau_{ij} = \alpha_{ij}\omega_T$ with the adaptive threshold constant α_{ij} proposed by Cai and Liu (2011).

6 Conclusion

We study the estimation of a high dimensional approximate factor model in the presence of cross sectional dependence and heteroskedasticity. The classical PCA method does not efficiently estimate the factor loadings or common factors because it essentially treats the idiosyncratic error to be homoskedastic and cross sectionally uncorrelated. For the efficient estimation, it is essential to estimate a large error covariance matrix.

We assume the model be conditionally sparse in the sense that after the common factors are taken out, the idiosyncratic components have a sparse covariance matrix. This enables us to combine the merits of both sparsity and high dimensional factor analysis. Two maximum-likelihood-based approaches are proposed to estimate the common factors and factor loadings, both involve regularizing a large covariance sparse matrix. Extensive asymptotic analysis has been carried out. In particular, we develop the inferential theory for the two-step estimation.

It remains to derive the limiting distribution as well as the optimal rates of convergence for the estimators by the joint-estimation method. This extends the consistency results obtained in the current paper. In the presence of a covariance $\Lambda_0\Lambda_0'$ that has fast-diverging eigenvalues, the task is difficult because it requires the consistency of the penalized covariance estimator under the operator norm. We intend to address this issue in future research.

A   Proofs for generic estimators

We need to establish the results for two sets of estimators: the two-step estimator and the joint estimator, whose proofs for consistency share some similarities. Therefore in this section we establish some preliminary results for generic estimators that can be used for both cases. We denote by $(\widehat{\Lambda}, \widehat{\Sigma}_u)$ as a generic estimator for (Λ_0, Σ_{u0}), which can be either $(\widehat{\Lambda}^{(1)}, \widehat{\Sigma}_u^{(1)})$ or $(\widehat{\Lambda}^{(2)}, \widehat{\Sigma}_u^{(2)})$. Define

$$Q_2(\Lambda, \Sigma_u) = \frac{1}{N}\mathrm{tr}(\Lambda_0'\Sigma_u^{-1}\Lambda_0 - \Lambda_0'\Sigma_u^{-1}\Lambda(\Lambda'\Sigma_u^{-1}\Lambda)^{-1}\Lambda'\Sigma_u^{-1}\Lambda_0), \tag{A.1}$$

$$Q_3(\Lambda, \Sigma_u) = \frac{1}{N}\log|\Lambda\Lambda'+\Sigma_u|+\frac{1}{N}\mathrm{tr}(S_y(\Lambda\Lambda'+\Sigma_u)^{-1})-\frac{1}{N}\mathrm{tr}(S_u\Sigma_u^{-1})-\frac{1}{N}\log|\Sigma_u|-Q_2(\Lambda, \Sigma_u). \tag{A.2}$$

Define the set

$$\Xi_\delta = \{(\Lambda, \Sigma_u) : \quad \delta^{-1} < \lambda_{\min}(N^{-1}\Lambda'\Lambda) \leq \lambda_{\max}(N^{-1}\Lambda'\Lambda) < \delta,$$
$$\delta^{-1} < \lambda_{\min}(\Sigma_u) \leq \lambda_{\max}(\Sigma_u) < \delta\}$$

We first present a lemma that will be needed throughout the proof.

Lemma A.1. *(i)* $\max_{i,j\leq r}|\frac{1}{T}\sum_{t=1}^T f_{it}f_{jt} - Ef_{it}f_{jt}| = O_p(\sqrt{1/T})$.
(ii) $\max_{i,j\leq N}|\frac{1}{T}\sum_{t=1}^T u_{it}u_{jt} - Eu_{it}u_{jt}| = O_p(\sqrt{(\log N)/T})$.
(iii) $\max_{i\leq r,j\leq N}|\frac{1}{T}\sum_{t=1}^T f_{it}u_{jt}| = O_p(\sqrt{(\log N)/T})$.

Proof. See Lemmas A.3 and B.1 in Fan, Liao and Mincheva (2011). □

Lemma A.2. *Under Assumption 3.2, for any $\delta > 0$,*

$$\sup_{(\Lambda,\Sigma_u)\in\Xi_\delta} |Q_3(\Lambda, \Sigma_u)| = O\left(\frac{\log N}{N} + \sqrt{\frac{\log N}{T}}\right).$$

Therefore we can write

$$\frac{1}{N}\log|\Lambda\Lambda' + \Sigma_u| + \frac{1}{N}\mathrm{tr}(S_y(\Lambda\Lambda' + \Sigma_u)^{-1})$$
$$= \frac{1}{N}\mathrm{tr}(S_u\Sigma_u^{-1}) + \frac{1}{N}\log|\Sigma_u| + Q_2(\Lambda, \Sigma_u) + O\left(\frac{\log N}{N} + \sqrt{\frac{\log N}{T}}\right). \tag{A.3}$$

Proof. First of all, note that $|\Lambda\Lambda' + \Sigma_u| = |\Sigma_u| \times |I_r + \Lambda'\Sigma_u^{-1}\Lambda|$, and $\sup_{(\Lambda,\Sigma_u)\in\Xi_\delta} \frac{1}{N}\log|I_r +$

$\Lambda'\Sigma_u^{-1}\Lambda| = O\left(\frac{\log N}{N}\right)$, hence we have

$$\frac{1}{N}\log|\Lambda\Lambda' + \Sigma_u| = \frac{1}{N}\log|\Sigma_u| + O\left(\frac{\log N}{N}\right), \tag{A.4}$$

where $O(\cdot)$ is uniform in Ξ_δ. Equation (A.4) will be used later in the proof.

We now consider the term $N^{-1}\mathrm{tr}(S_y(\Lambda\Lambda' + \Sigma_u)^{-1})$. With the identification condition $\frac{1}{T}\sum_{t=1}^{T} f_t f_t' = I_r$, $\bar{f} = 0$, and $S_u = \frac{1}{T}\sum_{t=1}^{T} u_t u_t'$,

$$S_y = \frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y})(y_t - \bar{y})' = \Lambda_0\Lambda_0' + S_u + \Lambda_0\frac{1}{T}\sum_{t=1}^{T} f_t u_t' + (\Lambda_0\frac{1}{T}\sum_{t=1}^{T} f_t u_t')' - \bar{u}\bar{u}'.$$

By the matrix inversion formula $(\Lambda\Lambda' + \Sigma_u)^{-1} = \Sigma_u^{-1} - \Sigma_u^{-1}\Lambda(I_r + \Lambda'\Sigma_u^{-1}\Lambda)^{-1}\Lambda'\Sigma_u^{-1}$,

$$\frac{1}{N}\mathrm{tr}(S_y(\Lambda\Lambda' + \Sigma_u)^{-1}) = \frac{1}{N}\mathrm{tr}(\Lambda_0'\Sigma_u^{-1}\Lambda_0) + \frac{1}{N}\mathrm{tr}(S_u\Sigma_u^{-1}) - A_1 + A_2 + A_3 - A_4 - A_5, \tag{A.5}$$

where $A_1 = N^{-1}\mathrm{tr}(\Lambda_0\Lambda_0'\Sigma_u^{-1}\Lambda(I_r+\Lambda'\Sigma_u^{-1}\Lambda)^{-1}\Lambda'\Sigma_u^{-1})$, $A_2 = \frac{1}{N}\mathrm{tr}(\frac{1}{T}\sum_{t=1}^{T}\Lambda_0 f_t u_t'(\Lambda\Lambda'+\Sigma_u)^{-1})$, $A_3 = \frac{1}{N}\mathrm{tr}(\frac{1}{T}\sum_{t=1}^{T} u_t f_t'\Lambda_0'(\Lambda\Lambda'+\Sigma_u)^{-1})$, and $A_4 = \frac{1}{N}\mathrm{tr}(S_u\Sigma_u^{-1}\Lambda(I_r+\Lambda'\Sigma_u^{-1}\Lambda)^{-1}\Lambda'\Sigma_u^{-1})$. Term $A_5 = N^{-1}\mathrm{tr}(\bar{u}\bar{u}'(\Lambda\Lambda'+\Sigma_u)^{-1}) = O_p((\log N)/T)$ uniformly in the parameter space, and hence can be ignored.

Let us look at terms A_1, A_2, A_3 and A_4 subsequently. Note that $\lambda_{\max}(\Sigma_u)$ and $N\lambda_{\min}^{-1}(\Lambda'\Lambda)$ are both bounded from above uniformly in Ξ_δ, we have,

$$\sup_{(\Lambda,\Sigma_u)\in\Xi_\delta} \lambda_{\max}[(\Lambda'\Sigma_u^{-1}\Lambda)^{-1}] \leq \sup_{(\Lambda,\Sigma_u)\in\Xi_\delta} \frac{\lambda_{\max}(\Sigma_u)}{\lambda_{\min}(\Lambda'\Lambda)} = O(N^{-1}), \tag{A.6}$$

$$\sup_{(\Lambda,\Sigma_u)\in\Xi_\delta} \lambda_{\max}[(I_r + \Lambda'\Sigma_u^{-1}\Lambda)^{-1}] \leq \sup_{(\Lambda,\Sigma_u)\in\Xi_\delta} \lambda_{\max}[(\Lambda'\Sigma_u^{-1}\Lambda)^{-1}] = O(N^{-1}). \tag{A.7}$$

In addition, $\|\Lambda\|_F = O(\sqrt{N})$, $\lambda_{\max}(\Sigma_u^{-1}) = O(1)$ uniformly in Ξ_δ, and $\|\Lambda_0\|_F = O(\sqrt{N})$. Applying the matrix inversion formula yields

$$
\begin{aligned}
A_1 &= \frac{1}{N}\mathrm{tr}(\Lambda_0'\Sigma_u^{-1}\Lambda(\Lambda'\Sigma_u^{-1}\Lambda)^{-1}\Lambda'\Sigma_u^{-1}\Lambda_0) - \frac{1}{N}\mathrm{tr}(\Lambda_0'\Sigma_u^{-1}\Lambda(\Lambda'\Sigma_u^{-1}\Lambda)^{-1}(I_r + \Lambda'\Sigma_u^{-1}\Lambda)^{-1}\Lambda'\Sigma_u^{-1}\Lambda_0) \\
&= \frac{1}{N}\mathrm{tr}(\Lambda_0'\Sigma_u^{-1}\Lambda(\Lambda'\Sigma_u^{-1}\Lambda)^{-1}\Lambda'\Sigma_u^{-1}\Lambda_0) + O\left(\frac{1}{N}\right),
\end{aligned} \tag{A.8}
$$

where $O(\cdot)$ is uniform over $(\Lambda, \Sigma_u) \in \Xi_\delta$. In the second equality above we applied (A.6) and (A.7) and the following inequality:

$$\frac{1}{N}\mathrm{tr}(\Lambda_0'\Sigma_u^{-1}\Lambda(\Lambda'\Sigma_u^{-1}\Lambda)^{-1}(I_r + \Lambda'\Sigma_u^{-1}\Lambda)^{-1}\Lambda'\Sigma_u^{-1}\Lambda_0)$$

$$\leq \frac{1}{N}\|\Lambda_0'\Sigma_u^{-1}\Lambda\|_F^2 \lambda_{\max}[(\Lambda'\Sigma_u^{-1}\Lambda)^{-1}]\lambda_{\max}[(I_r + \Lambda'\Sigma_u^{-1}\Lambda)^{-1}]$$
$$\leq O(N^{-3})\|\Lambda_0\|_F^2\|\Lambda\|_F^2 \lambda_{\max}(\Sigma_u^{-1}) = O(N^{-1}).$$

By Lemma A.1(iii), and $\lambda_{\max}((\Lambda\Lambda' + \Sigma_u)^{-1}) \leq \lambda_{\max}(\Sigma_u^{-1}) = O(1)$ uniformly in Ξ_δ,

$$\sup_{(\Lambda,\Sigma_u)\in\Xi_\delta} |A_2| \quad \leq \quad \frac{1}{N}\|\Lambda_0'(\Lambda\Lambda' + \Sigma_u)^{-1}\|_F \left\|\frac{1}{T}\sum_{t=1}^{T} f_t u_t'\right\|_F = O_p(\sqrt{\frac{\log N}{T}}). \qquad (A.9)$$

Similarly, $\sup_{(\Lambda,\Sigma_u)\in\Xi_\delta} |A_3| = O_p(\sqrt{\frac{\log N}{T}})$. Again by the matrix inversion formula,

$$A_4 = \frac{1}{N}\mathrm{tr}(S_u\Sigma_u^{-1}\Lambda(\Lambda'\Sigma_u^{-1}\Lambda)^{-1}\Lambda'\Sigma_u^{-1}) - \frac{1}{N}\mathrm{tr}(S_u\Sigma_u^{-1}\Lambda(\Lambda'\Sigma_u^{-1}\Lambda)^{-1}(I + \Lambda'\Sigma_u^{-1}\Lambda)^{-1}\Lambda'\Sigma_u^{-1}).$$

The second term on the right hand side is of smaller order (uniformly) than the first term, because it has an additional term $(I + \Lambda'\Sigma_u^{-1}\Lambda)^{-1}$, whose maximum eigenvalue is $O(N^{-1})$ uniformly by (A.7). The first term is bounded by (uniformly in Ξ_δ):

$$\frac{c}{N}\|S_u\Sigma_u^{-1}\Lambda\|_F O(N^{-1})\|\Lambda'\Sigma_u^{-1}\|_F \leq O(N^{-1})\lambda_{\max}(S_u) = O(\sqrt{\frac{\log N}{T}} + \frac{1}{N}).$$

Hence $\sup_{(\Lambda,\Sigma_u)\in\Xi_\delta} |A_4| = O(T^{-1/2}(\log N)^{1/2} + N^{-1})$. Results (A.4) and (A.5) then yield

$$\frac{1}{N}\log|\Lambda\Lambda' + \Sigma_u| + \frac{1}{N}\mathrm{tr}(S_y(\Lambda\Lambda' + \Sigma_u)^{-1})$$
$$= \frac{1}{N}\mathrm{tr}(\Lambda_0'\Sigma_u^{-1}\Lambda_0) + \frac{1}{N}\mathrm{tr}(S_u\Sigma_u^{-1}) + \frac{1}{N}\log|\Sigma_u| - \frac{1}{N}\mathrm{tr}(\Lambda_0'\Sigma_u^{-1}\Lambda(\Lambda'\Sigma_u^{-1}\Lambda)^{-1}\Lambda'\Sigma_u^{-1}\Lambda_0)$$
$$+ O\left(\frac{\log N}{N} + \sqrt{\frac{\log N}{T}}\right)$$
$$= \frac{1}{N}\mathrm{tr}(S_u\Sigma_u^{-1}) + \frac{1}{N}\log|\Sigma_u| + Q_2(\Lambda, \Sigma_u) + O\left(\frac{\log N}{N} + \sqrt{\frac{\log N}{T}}\right).$$

$\square$

Throughout the proofs, we note that the consistency depends crucially on the consistency of the following quantities:

$$J = (\widehat{\Lambda} - \Lambda_0)'\widehat{\Sigma}_u^{-1}\widehat{\Lambda}(\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}\widehat{\Lambda})^{-1}$$

We state the following lemma for the generic estimators.

Lemma A.3. *(i)* $\Lambda_0'\Sigma_{u0}^{-1}\Lambda_0 - (I_r - J)\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}\widehat{\Lambda}(I_r - J)' = o_p(N)$

28

(ii) First order condition: $\widehat{\Lambda}'(\widehat{\Lambda}\widehat{\Lambda}' + \widehat{\Sigma}_u)^{-1}(S_y - \widehat{\Lambda}\widehat{\Lambda}' - \widehat{\Sigma}_u) = 0.$

We will prove Lemma A.3 for both $(\widehat{\Lambda}^{(1)}, \widehat{\Sigma}_u^{(1)})$ and $(\widehat{\Lambda}^{(2)}, \widehat{\Sigma}_u^{(2)})$ later when we deal with these two estimators individually.

Lemma A.4. *Suppose Lemma A.3 holds, then*
(i) $\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}(S_y - \widehat{\Lambda}\widehat{\Lambda}' - \widehat{\Sigma}_u) = 0.$
(ii) $(J - I_r)'(J - I_r) - I_r = O_p(N^{-1} + T^{-1/2}(\log N)^{1/2}).$

Proof. (i) Using the matrix inverse formula, the same argument of Bai and Li (2012)'s (A.2) implies $\widehat{\Lambda}'(\widehat{\Lambda}\widehat{\Lambda}' + \widehat{\Sigma}_u)^{-1} = (I_r + \widehat{\Lambda}'\widehat{\Sigma}_u^{-1}\widehat{\Lambda})^{-1}\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}.$ Thus part (i) follows from the first order condition in Lemma A.3.

(ii) Let $H = (\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}\widehat{\Lambda})^{-1}.$ Part (i) can be equivalently written as $J + J' - J'J + K = 0$ where

$$K = J'\frac{1}{T}\sum_{t=1}^{T} f_t u_t'\widehat{\Sigma}_u^{-1}\widehat{\Lambda}H + H\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}\frac{1}{T}\sum_{t=1}^{T} u_t f_t' J - \frac{1}{T}\sum_{t=1}^{T} f_t u_t'\widehat{\Sigma}_u^{-1}\widehat{\Lambda}H - H\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}\frac{1}{T}\sum_{t=1}^{T} u_t f_t'$$

$$-H\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}(S_u - \widehat{\Sigma}_u)\widehat{\Sigma}_u^{-1}\widehat{\Lambda}H.$$

Note that for $(\widehat{\Lambda}, \widehat{\Sigma}_u) \in \Xi_\delta$, $H = O_p(N^{-1})$, $J = O_p(1)$ for each element, $\|\widehat{\Sigma}_u^{-1}\| = O_p(1)$, $\|\widehat{\Lambda}\|_F = O_p(\sqrt{N})$, hence

$$\|\frac{1}{T}\sum_{t=1}^{T} f_t u_t'\widehat{\Sigma}_u^{-1}\widehat{\Lambda}H\|_F \leq O_p(1)\|\frac{1}{NT}\sum_{t=1}^{T} f_t u_t'\|_F\|\widehat{\Sigma}_u^{-1}\|\|\widehat{\Lambda}\|_F = O_p(\frac{1}{N}\sqrt{\frac{N\log N}{T}}\sqrt{N}) = O_p(\sqrt{\frac{\log N}{T}})$$

Moreover, for the empirical covariance $\|S_u\|^2 \leq 2\sum_{i,j\leq N}(T^{-1}\sum_{t=1}^{T} u_{it}u_{jt} - \sigma_{u0,ij})^2 + 2\|\Sigma_{u0}\|^2 = O_p(T^{-1}N^2\log N + 1)$ by Lemma A.1, which implies $H\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}S_u\widehat{\Sigma}_u^{-1}\widehat{\Lambda}H = O_p(N^{-1} + T^{-1/2}(\log N)^{1/2}).$ Also, $H\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}\widehat{\Sigma}_u\widehat{\Sigma}_u^{-1}\widehat{\Lambda}H = H = O_p(N^{-1}).$ Therefore $K = O_p(N^{-1} + T^{-1/2}(\log N)^{1/2}).$ It then implies (ii).

□

Lemma A.5. *Suppose Lemma A.3 holds, then $J = o_p(1).$*

Proof. By our assumption, both $\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}\widehat{\Lambda}$ and $\Lambda_0'\Sigma_{u0}^{-1}\Lambda$ are diagonal. Moreover, the eigenvalues of $N^{-1}\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}\widehat{\Lambda}$ and $N^{-1}\Lambda_0'\Sigma_{u0}^{-1}\Lambda$ are bounded away from zero. Therefore by Lemma A.3(i) and Lemma A.4(ii), there are two diagonal matrices M_1 and M_2 whose eigenvalues are all bounded away from zero, such that

$$(I_r - J)M_1(I_r - J)' = M_2 + o_p(1), \quad (J - I_r)'(J - I_r) = I_r + o_p(1) \tag{A.10}$$

29

Applying Lemma A.1 of Bai and Li (2012), we have $J = o_p(1)$ and $M_1 = M_2 + o_p(1)$. We also assumed $\widehat{\Lambda}$ and Λ_0 have the same column signs, as a part of identification condition.

$\square$

B   Proofs for Section 3

In this section, $(\widehat{\Lambda}, \widehat{\Sigma}_u) = (\widehat{\Lambda}^{(1)}, \widehat{\Sigma}_u^{(1)})$ and $J = (\widehat{\Lambda}^{(1)} - \Lambda_0)(\widehat{\Sigma}_u^{(1)})^{-1}\widehat{\Lambda}^{(1)}(\widehat{\Lambda}^{(1)\prime}(\widehat{\Sigma}_u^{(1)})^{-1}\widehat{\Lambda}^{(1)})^{-1}$. Throughout Appendix B, we will let $H = (\widehat{\Lambda}^{(1)\prime}(\widehat{\Sigma}_u^{(1)})^{-1}\widehat{\Lambda}^{(1)})^{-1}$. For notational simplicity, we let

$$\omega_T = \frac{1}{\sqrt{N}} + \sqrt{\frac{\log N}{T}}.$$

We first cite a result from Fan et al. (2012):

Theorem B.1 (Theorem 3.1 in Fan et al. (2012)). *Suppose* $(\log N)^{6/\gamma} = o(T)$ *and* $\sqrt{T} = o(N)$, *then under Assumptions 3.1- 3.5,*

$$\|\widehat{\Sigma}_u^{(1)} - \Sigma_{u0}\| = O_p\left(m_N \omega_T^{1-q}\right) = \|(\widehat{\Sigma}_u^{(1)})^{-1} - \Sigma_{u0}^{-1}\|.$$

Proof. The sufficient conditions of this theorem are satisfied by our assumptions. See Fan et al. (2012). $\square$

We then prove Lemma A.3, which then enables us to apply Lemmas A.4 and A.5. Under Assumptions 3.1- 3.3, there is $\delta > 0$ such that $(\Lambda_0, \Sigma_{u0}) \in \Xi_\delta$ and $(\widehat{\Lambda}^{(1)}, \widehat{\Sigma}_u^{(1)}) \in \Xi_\delta$ with probability approaching one for Ξ_δ in Appendix A.

Lemma B.1. *For* $(\widehat{\Lambda}, \widehat{\Sigma}_u) = (\widehat{\Lambda}^{(1)}, \widehat{\Sigma}_u^{(1)})$, *Lemma A.3 is satisfied.*

Proof. The first order condition with respect to $\widehat{\Lambda}^{(1)}$ in (ii) is easy to verify, which is the same as that in Bai and Li (2012). We only show part (i).

By definition, $L_1(\widehat{\Lambda}^{(1)}) \leq L_1(\Lambda_0)$. Also the representation defined in Lemma A.2 yields

$$Q_3(\Lambda, \Sigma_u) + Q_2(\Lambda, \Sigma_u) = L_1(\Lambda) - N^{-1}\mathrm{tr}(S_u(\widehat{\Sigma}_u^{(1)})^{-1}) + N^{-1}\log|\widehat{\Sigma}_u^{(1)}|.$$

Thus

$$Q_2(\widehat{\Lambda}^{(1)}, \widehat{\Sigma}_u^{(1)}) + Q_3(\widehat{\Lambda}^{(1)}, \widehat{\Sigma}_u^{(1)}) \leq Q_2(\Lambda_0, \widehat{\Sigma}_u^{(1)}) + Q_3(\Lambda_0, \widehat{\Sigma}_u^{(1)})$$

Note that Q_2 is always nonnegative and $Q_2(\Lambda_0, \widehat{\Sigma}_u^{(1)}) = 0$. Therefore by Lemma A.2, $0 \leq Q_2(\widehat{\Lambda}^{(1)}, \widehat{\Sigma}_u^{(1)}) = o_p(1)$. Moreover, the matrix in the trace operation of Q_2 is semi-positive

definite, hence

$$\frac{1}{N}\Lambda_0'(\widehat{\Sigma}_u^{(1)})^{-1}\Lambda_0 - (I_r - J)\frac{1}{N}\widehat{\Lambda}^{(1)'}(\widehat{\Sigma}_u^{(1)})^{-1}\widehat{\Lambda}^{(1)}(I_r - J)' = o_p(1). \tag{B.1}$$

It remains to show that $N^{-1}\Lambda_0'((\widehat{\Sigma}_u^{(1)})^{-1} - \Sigma_{u0}^{-1})\Lambda_0 = o_p(1)$, which follows immediately from Theorem B.1 and that $m_N\omega_T^{1-q} = o(1)$.

$\square$

B.1 Proof of Theorem 3.1

B.1.1 Consistency for $\widehat{\Lambda}^{(1)}$

The equality (B.1) implies

$$\frac{1}{N}(\widehat{\Lambda}^{(1)} - \Lambda_0)'(\widehat{\Sigma}_u^{(1)})^{-1}(\widehat{\Lambda}^{(1)} - \Lambda_0) - \frac{1}{N}J\widehat{\Lambda}^{(1)'}(\widehat{\Sigma}_u^{(1)})^{-1}\widehat{\Lambda}^{(1)}J' = o_p(1).$$

The second term is bounded by $N^{-1}\|J\|_F^2\|\widehat{\Lambda}^{(1)}\|_F^2\|(\widehat{\Sigma}_u^{(1)})^{-1}\| = O_p(\|J\|_F^2)$. Lemma A.5 then implies the second term is $o_p(1)$, which then implies that the first term is $o_p(1)$. Because $(\widehat{\Sigma}_u^{(1)})^{-1}$ has eigenvalues bounded away from zero asymptotically, we have $N^{-1}\|\widehat{\Lambda}^{(1)} - \Lambda_0\|_F^2 = o_p(1)$.

B.1.2 Consistency for $\widehat{\lambda}_j^{(1)}$

Lemma A.4 (i) can be equivalently written as: for any $j \leq N$,

$$\widehat{\lambda}_j^{(1)} - \lambda_{0j} = -J'\lambda_{0j} + H\widehat{\Lambda}^{(1)'}(\widehat{\Sigma}_u^{(1)})^{-1}a_j \tag{B.2}$$

where $\widehat{\Sigma}_{u,j}^{(1)}$ denotes the jth column of $\widehat{\Sigma}_u^{(1)}$, and a_j is an $N \times 1$ vector

$$a_j = \Lambda_0 T^{-1}\sum_{t=1}^T f_t u_{jt} + T^{-1}\sum_{t=1}^T (u_t u_{jt} - \widehat{\Sigma}_{u,j}^{(1)}) + T^{-1}\sum_{t=1}^T u_t f_t' \lambda_{0j} - \bar{u}\bar{u}_j.$$

The consistency of $\max_{j\leq N}\|\widehat{\lambda}_j^{(1)} - \lambda_{0j}\|$ follows from Lemma A.5 and the following Lemma B.2.

Lemma B.2. $\max_{j\leq N}\|H\widehat{\Lambda}^{(1)'}(\widehat{\Sigma}_u^{(1)})^{-1}a_j\| = O_p(m_N N^{-1/2}\omega_T^{1-q} + T^{-1/2}(\log N)^{1/2})$.

Proof. By Lemma A.1, uniformly in $j \leq N$,

$$H\widehat{\Lambda}^{(1)\prime}(\widehat{\Sigma}_u^{(1)})^{-1}(\frac{1}{T}\sum_{t=1}^{T} u_t f_t' \lambda_{0j} + \Lambda_0 \frac{1}{T}\sum_{t=1}^{T} f_t u_{jt}) = O_p(\frac{\sqrt{N}}{N}(2\sqrt{N}\sqrt{\frac{\log N}{T}})) = O_p(\sqrt{\frac{\log N}{T}}).$$

$$H\widehat{\Lambda}^{(1)\prime}(\widehat{\Sigma}_u^{(1)})^{-1}(\frac{1}{T}\sum_{t=1}^{T} u_t u_{jt} - \Sigma_{u0,j}) = O_p(\frac{\sqrt{N}}{N}\sqrt{N}\sqrt{\frac{\log N}{T}}) = O_p(\sqrt{\frac{\log N}{T}}).$$

$$H\widehat{\Lambda}^{(1)\prime}(\widehat{\Sigma}_u^{(1)})^{-1}(\widehat{\Sigma}_{u,j}^{(1)} - \Sigma_{u0,j}) = O_p(\frac{\sqrt{N}}{N}m_N\omega_T^{1-q}) = O_p(\frac{m_N}{\sqrt{N}}\omega_T^{1-q}).$$

Finally, $\max_{j \leq N} \|H\widehat{\Lambda}^{(1)}(\widehat{\Sigma}_u^{(1)})^{-1}\bar{u}\bar{u}_j\| = O_p(\log N/T)$. The result then follows from a triangular inequality and that $m_N\omega_T^{1-q} = o(1)$. $\qquad\square$

B.2 Proof of Theorem 3.2

B.2.1 Uniform rate for $\widehat{\lambda}_j^{(1)}$

By (B.2), the uniform rate of convergence follows from Lemma B.2 and the following Lemma B.3.

Lemma B.3. $J = O_p(m_N\omega_T^{1-q})$.

Proof. The first order condition in Lemma A.4 (i) is equivalent to:

$$J'J + J' + J + H\widehat{\Lambda}^{(1)\prime}(\widehat{\Sigma}_u^{(1)})^{-1}B(\widehat{\Sigma}_u^{(1)})^{-1}\widehat{\Lambda}^{(1)}H = 0 \qquad (B.3)$$

where $B = \Lambda_0 T^{-1}\sum_{t=1}^{T} f_t u_t' + (\Lambda_0 T^{-1}\sum_{t=1}^{T} f_t u_t')' + S_u - \widehat{\Sigma}_u^{(1)} - \bar{u}\bar{u}'$. We have, $\|\Lambda_0\|_F = O(\sqrt{N})$, $\bar{u}\bar{u}' = O_p(N\log N/T)$, and $\|S_u - \widehat{\Sigma}_u^{(1)}\| \leq \|\widehat{\Sigma}_u^{(1)} - \Sigma_{u0}\| + \|S_u - \Sigma_{u0}\| = O_p(NT^{-1/2}(\log N)^{1/2} + m_N\omega_T^{1-q})$. Therefore $H\widehat{\Lambda}^{(1)\prime}(\widehat{\Sigma}_u^{(1)})^{-1}B(\widehat{\Sigma}_u^{(1)})^{-1}\widehat{\Lambda}^{(1)}H = O_p(T^{-1/2}(\log N)^{1/2} + m_N N^{-1}\omega_T^{1-q})$. Since $J = o_p(1)$, $J'J$ can be ignored. It follows from (B.3) that

$$J' + J = O_p(\sqrt{\frac{\log N}{T}} + \frac{m_N\omega_T^{1-q}}{N}). \qquad (B.4)$$

Let J_{ij} denote the $(i,j)$the entry of J. It then follows that $J_{ii} = O_p(T^{-1/2}(\log N)^{1/2} + m_N N^{-1}\omega_T^{1-q})$ for all $i \leq r$. It is also not hard to verify that $\sqrt{(\log N)/T} = O(m_N\omega_T^{1-q})$ for any $0 \leq q < 1$ since $m_N \geq 1$.

On the other hand, due to the identification condition, both $\Lambda_0'\Sigma_{u0}^{-1}\Lambda_0$ and $\widehat{\Lambda}^{(1)\prime}(\widehat{\Sigma}_u^{(1)})^{-1}\widehat{\Lambda}^{(1)}$ are diagonal. Let $\text{ndg}(M)$ denote the off-diagonal elements of M. Then

$\mathrm{ndg}(\Lambda_0'\Sigma_{u0}^{-1}\Lambda_0) = \mathrm{ndg}(\widehat{\Lambda}^{(1)'}(\widehat{\Sigma}_u^{(1)})^{-1}\widehat{\Lambda}^{(1)}) = 0$ is equivalent to

$$\mathrm{ndg}\{(\widehat{\Lambda}^{(1)} - \Lambda_0)'(\widehat{\Sigma}_u^{(1)})^{-1}\widehat{\Lambda}^{(1)} + \widehat{\Lambda}^{(1)'}(\widehat{\Sigma}_u^{(1)})^{-1}(\widehat{\Lambda}^{(1)} - \Lambda_0)\}$$

$$= \mathrm{ndg}\{-\Lambda_0'((\widehat{\Sigma}_u^{(1)})^{-1} - \Sigma_{u0}^{-1})\Lambda_0 + (\widehat{\Lambda}^{(1)} - \Lambda_0)'(\widehat{\Sigma}_u^{(1)})^{-1}(\widehat{\Lambda}^{(1)} - \Lambda_0)\}$$

Note that if $\mathrm{ndg}\{M_1\} = \mathrm{ndg}\{M_2\}$ then $\mathrm{ndg}\{HM_1H\} = \mathrm{ndg}\{HM_2H\}$ for two matrices M_1 and M_2 since H is diagonal. Also, $(\widehat{\Lambda}^{(1)} - \Lambda_0)'(\widehat{\Sigma}_u^{(1)})^{-1}\widehat{\Lambda}^{(1)}H = J$. The above identification condition implies

$$\mathrm{ndg}\{HJ + J'H\} = \mathrm{ndg}\{-H\Lambda_0'((\widehat{\Sigma}_u^{(1)})^{-1} - \Sigma_{u0}^{-1})\Lambda_0 H + H(\widehat{\Lambda}^{(1)} - \Lambda_0)'(\widehat{\Sigma}_u^{(1)})^{-1}(\widehat{\Lambda}^{(1)} - \Lambda_0)H\} \tag{B.5}$$

Note that $H\Lambda_0'((\widehat{\Sigma}_u^{(1)})^{-1} - \Sigma_{u0}^{-1})\Lambda_0 H = O_p(m_N N^{-1}\omega_T^{1-q})$. Let h_{ii} denote the ith diagonal entry of H. Let $X = (\widehat{\Lambda}^{(1)} - \Lambda_0)'(\widehat{\Sigma}_u^{(1)})^{-1}(\widehat{\Lambda}^{(1)} - \Lambda_0)$. Then for $i \neq j$, (B.4) and (B.5) imply that

$$J_{ji} + J_{ij} = O_p\left(\sqrt{\frac{\log N}{T}} + \frac{m_N\omega_T^{1-q}}{N}\right),$$

$$h_{ii}J_{ij} + h_{jj}J_{ji} = O_p\left(\frac{m_N\omega_T^{1-q}}{N}\right) + h_{ii}h_{jj}X_{ij}.$$

By assumption, with probability one, there is $\delta > 0$ such that $(N\delta)^{-1} < h_{ii} < N^{-1}\delta$, and $h_{ii} \neq h_{jj}$ for $i \neq j$. Moreover, since all the eigenvalues of $\widehat{\Sigma}_u$ are bounded away from zero and infinity, wpa1, $\|\widehat{\Lambda} - \Lambda_0\|_F^2 \geq c\|X\|_F$ for some $c > 0$. Then the above two equations imply that for any $i \neq j$, $J_{ij} = O_p(m_N\omega_T^{1-q}) + O_p(N^{-1})X_{ji}$ (since $\sqrt{\log N/T} = O(m_N\omega_T^{1-q})$). Then

$$\|J\|_F^2 = O_p\left(m_N^2\omega_T^{2-2q} + \frac{1}{N^2}\|X\|_F^2\right). \tag{B.6}$$

Moreover, by Lemma B.2, $\max_{j \leq N}\|H\widehat{\Lambda}^{(1)}(\widehat{\Sigma}_u^{(1)})^{-1}a_j\| = O_p(m_N\omega_T^{1-q})$.

We now show that $J = O_p(m_N\omega_T^{1-q})$. Suppose this does not hold, then (B.6) implies $J = O_p(N^{-1}X)$. By the definition

$$X = (\widehat{\Lambda}^{(1)} - \Lambda_0)'(\widehat{\Sigma}_u^{(1)})^{-1}(\widehat{\Lambda}^{(1)} - \Lambda_0),$$

$\|X\|_F = O_p(\|\widehat{\Lambda}^{(1)} - \Lambda_0\|_F^2)$. Therefore $J = O_p(N^{-1}X)$ yields $\|J\|_F^2 = O_p(N^{-2}\|\widehat{\Lambda} - \Lambda_0\|_F^4)$. The first order condition (B.2) also yields

$$\max_{j \leq N}\|\widehat{\lambda}_j^{(1)} - \lambda_{0j}\|^2 = O_p(\|J\|_F^2) = O_p(N^{-2}\|\widehat{\Lambda}^{(1)} - \Lambda_0\|_F^4),$$

which implies $\|\widehat{\Lambda}^{(1)} - \Lambda_0\|_F^2 = \sum_{j=1}^{N} \|\widehat{\lambda}_j^{(1)} - \lambda_{0j}\|^2 = O_p(N^{-1}\|\widehat{\Lambda}^{(1)} - \Lambda_0\|_F^4)$. Therefore

$$\frac{1}{N^{-1}\|\widehat{\Lambda}^{(1)} - \Lambda_0\|_F^2} = \frac{\|\widehat{\Lambda}^{(1)} - \Lambda_0\|_F^2}{N^{-1}\|\widehat{\Lambda}^{(1)} - \Lambda_0\|_F^4} = O_p(1),$$

which contradicts with the consistency $N^{-1}\|\widehat{\Lambda}^{(1)} - \Lambda_0\|_F^2 = o_p(1)$. This concludes the proof.

$\square$

Therefore, (B.2) gives $\max_{j \leq N} \|\widehat{\lambda}_j^{(1)} - \lambda_{0j}\| = O_p(\|J\|_F) = O_p(m_N \omega_T^{1-q})$. The rate of convergence for $N^{-1/2}\|\widehat{\Lambda}^{(1)} - \Lambda_0\|_F$ then follows immediately since it is bounded by $\max_{j \leq N} \|\widehat{\lambda}_j^{(1)} - \lambda_{0j}\|$.

B.3 Proof of Theorem 3.3

By the definition of the covariance estimator in the first step, $\widehat{\Sigma}_u^{(1)} = (s_{ij}(R_{ij}))_{N \times N}$, where s_{ij} is a chosen thresholding function. It was shown by Fan et al. (2012, Theorem 2.1) that R_{ij} is the PCA estimator of $T^{-1}\sum_{t=1}^{T} u_{it}u_{jt}$, that is, $R_{ij} = T^{-1}\sum_{t=1}^{T} \widehat{u}_{it}^{PCA}\widehat{u}_{jt}^{PCA}$.

Lemma B.4. *For any $\epsilon > 0$, and any constant $M > 0$, for all large enough N, T,*

$$P(|R_{ij}| > M\tau_{ij}, \forall(i,j) \in S_U) > 1 - \epsilon.$$

Proof. We have, $|R_{ij}| \geq |\Sigma_{u0,ij}| - |\Sigma_{u0,ij} - R_{ij}|$. Thus for all large enough N, T,

$$\begin{aligned} P(|R_{ij}| > M\tau_{ij}, \forall(i,j) \in S_U) &\geq P(|\Sigma_{u0,ij}| > M\tau_{ij} + |\Sigma_{u0,ij} - R_{ij}|, \forall(i,j) \in S_U) \\ &\geq P(|\Sigma_{u0,ij}|/2 > |\Sigma_{u0,ij} - R_{ij}|, \forall(i,j) \in S_U) > 1 - \epsilon, \end{aligned}$$

where in the second and last inequalities we used the assumption that $\omega_T = o(\min_{(i,j)\in S_U} |\Sigma_{u0,ij}|)$ and the fact that $\max_{ij} |\Sigma_{u0,ij} - R_{ij}| = O_p(\omega_T)$.

$\square$

Proof of Theorem 3.3

By Fan et al. (2012), $\max_{i,j} |R_{ij} - \Sigma_{u0,ij}| = O_p(\omega_T)$, which implies for any $\epsilon > 0$, there is $C > 0$ such that $P(\max_{i,j} |R_{ij} - \Sigma_{u0,ij}| > C\omega_T) < \epsilon/2$. For some universal $M > 0$, we set the threshold $\tau_{ij} = M\alpha_{ij}\omega_T$ at entry (i,j), where α_{ij} is a data-dependent value that satisfies, for any $\epsilon > 0$, there is $C_1 > 0$ such that $P(\alpha_{ij} > C_1, \forall i \neq j) > 1 - \epsilon/2$. Then as long as the constant M in the definition of the threshold is larger than $2C/C_1$,

$$P(\max_{i,j} |R_{ij} - \Sigma_{u0,ij}| > \min_{ij} \tau_{ij}/2) < P(\max_{i,j} |R_{ij} - \Sigma_{u0,ij}| > MC_1\omega_T/2) + \epsilon/2 < \epsilon.$$

Note also that if $s_{ij}(R_{ij}) \equiv \widehat{\Sigma}_{ij}^{(1)} \neq 0$, then $|R_{ij}| > \tau_{ij}$, by the definition of s_{ij}. This implies,

$$P(\widehat{\Sigma}_{ij}^{(1)} \neq 0, \exists (i,j) \in S_L) \leq P(|R_{ij}| > \tau_{ij}, \exists (i,j) \in S_L) \leq P(\max_{(i,j) \in S_L} |R_{ij}| > \min_{ij} \tau_{ij})$$

$$\leq P(\max_{i,j} |R_{ij} - \Sigma_{u0,ij}| + \max_{(i,j) \in S_L} |\Sigma_{u0,ij}| > \min_{ij} \tau_{ij}).$$

Since $\max_{(ij) \in S_L} |\Sigma_{u0,ij}| = o(\omega_T)$ by assumption, for all large T, N

$$P(\widehat{\Sigma}_{ij}^{(1)} \neq 0, \exists (i,j) \in S_L) \leq P(\max_{i,j} |R_{ij} - \Sigma_{u0,ij}| > \min_{ij} \tau_{ij}/2) < \epsilon.$$

On the other hand, for arbitrarily small $\epsilon > 0$, $P(\max_{ij} \tau_{ij} \leq K\omega_T) > 1 - \epsilon/2$ for some $K > 0$, which implies $P(|R_{ij}| \geq M\omega_T + K\omega_T, \forall (i,j) \in S_U) \leq P(|R_{ij}| \geq M\omega_T + \tau_{ij}, \forall (i,j) \in S_U) + \epsilon/2$. By the definition of s_{ij}, $|s_{ij}(z) - z| \leq \tau_{ij}$ for all $z \in \mathbb{R}$. Therefore $|R_{ij} - \widehat{\Sigma}_{u,ij}^{(1)}| = |R_{ij} - s_{ij}(R_{ij})| \leq \tau_{ij}$, hence for arbitrarily large $M > 0$,

$$P(|\widehat{\Sigma}_{u,ij}^{(1)}| > M\omega_T, \forall (i,j) \in S_U) \geq P(|R_{ij}| \geq M\omega_T + |R_{ij} - \widehat{\Sigma}_{u,ij}^{(1)}|, \forall (i,j) \in S_U)$$

$$\geq P(|R_{ij}| \geq (M+K)\omega_T, \forall (i,j) \in S_U) - \epsilon/2 \geq 1 - \epsilon$$

where the last inequality follows from Lemma B.4.

B.4 Proof of Theorems 3.4 and 3.5

B.4.1 Proof of Theorem 3.4

A simple derivation implies that $\|N^{-1}\Lambda_0'((\widehat{\Sigma}_u^{(1)})^{-1} - \Sigma_{u0}^{-1})\Lambda_0\|_F \leq N^{-1}\|\Lambda_0\|_F^2 \|(\widehat{\Sigma}_u^{(1)})^{-1} - \Sigma_{u0}^{-1}\| = O_p(m_N\omega_T^{1-q})$. This rate is not tight enough for the $\sqrt{T}$-consistency and limiting distribution $\widehat{\lambda}_j^{(1)}$. A more refined rate of $N^{-1}\Lambda_0'((\widehat{\Sigma}_u^{(1)})^{-1} - \Sigma_{u0}^{-1})\Lambda_0$ depends on the convergence properties of the PCA estimator. We begin by citing some results proved by Fan et al. (2012). Recall that R_{ij} denotes the (i,j)th entry of the orthogonal complement covariance in the sample covariance's spectrum decomposition, and $\widehat{\Sigma}_{u,ij}^{(1)} = s_{ij}(R_{ij})$.

Let $\{\widehat{u}_{it}\}_{i \leq N, t \leq T}$ be the PCA estimates of $\{u_{it}\}_{i \leq N, t \leq T}$. Let $\widehat{\lambda}_j^{PCA}$ and $\widehat{f}_t^{PCA}$ denote the PCA estimators of the factor loadings and factors.

Lemma B.5. (i) For any i, j, with probability one $R_{ij} = T^{-1}\sum_{t=1}^T \widehat{u}_{it}\widehat{u}_{jt}$,
(ii) $\max_{i \leq N} T^{-1}\sum_{t=1}^T (\widehat{u}_{it} - u_{it})^2 = O_p(\omega_T^2)$.
(iii) There is a nonsingular matrix $\bar{H}$ such that $T^{-1}\sum_{t=1}^T \|\widehat{f}_t^{PCA} - \bar{H}f_t\|^2 = O_p(T^{-1} + N^{-1})$ and $\max_j \|\widehat{\lambda}_j^{PCA} - \bar{H}'^{-1}\lambda_{0j}\| = O_p(\omega_T)$.
(iv) $\max_{i,j \leq N} |R_{ij} - \Sigma_{u0,ij}| = O_p(\omega_T)$.

35

Proof. See Theorem 2.1 and Lemma C.11 of Fan et al. (2012). □

Lemma B.6. $\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}u_{it}\lambda'_{0i}\bar{H}^{-1}(\hat{f}_t^{PCA}-\bar{H}f_t)\xi_i\xi'_i = O_p(\frac{1}{\sqrt{NT}}+\frac{1}{T}+\frac{1}{N})$.

Proof. By Bai (2003), there are two $r \times r$ matrices $\bar{H}$ and V, $\|V\|_F = O_p(1)$, $\|\bar{H}\|_F = O_p(1)$ such that $\hat{f}_t^{PCA}-\bar{H}f_t = V(NT)^{-1}\sum_{s=1}^{T}\hat{f}_s^{PCA}[u'_s u_t + f'_s\sum_{j=1}^{N}\lambda_{0j}u_{jt} + f'_t\sum_{j=1}^{N}\lambda_{0j}u_{js}]$. The desired result then follows from the following Lemma B.7.

□

Lemma B.7. *(i)* $\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}u_{it}\lambda'_{0i}\bar{H}^{-1}(NT)^{-1}\sum_{s=1}^{T}\hat{f}_s^{PCA}u'_s u_t\xi_i\xi'_i = O_p(\frac{1}{\sqrt{NT}}+\frac{1}{T}+\frac{1}{N})$
(ii) $\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}u_{it}\lambda'_{0i}\bar{H}^{-1}(NT)^{-1}\sum_{s=1}^{T}\hat{f}_s^{PCA}f'_s\sum_{j=1}^{N}\lambda_{0j}u_{jt}\xi_i\xi'_i = O_p(\frac{1}{\sqrt{NT}}+\frac{1}{N})$
(iii) $\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}u_{it}\lambda'_{0i}\bar{H}^{-1}(NT)^{-1}\sum_{s=1}^{T}\hat{f}_s^{PCA}f'_t\sum_{j=1}^{N}\lambda_{0j}u_{js}\xi_i\xi'_i = O_p(\frac{1}{\sqrt{NT}}+\frac{1}{T})$.

Proof. (i) We have,

$$\|\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}u_{it}\frac{1}{NT}\sum_{s=1}^{T}\hat{f}_s^{PCA'}u'_s u_t\bar{H}^{-1'}\lambda_{0i}\xi_i\xi'_i\| \leq \|\frac{1}{N^2T^2}\sum_{t=1}^{T}\sum_{i=1}^{N}u_{it}\sum_{s=1}^{T}f'_s\bar{H}'u'_s u_t\bar{H}^{-1'}\lambda_{0i}\xi_i\xi'_i\|$$
$$+\|\frac{1}{N^2T^2}\sum_{t=1}^{T}\sum_{i=1}^{N}u_{it}\sum_{s=1}^{T}(\hat{f}_s^{PCA'}-f'_s\bar{H}')u'_s u_t\bar{H}^{-1'}\lambda_{0i}\xi_i\xi'_i\| = a + b. \tag{B.7}$$

We bound a, b separately. Here a is upper bounded by $a_1 + a_2$, where by Cauchy-Schwarz,

$$\begin{aligned}
a_1 &= \|\frac{1}{N^2T^2}\sum_{t=1}^{T}\sum_{i=1}^{N}u_{it}\sum_{s=1}^{T}f'_s\bar{H}'(u'_s u_t - Eu'_s u_t)\bar{H}^{-1'}\lambda_{0i}\xi_i\xi'_i\| \\
&\leq \max_{i\leq N}\|\lambda_{0i}\xi_i\xi'_i\|(\frac{1}{T}\sum_{t=1}^{T}u_{it}^2)^{1/2}\|\frac{1}{N}(\frac{1}{T}\sum_{t=1}^{T}\|\frac{1}{T}\sum_{s=1}^{T}f_s(u'_s u_t - Eu'_s u_t)\|^2)^{1/2} \\
&\leq O_p(1)(\frac{1}{T}\sum_{t=1}^{T}\|\frac{1}{TN}\sum_{s=1}^{T}f_s(u'_s u_t - Eu'_s u_t)\|^2)^{1/2}. \tag{B.8}
\end{aligned}$$

Note that $E\frac{1}{T}\sum_{t=1}^{T}\|\frac{1}{TN}\sum_{s=1}^{T}f_s(u'_s u_t - Eu'_s u_t)\|^2 = E\|\frac{1}{TN}\sum_{s=1}^{T}f_s(u'_s u_t - Eu'_s u_t)\|^2$, which is $O(T^{-1}N^{-1})$ by Assumption 3.9. Hence $a_1 = O_p((NT)^{-1/2})$.

$$a_2 = \|\frac{1}{N^2T^2}\sum_{t=1}^{T}\sum_{i=1}^{N}u_{it}\sum_{s=1}^{T}f'_s\bar{H}'Eu'_s u_t\bar{H}^{-1'}\lambda_{0i}\xi_i\xi'_i\| \leq \max_{i\leq N}\frac{1}{T}\sum_{t=1}^{T}|u_{it}|O(1)\frac{1}{TN}\sum_{s=1}^{T}\|f_s Eu'_s u_t\| \tag{B.9}$$

Since $\max_{t\leq T}E(T^{-1}N^{-1}\sum_{s=1}^{T}\|f_s Eu'_s u_t\|) \leq O(T^{-1})\max_t\sum_{s=1}^{T}|Eu'_s u_t|/N = O(T^{-1})$ by the strong mixing condition (Lemma C.5 of Fan Liao and Mincheva 2012), we have $a_2 = O_p(T^{-1})$. This implies $a = O_p(N^{-1/2}T^{-1/2} + T^{-1})$.

Now we bound b. Using Cauchy Schwarz inequality, we have $b \le b_1 + b_2$ where

$$
\begin{aligned}
b_1 &= \| \frac{1}{N^2 T^2} \sum_{t=1}^{T} \sum_{i=1}^{N} u_{it} \sum_{s=1}^{T} (\widehat{f}_s^{PCA'} - f_s' \bar{H}')(u_s' u_t - E u_s' u_t) \bar{H}^{-1'} \lambda_{0i} \xi_i \xi_i' \| \\
&\le O_p(1) \frac{1}{N^2 T} \sum_{t=1}^{T} \sum_{i=1}^{N} |u_{it}| \left(\frac{1}{T} \sum_{s=1}^{T} \| \widehat{f}_s^{PCA} - \bar{H} f_s \|^2 \right)^{1/2} \left(\frac{1}{T} \sum_{s=1}^{T} |u_s' u_t - E u_s' u_t|^2 \right)^{1/2} \\
&\le O_p(\frac{1}{N}) O_p(\frac{1}{\sqrt{T}} + \frac{1}{\sqrt{N}}) O_p(\sqrt{N}) = O_p(\frac{1}{N} + \frac{1}{\sqrt{NT}}),
\end{aligned}
\tag{B.10}
$$

where the second inequality follows from $E T^{-1} \sum_{s=1}^{T} |u_s' u_t - E u_s' u_t|^2 = O(N)$. Using Cauchy-Schwarz inequality, we also obtain

$$
\begin{aligned}
b_2 &= \| \frac{1}{N^2 T^2} \sum_{t=1}^{T} \sum_{i=1}^{N} u_{it} \sum_{s=1}^{T} (\widehat{f}_s^{PCA'} - f_s' \bar{H}')(E u_s' u_t) \bar{H}^{-1'} \lambda_{0i} \xi_i \xi_i' \| \\
&\le O_p(1)(\frac{1}{T} \sum_{s=1}^{T} \| \widehat{f}_s^{PCA} - \bar{H} f_s \|^2)^{1/2} \left(\frac{1}{T} \sum_{s=1}^{T} |E u_s' u_t / N|^2 \right)^{1/2} \\
&= O_p(\frac{1}{\sqrt{NT}} + \frac{1}{T}).
\end{aligned}
\tag{B.11}
$$

(ii) Let $d_{i,kl}$ be the (k,l)th element of $\xi_i \xi_i'$. Then the (k,l)th element of the object of interest is bounded by $d_1 + d_2$, where, by Cauchy Schwarz inequality,

$$
\begin{aligned}
d_1 &= | \frac{1}{(NT)^2} \sum_{t=1}^{T} \sum_{s=1}^{T} \sum_{j=1}^{N} \sum_{i=1}^{N} (u_{it} u_{jt} - E u_{it} u_{jt}) \lambda_{0i}' \bar{H}^{-1} \widehat{f}_s^{PCA} f_s' \lambda_{0j} d_{i,kl} | \\
&\le O_p(1)(\frac{1}{T} \sum_{s=1}^{T} \| \widehat{f}_s^{PCA} \|^2)^{1/2} (\frac{1}{T} \sum_{s=1}^{T} \| \widehat{f}_s^{PCA} \|^2)^{1/2} \| \frac{1}{N^2 T} \sum_{j=1}^{N} \sum_{i=1}^{N} \sum_{t=1}^{T} (u_{it} u_{jt} - E u_{it} u_{jt}) \lambda_{0i} \lambda_{0j}' d_{i,kl} \| \\
&= O_p(\frac{1}{\sqrt{NT}}).
\end{aligned}
\tag{B.12}
$$

The last equality follows from Assumption 3.9. Also, $\sum_{i,j \le N} |E u_{it} u_{jt}| = \sum_{(i,j) \in S_U} |\Sigma_{u0,ij}| + \sum_{(i,j) \in S_L} |\Sigma_{u0,ij}| = O(N)$. Thus

$$
\begin{aligned}
d_2 &= | \frac{1}{(NT)^2} \sum_{t=1}^{T} \sum_{s=1}^{T} \sum_{j=1}^{N} \sum_{i=1}^{N} (E u_{it} u_{jt}) \lambda_{0i}' \bar{H}^{-1} \widehat{f}_s^{PCA} f_s' \lambda_{0j} d_{i,kl} | \\
&\le O_p(1) \frac{1}{N^2 T} \sum_{s=1}^{T} \| \widehat{f}_s^{PCA} \| \| f_s \| \sum_{i,j \le N} |E u_{it} u_{jt}| = O_p(\frac{1}{N}).
\end{aligned}
\tag{B.13}
$$

37

(iii) The object of interest is bounded by $e_1 + e_2$, where

$$e_1 = \|\frac{1}{N^2T^2}\sum_{s=1}^{T}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{i=1}^{N} u_{it}\lambda_{0i}'\bar{H}^{-1}(\widehat{f}_s^{PCA} - \bar{H}f_s)f_t'\lambda_{0j}u_{js}\xi_i\xi_i'\| = O_p(\frac{1}{\sqrt{NT}} + \frac{1}{T}), \quad (B.14)$$

and we used the fact that $\frac{1}{T}\sum_{t=1}^{T}\|\widehat{f}_t^{PCA} - \bar{H}f_t\|^2 = O_p(T^{-1} + N^{-1})$ from Lemma B.5, and that $N^{-1}\sum_{i=1}^{N}\|\frac{1}{T}\sum_{t=1}^{T}f_t u_{it}\| = O_p(T^{-1/2})$.[1]

$$e_2 = \|\frac{1}{N^2T^2}\sum_{s=1}^{T}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{i=1}^{N} u_{it}\lambda_{0i}'f_s f_t'\lambda_{0j}u_{js}\xi_i\xi_i'\| = O_p(\frac{1}{T}). \quad (B.15)$$

$\square$

Lemma B.8. *For S_U in the partition $\{(i,j) : i, j \leq N\} = S_L \cup S_U$,*

(i) $\frac{1}{NT}\sum_{t=1}^{T}\sum_{i\neq j,(i,j)\in S_U} u_{it}\lambda_{0j}'\bar{H}^{-1}(NT)^{-1}\sum_{s=1}^{T}\widehat{f}_s^{PCA}u_s'u_t\xi_i\xi_j' = O_p(\frac{1}{\sqrt{NT}} + \frac{1}{T} + \frac{1}{N})$

(ii) $\frac{1}{NT}\sum_{t=1}^{T}\sum_{i\neq j,(i,j)\in S_U} u_{it}\lambda_{0j}'\bar{H}^{-1}(NT)^{-1}\sum_{s=1}^{T}\widehat{f}_s^{PCA}f_s'\sum_{v=1}^{N}\lambda_{0v}u_{vt}\xi_i\xi_j' = O_p(\frac{1}{\sqrt{NT}} + \frac{m_N}{N})$

(iii) $\frac{1}{NT}\sum_{t=1}^{T}\sum_{i\neq j,(i,j)\in S_U} u_{it}\lambda_{0j}'\bar{H}^{-1}(NT)^{-1}\sum_{s=1}^{T}\widehat{f}_s^{PCA}f_t'\sum_{v=1}^{N}\lambda_{0v}u_{vs}\xi_i\xi_j' = O_p(\sqrt{\frac{\log N}{NT}} + \frac{\log N}{T})$.

Proof. (i) The term of interest is bounded by $a + b$, where

$$a = \|\frac{1}{N^2T^2}\sum_{t=1}^{T}\sum_{(i,j)\in S_U,i\neq j} u_{it}\sum_{s=1}^{T}f_s'\bar{H}'u_s'u_t\bar{H}^{-1'}\lambda_{0j}\xi_i\xi_j'\|,$$

$$b = \|\frac{1}{N^2T^2}\sum_{t=1}^{T}\sum_{(i,j)\in S_U,i\neq j} u_{it}\sum_{s=1}^{T}(\widehat{f}_s^{PCA'} - f_s'\bar{H}')u_s'u_t\bar{H}^{-1'}\lambda_{0j}\xi_i\xi_j'\|.$$

Here a is upper bounded by $a_1 + a_2$, where

$a_1 = \|\frac{1}{N^2T^2}\sum_{t=1}^{T}\sum_{(i,j)\in S_U,i\neq j} u_{it}\sum_{s=1}^{T}f_s'\bar{H}'(u_s'u_t - Eu_s'u_t)\bar{H}^{-1'}\lambda_{0j}\xi_i\xi_j'\|$, and

$a_2 = \|\frac{1}{N^2T^2}\sum_{t=1}^{T}\sum_{(i,j)\in S_U,i\neq j} u_{it}\sum_{s=1}^{T}f_s'\bar{H}'Eu_s'u_t\bar{H}^{-1'}\lambda_{0j}\xi_i\xi_j'\|$. Note that a_1 and a_2 can be bounded in the same way as (B.8) and (B.9). The only difference is that $N^{-1}\sum_{i=1}^{N}$ is replaced by a double sum $N^{-1}\sum_{(i,j)\in S_U,i\neq j}$. By the assumption, $N^{-1}\sum_{(i,j)\in S_U,i\neq j}1 = O(1)$. The result of the proof is exactly the same, so is omitted. We conclude that $a = O_p(N^{-1/2}T^{-1/2} + T^{-1})$.

On the other hand, $b \leq b_1 + b_2$ where

$b_1 = \|\frac{1}{N^2T^2}\sum_{t=1}^{T}\sum_{(i,j)\in S_U,i\neq j} u_{it}\sum_{s=1}^{T}(\widehat{f}_s^{PCA'} - f_s'\bar{H}')(u_s'u_t - Eu_s'u_t)\bar{H}^{-1'}\lambda_{0j}\xi_i\xi_j'\|$, and

---

[1] We have $(N^{-1}\sum_{i=1}^{N}\|\frac{1}{T}\sum_{t=1}^{T}f_t u_{it}\|)^2 \leq N^{-1}\sum_{i=1}^{N}\|\frac{1}{T}\sum_{t=1}^{T}f_t u_{it}\|^2 = N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{r}(\frac{1}{T}\sum_{t=1}^{T}f_t u_{it})^2$, whose expectation is $N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{r}\text{var}(\frac{1}{T}\sum_{t=1}^{T}f_{jt}u_{it})$. Note that $\text{var}(\frac{1}{T}\sum_{t=1}^{T}f_{jt}u_{it}) = O(T^{-1})$ uniformly in $i \leq N$.

$b_2 = \|\frac{1}{N^2T^2}\sum_{t=1}^{T}\sum_{(i,j)\in S_U, i\neq j} u_{it}\sum_{s=1}^{T}(\widehat{f}_s^{PCA'} - f_s'\bar{H}')(Eu_s'u_t)\bar{H}^{-1'}\lambda_{0j}\xi_i\xi_j'\|$. Using Cauchy-Schwarz inequality and the strong mixing condition, b_1 and b_2 can be also bounded in an exactly the same way of (B.10) and (B.11). We conclude that $b = O_p(N^{-1}+T^{-1}+(NT)^{-1/2})$.

(ii) Let $d_{ij,kl}$ be the (k,l)th element of $\xi_i\xi_j'$. Then the (k,l)th element of the object of interest is bounded by $d_1 + d_2$, where
$d_1 = |\frac{1}{(NT)^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\sum_{(i,j)\in S_U, i\neq j}\sum_{v=1}^{N}(u_{it}u_{vt} - Eu_{it}u_{vt})\lambda_{0j}'\bar{H}^{-1}\widehat{f}_s^{PCA}f_s'\lambda_{0v}d_{ij,kl}|$, and
$d_2 = |\frac{1}{(NT)^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\sum_{(i,j)\in S_U, i\neq j}\sum_{v=1}^{N}(Eu_{it}u_{vt})\lambda_{0j}'\bar{H}^{-1}\widehat{f}_s^{PCA}f_s'\lambda_{0v}d_{ij,kl}|$. Bounding d_1, d_2
is slightly different from (B.12) and (B.13), and we give the detail here. By Cauchy Schwarz inequality,

$$d_1 \leq O_p(1)(\frac{1}{T}\sum_{s=1}^{T}\|\widehat{f}_s^{PCA}\|^2)^{1/2}(\frac{1}{T}\sum_{s=1}^{T}\|f_s\|^2)^{1/2}\|\frac{1}{N^2T}\sum_{i\neq j,(i,j)\in S_U}\sum_{t=1}^{T}\sum_{v=1}^{N}(u_{it}u_{vt}-Eu_{it}u_{vt})\lambda_{0j}\lambda_{0v}'d_{ij,kl}\|$$

which is $O_p((NT)^{-1/2})$ by Assumption 3.9. On the other hand,
$d_2 \leq O_p(N^{-2})\sum_{i\neq j,(i,j)\in S_U}\sum_{k=1}^{N}|\Sigma_{u0,ik}|$. Note that $\|\Sigma_{u0}\|_1 = O(m_N)$, where m_N is as defined in Assumption 3.1. Thus $d_2 = O_p(N^{-1}m_N)$.

(iii) The object of interest is bounded by $e_1 + e_2$, where
$e_1 = \|\frac{1}{N^2T^2}\sum_{s=1}^{T}\sum_{i\neq j,(i,j)\in S_U}\sum_{t=1}^{T}\sum_{v=1}^{N}u_{it}\lambda_{0j}'\bar{H}^{-1}(\widehat{f}_s^{PCA} - \bar{H}f_s)f_t'\lambda_{0v}u_{vs}\xi_i\xi_j'\|$,
$e_1 = \|\frac{1}{N^2T^2}\sum_{s=1}^{T}\sum_{i\neq j,(i,j)\in S_U}\sum_{t=1}^{T}\sum_{v=1}^{N}u_{it}\lambda_{0j}'f_sf_t'\lambda_{0v}u_{vs}\xi_i\xi_j'\|$.

Since $\max_{i\leq N}\|T^{-1}\sum_{t=1}^{T}f_tu_{it}\| = O_p(\sqrt{\log N/T})$, we conclude that $e_1 = O_p(\frac{\sqrt{\log N}}{T} + \frac{\sqrt{\log N}}{\sqrt{NT}})$, and $e_2 = O_p(\frac{\log N}{T})$. $\qquad\square$

From Lemma B.8, immediately we have the following result.

Lemma B.9. $\frac{1}{NT}\sum_{t=1}^{T}\sum_{i\neq j,(i,j)\in S_U}u_{it}\lambda_{0j}'\bar{H}^{-1}(\widehat{f}_t^{PCA} - \bar{H}f_t)\xi_i\xi_j' = O_p(\omega_T^2 + m_N/N)$.

Proof. Note that results (i)(ii)(iii) in Lemma B.8 sum up to $O_p(\omega_T^2 + m_N/N)$. Hence Lemma B.9 follows from the equality $\widehat{f}_t^{PCA} - \bar{H}f_t = V(NT)^{-1}\sum_{s=1}^{T}\widehat{f}_s^{PCA}[u_s'u_t + f_s'\sum_{j=1}^{N}\lambda_{0j}u_{jt} + f_t'\sum_{j=1}^{N}\lambda_{0j}u_{js}]$. $\qquad\square$

The following lemma strengthens the results of Bai (2003) when Σ_{u0} is sparse.

Lemma B.10. *For the PCA estimator,*
(i) $N^{-1}\sum_{i=1}^{N}(R_{ii} - \Sigma_{u0,ii})\xi_i\xi_i' = O_p(\omega_T^2)$.
(ii) $N^{-1}\sum_{i\neq j,(i,j)\in S_U}(R_{ij} - \Sigma_{u0,ij})\xi_i\xi_j' = O_p(\omega_T^2 + m_N/N)$.

Proof. (i) $N^{-1}\sum_{i=1}^{N}(R_{ii} - \Sigma_{u0,ii})\xi_i\xi_i' = \sum_{i=1}^{N}(R_{ii} - S_{u,ii})\xi_i\xi_i'/N + \sum_{i=1}^{N}(S_{u,ii} - \Sigma_{u0,ii})\xi_i\xi_i'/N$.
By Assumption 3.9, $\sum_{i=1}^{N}(S_{u,ii} - \Sigma_{u0,ii})\xi_i\xi_i'/N = \sum_{i=1}^{N}\sum_{t=1}^{T}(u_{it}^2 - \Sigma_{u0,ii})\xi_i\xi_i'/(NT) = $

$O_p(1/\sqrt{NT})$. On the other hand, $\frac{1}{N}\sum_{i=1}^{N}(R_{ii}-S_{u,ii})\xi_i\xi_i'$ is equal to

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\widehat{u}_{it}^2-u_{it}^2)\xi_i\xi_i'=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\widehat{u}_{it}-u_{it})^2\xi_i\xi_i'+\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}u_{it}(\widehat{u}_{it}-u_{it})\xi_i\xi_i'.$$

The first term on the right hand side is $O_p(\omega_T^2)$. We now work on the second term. By Bai (2003), there is a nonsingular matrix $\bar{H}$ such that

$$\widehat{u}_{jt}-u_{jt}=\lambda_{0j}'\bar{H}^{-1}(\widehat{f}_t^{PCA}-\bar{H}f_t)+(\widehat{\lambda}_j^{PCA}-\bar{H}'^{-1}\lambda_{0j})'(\widehat{f}_t^{PCA}-\bar{H}f_t)+(\widehat{\lambda}_j^{PCA}-\bar{H}'^{-1}\lambda_{0j})\bar{H}f_t.$$
(B.16)

By Lemma B.6 $\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}u_{it}\lambda_{0i}'\bar{H}^{-1}(\widehat{f}_t^{PCA}-\bar{H}f_t)\xi_i\xi_i'=O_p(\frac{1}{\sqrt{NT}}+\frac{1}{T}+\frac{1}{N})$. In addition, for each element $d_{i,kl}$ of $\xi_i\xi_i'$,

$$\frac{1}{NT}\sum_{j=1}^{N}\sum_{t=1}^{T}u_{jt}(\widehat{\lambda}_j^{PCA}-\bar{H}'^{-1}\lambda_{0j})\bar{H}f_t d_{j,kl}\leq\frac{1}{N}\sum_{j=1}^{N}\|d_{j,kl}\frac{1}{T}\sum_{t=1}^{T}u_{jt}f_t'\bar{H}'\|\max_j\|\widehat{\lambda}_j^{PCA}-\bar{H}'^{-1}\lambda_{0j}\|,$$

which is $O_p(\omega_T\sqrt{\frac{\log N}{T}})$. Also,

$$\frac{1}{NT}\sum_{j=1}^{N}\sum_{t=1}^{T}u_{jt}(\widehat{\lambda}_j^{PCA}-\bar{H}'^{-1}\lambda_{0j})'(\widehat{f}_t^{PCA}-\bar{H}f_t)d_{j,kl}=\frac{1}{T}\sum_{t=1}^{T}(\widehat{f}_t^{PCA}-\bar{H}f_t)'\frac{1}{N}\sum_{j=1}^{N}u_{jt}(\widehat{\lambda}_j^{PCA}-\bar{H}'^{-1}\lambda_{0j})d_{j,kl}$$

$$\leq\left(\frac{1}{T}\sum_{t=1}^{T}\|\widehat{f}_t^{PCA}-\bar{H}f_t\|^2\max_j\|\widehat{\lambda}_j^{PCA}-\bar{H}'^{-1}\lambda_{0j}\|^2\frac{1}{T}\sum_{t=1}^{T}[\frac{1}{N}\sum_{j=1}^{N}|u_{jt}d_{j,kl}|]^2\right)^{1/2}=O_p(\frac{\omega_T}{\sqrt{T}}+\frac{\omega_T}{\sqrt{N}}).$$

(ii) Since $R_{ij}=T^{-1}\sum_{t=1}^{T}\widehat{u}_{it}\widehat{u}_{jt}$, the term of interest equals

$$\frac{2}{N}\sum_{i\neq j,(i,j)\in S_U}\frac{1}{T}\sum_{t=1}^{T}u_{it}(\widehat{u}_{jt}-u_{jt})\xi_i\xi_j'+\frac{1}{N}\sum_{i\neq j,(i,j)\in S_U}\frac{1}{T}\sum_{t=1}^{T}(\widehat{u}_{it}-u_{it})(\widehat{u}_{jt}-u_{jt})\xi_i\xi_j'$$

$$+\frac{1}{NT}\sum_{i\neq j,(i,j)\in S_U}\sum_{t=1}^{T}(u_{it}u_{jt}-\Sigma_{u0,ij})\xi_i\xi_j'.$$

By Assumption 3.9, the third term is $O_p((NT)^{-1/2})$. By the assumption that $\sum_{i\neq j,(i,j)\in S_U}1=O(N)$ and Cauchy Schwarz inequality, the second term is $O_p(\omega_T^2)$. We now work out the first term. Again we use the equality $\widehat{u}_{jt}-u_{jt}=\lambda_{0j}'\bar{H}^{-1}(\widehat{f}_t^{PCA}-\bar{H}f_t)+(\widehat{\lambda}_j^{PCA}-\bar{H}'^{-1}\lambda_{0j})'(\widehat{f}_t^{PCA}-$

$\bar{H}f_t) + (\widehat{\lambda}_j^{PCA} - \bar{H}'^{-1}\lambda_{0j})'\bar{H}f_t$. Lemma B.9 gives

$$\frac{1}{N}\sum_{i\neq j,(i,j)\in S_U}\frac{1}{T}\sum_{t=1}^{T}u_{jt}\lambda_{0i}'\bar{H}^{-1}(\widehat{f}_t^{PCA} - \bar{H}f_t)\xi_i\xi_j' = O_p(\omega_T^2 + \frac{m_N}{N}).$$

On the other hand, $\frac{2}{N}\sum_{i\neq j,(i,j)\in S_U}\frac{1}{T}\sum_{t=1}^{T}u_{jt}(\widehat{\lambda}_i^{PCA} - \bar{H}^{-1'}\lambda_{0i})'\bar{H}f_t\xi_i\xi_j'$ is bounded by,

$$\max_{i\leq N}\|\xi_i\|^2\frac{1}{N}\sum_{i\neq j,(i,j)\in S_U}\|\frac{1}{T}\sum_{t=1}^{T}u_{jt}f_t'\bar{H}'\|\max_j\|\widehat{\lambda}_j^{PCA} - \bar{H}'^{-1}\lambda_{0j}\| = O_p(\omega_T\sqrt{\frac{\log N}{T}}),$$

since $\max_{i\leq N}\|\xi_i\| = O(1)$. Also, $\frac{1}{N}\sum_{i\neq j,(i,j)\in S_U}\frac{1}{T}\sum_{t=1}^{T}u_{jt}(\widehat{\lambda}_i^{PCA} - \bar{H}^{-1'}\lambda_{0i})'(\widehat{f}_t^{PCA} - \bar{H}f_t)\xi_i\xi_j'$ is bounded by

$$O(1)\max_{i\leq N}\|\hat{b}_i - H^{-1'}\lambda_{0i}\|\left(\frac{1}{T}\sum_{t=1}^{T}\|\hat{f}_t - \bar{H}f_t\|^2\right)^{1/2}\left(\frac{1}{T}\sum_{t=1}^{T}[\frac{1}{N}\sum_{i\neq j,(i,j)\in S_U}|d_{ij,kl}u_{it}|]^2\right)^{1/2}$$

which is $O_p(\frac{\omega_T}{\sqrt{T}} + \frac{\omega_T}{\sqrt{N}})$.

$\square$

Proof of Theorem 3.4 $N^{-1}\Lambda_0'((\widehat{\Sigma}_u^{(1)})^{-1} - \Sigma_{u0}^{-1})\Lambda_0 = O_p(\omega_T^{2-2q}m_N^2)$

Proof. By the triangular inequality, the left-hand-side is bounded by

$$\frac{1}{N}\|\Lambda_0'((\widehat{\Sigma}_u^{(1)})^{-1} - \Sigma_{u0}^{-1})(\Sigma_{u0} - \widehat{\Sigma}_u^{(1)})\Sigma_{u0}^{-1}\Lambda_0\|_F + \frac{1}{N}\|\Lambda_0'\Sigma_{u0}^{-1}(\Sigma_{u0} - \widehat{\Sigma}_u^{(1)})\Sigma_{u0}^{-1}\Lambda_0\|_F.$$

The first term is $O_p(\omega_T^{2-2q}m_N^2)$. We now bound the second term, which is

$$\frac{1}{N}\Xi(\widehat{\Sigma}_u^{(1)} - \Sigma_{u0})\Xi' = \frac{1}{N}\sum_{i=1}^{N}(R_{ii} - \Sigma_{u0,ii})\xi_i\xi_i' + \frac{1}{N}\sum_{i\neq j,(i,j)\in S_U}(\widehat{\Sigma}_{u,ij}^{(1)} - \Sigma_{u0,ij})\xi_i\xi_j'$$
$$+ \frac{1}{N}\sum_{(i,j)\in S_L}(\widehat{\Sigma}_{u,ij}^{(1)} - \Sigma_{u0,ij})\xi_i\xi_j',$$

where $\Xi = \Lambda_0'\Sigma_{u0}^{-1}$. The first term on the right hand side is $O_p(\omega_T^2)$ by Lemma B.10. The third term is dominated by, $O(N^{-1})(\sum_{S_L}|\Sigma_{u0,ij}| + \sum_{S_L}|\widehat{\Sigma}_{u,ij}^{(1)}|) = O(N^{-1}) + O(N^{-1})\sum_{S_L}|\widehat{\Sigma}_{u,ij}^{(1)}|$. By Theorem 3.3, for any $\epsilon > 0$ and any $M > 0$, $P(\frac{1}{N}\sum_{(i,j)\in S_L}|\widehat{\Sigma}_{u,ij}^{(1)}| > M\omega_T^2) \leq P(\exists(i,j) \in$

$S_L, \widehat{\Sigma}^{(1)}_{u,ij} \neq 0) \leq \epsilon$. This implies the third term is $O_p(\omega_T^2)$. The second term equals

$$\frac{1}{N} \sum_{i \neq j, (i,j) \in S_U} (\widehat{\Sigma}^{(1)}_{u,ij} - R_{ij}) \xi_i \xi_j' + \frac{1}{N} \sum_{i \neq j, (i,j) \in S_U} (R_{ij} - \Sigma_{u0,ij}) \xi_i \xi_j'.$$

By Lemma B.10 (ii), $N^{-1} \sum_{i \neq j, (i,j) \in S_U} (R_{ij} - \Sigma_{u0,ij}) \xi_i \xi_j' = O_p(\omega_T^2 + m_N/N)$. On the other hand, recall that $|s_{ij}(z) - z| \leq a\tau_{ij}^2$ when $|z| > b\tau_{ij}$ (Section 3.1),

$$\|\frac{1}{N} \sum_{i \neq j, (i,j) \in S_U} (\widehat{\Sigma}^{(1)}_{u,ij} - R_{ij}) \xi_i \xi_j'\| = \|\frac{1}{N} \sum_{i \neq j, (i,j) \in S_U, |R_{ij}| > b\tau_{ij}} (s_{ij}(R_{ij}) - R_{ij}) \xi_i \xi_j'$$
$$+ \frac{1}{N} \sum_{i \neq j, (i,j) \in S_U, |R_{ij}| > b\tau_{ij}} (s_{ij}(R_{ij}) - R_{ij}) \xi_i \xi_j'\| \leq O_p(\omega_T^2) + \|\frac{1}{N} \sum_{i \neq j, (i,j) \in S_U, |R_{ij}| \leq b\tau_{ij}} (s_{ij}(R_{ij}) - R_{ij}) \xi_i \xi_j'\|.$$

Write $v = \|N^{-1} \sum_{i \neq j, (i,j) \in S_U, |R_{ij}| \leq b\tau_{ij}} (s_{ij}(R_{ij}) - R_{ij}) \xi_i \xi_j'\|$, then for any $C > 0$, and $\epsilon > 0$, Lemma B.4 implies $P(v > M\omega_T^2) \leq P(\exists (i,j) \in S_U, |R_{ij}| \leq b\tau_{ij}) < \epsilon$, which yields $v = O_p(\omega_T^2)$. Therefore $\frac{1}{N} \sum_{i \neq j, (i,j) \in S_U} (\widehat{\Sigma}^{(1)}_{u,ij} - \Sigma_{u0,ij}) \xi_i \xi_j' = O_p(\omega_T^2)$. This implies $N^{-1} \Lambda_0'((\widehat{\Sigma}^{(1)}_u)^{-1} - \Sigma_{u0}^{-1}) \Lambda_0 = O_p(\omega_T^2 + \omega_T^{2-2q} m_N^2 + m_N/N) = O_p(\omega_T^{2-2q} m_N^2)$.

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ □

B.4.2 Convergence rate for J

We now improve the rate in Lemma B.3.

Lemma B.11. *(i)* $H\widehat{\Lambda}^{(1)'}(\widehat{\Sigma}^{(1)}_u)^{-1}[\Lambda_0 \frac{1}{T} \sum_{t=1}^T f_t u_t' + (\Lambda_0 \frac{1}{T} \sum_{t=1}^T f_t u_t')'](\widehat{\Sigma}^{(1)}_u)^{-1} \widehat{\Lambda}^{(1)} H = O_p(m_N T^{-1/2} (\log N)^{1/2} \omega_T^{-q})$.
(ii) $H\widehat{\Lambda}^{(1)'} \widehat{\Sigma}_u^{-1} (S_u - \widehat{\Sigma}^{(1)}_u)(\widehat{\Sigma}^{(1)}_u)^{-1} \widehat{\Lambda}^{(1)} H = O_p(m_N \omega_T^{1-q} T^{-1/2} (\log N)^{1/2} + m_N \omega_T^{1-q} N^{-1})$.

Proof. (i) By Theorem 3.2,

$$\|\widehat{\Lambda}^{(1)} - \Lambda_0\|_F = O_p(\sqrt{N} m_N \omega_T^{1-q}) = \|\widehat{\Lambda}^{(1)'}(\widehat{\Sigma}^{(1)}_u)^{-1} - \Lambda_0' \Sigma_{u0}^{-1}\|_F. \qquad (B.17)$$

Therefore the RHS of part (i) equals

$$H\Lambda_0' \Sigma_{u0}^{-1} \left[\Lambda_0 \frac{1}{T} \sum_{t=1}^T f_t u_t' + (\Lambda_0 \frac{1}{T} \sum_{t=1}^T f_t u_t')' \right] \Sigma_{u0}^{-1} \Lambda_0 H + O_p(m_N \sqrt{\frac{\log N}{T}} \omega_T^{1-q}). \qquad (B.18)$$

Now it follows from Assumption 3.10 that

$$\frac{1}{NT} \sum_{t=1}^T f_t u_t' \Sigma_{u0}^{-1} \Lambda_0 = \frac{1}{NT} \sum_{t=1}^T \sum_{i=1}^N f_t u_{it} \xi_i' = O_p(\frac{1}{\sqrt{NT}}) = O_p(m_N \sqrt{\frac{\log N}{T}} \omega_T^{1-q}),$$

which then yields the desired result.

(ii) Recall that $\|S_u - \widehat{\Sigma}_u^{(1)}\| = O_p(NT^{-1/2}(\log N)^{1/2} + m_N\omega_T^{1-q})$ and that $\|\widehat{\Lambda}^{(1)\prime}(\widehat{\Sigma}_u^{(1)})^{-1} - \Lambda_0'\Sigma_{u0}^{-1}\|_F = O_p(\sqrt{N}m_N\omega_T^{1-q})$. By Theorem B.1, the RHS of (ii) equals

$$H\Lambda_0'\Sigma_{u0}^{-1}(S_u - \Sigma_{u0})\Sigma_{u0}^{-1}\Lambda_0 H + O_p(m_N\omega_T^{1-q}\sqrt{\frac{\log N}{T}} + \frac{m_N\omega_T^{1-q}}{N}).$$

By Assumption 3.10 (note that $H = O_p(N^{-1})$),

$$H\Lambda_0'\Sigma_{u0}^{-1}(S_u - \Sigma_{u0})\Sigma_{u0}^{-1}\Lambda_0 H = \frac{1}{T}H\sum_{i\leq N,j\leq N}\sum_{t\leq T}(u_{it}u_{jt} - Eu_{it}u_{jt})\xi_i\xi_j' H = O_p(\frac{1}{\sqrt{NT}}).$$

$\square$

Lemma B.12. $J = O_p(m_N^2\omega_T^{2-2q})$.

Proof. By (B.3) and Lemma B.11, ignoring the smaller order $J'J$, we have

$$J + J' = O_p(m_N\omega_T^{1-q}\sqrt{\frac{\log N}{T}} + \frac{m_N\omega_T^{1-q}}{N}).$$

This implies that $J_{ii} = O_p(m_N\omega_T^{1-q}(T^{-1/2}(\log N)^{1/2} + N^{-1}))$.

Moreover, since $H(\widehat{\Lambda}^{(1)} - \Lambda_0)'(\widehat{\Sigma}_u^{(1)})^{-1}(\widehat{\Lambda}^{(1)} - \Lambda_0)H = O_p(N^{-1}m_N^2\omega_T^{2-2q})$, (B.5) and Theorem 3.4 imply $\text{ndg}\{HJ + J'H\} = O_p(N^{-1}m_N^2\omega_T^{2-2q})$. Therefore for $i \neq j$, $J_{ij} = O_p(m_N^2\omega_T^{2-2q} + m_N\omega_T^{1-q}(T^{-1/2}(\log N)^{1/2} + N^{-1})) = O_p(m_N^2\omega_T^{2-2q})$. The desired result follows immediately. $\square$

B.4.3 Improved rate for $\widehat{\lambda}_j^{(1)}$

Lemma B.13. *(i)* $H\widehat{\Lambda}^{(1)\prime}(\widehat{\Sigma}_u^{(1)})^{-1}T^{-1}\sum_{t=1}^{T}(u_tu_{jt} - \widehat{\Sigma}_{u,j}^{(1)}) = O_p(m_N\omega_T^{2-q} + m_N^2\omega_T^{2-2q}N^{-1/2})$.
(ii) $H\widehat{\Lambda}^{(1)\prime}(\widehat{\Sigma}_u^{(1)})^{-1}T^{-1}\sum_{t=1}^{T}u_tf_t'\lambda_{0j} = O_p(m_N\omega_T^{1-q}T^{-1/2}(\log N)^{1/2})$.

Proof. (i) We have, $\|\widehat{\Lambda}^{(1)\prime}(\widehat{\Sigma}_u^{(1)})^{-1} - \Lambda_0'\Sigma_{u0}^{-1}\|_F = O_p(\sqrt{N}m_N\omega_T^{1-q})$. Hence $H(\widehat{\Lambda}^{(1)\prime}(\widehat{\Sigma}_u^{(1)})^{-1} - \Lambda_0'\Sigma_{u0}^{-1})T^{-1}\sum_{t=1}^{T}(u_tu_{jt} - \widehat{\Sigma}_{u,j}^{(1)}) = O_p(m_N\omega_T^{1-q}T^{-1/2}(\log N)^{1/2} + N^{-1/2}m_N^2\omega_T^{2-2-q})$. Hence part (i) equals

$$H\Lambda_0'\Sigma_{u0}^{-1}T^{-1}\sum_{t=1}^{T}(u_tu_{jt} - E(u_tu_{jt})) + O_p(m_N\omega_T^{2-q} + \frac{m_N^2\omega_T^{2-2q}}{\sqrt{N}})$$

where $O_p(\cdot)$ is uniform in $j \leq N$. By Assumption 3.10, for each $j \leq N$,

$$H\Lambda_0'\Sigma_{u0}^{-1}T^{-1}\sum_{t=1}^{T}(u_tu_{jt} - E(u_tu_{jt})) = H\frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T}\xi_i(u_{it}u_{jt} - E(u_tu_{jt})) = O_p((NT)^{-1/2}).$$

43

(ii) We have $\|H(\widehat{\Lambda}^{(1)\prime}(\widehat{\Sigma}_u^{(1)})^{-1} - \Lambda_0'\Sigma_{u0}^{-1})T^{-1}\sum_{t=1}^T u_t f_t'\|_F = O_p(m_N\omega_T^{1-q}T^{-1/2}(\log N)^{1/2})$.
Hence (ii) equals

$$H\Lambda_0'\Sigma_{u0}^{-1}\frac{1}{T}\sum_{t=1}^T u_t f_t'\lambda_{0j} + O_p(m_N\omega_T^{1-q}\sqrt{\frac{\log N}{T}}).$$

By Assumption 3.10, the first term equals $NH(NT)^{-1}\sum_{i=1}^N\sum_{t=1}^T \xi_i u_{it} f_t'\lambda_{0j} = O_p((NT)^{-1/2})$, which yields the desired result. $\qquad\square$

Lemma B.14. *For each fixed $j \le N$,*

$$\widehat{\lambda}_j^{(1)} - \lambda_{0j} = H\widehat{\Lambda}^{(1)\prime}(\widehat{\Sigma}_u^{(1)})^{-1}\Lambda_0\frac{1}{T}\sum_{t=1}^T f_t u_{jt} + O_p(m_N^2\omega_T^{2-2q}).$$

Proof. Note that those two terms in Lemma B.13 (i) (ii) are dominated by $O_p(m_N^2\omega_T^{2-2q})$. Therefore, the desired expansion follows from the first order condition (B.2) and Lemma B.12. $\qquad\square$

B.4.4   Proof of Theorem 3.5

By Lemma B.14, and (B.17)

$$
\begin{aligned}
\widehat{\lambda}_j^{(1)} - \lambda_{0j} &= H\widehat{\Lambda}^{(1)\prime}(\widehat{\Sigma}_u^{(1)})^{-1}(\Lambda_0 - \widehat{\Lambda}^{(1)})\frac{1}{T}\sum_{t=1}^T f_t u_{jt} + \frac{1}{T}\sum_{t=1}^T f_t u_{jt} + O_p(m_N^2\omega_T^{2-2q}) \\
&= \frac{1}{T}\sum_{t=1}^T f_t u_{jt} + O_p(m_N^2\omega_T^{2-2q} + m_N\omega_T^{1-q}\sqrt{\frac{\log N}{T}}) = \frac{1}{T}\sum_{t=1}^T f_t u_{jt} + O_p(m_N^2\omega_T^{2-2q}).
\end{aligned}
$$

By the assumption that $m_N^2\omega_T^{2-2q} = o(T^{-1/2})$, we have $\sqrt{T}(\widehat{\lambda}_j^{(1)} - \lambda_{0j}) = T^{-1/2}\sum_{t=1}^T f_t u_{jt} + o_p(1)$. The limiting distribution follows since

$$T^{-1/2}\sum_{t=1}^T f_t u_{jt} \to^d N_r(0, E(u_{jt}f_t f_t')).$$

B.5   Proof of Theorem 3.6

For any $t \le T$, $y_t - \bar{y} = \Lambda_0 f_t + u_t - \bar{u}$. Hence

$$\widehat{f}_t^{(1)} - f_t = -J'f_t + (\widehat{\Lambda}^{(1)\prime}(\widehat{\Sigma}_u^{(1)})^{-1}\widehat{\Lambda}^{(1)})^{-1}\widehat{\Lambda}^{(1)\prime}(\widehat{\Sigma}_u^{(1)})^{-1}(u_t - \bar{u}). \qquad (B.19)$$

B.5.1 Convergence rate

Since both f_t and u_t have exponential tails, using Bonferroni's method we have, $\max_t \|f_t\| = O_p((\log T)^{1/r_2})$ and $\max_t \|u_t\| = O_p(\sqrt{N}(\log T)^{1/r_1})$. Thus by Lemma B.12, $\max_{t \leq T} \|J'f_t\| = O_p(m_N^2 \omega_T^{2-2q}(\log T)^{1/r_2})$. The term with $\bar{u}$ in (B.19) is of smaller order hence is negligible. Also $\|(\widehat{\Lambda}^{(1)\prime}(\widehat{\Sigma}_u^{(1)})^{-1}\widehat{\Lambda}^{(1)})^{-1}(\widehat{\Lambda}^{(1)\prime}(\widehat{\Sigma}_u^{(1)})^{-1} - \Lambda_0'\Sigma_{u0}^{-1})\|_F = O_p(N^{-1/2}m_N\omega_T^{1-q})$, where we used $\|\widehat{\Lambda}^{(1)\prime}(\widehat{\Sigma}_u^{(1)})^{-1} - \Lambda_0'\Sigma_{u0}^{-1}\|_F = O_p(\sqrt{N}m_N\omega_T^{1-q})$. Hence

$$\max_{t \leq T}(\widehat{\Lambda}^{(1)\prime}(\widehat{\Sigma}_u^{(1)})^{-1}\widehat{\Lambda}^{(1)})^{-1}\widehat{\Lambda}^{(1)\prime}(\widehat{\Sigma}_u^{(1)})^{-1}(u_t - \bar{u}) = O_p(\frac{1}{N})\Lambda_0'\Sigma_{u0}^{-1}u_t$$

$$+O_p(m_N^2\omega_T^{2-2q}(\log T)^{1/r_2}+m_N\omega_T^{1-q}(\log T)^{1/r_1}) = O_p(\frac{1}{N})\Lambda_0'\Sigma_{u0}^{-1}u_t+O_p(m_N\omega_T^{1-q}(\log T)^{1/r_1+1/r_2}).$$

Finally, because $E(\frac{1}{N}\Lambda_0'\Sigma_{u0}^{-1}u_t u_t'\Sigma_{u0}^{-1}\Lambda_0) = \frac{1}{N}\Lambda_0'\Sigma_{u0}^{-1}\Lambda_0$, whose eigenvalues are bounded. Hence $\frac{1}{\sqrt{N}}\Lambda_0'\Sigma_{u0}^{-1}u_t = O_p(1)$. Also, $O_p(N^{-1/2})$ is of smaller order than $O_p(m_N\omega_T^{1-q}(\log T)^{1/r_1+1/r_2+1})$. This implies

$$\|\widehat{f}_t^{(1)} - f_t\| = O_p(m_N\omega_T^{1-q}(\log T)^{1/r_1+1/r_2+1}).$$

The above proof also shows that the rate can be made uniform if $\max_{t \leq T} \|\frac{1}{\sqrt{N}}\Lambda_0'\Sigma_{u0}^{-1}u_t\| = O_p(\log T)$.

B.5.2 Asymptotic normality

Recall that $\Xi = \Lambda_0'\Sigma_{u0}^{-1}$ and $\beta_t = \Sigma_{u0}^{-1}u_t$.

Lemma B.15. *For any fixed $t \leq T$, $N^{-1/2}(\widehat{\Lambda}^{(1)} - \Lambda_0)'\Sigma_{u0}^{-1}u_t = o_p(1)$.*

Proof. We expand $\widehat{\Lambda}^{(1)} - \Lambda_0$ using the first order condition

$$(\widehat{\Lambda}^{(1)} - \Lambda_0)' = J\Lambda_0' + H\widehat{\Lambda}^{(1)\prime}(\widehat{\Sigma}_u^{(1)})^{-1}[\Lambda_0\frac{1}{T}\sum_{s=1}^T f_s u_s' + \frac{1}{T}\sum_{s=1}^T u_s f_s'\Lambda_0' + S_u - \widehat{\Sigma}_u^{(1)}] \quad (\text{B.20})$$

and investigate each term separately. First of all, since $J = O_p(m_N^2\omega_T^{2-2q})$, and by assumption that $\Lambda_0'\Sigma_{u0}^{-1}u_t = \sum_{i=1}^N \xi_i u_{it} = O_p(\sqrt{N})$, we have $N^{-1/2}J\Lambda_0'\Sigma_{u0}^{-1}u_t = O_p(m_N^2\omega_T^{2-2q})$. Second, by the assumption that $(TN)^{-1/2}\sum_{s=1}^T f_s u_s'\Sigma_{u0}^{-1}u_t = O_p(1)$, we have

$$\frac{1}{\sqrt{N}}H\widehat{\Lambda}^{(1)\prime}(\widehat{\Sigma}_u^{(1)})^{-1}\Lambda_0\frac{1}{T}\sum_{s=1}^T f_s u_s'\Sigma_{u0}^{-1}u_t = O_p(\frac{1}{\sqrt{T}}).$$

Third, $N^{-1/2}H\widehat{\Lambda}^{(1)\prime}(\widehat{\Sigma}_u^{(1)})^{-1}\frac{1}{T}\sum_{s=1}^T u_s f_s'\Lambda_0'\Sigma_{u0}^{-1}u_t = O_p(\sqrt{\log N/T})$. Moreover,

$N^{-1/2}H(\widehat{\Lambda}^{(1)'}(\widehat{\Sigma}_u^{(1)})^{-1} - \Lambda_0'\Sigma_{u0}^{-1})(S_u - \Sigma_{u0})\Sigma_{u0}^{-1}u_t = O_p(m_N\omega_T^{1-q}\sqrt{N\log N/T}) = o_p(1)$. Therefore, by the assumption that $(NT\sqrt{N})^{-1}\sum_{i=1}^N\sum_{s=1}^T\xi_i(u_{is}u_s' - Eu_{is}u_s')\beta_t = o_p(1)$, we have,

$$\frac{1}{\sqrt{N}}H\widehat{\Lambda}^{(1)'}(\widehat{\Sigma}_u^{(1)})^{-1}(S_u - \Sigma_{u0})\Sigma_{u0}^{-1}u_t = \frac{1}{\sqrt{N}}H\Lambda_0'\Sigma_{u0}^{-1}(S_u - \Sigma_{u0})\Sigma_{u0}^{-1}u_t + o_p(1)$$

$$= \frac{1}{T\sqrt{N}}H\sum_{i=1}^N\sum_{j=1}^N\sum_{s=1}^T\xi_i(u_{is}u_{js} - Eu_{is}u_{js})\beta_{jt} = o_p(1).$$

Finally, $N^{-1/2}H\widehat{\Lambda}^{(1)'}(\widehat{\Sigma}_u^{(1)})^{-1}(\widehat{\Sigma}_u^{(1)} - \Sigma_{u0})\Sigma_{u0}^{-1}u_t = O_p(\frac{1}{\sqrt{N}}m_N\omega_N^{1-q})$. $\qquad\square$

Lemma B.16. *For any fixed $t \leq T$, $N^{-1/2}\Lambda_0'((\widehat{\Sigma}_u^{(1)})^{-1} - \Sigma_{u0}^{-1})u_t = o_p(1)$*

Proof. We note that, $N^{-1/2}\Lambda_0'((\widehat{\Sigma}_u^{(1)})^{-1} - \Sigma_{u0}^{-1})u_t = N^{-1/2}\Xi(\widehat{\Sigma}_u^{(1)} - \Sigma_{u0})\beta_t + O_p(\sqrt{N}m_N^2\omega_T^{2-2q})$. On the other hand,

$$\frac{1}{\sqrt{N}}\Xi(\widehat{\Sigma}_u^{(1)} - \Sigma_{u0})\beta_t = \frac{1}{\sqrt{N}}\sum_{i=1}^N(R_{ii} - \Sigma_{u0,ii})\xi_i\beta_{it} + \frac{1}{\sqrt{N}}\sum_{i\neq j,(i,j)\in S_U}(\widehat{\Sigma}_{u,ij}^{(1)} - \Sigma_{u0,ij})\xi_i\beta_{jt}$$

$$+ \frac{1}{\sqrt{N}}\sum_{(i,j)\in S_L}(\widehat{\Sigma}_{u,ij}^{(1)} - \Sigma_{u0,ij})\xi_i\beta_{jt}.$$

The result of the proof is very similar to that of Lemmas B.10 and Theorem l3.1, based on the expansion (B.10) and Theorem 3.3, hence is omitted. $\qquad\square$

Proof of asymptotic normality

We now fix t, then Lemma B.12 gives $J'f_t = O_p(m_N^2\omega_T^{2-2q})$. Hence $\sqrt{N}J'f_t$ is negligible as $\sqrt{N}m_N^2\omega_T^{2-2q} = o(1)$. Moreover, $(\widehat{\Lambda}^{(1)'}(\widehat{\Sigma}_u^{(1)})^{-1}\widehat{\Lambda}^{(1)})^{-1}\widehat{\Lambda}^{(1)'}(\widehat{\Sigma}_u^{(1)})^{-1}\bar{u}$ is of smaller order of $(\widehat{\Lambda}^{(1)'}(\widehat{\Sigma}_u^{(1)})^{-1}\widehat{\Lambda}^{(1)})^{-1}\widehat{\Lambda}^{(1)'}(\widehat{\Sigma}_u^{(1)})^{-1}u_t$, hence is negligible. Next,

$$\sqrt{N}(\widehat{\Lambda}^{(1)'}(\widehat{\Sigma}_u^{(1)})^{-1}\widehat{\Lambda}^{(1)})^{-1}\widehat{\Lambda}^{(1)'}(\widehat{\Sigma}_u^{(1)})^{-1}u_t = \sqrt{N}(\Lambda_0'\Sigma_{u0}^{-1}\Lambda_0)^{-1}\Lambda_0'\Sigma_{u0}^{-1}u_t$$

$$+ O_p(N^{-1/2})(\widehat{\Lambda}^{(1)'}(\widehat{\Sigma}_u^{(1)})^{-1} - \Lambda_0'\Sigma_{u0}^{-1})u_t + O_p(m_N\omega_T^{1-q})$$

where we used $(\widehat{\Lambda}^{(1)'}(\widehat{\Sigma}_u^{(1)})^{-1}\widehat{\Lambda}^{(1)'})^{-1} - (\Lambda_0'\Sigma_{u0}^{-1}\Lambda_0)^{-1} = O_p(N^{-1}m_N\omega_T^{1-q})$. By Lemmas B.15 and B.16, $N^{-1/2}(\widehat{\Lambda}^{(1)'}(\widehat{\Sigma}_u^{(1)})^{-1} - \Lambda_0'\Sigma_{u0}^{-1})u_t = o_p(1)$. This implies, for each fixed t,

$$\sqrt{N}(\widehat{f}_t^{(1)} - f_t) = \sqrt{N}(\Lambda_0'\Sigma_{u0}^{-1}\Lambda_0)^{-1}\Lambda_0'\Sigma_{u0}^{-1}u_t + O_p(\sqrt{N}m_N^2\omega_T^{2-2q} + m_N\omega_T^{1-q})$$

$$= \sqrt{N}(\Lambda_0'\Sigma_{u0}^{-1}\Lambda_0)^{-1}\Lambda_0'\Sigma_{u0}^{-1}u_t + o_p(1).$$

The asymptotic normality then follows from the fact that

$$N^{-1/2}\Lambda_0'\Sigma_{u0}^{-1}u_t = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\xi_i u_{it} \to^d N(0, Q).$$

C   Proofs of Section 4

C.1   Proof of Theorem 4.1

Define

$$
\begin{aligned}
Q_1(\Sigma_u) &= \frac{1}{N}\log|\Sigma_u| + \frac{1}{N}\text{tr}(S_u\Sigma_u^{-1}) + \frac{\mu_T}{N}\sum_{i\neq j}w_{ij}|\Sigma_{u,ij}| \\
&\quad - \frac{1}{N}\log|\Sigma_{u0}| - \frac{1}{N}\text{tr}(S_u\Sigma_{u0}^{-1}) - \frac{\mu_T}{N}\sum_{i\neq j}w_{ij}|\Sigma_{u0,ij}|,
\end{aligned}
$$

Let $L_c(\Lambda, \Sigma_u) = L_2(\Lambda, \Sigma_u) - N^{-1}\log|\Sigma_{u0}| - N^{-1}\text{tr}(S_u\Sigma_{u0}^{-1}) - N^{-1}\mu_T\sum_{i\neq j}w_{ij}|\Sigma_{u0,ij}|$. Then the minimizer of L_c is the same as that of L_2. This implies $L_c(\widehat{\Lambda}^{(2)}, \widehat{\Sigma}_u^{(2)}) \leq L_c(\Lambda_0, \Sigma_{u0})$. Recall the definitions of $Q_2(\Lambda, \Sigma_u)$ and $Q_3(\Lambda, \Sigma_u)$. Then

$$L_c(\Lambda, \Sigma_u) = Q_1(\Sigma_u) + Q_2(\Lambda, \Sigma_u) + Q_3(\Lambda, \Sigma_u).$$

Lemma C.1. *There is a nonnegative stochastic sequence* $0 \leq d_T = O_p(N^{-1}\log N + T^{-1/2}(\log N)^{1/2})$ *such that* $Q_1(\widehat{\Sigma}_u^{(2)}) \leq d_T$ *with probability one.*

Proof. We have $Q_2(\widehat{\Lambda}^{(2)}, \widehat{\Sigma}_u^{(2)}) \geq 0$. In addition, $Q_2(\Lambda_0, \Sigma_{u0}) = Q_1(\Sigma_{u0}) = 0$. Hence

$$
\begin{aligned}
Q_1(\widehat{\Sigma}_u^{(2)}) &= L_c(\widehat{\Lambda}^{(2)}, \widehat{\Sigma}_u^{(2)}) - Q_2(\widehat{\Lambda}^{(2)}, \widehat{\Sigma}_u^{(2)}) - Q_3(\widehat{\Lambda}^{(2)}, \widehat{\Sigma}_u^{(2)}) \\
&\leq L_c(\widehat{\Lambda}^{(2)}, \widehat{\Sigma}_u^{(2)}) - Q_3(\widehat{\Lambda}^{(2)}, \widehat{\Sigma}_u^{(2)}) \leq L_c(\Lambda_0, \Sigma_{u0}) - Q_3(\widehat{\Lambda}^{(2)}, \widehat{\Sigma}_u^{(2)}) \\
&= Q_3(\Lambda_0, \Sigma_{u0}) - Q_3(\widehat{\Lambda}^{(2)}, \widehat{\Sigma}_u^{(2)}).
\end{aligned}
$$

By the definition of $\Theta_\lambda \times \Gamma$, there is $\delta > 0$ such that $\Theta_\lambda \times \Gamma \subset \Xi_\delta$. The result then holds for $d_T = |Q_3(\Lambda_0, \Sigma_{u0})| + |Q_3(\widehat{\Lambda}^{(2)}, \widehat{\Sigma}_u^{(2)})|$ by Lemma A.2.

$\square$

Throughout, let (recall that $D = \sum_{i\neq j,(i,j)\in S_U} 1$.)

$$\Delta = (\widehat{\Sigma}_u^{(2)})^{-1} - \Sigma_{u0}^{-1}, \quad K_T = \sum_{(i,j)\in S_L}|\Sigma_{u0,ij}|.$$

Lemma C.2. *For all large enough T and N,*

$$NQ_1(\widehat{\Sigma}_u^{(2)}) \geq \frac{1}{2}\mu_T \min_{(i,j)\in S_L} w_{ij} \sum_{(i,j)\in S_L} |\widehat{\Sigma}_{u,ij} - \Sigma_{u0,ij}| + c\|\Delta\|_F^2 - 2\mu_T \max_{(i,j)\in S_L} w_{ij}K_T$$
$$- \left(O_p(\sqrt{\frac{\log N}{T}})\sqrt{N+D} + \mu_T \max_{i\neq j,(i,j)\in S_U} w_{ij}\sqrt{D} \right)\|\Delta\|_F.$$

Proof. Let $\Omega_0 = \Sigma_{u0}^{-1}$, $\widehat{\Omega} = (\widehat{\Sigma}_u^{(2)})^{-1}$. For any Σ_u, let $\Omega = \Sigma_u^{-1}$. Define a function $f(t) = -\log|\Omega_0 + t\Delta| + \mathrm{tr}(S_u(\Omega_0 + t\Delta))$, $t \geq 0$. Then $-\log|\widehat{\Omega}| + \mathrm{tr}(S_u\widehat{\Omega}) = f(1)$; $-\log|\Omega_0| + \mathrm{tr}(S_u\Omega_0) = f(0)$; and

$$NQ_1(\widehat{\Sigma}_u^{(2)}) = f(1) - f(0) + \mu_T \sum_{i\neq j} w_{ij}|\widehat{\Sigma}_{u,ij}| - \mu_T \sum_{i\neq j} w_{ij}|\Sigma_{u0,ij}| \qquad (C.1)$$

By the integral remainder Taylor expansion, $f(1) - f(0) = f'(0) + \int_0^1 (1-t)f''(t)dt$. We now calculate $f'(0)$ and $f''(t)$. Using the matrix differentiation formula, we have, $f'(t) = \mathrm{tr}(S_u\Delta) - \mathrm{tr}((\Omega_0 + t\Delta)^{-1}\Delta)$, which implies,

$$
\begin{aligned}
f'(0) &= \mathrm{tr}((S_u - \Sigma_{u0})(\widehat{\Omega} - \Omega_0)) = \mathrm{tr}(\Omega_0(S_u - \Sigma_{u0})\widehat{\Omega}(\Sigma_{u0} - \widehat{\Sigma}_u^{(2)})) \\
&= \sum_{ij} (\Omega_0(S_u - \Sigma_{u0})\widehat{\Omega})_{ij}(\Sigma_{u0} - \widehat{\Sigma}_u^{(2)})_{ij}.
\end{aligned}
$$

Note that both $\|\Omega_0\|_1$ and $\|\widehat{\Omega}\|_1$ are bounded from above for $\Sigma_{u0}, \widehat{\Sigma}_u^{(2)} \in \Gamma$. By Lemma A.1(ii), $\max_{ij} |(\Omega_0(S_u - \Sigma_{u0})\widehat{\Omega})_{ij}| \leq \max_{ij} |(S_u - \Sigma_{u0})_{ij}|\|\Omega_0\|_1\|\widehat{\Omega}\|_1 = O_p(\sqrt{\log N/T})$. Therefore, $|f'(0)| = O_p(\sqrt{\log N/T}) \sum_{ij} |\Sigma_{u0,ij} - \widehat{\Sigma}_{u,ij}|$. In addition,

$$f''(t) = \mathrm{tr}((\Omega_0 + t\Delta)^{-1}\Delta(\Omega_0 + t\Delta)^{-1}\Delta) = vec(\Delta)(\Omega_0 + t\Delta)^{-1} \otimes (\Omega_0 + t\Delta)^{-1}vec(\Delta),$$

where vec dentoes the vectorization operator and $\otimes$ denotes the Kronecker product. Since both $(\widehat{\Lambda}^{(2)}, \widehat{\Sigma}_u^{(2)})$ and (Λ_0, Σ_{u0}) are inside $\Theta_\lambda \times \Gamma$, $\sup_{0\leq t\leq 1} \lambda_{\max}(t(\widehat{\Sigma}_u^{(2)})^{-1} + (1-t)\Sigma_{u0}^{-1})$ is bounded from above, which then implies $\inf_{0\leq t\leq 1} \lambda_{\min}[(\Omega_0 + t\Delta)^{-1}] = \inf_{0\leq t\leq 1} \lambda_{\max}^{-1}(t(\widehat{\Sigma}_u^{(2)})^{-1} + (1-t)\Sigma_{u0}^{-1})$ is bounded below by a positive constant c. Hence $\inf_{0\leq t\leq 1} f''(t) \geq c\|\Delta\|_F^2$. From (C.1) and $f(1) - f(0) \geq -|f'(0)| + c\|\Delta\|_F^2$, we have

$$
\begin{aligned}
NQ_1(\widehat{\Sigma}_u^{(2)}) &\geq \mu_T \sum_{i\neq j} w_{ij}|\widehat{\Sigma}_{u,ij}| - \mu_T \sum_{i\neq j} w_{ij}|\Sigma_{u0,ij}| + c\|\Delta\|_F^2 - O_p(\sqrt{\frac{\log N}{T}}) \sum_{ij} |\Sigma_{u0,ij} - \widehat{\Sigma}_{u,ij}| \\
&= \mu_T \sum_{(i,j)\in S_L} w_{ij}|\widehat{\Sigma}_{u,ij}| + \mu_T \sum_{i\neq j,(i,j)\in S_U} w_{ij}|\widehat{\Sigma}_{u,ij}| - \mu_T \sum_{i\neq j} w_{ij}|\Sigma_{u0,ij}| + c\|\Delta\|_F^2
\end{aligned}
$$

48

$$-O_p(\sqrt{\frac{\log N}{T}}) \sum_{\Sigma_{u0,ij} \in S_U} |\Sigma_{u0,ij} - \widehat{\Sigma}_{u,ij}| - O_p(\sqrt{\frac{\log N}{T}}) \sum_{(i,j) \in S_L} |\Sigma_{u0,ij} - \widehat{\Sigma}_{u,ij}|.$$

Since $|\widehat{\Sigma}_{u,ij}| \geq |\widehat{\Sigma}_{u,ij} - \Sigma_{u0,ij}| - |\Sigma_{u0,ij}|$, and $\sum_{i \neq j} w_{ij} |\Sigma_{u0,ij}| = \sum_{i \neq j, (i,j) \in S_U} w_{ij} |\Sigma_{u0,ij}| + \sum_{(i,j) \in S_L} w_{ij} |\Sigma_{u0,ij}|$. It follows that

$$
\begin{aligned}
NQ_1(\widehat{\Sigma}_u^{(2)}) &\geq \mu_T \sum_{(i,j) \in S_L} w_{ij} |\widehat{\Sigma}_{u,ij} - \Sigma_{u0,ij}| - O_p(\sqrt{\frac{\log N}{T}}) \sum_{(i,j) \in S_L} |\Sigma_{u0,ij} - \widehat{\Sigma}_{u,ij}| + c\|\Delta\|_F^2 \\
&\quad - \mu_T \sum_{(i,j) \in S_L} w_{ij} |\Sigma_{u0,ij}| - O_p(\sqrt{\frac{\log N}{T}}) \sum_{\Sigma_{u0,ij} \in S_U} |\Sigma_{u0,ij} - \widehat{\Sigma}_{u,ij}| \\
&\quad - \mu_T \sum_{i \neq j, (i,j) \in S_U} w_{ij} [|\Sigma_{u0,ij}| - |\widehat{\Sigma}_{u,ij}|] - \mu_T \sum_{(i,j) \in S_L} w_{ij} |\Sigma_{u0,ij}| \\
&\geq (\mu_T \min_{(i,j) \in S_L} w_{ij} - O_p(\sqrt{\frac{\log N}{T}})) \sum_{(i,j) \in S_L} |\widehat{\Sigma}_{u,ij} - \Sigma_{u0,ij}| + c\|\Delta\|_F^2 \\
&\quad - 2\mu_T \sum_{(i,j) \in S_L} w_{ij} |\Sigma_{u0,ij}| - O_p(\sqrt{\frac{\log N}{T}}) \sum_{\Sigma_{u0,ij} \in S_U} |\Sigma_{u0,ij} - \widehat{\Sigma}_{u,ij}| \\
&\quad - \mu_T \max_{i \neq j, (i,j) \in S_U} w_{ij} \sum_{i \neq j, (i,j) \in S_U} |\Sigma_{u0,ij} - \widehat{\Sigma}_{u,ij}| \\
&\geq \frac{1}{2} \mu_T \min_{(i,j) \in S_L} w_{ij} \sum_{(i,j) \in S_L} |\widehat{\Sigma}_{u,ij} - \Sigma_{u0,ij}| + c\|\Delta\|_F^2 - 2\mu_T \max_{(i,j) \in S_L} w_{ij} K_T \\
&\quad - O_p(\sqrt{\frac{\log N}{T}})\sqrt{N+D}\|\Delta\|_F - \mu_T \max_{i \neq j, (i,j) \in S_U} w_{ij} \|\Delta\|_F \sqrt{D},
\end{aligned}
$$

which implies the desired result. $\qquad\square$

Lemma C.3.

$$
\begin{aligned}
\frac{1}{N}\|\Sigma_u - \widehat{\Sigma}_u^{(2)}\|_F^2 &= O_p\left(\frac{1}{N}\left(\mu_T \max_{(i,j) \in S_L} w_{ij} K_T + \log N + \mu_T^2 \max_{i \neq j, (i,j) \in S_U} w_{ij}^2 D\right)\right) \\
&\quad + O_p(\frac{D \log N}{NT} + \sqrt{\frac{\log N}{T}}).
\end{aligned}
$$

Proof. Lemma C.2 implies

$$NQ_1(\widehat{\Sigma}_u^{(2)}) \geq c\|\Delta\|_F^2 - 2\mu_T \max_{(i,j) \in S_L} w_{ij} K_T - \left(O_p(\sqrt{\frac{\log N}{T}})\sqrt{N+D} + \mu_T \max_{i \neq j, (i,j) \in S_U} w_{ij}\sqrt{D}\right)\|\Delta\|_F.$$

Lemma C.1 gives $NQ_1(\widehat{\Sigma}_u^{(2)}) \le O_p(\log N + N\sqrt{\log N/T})$. Hence we have

$$
\begin{aligned}
\|\Delta\|_F^2 &= O_p((\sqrt{\frac{(N+D)\log N}{T}} + \mu_T \max_{i \ne j,(i,j)\in S_U} w_{ij}\sqrt{D})^2) \\
&\quad + O_p(\mu_T \max_{(i,j)\in S_L} w_{ij}K_T + \log N + N\sqrt{\log N/T}) \\
&= O_p(\frac{(N+D)\log N}{T} + \mu_T^2 \max_{i \ne j,(i,j)\in S_U} w_{ij}^2 D + \mu_T \max_{(i,j)\in S_L} w_{ij}K_T + \log N + N\sqrt{\log N/T}) \\
&= O_p(\frac{D\log N}{T} + \mu_T^2 \max_{i \ne j,(i,j)\in S_U} w_{ij}^2 D + \mu_T \max_{(i,j)\in S_L} w_{ij}K_T + \log N + N\sqrt{\log N/T}).
\end{aligned}
$$

Note that $\Sigma_{u0} - \widehat{\Sigma}_u^{(2)} = \widehat{\Sigma}_u^{(2)}\Delta\Sigma_{u0}$. Hence the desired result follows from $\|\widehat{\Sigma}_u^{(2)}\| < M$ wp1 and $\|\Sigma_{u0}\| < M$.

$\square$

Lemma C.4. $N^{-1}\sum_{(i,j)\in S_L} |\widehat{\Sigma}_{u,ij} - \Sigma_{u0,ij}| = o_p(1)$.

Proof. Lemma C.2 implies

$$
\begin{aligned}
\frac{1}{2}\mu_T \min_{(i,j)\in S_L} w_{ij} \sum_{(i,j)\in S_L} |\widehat{\Sigma}_{u,ij} - \Sigma_{u0,ij}| &\le NQ_1(\widehat{\Sigma}_u^{(2)}) + 2\mu_T \max_{(i,j)\in S_L} w_{ij}K_T \\
&\quad + \left(O_p(\sqrt{\frac{\log N}{T}})\sqrt{N+D} + \mu_T \max_{i \ne j,(i,j)\in S_U} w_{ij}\sqrt{D} \right) \|\Delta\|_F.
\end{aligned}
$$

We have $NQ_1(\widehat{\Sigma}_u^{(2)}) \le O_p(\log N + N\sqrt{\log N/T})$. By Lemma C.3,

$$
\begin{aligned}
\|\Delta\|_F &= O_p(\sqrt{\frac{D\log N}{T}} + \mu_T \max_{i \ne j,(i,j)\in S_U} w_{ij}\sqrt{D}) \\
&\quad + O_p(\sqrt{\mu_T \max_{(i,j)\in S_L} w_{ij}K_T} + \sqrt{\log N} + \sqrt{N}(\frac{\log N}{T})^{1/4}).
\end{aligned}
$$

which implies the desired result under Assumption 4.2.

$\square$

Lemma C.5. $N^{-1}\Lambda_0'((\widehat{\Sigma}_u^{(2)})^{-1} - \Sigma_{u0}^{-1})\Lambda_0 = o_p(1)$.

Proof. Let $\Delta_1 = \widehat{\Sigma}_u^{(2)} - \Sigma_{u0}$, $\Xi = \Lambda_0'\Sigma_{u0}^{-1} = (\xi_1, ..., \xi_N)$, and $\widehat{V} = (\widehat{\Sigma}_u^{(2)})^{-1}\Lambda_0$. Since the l_1 norms of $(\widehat{\Sigma}_u^{(2)})^{-1}$ and Σ_{u0}^{-1} are bounded away from infinity, we have, $\sup_{i\le N} \|\widehat{V}_i\| = O_p(1)$ and $\sup_{i\le N} \|\xi_i\| = O(1)$. Then

$$
\frac{1}{N}\Lambda_0'(\Sigma_{u0}^{-1} - (\widehat{\Sigma}_u^{(2)})^{-1})\Lambda_0 = \frac{1}{N}\Xi\Delta_1\widehat{V} = \frac{1}{N}\sum_{(i,j)\in S_L} \xi_i\widehat{V}_j'\Delta_{1,ij} + \frac{1}{N}\sum_{\Sigma_{u0,ij}\in S_U} \xi_i\widehat{V}_j'\Delta_{1,ij}
$$

50

$$\leq \quad O_p(\frac{1}{N}) \sum_{(i,j)\in S_L} |\Delta_{1,ij}| + O_p(\frac{1}{N}) \sum_{\Sigma_{u0,ij}\in S_U} |\Delta_{1,ij}|.$$

The first term on the right hand side is $o_p(1)$ by Lemma C.4, and the second is bounded by $N^{-1}\|\widehat{\Sigma}_u^{(2)} - \Sigma_{u0}\|\sqrt{N+D}$ (using Cauchy-Schwarz inequality), which is also $o_p(1)$ by Lemma C.3 and Assumption 4.2.

<div style="text-align:right">□</div>

Lemma C.6. *For* $(\widehat{\Lambda}, \widehat{\Sigma}) = (\widehat{\Lambda}^{(2)}, \widehat{\Sigma}_u^{(2)})$, *Lemma A.3 is satisfied.*

Proof. We first show part (i) of Lemma A.3. Since $L_c(\widehat{\Lambda}^{(2)}, \widehat{\Sigma}_u^{(2)}) \leq L_c(\Lambda_0, \Sigma_{u0})$, and $Q_1(\Sigma_{u0}) = Q_2(\Lambda_0, \Sigma_{u0}) = 0$, there is a nonnegative sequence $d_n = O_p(N^{-1}\log N + T^{-1/2}(\log N)^{1/2})$ such that $Q_1(\widehat{\Sigma}_u^{(2)}) + Q_2(\widehat{\Lambda}^{(2)}, \widehat{\Sigma}_u^{(2)}) \leq d_n$. Lemma C.2 then implies $0 \leq Q_2(\widehat{\Sigma}_u^{(2)}, \widehat{\Lambda}^{(2)}) = o_p(1)$. On the other hand,

$$Q_2(\widehat{\Sigma}_u^{(2)}, \widehat{\Lambda}^{(2)}) = \frac{1}{N}\text{tr}\left[\Lambda_0'(\widehat{\Sigma}_u^{(2)})^{-1}\Lambda_0 - \Lambda_0'(\widehat{\Sigma}_u^{(2)})^{-1}\widehat{\Lambda}^{(2)}(\widehat{\Lambda}'(\widehat{\Sigma}_u^{(2)})^{-1}\widehat{\Lambda}^{(2)})^{-1}\widehat{\Lambda}'(\widehat{\Sigma}_u^{(2)})^{-1}\Lambda_0\right].$$

The matrix in the bracket is semi-positive definite. Hence

$$\frac{1}{N}\Lambda_0'(\widehat{\Sigma}_u^{(2)})^{-1}\Lambda_0 - (I_r - J)\frac{1}{N}\widehat{\Lambda}'(\widehat{\Sigma}_u^{(2)})^{-1}\widehat{\Lambda}^{(2)}(I_r - J)' = o_p(1). \tag{C.2}$$

Finally, the desired result follows from Lemma C.5.

The first order condition in part (ii) is easy to derive and is the same as that in Bai and Li (2012).

<div style="text-align:right">□</div>

Proof of Theorem 4.1

$N^{-1}\|\widehat{\Sigma}_u^{(2)} - \Sigma_{u0}\|_F^2 = o_p(1)$ follows from Lemma C.3 and Assumption 4.2. On the other hand, equation (C.2) also implies

$$\frac{1}{N}(\widehat{\Lambda}^{(2)} - \Lambda_0)'\widehat{\Sigma}_u^{-1}(\widehat{\Lambda}^{(2)} - \Lambda_0) - J\frac{1}{N}H^{-1}J' = o_p(1).$$

By Lemma A.5, $N^{-1}JH^{-1}J' = o_p(1)$. Hence $N^{-1}(\widehat{\Lambda}^{(2)} - \Lambda_0)'\widehat{\Sigma}_u^{-1}(\widehat{\Lambda} - \Lambda_0) = o_p(1)$, which implies the consistency $N^{-1}\|\widehat{\Lambda} - \Lambda_0\|^2 = o_p(1)$ because the eigenvalues of $\widehat{\Sigma}_u^{-1}$ are bounded away from zero. Q.E.D.

To prove the consistency of $\widehat{f}_t^{(2)}$, we note that the expansion (B.19) still holds for $\widehat{f}_t^{(2)}$. Since $J = o_p(1)$ by Lemma A.5, and $\bar{u}$ is of smaller order than u_t for each fixed t. Hence $\widehat{f}_t^{(2)} - f_t = O_p(N^{-1})\widehat{\Lambda}^{(2)'}(\widehat{\Sigma}_u^{(2)})^{-1}u_t + o_p(1)$. Moreover, since $\|(\widehat{\Sigma}_u^{(2)})^{-1}\|$ and $\|\widehat{\Sigma}_u^{(2)}\|$ are both $O_p(1)$ and $\|\widehat{\Lambda}^{(2)}\|_F = O_p(\sqrt{N})$ by the restriction of the parameter space $\Theta_\lambda \times \Gamma$, we have

$N^{-1}\|\widehat{\Lambda}^{(2)\prime}(\widehat{\Sigma}_u^{(2)})^{-1} - \Lambda_0\Sigma_{u0}^{-1}\|_F = O_p(N^{-1/2}\|\widehat{\Lambda}^{(2)} - \Lambda_0\|_F + N^{-1/2}\|\widehat{\Sigma}_u^{(2)} - \Sigma_{u0}\|_F)$, which is $o_p(1)$ as proved above. Therefore, since $N^{-1}\Lambda_0'\Sigma_{u0}^{-1}u_t = N^{-1}\sum_{i=1}^N \xi_i u_{it} = O_p(N^{-1/2})$,

$$\widehat{f}_t^{(2)} - f_t = O_p(N^{-1})\Lambda_0'\Sigma_{u0}^{-1}u_t + o_p(1) = o_p(1).$$

C.2   Proof of Theorem 4.2

We now verify Assumption 4.2 for the Adaptive Lasso.

Lemma C.7. *For adaptive lasso,*
(i) $\min_{i\neq j,(i,j)\in S_U} |\Sigma_{u0,ij}|^\gamma \max_{i\neq j,(i,j)\in S_U} w_{ij} = O_p(1)$.
(ii) $\delta_T^\gamma \max_{(i,j)\in S_L} w_{ij} = O_p(1)$,
(iii) $\omega_T^{-\gamma}(\min_{(i,j)\in S_L} w_{ij})^{-1} = O_p(1)$ (recall that $\omega_T = N^{-1/2} + T^{-1/2}(\log N)$).

Proof. By Lemma B.5 $\max_{i\leq N, j\leq N}|\widehat{\Sigma}_{u,ij}^* - \Sigma_{u0,ij}| = O_p(\tau)$. Given this result and the assumption that $\min_{(i,j)\in S_U} |\Sigma_{u0,ij}| \gg \omega_T$, we have result (i). For any $(i,j) \in S_L$, the following inequality holds: $\delta_T^{-\gamma} \leq w_{ij}^{-1} \leq (|\Sigma_{u0,ij}| + |\Sigma_{u0,ij} - \widehat{\Sigma}_{u,ij}| + \delta_T)^\gamma$, which then implies results (ii) and (iii), due to the assumptions that $\delta_T = o(\omega_T)$, and $\Sigma_{u0,ij} = O(\omega_T)$. □

Proof of Assumption 4.2 for Adaptive Lasso
It follows from the previous lemma that $\alpha_T = O_p(\omega_T^\gamma(\min_{i\neq j, \Sigma_{u0,ij}\in S_U} |\Sigma_{u0,ij}|)^{-\gamma}) = o_p(1)$, and $\beta_T = O_p((\omega_T/\delta_T)^\gamma)$. By the assumption that $D = O(N)$,

$$\zeta = \min\left\{\sqrt{\frac{T}{\log N}}\frac{N}{D}, \left(\frac{T}{\log N}\right)^{1/4}\sqrt{\frac{N}{D}}, \frac{N}{\sqrt{D\log N}}\right\} \gg \min\left\{\left(\frac{T}{\log N}\right)^{1/4}, \sqrt{\frac{N}{\log N}}\right\}.$$

Hence $\alpha_T = O_p(\zeta)$. This together with the lower bound assumption on δ_T yields Assumption 4.2 (i).

For part (ii), note that $\alpha_T = o_p(1)$ implies that with probability approaching one,

$$\min\{N, \frac{N^2}{D}, \frac{N^2}{D}\alpha_T^{-2}\} = N, \quad \min\{\frac{N}{D}, \sqrt{\frac{N}{D}}, \frac{N}{D}\alpha_T^{-1}\} = \sqrt{\frac{N}{D}}.$$

By Lemma C.7(ii), (recall that $K_T = \sum_{(i,j)\in S_L} |\Sigma_{u0,ij}|$) and the lower bound $\delta_T \gg \omega_T(K_T/N)^{1/\gamma}$, $\mu_T \max_{(i,j)\in S_L} w_{ij} K_T = O_p(\mu_T \delta_T^{-\gamma} K_T) = o_p(N)$.

By Lemma C.7(i) and the assumptions that $D = O(N)$ and $\min_{i\neq j,(i,j)\in S_U} |\Sigma_{u0,ij}| \gg \omega_T$, we have $\mu_T \max_{i\neq j,(i,j)\in S_U} w_{ij} = O_p\mu_T(\min_{i\neq j,(i,j)\in S_U} |\Sigma_{u0,ij}|^\gamma)^{-1} = o_p(\sqrt{N/D})$, due to the upper bound on $\mu_T = o(\omega_T^\gamma)$. Finally, by Lemma C.7(iii) and the assumption that $\mu_T \gg \omega_T^{1+\gamma}$, we have $\mu_T \min_{(i,j)\in S_L} w_{ij} \gg \omega_T$.

Proof of Assumption 4.2 for SCAD

Since $\mu_T / \min_{i \neq j, (i,j) \in S_U} |R_{ij}| = o_p(1)$ and $\max_{(i,j) \in S_L} |R_{ij}| = o_p(\mu_T)$, it is easy to verify that with probability approaching one, $\max_{i \neq j, (i,j) \in S_U} w_{ij} = 0$, $\min_{(i,j) \in S_L} w_{ij} = \max_{(i,j) \in S_L} w_{ij} = \mu_T$. Hence $\alpha_T = 0$ and $\beta_T = 1$. This immediately implies the desired result.

References

ALESSI, L., BARIGOZZI, M. and CAPASSOC, M. (2010). Improved penalization for determining the number of factors in approximate factor models. *Statistics and Probability Letters*, **80**, 1806-1813.

BAI, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*. **71** 135-171.

BAI, J. and LI, K. (2012). Statistical analysis of factor models of high dimension. *Ann. Statist.* **40**, 436-465.

BAI, J. and NG, S.(2002). Determining the number of factors in approximate factor models. *Econometrica*. **70** 191-221.

BICKEL, P. and LEVINA, E. (2008a). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577-2604.

BICKEL, P. and LEVINA, E. (2008b). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199-227.

BIEN, J. and TIBSHIRANI, R. (2011). Sparse estimation of a covariance matrix. *Biometrika*. **98**, 807-820.

BREITUNG, J. and TENHOFEN, J. (2011). GLS estimation of dynamic factor models. *J. Amer. Statist. Assoc.* **106**, 11501166.

CAI, T. and LIU, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.* **106**, 672-684.

CAI, T. and YUAN, M. (2012). Adaptive covariance matrix estimation through block thresholding. Forthcoming in *Ann. Statist.*

CAI, T. and ZHOU, H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. Forthcoming in *Ann. Statist.*

CANER, M. and FAN, M. (2011). A near minimax risk bound: adaptive lasso with heteroskedastic data in instrumental variable selection. *Manuscript*. North Carolina State University.

CHAMBERLAIN, G. and ROTHSCHILD, M. (1983). Arbitrage, factor structure and mean-variance analyssi in large asset markets. *Econometrica*. **51** 1305-1324.

CHOI, I. (2012). Efficient estimation of factor models. *Econometric Theory.* **28** 274-308.

DENG, X. and TSUI, K. (2010). Penalized covariance matrix estimation using a matrix-logarithm transformation. *Manuscript*, University of Wisconsin-Madison

DIAS, F., PINHERIO, M. and RUA, A. (2008). Determining the number of factors in approximate factor models with global and group-specific factors. *Manuscript.* Technical University of Lisbon.

FAN, J., FAN, Y. and LV, J. (2008). High dimensional covariance matrix estimation using a factor model. *J. Econometrics.* **147**, 186-197.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348-1360

FAN, J., LIAO, Y. and MINCHEVA, M. (2012). Large covariance estimation by thresholding principal orthogonal complements. *Manuscript.* Princeton University.

FORNI, M., HALLIN, M., LIPPI, M. and REICHLIN, L. (2000). The generalized dynamic factor model: identification and estimation. *Review of Economics and Statistics.* **82** 540-554.

FORNI, M. and LIPPI, M. (2001). The generalized dynamic factor model: representation theory. *Econometric Theory.* **17** 1113-1141.

VAN DE GEER, S., BÜHLMANN, P. and ZHOU, S. (2011). The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). *Electronic Journal of Statistics.* **5**, 688-749.

HALLIN, M. and LIŠKA, R. (2007). Determining the number of factors in the general dynamic factor model. *J. Amer. Statist. Assoc.* **102**, 603-617.

HAN, X. (2012). Determining the number of factors with potentially strong cross-sectional correlation in idiosyncratic shocks. *Manuscript.* North Carolina State University

HUANG, J., MA, S. and ZHANG, C. (2006). Adaptive lasso for sparse high-dimensional regression models. *Manuscript.* University of Iowa.

JOHNSTONE, I.M. and LU, A.Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Jour. Ameri. Statist. Assoc.*, **104**, 682-693.

JUNG, S. and MARRON, J.S. (2009). PCA consistency in high dimension, low sample size context. *Ann. Statist.*, **37**, 4104-4130.

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

EL KAROUI, N. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Ann. Statist.* **36** 2757-2790.

LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254-4278.

LAM, C. and YAO, Q. (2012). Factor modelling for high-dimensional time series: inference for the number of factors. Forthcoming in *Ann. Statist.*

LAWLEY, D. and MAXWELL, A. (1971). *Factor analysis as a statistical method.* Second ed. London, Butterworths.

LEDOIT, O. and WOLF, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Ann. Statist.* **40**, 1024-1060.

ONATSKI, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics.* **92**, 1004-1016.

PATI, D., BHATTACHARYA, A., PILLAI, N. and DUNSON, D. (2012) Posterior contraction in sparse Bayesian factor models for massive covariance matrices. *Manuscript*, Duke University

RAVIKUMAR, P., WAINWRIGHT, M., RASKUTTI, G. and YU, B. (2011), High-dimensional covariance estimation by minimizing l_1-penalized log-determinant divergence, *Electronic Journal of Statistics.* **5** 935-980.

ROHDE, A. and TSYBAKOV, A. (2011), Estimation of high-dimensional low-rank matrices. *Ann. Statist.* **39** 887-930.

ROTHMAN, A., LEVINA, E. and ZHU, J. (2009). Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.* **104** 177-186.

STOCK, J. and WATSON, M. (1998). Diffusion Indexes, NBER Working Paper 6702.

STOCK, J. and WATSON, M. (2002). Forecasting using principal components from a large number of predictors. *J. Amer. Statist. Assoc.* **97**, 1167-1179.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society,* Ser. B, **58** 267-288

WANG, P. (2009). Large dimensional factor models with a multi-level factor structure: identification, estimation and inference. *Manuscript.* Hong Kong University of Science and Technology.

WITTEN, D.M., TIBSHIRANI, R. and HASTIE, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, **10**, 515-534.

XUE, L. and ZOU, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. Forthcoming in *Ann. Statist.*

YUAN, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Machine Learning Research.* **2010**, 2261-2286

ZHANG, C. (2010). Nearly unbiased variable selection under minimax concave penalty, *Ann. Statist.*, **38** 894-942

ZHOU, S., RÜTIMANN, P., XU, M. and BÜHLMANN, P. (2011), High-dimensional covariance estimation based on Gaussian graphical models. *Journal of Machine Learning Research.* **12**, 2975-3026.

ZOU, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418-1429