

Likelihood approach to small T dynamic panel models with interactive effects

Jushan Bai
Preliminary and incomplete*

June, 2009

Abstract

This paper considers dynamic panel models with a factor analytic error structure that is correlated with the regressors. The model is rooted in both micro and macro econometrics. In microeconomics, for example, the observed wage is a function of observable characteristics and unobserved innate ability. The innate ability is potentially correlated with the observables, and moreover, is priced at each period (or is repriced occasionally) so its effect on wage is time varying. In macroeconomics, the motivation is that unobserved common shocks have heterogeneous effects on the cross-section units, and resources employed in the production are also affected by the common shocks. Each consideration leads to a factor analytic error structure that will be correlated with the explanatory variables.

A dynamic panel model constitutes a simultaneous equations system with T equations. With a small number of periods (T), consistent estimation requires a suitable formulation of the reduced form for the first observation and appropriate control for the correlation between the effects and the regressors. We use the method of Mundlak (1978) and Chamberlain (1982) to control the latter correlation. Under the factor error structure, the system implies parameter constraints between the mean vector and the covariance matrix. We explore the constraints through a quasi-FIML approach. The EM algorithm and its variants ECM and ECME and the associated closed-form solutions are elaborated for the maximization problem. With a small T , the factor process is treated as parameters and it can have arbitrary dynamics. The model also permits unit roots.

Key words and phrases: factor structure, dynamic panel, heterogeneity, interactive effects, exogeneity, FIML.

*I thank seminar participants at Princeton University for helpful comments. This paper is supported in part by an NSF grant (No. SES 0551275)

1 Introduction

In this paper we examine a panel data model of the form

$$y_{it} = \alpha y_{it-1} + x'_{it}\beta + \delta_t + \lambda'_i f_t + \varepsilon_{it}$$
$$i = 1, 2, \dots, N; t = 1, 2, \dots, T$$

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 microeconomics, 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 such as education. It is further assumed that the innate ability is priced at each period (or is repriced occasionally) so its effect on wage is time varying, as captured by f_t . In macroeconomics, the vector f_t would be regarded as common shocks, and they have heterogeneous effects on each cross-section unit via the individual-specific coefficient λ_i . The consequence for each motivation is a factor analytic error structure that is correlated with the regressors. We consider multiple factors to allow wages to be affected by other unobservable individual traits such as dedication and perseverance in the earnings study, and more than one common shock in the macro setting.

This paper focuses on the situation in which the number of cross-section units (N) is large and the number of time periods is fixed.¹ Under large N and large T , Bai (2005) considers the nonlinear least squares estimation of the model, treating both λ_i and f_t as parameters. Even though x_{it} is correlated with both λ_i and f_t , \sqrt{NT} consistency for β is obtainable. The nonlinear least squares approach has a within-group interpretation, and it is effective (and computationally easy) in controlling the unobserved effects under large N and large T . However, it is shown in Bai (2005) that the bias associated with serial correlation and time series heteroskedasticity in ε_{it} is of $O_p(1/T)$, and thus under fixed T , the estimator is inconsistent, also see Ahn et al (2001). This paper considers a small T framework that is robust to heteroskedasticity and serial correlation. As in Alvarez and Arellano (2005), heteroskedasticity itself may be an object of interest, and we provide a direct estimate of the changing variance.

¹The procedure developed in this paper is equally valid for large T , but the elaborate treatment for the initial conditions becomes less crucial.

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, Lee and Schmidt (2006) also consider the quasi-difference approach and GMM for general r . These methods are consistent under fixed T . With a moderate T , the number of moments can be large and increases rapidly as T increases. The likelihood approach considered here implicitly makes use of efficient combinations, at least under normality, of a large number of moments. Also, the maximum likelihood approach effectively explores many of the restrictions implied by the model.

We consider a quasi-FIML approach that treats the dynamic panel as a simultaneous equations system with T equations, assuming λ_i to be random and conditionally normal (conditional on the regressors). The approach is consistent despite nonnormality. A key feature of the model 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. In addition, because T is small, it is desirable to treat f_t as parameters instead of treating λ_i as parameters. Furthermore, as parameters, the process generating f_t can have arbitrary dynamics, be either a linear or a broken trend itself, or a random walk process, or a stationary process. The procedure is still valid under large T , but requires more computing time than the nonlinear least squares approach.

We provide a careful treatment of the initial observation, as it is key to consistent estimation with a small T . Both the reduced-form and the conditional approaches are considered. We study systems with strictly exogenous, weakly exogenous, and predetermined regressors and examine the regressors' relationships with the unobservable individual effects and the way in which the relationship affects the likelihood function.

The rest of the paper is organized as follows. Section 2 considers a panel model without lagged dependent variables. Section 3 introduces lagged dependent variables but additional regressors are strictly exogenous with respect to the disturbances. Section 4 considers dynamic panel with predetermined regressors. All models allow correlation between the effects and the regressors. Heteroskedasticity is maintained throughout. Section 5 examines an iterated GLS method for estimation and Section 6 studies the FIML procedure, implemented via the EM algorithm and its extensions. Simulation experiments are reported in Section 7 and the last section concludes.

2 Model basics and Assumptions

We first examine the model in the absence of lagged dependent variables, and focus our attention on controlling the correlation between the regressors and the effects. Consider

$$y_{it} = \delta_t + x'_{it}\beta + \lambda'_i f_t + \varepsilon_{it}$$

Because the $r \times 1$ vectors λ_i and f_t are unobservable and enter the model in a multiplicative way we need to impose r^2 restrictions. Let F be the $T \times r$ factor matrix

$$F = (f_1, f_2, \dots, f_T)'$$

We normalize F so its first $r \times r$ block is an identity matrix,

$$F = \begin{bmatrix} I_r \\ F_2 \end{bmatrix}$$

implying exactly r^2 restrictions. There is no loss of generality by assuming

$$E(\lambda_i) = 0$$

for if $E(\lambda_i) = \mu \neq 0$, rewrite $\lambda'_i f_t = (\lambda_i - \mu)' f_t + \mu' f_t$ and combine δ_t with $\mu' f_t$, and rename $\lambda_i - \mu$ as λ_i . Throughout, we assume $T \geq r$ and r is fixed and is given.

By stacking the observations over t , we can rewrite as

$$y_i = \delta + x_i \beta + F \lambda_i + \varepsilon_i$$

where $y_i = (y_{i1}, \dots, y_{iT})'$ is $T \times 1$, $x_i = (x_{i1}, \dots, x_{iT})'$ is $T \times p$, and ε_i is similarly define. When λ_i is uncorrelated with the regressors, we can use the generalized least squares (GLS) to estimate the model, treating $F \lambda_i + \varepsilon_i$ as the aggregate error. Because of the correlation between λ_i and the regressor (the primary focus of the paper) GLS is inconsistent. Following Chamberlain (1982), we assume

$$E(\lambda_i | x_{i1}, x_{i2}, \dots, x_{iT}) = \lambda + \sum_{s=1}^T \phi_s x_{is} \quad (1)$$

where λ is $r \times 1$ and ϕ_s is an $r \times p$ matrix ($s \geq 1$). We can view the above as a linear projection, the argument still goes through. The above full-projection on the entire path of x has many parameters to estimate unless T is small. We also consider a restricted projection as in Mundlak (1978),

$$E(\lambda_i | x_{i1}, x_{i2}, \dots, x_{iT}) = \lambda + \phi \bar{x}_i \quad (2)$$

where $\bar{x}_i = \frac{1}{T} \sum_{t=1}^T x_{it}$, and ϕ is $r \times p$. This is equivalent to $\phi_t = \phi/T$ for all t . Depending on which projection to use, we write

$$\lambda_i = \lambda + \sum_{s=1}^T \phi_s x_{is} + \eta_i$$

or

$$\lambda_i = \lambda + \phi \bar{x}_i + \eta_i$$

where by definition, $E(\eta_i | x_{i1}, \dots, x_{iT}) = 0$. We further assume the conditional variance of η_i is constant, denoted by Φ . It is possible to allow conditional heteroskedasticity $\text{var}(\eta_i | x_{i1}, \dots, x_{iT}) = \Phi(x_i, \tau)$, with a finite number of parameters τ . But for simplicity, constant conditional variance is assumed.

Under the Mundlak projection, the model can be rewritten as

$$y_{it} = (\delta_t + \lambda f_t) + x'_{it} \beta + \bar{x}'_i \phi' f_t + \eta'_i f_t + \varepsilon_{it}$$

Still denote $\delta_t + \lambda f_t$ as δ_t to economize the notation and stack the observations over t we have

$$y_i = \delta + x_i \beta + F \phi \bar{x}_i + F \eta_i + \varepsilon_i$$

The model is nonlinear in parameters because of the term $F \phi$.

We now state the assumptions imposed on the model

Assumption A. $(x_i, \eta_i, \varepsilon_i)$ are iid over i . The rank of the matrix $E(x'_i x_i)$ is of p , the number of regressors.

Assumption B. η_i and ε_i are independent normal random vectors, and they are independent of x_i .

It is reasonable to make an iid assumption over the cross sections for many data collection schemes, see Stock and Watson (2003, p105). While λ_i is correlated with the regressors, η_i as the projection error is uncorrelated with the regressors. So Assumption B is in fact not very strong. But assuming the regressors are uncorrelated with the vector ε_i rules out dynamic models, which will be studied in the next section.

With normality, we denote

$$\eta_i \sim N(0, \Phi), \quad \varepsilon_i \sim N(0, D)$$

where D is a positive definite matrix. In general, D is diagonal, or a parametrized according to an ARMA process for ε_{it} . The matrix D can be completely unrestricted if the parameter

of interest is β only (not F and ϕ). With a factor structure, we can provide a more efficient estimation of β . For simplicity, we assume D is diagonal throughout. The conditional mean of y_i is

$$E(y_i|x_i) = \delta + x_i\beta + F\phi\bar{x}_i$$

and the conditional variance is

$$\text{var}(y_i|x_i) = \Omega := F\Phi F' + D$$

Let $\theta = (\delta, \beta, \phi, F, \Phi, D)$, the likelihood function is

$$\ell(\theta) = -\frac{N}{2} \log \det(\Omega) - \frac{1}{2} \sum_{i=1}^N u_i' \Omega^{-1} u_i \quad (3)$$

where

$$u_i = y_i - \delta - x_i\beta - F\phi\bar{x}_i$$

Directly maximizing the above likelihood function is nontrivial even when u_i is observable, see Lawley and Maxwell (1973) and Anderson (1984). In our case, u_i are residuals with unknown parameters, and there exist parameter constraints between the mean and the variance since F occurs in both. Furthermore, nonlinearity in parameters occurs in the mean. In a later section, we consider two estimation schemes, one is iterated GLS, and the other is quasi-FIML, implemented via the EM algorithm.

We note that under the factor errors, the coefficients of time invariant regressors can be consistently estimated. Under the usual additive fixed effects, consistent estimation of these coefficients requires a two-step procedure with suitable instruments, see Hausman and Taylor (1981).

Remark 1 Suppose that x_{it} is correlated with λ_i such that $x_{it} = g_t'\lambda_i + \xi_{it}$, where ξ_{it} are iid over t . Then the Mundlak projection under fixed T fails to account for the correlation between the effects and the regressors, unless g_t is time invariant such that $g_t = g$ for all t . To see this, suppose that λ_i is scalar, $\lambda_i \sim N(0, \sigma_\lambda^2)$ and x_{it} is also a scalar with $\xi_{it} \sim N(0, \sigma_\xi^2)$. From the joint normality, we have $E(\lambda_i|x_i) = \sigma_\lambda^2 G'(\sigma_\lambda^2 G G' + \sigma_\xi^2 I_T)^{-1} x_i$, where $G = (g_1, g_2, \dots, g_T)'$. Unless G is a scalar multiple of $\nu_T = (1, 1, \dots, 1)'$, the conditional expectation cannot be written in terms of \bar{x}_i . This means that for any ϕ , $\eta_i = \lambda_i - \phi\bar{x}_i$ will still be correlated with x_i . Under large T , using the assumption $T^{-1/2} \sum_{t=1}^T (g_t - \bar{g}) = O_p(1)$, the conditional expectation can be approximated by $\phi\bar{x}_i$ for some ϕ with an $O_p(T^{-1/2})$ approximating error, implying large- T consistency of model estimation. In comparison,

the Chamberlain projection works under either fixed or large T when x_{it} follows a factor structure.

Despite Remark 1, for notational simplicity, we shall use the Mundlak projection in our exposition with the understanding that it may stand for the Chamberlain projection when necessary.

Joint modeling of y and x . If one is willing to specify the data generating process for the regressors x , then the y and x equations can be modeled jointly without the need for the Mundlak or Chamberlain projection. Suppose that

$$x_{it} = \delta_{xt} + g_t' \lambda_i + \xi_{it}$$

where δ_{xt} (vector) and g_t ($p \times r$ matrix) are time-varying parameters; ξ_{is} and ε_{it} are independent for all t and s so that x_{it} remains to be strictly exogenous with respect to the regression errors ε_{it} . We assume the individual effects λ_i are common for the y and x equations. This is for notational simplicity. Otherwise, let λ_i denote the pooled individual effects from the y and x equations, then some of elements of f_t and of g_t are zeros, which can be imposed when estimating the model. Let $z_{it} = (y_{it}, x_{it}')'$, $\delta_{zt} = (\delta_t, \delta_{xt}')'$, $\zeta_{it} = (\varepsilon_{it}, \xi_{it}')'$ and $\pi_t = (f_t, g_t)$, we have

$$\Gamma z_{it} = \delta_{zt} + \pi_t' \lambda_i + \zeta_{it}.$$

This is a special simultaneous equations system with

$$\Gamma = \begin{bmatrix} 1 & -\beta' \\ 0 & I_p \end{bmatrix}.$$

Stacking the observations over t , we have

$$(I_T \otimes \Gamma) z_i = \delta_z + \Pi \lambda_i + \zeta_i$$

where z_i stacks z_{it} over t , and Π stacks π_t' over t , and similarly for δ_z and ζ_i . The likelihood function of the parameters is

$$-\frac{N}{2} \log |\Omega_z| - \frac{1}{2} \sum_{i=1}^N u_i' \Omega_z^{-1} u_i$$

where $u_i = (I_T \otimes \Gamma) z_i - \delta_z$, $\Omega_z = \Pi \Phi \Pi' + D_z$, $\Phi = \text{var}(\lambda_i)$, and D_z is the covariance matrix of ζ_i , a diagonal matrix. The determinant of $I_T \otimes \Gamma$ is one, so the Jacobian term does not enter.

3 Dynamic panel with strictly exogenous regressors

With the exception of the interactive effects, the model considered in this section is identical to the models of Chapter 7 in Arellano (2003) with $T + 1$ observations

$$y_{it} = \alpha y_{i,t-1} + \delta_t + x'_{it}\beta + f'_t\lambda_i + \varepsilon_{it}$$

$$t = 0, 1, 2, \dots, T; \quad i = 1, 2, \dots, N$$

We do not assume a known data generating process for x_{it} . The main assumption is that

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

which implies the strict exogeneity of x_{it} with respect to ε_{it} ; but x_{it} is still allowed to be correlated with the effects λ_i . This assumption does not rule out serial correlation in ε_{it} . However, for simplicity, we assume ε_{it} is serially uncorrelated. If necessary, more lags of y_{it} can be added to account for the serial correlation.

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.

3.1 Regressors uncorrelated with the effects

This section considers the situation in which λ_i is uncorrelated with x_{it} for all t . We relabel λ_i as η_i to be consistent with previous notations and signify the lack of correlation. Even though the regressors x_{it} are uncorrelated with the fixed effects η_i , the lagged dependent variables y_{it} are still correlated with the effects.

We first derive the reduced form for the first observation

$$y_{i0} = \frac{1}{1 - \alpha} \delta_0 + \sum_{j=0}^{\infty} \alpha^j x'_{i,-j} \beta + \sum_{j=0}^{\infty} \alpha^j f'_{-j} \eta_i + \sum_{j=0}^{\infty} \alpha^j \varepsilon_{i,-j}$$

We project the second term on the right hand side onto x_i

$$E\left(\sum_{j=0}^{\infty} \alpha^j x'_{i,-j} \beta | x_{i1}, \dots, x_{iT}\right) = \sum_{s=1}^T x'_{i,s} \psi_{0,s}$$

so we can rewrite the equation for y_{i0} as

$$y_{i0} = \delta_0^* + \sum_{s=0}^T x'_{i,s} \psi_{0,s} + f_0^* \eta_i + \varepsilon_{i,0}^* = \delta_0^* + w'_i \psi_0 + f_0^* \eta_i + \varepsilon_{i,0}$$

where $\delta_0^* = \delta_0/(1 - \alpha)$ is a free parameter, and $f_0^* = \sum_{j=0}^{\infty} \alpha^j f_{-j}$, and $\varepsilon_{i,0}^*$ is summation of the projection error and $\sum_{j=0}^{\infty} \alpha^j \varepsilon_{i,-j}$,

$$w_i = \text{vec}(x'_i), \quad \psi_0 = (\psi'_{0,0}, \dots, \psi'_{0,T})'$$

The system of $T + 1$ equations becomes

$$\begin{aligned} y_{i0} &= \delta_0^* + w'_i \psi_0 + f_0^* \eta_i + \varepsilon_{i,0}^* \\ y_{it} &= \alpha y_{i,t-1} + \delta_t + x'_{it} \beta + f'_t \eta_i + \varepsilon_{it}, \quad t = 1, 2, \dots, T \end{aligned}$$

The second block of equations can be written as

$$B_1 y_i^+ = \delta + x_i \beta + F \eta_i + \varepsilon_i$$

where B_1 is the matrix B defined below with the first row deleted. Since $x_i \beta = (I_T \otimes \beta') \text{vec}(x'_i) = (I_T \otimes \beta') w_i$ (e.g., Magnus and Neudecker, 1999, p. 31), the whole system can be written as

$$B y_i^+ = \Gamma w_i + \delta^+ + F^+ \eta_i + \varepsilon_i^+ \tag{4}$$

where

$$B = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ -\alpha & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & -\alpha & 1 \end{bmatrix}, \quad \Gamma = \begin{bmatrix} \psi'_0 \\ I_T \otimes \beta' \end{bmatrix}$$

Again we normalize the first $r \times r$ block of the factors as an identity matrix, $F^+ = (I_r, F_2^+)'$. Introduce

$$\Omega^+ = F^+ \Phi F^{+'} + D^+$$

and let

$$u_i^+ = B y_i^+ - \Gamma w_i - \delta^+$$

the log-likelihood function for $(y_{i0}, y_{i1}, \dots, y_{iT})$ is

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

Since the determinant of B is equal to 1 the Jacobian term does not enter. The EM algorithm for this model is discussed in the next subsection as the model here can be considered as a special case.

3.2 Regressors correlated with the effects

Projecting λ_i on x_i ,

$$\lambda_i = \lambda + \phi \bar{x}_i + \eta_i.$$

The y_{i0} equation remains the same (by renaming the parameters since all are free parameters). That is,

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

The rest of the equations are

$$y_{it} = \alpha y_{it-1} + x_{it}' \beta + F \phi \bar{x}_i + \delta_t + F \eta_i + \varepsilon_{it}, \quad t \geq 1$$

Since $F \phi \bar{x}_i = F \phi \frac{1}{T} x_i' \iota_T = (\iota_T' \otimes F \phi \frac{1}{T}) w_i$ with $w_i = \text{vec}(x_i')$, the model has the same form as (4), namely

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

but here

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

Under Chamberlain's full projection, we replace $\iota_T' \otimes F \phi \frac{1}{T}$ by $F(\phi_1, \phi_2, \dots, \phi_T)$, a $T \times (rp)$ matrix. The likelihood function for this $(T+1)$ simultaneous equations system has the same form as (5).

It is interesting to examine the complete data likelihood, which is used for the EM algorithm later on. Assuming η_i is observable, the complete data likelihood function (ignore the likelihood of η_i itself) is

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

$$\begin{aligned}
&= -\frac{N}{2} \ln |D| - \frac{1}{2} \sum_{i=1}^N (u_i - F\eta_i)' D^{-1} (u_i - F\eta_i) \\
&\quad - \frac{N}{2} \ln \sigma_0^{*2} - \frac{1}{2} \frac{1}{\sigma_0^{*2}} \sum_{i=1}^N (u_{i0}^* - f_0^{*'} \eta_i)^2
\end{aligned}$$

where

$$u_i = y_i - y_{i(-1)}\alpha - \delta - x_i\beta - F\phi\bar{x}_i$$

and

$$u_{i0}^* = y_{i0} - \delta_0^* - w_i'\psi_0$$

The first term is otherwise identical to the complete data likelihood under strictly exogenous regressors (except the presence of an additional regressor $y_{i(-1)}$). The first term alone will not produce consistent estimators under fixed T since y_{i0} is correlated with η_i so y_{i0} is endogenous and enters into the y_{i1} equation as a regressor. The second term provides the bias correction so that the estimator will be consistent under fixed T .

3.3 Likelihood conditional on y_{i0}

An alternative approach to the full likelihood for the entire sequence $(y_{i0}, y_{i1}, \dots, y_{iT})$ 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 pure time series analysis by conditioning the first observation owing to the presence of individual effects. Since λ_i is correlated with y_{i0} , we project λ_i on y_{i0} in addition to \bar{x}_i such that

$$\lambda_i = \lambda + \phi\bar{x}_i + \phi_0 y_{i0} + \eta_i^*$$

The model can be written as ($t = 1, 2, \dots, T$)

$$y_{it} = \alpha y_{it-1} + x_{it}'\beta + f_t'\phi\bar{x}_i + f_t\phi_0 y_{i0} + \delta_t + f_t'\eta_i^* + \varepsilon_{it}$$

Using the notation as in Arellano (2003, page 99), we have

$$B^* y_i = \alpha y_{i0} e_1 + x_i'\beta + F\phi\bar{x}_i + F\phi_0 y_{i0} + \delta + F\eta_i^* + \varepsilon_{it}$$

where B^* is equal to B with the first row and first column deleted, and $e_1 = (1, 0, \dots, 0)'$. 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

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

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

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

and

$$u_{it} = y_{it} - \alpha y_{it-1} - x'_{it}\beta - f'_t\phi\bar{x}_i - f'_t\phi_0 y_{i0} - \delta_t$$

The complete likelihood is the same as the earlier case with two more regressors $y_{i,t-1}$ and y_{i0} . We can rewrite $f'_t\phi\bar{x}_i + f'_t\phi_0 y_{i0}$ as $f'_t\phi^*\bar{x}_i^*$ with $\phi^* = (\phi, \phi_0)$, and $\bar{x}_i^* = (\bar{x}'_i, y_{i0})'$. The residuals are

$$u_{it} = y_{it} - x'^*_{it}\beta^* - f'_t\phi^*\bar{x}_i^* - \delta_t$$

where $x^*_{it} = (y_{it-1}, x'_{it})'$ and $\beta^* = (\alpha, \beta)'$. The likelihood is identical to the case of Section 2, except that we have lagged dependent variable as one of regressors, and that we have the first observation y_{i0} as an extra regressor on its own.

4 Dynamic panel with predetermined regressors

This section considers the model

$$y_{it} = \alpha y_{i,t-1} + x'_{it}\beta + f'_t\lambda_i + \varepsilon_{it}$$

under the assumption that

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

where $y_i^t = (y_{i0}, \dots, y_{it})'$ and $x_i^t = (x_{i1}, \dots, x_{it})'$. 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 . It should be clear that ε_{it} are necessarily uncorrelated over time. The model extends that of Arellano (2003, Chapter 8), who does not consider the maximum likelihood estimation with predetermined regressors.

4.1 Weakly exogenous dynamic regressors

The concept of weak exogeneity is carefully 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. 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

$$x_{it} = \alpha_x x_{i,t-1} + \beta_x y_{i,t-1} + g_t' \tau_i + \xi_{it} \quad (6)$$

where α_x ($p \times p$) and β_x ($p \times 1$) are unknown parameters (but not the parameters of the interest). In addition, τ_i and λ_i are independent,² ε_{it} is independent of ξ_{it} , and 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 $y_{i,t-1}$. We also allow arbitrary correlation between x_{i0} (initial endowment) and the individual effect λ_i . Obviously, more lags can be entertained, but the key idea is the same.

It is not difficult to show that x_{it} in (6) is weakly exogenous with respect to the parameters in the y equation. The part of joint density function³ of (y_i, x_i) that involves the parameter of interest is given by

$$\ell_1(y_i, x_i | y_{i0}, x_{i0}) = -\frac{N}{2} \ln |\Omega^*| - \frac{1}{2} \sum_{i=1}^N u_i' \Omega^{*-1} u_i \quad (7)$$

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

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

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

($t = 1, 2, \dots, T$). The likelihood function is similar to that of Section 3.3, here the individual effects λ_i are projected onto the initial value of x_{i0} instead of the entire path $(x_{i0}, x_{i1}, \dots, x_{iT})$. Again, the factor process F occurs in both the mean and variance. The parameters can be estimated by the EM algorithm.

To verify (7), let $w_i = \text{vec}(x_i')$, a vector that stacks up x_{it} ($t = 1, 2, \dots, T$). 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^* was introduced earlier and

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

²It is sufficient to assume conditional independence, conditional on the initial observation (y_{i0}, x_{i0}) .

³More specifically, the conditional joint density, conditional on the initial observation.

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

Since the matrix

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

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 $((F\lambda_i + \varepsilon_i)', (G\tau_i + \xi_i)')$, 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

$$f(y_i, x_i | y_{i0}, x_{i0}) = f(F\lambda_i + \varepsilon_i | y_{i0}, x_{i0}) \cdot f(G\tau_i + \xi_i | y_{i0}, x_{i0}) \quad (8)$$

where f denotes a density function. Equation (7) 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 2 Equation (7) 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 (7) is the conditional density. See Engle et al (1983).

Remark 3 The likelihood function (7) is simple, and is in fact simpler than that of Section 3 under the strict exogeneity assumption. This is because under strict exogeneity, the process of x_{it} is unspecified, and to account for the correlation between the effects and the regressors, full path projection of λ_i on x_i is required. Under weak exogeneity together with the specification of the data generating process for x_{it} , it is sufficient to account for the correlation between the effects (λ_i) and the initial observation x_{i0} only.

4.2 Non-weakly exogenous dynamic regressors

We consider a similar process for x_{it} . However, we now permit correlation between λ_i and τ_i and correlation between ε_{it} and ξ_{it} . For notional simplicity, we assume τ_i and λ_i are identical. In order for the regressor x_{it} to be predetermined, we rewrite the y equation by lagging the x by one period such that

$$y_{it} = \alpha y_{i,t-1} + \beta' x_{i,t-1} + f'_t \lambda_i + \varepsilon_{it}$$

and

$$x_{it} = \alpha_x x_{i,t-1} + \beta_x y_{i,t-1} + g'_t \lambda_i + \xi_{it}$$

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. The VAR approach is most suitable for this setup. Let

$$z_{it} = \begin{bmatrix} y_{it} \\ x_{it} \end{bmatrix}, \quad A = \begin{bmatrix} \alpha & \beta' \\ \beta_x & \alpha_x \end{bmatrix}, \quad \pi'_t = \begin{bmatrix} f'_t \\ g'_t \end{bmatrix}, \quad \zeta_{it} = \begin{bmatrix} \varepsilon_{it} \\ \xi_{it} \end{bmatrix}$$

Then

$$z_{it} = Az_{it-1} + \pi'_t \lambda_i + \zeta_{it}$$

This model is a direct extension 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} ($t = 1, 2, \dots, T$) 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 $\Omega^* = \Pi \Phi^* \Pi' + \Sigma$ with $\Phi^* = \text{var}(\eta_i^*)$, and Σ is block diagonal such that $\Sigma = \text{diag}(\Sigma_1, \dots, \Sigma_T)$, $e_i = (e'_{i1}, e'_{i2}, \dots, e'_{iT})'$ with

$$e_{it} = z_{it} - Az_{it-1} - \pi'_t \phi_0 z_{i0}$$

($t = 1, 2, \dots, T$). Like all previous cases, the expanded factor matrix Π appears in both the mean and variance. This conditional likelihood function is the simplest, at least in form, among those discussed so far in this paper. 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 estimation is more time consuming in comparison with the weakly exogenous case as a result of parameter estimation for the x equations. This will, of course, be desirable when the parameters of the x equations are of direct interest.

5 Estimation via Iterated GLS

Given the appropriate handling of the first observation (for dynamic panels) and the control of correlation between the effects and regressors, the model can be estimated by iterated generalized least squares (GLS). The GLS approach is appealing since it does not rely on distributional (density) assumptions.

The requirement for iterations stems from the covariance matrix being unknown and nonlinearity in parameters. We consider a special iterated GLS that separately estimates the mean parameters and the covariance parameters. The separation simplifies the computation and accelerates the convergence. Consider the model of Section 2,

$$y_i = \delta + x_i\beta + F\phi\bar{x}_i + F\eta_i + \varepsilon_i$$

Suppose that F and $\Omega = F\Phi F' + D$ are known, then the mean parameters $\gamma = (\delta', \beta', \text{vec}(\phi)')$ can be estimated by

$$\hat{\gamma} = \sum_{i=1}^N \left(X_i' \Omega^{-1} X_i \right)^{-1} \sum_{i=1}^N X_i \Omega^{-1} y_i$$

where $X_i = (I_T, x_i, (\bar{x}_i' \otimes F))$, a matrix of $T \times [(T + p) + r \cdot p]$ that depends on F . The corresponding residuals are

$$\hat{u}_i = y_i - X_i \hat{\gamma}$$

which provide estimates for $u_i = F\eta_i + \varepsilon_i$ $i = 1, 2, \dots, N$). Using the residuals, we estimate F and the covariance matrix Ω by the factor analysis. The new estimated F and Ω are used to estimate γ again, and this procedure is iterated until convergence. The above approach is a combination of GLS and factor analysis. To estimate the factor model, we use a recently proposed procedure by Zhao et al (2008), who provide a fast and stable algorithm for the maximum likelihood estimation of pure factor models.

The estimation of F does not take into account any information contained in the mean and is based entirely on the covariance matrix. Therefore, this approach is not fully efficient. On the other hand, if $\phi = 0$ (this is the case when the effects λ_i are uncorrelated with the regressors), the mean contains no information about F , no efficiency is lost. Of course, if it is known that the regressors are uncorrelated with the effects ($\phi = 0$), we would set $\phi = 0$.

Remark 4 Because the mean and the variance contain common parameters, some version of continuously updated GLS may be used. This issue remains to be investigated. A less efficient way of estimation is to let $H = F\phi$ ($T \times p$), a matrix of free parameters (without linking to F) and to use GLS to obtain estimates for β , δ and H . One can use the covariance matrix $\hat{\Omega} = \frac{1}{N} \sum_{i=1}^N \hat{u}_i' \hat{u}_i$ without resort to factor analysis. Under the full projection (1), the matrix H will contain $O(T^2)$ parameters. The large number of parameters leads to efficiency loss unless T is very small.

The next section considers the FIML approach that takes into account constraints between the mean and variance. The GLS obtained here provides starting value for the FIML.

6 Estimation via the Maximum Likelihood

The FIML approach here is different from the GLS since they use different first order conditions. 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 Dempster 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. For ease of exposition, the ECM procedure is elaborated for the model of Section 2 and the likelihood function (3). All other models are similar.

The complete data likelihood is (assuming η_i is observable)

$$\begin{aligned} L(\theta) &= -\frac{N}{2} \ln |D| - \frac{1}{2} \sum_{i=1}^N (u_i - F\eta_i)' D^{-1} (u_i - F\eta_i) \\ &\quad - \frac{N}{2} \ln |\Phi| - \frac{1}{2} \sum_{i=1}^N \text{tr}(\Phi^{-1} \eta_i \eta_i') \\ &= -\frac{N}{2} \ln |D| - \frac{1}{2} \sum_{i=1}^N \left[u_i' D^{-1} u_i - 2u_i' D^{-1} F \eta_i + \text{tr}(F' D^{-1} F \eta_i \eta_i') \right] \\ &\quad - \frac{N}{2} \ln |\Phi| - \frac{1}{2} \sum_{i=1}^N \text{tr}(\Phi^{-1} \eta_i \eta_i') \end{aligned}$$

where $u_i = y_i - \delta - x_i \beta - F \phi \bar{x}_i$. Evaluated at the true parameters

$$u_i = F\eta_i + \varepsilon_i$$

The $T + r$ vector $(u_i', \eta_i')'$ is normally distributed such that

$$\begin{pmatrix} u_i \\ \eta_i \end{pmatrix} = N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} F\Phi F' + D & F\Phi \\ \Phi F' & \Phi \end{pmatrix} \right]$$

Thus

$$E(\eta_i | u_i) = \Phi F' \Omega^{-1} u_i$$

and

$$E(\eta_i \eta_i' | u_i) = E(\eta_i | u_i) E(\eta_i | u_i)' + \Phi - \Phi F' \Omega^{-1} F \Phi$$

The expected likelihood is

$$Q(\theta) = -\frac{N}{2} \ln |D| - \frac{1}{2} \sum_{i=1}^N \left[u_i' D^{-1} u_i - 2u_i' D^{-1} F E(\eta_i | u_i) + \text{tr}[F' D^{-1} F E(\eta_i \eta_i' | u_i)] \right]$$

$$-\frac{N}{2} \ln |\Phi| - \frac{1}{2} \sum_{i=1}^N \text{tr}[\Phi^{-1} E(\eta_i \eta_i' | u_i)]$$

Denote $\widehat{\eta}_i = E(\eta_i | u_i)$ and $\widehat{\eta_i \eta_i'} = E(\eta_i \eta_i' | u_i)$. We write the Q function as $Q(\theta | \bar{\theta})$ when the expectations are taken assuming $\bar{\theta}$ is the true parameter.

First order conditions. Note that

$$u_i' D^{-1} u_i = v_i' D^{-1} v_i - 2v_i' D^{-1} F \phi \bar{x}_i + \text{tr}(F' D^{-1} F \phi \bar{x}_i \bar{x}_i' \phi')$$

where

$$v_i = y_i - \delta - x_i \beta$$

Thus

$$\begin{aligned} \frac{\partial u_i' D^{-1} u_i}{\partial F} &= -2D^{-1} v_i \bar{x}_i' \phi' + 2D^{-1} F (\phi \bar{x}_i \bar{x}_i' \phi') \\ -2 \frac{\partial u_i' D^{-1} F \widehat{\eta}_i}{\partial F} &= 2D^{-1} F \widehat{\eta}_i \bar{x}_i' \phi' - 2D^{-1} u_i \widehat{\eta}_i' \\ &= 2D^{-1} F (\widehat{\eta}_i \bar{x}_i' \phi' + \phi \bar{x}_i \widehat{\eta}_i') - 2D^{-1} v_i \widehat{\eta}_i' \\ \frac{\partial \text{tr}(F' D^{-1} F \widehat{\eta_i \eta_i'})}{\partial F} &= 2D^{-1} F \widehat{\eta_i \eta_i'} \end{aligned}$$

This gives

$$\frac{\partial Q}{\partial F} = \sum_{i=1}^N D^{-1} \left[v_i (\bar{x}_i' \phi' + \widehat{\eta}_i') - F (\phi \bar{x}_i \bar{x}_i' \phi' + \widehat{\eta}_i \bar{x}_i' \phi' + \phi \bar{x}_i \widehat{\eta}_i' + \widehat{\eta_i \eta_i'}) \right]$$

$$\frac{\partial Q}{\partial D^{-1}} = \frac{N}{2} D - \frac{1}{2} \sum_{i=1}^N \left[u_i u_i' - 2F \widehat{\eta}_i u_i' + F (\widehat{\eta_i \eta_i'}) F' \right]$$

$$\frac{\partial Q}{\partial \Phi^{-1}} = \frac{N}{2} \Phi - \frac{1}{2} \sum_{i=1}^N \widehat{\eta_i \eta_i'}$$

$$\frac{\partial Q}{\partial \delta} = D^{-1} \sum_{i=1}^N (u_i - F \widehat{\eta}_i)$$

Rewrite

$$F \phi \bar{x}_i = (\bar{x}_i' \otimes F) \text{vec}(\phi)$$

Let

$$\theta_1 = (\beta', \text{vec}(\phi)')$$

we have

$$u_i = y_i - \delta - x_i\beta - F\phi\bar{x}_i = y_i - \delta - [x_i, (\bar{x}_i' \otimes F)]\theta_1 = y_i - \delta - X_i\theta_1$$

$$\frac{\partial Q}{\partial \theta_1} = \sum_{i=1}^N \left[X_i' D^{-1} (y_i - \delta - X_i\theta_1) - X_i' D^{-1} F \widehat{\eta}_i \right]$$

Setting the first order conditions to zero, we obtain

$$\begin{aligned} F &= \sum_{i=1}^N v_i (\bar{x}_i' \phi' + \widehat{\eta}_i') \left[\sum_{i=1}^N \left(\phi \bar{x}_i \bar{x}_i' \phi' + \widehat{\eta}_i \bar{x}_i' \phi' + \phi \bar{x}_i \widehat{\eta}_i' + \widehat{\eta}_i \widehat{\eta}_i' \right) \right]^{-1} \\ \delta &= \frac{1}{N} \sum_{i=1}^N (y_i - x_i\beta - F\phi\bar{x}_i - F\widehat{\eta}_i) \\ D &= \text{diag} \left[\frac{1}{N} \sum_{i=1}^N \left(u_i u_i' - 2F\widehat{\eta}_i u_i' + F\widehat{\eta}_i \widehat{\eta}_i' F' \right) \right] \\ \Phi &= \frac{1}{N} \sum_{i=1}^N \widehat{\eta}_i \widehat{\eta}_i' \end{aligned}$$

and

$$\theta_1 = \left[\sum_{i=1}^N X_i' D^{-1} X_i \right]^{-1} \sum_{i=1}^N \left[X_i' D^{-1} (y_i - \delta - F\widehat{\eta}_i) \right].$$

The first $r \times r$ block of F is set to I_r due to the restrictions.

Conditional Maximization. The first order solutions for $\theta = (F, D, \Phi, \delta, \beta, \phi)$ 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 since the sequential conditional maximizations have closed form solutions. Suppose the parameters are divided into two groups $\theta = (\theta_1, \theta_2)$. The objective 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 that of θ_2 alone. The problem becomes a constrained/conditional maximization (CM)

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

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

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

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 this extra expectation step does not necessarily accelerate the convergence, and for some cases, it can be detrimental. 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 $\theta_3 = (F, \Phi)$, $\theta_2 = (\delta, D)$, and $\theta_1 = (\beta', \text{vec}(\phi)')$. 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 k th step solution $\theta^{(k)}$, the ECM solution for $\theta^{(k+1)}$ can now be stated:

$$F^{(k+1)} = \sum_{i=1}^N v_i^{(k)} \left(\bar{x}_i' \phi^{(k)'} + \widehat{\eta}_i \right) \left[\sum_{i=1}^N \left(\phi^{(k)} \bar{x}_i \bar{x}_i' \phi^{(k)'} + \widehat{\eta}_i \bar{x}_i' \phi^{(k)'} + \phi^{(k)} \bar{x}_i \widehat{\eta}_i + \widehat{\eta}_i \widehat{\eta}_i' \right) \right]^{-1}$$

$$\Phi^{(k+1)} = \frac{1}{N} \sum_{i=1}^N \widehat{\eta}_i \widehat{\eta}_i'$$

$$\delta^{(k+1)} = \frac{1}{N} \sum_{i=1}^N \left(y_i - x_i \beta^{(k)} - F^{(k+1)} \phi^{(k)} \bar{x}_i - F^{(k+1)} \widehat{\eta}_i \right)$$

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

where $u_i^{(k+1/2)}$ is the updated residual after the CM1 step,

$$u_i^{(k+1/2)} = y_i - \delta^{(k+1)} - x_i \beta^{(k)} - F^{(k+1)} \phi^{(k)} \bar{x}_i$$

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-form solution is

$$\theta_1^{(k+1)} = \left[\sum_{i=1}^N X_i^{(k+1)'} (D^{(k+1)})^{-1} X_i^{(k+1)} \right]^{-1} \sum_{i=1}^N \left[X_i^{(k+1)'} (D^{(k+1)})^{-1} \left(y_i - \delta^{(k+1)} - F^{(k+1)} \widehat{\eta}_i \right) \right]$$

where

$$X_i^{(k+1)} = [x_i, \bar{x}_i' \otimes F^{(k+1)}].$$

The conditional expectations $\widehat{\eta}_i$ and $\widehat{\eta}_i \widehat{\eta}_i'$ are taken with respect to $\theta^{(k)}$:

$$\begin{aligned} \widehat{\eta}_i &= E(\eta_i | u_i^{(k)}, \theta^{(k)}) = \Phi^{(k)} F^{(k)'} (\Omega^{(k)})^{-1} u_i^{(k)} \\ \widehat{\eta}_i \widehat{\eta}_i' &= \widehat{\eta}_i \widehat{\eta}_i' + \Phi^{(k)} - \Phi^{(k)} F^{(k)'} (\Omega^{(k)})^{-1} F^{(k)} \Phi^{(k)} \end{aligned}$$

with

$$u_i^{(k)} = y_i - \delta^{(k)} - x_i \beta^{(k)} - F^{(k)} \phi^{(k)} \bar{x}_i$$

Having obtained $\theta^{(k+1)}$, we can compute $u^{(k+1)}$ and the conditional expectations $E(\eta_i | u^{(k+1)}, \theta^{(k+1)})$ and $E(\eta_i \eta_i' | u^{(k+1)}, \theta^{(k+1)})$, and then $\theta^{(k+2)}$. The process is continued until convergence.

Remark 5. 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

$$\theta_1^{(k+1)} = \left[\sum_{i=1}^N X_i^{(k+1)'} (\Omega^{(k+1)})^{-1} X_i^{(k+1)} \right]^{-1} \sum_{i=1}^N \left[X_i^{(k+1)'} (\Omega^{(k+1)})^{-1} \left(y_i - \delta^{(k+1)} \right) \right].$$

Remark 6. We could divide the parameters into two groups by combining θ_3 and θ_2 . This requires joint maximization over F and δ (note D will also be obtained given F and δ , and Φ does not depend on F and δ). Joint maximization is achieved by expanding the factor space and factor loadings, $F^+ = (\delta, F)$, and $\eta_i^+ = (1, \eta_i')'$. We can easily solve for F^+ from the original first order conditions for F and δ . The solution for F^+ depends on the conditional mean and the conditional second moments of η_i^+ , which are $\widehat{\eta}_i^+ = (1, \widehat{\eta}_i')'$ and

$$\widehat{\eta}_i^+ \widehat{\eta}_i^{+'} = \begin{bmatrix} 1 & \widehat{\eta}_i' \\ \widehat{\eta}_i & \widehat{\eta}_i \widehat{\eta}_i' \end{bmatrix}$$

respectively.

Starting values for iterations There are a number of ways of starting the iterations. One way is to use the iterated GLS estimators. The other is to use the simple least squares to obtain δ and β , and the residuals. Then use the principal components method to obtain F , Φ , and D from the residuals. The third way is as follows. The first r equations are

$$\begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{ir} \end{pmatrix} = \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_r \end{pmatrix} + \begin{pmatrix} x'_{i1} \\ x'_{i2} \\ \vdots \\ x'_{ir} \end{pmatrix} \beta + \phi \bar{x}_i + \eta_i + \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{ir} \end{pmatrix}$$

Cross-section regression using the first r equations with the SUR of Zellner (1960), we can obtain β and ϕ , only ϕ will be retained. Then using OLS to estimate the equation

$$y_i = \delta + x_i \beta + F(\phi \bar{x}_i) + u_i$$

to obtain δ , β , and F , and the residuals (F will be discarded). Using the residual

$$u_i = F \eta_i + \varepsilon_i$$

and the principal component method to obtain F , and D and $\Phi = \text{var}(\eta_i)$. Another method is to use GLS to estimate δ, β and H (see Remark 4) and using the residuals to estimate F and Φ and D , and finally $\phi = (F'F)^{-1}F'G$.

The EM algorithm (Dempster et al, 1977) has been widely used for estimating factor models, e.g., Rubin and Thayer (1982), Watson and Engle (1983), and Quah and Sargent (1993). Also see the monograph by McLachlan and Krishnan (1997). Most recent applications of the EM algorithm for factor models include Doz et al (2008), Proietti (2008), and Jungbacker and Koopman (2008). Most of these studies either do not consider the presence of regressors or do not consider potential correlations between the regressors and the effects. These studies also assume large T so that initial conditions are less important and thus either not modeled or stationarity is assumed. Focusing on the dynamic property of f_t (via Kalman smoother) and treating factor loadings as parameters will only give consistent estimation under large T . In fact, when the factor loadings are considered as parameters (there are many of them), under fixed T and large N , incidental parameter problem will occur. The framework taken here is appropriate for panel data with a small T although it also works for large T . It also allows arbitrary dynamic process for f_t .

There exists a large literature on factor models we have omitted for discussion. Interested readers are referred to Forni et al. (2000) and Stock and Watson (2002). This literature

focuses on consistent extraction of the common components rather than consistent estimation of model parameters. Large N and large T are assumed.

7 Preliminary Numerical results

Data are generated according to ($r = 1$):

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

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

$$k = 1, 2$$

where $\lambda_i, f_t, \varepsilon_{it}, \xi_{it}$ are all iid $N(0,1)$; $\beta_1 = 1, \beta_2 = 2, \sigma_t = \sqrt{t}$ so that the D matrix is

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

We set δ_t to zero, but time dummies are included in the model. Three estimators are considered:

1. OLS with time dummies without taking into account the correlation between the effects λ_i and x_{it} .

2. GLS with time dummies, under Chamberlain's full projection (1). The estimator is based on the method in Remark 4.

3. MLE with time dummies, with Chamberlain's full projection.

The results reported in the table below are based on 1000 repetitions.

N	T	OLS		GLS		MLE	
		$\beta_1 = 1$	$\beta_2 = 2$	$\beta_1 = 1$	$\beta_2 = 2$	$\beta_1 = 1$	$\beta_2 = 2$
100	5	1.2360	2.2397	1.0166	2.0094	1.0365	2.0343
		0.0700	0.0697	0.2461	0.2537	0.1226	0.1225
500	5	1.4112	2.4097	1.0049	2.0045	1.0105	2.0081
		0.0280	0.0283	0.1014	0.0987	0.0710	0.0718
100	10	1.2644	2.2636	1.0265	2.0136	1.0238	2.0233
		0.0715	0.0678	0.3747	0.3772	0.1053	0.0965
500	10	1.2124	2.2119	1.0107	2.0094	1.0001	1.9992
		0.0299	0.0308	0.1436	0.1412	0.0268	0.0276

For each N and T configuration, the first row displays the sample mean from the 1000 repetitions, and the second row gives the standard errors (not mean squared errors). The OLS is inconsistent as expected since it does not take into account the correlation between the regressors and the effects. GLS is consistent but it is less efficient than MLE. This can be seen from the standard errors. It is interesting to note that as T increases the GLS becomes less efficient since the standard errors also grow. The large number of parameters being estimated (T^2) under GLS is responsible for its deteriorating performance. This GLS estimator is based on the approach mentioned in Remark 4.

The MLE estimates become more precise as either N or T increases. The data generating process for x_{it} (also follows a factor structure) requires full projection of λ_i on the entire path of x . The Mundlak projection is inconsistent, see Remark 1.

Dynamic panel. The y process is generated as

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

all other variables are generated the same way as in the non-dynamic case. However, for each T , we simulate $2T$ observations and then discard the first half. The variance of ε_{it} is set to 1 for the first half, and for the second half, its variance is set to

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

So the retained sample is heteroskedastic. Two different values for α are considered: $\alpha = 0.5$ and $\alpha = 1.0$. The table below is based on 200 repetitions.

			OLS			GLS			MLE		
α	N	T	α	$\beta_1 = 1$	$\beta_2 = 2$	α	$\beta_1 = 1$	$\beta_2 = 2$	α	$\beta_1 = 1$	$\beta_2 = 2$
0.5	100	5	0.528	1.184	2.191	0.514	1.121	2.122	0.513	1.054	2.062
			0.019	0.071	0.078	0.064	0.228	0.222	0.037	0.085	0.100
	500	5	0.499	1.308	2.308	0.497	1.005	1.997	0.499	1.001	2.003
			0.008	0.030	0.030	0.025	0.101	0.106	0.012	0.032	0.035
	100	10	0.495	1.288	2.282	0.498	1.026	1.990	0.501	1.010	2.010
			0.016	0.073	0.067	0.039	0.378	0.384	0.023	0.081	0.080
	500	10	0.501	1.330	2.336	0.500	1.022	2.022	0.498	1.000	2.004
			0.006	0.027	0.027	0.018	0.150	0.142	0.009	0.028	0.028
1.0	100	5	0.989	1.174	2.180	1.003	1.077	2.081	1.003	1.036	2.040
			0.005	0.073	0.078	0.048	0.234	0.232	0.027	0.080	0.091
	500	5	1.002	1.308	2.308	1.001	1.006	1.997	1.000	1.001	2.003
			0.005	0.030	0.030	0.020	0.101	0.105	0.006	0.032	0.035
	100	10	0.994	1.282	2.275	0.998	1.022	1.993	1.000	1.014	2.015
			0.005	0.074	0.068	0.024	0.380	0.378	0.008	0.086	0.087
	500	10	0.995	1.326	2.332	1.002	1.022	2.021	1.000	1.000	2.004
			0.003	0.026	0.027	0.010	0.150	0.141	0.003	0.028	0.028

Again, for each N and T combination, the first row shows the mean and the second row shows the standard deviation (from 200 repetitions). It is interesting to note that the OLS estimator for the autoregressive coefficient α exhibits no sign of bias. This is probably due to the large magnitude of the lag regressor, which is in turn due to the large heteroskedasticity. The OLS estimated slope coefficients for the exogenous regressors are highly biased. The GLS are consistent, showing no bias for all coefficients, but the standard deviations are huge (this particular GLS is not efficient). For the MLE, α is again estimated well. For $N = 100$, there are some small bias for the slope coefficients, but for large N , biases are reduced even for $T = 5$. As T increases, not much improvement is seen for the estimated slope coefficients (in terms of standard errors) largely due to the increased heteroskedasticity.

Estimated heteroskedasticities for the dynamic model

We only reported the case of $T = 10$. The case of $T = 5$ is estimated less well, especially for $N = 100$ and $T = 5$ (there are some outliers). The true value for σ_t^2 is equal to t , as given in the first column of the table. The values in parentheses are standard deviations.

σ_t^2	$\alpha = 0.5$				$\alpha = 1$			
	$N = 100$		$N = 500$		$N = 100$		$N = 500$	
1	1.001	(0.175)	0.996	(0.066)	0.991	(0.180)	1.003	(0.079)
2	1.985	(0.304)	2.115	(0.172)	1.972	(0.307)	2.085	(0.184)
3	2.953	(0.428)	3.038	(0.187)	2.928	(0.438)	3.027	(0.198)
4	3.965	(0.583)	4.006	(0.261)	3.949	(0.573)	4.009	(0.261)
5	4.886	(0.731)	5.020	(0.336)	4.889	(0.733)	5.023	(0.337)
6	5.834	(0.874)	6.041	(0.395)	5.842	(0.877)	6.036	(0.395)
7	6.611	(1.051)	7.028	(0.473)	6.654	(1.036)	7.018	(0.475)
8	7.858	(1.152)	8.022	(0.466)	7.868	(1.135)	8.030	(0.466)
9	8.907	(1.220)	8.999	(0.581)	8.892	(1.231)	9.003	(0.581)
10	9.818	(1.417)	9.935	(0.621)	9.814	(1.407)	9.941	(0.629)

A note on computation. Much improvements can be made for the program used in this computation: different starting values, the maximum number of EM iterations (currently set at 1000), different number of parameter groups in the conditional maximization (CM), and the use of ECME.

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 (2006): "Panel Data Models with Multiple Time-varying Effects," mimeo, Arizona State University.
- [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 (2005): Robust likelihood estimation of dynamic panel data models. Unpublished manuscript, CEMFI.
- [5] Anderson, T. W. 1984, *An Introduction to Multivariate Statistical Analysis*, Wiley, New York.
- [6] Anderson, T.W., and C. Hsiao (1982): "Formulation and estimation of dynamic Models with Error Components," *Journal of Econometrics*, 76, 598-606.
- [7] 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.
- [8] Andrews, D. W. K. (2005): "Cross-section Regression with Common Shocks." *Econometrica*, 73, 1551-1585.
- [9] Arellano, M. (2003): *Panel Data Econometrics*, Oxford University Press.
- [10] Arellano, M., and B. Honore (2001): "Panel Data Models: Some Recent Developments," in *Handbook of Econometrics*, Vol. 5, ed. by J. J. Heckman and E. Leamer. Amsterdam: North-Holland.
- [11] Bai, J. (2005): "Panel data models with interactive fixed effects," forthcoming in *Econometrica*.
- [12] Baltagi, B.H. (2005): *Econometric Analysis of Panel Data*, John Wiley: Chichester.
- [13] Bhargava, A. and J.D. Sargan (1983): "Estimating Dynamic Random Effects Models from Panel Data Covering Short Time Periods," *Econometrica*, 51, 1635-1659.
- [14] Chamberlain, G. (1980): "Analysis of Covariance with Qualitative Data," *Review of Economic Studies*, 47, 225-238.
- [15] Chamberlain, G. (1982): "Multivariate regression models for panel data" *Journal of Econometrics*, 18, 5-46.
- [16] Chamberlain, G. (1984): "Panel Data," in *Handbook of Econometrics*, Vol. 2, ed. by Z. Griliches and M. Intriligator. Amsterdam: North-Holland
- [17] Doz, C., Giannone, D. and L Reichlin (2008). A quasi maximum likelihood approach for large approximate dynamic factor models. ECARES and CEPR.

- [18] 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.
- [19] Engle, R., D.F. Hendry, and J.F. Richard (1983): “Exogeneity,” *Econometrica*, 51, 277-304.
- [20] Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2000), “The Generalized Dynamic Factor Model: Identification and Estimation,” *Review of Economics and Statistics* 82, 540–554.
- [21] Goldberger, A.S. (1972): “Structural Equations Methods in the Social Sciences,” *Econometrica*, 40, 979-1001.
- [22] Hausman, J.A and W.E. Taylor (1981): “Panel data and unobservable individual effects,” *Econometrica*, 49,1377-1398.
- [23] Holtz-Eakin, D., W. Newey, and H. Rosen (1988): “Estimating Vector Autoregressions with Panel Data”, *Econometrica*, 56, 1371-1395.
- [24] Hsiao, C. (2003): *Analysis of Panel Data*. Cambridge University Press, New York.
- [25] Jöreskog, K.G., and Goldberger, A.S. (1975): “Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable,” *Journal of the American Statistical Association*, 70, 631-639.
- [26] Jungbacker, B. and S.J. Koopman (2008). Likelihood-based analysis for dynamic factor models, memo.
- [27] Kiviet, J. (1995): “On Bias, Inconsistency, and Efficiency of Various Estimators in Dynamic Panel Data Models”, *Journal of Econometrics*, 68, 53-78.
- [28] Kneip, A., R. Sickles, and W. Song (2008): “A new panel data treatment for heterogeneity in time trends”, unpublished manuscript, Department of Economics, Rice University.
- [29] Lawley, D.N. and A.E. Maxwell (1971): *Factor Analysis as a Statistical Method*, London: Butterworth.
- [30] 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 (1994)
- [31] MaCurdy, T. (1982): “The Use of Time Series Processes to Model the Error Structure of Earnings in a Longitudinal Data Analysis,” *Journal of Econometrics* 18, 83-114.
- [32] Magnus, J.R. and H. Neudecker (1999): *Matrix Differential Calculus with Applications in Statistics and Econometrics*, Wiley: New York.
- [33] McLachlan, G.J, and T. Krishnan (1996): *The EM Algorithm and Extensions*, Wiley, New York.
- [34] Meng, X.L and D.B. Rubin (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* 80(2), 267-278.

- [35] Mundlak, Y. (1978): “On the pooling of time series and cross section data,” *Econometrica*, 46, 69-85.
- [36] Neyman, J., and E. L. Scott (1948): “Consistent Estimates Based on Partially Consistent Observations,” *Econometrica*, 16, 1-32.
- [37] Nickell, S. (1981): “Biases in Dynamic Models with Fixed Effects,” *Econometrica*, 49, 1417-1426.
- [38] Pesaran, M. H. (2006): “Estimation and Inference in Large Heterogeneous panels with a Multifactor Error Structure,” *Econometrica*, 74, 967-1012.
- [39] Proietti, T. (2008). Estimation of common factors under cross-sectional and temporal aggregation constraints: nowcasting monthly GDP and its main components. Manuscript, University of Rome ”Tor Vergata.”
- [40] Quad, Q. and T. Sargent (1993). A Dynamic Index Model for Large Cross Sections. CEP Discussion Paper No. 0132.
- [41] Rubin, D.B. and D.T. Thayer (1982). EM algorithm for ML factor analysis. *Psychometrika*, 47 69-76.
- [42] Stock, J.H. and M.W. Watson (2003). *Introduction to Econometrics*, Addison-Wesley: New York.
- [43] Stock, J. H. and M. W. Watson (2002): “Forecasting Using Principal Components from a Large Number of Predictors,” *Journal of the American Statistical Association* **97**, 1167–1179.
- [44] 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.
- [45] Zhao, J.H., L.H. Yu, and Q. Jiang (2008), ML estimation for factor analysis: EM or non-EM? *Statistics and computing*, 18, 109-123.