Journal of Econometrics 200 (2017) 59–78

Contents lists available at ScienceDirect

Journal of Econometrics

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

Inferences in panel data with interactive effects using large covariance matrices

Jushan Bai[a,c,*], Yuan Liao[b]

[a] Department of Economics, Columbia University, New York, NY 10027, United States
[b] Department of Economics, Rutgers University, New Brunswick, NJ 08901, United States
[c] School of Finance, Nankai University, Tianjin, China

ARTICLE INFO

Article history:
Received 12 March 2016
Received in revised form 30 January 2017
Accepted 2 May 2017
Available online 24 June 2017

Keywords:
High dimensionality
Unknown factors
Conditional sparsity
Thresholding
Cross-sectional correlation
Heteroskedasticity
Optimal weight matrix

ABSTRACT

We consider efficient estimation of panel data models with interactive effects, which relies on a high-dimensional inverse covariance matrix estimator. By using a consistent estimator of the error covariance matrix, we can take into account both cross-sectional correlations and heteroskedasticity. In the presence of cross-sectional correlations, the proposed estimator eliminates the cross-sectional correlation bias, and is more efficient than the existing methods. The rate of convergence is also improved. In addition, we find that when the statistical inference involves estimating a high-dimensional inverse covariance matrix, the minimax convergence rate on large covariance estimations is not sufficient for inferences. To address this issue, a new "doubly weighted convergence" result is developed. The proposed method is applied to the US divorce rate data. We find that our more efficient estimator identifies the significant effects of divorce-law reforms on the divorce rate, and provides tighter confidence intervals than existing methods. This provides a confirmation for the empirical findings of Wolfers (2006) under more general unobserved heterogeneity.

1. Introduction

This paper studies a panel data model with an interactive effect:

$$y_{it} = X_{it}'\beta_0 + \varepsilon_{it}, \quad \varepsilon_{it} = \lambda_i'f_t + u_{it}, \quad i \le N, t \le T, \tag{1.1}$$

where X_{it} is a $d \times 1$ vector of regressors; β_0 is a $d \times 1$ vector of unknown coefficients; f_t is an $r \times 1$ vector of common factors, λ_i is a vector of factor loadings, and u_{it} represents the error term, often known as the *idiosyncratic component*. If we denote $Y_t = (y_{1t}, \ldots, y_{Nt})'$, $\Lambda = (\lambda_1, \ldots, \lambda_N)'$, $X_t = (X_{1t}, \ldots, X_{Nt})'$, $(N \times d)$, and $u_t = (u_{1t}, \ldots, u_{Nt})'$, model (1.1) can be written as

$$Y_t = \Lambda f_t + X_t\beta_0 + u_t, \quad t \le T.$$

In the model, the only observables are (y_{it}, X_{it}). The goal is to estimate the parameter β_0, whose dimension is fixed. In this model, the factor component $\lambda_i'f_t$ is regarded as an *interactive effect* of the individual and time effects. Because the regressors and factors can be correlated, simply regressing y_{it} on X_{it} is not consistent.

We focus on the *efficient estimation* of β_0, using an estimated high-dimensional covariance matrix. We propose to estimate β_0 via:

$$\widehat{\beta}(W) = \arg\min_{\beta}\min_{\Lambda,f_t}\sum_{t=1}^{T}(Y_t - X_t'\beta - \Lambda f_t)' \times W(Y_t - X_t'\beta - \Lambda f_t), \tag{1.2}$$

with a high-dimensional weight matrix W. One of the most popular estimators is studied in Bai (2009) and Moon and Weidner (2015), which uses $W = I_N$. Using the identity weight matrix, however, essentially treats $\{u_{it}\}_{i\le N}$ to be cross-sectionally independent and homoskedastic. Therefore, the estimation efficiency is lost in general. An important motivation for the paper is that efficiency can be gained by exploring the correlation structure in u_t. We allow a consistent estimator for $W = \Sigma_u^{-1}$ as the optimal weight matrix, where Σ_u is the $N \times N$ covariance matrix of u_t, assumed to be time-invariant. Note that often Σ_u is not a diagonal matrix. For instance, in financial applications, Y_t is a vector of firms' excess returns; X_t and f_t respectively represent observable and unobservable factors. Then u_t captures firms' idiosyncratic returns, which may be cross-sectionally correlated (Chamberlain and Rothschild, 1983).

We shall assume Σ_u to be a sparse covariance matrix. Roughly speaking, it means many off-diagonal entries of Σ_u are nearly zero. Except for the sparsity, the off-diagonal structure of Σ_u is unknown. We apply the "thresholding" method as in Fan et al. (2013), by directly estimating the small entries to be exactly zero.

* Corresponding author at: Department of Economics, Columbia University, New York, NY 10027, United States.
E-mail address: jb3064@columbia.edu (J. Bai).

http://dx.doi.org/10.1016/j.jeconom.2017.05.014
0304-4076/

We respectively use $\|A\| = \sqrt{\nu_{\max}(A'A)}$, $\|A\|_1 = \max_i \sum_j |A_{ij}|$ and $\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^2}$ as the operator norm, ℓ_1-norm and Frobenius norm of a matrix A. Here $\nu_{\max}(A'A)$ denotes the maximum eigenvalue of $A'A$. Note that if A is a vector, $\|A\| = \|A\|_F$ is equal to the Euclidean norm.

1.1. Overview of the results

The main result of this paper is to prove the following limiting distribution of the proposed estimator: with a consistent covariance estimator $\widehat{\Sigma}_u^{-1}$ as the weight matrix, $\widehat{\beta} = \widehat{\beta}(\widehat{\Sigma}_u^{-1})$ satisfies

$$\sqrt{NT}(\widehat{\beta} - \beta_0) \to^d \mathcal{N}(0, \Gamma^{-1}), \tag{1.3}$$

where the asymptotic covariance Γ^{-1} is first-order optimal among all choices of the weight matrix W. We focus on the case when $\{u_t\}$ are serially uncorrelated, and show that the rate of convergence is $O_P(\frac{1}{\sqrt{NT}})$ regardless of $N/T \in [0, \infty]$. In contrast, the estimator $\widehat{\beta}(I_N)$ of Bai (2009) has a rate of convergence (in the presence of cross-sectional correlations):

$$\|\widehat{\beta}(I_N) - \beta_0\| = O_P\left(\frac{1}{\sqrt{NT}} + \frac{1}{N}\right).$$

Hence the convergence rate is improved when $N = o(T)$. This is because essentially the efficient estimation works with the transformed data (by multiplying $\Sigma_u^{-1/2}$), which eliminates the bias due to cross-sectional correlations of Bai (2009)'s estimator. Therefore, no bias-correction is needed for $\widehat{\beta}$. It is also important to note that for technical simplicity, serial correlations of u_t are ruled out by our assumptions[1], but were previously allowed in the literature (e.g., Bai, 2009). In the presence of serial correlations, our estimator is also asymptotically biased. We discuss the remedies and possible extensions to this case in Section 3.4. In summary, the major advantages gained by using the proposed estimators are: (i) it eliminates cross-sectional correlation bias; (ii) it has a faster rate of convergence when $N = o(T)$, and (iii) it leads to a smaller asymptotic variance.

Why do we need a more efficient estimator? To demonstrate the potential loss of efficiency without incorporating Σ_u^{-1}, we present a real-data application in Section 6, which studies the effect of divorce law reforms on the change of divorce rates. Our estimator is applied to the year–state divorce rate data of U.S. during 1956–1985, and captures the significant (negative) effects from nine to twelve years after the law was reformed, consistent with the previous empirical findings in the literature. In contrast, existing methods without incorporating Σ_u^{-1} would result in wider confidence intervals and potentially conservative conclusions. Numerically, we find an average of 46% efficiency gained using the proposed estimator, relative to the existing method.

1.2. Technical contributions

Technically, we find that the "absolute convergence" $\|\widehat{\Sigma}_u^{-1} - \Sigma_u^{-1}\|$ is not sufficient for (1.3) when the dimension is larger than the sample size, even if its optimal convergence rate is achieved. Instead, for statistical inferences, a more relevant notion of covariance convergence should be the "doubly weighted convergence" of the form: $\frac{1}{\sqrt{NT}} Z'[(\widehat{\Sigma}_u^{-1} - \Sigma_u^{-1}) \otimes I_T]U = o_P(1)$, where I_T is a T-dimensional identity matrix and Z and U are stochastic matrices whose dimensions are $NT \times \dim(\beta)$ and $NT \times 1$ respectively. While using a feasible estimator $\widehat{\beta}(\widehat{\Sigma}_u^{-1})$, we need to show:

$$\sqrt{NT}(\widehat{\beta}(\Sigma_u^{-1}) - \widehat{\beta}(\widehat{\Sigma}_u^{-1})) = o_P(1). \tag{1.4}$$

[1] We do allow for serial correlations in (f_t, X_t).

The left hand side can be written as a weighted covariance estimation error:

$$\sqrt{NT}(\widehat{\beta}(\Sigma_u^{-1}) - \widehat{\beta}(\widehat{\Sigma}_u^{-1})) = \frac{1}{\sqrt{NT}} Z'[(\widehat{\Sigma}_u^{-1} - \Sigma_u^{-1}) \otimes I_T]U + o_P(1).$$

Hence it is crucial to establish the following convergence:

$$\frac{1}{\sqrt{NT}} Z'[(\widehat{\Sigma}_u^{-1} - \Sigma_u^{-1}) \otimes I_T]U = o_P(1).$$

However, the optimal rate of convergence (under the spectral norm) for estimating Σ_u^{-1} is Cai and Zhou (2012):

$$\|\widehat{\Sigma}_u^{-1} - \Sigma_u^{-1}\| = O_P\left(\sqrt{\frac{\log N}{T}}\right). \tag{1.5}$$

If we apply (1.5), we proceed as:

$$\left\| \frac{1}{\sqrt{NT}} Z'[(\widehat{\Sigma}_u^{-1} - \Sigma_u^{-1}) \otimes I_T]U \right\| \leq \frac{1}{\sqrt{NT}} \|Z\| \|\widehat{\Sigma}_u^{-1} - \Sigma_u^{-1}\| \|U\|.$$

In the current context, $\|Z\| = O_P(\sqrt{NT}) = \|U\|$, $i = 1, 2$, thus the upper bound does not converge. Hence the weighted convergence cannot be directly implied from the absolute convergence of covariance matrices. On the other hand, it is intuitively clear that the weights "average out" the estimation errors, and therefore the impact of estimating large inverse covariances should be asymptotically negligible. We develop a new technical argument to fulfill this intuition.

1.3. The literature

There has been a growing literature on panel data models with interactive effects: (Ahn et al., 2001; Pesaran, 2006; Moon and Weidner, 2015; Su et al., 2015; Harding and Lamarche, 2014; Westerlund and Urbain, 2013; Kao et al., 2012; Freyberger, 2012), etc. While we focus on the homogeneous panel with a common slope, several researchers also allow for heterogeneous slope coefficients β_i, such as Pesaran (2006) and Avarucci and Zaffaroni (2012) and Su and Chen (2013). In particular, Avarucci and Zaffaroni (2012) proposed GLS estimators to improve the estimation efficiency, and their method is different from ours in several aspects. For heterogeneous models, studying the efficient estimation using large covariance matrices is likely to be easier, due to a slower rate of convergence of the estimated β_i. In fact, the effect of estimating the optimal weight matrix is negligible as long as it converges faster than the parameters' rate of convergence. In heterogeneous models, the rate of convergence for each β_i is $O_P(\frac{1}{\sqrt{T}})$, while in homogeneous models the rate becomes $O_P(\frac{1}{\sqrt{NT}})$. Hence heterogeneous models often lead to weaker requirements on the effect of $\widehat{\Sigma}_u^{-1} - \Sigma_u^{-1}$, and are likely to be directly verifiable.

Our work complements papers that provide statistical inferences in models that involve high-dimensional nuisance parameters, e.g., (Belloni et al., 2014a,b; Zhang and Zhang, 2014; Van de Geer et al., 2014), among others. There, the high-dimensional nuisance parameter is often not covariance matrices, then as long as the "orthogonality condition" (e.g., Belloni et al., 2014a) holds, the minimax convergence rate is sufficient to ensure that the effect of estimating these nuisance parameters is negligible. In contrast, in the current context, even though the nuisance parameter Σ_u^{-1} is indeed orthogonal to β_0, the minimax convergence for estimating covariance matrices is no longer sufficient. The major reason causing such a distinctive feature is that, as introduced above, a "doubly-weighted" estimation error is relevant, where the weights are also high-dimensional. In contrast, the aforementioned works do not have this issue.[2] Furthermore, in the literature on factor models, Choi (2012) and Breitung and Tenhofen (2011) studied efficient estimations of factors and loadings using weighted least squares. More recently, Iwakura and Okui (2014) examined the asymptotic efficiency of estimating the factors and loadings. Other related works in the literature on static and dynamic factor models include, Tsai and Tsay (2010); Doz et al. (2011); Forni et al. (2000, 2005); Forni and Lippi (2001), among others. In addition, estimating a high-dimensional covariance matrix has been an active research area in the recent literature: El Karoui (2008); Bickel and Levina (2008a); Ledoit and Wolf (2012); Fan et al. (2013), etc.

The rest of the paper is organized as follows. Section 2 describes the problem and proposes the estimator. Section 3 introduces the regularity conditions. In particular, the sparsity condition on the error covariance matrix is defined. Section 4 presents the limiting distribution. Section 5 provides simulated studies, which also cover examples that are ruled out by our regularity conditions for robustness checks. Section 6 applies the method to a real data problem of divorce rate study. Finally, Section 7 concludes. All proofs are given in the Appendix.

Throughout the paper, we use $\nu_{\min}(A)$ and $\nu_{\max}(A)$ to denote the minimum and maximum eigenvalues of matrix A. For two sequences, we write $a_T \ll b_T$ (and equivalently $b_T \gg a_T$) if $a_T = o(b_T)$ as $N, T \to \infty$.

2. Estimation using sparse covariance estimator

2.1. The main algorithm

We estimate β_0 via

$$\widehat{\beta}(W) = \arg\min_{\beta} \min_{\{f_t\},\Lambda} \sum_{t=1}^{T} (Y_t - \Lambda f_t - X_t\beta)'W(Y_t - \Lambda f_t - X_t\beta),$$

$$\text{subject to } \frac{1}{T}\sum_{t=1}^{T} f_t f_t' = I_r \text{ and } \Lambda' W \Lambda \text{ is diagonal.} \tag{2.1}$$

Here W is an $N \times N$ weight matrix. The first-order asymptotic optimal weight matrix is taken as:

$$W = \Sigma_u^{-1}.$$

The estimator is feasible only when we have a consistent estimator $\widehat{\Sigma}_u^{-1}$ for Σ_u^{-1}. Our estimator of β is defined by:

$$\widehat{\beta} = \widehat{\beta}(\widehat{\Sigma}_u^{-1}). \tag{2.2}$$

The estimator can be computed as follows: the estimated β for each given (Λ, F) (where $F = (f_1, \ldots, f_T)'$) is simply

$$\beta(\Lambda, F, W) = \left(\sum_{t=1}^{T} X_t' W X_t\right)^{-1} \sum_{t=1}^{T} X_t' W (Y_t - \Lambda f_t).$$

On the other hand, given β, the variable $Y_t - X_t\beta$ has a factor structure: $Y_t - X_t\beta = \Lambda f_t + u_t$, from which we can estimate Λ and f_t using the principal components estimator. Specifically, let $X(\beta)$ be the $N \times T$ matrix $X(\beta) = (X_1\beta, \ldots, X_T\beta)$; let Y be the $N \times T$ matrix of y_{it}. Then the $T \times r$ matrix $F/\sqrt{T} = (f_1, \ldots, f_T)'/\sqrt{T}$ is estimated by $F(\beta, W)/\sqrt{T}$, whose columns are the eigenvectors corresponding to the largest r eigenvalues of the $T \times T$ matrix $(Y - X(\beta))'W(Y - X(\beta))$, and Λ is estimated by

$$\Lambda(\beta, W) = T^{-1}(Y - X(\beta))F(\beta, W).$$

[2] These authors considered inference procedures that are uniformly valid over a fairly broad class of data generating processes (DGP), while we focus on a fixed DGP. Nevertheless, they do not suffer from noise accumulations when applying the minimax rate of convergence for the high-dimensional nuisance parameter.

Therefore, given (Λ, F), we can estimate β, and given β, we can estimate (Λ, F). So starting from an initial value $\widehat{\beta}_{\text{in}}$ and the weight matrix W, $\widehat{\beta}(W)$ can be simply obtained by iterations. The following algorithm states the iterations, given $(\widehat{\beta}_{\text{in}}, W)$.

Algorithm A $(\widehat{\beta}_{\text{in}}, W)$
Step 1: Initialize $k = 0$ and $\widehat{\beta}_k = \widehat{\beta}_{\text{in}}$.
Step 2: Define $F_k = F(\widehat{\beta}_k, W)$, and $\Lambda_k = \Lambda(\widehat{\beta}_k, W)$.
Step 3: Let $\widehat{\beta}_{k+1} = \beta(\Lambda_k, F_k, W)$, and let $k \leftarrow k + 1$.
Step 4: Iterate steps 2, 3 until convergence.

Importantly, the iteration scheme only requires two matrix inverses: $\widehat{\Sigma}_u^{-1}$ (when $W = \widehat{\Sigma}_u^{-1}$) and $(\sum_{t=1}^{T} X_t' W X_t)^{-1}$, which do not update during iterations.

2.2. Computing $\widehat{\beta}_{in}$ and $W = \widehat{\Sigma}_u^{-1}$

Our estimator is defined using the asymptotically optimal weight $W = \widehat{\Sigma}_u^{-1}$, and the initial value $\widehat{\beta}_{\text{in}} = \widehat{\beta}(I_N)$, where $\widehat{\beta}(I_N)$ is the estimator that takes $W = I_N$ in (2.1). Under regularity conditions, $\widehat{\beta}(I_N)$ is consistent (e.g., Bai, 2009; Moon and Weidner, 2015; Westerlund and Urbain, 2013). It can be computed using Algorithm A $(0, I_N)$, that is, initializing with $\widehat{\beta}_{\text{in}} = 0$ and using the identity matrix as the weight matrix.

In our proposal, $\widehat{\beta}(I_N)$ is used for two purposes: (i) as an initial estimator for the iteration, and (ii) to construct $\widehat{\Sigma}_u^{-1}$. As for (ii), we define a covariance estimator that is similar to Fan et al. (2013): Apply the eigenvalue decomposition to

$$\frac{1}{T}\sum_{t=1}^{T}(Y_t - X_t\widehat{\beta}(I_N))(Y_t - X_t\widehat{\beta}(I_N))' = \sum_{i=1}^{N} \nu_i g_i g_i', \tag{2.3}$$

where $(\nu_j, g_j)_{j=1}^N$ are the eigenvalues–eigenvectors of $\frac{1}{T}\sum_{t=1}^{T}(Y_t - X_t\widehat{\beta}(I_N))(Y_t - X_t\widehat{\beta}(I_N))'$ in a decreasing order such that $\nu_1 \geq \nu_2 \geq \cdots \geq \nu_N$. Then $\widehat{\Sigma}_u = (\widehat{\Sigma}_{u,ij})_{N\times N}$,

$$\widehat{\Sigma}_{u,ij} = \begin{cases} \widetilde{R}_{ii}, & i = j \\ s_{ij}(\widetilde{R}_{ij}), & i \neq j \end{cases}, \quad \widetilde{R} = (\widetilde{R}_{ij})_{N\times N} = \sum_{i=r+1}^{N} \nu_i g_i g_i', \tag{2.4}$$

where $s_{ij}(\cdot) : \mathbb{R} \to \mathbb{R}$ is a thresholding function 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: $s_{ij}(z) = z1\{|z| > \tau_{ij}\}$, where $1\{|z| > \tau_{ij}\} = 1$ if $|z| > \tau_{ij}$, and zero otherwise; SCAD (Fan and Li, 2001), MPC (Zhang, 2010) etc. As for the threshold value, we specify

$$\tau_{ij} = C\sqrt{\tilde{R}_{ii}\tilde{R}_{jj}}\,\omega_{T,N}, \text{ where } \omega_{T,N} = \sqrt{\frac{\log N}{T}} + \frac{1}{\sqrt{N}} \tag{2.5}$$

for some pre-determined universal $C > 0$.

Fan et al. (2013) restricted themselves to the pure factor model without a regression effect $X_{it}'\beta$ in the model. In contrast, we focus on the first order asymptotics of $\widehat{\beta}(\widehat{\Sigma}_u^{-1})$, and show its advantage compared to estimators $\widehat{\beta}(W)$ with non-optimal weight matrices. It will be clear in the subsequent section that the rate of convergence for estimating Σ_u in Fan et al. (2013) is not sufficient to show that the effect of estimating Σ_u^{-1} is negligible.

The complete algorithm can be summarized as follows.

Algorithm B: two-step

1. Run [Algorithm A $(0, I_N)$], using 0 as the initial value and I_N as the weight matrix, obtain $\widehat{\beta}(I_N)$.
2. Compute $\widehat{\Sigma}_u^{-1}$ by (2.4).

3. Run [Algorithm A $(\widehat{\beta}(I_N), \widehat{\Sigma}_u^{-1})$], using $\widehat{\beta}(I_N)$ as the initial value and $\widehat{\Sigma}_u^{-1}$ as the weight matrix, obtain $\widehat{\beta}(\widehat{\Sigma}_u^{-1})$.

We shall call our proposed estimator $\widehat{\beta}(\widehat{\Sigma}_u^{-1})$ to be "EPC" (efficient principal components), and simply denote it by $\widehat{\beta}$. In contrast, we call $\widehat{\beta}(I_N)$ as "PC" (principal components).

2.3. A further refined iterative scheme

Once $\widehat{\beta}(\widehat{\Sigma}_u^{-1})$ is obtained, we can in principle update $\widehat{\Sigma}_u^{-1}$ using the thresholding method with $\widehat{\beta}(\widehat{\Sigma}_u^{-1})$ in place of $\widehat{\beta}(I_N)$ in (2.3). Intuitively, with an improved estimator $\widehat{\beta}(\widehat{\Sigma}_u^{-1})$, the estimation for Σ_u^{-1} is also improved. We then further update $\widehat{\beta}$ by using Algorithm A $(\widehat{\beta}(\widehat{\Sigma}_u^{-1}), \widehat{\Sigma}_u^{-1})$ with the improved covariance estimator and $\widehat{\beta}(\widehat{\Sigma}_u^{-1})$ as the initial value. Our theory admits this estimator as well. Indeed, we show that as long as the initial estimator has a rate of convergence $O_P(\frac{1}{\sqrt{NT}} + \frac{1}{T} + \frac{1}{N})$, and is used in (2.3) for the thresholded covariance estimation, the achieved estimated β is first-order asymptotically equivalent to (2.1) with $W = \Sigma_u^{-1}$ as the optimal weight matrix.

This motivates the following multistep algorithm.[3]

Algorithm C: multistep

1. Run [Algorithm B: two-step], compute $\widehat{\beta}(\widehat{\Sigma}_u^{-1})$.
2. Update $\widehat{\Sigma}_u^{-1}$ by (2.4), with $\widehat{\beta}(\widehat{\Sigma}_u^{-1})$ in place of $\widehat{\beta}(I_N)$.
3. Run [Algorithm A $(\widehat{\beta}(\widehat{\Sigma}_u^{-1}), \widehat{\Sigma}_u^{-1})$], using $\widehat{\beta}(\widehat{\Sigma}_u^{-1})$ as the initial value and the updated $\widehat{\Sigma}_u^{-1}$ as the weight matrix. Obtain the updated $\widehat{\beta}(\widehat{\Sigma}_u^{-1})$.
4. Repeat steps 2–3 until convergence.

Based on our numerical experience, while two-step and multistep algorithms give very close solutions, the two-step algorithm performs sufficiently well and runs much faster.

3. Assumptions

3.1. Regularity conditions

To avoid confusions, we shall use the $T \times r$ matrix $F_0 = (f_{10}, \ldots, f_{T0})'$ to denote the true unknown factors, and use $F = (f_1, \ldots, f_T)'$ as a generic matrix notation for factors (which may not be the true value). But we shall still use Σ_u and Λ to denote the true covariance matrix and loading matrix, which should not cause confusions.

We allow for serial dependence on the factors and explanatory variables, by requiring an (exponential) α-mixing condition.

Assumption 3.1. (i) $\{f_{t0}, u_t, X_t\}_{t\geq 1}$ is strictly stationary; $\{u_t\}_{t\geq 1}$ is serially uncorrelated, that is, $E(u_t u_s') = 0$ if $t \neq s$, for $t, s \leq T$. $Eu_t = 0$, and u_t is independent of (f_{t0}, X_t) for all $t \leq T$.

(ii) There exist constants $c_1, c_2 > 0$ such that $\nu_{\min}(\Sigma_u) > c_2$, $\max_{j\leq N}\|\lambda_j\| < c_1$, and $c_2 < \nu_{\min}(\mathrm{cov}(f_{t0})) \leq \nu_{\max}(\mathrm{cov}(f_{t0})) < c_1$.

(iii) Exponential tail: There exist $r_1, r_2 > 0$ and $b_1, b_2 > 0$, such that for any $s > 0$, $i \leq N$ and $j \leq r$, $P(|u_{it}| > s) \leq \exp(-(s/b_1)^{r_1})$, and $P(|f_{t0,j}| > s) \leq \exp(-(s/b_2)^{r_2})$.

(iv) Strong mixing: There exist $r_3, C > 0$ such that for all $T > 0$, $r_1^{-1} + r_2^{-1} + r_3^{-1} > 1$

$$\sup_{A\in\mathcal{F}_{-\infty}^{0}, B\in\mathcal{F}_T^{\infty}} |P(A)P(B) - P(AB)| < \exp(-CT^{r_3}),$$

where $\mathcal{F}_{-\infty}^{0}$ and $\mathcal{F}_T^{\infty}$ denote the σ-algebras generated by $\{(f_{t0}, X_t) : t \leq 0\}$ and $\{(f_{t0}, X_t) : t \geq T\}$ respectively.

Remark 3.1. It is interesting to study the efficient estimation of β in the presence of either cross-sectional correlation or serial correlation, or both. For the technical simplicity, in this paper we only allow for the cross-sectional correlations, and require $\{u_t\}$ be serially uncorrelated. As explained in Bai (2009), allowing for serial correlations may lead to serial biases. One of the possible ways of allowing the serial correlations is to flip the role of N and T, working with the $T \times T$ covariance matrix of $\{u_{it}\}$ for the efficient estimation. But this approach does not allow cross-sectional correlations. See Section 3.4 below for more detailed discussions.

Remark 3.2. The exponential tail assumption is needed to uniformly bound terms $\max_{i\leq N}\|\frac{1}{T}\sum_{t=1}^{T} f_{t0}\widehat{u}_{it}\|$ and $\max_{i,j\leq N}|\frac{1}{T}\sum_{t=1}^{T}(\widehat{u}_{it}\widehat{u}_{jt} - Eu_{it}u_{jt})|$ for the convergence of the error covariance matrix. The fact that N possibly grows faster than T requires us apply the concentration (exponential-type) inequalities. These inequalities require the underlying distributions be thin-tailed. On the other hand, allowing for heavy-tailed distributions is also an important question, which would require a very different estimation procedure, and is out of the scope of this paper.

Assumption 3.2. As $N \to \infty$, all the eigenvalues of the $r \times r$ matrix $\Lambda'\Lambda/N$ are bounded away from both zero and infinity.

Remark 3.3. We consider strong factors only, guaranteed by the above assumption. In fact, it is possible to additionally allow several weak factors. For instance, if there is only one factor but two are estimated, then by definition, the second estimated factor is weak. As is shown in Moon and Weidner (2015), we can still obtain the consistency. Yet, the asymptotic impact on estimating the high-dimensional error covariance matrix with additional weak factors is an open question.

Rearrange the design matrix

$$Z = (X_{11}, \ldots, X_{1T}, X_{21}, \ldots, X_{2T}, \ldots, X_{N1}, \ldots, X_{NT})', \quad NT \times d.$$

For the generic notation F of factors, which may not be the true values, let $M_F = I_T - F(F'F)^{-1}F'/T$. The following matrices play an important role in the identification and asymptotic analysis:

$$A_F = \left[\Sigma_u^{-1} - \Sigma_u^{-1}\Lambda\left(\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\Lambda'\Sigma_u^{-1}\right] \otimes M_F,$$

$$V(F) = \frac{1}{NT}Z'A_F Z, \tag{3.1}$$

where (Λ, Σ_u^{-1}) in the above represent the true loading matrix and inverse error covariance in the data generating process, and $\otimes$ denotes the Kronecker product.

Assumption 3.3. With probability approaching one,

$$\inf_{F:F'F/T=I_r} \nu_{\min}(V(F)) > 0.$$

This assumption assumes there are sufficient variations in the regressors, as in Bai (2009) and Moon and Weidner (2015). If we write $B_F = \left[\Sigma_u^{-1/2} - \Sigma_u^{-1}\Lambda\left(\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\Lambda'\Sigma_u^{-1/2}\right] \otimes M_F$, then $A_F = B_F B_F'$. So $V(F)$ is at least positive semi-definite. Also, summing over NT rows of Z should lead to a strictly positive definite matrix $V(F)$. As a sufficient condition, if X_{it} depends on the factors and loadings through:

$$X_{it} = \tau_i + \theta_t + \sum_{k=1}^{r} a_k\lambda_{ik} + \sum_{k=1}^{r} b_k f_{kt} + \sum_{k=1}^{r} c_k\lambda_{ik}f_{kt} + \eta_{it}$$

where a_k, b_k, c_k are constants (can be zero) and η_{it} is i.i.d. over both i and t, then Assumption 3.3 is satisfied (see Bai, 2009).

Let $U = (u_{11}, \ldots, u_{1T}, u_{21}, \ldots, u_{2T}, \ldots, u_{N1}, \ldots, u_{NT})'$, and F_0 be the $T \times r$ matrix of true factors.

[3] We thank a referee for motivating this more refined algorithm.

Assumption 3.4. There is a $\dim(\beta_0)\times\dim(\beta_0)$ positive definite matrix Γ such that

$$V(F_0)\to^P\Gamma,\quad \frac{1}{\sqrt{NT}}Z'A_{F_0}U\to^d\mathcal{N}(0,\Gamma).$$

This assumption is required for the asymptotic normality of $\widehat{\beta}:=\widehat{\beta}(\widehat{\Sigma}_u^{-1})$, because it can be shown that,

$$\sqrt{NT}(\widehat{\beta}-\beta_0)=V(F_0)^{-1}\frac{1}{\sqrt{NT}}Z'A_{F_0}U+o_P(1).$$

Hence the asymptotic normality depends on that of $\frac{1}{\sqrt{NT}}Z'A_{F_0}U$. Assumption 3.4 is not stringent because if we write $B_{F_0}'U=(\widetilde{u}_{11},\dots,\widetilde{u}_{1T},\widetilde{u}_{21},\dots,\widetilde{u}_{NT})'$, and $Z'B_{F_0}=(\widetilde{Z}_{11},\dots,\widetilde{Z}_{1T},\widetilde{Z}_{21},\dots,\widetilde{Z}_{NT})$, then $\frac{1}{\sqrt{NT}}Z'A_{F_0}U=\frac{1}{\sqrt{NT}}\sum_{t=1}^T\sum_{i=1}^N\widetilde{Z}_{it}\widetilde{u}_{it}$ is a standardized summation. We can further write

$$\sqrt{NT}(\widehat{\beta}-\beta_0)=\left(\frac{1}{NT}\sum_{t=1}^T\sum_{i=1}^N\widetilde{Z}_{it}\widetilde{Z}_{it}'\right)^{-1}\times\frac{1}{\sqrt{NT}}\sum_{t=1}^T\sum_{i=1}^N\widetilde{Z}_{it}\widetilde{u}_{it}+o_P(1).$$

Hence the second statement of Assumption 3.4 is a central limit theorem (CLT) for $\frac{1}{\sqrt{NT}}\sum_{t=1}^T\sum_{i=1}^N\widetilde{Z}_{it}\widetilde{u}_{it}$ on both cross-sectional and time domains. Note that the CLT permits variables to be correlated and heteroskedastic, see Davidson (1994).

In addition, in the absence of serial correlations, the conditional covariance of $\frac{1}{\sqrt{NT}}Z'A_{F_0}U$ given Z and F_0 equals $\frac{1}{NT}Z'A_{F_0}(\Sigma_u\otimes I_T)A_{F_0}Z=V(F_0)$.[4] This implies that the asymptotic variance of $\sqrt{NT}(\widehat{\beta}-\beta_0)$ is simply Γ^{-1}.

3.2. Sparsity condition on Σ_u

We assume Σ_u to be a sparse matrix and impose similar conditions as those in Rothman et al. (2008) and Lam and Fan (2009). Consider the notion of generalized sparsity: write $\Sigma_u=(\Sigma_{u,ij})_{N\times N}$. For some $q\in[0,1/2)$, define

$$m_N=\max_{i\le N}\sum_{j=1}^N|\Sigma_{u,ij}|^q. \tag{3.2}$$

In particular, when $q=0$, $m_N=\max_{i\le N}\sum_{j=1}^N 1\{\Sigma_{u,ij}\neq 0\}$. The "row (column)-sparse" structure on Σ_u requires, there is $q\in[0,1/2)$, such that

$$\sqrt{NT}m_N\omega_{T,N}^{3-q}=o(1),\quad m_N\omega_{T,N}^{1-q}=o(1), \tag{3.3}$$

where

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

Remark 3.4. To understand (3.3) in a simple case, consider $q=0$. Then (3.3) reduces to

$$N(\log N)^3m_N^2=o(T^2),\ Tm_N^2=o(N^2),\\ m_N^2\log N=o(\min\{N\log N,T\}). \tag{3.4}$$

Hence implicitly we require $N\log^3N=o(T^2)$ and $T=o(N^2)$. But we do not make specific conditions on N/T. If in addition, k $m_N=O(1)$, then these conditions are indeed further simplified to $N\log^3N=o(T^2)$ and $T=o(N^2)$. In general, we allow $m_N\to\infty$, which means the number of nonzero entries can grow slowly as the dimension increases.

While (3.3) is sufficient for the $\|.\|_1$-convergence of estimating Σ_u and its inverse (as in Fan et al., 2013), it is not sufficient for the "doubly weighted convergence" in (3.7) below. However, the latter notion of convergence is desirable for the statistical inference in our context. We need to strengthen the row-sparse condition. Let S_s and S_l denote two disjoint sets and respectively contain the indices of small and large elements of Σ_u in absolute value, and

$$\{(i,j):i\le N,j\le N\}=S_s\cup S_l. \tag{3.5}$$

Roughly speaking, a "small" entry is defined as an element of Σ_u whose magnitude is less than $O(\omega_{T,N})$; a "large" entry is defined as an element whose magnitude is greater than $O(\omega_{T,N})$. We assume $(i,i)\in S_l$ for all $i\le N$ so that all the diagonal entries are "large". The sparsity condition assumes that most of the indices (i,j) belong to S_s when $i\neq j$. The partition need not be unique, and our analysis suffices as long as such a partition exists. One does not need to know which elements belong to S_s or which elements belong to S_l.

Assumption 3.5. (i) There is $q\in[0,1/2)$ such that (3.3) holds.

(ii) There is a partition $\{(i,j):i\le N,j\le N\}=S_s\cup S_l$ such that the number of elements in S_l ("large entries") is $O(N)$ and $\sum_{(i,j)\in S_s}|\Sigma_{u,ij}|=O(1)$. In addition,

(iii)

$$\max_{(i,j)\in S_s}|\Sigma_{u,ij}|=O(\omega_{T,N}),\quad \omega_{T,N}=O(\min_{(i,j)\in S_l}|\Sigma_{u,ij}|).$$

(iv) $\max_{i\le N}\sum_{j=1}^N|\Sigma_{u,ij}|$ is bounded away from infinity.

In general, we allow the total number of large off-diagonal entries to grow at rate $O(N)$, and the remaining $O(N^2)-O(N)$ entries should be "small". Furthermore, most of the "small entries" should be sufficiently close to zero to have a bounded absolute sum (condition (ii)). Condition (iv) is a standard "idiosyncratic" assumption in the literature, e.g., Bai (2003) and Fan et al., (2013), which together with Assumption 3.2, guarantees that the common shocks $\lambda_i'f_t$ are asymptotically identified from u_{it}, as $N\to\infty$.

To see that Assumption 3.5 is indeed reasonable, consider a special case when Σ_u is strictly sparse, in the sense that its elements in small magnitudes (S_s) are exactly zero. In this case $\sum_{(i,j)\in S_s}|\Sigma_{u,ij}|=0$. Then S_l consists of all the nonzero entries. There are three important examples of the strictly sparse case under which all conditions of Assumption 3.5 are satisfied: (a) Block diagonal: Σ_u is a block covariance matrix with fixed block sizes. Then the number of nonzeros is naturally $O(N)$. (b) Strict factor model: Σ_u is a diagonal matrix. (c) Banded matrix, which means there is a fixed integer k such that

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

Then $S_s=\{(i,j):|i-j|>k\}$ and $S_l=\{(i,j):|i-j|\le k\}$. Then the number of nonzeros is no more than $O(kN)$. On the other hand, however, if k also grows with N, condition (ii) would be violated.

Remark 3.5 (*On Condition* (iii)). This condition requires the elements in S_s and S_l be well-separable. Often such a separability condition is used to prove the "oracle property" of covariance estimators (that is, the nonzero and zero entries can both be consistently identified), but are not needed to derive the absolute rate of convergence. As we have described, however, the inference problem under the current context encounters the weighted convergence, which turns out to be adverse to the accumulation of estimation errors. Without the lower bound condition on $\min_{(i,j)\in S_l}|\Sigma_{u,ij}|$, many of the estimated "large entries" can be thresholded off, introducing thresholding biases. Such a well-separability condition helps eliminate thresholding biases.

[4] The proof is a straightforward calculation: since $H:=I-\Sigma_u^{-1}\Lambda(\Lambda'\Sigma_u^{-1}\Lambda)^{-1}\Lambda'$ is idempotent, we have $A_F(\Sigma_u\otimes I)A_F=[(H\Sigma_u^{-1})\otimes M_F](\Sigma_u\otimes I)[(H\Sigma_u^{-1})\otimes M_F]=[H\otimes M_F][(H\Sigma_u^{-1})\otimes M_F]=(H\Sigma_u^{-1})\otimes M_F=A_F$.

3.3. Weighted convergence for estimating the weight matrix

With considerable technical analysis on the first order condition of $\widehat{\beta}$, it can be shown that

$$\sqrt{NT}(\widehat{\beta}-\beta_0)=V(F_0)^{-1}\frac{1}{\sqrt{NT}}Z'\widehat{A}U+o_P(1),$$

where $\widehat{A}$ is as $A_{\widehat{F}}$ in (3.1) with Σ_u^{-1} replaced with $\widehat{\Sigma}_u^{-1}$ and F replaced with $\widehat{F}$. Therefore, to reach the desirable asymptotic expansion

$$\sqrt{NT}(\widehat{\beta}-\beta_0)=V(F_0)^{-1}\frac{1}{\sqrt{NT}}Z'A_{F_0}U+o_P(1),$$

we need to show

$$\frac{1}{\sqrt{NT}}Z'(\widehat{A}-A_{F_0})U=o_P(1). \tag{3.6}$$

(3.6) requires the following convergence of the covariance estimation:

$$\frac{1}{\sqrt{NT}}Z'[(\widehat{\Sigma}_u^{-1}-\Sigma_u^{-1})\otimes I_T]U=o_P(1). \tag{3.7}$$

In other words, estimating Σ_u^{-1} should not affect the limiting distribution. However, proving (3.7) is surprisingly difficult. As is discussed in Section 1, simply applying the optimal rate of convergence for $\|\widehat{\Sigma}_u^{-1}-\Sigma_u^{-1}\|=O_P(\sqrt{\frac{\log N}{T}})$ does not work.[5]

To address this difficulty, we treat (3.7) as a "doubly weighted" estimation error of the covariance matrix estimations, with weights $\frac{1}{\sqrt{NT}}Z$ and U respectively. We name the convergence of this type of quantity to be "doubly weighted convergence", because both Z and U are high-dimensional weights on the covariance estimation error, so $(\widehat{\Sigma}_u^{-1}-\Sigma_u^{-1})\otimes I_T$ is weighted on both sides. When dealing with the weighted convergence, we should not separate the covariance estimation error from the weights, as they help "average out" the estimation errors.

The key assumption is stated as follows. For $l\leq \dim(\beta)$ and $t\leq T$, let $X_{l,t}=(X_{1t,l},\ldots,X_{Nt,l})'$ be $N\times 1$ and $\Sigma_{u,j}^{-1}$ denote the jth column of Σ_u^{-1}. In addition, let $e_t=\Sigma_u^{-1}u_t$.

Assumption 3.6. For each $l\leq \dim(\beta)$, define $Q_{jt}=\Sigma_{u,j}^{-1'}(EX_{l,t}f_{t0}')(Ef_{t0}f_{t0}')^{-1}f_{t0}$ and $\widetilde{Q}_{jt}=\Sigma_{u,j}^{-1'}X_{l,t}$ (here Q_{jt} and $\widetilde{Q}_{jt}$ are scalar). Then the following conditions hold for $w_{ji}\in\{\frac{1}{\sqrt{T}}\sum_{t=1}^T Q_{jt}e_{it},\frac{1}{\sqrt{T}}\sum_{t=1}^T\widetilde{Q}_{jt}e_{it}\}$:

$$\frac{1}{\sqrt{NT}}\sum_{i=1}^N\sum_{s=1}^T(u_{is}^2-Eu_{is}^2)w_{ii}=o_P(\sqrt{T}), \tag{3.8}$$

$$\frac{1}{\sqrt{NT}}\sum_{i\neq j,(i,j)\in S_l}\sum_{s=1}^T(u_{is}u_{js}-Eu_{is}u_{js})w_{ji}=o_P(\sqrt{T}). \tag{3.9}$$

This assumption is concerned about the weighted convergence of the empirical processes. Note that for each (i,j), $Ew_{ji}=0$, and $\mathrm{var}(w_{ji})=O(1)$ under the strong mixing condition. Hence the weights $\{w_{ji}\}$ are stochastically bounded. These conditions are not stringent because the left-hand-sides are in fact weighted double sums of the centered noises $(u_{is}u_{js}-Eu_{is}u_{js})$, with stochastically bounded weights. Condition (3.9) is new in the literature because it explicitly relates to the sparsity condition. It is however very similar to (3.8), where the index set of the cross-sectional sum is changed from $\{(i,j):i=j\}$ to $\{(i,j):i\neq j,(i,j)\in S_l\}$. Recall that S_l is defined as the set of "large entry" indices, whose number of elements is $O(N)$. So this condition is still reasonable.

[5] We would face the same technical problem even if the $\|.\|_1$ norm of $\widehat{\Sigma}_u^{-1}-\Sigma_u^{-1}$ is bounded, and the element-wise norm for either Z or U is used.

Remark 3.6. Assumption 3.6 can also be directly verified for concrete distributions such as normality. The details are given in Appendix A.6.

In summary, some of our conditions are similar to those of Bai (2009) (Assumptions 3.2–3.4), while others are stronger (Assumptions 3.1 and 3.5, which require serial independence, exponential tails, and the sparsity), due to the necessity of estimating Σ_u^{-1}. Finally, the new Assumption 3.6 is required to prove that the effect of estimating Σ_u^{-1} is negligible.

3.4. Discussion on serial correlation

In this subsection we discuss possible extensions to allow for serial correlations without asymptotic bias. The issue of controlling the effects of estimating large weight matrices is still present. In principle, this challenge can be resolved using our doubly weighted convergence technique, but we expect the technical analysis to be heavily involved, and do not get into technical details.

3.4.1. Serial correlations only

Suppose $\{u_{it}\}$ is weakly correlated across t, but independent across i: $\max_{i\leq N}|Eu_{it}u_{is}|\to 0$ fast as $|t-s|\to\infty$. Let y_i be the $T\times 1$ vector of y_{it}; X_i be the $T\times\dim(\beta)$ matrix of X_{it}, and F be the $T\times K$ matrix of f_t. We can efficiently estimate β via:

$$\min_{F,\Lambda,\beta}\sum_{i=1}^N(y_i-X_i\beta-F\lambda_i)'\Sigma_i^{-1}(y_i-X_i\beta-F\lambda_i),$$

where $\Sigma_i:=(\sigma_{i,ts})$ is a $T\times T$ matrix with $\sigma_{i,ts}=Eu_{it}u_{is}$. Due to the serial stationarity and strong mixing condition, $\sigma_{i,ts}$ depends on (t,s) only through $|t-s|$, and decays fast uniformly in $i\leq N$ as $|t-s|\to 0$. Therefore, we can employ the "banding" procedure of Bickel and Levina (2008b) to consistently estimate Σ_i:

$$\widehat{\sigma}_{i,ts}=\begin{cases}\frac{1}{T}\sum_{t'=1}^{T-h}\widehat{u}_{it'}\widehat{u}_{it'+h}, & \text{if } |t-s|=h,\quad h<l_T\\ 0, & \text{if } |t-s|\geq l_T.\end{cases}$$

The bandwidth $l_T\to\infty$ needs to be chosen carefully to maintain the positive definiteness of $\widehat{\Sigma}_i=(\widehat{\sigma}_{i,ts})$. In the absence of cross-sectional correlations, essentially we are working on the transformed data $\Sigma_i^{-1/2}y_i$ and $\Sigma_i^{-1/2}X_i$; thus the bias due to serial correlations can be removed. However, when the cross-sectional correlations are present this estimator can result in cross-sectional biases.

3.4.2. Both serial and cross-sectional correlations

When both serial and cross-sectional correlations are present, we can define an asymptotically unbiased efficient estimator as follows. Let $\tilde{Y}=(Y_1',\ldots,Y_T')'$ be the $(NT)\times 1$ vector of y_{it} with each Y_t being an $N\times 1$ vector ; $\tilde{X}=(X_1',\ldots,X_T')'$ be the $(NT)\times\dim(\beta_0)$ matrix of X_{it} with each X_t being $N\times\dim(\beta_0)$. We can efficiently estimate β_0 via the following GLS:

$$\min_{F,\Lambda,\beta}(\tilde{Y}-\tilde{X}\beta-\mathrm{vec}(\Lambda F'))'\Sigma_U^{-1}(\tilde{Y}-\tilde{X}\beta-\mathrm{vec}(\Lambda F')),$$

where $\Sigma_U=(Eu_tu_s')$ is an $(NT)\times(NT)$ matrix, consisting of many "blocks". The (t,s) th block is an $N\times N$ matrix Eu_tu_s'. Due to the serial stationarity and strong mixing condition, Eu_tu_s' depends on (t,s) only through $|t-s|$, and quickly converges to a zero matrix as $|t-s|\to 0$. Therefore, we can combine the thresholding with "banding" to consistently estimate Σ_U by $\widehat{\Sigma}_U=(\widehat{\Sigma}_{t,s})$, where each block $\widehat{\Sigma}_{t,s}$ is an $N\times N$ matrix defined as:

$$\widehat{\Sigma}_{t,s}=\begin{cases}(\widetilde{\sigma}_{h,ij})_{N\times N}, & \text{if } |t-s|=h,\quad h<l_T\\ 0, & \text{if } |t-s|\geq l_T,\end{cases}$$

and $(\widetilde{\sigma}_{h,ij})_{N\times N}$ is a thresholded autocovariance estimator, with the (i,j)th element $\widetilde{\sigma}_{h,ij} = s_{ij}(\frac{1}{T}\sum_{t'=1}^{T-h}\widehat{u}_{it'}\widehat{u}_{jt'+h})$. Hence we are using thresholding to regulate the estimation of cross-sectional correlations, and using banding to regulate the estimation of serial correlations. Still, we need to choose $l_T \to \infty$ carefully for both theoretical and practical purposes. This procedure allows both serial and cross-sectional correlations for u_{it}.

4. Limiting distribution

The following proposition shows that the effect of estimating the high-dimensional weight matrix is indeed negligible.

Proposition 4.1. *Suppose the initial estimator $\widehat{\beta}_{in}$ satisfies: $\|\widehat{\beta}_{in} - \beta_0\| = O_P(\frac{1}{\sqrt{NT}} + \frac{1}{N} + \frac{1}{T})$. Then under Assumption 3.1–3.6, as $N, T \to \infty$ we have the weighted convergence:*

$$\frac{1}{\sqrt{NT}} Z'(\widehat{A} - A_{F_0})U = o_P(1).$$

We have the following limiting distribution.

Theorem 4.1. *Under the assumptions of Proposition 4.1, the asymptotic limiting distribution of $\widehat{\beta}$ is the same when either $W = \Sigma_u^{-1}$ or the feasible weight $W_T = \widehat{\Sigma}_u^{-1}$ is used, and is given by, as $T, N \to \infty$,*

$$\sqrt{NT}(\widehat{\beta} - \beta_0) \to^d \mathcal{N}(0, \Gamma^{-1}),$$

where Γ is defined in Assumption 3.4.

Remark 4.1. As described in Section 2, the initial estimator can be chosen as the estimator of Bai (2009) $\widehat{\beta}_{\rm in} = \widehat{\beta}(I_N)$. Then the sufficient conditions of $\|\widehat{\beta}(I_N) - \beta_0\| = O_P(\frac{1}{\sqrt{NT}} + \frac{1}{N} + \frac{1}{T})$ are given in Bai (2009) and Moon and Weidner (2015). In addition, our results also allow for a multi-step updated estimator through Algorithm C, using the estimator obtained in Algorithm B as the initial estimator. By Theorem 4.1, these estimators are asymptotically equivalent, while Algorithm C is more computationally extensive.

Remark 4.2. In the absence of serial correlations and the heteroskedasticity, Theorem 4.1 shows that the rate of convergence is $O_P(\frac{1}{\sqrt{NT}})$ regardless of $N/T \in [0, \infty]$. In contrast, the estimator $\widehat{\beta}(I_N)$ of Bai (2009) has a rate of convergence (in the absence of serial correlations and the heteroskedasticity):

$$\|\widehat{\beta}(I_N) - \beta_0\| = O_P\left(\frac{1}{\sqrt{NT}} + \frac{1}{N}\right).$$

We see that the rate of $\widehat{\beta}$ does not have the term $\frac{1}{N}$, which reflects the usual bias caused by cross-sectional correlations. Hence the convergence rate is improved when $N = o(T)$. This is because essentially the efficient estimation works with the transformed data:

$$\widehat{\Sigma}_u^{-1/2} Y_t = \widehat{\Sigma}_u^{-1/2}\Lambda f_t + \widehat{\Sigma}_u^{-1/2} X_t \beta_0 + \widehat{\Sigma}_u^{-1/2} u_t,$$

hence eliminating the bias due to cross-sectional correlations. In the presence of serial correlations, our estimator would be also asymptotically biased.

Remark 4.3. An important motivation for the paper is that efficiency can be gained by exploring the correlation structure in u_t. To see this, note that under the same set of conditions, the estimator $\widehat{\beta}(W)$ with a general weight matrix W gives an asymptotic conditional covariance (given Z, F_0) of the sandwich-formula:

$$V_2 \equiv \left(\frac{1}{NT}Z'GZ\right)^{-1} \frac{1}{NT} Z'G(\Sigma_u \otimes I_T)GZ \left(\frac{1}{NT}Z'GZ\right)^{-1},$$

where G is defined as A_{F_0} with Σ_u^{-1} replaced with W. The first-order asymptotic optimal choice is $W = \Sigma_u^{-1}$ since the asymptotic variance Γ^{-1} is the limit of $V(F_0)^{-1}$ and $V_2 - V(F_0)^{-1}$ is positive semi-definite.

To estimate the asymptotic variance of $\widehat{\beta}$, let $\widehat{A}$ equal A_F with F, Λ and Σ_u^{-1} replaced with $\widehat{F}$, $\widehat{\Lambda}$ and $\widehat{\Sigma}_u^{-1}$. Define $\widehat{\Gamma} = \frac{1}{NT}Z'\widehat{A}Z$. The following result enables us to conduct inferences for β under large samples.

Theorem 4.2. *Under the assumptions of Theorem 4.1,*

$$\widehat{\Gamma}^{-1} \to^P \Gamma^{-1}.$$

Remark 4.4. (Unknown Number of Factors) The estimation is robust to over-estimating r. Moon and Weidner (2015) showed that under certain assumptions, the limiting distribution is independent of the number of factors used in the estimation, as long as this number is not underestimated. We can also start with a consistent estimator $\widehat{r}$ using a similar method of Bai and Ng (2002); Bai (2009); Ahn and Horenstein (2013); Hallin and Liška (2007); Li et al. (2017), among others. We then can apply the estimator $\widehat{r}$ to construct the EPC estimator, and achieve the same limiting distributions.

5. Simulated experiments

We consider the model $y_{it} = X_{it}\beta_0 + \lambda_{i1}f_{1,t} + \lambda_{i,2}f_{2t} + u_{it}$, where the true $\beta_0 = 1$. The regressors are generated to be dependent on (f_t, λ_i):

$$X_{it} = 2.5\lambda_{i1}f_{1,t} - 0.2\lambda_{i2}f_{2,t} - 1 + \eta_{it},$$

where η_{it} is i.i.d. standard normal. Let the two factors $\{f_{1t}, f_{2t}\}$ be i.i.d. $\mathcal{N}(0, 1)$, and $\{\lambda_{i,1}, \lambda_{i,2}\}_{i\le N}$ be uniform on [0, 1]. We generate the idiosyncratic errors as follows: let $\{\epsilon_{it}\}_{i\le N, t\le T}$ be i.i.d. $\mathcal{N}(0, 1)$. Let $\zeta_{1t} = \epsilon_{1t}$, $\zeta_{2t} = \epsilon_{2t} + a_1\epsilon_{1t}$, $\zeta_{3t} = \epsilon_{3t} + a_2\epsilon_{2t} + b_1\epsilon_{1t}$, and

$$\zeta_{i+1,t} = \epsilon_{i+1,t} + a_i\epsilon_{it} + b_{i-1}\epsilon_{i-1,t} + c_{i-2}\epsilon_{i-2,t},$$

where the constants $\{a_i, b_i, c_i\}_{i=1}^N$ are i.i.d. $\mathcal{N}(0, \sqrt{5})$. Then the correlation matrix of $\zeta_t = (\zeta_{1t}, \dots, \zeta_{Nt})'$ is banded, denoted by $R_\zeta = (R_{\zeta,ij})_{N\times N}$. We keep the cross-sectional correlations of ζ_t but introduce cross-sectional heteroskedasticity as follows: let $D = \text{diag}\{d_i\}$, where $\{d_i\}_{i\le N}$ are i.i.d. Uniform$(0, m)$. Finally, define the covariance matrix of u_t as $\Sigma_u = DR_\zeta D$, and generate $\{u_t\}_{t\le T}$ as i.i.d. $\mathcal{N}(0, \Sigma_u)$. Varying $m > 0$ enables us to control for the cross-sectional heteroskedasticity. In the studies below, we report results when $m = \sqrt{3}$.

We compare the proposed estimator with that of Bai (2009). To study the robustness of over-estimating the number of factors, we apply our method with the number of "working factors" being two through five, while the true number is two. The simulation is replicated for one thousand times. We calculate the bias, normalized standard error (Normal. SE) of EPC and PC, as well as the relative mean squared error (RMSE). Results are summarized in Table 1.

We see that both methods are almost unbiased, while the EPC indeed has significantly smaller standard errors than PC. In addition, it is clear that for all the choices of the working number of factors, the performances are quite similar. Hence increasing the working number of factors does not deteriorate the performance for both methods; yet the normalized standard error slightly increases. In all cases the RMSE is significantly smaller than one.

Our main theoretical studies require serial independence on u_t. We now numerically study the effects of serial dependences. Only the idiosyncratic errors are generated differently. Specifically, we define a $(TN) \times (TN)$ covariance matrix $\Sigma_U = (Eu_t u_s')$, where the (t, s) th block is an $N \times N$ covariance matrix, given by

$$Eu_t u_s' = DR_\zeta D\rho^{|t-s|}, \quad \rho \in (0, 1).$$

Table 1
Performances of estimated β_0; true $\beta_0 = 1$; serial independence.

N	T	Mean		Normal. SE		RMSE	Mean		Normal. SE		RMSE
		EPC	PC	EPC	PC		EPC	PC	EPC	PC	
		$r = 2$					$r = 3$				
100	100	1.003	1.006	0.682	1.128	0.351	1.003	1.006	0.692	1.120	0.369
	150	1.002	1.004	0.611	1.101	0.301	1.002	1.004	0.617	1.116	0.301
	200	1.002	1.003	0.539	1.073	0.252	1.002	1.003	0.547	1.071	0.261
150	100	1.003	1.005	0.759	1.139	0.427	1.003	1.005	0.773	1.144	0.440
	150	1.002	1.003	0.616	0.986	0.356	1.002	1.003	0.617	0.993	0.355
	200	1.001	1.003	0.594	0.920	0.355	1.001	1.003	0.601	0.926	0.360
		$r = 4$					$r = 5$				
100	100	1.003	1.006	0.707	1.124	0.387	1.003	1.005	0.723	1.126	0.404
	150	1.002	1.004	0.624	1.126	0.304	1.002	1.004	0.630	1.129	0.310
	200	1.002	1.003	0.554	1.069	0.272	1.002	1.003	0.563	1.071	0.281
150	100	1.003	1.005	0.786	1.158	0.446	1.003	1.005	0.794	1.165	0.450
	150	1.002	1.003	0.620	0.998	0.356	1.002	1.003	0.622	1.002	0.355
	200	1.001	1.003	0.606	0.926	0.367	1.001	1.003	0.610	0.927	0.373

EPC (with weight $\widehat{\Sigma}_u^{-1}$) and PC (existing method) comparison. "Mean" is the average of the estimators; "Normal. SE" is the standard error of the estimators multiplied by $\sqrt{NT}$. RMSE is the ratio of the mean squared error of EPC to that of PC; r represents the working number of factors, with the true value being 2. Reported results are based on 1000 replications.

Table 2
Performances of estimated β_0; true $\beta_0 = 1$; serial dependence.

N	T	Mean		Normal. SE		RMSE	Mean		Normal. SE		RMSE
		EPC	PC	EPC	PC		EPC	PC	EPC	PC	
		$r = 2$					$r = 3$				
100	100	1.003	1.007	0.740	1.452	0.233	1.002	1.006	0.746	1.415	0.261
100	150	1.002	1.006	0.647	1.549	0.156	1.002	1.006	0.653	1.542	0.164
100	200	1.001	1.003	0.477	1.406	0.117	1.001	1.003	0.483	1.401	0.122
150	100	1.004	1.008	1.177	1.680	0.438	1.004	1.007	1.167	1.657	0.462
150	150	1.002	1.004	0.610	1.441	0.186	1.002	1.004	0.614	1.419	0.196
150	200	1.001	1.003	0.569	1.223	0.211	1.001	1.003	0.576	1.229	0.217
		$r = 4$					$r = 5$				
100	100	1.002	1.005	0.747	1.388	0.278	1.002	1.005	0.756	1.389	0.291
100	150	1.002	1.005	0.665	1.539	0.174	1.002	1.005	0.675	1.529	0.183
100	200	1.001	1.003	0.487	1.407	0.123	1.001	1.003	0.493	1.414	0.125
150	100	1.004	1.006	1.174	1.663	0.477	1.004	1.006	1.170	1.662	0.484
150	150	1.002	1.004	0.617	1.415	0.200	1.002	1.004	0.621	1.414	0.204
150	200	1.001	1.003	0.579	1.230	0.220	1.001	1.003	0.581	1.231	0.223

EPC (with weight $\widehat{\Sigma}_u^{-1}$) and PC (existing method) comparison. "Mean" is the average of the estimators; "Normal. SE" is the standard error of the estimators multiplied by $\sqrt{NT}$. RMSE is the ratio of the mean squared error of EPC to that of PC; r represents the working number of factors, with the true value being 2. Reported results are based on 1000 replications.

Here $DR_\zeta D$ is the same base-covariance as in the independent case. Hence $\Sigma_U = (\rho^{|t-s|})_{T\times T} \otimes DR_\zeta D$; the (t, s)th block covariance decays exponentially as $|t - s|$ increases. We generate the $NT \times 1$ vector $(u_1', \dots, u_T')' = \Sigma_U^{1/2}\eta$,[6] where η is an $NT \times 1$ vector, whose entries are generated as i.i.d. $\mathcal{N}(0, 1)$. The estimation results are reported in Table 2 when $\rho = 0.7$. Our previous numerical findings continue to hold in the serial dependent case. Interestingly, we do not observe significant serial correlation biases for either estimator in this study.

Our last simulation considers weak factors. We focus on the serial independent case as in the first study, but now either one or both of the factors are weak. Specifically, we consider two examples. In the first example, $\{\lambda_{i,1}\}_{i\leq N}$ are generated independently from Uniform(0, 1), while $\{\lambda_{i,2}\}_{i\leq N}$ are from $\mathcal{N}(0, \sigma_\lambda^2)$, where σ_λ is close to zero. In the second example, both $\lambda_{i,1}$ and $\lambda_{i,2}$ are independently generated from $\mathcal{N}(0, \sigma_\lambda^2)$. Note that σ_λ controls the strength of the factor since $\frac{1}{N}\sum_i \lambda_{i,2}^2 \approx \sigma_\lambda^2$. When σ_λ is sufficiently small, the "effective" number of factors is just one in the first example, and is zero in the second example. Then essentially we are over-estimating the number of factors, and the previous numerical studies demonstrate that our estimation should be robust to this case. Indeed, Table 3 shows that the numerical results look very similar as before, where we set $\sigma_\lambda^2 = 0.01$. In general, the RMSE demonstrates that the efficiency gain is more substantial when both factors are weak.

6. Empirical study : Effects of divorce law reforms

This section demonstrates the gain of incorporating the estimated Σ_u in the panel data estimation and the efficiency gains compared to the usual PC method.

6.1. The background

An important question in sociology is the cause of the sharp increase in the U.S. divorce rate in the 1960s and 1970s. The association between divorce rates and divorce law reforms has been considered a potential key, and during 1970s, about three quarters of states in the U.S. liberalized their divorce system, so-called "no-fault revolution". There is plenty of empirical research regarding the effects of divorce law reforms on the divorce rates (e.g., Peters, 1986; Allen, 1992), and statistical significance of these

[6] While $((\rho^{|t-s|})\otimes DR_\zeta D)^{1/2} = (\rho^{|t-s|})^{1/2}\otimes(DR_\zeta D)^{1/2}$, it is much easier to compute $(\rho^{|t-s|})^{1/2} \otimes (DR_\zeta D)^{1/2}$ than to directly compute $((\rho^{|t-s|}) \otimes DR_\zeta D)^{1/2}$, because the latter is a much larger covariance matrix.

Table 3
True $\beta_0 = 1$; serial independence; weak factors.

N	T	Mean		Normal. SE		RMSE	Mean		Normal. SE		RMSE
		EPC	PC	EPC	PC		EPC	PC	EPC	PC	
		$r = 2$					$r = 3$				
		The second factor is weak									
100	100	1.001	1.005	0.504	0.917	0.250	1.001	1.005	0.509	0.911	0.262
100	200	1.001	1.002	0.308	0.819	0.134	1.001	1.002	0.314	0.827	0.141
150	100	1.003	1.004	0.729	1.005	0.491	1.003	1.004	0.735	1.007	0.498
150	200	1.001	1.002	0.686	0.919	0.554	1.001	1.002	0.694	0.915	0.573
		Both factors are weak									
100	100	1.003	1.011	0.226	0.710	0.068	1.002	1.011	0.208	0.704	0.051
100	200	1.004	1.011	0.281	0.699	0.163	1.004	1.010	0.281	0.687	0.144
150	100	1.001	1.010	0.128	0.761	0.012	1.001	1.009	0.120	0.744	0.010
150	200	1.001	1.010	0.256	0.791	0.023	1.001	1.010	0.226	0.797	0.018

Top panel: the loading of the second factor is generated from $\mathcal{N}(0, \sigma_\lambda^2)$; lower panel: the loadings of both factors are generated from $\mathcal{N}(0, \sigma_\lambda^2)$ with $\sigma_\lambda^2 = 0.01$.

effects has been found (e.g., Friedberg, 1998). In other words, states' law reforms are found to have significantly contributed to the increase in state-level divorce rates within the first eight years following reforms. In particular, Friedberg (1998) controlled for state and year fixed effects, and suggested that the adoption of unilateral divorce laws accounts for about one-sixth of the rise in the divorce rate since the late 1960s.

On the other hand, there has been a puzzle about longer effects. Empirical evidence also illustrates the subsequent decrease of the divorce rates starting from (around) 1975, which is between nine and fourteen years after the law reforms in most states. So whether law reforms continue to contribute to the rate decrease has been an interesting question. Wolfers (2006) studied a treatment effect panel data model, and identified negative effects for the subsequent years. Specifically, he studied the following model

$$y_{it} = \sum_{k=1}^{K} X_{it,k}\beta_k + g_{it} + f(\delta_i, t) + u_{it}, \tag{6.1}$$

where y_{it} is the divorce rate for state i in year t; $X_{it,k}$ is a binary regressor, representing the treatment effect $2k$ years after the reform. Specifically, we observe the law reform year T_i for each state. Then $X_{it,k} = 1$ if $2k - 1 \le t - T_i \le 2k$, and zero otherwise. Here g_{it} captures the unobserved heterogeneity, modeled as

$$g_{it} = \mu_i + \alpha_t \tag{6.2}$$

where μ_i and α_t are the state and time fixed effects. In addition, $f(\delta_i, t)$ is the time trend. For instance, the linear trend defines $f(\delta_i, t) = \delta_i t$ with unknown coefficient δ_i. Empirically Wolfers (2006) found that "the divorce rate rose sharply following the adoption of unilateral divorce laws, but this rise was reversed within about a decade". He concluded that "15 years after reform the divorce rate is lower as a result of the adoption of unilateral divorce, although it is hard to draw any strong conclusions about long-run effects". His result suggests that the increase in divorce following reform and the subsequent decrease may be two sides of the same treatment: after earlier dissolution of bad matches after law reforms, marital relations were gradually affected and changed.

However, there has been doubts on the robustness of Wolfers (2006), and it has been argued that his approach may fail to control for complex unobserved heterogeneity correlated with divorce law reforms, because the additive specification (6.2) is not flexible enough to capture factors varying across time and state. For instance, the endogeneity problem can arise in omitting social and cultural factors, such as the stigma of divorce, religious belief, and female participation in the work force, for most of which we do not have data or appropriate proxy variables. Friedberg (1998) discussed several factors that change across states and time and may affect both divorce rates and divorce law reform.

The heterogeneity may exist through an interactive effect, where unobserved common factors may change over time. The interactive fixed effects can effectively control for the remaining unobserved heterogeneity in the error term, and therefore help resolve the issue of the endogeneity due to omitted variables. Kim and Oka (2014) pioneered using interactive effect model for the study, where g_{it} in (6.1) is modeled as:

$$g_{it} = \lambda_i' f_t + \mu_i + \alpha_t. \tag{6.3}$$

Kim and Oka (2014) estimated the model using the PC method without estimating Σ_u^{-1}. Their first result confirms the significance of $(\beta_1, \ldots, \beta_4)$, which is in line with that of Wolfers (2006). Different from Wolfers (2006) though, they concluded insignificant $(\beta_5, \ldots, \beta_8)$, that is, the divorce rates after eight years and beyond are not affected by the reforms. However, we argue that estimating β without incorporating Σ_u^{-1} may lose efficiency because it ignores the remaining cross-sectional dependences (if any) that cannot be explained by the factors. So data used in the estimation is not "optimally weighted". As a result, this can lead to wide confidence intervals and possibly conservative conclusions.

6.2. Real data application

We re-estimate the model of Kim and Oka (2014) using the proposed method. As a first step, we rewrite the model to fit in the form being considered in this paper. Introduce the conventional notation: $\dot{y}_{it} = y_{it} - \frac{1}{T}\sum_{t=1}^{T} y_{it} - \frac{1}{N}\sum_{i=1}^{N} y_{it} + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} y_{it}$. Let $\dot{X}_{it,k}$, $\dot{u}_{it}$ be defined similarly. In addition, let $\dot{\lambda}_i = \lambda_i - \frac{1}{N}\sum_{i=1}^{N}\lambda_i$, and $\dot{f}_t = f_t - \frac{1}{T}\sum_{t=1}^{T} f_t$. If the time trend $f(\delta_i, t)$ is not present,[7] we have $\dot{y}_{it} = \dot{X}_{it}'\beta + \dot{\lambda}_i'\dot{f}_t + \dot{u}_{it}$.

The same data as in Wolfers (2006) and Kim and Oka (2014) are used, which contain the divorce rates, state-level reform years and binary regressors from 1956 to 1988 ($T = 33$) over 48 states. We fit the models both with and without linear time trend, and apply regular PC and our proposed EPC in each model to estimate β with confidence intervals. The number of factors is selected in a data-driven way as in Bai (2009). His IC and CP both suggested ten factors.[8] Moreover, for the EPC, the threshold value in the estimated covariance is obtained using the suggested cross-validation

[7] When the time trend is present, we can do a simple projection to eliminate the time trend, and still estimate the untransformed β from the familiar interactive effect model. For instance, suppose $f(\delta_i, t) = \delta_i t$. Let $M = (1, 2, \ldots, T)'$ and $P_M = I_T - M(M'M)^{-1}M'$. We can define $\widetilde{Y}_i = P_M(y_{i1}, \ldots, y_{iT})'$, and $\widetilde{X}_i = P_M(X_{i1}, \ldots, X_{iT})'$, and define $\widetilde{y}_{it}$ and $\widetilde{X}_{it}$ accordingly from $\widetilde{y}_{it}$ and $\widetilde{X}_{it}$.

[8] This is the same as in Kim and Oka (2014). We also tried a few larger values for r, and the estimates are similar, consistent with previous findings that the estimation is robust to over-estimating r.

Table 4
Method comparison in effects of divorce law reform: real data.

	Interactive effect				
	EPC		PC		var(EPC)/var(PC)
	Estimate	Confidence interval	Estimate	Confidence interval	
First 2 years	0.014	[0.007, 0.021]*	0.018	[0.0091, 0.028]*	0.59
3–4 years	0.034	[0.027, 0.041]*	0.042	[0.032, 0.053]*	0.59
5–6 years	0.025	[0.017, 0.032]*	0.032	[0.022, 0.042]*	0.58
7–8 years	0.015	[0.007, 0.023]*	0.030	[0.019, 0.04]*	0.56
9–10 years	−0.006	[-0.014, 0.001]	0.008	[−0.002, 0.018]	0.56
11–12 years	−0.008	[-0.015, -0.001]*	0.010	[−0.001, 0.02]	0.53
13–14 years	−0.009	[−0.017, −0.001]*	0.005	[−0.005, 0.016]	0.53
15 years+	0.009	[0.001, 0.017]*	0.031	[0.020, 0.042]*	0.55
	Interactive effect+linear trend				
	EPC		PC		var(EPC)/var(PC)
	Estimate	Confidence interval	Estimate	Confidence interval	
First 2 years	0.014	[0.006, 0.021]*	0.016	[0.006, 0.026]*	0.55
3–4 years	0.032	[0.024, 0.039]*	0.037	[0.026, 0.047]*	0.54
5–6 years	0.018	[0.010, 0.026]*	0.024	[0.012, 0.035]*	0.54
7–8 years	0.006	[−0.002, 0.014]	0.017	[0.005, 0.028]*	0.52
9–10 years	−0.017	[−0.025, −0.008]*	−0.007	[−0.019, 0.005]	0.52
11–12 years	−0.019	[−0.028, −0.010]*	−0.006	[−0.018, 0.006]	0.51
13–14 years	−0.021	[−0.030, −0.012]*	−0.012	[−0.025, 0.001]	0.50
15 years+	−0.003	[−0.012, 0.006]	0.014	[0.000, 0.028]*	0.46

95% confidence intervals are reported; intervals with * are significant. Relative efficiency is referred to EPC relative to PC, as estimated var(*EPC*)/var(*PC*).

procedure in Fan et al. (2013). The estimated $(\beta_1, \dots, \beta_8)$ and their confidence intervals are summarized in Table 4.

Both models produce similar estimates. Interestingly, EPC confirms that the law reforms significantly contribute to the subsequent decrease of the divorce rates, more specifically, 9–14 years after the reform in the model with linear time trends, and 11–14 years after in the model without linear time trends. In contrast, the regular PC reaches a more conservative conclusion as it does not capture these significant negative effects. We also notice that when the linear trend is added, the effect of 15 years+ estimated by EPC becomes insignificant, but it is significant if estimated by PC. We argue that while it is possible that the time trend is able to control for the effects in 15 years or longer, it is hard to draw any strong conclusions about long-run effects. Moreover, both methods show that the effect on the increase of divorce rates for the first 6 years are significant, which is consistent with previous findings in this literature.

We also report the var(EPC)/var(PC), where var(A) calculates the estimated variance of the estimator using method A. It is clear from the table that EPC achieves almost 50% of efficiency gain relative to the regular PC method.

7. Conclusion

This paper studies the efficient estimation and inference of panel data models that involve estimating a large sparse covariance matrix. We propose an estimator that takes into account both heteroskedasticity and cross-sectional dependence, and is shown to be more efficient than existing methods in the literature. We require no serial correlations be present, and focus on the effect of cross-sectional correlations. The estimator is asymptotically unbiased with an improved rate of convergence when $N = o(T)$, compared with the usual estimator.

The literature on estimating high-dimensional sparse covariance matrices has targeted on the covariance and inverse covariance directly, and the theoretical results are mostly in an absolute convergence form. We see that the absolute convergence, even though achieving the minimax optimal rate, is not suitable for statistical inference in our case. Thus using an estimated large covariance as the optimal weight matrix is a nontrivial issue. We study a new notion of "doubly weighted convergence", and achieve the first-order asymptotic results as if the large inverse covariance were known.

We also discussed possible extensions to allow for serial correlations without introducing biases. Formal theoretical studies on this issue is left to the future work.

Acknowledgments

We are grateful to four anonymous referees, an Associate Editor and the Co-Editor for their insightful comments, which helped generate a substantially improved paper.

Appendix A. proofs

A.1. Estimation error of the inverse thresholding estimator

Throughout the proof, we denote $w_{ij} = (\Sigma_u^{-1})_{ij}$. We first prove that the estimated covariance matrix is consistent. The following theorem extends the result of Fan et al. (2013) to the panel data model:

Theorem A.1. *Under the Assumptions 3.1, 3.2 and 3.5, when $\|\Sigma_u^{-1}\|_1 = O(1)$, for $\omega_{T,N} = \sqrt{\frac{\log N}{T}} + \frac{1}{\sqrt{N}}$,*

$$\|\widehat{\Sigma}_u^{-1} - \Sigma_u^{-1}\|_1 = O_P(m_N \omega_{T,N}^{1-q}) = \|\widehat{\Sigma}_u - \Sigma_u\|_1.$$

Proof. By the assumption that the initial value $\widehat{\beta}_{\text{in}}$ satisfies $\|\widehat{\beta}_{\text{in}} - \beta_0\| = O_P(\frac{1}{\sqrt{NT}} + \frac{1}{T} + \frac{1}{N})$, it is not hard to show that when applying the PC method on $(Y_t - X_t\widehat{\beta}_{\text{in}})$ to estimate $\lambda_i' f_t$, the effect of estimating β_0 is asymptotically negligible. Hence the same proofs as those of Fan et al. (2013) yield, for $\omega_{T,N} = \sqrt{\frac{\log N}{T}} + \frac{1}{\sqrt{N}}$,

$$\max_{i\le N, j\le N} |\widetilde{R}_{ij} - \Sigma_{u,ij}| = \max_{i\le N, j\le N} \left| \frac{1}{T}\sum_{t=1}^{T} \widehat{u}_{it}\widehat{u}_{jt} - \Sigma_{u,ij} \right| = O_P(\omega_{T,N}). \quad (A.1)$$

By Theorem A.1 of Fan et al. (2013), we then have $\|\widehat{\Sigma}_u - \Sigma_u\|_1 = O_P(m_N\omega_{T,N}^{1-q})$. We now show the first statement. Note that

$$\begin{aligned}\|\widehat{\Sigma}_u^{-1} - \Sigma_u^{-1}\|_1 &\le \|(\widehat{\Sigma}_u^{-1} - \Sigma_u^{-1})(\widehat{\Sigma}_u - \Sigma_u)\Sigma_u^{-1}\|_1 \\ &\quad + \|\Sigma_u^{-1}(\widehat{\Sigma}_u - \Sigma_u)\Sigma_u^{-1}\|_1 \equiv a + b.\end{aligned}$$

We have

$$
\begin{aligned}
a &\leq \max_{j\leq N}\sum_{i,k,l\leq N}|(\widehat{\Sigma}_u^{-1}-\Sigma_u^{-1})_{ik}\|\Sigma_{u,kl}-\widehat{\Sigma}_{u,kl}\|\Sigma_{u,lj}^{-1}| \\
&\leq \max_{l}\sum_{i,k\leq N}|(\widehat{\Sigma}_u^{-1}-\Sigma_u^{-1})_{ik}||\Sigma_{u,kl}-\widehat{\Sigma}_{u,kl}|\max_{j\leq N}\sum_{l}|\Sigma_{u,lj}^{-1}| \\
&\leq \max_{l}\max_{k}\sum_{i\leq N}|(\widehat{\Sigma}_u^{-1}-\Sigma_u^{-1})_{ik}|\sum_{k\leq N}|\Sigma_{u,kl}-\widehat{\Sigma}_{u,kl}|\|\Sigma_u^{-1}\|_1 \\
&\leq \|\widehat{\Sigma}_u^{-1}-\Sigma_u^{-1}\|_1\|\Sigma_u^{-1}\|_1\max_{l}\sum_{k\leq N}|\Sigma_{u,kl}-\widehat{\Sigma}_{u,kl}| \\
&= \|\widehat{\Sigma}_u^{-1}-\Sigma_u^{-1}\|_1\|\Sigma_u^{-1}\|_1\|\Sigma_u-\widehat{\Sigma}_u\|_1 \\
&= O_P(m_N\omega_{T,N}^{1-q})\|\widehat{\Sigma}_u^{-1}-\Sigma_u^{-1}\|_1.
\end{aligned}
$$

In addition,

$$
\begin{aligned}
b &\leq \max_{j\leq N}\sum_{i,k,l\leq N}|\Sigma_{u,ik}^{-1}\|\Sigma_{u,kl}-\widehat{\Sigma}_{u,kl}\|\Sigma_{u,lj}^{-1}| \\
&\leq \|\Sigma_u^{-1}\|_1^2\|\widehat{\Sigma}_u-\Sigma_u\|_1 = O_P(m_N\omega_{T,N}^{1-q}).
\end{aligned}
$$

Hence we have $(1+o_P(1))\|\widehat{\Sigma}_u^{-1}-\Sigma_u^{-1}\|_1 = O_P(m_N\omega_{T,N}^{1-q})$, which implies the result. □

A.2. Consistency of $\widehat{\beta}$

Lemma A.1. (i) $\sup_{F'F/T=I_r}\frac{1}{NT}|\mathrm{tr}[\Sigma_u^{-1}u'(F_0F_0'-FF')u/T]| = o_P(1)$.
$\sup_{F'F/T=I_r}\frac{1}{NT}|\mathrm{tr}[\widehat{\Sigma}_u^{-1}u'(F_0F_0'-FF')u/T]| = o_P(1)$
(ii) $\sup_{F'F/T=I_r}\frac{1}{NT}\|vec(u)'(\Sigma_u^{-1}\otimes M_F)Z\| = o_P(1)$,
$\sup_{F'F/T=I_r}\frac{1}{NT}\|vec(u)'(\widehat{\Sigma}_u^{-1}\otimes M_F)Z\| = o_P(1)$.
(iii) $\sup_{F'F/T=I_r}\frac{1}{NT}|\mathrm{tr}(\Sigma_u^{-1}u'M_FF_0\Lambda_0')| = o_P(1)$,
$\sup_{F'F/T=I_r}\frac{1}{NT}|\mathrm{tr}(\widehat{\Sigma}_u^{-1}u'M_FF_0\Lambda_0')| = o_P(1)$.

Proof. The proof of Lemma A.1 is given in Appendix A.7.

Lemma A.2. *Under the assumptions of Theorem 4.1,*

$\|\widehat{\beta}-\beta_0\| = o_P(1)$.

Proof. Let $u=(u_1,\ldots,u_T)'$, and (F_0,Λ_0) denote the true factor and loading matrices. Concentrating out Λ, it can be shown that the estimated $\widehat{\beta}$ and $\widehat{F}$ satisfy:

$$
\begin{aligned}
(\widehat{\beta},\widehat{F}) &= \arg\min_{\beta,F'F=TI_r}\frac{1}{NT}\mathrm{tr}[\widehat{\Sigma}_u^{-1}(Y-X(\beta)')'M_F(Y-X(\beta)')] \\
&\quad -\frac{1}{NT}\mathrm{tr}(\widehat{\Sigma}_u^{-1}u'M_{F_0}u) \\
&= \arg\min_{\beta,F'F=TI_r} S(\beta,F)+R(\beta,F);
\end{aligned}
$$

here $X(\beta)'$ is a $T\times N$ matrix with elements of $X_{it}'\beta$, where β is a generic parameter in the parameter space of β_0. Also let

$$
\begin{aligned}
S(\beta,F) &= \frac{1}{NT}(\beta-\beta_0)'Z'(\widehat{\Sigma}_u^{-1}\otimes M_F)Z(\beta-\beta_0) \\
&\quad +\frac{2}{NT}\sum_{i,j\leq N}\widehat{\Sigma}_{u,ij}^{-1}F_0\lambda_{0j}'M_FX_i'(\beta-\beta_0) \\
&\quad +\frac{1}{NT}\mathrm{tr}(\widehat{\Sigma}_u^{-1}\Lambda_0F_0'M_FF_0\Lambda_0'), \\
R(\beta,F) &= \frac{2}{NT}\mathrm{tr}(\widehat{\Sigma}_u^{-1}u'M_FF_0\Lambda_0') \\
&\quad +\frac{2}{NT}vec(u)'(\widehat{\Sigma}_u^{-1}\otimes M_F)Z(\beta-\beta_0) \\
&\quad +\frac{1}{NT}\mathrm{tr}(\widehat{\Sigma}_u^{-1}u'(F_0F_0'-FF')/Tu).
\end{aligned}
$$

It can be further verified that, with $V(F)$ as defined in (3.1),

$$
\begin{aligned}
S(\beta,F) &= (\beta-\beta_0)'V(F)(\beta-\beta_0) \\
&\quad +(\eta+B^{-1}C(\beta-\beta_0))'B(\eta+B^{-1}C(\beta-\beta_0))\geq 0
\end{aligned}
$$

where $\eta=vec(M_FF_0)$, $B=(\frac{1}{N}\Lambda_0'\widehat{\Sigma}_u^{-1}\Lambda_0)\otimes I_T$, and

$$
C=\frac{1}{NT}\sum_{i,j\leq N}\widehat{\Sigma}_{u,ij}^{-1}[\lambda_{0j}\otimes M_F]X_i.
$$

By Lemma A.1, $\sup_{\beta,F'F=TI_r}|R(\beta,F)| = o_P(1)$. Since $S(\beta_0,F_0)=0$, we have

$$
S(\widehat{\beta},\widehat{F})\leq o_P(1)+S(\beta_0,F_0)=o_P(1),
$$

which implies $(\widehat{\beta}-\beta_0)'V(\widehat{F})(\widehat{\beta}-\beta_0)=o_P(1)$. The consistency of $\widehat{\beta}$ follows since $\inf_{F'F=TI_r}\lambda_{\min}(V(F))$ is bounded away from zero in probability.

A.3. Preliminary analysis for the limiting distribution of $\widehat{\beta}$

We can write

$$
\widehat{\beta}=\left(\sum_{i,j\leq N}\widehat{\Sigma}_{u,ij}^{-1}X_i'M_{\widehat{F}}X_j\right)^{-1}\left(\sum_{i,j\leq N}\widehat{\Sigma}_{u,ij}^{-1}X_i'M_{\widehat{F}}Y_j\right).
$$

Note that

$$
\sum_{i,j\leq N}\widehat{\Sigma}_{u,ij}^{-1}X_i'M_{\widehat{F}}X_j = Z'(\widehat{\Sigma}_u^{-1}\otimes M_{\widehat{F}})Z,
$$

hence with $Y_j=X_j\beta_0+F_0\lambda_{0j}+u_j$,

$$
\begin{aligned}
&\frac{1}{NT}Z'(\widehat{\Sigma}_u^{-1}\otimes M_{\widehat{F}})Z(\widehat{\beta}-\beta_0) \\
&\quad =\frac{1}{NT}\sum_{i,j\leq N}\widehat{\Sigma}_{u,ij}^{-1}X_i'M_{\widehat{F}}F_0\lambda_{0j}+\frac{1}{NT}\sum_{i,j\leq N}\widehat{\Sigma}_{u,ij}^{-1}X_i'M_{\widehat{F}}u_j \\
&\quad = I+II.
\end{aligned}
\tag{A.2}
$$

We evaluate I and II separately. From now on, we use Λ for Λ_0 to denote the true matrix of loading, without causing any confusion. Let

$$
A=\left(\frac{1}{NT}\Lambda'\widehat{\Sigma}_u^{-1}\Lambda F_0'\widehat{F}\right)^{-1}
$$

and V be a diagonal matrix of the r largest eigenvalues of

$$
\frac{1}{NT}(Y-X(\widehat{\beta}))'\widehat{\Sigma}_u^{-1}(Y-X(\widehat{\beta})),
$$

where $X(\widehat{\beta})$ is an $N\times T$ matrix $X(\widehat{\beta})=(X_1\widehat{\beta},\ldots,X_T\widehat{\beta})$. Since $M_{\widehat{F}}\widehat{F}=0$, we have

$$
I=\sum_{i,j\leq N}\widehat{\Sigma}_{u,ij}^{-1}X_i'M_{\widehat{F}}(F_0-\widehat{F}VA)\lambda_{0j}.
$$

Next, by the definition of the eigenvalues,

$$
\frac{1}{NT}(Y-X(\widehat{\beta}))'\widehat{\Sigma}_u^{-1}(Y-X(\widehat{\beta}))\widehat{F}=\widehat{F}V.
$$

We thus have

$$
\begin{aligned}
\widehat{F}VA-F_0 &= \frac{1}{NT}\big\{[X(\beta_0-\widehat{\beta})]'\widehat{\Sigma}_u^{-1}[X(\beta_0-\widehat{\beta})]\widehat{F} \\
&\quad +[X(\beta_0-\widehat{\beta})]'\widehat{\Sigma}_u^{-1}\Lambda F_0'\widehat{F}+[X(\beta_0-\widehat{\beta})]'\widehat{\Sigma}_u^{-1}u'\widehat{F} \\
&\quad +F_0\Lambda'\widehat{\Sigma}_u^{-1}[X(\beta_0-\widehat{\beta})]\widehat{F}+u\widehat{\Sigma}_u^{-1}[X(\beta_0-\widehat{\beta})]\widehat{F} \\
&\quad +F_0\Lambda'\widehat{\Sigma}_u^{-1}u'\widehat{F} \\
&\quad +u\widehat{\Sigma}_u^{-1}\Lambda F_0'\widehat{F}+u\widehat{\Sigma}_u^{-1}u'\widehat{F}\big\}A,
\end{aligned}
\tag{A.3}
$$

where $[X(\beta_0-\widehat{\beta})]$ is a $N\times T$ matrix with elements of $X_{it}'(\beta_0-\widehat{\beta})$.

Substituting into I, we thus have

$$I=\sum_{i=1}^{8}J_i.$$

We define and bound each J_i in the following lemmas (Lemmas A.3 and A.4). Substituting Lemmas A.3 and A.4 to (A.2), we obtain

$$\begin{aligned}&\frac{\sqrt{NT}}{NT}Z'(\widehat{\Sigma}_u^{-1}\otimes M_{\widehat{F}})Z(\widehat{\beta}-\beta_0)\\&=\frac{\sqrt{NT}}{NT}Z'\left[\left(\frac{1}{N}\widehat{\Sigma}_u^{-1}\Lambda\left(\frac{\Lambda'\widehat{\Sigma}_u^{-1}\Lambda}{N}\right)^{-1}\Lambda'\widehat{\Sigma}_u^{-1}\right)\otimes M_{\widehat{F}}\right]\\&\quad\times Z(\widehat{\beta}-\beta_0)\\&\quad-\frac{\sqrt{NT}}{NT}Z'\left[\left(\frac{1}{N}\widehat{\Sigma}_u^{-1}\Lambda\left(\frac{\Lambda'\widehat{\Sigma}_u^{-1}\Lambda}{N}\right)^{-1}\Lambda'\widehat{\Sigma}_u^{-1}\right)\otimes M_{\widehat{F}}\right]U\\&\quad+\frac{\sqrt{NT}}{NT}Z[\widehat{\Sigma}_u^{-1}\otimes M_{\widehat{F}}]U+O_P(\sqrt{NT}m_N\omega_{T,N}^{3-q})\\&\quad+o_P(\sqrt{NT}\|\widehat{\beta}-\beta_0\|).\end{aligned}$$

It follows from $\|\widehat{\Sigma}_u^{-1}-\Sigma_u^{-1}\|=O_P(m_N\omega_{T,N}^{1-q})$, $\|\Lambda\|=O(\sqrt{N})$, $\|Z\|^2=O_P(NT)$ that

$$\left\|\frac{\sqrt{NT}}{NT}Z'((\Sigma_u^{-1}-\widehat{\Sigma}_u^{-1})\otimes M_{\widehat{F}})Z(\widehat{\beta}-\beta_0)\right\|=o_P(\sqrt{NT}\|\widehat{\beta}-\beta_0\|).$$

$$\begin{aligned}&\frac{\sqrt{NT}}{NT}Z'\left[\left(\frac{1}{N}\Sigma_u^{-1}\Lambda\left(\frac{\Lambda'\Sigma_u^{-1}\Lambda}{N}\right)^{-1}\Lambda'\Sigma_u^{-1}\right)\otimes M_{\widehat{F}}\right]Z(\widehat{\beta}-\beta_0)\\&=\frac{\sqrt{NT}}{NT}Z'\left[\left(\frac{1}{N}\widehat{\Sigma}_u^{-1}\Lambda\left(\frac{\Lambda'\widehat{\Sigma}_u^{-1}\Lambda}{N}\right)^{-1}\Lambda'\widehat{\Sigma}_u^{-1}\right)\otimes M_{\widehat{F}}\right]\\&\quad\times Z(\widehat{\beta}-\beta_0)+o_P(\sqrt{NT}\|\widehat{\beta}-\beta_0\|).\end{aligned}$$

Therefore we have:

$$\begin{aligned}&\frac{\sqrt{NT}}{NT}Z'(\Sigma_u^{-1}\otimes M_{\widehat{F}})Z(\widehat{\beta}-\beta_0)\\&=\frac{\sqrt{NT}}{NT}Z'\left[\left(\frac{1}{N}\Sigma_u^{-1}\Lambda\left(\frac{\Lambda'\Sigma_u^{-1}\Lambda}{N}\right)^{-1}\Lambda'\Sigma_u^{-1}\right)\otimes M_{\widehat{F}}\right]\\&\quad\times Z(\widehat{\beta}-\beta_0)\\&\quad-\frac{\sqrt{NT}}{NT}Z'\left[\left(\frac{1}{N}\widehat{\Sigma}_u^{-1}\Lambda\left(\frac{\Lambda'\widehat{\Sigma}_u^{-1}\Lambda}{N}\right)^{-1}\Lambda'\widehat{\Sigma}_u^{-1}\right)\otimes M_{\widehat{F}}\right]U\\&\quad+\frac{\sqrt{NT}}{NT}Z'[\widehat{\Sigma}_u^{-1}\otimes M_{\widehat{F}}]U+O_P(\sqrt{NT}m_N\omega_{T,N}^{3-q})\\&\quad+o_P(\sqrt{NT}\|\widehat{\beta}-\beta_0\|)\\&=\frac{\sqrt{NT}}{NT}Z'\left[\left(\frac{1}{N}\Sigma_u^{-1}\Lambda\left(\frac{\Lambda'\Sigma_u^{-1}\Lambda}{N}\right)^{-1}\Lambda'\Sigma_u^{-1}\right)\otimes M_{\widehat{F}}\right]\\&\quad\times Z(\widehat{\beta}-\beta_0)\\&\quad+\frac{1}{\sqrt{NT}}Z'A(\widehat{\Sigma}_u^{-1},\widehat{F})U+O_P(\sqrt{NT}m_N\omega_{T,N}^{3-q})\\&\quad+o_P(\sqrt{NT}\|\widehat{\beta}-\beta_0\|)\qquad\text{(A.4)}\end{aligned}$$

where

$$A(\widehat{\Sigma}_u^{-1},\widehat{F})=\left[\widehat{\Sigma}_u^{-1}-\widehat{\Sigma}_u^{-1}\Lambda\left(\Lambda'\widehat{\Sigma}_u^{-1}\Lambda\right)^{-1}\Lambda'\widehat{\Sigma}_u^{-1}\right]\otimes M_{\widehat{F}}.$$

Subtracting $\frac{\sqrt{NT}}{NT}Z'[(\frac{1}{N}\Sigma_u^{-1}\Lambda\left(\frac{\Lambda'\Sigma_u^{-1}\Lambda}{N}\right)^{-1}\Lambda'\Sigma_u^{-1})\otimes M_{\widehat{F}}]Z(\widehat{\beta}-\beta_0)$ on both sides of (A.4), we have

$$\begin{aligned}&\frac{\sqrt{NT}}{NT}Z'[(\Sigma_u^{-1}-\Sigma_u^{-1}\Lambda\left(\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\Lambda'\Sigma_u^{-1})\otimes M_{\widehat{F}}]Z(\widehat{\beta}-\beta_0)\\&=\frac{1}{\sqrt{NT}}Z'A(\widehat{\Sigma}_u^{-1},\widehat{F})U+O_P(\sqrt{NT}m_N\omega_{T,N}^{3-q})+o_P(\sqrt{NT}\|\widehat{\beta}-\beta_0\|).\end{aligned}$$

Note that $A_{\widehat{F}}=A(\Sigma_u^{-1},\widehat{F})=[\Sigma_u^{-1}-\Sigma_u^{-1}\Lambda\left(\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\Lambda'\Sigma_u^{-1}]\otimes M_{\widehat{F}}$. So

$$\begin{aligned}\sqrt{NT}\frac{Z'A_{\widehat{F}}Z}{NT}(\widehat{\beta}-\beta_0)&=\frac{1}{\sqrt{NT}}Z'A(\widehat{\Sigma}_u^{-1},\widehat{F})U\\&\quad+O_P(\sqrt{NT}m_N\omega_{T,N}^{3-q})+o_P(\sqrt{NT}\|\widehat{\beta}-\beta_0\|).\qquad\text{(A.5)}\end{aligned}$$

We need to show that the effect of replacing $\widehat{A}:=A(\widehat{\Sigma}_u^{-1},\widehat{F})$ with $A_{F_0}\equiv A(\Sigma_u^{-1},F_0)$ on the right hand side of (A.5) is asymptotically negligible. This is achieved by Proposition 4.1, to be proved below.

Lemma A.3. *We have,*

(i) $J_1=-\frac{1}{N^2T^2}\sum_{i,j\le N}\widehat{\Sigma}_{u,ij}^{-1}X_i'M_{\widehat{F}}[X(\beta_0-\widehat{\beta})]'\widehat{\Sigma}_u^{-1}[X(\beta_0-\widehat{\beta})]\widehat{F}A\lambda_{0j}=O_P(\|\beta_0-\widehat{\beta}\|^2)$

(ii) $J_4=-\frac{1}{N^2T^2}\sum_{i,j\le N}\widehat{\Sigma}_{u,ij}^{-1}X_i'M_{\widehat{F}}F_0\Lambda'\widehat{\Sigma}_u^{-1}[X(\beta_0-\widehat{\beta})]\widehat{F}A\lambda_{0j}=o_P(\|\beta_0-\widehat{\beta}\|)$

(iii) $J_5=-\frac{1}{N^2T^2}\sum_{i,j\le N}\widehat{\Sigma}_{u,ij}^{-1}X_i'M_{\widehat{F}}u\widehat{\Sigma}_u^{-1}[X(\beta_0-\widehat{\beta})]\widehat{F}A\lambda_{0j}=o_P(\|\beta_0-\widehat{\beta}\|)$

(iv) $J_3=-\frac{1}{N^2T^2}\sum_{i,j\le N}\widehat{\Sigma}_{u,ij}^{-1}X_i'M_{\widehat{F}}[X(\beta_0-\widehat{\beta})]'\widehat{\Sigma}_u^{-1}u'\widehat{F}A\lambda_{0j}=o_P(\|\beta_0-\widehat{\beta}\|)$

(v) $J_8=-\frac{1}{N^2T^2}\sum_{i,j\le N}\widehat{\Sigma}_{u,ij}^{-1}X_i'M_{\widehat{F}}u\widehat{\Sigma}_u^{-1}u'\widehat{F}A\lambda_{0j}=o_P(\|\widehat{\beta}-\beta_0\|+\frac{1}{\sqrt{N}}m_N\omega_{T,N}^{2-q})$,

(vi) $J_6=-\frac{1}{N^2T^2}\sum_{i,j\le N}\widehat{\Sigma}_{u,ij}^{-1}X_i'M_{\widehat{F}}F_0\Lambda'\widehat{\Sigma}_u^{-1}u'\widehat{F}A\lambda_{0j}=O_P(m_N\omega_{T,N}^{1-q}(\|\widehat{\beta}-\beta_0\|^2+\frac{1}{N}+\frac{1}{T}))$.

Proof. (i) It follows immediately from that $\|\widehat{\Sigma}_u^{-1}\|=O_P(1)$, $\|X\|_F=O_P(\sqrt{NT})$, $\|A\|_F=O_P(1)$, $\|\widehat{F}\|_F=O_P(\sqrt{T})$ and $\|M_{\widehat{F}}\|_F=O_P(1)$.

(ii) Note that $M_{\widehat{F}}F_0=M_{\widehat{F}}(F_0-\widehat{F}VA)$ due to $M_{\widehat{F}}\widehat{F}=0$. Using the same proof as that of Proposition A.1 in Bai (2009) to investigate (A.3), we have $\frac{1}{T}\|F_0-\widehat{F}VA\|_F^2=O_P(\|\widehat{\beta}-\beta_0\|^2+\frac{1}{N}+\frac{1}{T})$, which implies $\|M_{\widehat{F}}F_0\|_F=o_P(\sqrt{T})$, and the desired result.

(iii) Note that

$$\begin{aligned}J_5&=-\frac{1}{N^2T^2}(X_1M_{\widehat{F}}u,\ldots,X_NM_{\widehat{F}}u)\\&\quad\times\{\widehat{\Sigma}_u^{-1}\otimes(\widehat{\Sigma}_u^{-1}[X(\beta_0-\widehat{\beta})]\widehat{F}A)\}vec(\Lambda).\end{aligned}$$

$\|vec(\Lambda)\|=O(\sqrt{N})$, $\|\widehat{\Sigma}_u^{-1}\otimes(\widehat{\Sigma}_u^{-1}[X(\beta_0-\widehat{\beta})]\widehat{F}A)\|=O_P(T\sqrt{N}\|\widehat{\beta}-\beta_0\|)$. For each $i\le N$,

$$\begin{aligned}\frac{1}{T}X_i'M_{\widehat{F}}u&=\frac{1}{T}\sum_{t=1}^{T}X_{it}u_t'-\frac{1}{T}X_i'\widehat{F}\frac{1}{T}\sum_{t=1}^{T}(\widehat{F}_t-(VA)^{-1}f_{0t})u_t'\\&\quad-\frac{1}{T}X_i'\widehat{F}\frac{1}{T}\sum_{t=1}^{T}(VA)^{-1}f_{0t}u_t'.\end{aligned}$$

In addition, $\max_{i,j\le N}\|\frac{1}{T}\sum_{t=1}^{T}X_{it}u_{jt}\|=O_P(\sqrt{\frac{\log N}{T}})=\max_{i,j\le N}\|\frac{1}{T}\sum_{t=1}^{T}f_{0t}u_{jt}\|$.

Hence $\|(X_1M_{\widehat{F}}u,\ldots,X_NM_{\widehat{F}}u)\|_F=O_P(TN(\|\widehat{\beta}-\beta_0\|+\omega_{T,N}))$, and

$$J_5=O_P(\|\widehat{\beta}-\beta_0\|^2+\omega_{T,N}\|\widehat{\beta}-\beta_0\|).$$

This then yields the desired result.

(iv) We have

$$\begin{aligned}\|u'\widehat{F}\|_F&\le\|u'F_0(VA)^{-1}\|_F+\|u'(\widehat{F}-F_0(VA)^{-1})\\&=O_P(T\sqrt{N}\|\widehat{\beta}-\beta_0\|+T+\sqrt{NT})\end{aligned}$$

which implies that $J_3 = O_P(\|\widehat{\beta}-\beta_0\|^2 + \|\widehat{\beta}-\beta_0\|(\frac{1}{\sqrt{N}}+\frac{1}{\sqrt{T}}))$.

(v) First, let

$$J_{80} = -\frac{1}{N^2T^2}\sum_{i,j\le N}\widehat{\Sigma}_{u,ij}^{-1}X_i'M_{\widehat{F}}u\Sigma_u^{-1}u'\widehat{F}A\lambda_{0j}.$$

Then due to $\frac{1}{T}\|F_0-\widehat{F}VA\|_F^2 = O_P(\|\widehat{\beta}-\beta_0\|^2+\frac{1}{N}+\frac{1}{T})$, we have

$$J_8 = J_{80} + o_P\left(\|\widehat{\beta}-\beta_0\| + \frac{1}{\sqrt{N}}m_N\omega_{T,N}^{2-q}\right).$$

Also, $Eu\Sigma_u^{-1}u' = NI_T$, due to the serial uncorrelation, so

$$J_{80} = -\frac{1}{N^2T^2}\sum_{i,j\le N}\widehat{\Sigma}_{u,ij}^{-1}X_i'M_{\widehat{F}}(u\Sigma_u^{-1}u' - Eu\Sigma_u^{-1}u')\widehat{F}A\lambda_{0j}.$$

The similar proof to that of Lemma A.5 of Bai (2009) yields

$$J_{80} = o_P(\|\beta_0-\widehat{\beta}\|) + O_P\left(\frac{1}{N\sqrt{N}}+\frac{1}{T\sqrt{N}}+\frac{1}{N\sqrt{T}}\right)$$

which implies the result.

(vi) By $\frac{1}{T}\|F_0-\widehat{F}VA\|_F^2 = O_P(\|\widehat{\beta}-\beta_0\|^2+\frac{1}{N}+\frac{1}{T})$, we have

$$\|M_{\widehat{F}}F_0\|_F = O_P\left(\sqrt{T}\|\widehat{\beta}-\beta_0\| + \sqrt{\frac{T}{N}}+1\right).$$

On the other hand, $\|\Lambda'\Sigma_u^{-1}u'\|_F = O_P(\sqrt{NT})$, $\|u'F_0\|_F = O_P(\sqrt{NT})$, and $\|\Lambda'\Sigma_u^{-1}u'F_0\|_F = O_P(\sqrt{NT})$ because $\|\frac{1}{\sqrt{NT}}\sum_{i\le N,t\le T}\lambda_i(\Sigma_u^{-1}u_t)_if_t'\|_F = O_P(1)$. We thus have

$$\begin{aligned}\|\Lambda'\widehat{\Sigma}_u^{-1}u'\widehat{F}\|_F &\le \|\Lambda'(\widehat{\Sigma}_u^{-1}-\Sigma_u^{-1})u'\widehat{F}\|_F + \|\Lambda'\Sigma_u^{-1}u'\widehat{F}\|_F\\ &\le \|\Lambda'(\widehat{\Sigma}_u^{-1}-\Sigma_u^{-1})u'(\widehat{F}-F_0(VA)^{-1})\|_F\\ &\quad + \|\Lambda'(\widehat{\Sigma}_u^{-1}-\Sigma_u^{-1})u'F_0(VA)^{-1}\|_F\\ &\quad + \|\Lambda'\Sigma_u^{-1}u'(\widehat{F}-F_0(VA)^{-1})\|_F + \|\Lambda'\Sigma_u^{-1}u'F_0(VA)^{-1}\|_F\\ &= O_P\left(N\sqrt{T}m_N\omega_{T,N}^{1-q}\left(\sqrt{T}\|\widehat{\beta}-\beta_0\|+\sqrt{\frac{T}{N}}+1\right)\right).\end{aligned}$$

This implies the desired result. □

In the lemma below, recall that II was defined in (A.2).

Lemma A.4. (i)

$$\begin{aligned}J_2 &= -\frac{1}{N^2T^2}\sum_{i,j\le N}\widehat{\Sigma}_{u,ij}^{-1}X_i'M_{\widehat{F}}[X(\beta_0-\widehat{\beta})]'\widehat{\Sigma}_u^{-1}\Lambda F_0'\widehat{F}A\lambda_{0j}\\ &= \frac{1}{NT}Z'\left[\left(\frac{1}{N}\widehat{\Sigma}_u^{-1}\Lambda\left(\frac{\Lambda'\widehat{\Sigma}_u^{-1}\Lambda}{N}\right)^{-1}\Lambda'\widehat{\Sigma}_u^{-1}\right)\otimes M_{\widehat{F}}\right]Z(\widehat{\beta}-\beta_0)\end{aligned}$$

(ii)

$$\begin{aligned}J_7 &= -\frac{1}{N^2T^2}\sum_{i,j\le N}\widehat{\Sigma}_{u,ij}^{-1}X_i'M_{\widehat{F}}u\widehat{\Sigma}_u^{-1}\Lambda F_0'\widehat{F}A\lambda_{0j}\\ &= \frac{-1}{NT}Z'\left[\left(\frac{1}{N}\widehat{\Sigma}_u^{-1}\Lambda\left(\frac{\Lambda'\widehat{\Sigma}_u^{-1}\Lambda}{N}\right)^{-1}\Lambda'\widehat{\Sigma}_u^{-1}\right)\otimes M_{\widehat{F}}\right]U\end{aligned}$$

(iii) $II \equiv \frac{1}{NT}\sum_{i,j\le N}\widehat{\Sigma}_{u,ij}^{-1}X_i'M_{\widehat{F}}u_j = \frac{1}{NT}Z[\widehat{\Sigma}_u^{-1}\otimes M_{\widehat{F}}]U$.

Proof. The proofs are just straightforward calculations. □

A.4. Proof of Proposition 4.1

For each $q\le\dim(\beta)$, let $X_q = (X_{it,q})_{N\times T}$,

$$R = I_T - \frac{1}{T}F_0(Ef_tf_t')^{-1}F_0',\quad G = \frac{1}{T}F^*F^{*'},\quad F^* = F_0(VA)^{-1}.$$

Lemma A.5. *When* $Q_j\in\{\Sigma_{u,j}^{-1'}X, -\Sigma_{u,j}^{-1'}(EX_tf_t')(Ef_tf_t')^{-1}F_0'\}$,

$$\frac{1}{\sqrt{NT}}\sum_{i=1}^NQ_ie_i\frac{1}{T}\sum_{t=1}^T(\widehat{u}_{it}^2-Eu_{it}^2) = o_P(1).$$

Proof. The proof of Lemma A.5 is given in Appendix A.7.

Lemma A.6. *When* $Q_j\in\{\Sigma_{u,j}^{-1'}X, -\Sigma_{u,j}^{-1'}(EX_tf_t')(Ef_tf_t')^{-1}F_0'\}$,

$$\frac{1}{\sqrt{NT}}\sum_{i\ne j,(i,j)\in S_l}Q_je_i(\widehat{\Sigma}_{u,ij}-\Sigma_{u,ij}) = o_P(1).$$

Proof. The proof of Lemma A.6 is given in Appendix A.7.

Lemma A.7. *For each* $q\le d=\dim(\beta)$ *and* $X_{q,i}' = (X_{i1,q},\ldots,X_{iT,q})$, $\frac{1}{\sqrt{NT}}\mathrm{tr}[\Sigma_u^{-1}(\Sigma_u-\widehat{\Sigma}_u)\Sigma_u^{-1}X_qRu'] = o_P(1)$.

Proof. To simplify notation, we assume $d=1$ and write $X=X_q=(X_{it})_{N\times T}$ without loss of generality. Let e_i' and $\Sigma_{u,j}^{-1}$ denote the ith row of $\Sigma_u^{-1}u$ and the jth column of Σ_u^{-1} respectively. Then

$$\begin{aligned}L &= \frac{1}{\sqrt{NT}}\mathrm{tr}[\Sigma_u^{-1}(\Sigma_u-\widehat{\Sigma}_u)\Sigma_u^{-1}X_qRu']\\ &= \frac{1}{\sqrt{NT}}\sum_{i=1}^N\sum_{j=1}^N(\Sigma_u-\widehat{\Sigma}_u)_{ij}\Sigma_{u,j}^{-1'}XRe_i\\ &= \frac{1}{\sqrt{NT}}\sum_{i=1}^N\sum_{j=1}^N(\Sigma_u-\widehat{\Sigma}_u)_{ij}\Sigma_{u,j}^{-1'}Xe_i\\ &\quad -\frac{1}{\sqrt{NT}}\sum_{i=1}^N\sum_{j=1}^N(\Sigma_u-\widehat{\Sigma}_u)_{ij}\Sigma_{u,j}^{-1'}(EX_tf_t')(Ef_tf_t')^{-1}F_0'e_i\\ &\quad +\frac{1}{\sqrt{NT}}\sum_{i=1}^N\sum_{j=1}^N(\Sigma_u-\widehat{\Sigma}_u)_{ij}\Sigma_{u,j}^{-1'}\left(EX_tf_t'\right.\\ &\quad \left.-\frac{1}{T}\sum_{t=1}^TX_tf_t'\right)(Ef_tf_t')^{-1}\sum_{s=1}^Tf_se_{is}\\ &= L_1+L_2+L_3.\end{aligned}$$

Let $\widetilde{X}_{jt} = \Sigma_{u,j}^{-1'}X_t$, then $\max_{j\le N}\|\frac{1}{T}\sum_{t=1}^T\widetilde{X}_{jt}f_t - E\widetilde{X}_{jt}f_t\| = O_P(\sqrt{\frac{\log N}{T}})$ because $\|\Sigma_u^{-1}\|_1 = O(1)$.

$$\begin{aligned}L_3 &\le O\left(\frac{N}{\sqrt{NT}}\right)\max_i\left\|\sum_{s=1}^Tf_se_{is}\right\|\max_{j\le N}\left\|\frac{1}{T}\sum_{t=1}^T\widetilde{X}_{jt}f_t\right.\\ &\quad \left.-E\widetilde{X}_{jt}f_t\right\|\|\Sigma_u-\widehat{\Sigma}_u\|_1\\ &= O_P\left((\log N)\sqrt{\frac{N}{T}}m_N\omega_{T,N}^{1-q}\right) = o_P(1).\end{aligned}$$

On the other hand, both L_1 and L_2 are of the form: for some $1\times T$ vector Q_j

$$L_{1,2} = \frac{1}{\sqrt{NT}}\sum_{i=1}^N\sum_{j=1}^N(\Sigma_u-\widehat{\Sigma}_u)_{ij}Q_je_i$$

where $Q_j = \Sigma_{u,j}^{-1'}X$ for L_1 and $Q_j = -\Sigma_{u,j}^{-1'}(EX_tf_t')(Ef_tf_t')^{-1}F_0'$ for L_2. Because $\|\Sigma_u^{-1}\| = O(1)$,

$$\max_{i,j}|Q_je_i| \le \max_{ij}\left|\sum_{t=1}^TQ_{jt}e_{it}\right| = O_P(\sqrt{T\log N}).$$

By definition, when $i \neq j$, $\widehat{\Sigma}_{u,ij} = 0$ if $|\frac{1}{T}\sum_{t=1}^{T}\widehat{u}_{it}\widehat{u}_{jt}| \leq \tau_{ij}\omega_{T,N}$, where τ_{ij} is the threshold constant, bounded away from both zero and infinity with probability approaching one. For any $C > 0$, one can pick up a threshold constant in τ_{ij} such that $P(\tau_{ij} > C) \to 1$.

$$\begin{aligned}L_{1,2} &= \frac{1}{\sqrt{NT}}\sum_{i=1}^{N} Q_i e_i \frac{1}{T}\sum_{t=1}^{T}(\widehat{u}_{it}^2 - Eu_{it}^2) \\ &+ \frac{1}{\sqrt{NT}}\sum_{i\neq j,(i,j)\in S_l} Q_j e_i(\widehat{\Sigma}_{u,ij} - \Sigma_{u,ij}) \\ &+ \frac{1}{\sqrt{NT}}\sum_{(i,j)\in S_s} Q_j e_i(\widehat{\Sigma}_{u,ij} - \Sigma_{u,ij}).\end{aligned}$$

The first and second terms are bounded in Lemmas A.5 and A.6 below, which are $o_P(1)$. We now look at the third term. On one hand,

$$\begin{aligned}\frac{1}{\sqrt{NT}}\sum_{(i,j)\in S_s}|Q_j e_i||\Sigma_{u,ij}| &= O_P\left(\sqrt{\frac{T\log N}{NT}}\right)\sum_{(i,j)\in S_s}|\Sigma_{u,ij}| \\ &= O_P\left(\sqrt{\frac{\log N}{N}}\right).\end{aligned}$$

On the other hand, because $\max_{ij}|\frac{1}{T}\sum_{t=1}^{T}\widehat{u}_{it}\widehat{u}_{jt} - Eu_{it}u_{jt}| = O_P(\omega_{T,N})$ (see (A.1)), and $\max_{(i,j)\in S_s}|\Sigma_{u,ij}| = o(\omega_{T,N})$, then for any $\epsilon > 0$, one can pick up large enough $C > 0$ so that

$$\begin{aligned}&xP\left(|\frac{1}{\sqrt{NT}}\sum_{(i,j)\in S_s}|Q_j e_i||\widehat{\Sigma}_{u,ij}| > T^{-1}\right) \\ &\leq P\left(\max_{(i,j)\in S_s}|\widehat{\Sigma}_{u,ij}| > 0\right) \leq P(\exists(i,j)\in S_s, \left|\frac{1}{T}\sum_{t=1}^{T}\widehat{u}_{it}\widehat{u}_{jt}\right| > \tau_{ij}\omega_{T,N}) \\ &\leq P\left(\max_{ij}\left|\frac{1}{T}\sum_{t=1}^{T}\widehat{u}_{it}\widehat{u}_{jt}\right| > \omega_{T,N}C\right) + o(1) \\ &\leq P\left(\max_{ij}\left|\frac{1}{T}\sum_{t=1}^{T}\widehat{u}_{it}\widehat{u}_{jt} - \Sigma_{u,ij}\right| \right. \\ &\quad \left. + \max_{(i,j)\in S_s}|\Sigma_{u,ij}| > \omega_{T,N}C\right) + o(1) < \epsilon,\end{aligned}$$

which implies $|\frac{1}{\sqrt{NT}}\sum_{(i,j)\in S_s}|Q_j e_i||\widehat{\Sigma}_{u,ij}| = O_P(\frac{1}{T})$. Hence

$$\begin{aligned}&\left|\frac{1}{\sqrt{NT}}\sum_{(i,j)\in S_s} Q_j e_i(\widehat{\Sigma}_{u,ij} - \Sigma_{u,ij})\right| \\ &\leq \frac{1}{\sqrt{NT}}\sum_{(i,j)\in S_s}|Q_j e_i|(|\widehat{\Sigma}_{u,ij}| + |\Sigma_{u,ij}|) = o_P(1).\end{aligned}$$

Therefore, by Lemmas A.5 and A.6, we have $L_{1,2} = o_P(1)$ when either $Q_j = \Sigma_{u,j}^{-1'}X$ or $Q_j = -\Sigma_{u,j}^{-1'}(EX_t f_t')(Ef_t f_t')^{-1}F_0'$. This proves $L = o_P(1)$. □

Recall $H = I_T - \frac{1}{T}F_0(Ef_t f_t')^{-1}F_0'$, and $G = \frac{1}{T}F^* F^{*'}$ for $F^* = F_0(VA)^{-1}$.

Lemma A.8. *For each $q \leq d = \dim(\beta)$ and $X_{q,i}' = (X_{i1,q}, \ldots, X_{iT,q})$,*
(i) $\max_{i,j\leq N}|X_{q,i}' M_{\widehat{F}} u_j| = O_P(\sqrt{T\log N} + T\|\widehat{\beta} - \beta_0\| + \frac{T}{\sqrt{N}})$
(ii) $\max_{i,j}|X_{q,i}'(R - G)u_j| = O_P(\sqrt{T\log N}(\|\widehat{\beta} - \beta_0\| + \omega))$.

Proof. (i) The proof is a straightforward calculation, and very similar to that of Lemma A.3(iii).

(ii) Because $I_r = \frac{1}{T}\widehat{F}'\widehat{F}$, $Ef_t f_t' = O_P(\frac{1}{\sqrt{T}}) + \frac{1}{T}F_0'F_0$, and

$$\frac{1}{\sqrt{T}}\|\widehat{F} - F_0(VA)^{-1}\| = O_P\left(\|\widehat{\beta} - \beta_0\| + \frac{1}{\sqrt{N}} + \frac{1}{\sqrt{T}}\right),$$

we have

$$\begin{aligned}H - G &= \frac{1}{T}F_0((VA)^{-1}((VA)')^{-1} - (Ef_t f_t')^{-1})F_0' \\ &= O_P(\|\widehat{\beta} - \beta_0\| + \omega_{T,N})\frac{1}{\sqrt{T}}F_0'\end{aligned}$$

which implies the result since $\max_j \frac{1}{\sqrt{T}}\|F_0'u_j\| = O_P(\sqrt{\log N})$. □

Lemma A.9. *For each $q \leq d = \dim(\beta)$ and $X_{q,i}' = (X_{i1,q}, \ldots, X_{iT,q})$,*
(i) $\frac{1}{\sqrt{NT}}\mathrm{tr}\left[\left(\widehat{\Sigma}_u^{-1} - \Sigma_u^{-1}\right)\left(\Sigma_u - \widehat{\Sigma}_u\right)\Sigma_u^{-1}X_q M_{\widehat{F}}u'\right] = o_P(1) + o_P(\sqrt{NT}\|\widehat{\beta} - \beta_0\|)$
(ii) $\frac{1}{\sqrt{NT}}\mathrm{tr}[\Sigma_u^{-1}(\Sigma_u - \widehat{\Sigma}_u)\Sigma_u^{-1}X_q(G - R)u'] = o_P(1) + o_P(\sqrt{NT}\|\widehat{\beta} - \beta_0\|)$
(iii) $\frac{1}{\sqrt{NT}}\mathrm{tr}[\Sigma_u^{-1}(\Sigma_u - \widehat{\Sigma}_u)\Sigma_u^{-1}X_q(G - M_{\widehat{F}})u'] = o_P(1) + o_P(\sqrt{NT}\|\widehat{\beta} - \beta_0\|)$.

Proof. (i) By Theorem A.1,
$\|\widehat{\Sigma}_u - \Sigma_u\|_1 = O_P(m_N\omega_{T,N}^{1-q}) = \|\widehat{\Sigma}_u^{-1} - \Sigma_u^{-1}\|_1$. The term of interest is

$$\begin{aligned}&\frac{1}{\sqrt{NT}}\sum_{i,j,k\leq N}(\widehat{\Sigma}_u - \Sigma_u)_{ik}(\widehat{\Sigma}_u^{-1} - \Sigma_u^{-1})_{kj}(\Sigma_u^{-1}X_q M_{\widehat{F}}u')_{ji} \\ &\leq \|\Sigma_u^{-1}X_q M_{\widehat{F}}u'\|_{\max}\|\widehat{\Sigma}_u - \Sigma_u\|_1\|\widehat{\Sigma}_u^{-1} - \Sigma_u^{-1}\|_1\frac{N}{\sqrt{NT}} \\ &= O_P((\sqrt{N\log N} + \sqrt{NT}(\widehat{\beta} - \beta_0) + \sqrt{T})m_N^2\omega_{T,N}^{2-2q}),\end{aligned} \tag{A.6}$$

where we used $\|\Sigma_u^{-1}X_q M_{\widehat{F}}u'\|_{\max} = O_P(\sqrt{T\log N} + T\|\widehat{\beta} - \beta_0\| + \frac{T}{\sqrt{N}})$ by Lemma A.8 and $\|\Sigma_u^{-1}\|_1 = O(1)$. The desired result follows.

(ii) The objective is bounded by

$$\frac{N}{\sqrt{NT}}\|\Sigma_u^{-1}\|_1^2\|\Sigma_u - \widehat{\Sigma}_u\|_1\max_{i,j}|X_{q,i}'(R - G)u_j|.$$

The result then follows from Lemma A.8.

(iii) Recall the notation $e = \Sigma_u^{-1}u$ and $Q_j = \Sigma_{u,j}^{-1'}X_q$. We have

$$\max_{ij}\left|Q_j\frac{1}{T}(\widehat{F} - F^*)(\widehat{F} - F^*)'e_i\right| = O_P\left(T\|\widehat{\beta} - \beta_0\|^2 + \frac{T}{N} + 1\right),$$

$$\max_{ij}\left|Q_j\frac{1}{T}(\widehat{F} - F^*)(A'V)^{-1}F_0'e_i\right| = O_P\left(\sqrt{T}\|\widehat{\beta} - \beta_0\| + \sqrt{\frac{T}{N}} + 1\right).$$

Substituting

$$G - M_{\widehat{F}} = \frac{1}{T}(F^* - \widehat{F})F^{*'} - \frac{1}{T}(F^* - \widehat{F})(F^* - \widehat{F})' + F^*\frac{1}{T}(F^* - \widehat{F}),$$

and noting that $\|\widehat{\Sigma}_u - \Sigma_u\|_1 = O_P(m_N\omega_{T,N}^{1-q})$, we obtain

$$\begin{aligned}&\frac{1}{\sqrt{NT}}\mathrm{tr}[\Sigma_u^{-1}(\Sigma_u - \widehat{\Sigma}_u)\Sigma_u^{-1}X_q(G - M_{\widehat{F}})u'] \\ &= \frac{1}{\sqrt{NT}}\sum_{ij}(\Sigma_{u,ij} - \widehat{\Sigma}_{u,ij})Q_j(G - M_{\widehat{F}})e_i \\ &= \frac{1}{\sqrt{NT}}\sum_{ij}(\Sigma_{u,ij} - \widehat{\Sigma}_{u,ij})Q_j\frac{1}{T}F^*(F^* - \widehat{F})e_i \\ &\quad + o_P(1) + o_P(\sqrt{NT}\|\widehat{\beta} - \beta_0\|) \\ &\equiv B + o_P(1) + o_P(\sqrt{NT}\|\widehat{\beta} - \beta_0\|),\end{aligned}$$

where

$$B = \frac{1}{\sqrt{NT}}\sum_{ij}(\Sigma_{u,ij} - \widehat{\Sigma}_{u,ij})Q_j\frac{1}{T}F^*(F^* - \widehat{F})e_i.$$

We analyze $F^*-\widehat{F}$ in B using (A.3), and study it term by term. It is not difficult to obtain

$$Q_j\frac{1}{T}F^*(F^*-\widehat{F})e_i=-\frac{1}{T}\sum_{t=1}^{T}Q_{jt}f_t\frac{1}{NT}V^{-1}[\widehat{F}'F_0\Lambda'\widehat{\Sigma}_u^{-1}u'e_i+\widehat{F}'u\widehat{\Sigma}_u^{-1}u'e_i]+O_P(T\|\widehat{\beta}-\beta_0\|+\log N)$$
$$=B_1+B_2+O_P(T\|\widehat{\beta}-\beta_0\|+\log N),$$

where the $O_P(\cdot)$ term is uniform in $j,i\le N$. Term B_1 equals

$$-\frac{1}{T}\sum_{t=1}^{T}Q_{jt}f_t\frac{1}{NT}V^{-1}\widehat{F}'F_0\Lambda'\Big[(\widehat{\Sigma}_u^{-1}-\Sigma_u^{-1})(u'e_i-Eu'e_i)+(\widehat{\Sigma}_u^{-1}-\Sigma_u^{-1})Eu'e_i+\Sigma_u^{-1}(u'e_i-Eu'e_i)+\Sigma_u^{-1}Eu'e_i\Big]=\sum_{i=1}^{4}B_{1i}.$$

For $e'=\Sigma_u^{-1}u'$, the key observation is that $Eu'e_i=(0,\dots,T,\dots,0)'$, with the ith element being T and others being zero. Hence $\Lambda'\Sigma_u^{-1}Eu'e_i=O(T)$, which implies $B_{12}+B_{14}=O_P(\frac{T}{N}+\frac{T}{\sqrt{N}}m_N\omega_{T,N}^{1-q})$, and $B_{11}=O_P(m_N\omega_{T,N}^{1-q}\sqrt{T\log N})$, where the $O_P(\cdot)$ term is uniform in $j,i\le N$. Term B_2 can be treated similarly, and is easier. Combining these intermediate results (carrying over B_{13}), we obtain

$$\frac{1}{\sqrt{NT}}\mathrm{tr}[\Sigma_u^{-1}(\Sigma_u-\widehat{\Sigma}_u)\Sigma_u^{-1}X_q(G-M_{\widehat{F}})u']$$
$$=\frac{-1}{\sqrt{NT}}\sum_{ij}(\Sigma_{u,ij}-\widehat{\Sigma}_{u,ij})\frac{1}{T}\sum_{t=1}^{T}Q_{jt}f_t\times\frac{1}{NT}V^{-1}\widehat{F}'F_0\Lambda'\Sigma_u^{-1}(u'e_i-Eu'e_i)$$
$$+o_P\left(\sqrt{NT}\|\widehat{\beta}-\beta_0\|\right)+O_P\left(m_N\omega_{T,N}^{1-q}\sqrt{\frac{T}{N}}+m_N^2\omega_{T,N}^{2-2q}(\sqrt{T}+\sqrt{N\log N})\right)$$
$$\le O_P\left(\frac{m_N\omega_{T,N}^{1-q}}{N}\right)\sum_i\left\|\frac{1}{\sqrt{NT}}\Lambda'\Sigma_u^{-1}(u'e_i-Eu'e_i)\right\|+o_P(1)+o_P(\sqrt{NT}\|\widehat{\beta}-\beta\|).$$

Because $E\|\frac{1}{\sqrt{NT}}\Lambda'\Sigma_u^{-1}(u'e_i-Eu'e_i)\|^2=O(1)$, we complete the proof. □

Lemma A.10. *We have*

(i) $\|\frac{1}{\sqrt{NT}}Z'[(\widehat{\Sigma}_u^{-1}-\Sigma_u^{-1})\otimes M_{\widehat{F}}]U\|=o_P(1)+o_P(\sqrt{NT}\|\widehat{\beta}-\beta_0\|)$.

(ii) $\|\frac{1}{\sqrt{NT}}Z'\left\{[\frac{1}{N}\widehat{\Sigma}_u^{-1}\Lambda\left(\frac{\Lambda'\widehat{\Sigma}_u^{-1}\Lambda}{N}\right)^{-1}\Lambda'\widehat{\Sigma}_u^{-1})-\frac{1}{N}\Sigma_u^{-1}\Lambda\left(\frac{\Lambda'\Sigma_u^{-1}\Lambda}{N}\right)^{-1}\Lambda'\Sigma_u^{-1})]\otimes M_{\widehat{F}}\right\}U\|=o_P(1)+o_P(\sqrt{NT}\|\widehat{\beta}-\beta_0\|)$.

Proof. Consider part (i). The qth row ($q\le d$) of $\frac{1}{\sqrt{NT}}Z[(\widehat{\Sigma}_u^{-1}-\Sigma_u^{-1})\otimes M_{\widehat{F}}]U$ can be written as $\frac{1}{\sqrt{NT}}\mathrm{tr}[(\widehat{\Sigma}_u^{-1}-\Sigma_u^{-1})X_qM_{\widehat{F}}u']$ for $u=(u_{it})_{N\times T}$. In addition,

$$\frac{1}{\sqrt{NT}}\mathrm{tr}[(\widehat{\Sigma}_u^{-1}-\Sigma_u^{-1})X_qM_{\widehat{F}}u']$$
$$=\frac{1}{\sqrt{NT}}\mathrm{tr}[\Sigma_u^{-1}(\Sigma_u-\widehat{\Sigma}_u)\Sigma_u^{-1}X_qM_{\widehat{F}}u']+\frac{1}{\sqrt{NT}}\mathrm{tr}[(\widehat{\Sigma}_u^{-1}-\Sigma_u^{-1})(\Sigma_u-\widehat{\Sigma}_u)\Sigma_u^{-1}X_qM_{\widehat{F}}u']$$
$$=\frac{1}{\sqrt{NT}}\mathrm{tr}[\Sigma_u^{-1}(\Sigma_u-\widehat{\Sigma}_u)\Sigma_u^{-1}X_qRu']+\frac{1}{\sqrt{NT}}\mathrm{tr}[(\widehat{\Sigma}_u^{-1}-\Sigma_u^{-1})(\Sigma_u-\widehat{\Sigma}_u)\Sigma_u^{-1}X_qM_{\widehat{F}}u']$$
$$+\frac{1}{\sqrt{NT}}\mathrm{tr}[\Sigma_u^{-1}(\Sigma_u-\widehat{\Sigma}_u)\Sigma_u^{-1}X_q(M_{\widehat{F}}-G)u']+\frac{1}{\sqrt{NT}}\mathrm{tr}[\Sigma_u^{-1}(\Sigma_u-\widehat{\Sigma}_u)\Sigma_u^{-1}X_q(G-R)u'].$$

It follows from Lemmas A.7 and A.9 that the four terms on the right hand side are all $o_P(1)+o_P(\sqrt{NT}\|\widehat{\beta}-\beta_0\|)$, which concludes the proof for part (i). The proof of part (ii) is very similar to that of part (i). □

Recall the notation $A(\Sigma_u^{-1},\widehat{F})=A_{\widehat{F}}$, $A(\widehat{\Sigma}_u^{-1},\widehat{F})=\widehat{A}$ and $A(\Sigma_u^{-1},F_0)=A_{F_0}$.

Lemma A.11. $\frac{1}{\sqrt{NT}}Z'A_{\widehat{F}}U=\frac{1}{\sqrt{NT}}Z'A_{F_0}U+o_P(1)$.

Proof. We can prove the result for each row of Z'. Hence without loss of generality, we can simply assume $\dim(\beta)=1$, that is, Z' is a $1\times NT$ vector. Let $V=\Sigma_u^{-1/2}(u_1,\dots,u_t,\dots,u_T)$; $\widetilde{X}=\Sigma_u^{-1/2}(X_1,\dots,X_t,\dots,X_T)$. Let $f_t=F_{0t}$ denote the true vector of factors. Finally, for $P_F=F(F'F)^{-1}F'$, $K=I-\Sigma_u^{-1/2}\Lambda(\Lambda'\Sigma_u^{-1}\Lambda)^{-1}\Lambda'\Sigma_u^{-1/2}$. Then $A_{\widehat{F}}-A_{F_0}=(\Sigma_u^{-1/2}K\Sigma_u^{-1/2})\otimes(P_{F_0}-P_{\widehat{F}})$. Using the formula for Kronecker product $vec(ABC)=(C'\otimes A)vec(B)$ and $vec(A)'vec(B)=\mathrm{tr}(A'B)$, we reach

$$\frac{1}{\sqrt{NT}}Z'(A_{\widehat{F}}-A_{F_0})U=\frac{1}{\sqrt{NT}}\mathrm{tr}((P_{F_0}-P_{\widehat{F}})V'K\widetilde{X}).$$

Let v_i' and X_i' be the ith row of V and $\widetilde{X}$, $i=1,\dots,N$. Then v_i's are uncorrelated across i and $Ev_{it}v_{kt}=1\{i=k\}$. Note that

$$\frac{1}{\sqrt{NT}}\mathrm{tr}((P_{F_0}-P_{\widehat{F}})V'\widetilde{X})=\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\widetilde{X}_i'(P_{F_0}-P_{\widehat{F}})v_i.$$

The same argument as in the proof of Lemma A.8 in Bai (2009) yields

$$\frac{1}{\sqrt{NT}}\mathrm{tr}((P_{F_0}-P_{\widehat{F}})V'\widetilde{X})=\sqrt{\frac{T}{N}}\frac{1}{N}\sum_{i=1}^{N}\frac{\widetilde{X}_i'F_0}{T}\left(\frac{F_0'F_0}{T}\right)^{-1}\left(\frac{\widetilde{\Lambda}'\widetilde{\Lambda}}{N}\right)^{-1}\widetilde{\lambda}_i+o_P(1)$$

where $\widetilde{\Lambda}=\Sigma_u^{-1/2}\Lambda=(\widetilde{\lambda}_1,\dots,\widetilde{\lambda}_N)'$.

In addition, still by the same argument as in the proof of Lemma A.8 in Bai (2009),

$$\frac{1}{\sqrt{NT}}\mathrm{tr}((P_{F_0}-P_{\widehat{F}})V'\Sigma_u^{-1/2}\Lambda(\Lambda'\Sigma_u^{-1}\Lambda)^{-1}\Lambda'\Sigma_u^{-1/2}\widetilde{X})=\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}w_i'(P_{F_0}-P_{\widehat{F}})v_i$$
$$=\sqrt{\frac{T}{N}}\frac{1}{N}\sum_{i=1}^{N}\frac{w_i'F_0}{T}\left(\frac{F_0'F_0}{T}\right)^{-1}\left(\frac{\widetilde{\Lambda}'\widetilde{\Lambda}}{N}\right)^{-1}\widetilde{\lambda}_i+o_P(1),$$

where $w_i=\frac{1}{N}\sum_{k=1}^{N}a_{ik}\widetilde{X}_k$, $a_{ik}=\widetilde{\lambda}_i'(\frac{\widetilde{\Lambda}'\widetilde{\Lambda}}{N})^{-1}\widetilde{\lambda}_k$. Let $G=\frac{1}{T}F_0(\frac{F_0'F_0}{T})^{-1}(\frac{\widetilde{\Lambda}'\widetilde{\Lambda}}{N})^{-1}$. Importantly, note that $\frac{1}{N}\sum_i\widetilde{\lambda}_ia_{ik}=\widetilde{\lambda}_k$. Hence

$\sum_i w_i' G\widetilde{\lambda}_i = \sum_k \widetilde{X}_k' G\widetilde{\lambda}_k$. This proves that

$$\sqrt{\frac{T}{N}}\frac{1}{N}\sum_{i=1}^N \frac{\widetilde{X}_i'F_0}{T}\left(\frac{F_0'F_0}{T}\right)^{-1}\left(\frac{\widetilde{\Lambda}'\widetilde{\Lambda}}{N}\right)^{-1}\widetilde{\lambda}_i - \sqrt{\frac{T}{N}}\frac{1}{N}\sum_{i=1}^N \frac{w_i'F_0}{T}\left(\frac{F_0'F_0}{T}\right)^{-1}\left(\frac{\widetilde{\Lambda}'\widetilde{\Lambda}}{N}\right)^{-1}\widetilde{\lambda}_i = 0.$$

Therefore, $\frac{1}{\sqrt{NT}}\mathrm{tr}((P_{F_0} - P_{\widehat{F}})V'K\widetilde{X}) = o_P(1)$.

Proof of Proposition 4.1. It follows from Lemmas A.10 and A.11 that

$$\frac{1}{\sqrt{NT}}Z'(\widehat{A} - A_{\widehat{F}})U = o_P(1) + o_P(\sqrt{NT}\|\widehat{\beta} - \beta_0\|),$$

$$\frac{1}{\sqrt{NT}}Z'(A_{\widehat{F}} - A_{F_0})U = o_P(1).$$

Hence

$$\frac{1}{\sqrt{NT}}Z'(\widehat{A} - A_{F_0})U = o_P(1) + o_P(\sqrt{NT}\|\widehat{\beta} - \beta_0\|). \tag{A.7}$$

It then follows from (A.5) that, (note that $V(\widehat{F}) = \frac{1}{NT}Z'A_{\widehat{F}}Z$)

$$\sqrt{NT}V(\widehat{F})(\widehat{\beta} - \beta_0) = \frac{1}{\sqrt{NT}}Z'A_{F_0}U + o_P(1) + o_P(\sqrt{NT}\|\widehat{\beta} - \beta_0\|),$$

which implies, by Assumption 3.3,

$$\sqrt{NT}(\widehat{\beta} - \beta_0) = V(\widehat{F})^{-1}\frac{1}{\sqrt{NT}}Z'A_{F_0}U + o_P(1) + o_P(\sqrt{NT}\|\widehat{\beta} - \beta_0\|). \tag{A.8}$$

This also implies that, still by Assumption 3.3, there is $C > 0$ so that

$$(1 + o_P(1))\sqrt{NT}\|\widehat{\beta} - \beta_0\| \le \left\|\frac{C}{\sqrt{NT}}Z'A_{F_0}U\right\| + o_P(1).$$

Because $\frac{1}{\sqrt{NT}}Z'A_{F_0}U = O_P(1)$ by Assumption 3.4, hence $\sqrt{NT}(\widehat{\beta} - \beta_0) = O_P(1)$. It then follows from (A.7) that $\frac{1}{\sqrt{NT}}Z'(\widehat{A} - A_{F_0})U = o_P(1)$.

A.5. Proof of Theorems 4.1 and 4.2

By (A.8), $\sqrt{NT}(\widehat{\beta} - \beta_0) = V(\widehat{F})^{-1}\frac{1}{\sqrt{NT}}Z'A_{F_0}U + o_P(1)$. In addition, the same proof of Lemma A.9(i) in Bai (2009) implies that $V(\widehat{F})^{-1} \to^P V(F_0)^{-1}$. We have

$$\sqrt{NT}(\widehat{\beta} - \beta_0) = V(F_0)^{-1}\frac{1}{\sqrt{NT}}Z'A_{F_0}U + o_P(1).$$

Hence $\|\widehat{\beta} - \beta_0\| = O_P(\frac{1}{\sqrt{NT}})$. When $T = o(N)$, the limiting distribution then follows immediately from Assumption 3.4. □

To prove $\widehat{\Gamma}^{-1} \to^P \Gamma^{-1}$ where $\widehat{\Gamma} = \frac{1}{NT}Z'\widehat{A}Z$, we recall (3.1):

$$A_F = \left[\Sigma_u^{-1} - \Sigma_u^{-1}\Lambda\left(\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\Lambda'\Sigma_u^{-1}\right] \otimes M_F,$$

$$V(F) = \frac{1}{NT}Z'A_FZ, \quad V(F_0) \to \Gamma.$$

We first show $\|\widehat{A} - A_{F_0}\| = o_P(1)$. First, it follows from the same arguments of those of Bai (2009) that $\|M_{\widehat{F}} - M_{F_0}\|_F = o_P(1)$ and $\frac{1}{N}\|\widehat{\Lambda} - \Lambda H\|^2 = o_P(1)$. Also, $\|M_F\| \le 1$ for any F because M_F is idempotent. And $\left[\Sigma_u^{-1} - \Sigma_u^{-1}\Lambda\left(\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\Lambda'\Sigma_u^{-1}\right] = \Sigma_u^{-1/2}B\Sigma_u^{-1/2}$, where $B = I - \Sigma_u^{-1/2}\Lambda\left(\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\Lambda'\Sigma_u^{-1/2}$ is also idempotent. Hence $\|\Sigma_u^{-1} - \Sigma_u^{-1}\Lambda\left(\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\Lambda'\Sigma_u^{-1}\| \le \|\Sigma_u^{-1}\|$, which is also bounded. Furthermore, by Theorem A.1, $\|\widehat{\Sigma}_u^{-1} - \Sigma_u^{-1}\| = o_P(1)$. Therefore,

$$\begin{aligned}
\|\widehat{A} - A_{F_0}\| &\le \Big\|\left[\widehat{\Sigma}_u^{-1} - \widehat{\Sigma}_u^{-1}\widehat{\Lambda}\left(\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}\widehat{\Lambda}\right)^{-1}\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}\right] \\
&\quad - \left[\Sigma_u^{-1} - \Sigma_u^{-1}\Lambda\left(\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\Lambda'\Sigma_u^{-1}\right]\Big\| \\
&\quad + \|M_{\widehat{F}} - M_{F_0}\|_F \le o_P(1) + \|\widehat{\Sigma}_u^{-1}\widehat{\Lambda}\left(\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}\widehat{\Lambda}\right)^{-1}\widehat{\Lambda}'\widehat{\Sigma}_u^{-1} \\
&\quad - \Sigma_u^{-1}\Lambda\left(\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}\Lambda'\Sigma_u^{-1}\| \\
&\le \frac{1}{\sqrt{N}}\|\widehat{\Sigma}_u^{-1}\widehat{\Lambda} - \Sigma_u^{-1}\Lambda H\|\left\|\left(\frac{1}{N}\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}\widehat{\Lambda}\right)^{-1}\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}\right\|\frac{1}{\sqrt{N}} \\
&\quad + \left\|\frac{1}{\sqrt{N}}\Sigma_u^{-1}\Lambda H\right\|\Big\|\left(\frac{1}{N}\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}\widehat{\Lambda}\right)^{-1} \\
&\quad - H^{-1}\left(\frac{1}{N}\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}H'^{-1}\Big\|\,\|\widehat{\Lambda}'\widehat{\Sigma}_u^{-1}\|\frac{1}{\sqrt{N}} \\
&\quad + \left\|\frac{1}{\sqrt{N}}\Sigma_u^{-1}\Lambda\left(\frac{1}{N}\Lambda'\Sigma_u^{-1}\Lambda\right)^{-1}H'^{-1}\right\| \\
&\quad \times \left\|\frac{1}{\sqrt{N}}\right\|\widehat{\Sigma}_u^{-1}\widehat{\Lambda} - \Sigma_u^{-1}\Lambda H\| = o_P(1).
\end{aligned}$$

Hence

$$\|V(F_0) - \widehat{\Gamma}\| = \left\|\frac{1}{NT}Z'(\widehat{A} - A_{F_0})Z\right\|_F \le O_P(1)\|\widehat{A} - A_{F_0}\| = o_P(1).$$

It then follows from $V(F_0) \to \Gamma$ that $\widehat{\Gamma} \to^P \Gamma$. Thus $\widehat{\Gamma}^{-1} \to^P \Gamma^{-1}$.

A.6. Proof of Remark 3.6

We now prove Remark 3.6, which is restated as follows.

Suppose u_t is distributed as $\mathcal{N}(0, \Sigma_u)$, and $\{u_t, f_t, X_t\}$ are serially independent across t. Then, under $N = o(T^2)$, we have

$$\frac{1}{T\sqrt{NT}}\sum_{i=1}^N\sum_{s=1}^T(u_{is}^2 - Eu_{is}^2)\sum_{t=1}^T Q_{it}e_{it} = o_P(1), \tag{A.9}$$

$$\frac{1}{T\sqrt{NT}}\sum_{i\neq j,(i,j)\in S_l}\sum_{s=1}^T(u_{is}u_{js} - Eu_{is}u_{js})\sum_{t=1}^T Q_{jt}e_{it} = o_P(1). \tag{A.10}$$

The proof below holds for any initial estimator $\widehat{\beta}_{\text{in}}$ such that $\|\widehat{\beta}_{\text{in}} - \beta\| = O_P(\frac{1}{\sqrt{NT}} + \frac{1}{N} + \frac{1}{T})$, which holds for both the two-step algorithm and the multi-step algorithm.

A.6.1. Proof of (A.9)

Let

$$G = \frac{1}{T\sqrt{NT}}\sum_{i=1}^N\sum_{s=1}^T(u_{is}^2 - Eu_{is}^2)\sum_{t=1}^T Q_{it}e_{it}.$$

We respectively show that $|EG| = o(1)$ and $\mathrm{var}(G) = o(1)$, which will then imply $G = o_P(1)$.

Expectation

Because of the normality, we have $Eu_{is}^2e_{is} = 0$. Also Q_{is} and u_s are independent. Hence

$$EG = \frac{1}{T\sqrt{NT}}\sum_{i=1}^N\sum_{s=1}^T E(u_{is}^2 - Eu_{is}^2)Q_{is}e_{is} = \frac{1}{T\sqrt{NT}}\sum_{i=1}^N\sum_{s=1}^T E(u_{is}^2 - Eu_{is}^2)e_{is}EQ_{is} = 0.$$

Variance

$$\mathrm{var}(G) = \frac{1}{T^3N}\sum_{i=1}^N \mathrm{var}\left[\sum_{s=1}^T (u_{is}^2 - Eu_{is}^2)\sum_{t=1}^T Q_{it}e_{it}\right] + \frac{1}{T^3N}\sum_{i\neq j}\mathrm{cov}\left(\sum_{s=1}^T (u_{is}^2 - Eu_{is}^2)\sum_{t=1}^T Q_{it}e_{it}, \sum_{s=1}^T (u_{js}^2 - Eu_{js}^2) \times \sum_{t=1}^T Q_{jt}e_{jt}\right) \equiv A_1 + A_2.$$

By the Cauchy–Schwarz inequality,

$$A_1 \leq \frac{1}{T^3N}\sum_{i=1}^N E\left[\sum_{s=1}^T (u_{is}^2 - Eu_{is}^2)\sum_{t=1}^T Q_{it}e_{it}\right]^2$$

$$\leq \frac{1}{TN}\sum_{i=1}^N \left[E\left(\frac{1}{\sqrt{T}}\sum_{s=1}^T (u_{is}^2 - Eu_{is}^2)\right)^4\right]^{1/2} \times \left[E\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T Q_{it}e_{it}\right)^4\right]^{1/2} = O\left(\frac{1}{T}\right).$$

$$A_2 = \frac{1}{T^3N}\sum_{i\neq j}\sum_{s,t,k,l\leq T}\mathrm{cov}((u_{is}^2 - Eu_{is}^2)Q_{it}e_{it}, (u_{jk}^2 - Eu_{jk}^2)Q_{jl}e_{jl})$$

$$= \frac{1}{T^3N}\sum_{i\neq j}\sum_{s,t,k,l\leq T} EQ_{it}Q_{jl}\mathrm{cov}((u_{is}^2 - Eu_{is}^2)e_{it}, (u_{jk}^2 - Eu_{jk}^2)e_{jl})$$

$$\equiv \frac{1}{T^3N}\sum_{i\neq j}\sum_{s,t,k,l\leq T} EQ_{it}Q_{jl}C_{ij,stkl}$$

where $C_{ij,stkl} = \mathrm{cov}((u_{is}^2 - Eu_{is}^2)e_{it}, (u_{jk}^2 - Eu_{jk}^2)e_{jl})$. We have, $Ee_{it}u_{js}^2 = 0$ for any $i, j \leq N, t, s \leq T$. Also, Q_{jt} is independent of (u_t, e_t), and $\{Q_t, u_t\}_{t\leq T}$ is serially independent. Therefore, it is easy to verify that for fixed four integers s, t, k, l, if the set $\{s, t, k, l\}$ contains more than two distinct elements, $C_{ij,stkl} = 0$. Hence if we denote Θ as the set of (s, t, k, l) such that $\{s, t, k, l\}$ contains no more than two distinct elements, then its cardinality satisfies $|\Theta|_0 = O(T^2)$, and

$$\sum_{s,t,k,l\leq T} C_{ij,stkl} = \sum_{(s,t,k,l)\in\Theta} C_{ij,stkl}.$$

Let us partition Θ into $\Theta_1 \cup \Theta_2$ where each element $(s, t, k, l) \in \Theta_1$ contains exactly two distinct integers and each element in Θ_2 contains just one integer (that is, $s = t = k = l$ if $(s, t, k, l) \in \Theta_2$). We know that $\sum_{(s,t,k,l)\in\Theta_2} C_{ij,stkl} = O(T)$. Hence

$$A_2 = \frac{1}{T^3N}\sum_{i\neq j}\sum_{(s,t,k,l)\in\Theta_1} C_{ij,stkl} + O\left(\frac{N}{T^2}\right).$$

Because $Eu_{is}^2e_{js} = 0$ regardless of (i, j), so

$$A_2 = \frac{1}{T^3N}\sum_{i\neq j}\sum_{s=1}^T\sum_{t=1}^T [E(u_{is}^2 - Eu_{is}^2)(u_{js}^2 - Eu_{js}^2)] \times Ee_{it}e_{jt}EQ_{it}Q_{jt} + O\left(\frac{N}{T^2}\right).$$

Note that $Ee_{it}e_{jt} = (\Sigma_u^{-1})_{ij}$, and $\|\Sigma_u^{-1}\|_1 = O(1)$. In addition, by Assumption 3.5(i), (3.3) should hold. It then implies $N = o(T^2)$ and $T = o(N^2)$ (see Remark 3.4). Hence $A_2 = O(\frac{T+N}{T^2}) = o(1)$. This implies $\mathrm{var}(G) = o(1)$, and hence $G = o_P(1)$.

A.6.2. Proof of (A.10)

Let

$$M = \frac{1}{T\sqrt{NT}}\sum_{i\neq j,(i,j)\in S_l}\sum_{s=1}^T (u_{is}u_{js} - Eu_{is}u_{js})\sum_{t=1}^T Q_{jt}e_{it}.$$

Expectation For Gaussian errors, $Eu_{is}u_{js}e_{is} = 0$. Hence $EM = 0$.

Variance Let $\alpha_{ijs} = u_{is}u_{js} - Eu_{is}u_{js}$. We have,

$$\mathrm{var}(M) = \frac{1}{T^3N}\sum_{i\neq j,(i,j)\in S_l}\sum_{m\neq n,(m,n)\in S_l,(m,n)\neq(i,j)}\mathrm{cov}\left(\sum_{s=1}^T\alpha_{ijs}\sum_{t=1}^T Q_{jt}e_{it}, \times\sum_{s=1}^T\alpha_{mns}\sum_{t=1}^T Q_{mt}e_{nt}\right) + \frac{1}{T^3N}\sum_{i\neq j,(i,j)\in S_l}\mathrm{var}\left(\sum_{s=1}^T\alpha_{ijs}\sum_{t=1}^T Q_{jt}e_{it}\right) \equiv B_2 + B_1.$$

Using the Cauchy–Schwarz inequality like in the proof of the first statement. Similarly we can show $B_1 = O_P(\frac{1}{T})$. For B_2, let

$$C_{ijmn,stkl} = \mathrm{cov}(\alpha_{ijs}Q_{jt}e_{it}, \alpha_{mnk}Q_{ml}e_{nl}),$$

$$B_2 = \frac{1}{T^3N}\sum_{i\neq j,(i,j)\in S_l}\sum_{m\neq n,(m,n)\in S_l,(m,n)\neq(i,j)}\sum_{stkl\leq T} C_{ijmn,stkl}.$$

It is straightforward to check that when $\{s, t, k, l\}$ contains more than two distinct elements, $C_{ijmn,stkl} = 0$. In addition, $\sum_{i\neq j,(i,j)\in S_l} 1 = O(N)$. Define Θ_1 as the set of (s, t, k, l) such that $\{s, t, k, l\}$ contains exactly two distinct integers. Then

$$B_2 = \frac{1}{T^3N}\sum_{i\neq j,(i,j)\in S_l}\sum_{m\neq n,(m,n)\in S_l,(m,n)\neq(i,j)}\sum_{(s,t,k,l)\in\Theta_1} C_{ijmn,stkl} + O\left(\frac{N}{T^2}\right).$$

Moreover, because $\{u_t, Q_t\}_{t\leq T}$ is serially independent, $EQ_{is}e_{js} = 0$ and $Eu_{is}u_{js}e_{ns} = 0$ for all $i, j, n \leq N, s \leq T$, and $m_N = \max_{i\leq N}\sum_{j=1}^N 1_{\Sigma_{u,ij}\neq 0} = \max_{i\leq N}\sum_{j:(i,j)\in S_l} 1$, we have

$$B_2 = \frac{1}{T^3N}\sum_{i\neq j,(i,j)\in S_l}\sum_{m\neq n,(m,n)\in S_l,(m,n)\neq(i,j)}\sum_{s=1}^T\sum_{t=1}^T (E\alpha_{ijs}\alpha_{mns}) \times (Ee_{it}e_{nt})(EQ_{mt}Q_{jt}) + o(1)$$

$$\leq \frac{1}{TN}\max_{ijmnst}|E\alpha_{ijs}\alpha_{mns}||EQ_{mt}Q_{jt}|\sum_{i=1}^N\sum_{n=1}^N |(\Sigma_u^{-1})_{in}| \times \sum_{m:(m,n)\in S_l}\sum_{j:(i,j)\in S_l} 1 + o(1)$$

$$\leq O\left(\frac{m_N^2 N}{TN}\right)\|\Sigma_u^{-1}\| + o(1) = O\left(\frac{m_N^2}{T}\right) + o(1) = o(1).$$

Therefore, $\mathrm{var}(M) = B_1 + B_2 = o(1)$. This then implies (with $EM = 0$) that $M = o_P(1)$.

A.7. Proofs of technical lemmas

Proof of Lemma A.1. (i)

$$\left(\frac{1}{NT}\|u\Sigma_u^{-1}u'\|_F\right)^2 \leq \frac{2}{N^2T^2}\sum_{s,t\leq T}(u_t'\Sigma_u^{-1}u_s - Eu_t'\Sigma_u^{-1}u_s)^2 + \frac{2}{N^2T^2}\sum_{s,t\leq T}(Eu_t'\Sigma_u^{-1}u_s)^2.$$

With $W = \Sigma_u^{-1}$, $\frac{2}{N^2T^2}\sum_{s,t\le T}(u_t'\Sigma_u^{-1}u_s - Eu_t'\Sigma_u^{-1}u_s)^2 = O_P(\frac{1}{N})$. Also,

$$\frac{2}{N^2T^2}\sum_{s,t\le T}(Eu_t'\Sigma_u^{-1}u_s)^2 \le \frac{1}{T^2}\sum_{s,t\le T}\left|\frac{1}{N}\sum_{i,j\le N}w_{ij}Eu_{jt}u_{is}\right|^2$$
$$\le \frac{1}{T^2}\|\Sigma_u^{-1}\|_1^2\max_{i,j,s,t}|Eu_{jt}u_{is}|\max_{i,j\le N}\sum_{s,t\le T}|Eu_{jt}u_{is}| = O\left(\frac{1}{T}\right)$$

which is due to $\|\Sigma_u^{-1}\|_1 < \infty$ and $\max_{t\le T,i,j\le N}\sum_{s=1}^T|Eu_{jt}u_{is}| < \infty$. Therefore, using the inequality $|\mathrm{tr}(AB)| \le \|A\|_F\|B\|_F$, we have

$$\sup_{F'F/T=I_r}\frac{1}{NT}\mathrm{tr}[\Sigma_u^{-1}u'(F_0F_0' - FF')u/T]$$

$$\le \frac{1}{NT}\|u\Sigma_u^{-1}u'\|_F\sup_{F'F/T=I_r}\left\|\frac{1}{T}FF'\right\|_F = O_P\left(\frac{1}{\sqrt{N}}+\frac{1}{\sqrt{T}}\right).$$

For the second statement, since $\|\Sigma_u^{-1} - \widehat{\Sigma}_u^{-1}\| = o_P(1)$, it then follows that

$$\sup_{F'F/T=I_r}\frac{1}{NT}|\mathrm{tr}[(\Sigma_u^{-1} - \widehat{\Sigma}_u^{-1})u'(F_0F_0' - FF')u/T]| = o_P(1),$$

which yields the result.

(ii) Recall $e_t = \Sigma_u^{-1}u_t$. Let $e_i = (e_{i1},\ldots,e_{iT})'$ for $i \le N$. Then

$$\frac{1}{NT}\|vec(u)'(\Sigma_u^{-1}\otimes M_F)Z\| = \frac{1}{NT}\left\|\sum_{i=1}^N e_i'M_FX_i\right\|.$$

Under the assumption that $E|\frac{1}{\sqrt{N}}(e_t'e_s - Ee_t'e_s)|^2 < \infty$, the same proof of that of Lemma A.1 in Bai (2009) still goes through, which yields the result. The second statement follows immediately from $\|\Sigma_u^{-1} - \widehat{\Sigma}_u^{-1}\| = o_P(1)$.

(iii) By the definition of M_F, we bound

$$a_1 = \sup_{F'F/T=I_r}\frac{1}{NT}|\mathrm{tr}(\Sigma_u^{-1}u'F_0\Lambda_0')|$$

and

$$a_2 = \sup_{F'F/T=I_r}\frac{1}{NT}|\mathrm{tr}(\Sigma_u^{-1}u'FF'/TF_0\Lambda')|.$$

First, $a_1 \le \sup_{F'F=TI_r}\frac{1}{NT}\|\Lambda_0'\Sigma_u^{-1}\|\|u'F_0\|_F$, which is $o_P(1)$ since $\max_{i\le N}\|\frac{1}{T}\sum_{t=1}^T u_{it}f_t\| = O_P(\sqrt{\frac{\log N}{T}})$. On the other hand, a_2 is bounded by $O_P(\frac{1}{N\sqrt{T}})\|\Lambda_0'\Sigma_u^{-1}u'\|_F$, which is $o_P(1)$. Again, we conclude the proof by noting that $\|\Sigma_u^{-1} - \widehat{\Sigma}_u^{-1}\| = o_P(1)$. □

Proof of Lemma A.5. First we emphasize that $\widehat{u}_{it} = y_{it} - X_{it}'\widehat{\beta}_{\mathrm{in}} - \widehat{\lambda}_i'\widehat{f}_t$, where $(\widehat{\beta}_{\mathrm{in}},\widehat{\lambda}_i,\widehat{f}_t)$ are obtained in the first-step estimation (that is, by the method of Bai, 2009). Throughout Lemmas A.5 and A.6, these notation have the same meanings, without causing confusions. It is easy to show: there is an invertible matrix H so that (for $\omega_{T,N} = \frac{1}{\sqrt{N}} + \sqrt{\frac{\log N}{T}}$)

$$\frac{1}{T}\sum_{t=1}^T(\widehat{f}_t - Hf_t)^2 = O_P(\omega_{T,N}^2),$$

$$\max_{i\le N}\frac{1}{T}\sum_{t=1}^T(\widehat{u}_{it} - u_{it})^2 = O_P(\omega_{T,N}^2),$$

$$\max_{i\le N}|\widehat{\lambda}_i - H'^{-1}\lambda_i| = O_P(\omega_{T,N}),$$

$$\widehat{f}_s - Hf_s = \frac{1}{TN}\sum_{t=1}^T\widehat{f}_t(u_t'u_t + f_t\Lambda'u_s + f_s'\Lambda'u_t) + R_s \tag{A.11}$$

where the remaining term R_s depends on $\widehat{\beta}_{\mathrm{in}} - \beta$, which can be negligible because it is $O_P(\frac{1}{\sqrt{NT}} + \frac{1}{N})$ uniformly in s. The proof for the above results follows exactly the same lines as those of Fan et al. (2013), noting that the effect of estimating β_0 by $\widehat{\beta}_{\mathrm{in}}$ is asymptotically negligible because $\|\widehat{\beta}_{\mathrm{in}} - \beta\| = O_P(\frac{1}{\sqrt{NT}} + \frac{1}{N})$ in the absence of serial correlations, according to Bai (2009).

Now $\frac{1}{\sqrt{NT}}\sum_{i=1}^N Q_ie_i\frac{1}{T}\sum_{t=1}^T(\widehat{u}_{it}^2 - Eu_{it}^2)$ is bounded by

$$\left|\frac{1}{\sqrt{NT}}\sum_{i=1}^N Q_ie_i\frac{1}{T}\sum_{t=1}^T(\widehat{u}_{it}^2 - u_{it}^2)\right| + \left|\frac{1}{\sqrt{NT}}\sum_{i=1}^N Q_ie_i\frac{1}{T}\sum_{t=1}^T(u_{it}^2 - Eu_{it}^2)\right|$$
$$\equiv B_1 + B_2.$$

Term $B_2 = o_P(1)$ follows from Assumption 3.6. Term B_1 is bounded by

$$\left|\frac{2}{\sqrt{NT}}\sum_{i=1}^N Q_ie_i\frac{1}{T}\sum_{s=1}^T(\widehat{u}_{is} - u_{is})u_{is}\right|$$
$$+\left|\frac{1}{\sqrt{NT}}\sum_{i=1}^N Q_ie_i\frac{1}{T}\sum_{t=1}^T(\widehat{u}_{it} - u_{it})^2\right| \equiv B_{11} + B_{12}.$$

Note that $\max_i|Q_ie_i| = \max_i|\sum_t Q_{it}e_{it}| = O_P(\sqrt{T\log N})$. So we have $B_{12} = O_P(\sqrt{N\log N}\omega_{T,N}^2) = o(1)$ given $N\log N = o(T^2)$.

It then suffices to show $B_{11} = o(1)$. This part is difficult, and we separate it into a number of steps. Note that with the initial estimator $\widehat{\beta}_{\mathrm{in}}$,

$$u_{is} - \widehat{u}_{is} = (\widehat{f}_s - Hf_s)'(\widehat{\lambda}_i - H'^{-1}\lambda_i) + (\widehat{f}_s - Hf_s)'H'^{-1}\lambda_i$$

$$+f_s'H'(\widehat{\lambda}_i - H'^{-1}\lambda_i) + X_{it}'(\widehat{\beta}_{\mathrm{in}} - \beta).$$

We consider these terms one by one. By Cauchy–Schwarz inequality,

$$\left|\frac{2}{\sqrt{NT}}\sum_{i=1}^N Q_ie_i\frac{1}{T}\sum_{s=1}^T(\widehat{f}_s - Hf_s)'(\widehat{\lambda}_i - H'^{-1}\lambda_i)u_{is}\right|$$

$$\le \max_i|\widehat{\lambda}_i - H'^{-1}\lambda_i|\max_i|Q_ie_i|\left(\frac{1}{T}\sum_s(\widehat{f}_s - Hf_s)^2\right)^{1/2}$$
$$\times\left(\frac{1}{T}\sum_s u_{is}^2\right)^{1/2}\frac{\sqrt{N}}{\sqrt{T}} = o_P(1).$$

Second, $|\frac{2}{\sqrt{NT}}\sum_{i=1}^N Q_ie_i\frac{1}{T}\sum_{s=1}^T u_{is}f_s'H'(\widehat{\lambda}_i - H'^{-1}\lambda_i)| = o_P(1)$ because $\max_i\|\frac{1}{T}\sum_{s=1}^T u_{is}f_s\| = O_P(\sqrt{\frac{\log N}{T}}) = o_P(1)$. Note that (3.3) (assumed in Assumption 3.5(i)) implies $N = o(T^2)$ and $T = o(N^2)$ (see Remark 3.4). Hence $\log N = o(T)$.

Also, the term involving $X_{it}'(\widehat{\beta}_{\mathrm{in}} - \beta)$ is negligible under $N = o(T^2)$ and $T = o(N^2)$ (because $\|\widehat{\beta}_{\mathrm{in}} - \beta\| = O_P(\frac{1}{\sqrt{NT}} + \frac{1}{N} + \frac{1}{T})$). We now work on the term of $(\widehat{f}_s - Hf_s)'H'^{-1}\lambda_i$. By the formula $\widehat{f}_s - Hf_s = \frac{1}{TN}\sum_{t=1}^T\widehat{f}_t(u_s'u_t + f_t\Lambda'u_s + f_s'\Lambda'u_t) + R_s$,

$$\left|\frac{2}{\sqrt{NT}}\sum_{i=1}^N Q_ie_i\frac{1}{T}\sum_{s=1}^T(\widehat{f}_s - Hf_s)'H'^{-1}\lambda_iu_{is}\right| \le \sum_{i=1}^4 C_i.$$

Using (A.11) and by adding and subtracting terms,

$$C_1 = \left|\frac{2}{\sqrt{NT}}\sum_{i=1}^N Q_ie_i\frac{1}{T}\sum_{s=1}^T\lambda_i'H^{-1}\frac{1}{TN}\sum_{t=1}^T\widehat{f}_tu_s'u_tu_{is}\right|$$

$$\le\left|\frac{2}{\sqrt{NT}}\sum_{i=1}^N Q_ie_i\lambda_i'H^{-1}\frac{1}{T^2N}\sum_{s=1}^T(\widehat{f}_s - Hf_s)u_{is}E(u_s'u_s)\right|$$

$$+\left|\frac{2}{\sqrt{NT}}\sum_{i=1}^{N}Q_ie_i\lambda_i'\frac{1}{TN}\frac{1}{T}\sum_{s=1}^{T}f_su_{is}E(u_s'u_s)\right|$$

$$+\left|\frac{2}{\sqrt{NT}}\sum_{i=1}^{N}Q_ie_i\frac{1}{T}\sum_{s=1}^{T}\lambda_i'u_{is}\frac{1}{TN}\sum_{t=1}^{T}f_t(u_s'u_t-Eu_s'u_t)\right|$$

$$+\left|\frac{2}{\sqrt{NT}}\sum_{i=1}^{N}Q_ie_i\frac{1}{T}\sum_{s=1}^{T}\lambda_i'H^{-1}u_{is}\right.$$

$$\left.\times\frac{1}{TN}\sum_{t=1}^{T}(\widehat{f}_t-Hf_t)(u_s'u_t-Eu_s'u_t)\right|$$

$$=\sum_{i=1}^{4}C_{1i}.$$

By Cauchy–Schwarz inequality, $C_{11}, C_{12}=o_P(1)$. Also,

$$C_{13}\le\max_{is}|u_{is}|O_P\left(\frac{N\sqrt{T}\log N}{NT}\right)$$

$$\frac{1}{T}\sum_{s=1}^{T}\left|\frac{1}{\sqrt{TN}}\sum_{t=1}^{T}f_t(u_s'u_t-Eu_s'u_t)\right|.$$

Note that

$$E\frac{1}{T}\sum_{s=1}^{T}\left|\frac{1}{\sqrt{TN}}\sum_{t=1}^{T}f_t(u_s'u_t-Eu_s'u_t)\right|$$

$$\le\left(E\left|\frac{1}{\sqrt{TN}}\sum_{t=1}^{T}f_t(u_s'u_t-Eu_s'u_t)\right|^2\right)^{1/2}=O(1).$$

So $C_{13}=O_P(\sqrt{\frac{\log N}{T}}(\log NT))=o_P(1)$. Since $E(\frac{1}{\sqrt{N}}(u_s'u_t-Eu_s'u_t)^2)=O(1)$, by Cauchy–Schwarz inequality, $C_{14}=o_P(1)$.

$$C_2=\left|\frac{2}{\sqrt{NT}}\sum_{i=1}^{N}Q_ie_i\frac{1}{T}\sum_{s=1}^{T}\lambda_i'H^{-1}\frac{1}{TN}\sum_{t=1}^{T}\widehat{f}_tf_t'\Lambda'u_su_{is}\right|$$

$$\le\left|\frac{2}{TN}\sum_{i=1}^{N}Q_ie_i\lambda_i'H^{-1}\frac{1}{T}\sum_{t=1}^{T}\widehat{f}_tf_t'\frac{1}{\sqrt{NT}}\sum_{s=1}^{T}\sum_{j=1}^{N}\lambda_j(u_{js}u_{is}-Eu_{js}u_{is})\right|$$

$$+\left|\frac{2}{\sqrt{NT}}\sum_{i=1}^{N}Q_ie_i\lambda_i'H^{-1}\frac{1}{TN}\sum_{t=1}^{T}\widehat{f}_tf_t'\sum_{j=1}^{N}\lambda_jEu_{js}u_{is}\right|=o_P(1).$$

The first term is $o_P(1)$ because $E\frac{1}{N}\sum_{i=1}^{N}\|\frac{1}{\sqrt{NT}}\sum_{s=1}^{T}\sum_{j=1}^{N}\lambda_j(u_{js}u_{is}-Eu_{js}u_{is})\|^2=O(1)$ and $\max_i|Q_ie_i|=O_P(\sqrt{T\log N})$. The second term is $o_P(1)$ because

$$\max_i\sum_j|Eu_{js}u_{is}|=\|\Sigma_u\|_1=O(1).$$

$$C_3=\left|\frac{2}{\sqrt{NT}}\sum_{i=1}^{N}Q_ie_i\frac{1}{T}\sum_{s=1}^{T}\lambda_i'H^{-1}\frac{1}{TN}\sum_{t=1}^{T}\widehat{f}_tu_t'\Lambda f_su_{is}\right|$$

$$\le\left|\frac{2}{\sqrt{NT}}\sum_{i=1}^{N}Q_ie_i\lambda_i'\frac{1}{TN}\sum_{t=1}^{T}f_tu_t'\Lambda\frac{1}{T}\sum_{s=1}^{T}f_su_{is}\right|$$

$$+\left|\frac{2}{\sqrt{NT}}\sum_{i=1}^{N}Q_ie_i\frac{1}{T}\sum_{s=1}^{T}\lambda_i'H^{-1}\right.$$

$$\left.\times\frac{1}{TN}\sum_{t=1}^{T}(\widehat{f}_t-Hf_t)u_t'\Lambda f_su_{is}\right|=o_P(1).$$

The last term involving R_s is negligible. This concludes the proof. □

Proof of Lemma A.6. The term of interest is bounded by

$$\left|\frac{1}{\sqrt{NT}}\sum_{i\neq j,(i,j)\in S_l}Q_je_i\left(\frac{1}{T}\sum_{t=1}^{T}u_{it}u_{jt}-\Sigma_{u,ij}\right)\right|$$

$$+\left|\frac{1}{\sqrt{NT}}\sum_{i\neq j,(i,j)\in S_l}Q_je_i\left(\frac{1}{T}\sum_{t=1}^{T}\widehat{u}_{it}\widehat{u}_{jt}-u_{it}u_{jt}\right)\right|$$

$$+\left|\frac{1}{\sqrt{NT}}\sum_{i\neq j,(i,j)\in S_l}Q_je_i\left(\frac{1}{T}\sum_{t=1}^{T}\widehat{u}_{it}\widehat{u}_{jt}-\widehat{\Sigma}_{u,ij}\right)\right|\equiv D_1+D_2+D_3.$$

Term $D_1=o_P(1)$ follows from Assumption 3.6. From now on, we consider the hard-thresholding, that is, for $i\neq j$,

$$\widehat{\Sigma}_{u,ij}=\frac{1}{T}\sum_{t=1}^{T}\widehat{u}_{it}\widehat{u}_{jt}I\left(\left|\frac{1}{T}\sum_{t=1}^{T}\widehat{u}_{it}\widehat{u}_{jt}\right|>\tau_{ij}\omega_{T,N}\right)$$

where τ_{ij} is the threshold constant such that $P(\tau_{ij}<C_1)\to 1$ for some $C_1>0$. General thresholding functions can be treated very similarly. For D_3, we have, for any $\epsilon>0$,

$$P(D_3>T^{-1})\le P\left(\max_{i\neq j,(i,j)\in S_l}\left|\frac{1}{T}\sum_{t=1}^{T}\widehat{u}_{it}\widehat{u}_{jt}-\widehat{\Sigma}_{u,ij}\right|>0\right)$$

$$\le P\left(\exists(i,j)\in S_l,\left|\frac{1}{T}\sum_{t=1}^{T}\widehat{u}_{it}\widehat{u}_{jt}\right|\le\tau_{ij}\omega_{T,N}\right)$$

$$\le P\left(\min_{(i,j)\in S_l}\left|\frac{1}{T}\sum_{t=1}^{T}\widehat{u}_{it}\widehat{u}_{jt}\right|\le C_1\omega_{T,N}\right)+o(1)$$

$$\le P\left(\min_{(ij)\in S_l}\left|\Sigma_{u,ij}\right|-\max_{ij}\left|\frac{1}{T}\sum_{t=1}^{T}\widehat{u}_{it}\widehat{u}_{jt}-\Sigma_{u,ij}\right|\right.$$

$$\left.\le C_1\omega_{T,N}\right)+o(1)\le\epsilon+o(1),$$

where we use the assumption that $\omega_{T,N}=o(\min_{(ij)\in S_l}|\Sigma_{u,ij}|)$. This proves $D_3=O_P(\frac{1}{T})$. The proof of D_2 follows the same lines of that of term B_1 in Lemma A.5, hence is omitted. □

References

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

Ahn, S., Lee, Y., Schmidt, P., 2001. GMM estimation of linear panel data models with time-varying individual effects. J. Econometrics 101, 219–255.

Allen, D.W., 1992. Marriage and divorce: comment. Amer. Econ. Rev. 82, 679–685.

Avarucci, M., Zaffaroni, P., 2012. Generalized least squares estimation of panel with common shocks.

Bai, J., 2003. Inferential theory for factor models of large dimensions. Econometrica 71, 135–171.

Bai, J., 2009. Panel data models with interactive fixed effects. Econometrica 77, 1229–1279.

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

Belloni, A., Chernozhukov, V., Hansen, C., 2014a. High-dimensional methods and inference on structural and treatment effects. J. Econ. Perspect. 29–50.

Belloni, A., Chernozhukov, V., Hansen, C., 2014b. Inference on treatment effects after selection among high-dimensional controls. Rev. Econom. Stud. 81 (2), 608–650.

Bickel, P., Levina, E., 2008a. Covariance regularization by thresholding. Ann. Statist. 36 (6), 2577–2604.

Bickel, P., Levina, E., 2008b. Regularized estimation of large covariance matrices. Ann. Statist. 36 (1), 199–227.

Breitung, Tenhofen, 2011. GLS estimation of dynamic factor models. J. Amer. Statist. Assoc. 106, 1150–1166.

Cai, T., Zhou, H., 2012. Optimal rates of convergence for sparse covariance matrix estimation. Ann. Statist. 40 (5), 2389–2420.

Chamberlain, G., 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.

Davidson, J., 1994. Stochastic Limit Theory: An Introduction for Econometricians: An Introduction for Econometricians. Oxford University Press, UK.

Doz, C., Giannone, D., Reichlin, L., 2011. A two-step estimator for large approximate dynamic factor models based on Kalman filtering. J. Econometrics 164 (1), 188–205.

El Karoui, N., 2008. Operator norm consistent estimation of large dimensional sparse covariance matrices. Ann. Statist. 36 (6), 2717–2756.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96 (456), 1348–1360.

Fan, J., Liao, Y., Mincheva, M., 2013. Large covariance estimation by thresholding principal orthogonal complements (with discussion). J. R. Stat. Soc. Ser. B Stat. Methodol. 75, 603–680.

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

Forni, M., Hallin, M., Lippi, M., Reichlin, L., 2005. The generalized dynamic factor model: one-sided estimation and forecasting. J. Amer. Statist. Assoc. 830–840.

Forni, M., Lippi, M., 2001. The generalized dynamic factor model: representation theory. Econometric Theory 17 (06), 1113–1141.

Freyberger, J., 2012. Nonparametric Panel Data Models with Interactive Fixed Effects, Working Paper.

Friedberg, L., 1998. Did unilateral divorce raise divorce rates? Evidence from panel data. Amer. Econ. Rev. 88, 608–627.

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

Harding, M., Lamarche, C., 2014. Estimating and testing a quantile regression model with interactive effects. J. Econometrics 178, 101–113.

Iwakura, H., Okui, R., 2014. Asymptotic efficiency in factor models and dynamic panel data models. Available at SSRN 2395722.

Kao, C., Trapani, L., Urga, G., 2012. Asymptotics for panel models with common shocks. Econometric Rev. 31 (4), 390–439.

Kim, D., Oka, T., 2014. Divorce law reforms and divorce rates in the U.S.: an interactive fixed effects approach. J. Appl. Econometrics 29, 231–245.

Lam, C., Fan, J., 2009. Sparsistency and rates of convergence in large covariance matrix estimation. Ann. Statist. 37, 42–54.

Ledoit, O., Wolf, M., 2012. Nonlinear shrinkage estimation of large-dimensional covariance matrices. Ann. Statist. 40, 1024–1060.

Li, H., Li, Q., Shi, Y., 2017. Determining the number of factors when the number of factors can increase with sample size. J. Econometrics 197 (1), 76–86.

Moon, R., Weidner, M., 2015. Linear regression for panel with unknown number of factors as interactive fixed effects. Econometrica 83, 1543–1579.

Pesaran, H., 2006. Estimation and inference in large heterogeneous panels with a multifactor error structure. Econometrica 74, 967–1012.

Peters, H., 1986. Marriage and divorce: informational constraints and private contracting. Amer. Econ. Rev. 76, 437–454.

Rothman, A., Bickel, P., Levina, E., Zhu, J., 2008. Sparse permutation invariant covariance estimation. Electron. J. Stat. 2, 494–515.

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

Su, L., Jin, S., Zhang, Y., 2015. Specification test for panel data models with interactive fixed effects. J. Econometrics 186 (1), 222–244.

Tsai, H., Tsay, R.S., 2010. Constrained factor models. J. Amer. Statist. Assoc. 105 (492), 1593–1605.

Van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., et al., 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. Ann. Statist. 42 (3), 1166–1202.

Westerlund, J., Urbain, J.-P., 2013. On the estimation and inference in factor-augmented panel regressions with correlated loadings. Econom. Lett. 119 (3), 247–250.

Wolfers, J., 2006. Did unilateral divorce raise divorce rates? A reconciliation and new results. Amer. Econ. Rev. 96, 1802–1820.

Zhang, C., 2010. Nearly unbiased variable selection under minimax concave penalty. Ann. Statist. 38 (2), 894–942.

Zhang, C.-H., Zhang, S.S., 2014. Confidence intervals for low dimensional parameters in high dimensional linear models. J. R. Stat. Soc. Ser. B Stat. Methodol. 76 (1), 217–242.