The Annals of Statistics
2014, Vol. 42, No. 1, 142–170
DOI: 10.1214/13-AOS1183

THEORY AND METHODS OF PANEL DATA MODELS WITH INTERACTIVE EFFECTS

BY JUSHAN BAI[1] AND KUNPENG LI[2]

Columbia University and Central University of Economics and Finance, Capital University of Economics and Business and Tsinghua University

This paper considers the maximum likelihood estimation of panel data models with interactive effects. Motivated by applications in economics and other social sciences, a notable feature of the model is that the explanatory variables are correlated with the unobserved effects. The usual within-group estimator is inconsistent. Existing methods for consistent estimation are either designed for panel data with short time periods or are less efficient. The maximum likelihood estimator has desirable properties and is easy to implement, as illustrated by the Monte Carlo simulations. This paper develops the inferential theory for the maximum likelihood estimator, including consistency, rate of convergence and the limiting distributions. We further extend the model to include time-invariant regressors and common regressors (cross-section invariant). The regression coefficients for the time-invariant regressors are time-varying, and the coefficients for the common regressors are cross-sectionally varying.

1. Introduction. This paper studies the following panel data models with unobservable interactive effects:

$$y_{it} = \alpha_i + x_{it}\beta + \lambda_i' f_t + e_{it}, \qquad i = 1, \dots, N, t = 1, 2, \dots, T;$$

where y_{it} is the dependent variable; $x_{it} = (x_{it1}, \dots, x_{itK})$ is a row vector of explanatory variables; α_i is an intercept; the term $\lambda_i' f_t + e_{it}$ is unobservable and has a factor structure, λ_i is an $r \times 1$ vector of factor loadings, f_t is a vector of factors and e_{it} is the idiosyncratic error. The interactive effects ($\lambda_i' f_t$) generalize the usual additive individual and time effects; for example, if $\lambda_i \equiv 1$, then $\alpha_i + \lambda_i' f_t = \alpha_i + f_t$.

A key feature of the model is that the regressors x_{it} are allowed to be correlated with $(\alpha_i, \lambda_i, f_t)$. This situation is commonly encountered in economics and other social sciences, in which some of the regressors x_{it} are decision variables that are influenced by the unobserved individual heterogeneities. The practical relevance of the model will be further discussed below. The objective of this paper is to obtain

Received December 2012; revised October 2013.
[1]Supported by the NSF (SES-0962410).
[2]Supported by NSFC (71201031) and Humanities and Social Sciences of Chinese Ministry of Education (12YJCZH109).
MSC2010 subject classifications. Primary 60F12, 60F30; secondary 60H12.
Key words and phrases. Factor error structure, factors, factor loadings, maximum likelihood, principal components, within-group estimator, simultaneous equations.

consistent and efficient estimation of β in the presence of correlations between the regressors and the factor loadings and factors.

The usual pooled least squares estimator or even the within-group estimator is inconsistent for β. One method to obtain a consistent estimator is to treat $(\alpha_i, \lambda_i, f_t)$ as parameters and estimate them jointly with β. The idea is "controlling through estimating" (controlling the effects by estimating them). This is the approach used in [8, 23] and [30]. While there are some advantages, an undesirable consequence of this approach is the incidental parameters problem. There are too many parameters being estimated, and the incidental parameters bias arises; see [26]. In [1, 2] and [17] the authors consider the generalized method of moments (GMM) method. The GMM method is based on a nonlinear transformation known as quasi-differencing that eliminates the factor errors. Quasi-differencing increases the nonlinearity of the model especially with more than one factor. The GMM method works well with a small T. When T is large, the number of moment equations will be large, and the so called many-moment bias arises. In [27], the author considers an alternative method by augmenting the model with additional regressors $\bar{y}_t$ and $\bar{x}_t$, which are the cross-sectional averages of y_{it} and x_{it}. These averages provide an estimate for f_t. The estimator of [27] becomes inconsistent when the factor loadings in the y equation are correlated with those in the x equation, as shown in [32]. A further approach to controlling the correlation between the regressors and factor errors is to use the Mundlak–Chamberlain projection ([24] and [15]). The latter method projects α_i and λ_i onto the regressors such that $\lambda_i = c_0 + c_1 x_{i1} + \cdots + c_T x_{iT} + \eta_i$, where c_s $(s = 0, 1, \ldots, T)$ are parameters to be estimated, and η_i is the projection residual (a similar projection is done for α_i). The projection residuals are uncorrelated with the regressors so that a variety of approaches can be used to estimate the model. This framework is designed for small T and is studied by [9].

In this paper we consider the pseudo-Gaussian maximum likelihood method under large N and large T. The theory does not depend on normality. In view of the importance of the MLE in the statistical literature, it is of both practical and theoretical interest to examine the MLE in this context. We develop a rigorous theory for the MLE. We show that there is no incidental parameters bias for β.

We allow time-invariant regressors such as education, race and gender in the model. The corresponding regression coefficients are time-dependent. Similarly, we allow common regressors, which do not vary across individuals, such as prices and policy variables. The corresponding regression coefficients are individual-dependent so that individuals respond differently to policy or price changes. In our view, this is a sensible way to incorporate time-invariant and common regressors. For example, wages associated with education and with gender are more likely to change over time rather than remain constant. In our analysis, time invariant regressors are treated as the components of λ_i that are observable, and common regressors as the components of f_t that are observable. This view fits naturally into the factor framework in which part of the factor loadings and factors are observable, and the maximum likelihood method imposes the corresponding loadings and factors at their observed values.

While the theoretical analysis of MLE is demanding, the limiting distributions of the MLE are simple and have intuitive interpretations. The computation is also easy and can be implemented by adapting the ECM (expectation and constrained maximization) of [22]. In addition, the maximum likelihood method allows restrictions to be imposed on λ_i or on f_t to achieve more efficient estimation. These restrictions can take the form of known values, being either zeros, or other fixed values. Part of the rigorous analysis includes setting up the constrained maximization as a Lagrange multiplier problem. This approach provides insight into which kinds of restrictions provide efficiency gain and which kinds do not.

Panel data models with interactive effects have wide applicability in economics. In macroeconomics, for example, y_{it} can be the output growth rate for country i in year t; x_{it} represents production inputs, and f_t is a vector of common shocks (technological progress, financial crises); the common shocks have heterogenous impacts across countries through the different factor loadings λ_i; e_{it} represents the country-specific unmeasured growth rates. In microeconomics, and especially in earnings studies, y_{it} is the wage rate for individual i for period t (or for cohort t), x_{it} is a vector of observable characteristics such as marital status and experience; λ_i is a vector of unobservable individual traits such as ability, perseverance, motivation and dedication; the payoff to these individual traits is not constant over time, but time varying through f_t; and e_{it} is idiosyncratic variations in the wage rates. In finance, y_{it} is stock i's return in period t, x_{it} is a vector of observable factors, f_t is a vector of unobservable common factors (systematic risks) and λ_i is the exposure to the risks; e_{it} is the idiosyncratic returns. Factor error structures are also used as a flexible trend modeling as in [20]. Most of panel data analysis assumes cross-sectional independence; see, for example, [6, 13] and [18]. The factor structure is also capable of capturing the cross-sectional dependence arising from the common shocks f_t. Further motivation can be found in [7, 28, 29].

Throughout the paper, the norm of a vector or matrix is that of Frobenius, that is, $\|A\| = [\text{tr}(A'A)]^{1/2}$ for matrix A; $\text{diag}(A)$ is a column vector consisting of the diagonal elements of A when A is matrix, but $\text{diag}(A)$ represents a diagonal matrix when A is a vector. In addition, we use $\dot{v}_t$ to denote $v_t - \frac{1}{T}\sum_{t=1}^{T} v_t$ 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.

The rest of the paper is organized as follows. Section 2 introduces a common shock model and the maximum likelihood estimation. Consistency, rate of convergence and the limiting distributions of the MLE are established. Section 3 shows that if some factors do not affect the y equation but only the x equation, more efficient estimation can be obtained. Section 4 extends the analysis to time-invariant regressors and common regressors; the corresponding coefficients are time varying and cross-section varying, respectively. Computing algorithm is discussed in Section 5, and simulations results are reported in Section 6. The last section concludes. The theoretical proofs are provided in the supplementary document [11].

2. A common shock model. In the common-shock model, we assume that both y_{it} and x_{it} are impacted by the common shocks f_t so the model takes the form

$$\begin{aligned} y_{it} &= \alpha_i + x_{it1}\beta_1 + x_{it2}\beta_2 + \cdots + x_{itK}\beta_K + \lambda_i' f_t + e_{it}, \\ x_{itk} &= \mu_{ik} + \gamma_{ik}' f_t + v_{itk} \end{aligned} \tag{2.1}$$

for $k = 1, 2, \ldots, K$. In across-country output studies, for example, output y_{it} and inputs x_{it} (labor and capital) are both affected by the common shocks.

The parameter of interest is $\beta = (\beta_1, \ldots, \beta_K)'$. We also estimate $\alpha_i, \lambda_i, \mu_{ik}$ and γ_{ik} $(k = 1, 2, \ldots, K)$. By treating the latter as parameters, we also allow arbitrary correlations between (α_i, λ_i) and (μ_{ik}, γ_{ik}). Although we also treat f_t as fixed parameters, there is no need to estimate the individual f_t, but only the sample covariance of f_t. This is an advantage of the maximum likelihood method, which eliminates the incidental parameters problem in the time dimension. This kind of the maximum likelihood method was used for pure factor models in [3, 4] and [10]. By symmetry, we could also estimate individuals f_t, but then we only estimate the sample covariance of the factor loadings. The idea is that we do not simultaneously estimate the factor loadings and the factors f_t (which would be the case for the principal components method). This reduces the number of parameters considerably. If N is much smaller than T $(N \ll T)$, treating factor loadings as parameters is preferable since there are fewer parameters.

Because of the correlation between the regressors and regression errors in the y equation, the y and x equations form a simultaneous equation system; the MLE jointly estimates the parameters in both equations. The joint estimation avoids the Mundlak–Chamberlain projection and thus is applicable for large N and large T.

We assume the number of factors r is fixed and known. Determining the number of factors is discussed in Section 6, where a modified information criterion proposed by [12] is used. Let $x_{it} = (x_{it1}, x_{it2}, \ldots, x_{itK})$, $\gamma_{ix} = (\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{iK})$, $v_{itx} = (v_{it1}, v_{it2}, \ldots, v_{itK})'$ and $\mu_{ix} = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{iK})'$. The second equation of (2.1) can be written in matrix form as

$$x_{it}' = \mu_{ix} + \gamma_{ix}' f_t + v_{itx}.$$

Further let $\Gamma_i = (\lambda_i, \gamma_{ix})$, $z_{it} = (y_{it}, x_{it})'$, $\varepsilon_{it} = (e_{it}, v_{itx}')'$, $\mu_i = (\alpha_i, \mu_{ix}')'$. Then model (2.1) can be written as

$$\begin{bmatrix} 1 & -\beta' \\ 0 & I_K \end{bmatrix} z_{it} = \mu_i + \Gamma_i' f_t + \varepsilon_{it}.$$

Let B denote the coefficient matrix of z_{it} in the preceding equation. Let $z_t = (z'_{1t}, z'_{2t}, \dots, z'_{Nt})'$, $\Gamma = (\Gamma_1, \Gamma_2, \dots, \Gamma_N)'$, $\varepsilon_t = (\varepsilon'_{1t}, \varepsilon'_{2t}, \dots, \varepsilon'_{Nt})'$ and $\mu = (\mu'_1, \mu'_2, \dots, \mu'_N)'$. Stacking the equations over i, we have

$$(I_N \otimes B) z_t = \mu + \Gamma f_t + \varepsilon_t. \tag{2.2}$$

To analyze this model, we make the following assumptions.

2.1. *Assumptions.*

ASSUMPTION A. The factor process f_t is a sequence of constants. Let $M_{ff} = T^{-1}\sum_{t=1}^{T} \dot{f}_t \dot{f}'_t$, where $\dot{f}_t = f_t - \frac{1}{T}\sum_{t=1}^{T} f_t$. We assume that $\overline{M}_{ff} = \lim_{T\to\infty} M_{ff}$ is a strictly positive definite matrix.

REMARK 2.1. The nonrandomness assumption for f_t is not crucial. In fact, f_t can be a sequence of random variables such that $E(\|f_t\|^4) \le C < \infty$ uniformly in t, and f_t is independent of ε_s for all s. The fixed f_t assumption conforms with the usual fixed effects assumption in panel data literature and, in certain sense, is more general than random f_t.

ASSUMPTION B. The idiosyncratic errors $\varepsilon_{it} = (e_{it}, v'_{itx})'$ are such that:

(B.1) The e_{it} is independent and identically distributed over t and uncorrelated over i with $E(e_{it}) = 0$ and $E(e_{it}^4) \le \infty$ for all $i = 1, \dots, N$ and $t = 1, \dots, T$. Let Σ_{iie} denote the variance of e_{it}.

(B.2) v_{itx} is also independent and identically distributed over t and uncorrelated over i with $E(v_{itx}) = 0$ and $E(\|v_{itx}\|^4) \le \infty$ for all $i = 1, \dots, N$ and $t = 1, \dots, T$. We use Σ_{iix} to denote the variance matrix of v_{itx}.

(B.3) e_{it} is independent of v_{jsx} for all (i, j, t, s). Let Σ_{ii} denote the variance matrix ε_{it}. So we have $\Sigma_{ii} = \operatorname{diag}(\Sigma_{iie}, \Sigma_{iix})$, a block-diagonal matrix.

REMARK 2.2. Let $\Sigma_{\varepsilon\varepsilon}$ denote the variance of $\varepsilon_t = (\varepsilon'_{1t}, \dots, \varepsilon'_{Nt})'$. Due to the uncorrelatedness of ε_{it} over i, we have $\Sigma_{\varepsilon\varepsilon} = \operatorname{diag}(\Sigma_{11}, \Sigma_{22}, \dots, \Sigma_{NN})$, a block-diagonal matrix. Assumption B is more general than the usual assumption in the factor analysis. In a traditional factor model, the variances of the idiosyncratic error terms are assumed to be a diagonal matrix. In the present setting, the variance of ε_t is a block-diagonal matrix. Even without explanatory variables, this generalization is of interest. The factor analysis literature has a long history to explore the block-diagonal idiosyncratic variance, known as multiple battery factor analysis; see [31]. The maximum likelihood estimation theory for high-dimensional factor models with block diagonal covariance matrix has not been previously studied. The asymptotic theory developed in this paper not only provides a way of analyzing the coefficient β, but also a way of analyzing the factors and loadings in the multiple battery factor models. This framework is of independent interest.

ASSUMPTION C. There exists a $C > 0$ sufficiently large such that:

(C.1) $\|\Gamma_j\| \le C$ for all $j = 1, \dots, N$;
(C.2) $C^{-1} \le \tau_{\min}(\Sigma_{jj}) \le \tau_{\max}(\Sigma_{jj}) \le C$ for all $j = 1, \dots, N$, where $\tau_{\min}(\Sigma_{jj})$ and $\tau_{\max}(\Sigma_{jj})$ denote the smallest and largest eigenvalues of the matrix Σ_{jj}, respectively;
(C.3) there exists an $r \times r$ positive matrix Q such that

$$Q = \lim_{N\to\infty} N^{-1}\Gamma'\Sigma_{\varepsilon\varepsilon}^{-1}\Gamma,$$

where Γ is defined earlier.

ASSUMPTION D. The variances Σ_{ii} for all i and M_{ff} are estimated in a compact set, that is, all the eigenvalues of $\widehat{\Sigma}_{ii}$ and $\widehat{M}_{ff}$ are in an interval $[C^{-1}, C]$ for a sufficiently large constant C.

2.2. *Identification restrictions*. It is a well-known result in factor analysis that the factors and loadings can only be identified up to a rotation; see, for example, [5, 21]. The models considered in this paper can be viewed as extensions of the factor models. As such they inherit the same identification problem. We show that identification conditions can be imposed on the factors and loadings without loss of generality. To see this, model (2.2) can be rewritten as

$$(I_N \otimes B)z_t = (\mu + \Gamma\bar{f}) + \big[\Gamma M_{ff}^{1/2} R\big]\big[R' M_{ff}^{-1/2}(f_t - \bar{f})\big] + \varepsilon_t, \tag{2.3}$$

where R is an orthogonal matrix, which we choose to be the matrix consisting of the eigenvectors of $M_{ff}^{1/2}\Gamma'\Sigma_{\varepsilon\varepsilon}^{-1}\Gamma M_{ff}^{1/2}$ associated with the eigenvalues arranged in descending order. Treating $\mu + \Gamma\bar{f}$ as the new $\mu^\star$, $\Gamma M_{ff}^{1/2} R$ as the new $\Gamma^\star$ and $R' M_{ff}^{-1/2}(f_t - \bar{f})$ as the new $f_t^\star$, we have

$$(I_N \otimes B)z_t = \mu^\star + \Gamma^\star f_t^\star + \varepsilon_t$$

with $\frac{1}{T}\sum_{t=1}^T f_t^\star = 0$, $\frac{1}{T}\sum_{t=1}^T f_t^\star f_t^{\star\prime} = I_r$ and $\frac{1}{N}\Gamma^{\star\prime}\Sigma_{\varepsilon\varepsilon}^{-1}\Gamma^\star$ being a diagonal matrix. Thus we impose the following restrictions for model (2.2), which we refer to as IB (*identification* restrictions for *Basic* models).

(IB1) $M_{ff} = I_r$;
(IB2) $\frac{1}{N}\Gamma'\Sigma_{\varepsilon\varepsilon}^{-1}\Gamma = D$, where D is a diagonal matrix with its diagonal elements distinct and arranged in descending order;
(IB3) $\bar{f} = \frac{1}{T}\sum_{t=1}^T f_t = 0$.

2.3. *Estimation*. The objective function considered in this section is

$$\ln L(\theta) = -\frac{1}{2N}\ln|\Sigma_{zz}| - \frac{1}{2N}\operatorname{tr}\big[(I_N \otimes B)M_{zz}(I_N \otimes B')\Sigma_{zz}^{-1}\big], \tag{2.4}$$

where $\Sigma_{zz} = \Gamma M_{ff} \Gamma' + \Sigma_{\varepsilon\varepsilon}$ and $M_{zz} = \frac{1}{T}\sum_{t=1}^{T} \dot{z}_t \dot{z}_t'$. The latter is the data matrix. The parameters are $\theta = (\beta, \Gamma, M_{ff}, \Sigma_{\varepsilon\varepsilon})$. The MLE is defined as

$$\hat{\theta} = \operatorname*{argmax}_{\theta \in \Theta} \ln L(\theta),$$

where the parameter space Θ is defined to be a closed and bounded subset containing the true parameter θ^* as an interior point; $\Sigma_{\varepsilon\varepsilon}$ and M_{ff} are positive definite matrices, as in Assumption D. The boundedness of Θ implies that the elements of β and Γ are bounded. This is for theoretical purpose and is usually assumed for nonconvex optimizations, as in [19] and [25]. In actual computation with the EM algorithm, we do not find the need to impose an upper or lower bound for the parameter values. The likelihood function involves simple functions and are continuous on Θ (in fact differentiable), so the MLE $\hat{\theta}$ exists because a continuous function achieves its extreme value on a closed and bounded subset.

Note that the determinant of $I_N \otimes B$ is 1, so the Jacobian term does not depend on B. If ε_t and f_t are independent and normally distributed, the likelihood function for the observed data has the form of (2.4). Here recall that f_t are fixed constants, and ε_t are not necessarily normal; (2.4) is a pseudo-likelihood function.

For further analysis, we partition the matrix Σ_{zz} and M_{zz} as

$$\Sigma_{zz} = \begin{pmatrix} \Sigma_{zz}^{11} & \Sigma_{zz}^{12} & \cdots & \Sigma_{zz}^{1N} \\ \Sigma_{zz}^{21} & \Sigma_{zz}^{22} & \cdots & \Sigma_{zz}^{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{zz}^{N1} & \Sigma_{zz}^{N2} & \cdots & \Sigma_{zz}^{NN} \end{pmatrix}, \qquad M_{zz} = \begin{pmatrix} M_{zz}^{11} & M_{zz}^{12} & \cdots & M_{zz}^{1N} \\ M_{zz}^{21} & M_{zz}^{22} & \cdots & M_{zz}^{2N} \\ \vdots & \vdots & \ddots & \vdots \\ M_{zz}^{N1} & M_{zz}^{N2} & \cdots & M_{zz}^{NN} \end{pmatrix},$$

where for any (i, j), Σ_{zz}^{ij} and M_{zz}^{ij} are both $(K+1) \times (K+1)$ matrices.

Let $\hat{\beta}, \widehat{\Gamma}$ and $\widehat{\Sigma}_{\varepsilon\varepsilon}$ denote the MLE. The first order condition for β satisfies

$$\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \widehat{\Sigma}_{iie}^{-1} \left\{ (\dot{y}_{it} - \dot{x}_{it}\hat{\beta}) - \hat{\lambda}_i' \widehat{G} \sum_{j=1}^{N} \widehat{\Gamma}_j \widehat{\Sigma}_{jj}^{-1} \begin{bmatrix} \dot{y}_{jt} - \dot{x}_{jt}\hat{\beta} \\ \dot{x}_{jt}' \end{bmatrix} \right\} \dot{x}_{it} = 0, \tag{2.5}$$

where $\widehat{G} = (\widehat{M}_{ff}^{-1} + \widehat{\Gamma}' \widehat{\Sigma}_{\varepsilon\varepsilon}^{-1} \widehat{\Gamma})^{-1}$. The first order condition for Γ_j satisfies

$$\sum_{i=1}^{N} \widehat{\Gamma}_i \widehat{\Sigma}_{ii}^{-1} (\widehat{B} M_{zz}^{ij} \widehat{B}' - \widehat{\Sigma}_{zz}^{ij}) = 0. \tag{2.6}$$

Post-multiplying $\widehat{\Sigma}_{jj}^{-1} \widehat{\Gamma}_j'$ on both sides of (2.6) and then taking summation over j, we have

$$\sum_{i=1}^{N} \sum_{j=1}^{N} \widehat{\Gamma}_i \widehat{\Sigma}_{ii}^{-1} (\widehat{B} M_{zz}^{ij} \widehat{B}' - \widehat{\Sigma}_{zz}^{ij}) \widehat{\Sigma}_{jj}^{-1} \widehat{\Gamma}_j' = 0. \tag{2.7}$$

The first order condition for Σ_{ii} satisfies

$$\widehat{B} M_{zz}^{ii} \widehat{B}' - \widehat{\Sigma}_{zz}^{ii} = \mathbb{W}, \tag{2.8}$$

where $\mathbb{W}$ is a $(K+1)\times(K+1)$ matrix such that its upper-left 1×1 and lower-right $K\times K$ submatrices are both zero, but the remaining elements are undetermined. The undetermined elements correspond to the zero elements of Σ_{ii}. These first order conditions are needed for the asymptotic representation of the MLE.

2.4. *Asymptotic properties of the MLE.* Theorem 2.1 states the convergence rates of the MLE. The consistency is implied by the theorem.

THEOREM 2.1 (Convergence rate). *Let $\hat{\theta} = (\hat{\beta}, \widehat{\Gamma}, \widehat{\Sigma}_{\varepsilon\varepsilon})$ be the solution by maximizing* (2.4). *Under Assumptions* A–D *and the identification conditions* IB, *we have*

$$\hat{\beta} - \beta = O_p(N^{-1/2}T^{-1/2}) + O_p(T^{-1}),$$

$$\frac{1}{N}\sum_{i=1}^{N} \|\widehat{\Sigma}_{ii}^{-1}\| \cdot \|\widehat{\Gamma}_i - \Gamma_i\|^2 = O_p(T^{-1}), \qquad \frac{1}{N}\sum_{i=1}^{N} \|\widehat{\Sigma}_{ii} - \Sigma_{ii}\|^2 = O_p(T^{-1}).$$

REMARK 2.3. Bai [8] considers an iterated principal components estimator for model (2.1). His derivation shows that, in the presence of heteroscedasticities over the cross section, the PC estimator for β has a bias of order $O_p(N^{-1})$. As a comparison, Theorem 2.1 shows that the MLE is robust to the heteroscedasticities over the cross section. So if N is fixed, the estimator in [8] is inconsistent unless there is no heteroskedasticity, but the estimator here is still consistent.

Let $\mathcal{M}(\mathbb{X})$ denote the project matrix onto the space orthogonal to $\mathbb{X}$, that is, $\mathcal{M}(\mathbb{X}) = I - \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$. We have

THEOREM 2.2 (Asymptotic representation). *Under the assumptions of Theorem* 2.1, *we have*

$$\hat{\beta} - \beta = \Omega^{-1}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Sigma_{iie}^{-1} e_{it} v_{itx} + O_p(T^{-3/2}) + O_p(N^{-1}T^{-1/2}) + O_p(N^{-1/2}T^{-1}),$$

where Ω is a $K\times K$ matrix whose (p,q) element $\Omega_{pq} = \frac{1}{N}\sum_{i=1}^{N}\Sigma_{iie}^{-1}\Sigma_{iix}^{(p,q)}$ with $\Sigma_{iix}^{(p,q)}$ being the (p,q) element of matrix Σ_{iix}.

REMARK 2.4. In Appendix A.3 of the supplement [11], we show that the asymptotic expression of $\hat{\beta} - \beta$ can be alternatively expressed as

$$\hat{\beta} - \beta = \begin{pmatrix} \mathrm{tr}[\ddot{M} X_1 \mathcal{M}(\overline{\mathbb{F}}) X_1'] & \cdots & \mathrm{tr}[\ddot{M} X_1 \mathcal{M}(\overline{\mathbb{F}}) X_K'] \\ \vdots & \vdots & \vdots \\ \mathrm{tr}[\ddot{M} X_K \mathcal{M}(\overline{\mathbb{F}}) X_1'] & \cdots & \mathrm{tr}[\ddot{M} X_K \mathcal{M}(\overline{\mathbb{F}}) X_K'] \end{pmatrix}^{-1}$$

$$\times \begin{pmatrix} \mathrm{tr}[\ddot{M} X_1 \mathcal{M}(\overline{\mathbb{F}}) e'] \\ \vdots \\ \mathrm{tr}[\ddot{M} X_K \mathcal{M}(\overline{\mathbb{F}}) e'] \end{pmatrix} \tag{2.9}$$

$$+ O_p(T^{-3/2}) + O_p(N^{-1}T^{-1/2}) + O_p(N^{-1/2}T^{-1}),$$

where $X_k = (x_{itk})$ is $N \times T$ (the data matrix for the kth regressor, $k = 1, 2, \dots, K$); $e = (e_{it})$ is $N \times T$; $\ddot{M} = \Sigma_{ee}^{-1/2} \mathcal{M}(\Sigma_{ee}^{-1/2} \Lambda) \Sigma_{ee}^{-1/2}$ with $\Sigma_{ee} = \mathrm{diag}\{\Sigma_{11e}, \Sigma_{22e}, \dots, \Sigma_{NNe}\}$ and $\Lambda = (\lambda_1, \lambda_2, \dots, \lambda_N)'$; $\mathbb{F} = (f_1, f_2, \dots, f_T)'$; $\overline{\mathbb{F}} = (1_T, \mathbb{F})$ where 1_T is a $T \times 1$ vector with all 1's.

REMARK 2.5. Theorem 2.2 shows that the asymptotic expression of $\hat{\beta} - \beta$ only involves variations in e_{it} and v_{itx}. Intuitively, this is due to the fact that the error terms of the y equation share the same factors with the explanatory variables. The variations from the common factor part of x_{itk} (i.e., $\gamma_{ik}' f_t$) do not provide information for β since this part of information is offset by the common factor part of the error terms (i.e., $\lambda_i' f_t$) in the y equation.

COROLLARY 2.1 (Limiting distribution). *Under the assumptions of Theorem 2.2, if $\sqrt{N}/T \to 0$, we have*

$$\sqrt{NT}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \overline{\Omega}^{-1}),$$

where $\overline{\Omega} = \lim_{N,T\to\infty} \Omega$, and $\overline{\Omega}$ is also the limit of

$$\overline{\Omega} = \operatorname*{plim}_{N,T\to\infty} \frac{1}{NT} \begin{pmatrix} \mathrm{tr}[\ddot{M} X_1 \mathcal{M}(\overline{\mathbb{F}}) X_1'] & \cdots & \mathrm{tr}[\ddot{M} X_1 \mathcal{M}(\overline{\mathbb{F}}) X_K'] \\ \vdots & \vdots & \vdots \\ \mathrm{tr}[\ddot{M} X_K \mathcal{M}(\overline{\mathbb{F}}) X_1'] & \cdots & \mathrm{tr}[\ddot{M} X_K \mathcal{M}(\overline{\mathbb{F}}) X_K'] \end{pmatrix}.$$

REMARK 2.6. Matrix $\overline{\Omega}$ can be consistently estimated by

$$\frac{1}{NT} \begin{pmatrix} \mathrm{tr}[\widehat{\ddot{M}} X_1 \mathcal{M}(\widehat{\overline{\mathbb{F}}}) X_1'] & \cdots & \mathrm{tr}[\widehat{\ddot{M}} X_1 \mathcal{M}(\widehat{\overline{\mathbb{F}}}) X_K'] \\ \vdots & \vdots & \vdots \\ \mathrm{tr}[\widehat{\ddot{M}} X_K \mathcal{M}(\widehat{\overline{\mathbb{F}}}) X_1'] & \cdots & \mathrm{tr}[\widehat{\ddot{M}} X_K \mathcal{M}(\widehat{\overline{\mathbb{F}}}) X_K'] \end{pmatrix},$$

where X_k is the $N \times T$ data matrix for the kth regressor,

$$\widehat{\ddot{M}} = \widehat{\Sigma}_{ee}^{-1} - \widehat{\Sigma}_{ee}^{-1}\widehat{\Lambda}(\widehat{\Lambda}'\widehat{\Sigma}_{ee}^{-1}\widehat{\Lambda})^{-1}\widehat{\Lambda}'\widehat{\Sigma}_{ee}^{-1}; \tag{2.10}$$

$\widehat{\widehat{\mathbb{F}}} = (1_T, \widehat{\mathbb{F}})$ with $\widehat{\mathbb{F}} = (\hat{f}_1, \hat{f}_2, \dots, \hat{f}_T)'$ and

$$\hat{f}_t = \left(\sum_{i=1}^N \widehat{\Gamma}_i \widehat{\Sigma}_{ii}^{-1}\widehat{\Gamma}_i'\right)^{-1}\left(\sum_{i=1}^N \widehat{\Gamma}_i \widehat{\Sigma}_{ii}^{-1}\widehat{B}\dot{z}_{it}\right). \tag{2.11}$$

Here $\widehat{\Gamma}$, $\widehat{\Lambda}$, $\widehat{\Sigma}_{ii}$, $\widehat{\Sigma}_{ee}$ and $\widehat{B}$ are the maximum likelihood estimators.

3. Common shock models with zero restrictions. The basic model in Section 2 assumes that the explanatory variables x_{it} share the same factors with y_{it}. This section relaxes this assumption. We assume that the regressors are impacted by additional factors that do not affect the y equation. An alternative view is that some factor loadings in the y equation are restricted to be zero. Consider the following model:

$$\begin{aligned} y_{it} &= \alpha_i + x_{it1}\beta_1 + x_{it2}\beta_2 + \cdots + x_{itK}\beta_K + \psi_i' g_t + e_{it}, \\ x_{itk} &= \mu_{ik} + \gamma_{ik}^{g\prime} g_t + \gamma_{ik}^{h\prime} h_t + v_{itk} \end{aligned} \tag{3.1}$$

for $k = 1, 2, \dots, K$, where g_t is an $r_1 \times 1$ vector representing the shocks affecting both y_{it} and x_{it}, and h_t is an $r_2 \times 1$ vector representing the shocks affecting x_{it} only. Let $\lambda_i = (\psi_i', 0_{r_2\times 1}')'$, $\gamma_{ik} = (\gamma_{ik}^{g\prime}, \gamma_{ik}^{h\prime})'$ and $f_t = (g_t', h_t')'$, the above model can be written as

$$\begin{aligned} y_{it} &= \alpha_i + x_{it1}\beta_1 + x_{it2}\beta_2 + \cdots + x_{itK}\beta_K + \lambda_i' f_t + e_{it}, \\ x_{itk} &= \mu_{ik} + \gamma_{ik}' f_t + v_{itk}, \end{aligned}$$

which is the same as model (2.1) except that r_2 elements of λ_i are restricted to be zeros. For further analysis, we introduce some notation. We define

$$\begin{aligned} \Gamma_i^g &= (\psi_i, \gamma_{i1}^g, \dots, \gamma_{iK}^g), & \Gamma_i^h &= (0_{r_2\times 1}, \gamma_{i1}^h, \dots, \gamma_{iK}^h), \\ \Gamma^g &= (\Gamma_1^g, \Gamma_2^g, \dots, \Gamma_N^g)', & \Gamma^h &= (\Gamma_1^h, \Gamma_2^h, \dots, \Gamma_N^h)'. \end{aligned}$$

We also define $\mathbb{G}$ and $\mathbb{H}$ similarly as $\mathbb{F}$, that is, $\mathbb{G} = (g_1, g_2, \dots, g_T)'$, $\mathbb{H} = (h_1, h_2, \dots, h_T)'$. This implies that $\mathbb{F} = (\mathbb{G}, \mathbb{H})$. The presence of zero restrictions in (3.1) requires different identification conditions.

3.1. *Identification conditions.* Zero loading restrictions alleviate rotational indeterminacy. Instead of $r^2 = (r_1 + r_2)^2$ restrictions, we only need to impose $r_1^2 + r_1 r_2 + r_2^2$ restrictions. These restrictions are referred to as IZ restrictions (*Identification* conditions with *Zero* restrictions). They are:

(IZ1) $M_{ff} = I_r$;

(IZ2) $\frac{1}{N}\Gamma^{g\prime}\Sigma_{\varepsilon\varepsilon}^{-1}\Gamma^g = D_1$ and $\frac{1}{N}\Gamma^{h\prime}\Sigma_{\varepsilon\varepsilon}^{-1}\Gamma^h = D_2$, where D_1 and D_2 are both diagonal matrices with distinct diagonal elements in descending order;

(IZ3) $1_T'\mathbb{G} = 0$ and $1_T'\mathbb{H} = 0$.

In addition, we need an additional assumption for our analysis.

ASSUMPTION E. $\Psi = (\psi_1', \psi_2', \dots, \psi_N')'$ is of full column rank.

Identification conditions IZ are less stringent than IB of the previous section. Assumption E says that the factors g_t are pervasive for the y equation. In Appendix B of the supplement [11], we explain why $r_1^2 + r_1 r_2 + r_2^2$ restrictions are sufficient.

3.2. *Estimation*. The likelihood function is now maximized under three sets of restrictions, that is, $\frac{1}{N}\Gamma^{g\prime}\Sigma_{\varepsilon\varepsilon}^{-1}\Gamma^g = D_1$, $\frac{1}{N}\Gamma^{h\prime}\Sigma_{\varepsilon\varepsilon}^{-1}\Gamma^h = D_2$ and $\Phi = 0$ where Φ denotes the zero factor loading matrix in the y equation. The likelihood function with the Lagrange multipliers is

$$\begin{aligned}\ln L = &-\frac{1}{2N}\ln|\Sigma_{zz}| - \frac{1}{2N}\operatorname{tr}\big[(I_N \otimes B)M_{zz}(I_N \otimes B')\Sigma_{zz}^{-1}\big] \\ &+ \operatorname{tr}\Big[\Upsilon_1\Big(\frac{1}{N}\Gamma^{g\prime}\Sigma_{\varepsilon\varepsilon}^{-1}\Gamma^g - D_1\Big)\Big] + \operatorname{tr}\Big[\Upsilon_2\Big(\frac{1}{N}\Gamma^{h\prime}\Sigma_{\varepsilon\varepsilon}^{-1}\Gamma^h - D_2\Big)\Big] \\ &+ \operatorname{tr}\big[\Upsilon_3'\Phi\big],\end{aligned}$$

where $\Sigma_{zz} = \Gamma\Gamma' + \Sigma_{\varepsilon\varepsilon}$; Υ_1 is $r_1 \times r_1$ and Υ_2 is $r_2 \times r_2$, both are symmetric Lagrange multipliers matrices with zero diagonal elements; Υ_3 is a Lagrange multiplier matrix of dimension $r_2 \times N$.

Let $\mathbb{U} = \widehat{\Sigma}_{zz}^{-1}[(I_N \otimes \widehat{B})M_{zz}(I_N \otimes \widehat{B}') - \widehat{\Sigma}_{zz}]\widehat{\Sigma}_{zz}^{-1}$. Notice $\mathbb{U}$ is a symmetric matrix. The first order condition on $\widehat{\Gamma}^g$ gives

$$\frac{1}{N}\widehat{\Gamma}^{g\prime}\mathbb{U} + \Upsilon_1\frac{1}{N}\widehat{\Gamma}^{g\prime}\widehat{\Sigma}_{\varepsilon\varepsilon}^{-1} = 0.$$

Post-multiplying $\widehat{\Gamma}^g$ yields

$$\frac{1}{N}\widehat{\Gamma}^{g\prime}\mathbb{U}\widehat{\Gamma}^g + \Upsilon_1\frac{1}{N}\widehat{\Gamma}^{g\prime}\widehat{\Sigma}_{\varepsilon\varepsilon}^{-1}\widehat{\Gamma}^g = 0.$$

Since $\frac{1}{N}\widehat{\Gamma}^{g\prime}\mathbb{U}\widehat{\Gamma}^g$ is a symmetric matrix, the above equation implies that $\Upsilon_1\frac{1}{N}\widehat{\Gamma}^{g\prime}\widehat{\Sigma}_{\varepsilon\varepsilon}^{-1}\widehat{\Gamma}^g$ is also symmetric. But $\frac{1}{N}\widehat{\Gamma}^{g\prime}\widehat{\Sigma}_{\varepsilon\varepsilon}^{-1}\widehat{\Gamma}^g$ is a diagonal matrix. So the (i, j)th element of $\Upsilon_1\frac{1}{N}\widehat{\Gamma}^{g\prime}\widehat{\Sigma}_{\varepsilon\varepsilon}^{-1}\widehat{\Gamma}^g$ is $\Upsilon_{1,ij}d_{1j}$, where $\Upsilon_{1,ij}$ is the (i, j)th element of Υ_1 and d_{1j} is the jth diagonal element of $\widehat{D}_1$. Given $\Upsilon_1\frac{1}{N}\widehat{\Gamma}^{g\prime}\widehat{\Sigma}_{\varepsilon\varepsilon}^{-1}\widehat{\Gamma}^g$ is symmetric, we have $\Upsilon_{1,ij}d_{1j} = \Upsilon_{1,ji}d_{1i}$ for all $i \neq j$. However, Υ_1 is also symmetric, so $\Upsilon_{1,ij} = \Upsilon_{1,ji}$. This gives $\Upsilon_{1,ij}(d_{1j} - d_{1i}) = 0$. Since $d_{1j} \neq d_{1i}$ by IZ2, we have $\Upsilon_{1,ij} = 0$ for all $i \neq j$. This implies $\Upsilon_1 = 0$ since the diagonal elements of Υ_1 are all zeros.

Let $\Gamma_x^h = (\gamma_{1x}^h, \gamma_{2x}^h, \dots, \gamma_{Nx}^h)'$ with $\gamma_{ix}^h = (\gamma_{i1}^h, \gamma_{i2}^h, \dots, \gamma_{iK}^h)$, and $\Sigma_{xx} = \operatorname{diag}\{\Sigma_{11x}, \Sigma_{22x}, \dots, \Sigma_{NNx}\}$, a block diagonal matrix of $NK \times NK$ dimension.

We partition the matrix $\mathbb{U}$ and define the matrix $\overline{\mathbb{U}}$ as

$$\mathbb{U} = \begin{pmatrix} \mathbb{U}_{11} & \mathbb{U}_{12} & \cdots & \mathbb{U}_{1N} \\ \mathbb{U}_{21} & \mathbb{U}_{22} & \cdots & \mathbb{U}_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{U}_{N1} & \mathbb{U}_{N2} & \cdots & \mathbb{U}_{NN} \end{pmatrix}, \qquad \overline{\mathbb{U}} = \begin{pmatrix} \overline{\mathbb{U}}_{11} & \overline{\mathbb{U}}_{12} & \cdots & \overline{\mathbb{U}}_{1N} \\ \overline{\mathbb{U}}_{21} & \overline{\mathbb{U}}_{22} & \cdots & \overline{\mathbb{U}}_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{\mathbb{U}}_{N1} & \overline{\mathbb{U}}_{N2} & \cdots & \overline{\mathbb{U}}_{NN} \end{pmatrix},$$

where $\mathbb{U}_{ij}$ is a $(K+1)\times(K+1)$ matrix, and $\overline{\mathbb{U}}_{ij}$ is the lower-right $K \times K$ block of $\mathbb{U}_{ij}$. Notice $\overline{\mathbb{U}}$ is also a symmetric matrix. Then the first order condition on Γ_x^h gives

$$\frac{1}{N}\widehat{\Gamma}_x^{h\prime}\overline{\mathbb{U}} + \Upsilon_2 \frac{1}{N}\widehat{\Gamma}_x^{h\prime}\widehat{\Sigma}_{xx}^{-1} = 0.$$

Post-multiplying $\widehat{\Gamma}_x^h$ yields

$$\frac{1}{N}\widehat{\Gamma}_x^{h\prime}\overline{\mathbb{U}}\widehat{\Gamma}_x^h + \Upsilon_2 \frac{1}{N}\widehat{\Gamma}_x^{h\prime}\widehat{\Sigma}_{xx}^{-1}\widehat{\Gamma}_x^h = 0.$$

Notice $\frac{1}{N}\widehat{\Gamma}_x^{h\prime}\widehat{\Sigma}_{xx}^{-1}\widehat{\Gamma}_x^h = \frac{1}{N}\widehat{\Gamma}^{h\prime}\widehat{\Sigma}_{\varepsilon\varepsilon}^{-1}\widehat{\Gamma}^h = \widehat{D}_2$. By the similar arguments in deriving $\Upsilon_1 = 0$, we have $\Upsilon_2 = 0$. The interpretation for the zero Lagrange multipliers is that these constraints do not affect the optimal value of the likelihood function nor the efficiency of $\hat{\beta}$. In contrast, we cannot show Υ_3 to be zero. Thus the restriction $\Phi = 0$ affects the optimal value of the likelihood function and the efficiency of $\hat{\beta}$. In Section 2, we did not use the Lagrange multiplier approach to analyze the identification restrictions. Had this been done, we would have obtained zero valued Lagrange multipliers. This is another view of why these restrictions do not affect the limiting distribution of $\hat{\beta}$. But these restrictions are needed to remove the rotational indeterminacy.

Now the likelihood function is simplified as

$$\text{(3.2)} \qquad \ln L = -\frac{1}{2N}\ln|\Sigma_{zz}| - \frac{1}{2N}\operatorname{tr}\big[(I_N \otimes B)M_{zz}(I_N \otimes B')\Sigma_{zz}^{-1}\big] + \operatorname{tr}\big[\Upsilon_3'\Phi\big].$$

The first order condition on Γ is

$$\text{(3.3)} \qquad \widehat{\Gamma}'\widehat{\Sigma}_{zz}^{-1}\big[(I_N \otimes \widehat{B})M_{zz}(I_N \otimes \widehat{B}') - \widehat{\Sigma}_{zz}\big]\widehat{\Sigma}_{zz}^{-1} = W',$$

where W is a matrix having the same dimension as Γ, whose element is zero if the counterpart of Γ is not specified to be zero, otherwise undetermined (containing the Lagrange multipliers). Post-multiplying $\widehat{\Gamma}$ gives

$$\widehat{\Gamma}'\widehat{\Sigma}_{zz}^{-1}\big[(I_N \otimes \widehat{B})M_{zz}(I_N \otimes \widehat{B}') - \widehat{\Sigma}_{zz}\big]\widehat{\Sigma}_{zz}^{-1}\widehat{\Gamma} = W'\widehat{\Gamma}.$$

By the special structure of W and $\widehat{\Gamma}$, it is easy to verify that $W'\widehat{\Gamma}$ has the form

$$\begin{bmatrix} 0_{r_1\times r_1} & 0_{r_1\times r_2} \\ \times & 0_{r_2\times r_2} \end{bmatrix}.$$

However, the left-hand side of the preceding equation is a symmetric matrix, and so is the right-hand side. It follows that the subblock "$\times$" is zero, that is, $W'\widehat{\Gamma} = 0$.

Thus, $\widehat{\Gamma}'\widehat{\Sigma}_{zz}^{-1}[(I_N \otimes \widehat{B})M_{zz}(I_N \otimes \widehat{B}') - \widehat{\Sigma}_{zz}]\widehat{\Sigma}_{zz}^{-1}\widehat{\Gamma} = 0$. (This equation would be the first order condition for M_{ff} if it were unknown.) This equality can be simplified as

$$\widehat{\Gamma}'\widehat{\Sigma}_{\varepsilon\varepsilon}^{-1}[(I_N \otimes \widehat{B})M_{zz}(I_N \otimes \widehat{B}') - \widehat{\Sigma}_{zz}]\widehat{\Sigma}_{\varepsilon\varepsilon}^{-1}\widehat{\Gamma} = 0 \tag{3.4}$$

because $\widehat{\Gamma}'\widehat{\Sigma}_{zz}^{-1} = \widehat{G}\widehat{\Gamma}'\widehat{\Sigma}_{\varepsilon\varepsilon}^{-1}$ with $\widehat{G} = (I + \widehat{\Gamma}'\widehat{\Sigma}_{\varepsilon\varepsilon}^{-1}\widehat{\Gamma})^{-1}$. Next, we partition the matrix $\widehat{G} = (I + \widehat{\Gamma}'\widehat{\Sigma}_{\varepsilon\varepsilon}^{-1}\widehat{\Gamma})^{-1}$ and $\widehat{H} = (\widehat{\Gamma}'\widehat{\Sigma}_{\varepsilon\varepsilon}^{-1}\widehat{\Gamma})^{-1}$ as follows:

$$\widehat{G} = \begin{bmatrix} \widehat{G}_1 \\ \widehat{G}_2 \end{bmatrix} = \begin{bmatrix} \widehat{G}_{11} & \widehat{G}_{12} \\ \widehat{G}_{21} & \widehat{G}_{22} \end{bmatrix}, \qquad \widehat{H} = \begin{bmatrix} \widehat{H}_1 \\ \widehat{H}_2 \end{bmatrix} = \begin{bmatrix} \widehat{H}_{11} & \widehat{H}_{12} \\ \widehat{H}_{21} & \widehat{H}_{22} \end{bmatrix},$$

where $\widehat{G}_{11}, \widehat{H}_{11}$ are $r_1 \times r_1$, while $\widehat{G}_{22}, \widehat{H}_{22}$ are $r_2 \times r_2$.

Notice $\widehat{\Sigma}_{zz}^{-1} = \widehat{\Sigma}_{\varepsilon\varepsilon}^{-1} - \widehat{\Sigma}_{\varepsilon\varepsilon}^{-1}\widehat{\Gamma}\widehat{G}\widehat{\Gamma}'\widehat{\Sigma}_{\varepsilon\varepsilon}^{-1}$ and $\widehat{\Gamma}'\widehat{\Sigma}_{zz}^{-1} = \widehat{G}\widehat{\Gamma}'\widehat{\Sigma}_{\varepsilon\varepsilon}^{-1}$. Substitute these results into (3.3), and use (3.4). The first order condition for ψ_i can be simplified as

$$\widehat{G}_1 \sum_{i=1}^N \widehat{\Gamma}_i \widehat{\Sigma}_{ii}^{-1}(\widehat{B}M_{zz}^{ij}\widehat{B}' - \widehat{\Sigma}_{zz}^{ij})\widehat{\Sigma}_{jj}^{-1} I_{K+1}^1 = 0, \tag{3.5}$$

where I_{K+1}^1 is the first column of the identity matrix of dimension $K + 1$.

Similarly, the first order condition for $\gamma_{jx} = (\gamma_{j1}, \gamma_{j2}, \dots, \gamma_{jK})$ is

$$\sum_{i=1}^N \widehat{\Gamma}_i \widehat{\Sigma}_{ii}^{-1}(\widehat{B}M_{zz}^{ij}\widehat{B}' - \widehat{\Sigma}_{zz}^{ij})\widehat{\Sigma}_{jj}^{-1} I_{K+1}^- = 0, \tag{3.6}$$

where I_{K+1}^- is a $(K + 1) \times K$ matrix, obtained by deleting the first column of the identity matrix of dimension $K + 1$.

The first order condition for Σ_{jj} is

$$\begin{aligned} &\widehat{B}M_{zz}^{jj}\widehat{B}' - \widehat{\Sigma}_{zz}^{jj} - \widehat{\Gamma}_j'\widehat{G}\sum_{i=1}^N \widehat{\Gamma}_i \widehat{\Sigma}_{ii}^{-1}(\widehat{B}M_{zz}^{ij}\widehat{B}' - \widehat{\Sigma}_{zz}^{ij}) \\ &\quad - \sum_{i=1}^N (\widehat{B}M_{zz}^{ji}\widehat{B}' - \widehat{\Sigma}_{zz}^{ji})\widehat{\Sigma}_{ii}^{-1}\widehat{\Gamma}_i'\widehat{G}\widehat{\Gamma}_j = \mathbb{W}, \end{aligned} \tag{3.7}$$

where $\mathbb{W}$ is defined following (2.8).

The first order condition for β is

$$\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \widehat{\Sigma}_{iie}^{-1}\left\{(\dot{y}_{it} - \dot{x}_{it}\hat{\beta}) - \hat{\lambda}_i'\widehat{G}\sum_{j=1}^N \widehat{\Gamma}_j \widehat{\Sigma}_{jj}^{-1}\begin{bmatrix} \dot{y}_{jt} - \dot{x}_{jt}\hat{\beta} \\ \dot{x}_{jt}' \end{bmatrix}\right\}\dot{x}_{it} = 0, \tag{3.8}$$

which is the same as in Section 2.

We need an additional identity to study the properties of the MLE. Recall that, by the special structures of W and $\widehat{\Gamma}$, the three submatrices of $W'\widehat{\Gamma}$ can be directly derived to be zeros. The remaining submatrix is also zero, as shown earlier.

However, this submatrix being zero yields the following equation (the detailed derivation is delivered in Appendix F):

$$\frac{1}{N}\widehat{G}_2\sum_{i=1}^{N}\sum_{j=1}^{N}\widehat{\Gamma}_i\widehat{\Sigma}_{ii}^{-1}(\widehat{B}M_{zz}^{ij}\widehat{B}'-\widehat{\Sigma}_{zz}^{ij})\widehat{\Sigma}_{jj}^{-1}I_{K+1}^{1}\hat{\psi}_j'=0. \tag{3.9}$$

These identities are used to derive the asymptotic representations.

3.3. *Asymptotic properties of the MLE.* The results on consistency and the rate of convergence are similar to those in the previous section, which are presented in Appendixes B.1 and B.2. For simplicity, we only state the asymptotic representation for the MLE here.

PROPOSITION 3.1 (Asymptotic representation). *Under Assumptions A–E and the identification restriction IZ, we have*

$$\begin{aligned}\mathcal{P}^0(\hat{\beta}-\beta)&=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Sigma_{iie}^{-1}e_{it}v_{itx}+\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Sigma_{iie}^{-1}\gamma_{ix}^{h\prime}h_te_{it}\\&\quad-\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Sigma_{iie}^{-1}\psi_i'\Pi_{\psi\psi}^{-1}\left(\frac{1}{N}\sum_{j=1}^{N}\psi_j\Sigma_{jje}^{-1}\gamma_{jx}^{h\prime}\right)h_te_{it}\\&\quad+O_p(T^{-3/2})+O_p(N^{-1}T^{-1/2})+O_p(N^{-1/2}T^{-1}),\end{aligned}$$

where $\mathcal{P}^0$ is a $K\times K$ symmetric matrix with its (p,q) element equal to $\frac{1}{N}\operatorname{tr}(\Gamma_p^{h\prime}\ddot{M}\Gamma_q^h)+\frac{1}{N}\sum_{i=1}^{N}\Sigma_{iie}^{-1}\Sigma_{iix}^{(p,q)}$; $\Gamma_p^h=[\gamma_{1p}^h,\gamma_{2p}^h,\dots,\gamma_{Np}^h]'$; $\gamma_{jx}^h=[\gamma_{j1}^h,\dots,\gamma_{jK}^h]$; $\Pi_{\psi\psi}=\frac{1}{N}\sum_{i=1}^{N}\psi_i\Sigma_{iie}^{-1}\psi_i'$ and $\ddot{M}=\Sigma_{ee}^{-1/2}\mathcal{M}(\Sigma_{ee}^{-1/2}\Psi)\Sigma_{ee}^{-1/2}$.

Proposition 3.1 is derived under the identification conditions IZ. In Appendix B.3 of the supplement [11], we show that for any set of factors and factor loadings $(\psi_i,\gamma_{ik},g_t,h_t)$, it can always be transformed into a new set $(\psi_i^\star,\gamma_{ik}^\star,g_t^\star,h_t^\star)$, which satisfies IZ, and at the same time, leaving $\Phi=0$ intact. Given the asymptotic representation in Proposition 3.1, together with the relationship between the two sets, we have the following theorem, which does not depend on IZ.

THEOREM 3.1. *Under Assumptions A–E, we have*

$$\begin{aligned}\mathcal{P}(\hat{\beta}-\beta)&=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Sigma_{iie}^{-1}e_{it}v_{itx}+\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Sigma_{iie}^{-1}\gamma_{ix}^{h\prime}h_t^\star e_{it}\\&\quad-\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Sigma_{iie}^{-1}\psi_i'\Pi_{\psi\psi}^{-1}\left(\frac{1}{N}\sum_{j=1}^{N}\psi_j\Sigma_{jje}^{-1}\gamma_{jx}^{h\prime}\right)h_t^\star e_{it}\\&\quad+O_p(T^{-3/2})+O_p(N^{-1}T^{-1/2})+O_p(N^{-1/2}T^{-1}),\end{aligned}$$

where

$$h_t^{\star} = \dot{h}_t - \dot{\mathbb{H}}'\dot{\mathbb{G}}(\dot{\mathbb{G}}'\dot{\mathbb{G}})^{-1}\dot{g}_t;$$

$\mathcal{P}$ is a $K \times K$ symmetric matrix with its (p,q) element equal to

$$\frac{1}{NT}\operatorname{tr}\big[\ddot{M}\Gamma_q^h \mathbb{H}'\mathcal{M}(\overline{\mathbb{G}})\mathbb{H}\Gamma_p^{h\prime}\big] + \frac{1}{N}\sum_{i=1}^{N}\Sigma_{iie}^{-1}\Sigma_{iix}^{(p,q)},$$

where $\overline{\mathbb{G}} = (1_T, \mathbb{G})$; $\Pi_{\psi\psi} = \frac{1}{N}\sum_{i=1}^{N}\psi_i\Sigma_{iie}^{-1}\psi_i'$; $\ddot{M} = \Sigma_{ee}^{-1/2}\mathcal{M}(\Sigma_{ee}^{-1/2}\Psi)\Sigma_{ee}^{-1/2}$, $\Gamma_p^h = (\gamma_{1p}^h, \gamma_{2p}^h, \dots, \gamma_{Np}^h)'$.

REMARK 3.1. In Appendix B.3, we show that the asymptotic expression of $\hat{\beta} - \beta$ in Theorem 3.1 can be expressed alternatively as

$$\begin{aligned}\hat{\beta} - \beta = {} & \begin{pmatrix} \operatorname{tr}[\ddot{M}X_1\mathcal{M}(\overline{\mathbb{G}})X_1'] & \cdots & \operatorname{tr}[\ddot{M}X_1\mathcal{M}(\overline{\mathbb{G}})X_K'] \\ \vdots & \vdots & \vdots \\ \operatorname{tr}[\ddot{M}X_K\mathcal{M}(\overline{\mathbb{G}})X_1'] & \cdots & \operatorname{tr}[\ddot{M}X_K\mathcal{M}(\overline{\mathbb{G}})X_K'] \end{pmatrix}^{-1} \\ & \times \begin{pmatrix} \operatorname{tr}[\ddot{M}X_1\mathcal{M}(\overline{\mathbb{G}})e'] \\ \vdots \\ \operatorname{tr}[\ddot{M}X_K\mathcal{M}(\overline{\mathbb{G}})e'] \end{pmatrix} \\ & + O_p(T^{-3/2}) + O_p(N^{-1}T^{-1/2}) + O_p(N^{-1/2}T^{-1}),\end{aligned}$$

where X_k and e are defined below (2.9) and $\overline{\mathbb{G}} = (1_T, \mathbb{G})$. Notice $\ddot{M}$ is defined as $\Sigma_{ee}^{-1/2}\mathcal{M}(\Sigma_{ee}^{-1/2}\Psi)\Sigma_{ee}^{-1/2}$, which is equal to $\Sigma_{ee}^{-1/2}\mathcal{M}(\Sigma_{ee}^{-1/2}\Lambda)\Sigma_{ee}^{-1/2}$ since $\Lambda = (\Psi, 0_{N\times r_2})$ in the present context. In Appendix B.3 of the supplement [11], we also provide an intuitive explanation for this alternative expression.

Given Theorem 3.1 and Remark 3.1 we have the following corollary.

COROLLARY 3.1 (Limiting distribution). *Under Assumptions A–E, if $\sqrt{N}/T \to 0$, we have*

$$\sqrt{NT}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \overline{\mathcal{P}}^{-1}),$$

where $\overline{\mathcal{P}} = \lim_{N,T\to\infty}\mathcal{P}$, and $\overline{\mathcal{P}}$ is also the probability limit of

$$\overline{\mathcal{P}} = \operatorname*{plim}_{N,T\to\infty}\frac{1}{NT}\begin{pmatrix} \operatorname{tr}[\ddot{M}X_1\mathcal{M}(\overline{\mathbb{G}})X_1'] & \cdots & \operatorname{tr}[\ddot{M}X_1\mathcal{M}(\overline{\mathbb{G}})X_K'] \\ \vdots & \vdots & \vdots \\ \operatorname{tr}[\ddot{M}X_K\mathcal{M}(\overline{\mathbb{G}})X_1'] & \cdots & \operatorname{tr}[\ddot{M}X_K\mathcal{M}(\overline{\mathbb{G}})X_K'] \end{pmatrix}.$$

REMARK 3.2. Compared with the model in Section 2, $\hat{\beta}$ is more efficient under the zero loading restrictions. The reason is intuitive. In the previous model, only variations in v_{itx} provide information for β. But in the present case, variations in $\gamma_{ik}^{h\prime} h_t$ of x_{it} also provide information for β. This can also be seen by comparing the limiting variances of Corollaries 2.1 and 3.1. Notice the projection matrix now only involves $\overline{\mathbb{G}}$ instead of $\overline{\mathbb{F}}$; and $\overline{\mathbb{G}}$ is a submatrix of $\overline{\mathbb{F}}$. In addition, the covariance matrix $\overline{\mathcal{P}}$ can be estimated by the same method as in estimating $\overline{\Omega}$; see Remark 2.6.

4. Models with time-invariant regressors and common regressors. In this section, we extend the basic model in Section 2 to include time-invariant regressors and common regressors. Examples of time-invariant regressors include gender, race and education; and examples for common regressors include price variables, unemployment rate, or macroeconomic policy variables. These types of regressors are important for empirical applications.

We first consider the model with only time-invariant regressors,

$$
\begin{aligned}
y_{it} &= \alpha_i + x_{it1}\beta_1 + x_{it2}\beta_2 + \cdots + x_{itK}\beta_K + \psi_i' g_t + \phi_i' h_t + e_{it}, \\
x_{itk} &= \mu_{ik} + \gamma_{ik}^{g\prime} g_t + \gamma_{ik}^{h\prime} h_t + v_{itk}
\end{aligned} \tag{4.1}
$$

for $k = 1, 2, \ldots, K$, where g_t is an r_1-dimensional vector, and h_t is an r_2-dimensional vector. Let $f_t = (g_t', h_t')'$, an r-dimensional vector. The key point of model (4.1) is that the ϕ_i's are known (but not zeros). We treat ϕ_i as new added time-invariant regressors, whose coefficient h_t is allowed to be time-varying. The parameter of interest is still β. The inference for h_t is provided in Appendix C.4 of the supplement [11]. The model in the previous section can be viewed as $\Phi = 0$, where $\Phi = (\phi_1, \phi_2, \ldots, \phi_N)'$. However, the earlier derivation is not applicable here because now Φ is a general matrix with full column rank, which provides more information (restrictions) on the rotation matrix. Thus the number of restrictions required to eliminate rotational indeterminacy is even fewer than in Section 3. This point can be seen in the next subsection.

We define the following notation for further analysis:

$$
\begin{aligned}
&\Gamma_i^g = (\psi_i, \gamma_{i1}^g, \ldots, \gamma_{iK}^g), \qquad \Gamma_i^h = (\phi_i, \gamma_{i1}^h, \ldots, \gamma_{iK}^h), \qquad \Gamma_i = (\Gamma_i^{g\prime}, \Gamma_i^{h\prime})', \\
&\Phi = (\phi_1, \phi_2, \ldots, \phi_N)', \qquad \Psi = (\psi_1, \psi_2, \ldots, \psi_N)', \qquad \lambda_i = (\psi_i', \phi_i')', \\
&\Lambda = (\lambda_1, \lambda_2, \ldots, \lambda_N)'.
\end{aligned}
$$

Then equation (4.1) has the same matrix expression as (2.2). Note that $\Lambda = [\Psi, \Phi]$ is the factor loading matrix for the $N \times 1$ vector $(y_{1t}, y_{2t}, \ldots, y_{Nt})'$.

4.1. *Identification conditions.* We make the following identification conditions, which we refer to as IO (*Identification* conditions with partial *Observable* fixed effects), to emphasize the observed fixed effects:

(IO1) We partition the matrix M_{ff} as

$$M_{ff} = \begin{bmatrix} M_{gg} & M_{gh} \\ M_{hg} & M_{hh} \end{bmatrix}$$

and impose $M_{gh} = 0$ and $M_{gg} = I_{r_1}$;

(IO2) $\frac{1}{N}\Gamma^{g\prime}\Sigma_{\varepsilon\varepsilon}^{-1}\Gamma^g = D$, where D is a diagonal matrix with its diagonal elements distinct and arranged in descending order;

(IO3) $1_T'\mathbb{G} = 0$ and $1_T'\mathbb{H} = 0$.

In Appendix C, we show that IO is sufficient for identification. These restrictions can be imposed without loss of generality, as argued formally in Appendix C.3. In addition, we make the following assumption.

ASSUMPTION F. The loading matrix $\Lambda = [\Psi, \Phi]$ is of full column rank.

4.2. *Estimation*. For clarity, in this subsection, we use Φ^* to denote the observed value for Φ. Recall that $\Sigma_{zz} = \Gamma M_{ff}\Gamma' + \Sigma_{\varepsilon\varepsilon}$, where Γ contains the factor loading coefficients (including Φ); M_{ff} contains the sub-blocks M_{gg}, M_{gh} and M_{hh}; $\Sigma_{\varepsilon\varepsilon}$ contains the heteroskedasticity coefficients. The regression coefficient β is contained in matrix B. The maximization of the likelihood function is now subject to four sets of restrictions, $M_{gh} = 0$, $M_{gg} = I_{r_1}$, $\Phi = \Phi^*$ and $\frac{1}{N}\Gamma^{g\prime}\Sigma_{\varepsilon\varepsilon}^{-1}\Gamma^g = D$. The likelihood function augmented with the Lagrange multipliers is

$$\begin{aligned} \ln L = &-\frac{1}{2N}\ln|\Sigma_{zz}| - \frac{1}{2N}\operatorname{tr}\big[(I_N \otimes B)M_{zz}(I_N \otimes B')\Sigma_{zz}^{-1}\big] + \operatorname{tr}[\Upsilon_1 M_{gh}] \\ &+ \operatorname{tr}\big[\Upsilon_2(M_{gg} - I_{r_1})\big] + \operatorname{tr}\bigg[\Upsilon_3\bigg(\frac{1}{N}\Gamma^{g\prime}\Sigma_{\varepsilon\varepsilon}^{-1}\Gamma^g - D\bigg)\bigg] + \operatorname{tr}\big[\Upsilon_4(\Phi - \Phi^*)\big], \end{aligned}$$

where $\Upsilon_1, \Upsilon_2, \Upsilon_3$ and Υ_4 are all Lagrange multipliers matrices; Υ_1 is an $r_2 \times r_1$ matrix; Υ_2 is an $r_1 \times r_1$ symmetric matrix; Υ_3 is an $r_1 \times r_1$ symmetric matrix with all diagonal elements zeros; Υ_4 is an $r_2 \times N$ matrix; and $\Sigma_{zz} = \Gamma M_{ff}\Gamma' + \Sigma_{\varepsilon\varepsilon}$. Using the same arguments in deriving $\Upsilon_1 = 0$ in Section 3, we have $\Upsilon_3 = 0$. Then the likelihood function is simplified as

$$\begin{aligned} \ln L = &-\frac{1}{2N}\ln|\Sigma_{zz}| - \frac{1}{2N}\operatorname{tr}\big[(I_N \otimes B)M_{zz}(I_N \otimes B')\Sigma_{zz}^{-1}\big] \\ &+ \operatorname{tr}[\Upsilon_1 M_{gh}] + \operatorname{tr}\big[\Upsilon_2(M_{gg} - I_{r_1})\big] + \operatorname{tr}\big[\Upsilon_4(\Phi - \Phi^*)\big]. \end{aligned} \tag{4.2}$$

The first order condition for Γ gives

$$\widehat{M}_{ff}\widehat{\Gamma}'\widehat{\Sigma}_{zz}^{-1}\big[(I_N \otimes \widehat{B})M_{zz}(I_N \otimes \widehat{B}') - \widehat{\Sigma}_{zz}\big]\widehat{\Sigma}_{zz}^{-1} = W',$$

where W is defined in (3.3). Pre-multiplying $\widehat{M}_{ff}^{-1}$ and post-multiplying $\widehat{\Gamma}$, and by the special structures of W and $\widehat{\Gamma}$, we have

$$\frac{1}{N}\widehat{\Gamma}'\widehat{\Sigma}_{zz}^{-1}[(I_N\otimes\widehat{B})M_{zz}(I_N\otimes\widehat{B}')-\widehat{\Sigma}_{zz}]\widehat{\Sigma}_{zz}^{-1}\widehat{\Gamma}$$
$$=-\begin{bmatrix} 0_{r_1\times r_1} & 0_{r_1\times r_2} \\ \frac{1}{N}\widehat{M}_{hh}^{-1}\Upsilon_4'\widehat{\Psi} & \frac{1}{N}\widehat{M}_{hh}^{-1}\Upsilon_4'\Phi \end{bmatrix}.$$

But the first order condition for M_{ff} gives

$$\frac{1}{N}\widehat{\Gamma}'\widehat{\Sigma}_{zz}^{-1}[(I_N\otimes\widehat{B})M_{zz}(I_N\otimes\widehat{B}')-\widehat{\Sigma}_{zz}]\widehat{\Sigma}_{zz}^{-1}\widehat{\Gamma}=\begin{bmatrix}\Upsilon_2 & \Upsilon_1' \\ \Upsilon_1 & 0_{r_2\times r_2}\end{bmatrix}. \tag{4.3}$$

Comparing the proceeding two results and noting that the left-hand side is a symmetric matrix, we have $\widehat{\Gamma}'\widehat{\Sigma}_{zz}^{-1}[(I_N\otimes\widehat{B})M_{zz}(I_N\otimes\widehat{B}')-\widehat{\Sigma}_{zz}]\widehat{\Sigma}_{zz}^{-1}\widehat{\Gamma}=0$. But $\widehat{\Gamma}'\widehat{\Sigma}_{zz}^{-1}$ can be replaced by $\widehat{\Gamma}'\widehat{\Sigma}_{\varepsilon\varepsilon}^{-1}$; see (S.2) in the Appendix. Thus

$$\widehat{\Gamma}'\widehat{\Sigma}_{\varepsilon\varepsilon}^{-1}[(I_N\otimes\widehat{B})M_{zz}(I_N\otimes\widehat{B}')-\widehat{\Sigma}_{zz}]\widehat{\Sigma}_{\varepsilon\varepsilon}^{-1}\widehat{\Gamma}=0. \tag{4.4}$$

The above result implies that $\Upsilon_1=0$, $\Upsilon_2=0$, $\Upsilon_4'\widehat{\Psi}=0$ and $\Upsilon_4'\Phi=0$.

The first order condition for Σ_{ii} is the same as (3.7), that is,

$$\begin{aligned}&\widehat{B}M_{zz}^{jj}\widehat{B}'-\widehat{\Sigma}_{zz}^{jj}-\widehat{\Gamma}_j'\widehat{G}\sum_{i=1}^N\widehat{\Gamma}_i\widehat{\Sigma}_{ii}^{-1}(\widehat{B}M_{zz}^{ij}\widehat{B}'-\widehat{\Sigma}_{zz}^{ij})\\ &\quad-\sum_{i=1}^N(\widehat{B}M_{zz}^{ji}\widehat{B}'-\widehat{\Sigma}_{zz}^{ji})\widehat{\Sigma}_{ii}^{-1}\widehat{\Gamma}_i'\widehat{G}\widehat{\Gamma}_j=\mathbb{W},\end{aligned} \tag{4.5}$$

where $\mathbb{W}$ is defined following (2.8).

The first order condition on β is the same as (3.8), that is,

$$\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T\widehat{\Sigma}_{iie}^{-1}\left\{(\dot{y}_{it}-\dot{x}_{it}\hat{\beta})-\hat{\lambda}_i'\widehat{G}\sum_{j=1}^N\widehat{\Gamma}_j\widehat{\Sigma}_{jj}^{-1}\begin{bmatrix}\dot{y}_{jt}-\dot{x}_{jt}\hat{\beta}\\ \dot{x}_{jt}'\end{bmatrix}\right\}\dot{x}_{it}=0. \tag{4.6}$$

We need an additional identify for the theoretical analysis in the Appendix. The preceding analysis shows that $\frac{1}{N}\Upsilon_4'\widehat{\Psi}=0$ and $\frac{1}{N}\Upsilon_4'\Phi=0$. They imply

$$\frac{1}{N}\sum_{i=1}^N\sum_{j=1}^N\widehat{G}_2\widehat{\Gamma}_i\widehat{\Sigma}_{ii}^{-1}(\widehat{B}M_{zz}^{ij}\widehat{B}'-\widehat{\Sigma}_{zz}^{ij})\widehat{\Sigma}_{jj}^{-1}I_{K+1}^1\hat{\lambda}_j'=0, \tag{4.7}$$

where $\hat{\lambda}_j=(\hat{\psi}_i',\phi_i')'$.

4.3. *Asymptotic properties*. The asymptotic representation for $\hat{\beta}-\beta$ is:

PROPOSITION 4.1. *Under Assumptions A–D and F, and under the identification condition* IO, *we have*

$$\mathcal{Q}^0(\hat{\beta}-\beta)=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Sigma_{iie}^{-1}e_{it}v_{itx}+\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Sigma_{iie}^{-1}\gamma_{ix}^{h\prime}h_te_{it}$$
$$-\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Sigma_{iie}^{-1}\lambda_i'\Pi_{\lambda\lambda}^{-1}\left(\frac{1}{N}\sum_{j=1}^{N}\lambda_j'\Sigma_{jje}^{-1}\gamma_{jx}^{h\prime}\right)h_te_{it}$$
$$+O_p(T^{-3/2})+O_p(N^{-1}T^{-1/2})+O_p(N^{-1/2}T^{-1}),$$

where $\mathcal{Q}^0$ *is a* $K\times K$ *symmetric matrix with its* (p,q) *element equal to* $\frac{1}{N}\operatorname{tr}[M_{hh}\Gamma_p^{h\prime}\ddot{M}\Gamma_q^h]+\frac{1}{N}\sum_{i=1}^N\Sigma_{iie}^{-1}\Sigma_{iix}^{(p,q)}$; $\ddot{M}=\Sigma_{ee}^{-1/2}\mathcal{M}(\Sigma_{ee}^{-1/2}\Lambda)\Sigma_{ee}^{-1/2}$; $\Gamma_p^h=[\gamma_{1p}^h,\gamma_{2p}^h,\dots,\gamma_{Np}^h]'$; $\Pi_{\lambda\lambda}=\frac{1}{N}\sum_{i=1}^N\lambda_i\Sigma_{iie}^{-1}\lambda_i'$; *and* $\gamma_{jx}^h=[\gamma_{j1}^h,\gamma_{j2}^h,\dots,\gamma_{jK}^h]$.

Proposition 4.1 is derived under the identification conditions IO. In Appendix C.3, we show that for any set of factors and factor loadings $(\psi_i,\gamma_{ik},g_t,h_t)$, we can always transform it to another set $(\psi_i^\star,\gamma_{ik}^\star,g_t^\star,h_t^\star)$ which satisfies IO, and at the same time, still maintains the observability of Φ (i.e., Φ is untransformed). This is in agreement with the Lagrange multiplier analysis, in which $\Upsilon_j=0$ $(j=1,2,3)$, but the multiplier for $\Phi=\Phi^*$ is nonzero. Using the relationship between the two sets, we can generalize Proposition 4.1 into the following theorem, which does not depend on IO.

THEOREM 4.1. *Under Assumptions A–D and F, we have*

$$\mathcal{Q}(\hat{\beta}-\beta)=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Sigma_{iie}^{-1}e_{it}v_{itx}+\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Sigma_{iie}^{-1}\gamma_{ix}^{h\prime}h_t^\star e_{it}$$
$$-\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Sigma_{iie}^{-1}\lambda_i'\Pi_{\lambda\lambda}^{-1}\left(\frac{1}{N}\sum_{j=1}^{N}\lambda_j'\Sigma_{jje}^{-1}\gamma_{jx}^{h\prime}\right)h_t^\star e_{it}$$
$$+O_p(T^{-3/2})+O_p(N^{-1}T^{-1/2})+O_p(N^{-1/2}T^{-1}),$$

where

$$h_t^\star=\dot{h}_t-\dot{\mathbb{H}}'\dot{\mathbb{G}}(\dot{\mathbb{G}}'\dot{\mathbb{G}})^{-1}\dot{g}_t;$$

$\mathcal{Q}$ *is a* $K\times K$ *symmetric matrix with its* (p,q) *element equal to*

$$\frac{1}{NT}\operatorname{tr}[\ddot{M}\Gamma_q^h\mathbb{H}'\mathcal{M}(\overline{\mathbb{G}})\mathbb{H}\Gamma_p^{h\prime}]+\frac{1}{N}\sum_{i=1}^{N}\Sigma_{iie}^{-1}\Sigma_{iix}^{(p,q)}$$

and $\ddot{M}$, Γ_p^h *and* $\Pi_{\lambda\lambda}$ *are defined in Proposition 4.1.*

REMARK 4.1. In Appendix C.3 we show that the asymptotic expression of $\hat{\beta} - \beta$ in Theorem 4.1 can be expressed alternatively as

$$\hat{\beta} - \beta = \begin{pmatrix} \operatorname{tr}[\ddot{M}X_1\mathcal{M}(\overline{\mathbb{G}})X_1'] & \cdots & \operatorname{tr}[\ddot{M}X_1\mathcal{M}(\overline{\mathbb{G}})X_K'] \\ \vdots & \vdots & \vdots \\ \operatorname{tr}[\ddot{M}X_K\mathcal{M}(\overline{\mathbb{G}})X_1'] & \cdots & \operatorname{tr}[\ddot{M}X_K\mathcal{M}(\overline{\mathbb{G}})X_K'] \end{pmatrix}^{-1}$$
$$\times \begin{pmatrix} \operatorname{tr}[\ddot{M}X_1\mathcal{M}(\overline{\mathbb{G}})e'] \\ \vdots \\ \operatorname{tr}[\ddot{M}X_K\mathcal{M}(\overline{\mathbb{G}})e'] \end{pmatrix} + O_p(T^{-3/2})$$
$$+ O_p(N^{-1}T^{-1/2}) + O_p(N^{-1/2}T^{-1}),$$

where X_k and e are defined below (2.9) and $\overline{\mathbb{G}} = (1_T, \mathbb{G})$. We also show in Appendix C.3 that this alternative expression has an intuitive explanation.

From Theorem 4.1, we obtain the following corollary.

COROLLARY 4.1. *Under the conditions of Theorem* 4.1, *if* $\sqrt{N}/T \to 0$, *we have*

$$\sqrt{NT}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \overline{\mathcal{Q}}^{-1}),$$

where $\overline{\mathcal{Q}} = \lim_{N,T\to\infty} \mathcal{Q}$, *which has an alternative expression*

$$\overline{\mathcal{Q}} = \operatorname*{plim}_{N,T\to\infty} \frac{1}{NT} \begin{pmatrix} \operatorname{tr}[\ddot{M}X_1\mathcal{M}(\overline{\mathbb{G}})X_1'] & \cdots & \operatorname{tr}[\ddot{M}X_1\mathcal{M}(\overline{\mathbb{G}})X_K'] \\ \vdots & \vdots & \vdots \\ \operatorname{tr}[\ddot{M}X_K\mathcal{M}(\overline{\mathbb{G}})X_1'] & \cdots & \operatorname{tr}[\ddot{M}X_K\mathcal{M}(\overline{\mathbb{G}})X_K'] \end{pmatrix}.$$

REMARK 4.2. Compared with the model in Section 2, $\hat{\beta}$ is more efficient with observable fixed effects (time-invariant regressors). The reason is provided in Remark 3.2.

4.4. *Models with time-invariant regressors and common regressors*. In this subsection, we consider the joint presence of time-invariant regressors and common regressors. Consider the following model:

$$\begin{aligned} y_{it} &= x_{it1}\beta_1 + x_{it2}\beta_2 + \cdots + x_{itK}\beta_K + \psi_i' g_t + \phi_i' h_t + \kappa_i' d_t + e_{it}, \\ x_{itk} &= \gamma_{ik}^{g\prime} g_t + \gamma_{ik}^{h\prime} h_t + \gamma_{ik}^{d\prime} d_t + v_{itk} \end{aligned} \tag{4.8}$$

for $k = 1, 2, \ldots, K$, where g_t, h_t and d_t are $r_1 \times 1$, $r_2 \times 1$ and $r_3 \times 1$ vectors, respectively. A key feature of model (4.8) is that d_t and ϕ_i are observable for all i and t. We call ϕ_i the time-invariant regressors because they are invariant over time and d_t the common regressors because they are the same for all the cross-sectional units. In this model, the time-invariant regressors have time-varying coefficients, and the common regressors have heterogeneous (individual-dependent) coefficients. If $d_t \equiv 1$, κ_i plays the role of α_i in (4.1). So the model here is more general.

Similar to the previous subsection, we make the following assumption:

ASSUMPTION G. The matrices (Ψ, Φ, K) and $(\mathbb{G}, \mathbb{H}, \mathbb{D})$ are both of full column rank, where $\mathrm{K} = (\kappa_1, \kappa_2, \ldots, \kappa_N)'$ and $\mathbb{D} = (d_1, d_2, \ldots, d_T)'$.

Let $\lambda_i = (\psi_i', \phi_i')'$, $\gamma_{ik} = (\gamma_{ik}^{g\prime}, \gamma_{ik}^{h\prime})'$ and $\delta_i = (\kappa_i, \gamma_{ik}^d)$. The model can be written as

$$\begin{bmatrix} 1 & -\beta' \\ 0 & I_K \end{bmatrix} z_{it} = \Gamma_i' f_t + \delta_i' d_t + \varepsilon_{it},$$

where $z_{it}, \Gamma_i, \varepsilon_{it}$ are defined in Section 2; Let $\Delta = (\delta_1, \delta_2, \ldots, \delta_N)'$. Then

$$(I_N \otimes B) z_t - \Delta d_t = \Gamma f_t + \varepsilon_t, \tag{4.9}$$

where the symbols $\Gamma, z_t, B, \varepsilon_t$ are defiend in Section 2.

The likelihood function can be written as

$$\ln L = -\frac{1}{2N} \ln |\Sigma_{zz}| - \frac{1}{2NT} \sum_{t=1}^{T} \big[(I_N \otimes B) z_t - \Delta d_t\big]' \Sigma_{zz}^{-1} \big[(I_N \otimes B) z_t - \Delta d_t\big].$$

Take Σ_{zz} and β as given. Δ maximizes the above function at

$$\widehat{\Delta} = (I_N \otimes B) \left(\sum_{s=1}^{T} z_s d_s' \right) \left(\sum_{s=1}^{T} d_s d_s' \right)^{-1}.$$

Substituting $\widehat{\Delta}$ into the above likelihood function, we obtain the concentrated likelihood function

$$\ln L = -\frac{1}{2N} \ln |\Sigma_{zz}| - \frac{1}{2NT} \operatorname{tr}\big[(I_N \otimes B) Z \mathcal{M}(\mathbb{D}) Z' (I_N \otimes B') \Sigma_{zz}^{-1}\big],$$

where $Z = (z_1, z_2, \ldots, z_T)$, $\mathbb{D} = (d_1, d_2, \ldots, d_T)'$ and $\mathcal{M}(\mathbb{D}) = I_T - \mathbb{D}(\mathbb{D}'\mathbb{D})^{-1}\mathbb{D}'$, a projection matrix. Consider (4.9), which is equivalent to

$$(I_N \otimes B) Z = \Gamma \mathbb{F}' + \Delta \mathbb{D}' + \varepsilon,$$

where $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_T)$. Post-multiplying $\mathcal{M}(\mathbb{D})$ on both sides, we have

$$(I_N \otimes B) Z \mathcal{M}(\mathbb{D}) = \Gamma \mathbb{F}' \mathcal{M}(\mathbb{D}) + \varepsilon \mathcal{M}(\mathbb{D}).$$

If we treat $Z\mathcal{M}(\mathbb{D})$ as the new observable data, $\mathbb{F}'\mathcal{M}(\mathbb{D})$ as the new unobservable factors, the preceding equation can be viewed as a special case of (4.1). Invoking Theorem 4.1, which does not need IO [the factors $\mathbb{F}'\mathcal{M}(\mathbb{D})$ may not satisfy IO], we have the following theorem:

THEOREM 4.2. *Under Assumptions A–D and G, the asymptotic representation of $\hat{\beta}$ in the presence of time invariant and common regressors is*

$$\mathcal{R}(\hat{\beta}-\beta)=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Sigma_{iie}^{-1}e_{it}v_{itx}+\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Sigma_{iie}^{-1}\gamma_{ix}^{h\prime}h_t^{\star}e_{it}$$
$$-\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Sigma_{iie}^{-1}\lambda_i'\Pi_{\lambda\lambda}^{-1}\frac{1}{N}\sum_{j=1}^{N}\lambda_j'\Sigma_{jje}^{-1}\gamma_{jx}^{h\prime}h_t^{\star}e_{it}$$
$$+O_p(T^{-3/2})+O_p(N^{-1}T^{-1/2})+O_p(N^{-1/2}T^{-1}),$$

where

$$h_t^{\star}=h_t-\mathbb{H}'\mathbb{D}(\mathbb{D}'\mathbb{D})^{-1}d_t-\mathbb{H}'\mathcal{M}(\mathbb{D})\mathbb{G}[\mathbb{G}'\mathcal{M}(\mathbb{D})\mathbb{G}]^{-1}(g_t-\mathbb{G}'\mathbb{D}(\mathbb{D}'\mathbb{D})^{-1}d_t);$$

$\mathcal{R}$ *is a* $K\times K$ *symmetric matrix with its* (p,q) *element equal to*

$$\frac{1}{NT}\operatorname{tr}[\ddot{M}\Gamma_q^h\mathbb{H}'\mathcal{M}(\mathbb{B})\mathbb{H}\Gamma_p^{h\prime}]+\frac{1}{N}\sum_{i=1}^{N}\Sigma_{iie}^{-1}\Sigma_{iix}^{(p,q)},$$

where $b_t=(g_t',d_t')'$ *and* $\mathbb{B}=(b_1,b_2,\dots,b_T)'=(\mathbb{G},\mathbb{D})$, *a matrix of* $T\times(r_1+r_3)$ *dimension*; $\ddot{M}=\Sigma_{ee}^{-1/2}\mathcal{M}(\Sigma_{ee}^{-1/2}\Lambda)\Sigma_{ee}^{-1/2}$; $\Gamma_p^h=(\gamma_{1p}^h,\gamma_{2p}^h,\dots,\gamma_{Np}^h)'$; $\Pi_{\lambda\lambda}=\frac{1}{N}\sum_{i=1}^{N}\lambda_i\Sigma_{iie}^{-1}\lambda_i'$.

REMARK 4.3. The asymptotic expression of $\hat{\beta}-\beta$ can be alternatively expressed as

$$\hat{\beta}-\beta=\begin{pmatrix}\operatorname{tr}[\ddot{M}X_1\mathcal{M}(\mathbb{B})X_1'] & \cdots & \operatorname{tr}[\ddot{M}X_1\mathcal{M}(\mathbb{B})X_K']\\ \vdots & \vdots & \vdots\\ \operatorname{tr}[\ddot{M}X_K\mathcal{M}(\mathbb{B})X_1'] & \cdots & \operatorname{tr}[\ddot{M}X_K\mathcal{M}(\mathbb{B})X_K']\end{pmatrix}^{-1}$$
$$\times\begin{pmatrix}\operatorname{tr}[\ddot{M}X_1\mathcal{M}(\mathbb{B})e']\\ \vdots\\ \operatorname{tr}[\ddot{M}X_K\mathcal{M}(\mathbb{B})e']\end{pmatrix}$$
$$+O_p(T^{-3/2})+O_p(N^{-1}T^{-1/2})+O_p(N^{-1/2}T^{-1}).$$

If $\mathbb{D}=1_T$, the above asymptotic result reduces to the one in Theorem 4.1 since $\mathbb{B}=(1_T,\mathbb{G})=\overline{\mathbb{G}}$.

Given Theorem 4.2 and Remark 4.3, we have the following corollary.

COROLLARY 4.2. *Under Assumptions A–D and G, if* $\sqrt{N}/T\to 0$, *then*

$$\sqrt{NT}(\hat{\beta}-\beta)\xrightarrow{d}N(0,\overline{\mathcal{R}}^{-1}),$$

where $\overline{\mathcal{R}} = \lim_{N,T\to\infty} \mathcal{R}$, *and* $\overline{\mathcal{R}}$ *can also be expressed as*

$$\overline{\mathcal{R}} = \operatorname*{plim}_{N,T\to\infty} \frac{1}{NT} \begin{pmatrix} \operatorname{tr}[\ddot{M} X_1 \mathcal{M}(\mathbb{B}) X_1'] & \cdots & \operatorname{tr}[\ddot{M} X_1 \mathcal{M}(\mathbb{B}) X_K'] \\ \vdots & \vdots & \vdots \\ \operatorname{tr}[\ddot{M} X_K \mathcal{M}(\mathbb{B}) X_1'] & \cdots & \operatorname{tr}[\ddot{M} X_K \mathcal{M}(\mathbb{B}) X_K'] \end{pmatrix}.$$

5. Computing algorithm. To estimate the model by the maximum likelihood method, we adapt the ECM (expectation and conditional maximization) procedures of [22]. More specifically, in the M-step we split the parameter $\theta = (\beta, \Gamma, \Sigma_{\varepsilon\varepsilon}, M_{ff})$ into two blocks, $\theta_1 = (\Gamma, \Sigma_{\varepsilon\varepsilon}, M_{ff})$ and $\theta_2 = \beta$, and update $\theta_1^{(k)}$ to $\theta_1^{(k+1)}$ given $\theta_2^{(k)}$ and then update $\theta_2^{(k)}$ to $\theta_2^{(k+1)}$ given $\theta_1^{(k+1)}$, where $\theta^{(k)}$ is the estimated value at the kth iteration. In this section, we only state the iterating formulas for basic models. The iterating formulas for the models in Sections 3 and 4 can be found in Appendix E of [11]. In Appendix E, we also show that the iterated EM solutions satisfy the first order conditions. So the EM estimators are at least locally optimal.

In the basic model, $M_{ff} = I_r$. So the parameters to be estimated reduce to $\theta = (\beta, \Gamma, \Sigma_{\varepsilon\varepsilon})$. Let $\theta^{(k)} = (\beta^{(k)}, \Gamma^{(k)}, \Sigma_{\varepsilon\varepsilon}^{(k)})$ be the estimated value at the kth iteration. We update $\Gamma^{(k)}$ according to

$$\Gamma^{(k+1)} = \left[\frac{1}{T}\sum_{t=1}^{T} E(z_t f_t' | Z, \theta^{(k)})\right]\left[\frac{1}{T}\sum_{t=1}^{T} E(f_t f_t' | Z, \theta^{(k)})\right]^{-1}, \tag{5.1}$$

where

$$\begin{aligned} &\frac{1}{T}\sum_{t=1}^{T} E(f_t f_t' | Z, \theta^{(k)}) \\ &\quad = I_r - \Gamma^{(k)\prime}(\Sigma_{zz}^{(k)})^{-1}\Gamma^{(k)} \\ &\qquad + \Gamma^{(k)\prime}(\Sigma_{zz}^{(k)})^{-1}(I_N \otimes B^{(k)}) M_{zz} (I_N \otimes B^{(k)\prime})(\Sigma_{zz}^{(k)})^{-1}\Gamma^{(k)}, \end{aligned} \tag{5.2}$$

$$\frac{1}{T}\sum_{t=1}^{T} E(z_t f_t' | Z, \theta^{(k)}) = (I_N \otimes B^{(k)}) M_{zz} (I_N \otimes B^{(k)\prime})(\Sigma_{zz}^{(k)})^{-1}\Gamma^{(k)} \tag{5.3}$$

with $\Sigma_{zz}^{(k)} = \Gamma^{(k)}\Gamma^{(k)\prime} + \Sigma_{\varepsilon\varepsilon}^{(k)}$. We update $\Sigma_{\varepsilon\varepsilon}^{(k)}$ and $\beta^{(k)}$ according to

$$\begin{aligned} \Sigma_{\varepsilon\varepsilon}^{(k+1)} = \operatorname{Dg}\{&(I_{N(K+1)} - \Gamma^{(k+1)}\Gamma^{(k)\prime}(\Sigma_{zz}^{(k)})^{-1}) \\ &\times (I_N \otimes B^{(k)}) M_{zz} (I_N \otimes B^{(k)\prime})\}, \end{aligned} \tag{5.4}$$

$$\begin{aligned} \beta^{(k+1)} = &\left(\sum_{i=1}^{N}\sum_{t=1}^{T} \dot{x}_{it}' (\Sigma_{iie}^{(k+1)})^{-1} \dot{x}_{it}\right)^{-1} \\ &\times \left(\sum_{i=1}^{N}\sum_{t=1}^{T} \dot{x}_{it}' (\Sigma_{iie}^{(k+1)})^{-1} (\dot{y}_{it} - \lambda_i^{(k+1)\prime} f_t^{(k)})\right), \end{aligned} \tag{5.5}$$

where $f_t^{(k)}$ is the transpose of the tth row of

$$\mathbb{F}^{(k)} = E(\mathbb{F}|Z, \theta^{(k)}) = \dot{Z}'(I_N \otimes B^{(k)\prime})(\Sigma_{zz}^{(k)})^{-1}\Gamma^{(k)},$$

where $\dot{Z} = (\dot{z}_1, \dot{z}_2, \ldots, \dot{z}_T)$ with $\dot{z}_t = z_t - \frac{1}{T}\sum_{s=1}^{T} z_s$; $\mathrm{Dg}(\cdot)$ is the operator that sets the entries of its argument to zeros if the counterparts of $E(\varepsilon_t\varepsilon_t')$ are zeros.

Putting together, we obtain $\theta^{(k+1)} = (\Gamma^{(k+1)}, \beta^{(k+1)}, \Sigma_{\varepsilon\varepsilon}^{(k+1)})$. The above iteration continues until $\|\theta^{(k+1)} - \theta^{(k)}\|$ is smaller than a preset error tolerance. The initial values use the iterated PC estimators of [8].

6. Finite sample properties. In this section, we consider the finite sample properties of the MLE. Data are generated according to

$$\begin{aligned} y_{it} &= \alpha_i + x_{it1}\beta_1 + x_{it2}\beta_2 + \psi_i' g_t + \phi_i' h_t + \kappa_i' d_t + e_{it}, \\ x_{itk} &= \mu_{ik} + \gamma_{ik}^{g\prime} g_t + \gamma_{ik}^{h\prime} h_t + \gamma_{ik}^{d\prime} d_t + v_{itk}, \qquad k = 1, 2. \end{aligned} \tag{6.1}$$

The dimensions of g_t, h_t, d_t are each fixed to 1. We set $\beta_1 = 1$ and $\beta_2 = 2$. We consider four types of DGP (data generating process), which correspond to the four models considered in the paper.

DGP1: $\phi_i, \kappa_i, \gamma_{ik}^h$ and γ_{ik}^d are fixed to zeros; $\alpha_i, \mu_{ik}, \psi_i$ and g_t are generated from $N(0, 1)$ and $\gamma_{ik}^g = \psi_i + N(0, 1)$.

DGP2: ϕ_i, κ_i and γ_{ik}^d are fixed to zeros; $\alpha_i, \mu_{ik}, \psi_i, \gamma_{ik}^h, g_t$ and h_t are generated from $N(0, 1)$; $\gamma_{ik}^g = \psi_i + N(0, 1)$.

DGP3: κ_i and γ_{ik}^d are fixed to zeros; $\alpha_i, \mu_{ik}, \psi_i, \phi_i, g_t$ and h_t are generated from $N(0, 1)$; $\gamma_{ik}^g = \psi_i + N(0, 1)$ and $\gamma_{ik}^h = \phi_i + N(0, 1)$. Here ϕ_i is observable.

DGP4: $\alpha_i, \mu_{ik}, \psi_i, \phi_i, \kappa_i, g_t$ and h_t are generated from $N(0, 1)$; $d_t = 1 + N(0, 1)$, $\gamma_{ik}^g = \psi_i + N(0, 1)$, $\gamma_{ik}^h = \phi_i + N(0, 1)$ and $\gamma_{ik}^d = \kappa_i + N(0, 1)$. Here ϕ_i and d_t are observable.

Using the method of writing (2.2), we can rewrite (6.1) as

$$(I_N \otimes B)z_t = \mu + L\varsigma_t + \varepsilon_t,$$

where $\varsigma_t = g_t$ for DGP1; $\varsigma_t = (g_t, h_t)'$ for DGP2 and DGP3; $\varsigma_t = (g_t, h_t, d_t)'$ for DGP4, and L is the corresponding loadings matrix. Let ι_i' be the ith row of L. We generate the cross-sectional heteroscedasticity Ξ, an $N(K+1) \times 1$ vector, according to $\Xi_i = \frac{\eta_i}{1-\eta_i}\iota_i'\iota_i$, $i = 1, 2, \ldots, N(K+1)$, where η_i is drawn from $U[u, 1-u]$ with $u = 0.1$. A similar way of generating heteroscedasticity is also used in [14] and [16]. Let $\Upsilon = \mathrm{diag}(\Upsilon_1, \Upsilon_2, \ldots, \Upsilon_N)$ be an $N(K+1) \times N(K+1)$ block diagonal matrix, in which $\Upsilon_i = \mathrm{diag}\{1, (M_i'M_i)^{-1/2}M_i\}$ with M_i being a $K \times K$ standard normal random matrix for each i. Once Υ is generated, the error term ε_t, which is defined as $(\varepsilon_{1t}', \varepsilon_{2t}', \ldots, \varepsilon_{Nt}')'$ with $\varepsilon_{it} = (e_{it}, v_{it1}, v_{it2})'$, is calculated by $\varepsilon_t = \sqrt{\mathrm{diag}(\Xi)}\Upsilon\epsilon_t$, where ϵ_t is an $N(K+1) \times 1$ vector with all its elements being i.i.d. $(\chi_2^2 - 2)/2$, where χ_2^2 denotes the chi-squared distribution with two freedom degrees, which is normalized to mean zero and variance one. Additional simulation results for normal and student-t errors are given in Appendix D. Once ε_t is obtained, we use

$$z_t = (I_N \otimes B)^{-1}(\mu + L\varsigma_t + \varepsilon_t)$$

to yield the observable data.

In the basic model, the number of factors is determined by

$$\hat{r} = \underset{0 \le m \le r_{\max}}{\operatorname{argmin}} \ \mathrm{IC}(m) \tag{6.2}$$

with

$$\mathrm{IC}(m) = \frac{1}{N\overline{K}} \ln\big|\widehat{\Gamma}^m \widehat{\Gamma}^{m\prime} + \widehat{\Sigma}^m_{\varepsilon\varepsilon}\big| + m \frac{N\overline{K} + T}{N\overline{K}T} \ln\big(\min(N\overline{K}, T)\big),$$

where $\widehat{\Gamma}^m$ and $\widehat{\Sigma}^m_{\varepsilon\varepsilon}$ are the respective estimators of Γ and $\Sigma_{\varepsilon\varepsilon}$ when the factor number is set to m and $\overline{K} = K + 1$. In the simulation, we set $r_{\max} = 4$. For the model with zero restrictions, we consider a two-step method to determine r_1 and r_2. First, we use (6.2) to estimate the total number $r = r_1 + r_2$, denoted by $\hat{r}$, and obtain $\hat{\beta}^{\hat{r}}$ by the method of the basic model under $\hat{r}$. Then we calculate the matrix $\mathscr{R} = (\mathscr{R}_{it})$ with $\mathscr{R}_{it} = \dot{y}_{it} - \dot{x}_{it}\hat{\beta}^{\hat{r}}$ and use the information criterion proposed by [12] to determine the factor number in $\mathscr{R}$, which we use $\hat{r}_1$ to denote. In the second step, the upper bound of the factor number is set to $\hat{r}$. Then $\hat{r}_2 = \hat{r} - \hat{r}_1$. For models in Section 4, even though there are observable common regressors and time invariant regressors in the y equation, we treat them as part of the unknown factor structure when estimating the total number of factors. Once the total number of factors are obtained, the dimension of g_t is obtained by subtracting the dimension of ϕ_i and that of d_t because ϕ_i and d_t are observable in Section 4. This approach works very well. Other methods may also be considered.

We consider an unified way to estimate the model in Section 2 and the model in Section 3 (with zero restrictions). More specifically, for a given data set, we calculate r and r_1. If $\hat{r} = \hat{r}_1$, we turn to the basic model; if $\hat{r} > \hat{r}_1$, we turn to the model with zero restrictions.

Tables 1–2 report the simulation results based on 1000 repetitions. Bias and root mean square error (RMSE) are computed to measure the performance of the estimators. The percentage that the factor number is correctly estimated by the above procedure is given in the third column of each table. For comparison, we also report the performance of the within-group (WG) estimators and Bai's iterated principal components estimators (PC). Simulations for the models in Section 4 are provided in the supplement [11].

From the tables, we can see that the factor number can be correctly estimated with very high probability. It is also seen from the simulations that the WG estimators are inconsistent. The bias of the WG estimators shows no signs of decreasing as the sample size grows. The iterated PC estimators are consistent, but biased. As

TABLE 1

The performance of WG, PC and ML estimators in the basic model

N	**T**	**%** $\hat{r}=r$	**WG**				**PC**				**MLE**			
			β_1		β_2		β_1		β_2		β_1		β_2	
			Bias	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**
50	75	99.9	0.1562	0.1616	0.1550	0.1600	0.0174	0.0405	0.0171	0.0411	−0.0001	0.0020	0.0000	0.0034
100	75	100.0	0.1539	0.1568	0.1558	0.1587	0.0061	0.0228	0.0062	0.0224	0.0000	0.0011	0.0000	0.0010
150	75	100.0	0.1534	0.1556	0.1540	0.1561	0.0029	0.0168	0.0028	0.0146	0.0000	0.0007	0.0000	0.0007
50	125	100.0	0.1559	0.1605	0.1588	0.1636	0.0182	0.0389	0.0184	0.0409	0.0000	0.0017	0.0000	0.0016
100	125	100.0	0.1561	0.1586	0.1554	0.1579	0.0050	0.0167	0.0052	0.0167	0.0000	0.0009	0.0000	0.0008
150	125	100.0	0.1546	0.1565	0.1551	0.1570	0.0025	0.0108	0.0025	0.0106	0.0000	0.0006	0.0000	0.0005

TABLE 2

The performance of WG, PC and ML estimators in the model with zero restrictions

N	**T**	**%** $\hat{r}=r$	**WG**				**PC**				**MLE**			
			β_1		β_2		β_1		β_2		β_1		β_2	
			Bias	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**
50	75	99.7	0.1098	0.1137	0.1095	0.1135	0.0097	0.0245	0.0099	0.0246	0.0000	0.0012	0.0000	0.0011
100	75	100.0	0.1088	0.1111	0.1092	0.1114	0.0038	0.0140	0.0038	0.0140	0.0000	0.0006	0.0000	0.0006
150	75	100.0	0.1086	0.1102	0.1083	0.1099	0.0011	0.0075	0.0015	0.0076	0.0000	0.0004	0.0000	0.0004
50	125	99.7	0.1089	0.1121	0.1097	0.1130	0.0076	0.0199	0.0077	0.0196	0.0000	0.0009	0.0000	0.0009
100	125	100.0	0.1088	0.1107	0.1087	0.1106	0.0029	0.0104	0.0026	0.0100	0.0000	0.0005	0.0000	0.0004
150	125	100.0	0.1086	0.1099	0.1076	0.1090	0.0011	0.0055	0.0010	0.0054	0.0000	0.0003	0.0000	0.0003

the sample size becomes large, the bias decreases noticeably. However, when the sample size is moderate, the bias of the iterated PC estimators is still pronounced. In comparison, the ML estimators are consistent and unbiased. For all the sample sizes, the biases of the ML estimators are very small and negligible. In addition, the RMSEs of the ML estimators are always the smallest among the three estimators, illustrating the efficiency of the ML method. The same patten is observed for all of the four models considered.

7. Conclusion. This paper considers estimating panel data models with interactive effects, in which explanatory variables are correlated with the unobserved effects. Standard panel data methods (such as the within-group estimator) are not suitable for this type of models. We study the maximum likelihood method and provide a rigorous analysis for the asymptotic theory. While the analysis is difficult, the limiting distributions of the MLE are simple and have intuitive interpretations. The maximum likelihood method can incorporate parameter restrictions to gain efficiency, a useful feature in view of the large number of parameters under large N and large T. We analyze the restrictions via the Lagrange multiplier approach, which is capable of revealing what kinds of restrictions lead to efficiency gain. We allow the model to include time invariant regressors and common regressors. The coefficients of the time invariant regressors are time dependent, and the coefficients of the common regressors are cross-section dependent. This is a sensible way for modeling the effects of such variables in panel data context and fits naturally into the framework of interactive effects. The likelihood method is easy to implement and performs very well, as demonstrated by the Monte Carlo simulations.

Acknowledgments. The authors thank two anonymous referees, an Associate Editor and an Editor for constructive comments.

SUPPLEMENTARY MATERIAL

Supplement to "Theory and methods of panel data models with interactive effects" (DOI: 10.1214/13-AOS1183SUPP; .pdf). This supplement provides detailed technical proofs. Inferential theory for the estimated coefficients of time-invariant and common regressors is given. The EM solutions are shown to have local optimality property. Additional simulation results are presented.

REFERENCES

[1] AHN, S. C., LEE, Y. H. and SCHMIDT, P. (2001). GMM estimation of linear panel data models with time-varying individual effects. *J. Econometrics* **101** 219–255. MR1820251

[2] AHN, S. C., LEE, Y. H. and SCHMIDT, P. (2013). Panel data models with multiple time-varying individual effects. *J. Econometrics* **174** 1–14. MR3036957

[3] AMEMIYA, Y., FULLER, W. A. and PANTULA, S. G. (1987). The asymptotic distributions of some estimators for a factor analysis model. *J. Multivariate Anal.* **22** 51–64. MR0890881

[4] ANDERSON, T. W. and AMEMIYA, Y. (1988). The asymptotic normal distribution of estimators in factor analysis under general conditions. *Ann. Statist.* **16** 759–771. MR0947576

[5] ANDERSON, T. W. and RUBIN, H. (1956). Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability: Contributions to The Theory of Statistics*. Univ. California Press, Berkeley. MR0084943

[6] ARELLANO, M. (2003). *Panel Data Econometrics*. Oxford Univ. Press, Oxford. MR2060514

[7] BAI, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* **71** 135–171. MR1956857

[8] BAI, J. (2009a). Panel data models with interactive fixed effects. *Econometrica* **77** 1229–1279. MR2547073

[9] BAI, J. (2009b). Likelihood approach to small T dynamic panel models with interactive effects. Unpublished manuscript. Columbia Univ.

[10] BAI, J. and LI, K. (2012). Statistical analysis of factor models of high dimension. *Ann. Statist.* **40** 436–465. MR3014313

[11] BAI, J. and LI, K. (2014). Supplement to "Theory and methods of panel data models with interactive effects." DOI:10.1214/13-AOS1183SUPP.

[12] BAI, J. and NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221. MR1926259

[13] BALTAGI, B. H. (2005). *Econometric Analysis of Panel Data*. Wiley, Chichester.

[14] BREITUNG, J. and TENHOFEN, J. (2011). GLS estimation of dynamic factor models. *J. Amer. Statist. Assoc.* **106** 1150–1166. MR2894771

[15] CHAMBERLAIN, G. (1984). Panel data. In *Handbook of Econometrics* (Z. Griliches and M. Intriligator, eds.) **2** 1247–1318. North-Holland, Amsterdam.

[16] DOZ, C., GIANNONE, D. and REICHLIN, L. (2012). A qausi-maximum likelihood approach for large approximate dynamic factor models. *Rev. Econom. Statist.* **94** 1014–1024.

[17] HOLTZ-EAKIN, D., NEWEY, W. and ROSEN, H. S. (1988). Estimating vector autoregressions with panel data. *Econometrica* **56** 1371–1395.

[18] HSIAO, C. (2003). *Analysis of Panel Data*, 2nd ed. Cambridge Univ. Press, Cambridge. MR1962511

[19] JENNRICH, R. I. (1969). Asymptotic properties of nonlinear least squares estimators. *Ann. Math. Statist.* **40** 633–643. MR0238419

[20] KNEIP, A., SICKLES, R. C. and SONG, W. (2012). A new panel data treatment for heterogeneity in time trends. *Econometric Theory* **28** 590–628. MR2927921

[21] LAWLEY, D. N. and MAXWELL, A. E. (1971). *Factor Analysis as a Statistical Method*, 2nd ed. American Elsevier Publishing Co., Inc., New York. MR0343471

[22] MENG, X.-L. and RUBIN, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80** 267–278. MR1243503

[23] MOON, H. and WEIDNER, M. (2009). Likelihood expansion for panel regression models with factors. Unpublished manuscript. Univ. Southern California.

[24] MUNDLAK, Y. (1978). On the pooling of time series and cross section data. *Econometrica* **46** 69–85. MR0478489

[25] NEWEY, W. K. and MCFADDEN, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics* (R. F. Engle and D. McFadden, eds.). North-Holland, Amsterdam. MR1315971

[26] NEYMAN, J. and SCOTT, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16** 1–32. MR0025113

[27] PESARAN, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica* **74** 967–1012. MR2238209

[28] ROSS, S. A. (1976). The arbitrage theory of capital asset pricing. *J. Econom. Theory* **13** 341–360. MR0429063

[29] STOCK, J. H. and WATSON, M. W. (2002). Forecasting using principal components from a large number of predictors. *J. Amer. Statist. Assoc.* **97** 1167–1179. MR1951271

[30] SU, L., JIN, S. and ZHANG, Y. (2012). Specification test for panel data models with interactive fixed effects. Unpublished manuscript. Singapore Management Univ.

[31] TUCKER, L. R. (1958). An inter-battery method of factor analysis. *Psychometrika* **23** 111–136. MR0099737

[32] WESTERLUND, J. and URBAIN, J.-P. (2013). On the estimation and inference in factor-augmented panel regressions with correlated loadings. *Econom. Lett.* **119** 247–250. MR3053660

DEPARTMENT OF ECONOMICS
COLUMBIA UNIVERSITY
420, WEST 118TH STREET
NEW YORK, NEW YORK 10027
USA
E-MAIL: jushan.bai@columbia.edu

INTERNATIONAL SCHOOL OF ECONOMICS AND MANAGEMENT
CAPITAL UNIVERSITY OF ECONOMICS AND BUSINESS
HUAXIANG TOWN, FENGTAI DISTRICT
BEIJING 100070
CHINA
E-MAIL: likp.07@sem.tsinghua.edu.cn