Federal Reserve Bank of Boston®

Supervisory Research and Analysis Unit

Working Paper | SRA 20-01 | March 23, 2020

Efficient Estimation of Heterogeneous Coefficients in Panel Data Models with Common Shocks

Kunpeng Li, Guowei Cui, and Lina Lu

Public Service That Makes a Difference®

bostonfed.org

Efficient estimation of heterogeneous coefficients in panel data models with common shocks*

Kunpeng Li† and Guowei Cui‡ and Lina Lu§

Abstract

This paper investigates the estimation and inference issues of heterogeneous coefficients in panel data models with common shocks. We propose a novel two-step method to estimate the heterogeneous coefficients. We establish the asymptotic theory of our estimators, including consistency, asymptotic representation, and limiting distribution. Our two-step method can effectively address the limitations of the existing methods, such as the common correlated effects method proposed by Pesaran (2006, Econometrica) and the iterated principal components method proposed by Song (2013). The two-step estimator is as efficient as the two existing competitors in the basic model, and more efficient in the model with zero restrictions. Intensive Monte Carlo simulations show that the proposed estimator performs robustly in a variety of data setups.

Key Words: Factor analysis; Block diagonal covariance; Panel data models; Common shocks; Maximum likelihood estimation; heterogeneous coefficients; Inferential theory; Extreme value theory.

JEL classification: C33, C35

*We thank the editor Han Hong, an associate editor and three anonymous referees for their constructive comments which greatly improve the presentation of this paper. We also thank Jushan Bai, Bernard Salanié, Serena Ng, Qi Li, and participants in seminars at Columbia University, Huazhong University of Science and Technology, Renmin University of China, and University of International Business and Economics for their helpful comments on an early version of this paper. Kunpeng Li gratefully acknowledges the financial supports from China NSF No. 71722011, No. 71571122. Guowei Cui gratefully acknowledges the financial support from China NSF No. 71703046. During the revisions of this paper, our co-author, Guowei Cui, passed away suddenly in June, 2019. His kindness, diligence and talent will be deeply missed.

†International School of Economics and Management, Capital University of Economics and Business, Beijing, China, 100070. Corresponding author. Email: likp.07@gmail.com.

‡School of Economics, Huazhong University of Science and Technology, Wuhan, China, 430074. Email: cuigw610@163.com.

§Federal Reserve Bank of Boston, Department of Supervision, Regulation and Credit, 600 Atlantic Avenue, Boston, MA 02210, USA. The views expressed in the paper are those of the authors and should not be interpreted as reflecting the views of the Federal Reserve Bank of Boston, or any other person associated with the Federal Reserve System. Email: Lina.Lu@bos.frb.org.

1 Introduction

It has been long recognized and well documented in the literature that a small number of factors can explain a large fraction of the comovement of financial, macroeconomic and sectorial variables, see, for example, Ross (1976), Sargent and Sims (1977), Geweke (1977) and Stock and Watson (1998). Based on this fact, recent econometric literature places particular focus on panel data models with common shocks. These models specify that the dependent variable and explanatory variables both have a factor structure. A typical example can be written as

$$\begin{aligned} y_{it} &= \alpha_i + x_{it}'\beta_i + \lambda_i' f_t + \epsilon_{it}, \\ x_{it} &= \nu_i + \gamma_i' f_t + v_{it}, \qquad i = 1, 2, \dots, N; \ t = 1, 2, \dots, T. \end{aligned} \tag{1.1}$$

where y_{it} denotes the dependent variable; x_{it} denotes a $k \times 1$ vector of explanatory variables; and f_t is an $r \times 1$ vector of unknown factors, which represents the unobserved economic shocks. The factor loadings γ_i and λ_i capture the heterogeneous responses to the shocks. A salient feature of this paper is that the coefficients of x_{it} are assumed to be individual-dependent. Throughout the paper, we assume that the number of factors is fixed. For the case where the number of factors increases with the sample size, see Li, Li and Shi (2017).

Due to the presence of factors f_t, the error term of the y equation (i.e., $\lambda_i' f_t + \epsilon_{it}$) is correlated with the explanatory variables. The usual methods, such as ordinary least squares method, lead to inconsistent estimation. The instrumental variables (IV) method appears to be an intuitive way to address this issue, but the validity of IV is difficult to justify in practice. A remarkable result from recent studies is that, even without IV, model (1.1) can still be consistently estimated. For related studies, see Ahn, Lee and Schmidt (2001, 2013), Bai (2009), Bai and Li (2014), Moon and Weidner (2009, 2015, 2017), Pesaran (2006), Su, Jin and Zhang (2015) and Song (2013), among others.

Bai (2009) proposes an iterated principal components (PC) method to estimate a model with homogeneous coefficients. His analysis has been reexamined and extended by the perturbation theory in Moon and Weidner (2009, 2015, 2017). Su, Jin and Zhang (2015) propose a statistic to test the linearity specification of the model. These studies find that a bias arises from cross-sectional heteroscedasticity. Bai and Li (2014) therefore consider the quasi maximum likelihood method to eliminate this bias from the estimator. All the aforementioned studies are based on the assumption of a homogeneous coefficient. If the underlying coefficients are heterogeneous, misspecification of homogeneity would lead to inconsistent estimation (see the simulation of Kapetanios, Pesaran and Yamagata (2011)).

There are several studies on the estimation of heterogeneous coefficients. Pesaran (2006) proposes the common correlated effect (CCE) estimation method to estimate the heterogeneous coefficients in (1.1). The intuition of his method is to approximate the unknown projection space of the factors f_t by the space spanned by the cross-sectional average of the observations $(y_{it}, x_{it}')'$. To this end, some rank condition is needed. Song (2013) alternatively considers the iterated PC method, which extends the analysis of Bai

(2009) to the case of heterogeneous coefficients. In this paper, we propose a novel method to estimate model (1.1) and its extensions. Our estimation method is motivated by limitations of Pesaran's and Song's methods in estimating the heterogeneous coefficients for some particular data setups. The CCE estimator has a reputation for computational simplicity and excellent finite sample properties. However, we note that in some cases rank condition alone is not enough for a good approximation. When a good approximation breaks down, the CCE estimator would perform poorly. Song's method has a remarkable advantage that it does not need the factor structure specification on X, which makes it more flexible in applications. Although this advantage is very attractive to applied studies, Song's method also has the issue that the minimizer of the objective function is not easily obtained. In addition, the limiting distribution of the PC estimator relies on the assumption of cross sectional independence. The limitations of the CCE method and the PC method are manifested by simulations in Section 5.

Our estimation method consists of two steps. In the first step, we use the maximum likelihood (ML) method to estimate a pure factor model. Next, the heterogeneous coefficients are estimated by using relations implied by the model and replacing the unknown parameters with their ML estimators. Our two-step method can effectively address the aforementioned limitations. In addition, our two-step estimator has a striking advantage that it is as efficient as the CCE and PC estimators in model (1.1), and more efficient than the two competitors in model (4.1) below. The comparison of our two-step estimator and the two existing ones are discussed in details in Sections 3 and 4.

The proposed method strikes a balance between efficiency and computational economy. The computational burden in model (1.1) is not an ignorable issue due to a great number of βs being estimated, especially when N is large. This problem is made worse because we can only compute β_i ($i = 1, 2, \ldots, N$) sequentially, instead of all β_i simultaneously by matrix algebra. The PC method of Song (2013) seems less attractive in this perspective because it updates β_i sequentially in each iteration, which inevitably entails heavy computational burden. Our estimation method overcomes this problem by invoking the iterated computation method only in the first step to estimate a pure factor model. In the second step, we compute the estimators of β_i with a closed form, which greatly reduces the computation costs. In addition, the number of iterations can be substantially reduced in the first step if the PC estimates are chosen as the initial values. Nevertheless, our computation cost is still considerably larger than the CCE method.

Allowance of coefficients heterogeneity is one key ingredient of our specification, which has both theoretical and practical relevance. A recent development in econometrics is to identify the latent structure of panel data, see Su, Shi and Phillips (2016), Li, Qian and Su (2016), Su and Ju (2017), Lin and Ng (2012) and so forth. In these studies, cross sectional units sharing the same regression coefficients are classified into one group. To determine the membership of each unit, the first step is to estimate the cross sectional coefficients, assuming that they are unit-dependent. Next, based on the estimate of heterogeneous coefficients, the membership is determined by invoking some classification methods such as the K-means algorithm. Apparently, an efficient estimation of heterogeneous coefficients, the objective of this paper, is important to entail a better classification. In addition, if one is willing to specify the random coefficients as in Swamy (1970), the mean of coefficients can be easily estimated. This kind of work is considered in Pesaran (2006), and also pursued in the current paper. In regards to a practical aspect, Eberhardt, Helmers and Strauss (2013) point out that the existing studies on private return to R&D either fail to account for spill-over feature of knowledge in the Griliches knowledge production function, or, if account, specify the spill-over effects with some contentious spatial weights matrix. To overcome these weaknesses, Eberhardt, Helmers and Strauss (2013) consider an econometric model which is the same with model (4.1) below.[①] Their empirical results indicate that the private return to R&D is much smaller than what existing studies have found.

This paper contributes to the related literature in several dimensions. First, we propose a new approach to estimate the heterogeneous panels with common shocks, which to a large extent complements the existing methods. Second, we consider the ML estimation and inference of high dimensional factor models under the misspeficiation that the covariance matrix of idiosyncratic errors is block-diagonal. This theoretical work has independent interests. In addition, we also present some maximal results. These results are of theoretical relevance when developing model extensions. Third, we propose new statistics to perform the hypothesis testing on the validity of moment conditions. The proposed statistics share the spirit of the Hansen-Sargan statistic, but have to accommodate the particular issue in the current setup. So the theoretical analysis is more complicated.

The rest of the paper is organized as follows. Section 2 presents some theoretical results of the factor model with a block-diagonal covariance matrix of idiosyncratic errors. These results constitute the basis of the subsequent analysis. Section 3 illustrates the key estimation idea and derives the theoretical results for the basic model. Section 4 considers the model with restrictions on the loadings. We first consider a model with zero restrictions on the loadings in the y equation. We show that when zero restrictions are present, the loadings contain information for β, which is helpful to improve the estimation efficiency. We propose a minimum distance estimator to achieve the efficiency. We also propose Hansen-Sargan statistics to test the validity of additional moment conditions. Random slope specification is also investigated and the asymptotic properties for the slope mean are established. We next consider a model with an observed loadings in the y equation and a model with time-invariant regressors. Section 5 conducts extensive simulations to investigate the finite sample properties of the proposed estimator and provides the comparisons with some competitors. Section 6 concludes.

2 Factor models

In this section, we investigate the ML estimation for an approximate factor (AF) model under the misspecification that the covariance matrix of idiosyncratic errors is block-diagonal. As will be seen below, this AF model is closely related with model (1.1). Bai and Li (2016) consider a similar issue but misspecified that the covariance matrix of

[①] In fact, Equations (8a) and (8b) in Eberhardt, Helmers and Strauss (2013) are the same as model (4.1).

idiosyncratic errors is diagonal. In the viewpoint of covariance structure, our analysis generalizes their results. The primary purpose of this section is to build up some basic results that are needed for the subsequent analysis. We note that the theoretical analysis of this section has independent interests. In literature, the factor model with a block-diagonal variance matrix of errors is called the "multi-battery factor model", which has a long history and can date back to Tucker (1958). Our theoretical results would contribute to this branch of the literature. Throughout the paper, we use the Frobenius norm for a matrix, i.e., $\|A\| = [\text{tr}(A'A)]^{1/2}$ for matrix A. We use the symbols with overbar to denote the corresponding sample mean over the time, for example, $\bar{f} = \frac{1}{T}\sum_{s=1}^{T} f_s$; use $\dot{v}_t$ to denote $v_t - \bar{v}$ for any column vector v_t and M_{wv} to denote $\frac{1}{T}\sum_{t=1}^{T} \dot{w}_t \dot{v}_t'$ for any vectors w_t and v_t.

Now consider model (1.1). Substituting the second equation of model (1.1) into the first one, we have

$$\underbrace{\begin{bmatrix} y_{it} \\ x_{it} \end{bmatrix}}_{z_{it}} = \underbrace{\begin{bmatrix} \beta_i' \nu_i + \alpha_i \\ \nu_i \end{bmatrix}}_{\mu_i} + \underbrace{\begin{bmatrix} \beta_i' \gamma_i' + \lambda_i' \\ \gamma_i' \end{bmatrix}}_{\Lambda_i'} f_t + \underbrace{\begin{bmatrix} \beta_i' v_{it} + \epsilon_{it} \\ v_{it} \end{bmatrix}}_{u_{it}}, \tag{2.1}$$

or equivalently

$$z_{it} = \mu_i + \Lambda_i' f_t + u_{it}, \qquad i = 1, 2, \ldots, N; \;\; t = 1, 2, \ldots, T. \tag{2.2}$$

According to the definitions in (2.1), we see that z_{it} is a $\bar{K} \times 1$ vector of observations with $\bar{K} = k + 1$; u_{it} is a $\bar{K} \times 1$ vector of error terms; Λ_i is an $r \times \bar{K}$ loading matrix; and f_t is an $r \times 1$ vector of factors. Let $z_t = (z_{1t}', z_{2t}', \ldots, z_{Nt}')'$, $\mu = (\mu_1', \mu_2', \ldots, \mu_N')'$, $\Lambda = (\Lambda_1, \Lambda_2, \ldots, \Lambda_N)'$ and $u_t = (u_{1t}', u_{2t}', \ldots, u_{Nt}')'$, then we can rewrite (2.2) as

$$z_t = \mu + \Lambda f_t + u_t. \tag{2.3}$$

Without loss of generality, we assume $\bar{f} = T^{-1}\sum_{t=1}^{T} f_t = 0$ in the paper because the model can always be written as $z_t = \mu + \Lambda \bar{f} + \Lambda(f_t - \bar{f}) + u_t = \mu^* + \Lambda f_t^* + u_t$ with $\mu^* = \mu + \Lambda \bar{f}$ and $f_t^* = f_t - \bar{f}$.

2.1 Estimation

We consider the estimation of model (2.2). Suppose that (i) f_t is normally distributed with mean zero and variance $M_{ff}^{\dagger}$, (ii) u_{it} is independent and identically normally distributed over t and independent over i with mean zero and variance $\Psi_i^{\dagger}$, (iii) f_t is independent with u_{is} for all i, t and s, it can be readily verified that the corresponding likelihood function after concentrating out the intercept μ is

$$\log \mathscr{L}(\theta) = -\frac{1}{2N} \log |\Sigma_{zz}| - \frac{1}{2N} \text{tr}[M_{zz} \Sigma_{zz}^{-1}] \tag{2.4}$$

where $\theta = (\Lambda^{\dagger}, \Psi^{\dagger}, M_{ff}^{\dagger})$ and $\Sigma_{zz}^{\dagger} = \Lambda^{\dagger} M_{ff}^{\dagger} \Lambda^{\dagger\prime} + \Psi^{\dagger}$; $M_{zz} = \frac{1}{T}\sum_{t=1}^{T} \dot{z}_t \dot{z}_t'$ is the data matrix with $\dot{z}_t = z_t - \frac{1}{T}\sum_{s=1}^{T} z_s$, and $\Psi^{\dagger} = \text{diag}(\Psi_1^{\dagger}, \ldots, \Psi_N^{\dagger})$ is a block-diagonal matrix with $\Psi_i^{\dagger}$ being a $\bar{K} \times \bar{K}$ symmetric positive definite matrix for each $i = 1, 2, \ldots, N$. Here we use the symbols with dagger to denote the inputs of the likelihood function. Note that for any given $\theta = (\Lambda^\dagger, \Psi^\dagger, M_{ff}^\dagger)$, we have $\log \mathscr{L}(\widetilde{\theta}) = \log \mathscr{L}(\theta)$ for $\widetilde{\theta} = (\Lambda^\dagger M_{ff}^{\dagger 1/2}, \Psi^\dagger, I_r)$. From this perspective, it is no loss of generality to normalize $M_{ff}^\dagger = I_r$. Therefore $\Sigma_{zz}^\dagger$ is simplified as $\Sigma_{zz}^\dagger = \Lambda^\dagger \Lambda^{\dagger\prime} + \Psi^\dagger$. Although the factors f_t are assumed to be fixed constants in Assumption A below, and the errors u_{it} may have weak correlations and heteroskedasticities over i and t, we still use the above objective function and call the maximizer $\widehat{\theta} = (\widehat{\Lambda}, \widehat{\Psi})$, defined by

$$\widehat{\theta} = \underset{\theta \in \mathrm{Par}(\theta)}{\operatorname{argmax}} \log \mathscr{L}(\theta),$$

the quasi maximum likelihood estimator, or the MLE, where $\mathrm{Par}(\theta)$ is the parameters space for θ defined by

$$\mathrm{Par}(\theta) = \Big\{ (\Lambda^\dagger, \Psi^\dagger) \Big| \Psi^\dagger = \mathrm{diag}(\Psi_1^\dagger, \ldots, \Psi_N^\dagger), 0 < c \leq \tau_{\min}(\Psi_i^\dagger) \leq \tau_{\max}(\Psi_i^\dagger) \leq C \text{ for each } i \Big\},$$

where $\tau_{\min}(\cdot)$ and $\tau_{\max}(\cdot)$ denote the respective smallest and largest eigenvalues of its input, and c and C are two constants given below. The parameters space specifies that $\Psi_i^\dagger$ is estimated in a compact set which is bounded away from zero. This specification is due to the possible appearance of the so-called Heywood case (see Lawley and Maxwell (1971)). To be specific, consider the factor model $z_{it} = \mu_i + \Lambda_i' f_t + u_{it}$, which has an undesirable alternative expression

$$z_{it} = \mu_i + \tilde{\Lambda}_i' f_t + 1(i=1) z_{1t} + \tilde{u}_{it} = \underbrace{\Big[\tilde{\Lambda}_i', \mathcal{I}_{\bar{K}}(i=1) \Big]}_{\Lambda_i^{*\prime}} \underbrace{\begin{bmatrix} f_t \\ z_{1t} \end{bmatrix}}_{f_t^*} + \tilde{u}_{it} = \mu_i + \Lambda_i^{*\prime} f_t^* + \tilde{u}_{it},$$

with

$$\tilde{\Lambda}_i = \Lambda_i \Big(1 - 1(i=1) \Big) = \begin{cases} 0 & \text{if } i = 1; \\ \Lambda_i & \text{if } i \neq 1. \end{cases} \qquad \tilde{u}_{it} = u_{it} \Big(1 - 1(i=1) \Big) = \begin{cases} 0 & \text{if } i = 1; \\ u_{it} & \text{if } i \neq 1. \end{cases}$$

where $1(i=1)$ is equal to 1 if $i = 1$, 0 otherwise; and $\mathcal{I}_{\bar{K}}(i=1)$ is equal to a $\bar{K}$-dimensional identity matrix if $i = 1$, and 0 otherwise. Although this peculiar case can be precluded asymptotically by correctly determining the number of factors (see the discussion on identification below), it does cause trouble in the finite sample, especially when implementing the ML estimation. As a response, we require that $\Psi_i^\dagger$ be estimated in a compact set bounded away from zero, see Assumption E below. In practice, one can choose $c = 0.1 \min_i \tilde{\sigma}_i^2, C = 10 \max_i \tilde{\sigma}_i^2$ where $\tilde{\sigma}_i^2$ is the estimator of idiosyncratic variance for unit i by the PC method as long as the number of factors is not underestimated. The estimation procedure for (2.4) is discussed in Appendix E, which is an extension of the EM algorithm of Bai and Li (2012).

Maximizing the objective function (2.4) with respect to Λ and Ψ gives the following two first order conditions.

$$\widehat{\Lambda}' \widehat{\Psi}^{-1} (M_{zz} - \widehat{\Sigma}_{zz}) = 0, \tag{2.5}$$

$$\mathrm{Bdiag}(M_{zz} - \widehat{\Sigma}_{zz}) = 0, \tag{2.6}$$

where Bdiag$(\cdot)$ is the block-diagonal operator, which puts the element of its argument to zero if the counterpart of Ψ is nonzero, otherwise unspecified. $\widehat{\Lambda}$ and $\widehat{\Psi}$ denote the ML estimators and $\widehat{\Sigma}_{zz} = \widehat{\Lambda}\widehat{\Lambda}' + \widehat{\Psi}$. The detailed derivations of the above two equations are given in Appendix A.

2.2 Assumptions

To analyze (2.3), we make the following assumptions. Hereafter we use C to denote a generic constant sufficiently large.

Assumption A: The factor f_t is a sequence of constants with $\|f_t\| \leq C$ for all $t = 1, 2, \ldots, T$. Let $M_{ff} = T^{-1}\sum_{t=1}^{T} f_t f_t'$. We assume that $\overline{M}_{ff} = \lim_{T\to\infty} M_{ff}$ is a positive definite matrix.

Assumption B: We make following assumptions on the idiosyncratic error u_{it}:

(B.1) $E(u_{it}) = 0$ and $E(\|u_{it}\|^{16}) \leq C$ for each i and t. In addition, $E(\|\sqrt{T}\bar{u}_i\|^4) \leq C$ for each i, where $\bar{u}_i = T^{-1}\sum_{t=1}^{T} u_{it}$

(B.2) Let $\Sigma_{ij,t} = E(u_{it}u_{jt}')$. We assume $\sum_{j=1}^{N} \mathbf{a}_{ij} \leq C$ for each i, where $\mathbf{a}_{ij} = \max_{t\leq T}\|\Sigma_{ij,t}\|$.

(B.3) Let $\Xi_{i,ts} = E(u_{it}u_{is}')$. We assume $\sum_{s=1}^{T} \mathbf{b}_{ts} \leq C$ for each t, where $\mathbf{b}_{ts} = \max_{i\leq N}\|\Xi_{i,ts}\|$.

(B.4) $\tau_{min}(\Sigma_{ii,t}) \geq c$ for each i and t, where c is a generic constant which is strictly greater than 0, and $\tau_{min}(\cdot)$ denotes the smallest eigenvalue of its argument.

Assumption C: We make following assumptions on the loadings:

(C.1) The loadings Λ_i are fixed values satisfying $\|\Lambda_i\| \leq C$ for all $i = 1, \cdots, N$.

(C.2) There exists an $r \times r$ positively definite matrix Q such that $Q = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \Lambda_i \bar{\Sigma}_{ii}^{-1}\Lambda_i'$, where $\bar{\Sigma}_{ii} = \frac{1}{T}\sum_{t=1}^{T} \Sigma_{ii,t}$ with $\Sigma_{ii,t}$ defined in Assumption B.2.

Assumption D: Let $u_{it} = (\epsilon_{it} + v_{it}'\beta, v_{it}')'$, where ϵ_{it} and v_{it} are the respective idiosyncratic errors of the Y and X equations in model (1.1). We have the following moments conditions and weak convergence.

(D.1) $E\left[\left\|\dfrac{1}{\sqrt{T}}\sum_{t=1}^{T} u_{it}u_{jt}' - E(u_{it}u_{jt}')\right\|^2\right] \leq C, \quad$ for each i and j,

(D.2) $E\left[\left\|\dfrac{1}{\sqrt{NT}}\sum_{j=1}^{N}\sum_{t=1}^{T} \Lambda_i\bar{\Sigma}_{ii}^{-1}[u_{jt}u_{it}' - E(u_{jt}u_{it}')]\right\|^2\right] \leq C, \quad$ for each i,

(D.3) $E\left[\left\|\dfrac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T} \Lambda_i\bar{\Sigma}_{ii}^{-1}u_{it}f_t'\right\|^2\right] \leq C,$

(D.4) $\dfrac{1}{\sqrt{T}}\sum_{t=1}^{T} \tilde{P}_{it}\epsilon_{it} \xrightarrow{d} N(0, J_i), \quad$ for each i, where $\tilde{P}_{it} = [h_t^{\star\prime}, v_{it}']'$ with $h_t^\star$ implicitly defined by $f_t^\star = [g_t^{\star\prime}, h_t^{\star\prime}]' = M_{ff}^{-1} f_t$, and $J_i = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \sum_{s=1}^T E(\tilde{P}_{it}\tilde{P}_{is}' \epsilon_{it}\epsilon_{is})$.

Assumption E: The variances $\bar{\Sigma}_{ii}$ for all i are estimated in a compact set; that is, all the eigenvalues of $\widehat{\Psi}_i$ (which we also use $\widehat{\Sigma}_{ii}$ to denote in the paper) are in an interval $[c, C]$ for two positive constants $c < C$.

Assumption F: We make following three assumptions to further restrict the idiosyncratic errors:

(F.1) For each i, the error sequence u_{it} admits a Wold representation: $u_{it} = \sum_{\ell=0}^\infty D_{i\ell}\varepsilon_{it-\ell}$ with $D_{i0} = I_{k+1}$. Let $\mathbb{D}_{im} = \sum_{\ell \geq m} \|D_{i\ell}\|$. We assume that $\max_{i\leq N} \mathbb{D}_{im} = O(m^{-5})$.

(F.2) For each i, ε_{it} is uniformly L_2 bounded, independent, continuous with probability density function $f_{\varepsilon_{it}}$ and

$$\sup_{t\leq T}\int_{-\infty}^{\infty} |f_{\varepsilon_{it}}(x+a) - f_{\varepsilon_{it}}(x)| dx \leq C|a|$$

whenever $|a| \leq c$ for some $c > 0$.

(F.3) $\sum_{\ell=1}^\infty D_{i\ell}x^\ell \neq 0$ for all complex numbers x with $|x| \leq 1$.

Assumption G: Let $f_t^\diamond = [f_t', 1]'$. We have the following moments conditions:

(G.1) $$E\left[\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^T f_t^\diamond u_{it}'\right\|^8\right] \leq C, \quad \text{for each } i;$$

(G.2) $$E\left[\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^T [u_{it}u_{jt}' - E(u_{it}u_{jt}')]\right\|^8\right] \leq C, \quad \text{for each } i, j;$$

(G.3) $$E\left[\left\|\frac{1}{\sqrt{NT}}\sum_{i=1}^N\sum_{t=1}^T \Lambda_i\bar{\Sigma}_{ii}^{-1}u_{it}f_t^{\diamond\prime}\right\|^8\right] \leq C;$$

(G.4) $$E\left[\left\|\frac{1}{\sqrt{NT}}\sum_{i=1}^N\sum_{t=1}^T \Lambda_i\bar{\Sigma}_{ii}^{-1}[u_{it}u_{jt}' - E(u_{it}u_{jt}')]\right\|^8\right] \leq C, \quad \text{for each } j.$$

We give some comments on the above assumptions. Assumption A is about factors. The factors are treated as parameters according to this assumption. Similar assumptions are made in Bai (2009), Bai and Li (2014), Moon and Weidner (2009), etc. In cases when factors are random, the analysis in this paper can be viewed as conditioning on a particular realization of factors. This assumption makes our factor model more flexible. For example, we can allow one factor to be normalized time trend, i.e., $f_{tk} = t/T$. For random factors, the counterpart of Assumption A is $E(\|f_t\|^8) < \infty$ for all t, and $\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^T E(f_t f_t')$ is positive definite. Assumption B is about idiosyncratic errors. It allows both cross-sectional and temporal correlations and heteroskedasticities in errors. So the model considered here is essentially an approximate factor model in the sense of Chamberlain and Rothschild (1983). Assumption B.1 requires that the finiteness of sixteenth moment, which is slightly more stringent than the corresponding assumption on the eighth moment in the literature. Assumptions B.2 and B.3 aim to control the correlation magnitudes over the cross section and time, which correspond to Assumptions C.2 and C.3 in Bai (2003). With these two assumptions, the correlations from the idiosyncratic part are weak enough that allow us to separate them from the common components part. Assumption B.4 requires that the variances of idiosyncratic errors be bounded away from zeros, which is standard. Note that Assumption B.2 implies that these variances are bounded from the above. Assumption C is about loadings, which is standard. We note that, under Assumptions A and C, the factors would have the so-called pervasive property. Assumption D imposes some moment conditions and assumes a weak convergence result, which corresponds to Assumption F in Bai (2003). This assumption is for the tractability of theoretical analysis.

Assumption E assumes that partial parameters (i.e., $\bar{\Sigma}_{ii}$) are estimated in a compact set, which corresponds to the parameters space specified above. In econometric or statistical literature, compactness is often made when dealing with nonlinear objective functions, see, for example, Newey and McFadden (1994), Jennirch (1969) and so forth. The objective function (2.4) in the current paper has severe nonlinearity. We therefore impose this assumption for theoretical tractability. In addition, Assumption E is also helpful to preclude the Heywood case mentioned above.

Assumptions F and G are strengthened versions of Assumptions B and D, respectively. These two assumptions are made mainly for the theoretical results in Subsections 4.3 and 4.4 below, where some uniform results are needed. Assumption F.1 assumes that idiosyncratic errors admit linear processes. This assumption is sufficiently general to allow a rich class of dependent processes and a similar assumption is also made in a number of related studies, such as Doz, Giannone and Reichlin (2011). With Assumption F.1, we can invoke the Komlós-Major-Tusnády approximation developed recently by Berkes, Liu and Wu (2014) to analyze Theorem 4.2 in Subsection 4.3. Assumptions F.2 and F.3 are borrowed from Theorem 14.9 of Davidson (1994), which guarantee that u_{it} is a α-mixing process. As pointed out in Davidson (1994), they are sufficiently weak. Assumption G imposes more tight moment conditions, which are helpful to control the inflating factors when deriving the maximal results. Similar assumptions are also made in Castagnetti, Rossi and Trapani (2015) and Fan, Liao and Mincheva (2013).

2.3 Asymptotic properties of the ML estimators

This section presents the asymptotic results of the ML estimators of (2.4). It is well known that the factors and loadings can only be identified up to a rotation. In this section, we adopt the treatment of Bai (2003), in which the rotational matrix appears in the asymptotic representation. This treatment has two advantages in the present context. First, it simplifies our analysis. Second, it clarifies that the estimation and inferential theory of β is invariant to the rotational matrix. Alternatively, we can impose some additional restrictions to uniquely fix the rotational matrix; see Anderson and Rudin (1956) and Bai and Li (2012) for full identification strategies. The following theorem, which serves as the base for the subsequent analysis, gives the asymptotic representations of the MLE.

Theorem 2.1 *Under Assumptions A-E, as $N, T \to \infty$, we have that for each i,*

$$\widehat{\Lambda}_i - R'\Lambda_i = R'M_{ff}^{-1}\frac{1}{T}\sum_{t=1}^{T} f_t u_{it}' + O_p(\frac{1}{N}) + O_p(\frac{1}{T})$$

$$\widehat{\Sigma}_{ii} - \bar{\Sigma}_{ii} = \frac{1}{T}\sum_{t=1}^{T}(u_{it}u_{it}' - \Sigma_{ii,t}) + O_p(\frac{1}{N}) + O_p(\frac{1}{T})$$

where $R = M_{ff}\Lambda'\widehat{\Psi}^{-1}\widehat{\Lambda}(\widehat{\Lambda}'\widehat{\Psi}^{-1}\widehat{\Lambda})^{-1}$.

From Theorem 2.1, we see that the rotational matrix R only enters in the asymptotic representation of $\widehat{\Lambda}_i$. This is consistent with only loadings and factors having rotational indeterminacy and idiosyncratic errors not having such a problem. In addition, we see $\widehat{\Lambda}_i - R'\Lambda_i = O_p(\frac{1}{\sqrt{T}}) + O_p(\frac{1}{N})$ and $\widehat{\Sigma}_{ii} - \bar{\Sigma}_{ii} = O_p(\frac{1}{\sqrt{T}}) + O_p(\frac{1}{N})$. The term $O(\frac{1}{N})$ is a bias term, which arises from misspecifying the variance matrix of idiosyncratic errors to be block-diagonal. So if the cross-sectional correlations only exists within group i, the model is therefore correctly specified and the ML estimator is simply $\sqrt{T}$-consistent. This result has a strong implication that the CV and LV estimators defined below would be consistent even when N is finite under this special case. However, the asymptotic representations will be more complicated when N is finite.

3 Asymptotics of the basic model

Now consider the basic model (1.1), one target of this paper. We make the following assumption for further analysis:

Assumption H: $E(v_{it}\epsilon_{it}) = 0$.

Assumption I: The coefficients β_i ($i = 1, 2, \ldots, N$) are fixed values with $\|\beta_i\| \leq C$ for some finite constant C.

Assumption H is crucial to the models with common shocks and is maintained by all the related studies; for example, Bai (2009), Bai and Li (2014), Pesaran (2006), and Moon and Weidner (2009, 2015, 2017). Assumption I is standard in the heterogeneous model.

Before presenting our estimation method, we discuss on the identification issue related to our model. We note that for any given $i^* \in \{1, 2, \ldots, N\}$, model (1.1) has an alternative expression

$$\begin{aligned} y_{it} &= \alpha_i + x_{it}'\Big(\beta_i - \mathbf{1}(i = i^*)(\beta_i - \beta^\dagger)\Big) + \lambda_i' f_t + \mathbf{1}(i = i^*)x_{it}'(\beta_i - \beta^\dagger) + \epsilon_{it} \\ &= \alpha_i + x_{it}'\Big(\beta_i - \mathbf{1}(i = i^*)(\beta_i - \beta^\dagger)\Big) + \lambda_i' f_t + \mathbf{1}(i = i^*)x_{i^*t}'(\beta_{i^*} - \beta^\dagger) + \epsilon_{it} \\ &= \alpha_i + x_{it}'\Big(\beta_i - \mathbf{1}(i = i^*)(\beta_i - \beta^\dagger)\Big) + \lambda_i^{\dagger\prime} f_t^\dagger + \epsilon_{it}, \end{aligned} \tag{3.1}$$

with

$$\lambda_i^\dagger = \begin{bmatrix} \lambda_i \\ \mathbf{1}(i = i^*) \end{bmatrix}, \qquad f_t^\dagger = \begin{bmatrix} f_t \\ x_{i^*t}'(\beta_{i^*} - \beta^\dagger) \end{bmatrix}.$$

Apparently, the equation for x_{it} has a similarly extended expression by setting the loadings of the newly-added factor $x'_{i^*t}(\beta_{i^*} - \beta^\dagger)$ to be zeros. Now we have an observationally equivalent model. This ill-posed representation entails an identification issue for the regression coefficient β_i which should be precluded in this paper. For model (3.1), the factor $f_t^\dagger$ now is a mixture of random and nonrandom factors. Its nonrandom factors satisfy Assumption A, and its random factor satisfies the random version of Assumption A, which is given in the comments of Assumption A. But the loading $\lambda_i^\dagger$ does not satisfy Assumption C. As a result, the term $\mathbf{1}(i = i^*)x'_{i^*t}(\beta_{i^*} - \beta^\dagger)$ should not enter into $\lambda_i' f_t$, but the remaining expression. There are cases that Assumption C is satisfied but Assumption A breaks down. In the literature, f_t is called strong factors if they satisfy Assumption A (or, if random, satisfy the random version of Assumption A) and their loadings satisfy Assumption C. Our strong factors assumption has two implications. First, the weak factors in the sense of Onatski (2012) should not be treated as factors in the current paper even they have signals stronger than the noise. Second, the method to determine the number of factors should be consistent with this treatment. Fortunately, most of the existing studies focus on determining the number of strong factors. So we have many choices at hand such as Bai and Ng (2002), Ahn and Horenstein (2013) and so forth. In this paper, we use the method of Ahn and Horenstein (2013) to determine the number of factors.

We now illustrate the idea of our estimation method. Let Ω_{it} be the covariance matrix of v_{it} and σ_{it}^2 the variance of ϵ_{it}. According to equation (2.1), the covariance of u_{it}, which is $\Sigma_{ii,t}$ by Assumption B.2, now is

$$\Sigma_{ii,t} = \begin{bmatrix} \Sigma_{ii,t,11} & \Sigma_{ii,t,12} \\ \Sigma_{ii,t,21} & \Sigma_{ii,t,22} \end{bmatrix} = \begin{bmatrix} \beta_i'\Omega_{it}\beta_i + \sigma_{it}^2 & \beta_i'\Omega_{it} \\ \Omega_{it}\beta_i & \Omega_{it} \end{bmatrix}. \tag{3.2}$$

This leads to $\Sigma_{ii,t,22}\beta_i = \Sigma_{ii,t,21}$. Taking average over t on both sides, we have

$$\bar{\Sigma}_{ii,22}\beta_i = \bar{\Sigma}_{ii,21}. \tag{3.3}$$

where $\bar{\Sigma}_{ii,22}$ and $\bar{\Sigma}_{ii,21}$ are implicitly defined in Assumption C.2. According to Theorem 2.1, $\widehat{\Sigma}_{ii}$ is a consistent estimator of $\bar{\Sigma}_{ii}$. So β_i can be estimated by $\widehat{\beta}_i = \widehat{\Sigma}_{ii,22}^{-1}\widehat{\Sigma}_{ii,21}$. We call this estimator *CoVariance estimator*, denoted by $\widehat{\beta}_i^{CV}$ since the estimation for β_i only involves the covariance of u_{it}.

Theorem 2.1 implies $\widehat{\Sigma}_{ii} = \bar{\Sigma}_{ii} + o_p(1)$, so the consistency of $\widehat{\beta}_i^{CV}$ is immediately obtained by the continuous mapping theorem. Furthermore, by Theorem 2.1,

$$\widehat{\Sigma}_{ii} - \bar{\Sigma}_{ii} = \frac{1}{T}\sum_{t=1}^{T}(u_{it}u_{it}' - \Sigma_{ii,t}) + O_p(\frac{1}{N}) + O_p(\frac{1}{T}).$$

Then it follows

$$\widehat{\Sigma}_{ii,21} - \bar{\Sigma}_{ii,21} = \frac{1}{T}\sum_{t=1}^{T}[v_{it}(\epsilon_{it} + v_{it}'\beta_i) - \Omega_{it}\beta_i] + O_p(\frac{1}{N}) + O_p(\frac{1}{T}); \tag{3.4}$$

$$\widehat{\Sigma}_{ii,22} - \bar{\Sigma}_{ii,22} = \frac{1}{T}\sum_{t=1}^{T}[v_{it}v_{it}' - \Omega_{it}] + O_p(\frac{1}{N}) + O_p(\frac{1}{T}). \tag{3.5}$$

Notice that

$$\begin{aligned}\widehat{\beta}_i^{CV}-\beta_i &= (\widehat{\Sigma}_{ii,22})^{-1}\widehat{\Sigma}_{ii,21}-\bar{\Sigma}_{ii,22}^{-1}\bar{\Sigma}_{ii,21}\\ &= (\widehat{\Sigma}_{ii,22})^{-1}\Big[(\widehat{\Sigma}_{ii,21}-\bar{\Sigma}_{ii,21})-(\widehat{\Sigma}_{ii,22}-\bar{\Sigma}_{ii,22})\bar{\Sigma}_{ii,22}^{-1}\bar{\Sigma}_{ii,21}\Big]\end{aligned} \tag{3.6}$$

Let $\bar{\Omega}_i = T^{-1}\sum_{t=1}^{T}\Omega_{it}$. Substituting (3.4) and (3.5) into (3.6) and noting that $\widehat{\Sigma}_{ii,22}\xrightarrow{p}\bar{\Omega}_i$ and $\beta_i=\bar{\Sigma}_{ii,22}^{-1}\bar{\Sigma}_{ii,21}$, we have the following theorem about $\widehat{\beta}_i^{CV}$.

Theorem 3.1 *Under Assumptions A-E, H and I, when $N,T\to\infty$ and $T/N^2\to 0$, we have that for each i,*

$$\sqrt{T}(\widehat{\beta}_i^{CV}-\beta_i)=\bar{\Omega}_i^{-1}\Big(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}v_{it}\epsilon_{it}\Big)+o_p(1). \tag{3.7}$$

Remark 3.1 Consider the equation $y_{it}=\alpha_i+x_{it}'\beta_i+\lambda_i' f_t+\epsilon_{it}$. Suppose that the factors are observed for all t, then the (infeasible) ordinary least square estimator is

$$\widehat{\beta}_i^{inf}=(X_i'M_{F^\diamond}X_i)^{-1}X_i'M_{F^\diamond}Y_i=(V_i'M_{F^\diamond}V_i)^{-1}V_i'M_{F^\diamond}Y_i.$$

where $F^\diamond=(\mathbf{1}_T,F)$ and $X_i=(x_{i1},x_{i2},\ldots,x_{iT})'$ is a $T\times k$ matrix; Y_i and V_i are defined similarly. Some straightforward computations show that the infeasible estimator $\widehat{\beta}_i^{inf}$ has the asymptotic representation:

$$\sqrt{T}(\widehat{\beta}_i^{inf}-\beta_i)=\bar{\Omega}_i^{-1}\Big(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}v_{it}\epsilon_{it}\Big)+o_p(1), \tag{3.8}$$

which is the same as that of Theorem 3.1. The above analysis indicates that the two-step method amounts to making the unobserved factors observable. In addition, we note that the CCE estimator of Pesaran (2006) and the iterated PC estimator of Song (2013) both have this asymptotic representation. So their asymptotic results can be interpreted in the same way. □

Remark 3.2 Our analysis indicates that if u_{it} is independent across i, the $O_p(\frac{1}{N})$ term in Theorem 2.1 would disappear. As a result, the CV estimator would be consistent in this special case even when N is finite. However, the asymptotic representation in this special case will be complicated since the $O_p(\frac{1}{\sqrt{NT}})$ terms, which are ignored in large-N setup, will enter into the asymptotic representation since they have the same magnitude with $O_p(\frac{1}{\sqrt{T}})$. □

Remark 3.3 Our estimation method and Song's method share the similarity of both estimating a factor model, they also have differences. The idea of Song's method is the so-called "controlling through estimating" method. Consider the Y equation

$$y_{it}=\alpha_i+x_{it}'\beta_i+\lambda_i' f_t+e_{it}.$$

The endogeneity issue of β_i is due to the presence of $\lambda_i' f_t$, so the PC method addresses this issue by estimating β_i simultaneously with λ_i and f_t to control the endogeneity.

Our estimation method is essentially the generalized method of moments (GMM). The key point of this paper is that we find some moment conditions for β_i, which is $\bar{\Sigma}_{ii,22}\beta_i = \bar{\Sigma}_{ii,21}$ in the basic model, and $\Delta_i\beta_i = \delta_i$ in the extended model below. However, an obstacle to apply these moment conditions is that the parameters $\bar{\Sigma}_{ii,22}, \bar{\Sigma}_{ii,21}, \Delta_i$ and δ_i are all unknown. Thus, we have to conduct the factor analysis to estimate these parameters first. Now we highlight the different purposes of factor analysis in the two methods. In Song's method, it aims to estimate $\lambda_i' f_t$ to control the endogeneity. In our method, it aims to estimate the unknown parameters in moment conditions. □

Corollary 3.1 *Under the assumptions of Theorem 3.1, we have that for each i,*

$$\sqrt{T}(\widehat{\beta}_i^{CV} - \beta_i) \xrightarrow{d} N\big(0, \bar{\Omega}_i^{-1}\Theta_i\bar{\Omega}_i^{-1}\big),$$

where

$$\Theta_i = \text{avar}\Big(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} v_{it}\epsilon_{it}\Big) = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T} E(v_{it}v_{is}'\epsilon_{it}\epsilon_{is}).$$

Remark 3.4 The limiting variance can be consistently estimated by $\widehat{\Sigma}_{ii,22}^{-1}\widehat{\Theta}_i\widehat{\Sigma}_{ii,22}^{-1}$, where

$$\widehat{\Theta}_i = \frac{1}{T}\sum_{t=1}^{T}\widehat{v}_{it}\widehat{v}_{it}'\widehat{\epsilon}_{it}^2 + \sum_{\nu=1}^{S_{i,T}} K(\nu, S_{i,T})\Big[\frac{1}{T}\sum_{t=\nu+1}^{T}(\widehat{v}_{it}\widehat{v}_{i,t-\nu}' + \widehat{v}_{i,t-\nu}\widehat{v}_{it}')\widehat{\epsilon}_{it}\widehat{\epsilon}_{i,t-\nu}\Big], \tag{3.9}$$

where $K(v, S_{i,T})$ is the kernel used to weight the information signals of temporal correlations of different orders and guarantee positive definiteness of the estimated covariance matrix. $S_{i,T}$ is the bandwidth (lag truncation) parameter for the unit i. If one uses the Bartlett kernel, then the estimator is

$$\widehat{\Theta}_i = \frac{1}{T}\sum_{t=1}^{T}\widehat{v}_{it}\widehat{v}_{it}'\widehat{\epsilon}_{it}^2 + \sum_{\nu=1}^{S_{i,T}}(1 - \frac{\nu}{S_{i,T}+1})\Big[\frac{1}{T}\sum_{t=\nu+1}^{T}(\widehat{v}_{it}\widehat{v}_{i,t-\nu}' + \widehat{v}_{i,t-\nu}\widehat{v}_{it}')\widehat{\epsilon}_{it}\widehat{\epsilon}_{i,t-\nu}\Big].$$

The truncated parameter $S_{i,T}$ can be determined by the data-driven method suggested by Andrews (1991). The consistency of the above estimator is given in Proposition 4.1 below. □

4 Models with restrictions

In this section, we consider the following restricted model:

$$\begin{aligned} y_{it} &= \alpha_i + x_{it}'\beta_i + \psi_i' g_t + \epsilon_{it} \\ x_{it} &= \nu_i + \gamma_i^{g\prime} g_t + \gamma_i^{h\prime} h_t + v_{it} \end{aligned} \tag{4.1}$$

where the dimensions of g_t and h_t are $r_1 \times 1$ and $r_2 \times 1$, respectively. A salient feature of model (4.1) is that the explanatory variables include more factors than the error of the y equation.

The restricted model has its empirical motivations. Consider the study on the relationship between the unemployment rate (UR) and foreign direct investment (FDI). Let

y_{it} be the UR for state i at time t, and one of x_{it} be FDI$_{it}$, i.e., the FDI for state i and time t. One may introduce other control variables such as inflation, government expenditure, infrastructure, etc. In the above specification, f_t can be interpreted as the domestic economic shocks such as the economic policy shocks of government over time, which both affect UR and FDI. However, we note that FDI may have its own particular shocks. For example, technical advancement of shipment, tariff changes, regional economic cooperation and development agreement, etc. Although we cannot assert with one hundred percents confidence that these shocks have no direct relations with the UR, these shocks are indeed loosely related with the local URs according to the economic theory. So it is plausible that one treats these particular shocks as h_t first. In the following subsections, we develop the estimation procedures to determine whether this treatment is appropriate or not. Beside the above application, Eberhardt, Helmers and Strauss (2013) recently use the same model to study the private returns to R&D and find that the private returns to R&D are overestimated in the previous studies. Their empirical work is also a good motivation for model (4.1).

The y equation of (4.1) can be written as

$$y_{it} = \alpha_i + x_{it}'\beta_i + \psi_i' g_t + \phi_i' h_t + \epsilon_{it}$$

with $\phi_i = 0$ for all i. Let $f_t = (g_t', h_t')'$, $\lambda_i = (\psi_i', \phi_i')'$ and $\gamma_i = (\gamma_i^{g\prime}, \gamma_i^{h\prime})'$, we have the same representation as (1.1). From this perspective, model (4.1) can be viewed as a restricted version of model (1.1). This implies that the two-step method proposed in Section 3 is applicable to (4.1). However, this estimation method is not efficient. Consider the ideal case that g_t is observable. To eliminate the endogenous ingredient $\psi_i' g_t$, we post-multiply $M_G = I - G(G'G)^{-1}G'$ on both sides of the y equation. The remaining part of x_{it} includes v_{it} and $\gamma_i^{h\prime}(h_t - H'G(G'G)^{-1}g_t)$, both of them provide the information for β. However, as shown in Theorem 3.1, only the variations of v_{it} are used to signal β_i in $\widehat{\beta}_i^{CV}$. Therefore, partial information is discarded and the two-step method in Section 3 is inefficient.

The preceding discussion provides some insights on the improvement of efficiency. To efficiently estimate model (4.1), we need to use information contained in the common components of x_{it}. Rewrite model (4.1) as

$$\begin{bmatrix} y_{it} \\ x_{it} \end{bmatrix} = \begin{bmatrix} \beta_i'\nu_i + \alpha_i \\ \nu_i \end{bmatrix} + \begin{bmatrix} \beta_i'\gamma_i^{g\prime} + \psi_i' & \beta_i'\gamma_i^{h\prime} \\ \gamma_i^{g\prime} & \gamma_i^{h\prime} \end{bmatrix} \begin{bmatrix} g_t \\ h_t \end{bmatrix} + \begin{bmatrix} \beta_i' v_{it} + \epsilon_{it} \\ v_{it} \end{bmatrix} \tag{4.2}$$

We use Λ_i' to denote the loadings matrix in font of $f_t = (g_t', h_t')'$. The symbols μ_i, z_{it} and u_{it} are defined the same as before. We then have the same equation as (2.2). Further partition the loadings matrix Λ_i into four blocks,

$$\Lambda_i = \begin{bmatrix} \Lambda_{i,11} & \Lambda_{i,12} \\ \Lambda_{i,21} & \Lambda_{i,22} \end{bmatrix} = \begin{bmatrix} \psi_i + \gamma_i^g \beta_i & \gamma_i^g \\ \gamma_i^h \beta_i & \gamma_i^h \end{bmatrix}. \tag{4.3}$$

So we have $\Lambda_{i,22}\beta_i = \Lambda_{i,21}$. This result together with (3.3) leads to

$$\begin{bmatrix} \Lambda_{i,22} \\ \Sigma_{ii,22} \end{bmatrix} \beta_i = \begin{bmatrix} \Lambda_{i,21} \\ \Sigma_{ii,21} \end{bmatrix} \tag{4.4}$$

Given the above structural relationship, a routine to estimate β_i is to replace $\Lambda_{i,22}, \Lambda_{i,21}, \Sigma_{ii,22}$ and $\Sigma_{ii,21}$ with their MLE and minimize the distance between the two sides of the equation with some prespecified weighting matrix. While this method is intuitive, it is not correct. The reason is that $\widehat{\Lambda}_{i,22}$ and $\widehat{\Lambda}_{i,21}$ are not consistent estimators of $\Lambda_{i,22}$ and $\Lambda_{i,21}$ due to the rotational indeterminacy, as shown in Theorem 2.1. Let $\Lambda_i^* = R'\Lambda_i$ represent the underlying parameters that the MLE corresponds to, where R is the rotation matrix defined in Theorem 2.1. Then

$$\Lambda_i^{*\prime} = \begin{bmatrix} \Lambda_{i,11}^{*\prime} & \Lambda_{i,21}^{*\prime} \\ \Lambda_{i,12}^{*\prime} & \Lambda_{i,22}^{*\prime} \end{bmatrix} = \Lambda_i' R = \begin{bmatrix} \Lambda_{i,11}' & \Lambda_{i,21}' \\ \Lambda_{i,12}' & \Lambda_{i,22}' \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} = \begin{bmatrix} \beta_i'\gamma_i^{g\prime} + \psi_i' & \beta_i'\gamma_i^{h\prime} \\ \gamma_i^{g\prime} & \gamma_i^{h\prime} \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}$$

implying

$$\Lambda_{i,21}^* = (R_{12}'\gamma_i^g + R_{22}'\gamma_i^h)\beta_i + R_{12}'\psi_i \tag{4.5}$$
$$\Lambda_{i,22}^* = R_{12}'\gamma_i^g + R_{22}'\gamma_i^h \tag{4.6}$$

From (4.5) and (4.6), we see that unless $\psi_i = 0$, $\Lambda_{i,22}^*\beta_i = \Lambda_{i,21}^*$ does not hold. But when $\psi_i = 0$, we see from (4.1) that the model is free of the endogeneity problem and the ordinary least squares method is applicable. The preceding analysis indicates that the existence of the rotational indeterminacy for loadings impedes the use of the underlying relation $\Lambda_{i,22}\beta_i = \Lambda_{i,21}$ in the estimation of β_i.

Although this result is a little disappointing, we now show that with some transformation, $\Lambda_{i,22}\beta_i = \Lambda_{i,21}$ can still be used to estimate β_i. First by $\Lambda_i^{*\prime} = \Lambda_i' R$,

$$\Lambda_{i,11}^* = (R_{11}'\gamma_i^g + R_{21}'\gamma_i^h)\beta_i + R_{11}'\psi_i \tag{4.7}$$
$$\Lambda_{i,12}^* = R_{11}'\gamma_i^g + R_{21}'\gamma_i^h \tag{4.8}$$

By the expressions (4.5)-(4.8), we have the following equation:

$$(\Lambda_{i,21}^* - \Lambda_{i,22}^*\beta_i) = R_{12}'R_{11}'^{-1}(\Lambda_{i,11}^* - \Lambda_{i,12}^*\beta_i) = V(\Lambda_{i,11}^* - \Lambda_{i,12}^*\beta_i) \tag{4.9}$$

where $V = R_{12}'R_{11}'^{-1}$, an $r_2 \times r_1$ rotational matrix. The preceding equation can be written as

$$(\Lambda_{i,22}^* - V\Lambda_{i,12}^*)\beta_i = \Lambda_{i,21}^* - V\Lambda_{i,11}^* \tag{4.10}$$

Given the above result, together with (3.3), we have

$$\begin{bmatrix} \Lambda_{i,22}^* - V\Lambda_{i,12}^* \\ \bar{\Sigma}_{ii,22} \end{bmatrix} \beta_i = \begin{bmatrix} \Lambda_{i,21}^* - V\Lambda_{i,11}^* \\ \bar{\Sigma}_{ii,21} \end{bmatrix} \tag{4.11}$$

If V is known, it suffices to replace $\Lambda_{i,11}^*, \Lambda_{i,12}^*, \Lambda_{i,21}^*, \Lambda_{i,22}^*$ with the corresponding estimates to perform the estimation, and β_i is efficiently estimated. Although V is unknown, it can be consistently estimated by (4.9) since β_i can be consistently (albeit not efficiently) estimated by $\widehat{\beta}_i^{CV} = \widehat{\Sigma}_{ii,22}^{-1}\widehat{\Sigma}_{ii,21}$. Given the above analysis, we propose the following estimation procedure:

1. *Use the maximum likelihood method to obtain the estimates $\widehat{\Sigma}_{ii}, \widehat{\Lambda}_i, \widehat{f}_t$ for all i and t.*

2. *Calculate* $\widehat{\beta}_i^{CV} = \widehat{\Sigma}_{ii,22}^{-1}\widehat{\Sigma}_{ii,21}$ *and*

$$\widehat{V} = \left[\sum_{i=1}^{N}(\widehat{\Lambda}_{i,21} - \widehat{\Lambda}_{i,22}\widehat{\beta}_i^{CV})(\widehat{\Lambda}_{i,11} - \widehat{\Lambda}_{i,12}\widehat{\beta}_i^{CV})'\right]\left[\sum_{i=1}^{N}(\widehat{\Lambda}_{i,11} - \widehat{\Lambda}_{i,12}\widehat{\beta}_i^{CV})(\widehat{\Lambda}_{i,11} - \widehat{\Lambda}_{i,12}\widehat{\beta}_i^{CV})'\right]^{-1}.$$

3. *Calculate* $\widehat{\beta}_i = (\widehat{\Delta}_i' W_i^{-1}\widehat{\Delta}_i)^{-1}\widehat{\Delta}_i' W_i^{-1}\widehat{\delta}_i$, *where* W_i *is a predetermined weighting matrix that is given in Subsection 4.1 below, and*

$$\widehat{\Delta}_i = \begin{bmatrix}\widehat{\Lambda}_{i,22} - \widehat{V}\widehat{\Lambda}_{i,12} \\ \widehat{\Sigma}_{ii,22}\end{bmatrix}, \qquad \widehat{\delta}_i = \begin{bmatrix}\widehat{\Lambda}_{i,21} - \widehat{V}\widehat{\Lambda}_{i,11} \\ \widehat{\Sigma}_{ii,21}\end{bmatrix} \tag{4.12}$$

where we call the resulting estimator the Loading-coVariance estimators, denoted by $\widehat{\beta}_i^{LV}$.

Remark 4.1 We can iterate the second and third steps by using the updated estimator of β_i to calculate $\widehat{V}$. We call the estimator resulting from this iterating procedure the *Iterated-LV* estimator, denoted by $\widehat{\beta}_i^{ILV}$. The iterated estimator has the same asymptotic representation as the LV estimator, but better finite sample performance; see the simulation results in Section 5. □

4.1 Optimal weighting matrix

To carry out the estimation procedure, we need to specify the weighting matrix W_i. According to Hansen (1982), the optimal weighting matrix is

$$W_i = T \cdot \text{avar}\left(\begin{bmatrix}\widehat{\Lambda}_{i,22} - \widehat{V}\widehat{\Lambda}_{i,12} \\ \widehat{\Sigma}_{ii,22}\end{bmatrix}\beta_i - \begin{bmatrix}\widehat{\Lambda}_{i,21} - \widehat{V}\widehat{\Lambda}_{i,11} \\ \widehat{\Sigma}_{ii,21}\end{bmatrix}\right).$$

where "avar" denotes the asymptotic variance. By the results in Theorem 2.1 and Lemma B.5 in the appendix, we can show that

$$W_i = \begin{bmatrix}R_{22\cdot1}' & 0 \\ 0 & I_K\end{bmatrix}\text{avar}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\begin{bmatrix}h_t^\star \\ v_{it}\end{bmatrix}\epsilon_{it}\right)\begin{bmatrix}R_{22\cdot1} & 0 \\ 0 & I_K\end{bmatrix} = \mathcal{R}'E\left[\frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\tilde{P}_{it}\tilde{P}_{is}'\epsilon_{it}\epsilon_{is}\right]\mathcal{R}.$$

where $\mathcal{R} = \text{diag}(R_{22\cdot1}, I_K)$ with $R_{22\cdot1} = R_{22} - R_{21}R_{11}^{-1}R_{12}$ and $\tilde{P}_{it} = [h_t^{\star\prime}, v_{it}']'$ with $h_t^\star$ implicitly defined by $f_t^\star = (g_t^{\star\prime}, h_t^{\star\prime})' = M_{ff}^{-1}f_t$. To estimate this optimal weighting matrix, we first estimate the factors by

$$\widehat{f}_t = \begin{bmatrix}\widehat{g}_t \\ \widehat{h}_t\end{bmatrix} = \Big(\sum_{i=1}^{N}\widehat{\Lambda}_i\widehat{\Sigma}_{ii}^{-1}\widehat{\Lambda}_i'\Big)^{-1}\Big(\sum_{i=1}^{N}\widehat{\Lambda}_i\widehat{\Sigma}_{ii}^{-1}z_{it}\Big).$$

Let $\widehat{p}_t = \widehat{h}_t - \widehat{V}\widehat{g}_t$. It can be shown that $\widehat{p}_t - R_{22\cdot1}'h_t^\star = o_p(1)$. Let $P_{it} = [h_t^{\star\prime}R_{22\cdot1}, v_{it}']' = \mathcal{R}'\tilde{P}_{it}$ and $\widehat{P}_{it} = [\widehat{p}_t', \widehat{v}_{it}']'$, where $\widehat{v}_{it}$ is implicitly given by

$$\widehat{u}_{it} = \begin{bmatrix}\widehat{\beta}_i^{CV\prime}\widehat{v}_{it} + \widehat{\epsilon}_{it} \\ \widehat{v}_{it}\end{bmatrix} = \dot{z}_{it} - \widehat{\Lambda}_i'\widehat{f}_t.$$

Given the above analysis, the optimal weight matrix W_i can be consistently estimated by

$$\widehat{W}_i = \frac{1}{T}\sum_{t=1}^{T}\widehat{P}_{it}\widehat{P}_{it}'\widehat{\epsilon}_{it}^2 + \sum_{\nu=1}^{S_{i,T}}K(\nu, S_{i,T})\left[\frac{1}{T}\sum_{t=\nu+1}^{T}(\widehat{P}_{it}\widehat{P}_{i,t-\nu}' + \widehat{P}_{i,t-\nu}\widehat{P}_{it}')\widehat{\epsilon}_{it}\widehat{\epsilon}_{i,t-\nu}\right].$$

where $K(v, S_{i,T})$ and $S_{i,T}$ are defined the same as in (3.9). If the Bartlett kernel is used, the estimator then is

$$\widehat{W}_i = \frac{1}{T}\sum_{t=1}^{T}\widehat{P}_{it}\widehat{P}_{it}'\widehat{\epsilon}_{it}^2 + \sum_{\nu=1}^{S_{i,T}}(1-\frac{\nu}{S_{i,T}+1})\Big[\frac{1}{T}\sum_{t=\nu+1}^{T}(\widehat{P}_{it}\widehat{P}_{i,t-\nu}' + \widehat{P}_{i,t-\nu}\widehat{P}_{it}')\widehat{\epsilon}_{it}\widehat{\epsilon}_{i,t-\nu}\Big]. \tag{4.13}$$

We have the following proposition on $\widehat{W}_i$.

Proposition 4.1 *Under Assumptions A-H, if $N/T \to \kappa \in (0,\infty)$ and $(\max S_{i,T})^{16}/T \to 0$, we have*

$$\max_{i\le N}\|\widehat{W}_i - W_i\| = o_p(1).$$

4.2 Asymptotic results

The following theorem gives the asymptotic representation of the LV estimator with some remarks following.

Theorem 4.1 *Under Assumptions A-E, H and I, when $N, T \to \infty$ and $T/N^2 \to 0$, we have that for each i,*

$$\sqrt{T}(\widehat{\beta}_i^{LV} - \beta_i) = \Big[D_i'J_i^{-1}D_i\Big]^{-1}D_i'J_i^{-1}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\begin{bmatrix}h_t^\star\\ v_{it}\end{bmatrix}\epsilon_{it} + o_p(1).$$

where $D_i = [\gamma_i^{h\prime}, \bar{\Omega}_i]'$ and

$$J_i = \text{avar}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\begin{bmatrix}h_t^\star\\ v_{it}\end{bmatrix}\epsilon_{it}\right).$$

Given Theorem 4.1, we have the following corollary.

Corollary 4.1 *Under the assumptions of Theorem 4.1, we have that for each i,*

$$\sqrt{T}(\widehat{\beta}_i^{LV} - \beta_i) \xrightarrow{d} N\Big(0, (D_i'J_i^{-1}D_i)^{-1}\Big).$$

The above asymptotic result can be alternatively written as

$$\sqrt{T}(\widehat{\beta}_i^{LV} - \beta_i) \xrightarrow{d} N\left(0, \Big[\operatorname*{plim}_{N,T\to\infty}\Delta_i'W_i^{-1}\Delta_i\Big]^{-1}\right).$$

So the limiting variance can be consistently estimated by $\big(\widehat{\Delta}_i'\widehat{W}_i^{-1}\widehat{\Delta}_i\big)^{-1}$.

Remark 4.2 We compare our LV estimator with the CCE and PC estimators in terms of limiting variance. By the fact that J_i is a diagonal matrix, the limiting variance $(D_i'J_i^{-1}D_i)^{-1}$ in Corollary 4.1 can be alternatively written as[②]

$$(D_i'J_i^{-1}D_i)^{-1} = \left[\gamma_i^{h\prime}\Big(\text{avar}(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}h_t^\star\epsilon_{it})\Big)^{-1}\gamma_i^h + \bar{\Omega}_i\Big(\text{avar}(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}v_{it}\epsilon_{it})\Big)^{-1}\bar{\Omega}_i\right]^{-1}$$

[②]For square matrices A and B, we say $A \geq B$ if $A - B$ is positive semidefinite.

$$\leq \left[\bar{\Omega}_i\Big(\text{avar}\big(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} v_{it}\epsilon_{it}\big)\Big)^{-1}\bar{\Omega}_i\right]^{-1} = \bar{\Omega}_i^{-1}\text{avar}\Big(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} v_{it}\epsilon_{it}\Big)\bar{\Omega}_i^{-1}.$$

The last expression is the limiting variance of the CV estimator, see Corollary 3.1. So the LV estimator is more efficient than the CV. Note that the CCE method, which uses the cross sectional average values of data as the instruments for unobserved factors, would deliver the same estimators as in model (1.1). In Remark 3.1, we point out that the CV estimator and the CCE estimator have the same limiting distribution. Given this, we conclude that the LV estimator is more efficient than the CCE in model (4.1).

Consider the "y" equation, which can be written as $Y_i = \alpha_i \mathbf{1}_T + X_i\beta_i + G\psi_i + \epsilon_i$. The PC estimator, according to Song (2013)[③], has the asymptotic expression

$$\widehat{\beta}_i^{PC} - \beta_i = (X_i' M_{G^\diamond} X_i)^{-1}(X_i' M_{G^\diamond}\epsilon_i),$$

where $G^\diamond = (\mathbf{1}_T, G)$. The PC estimator can be viewed a variant of our GMM estimator since it has the same limiting variance with the following GMM estimator

$$\widehat{\beta}_i^\dagger = \left[\widehat{\Delta}_i'\widehat{W}_i^{\dagger-1}\widehat{\Delta}_i\right]^{-1}\left[\widehat{\Delta}_i'\widehat{W}_i^{\dagger-1}\widehat{\delta}_i\right],$$

where $\widehat{W}_i^{\dagger-1}$ is a consistent estimator of $W_i^\dagger$ and

$$W_i^\dagger = \begin{bmatrix} R_{22\cdot1}' M_{hh\cdot g}^{-1} R_{22\cdot1} & 0 \\ 0 & \bar{\Omega}_i \end{bmatrix} = \mathcal{R}' \begin{bmatrix} M_{hh\cdot g}^{-1} & 0 \\ 0 & \bar{\Omega}_i \end{bmatrix} \mathcal{R}.$$

with $M_{hh\cdot g} = M_{hh} - M_{hg}M_{gg}^{-1}M_{gh}$. To see this, first note that

$$\widehat{\beta}_i^\dagger - \beta_i = \left[\widehat{\Delta}_i'\widehat{W}_i^{\dagger-1}\widehat{\Delta}_i\right]^{-1}\left[\widehat{\Delta}_i'\widehat{W}_i^{\dagger-1}(\widehat{\delta}_i - \widehat{\Delta}_i\beta_i)\right].$$

Second, by (B.16) in the appendix,

$$\widehat{\Delta}_i'\widehat{W}_i^{\dagger-1}\widehat{\Delta}_i = \Delta_i' W_i^{\dagger-1}\Delta_i + o_p(1) = \Big(\gamma_i^{h\prime} M_{hh\cdot g}\gamma_i^h + \bar{\Omega}_i\Big) + o_p(1) = \frac{1}{T}X_i' M_{G^\diamond} X_i + o_p(1).$$

Third, by (B.14) and (B.15), under $\sqrt{N}/T \to 0$,

$$\begin{aligned}\sqrt{T}\widehat{\Delta}_i'\widehat{W}_i^{\dagger-1}(\widehat{\delta}_i - \widehat{\Delta}_i\beta_i) &= \sqrt{T}\Delta_i' W_i^{\dagger-1}(\widehat{\delta}_i - \widehat{\Delta}_i\beta_i) + o_p(1) \\ &= \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\Big[\gamma_i^{h\prime} M_{hh\cdot g} h_t^\star + v_{it}\Big]\epsilon_{it} + o_p(1).\end{aligned}$$

The last term of the proceeding equation can be alternatively written as $\frac{1}{\sqrt{T}}X_i' M_{G^\diamond}\epsilon_i + o_p(1)$ since $M_{hh\cdot g}h_t^\star = h_t - M_{hg}M_{gg}^{-1}g_t$. Given this, we have

$$\sqrt{T}(\widehat{\beta}_i^{PC} - \beta_i) \xrightarrow{d} N\Big(0, (\Delta_i' W_i^{\dagger-1}\Delta_i)^{-1}\Delta_i' W_i^{\dagger-1} W_i W_i^{\dagger-1}\Delta_i(\Delta_i' W_i^{\dagger-1}\Delta_i)^{-1}\Big).$$

[③] We thank an anonymous referee to remind us that Assumption B(iii) in Song (2013) is problematic since it does not preclude the chance that the observed regressors for a particular unit might be equal to the unobserved factors.

According to Proposition 4.45 of White (2001),

$$(\Delta_i' W_i^{-1} \Delta_i)^{-1} \leq (\Delta_i' W_i^{\dagger -1} \Delta_i)^{-1} \Delta_i' W_i^{\dagger -1} W_i W_i^{\dagger -1} \Delta_i (\Delta_i' W_i^{\dagger -1} \Delta_i)^{-1},$$

where the equality holds if $W_i^{\dagger} = cW_i$ for some scalar constant c. So the LV estimator is also more efficient than the PC. Only when ϵ_{it} is independently and identically distributed over time, $W_i^{\dagger} = cW_i$ and the two estimators have the same limiting distribution.

We summarize the above discussion as follows. In the model with zero restrictions, our two-step method is more efficient than the CCE method because the two-step method uses more moment conditions, and is also more efficient than the PC method because the two-step method uses the optimal weights matrix. □

Remark 4.3 Consider the following model, in which zero restrictions exist in both the x equation and the y equation:

$$\begin{aligned} y_{it} &= \alpha_i + x_{it}' \beta_i + \psi_i' g_t + \epsilon_{it} \\ x_{it} &= \nu_i + \gamma_i^{h\prime} h_t + v_{it} \end{aligned} \tag{4.14}$$

where g_t and h_t are assumed to be correlated. Model (4.14) is a special case of (4.1) in view that γ_i^g is restricted to zero. We note that the LV estimator is still applicable in this case. To see this point, notice that Λ_i in model (4.14) is

$$\Lambda_i = \begin{bmatrix} \Lambda_{i,11} & \Lambda_{i,12} \\ \Lambda_{i,21} & \Lambda_{i,22} \end{bmatrix} = \begin{bmatrix} \psi_i & 0 \\ \gamma_i^h \beta_i & \gamma_i^h \end{bmatrix}.$$

The coefficient β_i can be estimated by the relations of $\Lambda_{i,21}$ and $\Lambda_{i,22}$, which is the same as Model (4.1). But for model

$$\begin{aligned} y_{it} &= \alpha_i + x_{it}' \beta_i + \psi_i' g_t + \phi_i' h_t + \epsilon_{it}, \\ x_{it} &= \nu_i + \gamma_i^{h\prime} h_t + v_{it}, \end{aligned}$$

the matrix Λ_i is

$$\Lambda_i = \begin{bmatrix} \Lambda_{i,11} & \Lambda_{i,12} \\ \Lambda_{i,21} & \Lambda_{i,22} \end{bmatrix} = \begin{bmatrix} \psi_i & 0 \\ \phi_i + \gamma_i^h \beta_i & \gamma_i^h \end{bmatrix}.$$

It is easy to see that there is no extra relation between $\Lambda_{i,21}$ and $\Lambda_{i,22}$ to identify β_i. So we cannot use the LV method to improve efficiency. □

4.3 Hypothesis testing on the over-identification

In the previous subsection, we have shown that the moment conditions implied in the loadings can improve the estimation efficiency. However, inclusion of these moment conditions, if they are not correct, would also lead to the inconsistency of the estimates. It is necessary to test the validity of these additional moment conditions. This subsection pursues this work. In the classical GMM framework, such a test is known as the Hansen-Sargan test. In the current setup, the Hansen-Sargan statistic can be formulated as

$$HS_i = T\left(\widehat{\Delta}_i \widehat{\beta}_i^{LV} - \widehat{\delta}_i\right)' \widehat{W}_i^{-1} \left(\widehat{\Delta}_i \widehat{\beta}_i^{LV} - \widehat{\delta}_i\right),$$

where

$$\widehat{\Delta}_i = \left[\frac{\widehat{\Lambda}_{i,22} - \widehat{V}\widehat{\Lambda}_{i,12}}{\widehat{\Sigma}_{ii,22}}\right], \qquad \widehat{\delta}_i = \left[\frac{\widehat{\Lambda}_{i,21} - \widehat{V}\widehat{\Lambda}_{i,11}}{\widehat{\Sigma}_{ii,21}}\right].$$

When the regressors contain additional factors h_t, each unit would have its validly added moment conditions. So, to test the validity of the added moment conditions as a whole, it is natural to take the maximum over these N statistics, i.e., we use maxHS = $\max_{i \le N} HS_i$ as the statistic. Since the null would be rejected if the Hansen-Sargan statistic is large, the maximum statistic would be favorable to the alternative. As a result, the maximum test "maxHS" would have most conservative size. To analyze the maxHS statistic, we make the following assumption:

Assumption J: For each pair (i, j) with $i \neq j$, $\log N[T^{-1} \sum_{t=1}^{T} \sum_{s=1}^{T} E(\epsilon_{it}\epsilon_{js})] \to 0$ as $N, T \to \infty$

Assumption J assumes asymptotical independence over cross section in the average sense uniformly on the pairs (i, j). This assumption is indispensable to use the extreme value theory. The same assumption is also made in Castagnetti, Rossi and Trapani (2015).

The following theorem gives the theoretical results on HS_i and maxHS.

Theorem 4.2 *Under Assumptions A-I, as $N, T \to \infty$, $N/T \to \kappa \in (0, \infty)$ and $(\max S_{i,T})^{16}/T \to 0$, we have*

(a) for each i, $HS_i \xrightarrow{d} \chi^2_{r_2}$, under $H_0 : \Delta_i\beta_i = \delta_i$; and $HS_i \longrightarrow \infty$ under $H_1 : \Delta_i\beta_i \neq \delta_i$.

(b) if Assumption J is further satisfied, under $H_0 : \Delta_i\beta_i = \delta_i$ for every i,

$$P\Big(\text{maxHS} \le 2x + 2B_N\Big) \le e^{-e^{-x}},$$

with

$$B_N = \log N + (\frac{r_2}{2} - 1)\log\log N - \log\Gamma(\frac{r_2}{2});$$

Under $H_1 : \Delta_i\beta_i \neq \delta_i$ for some i, maxHS $\longrightarrow \infty$.

Remark 4.4 The proof of result (a) is relatively easy, but the proof of result (b) requires considerable amount of work, which is essentially an application of the Extreme Value Theory (EVT). Recently, Castagnetti, Rossi and Trapani (2015) apply EVT to test factor structure in heterogeneous panels. The proof of result (b) relies critically on two mathematical tools: the maximal inequality and the Komlós-Major-Tusnády approximation. The maximal inequality says that for some convex, nondecreasing, nonzero function ψ satisfying $\psi(0) = 0$ and some boundedness conditions,

$$\Big\|\max_{1 \le i \le N} X_i\Big\|_{\psi} \le K\psi^{-1}(N) \max_{1 \le i \le N} \|X_i\|_{\psi},$$

where $\|\cdot\|_{\psi}$ denotes Orlicz norm, see Lemma 2.2.2 of Van Der Vaart and Wellner (1996). If $X_i = O_p(1)$ for each i, the above inequality implies that $\max_{1 \le i \le N} X_i = O_p(\psi^{-1}(N))$, so the magnitude would inflate with a factor $\psi^{-1}(N)$. As a result, we need to impose more stringent condition on N and T to obtain the desired result. In Theorem 4.2, we require $N/T \to \kappa \in (0,\infty)$, instead of $\sqrt{N}/T \to 0$ in the previous theorems. This condition, which is also made in a number of related studies such as Moon and Weidner (2009, 2015, 2017), seems plausible since N is usually comparable to or mildly larger than T in real data applications in economics and finance. We note that the condition $N/T \to \kappa$ can be relaxed by paying the cost of more stringent conditions on moments and correlations. For example, Fan, Liao and Mincheva (2011, 2013) shows that if the errors have exponential tails and the temporal correlation decays exponentially, the maximal results can be obtained with an inflating factor $\sqrt{\log N}$. Our analysis is closely related to Castagnetti, Rossi and Trapani (2015), but slightly different in that they make Burkholder-type conditions but we do not. As a consequence, our condition on N and T is more stringent than theirs.

Our analysis also involves the approximation of the partial sums of random variables to a gaussian random variable. This issue is first investigated by Komlós, Major and Tusnády (1975) under independently and identically distributed (i.i.d) case, and is known as Komlós-Major-Tusnády approximation hereafter in statistics. The results in the current paper are developed upon the recent work of Berkes, Liu and Wu (2014), which explore the approximation under dependence. Assumption F.1 guarantees that we can use Theorem 2.1 of Berkes, Liu and Wu (2014) in our analysis. □

Remark 4.5 Let $\mathcal{C}_{\alpha,N}$ be the critical value for a given significance level α, i.e., $P(\text{maxHS} \geq \mathcal{C}_{\alpha,N}) = \alpha$. According to Theorem 4.2, we see

$$\mathcal{C}_{\alpha,N} = 2B_N - \log|\log(1-\alpha)|^2.$$

In practice, we may use $F^{-1}_{\chi^2_{\tilde{r}_2}}(1-\frac{1}{N})$ instead of $2B_n$ to achieve a better finite sample performance, where $F_{\chi^2_r}(\cdot)$ is the cumulative distribution function of a chi-square variable with r degrees of freedom. We refer readers to Castagnetti, Rossi and Trapani (2015) for more discussions on this point. □

The LV estimation procedures need to specify the values of r_1 and r_2. For any prespecified pair $(\tilde{r}_1, \tilde{r}_2)$, one can use these two values to obtain the corresponding LV estimator, and next calculate the maxHS statistic. According to Theorem 4.2(b), one can test the null hypothesis that the prespecified pair $(\tilde{r}_1, \tilde{r}_2)$ is equal to the underlying true one (r_1, r_2). An undesirable feature of this hypothesis testing is that it is sensitive to outliers, an issue which is common in super-type statistics. To make the estimation more flexible and robust to outliers, we propose the following unified robust estimation method, which is based on the individual statistic HS_i:

1. Use the factor analysis to estimate the number of factors, denoted by $\widehat{r}$, and the covariance matrix of idiosyncratic errors in model (2.2).

2. Calculate the CV estimators and the corresponding t-statistics as suggested in Section 3.

3. Use factor analysis to determine the number of factors in the residuals $(y_{it} - x'_{it}\widehat{\beta}_i^{CV})_{N\times T}$ by setting the upper bound of the number of factors to be $\widehat{r}$. The estimated number of factors is denoted by $\widehat{r}_1$.

4. If $\widehat{r} = \widehat{r}_1$, the estimation is finished, otherwise goes to the next step.

5. Calculate the LV or ILV estimators and the corresponding t-statistics with $\widehat{r}_1$ and $\widehat{r}_2 = \widehat{r} - \widehat{r}_1$ according to the estimation procedures suggested in Section 4 above.

6. For each i, calculate the Hansen-Sargan statistic HS_i. If $HS_i \geq \mathcal{C}_{\widehat{r}_2}$, we use the CV estimator for i, otherwise use the LV or ILV estimator for i, where $\mathcal{C}_{\widehat{r}_2}$ is the critical value. The critical value can be $\widehat{r}_2 \log T$, which is borrowed from the bayesian information criterion[④]. Note that we cannot choose $\mathcal{C}_{\widehat{r}_2}$ to be the 95% quantile of $\chi^2_{\widehat{r}_2}$ since it would cause pre-test estimation issue.

The above estimation procedures aim to weaken the sensitiveness of our two-step estimators to the estimated number of factors. First, the simulations in the appendix indicate that the CV estimators still perform well if the number of factors is slightly overestimated. So the robustness is maintained to some extent in the CV estimators. Second, although the number of factors is likely to be wrongly determined for the extended model in a limited sample size which may cause bias issue of the estimators, the Hansen-Sargan statistic is helpful to remedy it. More specifically, if the added moment conditions are invalid, a direct consequence is that the Hansen-Sargan statistic would be large. With this signal, we would give up the LV or ILV estimators and move back to the CV estimator. So the robustness is also maintained in some sense. For these reasons, we call the above procedures the robust estimation method.

4.4 Estimation and inference for the mean of coefficient

Following Pesaran (2006), we make the following assumption on β_i.

Assumption I′: $\beta_i = \beta + v_i^\beta$, where v_i^β is independently and identically distributed over i with mean zero and variance Σ_β. Furthermore, v_i^β is independent with $(\epsilon_{jt}, v_{jt}')'$ for all i, j, t.

We can use the within group estimator $\widehat{\beta}^{WG} = \frac{1}{N}\sum_{i=1}^N \widehat{\beta}_i^{LV}$ to estimate β. Given that v_i^β is independent with $(\epsilon_{jt}, v_{jt}')'$ for all i, j, t, we can still use the theoretical results in the previous analysis by conditioning on a particular realization of $\beta_1, \beta_2, \ldots, \beta_N$. More specifically, for any $\varepsilon > 0$,

$$P\Big(\Big|\frac{1}{\sqrt{N}}\sum_{i=1}^N(\widehat{\beta}_i^{LV} - \beta_i)\Big| > \varepsilon\Big)$$
$$= \int P\Big(\Big|\frac{1}{\sqrt{N}}\sum_{i=1}^N(\widehat{\beta}_i^{LV} - \beta_i)\Big| > \epsilon \Big| \beta_1, \beta_2, \ldots, \beta_N\Big) g(\beta_1, \beta_2, \ldots, \beta_N) d\beta_1 \ldots d\beta_N,$$

[④] In finite sample applications, this diverging critical value would make our estimation procedure more inclined to the LV estimator. If $\|\phi_i\|$ is small but not equal to zero, the LV estimator would have a bias, which may make statistical inference size-distorted. If one is not comfortable with the bias, a smaller critical value can be chosen, say $2\widehat{r}_2 \log\log T$ suggested from the Hannan-Quinn information criterion. How to choose a critical value to best balance the bias and standard error is an interesting issue, but is beyond the scope of this paper. We left it as a future work.

where $g(\beta_1, \beta_2, \ldots, \beta_N)$ is the joint density function of $\beta_1, \beta_2, \ldots, \beta_N$. If we can show that for any $\varepsilon > 0$,

$$P\Big(\Big|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\widehat{\beta}_i^{LV} - \beta_i)\Big| > \epsilon \Big| \beta_1, \beta_2, \ldots, \beta_N\Big) \to 0,$$

together with the fact that probability is bounded by 1, we would have

$$P\Big(\Big|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\widehat{\beta}_i^{LV} - \beta_i)\Big| > \varepsilon\Big) \to 0$$

by the dominated convergence theorem. The following theorem, which is built upon the above result, gives the asymptotic property of $\widehat{\beta}^{WG}$.

Theorem 4.3 *Under Assumptions A-H and I′, as $N, T \to \infty$, $N/T \to \kappa \in (0, \infty)$ and $(\max S_{i,T})^{16}/T \to 0$, we have*

$$\sqrt{N}(\widehat{\beta}^{WG} - \beta) \xrightarrow{d} N(0, \Sigma_\beta).$$

Remark 4.6 For model (1.1), we can use $\widehat{\beta} = \frac{1}{N}\sum_{i=1}^{N}\widehat{\beta}_i^{CV}$ to estimate β. The limiting result is the same with Theorem 4.3. In addition, the limiting variance Σ_β can be consistently estimated by

$$\widehat{\Sigma}_\beta = \frac{1}{N}\sum_{i=1}^{N}(\widehat{\beta}_i^{LV} - \widehat{\beta}^{WG})(\widehat{\beta}_i^{LV} - \widehat{\beta}^{WG})'. \quad \square$$

4.5 Some extensions

A direct extension of model (4.1) is the following one

$$\begin{aligned} y_{it} &= \alpha_i + x_{it}'\beta_i + \psi_i' g_t + \phi_i' h_t + \epsilon_{it} \\ x_{it} &= \nu_i + \gamma_i^{g\prime} g_t + \gamma_i^{h\prime} h_t + v_{it} \end{aligned} \tag{4.15}$$

where ϕ_i's are observable loadings satisfying $\Phi = [\phi_1, \phi_2, \ldots, \phi_N]'$ is of full column rank. Compared with model (4.1), model (4.15) specifies that Φ is not a zero matrix, but a general observed data matrix. Now we show that our estimation idea can be used to estimate (4.15). As in the previous section, rewrite model (4.15) as

$$\begin{bmatrix} y_{it} \\ x_{it} \end{bmatrix} = \begin{bmatrix} \alpha_i + \beta_i'\nu_i \\ \nu_i \end{bmatrix} + \begin{bmatrix} \beta_i'\gamma_i^{g\prime} + \psi_i' & \beta_i'\gamma_i^{h\prime} + \phi_i' \\ \gamma_i^{g\prime} & \gamma_i^{h\prime} \end{bmatrix} \begin{bmatrix} g_t \\ h_t \end{bmatrix} + \begin{bmatrix} \beta_i' v_{it} + \epsilon_{it} \\ v_{it} \end{bmatrix}$$

Let Λ_i' be the loadings matrix in font of $f_t = (g_t', h_t')'$ and partition it into four blocks, we have

$$\Lambda_i = \begin{bmatrix} \Lambda_{i,11} & \Lambda_{i,12} \\ \Lambda_{i,21} & \Lambda_{i,22} \end{bmatrix} = \begin{bmatrix} \psi_i + \gamma_i^g \beta_i & \gamma_i^g \\ \phi_i + \gamma_i^h \beta_i & \gamma_i^h \end{bmatrix}$$

Let $\Lambda_i^* = R'\Lambda_i$ be the underlying parameters that the estimators correspond to. So we have

$$\Lambda_i^{*\prime} = \begin{bmatrix} \Lambda_{i,11}^{*\prime} & \Lambda_{i,21}^{*\prime} \\ \Lambda_{i,12}^{*\prime} & \Lambda_{i,22}^{*\prime} \end{bmatrix} = \Lambda_i' R = \begin{bmatrix} \Lambda_{i,11}' & \Lambda_{i,21}' \\ \Lambda_{i,12}' & \Lambda_{i,22}' \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix},$$

which leads to

$$\Lambda^*_{i,11} = (R'_{11}\gamma_i^g + R'_{21}\gamma_i^h)\beta_i + R'_{11}\psi_i + R'_{21}\phi_i, \qquad \Lambda^*_{i,12} = R'_{11}\gamma_i^g + R'_{21}\gamma_i^h,$$
$$\Lambda^*_{i,21} = (R'_{12}\gamma_i^g + R'_{22}\gamma_i^h)\beta_i + R'_{12}\psi_i + R'_{22}\phi_i, \qquad \Lambda^*_{i,22} = R'_{12}\gamma_i^g + R'_{22}\gamma_i^h,$$

where $\psi_i = R_{11}^{-1\prime}(\Lambda^*_{i,11} - \Lambda^*_{i,12}\beta_i - R'_{21}\gamma_i^h)$. Substituting this formula into the expressions on $\Lambda^*_{i,21}$ and $\Lambda^*_{i,22}$,

$$R'_{12}R'^{-1}_{11}(\Lambda^*_{i,11} - \Lambda^*_{i,12}\beta_i) + R'_{22\cdot 1}\phi_i = (\Lambda^*_{i,21} - \Lambda^*_{i,22}\beta_i) \tag{4.16}$$

where $R_{22\cdot 1} = R_{22} - R_{21}R_{11}^{-1}R_{12}$. Equation (4.16) together with $\Sigma_{ii,22}\beta_i = \Sigma_{ii,21}$ gives

$$\begin{bmatrix} \Lambda^*_{i,22} - V\Lambda^*_{i,12} \\ \Sigma_{ii,22} \end{bmatrix} \beta_i = \begin{bmatrix} \Lambda^*_{i,21} - V\Lambda^*_{i,11} - R'_{22\cdot 1}\phi_i \\ \Sigma_{ii,21} \end{bmatrix} \tag{4.17}$$

where $V = R'_{12}R'^{-1}_{11}$. If V and $R_{22\cdot 1}$ are known, we can use (4.17) to efficiently estimate β_i. Similarly as in the previous section, we can use $\widehat{\beta}_i^{CV}$ to get a preliminary estimators for V and $R_{22\cdot 1}$ through (4.16). This leads to the following estimation procedures:

1. *Use the maximum likelihood method to obtain the estimates* $\widehat{\Sigma}_{ii}$, $\widehat{\Lambda}_i$ *and* $\widehat{f}_t$ *for all* i *and* t.
2. *Calculate* $\widehat{\beta}_i^{CV} = \widehat{\Sigma}_{ii,22}^{-1}\widehat{\Sigma}_{ii,21}$ *and* $\widehat{V}$ *and* $\widehat{R}_{22\cdot 1}$ *by*

$$[\widehat{V}, \widehat{R}'_{22\cdot 1}] = \Big[\sum_{i=1}^N (\widehat{\Lambda}_{i,21} - \widehat{\Lambda}_{i,22}\widehat{\beta}_i^{CV})\widehat{\Pi}_i\Big]\Big[\sum_{i=1}^N \widehat{\Pi}_i\widehat{\Pi}'_i\Big]^{-1}$$

   where $\widehat{\Pi}_i = [(\widehat{\Lambda}_{i,11} - \widehat{\Lambda}_{i,12}\widehat{\beta}_i^{CV})', \phi'_i]'$.
3. *Calculate* $\widehat{\beta}_i^{LV} = (\widehat{\Delta}'_i\widehat{W}_i^{-1}\widehat{\Delta}_i)^{-1}\widehat{\Delta}'_i\widehat{W}_i^{-1}\widehat{\gamma}_i$, *where*

$$\widehat{\Delta}_i = \begin{bmatrix} \widehat{\Lambda}_{i,22} - \widehat{V}\widehat{\Lambda}_{i,12} \\ \widehat{\Sigma}_{ii,22} \end{bmatrix}, \qquad \widehat{\gamma}_i = \begin{bmatrix} \widehat{\Lambda}_{i,21} - \widehat{V}\widehat{\Lambda}_{i,11} - \widehat{R}'_{22\cdot 1}\phi_i \\ \widehat{\Sigma}_{ii,21} \end{bmatrix}$$

   and $\widehat{W}_i$ *is the predetermined weighting matrix, which is the same as (4.13).*

The asymptotic properties of the LV estimator for model (4.15) is presented in the following theorem.

Theorem 4.4 *Under Assumptions A-E, H and I, when* $N, T \to \infty$ *and* $\sqrt{T}/N \to 0$, *we have that for each* i,

$$\sqrt{T}(\hat{\beta}_i^{LV} - \beta_i) \xrightarrow{d} N(0, \digamma_{a,i}^{-1}),$$

where

$$\digamma_{a,i} = \gamma_i^{h\prime}\Big[\text{avar}\Big(\frac{1}{\sqrt{T}}\sum_{t=1}^T h_t^{\star}\epsilon_{it}\Big)\Big]^{-1}\gamma_i^h + \bar{\Omega}_i\Big[\text{avar}\Big(\frac{1}{\sqrt{T}}\sum_{t=1}^T v_{it}\epsilon_{it}\Big)\Big]^{-1}\bar{\Omega}_i.$$

The proof of Theorem 4.4 is similar as that of Theorem 4.1, and is sketched in Appendix C.

In some applications, it is of interest to include some time-invariant variables, such as gender, race, education, and so forth. To address this concern, consider the following model with time-invariant variables:

$$\begin{aligned} y_{it} &= \alpha_i + x_{it}'\beta_i + \psi_i' g_t + \phi_i' h_t^y + \epsilon_{it}, \\ x_{itp} &= \nu_{ip} + \gamma_{ip}^{g\prime} g_t + \phi_i' h_{tp}^x + v_{itp}, \quad \text{for } p = 1, 2, \ldots, k, \end{aligned} \tag{4.18}$$

where ϕ_i denotes the observed time-invariant features of unit i. In the original work of Pesaran (2006), he considers a heterogeneous-coefficients model with observed time-varying regressor (observed factors). The model considered here can be viewed as a mirror of his model. Let $x_{it} = (x_{it1}, x_{it2}, \ldots, x_{itk})'$, $\nu_i = [\nu_{i1}, \nu_{i2}, \ldots, \nu_{ik}]'$, $\gamma_i^g = [\gamma_{i1}^g, \gamma_{i2}^g, \ldots, \gamma_{ik}^g]$, $h_t^x = [h_{t1}^{x\prime}, h_{t2}^{x\prime}, \ldots, h_{tk}^{x\prime}]'$ and $v_{it} = [v_{it1}, v_{it2}, \ldots, v_{itk}]'$, the second equation now can be written as

$$x_{it} = \nu_i + \gamma_i^{g\prime} g_t + (I_k \otimes \phi_i)' h_t^x + v_{it}.$$

With some simple manipulations, we have

$$\begin{bmatrix} y_{it} \\ x_{it} \end{bmatrix} = \begin{bmatrix} \alpha_i + \beta_i'\nu_i \\ \nu_i \end{bmatrix} + \begin{bmatrix} \beta_i'\gamma_i^{g\prime} + \psi_i' & \phi_i' & \beta_i' \otimes \phi_i' \\ \gamma_i^{g\prime} & 0 & I_k \otimes \phi_i' \end{bmatrix} \begin{bmatrix} g_t \\ h_t^y \\ h_t^x \end{bmatrix} + \begin{bmatrix} \beta_i' v_{it} + \epsilon_{it} \\ v_{it} \end{bmatrix}$$

Similarly, we use Λ_i' to denote the loadings in font of $f_t = (g_t', h_t^{y\prime}, h_t^{x\prime})'$, which we partition as

$$\Lambda_i' = \begin{bmatrix} \Lambda_{i,11}' & \Lambda_{i,21}' & \Lambda_{i,31}' \\ \Lambda_{i,12}' & \Lambda_{i,22}' & \Lambda_{i,32}' \end{bmatrix} = \begin{bmatrix} \beta_i'\gamma_i^{g\prime} + \psi_i' & \phi_i' & \beta_i' \otimes \phi_i' \\ \gamma_i^{g\prime} & 0 & I_k \otimes \phi_i' \end{bmatrix}$$

Let $\Lambda_i^* = R'\Lambda_i$ be the limit of the ML estimator, where R is the rotational matrix. We partition matrix R into $R = [R_{11}\ R_{12}\ R_{13}\ ;\ R_{21}\ R_{22}\ R_{23}\ ;\ R_{31}\ R_{32}\ R_{33}]$. Let $\Lambda_{i,pq}^*$ be defined similarly as $\Lambda_{i,pq}$ for $p = 1, 2, 3$ and $q = 1, 2$. According to $\Lambda_i^* = R'\Lambda_i$, we have

$$\begin{aligned} \Lambda_{i,11}^{*\prime} &= (\psi_i' + \beta_i'\gamma_i^{g\prime})R_{11} + \phi_i' R_{21} + (\beta_i' \otimes \phi_i')R_{31}, & \Lambda_{i,12}^{*\prime} &= \gamma_i^{g\prime} R_{11} + (I_k \otimes \phi_i')R_{31}, \\ \Lambda_{i,21}^{*\prime} &= (\psi_i' + \beta_i'\gamma_i^{g\prime})R_{12} + \phi_i' R_{22} + (\beta_i' \otimes \phi_i')R_{32}, & \Lambda_{i,22}^{*\prime} &= \gamma_i^{g\prime} R_{12} + (I_k \otimes \phi_i')R_{32}, \\ \Lambda_{i,31}^{*\prime} &= (\psi_i' + \beta_i'\gamma_i^{g\prime})R_{13} + \phi_i' R_{23} + (\beta_i' \otimes \phi_i')R_{33}, & \Lambda_{i,32}^{*\prime} &= \gamma_i^{g\prime} R_{13} + (I_k \otimes \phi_i')R_{33}, \end{aligned}$$

implying

$$\begin{aligned} &\Lambda_{i,11}^* - \Lambda_{i,12}^*\beta_i = R_{11}'\psi_i + R_{21}'\phi_i, \qquad \Lambda_{i,21}^* - \Lambda_{i,22}^*\beta_i = R_{12}'\psi_i + R_{22}'\phi_i, \\ &\Lambda_{i,31}^* - \Lambda_{i,32}^*\beta_i = R_{13}'\psi_i + R_{23}'\phi_i. \end{aligned} \tag{4.19}$$

By the first equation of (4.19), $\psi_i = R_{11}^{-1\prime}(\Lambda_{i,11}^* - \Lambda_{i,12}^*\beta_i - R_{21}'\phi_i)$. Substituting this formula into the two remaining expressions to remove ψ_i, we have

$$\begin{bmatrix} \Lambda_{i,22}^* - V_a\Lambda_{i,12}^* \\ \Lambda_{i,32}^* - V_b\Lambda_{i,12}^* \end{bmatrix} \beta_i = \begin{bmatrix} \Lambda_{i,21}^* - V_a\Lambda_{i,11}^* - R_{22\cdot 1}'\phi_i \\ \Lambda_{i,31}^* - V_b\Lambda_{i,11}^* - R_{23\cdot 1}'\phi_i \end{bmatrix},$$

where $V_a = R'_{12}R_{11}^{-1\prime}$, $V_b = R'_{13}R_{11}^{-1\prime}$, $R_{22\cdot 1} = R_{22} - R_{21}R_{11}^{-1}R_{12}$ and $R_{23\cdot 1} = R_{23} - R_{21}R_{11}^{-1}R_{13}$. The above expression can be alternatively written as

$$\Lambda^*_{i,21} - \Lambda^*_{i,22}\beta_i = \begin{bmatrix} V_a & R'_{22\cdot 1} \end{bmatrix} \begin{bmatrix} \Lambda^*_{i,11} - \Lambda^*_{i,12}\beta_i \\ \phi_i \end{bmatrix}$$
$$\Lambda^*_{i,31} - \Lambda^*_{i,32}\beta_i = \begin{bmatrix} V_b & R'_{23\cdot 1} \end{bmatrix} \begin{bmatrix} \Lambda^*_{i,11} - \Lambda^*_{i,12}\beta_i \\ \phi_i \end{bmatrix}$$

As a result, the estimation procedure is as follows

1. *Use the maximum likelihood method to obtain the estimates* $\widehat{\Sigma}_{ii}$, $\widehat{\Lambda}_i$ *and* $\widehat{f}_t$ *for all* i *and* t.
2. *Calculate* $\widehat{\beta}_i^{CV} = \widehat{\Sigma}_{ii,22}^{-1}\widehat{\Sigma}_{ii,21}$ *and* $\widehat{V}_a$, $\widehat{V}_b$, $\widehat{R}_{22\cdot 1}$ *and* $\widehat{R}_{23\cdot 1}$ *by*

$$[\widehat{V}_a, \widehat{R}'_{22\cdot 1}] = \Big[\sum_{i=1}^{N}(\widehat{\Lambda}_{i,21} - \widehat{\Lambda}_{i,22}\widehat{\beta}_i^{CV})\widehat{\Gamma}'_i\Big]\Big[\sum_{i=1}^{N}\widehat{\Gamma}_i\widehat{\Gamma}'_i\Big]^{-1}$$
$$[\widehat{V}_b, \widehat{R}'_{23\cdot 1}] = \Big[\sum_{i=1}^{N}(\widehat{\Lambda}_{i,31} - \widehat{\Lambda}_{i,32}\widehat{\beta}_i^{CV})\widehat{\Gamma}'_i\Big]\Big[\sum_{i=1}^{N}\widehat{\Gamma}_i\widehat{\Gamma}'_i\Big]^{-1}$$

   where $\widehat{\Gamma}_i = [(\widehat{\Lambda}_{i,11} - \widehat{\Lambda}_{i,12}\widehat{\beta}_i^{CV})', \phi'_i]'$.
3. *Calculate* $\widehat{\beta}_i^{LV} = (\widehat{\Delta}'_i\widehat{W}_i^{-1}\widehat{\Delta}_i)^{-1}\widehat{\Delta}'_i\widehat{W}_i^{-1}\widehat{\gamma}_i$, *where*

$$\widehat{\Delta}_i = \begin{bmatrix} \widehat{\Lambda}_{i,22} - \widehat{V}_a\widehat{\Lambda}_{i,12} \\ \widehat{\Lambda}_{i,32} - \widehat{V}_b\widehat{\Lambda}_{i,12} \\ \widehat{\Sigma}_{ii,22} \end{bmatrix}, \qquad \widehat{\gamma}_i = \begin{bmatrix} \widehat{\Lambda}_{i,21} - \widehat{V}_a\widehat{\Lambda}_{i,11} - \widehat{R}'_{22\cdot 1}\phi_i \\ \widehat{\Lambda}_{i,31} - \widehat{V}_b\widehat{\Lambda}_{i,11} - \widehat{R}'_{23\cdot 1}\phi_i \\ \widehat{\Sigma}_{ii,21} \end{bmatrix}$$

   and $\widehat{W}_i$ *is the predetermined weighting matrix, which is the same as (4.13) if we treat* $(h_t^{y\prime}, h_t^{x\prime})'$ *as the* h_t *in Subsection 4.1.*

The asymptotic properties of the LV estimator for model (4.18) is presented in the following theorem.

Theorem 4.5 *Under Assumptions A-E, H and I, as* $N, T \to \infty$ *and* $\sqrt{T}/N \to 0$, *we have that for each* i,

$$\sqrt{T}(\hat{\beta}_i^{LV} - \beta_i) \xrightarrow{d} N(0, \mathcal{F}_{b,i}^{-1}),$$

where

$$\mathcal{F}_{b,i} = (I_k \otimes \phi_i)'\mathcal{I}'\Big[\text{avar}\big(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}h_t^{\star}\epsilon_{it}\big)\Big]^{-1}\mathcal{I}(I_k \otimes \phi_i) + \bar{\Omega}_i\Big[\text{avar}\big(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}v_{it}\epsilon_{it}\big)\Big]^{-1}\bar{\Omega}_i.$$

with $\mathcal{I} = [0_{kr_2\times r_2}, I_{kr_2}]'$.

The proof of Theorem 4.5 is similar as that of Theorem 4.4, and is therefore omitted.

5 Finite sample performance

In this section, we run Monte Carlo simulations to investigate the finite sample performance of the proposed estimators. The model considered in the simulation consists of three explanatory variables ($K = 3$) and two factors ($r = 2$), which can be presented as

$$\begin{aligned} y_{it} &= \alpha_i + x_{it}'\beta_i + \psi_i g_t + \phi_i h_t + \epsilon_{it} \\ &= \alpha_i + x_{it1}\beta_{i1} + x_{it2}\beta_{i2} + x_{it3}\beta_{i3} + \psi_i g_t + \phi_i h_t + \epsilon_{it}, \\ x_{itk} &= \nu_{ik} + \gamma_{ik}^g g_t + \gamma_{ik}^h h_t + v_{itk}, \qquad k = 1, 2, 3. \end{aligned}$$

where g_t and h_t are both scalars and $f_t = [g_t, h_t]'$ is generated according to $f_t = 0.8 f_{t-1} + 0.6 v_t^f$ with the elements of v_t^f drawn from $N(0,1)$ and $\text{var}(v_t^f) = I_2$. The other parameters including α_i, ν_{ik} are all generated independently from $N(0,1)$. The heterogeneous regression coefficients β_i are generated by

$$\beta_i = [0.5, 1, 1.5]' + [N(0, 0.04), N(0, 0.04), N(0, 0.04)]'$$

for each i. We consider the following different specifications on the models (M), loadings (L), errors (E):

M1 (Basic model): ψ_i and ϕ_i are random variables for all i;

M2 (Extended model): ϕ_i is zero for all i and ψ_i is a random variable.

L1 (Independent random loadings): ψ_i and ϕ_i (if not zero) are generated according to $\psi_i = 0.5 + N(0,1)$ and $\phi_i = 1 + N(0,1)$; similarly γ_{ik}^g and γ_{ik}^h are generated by $\gamma_{ik}^g = 1 + N(0,1)$ and $\gamma_{ik}^h = 0.5 + N(0,1)$.

L2 (Correlated random loadings): ψ_i and ϕ_i (if not zero) are generated from $N(0,1)$; γ_{ik}^g and γ_{ik}^h are generated according to $\gamma_{ik}^g = \psi_i + N(0,1)$ and $\gamma_{ik}^h = \phi_i + N(0,1)$.

E1 (Cross sectional homoskedasticity): φ_t is generated according to $\varphi_t = \sqrt{\text{diag}(\mathcal{U})}\widetilde{\varphi}_t$, where $\varphi_t = (\varphi_{1t}', \varphi_{2t}', \ldots, \varphi_{Nt}')'$ with $\varphi_{it} = (\epsilon_{it}, v_{it}')'$. Here $\mathcal{U}$ is the vector used to generate the cross sectional homoskedasticity or heteroskedasticity, which will be specified below. $\widetilde{\varphi}_t$, defined similarly as φ_t, is generated as follows. Let $\widetilde{v}_{\ell t}$ be the ℓ-th element of $\widetilde{v}_t$ with $\widetilde{v}_t = (\widetilde{v}_{1t}', \widetilde{v}_{2t}', \ldots, \widetilde{v}_{Nt}')'$. $\widetilde{\epsilon}_{it}$ and $\widetilde{v}_{\ell t}$ are generated separately according to

$$\begin{aligned} \widetilde{\epsilon}_{it} &= \rho_i^\epsilon \widetilde{\epsilon}_{it-1} + \varsigma_{it}^\epsilon + \sum_{h=\max(i-J,1)}^{i-1} \varrho_\epsilon \varsigma_{ht}^\epsilon + \sum_{h=i+1}^{\min(i+J,N)} \varrho_\epsilon \varsigma_{ht}^\epsilon, \qquad i = 1, \cdots, N \\ \widetilde{v}_{\ell t} &= \rho_i^v \widetilde{v}_{\ell t-1} + \varsigma_{\ell t}^v + \sum_{h=\max(\ell-J,1)}^{\ell-1} \varrho_v \varsigma_{ht}^v + \sum_{h=\ell+1}^{\min(\ell+J,NK)} \varrho_v \varsigma_{ht}^v, \qquad \ell = 1, \cdots, NK. \end{aligned}$$

All ς_{ht}^ϵ and ς_{ht}^v are drawn from $N(0,1)$. Following Ahn and Horenstein (2013), we set $\rho_i^\epsilon = \rho_i^v = 0.7$ for all i, $\varrho_\epsilon = \varrho_v = 0.3$ and $J = \min(10, N/20)$. Let $\mathcal{U}$ be a $N(K+1)$ dimensional vector with $\mathcal{U}_i$, the i-th element, equal to

$$\mathcal{U}_i = \begin{cases} \frac{26(1-\rho^2)}{4(1+2J\varrho^2)}, & \text{for } \mathbf{M1} \\ \frac{13(1-\rho^2)}{4(1+2J\varrho^2)}, & \text{for } \mathbf{M2} \end{cases}$$

E2 (Cross sectional heteroskedasticity): Let

$$L_i = \begin{bmatrix} \psi_i & \phi_i \\ \gamma_i^g & \gamma_i^h \end{bmatrix}, \qquad i = 1, 2, \ldots, N$$

and $L = (L_1', L_2', \ldots, L_N')'$ an $N(K+1) \times 2$ matrix. φ_t is generated as in **E1** except that

$$\mathcal{U}_i = \begin{cases} \frac{2(1-\rho^2)}{1+2J\varrho^2}\left(0.1 + \frac{\eta_i}{1-\eta_i}\iota_i'\iota_i\right) & \text{for } \mathbf{M1}, \\ \frac{(1-\rho^2)}{1+2J\varrho^2}\left(0.1 + \frac{\eta_i}{1-\eta_i}\iota_i'\iota_i\right) & \text{for } \mathbf{M2} \end{cases}$$

for $i = 1, 2, \cdots, N(K+1)$, where ι_i' is the ith row of L, and η_i is drawn independently from $U[u, 1-u]$ with $u = 0.1$.

Remark 5.1 Two specifications in M denotes the two models considered in the paper. M1 corresponds to the basic model, and M2 corresponds to the model with zero restrictions. We consider two different sets of loadings, L1 and L2. Under L1, the correlations among the loadings are due to non-zero means. If the means are removed, the remaining parts are actually independent. For this reason, we call L1 the independent random loadings. In contrast, the correlations of loadings under L2 are due to the fact that they share the same random ingredient. Both specifications give rise to the endogeneity problem in the y equation, but as will be seen below, the CCE estimator performs quite differently in the two setups.

We also consider the cross-sectional homoscedasticity and heteroscedasticity in the simulation, which correspond to E1 and E2, respectively. When generating heteroscedasticity, we add 0.1 to the expression, avoiding the variance being too close to zero. The variances of idiosyncratic errors are chosen to guarantee that the signal-to-noise ratio of regressors is on average 0.5 for M1 and 1 for M2. Our approach to generating the idiosyncratic errors is similar to Doz, Giannone and Reichlin (2012), Bai and Li (2014), and Ahn and Horenstein (2013). In both E1 and E2, weak cross-sectional and temporal correlations are generated.

In this section, we use the average of root mean square error (RMSE) and the average (empirical) size to evaluate the performance of estimators. We take β_1 as the example to illustrate how there two measures are calculated. The RMSE of β_1 is calculated by

$$\text{RMSE}(\beta_1) = \sqrt{\frac{1}{N\mathcal{S}} \sum_{s=1}^{\mathcal{S}} \sum_{i=1}^{N} (\widehat{\beta}_{i1}^{(s)} - \beta_{i1})^2},$$

where $\widehat{\beta}_{i1}^{(s)}$ is the estimator of β_{i1}, the first component of β_i, in the s-th experiment, and $\mathcal{S}$ is the number of repetitions. The average size of β_1 is defined by

$$\text{Average Size}(\beta_1) = \frac{1}{N\mathcal{S}} \sum_{s=1}^{\mathcal{S}} \sum_{i=1}^{N} \mathbf{1}\left(|\widehat{t}_{\beta_{i1}}^{(s)}| \leq z_{0.05}\right) \times 100\%$$

where $\widehat{t}_{\beta_{i1}}^{(s)}$ is the t-statistic for β_{i1} in the s-th experiment and $z_{0.05}$ is the two-side critical value of normal distribution under the 5% significance level. We set $\mathcal{S}$ to be 1000 in Subsections 5.1, 5.2 and 5.4, and 2000 in Subsection 5.3.

5.1 Finite sample performance of the CV estimator in the basic model

In this subsection, we examine the performance (average RMSE and average empirical size) of the two-step CV estimator in the basic model. For the purpose of comparison, we also calculate Pesaran's CCE estimator and Song's PC estimator. As pointed out in Remark 3.1, the CCE, PC and CV estimators have the same limiting distribution with an infeasible estimator, which can be obtained by the least square method, assuming that the factors are observed. For this reason, we also calculate this infeasible estimator to serve as the benchmark for comparison. Throughout the whole section, we assume that the number of factors is unknown and this value is determined by minimizing the growth ratio (GR) value of Ahn and Horenstein (2013). The largest possible value for the number of factors is set to 6.

[Insert Table 1 here]

The left panel of Table 1 presents the percentages of correctly estimating the number of factors in the basic model. Since both cross sectional and temporal correlations are present in our generated data, we see that the GR method performs poorly in the relatively small sample size, but when the sample size increases, the performance is improved dramatically.

[Insert Tables 2-5 here]

Tables 2-5 report the performance (average RMSE) of the CCE, PC, CV and infeasible (denoted by INF) estimators under different loadings and error choices in the basic model. In summary, we see that the CCE estimator performs well under independent random loadings (L1), but poorly under correlated random loadings (L2); the PC estimator performs moderately under cross sectional homoskedasticity (E1), but very poorly under cross sectional heteroskedasticity (E2); the CV estimator performs well under all setups.

First consider the different loadings choices. Under L1, the performance of the CCE estimator is considerably good in term of the RMSE. Impressively, the CCE estimator even defeats the infeasible in the small sample size such as $N = 50$ and $T = 50, 150, 250$. When the sample size grows larger, the CV and infeasible estimators catch up the CCE and perform better. In regards to the average empirical size, the best is the infeasible estimator, and the next is the CV. The CCE and PC estimators both have relatively larger size distortions. The performance of the CCE estimator is not satisfactory under L2. Not only does it have a large average RMSE, but it also has severe size distortions. In contrast, the CV estimator performs close to the infeasible estimator.

The reason for the different performance of the CCE estimator under different loading settings is that the space spanned by $\tilde{z}_t = \frac{1}{N}\sum_{i=1}^{N} \dot{z}_{it}$ with $\dot{z}_{it} = (\dot{y}_{it}, \dot{x}_{it}')'$ provides a good approximation to the space spanned by f_t under L1, but a poor approximation under L2. To see this point more clearly, consider (2.2), which can be written as $\dot{z}_{it} = \Lambda_i' f_t + \dot{u}_{it}$. Taking the average over i, we have $\tilde{z}_t = \tilde{\Lambda}' f_t + \tilde{u}_t$, where $\tilde{\Lambda}$ and $\tilde{u}_t$ are defined similarly to $\tilde{z}_t$. With some transformation, we have $f_t = (\tilde{\Lambda}\tilde{\Lambda}')^{-1}\tilde{\Lambda}(\tilde{z}_t - \tilde{u}_t)$. So a good approximation requires two conditions. First, $\tilde{z}_t$ dominates $\tilde{u}_t$ so that $\tilde{u}_t$ is negligible. Second, $\tilde{\Lambda}\tilde{\Lambda}'$ is invertible when N goes to infinity[5]. The loadings in L1 satisfy these two conditions, but the loadings in L2 violate the first one. In fact, the terms $\tilde{\Lambda}' f_t$ and $\tilde{u}_t$ are of the same magnitude under L2. So a good approximation fails. There are cases in which the second condition breaks down. For example, if all rows of Λ share the same mean, then $\tilde{\Lambda}$ is of rank one asymptotically, which in turn leads to $\tilde{\Lambda}'\tilde{\Lambda}$ being singular asymptotically. The simulation results confirm that the CCE estimator performs poorly in this case.

Consider then the different choices of the errors. Tables 4 and 5 show that the PC estimator performs poorly in the presence of cross-sectional heteroscedasticity (E2) in term of the RMSE. In addition, we find that the t-statistics for the PC estimates have severe size distortions in all the data setups. In Song (2013), he makes the assumption that the errors are independent over cross section to derive the final limiting distribution. We find in simulations that this assumption is important for the finite sample performance of the PC estimates. We note that the comparison with the PC method under the current simulation setup, where the data of X is generated with a factor structure, is a bit unfair to the PC method. As mentioned before, a remarkable advantage of the PC method is that it does not rely on the factor structure assumption on X. So when X fails to have a factor structure, we will see that the PC still works well but the CCE and our method break down.

Finally, we emphasize that although the CV estimates have better performance relative to the CCE and the PC, their t-statistics suffer mild size distortions even in the large sample size. As seen in Tables 2-5, the actual sizes, where we use the critical value for the 5% size of normal distribution, are all above 10% in all combinations of N and T. However, we note that this result is not related with our estimation method, but the poor performance of the Newey-West estimator. In fact, the t-statistics of the infeasible estimators do not perform well neither. In addition, the performance of the t-statistics of the CV is close to that of the infeasible estimator. The poor performance of the Newey-West estimator in finite sample has been well documented in the literature, see Kiefer and Vogelsang (2002), Müller (2007, 2014), Sun (2014) and reference therein. One reason for the bad performance is that the sampling uncertainty of the Newey-West estimator is ignored when we use the critical values of normal distribution, see Müller (2014).

5.2 Finite sample performance of the LV estimators in the extended model

This subsection examines the performance (average RMSE and average empirical size) of the two-step LV (ILV) estimators in the extended model. We also calculate the CCE, PC and CV estimators for comparison. But we do not calculate the infeasible estimator since its limiting distribution is different from that of the LV in the extended model. So the performance of the infeasible estimator cannot be viewed as the benchmark any more. The right panel of Table 1 presents the percentages of correctly determining the number of factors in X and Y in the extended model by the method of Ahn and Horenstein (2013). Again, we see that Ahn and Horenstein's method perform poorly in the small

[5] The rank condition in Pesaran (2006) is a necessary but not sufficient condition for invertibility of $\tilde{\Lambda}\tilde{\Lambda}'$.

sample, but well in the large sample.

[Insert Tables 6-9 here]

Tables 6-9 report the simulation results for the extended model with the number of factors determined by Ahn and Horenstein's method. Overall, these tables reaffirm what we found in the simulations of the previous subsection: the CCE does not perform well under L2, and the PC performs poorly under E2. Besides this result, there are several additional points worth mentioning. First, the CCE, PC and CV estimators are inefficient. In the panels of large size, we see that the RMSEs of the LV and ILV are significantly smaller than those of the CCE, PC and CV. This is consistent with our theoretical results in the previous section. Second, several iterations over the LV estimator are helpful to improve the performance of the *t*-statistics. Nearly in all the combinations of N and T, the empirical sizes of the *t*-statistics for the ILV are slightly smaller than the ones of the LV, which implies a mild improvement. Third, the CV dominates the LV and ILV in term of average size. This result is not surprising. As reflected in the estimation procedures, the LV and ILV estimators depend more closely on the Newey-West estimator. Their *t*-statistics have to pay the cost for this dependency.

5.3 Finite sample performance of the robust two-step estimators

A potential weakness of the simulations in the previous two subsections is that the data in each repetition is generated from the same type of model. In this subsection, we generate the data randomly from the basic model or from the extended model with equal probability. We investigate the performance of the robust estimation method suggested in Subsection 4.3.

[Insert Tables 10-13 here]

Tables 10-13 present the simulation results under different loadings type and different variance type. The results are obtained by 2000 repetitions. As seen in Tables 10-13, our estimators perform considerably well in all types of the data setup. The simulation results shown here are very similar as found in the previous subsections. So we will not repeat the analysis.

5.4 Empirical size of the over-identification tests

In this subsection, we investigate the finite sample performance of the maxHS statistic. One difficulty in this simulation is to generate errors satisfying Assumption J. In Castagnetti, Rossi and Trapani (2015), they address this issue by simply setting independence over the cross section and the time. In the current setup, we adopt the same treatment by letting $\rho_i^\epsilon = \rho_i^v = 0$ for all i and $\varrho_\epsilon = \varrho_v = 0$.

[Insert Table 14 here]

Table 14 presents the simulation results. As seen, the maxHS statistic suffers severe size distortion when the sample size is small, but its performance is much improved as the sample size grows large. The maxHS statistic calculated by the ILV estimator performs much better than the one of the LV. When the sample size is large, say $N = 150, T = 250$, the empirical size of the HSmax statistic calculated by the ILV is close to the nominal size (5%).[6]

6 Conclusion

This paper considers the estimation of heterogeneous coefficients in panel data models with common shocks. We propose a novel two-step method to estimate heterogeneous coefficients, in which the ML method is first used to estimate the loadings and variances of the idiosyncratic errors in a pure factor model, and heterogeneous coefficients are then estimated based on the estimates and structural relations implied by the model. Asymptotic properties of the proposed estimators including the asymptotic representations and limiting distributions are investigated and provided. We extend our method to the models with restrictions on the partial loadings in the Y equation. We point out that efficiency can be gained by using the information contained in the loadings. The asymptotic representation and limiting distribution of the new two-step estimator are studied. We also consider the model with time-invariant regressors.

Our two-step method depends on the assumption that the observed data has a factor structure. If the model is dynamic, this assumption would break down. So it is an open question whether our two-step method can be applied in the dynamic models. In addition, our two-step method is applicable to estimate the homogeneous coefficient with some appropriate modifications. This gives rise to a question that whether our two-step estimator is more efficient than the existing ones in a general correlations setup, particularly in comparison with the full information ML estimator proposed by Bai and Li (2014). We will investigate these two issues in the future work.

References

Ahn, S. C., and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. Econometrica, 81(3), 1203-1227.

Ahn, S. C., Lee, Y. H., and Schmidt, P. (2001). GMM estimation of linear panel data models with time-varying individual effects. Journal of Econometrics, 101(2), 219-255.

Ahn, S. C., Lee, Y. H., and Schmidt, P. (2013). Panel data models with multiple time-varying individual effects. Journal of Econometrics, 174(1), 1-14.

[6] As documented in the literature, the performance of the Newey-West estimator is related with the strength of autocorrelations. The stronger correlations over time, the worse performance of the Newey-West estimators. Since the idiosyncratic errors have no correlations in the current data generating process, we see that the performance of the statistic has been significantly improved.

Anderson, T. W. and Rubin, H. (1956). Statistical inference in factor analysis. In Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954-1955, Vol. V 111-150. University California Press, Berkeley.

Andrews, D. W. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica, 59(3), 817-858.

Bai, J. (2003). Inferential theory for factor models of large dimensions. Econometrica, 71(1), 135-171.

Bai, J. (2009). Panel data models with interactive fixed effects. Econometrica, 77(4), 1229–1279.

Bai, J. and Li, K. (2012). Statistical analysis of factor models of high dimension, The Annals of Statistics, 40(1), 436–465.

Bai, J. and Li, K. (2014). Theory and methods of panel data models with interactive effects, The Annals of Statistics, 42(1), 142–170.

Bai, J. and Li, K. (2016). Maximum likelihood estimation and inference of approximate factor mdoels of high dimension, Review of Economics and Statistics, 98(2), 298-309.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. Econometrica, 70(1), 191-221.

Berkes, I., Liu, W., and Wu, W. B. (2014). Komlós-Major-Tusnády approximation under dependence. The Annals of Probability, 42(2), 794-817.

Castagnetti, C., Rossi, E., and Trapani, L. (2015). Inference on factor structures in heterogeneous panels. Journal of Econometrics, 184(1), 145-157.

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

Davidson, J. (1994). Stochastic limit theory: An introduction for econometricians. Oxford University Press.

Doz, C., Giannone, D., and Reichlin, L. (2012). A two-step estimator for large approximate dynamic factor models based on Kalman filtering, Journal of Econometrics, 164, 188-205.

Doz, C., Giannone, D., and Reichlin, L. (2012). A quasi maximum likelihood approach for large approximate dynamic factor models. Review of Economics and Statistics, 94, 1014-1024.

Eberhardt, M., Helmers, C., and Strauss, H. (2013). Do spillovers matter when estimating private returns to R&D? Review of Economics and Statistics, 95(2), 436-448.

Fan, J., Liao, Y., and Mincheva, M. (2011). High dimensional covariance matrix estimation in approximate factor models. Annals of statistics, 39(6), 3320.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 75(4), 603-680.

Geweke, John. (1977). The dynamic factor analysis of economic time series, in Dennis J. Aigner and Arthur S. Goldberger (eds.) Latent Variables in Socio-Economic Models (Amsterdam: North-Holland).

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. Econometrica, 1029-1054.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. The Annals of Mathematical Statistics, 40(2), 633-643.

Kapetanios, G., Pesaran, M. H., and Yamagata, T. (2011). Panels with non-stationary multifactor error structures. Journal of Econometrics, 160(2), 326-348.

Kiefer, N. M., and Vogelsang, T. J. (2002). Heteroskedasticity-autocorrelation robust standard errors using the Bartlett kernel without truncation. Econometrica, 70(5), 2093-2095.

Komlós, J., Major, P., and Tusnády, G. (1975). An approximation of partial sums of independent RV'-s, and the sample DF. I. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, 32(1-2), 111-131.

Lawley, D. N. and Maxwell, A. E. (1971). Factor Analysis as a Statistical Method, 2nd ed. Elsevier, New York.

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

Li, D., Qian, J., and Su, L. (2016). Panel data models with interactive fixed effects and multiple structural breaks. Journal of the American Statistical Association, 111(516), 1804-1819.

Lin, C.-C and Ng, S. (2012). Estimation of panel data models with parameter heterogeneity when group memebership is unknown. Jounral of Econometric Method, 1, 42-55.

Moon, H., and Weidner, M. (2009). Likelihood expansion for panel regression models with factors. Unpubslished manuscript, USC.

Moon, H. R., and Weidner, M. (2015). Linear regression for panel with unknown number of factors as interactive fixed effects. Econometrica, 83(4), 1543-1579.

Moon, H. R., and Weidner, M. (2017). Dynamic linear panel regression models with interactive fixed effects. Econometric Theory, 33(1), 158-195.

Müller, U. K. (2007). A theory of robust long-run variance estimation. Journal of Econometrics, 141(2), 1331-1352.

Müller, U. K. (2014). HAC corrections for strongly autocorrelated time series. Journal of Business & Economic Statistics, 32(3), 311-322.

Newey, W. K., and McFadden, D. (1994). Large sample estimation and hypothesis testing. Handbook of Econometrics, 4, 2111-2245.

Newey, W. K., and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica, 55(3), 703-708.

Onatski, A. (2009). Testing hypotheses about the number of factors in large factor models. Econometrica, 77(5), 1447-1479.

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. Journal of Econometrics, 168(2), 244-258.

Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure, Econometrica, 74(4), 967–1012.

Ross, S. A. (1976). The arbitrage theory of capital asset pricing, Journal of Economic Theory,**13(3)**, 341–360.

Sargent, T. J., and Sims, C. A. (1977). Business cycle modeling without pretending to have too much a priori economic theory. New methods in business cycle research, 1, 145-168.

Stock, J. H., and Watson, M. W. (1998). Diffusion indexes (No. w6702). National Bureau of Economic Research.

Song, M. (2013). Asymptotic theory for dynamic heterogeneous panels with cross-sectional dependence and its applications. Manuscipt, Columbia University.

Su, L., and Jin, S. (2012). Sieve estimation of panel data models with cross section dependence. Journal of Econometrics, 169(1), 34-47.

Su, L., Jin, S., and Zhang, Y. (2013). Specification test for panel data models with interactive fixed effects. Journal of Econometrics, 186(1), 222-244.

Su, L. and G. Ju, (2018). Identifying latent grouped patterns in panel data models with interactive fixed effects, Journal of Econometrics, 206(2), 554-573.

Su, L., Shi, Z., and Phillips, P. C. (2016). Identifying latent structures in panel data. Econometrica, 84(6), 2215-2264.

Sun, Y. (2014). Fixed-smoothing asymptotics in a two-step generalized method of moments framework. Econometrica, 82(6), 2327-2370.

Swamy, P. A. (1970). Efficient inference in a random coefficient regression model. Econometrica, 38(2), 311-323.

Tucker, L. R. (1958). An inter-battery method of factor analysis. Psychometrika, 23(2), 111-136.

Van Der Vaart A. W., and Wellner J. A. (1996). Weak convergence and empirical processes with applications to statistics. Springer.

White, H. (2001). Asymptotic theory for econometricians. Emerald Group Publishing Limited.

Table 1: The percentage of correctly estimating the number of factors

		M1 (%)				M2 (%)			
N	**T**	**L1+E1**	**L2+E1**	**L1+E2**	**L2+E2**	**L1+E1**	**L2+E1**	**L1+E2**	**L2+E2**
50	50	27.2	55.5	35.0	50.0	25.6	22.8	45.5	41.4
100	50	25.7	55.9	35.5	54.1	25.6	19.8	51.3	47.1
150	50	25.6	59.4	37.9	58.5	23.8	17.4	46.0	41.4
50	150	63.6	96.1	75.9	91.5	87.8	72.3	93.5	82.2
100	150	71.4	98.8	84.1	97.5	95.1	83.3	98.8	95.3
150	150	68.2	98.6	85.6	97.6	93.2	81.7	98.8	96.3
50	250	84.4	99.4	89.5	97.3	97.5	90.4	96.3	84.2
100	250	94.9	99.9	97.8	99.9	99.5	96.8	100.0	97.6
150	250	91.1	99.9	97.2	99.9	98.6	97.2	100.0	99.2

Table 2: The performance (average RMSE and average size) of the four estimators in the setup of the basic model, independent random loadings and cross-sectional homoskedasticity

		CCE			PC			CV			INF		
N	T	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}
Average RMSE (×100)													
50	50	31.30	32.54	31.24	31.46	32.91	31.53	31.99	33.53	32.00	31.58	32.96	31.63
100	50	33.30	34.90	33.21	34.00	35.81	33.93	34.93	36.84	34.90	34.73	36.56	34.57
150	50	36.71	38.79	36.83	36.99	39.01	36.86	38.66	40.98	38.63	38.90	41.21	38.92
50	150	20.11	21.20	20.05	21.36	22.69	21.33	20.12	21.33	20.12	18.49	19.55	18.44
100	150	20.57	21.76	20.61	23.02	24.59	23.10	21.70	23.03	21.74	20.33	21.49	20.32
150	150	22.81	24.25	22.90	26.07	27.89	26.14	24.39	26.07	24.47	22.83	24.32	22.86
50	250	16.96	17.97	16.92	17.98	19.18	17.94	15.61	16.51	15.60	14.38	15.08	14.40
100	250	16.76	17.87	16.82	19.13	20.62	19.24	16.22	17.25	16.27	15.80	16.76	15.81
150	250	18.63	19.94	18.73	22.50	24.26	22.62	18.45	19.78	18.53	17.70	18.88	17.75
Average Size (%)													
50	50	26.95	26.65	26.72	31.92	31.84	31.62	26.09	26.11	26.03	22.85	22.54	23.17
100	50	25.59	25.59	25.47	29.86	30.06	29.46	25.28	25.48	25.39	23.16	22.86	22.89
150	50	25.58	25.71	25.61	29.78	30.46	29.87	25.36	25.60	25.29	22.98	22.92	23.04
50	150	20.75	21.12	20.62	25.77	26.45	26.02	18.19	18.38	18.17	14.03	14.12	13.78
100	150	18.11	18.48	18.18	24.65	25.72	24.88	17.15	17.37	17.31	13.87	13.88	13.85
150	150	18.36	18.75	18.54	26.63	27.78	26.88	17.49	17.95	17.55	13.94	14.07	14.02
50	250	20.45	20.92	20.50	25.74	26.74	25.37	14.53	14.93	14.54	11.14	11.09	11.29
100	250	16.72	17.34	16.85	24.14	25.45	24.53	12.78	12.97	12.88	11.26	11.39	11.34
150	250	16.98	17.81	17.22	27.69	29.50	28.06	13.22	13.56	13.27	11.15	11.29	11.34

Table 3: The performance (average RMSE and average size) of the four estimators in the setup of the basic model, correlated random loadings and cross-sectional homoskedasticity

		CCE			PC			CV			INF		
N	T	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}
Average RMSE (×100)													
50	50	31.86	33.16	31.53	31.28	32.60	31.18	31.74	33.41	31.69	31.65	33.29	31.67
100	50	33.66	35.42	33.59	33.46	35.14	33.30	34.55	36.47	34.37	34.85	36.83	34.74
150	50	37.05	38.98	36.85	36.20	38.11	36.13	38.26	40.49	38.13	38.79	41.14	38.70
50	150	21.76	22.61	21.82	21.75	22.67	21.71	18.92	19.90	18.87	18.61	19.58	18.59
100	150	22.03	23.26	22.01	23.44	24.93	23.54	20.51	21.76	20.51	20.37	21.61	20.35
150	150	24.28	25.74	24.33	26.14	27.71	26.12	22.84	24.27	22.84	22.74	24.13	22.73
50	250	19.01	19.91	18.91	18.95	20.10	18.91	14.59	15.45	14.56	14.34	15.16	14.30
100	250	18.66	19.79	18.66	20.73	22.05	20.69	15.91	16.84	15.90	15.80	16.70	15.79
150	250	20.74	22.17	20.83	23.53	25.08	23.61	17.82	18.90	17.77	17.73	18.78	17.68
Average Size (%)													
50	50	29.46	29.55	29.20	32.79	32.44	32.30	24.61	24.49	24.25	23.22	23.04	22.88
100	50	27.50	27.67	27.43	30.99	31.55	30.53	24.03	24.10	23.77	23.01	23.12	22.80
150	50	27.97	27.86	27.62	31.33	32.05	31.45	23.75	23.95	23.70	23.01	22.93	22.92
50	150	26.34	26.25	26.61	30.14	30.35	30.23	15.15	15.42	15.42	13.85	14.11	14.17
100	150	22.37	22.82	22.36	29.11	30.11	29.24	14.43	14.63	14.59	13.85	13.94	13.88
150	150	22.84	23.42	22.90	31.30	32.01	31.25	14.34	14.41	14.50	13.86	13.85	13.95
50	250	27.90	28.16	27.51	31.22	32.38	31.24	12.26	12.39	12.27	11.26	11.19	11.06
100	250	22.45	23.32	22.62	30.82	31.99	30.75	11.83	11.72	11.62	11.29	11.13	11.10
150	250	23.29	24.08	23.35	34.20	35.34	34.26	11.64	11.59	11.54	11.24	11.13	11.12

Table 4: The performance (average RMSE and average size) of the four estimators in the setup of the basic model, independent random loadings and cross-sectional heteroskedasticity

		CCE			PC			CV			INF		
N	T	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}
Average RMSE (×100)													
50	50	70.38	73.35	70.54	259.73	374.69	316.98	77.43	82.77	78.14	78.35	83.62	78.90
100	50	78.70	80.63	78.33	420.62	194.31	263.95	85.61	88.74	85.94	88.52	92.57	88.86
150	50	86.09	88.72	84.51	886.46	473.54	170.40	95.48	100.20	94.24	98.71	103.83	96.74
50	150	45.14	47.94	45.71	158.18	155.48	143.87	47.56	50.09	48.30	46.10	47.98	46.60
100	150	49.44	52.23	49.91	178.02	183.40	164.33	52.06	55.41	51.89	50.83	54.08	51.17
150	150	54.45	57.67	54.61	92.65	72.41	81.18	58.45	62.57	58.18	57.33	61.46	57.29
50	250	38.67	40.99	38.58	112.20	130.08	107.46	37.02	39.80	37.97	35.27	37.88	36.27
100	250	42.14	43.60	41.55	95.00	128.44	88.35	41.14	42.42	40.25	40.51	41.92	39.60
150	250	45.71	48.52	46.07	60.75	73.84	77.54	44.66	47.96	45.13	44.06	47.08	44.59
Average Size (%)													
50	50	29.00	29.72	29.43	35.32	36.05	35.63	26.62	26.55	26.43	22.85	22.50	23.14
100	50	29.08	29.21	29.17	34.60	35.11	34.51	26.24	26.53	26.20	23.14	22.82	22.89
150	50	29.64	29.95	29.59	35.13	35.84	35.31	26.33	26.50	26.20	22.97	22.91	23.03
50	150	25.64	26.41	25.53	34.87	35.66	34.69	17.99	18.44	17.97	14.03	14.10	13.76
100	150	25.04	25.80	25.00	34.21	35.44	34.38	16.55	16.75	16.50	13.86	13.88	13.85
150	150	26.01	27.06	26.09	35.59	37.13	35.79	16.56	16.89	16.61	13.91	14.07	14.02
50	250	26.74	27.39	26.91	37.54	38.51	37.76	14.35	14.57	14.44	11.16	11.13	11.30
100	250	25.53	26.81	25.73	36.78	38.32	36.90	12.47	12.64	12.57	11.27	11.42	11.34
150	250	26.95	28.41	27.12	38.89	40.27	39.04	12.44	12.73	12.46	11.16	11.34	11.37

Table 5: The performance (average RMSE and average size) of the four estimators
in the setup of the basic model, correlated random loadings and cross-sectional heteroskedasticity

		CCE			PC			CV			INF		
N	T	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}
Average RMSE (×100)													
50	50	48.49	49.09	47.77	208.64	173.96	184.80	50.70	53.10	50.12	50.73	53.05	50.59
100	50	51.79	53.52	51.65	104.47	163.92	262.64	54.98	57.04	54.20	56.00	58.12	55.33
150	50	55.84	58.77	56.53	147.61	173.54	95.07	60.03	63.41	60.64	61.20	65.06	61.74
50	150	34.80	36.79	34.92	92.33	114.59	98.79	30.31	32.84	30.19	29.53	32.09	29.45
100	150	37.30	38.79	37.19	86.70	95.21	82.04	32.79	34.53	32.49	32.46	34.15	32.02
150	150	40.64	43.15	41.03	72.97	81.29	52.90	36.36	38.87	36.06	36.11	38.61	35.80
50	250	31.30	32.38	30.81	96.47	90.10	80.93	23.10	24.62	23.45	22.56	24.04	22.95
100	250	33.51	35.14	33.60	102.57	68.14	148.82	25.39	26.95	25.26	25.15	26.65	25.07
150	250	36.78	38.84	36.96	45.35	48.59	45.10	28.39	30.02	28.02	28.24	29.77	27.82
Average Size (%)													
50	50	34.66	35.18	34.56	36.04	36.55	35.67	25.73	25.99	25.42	23.23	22.93	22.84
100	50	34.53	35.05	34.57	34.94	35.66	34.89	24.92	25.21	24.74	22.97	23.14	22.81
150	50	36.11	36.73	36.07	36.12	36.95	35.89	24.85	25.18	24.75	23.00	22.89	22.91
50	150	35.55	36.36	35.96	35.93	36.66	36.42	15.69	16.05	16.01	13.89	14.11	14.17
100	150	35.52	36.65	36.02	36.03	37.15	35.97	14.80	14.96	14.82	13.88	13.97	13.91
150	150	37.52	38.77	37.29	38.92	40.07	38.99	14.63	14.57	14.69	13.89	13.87	14.00
50	250	39.08	39.76	38.74	39.16	40.21	39.00	12.53	12.78	12.39	11.30	11.24	11.08
100	250	38.93	40.56	39.30	39.27	40.69	39.48	11.89	11.73	11.75	11.32	11.15	11.12
150	250	41.20	42.76	41.34	42.75	44.34	42.75	11.70	11.66	11.58	11.26	11.18	11.14

Table 6: The performance (average RMSE and average size) of the five estimators
in the setup of extended model, independent random loadings and cross-sectional homoskedasticity

		CCE			PC			CV			LV			ILV		
N	T	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}
Average RMSE (×100)																
50	50	29.01	30.08	28.87	28.46	29.53	28.53	31.33	32.72	31.26	30.95	32.33	30.94	31.47	32.78	31.36
100	50	31.82	33.17	31.71	30.76	31.94	30.62	33.77	35.55	33.70	33.21	34.88	33.17	33.72	35.46	33.73
150	50	35.12	36.98	35.23	33.17	34.65	33.13	37.90	40.07	37.88	37.25	39.25	37.27	37.68	39.64	37.67
50	150	17.91	18.82	17.95	17.97	18.89	17.99	18.77	19.84	18.69	17.43	18.31	17.36	17.45	18.32	17.38
100	150	19.13	20.12	19.17	19.20	20.22	19.21	20.51	21.69	20.53	18.80	19.72	18.82	18.79	19.72	18.82
150	150	21.23	22.42	21.28	21.72	22.84	21.76	22.93	24.42	22.95	20.81	21.80	20.82	20.78	21.76	20.77
50	250	14.75	15.46	14.73	14.75	15.43	14.72	14.58	15.33	14.60	13.39	13.93	13.42	13.38	13.89	13.40
100	250	15.33	16.20	15.36	15.49	16.36	15.59	15.90	16.86	15.91	14.31	15.04	14.39	14.29	15.01	14.36
150	250	16.92	18.01	17.01	17.83	18.90	17.92	17.79	18.99	17.85	15.84	16.65	15.88	15.80	16.61	15.84
Average Size (%)																
50	50	25.98	25.73	25.86	29.30	29.28	29.17	24.61	24.80	24.82	26.77	26.95	26.99	26.39	26.36	26.54
100	50	24.82	24.87	24.76	27.45	27.17	27.27	24.03	24.09	23.94	26.04	25.90	25.89	25.72	25.54	25.62
150	50	24.82	25.01	24.88	27.64	27.78	27.53	24.20	24.38	24.15	25.86	26.07	25.81	25.51	25.64	25.47
50	150	18.61	19.08	18.56	20.02	20.43	19.93	15.30	15.49	15.22	17.11	17.31	16.98	16.52	16.70	16.32
100	150	16.56	16.84	16.62	18.59	18.96	18.58	14.74	14.75	14.69	17.10	16.90	16.93	16.42	16.28	16.31
150	150	16.71	17.03	16.96	20.30	20.70	20.37	14.56	14.80	14.59	16.89	17.09	17.01	16.18	16.36	16.25
50	250	17.50	17.83	17.50	18.49	18.70	18.57	12.20	12.17	12.22	13.54	13.42	13.78	13.25	12.91	13.37
100	250	14.73	14.99	14.86	16.85	17.36	17.06	11.79	11.93	11.84	13.24	13.48	13.39	12.80	13.00	12.97
150	250	14.63	15.20	14.92	18.98	19.67	19.23	11.63	11.82	11.80	13.27	13.48	13.35	12.80	12.98	12.86

Table 7: The performance (average RMSE and average size) of the five estimators
in the setup of extended model, correlated random loadings and cross-sectional homoskedasticity

		CCE			PC			CV			LV			ILV		
N	T	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}
Average RMSE (×100)																
50	50	29.73	30.82	29.65	29.03	30.14	28.94	30.82	32.23	30.71	30.35	31.66	30.33	30.50	31.85	30.56
100	50	32.20	33.68	32.15	31.16	32.34	31.03	33.64	35.31	33.41	33.25	34.81	33.01	33.36	34.94	33.13
150	50	35.41	37.11	35.41	33.50	35.00	33.57	37.06	39.10	37.05	36.61	38.54	36.54	36.76	38.69	36.69
50	150	20.16	21.07	20.09	19.26	20.27	19.31	18.45	19.51	18.54	17.40	18.26	17.44	17.44	18.28	17.47
100	150	20.78	21.90	20.83	20.86	21.86	20.92	20.19	21.37	20.14	18.68	19.56	18.60	18.74	19.62	18.66
150	150	22.85	23.95	22.70	23.81	24.95	23.72	22.60	23.94	22.51	20.69	21.69	20.65	20.72	21.72	20.69
50	250	17.64	18.41	17.54	16.27	16.98	16.19	14.51	15.26	14.47	13.37	13.91	13.36	13.38	13.92	13.38
100	250	17.52	18.64	17.59	17.27	18.23	17.32	15.80	16.83	15.87	14.34	15.05	14.40	14.36	15.07	14.41
150	250	19.27	20.49	19.27	20.49	21.57	20.45	17.74	18.84	17.71	15.83	16.53	15.80	15.85	16.54	15.82
Average Size (%)																
50	50	28.91	29.00	28.83	30.68	30.76	30.41	24.23	24.03	23.78	25.32	25.17	24.95	25.16	24.93	24.73
100	50	27.00	27.41	27.17	29.60	29.54	29.31	23.48	23.66	23.39	24.30	24.49	24.12	24.08	24.27	23.93
150	50	27.29	27.30	27.12	30.19	30.49	29.92	23.28	23.34	23.33	24.09	24.20	24.07	23.96	24.00	23.87
50	150	24.89	25.51	24.73	23.80	24.40	23.80	14.84	14.95	14.91	16.02	16.27	16.15	15.72	15.93	15.88
100	150	21.30	21.76	21.52	23.23	23.57	23.30	14.40	14.50	14.45	15.95	16.05	16.00	15.65	15.68	15.62
150	150	21.44	21.69	21.26	26.22	26.57	26.08	14.52	14.55	14.49	16.18	16.12	16.05	15.74	15.75	15.68
50	250	26.13	26.37	25.96	22.73	23.11	22.72	11.92	12.11	11.87	12.94	13.08	13.06	12.67	12.91	12.79
100	250	20.90	21.81	21.04	21.50	22.04	21.53	11.62	11.74	11.72	12.99	13.09	13.00	12.73	12.80	12.72
150	250	21.32	22.01	21.09	25.96	26.55	25.87	11.51	11.69	11.48	12.83	13.06	12.87	12.54	12.74	12.62

Table 8: The performance (average RMSE and average size) of the five estimators
in the setup of extended model, independent random loadings and cross-sectional heteroskedasticity

		CCE			PC			CV			LV			ILV		
N	T	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}
Average RMSE (×100)																
50	50	41.83	44.37	41.90	331.42	533.52	219.16	49.37	52.48	49.63	47.47	50.83	47.67	47.95	51.24	48.16
100	50	47.10	47.90	46.54	278.27	212.75	265.08	53.86	55.98	54.00	51.29	52.30	51.48	51.98	52.71	51.81
150	50	51.04	52.49	50.56	161.27	256.35	316.73	60.25	63.38	60.30	56.87	59.13	56.63	57.34	59.69	57.30
50	150	26.36	27.95	27.21	118.85	125.09	120.60	28.89	30.65	29.98	25.89	27.03	26.54	25.69	27.00	26.28
100	150	28.80	30.05	29.09	108.31	127.13	93.95	32.20	34.36	32.01	27.85	29.38	28.06	27.60	29.21	27.92
150	150	32.00	33.59	31.85	113.45	87.07	117.23	36.36	38.96	36.16	30.88	32.46	30.49	30.63	32.24	30.36
50	250	22.49	23.74	22.60	105.29	116.74	129.38	23.06	24.40	23.11	20.03	21.29	20.06	19.87	21.15	19.88
100	250	24.23	25.16	24.24	62.06	107.36	85.64	25.71	26.47	24.93	21.75	22.15	21.22	21.66	22.06	21.14
150	250	26.43	27.65	26.52	83.50	94.99	123.63	27.82	29.77	28.30	23.22	24.66	23.87	23.01	24.45	23.70
Average Size (%)																
50	50	28.29	28.28	28.34	31.62	31.55	31.75	25.31	25.10	25.32	28.67	28.55	28.76	28.00	27.92	28.02
100	50	27.88	27.96	27.95	29.99	29.90	29.68	24.51	24.60	24.26	27.50	27.70	27.40	26.60	26.71	26.62
150	50	28.37	28.70	28.32	30.59	30.87	30.51	24.65	24.86	24.57	27.71	27.88	27.56	26.85	27.01	26.77
50	150	23.86	24.81	23.89	27.76	28.30	27.31	15.29	15.41	15.16	17.61	17.47	17.38	16.52	16.47	16.39
100	150	22.71	23.35	22.82	25.58	26.12	25.58	14.55	14.55	14.54	17.26	17.30	17.23	16.10	16.08	16.03
150	150	23.66	24.48	23.71	26.06	26.91	26.09	14.56	14.70	14.60	17.40	17.66	17.53	16.13	16.35	16.32
50	250	24.48	24.90	24.35	29.34	29.54	29.15	12.34	12.29	12.34	13.95	13.69	13.98	13.34	13.14	13.35
100	250	23.06	24.18	23.16	26.35	27.13	26.47	11.82	11.95	11.95	13.49	13.86	13.70	12.75	13.16	13.01
150	250	23.84	25.00	24.16	27.11	27.81	27.22	11.66	11.82	11.76	13.60	13.79	13.78	12.81	12.88	12.89

Table 9: The performance (average RMSE and average size) of the five estimators
in the setup of extended model, correlated random loadings and cross-sectional heteroskedasticity

		CCE			PC			CV			LV			ILV		
N	T	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}
Average RMSE (×100)																
50	50	37.59	39.11	37.53	227.41	294.27	153.09	38.77	40.90	37.94	37.62	39.03	36.42	37.80	39.18	36.82
100	50	41.48	42.45	41.30	178.55	153.24	164.86	42.38	44.66	42.57	40.11	42.19	40.43	40.32	42.44	40.63
150	50	45.44	46.98	45.29	148.43	78.28	105.04	47.44	49.49	46.89	44.94	46.41	44.33	45.06	46.71	44.39
50	150	28.78	30.86	29.48	131.38	108.23	132.53	23.01	25.08	23.71	21.00	22.33	21.47	20.98	22.26	21.44
100	150	30.99	32.59	31.17	97.96	142.40	90.30	25.46	27.03	25.78	21.96	22.94	22.44	21.96	22.88	22.49
150	150	34.33	35.92	34.64	124.16	90.41	89.23	28.62	30.72	29.22	24.29	25.56	24.74	24.24	25.51	24.67
50	250	26.84	28.22	27.12	111.85	118.67	92.93	18.40	19.48	18.38	16.38	17.39	16.40	16.36	17.38	16.37
100	250	28.58	30.12	28.54	83.03	63.78	86.87	19.97	21.21	20.02	17.07	17.93	17.14	17.04	17.92	17.15
150	250	32.25	33.43	32.01	70.55	67.81	67.37	22.40	23.80	22.55	18.76	19.61	18.82	18.75	19.61	18.81
Average Size (%)																
50	50	36.62	37.05	36.34	31.64	32.01	31.32	24.87	24.43	24.49	26.75	26.54	26.49	26.40	26.02	26.10
100	50	36.87	36.99	36.40	30.56	30.67	30.58	24.00	24.07	23.61	25.87	26.08	25.70	25.30	25.45	25.04
150	50	37.81	38.64	37.79	31.07	31.16	31.06	23.71	23.74	23.86	25.42	25.52	25.49	24.84	24.85	24.91
50	150	38.28	39.14	38.64	27.86	28.43	27.98	14.93	15.08	15.07	16.17	16.36	16.27	15.68	15.84	15.78
100	150	38.53	39.23	38.44	26.07	26.58	26.03	14.42	14.61	14.42	16.04	16.20	16.12	15.56	15.72	15.74
150	150	39.86	40.84	40.09	26.92	27.49	26.87	14.52	14.53	14.46	16.30	16.23	16.21	15.74	15.66	15.63
50	250	42.12	43.30	41.96	29.25	29.94	29.11	11.98	12.17	12.05	12.65	12.59	12.74	12.37	12.32	12.53
100	250	42.00	43.27	41.57	27.01	27.69	27.11	11.61	11.72	11.73	12.71	12.64	12.77	12.44	12.29	12.44
150	250	44.34	45.47	44.24	27.95	28.64	27.94	11.58	11.62	11.50	12.76	12.92	12.79	12.39	12.51	12.41

Table 10: The performance (average RMSE and average size) of the CCE, PC and robust two-step LV, ILV estimators in the setup of independent random loadings and cross-sectional homoskedasticity

		CCE			**PC**			**Robust LV**			**Robust ILV**		
N	T	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}
Average RMSE (×100)													
50	50	30.43	31.88	30.47	31.46	33.02	31.53	32.19	33.96	32.43	32.25	34.03	32.54
100	50	32.51	34.36	32.55	33.57	35.31	33.41	35.11	37.06	35.02	35.15	37.08	35.06
150	50	35.85	37.89	35.90	36.59	38.70	36.57	38.75	41.22	38.81	38.89	41.32	38.94
50	150	19.70	20.90	19.69	20.42	21.62	20.43	18.79	19.86	18.88	18.69	19.75	18.80
100	150	20.22	21.33	20.19	21.50	22.79	21.54	20.16	21.25	20.18	20.01	21.06	20.03
150	150	22.35	23.65	22.34	24.58	26.14	24.63	22.46	23.73	22.41	22.27	23.49	22.20
50	250	16.63	17.70	16.62	16.86	17.84	16.82	14.35	15.08	14.35	14.29	15.00	14.30
100	250	16.37	17.37	16.34	17.54	18.64	17.48	15.48	16.33	15.48	15.39	16.23	15.39
150	250	18.10	19.29	18.15	20.57	22.04	20.62	17.16	18.21	17.22	17.05	18.07	17.10
Average Size (%)													
50	50	26.75	27.04	26.93	30.81	30.95	30.74	26.68	26.95	26.72	26.73	27.01	26.79
100	50	25.13	25.61	25.39	29.38	29.91	29.38	26.22	26.45	26.12	26.14	26.37	26.04
150	50	25.47	25.89	25.67	29.98	30.59	29.89	26.11	26.53	26.17	26.06	26.51	26.15
50	150	20.93	21.61	20.78	23.47	24.13	23.59	17.26	17.57	17.49	17.00	17.32	17.28
100	150	18.09	18.39	17.95	21.89	22.43	21.81	16.46	16.70	16.62	16.16	16.25	16.23
150	150	18.32	18.60	18.22	24.11	24.75	24.23	16.60	16.72	16.61	16.18	16.24	16.14
50	250	20.50	21.28	20.40	22.31	23.06	22.14	13.78	14.01	13.81	13.64	13.80	13.71
100	250	16.34	16.87	16.26	20.23	20.94	19.99	13.26	13.41	13.20	13.03	13.15	12.92
150	250	16.53	17.24	16.73	23.19	24.21	23.33	13.21	13.40	13.24	12.90	13.07	12.91

Table 11: The performance (average RMSE and average size) of the CCE, PC and robust two-step LV, ILV estimators in the setup of correlated random loadings and cross-sectional homoskedasticity

		CCE			**PC**			**Robust LV**			**Robust ILV**		
N	T	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}
Average RMSE (×100)													
50	50	31.20	32.50	31.17	33.70	39.45	33.19	31.32	32.99	31.38	31.39	33.00	31.40
100	50	33.16	34.86	33.19	33.05	34.73	33.05	33.98	35.96	33.94	33.97	35.88	33.90
150	50	36.54	38.45	36.50	35.88	37.73	35.97	37.80	40.05	37.82	37.76	39.97	37.78
50	150	22.05	23.06	21.98	21.58	22.76	21.60	18.40	19.40	18.49	18.35	19.34	18.43
100	150	22.10	23.30	22.10	23.10	24.36	23.13	19.91	21.05	19.97	19.81	20.92	19.87
150	150	24.31	25.80	24.29	26.32	27.91	26.38	22.22	23.57	22.24	22.09	23.41	22.11
50	250	19.45	20.29	19.54	18.62	19.48	18.61	14.27	15.00	14.31	14.24	14.95	14.27
100	250	18.92	19.99	18.91	19.81	20.97	19.78	15.44	16.30	15.43	15.36	16.21	15.36
150	250	20.95	22.23	20.96	23.24	24.77	23.22	17.15	18.15	17.14	17.05	18.04	17.05
Average Size (%)													
50	50	30.44	30.62	30.47	32.99	33.36	32.86	25.08	25.26	25.13	25.14	25.29	25.10
100	50	28.18	28.64	28.08	31.55	32.03	31.48	24.35	24.65	24.15	24.30	24.52	24.07
150	50	28.45	28.68	28.35	32.43	33.16	32.51	24.08	24.31	24.13	24.00	24.19	24.03
50	150	28.42	28.89	28.14	28.38	29.13	28.31	16.46	16.55	16.47	16.36	16.44	16.38
100	150	23.52	23.96	23.64	27.17	27.72	27.09	15.79	16.01	15.84	15.57	15.72	15.62
150	150	23.75	24.58	23.93	30.49	31.53	30.72	15.97	16.11	15.97	15.76	15.84	15.75
50	250	30.08	30.32	30.29	28.49	28.80	28.41	13.46	13.44	13.47	13.37	13.31	13.39
100	250	23.63	24.41	23.79	26.82	27.61	26.70	12.97	13.05	12.92	12.76	12.81	12.75
150	250	24.34	25.02	24.25	31.45	32.39	31.28	12.84	12.93	12.93	12.62	12.65	12.67

Table 12: The performance (average RMSE and average size) of the CCE, PC and robust two-step LV, ILV estimators in the setup of independent random loadings and cross-sectional heteroskedasticity

		CCE			**PC**			**Robust LV**			**Robust ILV**		
N	T	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}
Average RMSE (×100)													
50	50	57.47	61.52	59.21	378.78	282.15	364.02	65.06	69.94	66.05	65.19	69.91	66.10
100	50	63.23	66.26	63.85	313.32	278.65	273.47	70.64	77.15	72.60	70.34	76.69	72.11
150	50	69.38	73.04	69.38	335.98	375.29	258.20	79.12	84.34	79.92	78.75	83.98	79.70
50	150	39.31	41.54	38.69	197.91	356.44	197.73	38.78	41.30	38.83	38.61	41.16	38.74
100	150	41.80	44.10	41.58	174.53	141.71	159.60	42.21	44.67	42.11	41.98	44.32	41.79
150	150	45.86	48.46	45.97	204.87	128.39	138.74	47.37	50.14	47.01	46.97	49.73	46.60
50	250	33.90	35.61	33.62	157.34	160.58	160.82	30.80	31.89	30.58	30.69	31.75	30.49
100	250	35.26	37.44	35.30	124.00	145.33	127.03	32.84	34.79	32.70	32.69	34.58	32.54
150	250	38.69	40.95	38.88	103.03	89.45	95.32	36.38	38.87	36.87	36.16	38.56	36.62
Average Size (%)													
50	50	29.56	30.43	29.94	34.56	35.53	34.70	27.24	27.85	27.62	27.37	27.76	27.57
100	50	29.04	29.66	28.97	33.63	34.19	33.51	26.81	27.16	26.74	26.55	26.87	26.48
150	50	29.72	30.26	29.83	34.32	35.04	34.55	26.73	27.23	26.88	26.39	26.84	26.55
50	150	26.73	27.34	26.43	33.22	34.19	33.44	17.18	17.34	17.29	16.86	17.05	17.02
100	150	25.44	26.35	25.55	31.72	32.51	31.68	16.48	16.54	16.50	15.99	16.02	16.00
150	150	26.48	27.38	26.43	33.18	34.00	33.01	16.49	16.63	16.55	15.88	15.87	15.90
50	250	27.96	28.79	27.80	35.66	36.24	35.30	13.89	14.08	13.82	13.59	13.81	13.68
100	250	25.98	27.04	26.10	33.28	34.31	33.33	13.31	13.45	13.33	12.95	13.02	12.95
150	250	27.27	28.46	27.35	35.01	36.20	35.03	13.21	13.44	13.34	12.79	12.91	12.86

Table 13: The performance (average RMSE and average size) of the CCE, PC and robust two-step LV, ILV estimators in the setup of correlated random loadings and cross-sectional heteroskedasticity

		CCE			**PC**			**Robust LV**			**Robust ILV**		
N	T	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}	β_{i1}	β_{i2}	β_{i3}
Average RMSE (×100)													
50	50	44.05	46.07	44.22	220.97	226.64	287.69	44.81	47.6	44.64	44.64	47.47	44.57
100	50	47.35	49.36	47.76	237.26	193.55	193.57	48.25	51.55	48.85	48.01	51.18	48.46
150	50	52.37	54.07	52.01	188.41	202.05	164.86	54.47	57.54	54.24	54.10	57.12	53.91
50	150	34.09	35.03	34.52	153.81	207.97	144.31	26.57	27.33	26.99	26.52	27.25	26.95
100	150	36.40	38.11	36.34	114.92	123.97	99.73	28.49	30.31	28.40	28.35	30.05	28.18
150	150	39.96	42.07	40.05	103.12	116.78	85.33	31.61	34.08	31.72	31.34	33.67	31.40
50	250	31.63	33.17	31.97	140.56	138.81	125.25	20.64	22.00	20.81	20.61	21.94	20.75
100	250	33.52	35.23	33.67	96.04	134.02	92.26	21.99	23.08	22.10	21.85	22.93	22.00
150	250	37.24	39.39	37.40	78.30	94.51	117.31	24.39	26.21	24.49	24.23	25.97	24.32
Average Size (%)													
50	50	38.14	38.46	37.90	35.25	35.74	35.20	26.05	26.39	26.21	25.85	26.28	26.01
100	50	38.24	38.89	38.07	34.54	35.11	34.36	25.29	25.50	25.28	24.95	25.21	24.91
150	50	39.55	40.14	39.46	35.37	36.11	35.47	25.11	25.41	25.06	24.81	25.06	24.79
50	150	40.29	40.96	40.57	33.72	34.38	33.57	16.40	16.59	16.24	16.17	16.39	16.04
100	150	40.41	41.56	40.37	32.25	33.20	32.12	16.03	16.12	16.07	15.65	15.69	15.68
150	150	42.51	43.59	42.41	34.60	35.64	34.57	16.03	16.13	15.87	15.52	15.58	15.36
50	250	44.64	45.88	44.75	35.35	36.71	35.91	13.13	13.43	13.22	12.97	13.22	13.09
100	250	44.37	45.69	44.60	34.08	35.12	34.40	12.89	12.99	12.88	12.64	12.74	12.68
150	250	47.07	48.32	46.88	36.77	37.87	36.65	12.82	12.92	12.82	12.54	12.58	12.49

Table 14: Empirical size of the over-identification tests (5% level)

		LV (%)				ILV (%)			
N	**T**	**L1+E1**	**L2+E1**	**L1+E2**	**L2+E2**	**L1+E1**	**L2+E1**	**L1+E2**	**L2+E2**
50	50	50.1	44.4	51.3	40.3	15.4	14.7	13.0	13.0
100	50	60.2	54.2	55.3	47.2	18.4	15.1	14.7	13.1
150	50	65.3	57.0	65.8	52.2	18.0	15.2	15.0	14.6
50	150	17.8	13.7	17.6	12.9	10.5	10.0	9.5	9.6
100	150	14.9	15.7	17.9	14.3	7.7	8.4	8.4	6.7
150	150	19.2	18.0	19.7	16.3	9.0	7.7	7.7	8.3
50	250	10.9	9.5	14.8	9.4	8.0	8.5	8.1	7.1
100	250	11.6	10.0	13.6	9.3	7.1	9.1	9.1	4.9
150	250	10.8	11.3	14.3	11.8	6.5	7.0	7.5	6.3