THEORY AND METHODS OF PANEL DATA MODELS
WITH INTERACTIVE EFFECTS

BY JUSHAN BAI1 AND KUNPENG LI2

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 ei-
ther designed for panel data with short time periods or are less efficient. The
maximum likelihood estimator has desirable properties and is easy to im-
plement, as illustrated by the Monte Carlo simulations. This paper develops
the inferential theory for the maximum likelihood estimator, including con-
sistency, rate of convergence and the limiting distributions. We further ex-
tend the model to include time-invariant regressors and common regressors
(cross-section invariant). The regression coefficients for the time-invariant re-
gressors 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 un-
observable interactive effects:

\[ y_{it} = \alpha_i + x_{it}\beta + \lambda_i'f_t + e_{it}, \quad i = 1, \ldots, N, t = 1, 2, \ldots, T; \]

where \( y_{it} \) is the dependent variable; \( x_{it} = (x_{it1}, \ldots, x_{itK}) \) is a row vector of ex-
planatory variables; \( \alpha_i \) is an intercept; the term \( \lambda_i'f_t + e_{it} \) is unobservable and has a
factor structure, \( \lambda_i \) 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 ad-
ditive 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.

1Supported by the NSF (SES-0962410).

2Supported 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 $\beta$ 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 $\beta$. One method to obtain a consistent estimator is to treat $(\alpha_i, \lambda_i, f_t)$ as parameters and estimate them jointly with $\beta$. 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 $\alpha_i$ and $\lambda_i$ 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 $\eta_i$ is the projection residual (a similar projection is done for $\alpha_i$). 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 $\beta$.

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 $\lambda_i$ 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 \( \lambda_i \) 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 \( \lambda_i \); \( e_{it} \) 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; \( \lambda_i \) 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 \( \lambda_i \) 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 \( \hat{v}_t \) to denote \( v_t - \frac{1}{T} \sum_{t=1}^{T} v_t \) for any column vector \( v_t \) and \( M_{w,v} \) to denote \( \frac{1}{T} \sum_{t=1}^{T} \hat{w}_t \hat{v}_t' \) 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

$$y_{it} = \alpha_i + x_{it1}\beta_1 + x_{it2}\beta_2 + \cdots + x_{itK}\beta_K + \lambda'_i f_t + e_{it},$$

(2.1)

$$x_{itk} = \mu_{ik} + \gamma'_{ik} f_t + \nu_{itk}$$

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 $\gamma_{ik}$ $(k = 1, 2, \ldots, K)$. By treating the latter as parameters, we also allow arbitrary correlations between $(\alpha_i, \lambda_i)$ and $(\mu_{ik}, \gamma_{ik})$. 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 $x_{it} = (x_{it1}, x_{it2}, \ldots, x_{itK}), y_{ix} = (y_{i1}, y_{i2}, \ldots, y_{iK}), v_{itx} = (v_{it1}, v_{it2}, \ldots, v_{itK})'$ 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 $\Gamma_i = (\lambda_i, \gamma_{ix}), z_{it} = (y_{it}, x_{it})', e_{it} = (e_{it}, v_{itx})'$, $\mu_i = (\alpha_i, \mu_{ix}')$. Then model (2.1) can be written as

$$\begin{bmatrix} 1 & -\beta' \\ 0 & I_K \end{bmatrix} z_{it} = \mu_i + \Gamma'_i f_t + e_{it}.$$
Let $B$ denote the coefficient matrix of $z_{it}$ in the preceding equation. Let $z_t = (z'_1t, z'_2t, \ldots, z'_Nt)'$, $\Gamma = (\Gamma_1, \Gamma_2, \ldots, \Gamma_N)'$, $\epsilon_t = (\epsilon'_1t, \epsilon'_2t, \ldots, \epsilon'_Nt)'$ and $\mu = (\mu'_1, \mu'_2, \ldots, \mu'_N)'$. Stacking the equations over $i$, we have

$$
(I_N \otimes B)z_t = \mu + \Gamma f_t + \epsilon_t.
$$

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 $M_{ff} = \frac{T}{T-1} \sum_{t=1}^{T} \dot{f}_t \dot{f}_t'$, 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 $E(\|f_t\|^4) \leq C < \infty$ uniformly in $t$, and $f_t$ is independent of $\epsilon_s$ 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 $\epsilon_{it} = (e_{it}, v'_{itx})'$ are such that:

1. $e_{it}$ is independent and identically distributed over $t$ and uncorrelated over $i$ with $E(e_{it}) = 0$ and $E(e_{it}^4) \leq \infty$ for all $i = 1, \ldots, N$ and $t = 1, \ldots, T$. Let $\Sigma_{ii}^e$ denote the variance of $e_{it}$.

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

3. $e_{it}$ is independent of $v_{jtx}$ for all $(i, j, t, s)$. Let $\Sigma_{ii}$ denote the variance matrix $e_{it}$. So we have $\Sigma_{ii} = \text{diag}(\Sigma_{ii}^e, \Sigma_{ii}^x)$, a block-diagonal matrix.

**REMARK 2.2.** Let $\Sigma_{\epsilon\epsilon}$ denote the variance of $\epsilon_t = (\epsilon'_1t, \ldots, \epsilon'_Nt)'$. Due to the uncorrelatedness of $\epsilon_{it}$ over $i$, we have $\Sigma_{\epsilon\epsilon} = \text{diag}(\Sigma_{11}, \Sigma_{22}, \ldots, \Sigma_{NN})$, 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 $\epsilon_t$ 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 $\beta$, 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) $\|\Gamma_j\| \leq C$ for all $j = 1, \ldots, N$;
(C.2) $C^{-1} \leq \tau_{\min}(\Sigma_{jj}) \leq \tau_{\max}(\Sigma_{jj}) \leq C$ for all $j = 1, \ldots, N$, where $\tau_{\min}(\Sigma_{jj})$ and $\tau_{\max}(\Sigma_{jj})$ denote the smallest and largest eigenvalues of the matrix $\Sigma_{jj}$, respectively;
(C.3) there exists an $r \times r$ positive matrix $Q$ such that

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

where $\Gamma$ is defined earlier.

ASSUMPTION D. The variances $\Sigma_{ii}$ for all $i$ and $M_{ff}$ are estimated in a compact set, that is, all the eigenvalues of $\hat{\Sigma}_{ii}$ and $\hat{M}_{ff}$ 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

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

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

$$(I_N \otimes B)z_t = \mu^* + \Gamma^* f_t^* + \epsilon_t$$

with $\frac{1}{T} \sum_{t=1}^T f_t^* = 0$, $\frac{1}{T} \sum_{t=1}^T f_t^* f_t^* = I_r$ and $\frac{1}{N} \Gamma^* \Sigma_{ee}^{-1} \Gamma^*$ 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) $\frac{1}{N} \Gamma^* \Sigma_{ee}^{-1} \Gamma = D$, 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

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

with $\Sigma_{zz}$ being the variance-covariance matrix of the error terms.
where $\Sigma_{zz} = \Gamma M_{ff} \Gamma' + \Sigma_{ee}$ and $M_{zz} = \frac{1}{T} \sum_{t=1}^{T} \hat{z}_t \hat{z}_t'$. The latter is the data matrix. The parameters are $\theta = (\beta, \Gamma, M_{ff}, \Sigma_{ee})$. The MLE is defined as

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \ln L(\theta),$$

where the parameter space $\Theta$ is defined to be a closed and bounded subset containing the true parameter $\theta^*$ as an interior point; $\Sigma_{ee}$ and $M_{ff}$ are positive definite matrices, as in Assumption D. The boundedness of $\Theta$ implies that the elements of $\beta$ and $\Gamma$ 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 $\Theta$ (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 $\varepsilon_t$ 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 $\varepsilon_t$ are not necessarily normal; (2.4) is a pseudo-likelihood function.

For further analysis, we partition the matrix $\Sigma_{zz}$ and $M_{zz}$ as

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

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

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

$$(2.5) \quad \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{\Sigma}_{ii}^{-1} \left\{ (\hat{y}_{it} - \hat{x}_{it} \hat{\beta}) - \hat{x}_{it}' \hat{\Sigma}_{jj}^{-1} \hat{\Sigma}_{jj}' \hat{x}_{jt} \right\} \hat{x}_{it} = 0,$$

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

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

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

$$(2.7) \quad \sum_{i=1}^{N} \sum_{j=1}^{N} \hat{\Gamma}_i \hat{\Sigma}_{ii}^{-1} (\hat{B} M_{ij}^{zz} \hat{\beta}' - \hat{\Sigma}_{ij}^{zz}) \hat{\Sigma}_{jj}^{-1} \hat{\Gamma}_j' = 0.$$
The first order condition for $\Sigma_{ii}$ satisfies

\[
(2.8) \qquad \hat{B} M_{zz}^{ii} \hat{B}' - \hat{\Sigma}_{zz}^{ii} = \mathbb{W},
\]

where $\mathbb{W}$ is a $(K + 1) \times (K + 1)$ matrix such that its upper-left $1 \times 1$ 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 $\Sigma_{ii}$. 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 $\hat{\theta} = (\hat{\beta}, \hat{\Gamma}, \hat{\Sigma}_{\varepsilon\varepsilon})$ be the solution by maximizing (2.4). Under Assumptions A–D and the identification conditions IB, we have

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

\[
\frac{1}{N} \sum_{i=1}^{N} \| \hat{\Sigma}_{ii}^{-1} \| \cdot \| \hat{\Gamma}_i - \Gamma_i \|^2 = O_p(T^{-1}), \quad \frac{1}{N} \sum_{i=1}^{N} \| \hat{\Sigma}_{ii} - \Sigma_{ii} \|^2 = O_p(T^{-1}).
\]

**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 $\beta$ 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 $\mathcal{M}(\bar{X})$ denote the project matrix onto the space orthogonal to $\bar{X}$, that is, $\mathcal{M}(\bar{X}) = I - \bar{X}(\bar{X}'\bar{X})^{-1}\bar{X}'$. 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 $\Omega$ 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 $\Sigma_{iix}$.
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

$$
\hat{\beta} - \beta = \left( \begin{array}{ccc}
\text{tr}[\hat{\mathcal{M}}X_1\mathcal{M}(\overline{F})X'_1] & \cdots & \text{tr}[\hat{\mathcal{M}}X_1\mathcal{M}(\overline{F})X'_K] \\
\vdots & \ddots & \vdots \\
\text{tr}[\hat{\mathcal{M}}X_K\mathcal{M}(\overline{F})X'_1] & \cdots & \text{tr}[\hat{\mathcal{M}}X_K\mathcal{M}(\overline{F})X'_K]
\end{array} \right)^{-1} 
\times \left( \begin{array}{c}
\text{tr}[\hat{\mathcal{M}}X_1\mathcal{M}(\overline{F})e'] \\
\vdots \\
\text{tr}[\hat{\mathcal{M}}X_K\mathcal{M}(\overline{F})e']
\end{array} \right) + O_p(T^{-3/2}) + O_p(N^{-1}T^{-1/2}) + O_p(N^{-1/2}T^{-1}),
$$

(2.9)

where $X_k = (x_{itk})$ is $N \times T$ (the data matrix for the $k$th regressor, $k = 1, 2, \ldots, K$); $e = (e_{it})$ is $N \times T$; $\hat{\mathcal{M}} = \Sigma_{ee}^{-1/2} \mathcal{M}(\Sigma_{ee}^{-1/2} \Lambda) \Sigma_{ee}^{-1/2}$ with $\Sigma_{ee} = \text{diag} \{\Sigma_{11e}, \Sigma_{22e}, \ldots, \Sigma_{NNe}\}$ and $\Lambda = (\lambda_1, \lambda_2, \ldots, \lambda_N)'$; $\overline{F} = (f_1, f_2, \ldots, f_T)'$; $\overline{F} = (1_T, \overline{F})$ 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 $v_{ixi}$. 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 $\beta$ 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 $\sqrt{N}/T \rightarrow 0$, we have

$$
\text{N}(0, \overline{\Omega}^{-1}),
$$

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

$$
\overline{\Omega} = \text{plim} \frac{1}{NT} \left( \begin{array}{ccc}
\text{tr}[\hat{\mathcal{M}}X_1\mathcal{M}(\overline{F})X'_1] & \cdots & \text{tr}[\hat{\mathcal{M}}X_1\mathcal{M}(\overline{F})X'_K] \\
\vdots & \ddots & \vdots \\
\text{tr}[\hat{\mathcal{M}}X_K\mathcal{M}(\overline{F})X'_1] & \cdots & \text{tr}[\hat{\mathcal{M}}X_K\mathcal{M}(\overline{F})X'_K]
\end{array} \right).
$$

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

$$
\frac{1}{NT} \left( \begin{array}{ccc}
\text{tr}[\hat{\mathcal{M}}X_1\mathcal{M}(\overline{F})X'_1] & \cdots & \text{tr}[\hat{\mathcal{M}}X_1\mathcal{M}(\overline{F})X'_K] \\
\vdots & \ddots & \vdots \\
\text{tr}[\hat{\mathcal{M}}X_K\mathcal{M}(\overline{F})X'_1] & \cdots & \text{tr}[\hat{\mathcal{M}}X_K\mathcal{M}(\overline{F})X'_K]
\end{array} \right),
$$

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

\[
(2.10) \quad \hat{M} = \hat{\Sigma}_{ee}^{-1} - \hat{\Sigma}_{ee}^{-1} \hat{\Lambda} (\hat{\Lambda}' \hat{\Sigma}_{ee}^{-1} \hat{\Lambda})^{-1} \hat{\Lambda}' \hat{\Sigma}_{ee}^{-1}.
\]

$\hat{P} = (1_T, \hat{P})$ with $\hat{P} = (\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_T)'$ and

\[
(2.11) \quad \hat{f}_i = \left( \sum_{i=1}^{N} \hat{\Gamma}_i \hat{\Sigma}_{ii}^{-1} \hat{\Gamma}_i' \right)^{-1} \left( \sum_{i=1}^{N} \hat{\Gamma}_i \hat{\Sigma}_{ii}^{-1} \hat{B} z_{it} \right).
\]

Here $\hat{\Gamma}$, $\hat{\Lambda}$, $\hat{\Sigma}_{ii}$, $\hat{\Sigma}_{ee}$ and $\hat{B}$ 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:

\[
y_{it} = \alpha_i + x_{it1} \beta_1 + x_{it2} \beta_2 + \cdots + x_{itK} \beta_K + \psi_i g_t + e_{it},
\]

\[
x_{itk} = \mu_{ik} + \psi_{ik} g_t + \gamma_{ik} h_t + \nu_{itk}
\]

for $k = 1, 2, \ldots, K$, 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 $\lambda_i = (\psi_i, 0_{r_2 \times 1})'$, $\gamma_{ik} = (\gamma_{ik}^g, \gamma_{ik}^h)'$ and $f_i = (g_i', h_i')'$, the above model can be written as

\[
y_{it} = \alpha_i + x_{it1} \beta_1 + x_{it2} \beta_2 + \cdots + x_{itK} \beta_K + \lambda_i' f_i + e_{it},
\]

\[
x_{itk} = \mu_{ik} + \gamma_{ik}^l f_t + \nu_{itk},
\]

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

\[
\Gamma_i^g = (\psi_i, \gamma_{i1}^g, \ldots, \gamma_{iK}^g), \quad \Gamma_i^h = (0_{r_2 \times 1}, \gamma_{i1}^h, \ldots, \gamma_{iK}^h),
\]

\[
\Gamma^g = (\Gamma_1^g, \Gamma_2^g, \ldots, \Gamma_N^g)', \quad \Gamma^h = (\Gamma_1^h, \Gamma_2^h, \ldots, \Gamma_N^h)'.
\]

We also define $\mathbb{G}$ and $\mathbb{H}$ similarly as $\mathbb{F}$, that is, $\mathbb{G} = (g_1, g_2, \ldots, g_T)'$, $\mathbb{H} = (h_1, h_2, \ldots, 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) $\frac{1}{N} \Gamma^g \Sigma_{ee}^{-1} \Gamma^g = D_1$ and $\frac{1}{N} \Gamma^h \Sigma_{ee}^{-1} \Gamma^h = D_2$, 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.** \( \Psi = (\psi_1', \psi_2', \ldots, \psi_N') \)' is of full column rank.

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

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

\[
\ln L = -\frac{1}{2N} \ln |\Sigma_{zz}| - \frac{1}{2N} \text{tr}[(I_N \otimes B)M_{zz}(I_N \otimes B')\Sigma_{zz}^{-1}]
\]

\[
+ \text{tr}\left[\gamma_1 \left(\frac{1}{N} \Gamma^g \Sigma_{ee}^{-1} \Gamma^g - D_1\right)\right] + \text{tr}\left[\gamma_2 \left(\frac{1}{N} \Gamma^h \Sigma_{ee}^{-1} \Gamma^h - D_2\right)\right]
\]

\[
+ \text{tr}[\gamma_3 \Phi],
\]

where \( \Sigma_{zz} = \Gamma' + \Sigma_{ee} \); \( \gamma_1 \) is \( r_1 \times r_1 \) and \( \gamma_2 \) is \( r_2 \times r_2 \), both are symmetric Lagrange multiplier matrices with zero diagonal elements; \( \gamma_3 \) is a Lagrange multiplier matrix of dimension \( r_2 \times N \).

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

\[
\frac{1}{N} \hat{\Gamma}^g \mathbb{U} + \gamma_1 \frac{1}{N} \hat{\Gamma}^g \Sigma_{ee}^{-1} = 0.
\]

Post-multiplying \( \hat{\Gamma}^g \) yields

\[
\frac{1}{N} \hat{\Gamma}^g \mathbb{U} \hat{\Gamma}^g + \gamma_1 \frac{1}{N} \hat{\Gamma}^g \Sigma_{ee}^{-1} \hat{\Gamma}^g = 0.
\]

Since \( \frac{1}{N} \hat{\Gamma}^g \mathbb{U} \hat{\Gamma}^g \) is a symmetric matrix, the above equation implies that \( \gamma_1 \frac{1}{N} \hat{\Gamma}^g \Sigma_{ee}^{-1} \hat{\Gamma}^g \) is also symmetric. But \( \frac{1}{N} \hat{\Gamma}^g \Sigma_{ee}^{-1} \hat{\Gamma}^g \) is a diagonal matrix. So the \((i, j)\)th element of \( \gamma_1 \frac{1}{N} \hat{\Gamma}^g \Sigma_{ee}^{-1} \hat{\Gamma}^g \) is \( \gamma_{1, ij} d_{1j} \), where \( \gamma_{1, ij} \) is the \((i, j)\)th element of \( \gamma_1 \) and \( d_{1j} \) is the \( j \)th diagonal element of \( \hat{D}_1 \). Given \( \gamma_1 \frac{1}{N} \hat{\Gamma}^g \Sigma_{ee}^{-1} \hat{\Gamma}^g \) is symmetric, we have \( \gamma_{1, ij} d_{1j} = \gamma_{1, ji} d_{1i} \) for all \( i \neq j \). However, \( \gamma_1 \) is also symmetric, so \( \gamma_{1, ij} = \gamma_{1, ji} \). This gives \( \gamma_{1, ij}(d_{1j} - d_{1i}) = 0 \). Since \( d_{1j} \neq d_{1i} \) by IZ2, we have \( \gamma_{1, ij} = 0 \) for all \( i \neq j \). This implies \( \gamma_1 = 0 \) since the diagonal elements of \( \gamma_1 \) are all zeros.

Let \( \Gamma^h = (\gamma_{1h}, \gamma_{2h}, \ldots, \gamma_{Nh})' \) with \( \gamma_{ih} = (\gamma_{i1h}, \gamma_{i2h}, \ldots, \gamma_{iKh}) \), and \( \Sigma_{xx} = \text{diag}\{\Sigma_{11x}, \Sigma_{22x}, \ldots, \Sigma_{Nlx}\} \), a block diagonal matrix of \( NK \times NK \) dimension.
We partition the matrix $\mathbf{U}$ and define the matrix $\overline{\mathbf{U}}$ as

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

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

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

$$
\frac{1}{N} \hat{\Gamma}^h_x \overline{\mathbf{U}} + \gamma_2 \frac{1}{N} \hat{\Gamma}^h_x \Sigma^{-1}_{xx} = 0.
$$

Post-multiplying $\hat{\Gamma}^h_x$ yields

$$
\frac{1}{N} \hat{\Gamma}^h_x \overline{\mathbf{U}} \hat{\Gamma}^h_x + \gamma_2 \frac{1}{N} \hat{\Gamma}^h_x \Sigma^{-1}_{xx} \hat{\Gamma}^h_x = 0.
$$

Notice $\frac{1}{N} \hat{\Gamma}^h_x \Sigma^{-1}_{xx} \hat{\Gamma}^h_x = \frac{1}{N} \hat{\Gamma}^h_x \Sigma^{-1}_{xx} \hat{\Gamma}^h_x = \hat{D}_2$. By the similar arguments in deriving $\gamma_1 = 0$, we have $\gamma_2 = 0$. 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 $\gamma_3$ 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

$$
\ln L = -\frac{1}{2N} \ln |\Sigma_{zz}| - \frac{1}{2N} \text{tr}[(I_N \otimes B) M_{zz} (I_N \otimes B') \Sigma_{zz}^{-1}] + \text{tr}[\gamma_3' \Phi].
$$

The first order condition on $\Gamma$ is

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

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

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

By the special structure of $W$ and $\hat{\Gamma}$, it is easy to verify that $W' \hat{\Gamma}$ 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 “×” is zero, that is, $W' \hat{\Gamma} = 0$. 

---

**Panel Data Models with Interactive Effects** 153
Thus, \( \hat{\Gamma}' \hat{\Sigma}_{zz}^{-1} [(I_N \otimes \hat{B}) M_{zz} (I_N \otimes \hat{B}')] - \hat{\Sigma}_{zz}^{-1} \hat{\Gamma} = 0 \). (This equation would be the first order condition for \( M_{ff} \) if it were unknown.) This equality can be simplified as

\[
(3.4) \quad \hat{\Gamma}' \hat{\Sigma}_{ee}^{-1} [(I_N \otimes \hat{B}) M_{zz} (I_N \otimes \hat{B}')] - \hat{\Sigma}_{zz}^{-1} \hat{\Gamma} = 0
\]

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

\[
\hat{G} = \begin{bmatrix} \hat{G}_1 \\
\hat{G}_21 \\
\hat{G}_22 \end{bmatrix} = \begin{bmatrix} \hat{G}_{11} & \hat{G}_{12} \\
\hat{G}_{21} & \hat{G}_{22} \end{bmatrix}, \quad \hat{H} = \begin{bmatrix} \hat{H}_1 \\
\hat{H}_21 \\
\hat{H}_22 \end{bmatrix} = \begin{bmatrix} \hat{H}_{11} & \hat{H}_{12} \\
\hat{H}_{21} & \hat{H}_{22} \end{bmatrix},
\]

where \( \hat{G}_{11}, \hat{H}_{11} \) are \( r_1 \times r_1 \), while \( \hat{G}_{22}, \hat{H}_{22} \) are \( r_2 \times r_2 \).

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

\[
(3.5) \quad \hat{G}_1 \sum_{i=1}^{N} \hat{\Gamma}_i \hat{\Sigma}_{ii}^{-1} (\hat{B} M_{jj}^{ii} \hat{B}' - \hat{\Sigma}_{zz}^{-1}) \hat{\Sigma}_{jj}^{-1} I_{K+1}^1 = 0,
\]

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

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

\[
(3.6) \quad \sum_{i=1}^{N} \hat{\Gamma}_i \hat{\Sigma}_{ii}^{-1} (\hat{B} M_{jj}^{ii} \hat{B}' - \hat{\Sigma}_{zz}^{-1}) \hat{\Sigma}_{jj}^{-1} I_{K+1}^- = 0,
\]

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 \( \Sigma_{jj} \) is

\[
\hat{B} M_{jj}^{ii} \hat{B}' - \hat{\Sigma}_{zz}^{jj} - \hat{\Gamma}_j \hat{G} \sum_{i=1}^{N} \hat{\Gamma}_i \hat{\Sigma}_{ii}^{-1} (\hat{B} M_{jj}^{ii} \hat{B}' - \hat{\Sigma}_{zz}^{jj})
\]

\[
- \sum_{i=1}^{N} (\hat{B} M_{jj}^{ii} \hat{B}' - \hat{\Sigma}_{zz}^{jj}) \hat{\Sigma}_{ii}^{-1} \hat{\Gamma}_i \hat{G} \hat{\Gamma}_j = \mathbb{W},
\]

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

The first order condition for \( \beta \) is

\[
(3.8) \quad \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{r=1}^{T} \hat{\Sigma}_{ie}^{-1} \left\{ \left( \hat{y}_{it} - \hat{x}_{it} \hat{\beta} \right) - \hat{x}_{it} \hat{G} \sum_{j=1}^{N} \hat{\Gamma}_j \hat{\Sigma}_{jj}^{-1} \left[ \hat{\beta}_j \hat{x}_{jt} \right] \right\} = 0,
\]

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 \( \hat{\Gamma} \), the three submatrices of \( W' \hat{\Gamma} \) 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):

\[
\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \hat{G}_i \hat{\Sigma}_{ij}^{-1} \left( \hat{B} M_{ij} \hat{B}' - \hat{\Sigma}_{ij} \right) \hat{I}_{K+1} \hat{\psi}' = 0.
\]

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

\[
\mathcal{P}^0(\hat{\beta} - \beta) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \Sigma_{iit}^{-1} e_{it} v_{itx} + \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \Sigma_{iit}^{-1} \gamma_{iht} h_{it}
\]

\[
- \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \Sigma_{iit}^{-1} \psi_i' \Pi_{\psi}^{-1} \left( \frac{1}{N} \sum_{j=1}^{N} \psi_j \Sigma_{jjt}^{-1} \gamma_{jhx} h_{it} \right) + O_p(T^{-3/2}) + O_p(N^{-1}T^{-1/2}) + O_p(N^{-1/2}T^{-1}),
\]

where \( \mathcal{P}^0 \) is a \( K \times K \) symmetric matrix with its \((p,q)\) element equal to

\[
\frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \Sigma_{iit}^{-1} \gamma_{ih} \gamma_{ih}' h_{it} + \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \Sigma_{iit}^{-1} \psi_i' \Pi_{\psi}^{-1} \left( \frac{1}{N} \sum_{j=1}^{N} \psi_j \Sigma_{jjt}^{-1} \gamma_{jhx} h_{it} \right),
\]

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 \((\psi_i^*, \gamma_{ik}^*, g_t^*, h_t^*)\), 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

\[
\mathcal{P}(\hat{\beta} - \beta) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \Sigma_{iit}^{-1} e_{it} v_{itx} + \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \Sigma_{iit}^{-1} \gamma_{iht} h_{it}
\]

\[
- \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \Sigma_{iit}^{-1} \psi_i' \Pi_{\psi}^{-1} \left( \frac{1}{N} \sum_{j=1}^{N} \psi_j \Sigma_{jjt}^{-1} \gamma_{jhx} h_{it} \right) + O_p(T^{-3/2}) + O_p(N^{-1}T^{-1/2}) + O_p(N^{-1/2}T^{-1}),
\]
where

\[ h_t^* = \dot{h}_t - \mathbb{H}(\mathbb{G}'\mathbb{G})^{-1} \dot{g}_t; \]

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

\[
\frac{1}{NT} \text{tr}[\dot{M} \Gamma_p h \mathbb{H}' \mathbb{M}(\mathbb{G}) \mathbb{H} \Gamma_p h'] + \frac{1}{N} \sum_{i=1}^{N} \Sigma_{ii}^{-1} \Sigma_{iix},
\]

where \( \mathbb{G} = (1_T, \mathbb{G}); \ \Pi_{\psi} = \frac{1}{N} \sum_{i=1}^{N} \psi_i \Sigma_{ii}^{-1} \psi_i; \ \dot{M} = \Sigma_{ee}^{-1/2} \mathcal{M}(\Sigma_{ee}^{-1/2} \Psi) \Sigma_{ee}^{-1/2}, \ \Gamma^h_p = (\gamma^h_1, \gamma^h_2, \ldots, \gamma^h_N)' \).

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

\[
\hat{\beta} - \beta = \left( \begin{array}{ccc}
\text{tr}[\dot{M} X_1 \mathcal{M}(\mathbb{G}) X_1'] & \cdots & \text{tr}[\dot{M} X_1 \mathcal{M}(\mathbb{G}) X_K'] \\
\vdots & \ddots & \vdots \\
\text{tr}[\dot{M} X_K \mathcal{M}(\mathbb{G}) X_1'] & \cdots & \text{tr}[\dot{M} X_K \mathcal{M}(\mathbb{G}) X_K']
\end{array} \right)^{-1} \\
\times \left( \begin{array}{c}
\text{tr}[\dot{M} X_1 \mathcal{M}(\mathbb{G}) e'] \\
\vdots \\
\text{tr}[\dot{M} X_K \mathcal{M}(\mathbb{G}) e']
\end{array} \right)
+ O_p(T^{-3/2}) + O_p(N^{-1}T^{-1/2}) + O_p(N^{-1/2}T^{-1}),
\]

where \( X_k \) and \( e \) are defined below (2.9) and \( \mathbb{G} = (1_T, \mathbb{G}). \) Notice \( \dot{M} \) 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

\[
\sqrt{NT}(\hat{\beta} - \beta) \overset{d}{\to} N(0, \overline{P}^{-1}),
\]

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

\[
\overline{P} = \text{plim}_{N,T\to\infty} \frac{1}{NT} \left( \begin{array}{ccc}
\text{tr}[\dot{M} X_1 \mathcal{M}(\mathbb{G}) X_1'] & \cdots & \text{tr}[\dot{M} X_1 \mathcal{M}(\mathbb{G}) X_K'] \\
\vdots & \ddots & \vdots \\
\text{tr}[\dot{M} X_K \mathcal{M}(\mathbb{G}) X_1'] & \cdots & \text{tr}[\dot{M} X_K \mathcal{M}(\mathbb{G}) X_K']
\end{array} \right).
\]
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 \( \nu_{it} \) provide information for \( \beta \). But in the present case, variations in \( \gamma_{ih} 'i \) of \( x_{it} \) also provide information for \( \beta \). This can also be seen by comparing the limiting variances of Corollaries 2.1 and 3.1. Notice the projection matrix now only involves \( \bar{G} \) instead of \( \bar{F} \); and \( \bar{G} \) is a submatrix of \( \bar{F} \). In addition, the covariance matrix \( \bar{P} \) can be estimated by the same method as in estimating \( \bar{\Omega} \); 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,

\[
\begin{align*}
  y_{it} &= \alpha_i + x_{it1} \beta_1 + x_{it2} \beta_2 + \cdots + x_{itK} \beta_K + \psi_i g_t + \phi_i h_t + e_{it}, \\
  x_{itk} &= \mu_{ik} + \gamma_{ik} g_t + \gamma_{ik} h_t + \nu_{itk}
\end{align*}
\]

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 \( \phi_i \)'s are known (but not zeros). We treat \( \phi_i \) as new added time-invariant regressors, whose coefficient \( h_t \) is allowed to be time-varying. The parameter of interest is still \( \beta \). 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 \( \Phi_1 = 0 \), where \( \Phi = (\phi_1, \phi_2, \ldots, \phi_N)' \). However, the earlier derivation is not applicable here because now \( \Phi \) 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:

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

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 $M_{gg} = I_r$;

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

(IO3) $1_T^T G = 0$ and $1_T^T H = 0$.

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 $\Phi^*$ to denote the observed value for $\Phi$. Recall that $\Sigma_{zz} = \Gamma M_{ff} \Gamma^T + \Sigma_{ee}$, where $\Gamma$ contains the factor loading coefficients (including $\Phi$); $M_{ff}$ contains the sub-blocks $M_{gg}, M_{gh}$ and $M_{hh}$; $\Sigma_{ee}$ contains the heteroskedasticity coefficients. The regression coefficient $\beta$ is contained in matrix $B$. The maximization of the likelihood function is now subject to four sets of restrictions, $M_{gh} = 0$, $M_{gg} = I_r$, $\Phi = \Phi^*$ and $\frac{1}{N} \Gamma^g \Sigma_{ee}^{-1} \Gamma^g = D$. The likelihood function augmented with the Lagrange multipliers is

$$\ln L = -\frac{1}{2N} \ln |\Sigma_{zz}| - \frac{1}{2N} \text{tr} [(I_N \otimes B) M_{zz} (I_N \otimes B^T) \Sigma_{zz}^{-1}] + \text{tr} [\Upsilon_1 M_{gh}]$$

$$+ \text{tr} [\Upsilon_2 (M_{gg} - I_r)] + \text{tr} [\Upsilon_3 \left( \frac{1}{N} \Gamma^g \Sigma_{ee}^{-1} \Gamma^g - D \right)] + \text{tr} [\Upsilon_4 (\Phi - \Phi^*)],$$

where $\Upsilon_1, \Upsilon_2, \Upsilon_3$ and $\Upsilon_4$ are all Lagrange multipliers matrices; $\Upsilon_1$ is an $r_2 \times r_1$ matrix; $\Upsilon_2$ is an $r_1 \times r_1$ symmetric matrix; $\Upsilon_3$ is an $r_1 \times r_1$ symmetric matrix with all diagonal elements zeros; $\Upsilon_4$ is an $r_2 \times N$ matrix; and $\Sigma_{zz} = \Gamma M_{ff} \Gamma^T + \Sigma_{ee}$. Using the same arguments in deriving $\Upsilon_1 = 0$ in Section 3, we have $\Upsilon_3 = 0$. Then the likelihood function is simplified as

$$\ln L = -\frac{1}{2N} \ln |\Sigma_{zz}| - \frac{1}{2N} \text{tr} [(I_N \otimes B) M_{zz} (I_N \otimes B^T) \Sigma_{zz}^{-1}]$$

$$+ \text{tr} [\Upsilon_1 M_{gh}] + \text{tr} [\Upsilon_2 (M_{gg} - I_r)] + \text{tr} [\Upsilon_4 (\Phi - \Phi^*)].$$

(4.2)

The first order condition for $\Gamma$ gives

$$\widehat{M}_{ff} \Gamma^T \hat{\Sigma}_{zz}^{-1} [(I_N \otimes \hat{B}) M_{zz} (I_N \otimes \hat{B}^T) - \hat{\Sigma}_{zz}] \hat{\Sigma}_{zz}^{-1} = W',$$

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

\[
\frac{1}{N} \hat{\Gamma}' \hat{\Sigma}_{zz}^{-1} \left[ (I_N \otimes \hat{B}) M_{zz} (I_N \otimes \hat{B}') - \hat{\Sigma}_{zz} \right] \hat{\Sigma}_{zz}^{-1} \hat{\Gamma} = - \begin{bmatrix} 0_{r_1 \times r_1} & 0_{r_1 \times r_2} \\ \frac{1}{N} \hat{M}_{hh}^{-1} \gamma_4' \hat{\Psi} & \frac{1}{N} \hat{M}_{hh}^{-1} \gamma_4' \Phi \end{bmatrix}.
\]

But the first order condition for \( M_{ff} \) gives

\[
\frac{1}{N} \hat{\Gamma}' \hat{\Sigma}_{zz}^{-1} \left[ (I_N \otimes \hat{B}) M_{zz} (I_N \otimes \hat{B}') - \hat{\Sigma}_{zz} \right] \hat{\Sigma}_{zz}^{-1} \hat{\Gamma} = \begin{bmatrix} \gamma_2 & \gamma_1' \\ \gamma_1 & 0_{r_2 \times r_2} \end{bmatrix}.
\]

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

\[
\frac{1}{N} \hat{\Gamma}' \hat{\Sigma}_{zz}^{-1} \left[ (I_N \otimes \hat{B}) M_{zz} (I_N \otimes \hat{B}') - \hat{\Sigma}_{zz} \right] \hat{\Sigma}_{zz}^{-1} \hat{\Gamma} = 0.
\]

The above result implies that \( \gamma_1 = 0, \gamma_2 = 0, \gamma_4' \hat{\Psi} = 0 \) and \( \gamma_4' \Phi = 0. \)

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

\[
\hat{B} M_{zz}^{ij} \hat{B}' - \hat{\Sigma}_{zz}^{jj} - \hat{\Gamma}_j' \hat{G} \sum_{i=1}^{N} \hat{\Gamma}_i \hat{\Sigma}_{ii}^{-1} (\hat{B} M_{zz}^{ij} \hat{B}' - \hat{\Sigma}_{zz}^{ij})
\]

\[
- \sum_{i=1}^{N} (\hat{B} M_{zz}^{ij} \hat{B}' - \hat{\Sigma}_{zz}^{ij}) \hat{\Sigma}_{ii}^{-1} \hat{\Gamma}_i' \hat{G} \hat{\Gamma}_j = \mathbb{W},
\]

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

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

\[
\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{\Sigma}_{ii}^{-1} \left\{ (\hat{y}_{it} - \hat{x}_{it} \hat{B}) - \hat{\lambda}_j' \hat{G} \sum_{j=1}^{N} \hat{\Gamma}_j \hat{\Sigma}_{jj}^{-1} \left[ \hat{y}_{jt} - \hat{x}_{jt} \hat{B} \right] \hat{x}_{jt} \right\} = 0.
\]

We need an additional identity for the theoretical analysis in the Appendix. The preceding analysis shows that \( \frac{1}{N} \gamma_4' \hat{\Psi} = 0 \) and \( \frac{1}{N} \gamma_4' \Phi = 0. \) They imply

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

where \( \hat{\lambda}_j = (\hat{\psi}_j', \phi_j'). \)
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

$$Q^0(\hat{\beta} - \beta) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \Sigma_{ii}\epsilon_{it}v_{itx} + \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \Sigma_{ii}\gamma_{ix}\hat{h}_t\epsilon_{it}$$

$$+ \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \Sigma_{ii}\gamma_{ix}\hat{h}_t\epsilon_{it}$$

$$- \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \Sigma_{ii}\lambda_i\Pi_{\lambda\lambda}^{-1}\left(\frac{1}{N} \sum_{j=1}^{N} \lambda'_{j}\Sigma_{jj}\gamma_{jx}\right)\hat{h}_t\epsilon_{it}$$

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

where $Q^0$ is a $K \times K$ symmetric matrix with its $(p, q)$ element equal to

$$\frac{1}{N} \text{tr}\left[M_{hh}^{\gamma} G_{p}\tilde{\Gamma}_{q}\right] + \frac{1}{N} \sum_{i=1}^{N} \Sigma_{ii}(p, q) \text{; } \tilde{\Gamma}_{p} = [\gamma_{h1}$, $\gamma_{h2}$, ..., $\gamma_{hp}]; \text{ and } \Pi_{\lambda\lambda} = \frac{1}{N} \sum_{i=1}^{N} \lambda_i \Sigma_{ii}\lambda_i';$$

and $\gamma_{jx}$ is untransformed. This is in agreement with the Lagrange multiplier analysis, in which $\gamma_{jx}$ is defined in Proposition 4.1, which satisfies IO.

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 $(\psi_i^*, \gamma_{ik}^*, g_t^*, h_t^*)$ which satisfies IO, and at the same time, still maintains the observability of $\Phi$ (i.e., $\Phi$ is untransformed). 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

$$Q(\hat{\beta} - \beta) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \Sigma_{ii}\epsilon_{it}v_{itx} + \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \Sigma_{ii}\gamma_{ix}\hat{h}_t^*\epsilon_{it}$$

$$- \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \Sigma_{ii}\lambda_i\Pi_{\lambda\lambda}^{-1}\left(\frac{1}{N} \sum_{j=1}^{N} \lambda'_{j}\Sigma_{jj}\gamma_{jx}\right)\hat{h}_t^*\epsilon_{it}$$

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

where $h_t^* = \hat{h}_t - \hat{\Gamma}_{q}\hat{\Gamma}_{q}\hat{\gamma}_{p}\hat{g}_t; Q$ is a $K \times K$ symmetric matrix with its $(p, q)$ element equal to

$$\frac{1}{NT} \text{tr}\left[M_{hh}\Gamma_{p}\tilde{\Gamma}_{q}\right] + \frac{1}{N} \sum_{i=1}^{N} \Sigma_{ii}\Sigma_{ii}\gamma_{i}(p, q)$$

and $\tilde{\Gamma}_{q}$, $\Gamma_{p}^h$ 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

$$\hat{\beta} - \beta = \left( \begin{array}{ccc} \text{tr}[\hat{M}X_1\mathcal{M}(\overline{G})X'_1] & \cdots & \text{tr}[\hat{M}X_1\mathcal{M}(\overline{G})X'_K] \\ \vdots & \ddots & \vdots \\ \text{tr}[\hat{M}X_K\mathcal{M}(\overline{G})X'_1] & \cdots & \text{tr}[\hat{M}X_K\mathcal{M}(\overline{G})X'_K] \end{array} \right)^{-1} \times \left( \begin{array}{c} \text{tr}[\hat{M}X_1\mathcal{M}(\overline{G})e'] \\ \vdots \\ \text{tr}[\hat{M}X_K\mathcal{M}(\overline{G})e'] \end{array} \right) + O_p(T^{-3/2})$$

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

where $X_k$ and $e$ are defined below (2.9) and $\overline{G} = (1_T, G)$. 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

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

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

$$\overline{Q} = \text{plim}_{N,T \to \infty} \frac{1}{NT} \left( \begin{array}{ccc} \text{tr}[\hat{M}X_1\mathcal{M}(\overline{G})X'_1] & \cdots & \text{tr}[\hat{M}X_1\mathcal{M}(\overline{G})X'_K] \\ \vdots & \ddots & \vdots \\ \text{tr}[\hat{M}X_K\mathcal{M}(\overline{G})X'_1] & \cdots & \text{tr}[\hat{M}X_K\mathcal{M}(\overline{G})X'_K] \end{array} \right).$$

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:

$$y_{it} = x_{it1}\beta_1 + x_{it2}\beta_2 + \cdots + x_{itK}\beta_K + \psi'_i g_t + \phi'_i h_t + \kappa'_i d_t + e_{it},$$

$$(4.8)$$

$$x_{itk} = \gamma_{ik}^g g_t + \gamma_{ik}^h h_t + \gamma_{ik}^d d_t + v_{itk}$$

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 $\phi_i$ are observable for all $i$ and $t$. We call $\phi_i$ 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 \( d_t \equiv 1, \kappa_i \) plays the role of \( \alpha_i \) in (4.1). So the model here is more general.

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

**ASSUMPTION G.** The matrices \((\Psi, \Phi, K)\) and \((G, H, D)\) are both of full column rank, where \( K = (\kappa_1, \kappa_2, \ldots, \kappa_N)' \) and \( D = (d_1, d_2, \ldots, d_T)' \).

Let \( \lambda_i = (\psi_i', \phi_i')' \), \( \gamma_{ik} = (\gamma_{ik}'^g, \gamma_{ik}'^h)' \) and \( \delta_i = (\kappa_i, \gamma_{id}^d) \). 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; \( \Delta = (\delta_1, \delta_2, \ldots, \delta_N)' \). Then (4.9)

\[
(I_N \otimes B)z_t - \Delta d_t = \Gamma f_t + \varepsilon_t,
\]

where the symbols \( \Gamma, z_t, B, \varepsilon_t \) are defined in Section 2.

The likelihood function can be written as

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

Take \( \Sigma_{zz} \) and \( \beta \) as given. \( \Delta \) maximizes the above function at

\[
\hat{\Delta} = (I_N \otimes B) \left( \sum_{s=1}^{T} z_s d_s \right) \left( \sum_{s=1}^{T} d_s d_s' \right)^{-1},
\]

Substituting \( \hat{\Delta} \) into the above likelihood function, we obtain the concentrated likelihood function

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

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

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

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

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

If we treat \( Z M(\mathbb{D}) \) as the new observable data, \( F' M(\mathbb{D}) \) 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 \( F' M(\mathbb{D}) \) 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

$$R(\hat{\beta} - \beta) = \frac{1}{NT} \sum_{t=1}^{N} \sum_{i=1}^{T} \sum_{i} \Sigma_{i}^{-1} e_{it} v_{it} + \frac{1}{NT} \sum_{t=1}^{N} \sum_{i=1}^{T} \sum_{i} \Sigma_{i}^{-1} \gamma_{i} \hat{h}_{i} e_{it}$$

$$- \frac{1}{NT} \sum_{t=1}^{N} \sum_{i=1}^{T} \sum_{i} \Sigma_{i}^{-1} \lambda_{i}^{j} \Pi_{i}^{-1} \frac{1}{N} \sum_{j=1}^{N} \lambda_{j}^{j} \Sigma_{j}^{-1} \gamma_{j} \hat{h}_{j} e_{it}$$

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

where

$$h_{i}^{*} = h_{i} - H'\bar{D}'(D'D)^{-1} d_{i} - H'M(D)G[G'M(D)G]^{-1}(g_{t} - G'D(D'D)^{-1} d_{i});$$

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

$$\frac{1}{NT} \text{tr}[\hat{M} \Gamma_{q}^{h} H' \bar{M}(B)H \Gamma_{p}^{h}] + \frac{1}{N} \sum_{t=1}^{N} \sum_{i} \Sigma_{i}^{-1} \Sigma_{i}^{(p,q)}$$

where $b_{t} = (g_{t}', d_{t}')'$ and $B = (b_{1}, b_{2}, \ldots, b_{T})' = (G, D)$, a matrix of $T \times (r_{1} + r_{3})$ dimension; $\hat{M} = \Sigma_{ee}^{-1/2} M(\Sigma_{ee}^{-1/2} \Lambda) \Sigma_{ee}^{-1/2}; \Gamma_{p}^{h} = (\gamma_{1p}^{h}, \gamma_{2p}^{h}, \ldots, \gamma_{Np}^{h})'$; $\Pi_{i} = \frac{1}{N} \sum_{j=1}^{N} \lambda_{j} \Sigma_{i}^{-1} \lambda_{j}^{j}$.

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

$$\hat{\beta} - \beta = \left( \begin{array}{c} \text{tr}[\hat{M} X_{1} \mathcal{M}(B) X_{1}^{'}] \\ \vdots \\ \text{tr}[\hat{M} X_{K} \mathcal{M}(B) X_{K}^{'}] \end{array} \right)^{-1} \left( \begin{array}{c} \text{tr}[\hat{M} X_{1} \mathcal{M}(B)e'] \\ \vdots \\ \text{tr}[\hat{M} X_{K} \mathcal{M}(B)e'] \end{array} \right)$$

$$+ O_{p}(T^{-3/2}) + O_{p}(N^{-1/2}T^{-1/2}) + O_{p}(N^{-1/2}T^{-1}).$$

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

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

$$\sqrt{NT}(\hat{\beta} - \beta) \overset{d}{\to} N(0, R^{-1}),$$
where $\overline{R} = \lim_{N,T \to \infty} R$, and $\overline{R}$ can also be expressed as

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

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 $k$th 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 $k$th iteration. We update $\Gamma^{(k)}$ according to

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

where

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

and

$$
\begin{align*}
\frac{1}{T} \sum_{t=1}^{T} E(z_t f_t'|Z, \theta^{(k)}) &= (I_N \otimes B^{(k)}) M_{zz}(I_N \otimes B^{(k)})' (\Sigma_{zz}^{(k)})^{-1} \Gamma^{(k)}
\end{align*}
$$

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

$$
\begin{align*}
\Sigma_{\varepsilon\varepsilon}^{(k+1)} &= \text{Dg}\{ (I_N(K+1) - \Gamma^{(k+1)} \Gamma^{(k)}(\Sigma_{zz}^{(k)})^{-1} ) \\
&\quad \times (I_N \otimes B^{(k)}) M_{zz}(I_N \otimes B^{(k)}) \}' 
\end{align*}
$$

and

$$
\begin{align*}
\beta^{(k+1)} &= \left( \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{x}_{it}' (\Sigma_{iie}^{(k+1)})^{-1} \hat{x}_{it} ight)^{-1} \\
&\quad \times \left( \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{x}_{it}' (\Sigma_{iie}^{(k+1)})^{-1} (\hat{y}_{it} - \lambda_i^{(k+1)} f_t) \right),
\end{align*}
$$

with $\lambda_i^{(k+1)} = \lambda_i^{(k)} + \lambda_i^{(k+1)} - \lambda_i^{(k+1)} f_t$.
where \( f_t^{(k)} \) is the transpose of the \( t \)th row of
\[
\hat{F}^{(k)} = E(\hat{F} | Z, \theta^{(k)}) = \hat{Z}'(I_N \otimes B^{(k)})^{-1}(\Sigma_{\hat{Z}Z}^{(k)})^{-1} \Gamma^{(k)},
\]
where \( \hat{Z} = (\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_T) \) with \( \hat{z}_t = z_t - \frac{1}{T} \sum_{s=1}^{T} z_s \); \( Dg(\cdot) \) is the operator that sets the entries of its argument to zeros if the counterparts of \( E(\epsilon_t \epsilon'_t) \) are zeros.

Putting together, we obtain \( \theta^{(k+1)} = (\Gamma^{(k+1)}, \beta^{(k+1)}, \Sigma_{\epsilon \epsilon}^{(k+1)}) \). 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
\[
L \xrightarrow{DGP4, \text{i.i.d.}} (\chi^2_1 + \chi^2_2),
\]
where \( \epsilon_t \) is generated from \( N(\mu, \Sigma) \).

Putting together, we obtain \( \theta^{(k+1)} = (\Gamma^{(k+1)}, \beta^{(k+1)}, \Sigma_{\epsilon \epsilon}^{(k+1)}) \). 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
\[
\begin{align*}
y_{it} &= \alpha_i + x_{it1} \beta_1 + x_{it2} \beta_2 + \psi_i' g_t + \phi_i' h_t + \kappa_i' d_t + \epsilon_{it}, \\
x_{itk} &= \mu_{ik} + \gamma_{ik}^g g_t + \gamma_{ik}^h h_t + \gamma_{ik}^d d_t + v_{itk}, \quad k = 1, 2.
\end{align*}
\]
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:** \( \phi_i, \kappa_i, \gamma_{ik}^h \text{ and } \gamma_{ik}^d \) are fixed to zeros; \( \alpha_i, \mu_{ik}, \psi_i \) and \( g_t \) are generated from \( N(0, 1) \) and \( \gamma_{ik}^g = \gamma_i + N(0, 1) \).

- **DGP2:** \( \phi_i, \kappa_i \) and \( \gamma_{ik}^d \) are fixed to zeros; \( \alpha_i, \mu_{ik}, \psi_i, \gamma_{ik}^h, g_t \) and \( h_t \) are generated from \( N(0, 1) \) and \( \gamma_{ik}^g = \gamma_i + N(0, 1) \).

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

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

Using the method of writing (2.2), we can rewrite (6.1) as
\[
(I_N \otimes B)z_t = \mu + L \xi_t + \epsilon_t,
\]
where \( \xi_t = g_t \) for DGP1; \( \xi_t = (g_t, h_t)' \) for DGP2 and DGP3; \( \xi_t = (g_t, h_t, d_t)' \) for DGP4, and \( L \) is the corresponding loadings matrix. Let \( i_t \) be the \( t \)th row of \( L \). We generate the cross-sectional heteroscedasticity \( \Xi \), an \( N(K + 1) \times 1 \) vector, according to \( \Xi_i = \frac{1}{1 - \eta_i} i_t', i = 1, 2, \ldots, N(K + 1) \), where \( \eta_i \) is drawn from \( U[0, 1] \) with \( u = 0.1 \). A similar way of generating heteroscedasticity is also used in [14] and [16]. Let \( \Upsilon = \text{diag}(\Upsilon_1, \Upsilon_2, \ldots, \Upsilon_N) \) be an \( N(K + 1) \times N(K + 1) \) block diagonal matrix, in which \( \Upsilon_i = \text{diag}\{1, (M_i M_i^{-1/2})M_i^{-1/2} \} \) with \( M_i \) being a \( K \times K \) standard normal random matrix for each \( i \). Once \( \Upsilon \) is generated, the error term \( \epsilon_t \), which is defined as \( (\epsilon_{1t}, \epsilon_{2t}, \ldots, \epsilon_{Nt})' \) with \( \epsilon_{it} = (\epsilon_{it}, v_{it1}, v_{it2})' \), is calculated by \( \epsilon_t = \sqrt{\text{diag}(\Xi)\Upsilon \epsilon_t} \), where \( \epsilon_t \) is an \( N(K + 1) \times 1 \) vector with all its elements being i.i.d. \( (\chi^2_2 - 2)/2 \), where \( \chi^2_2 \) 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 $\varepsilon_t$ is obtained, we use

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

to yield the observable data.

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

$$(6.2) \quad \hat{r} = \arg\min_{0 \leq m \leq r_{\max}} IC(m)$$

with

$$IC(m) = \frac{1}{NK} \ln |\hat{\Gamma}^m \hat{\Gamma}^{m'} + \hat{\Sigma}_{\varepsilon\varepsilon}^m| + m \frac{NK + T}{NK} \ln(\min(NK, T)),$$

where $\hat{\Gamma}^m$ and $\hat{\Sigma}_{\varepsilon\varepsilon}^m$ are the respective estimators of $\Gamma$ 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 $\hat{\beta}^r$ by the method of the basic model under $\hat{r}$. Then we calculate the matrix $\mathcal{R} = (\mathcal{R}_{ii})$ with $\mathcal{R}_{ii} = \hat{y}_{it} - \hat{x}_{it} \hat{\beta}^r$ and use the information criterion proposed by [12] to determine the factor number in $\mathcal{R}$, 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 $\phi_t$ and that of $d_t$ because $\phi_t$ 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              |                |
|-----|-----|-----|---------------|------------------|------------------|------------------|------------------|------------------|------------------|
|     |     |     |               | $\beta_1$       | $\beta_2$       | $\beta_1$       | $\beta_2$       | $\beta_1$       | $\beta_2$       |
|     |     |     |               | Bias | RMSE | Bias | RMSE | 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              |                |
|-----|-----|-----|---------------|------------------|------------------|------------------|------------------|------------------|------------------|
|     |     |     |               | $\beta_1$       | $\beta_2$       | $\beta_1$       | $\beta_2$       | $\beta_1$       | $\beta_2$       |
|     |     |     |               | Bias | RMSE | Bias | RMSE | 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 pattern 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 quasi-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