Likelihood approach to dynamic panel models with interactive effects

Jushan Bai*

October, 2013

Abstract

This paper considers dynamic panel models with a factor error structure that is correlated with the regressors. A large number of incidental parameters exist under the model. Both short panels (small T) and long panels (large T) are considered. Consistent estimation under a small T requires either a suitable formulation of the reduced form or an appropriate conditional equation for the first observation. A dynamic panel forms a simultaneous-equation system, and under the factor error structure, there exist constraints between the mean and the covariance matrix. We explore the constraints through a quasi-FIML approach, which does not estimate individual effects, even if they are fixed constants.

The factor process is treated as parameters and it can have arbitrary dynamics under both fixed and large T. The large T setting involves further incidental parameters because the number of parameters (including the time effects, the factor process, the heteroskedasticity parameters) increases with T. Even though an increasing number of parameters are estimated, we show that there is no incidental parameters bias to affect the limiting distributions; the estimator is centered at zero even scaled by the fast convergence rate of root-NT. We also show that the quasi-FIML approach is efficient under both fixed and large T, despite non-normality, heteroskedasticity, and incidental parameters. Finally, we develop a feasible and fast algorithm for computing the quasi-FIML estimators under interactive effects.

Key words and phrases: factor structure, interactive effects, incidental parameters, predetermined regressors, heterogeneity and endogeneity, quasi-FIML, efficiency.

*An earlier version of this paper has been circulated under the title "Likelihood approach to small T dynamic panel models with interactive effects." I thank seminar participants at Princeton University, MIT/Harvard, Boston University (at the Distinguished Visitor's Workshop), Triangle Area Econometrics workshops, Rochester University and University of Maryland for helpful comments. This paper was also presented at the Summer 2009 Cowles Foundation Econometrics Conference, and the 15th International Conference on Panel Data, Bonn, Germany. Partial results were also presented at the 10th World Congress of the Econometric Society, Shanghai, China, and at the African Econometrics Society, Nairobi, Kenya and also at the Pre-Conference of AMES at the Seoul National University, 2011. Financial support from the NSF grants SES 0551275 and SES-0962410 is acknowledged.

1 Introduction

In this paper we consider consistent and efficient estimation of dynamic panel data models with a factor error structure that is correlated with the regressors

$$y_{it} = \alpha\, y_{it-1} + x_{it}'\beta + \delta_t + \lambda_i' f_t + \varepsilon_{it}$$

$$i = 1, 2, ..., N; t = 1, 2, ..., T$$

where y_{it} is the dependent variable, and x_{it} $(p \times 1)$ is the regressor, β $(p \times 1)$ is the unknown coefficient, λ_i and f_t are each $r \times 1$ and both are unobservable, δ_t is the time effect, and ε_{it} is the error term.

The model considered here has its roots in both micro and macro econometrics. In microeconometrics, for example, the observed wage is a function of observable variables (x_{it}) and unobserved innate ability (λ_i). The innate ability is potentially correlated with the observed individual characteristics, and the effect of innate ability on wages is not constant over time, but time varying. In macroeconometrics, f_t is a vector of common shocks, and they have heterogeneous effects on each cross-sectional unit via the individual-specific coefficient λ_i. In finance, f_t represents a vector of systematic risks and λ_i is the exposure to the risks; asset return y_{it} is affected by both observable and nonobservable factors. Each motivation gives rise to a factor error structure that is correlated with the regressors. If $f_t \equiv 1$, then $\delta_t + \lambda_i' f_t = \delta_t + \lambda_i$, we have the additive effects model. An additive effects model does not allow multiple individual effects. Under interactive effects, wages can be affected by multiple unobservable individual traits such as motivation, dedication, and perseverance in the earnings study, and more than one common shock in the macroeconomic setting.

For the general case to be considered, we allow arbitrary correlation between x_{it} and $(\delta_t, \lambda_i, f_t)$. For panel data, T is usually much smaller than N. It is desirable to treat f_t as parameters. As such, f_t itself can have arbitrary dynamics (stationary or nonstationary process). Also, f_t can be a sequence of fixed constants such as linear or broken trend. We do not make any distributional assumptions on λ_i, they are not required to be i.i.d. In fact, λ_i can be a sequence of non-random constants. Even in the latter case, we do not estimate individual λ_i. The approach in this paper is that we only need to estimate their sample covariance matrix, which is of fixed dimension. This removes one source of incidental parameters problem.

We use data in levels as opposed to data in differences. Differencing the data tends to remove useful information. For dynamic panel, differencing also leads to lagged error terms, which are correlated with the lagged dependent variables. Under interactive effects, simple differencing is ineffective in removing the individual effects. Quasi-differencing as in Holtz-Eakin et al. (1988) introduces nonlinear transformation of parameters and the original parameters need to be recovered from the transformation. Also, the transformation becomes less tractable with more than one factor. We set up the problem as a simultaneous equations system with T equations and use the FIML approach to estimate the model.

Two sources of cross-sectional correlation are allowed in the model. One apparent source of correlation is the sharing of the common factors f_t by each cross-sectional unit. The other is implicit. The λ_i's can be correlated over i; permitting cross-sectional dependence through individual effects. This makes the analysis of FIML more challenging, but allows model's wider applicability.

For the case of a single factor ($r = 1$), Holtz-Eakin et al. (1988) suggest the quasi-difference approach to purge the factor structure, and use GMM to consistently estimate the model parameters. Ahn et al. (2001) also consider the quasi-difference approach and GMM estimation. Ahn et al. (2013) further generalize the method to the case of $r > 1$. These methods are consistent under fixed T. With a moderate T, the number of moments can be large and increases rapidly as T increases (order of $O(T^2)$). The likelihood approach considered here implicitly makes use of efficient combinations of a large number of moments, and it also effectively explores many of the restrictions implied by the model. While it has long been used, the likelihood approach to dynamic panel models has been emphasized more recently by Alveraz and Arellano (2003, 2004), Chamberlain and Moreira (2009), Kruiniger (2008), Moreira (2009), and Sims (2000).

Pesaran (2006) suggests adding the cross-sectional averages of the dependent and independent variables as additional regressors. The idea is that these averages provide an estimate for f_t. The limitation of the Pesaran method is discussed by Westerlund and Urbain (2013). Bai (2009), Moon and Weidner (2010a,b), Su and Chen (2013), and Su et al. (2013) treat both λ_i and f_t as parameters. While the latter estimator is consistent for (α, β) under large N and large T, these authors show that the estimator has bias, due to the incidental parameters problem and heteroskedasticity. The FIML approach in this paper does not estimate individual λ_is even if they are fixed effects.

The FIML approach treats the dynamic panel as a simultaneous equations system with T equations. We provide a careful treatment of the initial observation, as it is key to consistent estimation with a small T.[1] We consider two quasi-FIML formulations with respect to the first observation. One is the reduced-form formulation and the other is the conditional formulation. The argument for conditioning on y_{i0} is different from the time series analysis as y_{i0} is also correlated with the individual effects. Another notable feature of the model, as previously mentioned, is the existence of correlations between the effects λ_i and the regressors. We use the methods of Mundlak (1978) and Chamberlain (1982) to control for this correlation.

The FIML procedure also simultaneously estimates heteroskedasticities $(\sigma_1^2, \sigma_2^2, ..., \sigma_T^2)$, where $\sigma_t^2 = E(\varepsilon_{it}^2)$. A changing variance over time is an important empirical fact and the estimates for σ_t^2 are of economic interest (Moffitt and Gottschalk, 2002). Another important consideration is that if heteroskedasticity exists, but is not allowed in the estimation for dynamic model, the estimated parameters are inconsistent under fixed T. This important fact motivates the work of Alvarez and Arellano (2004). Allowing heteroskedasticity is not a matter of efficiency as researchers are accustomed to, but a matter of consistency for dynamic panel models. We demonstrate that allowing heteroskedasticity does not lose asymptotic efficiency under large T even if there is no

[1] For example, Anderson and Hsiao (1981,1982), Blundell and Bond (1998) and Blundell and Smith (1991).

heteroskedasticity.

Under fixed T, once the estimation problem is properly formulated under the quasi-FIML approach, consistency and asymptotically normality for the quasi-FIML estimator follow from existing theory for extremum estimation. Difficulty arises when T is also large because of the incidental parameters problem. The large T setting is practically relevant as many panel data sets nowdays have nonsmall T. Also, large T analysis provides a guidance for small T setting. One of the challenges is the consistency argument, which is nonstandard under an increasing number of parameters. Another difficulty lies in that, even scaled by the fast convergence rate $\sqrt{NT}$, we aim to demonstrate that the limiting distribution is centered at zero, and there are no asymptotic biases. We further aim to show that the quasi-FIML is asymptotically efficient despite the incidental parameters problem. We use the large dimensional factor analytical perspective to shed lights on these problems. This perspective has been used by Bai (2013) in the analysis of additive effects models, in which $f_t = 1$ is known and not estimated. The interactive-effect model in this paper has a non-degenerate factor structure, and allows multiple effects with $r > 1$. The analysis is very demanding, but the final result is simple and intuitive.

This paper also provides a feasible algorithm to compute the FIML estimators. Considerable amount of efforts have been devoted to the algorithm, which produces stable and quick estimates. In our simulated data, it takes a fraction of a second to produce the FIML estimator. Finite sample property of the estimator is documented by Monte Carlo simulations.

We describe, in Section 2, dynamic panels with strictly exogenous regressors x_{it}, either correlated or uncorrelated with the effects (λ_i, f_t). We consider two likelihood functions, joint or conditional with respect to the first observation. Section 3 considers predetermined regressors, either weakly exogenous or non-weakly exogenous regressors. Section 4 provides the inferential theory, consistency and the limiting distribution. We also establish some optimality property of the FIML estimator under both fixed and large T. Section 5 describes a feasible and fast computing algorithm. Simulation results are reported in Section 6, and the last section concludes. Technical proofs are provided in the appendix, and additional proofs are given in a supplementary document.

2 Dynamic panel with strictly exogenous regressors

We consider the following dynamic panel model with $T + 1$ observations

$$y_{it} = \delta_t + \alpha y_{i,t-1} + x_{it}'\beta + f_t'\lambda_i + \varepsilon_{it} \tag{1}$$
$$t = 0, 1, 2, ..., T; \quad i = 1, 2, ..., N$$

Strict exogeneity of x_{it} with respect to ε_{it} means

$$E(\varepsilon_{it}|x_{i1}, ..., x_{iT}, \lambda_i) = 0,$$

so that x_{is} is uncorrelated with ε_{it} but x_{it} is allowed to be correlated with λ_i or f_t, or both.

We do not require f_t to have zero mean. We treat f_t as parameters so that it can have arbitrary dynamics, either deterministic or random. For example, f_t can be a linear trend or broken trend, appropriately normalized so that $\frac{1}{T}\sum_{t=1}^{T} f_t f_t'$ converges to a positive matrix (e.g., in case of linear trend, f_t represents t/T).

The stability condition of $|\alpha| < 1$ is maintained throughout, while stationarity of the model is not assumed. In particular, the first observation y_{i0} is not necessarily originated from a stationary distribution. We follow Bhargava and Sargan (1983) to view the model as a simultaneous equations system with $T+1$ equations. For dynamic panel data, modeling the first observation is crucial for consistent estimation. Different assumptions on the initial conditions give rise to different likelihood functions, see Hsiao (2003, Chapter 4), although the impact of the initial condition diminishes to zero as T goes to infinity. When α is close to 1, the first observation is still important for large T.

Throughout the paper, we use the following notation:

$$y_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{bmatrix}, \quad x_i = \begin{bmatrix} x_{i1}' \\ \vdots \\ x_{iT}' \end{bmatrix}, \quad \delta = \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_T \end{bmatrix}, \quad F = \begin{bmatrix} f_1' \\ \vdots \\ f_T' \end{bmatrix}, \quad \varepsilon_i = \begin{bmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{iT} \end{bmatrix} \tag{2}$$

2.1 Regressors uncorrelated with the effects

When the effects are uncorrelated with the exogenous regressors (still correlated with the lagged dependent variables), it is easier to motivate and formulate the likelihood function by assuming that λ_i are random variables and are independent of x_i. First note that in the presence of time effects δ_t, it is without loss of generality to assume $E(\lambda_i) = 0$. If $\mu = E(\lambda_i) \neq 0$, we can write $\lambda_i' f_t = (\lambda_i - \mu)' f_t + \mu' f_t$, and we can absorb $\mu' f_t$ into δ_t.

The initial observation y_{i0} may or may not follow the dynamic process (1). In either case, y_{i0} requires a special consideration for dynamic models. Write the reduced form for y_{i0}, similar to Bhargava and Sargan (1983):

$$y_{i0} = \delta_0^* + \sum_{s=1}^{T} x_{i,s}' \psi_{0,s} + f_0^{*\prime} \lambda_i + \varepsilon_{i,0}^* = \delta_0^* + w_i' \psi_0 + f_0^{*\prime} \lambda_i + \varepsilon_{i,0}^*$$

where[2]

$$w_i = \text{vec}(x_i'), \quad \psi_0 = (\psi_{0,1}', ..., \psi_{0,T}')'$$

In general, we regard the reduced form as a projection of y_{i0} on $[1, w_i, \lambda_i]$. It will not affect the analysis by removing the asterisk from $(\delta_0^*, f_0^*, \varepsilon_{i0}^*)$; [3] the asterisk indicates that these variables are different from $(\delta_0, f_0, \varepsilon_{i0})$ that appears in the y_{i0} equation should y_{i0} also follow (1). For $t \geq 1$,

$$y_{it} = \alpha y_{i,t-1} + \delta_t + x_{it}' \beta + f_t' \lambda_i + \varepsilon_{it}, \quad t = 1, 2, ..., T$$

[2] If x_{i0} is observable, we should also include x_{i0} as a predictor.

[3] This is because we treat δ_t and f_t (for all t) as free parameters; δ_0^* and f_0^* are also free parameters, thus can be denoted as (δ_0, f_0). Similarly, we allow ε_{it} to be heteroskedastic we can use ε_{i0} for ε_{it}^* (or σ_0^2 for σ_0^{*2}).

Again, let $x_i = (x_{i1}, ..., x_{iT})'$. Since $x_i\beta = (I_T \otimes \beta')\text{vec}(x_i') = (I_T \otimes \beta')w_i$, the system of $T+1$ equations can be written as

$$B^+ y_i^+ = Cw_i + \delta^+ + F^+\lambda_i + \varepsilon_i^+ \tag{3}$$

where

$$y_i^+ = \begin{bmatrix} y_{i0} \\ y_i \end{bmatrix}, \quad \delta^+ = \begin{bmatrix} \delta_0^* \\ \delta \end{bmatrix}, \quad F^+ = \begin{bmatrix} f_0'^* \\ F \end{bmatrix}, \quad \varepsilon_i^+ = \begin{bmatrix} \varepsilon_{i0}^* \\ \varepsilon_i \end{bmatrix}$$

$$B^+ = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ -\alpha & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & -\alpha & 1 \end{bmatrix}, \quad C = \begin{bmatrix} \psi_0' \\ I_T \otimes \beta' \end{bmatrix} \tag{4}$$

and y_i, δ, F and ε_i are defined in (2). We normalize the first $r \times r$ block of the factors as an identity matrix, $F^+ = (I_r, F_2')'$ to remove the rotational indeterminacy. Introduce

$$\Omega^+ = F^+\Psi_\lambda F^{+\prime} + D^+$$

where $\Psi_\lambda = E(\lambda_i\lambda_i')$, and $D^+ = E(\varepsilon_i^+\varepsilon_i^{+\prime}) = \text{diag}(\sigma_0^{*2}, \sigma_1^2, ..., \sigma_T^2)$, and let

$$u_i^+ = B^+ y_i^+ - Cw_i - \delta^+$$

the quasi log-likelihood function for $(y_{i0}, y_{i1}, ..., y_{iT})$, conditional on w_i, is

$$-\frac{N}{2}\ln|\Omega^+| - \frac{1}{2}\sum_{i=1}^{N} u_i^{+\prime}(\Omega^+)^{-1}u_i^+ \tag{5}$$

Because the determinant of B^+ is equal to 1 the Jacobian term does not enter. With the factor structure, assuming D^+ is diagonal (ε_{it} are uncorrelated over t), the model is identifiable if $T \geq 2r+1$.

This likelihood function generalizes the classical likelihood function to include interactive effects. Anderson and Hsiao (1982), Bhargava and Sargan (1983), and Arellano and Alveraz (2004) are examples of classical likelihood analysis. The first two papers assume homoskedasticity.

Remark 1 Although the likelihood function is motivated by assuming λ_i being random variables, it is still valid when λ_i is a sequence of constants. In this case, we interpret Ψ_λ as $\Psi_n = \frac{1}{n}\sum_{i=1}^N(\lambda_i - \bar{\lambda})(\lambda_i - \bar{\lambda})'$, which is the sample variance of λ_i ($n = N-1$). This is a matter of recentering and rescaling the parameters $(\delta, F, \Psi_\lambda)$. To see this, if we concentrate out δ^+ from (5), the likelihood function involves $\dot{u}_i^+ = F^+\dot{\lambda}_i + \dot{\varepsilon}_i^+$, where $\dot{\lambda}_i = \lambda_i - \bar{\lambda}$ and $\dot{\varepsilon}_i^+ = \varepsilon_i^+ - \bar{\varepsilon}^+$. Then the expected value $E(\frac{1}{n}\sum_{i=1}^N \dot{u}_i^+\dot{u}_i^{+\prime}) = F^+\Psi_n F^{+\prime} + D^+$, where the expectation is taken assuming that λ_i are fixed constants. We then interpret the likelihood function as a distance measure between $\frac{1}{n}\sum_{i=1}^N \dot{u}_i^+\dot{u}_i^{+\prime}$ and $F^+\Psi_n F^{+\prime} + D^+$ (other distance can also be used). Note that the recentering does not affect

the key parameters (α, β, D^+). Recentering is used in classical factor analysis for non-random λ_i, but it works regardless of λ_i being random, see Amemiya et al. (1987) and Dahm and Fuller (1986). Recentering simplifies the statistical analysis and permits weaker conditions for asymptotic theory (Anderson and Amemiya, 1988). Also, the resulting likelihood function can be motivated from a decision theoretical framework (Chamberlain and Moreira, 2009, and Moreira, 2009) with an appropriate choice of prior information and loss function. □

The rest of the paper considers the general situation in which x_{it} is correlated with λ_i or f_t or both. This correlation is fundamental for panel data econometrics.

2.2 Regressors correlated with the effects

Projecting λ_i on $w_i = \text{vec}(x_i')$,

$$\lambda_i = \lambda + \phi_1 \, x_{i1} + \cdots + \phi_T \, x_{iT} + \eta_i \tag{6}$$

or write it more compactly as

$$\lambda_i = \lambda + \phi \, w_i + \eta_i$$

where λ is the intercept and η_i is the projection residual, and ϕ_i are matrices $(r \times p)$ of projection coefficients. This is known as the Mundlak-Chamberlain projection (Mundlak, 1978, and Chamberlain, 1982). By definition, $E(\eta_i) = 0$ and $E(x_{it}\eta_i) = 0$ for all t. This means that the factor errors $F\eta_i + \varepsilon_i$ will be uncorrelated with the regressors x_i.

Substitute the preceding projection into (1) and absorb $f_t'\lambda$ into δ_t, we have, for $t \geq 1$,

$$y_{it} = \alpha y_{it-1} + x_{it}'\beta + f_t'\phi \, w_i + \delta_t + f_t'\eta_i + \varepsilon_{it}, \quad t \geq 1.$$

The y_{i0} equation has the same form (by renaming the parameters since all are free parameters). That is, we can write y_{i0} as

$$y_{i0} = \delta_0^* + w_i'\psi_0 + f_0^{*\prime}\eta_i + \varepsilon_{i,0}^*$$

Stacking these equations, the model has the same form as (2.1), namely

$$B^+ y_i^+ = C w_i + \delta^+ + F^+ \eta_i + \varepsilon_i^+$$

but here

$$C = \begin{bmatrix} \psi_0' \\ I_T \otimes \beta' + F\phi \end{bmatrix}$$

The likelihood function for the $(T+1)$ simultaneous equations system has the same form as (5), with $\Omega^+ = F\Psi_\eta F' + D$, where we replace Ψ_λ by $\Psi_\eta = E(\eta_i\eta_i')$.

A special case of the Mundlak-Chamberlain projection is to assume $\phi_1 = \phi_2 = \cdots = \phi_T$. We then write the projection as $\lambda_i = \lambda + \phi \, \bar{x}_i + \eta_i$ with $\bar{x}_i = \frac{1}{T}\sum_{s=1}^T x_{is}$. Because $f_t'\phi \, \bar{x}_i = f_t'\phi \frac{1}{T} x_i' \iota_T =$

$(\iota_T' \otimes f_t'\phi\frac{1}{T})w_i$ with $w_i = \text{vec}(x_i')$ and $\iota_T = (1, 1, ..., 1)'$, the model is the same as above, but the coefficient matrix C becomes

$$C = \begin{bmatrix} \psi_0' \\ I_T \otimes \beta' + \iota_T' \otimes F\phi\frac{1}{T} \end{bmatrix}$$

Here we use the same notation ϕ, but its dimension is different in the special case. In general, this restricted projection may not lead to consistent estimation.

If we interpret the linear projection (6) as the conditional mean, the likelihood function can be interpreted as the conditional likelihood function, conditional on the regressors. But the quasi-FIML, which is based on the second moments of the data, still works under the projection interpretation.

Remark 2 When the projection (6) is considered as the population projection, the projection error η_i is uncorrelated with the predictor x_{it}, and we have $E(\eta_i) = 0$ and $E(x_{it}\eta_i) = 0$ for each t. We can also consider η_i as the least squares residuals and $(\lambda, \phi_1, ..., \phi_T)$ as the least squares estimated coefficients (so they depend on N). As the least squares residuals, the η_i's satisfy $\sum_{i=1}^N \eta_i = 0$, and $\sum_{i=1}^N x_{it}\eta_i = 0$ for $t = 1, 2, ..., T$. The different interpretation is a matter of recentering the nuisance parameters $\lambda, \phi_1, ..., \phi_T$ (and also η_i). These parameters are not the parameters of interest. The estimator for the key parameters $(\alpha, \beta, \sigma_0^2, \sigma_1^2, ..., \sigma_T^2)$ is not affected by the recentering of the nuisance parameters. The least squares interpretation is useful when λ_i is (or is treated as) a sequence of fixed constants. In this case, we interpret Ψ_η as $\Psi_n = \frac{1}{n}\sum_{i=1}^N (\eta_i - \bar{\eta})(\eta_i - \bar{\eta})'$ (with $n = N - 1$), the sample covariance of η_i. Also see Remark 1. □

2.3 Likelihood conditional on y_{i0}

An alternative approach to the full likelihood for the entire sequence $(y_{i0}, y_{i1}, ..., y_{iT})$ is the conditional likelihood, conditional on the initial observation. The conditional likelihood is less sensitive to the specification of initial conditions. The analysis is different from the conditional estimation in the pure time series context owing to the presence of individual effects. Since λ_i can be correlated with y_{i0}, we project λ_i on y_{i0} in addition to w_i such that

$$\lambda_i = \lambda + \phi_0\, y_{i0} + \phi\, w_i + \eta_i \tag{7}$$

where η_i denotes the projection residual. The model can be written as $(t = 1, 2, ..., T)$

$$y_{it} = \alpha y_{it-1} + x_{it}'\beta + f_t\phi_0 y_{i0} + f_t'\phi w_i + \delta_t + f_t'\eta_i + \varepsilon_{it}$$

In matrix form,

$$By_i = \alpha y_{i0}e_1 + x_i\beta + F\phi_0 y_{i0} + F\phi w_i + \delta + F\eta_i + \varepsilon_i$$

where B is equal to B^+ with the first row and first column deleted, and $e_1 = (1, 0, ..., 0)'$. Since the determinant of B is 1, the likelihood for $F\eta_i + \varepsilon_i$ is the same as the likelihood for y_i conditional on y_{i0} and x_i. Thus

$$\ell(y_i|y_{i0}, x_i) = -\frac{N}{2}\ln|\Omega| - \frac{1}{2}\sum_{i=1}^N u_i'\Omega^{-1}u_i \tag{8}$$

where $\Omega = F\Psi F' + D$ with $\Psi = \text{var}(\eta_i)$,

$$u_i = (u_{i1}, ..., u_{iT})'$$

with

$$u_{it} = y_{it} - \alpha y_{it-1} - x_{it}'\beta - f_t'\phi_0\, y_{i0} - f_t'\phi\, w_i - \delta_t.$$

The first observation y_{i0} appears in every equation (every t).

The conditional likelihood here is robust to assumptions made on the initial observations y_{i0}, whether it is from the stationary distribution, whether it is correlated with regressors or the effects. Again, λ_i can be a sequence of fixed constants as noted in the earlier remarks.

3 Dynamic panel with predetermined regressors

This section considers the model

$$y_{it} = \alpha\, y_{i,t-1} + x_{it}'\beta + f_t'\lambda_i + \varepsilon_{it} \tag{9}$$

under the assumption that

$$E(\varepsilon_{it}|y_i^{t-1}, x_i^t, \lambda_i) = 0$$

where $y_i^t = (y_{i0}, ..., y_{it})'$ and $x_i^t = (x_{i1}, ..., x_{it})'$. Under this assumption, x_{it} is allowed to be correlated with past ε_{it}, thus predetermined. This assumption also allows feedback from past y to current x. The model extends that of Arellano (2003, Chapter 8) to interactive effects and to the maximum likelihood estimation.

3.1 Weakly exogenous dynamic regressors

The concept of weak exogeneity is examined by Engle et al (1983). The basic idea is that inference for the parameter of interest can be performed conditional on weakly exogenous regressors without affecting efficiency. In this case, we show that the Mundlak-Chamberlain projection will not be necessary. The objective function given in (11) below is sufficient for consistent and efficient estimation. Under weak exogeneity the joint density for (y_{it}, x_{it}) (conditional on past data) can be written as the conditional density of y_{it}, (conditional on x_{it}) multiplied by the marginal density of x_{it} (all conditional on past data), where the latter is uninformative about the parameters of interest. To be concrete, we consider the following process

$$x_{it} = \alpha_x x_{i,t-1} + \beta_x y_{i,t-1} + g_t'\tau_i + \xi_{it} \tag{10}$$

where α_x $(p \times p)$ and β_x $(p \times 1)$ are unknown parameters (not necessarily the parameters of the interest). In addition, τ_i and λ_i are conditionally independent (conditional on the initial observation (y_{i0}, x_{i0})); ε_{it} is independent of ξ_{it}; f_t and g_t are free parameters. The regressor x_{it} is correlated

with past ε_{it}, thus predetermined; x_{it} is also correlated with λ_i and past f_t through $y_{i,t-1}$. Arbitrary correlation between λ_i and (x_{i0}, y_{i0}) (initial endowment) is also allowed.

Note that for the y equation, the regressor x_{it} is correlated with λ_i, even conditional on the included regressor y_{it-1}. This correlation originates from the correlation between x_{i0} and λ_i.

We next argue that x_{it} in (10) is weakly exogenous with respect to the parameters in the y equation. The part of joint density function[4] of (y_i, x_i) that involves the parameter of interest is given by

$$\ell_1(y_i, x_i | y_{i0}, x_{i0}) = -\frac{N}{2} \ln |\Omega^*| - \frac{1}{2} \sum_{i=1}^{N} u_i' \Omega^{*-1} u_i \tag{11}$$

where $\Omega^* = F\Psi^* F' + D$ with $\Psi^* = \text{var}(\eta_i^*)$ and $\eta_i^* = \lambda_i - \phi_0 y_{i0} - \psi_0 x_{i0}$,

$$y_i = (y_{i1}, y_{i2}, ..., y_{iT})', \quad x_i = (x_{i1}, x_{i2}, ..., x_{iT})', \quad u_i = (u_{i1}, ..., u_{iT})'$$

$$u_{it} = y_{it} - \alpha y_{it-1} - x_{it}'\beta - f_t'\phi_0 y_{i0} - f_t'\psi_0 x_{i0}$$

$(t = 1, 2, ..., T)$. The likelihood function is similar to that of Section 2.3, here the individual effects λ_i are projected onto the initial value of x_{i0} instead of the entire path $(x_{i0}, x_{i1}, ..., x_{iT})$. Again, the factor process F occurs in both the mean and variance.

The preceding likelihood function is simple. There is no need to estimate the parameters in the x equation, so the computation is relatively easy. The parameters can be easily estimated by the algorithm in Section 5.

To verify (11), let $w_i = \text{vec}(x_i')$, a vector that stacks up x_{it} $(t = 1, 2, ..., T)$. Then

$$\begin{bmatrix} B & -(I_T \otimes \beta') \\ C_1 & C_2 \end{bmatrix} \begin{bmatrix} y_i \\ w_i \end{bmatrix} = d_1 y_{i0} + d_2 x_{i0} + \begin{bmatrix} F\lambda_i + \varepsilon_i \\ G\tau_i + \xi_i \end{bmatrix}$$

where B has the same form as B^+ but with dimension $T \times T$, and

$$C_1 = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ -\beta_x & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & -\beta_x & 0 \end{bmatrix}, \quad C_2 = \begin{bmatrix} I_p & 0 & \cdots & 0 \\ -\alpha_x & I_p & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & -\alpha_x & I_p \end{bmatrix}, \quad d_1 = \begin{bmatrix} \alpha \\ 0 \\ \vdots \\ \beta_x \\ 0 \\ \vdots \end{bmatrix}, \quad d_2 = \begin{bmatrix} 0 \\ \vdots \\ \alpha_x \\ 0 \\ \vdots \end{bmatrix}$$

$G = (g_1, g_2, ..., g_T)'$ and $\xi_i = (\xi_{i1}, ..., x_{iT})'$. All elements of d_1 and d_2 are zero except those displayed.

It can be easily shown that

$$B^\dagger = \begin{bmatrix} B & -(I_T \otimes \beta') \\ C_1 & C_2 \end{bmatrix}$$

has a determinant equal to one, the joint density of (y_i, w_i) is equal to the joint density of $B^\dagger(y_i', w_i')'$. The latter is equal to, apart from a mean adjustment, the joint density of $((F\lambda_i + \varepsilon_i)', (G\tau_i + \xi_i)')'$,

[4] More specifically, the conditional joint density, conditional on the initial observation.

where all densities are conditional on the initial observation (y_{i0}, x_{i0}). Assuming λ_i and τ_i are conditionally independent (conditional on y_{i0} and x_{i0}), then $F\lambda_i + \varepsilon_i$ is conditionally independent of $G\tau_i + \xi_i$. Thus we have

$$f(y_i, x_i|y_{i0}, x_{i0}) = f(F\lambda_i + \varepsilon_i|y_{i0}, x_{i0}) \cdot f(G\tau_i + \xi_i|y_{i0}, x_{i0}) \tag{12}$$

where f denotes a density function. Equation (11) is equal to $\log f(F\lambda_i + \varepsilon_i|y_{i0}, x_{i0})$. The logarithm of the second term does not depend on the parameters of interest.

Remark 3 Equation (11) is neither the (log-valued) joint density of (y_i, x_i), nor the conditional density $f(y_i|x_i, y_{i0}, x_{i0})$. It is the term in the joint density that depends on the parameters of interest. When y does not Granger cause x (i.e., $\beta_x = 0$), then (11) is the conditional density. See Engle et al (1983). □

Remark 4 The likelihood function (11) is simpler than that of Section 2. The reason is that, under strict exogeneity of Section 2, the process of x_{it} is unspecified, and to account for the arbitrary correlation between the effects and the regressors, full path projection of λ_i on x_i is required. Under weak exogeneity together with a dynamically generated x_{it}, it is sufficient to account for the correlation between the effects (λ_i) and the initial observations y_{i0} and x_{i0} only. □

3.2 Non-weakly exogenous dynamic regressors

We consider a similar process for x_{it}. However, we now permit arbitrary correlation between λ_i and τ_i and arbitrary correlation between ε_{it} and ξ_{it}. Solely for notional simplicity, we assume τ_i and λ_i are identical. We also allow arbitrary correlation between f_t and g_t. We rewrite the y equation by lagging the x by one period (also for notational simplicity) so that

$$y_{it} = \alpha y_{i,t-1} + \beta' x_{i,t-1} + \delta_{yt} + f_t'\lambda_i + \varepsilon_{it}$$

and

$$x_{it} = \alpha_x x_{i,t-1} + \beta_x y_{i,t-1} + \delta_{xt} + g_t'\lambda_i + \xi_{it}$$

Because of the correlation between ε_{it} and ξ_{it}, and the common λ_i cross equations, the regressor x_{it} is no longer weakly exogenous, although predetermined with respect to $\{\varepsilon_{it}\}$. The x and y equations should be modeled jointly even though the parameters of interest are those in the y equation only. The VAR approach is most suitable for this setup. Let

$$z_{it} = \begin{bmatrix} y_{it} \\ x_{it} \end{bmatrix}, \quad A = \begin{bmatrix} \alpha & \beta' \\ \beta_x & \alpha_x \end{bmatrix}, \quad \delta_t = \begin{bmatrix} \delta_{yt} \\ \delta_{xt}, \end{bmatrix} \quad \pi_t' = \begin{bmatrix} f_t' \\ g_t' \end{bmatrix}, \quad \zeta_{it} = \begin{bmatrix} \varepsilon_{it} \\ \xi_{it} \end{bmatrix}$$

Then

$$z_{it} = Az_{it-1} + \delta_t + \pi_t'\lambda_i + \zeta_{it} \tag{13}$$

This formulation extends the model of Holtz-Eakin et al. (1988) to multiple factors. Let z_i be the $T(p+1)\times 1$ vector that stacks up z_{it} $(t=1,2,...,T)$ and Π be the $T(p+1)\times r$ matrix that stacks up the expanded factors π_t'. Under the assumption that ζ_{it} are independent normal over t, $N(0,\Sigma_t)$, the conditional likelihood function, conditional on z_{i0}, is given by

$$\ell(z_i|z_{i0}) = -\frac{N}{2}\ln|\Omega^*| - \frac{1}{2}\sum_{i=1}^{N} e_i'\Omega^{*-1}e_i$$

where $\Omega^* = \Pi\Psi^*\Pi' + \Sigma$ with $\Psi^* = \text{var}(\eta_i^*)$, with η_i^* being the projection residual in $\lambda_i = \lambda + \phi_0 z_{i0} + \eta_i^*$; Σ is block diagonal such that $\Sigma = \text{diag}(\Sigma_1,...,\Sigma_T)$, and $e_i = (e_{i1}', e_{i2}', ..., e_{iT}')'$ with

$$e_{it} = z_{it} - Az_{it-1} - \delta_t - \pi_t'\,\phi_0\,z_{i0}, \quad (t=1,2,...,T)$$

where δ_t absorbs $\pi_t'\lambda$ (λ is the intercept in the projection of λ_i onto $[1, z_{i0}]$). Note that z_{i0} appears as a regressor in every equation (i.e., each t) and appears twice for the first equation ($t=1$). The expanded factor matrix $\Pi = (\pi_1, \pi_2, ..., \pi_T)'$ appears in both the mean and variance. This conditional likelihood (conditional on z_{i0}) is the simplest, at least in form, among those discussed so far in this paper. The restricted aspect is that we need to model the x equation, in comparison with the weakly exogenous case. This will, of course, be desirable when the parameters of the x equations are also of interest.

In addition to the conditional likelihood (conditional on z_{i0}), the joint likelihood of $z_i^+ = (z_{i0}, z_{i1}, ..., z_{iT})'$ is easy to obtain. We obtain the reduced form for the first observation by projecting z_{i0} on $[1, \lambda_i]$ such that $z_{i0} = \delta_0 + \psi_0\lambda_i + \varepsilon_{i0}$. In either form, the Mundlak-Chamberlain projection is not required. The maximum likelihood estimation can be easily implemented by the algorithm in Section 5.

4 Inferential theory

4.1 Fixed T inferential theory

Despite the factor error structure, because we do not estimate individual heterogeneities (the factor loadings) but only their sample variance, this eliminates the incidental parameters problem. Under fixed T, there are only a fixed number of parameters so that the standard theory of the quasi-maximum likelihood applies. In particular, consistency and asymptotic normality hold. Let θ denote the vector of free and unknown parameters, that is, α, β, the lower triangular of Ψ (due to symmetry), the unknown elements in F, and the unknown elements in D. Let $\widehat{\theta}$ denote the quasi-FIML estimator. Standard theory implies the following result:

$$\sqrt{N}(\widehat{\theta} - \theta) \xrightarrow{d} N(0, V)$$

where

$$V = \text{plim}\, N\left(\frac{\partial^2\ell}{\partial\theta\partial\theta'}\right)^{-1}\left(E\frac{\partial\ell}{\partial\theta}\frac{\partial\ell}{\partial\theta'}\right)\left(\frac{\partial^2\ell}{\partial\theta\partial\theta'}\right)^{-1}$$

and the derivatives are evaluated at θ^0. So the estimator is consistent and asymptotically normal under fixed T. This result contrasts with the within-group estimator under fixed T (for additive effects) or the principal components estimator (for interactive effects). The latter estimators can be inconsistent under fixed T. Despite the sandwich formula for the covariance, we argue that the estimator is efficient in a later subsection.

4.2 Large T inferential theory

The large T analysis is quite different and is enormously difficult, though the final limiting results are simple. There are an infinite number of parameters in the limit. The usual argument of consistency as in Amemiya (1985) and Newey and McFadden (1994) no longer applies. The incidental parameters problem occurs because of time effects δ (T parameters), the factor process F ($T \times r$ parameters), and heteroskedasticity D (T parameters). We examine how the incidental parameters problem in the T dimension affects the limiting behavior of the estimators. One interesting question is whether higher order biases exist. Existing theory on incidental parameter problem, e.g., Neyman and Scott (1948), Nickell (1981), Kiviet (1995), Lancaster (2000, 2002), and Alvarez and Arellano (2003), suggests potential biases.

We consider the case without additional regressors other than the lag of the dependent variable:

$$y_{it} = \alpha\, y_{it-1} + \delta_t + \lambda_i' f_t + \varepsilon_{it}$$

The theory to be developed is applicable for the vector autoregressive model (13), in which α is replaced by matrix A, assuming that the eigenvalues of A are less than unity in absolute values.

Under large T, we shall assume $y_{i0} = 0$ for notational simplicity. A single observation will not affect the consistency and the limiting distribution under large T. Writing in vector-matrix notation

$$By_i = \delta + F\lambda_i + \varepsilon_i$$

where B ($T \times T$) has the form of B^+ in (4); y_i, δ, F, and ε_i are defined in (2). Rewrite the above as

$$y_i = \Gamma\delta + \Gamma F\lambda_i + \Gamma\varepsilon_i$$

where

$$\Gamma = B^{-1} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \alpha & 1 & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \alpha^{T-1} & \cdots & \alpha & 1 \end{bmatrix}$$

The idiosyncratic error $\Gamma\varepsilon_i$ has covariance matrix $\Gamma D\Gamma'$, which is not diagonal. Since $\Gamma\delta$ is a vector of free parameter, the MLE of $\Gamma\delta$ is equal to the sample mean $\bar{y} = \frac{1}{N}\sum_{i=1}^N y_i$. Let $S_n = \frac{1}{n}\sum_{i=1}^N (y_i - \bar{y})(y_i - \bar{y})'$ be the sample variance of y_i, and let

$$\Psi_n = \frac{1}{n}\sum_{i=1}^N (\lambda_i - \bar{\lambda})(\lambda_i - \bar{\lambda})'$$

be the sample variance of λ_i, with $n = N-1$. We consider the fixed effects setup so that λ_i and Ψ_n are nonrandom. Despite the fixed effects setup, we do not estimate the individual λ_i but only its sample covariance matrix Ψ_n. It is common in the factor literature to estimate the sample moments of the effects, whether they are random or deterministic, see, for instance, Amemiya et al. (1987), Anderson and Amemiya (1988). Estimating the sample moment instead of the effects themselves eliminates the incidental parameters problem in the cross-section. But under large T, we have new incidental parameters due to the increasing dimension of δ, F and D. Taking expectation, we obtain

$$E(S_n) = \Sigma(\theta) = \Gamma F \Psi_n F' \Gamma' + \Gamma D \Gamma' = \Gamma \Omega \Gamma'$$

where θ denotes the parameter vector consisting of α, the non-repetitive and free elements of Ψ_n, the free elements of F, and the diagonal elements of D. It is convenient to simply put $\theta = (\alpha, \Psi_n, F, D)$. In the above, $\Omega = \Omega(\theta) = F\Psi_n F' + D$.

The likelihood function after concentrating out δ becomes

$$\ell(\theta) = -\frac{n}{2} \log |\Sigma(\theta)| - \frac{n}{2} \mathrm{tr}[S_n \Sigma(\theta)^{-1}] \tag{14}$$

where $n = N - 1$. We make the following assumptions:

Assumption 1: ε_i are iid over i; $E(\varepsilon_{it}) = 0$, $\mathrm{var}(\varepsilon_{it}) = \sigma_t^2 > 0$, and ε_{it} are independent over t; $E\varepsilon_{it}^4 \leq M < \infty$ for all i and t; $E(\varepsilon_{it}|\lambda_1, ..., \lambda_N, F) = 0$, for $t \geq 1$; $|\alpha| < 1$.

Assumption 2: The λ_i are either random or fixed constants with $\Psi_n \to \Psi > 0$, as $N \to \infty$.

Assumption 3: There exist constants a and b such that $0 < a < \sigma_t^2 < b < \infty$ for all t; $\frac{1}{T} F' D^{-1} F = \frac{1}{T} \sum_{t=1}^T \sigma_t^{-2} f_t f_t' \to Q$ and $\frac{1}{T} \sum_{t=1}^T \sigma_t^{-4} (f_t f_t' \otimes f_t f_t') \to \Xi$, as $T \to \infty$, for some positive definite matrices Q and Ξ.

As explained in the previous sections, we need r^2 restrictions to remove the rotational indeterminacy for factor models. We consider two different sets of restrictions, referred to as IC1 and IC2. They are stated below:

IC1: Ψ_n is unrestricted, $F = (I_r, F_2')'$

IC2: Ψ_n is diagonal, and $T^{-1} F' D^{-1} F = I_r$.

Remark 5 A variation to IC2 is $\Psi_n = I_r$ and $\frac{1}{T} F' D^{-1} F = I_r$. IC2 or its variation is often used in the classical maximum likelihood estimation of pure factor models (e.g., Anderson and Rubin, 1956; Lawley and Maxwell, 1971). Whether IC1 or IC2 (or its variation) is used, the estimated parameters α and σ_t^2 $(t = 1, 2, ..., T)$ are numerically identical. □

Remark 6 Similar to classical factor analysis (e.g., Lawley and Maxwell, 1971), If IC1 or IC2 holds for the underlying parameters, we will be able to estimate the true parameters (F, Ψ_n) instead of rotations of them. If IC1 and IC2 are merely considered as a device to uniquely determine the estimates, then the estimated $\widehat{F}$ is a rotation of the true F, and $\widehat{\Psi}$ is a rotation of true Ψ_n. In this paper, we regard the restrictions hold for the true parameters so we are directly estimating

the true F and true Ψ_n without rotations. This interpretation in fact makes the analysis more challenging because we need to show that the rotation matrix is an identity matrix. Under either interpretation of the restrictions, the estimated parameters α and $(\sigma_1^2, ..., \sigma_T^2)$ are identical. □

If we let Ψ_n^0 denote the true sample variance of λ_i, it is a maintained assumption that $\Psi_n^0 > 0$ (positive definite). As a variable (an argument) of the likelihood function, Ψ_n is only required to be semi-positive definite. Assuming the diagonal matrix D is invertible, then $\Sigma(\theta)^{-1}$ exists provided that $\Psi_n \geq 0$.

The likelihood function is nonlinear. Like any nonlinear maximization, we need some restrictions on the parameter space. For technical reasons, we shall assume the maximization with respect to σ_t^2 (for all t) is taken over the set $[a, b]$ with a and b positive (though arbitrary), such that $\sigma_t^2 \in (a, b)$. We assume a stable dynamic process, that is, $\alpha \in [-\bar{\alpha}, \bar{\alpha}]$, a compact subset of $(-1, 1)$. We put no restrictions on F and Ψ_n other than the normalization restrictions. The consistency theory does require that the determinant $|I_r + F'D^{-1}F\Psi_n|$ be bounded by $O_p(T^k)$ for some $k \geq 1$. But this imposes essentially no restriction since k is arbitrarily given. Indeed, in actual computation, no restriction is imposed other than the normalization restrictions.

Let Θ denote the parameter space as just described. That is, $\alpha \in [-\bar{\alpha}, \bar{\alpha}]$ a compact subset of $(-1, 1)$; $\sigma_t^2 \in [a, b]$ for each t, Ψ_n is semi-positive definite, and the elements of F and those of Ψ_n are unrestricted except that $|I_r + F'D^{-1}F\Psi_n|$ is bounded by $O(T^k)$ for some $k \geq 1$.

Let $\widehat{\theta}$ be the maximum likelihood estimator over the parameter space Θ under a given set of identification restrictions (IC1 or IC2). That is, $\widehat{\theta} = \text{argmax}_{\theta \in \Theta} \ell(\theta)$. To establish consistency, we need to make a distinction between the true parameters and the variables in the likelihood function. Let $\theta^0 = (\alpha^0, \Psi_n^0, F^0, D^0)$ denote the true parameter, an interior point of Θ. Let $G^0 = \Gamma^0 F^0$, where Γ^0 is Γ evaluated at α^0. Also introduce two $T \times T$ matrices:

$$J_T = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 \end{bmatrix}, \quad L = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \\ \alpha & 1 & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \alpha^{T-2} & \cdots & \alpha & 1 & 0 \end{bmatrix} \tag{15}$$

Note $L = J_T\Gamma = J_T B^{-1}$; both J_T and L are $T \times T$.

Before preceding, we emphasize that under fixed T, it is easy to obtain consistency and asymptotic normality. Classical factor analysis relies crucially on the assumption that $\sqrt{N}(S_n - \Sigma(\theta^0))$ is asymptotically normal, as $N \to \infty$. This assumption combined with the delta method (Taylor expansion of the objection function) is sufficient for asymptotic normality. Under large T, however, the dimension of S_n increases, so the limit of $\sqrt{N}(S_n - \Sigma(\theta^0))$ is not well defined as $N, T \to \infty$. In addition, we have infinite number of parameters in the limit. So the classical approach fails to work. A new framework is needed. The analysis is very demanding primarily because we need to handle

large dimensional matrices and an infinite number of parameters. Our analysis of consistency and asymptotic normality is inevitably different from the classical analysis.

We start with the following lemma.

Lemma 1 *Under Assumptions 1-3 and under either IC1 or IC2, as $N, T \to \infty$, we have uniformly for $\theta = (\alpha, F, \Psi_n, D) \in \Theta$,*

$$\begin{aligned}\frac{1}{nT}\ell(\theta) &= -\frac{1}{2T}\Big[\sum_{t=1}^{T}\log(\sigma_t^2) + \frac{\sigma_t^{02}}{\sigma_t^2}\Big] - \frac{1}{2}(\alpha-\alpha^0)^2\,\frac{1}{T}tr\Big[L^0D^0L^{0\prime}D^{-1}\Big] \\ &- \frac{1}{2T}tr\Big[G^0\Psi_n^0G^{0\prime}\Sigma(\theta)^{-1}\Big] + o_p(1).\end{aligned}$$

where $D = \mathrm{diag}(\sigma_1^2, ..., \sigma_T^2)$, $\Sigma(\theta) = \Gamma(F\Psi_nF' + D)\Gamma'$, $L^0 = J_T\Gamma^0$; $G^0 = \Gamma^0F^0$; $o_p(1)$ is uniform in $\theta \in \Theta$.

Evaluate the the likelihood function at $\theta^0 = (\alpha^0, F^0, \Psi_n^0, D^0)$, we have

$$\begin{aligned}\frac{1}{nT}\ell(\theta^0) &= -\frac{1}{2T}\Big[\sum_{t=1}^{T}\log(\sigma_t^{02}) + 1\Big] - \frac{1}{2T}\mathrm{tr}\Big[G^0\Psi_n^0G^{0\prime}\Sigma(\theta^0)^{-1}\Big] + o_p(1) \\ &= -\frac{1}{2T}\Big[\sum_{t=1}^{T}\log(\sigma_t^{02}) + 1\Big] + o_p(1)\end{aligned}$$

the second equality follows from $T^{-1}\mathrm{tr}[G^0\Psi_n^0G^{0\prime}\Sigma(\theta^0)^{-1}] = O_p(T^{-1}) = o_p(1)$, which is easy to show as it does not involve any estimated parameters. Consider the centered-likelihood function

$$\begin{aligned}\frac{1}{nT}\ell(\theta) - \frac{1}{nT}\ell(\theta^0) &= -\frac{1}{2T}\Big[\sum_{t=1}^{T}\log(\sigma_t^2) + \frac{\sigma_t^{02}}{\sigma_t^2} - \log(\sigma_t^{02}) - 1\Big] \\ &-\frac{1}{2}(\alpha-\alpha^0)^2\frac{1}{T}\mathrm{tr}\Big[L^0D^0L^{0\prime}D^{-1}\Big] \\ &-\frac{1}{2T}\mathrm{tr}\Big[G^0\Psi_n^0G^{0\prime}\Sigma(\theta)^{-1}\Big] + o_p(1)\end{aligned}$$

A key observation is that the three terms on the right hand side are all non-positive for all values $\theta \in \Theta$. In particular, they are non-positive when evaluated at $\widehat{\theta}$. On the other hand, $\ell(\widehat{\theta}) - \ell(\theta^0) \geq 0$. This can only be possible if

$$(\widehat{\alpha}-\alpha^0)^2\frac{1}{T}\mathrm{tr}\Big[L^0D^0L^{0\prime}\widehat{D}^{-1}\Big] = o_p(1)$$

$$\frac{1}{T}\Big[\sum_{t=1}^{T}\log(\widehat{\sigma}_t^2) + \frac{\sigma_t^{02}}{\widehat{\sigma}_t^2} - \log(\sigma_t^{02}) - 1\Big] = o_p(1)$$

$$\frac{1}{T}\mathrm{tr}\Big[G^0\Psi_n^0G^{0\prime}\Sigma(\widehat{\theta})^{-1}\Big] = o_p(1) \tag{16}$$

The first equation implies the consistency of $\widehat{\alpha}$ because it can be shown that $\frac{1}{T}\text{tr}(L^0 D^0 L^{0\prime}\widehat{D}^{-1}) \geq c > 0$ for some c, not depending on T and N. So $\widehat{\alpha} = \alpha^0 + o_p(1)$. The second equation implies an average consistency in the sense that

$$\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{\sigma}_t^2 - \sigma_t^{02}\right)^2 = o_p(1). \tag{17}$$

This follows from the fact that the function $h(x) = \log(x) + \log(\frac{a_i}{x}) - \log(a_i) - 1$ satisfies $h(x) \geq c(x - a_i)^2$ for all $x, a_i \in [a, b]$, where $0 < a < b < \infty$, for some $c > 0$ only depending on a and b; also see Bai (2013). The consistency of $\widehat{\alpha}$ and the average consistency of $\widehat{\sigma}_t^2$ in (17) together with (16) imply that $\widehat{\Psi} = \Psi_n^0 + o_p(1)$, and $\widehat{\sigma}_t^2 = \sigma_t^{02} + o_p(1)$ and $\widehat{f}_t = f_t^0 + o_p(1)$ for each t, under either IC1 or IC2. The proof is given in the appendix. Therefore, we have

Proposition 1 *Under Assumptions 1-3 and under either IC1 or IC2, we have* $\widehat{\alpha} = \alpha^0 + o_p(1)$, $\widehat{\Psi} = \Psi_n^0 + o_p(1)$*; and for each* t, $\widehat{f}_t = f_t^0 + o_p(1)$ *and* $\widehat{\sigma}_t^2 = \sigma_t^{02} + o_p(1)$.

We next investigate the asymptotic representations for the estimators and derive their limiting distributions. Given consistency, it is no longer necessary to put a superscript "0" for the true parameters. We shall drop the superscript. Thus all parameters or variables without a "hat" represent the true values.

From the first order conditions, we can show that the estimator $\widehat{\alpha}$ is given by

$$\widehat{\alpha} = \left[\text{tr}\left(J_T S_n J_T' \Omega(\widehat{\theta})^{-1}\right)\right]^{-1}\text{tr}\left(J_T S_n \Omega(\widehat{\theta})^{-1}\right)$$

and the time series heteroskedasticity is estimated by

$$\widehat{D} = \text{diag}\left[\widehat{B} S_n \widehat{B}' - \widehat{F}\widehat{\Psi}\widehat{F}'\right].$$

Remark 7 The above expression says that to estimate σ_t^2, there is no need to estimate the individual residuals ε_{it}. If the individuals ε_{it} were to estimated, it would invariably need to estimate both F and $\Lambda = (\lambda_1, ..., \lambda_N)'$. This would lead to the incidental parameter problem, and thus biases and loss of efficiency. In fact, if T is fixed, λ_i cannot be consistently estimated, this means that individuals ε_{it} cannot be consistently estimated. This further implies that error variance σ_t^2 cannot be consistently estimated using the residuals $\widehat{\varepsilon}_{it}$. This is essentially the bias studied by Neyman and Scott (1948), though the latter paper assumes homoskedasticity. The FIML approach avoids estimating individuals λ_i even if they are fixed constants, and thus permits consistent estimation of σ_t^2. □

Whether IC1 or IC2 is used, the product $\widehat{F}\widehat{\Psi}\widehat{F}'$ is identical, and $\widehat{D}$ is also identical. So the matrix $\Omega(\widehat{\theta}) = \widehat{F}\widehat{\Psi}\widehat{F}' + \widehat{D}$ is identical under IC1 or IC2. This further implies that $\widehat{\alpha}$ is the same in view of the expression for $\widehat{\alpha}$.

The asymptotic representation of $\widehat{\alpha}$ is given in the following theorem:

Theorem 1 *Under Assumptions 1-3, and with either IC1 or IC2,*

$$\sqrt{NT}(\widehat{\alpha}-\alpha)=\Big(\frac{1}{T}tr(LDL'D^{-1})\Big)^{-1}\times\Big[\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\varepsilon_i'D^{-1}L\varepsilon_i\Big]+o_p(1) \tag{18}$$

where $o_p(1)$ *holds if* $N,T\to\infty$ *with* $T/N^2\to 0$ *and* $N/T^3\to 0$.

The interpretation of Theorem 1 is the following. Suppose that the dynamic panel model is such that there are no time effects and no factor structure: $y_{it}=\alpha\, y_{it-1}+\varepsilon_{it}$, and that the heteroskedasticities σ_t^2 are known. Then the generalized least squares method for α has the asymptotic representation given by Theorem 1. So the quasi-FIML method eliminates all these incidental parameters and as if σ_t^2 were known. The derivation of Theorem 1 is very demanding. In the appendix, we provide the key insights as to why the result holds, along with the necessary technical details. The requirement of $N/T^3\to 0$ is for the representation to be as simple as above. It is not a condition for bias removal because $\widehat{\alpha}$ is consistent (without bias) under even fixed T. Also, if N and T are comparable such that $N/T\to c<\infty$, then clearly $N/T^3\to 0$. Thus it is a mild condition for the above representation to hold.

To derive the limiting distribution, notice that the variance of $\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\varepsilon_i'D^{-1}L\varepsilon_i$ is equal to $\frac{1}{T}\mathrm{tr}(LDL'D^{-1})$, and

$$\frac{1}{T}\mathrm{tr}(LDL'D^{-1})=\frac{1}{T}\sum_{t=2}^{T}\frac{1}{\sigma_t^2}\Big(\sigma_{t-1}^2+\alpha^2\sigma_{t-2}^2+\cdots+\alpha^{2(t-2)}\sigma_1^2\Big)\longrightarrow\gamma>0 \tag{19}$$

where we assume the above limit exists. The representation of $\widehat{\alpha}$ implies $\sqrt{NT}(\widehat{\alpha}-\alpha)\xrightarrow{d}N\Big(0,1/\gamma\Big)$. The asymptotic representation of $\widehat{\sigma}_t^2$ is found to be

$$\widehat{\sigma}_t^2-\sigma_t^2=\frac{1}{N}\sum_{i=1}^{N}(\varepsilon_{it}^2-\sigma_t^2)+o_p(N^{-1/2})+O_p(1/T). \tag{20}$$

Summarizing the above results, we have

Theorem 2 *Under the assumptions of Theorem 1, we have*

$$\sqrt{NT}(\widehat{\alpha}-\alpha)\xrightarrow{d}N\Big(0,1/\gamma\Big),$$

and for each t*, let* κ_t *be the excess kurtosis of* ε_{it}*, then*

$$\sqrt{N}(\widehat{\sigma}_t^2-\sigma_t^2)\xrightarrow{d}N\big(0,(2+\kappa_t)\sigma_t^4\big).$$

The estimator is centered at zero despite incidental parameters in the time effects, in the factor structure, and in the heteroskedasticity. For additive effects models, the within group estimator of α has a bias of order $1/T$ and the GMM estimator has a bias of order $1/N$ (Alveraz and Arellano, 2003). Thus the FIML method has desirable theoretical properties.

Under homoskedasticity, (19) implies $\gamma = 1/(1-\alpha^2)$ so $1/\gamma = 1-\alpha^2$. Theorem 2 implies that $\sqrt{NT}(\widehat{\alpha}-\alpha) \to N(0, 1-\alpha^2)$. This is obtained without enforcing homoskedasticity. Thus there is no loss of asymptotic efficiency even under homoskedasticity. Enforcing homoskedasticity does not increase efficiency under large T, and will be inconsistent under fixed T when homoskedasticity does not hold. The FIML estimator is consistent under both fixed and large T.

4.3 Inference on $\widehat{F}$ and $\widehat{\Psi}$

The rate of convergence and the limiting distributions for $\widehat{F}$ and $\widehat{\Psi}$ are of independent interest as they can be useful for analysis such as diffusion index forecasting and factor-augmented vector autoregression (FAVAR). The estimators $\widehat{\Psi}$ and $\widehat{F}$ and their distributions depend on which restrictions are used. Under IC1, it can be shown that

$$\widehat{\Psi} = (\widehat{F}'\widehat{D}^{-1}\widehat{F})^{-1}(\widehat{F}'\widehat{D}^{-1}\widehat{B}S_n\widehat{B}'\widehat{D}^{-1}\widehat{F})(\widehat{F}'\widehat{D}^{-1}\widehat{F})^{-1} - (\widehat{F}'\widehat{D}^{-1}\widehat{F})^{-1}$$

$$\widehat{F}' = (I_r + \widehat{F}'\widehat{D}^{-1}\widehat{F}\widehat{\Psi})^{-1}\widehat{F}'\widehat{D}^{-1}\widehat{B}S_n\widehat{B}'$$

subject to the restriction that the first $(r \times r)$ block of $\widehat{F}$ is I_r. Under IC2,

$$\widehat{\Psi} = \text{diag}\left(T^{-2}[\widehat{F}'\widehat{D}^{-1}\widehat{B}S_n\widehat{B}'\widehat{D}^{-1}\widehat{F}] - T^{-1}I_r\right)$$

$$\widehat{F}' = (I_r + T\widehat{\Psi})^{-1}\widehat{F}'\widehat{D}^{-1}\widehat{B}S_n\widehat{B}'$$

and subject to the normalization $T^{-1}\widehat{F}'\widehat{D}^{-1}\widehat{F} = I_r$.

The rate of convergence for $\widehat{f_t}$ is $N^{1/2}$, the best rate possible even when the factor loadings λ_i $(i = 1, 2, ..., N)$ are observable. However, the rate for $\widehat{\Psi}$ depends on the identification restrictions. Under IC1, the rate is $N^{1/2}$, and under IC2, the rate is $(NT)^{1/2}$. The underlying reason is the following. The matrix Ψ_n contains a small number of parameters. Under IC2, the entire cross sections are used to identify and to estimate Ψ_n, so the convergence rate is faster. Under IC1, the first $r \times r$ block of F is restricted to be I_r in order to identify Ψ_n, we effectively redistribute the first block of F to Ψ_n. The rate for the newly defined $\widehat{\Psi}$ is dominated by the rate of $\widehat{f_t}$, which is $N^{1/2}$.

Under IC1, the asymptotic representation of $\widehat{\Psi}$ is found to be

$$\sqrt{N}(\widehat{\Psi} - \Psi_n) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\lambda_i - \bar{\lambda})\xi_i' + \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\xi_i(\lambda_i - \bar{\lambda})' + o_p(1) \tag{21}$$

where $\xi_i = (\varepsilon_{i1}, ..., \varepsilon_{ir})'$ and for $t = r+1, r+2, ..., T$,

$$\sqrt{N}(\widehat{f_t} - f_t) = -\Psi_n^{-1}\Big(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\lambda_i - \bar{\lambda})\xi_i'\Big)f_t + \Psi_n^{-1}\Big(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\lambda_i - \bar{\lambda})\varepsilon_{it}\Big) + o_p(1). \tag{22}$$

From the asymptotic representations, we find the limiting distributions:

Proposition 2 *Under Assumptions 1-3 and IC1, as $N, T \to \infty$, we have, for each $t > r$,*

$$\sqrt{N}(\widehat{f}_t - f_t) \xrightarrow{d} N\Big(0, \Psi^{-1}[f_t' D_r f_t + \sigma_t^2]\Big),$$

$$\sqrt{N}\,\mathrm{vech}(\widehat{\Psi} - \Psi_n) \xrightarrow{d} N\Big(0, 4\mathcal{D}_r^+ (D_r \otimes \Psi)\mathcal{D}_r^{+\prime}\Big)$$

where $D_r = \mathrm{diag}(\sigma_1^2, ..., \sigma_r^2)$, Ψ is the limit of Ψ_n, and $\mathcal{D}_r^+$ is the Moore-Penrose generalized inverse of the duplication matrix $\mathcal{D}_r$ associated with an $r \times r$ matrix.

Under IC2, the estimated factors has the following representation,

$$\sqrt{N}(\widehat{f}_t - f_t) = \Psi_n^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (\lambda_i - \bar{\lambda})\varepsilon_{it} + o_p(1) \tag{23}$$

for $t = 1, 2, ..., T$. The estimator has a simple interpretation. Since $\Psi_n = \frac{1}{N-1}\sum_{i=1}^{N}(\lambda_i - \bar{\lambda})(\widehat{\lambda}_i - \bar{\lambda})'$, the estimator $\widehat{f}_t$ is the least squares regression of $y_i - \alpha\, y_{i,-1}$ $(T \times 1)$ on Λ and a constant, as if Λ were known even though we never estimate Λ itself other than its sample variance. It is thus an interesting result. The central limit theorem (CLT) implies that $\sqrt{N}(\widehat{f}_t - f_t) \xrightarrow{d} N(0, \Psi^{-1}\sigma_t^2)$. The asymptotic variance of $\widehat{f}_t$ is consistently estimable because both $\widehat{\Psi}_n$ and $\widehat{\sigma}_t^2$ are consistent.

Under IC2, the convergence rate for $\widehat{\Psi}$ is much faster. However, there is a bias of order $O(1/N)$ arising from the estimation of f_t and σ_t^2. When scaled by the convergence rate $(NT)^{1/2}$, the bias is non-negligible unless $T/N \to 0$. The asymptotic representation for $\widehat{\Psi}$ is found to be

Lemma 2 *Under Assumptions 1-3 and IC2 and $T/N \to 0$,*

$$\begin{aligned}\sqrt{NT}\,\mathrm{diag}(\widehat{\Psi} - \Psi_n) = &-2\sqrt{NT}(\widehat{\alpha} - \alpha)\,\mathrm{diag}\left[\Psi_n \frac{1}{T}(F'L'D^{-1}F)\right] \\ &- \mathrm{diag}\left[\Psi_n \frac{1}{\sqrt{NT}} \sum_{i=1}^{N}\sum_{t=1}^{T} \frac{1}{\sigma_t^4}(\varepsilon_{it}^2 - \sigma_t^2) f_t f_t'\right] \\ &+ 2\,\mathrm{diag}\left[\frac{1}{\sqrt{NT}} \sum_{i=1}^{N}\sum_{t=1}^{T} \frac{1}{\sigma_t^2}(\lambda_i - \bar{\lambda}) f_t'\,\varepsilon_{it}\right] + o_p(1)\end{aligned} \tag{24}$$

where $\mathrm{diag}(A)$ denotes the vector formed from the diagonal elements of A.

Let Υ denote the limit of $\frac{1}{T}(F'L'D^{-1}F)$ as $T \to \infty$. Let h denote the $r \times 1$ vector of $\mathrm{diag}(\Psi\Upsilon)$. Let $\mathcal{P}_r$ be a diagonal selection matrix $(r \times r^2)$ such that $\mathrm{diag}(C) = \mathcal{P}_r \mathrm{vec}(C)$ for any $r \times r$ matrix C. The representations for $\widehat{f}_t$ and $\widehat{\Psi}_n$ imply

Proposition 3 *Under Assumptions 1-3 and IC2, as $N, T \to \infty$, we have, for each t,*

$$\sqrt{N}(\widehat{f}_t - f_t) \xrightarrow{d} N(0, \Psi^{-1}\sigma_t^2).$$

And if $T/N \to 0$ and ε_{it} are normal, then

$$\sqrt{NT}\,\mathrm{diag}(\widehat{\Psi} - \Psi_n) \xrightarrow{d} N\Big(0, 4hh'/\gamma + \mathcal{P}_r\Big[2(I_r \otimes \Psi)\Xi(I_r \otimes \Psi) + 4(Q \times \Psi)\Big]\mathcal{P}_r'\Big)$$

where Q and Ξ are given in Assumption 3, and $\mathcal{P}_r$ is a diagonal selection matrix $(r \times r^2)$.

The normality assumption of ε_{it} is only used for deriving the limiting variance of $\widehat{\Psi}$. It is also easy to find the limiting distribution of $\sqrt{NT}(\widehat{\Psi}-\Psi_n)$ under non-normality given its representation in (24). Here the condition T/N going to zero is needed under the fast scaling rate $\sqrt{NT}$. This condition is not needed for all other estimated parameters, and especially not needed for $\widehat{\alpha}$.

4.4 Efficiency

Efficiency of $\widehat{\theta}$ under fixed T. The objective is to show that the estimator $\widehat{\theta}=(\widehat{\alpha},\widehat{D},\widehat{F},\widehat{\Psi}_n)$ is efficient among all estimators that are based on the second moments of the data, regardless of normality. Let $s_n=\text{vech}(S_n)$ and $g(\theta)=\text{vech}(\Sigma(\theta))$. It is well known that the estimator $\widehat{\theta}$ based on the objection function (14) is asymptotically equivalent to the generalized method moments (GMM) estimator

$$\min_{\theta} n[s_n-g(\theta)]'W^{-1}[s_n-g(\theta)]$$

where $W=2\mathcal{D}^{+}[\Sigma(\theta^0)\otimes\Sigma(\theta^0)]\mathcal{D}^{+'}$ and $\mathcal{D}^{+}$, a matrix of $T(T+1)/2\times T^2$, is the generalized inverse of a duplication matrix $\mathcal{D}$, e.g., Chamberlain (1984) and Magnus and Neudecker (1999). The optimal GMM uses the inverse of $W_{opt}=\text{var}(\sqrt{n}[s_n-g(\theta^0)])$ as the weight matrix. Let $G=\partial g/\partial\theta'$, then the optimal GMM has the limiting distribution

$$\sqrt{n}(\widehat{\theta}_{opt}-\theta^0)\xrightarrow{d}N\big(0,\text{plim}(G'W_{opt}^{-1}G)^{-1}\big)$$

In comparison, the asymptotic variance of $\widehat{\theta}$ is the probability limit of

$$(G'W^{-1}G)^{-1}(G'W^{-1}W_{opt}W^{-1}G)(G'W^{-1}G)^{-1}$$

However, we will show that the preceding expression coincides with $(G'W_{opt}^{-1}G)^{-1}$. Thus we have

Theorem 3 *Under Assumptions 1-3, the quasi-FIML is asymptotically equivalent to the optimal GMM estimator based on the moments $E[s_n-g(\theta)]=0$, and*

$$\sqrt{n}(\widehat{\theta}-\theta^0)\xrightarrow{d}N\big(0,\text{plim}(G'W_{opt}^{-1}G)^{-1}\big).$$

It is interesting to note that the FIML does not explicitly estimate the optimal weighting matrix W_{opt}, but still achieves the efficiency. Estimation of W_{opt} would involve the fourth moments of the data. The number of elements in the optimal weighting matrix is large even with a moderate T (order of T^2 by T^2). Thus the estimate for W_{opt} can be unreliable. Also note that, we prove this proposition under the fixed effects setup without assuming λ_i to be iid random variables and without assuming ε_{it} to be normal. Moreover, FIML remains to be efficient under large T, as discussed in Theorem 4 below.

To prove Theorem 3, we first derive the analytical expression for W_{opt} under the fixed effects setup (λ_i are nonrandom)

$$\sqrt{n}[s_n-g(\theta^0)]=\sqrt{n}\ \text{vech}[H+H'+\Gamma(S_{\varepsilon\varepsilon}-D)\Gamma']$$

where

$$H = \Gamma \frac{1}{n} \sum_{i=1}^{N} (\varepsilon_i - \bar{\varepsilon})(\lambda_i - \bar{\lambda})' F' \Gamma', \quad S_{\varepsilon\varepsilon} = \frac{1}{n} \sum_{i=1}^{N} (\varepsilon_i - \bar{\varepsilon})(\varepsilon_i - \bar{\varepsilon})'$$

It is easy to show that

$$\text{var}[\sqrt{n}\,\text{vech}(H + H')] = 4\mathcal{D}^+ \Gamma (F \Psi_n F' \otimes D) \Gamma' \mathcal{D}^{+'}$$

It can also be shown that

$$\text{var}[\sqrt{n}\,\text{vech}(\Gamma S_{\varepsilon\varepsilon} \Gamma')] = \mathcal{D}^+ (\Gamma \otimes \Gamma) \mathcal{P}' V \mathcal{P} (\Gamma' \otimes \Gamma') \mathcal{D}^{+'} + 2\mathcal{D}^+ (\Gamma D \Gamma' \otimes \Gamma D \Gamma') \mathcal{D}^{+'}$$

where $\mathcal{P}$ $(T \times T^2)$ is a diagonal selection matrix such that $\mathcal{P}\text{vec}(A)$ gives the diagonal elements of A for a T-dimensional square matrix A, and V is a T-dimensional diagonal matrix with elements being $E(\varepsilon_{it}^4) - 3\sigma_t^4$ $(t = 1, 2, ..., T)$. The optimal weighting matrix W_{opt} is given by the sum of the two preceding equations.

Next, from $\Sigma(\theta^0) = \Gamma(F \Psi_n F' + D)\Gamma'$ and $W = 2\mathcal{D}^+ [\Sigma(\theta^0) \otimes \Sigma(\theta^0)] \mathcal{D}^{+'}$, it follows that

$$\begin{aligned} W = {} & 2\mathcal{D}^+ (\Gamma F \otimes \Gamma F)(\Psi_n \otimes \Psi_n)(F' \Gamma' \otimes F' \Gamma) \mathcal{D}^{+'} \\ & + 4\mathcal{D}^+ \Gamma (F \Psi_n F' \otimes D) \Gamma' \mathcal{D}^{+'} + 2\mathcal{D}^+ (\Gamma D \Gamma' \otimes \Gamma D \Gamma') \mathcal{D}^{+'} \end{aligned}$$

Thus

$$W = W_{opt} + 2\mathcal{D}^+ (\Gamma F \otimes \Gamma F)(\Psi_n \otimes \Psi_n)(F' \Gamma' \otimes F' \Gamma) \mathcal{D}^{+'} - \mathcal{D}^+ (\Gamma \otimes \Gamma) \mathcal{P}' V \mathcal{P} (\Gamma' \otimes \Gamma') \mathcal{D}^{+'} \quad (25)$$

Let ψ denote the free parameters in Ψ_n, and let ϕ denote the diagonal elements of D. Then

$$G_\psi = \frac{\partial g}{\partial \psi'} = \mathcal{D}^+ (\Gamma F \otimes \Gamma F) \mathcal{D}_r, \quad \text{under IC1}$$

$$G_\phi = \frac{\partial g}{\partial \phi'} = \mathcal{D}^+ (\Gamma \otimes \Gamma) \mathcal{P}'$$

The second term on the right-hand side of (25) is $G_\psi \mathcal{D}_r^+ (\Psi_n \otimes \Psi_n) \mathcal{D}_r^{+'} G_\psi'$ because $\Psi_n \otimes \Psi_n = \mathcal{D}_r \mathcal{D}_r^+ (\Psi_n \otimes \Psi_n) \mathcal{D}_r^{+'} \mathcal{D}_r'$. The last term of (25) is equal to $G_\phi V G_\phi'$. This means that

$$W = W_{opt} + GRG' \quad (26)$$

where R is a block diagonal matrix $R = \text{diag}(0, \mathcal{D}^+ (\Psi_N \otimes \Psi_n) \mathcal{D}^{+'}, -V)$; $G = \frac{\partial g}{\partial \theta'}$. From this relationship between W and W_{opt}, we can verify

$$(G' W^{-1} G)^{-1} (G' W^{-1} W_{opt} W^{-1} G)(G' W^{-1} G)^{-1} \equiv (G' W_{opt}^{-1} G)^{-1}$$

In fact, the above holds for an arbitrary symmetric R provided that $W_{opt} + GRG'$ is positive definite; see Shapiro (1986) and Rao and Mitra (1971, Chapter 8). This proves Theorem 3. It follows that

the quasi-FIML estimator with interactive effects is not only consistent but also efficient despite non-normality and despite the fixed effects setup.

Efficiency under large T. The estimator $\widehat{\alpha}$ is efficient under large T in the sense that it achieves the semiparametric efficiency bound. In the supplementary material, we derive the semiparametric efficiency bound in the sense of Hahn and Kuersteiner (2002) under normality of ε_{it}. The nonparametric components of the model include the time effects, the factor process f_t and the factor loadings λ_i and the heteroskedasticities σ_t^2. The bound is derived in the presence of a large number of incidental parameters.

Theorem 4 *Suppose that Assumptions 1-3 hold. Then we have: (i) under normality of* ε_{it}*, the semiparametric efficiency bound for regular estimators of* α *is* $1/\gamma$*, where* γ *is defined in (19); (ii) under the additional assumption that* λ_i *are iid normal and independent of* ε_{it}*, the semiparametric efficiency bound is also* $1/\gamma$*. Furthermore, the quasi-FIML approach achieves the semiparametric efficiency bound.*

For additive effects models (non-interactive) and under homoskedasticity, Hahn and Kuersteiner (2002) derive the semiparametric efficiency bound. They show that the within-group estimator achieves the semiparametric efficiency bound after a bias correction. The quasi-FIML estimator here achieves the semiparametric efficiency bound without the need of bias correction and without the need of homoskedasticity assumption. FIML achieves the efficiency bound under the more general setup of interactive effects.

Regular estimators rule out the superefficient ones, see Hahn and Kuersteiner (2002) and van der Vaart and Wellner (1996). The estimated variance $\widehat{\sigma}_t^2$ is also efficient since it has an asymptotic representation as if $\frac{1}{n}\sum_{i=1}^N \varepsilon_{it}^2$ were observable. That is, even if all individuals ε_{it} (for all i and t) were observable, the estimated variance based on the second moment would have the same representation. Throughout the process, we never estimate the individual ε_{it} (the residuals). Estimating individual residuals would entail estimating individual λ_i, in addition to f_t. That would lead to the incidental parameters biases and efficiency loss. A key to efficiency is the estimation of the sum of the squares, i.e., the whole term $\frac{1}{n}\sum_{i=1}^N \varepsilon_{it}^2$.

5 Computing the quasi-FIML estimator

We implement the FIML procedure by the ECM (expectation and conditional maximization) algorithm of Meng and Rubin (1993). The E-step in the ECM algorithm is identical to that of the EM algorithm of Dempter et al (1977), but the M-step is broken into a sequence of maximizations instead of simultaneously maximization over the full parameter space. Sequential maximization involves low dimensional parameters and often has closed-form solutions, as in our case. We elaborate the ECM procedure for the conditional likelihood, conditional on y_{i0} in (8). Other likelihood functions discussed earlier are also applicable with minor changes.

The complete data likelihood under normality is (assuming η_i is observable)

$$L(\theta) = -\frac{N}{2}\ln|D| - \frac{1}{2}\sum_{i=1}^{N}(u_i - F\eta_i)'D^{-1}(u_i - F\eta_i)$$

$$-\frac{N}{2}\ln|\Psi| - \frac{1}{2}\sum_{i=1}^{N}\mathrm{tr}(\Psi^{-1}\eta_i\eta_i')$$

$$= -\frac{N}{2}\ln|D| - \frac{1}{2}\sum_{i=1}^{N}\Big[u_i'D^{-1}u_i - 2u_i'D^{-1}F\eta_i + \mathrm{tr}(F'D^{-1}F\eta_i\eta_i')\Big]$$

$$-\frac{N}{2}\ln|\Psi| - \frac{1}{2}\sum_{i=1}^{N}\mathrm{tr}(\Psi^{-1}\eta_i\eta_i')$$

where $u_i = y_i - \delta - X_i\gamma - F\psi W_i$, with

$$X_i = [y_{i,-1}, x_i], \quad \gamma = \begin{bmatrix}\alpha\\ \beta\end{bmatrix}, \quad W_i = \begin{bmatrix}y_{i0}\\ w_i\end{bmatrix}, \quad \psi = (\phi_0, \phi).$$

Here $y_{i,-1} = (y_{i0}, y_{i1}, ..., y_{iT-1})'$ and θ collects all the unknown parameters, $\theta = (F, \delta, D, \alpha, \beta, \psi)$. For non-dynamic panels, u_i is defined as $u_i = y_i - \delta - x_i\beta - F\phi w_i$. And if the usual Mundlak projection is used, we replace w_i by $\bar{x}_i$.

The EM algorithm is an iterative procedure. The expectation step of the algorithm finds the condition expectation $Q(\theta|\bar{\theta}) = E(L(\theta)|data, \bar{\theta})$, conditional on the data and assuming $\bar{\theta}$ is the true parameter. The M step maximizes $Q(\theta|\bar{\theta})$ with respect to θ. The procedure iterates by replacing $\bar{\theta}$ with the newly obtained optimal value of θ. By solving the first order conditions, the optimal value of θ satisfies (the supplementary document contains the detailed derivation)

$$F = \sum_{i=1}^{N} v_i(W_i'\psi' + \widehat{\eta}_i')\Big[\sum_{i=1}^{N}\Big(\psi W_i W_i'\psi' + \widehat{\eta}_i W_i'\psi' + \psi W_i\widehat{\eta}_i' + \widehat{\eta_i\eta_i'}\Big)\Big]^{-1}$$

$$\delta = \frac{1}{N}\sum_{i=1}^{N}(y_i - X_i\gamma - F\psi W_i - F\widehat{\eta}_i)$$

$$D = \mathrm{diag}\Big[\frac{1}{N}\sum_{i=1}^{N}\Big(u_iu_i' - 2F\widehat{\eta}_iu_i' + F\widehat{\eta_i\eta_i'}F'\Big)\Big]$$

$$\Psi = \frac{1}{N}\sum_{i=1}^{N}\widehat{\eta_i\eta_i'}$$

and

$$\theta_1 = \Big[\sum_{i=1}^{N}\mathbb{X}_i'D^{-1}\mathbb{X}_i\Big]^{-1}\sum_{i=1}^{N}\Big[\mathbb{X}_i'D^{-1}\Big(y_i - \delta - F\widehat{\eta}_i\Big)\Big]$$

where $v_i = y_i - \delta - X_i\gamma$, $\theta_1 = (\gamma', vec(\psi)')'$ and $\mathbb{X}_i = [y_{i,-1}, x_i, (W_i' \otimes F)]$; $\widehat{\eta}_i$ and $\widehat{\eta_i\eta_i'}$ are the conditional mean and the conditional second moment of η_i.

The solutions for $\theta = (F, D, \Psi, \delta, \alpha, \beta, \psi)$ from the first order conditions are intertwined and they are functions of each other. In other words, there are no closed-form solutions. Therefore, maximization for the expected complete data likelihood itself requires iteration, in addition to the usual EM iterations. To avoid this iteration, the ECM of Meng and Rubin (1993) is pertinent because the sequential conditional maximizations have closed form solutions.

Conditional Maximization. Suppose the parameters are divided into two groups $\theta = (\theta_1, \theta_2)$. The expected complete likelihood function is

$$Q(\theta_1, \theta_2 | \theta_1^{(k)}, \theta_2^{(k)})$$

where the expectation is taken assuming $\theta^{(k)}$ is the true parameter. The sequential maximization sets θ_1 at $\theta_1^{(k)}$ so that the objective function is a function of θ_2 alone. The problem becomes a conditional/constrained maximization (CM)

$$CM1: \qquad \max_{\theta_2} \; Q(\theta_1^{(k)}, \theta_2 | \theta_1^{(k)}, \theta_2^{(k)})$$

Denote the optimal solution by $\theta_2^{(k+1)}$. The second step fixes θ_2 at $\theta_2^{(k+1)}$ so that the objective function is that of θ_1 alone. This is again a conditional/constrained maximization

$$CM2: \qquad \max_{\theta_1} \; Q(\theta_1, \theta_2^{(k+1)} | \theta_1^{(k)}, \theta_2^{(k)})$$

Denote the solution by $\theta_1^{(k+1)}$. Combining the solutions from the two steps, we obtain $\theta^{(k+1)} = (\theta_1^{(k+1)}, \theta_2^{(k+1)})$, which is used as input for computing the conditional expectations for the next round of iteration. Prior to the CM2 step, an expectation step can be taken so that the maximization problem becomes $Q(\theta_1, \theta_2^{(k+1)} | \theta_1^{(k)}, \theta_2^{(k+1)})$. Meng and Rubin (1993) reported that this extra expectation step does not necessarily accelerate the convergence. A further extension to the ECM algorithm is given by Liu and Rubin (1994), called ECME, which for some of the CM steps, the maximization is taken with respect to the actual likelihood function $\ell(\theta)$ rather than the expected complete data likelihood function $Q(\theta | \theta^{(k)})$. Both ECM and ECME share with the standard EM the monotone convergence property, and ECME can have substantially faster rate of convergence. The main advantage is that ECM and ECME in general have closed-form solutions.

In our application, we divide the parameters into three groups $\theta_3 = (F, \Psi)$, $\theta_2 = (\delta, D)$, and $\theta_1 = (\gamma', \text{vec}(\psi)')'$. The expected likelihood Q is maximized with respect to θ_3 first, followed by θ_2 and then by θ_1. Closed-form solutions exist with this division of the parameter space.

Given the kth step solution $\theta^{(k)}$, the ECM solution for $\theta^{(k+1)}$ can now be stated:

$$F^{(k+1)} = \sum_{i=1}^{N} v_i^{(k)} \Big(W_i' \psi^{(k)\prime} + \widehat{\eta_i'} \Big) \Big[\sum_{i=1}^{N} \Big(\psi^{(k)} W_i W_i' \psi^{(k)\prime} + \widehat{\eta}_i W_i' \psi^{(k)\prime} + \psi^{(k)} W_i \widehat{\eta_i'} + \widehat{\eta_i \eta_i'} \Big) \Big]^{-1}$$

$$\Psi^{(k+1)} = \frac{1}{N} \sum_{i=1}^{N} \widehat{\eta_i \eta_i'}$$

$$\delta^{(k+1)} = \frac{1}{N}\sum_{i=1}^{N}\Big(y_i - X_i\gamma^{(k)} - F^{(k+1)}\psi^{(k)}W_i - F^{(k+1)}\widehat{\eta}_i\Big)$$

$$D^{(k+1)} = \text{diag}\Big[\frac{1}{N}\sum_{i=1}^{N}\Big(u_i^{(k+1/2)}u_i^{(k+1/2)\prime} - 2F^{(k+1)}\widehat{\eta}_i u_i^{(k+1/2)\prime} + F^{(k+1)}(\widehat{\eta_i\eta_i'})F^{(k+1)\prime}\Big)\Big]$$

where $v_i^{(k)} = y_i - \delta^{(k)} - X_i\gamma^{(k)}$, and $u_i^{(k+1/2)}$ is the updated residual after the CM1 step,

$$u_i^{(k+1/2)} = y_i - \delta^{(k+1)} - X_i\gamma^{(k)} - F^{(k+1)}\psi^{(k)}W_i$$

The above gives the solutions for the first two CM steps. The third CM step maximizes the Q function with respect to θ_1 only. The closed-from solution is

$$\theta_1^{(k+1)} = \Big[\sum_{i=1}^{N}\mathbb{X}_i^{(k+1)\prime}(D^{(k+1)})^{-1}\mathbb{X}_i^{(k+1)}\Big]^{-1}\sum_{i=1}^{N}\Big[\mathbb{X}_i^{(k+1)\prime}(D^{(k+1)})^{-1}\Big(y_i - \delta^{(k+1)} - F^{(k+1)}\widehat{\eta}_i\Big)\Big]$$

where

$$\mathbb{X}_i^{(k+1)} = [X_i, W_i' \otimes F^{(k+1)}].$$

The conditional expectations $\widehat{\eta}_i$ and $\widehat{\eta_i\eta_i'}$ are taken assuming $\theta^{(k)}$ being the true parameter:

$$\widehat{\eta}_i = E(\eta_i|u_i^{(k)}, \theta^{(k)}) = \Psi^{(k)}F^{(k)\prime}(\Omega^{(k)})^{-1}u_i^{(k)},$$

$$\widehat{\eta_i\eta_i'} = \widehat{\eta}_i\,\widehat{\eta}_i' + \Psi^{(k)} - \Psi^{(k)}F^{(k)\prime}(\Omega^{(k)})^{-1}F^{(k)}\Psi^{(k)},$$

with $\Omega^{(k)} = F^{(k)}\Psi^{(k)}F^{(k)\prime} + D^{(k)}$ and

$$u_i^{(k)} = y_i - \delta^{(k)} - X_i\gamma^{(k)} - F^{(k)}\psi^{(k)}W_i.$$

Having obtained $\theta^{(k+1)}$, we can compute $u_i^{(k+1)}$ and the conditional expectations $E(\eta_i|u_i^{(k+1)}, \theta^{(k+1)})$ and $E(\eta_i\eta_i'|u^{(k+1)}, \theta^{(k+1)})$, and then $\theta^{(k+2)}$. The process is continued until convergence. The choice of starting values is discussed in the supplementary document.

Remark 8 If we replace the CM3 step by the ECME of Liu and Rubin (1994) by directly maximizing the *actual* likelihood function, a standard GLS problem, the solution is

$$\theta_1^{(k+1)} = \Big[\sum_{i=1}^{N}\mathbb{X}_i^{(k+1)\prime}(\Omega^{(k+1)})^{-1}\mathbb{X}_i^{(k+1)}\Big]^{-1}\sum_{i=1}^{N}\Big[\mathbb{X}_i^{(k+1)\prime}(\Omega^{(k+1)})^{-1}\Big(y_i - \delta^{(k+1)}\Big)\Big].$$

Our computer program allows this choice. □

Remark 9 We can also divide the parameters into two groups by combining θ_3 and θ_2. This requires the joint maximization over F and δ (note D will also be obtained given F and δ; Ψ does not depend on F and δ). Joint maximization is achieved by expanding the factor space and factor loadings, $F^\dagger = (\delta, F)$, and $\eta_i^\dagger = (1, \eta_i')'$. We can easily solve for $F^\dagger$ from the original first order

conditions for F and δ. The solution for $F^\dagger$ depends on the conditional mean and the conditional second moments of $\eta_i^\dagger$, which are $\widehat{\eta}_i^\dagger = (1, \widehat{\eta}_i')'$ and

$$\widehat{\eta_i^\dagger \eta_i^{\dagger\prime}} = \begin{bmatrix} 1 & \widehat{\eta}_i' \\ \widehat{\eta}_i & \widehat{\eta_i \eta_i'} \end{bmatrix}$$

respectively. □

Remark 10 (Speeding up the computation). To be concrete, consider computing $\theta_1^{(k+1)}$, which involves the matrix $\sum_{i=1}^N \mathbb{X}_i^{(k+1)\prime}(D^{(k+1)})^{-1}\mathbb{X}_i^{(k+1)}$, where $\mathbb{X}_i^{(k+1)} = [X_i, (W_i' \otimes F^{(k+1)})]$. Let $\mathbb{A}^{(k)}$ denote this matrix for a moment. It is important not to compute matrix $\mathbb{A}^{(k)}$ in brute force. By vectorizing $\mathbb{A}^{(k)}$, we can see that $\mathbb{A}^{(k)}$ depends on components such as $\sum_{i=1}^N (X_i \otimes X_i)$, $\sum_{i=1}^N (X_i \otimes W_i)$, etc. These components do not vary with k (do not depend on iterations). Dramatic computational savings is achieved by computing these non-updating components only once and store their values outside the iteration loops. This is especially important for large N (our program can handle very large N, for example, hundreds of thousands). Matrix $\mathbb{A}^{(k)}$ is then easily constructed from these non-updating components and $D^{(k+1)}$ and $F^{(k+1)}$. A similar treatment is applied to terms wherever applicable. This is chiefly responsible for the fast speed of our algorithm. □

6 Simulation results

Non-dynamic panel. Data are generated according to ($r = 2$):

$$y_{it} = \delta_t + \beta_1 x_{it,1} + \beta_2 x_{it,2} + \lambda_i' f_t + \varepsilon_{it}$$

$$x_{it,k} = \iota' \lambda_i + \iota' f_t + \lambda_i' f_t + \xi_{it,k}, \quad k = 1, 2$$

where $\varepsilon_{it} \sim N(0, \sigma_t^2)$ with $\sigma_t^2 = t$, independent over t and i; $\xi_{it,k}$ and the components of λ_i and f_t are all iid N(0,1); $\iota' = (1, 1)$; $\beta_1 = 1, \beta_2 = 2$. So the regressors are correlated with the loadings, the factors, and the their product. We consider heavy heteroskedasticity such that

$$D = \text{diag}(1, 2, ..., T)$$

We set δ_t to zero, but time effects are allowed in the estimation.

While the usual within-group estimator is consistent under additive effects for non-dynamic models, it is inconsistent under interactive effects. To see the extent of bias, we also report the within-group estimator. The simulation results are reported in Table 1.

The columns are either the sample means (under the β coefficients) or the standard deviations (under SD) from 5000 repetitions. The within-group estimator is inconsistent since it cannot remove the correlation between the factor errors and the regressors. The MLE is consistent and becomes more precise as either N or T increases. The data generating process for x_{it} (also admits a factor structure) requires the projection of λ_i onto the entire path of x_i. The usual Mundlak projection

Table 1: Estimated coefficients for the non-dynamic panel

T	N	Within-Group				MLE			
		$\beta_1 = 1$	SD	$\beta_2 = 2$	SD	$\beta_1 = 1$	SD	$\beta_2 = 2$	SD
5	100	1.382	0.088	2.382	0.088	1.046	0.144	2.045	0.141
5	200	1.385	0.071	2.382	0.070	1.017	0.089	2.016	0.090
5	500	1.383	0.060	2.383	0.060	1.004	0.049	2.003	0.050
10	100	1.393	0.071	2.391	0.072	1.029	0.104	2.030	0.102
10	200	1.391	0.054	2.392	0.056	1.007	0.058	2.006	0.057
10	500	1.393	0.040	2.392	0.041	1.001	0.033	2.000	0.033

on $\bar{x}_i$ is inconsistent. Not reported is the widely used Pesaran's estimator, which does not perform well. This corroborates the theory in Westerlund and Urbain (2013), who show that Pesaran's estimator becomes inconsistent when the factor loadings in the y equation are correlated with the factor loadings in the x equation. Our data generating process allows this correlation.

Dynamic panel. The y process is generated as ($r = 2$)

$$y_{it} = \delta_t + \alpha\, y_{it-1} + \beta_1 x_{it,1} + \beta_2 x_{it,2} + \lambda_i' f_t + \varepsilon_{it}$$

all other variables are generated the same way as in the non-dynamic case. We again set δ_t to zero but allow time effects in the estimation. For each T, we simulate $2T$ observations and then discard the first half (we could generate only T observations with an arbitrary y_{i0}). The variance of ε_{it} is set to 1 for the first half, and for the second half, its variance is set to

$$\text{var}(\varepsilon_{it}) = t; \quad t = 1, 2, ..., T$$

So the retained sample is heteroskedastic. Two different values of α are considered: $\alpha = 0.5$ and $\alpha = 1.0$. Table 2 reports the sample means and the standard deviations of the estimated slope coefficients from 5000 repetitions. The top panel is for $\alpha = 0.5$ and the bottom panel is for $\alpha = 1$.

For dynamic models, we project λ_i onto the entire path of the regressors x_{it} plus the initial value of y_{i0}. The FIML jointly estimates all coefficients, although only the slope coefficients are reported. Again, the within-group estimator is inconsistent and the FIML performs well. While our theoretical analysis focuses on the stable case $|\alpha| < 1$, simulations show that FIML also works for the case of $\alpha = 1$.

Estimated heteroskedasticities for the dynamic panel. The maximum likelihood approach also produces good estimates of heteroskedasticities for the idiosyncratic errors. Table 3 reports the estimates for the case of $T = 10$. The actual values are $\sigma_t^2 = t$ ($t = 1, 2, ..., 10$). The estimates have some downward biases (reported are the sample means without adjustment for the degrees of freedom). The precision increases as N increases, as expected. The standard errors become larger as t increases, also as expected, because the standard error is proportional to $\sigma_t^2 = t$ (according to the limiting distribution), The precision does not depend on the value of α. This is consistent with the theory.

Table 2: Estimated coefficients for dynamic panels ($\alpha = 0.5$ and $\alpha = 1.0$)

		Within-Group						MLE					
T	N	α	SD	$\beta_1 = 1$	SD	$\beta_2 = 2$	SD	α	SD	$\beta_1 = 1$	SD	$\beta_2 = 2$	SD
5	100	0.466	0.031	1.363	0.092	2.349	0.097	0.488	0.051	1.050	0.141	2.047	0.142
5	200	0.465	0.029	1.361	0.076	2.351	0.083	0.496	0.037	1.017	0.087	2.016	0.088
5	500	0.465	0.027	1.360	0.068	2.349	0.071	0.499	0.023	1.004	0.049	2.003	0.050
10	100	0.476	0.019	1.384	0.072	2.378	0.072	0.495	0.033	1.037	0.105	2.034	0.103
10	200	0.476	0.017	1.384	0.057	2.378	0.056	0.499	0.021	1.007	0.059	2.007	0.059
10	500	0.476	0.015	1.384	0.043	2.379	0.045	0.500	0.012	1.002	0.033	2.001	0.034
5	100	0.974	0.028	1.357	0.096	2.340	0.101	0.983	0.043	1.051	0.145	2.045	0.148
5	200	0.975	0.026	1.355	0.079	2.343	0.086	0.990	0.032	1.016	0.087	2.012	0.089
5	500	0.974	0.025	1.355	0.070	2.342	0.076	0.996	0.019	1.005	0.051	2.003	0.053
10	100	0.988	0.011	1.381	0.073	2.373	0.073	0.991	0.022	1.037	0.110	2.032	0.109
10	200	0.988	0.010	1.381	0.058	2.374	0.058	0.995	0.015	1.007	0.059	2.005	0.059
10	500	0.988	0.010	1.381	0.044	2.374	0.047	0.998	0.008	1.002	0.034	2.001	0.035

Table 3: Estimated heteroskedasticities for the dynamic panel ($T = 10$)

	$\alpha = 0.5$						$\alpha = 1$					
	$N = 100$		$N = 200$		$N = 500$		$N = 100$		$N = 200$		$N = 500$	
$\sigma_t^2 = t$	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	0.851	0.200	0.916	0.129	0.951	0.082	0.853	0.198	0.920	0.130	0.961	0.080
2	1.782	0.359	1.883	0.236	1.938	0.144	1.780	0.360	1.887	0.233	1.947	0.144
3	2.732	0.505	2.855	0.328	2.925	0.207	2.732	0.503	2.855	0.328	2.933	0.206
4	3.684	0.657	3.833	0.422	3.913	0.265	3.694	0.640	3.829	0.423	3.921	0.264
5	4.641	0.796	4.813	0.520	4.908	0.328	4.623	0.790	4.807	0.525	4.913	0.326
6	5.571	0.929	5.784	0.631	5.897	0.399	5.551	0.931	5.773	0.628	5.902	0.400
7	6.575	1.065	6.780	0.735	6.888	0.459	6.541	1.062	6.764	0.737	6.892	0.460
8	7.507	1.234	7.741	0.838	7.876	0.521	7.469	1.236	7.717	0.846	7.875	0.524
9	8.470	1.372	8.720	0.931	8.876	0.579	8.421	1.374	8.692	0.933	8.875	0.581
10	9.410	1.515	9.718	1.025	9.861	0.645	9.344	1.521	9.687	1.026	9.860	0.645

A note on computation. For each (N, T) combination in the tables, the average time it takes to obtain the final FIML estimator is 0.74 seconds of real time on a desktop PC with an Intel E6400 processor. It should take even less time on an up-to-date computer. The average number of EM iterations is 291; more iterations are required to achieve convergence for smaller sample sizes (e.g., the average number is 354 for N=100,T=5, and 203 for N=500, T=10).

7 Conclusion

This paper considers dynamic panel models with a factor analytic error structure, which is correlated with the regressors. A dynamic panel model constitutes a simultaneous equations system with T equations. We show how the FIML procedure can be used to estimate the system, and how to derive the likelihood function for dynamic panels with exogenous and predetermined regressors that are correlated with the factors and factor loadings or both. We examine consistency, limiting distribution, and efficiency of the FIML estimators.

Under fixed T, despite the interactive effects, consistency and asymptotic normality are a consequence of the standard theory for the quasi-FIML procedure because there are only a fixed number of parameters; FIML does not estimate individual λ_is even if they are fixed constants. The FIML estimator is also efficient. These results do not depend on the normality of errors, and hold whether λ_i and f_t are fixed constants or random variables. In particular, we do not assume λ_i to be iid.

We also consider the large T setting, establishing consistency, asymptotic normality, and efficiency. Under large T, an infinite number of parameters exist in the limit. Classical argument of consistency does not apply. Our analysis of consistency and the inferential theory is inevitably different from the existing literature. Moreover, even scaled by the fast rate of convergence $\sqrt{NT}$, the estimator exhibits no asymptotic bias and is asymptotically efficient despite incidental parameters.

Much efforts have also been devoted to the implementation of the FIML under interactive effects. A fast algorithm has been developed to compute the FIML estimators.

There exists a large literature on factor models. Interested readers are referred to the recent survey paper of Stock and Watson (2011). This strand of literature focuses on consistent extraction of the common components and forecasting instead of consistent and efficient estimation of model parameters. The present paper, with a different focus and analysis, is a careful treatment of interactive effects more from a microeconometric perspective (e.g., Arellano, 2003; Baltagi, 2005; and Hsiao, 2003) than from a macroeconometric one. The microeconometric perspective of panel analysis has had tremendous impact on empirical research and remains immensely popular with applied researchers. The interactive-effect models enrich the tools for analyzing dynamic panel data sets. The proposed estimator is shown to have desirable theoretical properties under both fixed and large T.

Appendix: Technical details

Proof of Lemma 1. From $y_i - \bar{y} = \Gamma F(\lambda_i - \bar{\lambda}) + \Gamma(\varepsilon_i - \bar{\varepsilon})$, and let $G = \Gamma F$, we have

$$\begin{aligned} S_n &= G\Psi_n G' + \Gamma D\Gamma' + G\frac{1}{n}\sum_{i=1}^{N}(\lambda_i - \bar{\lambda})(\varepsilon_i - \bar{\varepsilon})'\Gamma' \\ &+ \Gamma\frac{1}{n}\sum_{i=1}^{N}(\varepsilon_i - \bar{\varepsilon})(\lambda_i - \bar{\lambda})'G' + \Gamma\frac{1}{n}\sum_{i=1}^{N}[(\varepsilon_i - \bar{\varepsilon})(\varepsilon_i - \bar{\varepsilon})' - D]\Gamma' \end{aligned}$$

For consistency proof, we put superscript "0" for parameter matrices on the right hand side above (but no need for individual λ_i since it is not estimated), we have

$$\mathrm{tr}[S_n\Sigma(\theta)^{-1}] = \mathrm{tr}[\Sigma(\theta^0)\Sigma(\theta)^{-1}] + 2\mathrm{tr}\Big[G^{0\prime}\Sigma(\theta)^{-1}\Gamma^0\frac{1}{n}\sum_{i=1}^{N}(\varepsilon_i - \bar{\varepsilon})(\lambda_i - \bar{\lambda})'\Big]$$

$$+\mathrm{tr}\Big[\Gamma^{0\prime}\Sigma(\theta)^{-1}\Gamma^0\frac{1}{n}\sum_{i=1}^{N}\Big((\varepsilon_i - \bar{\varepsilon})(\varepsilon_i - \bar{\varepsilon})' - D^0\Big)\Big]$$

Lemma A.1 *Under Assumptions B1-B3, as $N \to \infty$, regardless of T*

(i) $\sup_{\theta\in\Theta}\frac{1}{T}tr\Big[G^{0\prime}\Sigma(\theta)^{-1}\Gamma^0\frac{1}{n}\sum_{i=1}^{N}(\varepsilon_i - \bar{\varepsilon})(\lambda_i - \bar{\lambda})'\Big] = o_p(1)$,
(ii) $\sup_{\theta\in\Theta}\frac{1}{T}tr\Big[\Gamma^{0\prime}\Sigma(\theta)^{-1}\Gamma^0\frac{1}{n}\sum_{i=1}^{N}\Big((\varepsilon_i - \bar{\varepsilon})(\varepsilon_i - \bar{\varepsilon})' - D^0\Big)\Big] = o_p(1)$,
where Θ is the parameter space, $\theta = (\Gamma, F, \Psi_n, D)$, and $\Sigma(\theta) = \Gamma(F\Psi_n F' + D)\Gamma'$.

The proof of this lemma is elementary, thus omitted. Bai and Li (2012) prove a similar result for the case $\Gamma = \Gamma^0 = I_T$. □

By Lemma A.1, we have

$$\frac{1}{nT}\ell(\theta) = -\frac{1}{2T}\log|\Sigma(\theta)| - \frac{1}{2T}\mathrm{tr}[\Sigma(\theta^0)\Sigma(\theta)^{-1}] + o_p(1). \tag{27}$$

From $\Sigma(\theta^0) = G^0\Psi_n^0 G^{0\prime} + \Gamma^0 D^0\Gamma^{0\prime}$, we have

$$\Sigma(\theta^0)\Sigma(\theta)^{-1} = G^0\Psi_n^0 G^{0\prime}\Sigma(\theta)^{-1} + \Gamma^0 D^0\Gamma^{0\prime}\Sigma(\theta)^{-1} \tag{28}$$

Lemma A.2

$$\frac{1}{T}tr[\Gamma^0 D^0\Gamma^{0\prime}\Sigma(\theta)^{-1}] = \frac{1}{T}tr(D^0 D^{-1}) + (\alpha - \alpha^0)^2\frac{1}{T}tr[L^0 D^0 L^{0\prime} D^{-1}] + o(1)$$

where $o(1)$ is in fact $O(T^{-1})$ and is uniform on Θ.

Proof: Let $V = \Gamma D\Gamma'$. Using $\Sigma(\theta)^{-1} = V^{-1} - V^{-1}G(\Psi_n^{-1} + G'V^{-1}G)^{-1}G'V^{-1}$,

$$\mathrm{tr}[\Gamma^0 D^0\Gamma^{0\prime}\Sigma(\theta)^{-1}] = \mathrm{tr}[\Gamma^0 D^0\Gamma^{0\prime}V^{-1}] - \mathrm{tr}[(\Psi_n^{-1} + G'V^{-1}G)^{-1}(G'V^{-1}\Gamma^0 D^0\Gamma^{0\prime}V^{-1}G)]$$

From $V^{-1} = \Gamma^{-1}D^{-1}\Gamma^{-1}$,

$$\text{tr}[\Gamma^0 D^0 \Gamma^{0'} V^{-1}] = \text{tr}[\Gamma^{-1}\Gamma^0 D^0 \Gamma^{0'}\Gamma^{-1}D^{-1}]$$

From $\Gamma^{-1}\Gamma^0 = I_T + (\alpha^0 - \alpha)L^0$, we have

$$\text{tr}[\Gamma^{-1}\Gamma^0 D^0 \Gamma^{0'}\Gamma^{-1}D^{-1}] \equiv \text{tr}(D^0 D^{-1}) + (\alpha^0 - \alpha)^2 \text{tr}(L^0 D^0 L^{0'} D^{-1})$$

we have used the fact that $\text{tr}(L^0 D^0 D^{-1}) = 0$ because L^0 is lower triangular and D^0 and D^{-1} are diagonal. We next show, uniformly on Θ,

$$\text{tr}[(\Psi_n^{-1} + G'V^{-1}G)^{-1}(G'V^{-1}\Gamma^0 D^0 \Gamma^{0'} V^{-1}G)] = O(1) \tag{29}$$

We shall use the fact that for semi-positive matrices A, B, C, and D, if $A \leq C$ and $B \leq D$, then $\text{tr}(AB) \leq \text{tr}(CD)$. Note

$$(\Psi_n^{-1} + G'V^{-1}G)^{-1} \leq (G'V^{-1}G)^{-1} = (F'D^{-1}F)^{-1}$$

We next shows

$$\text{tr}(G'V^{-1}\Gamma^0 D^0 \Gamma^{0'} V^{-1}G) \leq \frac{b}{a}\frac{8}{(1-|\alpha^0|)}(F'D^{-1}F) \tag{30}$$

The preceding two inequalities imply that (29) is bounded by $8(b/a)/(1-|\alpha^0|)\text{tr}(I_r) = 8r(b/a)/(1-|\alpha^0|)$. To prove (30), first $D^0 \leq bI_T$, thus $L^0 D^0 L^{0'} \leq bL^0 L^{0'}$. Since the largest eigenvalue of a symmetric matrix is bounded by the maximum of row sums (of absolute values), and for each row, the sum of absolute values in $L^0 L^{0'}$ is bounded by $2/(1-|\alpha^0|)$, it follows that $L^0 L^{0'} \leq 2/(1-|\alpha^0|)I_T$. Thus $\text{tr}(G'V^{-1}\Gamma^0 D^0 \Gamma^{0'} V^{-1}G) \leq 2b/(1-|\alpha^0|)\text{tr}(G'V^{-1}V^{-1}G)$. Next $G'V^{-1} = F'D^{-1}\Gamma^{-1}$. But $F'D^{-1}\Gamma^{-1}\Gamma'^{-1}D^{-1}F \leq (4/a)F'D^{-1}F$ because $\Gamma^{-1}\Gamma'^{-1} = BB' \leq 4I_T$ and $D^{-1} \leq (1/a)I_T$. This proves (30) and thus (29). Combing results we obtain Lemma A.2. □

By Lemma A.2 and (28) we have

$$\begin{aligned}\frac{1}{T}\text{tr}[\Sigma(\theta^0)\Sigma(\theta)^{-1}] &= \frac{1}{T}\text{tr}[G^0\Psi_n^0 G^{0'}\Sigma(\theta)^{-1}] \\ &\quad + \frac{1}{T}\text{tr}(D^0 D^{-1}) + (\alpha - \alpha^0)^2 \frac{1}{T}\text{tr}[L^0 D^0 L^{0'} D^{-1}] + o(1)\end{aligned} \tag{31}$$

Next, $|\Sigma(\theta)| = |D| \cdot |I_r + F'D^{-1}F\Psi_n|$. Also, $|I_r + F'D^{-1}F\Psi_n| = O(T^k)$ on Θ, we have

$$\frac{1}{T}\log|\Sigma(\theta)| = \frac{1}{T}\log|D| + O(\frac{\log T}{T}) = \frac{1}{T}\sum_{t=1}^{T}\log(\sigma_t^2) + o(1) \tag{32}$$

Combining (27), (31), (32), and $\text{tr}(D^0 D^{-1}) = \sum_{t=1}^{T}\sigma_t^{02}/\sigma_t^2$, we obtain Lemma 1. □

Proof of Proposition 1. The consistency of $\widehat{\alpha}$ and the average consistency of $\widehat{\sigma}_t^2$ in the sense of (17) are already given in the main text. Using $\widehat{\alpha} - \alpha^0 = o_p(1)$, we show that (16) implies

$$\frac{1}{T}\text{tr}\Big(F^0\Psi_n^0 F^{0'}\Omega(\widehat{\theta})^{-1}\Big) = o_p(1) \tag{33}$$

where $\Omega(\widehat{\theta}) = \widehat{F}\widehat{\Psi}\widehat{F}' + \widehat{D}$. Notice

$$\frac{1}{T}\text{tr}[G^0\Psi_n^0 G^{0\prime}\Sigma(\widehat{\theta})^{-1}] = \frac{1}{T}\text{tr}[\widehat{\Gamma}^{-1}\Gamma^0 F^0\Psi_n^0 F^{0\prime}\Gamma^{0\prime}\widehat{\Gamma}^{-1}\Omega(\widehat{\theta})^{-1}]$$

From $\widehat{\Gamma}^{-1}\Gamma^0 = I_T + (\alpha^0 - \widehat{\alpha})L^0$, and let $A^0 = F^0\Psi_n^0 F^{0\prime}$ for the moment, then

$$\frac{1}{T}\text{tr}[G^0\Psi_n^0 G^{0\prime}\Sigma(\widehat{\theta})^{-1}] = \frac{1}{T}\text{tr}[A^0\Omega(\widehat{\theta})^{-1}] + 2(\alpha^0 - \widehat{\alpha})\frac{1}{T}\text{tr}[L^0A^0\Omega(\widehat{\theta})^{-1}] + (\alpha^0 - \widehat{\alpha})^2\frac{1}{T}\text{tr}[L^0A^0L^{0\prime}\Omega(\widehat{\theta})^{-1}]$$

However, it is easy to show that

$$\frac{1}{T}\text{tr}[L^0A^0\Omega(\theta)] = O(1), \quad \text{and } \frac{1}{T}\text{tr}[L^0A^0L^{0\prime}\Omega(\theta)] = O(1)$$

uniformly on Θ, and so they are $O_p(1)$ when evaluated at $\theta = \widehat{\theta}$. Since $\widehat{\alpha} - \alpha^0 = o_p(1)$, we have

$$\frac{1}{T}\text{tr}[G^0\Psi_n^0 G^{0\prime}\Sigma(\widehat{\theta})^{-1}] = \frac{1}{T}\text{tr}[A^0\Omega(\widehat{\theta})^{-1}] + o_p(1)$$

The left hand side is $o_p(1)$ by (16), so $\frac{1}{T}\text{tr}[A^0\Omega(\widehat{\theta})^{-1}] = o_p(1)$, proving (33). Note that $\Omega(\theta)$ has a standard factor structure with diagonal idiosyncratic covariance matrix. Bai and Li (2012) show that, for factor models with diagonal idiosyncratic covariance matrix, (17) and (33) imply that $\widehat{\Psi} = \Psi_n^0 + o_p(1)$, and $\widehat{\sigma}_t^2 = \sigma_t^{02} + o_p(1)$ and $\widehat{f}_t = f_t^0 + o_p(1)$ for each t, under either IC1 or IC2. This completes the proof of Proposition 1. □.

Proof of Theorem 1. The technique details are involved. We first provide the key insights of the proof and then move on to the details. Even the details here only contain the key steps; a complete proof would require a much lengthy argument because of the complexity of the problem. The approach taken here should help readers see the key ideas.

The first order condition for $\widehat{\alpha}$ implies

$$\text{tr}\Big(J_T S_n J_T'\Omega(\widehat{\theta})^{-1}\Big)\widehat{\alpha} = \text{tr}\Big(J_T S_n \Omega(\widehat{\theta})^{-1}\Big)$$

where $\Omega(\widehat{\theta}) = \widehat{F}\widehat{F}' + \widehat{D}$. Here we use the normalization $\Psi_n = I_r$ and $T^{-1}F'D^{-1}F$ being diagonal. This is a variation to IC2. It does not affect the estimator of $\widehat{\alpha}$ (see Remark 5), but simplifies our analysis. From $I_T = B' + \alpha J_T'$, we have

$$\text{tr}\Big(J_T S_n \Omega(\widehat{\theta})^{-1}\Big) = \alpha\,\text{tr}\Big(J_T S_n J_T'\Omega(\widehat{\theta})^{-1}\Big) + \text{tr}\Big(J_T S_n B'\Omega(\widehat{\theta})^{-1}\Big).$$

Thus we can rewrite the estimator as

$$\text{tr}\Big(\frac{1}{T}J_T S_n J_T'\Omega(\widehat{\theta})^{-1}\Big)(\widehat{\alpha} - \alpha) = \text{tr}\Big(\frac{1}{T}J_T S_n B'\Omega(\widehat{\theta})^{-1}\Big) \tag{34}$$

Here we divide each side by T; Though not occurring explicitly, the right hand side also depends on $\widehat{\alpha} - \alpha$ due to the estimation of F and D. Their estimation affects the limiting distribution of $\widehat{\alpha} - \alpha$. While the term on the left $\text{tr}\Big(\frac{1}{T}J_T S_n J_T'\Omega(\widehat{\theta})^{-1}\Big)$ also implicitly depends on $\widehat{\alpha} - \alpha$, the components

that depend on $\widehat{\alpha}-\alpha$ do not affect the limiting distribution because their multiplication with $(\widehat{\alpha}-\alpha)$ already on the left is $O_p((\widehat{\alpha}-\alpha)^2) = o_p(\widehat{\alpha}-\alpha)$. The key is to analyze the right hand side.

The idea of the analysis is the following. Multiplying $\sqrt{NT}$ on both sides of (34) gives

$$\mathrm{tr}\Big(\frac{1}{T}J_T S_n J_T' \Omega(\widehat{\theta})^{-1}\Big)\sqrt{NT}(\widehat{\alpha}-\alpha) = \sqrt{\frac{N}{T}}\mathrm{tr}\Big(J_T S_n B' \Omega(\widehat{\theta})^{-1}\Big) \tag{35}$$

We shall decompose the right hand side term $(N/T)^{1/2}\mathrm{tr}(J_T S_n B' \Omega(\widehat{\theta})^{-1})$ into three groups. The first group has an asymptotic distribution. The second group is asymptotically negligible. The third group is of $(C_1 + o_p(1))\sqrt{NT}(\widehat{\alpha}-\alpha)$, and $C_1 \neq 0$, and must be combined with the left hand side (canceling out with the corresponding term on the left hand side). The resulting outcome is Theorem 1. The following contains the technical details.

Rewrite the right hand side term as

$$\mathrm{tr}\Big(J_T S_n B' \Omega(\widehat{\theta})^{-1}\Big) = \mathrm{tr}\Big(J_T S_n B' \Omega(\theta)^{-1}\Big) + \mathrm{tr}\Big(J_T S_n B' [\Omega(\widehat{\theta})^{-1} - \Omega(\theta)^{-1}]\Big) \tag{36}$$

and notice

$$\begin{aligned} J_T S_n B' &= L\Omega(\theta) + LF\frac{1}{n}\sum_{i=1}^{N}(\lambda_i - \bar{\lambda})(\varepsilon_i - \bar{\varepsilon})' \\ &+ L\frac{1}{n}\sum_{i=1}^{N}(\varepsilon_i - \bar{\varepsilon})(\lambda_i - \bar{\lambda})' F' + L\frac{1}{n}\sum_{i=1}^{N}[(\varepsilon_i - \bar{\varepsilon})(\varepsilon_i - \bar{\varepsilon})' - D]. \end{aligned} \tag{37}$$

Right multiply the preceding equation by $\Omega(\theta)^{-1}$ and noting $\mathrm{tr}(L) = 0$, it is not difficult to show (since there are no estimated parameters),

$$\begin{aligned} \sqrt{\frac{N}{T}}\mathrm{tr}(J_T S_n B' \Omega(\theta)^{-1}) &= \frac{1}{\sqrt{NT}}\sum_{i=1}^{N}(\varepsilon_i - \bar{\varepsilon})' D^{-1} L(\varepsilon_i - \bar{\varepsilon}) \\ &+ \frac{1}{\sqrt{NT}}\sum_{i=1}^{N}(\varepsilon_i - \bar{\varepsilon})' D^{-1} LF(\lambda_i - \bar{\lambda}) \\ &- \frac{1}{\sqrt{NT}}\sum_{i=1}^{N}(\varepsilon_i - \bar{\varepsilon})' D^{-1} F H^{-1}(F' D^{-1} LF)(\lambda_i - \bar{\lambda}) + o_p(1) \end{aligned} \tag{38}$$

where $H = I_r + F' D^{-1} F$. The analysis of the second term of (36) is quite involved, we state the result as a proposition.

Proposition A.1 *The second term of (36), multiplied by $\sqrt{N/T}$, satisfy*

$$\begin{aligned} \sqrt{\frac{N}{T}} tr\Big(J_T S_n B' [\Omega(\widehat{\theta})^{-1} - \Omega(\theta)^{-1}]\Big) &= -\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}(\varepsilon_i - \bar{\varepsilon})' D^{-1} LF(\lambda_i - \bar{\lambda}) \\ &+ \frac{1}{\sqrt{NT}}\sum_{i=1}^{N}(\varepsilon_i - \bar{\varepsilon})' D^{-1} F H^{-1}(F' D^{-1} LF)(\lambda_i - \bar{\lambda}) \\ &+ \frac{1}{T} tr\Big(F' L' \Omega(\widehat{\theta})^{-1} LF\Big)\sqrt{NT}(\widehat{\alpha}-\alpha) + o_p(1) \end{aligned} \tag{39}$$

The proof will be given later. The right hand side of (35) is the sum of (38) and (39):

$$\sqrt{\frac{N}{T}}\mathrm{tr}\Big(J_T S_n B'\Omega(\widehat{\theta})^{-1}\Big) = \frac{1}{\sqrt{NT}}\sum_{i=1}^{N}(\varepsilon_i-\bar{\varepsilon})'D^{-1}L(\varepsilon_i-\bar{\varepsilon}) \tag{40a}$$

$$+\frac{1}{T}\mathrm{tr}\Big(F'L'\Omega(\widehat{\theta})^{-1}LF\Big)\sqrt{NT}(\widehat{\alpha}-\alpha)+o_p(1) \tag{40b}$$

We will show that expression (40b) also appears on the left hand side of (35).

Proposition A.2 *The left hand side of (35) is*

$$tr\Big(\frac{1}{T}J_T S_n J_T'\Omega(\widehat{\theta})^{-1}\Big)\sqrt{NT}(\widehat{\alpha}-\alpha) = \frac{1}{T}tr\Big(LDL'D^{-1}\Big)\sqrt{NT}(\widehat{\alpha}-\alpha) \tag{41a}$$

$$+\frac{1}{T}tr\Big(F'L'\Omega(\widehat{\theta})^{-1}LF\Big)\sqrt{NT}(\widehat{\alpha}-\alpha)+o_p(1) \tag{41b}$$

Given the two propositions, Theorem 1 follows from (35), (40), and (41), noting that (40b) and (41b) cancel each other. The remaining task is to prove the two propositions.

Proof of Proposition A.1. Consider the second term of (36) without taking the trace. By (37),

$$J_T S_n B'[\Omega(\widehat{\theta})^{-1}-\Omega(\theta)^{-1}]$$

$$= L\Omega(\theta)[\Omega(\widehat{\theta})^{-1}-\Omega(\theta)^{-1}] \tag{42a}$$

$$+LF\frac{1}{n}\sum_{i=1}^{N}(\lambda_i-\bar{\lambda})(\varepsilon_i-\bar{\varepsilon})'[\Omega(\widehat{\theta})^{-1}-\Omega(\theta)^{-1}] \tag{42b}$$

$$+L\frac{1}{n}\sum_{i=1}^{N}(\varepsilon_i-\bar{\varepsilon})(\lambda_i-\bar{\lambda})'F'[\Omega(\widehat{\theta})^{-1}-\Omega(\theta)^{-1}] \tag{42c}$$

$$+L\frac{1}{n}\sum_{i=1}^{N}[(\varepsilon_i-\bar{\varepsilon})(\varepsilon_i-\bar{\varepsilon})'-D][\Omega(\widehat{\theta})^{-1}-\Omega(\theta)^{-1}] \tag{42d}$$

It can be shown that the trace of (42b)-(42d) multiplied by $\sqrt{N/T}$ is negligible.

Lemma A.3 *Under Assumptions 1-3,*

1. $\sqrt{N/T}\,\mathrm{tr}\Big(LF\frac{1}{n}\sum_{i=1}^{N}(\lambda_i-\bar{\lambda})(\varepsilon_i-\bar{\varepsilon})'[\Omega(\widehat{\theta})^{-1}-\Omega(\theta)^{-1}]\Big)=o_p(1)$
2. $\sqrt{N/T}\,\mathrm{tr}\Big(L\frac{1}{n}\sum_{i=1}^{N}(\varepsilon_i-\bar{\varepsilon})(\lambda_i-\bar{\lambda})'F'[\Omega(\widehat{\theta})^{-1}-\Omega(\theta)^{-1}]\Big)=o_p(1)$
3. $\sqrt{N/T}\,\mathrm{tr}\Big(L\frac{1}{n}\sum_{i=1}^{N}[(\varepsilon_i-\bar{\varepsilon})(\varepsilon_i-\bar{\varepsilon})'-D][\Omega(\widehat{\theta})^{-1}-\Omega(\theta)^{-1}]\Big)=o_p(1)$

The proof of this lemma will be omitted. We focus on (42a).

Using $\Omega(\widehat{\theta})^{-1}-\Omega(\theta)^{-1}=\Omega(\theta)^{-1}[\Omega(\theta)-\Omega(\widehat{\theta})]\Omega(\widehat{\theta})^{-1}$ and

$$\Omega(\theta)-\Omega(\widehat{\theta}) = FF'+D-(\widehat{F}\widehat{F}'+\widehat{D}) = F(F-\widehat{F})'+(F-\widehat{F})\widehat{F}'+D-\widehat{D}$$

we have

$$\Omega(\widehat{\theta})^{-1} - \Omega(\theta)^{-1} = \Omega(\theta)^{-1}[F(F-\widehat{F})' + (F-\widehat{F})\widehat{F}' + D - \widehat{D}]\Omega(\widehat{\theta})^{-1}$$

Thus

$$\begin{aligned}\mathrm{tr}\Big(L\Omega(\theta)[\Omega(\widehat{\theta})^{-1} - \Omega(\theta)^{-1}]\Big) &= \mathrm{tr}\Big(\Big[LF(F-\widehat{F})' + L(F-\widehat{F})\widehat{F}' + L(D-\widehat{D})\Big]\Omega(\widehat{\theta})^{-1}\Big) \\ &= \mathrm{tr}\Big((F-\widehat{F})'\Omega(\widehat{\theta})^{-1}LF\Big) + \mathrm{tr}\Big(\widehat{F}'\Omega(\widehat{\theta})^{-1}L(F-\widehat{F})\Big) + \mathrm{tr}\Big(L(D-\widehat{D})\Omega(\widehat{\theta})^{-1}\Big)\end{aligned}$$

Using $\Omega(\widehat{\theta})^{-1} = \widehat{D}^{-1} - \widehat{D}^{-1}\widehat{F}\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}$ with $\widehat{H} = I_r + \widehat{F}'\widehat{D}^{-1}\widehat{F}$, we expand the first exression into two, and from $\widehat{F}'\Omega(\widehat{\theta})^{-1} = \widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}$ we can simplify the second expression, and using $\mathrm{tr}(L(D-\widehat{D})\widehat{D}^{-1}) = 0$, we can simplify the third expression. These steps give

$$\begin{aligned}&\mathrm{tr}\Big[L\Omega(\theta)[\Omega(\widehat{\theta})^{-1} - \Omega(\theta)^{-1}]\Big] \\ &\qquad = \mathrm{tr}\Big[(F-\widehat{F})'\widehat{D}^{-1}LF\Big] &\text{(43a)} \\ &\qquad - \mathrm{tr}\Big[[(F-\widehat{F})'\widehat{D}^{-1}\widehat{F}](\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}LF)\Big] &\text{(43b)} \\ &\qquad + \mathrm{tr}\Big[\widehat{F}'\widehat{D}^{-1}L(F-\widehat{F})\widehat{H}^{-1}\Big] &\text{(43c)} \\ &\qquad - \mathrm{tr}\Big[\widehat{F}'L(D-\widehat{D})\widehat{D}^{-1}\widehat{F}\widehat{H}^{-1})\Big] &\text{(43d)}\end{aligned}$$

The last two terms can be shown to be negligible because they are weighted average of $\widehat{f}_t - f_t$ and $(\widehat{\sigma}_t^2 - \sigma_t^2)$ (note $\widehat{H}^{-1} = O_p(1/T)$). The first two terms also involves the weighted sum of $\widehat{f}_t - f_t$, but without the $1/T$ factor, thus dominating the last two terms. Note that $(\widehat{F}'\widehat{D}^{-1}LF\widehat{H}^{-1}) = O_p(1)$.

We next derive the representation of $\widehat{F}$. The first order condition for $\widehat{F}$ satisfies

$$\widehat{F}' = \widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}\widehat{B}S_n\widehat{B}'$$

From $\widehat{B} = B - (\widehat{\alpha} - \alpha)J_T$, we can rewrite

$$\widehat{F} = \widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}BS_nB' - (\widehat{\alpha}-\alpha)\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}J_TS_nB' - (\widehat{\alpha}-\alpha)\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}BS_nJ_T'$$

and here we ignore the term involving $(\widehat{\alpha} - \alpha)^2$. Denote

$$BS_nB' = (FF' + D) + \Upsilon$$

where Υ represents the random part of BS_nB', that is,

$$\begin{aligned}\Upsilon &= F\frac{1}{n}\sum_{i=1}^{N}(\lambda_i - \bar{\lambda})(\varepsilon_i - \bar{\varepsilon})' \\ &+ \frac{1}{n}\sum_{i=1}^{N}(\varepsilon_i - \bar{\varepsilon})(\lambda_i - \bar{\lambda})'F' + \frac{1}{n}\sum_{i=1}^{N}[(\varepsilon_i - \bar{\varepsilon})(\varepsilon_i - \bar{\varepsilon})' - D] \qquad (44)\end{aligned}$$

then

$$\begin{aligned}\widehat{F}' &= \widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}(FF'+D)+\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}\Upsilon \\ &- (\widehat{\alpha}-\alpha)\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}J_T S_n B' - (\widehat{\alpha}-\alpha)\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}BS_n J_T'\end{aligned}$$

But the first term on the right hand side is

$$\begin{aligned}&\widehat{H}^{-1}(\widehat{F}'\widehat{D}^{-1}F)F' + \widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}D \\ &= \widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}(F-\widehat{F})F' + \widehat{H}^{-1}(\widehat{F}'\widehat{D}^{-1}\widehat{F})F' + \widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}D \\ &= \widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}(F-\widehat{F})F' + F' - \widehat{H}^{-1}F' + \widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}D \\ &= F' + \widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}(F-\widehat{F})F' + \widehat{H}^{-1}\widehat{F}'(\widehat{D}^{-1}-D^{-1})D\end{aligned}$$

We can rewrite the representation of $\widehat{F}$ as

$$\begin{aligned}(\widehat{F}-F)' &= \widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}(F-\widehat{F})F' \\ &+ \widehat{H}^{-1}\widehat{F}'(\widehat{D}^{-1}-D^{-1})D \\ &+ \widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}\Upsilon \\ &- (\widehat{\alpha}-\alpha)\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}J_T S_n B' \\ &- (\widehat{\alpha}-\alpha)\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}BS_n J_T'\end{aligned} \tag{45}$$

Using the above representation, term (43a) can be written as (before taking trace),

$$\begin{aligned}-(\widehat{F}-F)'\widehat{D}^{-1}LF &= -\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}(F-\widehat{F})F'\widehat{D}^{-1}LF \\ &- \widehat{H}^{-1}\widehat{F}'(\widehat{D}^{-1}-D^{-1})D\widehat{D}^{-1}LF \\ &- \widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}\Upsilon\widehat{D}^{-1}LF \\ &+ (\widehat{\alpha}-\alpha)\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}J_T S_n B'\widehat{D}^{-1}LF \\ &+ (\widehat{\alpha}-\alpha)\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}BS_n J_T'\widehat{D}^{-1}LF\end{aligned} \tag{46}$$

Term (43b) can be written as (before taking the trace)

$$\begin{aligned}[(\widehat{F}-F)'\widehat{D}^{-1}\widehat{F}](\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}LF) = \\ &+ \widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}(F-\widehat{F})F'\widehat{D}^{-1}\widehat{F}(\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}LF) \\ &+ \widehat{H}^{-1}\widehat{F}'(\widehat{D}^{-1}-D^{-1})D\widehat{D}^{-1}\widehat{F}(\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}LF) \\ &+ \widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}\Upsilon\widehat{D}^{-1}\widehat{F}(\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}LF) \\ &- (\widehat{\alpha}-\alpha)\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}J_T S_n B'\widehat{D}^{-1}\widehat{F}(\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}LF) \\ &- (\widehat{\alpha}-\alpha)\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}BS_n J_T'\widehat{F}(\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}LF)\end{aligned} \tag{47}$$

The first term of (46) and that of (47) are canceled out. To see this, consider the first term of each equation. Their sum is

$$\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}(F-\widehat{F})[F'\widehat{D}^{-1}\widehat{F}\widehat{H}^{-1}-I_r](\widehat{F}'\widehat{D}^{-1}LF) \tag{48}$$

But

$$F'\widehat{D}^{-1}\widehat{F}\widehat{H}^{-1} - I_r = (F-\widehat{F})'\widehat{D}^{-1}\widehat{F}\widehat{H}^{-1} + \widehat{H}^{-1}$$

Thus if we let $A = (F-\widehat{F})'\widehat{D}^{-1}\widehat{F}\widehat{H}^{-1}$, then (48) is

$$A'A(\widehat{F}'\widehat{D}^{-1}LF) + A'\widehat{H}^{-1}(\widehat{F}'\widehat{D}^{-1}LF)$$

which is bounded by $\|A\|^2 O_p(T) + \|A\| O_p(1)$. Note that $\widehat{H}^{-1} = O_p(1/T)$, we can show that (similar to Lemma A.3 of Bai, 2009, note that H denotes a different object there)

$$A = (F-\widehat{F})'\widehat{D}^{-1}\widehat{F}\widehat{H}^{-1} = O_p(\frac{1}{N}) + O_p(\frac{1}{T}) + O_p(\widehat{\alpha}-\alpha) \tag{49}$$

This means that (48) is bounded by

$$[O_p(N^{-2}) + O_p(T^{-2}) + O_p((\widehat{\alpha}-\alpha)^2)]T + O_p(N^{-1}) + O_p(T^{-1}) + O_p(\widehat{\alpha}-\alpha)$$

Multiplied by $\sqrt{N/T}$, the above is of $O_p(\sqrt{T}/N^{3/2}) + O_p(\sqrt{N}/T^{3/2}) + O_p((NT)^{-1/2})$ plus $\sqrt{NT}O_p((\widehat{\alpha}-\alpha)^2) + \sqrt{N/T}(\widehat{\alpha}-\alpha)$. For former is negligible if $T/N^3 \to 0$ and $N/T^3 \to 0$. The latter is of smaller order than $\sqrt{NT}(\widehat{\alpha}-\alpha)$, thus also negligible. It can be shown that the second term of (46) and the second term of (47) are each $O_p((NT)^{-1/2})$. Thus, after multiplying by $\sqrt{N/T}$, they are each $O_p(1/T) = o_p(1)$, not influencing the limiting distribution.

We next study the third term of (46) and that of (47). Using (44), the third term of (46) is

$$\begin{aligned}
-\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}\Upsilon\widehat{D}^{-1}LF = &-\widehat{H}^{-1}(\widehat{F}'\widehat{D}^{-1}F)\frac{1}{n}\sum_{i=1}^{N}(\lambda_i-\bar{\lambda})(\varepsilon_i-\bar{\varepsilon})'\widehat{D}^{-1}LF \\
&- \widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}\frac{1}{n}\sum_{i=1}^{N}(\varepsilon_i-\bar{\varepsilon})(\lambda_i-\bar{\lambda})'(F'\widehat{D}^{-1}LF) \\
&- \widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}\frac{1}{n}\sum_{i=1}^{N}[(\varepsilon_i-\bar{\varepsilon})(\varepsilon_i-\bar{\varepsilon})' - D]\widehat{D}^{-1}LF
\end{aligned} \tag{50}$$

and the third term of (47) is

$$\begin{aligned}
&\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}\Upsilon\widehat{D}^{-1}\widehat{F}(\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}LF) \\
&\quad= \widehat{H}^{-1}(\widehat{F}'\widehat{D}^{-1}F)\frac{1}{n}\sum_{i=1}^{N}(\lambda_i-\bar{\lambda})(\varepsilon_i-\bar{\varepsilon})'\widehat{D}^{-1}\widehat{F}\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}LF \\
&\quad+ \widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}\frac{1}{n}\sum_{i=1}^{N}(\varepsilon_i-\bar{\varepsilon})(\lambda_i-\bar{\lambda})'(F'\widehat{D}^{-1}\widehat{F})\widehat{H}^{-1}(\widehat{F}'\widehat{D}^{-1}LF) \\
&\quad+ \widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}\frac{1}{n}\sum_{i=1}^{N}[(\varepsilon_i-\bar{\varepsilon})(\varepsilon_i-\bar{\varepsilon})' - D]\widehat{D}^{-1}\widehat{F}(\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}LF)
\end{aligned} \tag{51}$$

It can be shown that the third term of (50) and the third term of (51) do not affect the limiting distribution. Next, consider the second term of (50) and that of (51). Notice

$$(F'\widehat{D}^{-1}\widehat{F})\widehat{H}^{-1} = I_r + O_p(1/N) + O_p(1/T) + O_p(\widehat{\alpha} - \alpha) \tag{52}$$

That is, $(F'\widehat{D}^{-1}\widehat{F})\widehat{H}^{-1}$ is essentially an identity matrix. To see this,

$$(F'\widehat{D}^{-1}\widehat{F})\widehat{H}^{-1} = (F-\widehat{F})'\widehat{D}^{-1}\widehat{F}\widehat{H}^{-1} + \widehat{F}'\widehat{D}^{-1}\widehat{F}\widehat{H}^{-1} = (F-\widehat{F})'\widehat{D}^{-1}\widehat{F}\widehat{H}^{-1} + I_r - \widehat{H}^{-1}.$$

Thus, (52) follows from (49) and $\widehat{H}^{-1} = O_p(1/T)$. Thus the sum of the second term in (50) and that in (51) becomes

$$\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}\frac{1}{n}\sum_{i=1}^{N}(\varepsilon_i - \bar{\varepsilon})(\lambda_i - \bar{\lambda})'\Big[O_p(1)[\frac{1}{N} + \frac{1}{T}] + O_p(\widehat{\alpha} - \alpha)\Big](\widehat{F}'\widehat{D}^{-1}LF)$$

which is negligible after multiplying $\sqrt{N/T}$.

This means that the sum of (50) and (51) is equal to the sum of their first terms. Again, using (52), the sum of them is

$$-\frac{1}{n}\sum_{i=1}^{N}(\lambda_i - \bar{\lambda})(\varepsilon_i - \bar{\varepsilon})'\widehat{D}^{-1}LF + \frac{1}{n}\sum_{i=1}^{N}(\lambda_i - \bar{\lambda})(\varepsilon_i - \bar{\varepsilon})'\widehat{D}^{-1}\widehat{F}\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}LF + o_p(\sqrt{T/N})$$

Multiplying by $\sqrt{N/T}$ and taking trace, the previous equation becomes

$$-\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}(\varepsilon_i - \bar{\varepsilon})'\widehat{D}^{-1}LF(\lambda_i - \bar{\lambda}) + \frac{1}{\sqrt{NT}}\sum_{i=1}^{N}(\varepsilon_i - \bar{\varepsilon})'\widehat{D}^{-1}\widehat{F}\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}LF(\lambda_i - \bar{\lambda}) + o_p(1) \tag{53}$$

The above is also equal to the sum of the first three terms in (46) and those in (47), multiplied by $(N/T)^{1/2}$. Most importantly, the above is canceled out with the second and third term of (38). This result is stated in the following lemma:

Lemma A.4 *(i)* $\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}(\varepsilon_i - \bar{\varepsilon})'(D^{-1} - \widehat{D}^{-1})LF(\lambda_i - \bar{\lambda}) = o_p(1)$

(ii) $\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}(\varepsilon_i - \bar{\varepsilon})'\Big[D^{-1}F(H^{-1}F'D^{-1}LF) - \widehat{D}^{-1}\widehat{F}(\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}LF)\Big](\lambda_i - \bar{\lambda}) = o_p(1)$

Proof of (i). Let g_t denote the t-th row of LF. From $D^{-1} - \widehat{D}^{-1} = D^{-1}(\widehat{D} - D)\widehat{D}^{-1}$, the left side of (i) is

$$\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\sum_{i=1}^{N}(\varepsilon_{it} - \bar{\varepsilon}_i)(\widehat{\sigma}_t^2 - \sigma_t^2)g_t'(\lambda_i - \bar{\lambda})/(\sigma_t^2\widehat{\sigma}_t^2)$$

Using the representation for $\widehat{\sigma}_t^2 - \sigma_t^2$ in (20), ignore smaller order terms (also ignore $\bar{\varepsilon}_i$), the above can be written as

$$\frac{1}{\sqrt{NT}}\frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{k=1}^{N}\varepsilon_{it}(\varepsilon_{kt}^2 - \sigma_t^2)g_t'(\lambda_i - \bar{\lambda})/(\sigma_t^4)$$

This expression is $O_p(N^{-1/2})$. To see this, For $i \neq k$, the expected value of the above is zero. For $i = k$, let $\nu_t = E(\varepsilon_{it}^3)$, not necessarily being zero. The expected value of the above $\sum_{t=1}^T \sum_{i=1}^N \nu_t g_t'(\lambda_i - \bar{\lambda})/\sigma_t^4 = 0$ because $\sum_{i=1}^N (\lambda_i - \bar{\lambda}) = 0$. Thus if we let $a_t = \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_{it} g_t'(\lambda_i - \bar{\lambda})/\sigma_t^4$, and $b_t = \frac{1}{\sqrt{N}} \sum_{k=1}^N (\varepsilon_{kt}^2 - \sigma_t^2)$, the above is $N^{-1/2}T^{-1/2} \sum_{t=1}^T a_t b_t$ with $E(a_t b_t) = 0$. We have $T^{-1/2} \sum_{t=1}^T a_t b_t = O_p(1)$. Thus the whole expression is $O_p(N^{-1/2})$, proving (i). In the preceding proof, we have used (20), which holds for pure factor models, see Bai and Li (2012). Now $y_{it} - \alpha y_{it-1}$ is a pure factor model. The convergence rate for $\widehat{\sigma}_t^2 - \sigma_t^2$ is only $\sqrt{N}$, much slower then $\widehat{\alpha}$. Thus the estimation of α does not affect the limiting representation of $\widehat{\sigma}_t^2 - \sigma_t^2$, that is, the representation for pure factor models holds. A rigorous proof can be given (quite involved), we shall not pursue that here.

Proof of (ii). Write

$$D^{-1}F(H^{-1}F'D^{-1}LF) - \widehat{D}^{-1}\widehat{F}(\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}LF) = (D^{-1} - \widehat{D}^{-1})F(H^{-1}F'D^{-1}LF)$$

$$+\widehat{D}^{-1}(F - \widehat{F})(H^{-1}F'D^{-1}LF) + \widehat{D}^{-1}\widehat{F}^{-1}\Big[(H^{-1}F'D^{-1}LF) - (\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}LF)\Big]$$

The left hand side of (ii) can be written as the sum of three expressions corresponding to the above decomposition. The first expression being $o_p(1)$ can be proved in the same way as (i), we simply replace LF by F, also noting that $H^{-1}F'D^{-1}LF = O_p(1)$. For the second expression, replace $\widehat{D}^{-1}$ by D does not affect the analysis because $(\widehat{D}^{-1} - D^{-1})(F - \widehat{F})(H^{-1}F'D^{-1}LF)$ will be a smaller quantity. Now using the expression for $F - \widehat{F}$ in (45), we can show term by term that the second expression is also $o_p(1)$. Finally, consider the third expression. For this, we use the negligibility of

$$R = (H^{-1}F'D^{-1}LF) - (\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}LF) = O_p(1/N) + O_p(1/T) + O_p(\widehat{\alpha} - \alpha) = o_p(1)$$

The third expression is $\frac{1}{\sqrt{NT}} \sum_{i=1}^N (\varepsilon_i - \bar{\varepsilon})'\widehat{D}^{-1}\widehat{F}R(\lambda_i - \bar{\lambda})$, where R is defined as above. We can replace $\widehat{D}^{-1}\widehat{F}$ by $D^{-1}F$ without affecting the analysis. But then

$$\frac{1}{\sqrt{NT}} \sum_{i=1}^N [(\lambda_i - \bar{\lambda})' \otimes (\varepsilon_i - \bar{\varepsilon})'D^{-1}F] = \frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T (\lambda_i - \bar{\lambda})' \otimes f_t(\varepsilon_{it} - \bar{\varepsilon}_i)/\sigma_t^2 = O_p(1)$$

Because each element of R is $o_p(1)$, the third expression is $o_p(1)$, proving (ii). □

It remains to study the last two terms of (46) and (47). They depend on $(\widehat{\alpha} - \alpha)$. Upon multiplied by $\sqrt{N/T}$, terms that are $O_p(1)\sqrt{NT}(\widehat{\alpha} - \alpha)$ will affect the limiting distribution, and terms that are $o_p(1)\sqrt{NT}(\widehat{\alpha} - \alpha)$ are dominated by the left hand side of (35) and will not affect the limiting distribution.

Lemma A.5 *The last two expressions in (46), multiplied by $\sqrt{N/T}$, satisfy*

$$\sqrt{\frac{N}{T}}(\widehat{\alpha}-\alpha)\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}J_T S_n B'\widehat{D}^{-1}LF = \sqrt{NT}(\widehat{\alpha}-\alpha)\frac{1}{T}\widehat{H}^{-1}(\widehat{F}'\widehat{D}^{-1}LF)(F'\widehat{D}^{-1}LF)+o_p(1). \quad (54)$$

$$\sqrt{\frac{N}{T}}(\widehat{\alpha}-\alpha)\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}J_T B S_n J_T'\widehat{D}^{-1}LF = \sqrt{NT}(\widehat{\alpha}-\alpha)\frac{1}{T}\widehat{H}^{-1}(\widehat{F}'\widehat{D}^{-1}F)F'L'\widehat{D}^{-1}LF+o_p(1). \quad (55)$$

Proof: Rewrite (37) as $J_T S_n B' = L\Omega(\theta) + \Upsilon_1$, where Υ_1 represents its last three terms. Then

$$\sqrt{\frac{N}{T}}(\widehat{\alpha}-\alpha)\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}J_T S_n B'\widehat{D}^{-1}LF$$
$$= \sqrt{NT}(\widehat{\alpha}-\alpha)\frac{1}{T}\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}L\Omega(\theta)\widehat{D}^{-1}LF + \sqrt{NT}(\widehat{\alpha}-\alpha)\frac{1}{T}\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}\Upsilon_1\widehat{D}^{-1}LF \quad (56)$$

It can be shown that

$$\frac{1}{T}\text{tr}[\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}\Upsilon_1\widehat{D}^{-1}LF] = o_p(1)$$

thus the second term is $o_p(1)$. Consider the first term of (56), omitting $(\widehat{\alpha}-\alpha)$ for a moment,

$$\frac{1}{T}\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}L\Omega(\theta)\widehat{D}^{-1}LF = \frac{1}{T}\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}LFF'\widehat{D}^{-1}LF + \frac{1}{T}\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}LD\widehat{D}^{-1}LF$$

The second term is $O_p(1/T)$, thus negligible. This proves (54).

For (55), write $BS_nJ_T' = \Omega(\theta)L' + \Upsilon_1'$,

$$\sqrt{\frac{N}{T}}(\widehat{\alpha}-\alpha)\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}J_T BS_nJ_T'\widehat{D}^{-1}LF$$
$$= \sqrt{NT}(\widehat{\alpha}-\alpha)\frac{1}{T}\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}\Omega(\theta)L'\widehat{D}^{-1}LF + \sqrt{NT}(\widehat{\alpha}-\alpha)\frac{1}{T}\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}\Upsilon_1'\widehat{D}^{-1}LF$$

The second term can be shown to be $o_p(1)$, thus negligible. Consider the first term, omitting $(\widehat{\alpha}-\alpha)$ for a moment,

$$\frac{1}{T}\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}\Omega(\theta)L'\widehat{D}^{-1}LF = \frac{1}{T}\widehat{H}^{-1}(\widehat{F}'\widehat{D}^{-1}F)F'L'\widehat{D}^{-1}LF + \frac{1}{T}\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}DL'\widehat{D}^{-1}LF.$$

The second term is $O_p(1/T)$, thus negligible. This gives (55). □

Lemma A.6 *The last two expressions of (47), multiplied by $\sqrt{N/T}$, satisfy*

$$-\sqrt{\frac{N}{T}}(\widehat{\alpha}-\alpha)\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}J_T S_n B'\widehat{D}^{-1}\widehat{F}(\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}LF)$$
$$= -\sqrt{NT}(\widehat{\alpha}-\alpha)\frac{1}{T}\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}LF(F'\widehat{D}^{-1}\widehat{F})\widehat{H}^{-1}(\widehat{F}'\widehat{D}^{-1}LF) + o_p(1) \quad (57)$$

$$-\sqrt{\frac{N}{T}}(\widehat{\alpha}-\alpha)\frac{1}{T}\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}BS_nJ_T'\widehat{D}^{-1}\widehat{F}(\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}LF)$$
$$= -\sqrt{NT}(\widehat{\alpha}-\alpha)\frac{1}{T}\widehat{H}^{-1}(\widehat{F}'\widehat{D}^{-1}F)(F'L'\widehat{D}^{-1}\widehat{F})(\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}LF) + o_p(1) \quad (58)$$

Proof: The proof of (57) is similar to (54). The only difference is the replacement of $\widehat{D}^{-1}LF$ by $\widehat{D}^{-1}\widehat{F}(\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}LF)$ and noting $\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}LF = O_p(1)$. The proof of (58) is similar to (55). □

Corollary A.1 *The sum of (54), (55), (57), and (58) is given by*

$$\sqrt{NT}(\widehat{\alpha}-\alpha)\frac{1}{T}(LF)'\Omega(\widehat{\theta})^{-1}(LF) + o_p(1) \quad (59)$$

Proof: The right hand side of (54) and that of (57) are canceled each other; their sum is $o_p(1)$. This follows from (52), that is, $(F'\widehat{D}^{-1}\widehat{F})\widehat{H}^{-1}$ is essentially an identify matrix. The right hand side of (55) is equal to

$$\sqrt{NT}(\widehat{\alpha}-\alpha)\frac{1}{T}F'L'\widehat{D}^{-1}LF+o_p(1)$$

this is due to (52). And the right hand side of (58) is

$$-\sqrt{NT}(\widehat{\alpha}-\alpha)\frac{1}{T}(F'L'\widehat{D}^{-1}\widehat{F})(\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}LF)+o_p(1)(1)$$

again due to (52). The sum of the preceding two expressions gives the corollary. □

The final proof of Proposition A.1. Our entire analysis thus far shows that the left hand side of (39) is composed of the terms in (53) and in Corollary A.1, all other terms are negligible. Lemma A.4 shows that the terms in (53) are asymptotically equivalent to the last two terms on the right hand side of (38). This completes the proof of Proposition A.1. □

Proof of Proposition A.2. Notice

$$\begin{aligned}J_TS_nJ_T' &= L\Omega(\theta)L'+LF\frac{1}{n}\sum_{i=1}^{N}(\lambda_i-\bar{\lambda})(\varepsilon_i-\bar{\varepsilon})'L' \\ &+ L\frac{1}{n}\sum_{i=1}^{N}(\varepsilon_i-\bar{\varepsilon})(\lambda_i-\bar{\lambda})'F'L'+L\frac{1}{n}\sum_{i=1}^{N}[(\varepsilon_i-\bar{\varepsilon})(\varepsilon_i-\bar{\varepsilon})'-D]L'\end{aligned}$$

Let Υ_2 represents the last three terms. The left hand side of (35), omitting $(\widehat{\alpha}-\alpha)$, is

$$\frac{1}{T}\mathrm{tr}\Big(J_TS_nJ_T'\Omega(\widehat{\theta})^{-1}\Big)=\frac{1}{T}\mathrm{tr}\Big(L\Omega(\theta)L'\Omega(\widehat{\theta})^{-1}\Big)+\frac{1}{T}\mathrm{tr}\Big(\Upsilon_2\Omega(\widehat{\theta})^{-1}\Big)$$

It is easy to show that the second term on the right is dominated by the first. But

$$L\Omega(\theta)L'\Omega(\widehat{\theta})^{-1}=L(FF'+D)L'\Omega(\widehat{\theta})^{-1}=LFF'L'\Omega(\widehat{\theta})^{-1}+LDL'\Omega(\widehat{\theta})^{-1}$$

$$=LFF'L'\Omega(\widehat{\theta})^{-1}+LDL'\widehat{D}^{-1}-LDL'\widehat{F}\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1}$$

Taking trace and dividing by T, we have

$$\frac{1}{T}\mathrm{tr}(L\Omega(\theta)L'\Omega(\widehat{\theta})^{-1})=\frac{1}{T}\mathrm{tr}(F'L'\Omega(\widehat{\theta})^{-1}LF)+\frac{1}{T}\mathrm{tr}(LDL'\widehat{D}^{-1})+O_p(\frac{1}{T})$$

where we have used

$$\frac{1}{T}\mathrm{tr}(LDL'\widehat{F}\widehat{H}^{-1}\widehat{F}'\widehat{D}^{-1})=\frac{1}{T}\mathrm{tr}[(\widehat{F}'\widehat{D}^{-1}LDL'\widehat{F})H^{-1}]=O_p(1/T).$$

Note that $\frac{1}{T}\mathrm{tr}(LDL'\widehat{D}^{-1})=\frac{1}{T}\mathrm{tr}(LDL'D^{-1})+o_p(1)$. This proves the proposition. □

Proof of Theorem 2 The limit of $\sqrt{NT}(\widehat{\alpha}-\alpha)$ follows from representation (18). The expected value of $\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\varepsilon_i'D^{-1}L\varepsilon_i$ is equal to $(NT)^{-1/2}\mathrm{tr}(D^{-1}LD)=(NT)^{-1/2}\mathrm{tr}(L)=0$ since L is lower

triangular. The variance is the second moment, which is equal to $T^{-1}\text{tr}(LDL'D^{-1})$ and its limit is γ in (19). The theorem is a result of the central limit theorem.

The limiting distribution of $\sqrt{N}(\widehat{\sigma}_t^2 - \sigma_t^2)$ follows from the representation (20) and the Central Limit Theorem for the cross-section sequence $\varepsilon_{it}^2 - \sigma_t^2$ by Assumption B1. Note the variance of $\varepsilon_{it}^2 - \sigma_t^2$ is $\sigma_t^4(2 + \kappa_t)$. □.

Proof of (21), (22), and (23). These representations hold for pure factor models (no lagged dependent variable and with diagonal idiosyncratic covariance matrix), and they are derived in Bai and Li (2012). Noting that the residual $y_{it} - \alpha y_{it-1}$ follows a pure factor model. So if α is known, these representations hold. Since the convergence rate for $\widehat{\alpha} - \alpha$ is much faster than those of $\widehat{\sigma}_t^2 - \sigma_t^2$, $\widehat{f}_t - f_t$ and $\widehat{\Psi} - \Psi_n$, the estimation of α does not affect these representations. □

Proof of Proposition 2. We use representations (21) and (22). For $t \leq r$, f_t is known thus not estimated under IC1. For $t > r$, ε_{it} is independent of ξ_i, so the two terms on the right hand side of (22) are independent. The asymptotic variance of $\sqrt{N}(\widehat{f}_t - f_t)$ is the sum of the two asymptotic variances. From $\xi_i = (\varepsilon_{i1}, ..., \varepsilon_{ir})'$, the variance of $[\frac{1}{\sqrt{N}}\sum_{i=1}^N(\lambda_i - \bar{\lambda})\xi_i']f_t$ is equal to $\Psi_n f_t' D_r f_t$, where $D_r = \text{diag}(\sigma_1^2, ..., \sigma_r^2)$. Thus the variance of the first term on the right of (22) is $\Psi_n^{-1} f_t' D_r f_t$. The variance of the second term is $\Psi_n^{-1}\sigma_t^2$. Noting that $\Psi_n \to \Psi$, the limiting variance of $\sqrt{N}(\widehat{f}_t - f_t)$ is the sum of the two terms.

Next, let A be the first term on the right hand side of (21), the second term is A'. Using $\text{vech}(A + A') = \mathcal{D}_r^+\text{vec}(A + A') = 2\mathcal{D}_r^+\text{vec}(A)$ since $\mathcal{D}_r^+\text{vec}(A) = \mathcal{D}_r^+\text{vec}(A')$, we have $\sqrt{N}\,\text{vech}(\widehat{\Psi} - \Psi_n) = 2\mathcal{D}_r^+\text{vec}(A) + o_p(1)$. But $\text{vec}[(\lambda_i - \bar{\lambda})\xi_i'] = [I_r \otimes (\lambda_i - \bar{\lambda})]\xi_i$ and the limiting covariance matrix of $N^{-1/2}\sum_{i=1}^N[I_r \otimes (\lambda_i - \bar{\lambda})]\xi_i$ is $D_r \otimes \Psi$. This obtains the limiting distribution for $\sqrt{N}(\widehat{\Psi} - \Psi_n)$ as stated. □

Proof of Lemma 2. Recall that $\widehat{\Psi} = \text{diag}[T^{-2}\widehat{F}'\widehat{D}^{-1}\widehat{B}S_n\widehat{B}'\widehat{D}^{-1}\widehat{F} - T^{-1}I_r]$ under IC2. Here we shall provide the key insight instead of an elaborate proof, which is very lengthy. The last two terms on the right hand side of (24) are identical to the representations in Bai and Li (2012) for a pure factor model. The first term is due to the estimation of α. That is, if α, or equivalently B, is known, and if we define

$$\widetilde{\Psi} = \text{diag}[T^{-2}\widehat{F}'\widehat{D}^{-1}BS_nB'\widehat{D}^{-1}\widehat{F} - T^{-1}I_r]$$

then $\sqrt{NT}(\widetilde{\Psi} - \Psi_n)$ will be given by the last two terms of (24). It remains to argue that if B is replaced by $\widehat{B}$, we obtain the first term on the right side of (24). That is, we shall show

$$\text{diag}\left[T^{-2}\widehat{F}'\widehat{D}^{-1}\big(\widehat{B}S_n\widehat{B}' - BS_nB'\big)\widehat{D}^{-1}\widehat{F}\right] = -2(\widehat{\alpha} - \alpha)\,\text{diag}[\Psi_n(\frac{1}{T}F'L'D^{-1}F)] + (\widehat{\alpha} - \alpha)o_p(1).$$

Write $\widehat{B} = \widehat{B} - B + B = -(\widehat{\alpha} - \alpha)J_T + B$ and use $\text{diag}(A) = \text{diag}(A')$, we have

$$\text{diag}\left[T^{-2}\widehat{F}'\widehat{D}^{-1}\big(\widehat{B}S_n\widehat{B}' - BS_nB'\big)\widehat{D}^{-1}\widehat{F}\right] = -2(\widehat{\alpha} - \alpha)\,\text{diag}[T^{-2}\widehat{F}'\widehat{D}^{-1}BS_nJ_T'\widehat{D}^{-1}\widehat{F}]$$

$$+(\widehat{\alpha} - \alpha)^2\,\text{diag}[T^{-2}\widehat{F}\widehat{D}^{-1}J_TS_nJ_T'\widehat{D}^{-1}\widehat{F}]$$

The last expression is negligible since it involves $(\widehat{\alpha}-\alpha)^2$. Replacing $\widehat{F}$ and $\widehat{D}$ by F and D on the right hand side does not affect the analysis. So we need to show

$$T^{-2}F'D^{-1}BS_nJ_T'D^{-1}F = T^{-1}\Psi_nF'L'D^{-1}F + o_p(1)$$

The above equality does not involve any estimated quantity and can be verified with algebraic operation together with $\Gamma' J_T' = L'$ and $F'D^{-1}F/T = I_r$ (identification restriction). This completes the proof of Lemma 2. □.

Proof of Proposition 3 . The limit for $\widehat{f_t}$ is already discussed in the main text. Consider the limit of $\widehat{\Psi}$. Under normality, the three terms on the right of (24) are asymptotically independent, so we only need to find out the limiting covariance matrix for each term. The limiting variance of the first term is $4hh'/\gamma$ by Theorem 2 and the definition of h. For the second term, by the definition of $\mathcal{P}_r$, $\text{diag}(A) = \mathcal{P}_r\text{vec}(A)$ for all A, so $\text{diag}(\Psi_n f_t f_t'/\sigma_t^2) = \mathcal{P}_r(I\otimes\Psi_n)(f_t\otimes f_t)\frac{1}{\sigma_t^2}$. Also, the variance of $(\varepsilon_{it}^2-\sigma_t^2)/\sigma_t^2$ is equal to 2 under normality. So the covariance matrix of the second term is $2\mathcal{P}_r(I\otimes\Psi_n)[\frac{1}{T}\sum_{t=1}^T\sigma_t^{-4}(f_tf_t'\otimes f_tf_t')](I\otimes\Psi_n)\mathcal{P}_r'$. The third term is similar. From $\text{vec}[(\lambda_i-\bar{\lambda})f_t'/\sigma_t] = \sigma_t^{-1}f_t\otimes(\lambda_i-\bar{\lambda})$ and unit variance for $\varepsilon_{it}/\sigma_t$, the covariance matrix of the third term is $4\mathcal{P}_r[(\frac{1}{T}\sum_{t=1}^T\sigma_t^{-2}f_tf_t')\otimes\Psi_n]\mathcal{P}_r'$. Taking limits and using Assumption 3, we obtain the limiting covariance matrix in Proposition 3. □

Proof of Theorem 3. The proof is contained in the main text.

Proof of Theorem 4. The proof is lengthy and is given in the supplementary document.

References

[1] Ahn, S.G., Y.H. Lee and P. Schmidt (2001): "GMM Estimation of Linear Panel Data Models with Time-varying Individual Effects," *Journal of Econometrics*, 102, 219-255.

[2] Ahn, S.G., Y.H. Lee and P. Schmidt (2013): "Panel Data Models with Multiple Time-varying Effects," *Journal of Econometrics*, 174, 1-14.

[3] Alvarez, J. and M. Arellano (2003): "The Time Series and Cross-Section Asymptotics of Dynamic Panel Data Estimators." *Econometrica* 71, 1121-1159.

[4] Alvarez, J. and M. Arellano (2004): "Robust likelihood estimation of dynamic panel data models." Unpublished manuscript, CEMFI.

[5] Amemiya, T. (1985): *Advanced Econometrics*, Harvard University Press, Cambridge, MA.

[6] Amemiya, Y., W.A. Fuller, and S.G. Pantula (1987), The Asymptotic Distributions of Some Estimators for a Factor Analysis Model, *Journal of Multivariate Analysis*, 22, 51-64.

[7] Anderson, T.W. and Y. Amemiya (1988). The asymptotic normal distribution of estimators in factor analysis under general conditions, *Annals of Statistics*, 16 759-771

[8] Anderson, T.W., and C. Hsiao (1981): "Estimation of dynamic models with error components," *Journal of American Statistical Association*, 76, 598-606.

[9] Anderson, T.W., and C. Hsiao (1982): "Formulation and estimation of dynamic Models with Error Components," *Journal of Econometrics*, 76, 598-606.

[10] Anderson, T.W. and H. Rubin (1956): "Statistical Inference in Factor Analysis," in J. Neyman, ed., *Proceedings of Third Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Vol 5, 111-150.

[11] Arellano, M. (2003): *Panel Data Econometrics*, Oxford University Press.

[12] Bai, J. (2009): "Panel data models with interactive fixed effects," *Econometrica*, 77 1229-1279.

[13] Bai, J. (2013). Fixed effects dynamic panel data models, a factor analytical method, *Econometrica,* 81, 285-314.

[14] Bai, J. and K.P. Li (2012). Statistical analysis of factor models of high dimension. *Annals of Statistics,* 40, 436-465.

[15] Baltagi, B.H. (2005): *Econometric Analysis of Panel Data*, John Wiley: Chichester.

[16] Bhargava, A. and J.D. Sargan (1983): "Estimating Dynamic Random Effects Models from Panel Data Covering Short Time Periods," *Econometrica*, 51, 1635-1659.

[17] Blundell R. and S. Bond (1998). Initial conditions and moment restrictions in dynamic panel data models, *Journal of Econometrics*, 87 115-143.

[18] Blundell R. and R.J. Smith (1991). Initial conditions and efficient estimation in dynamic panel data models. *Annales d'Economie et de Statistique* 20/21, 109-123.

[19] Chamberlain, G. (1982): "Multivariate regression models for panel data," *Journal of Econometrics*, 18, 5-46.

[20] Chamberlain, G. (1984): "Panel Data," in *Handbook of Econometrics*, Vol. 2, ed. by Z. Griliches and M. Intriligator. Amsterdam: North-Holland

[21] Chamberlain, G. and M.J. Moreira (2009), Decision Theory Applied To A Linear Panel Data Model, *Econometrica* 77, 107-133.

[22] Dahm, P.F. and W.A. Fuller (1986): "Generalized Least Squares Estimation of the Functional Multivariate Linear Errors-in-variables Model," *Journal of Multivariate Analysis* 19, 132-141.

[23] Dempster, A.P. N.M. Laird, and D.B. Rubin (1977), Maximum likelihood from incomplete data via the EM algorithm. *Journal of The Royal Statistical Society B*, 39, 1-38.

[24] Doz, C., Giannone, D. and L Reichlin (2008). A quasi maximum likelihood approach for large approximate dynamic factor models. ECARES and CEPR.

[25] Engle, R., D.F. Hendry, and J.F. Richard (1983): "Exogeneity," *Econometrica,* 51, 277-304.

[26] Hahn, J., and G. Kuersteiner (2002): "Asymptotically Unbiased Inference for a Dynamic Panel Model with Fixed Effects when Both n and T Are Large," *Econometrica*, 70, 1639-1657.

[27] Holtz-Eakin, D., W. Newey, and H. Rosen (1988): "Estimating Vector Autoregressions with Panel Data", *Econometrica*, 56, 1371-1395.

[28] Hsiao, C. (2003): *Analysis of Panel Data.* Cambridge University Press, New York.

[29] Iwakura, H. and R. Okui (2012). Asymptotic Efficiency in Dynamic Panel Data Models with Factor Structure, unpublished manuscript, Institute of Economic Research, Kyoto University.

[30] Jungbacker, B. and S.J. Koopman (2008). Likelihood-based analysis for dynamic factor models, memio.

[31] Kiviet, J. (1995): "On Bias, Inconsistency, and Efficiency of Various Estimators in Dynamic Panel Data Models", *Journal of Econometrics,* 68, 53-78.

[32] Kruiniger, H. (2008). Maximum likelihood estimation and inference methods for the covariance stationary panel AR(1)/unit root model. *Journal of Econometrics*, 447-464.

[33] Lancaster, T. (2000): "The incidental parameter problem since 1948," *Journal of Econometrics*, 95 391-413.

[34] Lancaster, T. (2002): "Orthogonal parameters and panel data," *Review of Economics Studies.* 69 647-666.

[35] Lawley, D.N. and A.E. Maxwell (1971): *Factor Analysis as a Statistical Method*, London: Butterworth.

[36] Liu, C. and D.B. Rubin (1994). The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. *Biometrika*, 81, 633-648.

[37] Magnus, J.R. and H. Neudecker (1999): *Matrix Differential Calculus with Applications in Statistics and Econometrics*, Wiley: New York.

[38] McLachlan, G.J, and T. Krishnan (1996): *The EM Algorithm and Extensions*, Wiley, New York.

[39] Meng, X.L and D.B. Rubin (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika,* 80(2), 267-278.

[40] Moffitt, R., and P. Gottschalk (2002): "Trends in the Transitory Variance of Earnings in the United States," *The Economic Journal,* 112, C68-C73.

[41] Moon H.R. and M. Weidner (2010a). "Dynamic Linear Panel Regression Models with Interactive Fixed Effects", unpublished manuscript, USC.

[42] Moon, H.R. and M. Weidner (2010b). "Linear regression for panel with unknown number of factors as interactive fixed effects," unpublished manuscript, USC.

[43] Moreira, M.J. (2009), A Maximum Likelihood Method for the Incidental Parameter Problem, *Annals of Statistics*, 37, 3660-3696.

[44] Mundlak, Y. (1978): "On the pooling of time series and cross section data," *Econometrica,* 46, 69-85.

[45] Newey, W. and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in Engle, R.F. and D. McFadden (eds.) *Handbook of Econometrics,* North Holland.

[46] Neyman, J., and E. L. Scott (1948): "Consistent Estimates Based on Partially Consistent Observations," *Econometrica*, 16, 1-32.

[47] Nickell, S. (1981): "Biases in Dynamic Models with Fixed Effects," *Econometrica*, 49, 1417-1426.

[48] Pesaran, M. H. (2006): "Estimation and Inference in Large Heterogeneous panels with a Multifactor Error Structure," *Econometrica*, 74, 967-1012.

[49] Proietti, T. (2008). Estimation of common factors under cross-sectional and temporal aggregation constraints: nowcasting monthly GDP and its main components. MPRA Paper 6860, University Library of Munich, Germany.

[50] Quad, Q. and T. Sargent (1993). A Dynamic Index Model for Large Cross Sections. CEP Discussion Paper No. 0132.

[51] Rao, C.R. and S.K. Mitra (1971). *Generalized Inverse of Matrices and its Applications*, Wiley: New York.

[52] Rubin, D.B. and D.T. Thayer (1982). EM algorithm for ML factor analysis. *Psychometrika*, 47 69-76.

[53] Shapiro, A. (1986). Asymptotic theory of overparameterized structural models. *Journal of the American Statistical Association* 81, 142-149.

[54] Sims, C.A. (2000). Using a likelihood perspective to sharpen econometric discourse: Three examples. *Journal of Econometrics,* 95 443-462.

[55] Stock, J. H. and M. W. Watson (2011): Dynamic Factor Models, *The Oxford Handbook of Economic Forecasting*, Edited by M.P. Clements and D.F. Hendry, Oxford University Press.

[56] Su, L. and Q. Chen (2013): Testing homogeneity in panel data models with interactive fixed effects. Department of Economics, SMU, Singapore. Forthcoming in *Econometric Theory.*

[57] Su, L., S. Jin, and Y. Zhang (2013): Specification test for panel data models with interactive fixed effects. Department of Economics, SMU, Singapore.

[58] van der Vaart, A.W. and J.A. Wellner (1996): *Weak Convergence and Empirical Processes.* Springer, New York.

[59] Watson, M.W. and R.F. Engle (1983): Alternative algorithms for the estimation of the dynamic factor, MIMIC, and varying coefficient regression models. *Journal of Econometrics,* Vol. 23, pp. 385-400.

[60] Westerlund, J. and J.P. Urbain (2013): On the estimation and inference in factor-augmented panel regressions with correlated loadings. *Economics Letters*, 119(3), 247-250.

Note: the supplementary document can be found at

http://mpra.ub.uni-muenchen.de/50267/15/MPRA_paper_50267.pdf