cemmap

centre for microdata methods and practice

Likelihood inference and the role of initial conditions for the dynamic panel data model

Jose Diogo Barbosa
Marcelo J. Moreira

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP04/17

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

An ESRC Research Centre

Likelihood Inference and The Role of Initial Conditions for the Dynamic Panel Data Model

Jose Diogo Barbosa[1]
University of Michigan

Marcelo J. Moreira
FGV

This version: January 16, 2017

[1] The authors gratefully acknowledge the research support of CNPq and FAPERJ.

Abstract

Lancaster (2002) proposes an estimator for the dynamic panel data model with homoskedastic errors and zero initial conditions. In this paper, we show this estimator is invariant to orthogonal transformations, but is inefficient because it ignores additional information available in the data. The zero initial condition is trivially satisfied by subtracting initial observations from the data. We show that differencing out the data further erodes efficiency compared to drawing inference conditional on the first observations.

Finally, we compare the conditional method with standard random effects approaches for unobserved data. Standard approaches implicitly rely on normal approximations, which may not be reliable when unobserved data is very skewed with some mass at zero values. For example, panel data on firms naturally depend on the first period in which the firm enters on a new state. It seems unreasonable then to assume that the process determining unobserved data is known or stationary. We can instead make inference on structural parameters by conditioning on the initial observations.

Keywords: Autoregressive, Panel Data, Invariance, Efficiency.

JEL Classification Numbers: C12, C30.

1 Introduction

In an important paper, Lancaster (2002) studies, from a Bayesian perspective, estimation of the structural parameters of a dynamic panel data model with fixed effects and initial observations equal to zero. His method involves reparameterizing the model so that the information matrix is block diagonal, with the common parameters in one block and the incidental parameters in the other. His estimator is then defined as one of the local maxima of the integrated likelihood function, integrating with respect to the Lebesgue measure on $\mathbb{R}^N$. However, Lancaster leaves unanswered the question of how to uniquely determine the consistent root of his proposed methodology. Some authors, including Dhaene and Jochmans (2016) and Kruiniger (2014), have proposed different ways to find the consistent estimator in Lancaster's approach.

In this paper, we explain the shortcoming of Lancaster's estimator: it ignores available information in the model. In particular, Lancaster's posterior distribution uses only part of the maximal invariant statistic's log-likelihood function; when the full likelihood function is used in the estimation, a unique, consistent, and asymptotically normal efficient estimator is obtained; see Moreira (2009). The estimator obtained using the full likelihood is asymptotically more efficient than Lancaster's estimator. Therefore, trying to correct the nonuniqueness issue of Lancaster's estimator is unnecessary and leads to inefficient estimators.

Lancaster (2002) and Moreira (2009) study consistent estimation of dynamic panel models under the same set of assumptions and with initial observations equal to zero. The zero initial condition is trivially satisfied when the initial observations are subtracted from the autoregressive variables. We then show that efficiency is improved by conditioning on the initial observations instead of differencing out the data. The conditional argument is essentially a fixed-effects approach in which we make no further premises on unobserved data. This is in contrast with commonly-used estimators in the literature which make further assumptions; see Bai (2013a,b) on correlated random effects or Blundell and Bond (1998) on stationarity.

A potential advantage of conditioning on the first observation is robustness. As Blundell and Smith (1991) point out, asymptotic arguments are usually based on the average temporal effect, calculated on the individual dimension. Therefore, the importance of unobserved data does not disappear asymptotically for relatively short panels. For example, take data on firms or individual earnings. Cabral and Mata (2003) show that the distribution of firms is very skewed, with most of the mass being small firms, while Evans (1987b,a) and Hall (1987) show that Gilbrat's Law (independent firm size and growth) is rejected for small firms. As only a handful of large firms provide the bulk of the data, we should not expect estimators based on assumptions about unobserved data to be approximately normal. As the firms' entry states do not disappear asymptotically for relatively short panels, conditioning on the first observation would be preferable to assuming known processes for unobserved data.

The remainder of this paper is organized as follows. Section 2 introduces a simple dynamic panel data model without covariates and a zero initial condition. This section determines the maximal invariant statistic and summarizes the asymptotic theory for the maximum invariant likelihood estimator (MILE). Section 3 develops the asymptotic theory for Lancaster's estimator. Section 4 shows that this estimator is less efficient than MILE because it ignores relevant data. Section 5 shows that it is less efficient to difference out the first observation than to condition on it. Section 6 compares the conditional argument to standard random effects approaches for the unobserved data. Section 7 concludes and discusses how to extend the model to a more useful form.

2 The Model and the Maximal Invariant Likelihood

We consider a simple homoskedastic dynamic panel model with fixed effects and without covariates:

$$y_{i,t+1} = \rho y_{i,t} + \eta_i + \sigma u_{i,t},\ i = 1, \cdots, N;\ t = 1, \cdots, T \tag{1}$$

where $N \geq T+1$, $y_{i,t} \in \mathbb{R}$ are observable variables and $u_{i,t} \overset{iid}{\sim} N(0,1)$ are unobservable errors; $\eta_i \in \mathbb{R}$ are incidental parameters and $(\rho, \sigma^2) \in \mathbb{R} \times \mathbb{R}$ are structural parameters. We denote the true unknown parameters by $(\rho^*, \sigma^{*2}, \eta_i^*)$. We assume that the parameter space is a compact set and $\sigma^{*2} > 0$. For now, we assume that the initial *observed* condition is $y_{i,1} = 0$ as Lancaster (2002) does. We relax this assumption in Sections 5 and 6.

Solving model (1) recursively and writing it in matrix form yields

$$\begin{aligned} Y_T &= \eta 1_T' B_T' + \sigma U_T B_T',\ \text{where} \\ U_T &\sim N(0_{N\times T}, I_N \otimes I_T), \end{aligned} \tag{2}$$

$\eta = (\eta_1, \cdots, \eta_N)' \in \mathbb{R}^{N\times 1}$, $1_T = (1, \cdots, 1)' \in \mathbb{R}^{T\times 1}$,

$$Y_T = \begin{bmatrix} y_{1,2} & \cdots & y_{1,T+1} \\ \vdots & \ddots & \vdots \\ y_{N,2} & \cdots & y_{N,T+1} \end{bmatrix},\ U_T = \begin{bmatrix} u_{1,2} & \cdots & u_{1,T+1} \\ \vdots & \ddots & \vdots \\ u_{N,2} & \cdots & u_{N,T+1} \end{bmatrix},\ \text{and } B_T = \begin{bmatrix} 1 & & \\ \vdots & \ddots & \\ \rho^{T-1} & \cdots & 1 \end{bmatrix}.$$

When there is no confusion, we will omit the subscript from the matrices; e.g., Y instead of Y_T, B instead of B_T, etc.

The inverse of B has a simple form,

$$B^{-1} \equiv D = I_T - \rho J_T,\ \text{where } J_T = \begin{bmatrix} 0_{T-1}' & 0 \\ I_{T-1} & 0_{T-1} \end{bmatrix}$$

and 0_{T-1} is a $(T-1)$-dimensional column vector with zero entries.

If individuals i are treated equally, the coordinate system used to specify the vector $(y_{1,t},\cdots,y_{N,t})$ should not affect inference based on them. Therefore, it is reasonable to restrict attention to coordinate-free functions of $(y_{1,t},\cdots,y_{N,t})$. Chamberlain and Moreira (2009) and Moreira (2009) show that, indeed, orthogonal transformations preserve both the model (2) and the structural parameters (ρ,σ^2) and this yields a maximal invariant statistic, the $T\times T$ matrix $Y'Y$. So, if the researcher finds that it is reasonable to restrict attention to statistics that are invariant to orthogonal transformations, the maximal invariant statistic plays a crucial role: a statistic is invariant to orthogonal transformations if, and only if, it depends on the data through the maximal invariant statistic $Y'Y$.

The maximal invariant statistic $Y'Y$ has a noncentral Wishart distribution and depends only on ρ, σ^2, and $\omega_\eta^2\equiv\frac{\eta'\eta}{\sigma^2N}$. The noncentral Wishart distribution is the multivariate generalization of the noncentral Chi-squared distribution and it depends on the modified Bessel function of the first kind. We use uniform approximations of Bessel functions (see Abramowitz and Stegun (1965)), which allows us to write the density of the noncentral Wishart distribution in a more tractable form. Specifically, the log-likelihood of $Y'Y$ is, up to an $o_p\left(N^{-1}\right)$ term, proportional to

$$\begin{aligned} Q_N^M(\theta) &= -\frac{1}{2}\ln\left(\sigma^2\right)-\frac{1}{2\sigma^2}\frac{tr\left(DY'YD'\right)}{NT}-\frac{\omega_\eta^2}{2} \\ &\quad +\frac{\left(1+A_N^2\right)^{1/2}}{2T}-\frac{\ln\left(1+\left(1+A_N^2\right)^{1/2}\right)}{2T}, \end{aligned} \tag{3}$$

where $\theta=\left(\rho,\sigma^2,\omega_\eta^2\right)$ and $A_N=2\sqrt{\omega_\eta^2\frac{1_T'DY'YD'1_T}{\sigma^2N}}$.

The log-likelihood $Q_N^M(\theta)$ is free from the incidental parameter problem since $Y'Y$ is parametrized by the fixed dimensional vector of parameters θ. Although the dimension is fixed, the parameter

$$\omega_\eta^2\equiv\frac{\eta'\eta}{\sigma^2N}=\frac{\sum_{i=1}^N\eta_i^2}{\sigma^2N} \tag{4}$$

depends on the sample size N. For simplicity, we omit the dependence on N from the parameter ω_η^2. However, the asymptotic properties of the estimator will be derived under different sequences of ω_η^2. The asymptotic properties of the estimator obtained by maximizing the objective function $Q_N^M(\theta)$ are studied by Moreira (2009) and we reproduce these results here for convenience. The information matrix $\mathcal{I}_T(\theta^*)$ in Moreira (2009) contains typographical errors which are corrected here. Define the matrix

$$F_0=\begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ \rho & 0 & \cdots & 0 & 0 \\ \frac{1}{2}\rho^2 & \rho & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{1}{T-1}\rho^{T-1} & \frac{1}{T-2}\rho^{T-2} & \cdots & \rho & 0 \end{bmatrix}$$

and its derivatives:

$$F_j = \frac{d^j}{d\rho^j} F_0 \text{ and } F_j^* = \left. \frac{d^j}{d\rho^j} F_0 \right|_{\rho=\rho^*}.$$

Theorem 1: Let

$$\hat{\theta} = \underset{\theta \in \Theta}{\arg\max} Q_N^M(\theta). \tag{5}$$

(A.1) Under the assumption that $N \to \infty$ with T fixed, (i) if ω_η^2 is fixed at ω_η^{*2}, then $\hat{\theta}_M \to_p \theta^* = \left(\rho^*, \sigma^{*2}, \omega_\eta^{2*}\right)$; (ii) if $\omega_\eta^2 \to \omega_\eta^{*2}$, then $\hat{\theta}_M \to_p \theta^* = \left(\rho^*, \sigma^{*2}, \omega_\eta^{*2}\right)$; and (iii) if $\limsup \omega_\eta^{*2} < \infty$, then $\hat{\theta}_M = \theta^* + o_p(1)$, where $\theta^* = \left(\rho^*, \sigma^{*2}, \omega_\eta^{*2}\right)$.

(A.2) Under the assumption that $T \to \infty$ and $|\rho^*| < 1$, (i) if ω_η^2 is fixed at ω_η^{*2}, then $\hat{\theta}_M \to_p \theta^* = \left(\rho^*, \sigma^{*2}, \omega_\eta^{*2}\right)$; (ii) if $\omega_\eta^2 \to \omega_\eta^{*2}$, then $\hat{\theta}_M \to_p \theta^* = \left(\rho^*, \sigma^{*2}, \omega_\eta^{*2}\right)$; and (iii) if $\limsup \omega_\eta^{*2} < \infty$, then $\hat{\theta}_M = \theta^* + o_p(1)$, where $\theta^* = \left(\rho^*, \sigma^{*2}, \omega_\eta^{*2}\right)$.

(B) Assume that $\omega_\eta^{*2} > 0$ is fixed, and let the score statistic and the Hessian matrix be

$$S_N^M(\theta) = \frac{\partial Q_N^M(\theta)}{\partial \theta} \text{ and } H_N^M(\theta) = \frac{\partial^2 Q_N^M(\theta)}{\partial \theta \partial \theta'},$$

respectively, and define the matrix

$$\mathcal{I}_T^M(\theta^*) = \begin{bmatrix} h_T^M & \frac{\omega_\eta^{*4}}{\sigma^{*2}} \frac{1_T' F_1^* 1_T}{1+2\omega_\eta^{*2}T} & \frac{1+\omega_\eta^{*2}T}{1+2\omega_\eta^{*2}T} \frac{1_T' F_1^* 1_T}{T} \\ \frac{\omega_\eta^{*4}}{\sigma^{*2}} \frac{1_T' F_1^* 1_T}{1+2\omega_\eta^{*2}T} & \frac{1}{2(\sigma^{*2})^2}\left(1 + \frac{\omega_\eta^{*4}T}{1+2\omega_\eta^{*2}T}\right) & \frac{1}{2\sigma^{*2}} \frac{1+\omega_\eta^{*2}T}{1+2\omega_\eta^{*2}T} \\ \frac{1+\omega_\eta^{*2}T}{1+2\omega_\eta^{*2}T} \frac{1_T' F_1^* 1_T}{T} & \frac{1}{2\sigma^{*2}} \frac{1+\omega_\eta^{*2}T}{1+2\omega_\eta^{*2}T} & \frac{T}{2\left(1+2\omega_\eta^{*2}T\right)} \end{bmatrix},$$

where

$$h_T^M = \frac{tr\left(F_1^* F_1'^*\right)}{T} + \frac{\omega_\eta^{*4}T}{1+\omega_\eta^{*2}T}\left(\frac{1_T' F_1^* F_1'^* 1_T}{T} + \frac{1}{1+2\omega_\eta^{*2}T}\left(\frac{1_T' F_1^* 1_T}{T}\right)^2\right).$$

As $N \to \infty$ with T fixed, (i) $\sqrt{NT} S_N^M(\theta^*) \to_d N\left(0, \mathcal{I}_T^M(\theta^*)\right)$; (ii) $H_N^M(\theta^*) \to_p -\mathcal{I}_T^M(\theta^*)$; (iii) $\sqrt{NT}\left(\hat{\theta}_M - \theta^*\right) \to_d N\left(0, \mathcal{I}_T^M(\theta^*)^{-1}\right)$; and (iv) the log-likelihood ratio is

$$\begin{aligned} \Lambda_N\left(\theta^* + h \cdot (NT)^{-1/2}, \theta^*\right) &= NT\left(Q_N^M\left(\theta^* + h \cdot (NT)^{-1/2}\right) - Q_N^M(\theta^*)\right) \\ &= h'\sqrt{NT} S_N^M(\theta^*) - \frac{1}{2} h' \mathcal{I}_T(\theta^*) h + o_{Q_N^M(\theta^*)}(1), \end{aligned}$$

$\sqrt{NT} S_N^M(\theta^*) \to_d N\left(0, \mathcal{I}_T^M(\theta^*)\right)$ under $Q_N^M(\theta^*)$. Furthermore, $\hat{\theta}_M$ is asymptotically efficient within the class of regular invariant estimators for the differenced model (16) under large N, fixed T asymptotics.

Part (A) of the above theorem implies that $\hat{\rho}_M \to_p \rho^*$ and $\hat{\sigma}_M^2 \to_p \sigma^{*2}$ regardless of the growth rate of N and T as long as $NT \to \infty$. Part (B) derives the limiting distribution of $\hat{\theta}_M$. It shows, in particular, that $\hat{\rho}_M$ achieves the efficiency bound $\left(\mathcal{I}_T^M\left(\theta^*\right)^{-1}\right)_{11}$ for regular invariant estimators as $N \to \infty$. Regular estimators exclude superefficient estimators in the sense of Hodges-Le Cam and, heuristically, a regular estimator is one whose asymptotic distribution does not change in shrinking neighborhoods of the true parameter value (see Bickel, Klaassen, Ritov, and Wellner (1998) for more details).

Part (B) of Theorem 1 finds the asymptotic distribution of $\hat{\theta}_M$, assuming that ω_η^2 is fixed at ω_η^{*2}. A generalization of part (B) that allows for $\omega_\eta^2 \to \omega_\eta^{*2}$ follows from a simple application of Le Cam's third lemma.

Lemma 2: Assume $\omega_\eta^2 = \omega_\eta^{*2} + h/\sqrt{N}$, where $\omega_\eta^{*2} > 0$, $h \in \mathbb{R}$ and ω_η^{*2} is in the parameter space. As $N \to \infty$ with T fixed,

$$\sqrt{NT}\left(\hat{\theta}_M - \theta^*\right) \to_d N\left(\left(\begin{array}{c} 0 \\ 0 \\ h \end{array}\right), \mathcal{I}_T^M\left(\theta^*\right)^{-1}\right),$$

where $\mathcal{I}_T^M\left(\theta^*\right)^{-1}$ is defined in Theorem 1, part (B).

Lemma 2 shows that the asymptotic distribution of the structural parameters does not change, whether ω_η^2 is fixed at ω_η^{*2} or $\omega_\eta^2 \to \omega_\eta^{*2}$. For simplicity's sake, we assume throughout the paper that the sequence ω_η^2 (which can depend on the sample size N) is fixed at $\omega_\eta^{*2} > 0$.

3 Lancaster's Estimator

Lancaster (2002) proposes a Bayesian approach to estimate the structural parameters which involves reparameterizing model (2) so that the information matrix is block diagonal, with the common parameters in one block and the incidental parameters in the other. Then he defines his estimator as one of the local maxima of the integrated likelihood function, integrating with respect to the Lebesgue measure on $\mathbb{R}^N$. Dhaene and Jochmans (2016) give an alternative interpretation for Lancaster's estimator by showing that Lancaster's posterior can be obtained by adjusting the profile likelihood so that its score is free from asymptotic bias. In this sense, Lancaster's estimator is a bias-corrected estimator.

Lancaster's estimator seeks to maximize the following objective function:

$$Q_N^L\left(\rho, \sigma^2\right) = -\frac{1}{2}\ln\left(\sigma^2\right) + \frac{1_T' F_0 1_T}{T\left(T-1\right)} - \frac{1}{2\sigma^2}\frac{tr\left(DY'YD'H\right)}{N\left(T-1\right)}, \tag{6}$$

where the matrix H is defined as

$$H = I_T - \frac{1}{T}1_T 1_T'.$$

The posterior (6) is not a likelihood function and $\left(\rho^*, \sigma^{*2}\right)$ is not a global maximizer of $\underset{N\to\infty}{plim} Q_N^L\left(\rho, \sigma^2\right)$ as in standard maximum likelihood theory (see Dhaene and Jochmans (2016)). Nonetheless, Theorem 3, proved by Lancaster (2002), shows that the posterior (6) can be used to consistently estimate the structural parameters.

Theorem 3: Let $S_N^L\left(\rho, \sigma^2\right) = \frac{\partial Q_N^L\left(\rho,\sigma^2\right)}{\partial(\rho,\sigma^2)'}$ be the score of the posterior (6) and let Θ_N be the set of roots of

$$S_N^L\left(\rho, \sigma^2\right) = 0 \tag{7}$$

corresponding to the local maxima. If that set is empty, set Θ_N equals to $\{0\}$. Then, there is a consistent root of equation (7).

The usefulness of Theorem 3 is limited by the fact that it only states that one (of possibly many) of the local maxima of (6) is a consistent estimator of the structural parameters. However, it does not indicate how to find a consistent estimator. Therefore, even though the posterior can be used to consistently estimate the structural parameters of model (2), Lancaster leaves unanswered the question of how to uniquely determine the consistent root of (7). To our knowledge, two different methodologies to uniquely choose the consistent root of Lancaster's score have been proposed. The first approach, by Dhaene and Jochmans (2016), suggests as a consistent estimator of the structural parameters, the minimizer of the norm of Lancaster's score on an interval around the maximum likelihood estimator obtained from the maximization of the likelihood of (2). The second method, by Kruiniger (2014), uses as a consistent estimator, the minimizer of a quadratic form in Lancaster's score subjected to a condition on the Hessian matrix.

Lemma 4 finds the asymptotic variance of a consistent root of (7).

Lemma 4: Assume $\omega_\eta^{*2} > 0$ and let $\left(\hat{\rho}_L, \hat{\sigma}_L^2\right)$ be a consistent root of $S_N^L\left(\rho, \sigma^2\right)$. Let the Hessian matrix be

$$H_N^L\left(\rho, \sigma^2\right) = \frac{\partial^2 Q_N^L\left(\rho, \sigma^2\right)}{\partial\left(\rho, \sigma^2\right)' \partial\left(\rho, \sigma^2\right)},$$

and define the matrices

$$\mathcal{I}_T^L\left(\theta^*\right) = \left[\begin{array}{cc} h_T^L & -\frac{1}{\sigma^{*2}} \frac{1_T' F_1^* 1_T}{T(T-1)} \\ -\frac{1}{\sigma^{*2}} \frac{1_T' F_1^* 1_T}{T(T-1)} & \frac{1}{2(\sigma^{*2})^2} \end{array}\right]$$

and

$$\Sigma_T\left(\theta^*\right) = \frac{T}{T-1}\left[\begin{array}{cc} a_T^L & \mathcal{I}_T^L(1,2) \\ \mathcal{I}_T^L(1,2) & \mathcal{I}_T^L(2,2) \end{array}\right], \tag{8}$$

where

$$h_T^L = -\frac{1_T' F_2^* 1_T}{T(T-1)} + \frac{tr\left(F_1^* F_1^{*\prime}\right)}{T-1} + \frac{1_T' F_1^{*\prime} F_1^* 1_T}{T} \frac{\left(\omega_\eta^{*2} T - 1\right)}{T-1} - \frac{\omega_\eta^{*2} T}{T-1}\left(\frac{1_T' F_1^{*\prime} 1_T}{T}\right)^2,$$

$$a_T^L = 2h_T^L + 2\frac{1_T' F_2^* 1_T}{T\left(T-1\right)},$$

and $\mathcal{I}_T^L\left(i,j\right)$ is the $\left(i,j\right)$-th entry of $\mathcal{I}_T^L\left(\theta^*\right)$. When $\mathcal{I}_T^L\left(\theta^*\right)$ is nonsingular, as $N \to \infty$ with T fixed, (i) $\sqrt{NT}S_N^L\left(\rho^*, \sigma^{*2}\right) \to_d N\left(0, \Sigma_T\left(\theta^*\right)\right)$; (ii) $-H_N^L\left(\rho^*, \sigma^{*2}\right) \to_p \mathcal{I}_T^L\left(\theta^*\right)$; and (iii) $\sqrt{NT}\left(\begin{array}{c} \hat{\rho}_L - \rho^* \\ \hat{\sigma^2}_L - \sigma^{*2} \end{array}\right) \to_d N\left(0, \mathcal{I}_T^L\left(\theta^*\right)^{-1}\Sigma_T\left(\theta^*\right)\mathcal{I}_T^L\left(\theta^*\right)^{-1}\right)$.

Lemma 4 follows from standard asymptotic theory because of the assumption that the Hessian matrix $-H_N^L\left(\rho^*, \sigma^{*2}\right)$ converges in probability to a nonsingular matrix $\mathcal{I}_T^L\left(\theta^*\right)$. Dhaene and Jochmans (2016) prove that this assumption is violated, i.e. $\mathcal{I}_T^L\left(\theta^*\right)$ is singular, when $T = 2$ or $\rho^* = 1$.

4 Lancaster's Estimator and a Decomposition of the Maximal Invariant Statistic's Log-Likelihood

In contrast to Lancaster's estimator, the estimator defined in (5), based on the maximal invariant statistic's log-likelihood, is a standard maximum likelihood estimator in the sense that its objective function is a likelihood function and the limit of its objective function attains a unique global maximum at the vector of true parameters. This implies that the estimator that maximizes the maximal invariant statistic's log-likelihood is uniquely determined and no additional procedures are necessary to obtain a consistent and asymptotically normal estimator of the structural parameters. More than that, Theorem 1 shows that it attains the minimum variance bound for invariant regular estimators of the structural parameters of model (2). Therefore, the estimator defined in (5) efficiently uses all information available in the maximal invariant statistic $Y'Y$.

Lancaster's estimator is also invariant to orthogonal transformations since it depends on the data only through the maximal invariant statistic $Y'Y$; see equation (6). However, it is not uniquely determined and additional steps, such as those proposed by Dhaene and Jochmans (2016) and Kruiniger (2014), are necessary to choose the consistent root among the (possibly) many roots of Lancaster's score function. Furthermore, the limit of probability of the Hessian of $Q_N^L\left(\rho, \sigma^2\right)$ is not necessarily proportional to the asymptotic variance ($AVar$) of the score. Since the information equality does not hold, the asymptotic variance of Lancaster's consistent estimator cannot attain the lower bound for invariant regular estimators found in Theorem 1.

The explanation for the shortcomings of Lancaster's estimator is that, even though it is invariant, its objective function ignores information available in the maximal invariant by using only part of the maximal invariant statistic's log-likelihood function. Indeed, the maximal invariant statistic's log-likelihood function (3) concentrated out of σ^2 and ω_η^2 is

$$Q_N^M(\rho) = -\frac{1}{2}\frac{T-1}{T}\ln\left(\frac{tr\left(DY'YD'H\right)}{N\left(T-1\right)}\right) - \frac{1}{2T}\ln\left(\frac{1_T'DY'YD'1_T}{NT}\right), \tag{9}$$

while Lancaster's objective function concentrated out of σ^2 is

$$Q_N^L(\rho) = \frac{1_T'F_0 1_T}{T\left(T-1\right)} - \frac{1}{2}\ln\left(\frac{tr\left(DY'YD'H\right)}{N\left(T-1\right)}\right),$$

which allows us to write

$$\begin{aligned} Q_N^M(\rho) &= \frac{T-1}{T}Q_N^L(\rho) + Q_N^{M-L}(\rho), \text{ where} \\ Q_N^{M-L}(\rho) &= -\frac{1_T'F_0 1_T}{T^2} - \frac{1}{2T}\ln\left(\frac{1_T'DY'YD'1_T}{NT}\right). \end{aligned} \tag{10}$$

The respective score functions are given by

$$S_N^L(\rho) = \frac{1_T'F_1 1_T}{T\left(T-1\right)} + \frac{tr\left(J_T Y'YD'H\right)}{tr\left(DY'YD'H\right)}$$

and

$$\begin{aligned} S_N^M(\rho) &= \frac{T-1}{T}S_N^L(\rho) + S_N^{M-L}(\rho), \text{ where} \\ S_N^{M-L}(\rho) &= -\frac{1_T'F_1 1_T}{T^2} + \frac{1}{T}\frac{1_T'J_T Y'YD'1_T}{1_T'DY'YD'1_T}. \end{aligned} \tag{11}$$

The decomposition (10) implies an orthogonal decomposition of the concentrated score function:

Theorem 5: Let $S_N^M(\rho)$, $S_N^L(\rho)$ and $S_N^{M-L}(\rho)$ be the score functions associated with $Q_N^M(\rho)$, $Q_N^L(\rho)$ and $Q_N^{M-L}(\rho)$, respectively. Then,

$$S_N^M(\rho) = \frac{T-1}{T}S_N^L(\rho) + S_N^{M-L}(\rho), \tag{12}$$

where

$$S_N^{M-L}(\rho^*) \rightarrow_p 0 \tag{13}$$

and

$$\left(\begin{array}{c} \sqrt{NT}S_N^L(\rho^*) \\ \sqrt{NT}S_N^{M-L}(\rho^*) \end{array}\right) \rightarrow_d N\left(\left(\begin{array}{c} 0 \\ 0 \end{array}\right), \left(\begin{array}{cc} b_T & 0 \\ 0 & c_T \end{array}\right)\right), \tag{14}$$

with

$$b_T = a_T^L - 2\frac{T}{\left(T-1\right)^3}\left(\frac{1_T'F_1^* 1_T}{T}\right)^2 > 0,$$

the constant a_T^L as defined in Lemma 4, and

$$c_T = \frac{2T}{\left(1+\omega_\eta^{*2}T\right)}\left(\frac{1_T' F_1^{*'} F_1^* 1_T}{T} - \left(\frac{1_T' F_1^* 1_T}{T}\right)^2\right) > 0.$$

A consistent estimator for ρ^* is expected to be more efficient, the closer it stays to the maximum likelihood estimator $\hat{\rho}_M$. This implies that the consistent estimator $\hat{\rho}_L$ might have low efficiency because it only uses part of the full log-likelihood (10). One may use $\left|S_N^M(\hat{\rho}_L)\right|$ as a measure of how "close" any inflection point $\hat{\rho}_L$ is to $\hat{\rho}_M$ and, because $S_N^M(\hat{\rho}_L) = S_N^{M-L}(\hat{\rho}_L)$, $\hat{\rho}_L$ will be more efficient, the smaller $\left|S_N^{M-L}(\hat{\rho}_L)\right|$ is. The term $S_N^{M-L}(.)$ evaluated at the true value ρ^* or at a consistent estimator, converges in probability to zero. This suggests choosing the inflexion points based on Lancaster's score by looking at $S_N^{M-L}(\hat{\rho}_L)$ to uniquely determine the consistent root, akin to Dhaene and Jochmans (2016) and Kruiniger (2014).

Instead, we can use the information on $S_N^{M-L}(.)$ to increase the efficiency of $\hat{\rho}_L$. Simple algebra shows that

$$\left|S_N^{M-L}(\hat{\rho}_L)\right| = \left|(T-1)\frac{tr\left(J_T Y'Y\hat{D}_L' H\right)}{tr\left(\hat{D}_L Y'Y\hat{D}_L' H\right)} + \frac{1_T' J_T Y'Y\hat{D}_L' 1_T}{1_T'\hat{D}_L Y'Y\hat{D}_L' 1_T}\right|,$$

where $\hat{D}_L$ is the matrix D evaluated at $\hat{\rho}_L$. Notice that $\hat{\rho}_L$ is inefficient since $\sqrt{NT}S_N^{M-L}(\hat{\rho}_L)$ does not converge in probability to zero. Instead of choosing the consistent root following Lancaster's methodology, we could add the information in $S_N^{M-L}(.)$ to $S_N^L(.)$ by using (12). The associated estimator $\hat{\rho}_M$ uses the information in both $S_N^{M-L}(.)$ and $S_N^L(.)$ efficiently and is uniquely defined, making the estimation problem more simple and objective.

5 Initial Condition and Likelihood Inference

Theorem 1 shows that the estimator for the structural parameters (ρ, σ^2) attains the efficiency bound for model (1) when the first observation is $y_{i,1} = 0$. Lancaster (2002) suggests working with the differenced data $\tilde{y}_{i,t} = y_{i,t} - y_{i,1}$ so that we may continue working with an initial observation equal to zero. Instead, we draw inference on (ρ, σ^2) by conditioning on the first observation $y_{i,1}$. We will show the differencing method is indeed less efficient than the conditional method.

By using model (1) with the variables $y_{i,t}$ in levels, there are more observations that can be used to estimate the parameters. On the other hand, the number of parameters to be estimated increases. In this section we show that the estimator that maximizes the likelihood of the maximal invariant statistic of the original model (1) has asymptotic variance that is no larger than the variance of the estimator (5) and, under very general conditions, using $y_{i,t}$ in levels will be strictly more efficient.

When we allow for a nonzero initial condition, the model (1) becomes

$$\begin{aligned} Y_T &= y_1.\rho e_1' B_T' + \eta.1_T' B' + \sigma.U_T.B', \text{ where} \\ U_T &\sim N\left(0_{N\times T}, I_N \otimes I_T\right), \end{aligned} \tag{15}$$

$y_1 = (y_{1,1}, \cdots, y_{N,1})'$ is the vector of first observations and $e_1 = (1, 0, ..., 0)'$ is the canonical vector.

Lancaster (2002) works with a differenced version of model (1) given by

$$\begin{aligned} \widetilde{y}_{i,t+1} &= \rho \widetilde{y}_{i,t} + \widetilde{\eta}_i + \sigma u_{i,t}, \; i = 1, \cdots, N; \; t = 1, \cdots, T \\ \widetilde{y}_{i,t} &\equiv y_{i,t} - y_{i,1}, \; i = 1, \cdots, N; \; t = 1, \cdots, T+1 \\ \widetilde{\eta}_i &\equiv \eta_i - y_{i,1}(1-\rho), \; i = 1, \cdots, N \end{aligned} \tag{16}$$

and seeks inference based on $\widetilde{y}_{i,1} = 0$ for all i. In matrix form,

$$Y_{T+1} - y_1.1_{T+1}' = \left[0_{N\times 1} : Y_T - y_1.1_T'\right].$$

The second term equals

$$\begin{aligned} Y_T - y_1.1_T' &= y_1.\left(\rho.e_1' B' - 1_T'\right) + \eta.1_T' B' + \sigma.U.B' \\ &= \left[\eta - (1-\rho) y_1\right].1_T' B' + \sigma.U.B'. \end{aligned}$$

Defining differenced variables and the incidental parameters,

$$\widetilde{Y}_T = Y_T - y_1.1_T' \text{ and } \widetilde{\eta} = \eta - (1-\rho).y_1,$$

we are back to the model with the first observation equal to zero:

$$\widetilde{Y}_T = \widetilde{\eta}.1_T' B' + \sigma.U_T.B', \; U \sim N\left(0_{N\times T}, I_N \otimes I_T\right), \tag{17}$$

where the incidental parameter is $\tau = \widetilde{\eta}$.

Differencing eliminates one time period from the data. Instead, we will condition on the initial observation y_1 itself. It is convenient to work with the linear setup of Chamberlain and Moreira (2009):

$$Y = x.a(\gamma) + \tau.b(\gamma) + U.c(\gamma), \tag{18}$$

where $x \in \mathbb{R}^{N\times K}$ are explanatory variables, $\tau \in \mathbb{R}^{N\times J}$ are the incidental parameters, and $a(\cdot)$, $b(\cdot)$, $c(\cdot)$ are given functions of the unknown parameter of interest γ. Consider the group of transformations $g.Y$, where g are orthogonal matrices such that $g.x = x$. This group modifies the unknown incidental parameter τ, but preserves the model and the parameter γ of interest:

$$\begin{aligned} g.Y &= g.x.a(\gamma) + g.\tau.b(\gamma) + g.U.c(\gamma) \\ &= x.a(\gamma) + (g.\tau).b(\gamma) + U.c(\gamma) \end{aligned}$$

in distribution, because the law of $g.U$ is the same as the law of U. This group yields the maximal invariant statistic, which is given by the pair

$$Z_1 = \left(x'x\right)^{-1/2} x'Y \text{ and } Z_2'Z_2 = Y'M_xY, \tag{19}$$

where $Z_2 = q_2'Y$ for any matrix q_2 such that the matrix $q = \left[x\left(x'x\right)^{-1/2} : q_2\right]$ is orthogonal. Because q is an orthogonal matrix, we have $q_2q_2' = I_N - N_x = M_x$ for $N_x = x\left(x'x\right)^{-1} x'$. We refer the reader to Chamberlain and Moreira (2009) for more details.

Define the parameters

$$\omega_{\tau,x}^2 = \frac{\tau' M_x \tau}{\sigma^2 N} \text{ and } \delta_{\tau,x} = \left(x'x\right)^{-1} x'\tau. \tag{20}$$

The coefficient $\delta_{\tau,x}$ is the ordinary least squares (OLS) estimator and $\omega_{\tau,x}^2$ is the standardized average of sum of squared residuals from a fictitious regression of τ on x (if there are no covariates x, we define $\omega_\tau^2 = \left(\sigma^2 N\right)^{-1} \tau'\tau$).

For the model (18), the statistics Z_1 and $Z_2'Z_2$ are independently normal and noncentral Wishart distributed. Their distributions depend on $(\rho, \sigma^2, \delta_{\tau,x})$ and $\left(\rho, \sigma^2, \omega_{\tau,x}^2\right)$, respectively.

If we condition on the initial observation y_1, the model (15) for Y_T is a special case of (18), where $x = y_1$, $\tau = \eta$, and the regression coefficients are

$$a\left(\gamma\right) = \rho e_1' B',\ b\left(\gamma\right) = 1_T' B',\ \text{and } c\left(\gamma\right) = \sigma . B'.$$

Let $\theta_1 = \left(\rho, \sigma^2, \delta_{\eta,y_1}, \omega_{\eta,y_1}^2\right) \in \mathbb{R}^4$ be the parameters of the joint distribution of (19). The likelihood estimator that maximizes the joint likelihood of (19) does not have the incidental parameter problem since it is parametrized by θ_1, which has fixed dimension.

If we instead difference out the data, as Lancaster (2002) does, the model (17) for $\widetilde{Y}_T$ is a special case of (18), in which x is absent, the incidental parameter τ is $\widetilde{\eta} = \eta - \left(1-\rho\right).y_1$, and the regression coefficients are

$$b\left(\gamma\right) = 1_T' B' \text{ and } c\left(\gamma\right) = \sigma . B'.$$

We now show that Lancaster (2002)'s differencing method entails unnecessary efficiency loss. We begin by defining the estimator based on the likelihood of $(Z_1, Z_2'Z_2)$. Let

$$\hat{\theta}_{M1} = \underset{\theta_1 \in \Theta}{\arg\min} Q_N^{M1}\left(\theta_1\right), \tag{21}$$

where

$$\begin{aligned} Q_N^{M1}\left(\theta_1\right) \quad = \quad & -\frac{1}{2}\ln\sigma^2 - \frac{\omega_{\eta,y_1}^2}{2} - \frac{1}{2\sigma^2} tr\left(\frac{DZ_2'Z_2D'}{NT}\right) + \frac{1}{2T}\left(1 + A_{1,N}^2\right)^{1/2} \\ & - \frac{1}{2\sigma^2 NT}\left(Z_1D' - \left\|y_1\right\|\left(\rho e_1' + \delta_{\eta,y_1} 1_T'\right)\right)\left(DZ_1' - \left\|y_1\right\|\left(\rho e_1 + \delta_{\eta,y_1} 1_T\right)\right) \\ & - \frac{1}{2T}\ln\left(1 + \left(1 + A_{1,N}^2\right)^{1/2}\right), \end{aligned} \tag{22}$$

is, up to $o_p\left(N^{-1}\right)$ terms, proportional to the (conditional on y_1) log-likelihood of $(Z_1, Z_2' Z_2)$ and $A_{1,N} = 2\sqrt{\omega^2_{\eta,y_1}\frac{1_T' D Z_2' Z_2 D' 1_T}{\sigma^2 N}}$. The asymptotic behavior of the estimator $\hat{\theta}_{M1}$ is given next.

Lemma 6: Assume that the true parameters δ_{η,y_1} and ω^2_{η,y_1} are fixed, respectively, at δ^*_{η,y_1} and $\omega^{*2}_{\eta,y_1} > 0$ and that $0 < \lim\limits_{N\to\infty} \frac{\|y_1\|^2}{N} \equiv \overline{\|y_1\|}^2 < \infty$. As $N \to \infty$ with T fixed,

(A) $\hat{\theta}_{M1} \to_p \theta_1^* = \left(\rho^*, \sigma^{*2}, \delta^*_{\eta,y_1}, \omega^{*2}_{\eta,y_1}\right)$.

(B) Let $Q_N^{M1}\left(\rho, \sigma^2\right)$ be the objective function (22) concentrated out of δ_{η,y_1} and ω^2_{η,y_1} and denote the score statistic and the Hessian matrix by

$$S_N^{M1}\left(\rho, \sigma^2\right) = \frac{\partial Q_N^{M1}\left(\rho, \sigma^2\right)}{\partial\left(\rho, \sigma^2\right)'} \text{ and } H_N^{M1}\left(\rho, \sigma^2\right) = \frac{\partial^2 Q_N^{M1}\left(\rho, \sigma^2\right)}{\partial\left(\rho, \sigma^2\right)' \partial\left(\rho, \sigma^2\right)},$$

respectively. Also, define the matrix

$$\mathcal{I}_T^{M1}\left(\theta_1^*\right) = \left[\begin{array}{cc} d_T^{M1} & -\frac{1_T' F_1^* 1_T}{\sigma^{*2} T^2} \\ -\frac{1_T' F_1^* 1_T}{\sigma^{*2} T^2} & \frac{1}{2\left(\sigma^{*2}\right)^2}\frac{T-1}{T} \end{array}\right],$$

where

$$\begin{aligned} d_T^{M1} = {} & \frac{1}{T\left(1+\omega^{*2}_{\eta,y_1} T\right)}\left(\frac{1_T' F_1^{*'} F_1^* 1_T}{T} + \omega^{*2}_{\eta,y_1} T\left(\frac{1_T' F_1^* 1_T}{T}\right)^2\right) - \frac{2}{T}\left(\frac{1_T' F_1^* 1_T}{T}\right)^2 \\ & + \left(\omega^{*2}_{\eta,y_1} + \frac{\overline{\|y_1\|}^2}{\sigma^{*2}}\left(\delta^*_{\eta,y_1} + \rho^* - 1\right)^2\right)\left(\frac{1_T' F_1^{*'} F_1^* 1_T}{T} - \left(\frac{1_T' F_1^* 1_T}{T}\right)^2\right) \\ & + \frac{tr\left(F_1^* F_1^{*'}\right)}{T} - \frac{1_T' F_1^{*'} F_1^* 1_T}{T^2}. \end{aligned}$$

Then, (i) $\sqrt{NT} S_N^{M1}\left(\rho^*, \sigma^{*2}\right) \to_d N\left(0, \mathcal{I}_T^{M1}\left(\theta_1^*\right)\right)$; (ii) $H_N^{M1}\left(\rho^*, \sigma^{*2}\right) \to_p -\mathcal{I}_T^{M1}\left(\theta_1^*\right)$; and (iii) $\sqrt{NT}\left(\begin{array}{c} \hat{\rho}_{M1} - \rho^* \\ \hat{\sigma}^2_{M1} - \sigma^{*2} \end{array}\right) \to_d N\left(0, \mathcal{I}_T^{M1}\left(\theta_1^*\right)^{-1}\right)$.

Lemma 7 below shows that, in general, (21) estimates the structural parameters with strictly smaller variances than the original estimator that uses differenced data, as defined in (5).

Lemma 7: If $\delta^*_{\eta,y_1} + \rho^* \neq 1$, then $AVar(\hat{\rho}_M) > AVar(\hat{\rho}_{M1})$ and $AVar\left(\hat{\sigma}^2_M\right) > AVar\left(\hat{\sigma}^2_{M1}\right)$. If $\delta^*_{\eta,y_1} + \rho^* = 1$, then $AVar(\hat{\rho}_M) = AVar(\hat{\rho}_{M1})$ and $AVar\left(\hat{\sigma}^2_M\right) = AVar\left(\hat{\sigma}^2_{M1}\right)$.

6 Random Effects and Moment Conditions

In Section 5 we draw inference on the structural parameters by conditioning or by differencing the data based on the first *observed* initial condition $y_{i,1}$. Both methods are fixed-effect approaches, which do not rely on any further assumptions on the *unobserved* data such as $y_{i,0}$. We could instead consider random-effect approaches which rely on assumptions for the unobserved value $y_{i,0}$. These include Chamberlain's (correlated) random effects and Blundell and Bond (1998)'s stationarity assumptions for the first unobserved value $y_{i,0}$.

In this section, we compare the conditional method with the random-effect methods using the first moment of

$$\left(x'x\right)^{-1}x'Y \text{ and } \frac{Y'Y}{N}. \tag{23}$$

Define the function

$$\pi\left(\gamma,\delta_{\tau,x}\right)=a\left(\gamma\right)+\delta_{\tau,x}.b\left(\gamma\right).$$

The expectation of the maximal invariant is then given by

$$\begin{aligned} E\ \left(x'x\right)^{-1}x'Y &= \pi\left(\gamma,\delta_{\tau,x}\right) \text{ and} \\ E\ \frac{Y'Y}{N} &= \sigma^2\left[\pi\left(\gamma,\delta_{\tau,x}\right)'\omega_x^2\pi\left(\gamma,\delta_{\tau,x}\right)+b\left(\gamma\right)'\omega_{\tau,x}^2 b\left(\gamma\right)\right]+c\left(\gamma\right)'c\left(\gamma\right), \end{aligned}$$

For example, Lancaster (2002)'s differencing approach yields

$$E\ \frac{\widetilde{Y}_T'\widetilde{Y}_T}{N}=\sigma^2 B_T\left[\omega_{\widetilde{\eta}}^2 1_T 1_T'+I_T\right]B_T'.$$

Hence, we have $\left(T+1\right)T/2$ moments and three unknown parameters: ρ, σ^2, and $\omega_{\widetilde{\eta}}^2$. Conditioning on y_1 yields the following (conditional) moments based on the maximal invariant:

$$E\left[\left(y_1'y_1\right)^{-1}y_1'Y_T\middle|\, y_1\right]=\rho e_1'B_T'+\delta_{\eta,y_1}.1_T'B_T' \tag{24}$$

and

$$\begin{aligned} E\left[\frac{Y_T'Y_T}{N}\middle|\, y_1\right] &= \sigma^2 B_T\left\{\omega_{\eta,y_1}^2 1_T 1_T'+I_T\right\}B_T' \\ &+ \sigma^2\left[\rho B_T e_1+\delta_{\eta,y_1}.B_T 1_T\right]\omega_{y_1}^2\left[\rho e_1'B_T'+\delta_{\eta,y_1}.1_T'B_T'\right]. \end{aligned} \tag{25}$$

The unknown parameters are the autoregressive coefficient ρ, the error variance σ^2, the OLS coefficient δ_{η,y_1}, and the standardized squared residuals ω_{η,y_1}^2.

Invariance reduces the information to $T+\left(T+1\right)T/2=\left(T+1\right)\left(T+2\right)/2-1$ moment conditions which depend on only four parameters. Conditioning on the first observation yields T more moments than Lancaster (2002)'s differencing method and one additional parameter (four instead of three). This comparison helps to explain the efficiency gains from conditioning instead of differencing.

We can then explore the conditional moments (24) and (25) to find Minimum Distance (MD) estimators. We further note that Arellano and Bond (1991) and Ahn and Schmidt (1995) implicitly use linear combinations of the moments (24) and (25) to find moments for the autoregressive parameter ρ.

We instead explore connections between these moments, conditional on the first observed value $y_{i,1}$, and random effects assumptions for the unobserved value $y_{i,0}$. To simplify comparison, we continue working with the model without covariates and with homoskedastic errors.

6.1 Unobserved Initial Conditions as Incidental Parameters

As a preliminary to the correlated random effects assumption, consider the model written recursively to the time $t = 0$:

$$\begin{aligned} Y_{T+1} &= y_0.\rho e_1' B_{T+1}' + \eta.1_{T+1}' B_{T+1}' + \sigma.U_{T+1}.B_{T+1}', \text{ where} \\ U_{T+1} &\sim N\left(0_{N\times(T+1)}, I_N \otimes I_{(T+1)}\right), \end{aligned} \tag{26}$$

and y_0 is unobserved. This model is again a special case of the linear setup of Chamberlain and Moreira (2009) without covariates x. The incidental parameter $\tau = \left[\begin{array}{cc} y_0 & \eta \end{array}\right]$ would encapsulate the individual fixed effects and the initial conditions themselves. The regression coefficients would be given by

$$b(\gamma) = \left[\begin{array}{c} \rho e_1' \\ 1_{T+1}' \end{array}\right] B' \text{ and } c(\gamma) = \sigma.B'$$

(where we again omit the subscript, here $(T+1)$, from the matrices when there is no confusion).

The maximal invariant is $Y_{T+1}'Y_{T+1}$, which contains $(T+1)(T+2)/2$ nonredundant moments. Its expectation is given by

$$E\,\frac{Y_{T+1}'Y_{T+1}}{N} = \sigma^2 B_{T+1}\left\{\left[\begin{array}{cc} \rho.e_1 & 1_{T+1} \end{array}\right]\omega_\tau^2\left[\begin{array}{c} \rho.e_1' \\ 1_{T+1}' \end{array}\right] + I_{T+1}\right\} B_{T+1}'.$$

The total number of parameters are five: the autoregressive parameter ρ, the error variance σ^2, and the three nonredundant elements of ω_τ^2. The parameters need to be well-behaved as the sample size N grows, and inference is based on the asymptotic normality of their respective estimators. If different series start at different points in time (as typically happens with firms' data), then the parameter $\omega_{y_0}^2$ may not be well-behaved or its estimator may not be asymptotically normal. For example, moment equations depend on terms such as

$$\sum_{i=1}^{N} y_{i,0} u_{i,t}.$$

This term may not be approximately normal if the bulk of observations $y_{i,0}$ are close to zero. This happens because the variance ratio of some terms $y_{i,0}u_{i,t}$ to their sum,

$$\max_{j\leq N} \frac{Var\left(y_{j,0}u_{j,t}\right)}{\sum_{i=1}^{N} Var\left(y_{i,0}u_{i,t}\right)}$$

may not be negligible.[1] This problem is mitigated by shocks over time, hence conditioning on the observations $y_{i,1}$ is more robust.

6.2 The Correlated Random Effects Estimator

For us to make a connection to the CRE estimator, we need to include the vector of ones in $x = 1_N$. The reason is that the (correlated) random effects assumption

$$\left[\begin{array}{cc} y_0 & \eta \end{array}\right] |x \sim N\left(x.\iota, I_N \otimes \Phi\right), \text{ where } \iota = \left(\iota_1, \iota_2\right).$$

would need to be invariant to our group of transformations to be decomposed into an invariant uniform prior and an additional term (and we want to allow the random effects to have a nonzero mean even without additional regressors).

The setup is the same as treating the initial condition as an incidental parameter. However, we include $x = 1_N$ and define the regression coefficient on the vector of ones, as in Section 7 of Chamberlain and Moreira (2009). The maximal invariant is

$$1_N' Y_{T+1} \text{ and } Y_{T+1}' Y_{T+1}.$$

Their (conditional on $\tau = \left[\begin{array}{cc} y_0 & \eta \end{array}\right]$) expectation is given by

$$\begin{aligned} E\left[1_N' Y_{T+1}\right] &= \delta_{\tau,1_N} \left[\begin{array}{c} \rho.e_1' \\ 1_{T+1}' \end{array}\right] B_{T+1}' \text{ and} \\ E\left[\frac{Y_{T+1}' Y_{T+1}}{N}\right] &= B_{T+1} \left\{ \left[\begin{array}{cc} \rho.e_1 & 1_{T+1} \end{array}\right] \left(\sigma^2 \omega_{\tau,1_N}^2 + \delta_{\tau,1_N}' \delta_{\tau,1_N}\right) \left[\begin{array}{c} \rho.e_1' \\ 1_{T+1}' \end{array}\right] + \sigma^2 . I_{T+1} \right\} B_{T+1}'. \end{aligned}$$

The unknown parameters are the autoregressive coefficient ρ, the error variance σ^2, the sample averages $\delta_{\tau,1_N}$, and standardized squared deviations $\omega_{\tau,1_N}^2$. Hence, the invariance argument reduces the data space to $T + 1 + (T+1)(T+2)/2$ moment conditions, which depend on seven parameters. The parameter $\omega_{\tau,1_N}^2$ may not be well-behaved and its estimator may not be asymptotically normal.

Under the random effects assumption, the model becomes

$$Y_{T+1} = 1_N . a\left(\gamma\right) + U_{T+1} . c\left(\gamma\right),$$

[1] The Lindeberg condition holds if and only if the series is approximately normal and the terms are asymptotically negligible.

where the coefficients are given by

$$\begin{aligned} a(\gamma) &= \iota \begin{bmatrix} \rho e_1' \\ 1_{T+1}' \end{bmatrix} B_{T+1}' \text{ and} \\ c(\gamma)' c(\gamma) &= B_{T+1} \left\{ \begin{bmatrix} \rho.e_1 & 1_{T+1} \end{bmatrix} \Phi \begin{bmatrix} \rho e_1' \\ 1_{T+1}' \end{bmatrix} + \sigma^2 I_{T+1} \right\} B_{T+1}'. \end{aligned}$$

The (unconditional) expectation of the maximal invariant has the same functional form as the moments conditional on $\tau = \begin{bmatrix} y_0 & \eta \end{bmatrix}$:

$$\begin{aligned} E\left[\frac{1_N' Y_{T+1}}{N}\right] &= \iota \begin{bmatrix} \rho.e_1' \\ 1_{T+1}' \end{bmatrix} B_{T+1}' \text{ and} \\ E\left[\frac{Y_{T+1}' Y_{T+1}}{N}\right] &= B_{T+1} \left\{ \begin{bmatrix} \rho.e_1 & 1_{T+1} \end{bmatrix} (\Phi + \iota'\iota) \begin{bmatrix} \rho.e_1' \\ 1_{T+1}' \end{bmatrix} + \sigma^2.I_{T+1} \right\} B_{T+1}'. \end{aligned}$$

So the criticism on lack of robustness to the data-generating process for $y_{i,0}$ is applicable here as well.

6.3 Blundell and Bond (1998)

Blundell and Bond (1998) instead consider a different assumption, that y_0 does not deviate systematically from the stationary mean $\eta/(1-\rho)$. We start with the assumption

$$y_0 \sim N\left(\frac{\eta}{1-\rho}, \sigma_0^2.I_N\right).$$

We assume normality for the purposed of invariance. However, only the mean and variance are relevant for the expectation calculations derived below.

Under this assumption, the model is equivalent to

$$Y_{T+1} = \eta.b(\gamma) + U_{T+1}.c(\gamma),$$

where the coefficients are given by

$$\begin{aligned} b(\gamma) &= 1_{T+1}' B_{T+1}' + \frac{\rho}{1-\rho} e_1' B_{T+1} \text{ and} \\ c(\gamma)' c(\gamma) &= B_{T+1} \left(\sigma_0^2 \rho^2 e_1 e_1' + \sigma^2 I_{T+1}\right) B_{T+1}'. \end{aligned}$$

The maximal invariant is given by $Y_{T+1}' Y_{T+1}$:

$$\begin{aligned} E\,\frac{Y_{T+1}' Y_{T+1}}{N} &= B_{T+1} \left(\sigma_0^2 \rho^2 e_1 e_1' + \sigma^2 I_{T+1}\right) B_{T+1}' \\ &+\sigma^2 \omega_\eta^2 B_{T+1} \left[1_{T+1} 1_{T+1}' + \frac{\rho}{1-\rho}\left(e_1 1_{T+1}' + 1_{T+1} e_1'\right) + \frac{\rho^2}{(1-\rho)^2} e_1 e_1'\right] B_{T+1}'. \end{aligned}$$

This expectation depends on four parameters: ρ, σ^2, ω_η^2, and σ_0^2.

We can re-arrange some of these $(T+1)\,(T{+}2)/2$ moments so that the expectation is zero at the true autoregressive coefficient ρ. For example,

$$E\left[\left(y_{i,2}-y_{i,1}\right).\left(y_{i,3}-\rho y_{i,2}\right)\right]=0. \tag{27}$$

As Blundell and Bond (1998) also note, this expectation may fail to be zero if the stationarity assumption breaks down. For example, this moment condition breaks down if either $y_{i,1}=k$ or even if $y_{i,1}\overset{iid}{\sim}N\left(0,\sigma_0^2\right)$.

It is worthwhile making a connection between the stationarity assumption and inference conditional on the first observations. Since inference is conditional on y_1, we could stack the quantities together and look at the (conditional) expectation of

$$Y_{T+1}'Y_{T+1}=\left[\begin{array}{cc} y_1'y_1 & y_1'Y_T \\ Y_T'y_1 & Y_T'Y_T \end{array}\right].$$

As in the derivation based on that of Blundell and Bond (1998), we have exactly the same quantity $Y_{T+1}'Y_{T+1}$. However, we have five parameters (if we consider $y_1'y_1$ itself to be the parameter). Alternatively, we lose one moment (from removing $y_1'y_1$ itself) and four parameters. If the stationarity assumption is correct, there should be efficiency loss from making inference conditional on the first observation. On the other hand, conditional inference should be robust to different data-generating process for the initial data values.

7 Conclusion

This paper studies the Bayesian estimator of Lancaster (2002), which is invariant to natural rotations of the data. The likelihood of the maximal invariant can be divided into two parts, one of which is maximized to obtain Lancaster's estimator. The second part is not asymptotically negligible and, as such, can be used to attain a more efficient estimator. A natural conclusion is that it is unnecessary to use the data to choose the correct inflection point of Lancaster's score likelihood. Lancaster (2002)'s theory is based on differencing the data using the first observation. Instead, we can condition on the first observation itself to further improve efficiency. The conditional argument is essentially a fixed-effects approach in which we make no further assumptions on unobserved data.

Current practice in economics uses standard GMM methods based on data differencing; e.g., Athanasoglou, Brissimis, and Delis (2008) on bank profitability, Guiso, Pistaferri, and Schivardi (2005) on risk allocation, Konings and Vandenbussche (2005) on antidumping protection effects, and Topalova and Khandelwal (2011) on tariffs' change on firm productivity, among others. Instead, we advocate using moments based on invariant statistics from a model that allows for heteroskedasticity and factor structure. It is relatively straightforward to extend our method to allow for covariates in the linear model of Chamberlain and Moreira (2009). Bai (2013a) allows for general time-series heteroskedasticity as opposed to assuming specific moving average (MA) processes and data values lagged enough as instruments. Chamberlain and Moreira (2009) and Moon and Weidner (2015) suggest including factors which generalize individual and time effects. Conditioning on the first observation can then provide reliable inference in dynamic panel data models using moments based on invariant statistics for more general models.

References

Abramowitz, M., and I. A. Stegun (1965): *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables.*

Ahn, S., and P. Schmidt (1995): "Efficient Estimation of Models for Dynamic Panel Data," *Journal of Econometrics*, 68, 5–27.

Arellano, M., and S. Bond (1991): "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *Review of Economic Studies*, 58, 277–297.

Athanasoglou, P. P., S. N. Brissimis, and M. D. Delis (2008): "Bank-specific, industry-specific and macroeconomic determinants of bank profitability," *Journal of International Financial Markets, Institutions and Money*, 18, 121–136.

Bai, J. (2013a): "Fixed-Effects Dynamic Panel Models, A Factor Analytical Method," *Econometrica*, 81, 285–314.

——— (2013b): "Fixed-Effects Dynamic Panel Models, A Factor Analytical Method," *Econometrica*, 81, 185–314, Supplement.

Bickel, P. J., C. A. J. Klaassen, Y. Ritov, and J. A. Wellner (1998): *Efficient and Adaptive Estimation for Semiparametric Models.* Springer.

Blundell, R., and S. Bond (1998): "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models," *Journal of Econometrics*, 87, 115–143.

Blundell, R., and R. Smith (1991): "Conditions Initiales et Estimation Efficace dans les Modeles Dynamiques sur Donnees de Panel: Une Application au Comportement d'Investissement des Entreprises," *Annales d'Economie et de Statistique*, 20.

Cabral, L. M. B., and J. Mata (2003): "On the Evolution of the Firm Size Distribution: Facts and Theory," *American Economic Review*, 93, 1075–1090.

Chamberlain, G., and M. J. Moreira (2009): "Decision Theory Applied to a Linear Panel Data Model," *Econometrica*, 77, 107–133.

Dhaene, G., and K. Jochmans (2016): "Likelihood Inference in an Autoregression with Fixed Effects," *Econometric Theory*, 32, 1–38.

Evans, D. S. (1987a): "The Relationship Between Firm Growth, Size, and Age: Estimates for 100 Manufacturing Industries," *The Journal of Industrial Economics*, 35, 567–581.

——— (1987b): "Tests of Alternative Theories of Firm Growth," *Journal of Political Economy*, 95, 657–674.

Guiso, L., L. Pistaferri, and F. Schivardi (2005): "Insurance within the Firm," *Journal of Political Economy*, 113, 1054–1087.

Hall, B. H. (1987): "The Relationship between Firm Size and Firm Growth in the US Manufacturing Sector," *The Journal of Industrial Economics*, 35, 583–606.

Konings, J., and H. Vandenbussche (2005): "Antidumping Protection and Markups of Domestic Firms," *Journal of International Economics*, 65, 151–165.

Kruiniger, H. (2014): "A Further Look at Modified ML Estimation of the Panel AR(1) Model with Fixed Effects and Arbitrary Initial Conditions," Working Paper, Durham University.

Lancaster, T. (2002): "Orthogonal Parameters and Panel Data," *The Review of Economic Studies*, 69, 647–666.

Moon, H. R., and M. Weidner (2015): "Linear Regression for Panel with Unknown Number of Factors as Interactive Fixed Effects," *Econometrica*, 83, 1543–1579.

Moreira, M. J. (2009): "A Maximum Likelihood Method for the Incidental Parameter Problem," *Annals of Statistics*, 37, 3660–3696.

Topalova, P., and A. Khandelwal (2011): "Trade Liberalization and Firm Productivity: The Case of India," *The Review of Economics and Statistics*, 93, 995–1009.