Robust Likelihood Estimation of Dynamic Panel Data Models*

Javier Alvarez[1] Manuel Arellano[2]

This version: September 2004
Preliminary and Incomplete

Abstract

We develop likelihood-based estimators for autoregressive panel data models that are consistent in the presence of time series heteroskedasticity.

[1]Universidad de Alicante, Departamento de Fundamentos del Análisis Económico, 03071 Alicante, Spain (e-mail: alvarez@merlin.fae.ua.es).
[2]CEMFI, Casado del Alisal 5, 28014 Madrid, Spain. (e-mail: arellano@cemfi.es).

*We thank Olympia Bover and Enrique Sentana for helpful discussions, and Gema Zamarro for very able research assistance.

1. Introduction

The generalized method of moments (GMM) is routinely employed in the estimation of autoregressive models from short panels, because it provides simple estimates that are fixed-T consistent and optimally enforce the model's restrictions on the data covariance matrix. Yet they are known to frequently exhibit poor properties in finite samples and may be asymptotically biased if T is not treated as fixed.

There are also available in the literature fixed-T consistent maximum likelihood methods that are likely to have very different properties to GMM in finite samples and double asymptotics. This category includes random effects estimators of the type considered by Blundell and Smith (1991) and Alvarez and Arellano (2003), the conditional likelihood estimator in Lancaster (2002), and the estimators for first-differenced data in Hsiao, Pesaran, and Tahmiscioglu (2002). However, the existing likelihood-based estimators require that the error variances remain constant through time for fixed-T consistency. Lack of robustness to time series heteroskedasticity is an important limitation because the dispersion of the cross-sectional distribution of errors at each period may differ not only due to nonstationarity at the individual level but also as a result of aggregate effects.

In this paper we develop likelihood-based estimators of autoregressive models that are robust in the sense that remain consistent under the same assumptions as standard panel GMM procedures.[1] From a GMM perspective, likelihood-based estimation can be motivated as a way of reducing the number of moments available for estimation, and hence the extent of bias in second-order or double asymptotics.

We report numerical calculations of relative asymptotic variances, and provide an empirical illustration in the context of individual earnings dynamics.

[1] Cf. Holtz-Eakin, Newey, and Rosen (1988), Arellano and Bond (1991), Arellano and Bover (1995), and Ahn and Schmidt (1995).

2. Model and Assumptions

We consider an autoregressive model for panel data given by

$$y_{it} = \alpha y_{i(t-1)} + \eta_i + v_{it} \quad (t = 1, ..., T; i = 1, ..., N). \tag{2.1}$$

The variables $(y_{i0}, ..., y_{iT})$ are observed but η_i is an unobservable individual effect.[2] We abstract from additive aggregate effects by regarding y_{it} as a deviation from a time effect. It is convenient to introduce the notation $x_{it} = y_{i(t-1)}$ and write the model in the form:

$$y_i = \alpha x_i + \eta_i \iota + v_i \tag{2.2}$$

where $y_i = (y_{i1}, ..., y_{iT})'$, $x_i = (x_{i1}, ..., x_{iT})'$, ι is a $T \times 1$ vector of ones, and $v_i = (v_{i1}, ..., v_{iT})'$.

The following assumption will be maintained throughout:

Assumption A : $\{\eta_i, y_{i0}, y_{i1}, ..., y_{iT}\}_{i=1}^{N}$ is a random sample from a well defined joint distribution with finite second-order moments that satisfies $|\alpha| < 1$ and

$$E\left(v_{it} \mid \eta_i, y_{i0}, ..., y_{i(t-1)}\right) = 0 \quad (t = 1, ..., T). \tag{2.3}$$

This is our core condition in the sense that we wish to consider estimators that are consistent and asymptotically normal for fixed T and large N under Assumption A.

Note that neither time series or conditional heteroskedasticity are assumed. That is, the unconditional variances of the errors, denoted as

$$E\left(v_{it}^2\right) = \sigma_t^2, \tag{2.4}$$

are allowed to change with t and to differ from the conditional variances

$$E\left(v_{it}^2 \mid \eta_i, y_{i0}, ..., y_{i(t-1)}\right).$$

[2] For notational convenience we assume that y_{i0} is observed, so that the actual number of time series observations in the data is $T^o = T + 1$.

Time series homoskedasticity is a particularly restrictive assumption in the context of short micropanels, both because estimators that enforce homoskedasticity are inconsistent when the assumption fails, and because it can be easily violated if aggregate effects are present in the conditional variance of the process.

Also note that we assume stability of the process but not stationarity in mean. Let the covariance matrix of (η_i, y_{i0}) be denoted as

$$var\begin{pmatrix} \eta_i \\ y_{i0} \end{pmatrix} = \begin{pmatrix} \sigma_\eta^2 & \gamma_{\eta 0} \\ \gamma_{\eta 0} & \gamma_{00} \end{pmatrix}. \tag{2.5}$$

Model (2.1) can be written as

$$y_{it} = \left(\frac{1-\alpha^t}{1-\alpha}\right)\eta_i + \alpha^t y_{i0} + \left(v_{it} + \alpha v_{i(t-1)} + ... + \alpha^{t-1} v_{i1}\right). \tag{2.6}$$

Thus, for large t $E(y_{it} \mid \eta_i)$ tends to the steady state mean $\mu_i = \eta_i / (1-\alpha)$. If the process started in the distant past we would have

$$y_{i0} = \frac{\eta_i}{(1-\alpha)} + \sum_{j=0}^{\infty} \alpha^j v_{i(-j)}, \tag{2.7}$$

implying $\gamma_{\eta 0} = \sigma_\eta^2 / (1-\alpha)$ and $\gamma_{00} = \sigma_\eta^2 / (1-\alpha)^2 + \sum_{j=0}^{\infty} \alpha^{2j} \sigma_{-j}^2$.[3] However, here $\gamma_{\eta 0}$ and γ_{00} are treated as free parameters. Note that an implication of lack of stationarity in mean is that the data in first differences will generally depend on individual effects.

In a short panel, steady state assumptions about initial observations are also critical since estimators that impose them lose consistency if the assumptions fail. Moreover, there are relevant applied situations in which a stable process approximates well the dynamics of data, and yet there are theoretical or empirical grounds to believe that the distribution of initial observations does not coincide with the steady state distribution of the process (cf. Hause, 1980 or Barro and Sala-i-Martin, 1995, and discussion in Arellano, 2003).

[3] With the addition of homoskedasticity $\gamma_{00} = \sigma_\eta^2 / (1-\alpha)^2 + \sigma^2 / \left(1-\alpha^2\right)$.

3. Bias-Corrected Conditional Score Estimation

3.1. Normal Likelihood Given Initial Observations and Effects

Under the normality assumption

$$y_{it} \mid y_i^{t-1}, \eta_i \sim \mathcal{N}\left(\alpha y_{i(t-1)} + \eta_i, \sigma_t^2\right) \quad (t = 1, ..., T), \qquad (Assumption\ G1)$$

the log density of y_i conditioned on (y_{i0}, η_i) is given by

$$\ln f\left(y_i \mid y_{i0}, \eta_i\right) = -\frac{1}{2}\ln\det\Lambda - \frac{1}{2}v_i'\Lambda^{-1}v_i \tag{3.1}$$

where Λ is a diagonal matrix with elements $\left(\sigma_1^2, ..., \sigma_T^2\right)$.

The MLE of η_i for given $\alpha, \sigma_1^2, ..., \sigma_T^2$ that maximizes (3.1) is

$$\widehat{\eta}_i = \overline{y}_i - \overline{x}_i\alpha \tag{3.2}$$

where $\overline{y}_i$ denotes a weighted average $\overline{y}_i = \sum_{t=1}^T \varphi_t y_{it}$ with weights

$$\varphi_t = \frac{\sigma_t^{-2}}{\sigma_1^{-2} + ... + \sigma_T^{-2}}. \tag{3.3}$$

Concentrating the log likelihood function with respect to the individual effects we obtain

$$L^* = \frac{N}{2}\ln\det\Phi - \frac{NT}{2}\ln\omega_T - \frac{1}{2\omega_T}\sum_{i=1}^N v_i'\left(\Phi - \Phi\iota\iota'\Phi\right)v_i \tag{3.4}$$

where Φ is a diagonal matrix with elements $(\varphi_1, ..., \varphi_T)$ and ω_T is the variance of the weighted average error:

$$\omega_T = Var\left(\overline{v}_i\right) = \frac{1}{\sigma_1^{-2} + ... + \sigma_T^{-2}}. \tag{3.5}$$

It is useful at this point to note that the following identities hold:

$$v_i' D' \left(D\Lambda D'\right)^{-1} D v_i = \frac{1}{\omega_T} v_i' \left(\Phi - \Phi\iota\iota'\Phi\right) v_i = \sum_{t=1}^{T} \frac{\left(v_{it} - \overline{v}_i\right)^2}{\sigma_t^2} \tag{3.6}$$

$$\ln\det\left(D\Lambda D'\right) = -\ln\det\Phi + (T-1)\ln\omega_T \tag{3.7}$$

where D is the $(T-1)\times T$ first-difference matrix operator. Thus, L^* can be equally regarded as a function of the data in first differences or in deviations from (weighted) means. Note that with $T = 3$ (i.e. Four time series observations per unit), $D\Lambda D'$ is unrestricted:

$$D\Lambda D' = \begin{pmatrix} \sigma_1^2 + \sigma_2^2 & -\sigma_2^2 \\ -\sigma_2^2 & \sigma_2^2 + \sigma_3^2 \end{pmatrix}.$$

Moreover, the relationship between period-specific and within-group variances is given by

$$\sigma_t^2 = E\left[\left(v_{it} - \overline{v}_i\right)^2\right] + \omega_T \quad (t = 1, ...T). \tag{3.8}$$

The MLE of α for given weights is the following heteroskedastic within-groups estimator

$$\widehat{\alpha} = \left[\sum_{i=1}^{N}\sum_{t=1}^{T}\varphi_t\left(x_{it} - \overline{x}_i\right)^2\right]^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\varphi_t\left(x_{it} - \overline{x}_i\right)\left(y_{it} - \overline{y}_i\right), \tag{3.9}$$

which in first differences can also be written as

$$\widehat{\alpha} = \left[\sum_{i=1}^{N} x_i' D' \left(D\Lambda D'\right)^{-1} D x_i\right]^{-1}\sum_{i=1}^{N} x_i' D' \left(D\Lambda D'\right)^{-1} D y_i. \tag{3.10}$$

Finally, the MLE of ω_T for given weights is

$$\widehat{\omega}_T = \frac{1}{TN}\sum_{i=1}^{N}\sum_{t=1}^{T}\varphi_t\left(v_{it} - \overline{v}_i\right)^2.$$

Note that, in common with the situation under homoskedasticity, both $\widehat{\alpha}$ and $\widehat{\omega}_T$ suffer from the incidental parameters problem. Firstly, although x_{it} and v_{it} are orthogonal, their deviations, $(x_{it} - \overline{x}_i)$ and $(v_{it} - \overline{v_i})$, are not, leading to a bias in $\widehat{\alpha}$. Secondly, $\widehat{\omega}_T$ evaluated at the true errors and weights will be inconsistent for fixed T due to lack of degrees of freedom adjustment, as evidenced by the equality

$$\omega_T = E\left[\frac{1}{(T-1)}\sum_{t=1}^{T}\varphi_t\left(v_{it} - \overline{v}_i\right)^2\right]. \tag{3.11}$$

3.2. Likelihood Conditioned on the ML Estimates of the Effects

Provided $G1$ holds, the ML estimates of the effects at the true values of the common parameters $\widehat{\eta}_i = \eta_i + \overline{v}_i$ satisfy

$$\widehat{\eta}_i \mid y_{i0}, \eta_i \sim \mathcal{N}\left(\eta_i, \omega_T\right). \tag{3.12}$$

Moreover, the conditional log density of y_i given $y_{i0}, \eta_i, \widehat{\eta}_i$ is given by

$$\ln f\left(y_i \mid y_{i0}, \eta_i, \widehat{\eta}_i\right) = -\frac{1}{2}\ln\det\left(D\Lambda D'\right) - \frac{1}{2}v_i' D'\left(D\Lambda D'\right)^{-1} D v_i, \tag{3.13}$$

which is a within-group density that does not depend on η_i. Thus, (3.1) admits the decomposition

$$f\left(y_i \mid y_{i0}, \eta_i\right) = f\left(y_i \mid y_{i0}, \widehat{\eta}_i\right) f\left(\widehat{\eta}_i \mid y_{i0}, \eta_i\right), \tag{3.14}$$

which confines the dependence on η_i to the conditional density of $\widehat{\eta}_i$. Similarly, any marginal density for $y_i \mid y_{i0}$ which imposes a prior distribution on the effects can be written as

$$f\left(y_i \mid y_{i0}\right) = f\left(y_i \mid y_{i0}, \widehat{\eta}_i\right) f\left(\widehat{\eta}_i \mid y_{i0}\right). \tag{3.15}$$

The log likelihood conditioned on $\widehat{\eta}_i$ is therefore given by

$$L_C = \frac{N}{2}\ln\det\Phi - \frac{N(T-1)}{2}\ln\omega_T - \frac{1}{2\omega_T}\sum_{i=1}^{N} v_i'\left(\Phi - \Phi\iota\iota'\Phi\right)v_i \tag{3.16}$$

or

$$L_C = -\frac{N}{2}\ln\det\left(D\Lambda D'\right) - \frac{1}{2}\sum_{i=1}^{N} v_i' D' \left(D\Lambda D'\right)^{-1} Dv_i, \tag{3.17}$$

which is similar to the concentrated likelihood (3.4) except that it incorporates a correction for degrees of freedom. In a model with a strictly exogenous x_i, L_C coincides with the likelihood conditioned on sufficient statistics for the effects, which provides consistent estimates of both the regression and residual variance parameters. However, in the autoregressive situation, the estimator of α that maximizes L_C satisfies a heteroskedastic within-group equation of the same form as (3.9) and is therefore inconsistent for fixed T.

Inference from a likelihood conditioned on the ML estimates of the effects may lead to consistent estimates provided the scores of the common parameters and the effects are uncorrelated (Cox and Reid, 1987). Cox and Reid's approximate conditional likelihood approach was motivated by the fact that in an exponential family model, it is optimal to condition on sufficient statistics for the nuisance parameters, and these can be regarded as the MLE of nuisance parameters chosen in a form to be orthogonal to the parameters of interest. From this perspective, the inconsistency of within-groups in the autoregressive model results from lack of orthogonality between the scores of α and the effects.

In the homoskedastic case, Lancaster (2002) showed that a likelihood conditioned on the ML estimate of an orthogonalized effect led to a bias-corrected score and a consistent method-of-moments estimator under homoskedasticity. Following a similar approach, we construct a heteroskedasticity-consistent estimator as the solution to a bias corrected version of the first-order conditions from the likelihood conditioned on the MLE of the effects.

First-Order Conditions The derivatives of L_C with respect to α and $\theta = (\sigma_1^2 ... \sigma_T^2)'$ are given by

$$\frac{\partial L_C}{\partial \alpha} = \sum_{i=1}^{N} x_i' D' \left(D \Lambda D'\right)^{-1} D v_i. \tag{3.18}$$

$$\frac{\partial L_C}{\partial \theta} = \frac{1}{2} \sum_{i=1}^{N} K' \left(D\Lambda D' \otimes D\Lambda D'\right)^{-1} vec\left(D v_i v_i' D' - D\Lambda D'\right) \tag{3.19}$$

where K is a $(T-1)^2 \times T$ selection matrix such that $vec\left(D\Lambda D'\right) = K\theta$.

Maximizing L_C with respect to ω_T and $(\varphi_1 ... \varphi_T)$ for given α, subject to the adding-up restriction $\iota' \Phi \iota = 1$, the first-order conditions for variance parameters can also be written in a form analogous to (3.8) and (3.11) as

$$\sum_{i=1}^{N} \left[\frac{1}{(T-1)} v_i' \left(\Phi - \Phi \iota \iota' \Phi\right) v_i - \omega_T \right] = 0 \tag{3.20}$$

$$\sum_{i=1}^{N} \left[\left(v_{it} - \overline{v}_i\right)^2 - \left(v_{i(t-1)} - \overline{v}_i\right)^2 - \left(\sigma_t^2 - \sigma_{t-1}^2\right) \right] = 0 \quad (t = 2, ..., T). \tag{3.21}$$

Thus, the conditional MLE of α and θ solve, respectively, (3.10) and

$$\widehat{\theta} = \left(K' \Upsilon^{-1} K\right)^{-1} K' \Upsilon^{-1} \frac{1}{N} \sum_{i=1}^{N} vec\left(D v_i v_i' D'\right). \tag{3.22}$$

where $\Upsilon = D\Lambda D' \otimes D\Lambda D'$.

Bias corrected conditional ML scores Under Assumption A the expected conditional ML scores are given by

$$E\left[x_i' D' \left(D\Lambda D'\right)^{-1} D v_i \right] = -h_T(\alpha, \varphi) \tag{3.23}$$

$$E\left[K' \left(D\Lambda D' \otimes D\Lambda D'\right)^{-1} vec\left(D v_i v_i' D' - D\Lambda D'\right) \right] = 0 \tag{3.24}$$

where

$$h_T(\alpha, \varphi) = \sum_{t=1}^{T-1} \left(\frac{1-\alpha^t}{1-\alpha} \right) \varphi_{t+1}. \tag{3.25}$$

Under homoskedasticity $\varphi_t = T^{-1}$ for all t, in which case the bias function (3.25) boils down to

$$h_T(\alpha) = \frac{1}{(1-\alpha)} \left[1 - \frac{1}{T} \left(\frac{1-\alpha^T}{1-\alpha} \right) \right], \tag{3.26}$$

which corresponds to the homoskedastic expression in Nickell (1981) and Lancaster (2002).

In view of (3.23)-(3.24), heteroskedasticity-consistent GMM estimators can be obtained as the solution to the nonlinear estimating equations

$$\sum_{i=1}^{N} x_i' D' (D \Lambda D')^{-1} D v_i + N h_T(\alpha, \varphi) = 0 \tag{3.27}$$

$$K' (D \Lambda D' \otimes D \Lambda D')^{-1} vec \sum_{i=1}^{N} (D v_i v_i' D' - D \Lambda D') = 0. \tag{3.28}$$

Consistency of the bias-corrected score estimator (BCS) that solves (3.27)-(3.28) does not depend on normality nor on conditional or time-series homoskedasticity.

BCS estimation is not possible from a three-wave panel (i.e. $T = 2$) because in that case α is not identified from the expected scores, which are given by

$$E[(y_{i1} - y_{i0})(v_{i2} - v_{i1})] = -\sigma_1^2 \tag{3.29}$$

$$E\left[(v_{i2} - v_{i1})^2\right] = \sigma_1^2 + \sigma_2^2. \tag{3.30}$$

This situation is in contrast with Lancaster's BCS estimator that enforces time series homoskedasticity (hence achieving identification from (3.29)-(3.30)), or the bias-corrected within-group estimator considered in Kiviet (1995).

Modified Conditional Likelihood Interpretation If the weights φ are known, the method of moments estimators of α and ω_T based on the bias corrected scores

$$E\left[x_i' D' \left(D\Phi^{-1}D'\right)^{-1} D v_i\right] = -\omega_T h_T(\alpha, \varphi) \tag{3.31}$$

$$E\left[v_i' D' \left(D\Phi^{-1}D'\right)^{-1} D v_i\right] = (T-1)\,\omega_T \tag{3.32}$$

can be regarded as the maximizers of the criterion function

$$L_{CR} = L_C + N b_T(\alpha, \varphi) \tag{3.33}$$

where

$$b_T(\alpha, \varphi) = \sum_{t=1}^{T-1} \frac{\left(\varphi_{t+1} + ... + \varphi_T\right)}{t} \alpha^t, \tag{3.34}$$

which is the integral of $h_T(\alpha, \varphi)$ up to an arbitrary constant of integration that may depend on φ.

Following Lancaster (2002), L_{CR} can be interpreted as a Cox-Reid likelihood conditioned on the ML estimate $\widehat{\lambda}_i$ of an orthogonal effect λ_i (Arellano, 2003, p. 105)

$$L_{CR} = \sum_{i=1}^{N} \ln f\left(y_i \mid y_{i0}, \widehat{\lambda}_i\right), \tag{3.35}$$

or as an integrated likelihood

$$L_{CR} = \sum_{i=1}^{N} \ln f(y_i \mid y_{i0}) = \sum_{i=1}^{N} \ln f(y_i \mid y_{i0}, \widehat{\eta}_i) + \sum_{i=1}^{N} \ln f(\widehat{\eta}_i \mid y_{i0}) \tag{3.36}$$

in which the chosen prior distribution of the effects conditioned on y_{i0} is such that the marginal density of $\widehat{\eta}_i \mid y_{i0}$ satisfies:

$$f(\widehat{\eta}_i \mid y_{i0}) = \kappa_i(\varphi)\, e^{b_T(\alpha,\varphi)} \tag{3.37}$$

where $\kappa_i(\varphi)$ is a version of the constant of integration.

The first interpretation is based on a decomposition conditional on $\widehat{\lambda}_i$ similar to (3.14), whereas the second relies on factorization (3.15).

4. Random Effects Estimation

The analysis so far was conditional on y_{i0} and $\widehat{\eta}_i$. Conditioning on y_{i0} avoided steady state restrictions, but by conditioning on $\widehat{\eta}_i$ estimation is exclusively based on the data in first-differences. We now turn to explore marginal maximum likelihood estimation based on a normal prior distribution of the effects conditioned on y_{i0}, with linear mean and constant variance. A sufficient condition that we use for simplicity is:

Assumption G2: (η_i, y_{i0}) is jointly normally distributed with an unrestricted covariance matrix.

Normality of y_{i0} is unessential because its variance is a free parameter, so the following analysis can be regarded as conditional on y_{i0}. Clearly, assumptions $G1$ and $G2$ together imply that$(\eta_i, y_{i0}, y_{i1}, ..., y_{iT})$ are jointly normally distributed.

The random effects log likelihood Under $G2$,

$$\widehat{\eta}_i \mid y_{i0} \sim \mathcal{N}\left(\phi y_{i0}, \sigma_\varepsilon^2\right), \tag{4.1}$$

where $\phi = \gamma_{\eta 0}/\gamma_{00}$ and $\sigma_\varepsilon^2 = \omega_T + \sigma_\eta^2 - \gamma_{\eta 0}^2/\gamma_{00}$. So, using factorization (3.15), the density of y_i conditioned on y_{i0} but marginal on η_i is:

$$\begin{aligned} \ln f\left(y_i \mid y_{i0}\right) \propto & -\frac{1}{2}\ln\det\left(D\Lambda D'\right) - \frac{1}{2}v_i' D'\left(D\Lambda D'\right)^{-1} D v_i \\ & -\frac{1}{2}\ln\sigma_\varepsilon^2 - \frac{1}{2\sigma_\varepsilon^2}\left(\overline{y}_i - \alpha\overline{x}_i - \phi y_{i0}\right)^2. \end{aligned} \tag{4.2}$$

Thus, letting $\overline{u}_i = \overline{y}_i - \alpha\overline{x}_i$, the random effects log likelihood is a function of $(\alpha, \sigma_1^2, ..., \sigma_T^2, \phi, \sigma_\varepsilon^2)$ given by

$$L_R = L_C - \frac{N}{2}\ln\sigma_\varepsilon^2 - \frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{N}\left(\overline{u}_i - \phi y_{i0}\right)^2, \tag{4.3}$$

with scores:

$$\frac{\partial L_R}{\partial \alpha} = \frac{\partial L_C}{\partial \alpha} + \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{N} \overline{x}_i \left(\overline{u}_i - \phi y_{i0}\right) \tag{4.4}$$

$$\frac{\partial L_R}{\partial \theta} = \frac{\partial L_C}{\partial \theta} + \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{N} \Phi D' \left(D\Lambda D'\right)^{-1} Du_i \left(\overline{u}_i - \phi y_{i0}\right) \tag{4.5}$$

$$\frac{\partial L_R}{\partial \phi} = \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{N} y_{i0} \left(\overline{u}_i - \phi y_{i0}\right) \tag{4.6}$$

$$\frac{\partial L_R}{\partial \sigma_\varepsilon^2} = \frac{1}{2\sigma_\varepsilon^4}\sum_{i=1}^{N} \left[\left(\overline{u}_i - \phi y_{i0}\right)^2 - \sigma_\varepsilon^2\right]. \tag{4.7}$$

Under Assumption A the expectations of the second terms in the scores for α and θ at true values are:

$$E\left[\frac{1}{\sigma_\varepsilon^2}\overline{x}_i \left(\overline{u}_i - \phi y_{i0}\right)\right] = h_T\left(\alpha, \varphi\right) \tag{4.8}$$

and

$$E\left[\frac{1}{\sigma_\varepsilon^2}\Phi D' \left(D\Lambda D'\right)^{-1} Dv_i \left(\overline{u}_i - \phi y_{i0}\right)\right] = 0. \tag{4.9}$$

Therefore, in view of (3.23) and (3.24), under Assumption A the expected scores evaluated at the true values of the parameters are equal to zero:

$$E\left[x_i' D' \left(D\Lambda D'\right)^{-1} Dv_i + \frac{1}{\sigma_\varepsilon^2}\overline{x}_i \left(\overline{u}_i - \phi y_{i0}\right)\right] = 0$$

$$E\left[\frac{1}{2}K' \left(D\Lambda D' \otimes D\Lambda D'\right)^{-1} vec\left(Dv_i v_i' D' - D\Lambda D'\right) + \frac{1}{\sigma_\varepsilon^2}\Phi D' \left(D\Lambda D'\right)^{-1} Dv_i \left(\overline{u}_i - \phi y_{i0}\right)\right] = 0$$

$$E\left[y_{i0}\left(\overline{u}_i - \phi y_{i0}\right)\right] = 0$$

$$E\left[\left(\overline{u}_i - \phi y_{i0}\right)^2 - \sigma_\varepsilon^2\right] = 0.$$

The random effects maximum likelihood estimator (RML) solves the estimating equations (4.4)-(4.7) and is consistent and asymptotically normal under assumption A regardless of non-normality or conditional heteroskedasticity.

In a three-wave panel ($T = 2$) the model is just-identified and the RML estimator coincides with the Anderson-Hsiao (1981) estimator based on the instrumental-variable condition

$$E\left[y_{i0}\left(\Delta y_{i2} - \alpha \Delta y_{i1}\right)\right] = 0. \tag{4.10}$$

Random effects likelihood functions for homoskedastic autoregressive models under the normality assumption $G2$ have been considered in Blundell and Smith (1991), Sims (2000), and Alvarez and Arellano (2003).

Efficiency Comparisons In order to compare the relative efficiency of the BCS and RML estimators, it is useful to notice that RML is asymptotically equivalent to an overidentified GMM estimator that uses the moment conditions:

$$E\left[x_i' D' \left(D \Lambda D'\right)^{-1} D v_i\right] = -h_T\left(\alpha, \varphi\right) \tag{4.11}$$

$$E\left[K' \left(D\Lambda D' \otimes D\Lambda D'\right)^{-1} vec\left(D v_i v_i' D' - D\Lambda D'\right)\right] = 0 \tag{4.12}$$

$$E\left[\frac{1}{\sigma_\varepsilon^2} \overline{x}_i \left(\overline{u}_i - \phi y_{i0}\right)\right] = h_T\left(\alpha, \varphi\right) \tag{4.13}$$

$$E\left[D' \left(D\Lambda D'\right)^{-1} D v_i \left(\overline{u}_i - \phi y_{i0}\right)\right] = 0 \tag{4.14}$$

$$E\left[y_{i0}\left(\overline{u}_i - \phi y_{i0}\right)\right] = 0 \tag{4.15}$$

$$E\left[\left(\overline{u}_i - \phi y_{i0}\right)^2 - \sigma_\varepsilon^2\right] = 0. \tag{4.16}$$

and a weight matrix calculated under the assumption of normality.

BCS is based on moments (4.11) and (4.12), but RML is also using the information from the data in levels contained in (4.13) and (4.14). Moment (4.13) gives the between-group covariance between the regressor and the error, in the same way as the BCS moment (4.11) specified the within-group covariance. The moments in (4.14) state the orthogonality between within-group and between-group errors (partialling out the initial observation). Finally, (4.15) and (4.16) are unrestricted moments that determine ϕ and σ_ε^2.

Therefore, if the data are normally distributed RML is asymptotically more efficient than BCS. Otherwise, they cannot be ordered. Nevertheless, a GMM estimator based on (4.11)-(4.16) and a robust weight matrix that remains optimal under nonnormality will never be less efficient asymptotically than BCS, and may achieve a significant reduction in the number of moments relative to standard GMM procedures.

The concentrated random effects log-likelihood Concentrating L_R with respect to σ_ε^2 and ϕ we obtain the following criterion function that only depends on α and θ:

$$L_R^* = L_C - \frac{N}{2} \ln \left[(\overline{y} - \alpha \overline{x})' S_0 (\overline{y} - \alpha \overline{x}) \right] \tag{4.17}$$

where $S_0 = I_N - y_0 y_0' / (y_0' y_0)$, and $y_0 = (y_{10}, ..., y_{N0})'$.

L_R^* can be regarded as a modified heteroskedastic within-group criterion with a correction term that becomes less important as T increases.

5. Estimation from the Data in Differences

Until now, the starting point was an interest in the conditional distribution of $(y_{i1}, ..., y_{iT})$ given y_{i0} and η_i under the assumption that y_{i0} was observed but η_i was not. That is, that the data consisted on a random sample of the vectors $(y_{i0}, y_{i1}, ..., y_{iT})$. In this section we maintain the interest in the same conditional distribution as before, but assume that only changes of the y_{it} variables are observed, so that the data on individual i is $(\Delta y_{i1}, ..., \Delta y_{iT})$. This situation is clearly relevant when the data source only provides information on changes, but it may also be interesting if it is thought that an analysis based on changes is more "robust" than one based on levels. An objective of this and the next section is to discuss the content of this intuition by relating ML in differences to the previous conditional and marginal methods. Maximum likelihood estimation of autoregressive models using first-differences has been considered by Hsiao, Pesaran, and Tahmiscioglu (2002).

As a matter of notation, note that observability of $(\Delta y_{i1}, ..., \Delta y_{iT})$ is equivalent to observing $y_i^\dagger = \left(y_{i1}^\dagger, ..., y_{iT}^\dagger\right)' = (y_{i1} - y_{i0}, ..., y_{iT} - y_{i0})'$, since $(\Delta y_{i1}, ..., \Delta y_{iT})' = D^\dagger y_i^\dagger$ where $D^\dagger$ is the $T \times T$ nonsingular transformation matrix

$$D^\dagger = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ & & D & & \end{pmatrix}$$

with $\det\left(D^\dagger\right) = 1$. Also note that by construction $y_{i0}^\dagger = y_{i0} - y_{i0} = 0$.

We shall also use the notation $x_i^\dagger = \left(y_{i0}^\dagger, ..., y_{i(T-1)}^\dagger\right)'$, so that $y_i^\dagger = y_i - y_{i0}\iota$ and $x_i^\dagger = x_i - y_{i0}\iota$. Similarly, $\overline{y}_i^\dagger = y_i^{\dagger\prime}\Phi\iota = \overline{y}_i - y_{i0}$, etc. The following is an expression of $\overline{y}_i^\dagger$ that makes explicit the connection to the data in differences:

$$\overline{y}_i^\dagger = \iota'\Phi D^{\dagger-1}\left(\Delta y_{i1}, ..., \Delta y_{iT}\right)' = \sum_{t=1}^{T}\left(\sum_{s=t}^{T}\varphi_s\right)\Delta y_{it}. \qquad (5.1)$$

Under homoskedasticity $\varphi_t = 1/T$ and (5.1) reduces to

$$\overline{y}_i^{\dagger} = \sum_{t=1}^{T} \left(\frac{T-t+1}{T} \right) \Delta y_{it}. \tag{5.2}$$

The original model can be written as

$$y_{i1} - y_{i0} = [\eta_i - (1-\alpha) y_{i0}] + v_{i1} \tag{5.3}$$

$$y_{it} - y_{i0} = \alpha (y_{it} - y_{i0}) + [\eta_i - (1-\alpha) y_{i0}] + v_{it} \quad (t = 2, ..., T). \tag{5.4}$$

Thus, the model for the deviations $y_{it}^{\dagger}$ can be regarded as a version of the original model in which $y_{i0}^{\dagger} = 0$ for all individuals and the effect is given by

$$\eta_i^{\dagger} = \eta_i - (1-\alpha) y_{i0}. \tag{5.5}$$

From the point of view of this section, bundling together y_{i0} and η_i into $\eta_i^{\dagger}$ makes sense because they are both unobserved. The usefulness of this notation is that it allows us to easily obtain densities for the variables in first differences relying on the previous results for the levels

Since the shocks v_{it} remain the same in representation (5.3)-(5.4), applying (3.13) we have

$$\ln f\left(y_i^{\dagger} \mid y_{i0}^{\dagger}, \eta_i^{\dagger}, \widehat{\eta}_i^{\dagger}\right) = -\frac{1}{2} \ln \det (D\Lambda D') - \frac{1}{2} v_i' D' (D\Lambda D')^{-1} D v_i \tag{5.6}$$

where at true values

$$\widehat{\eta}_i^{\dagger} = \overline{y}_i^{\dagger} - \overline{x}_i^{\dagger} \alpha = \eta_i^{\dagger} + \overline{v}_i = \overline{u}_i - (1-\alpha) y_{i0}, \tag{5.7}$$

and following (3.12):

$$\widehat{\eta}_i^{\dagger} \mid y_{i0}^{\dagger}, \eta_i^{\dagger} \sim \mathcal{N}\left(\eta_i^{\dagger}, \omega_T\right). \tag{5.8}$$

Also, mimicking the marginal density decomposition in (3.15):

$$f\left(y_i^{\dagger} \mid y_{i0}^{\dagger}\right) = f\left(y_i^{\dagger} \mid y_{i0}^{\dagger}, \widehat{\eta}_i^{\dagger}\right) \int f\left(\widehat{\eta}_i^{\dagger} \mid y_{i0}^{\dagger}, \eta_i^{\dagger}\right) dG\left(\eta_i^{\dagger} \mid y_{i0}^{\dagger}\right). \tag{5.9}$$

Moreover, since $y_{i0}^{\dagger}=0$ with probability one, $f\left(y_i^{\dagger}\right)=f\left(y_i^{\dagger}\mid y_{i0}^{\dagger}\right)$ and

$$\widehat{\eta}_i^{\dagger}\mid\eta_i^{\dagger}\sim\mathcal{N}\left(\eta_i^{\dagger},\omega_T\right), \tag{5.10}$$

so that

$$f\left(y_i^{\dagger}\right)=f\left(y_i^{\dagger}\mid\widehat{\eta}_i^{\dagger}\right)\int f\left(\widehat{\eta}_i^{\dagger}\mid\eta_i^{\dagger}\right)dG\left(\eta_i^{\dagger}\right)=f\left(y_i^{\dagger}\mid\widehat{\eta}_i^{\dagger}\right)f\left(\widehat{\eta}_i^{\dagger}\right). \tag{5.11}$$

Recall that the density $f\left(y_i^{\dagger}\right)$ is also the density of the first-differences of the data $(\Delta y_{i1},...,\Delta y_{iT})$, which we are expressing as the product of the usual within-group conditional density and the marginal density of $\widehat{\eta}_i^{\dagger}$. Therefore, in the absence of steady state assumptions about initial conditions, the form of the density of panel AR(1) data in first differences depends on the distribution of the effects. In the next section we shall see that this dependence vanishes under the assumption of mean stationarity.

Let $\sigma_{\varepsilon}^{2\dagger}$ denote the variance of $\widehat{\eta}_i^{\dagger}$, which in general satisfies

$$\sigma_{\varepsilon}^{2\dagger}=\sigma_{\eta}^{2}+(1-\alpha)^{2}\gamma_{00}-2(1-\alpha)\gamma_{0\eta}+\omega_T. \tag{5.12}$$

Under the assumption that $\eta_i^{\dagger}$ is normally distributed (as implied by $G2$, or by the assumption that the marginal distribution of Δy_{i1} is normal)

$$\widehat{\eta}_i^{\dagger}\sim\mathcal{N}\left(0,\sigma_{\varepsilon}^{2\dagger}\right),$$

we have the following "random effects" marginal log density for the data in first differences

$$\begin{aligned}\ln f(\Delta y_{i1},...,\Delta y_{iT}) &= -\frac{1}{2}\ln\det\left(D\Lambda D'\right)-\frac{1}{2}v_i'D'\left(D\Lambda D'\right)^{-1}Dv_i \\ &\quad -\frac{1}{2}\ln\sigma_{\varepsilon}^{2\dagger}-\frac{1}{2\sigma_{\varepsilon}^{2\dagger}}\left(\overline{y}_i^{\dagger}-\overline{x}_i^{\dagger}\alpha\right)^2 \end{aligned} \tag{5.13}$$

where the error in the last term can be expressed in first-differences as $\overline{y}_i^{\dagger}-\overline{x}_i^{\dagger}\alpha=\sum_{t=1}^{T}\left(\sum_{s=t}^{T}\varphi_s\right)\Delta v_{it}$.

Therefore, the random effects log likelihood for the data in first-differences is a function of $\left(\alpha, \sigma_1^2, ..., \sigma_T^2, \sigma_\varepsilon^{2\dagger}\right)$ given by

$$L_{RD} = L_C - \frac{N}{2} \ln \sigma_\varepsilon^{2\dagger} - \frac{1}{2\sigma_\varepsilon^{2\dagger}} \sum_{i=1}^{N} \left(\overline{y}_i^\dagger - \overline{x}_i^\dagger \alpha\right)^2. \tag{5.14}$$

Concentrating L_{RD} with respect to $\sigma_\varepsilon^{2\dagger}$ we obtain the following criterion function that only depends on α and θ:

$$L_{RD}^* = L_C - \frac{N}{2} \ln \left(\overline{y}^\dagger - \alpha \overline{x}^\dagger\right)' \left(\overline{y}^\dagger - \alpha \overline{x}^\dagger\right), \tag{5.15}$$

which, in common with (4.17), can be regarded as a modified heteroskedastic within-group criterion with a small T correction term.

The random effects ML estimator in first-differences (RML-dif) maximizes L_{RD}^* and is consistent and asymptotically normal under assumption A regardless of nonnormality or conditional heteroskedasticity.

Underidentification in a Three-Wave Panel ($T = 2$) In common with BCS, RML-dif estimation is not possible from a three-wave panel because α is not identified from the expected scores of L_{RD}. In contrast, RML achieves identification by relying on the data in levels. The relationship between the two procedures is best seen by examining the covariance matrix of the transformed series

$$Var \begin{pmatrix} y_{i0} \\ \Delta y_{i1} \\ \Delta y_{i2} \end{pmatrix} = \Omega^* = \begin{pmatrix} \gamma_{00} & \gamma_{0\Delta 1} & \gamma_{0\Delta 2} \\ \gamma_{0\Delta 1} & & \\ \gamma_{0\Delta 2} & & \Omega_\Delta \end{pmatrix},$$

where Ω^* is a non-singular transformation of the covariance matrix in levels and Ω_Δ is the covariance matrix in first-differences. Thus, a model of Ω_Δ is equivalent to a model of Ω^* that leaves the coefficients γ_{00}, $\gamma_{0\Delta 1}$ and $\gamma_{0\Delta 2}$ unrestricted (Arellano, 2003, p. 67). With $T = 2$, the only identifying information about α is precisely the restriction $\gamma_{0\Delta 2} = \alpha \gamma_{0\Delta 1}$ satisfied by those coefficients, hence lack of identification from Ω_Δ. Under time series homoskedasticity, α is identifiable from Ω_Δ when $T = 2$, but in that case all the information comes from the homoskedasticity assumption.

Efficiency Comparisons If the data are normally distributed RML is asymptotically more efficient than RML-dif, which in turn is more efficient than BCS. The relative efficiency of RML-dif with respect to BCS under normality is a consequence of the fact that both are statistics of the first differenced data, but the former is the maximum likelihood estimator.

In the absence of normality, the estimators cannot be ranked. However, regardless of normality, under Assumption A estimates based on first-differences alone will never be more efficient than an optimal GMM estimator based on the full covariance structure for the data in levels.

6. Estimation Under Stationarity in Mean

In this section we consider conditional and marginal maximum likelihood estimators that allow for time series heteroskedasticity but exploit the stationarity in mean condition discussed in Section 2. Namely, that for every t the mean of y_{it} conditioned on η_i coincides with the steady state mean of the process $\mu_i = \eta_i / (1-\alpha)$. Specifically, we assume:

$$\gamma_{\eta 0} = \frac{\sigma_\eta^2}{(1-\alpha)}. \qquad (Assumption\ B)$$

Under assumptions A and B the correlation between y_{it} and η_i does not depend on t, so that the first differenced data are orthogonal to the effects. This situation led to orthogonality conditions for errors in levels used in the "system" GMM methods considered by Arellano and Bover (1995) and Blundell and Bond (1998). System GMM remained consistent in the presence of time series heteroskedasticity, and the random effects estimator discussed below can be regarded as likelihood-based counterpart to these procedures.

6.1. Conditional Maximum Likelihood Estimation

In order to construct a likelihood conditioned on the ML estimator of the effects under mean stationarity, we consider the following conditional normality assumption for y_{i0} given the effects:

$$y_{i0} \mid \mu_i \sim \mathcal{N}\left(\mu_i, \overline{\sigma}_0^2\right) \qquad (Assumption\ G3)$$

where $\overline{\sigma}_0^2$ can be regarded as $\sum_{j=0}^{\infty} \alpha^{2j} \sigma_{-j}^2$ and satisfies $\overline{\sigma}_0^2 = \gamma_{00} - \sigma_\eta^2 / (1-\alpha)^2$.

Under assumptions $G1$ and $G3$

$$y_i^T \mid \mu_i \sim \mathcal{N}\left(\mu_i \iota, V\right) \qquad (6.1)$$

where $y_i^T = (y_{i0}, y_{i1}, ..., y_{iT})'$, $\overline{\iota}$ denotes a vector of ones of order $(T+1)$, and

$$V = \Gamma \Lambda^{\dagger} \Gamma' \tag{6.2}$$

with $\Lambda^{\dagger} = diag\left(\overline{\sigma}_0^2, \sigma_1^2, ..., \sigma_T^2\right)$ and

$$\Gamma = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ \alpha & 1 & \dots & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ \alpha^T & \alpha^{T-1} & \dots & \alpha & 1 \end{pmatrix}.$$

Thus

$$\ln f\left(y_i^T \mid \mu_i\right) = -\frac{1}{2} \ln \det V - \frac{1}{2}\left(y_i^T - \mu_i \overline{\iota}\right)' V^{-1}\left(y_i^T - \mu_i \overline{\iota}\right). \tag{6.3}$$

The MLE of μ_i for given α and $\Lambda^{\dagger}$ is

$$\widehat{\mu}_i = \left(\overline{\iota}' V^{-1} \overline{\iota}\right)^{-1} \overline{\iota}' V^{-1} y_i^T. \tag{6.4}$$

Next, to obtain the density of y_i^T conditioned on $\widehat{\mu}_i$ (at true values of α and $\Lambda^{\dagger}$), it is simpler to use the transformation matrix

$$\mathcal{H} = \begin{pmatrix} \left(\overline{\iota}' V^{-1} \overline{\iota}\right)^{-1} \overline{\iota}' V^{-1} \\ \overline{D} \end{pmatrix}, \tag{6.5}$$

which transforms y_i^T into $\left(\widehat{\mu}_i, \overline{D} y_i^T\right)$, where $\overline{D}$ denotes the $T \times (T+1)$ first-difference matrix operator. Since $y_i^T \mid \mu_i$ is normal so is $\mathcal{H} y_i^T \mid \mu_i$. Moreover,

$$Var\left(\mathcal{H} y_i^T \mid \mu_i\right) = \begin{pmatrix} \left(\overline{\iota}' V^{-1} \overline{\iota}\right)^{-1} & 0 \\ 0 & \overline{D} V \overline{D}' \end{pmatrix} \tag{6.6}$$

so that $\widehat{\mu}_i$ and $\overline{D} y_i^T$ are conditionally independent. Therefore,

$$f\left(y_i^T \mid \mu_i\right) = f\left(\mathcal{H} y_i^T \mid \mu_i\right) |\det \mathcal{H}| = f\left(\overline{D} y_i^T\right) f\left(\widehat{\mu}_i \mid \mu_i\right). \tag{6.7}$$

This is so because $\overline{D} y_i^T$ is independent of μ_i and the fact that $|\det \mathcal{H}| = 1$ (Arellano, 2003, p. 94).

Therefore, the density of y_i^T conditional on $\widehat{\mu}_i$ does not depend on μ_i and coincides with the density for the data in first differences:

$$f\left(y_i^T \mid \widehat{\mu}_i, \mu_i\right) = \frac{f\left(y_i^T \mid \mu_i\right)}{f\left(\widehat{\mu}_i \mid \mu_i\right)} = f\left(\overline{D} y_i^T\right), \tag{6.8}$$

which is

$$\ln f\left(\overline{D} y_i^T\right) = -\frac{1}{2} \ln \det\left(\overline{D} V \overline{D}'\right) - \frac{1}{2} y_i^{T\prime} \overline{D}' \left(\overline{D} V \overline{D}'\right)^{-1} \overline{D} y_i^T. \tag{6.9}$$

This result is similar to the one discussed by Lancaster (2002) for a homoskedastic stationary model.

Comparison with the Marginal Likelihood for First Differenced Data Thus, the log likelihood conditioned on the ML estimates of the effects under mean stationarity is a function of $(\alpha, \sigma_1^2, ..., \sigma_T^2, \overline{\sigma}_0^2)$ given by

$$L_{CS} = -\frac{N}{2} \ln \det\left(\overline{D} V \overline{D}'\right) - \frac{1}{2} \sum_{i=1}^{N} y_i^{T\prime} \overline{D}' \left(\overline{D} V \overline{D}'\right)^{-1} \overline{D} y_i^T. \tag{6.10}$$

In the previous section we obtained a random effects likelihood (5.14) for data in first-differences without assuming mean stationarity as a function of $\left(\alpha, \sigma_1^2, ..., \sigma_T^2, \sigma_\varepsilon^{2\dagger}\right)$. In general, $\sigma_\varepsilon^{2\dagger}$ satisfies expression (5.12), which under mean stationarity becomes

$$\sigma_\varepsilon^{2\dagger} = (1-\alpha)^2 \overline{\sigma}_0^2, \tag{6.11}$$

but remains a free parameter because so is $\overline{\sigma}_0^2$.

Thus, the restriction of mean stationarity is immaterial to the data in first differences. L_{RD} and L_{CS} are different parameterizations of the same criterion. Depending on ones taste it can be regarded as a mean-stationary conditional likelihood or as a nonstationary random effects likelihood for the first differenced data. In particular the estimator that maximizes L_{CS} (or L_{RD}) will be consistent under Assumption A regardless of mean stationarity.

Note that under homoskedasticity or covariance stationarity the situation is different because $\overline{\sigma}_0^2$ is no longer a free parameter, but tied to the common variance σ^2 through $\overline{\sigma}_0^2 = \sigma^2 / \left(1 - \alpha^2\right)$.

6.2. Random Effects

If in addition to assumptions $G1$ and $G3$ we assume that μ_i is normally distributed (as implied by $G2$), we can obtain the integrated density marginal on μ_i:

$$f\left(y_i^T\right) = \int f\left(y_i^T \mid \mu_i\right) dG\left(\mu_i\right) \tag{6.12}$$

whose log is given by

$$\ln f\left(y_i^T\right) = -\frac{1}{2}\ln\det\Omega - \frac{1}{2} y_i^{T\prime}\Omega^{-1} y_i^T \tag{6.13}$$

with

$$\Omega = \sigma_\mu^2 \iota\iota' + V. \tag{6.14}$$

Therefore, the random effects log likelihood under mean stationarity is a function of $\left(\alpha, \sigma_1^2, ..., \sigma_T^2, \overline{\sigma}_0^2, \sigma_\eta^2\right)$ given by

$$L_{RS} = -\frac{N}{2}\ln\det\Omega - \frac{1}{2}\sum_{i=1}^{N} y_i^{T\prime}\Omega^{-1} y_i^T. \tag{6.15}$$

The random effects ML estimator subject to mean stationarity (RML-s) maximizes L_{RS} and is consistent and asymptotically normal under assumptions A and B regardless of non-normality or conditional heteroskedasticity.

In a three-wave panel ($T = 2$), the mean stationarity assumption imposes one restriction in the data covariance matrix Ω, which corresponds to the orthogonality conditions for the system GMM estimator simulated in Arellano and Bover (1995):

$$\begin{aligned} E\left[y_{i0}\left(\Delta y_{i2} - \alpha\Delta y_{i1}\right)\right] &= 0 \\ E\left[\Delta y_{i1}\left(y_{i2} - \alpha y_{i1}\right)\right] &= 0. \end{aligned}$$

RML-s provides a one-step estimator based on $T+4$ moment conditions that is asymptotically equivalent to two-step GMM system estimators under conditional homoskedasticity, and more efficient than standard one-step system estimators under time series heteroskedasticity.

As in the previous sections, the comparison between conditional and marginal ML estimates under stationarity can be understood as a straightforward comparison between covariance matrices of data in levels and first-differences

Relation to RML without Mean Stationarity Equation (4.3) in Section 4 gave the random effects log likelihood conditioned on y_{i0}. Adding to this expression the likelihood of y_{i0}, we can write the likelihood of y_i^T in the absence of mean stationarity as a function of $(\alpha, \sigma_1^2, ..., \sigma_T^2, \phi, \sigma_\varepsilon^2, \gamma_{00})$ given by

$$L_{RU} = L_R - \frac{N}{2} \ln \gamma_{00} - \frac{1}{2\gamma_{00}} y_0' y_0. \tag{6.16}$$

In the parameterization of L_{RU}, mean stationarity can be expressed as the restriction

$$\sigma_\varepsilon^2 = (1-\alpha)\,\phi\,(1-\phi)\,\gamma_{00} + \omega_T. \tag{6.17}$$

Thus, RML-s can also be obtained maximizing L_{RU} subject to (6.17).

7. Calculations of Relative Asymptotic Variances

We perform numerical calculations of the asymptotic variances for various estimators of the autoregressive coefficient. In this draft we report the asymptotic variances of the homoskedastic BCS and RML-dif estimators relative to the homoskedastic RML in levels, calculated under the assumption of normality. Formulae for the asymptotic variances are derived in Appendix A.

The interest of the exercise is in providing information on the efficiency gains that can be expected from the levels of the data, relative to only using first-differences, when RML is the maximum likelihood estimator, and stationarity restrictions are not enforced. In addition, we also get to know about the magnitude of the asymptotic inefficiency of BCS relative to RML-dif under normality.

Figures 1 and 2 show values of the asymptotic standard deviations of the homoskedastic BCS and RML-dif estimators relative to the standard deviation of RML, for non-negative values of α. The calculations are for $T = 2$, 3, and 9, under stationarity and $\sigma^2 = 1$.

The $T = 2$ case is special because in that situation BCS and RML-dif coincide and their ability to identify α rests exclusively on the homoskedasticity restriction.

In Figure 1 the variance of the effects has been set to zero ($\lambda = \sigma_\eta^2/\sigma^2 = 0$), whereas in Figure 2 σ_η^2 and σ^2 are equal ($\lambda = 1$). The relative inefficiency of both estimators increases monotonically with α and decreases with λ and T. Figure 1 shows potentially important efficiency gains from using the levels when $T = 3$ and α is large, but the gains become much smaller when $\lambda = 1$, as shown in Figure 2.

Finally in Figure 3 we explore the impact of nonstationarity. We calculate the same relative inefficiency measures as in the previous figures for different values of the ratio of the actual to the steady state standard deviations of y_0. Thus, under stationarity $\kappa = 1$, and a value of $\kappa = 2$ means that the standard deviation of initial conditions is twice the standard deviation of the steady state standard deviation of the process. We set $T = 3$, $\lambda = 0$, and $\alpha = 0.9$, so that we essentially calculate the maximal inefficiencies for each value of κ. For $\kappa < 1$, the inefficiency of BCS can be enormous, whereas the inefficiency of RML-dif is much smaller and shows a non-monotonic pattern.

8. Empirical Illustration: Individual Earnings Dynamics

In order to illustrate the properties of some of the previous methods, we estimate first-order autoregressive equations for individual labour income using two different samples. The first one is a sample of Spanish men from the European Community Household Panel (ECHP) for the period 1994-1999. The second is a sample from PSID for the period 1977-1983 taken from Alvarez, Browning, and Ejrnæs (2001).

There are 632 individuals in the Spanish data set and 792 in the PSID sample. All individuals in both data sets are married males, who are aged 20-65 during the sample period, heads of household, and continuously employed. The earnings variable is similarly defined in the two samples as total annual labour income of the head.

The variables that we use in the estimation are log earnings residuals from first-stage regressions on age, age squared, education and year dummies (see Alvarez, Browning, and Ejrnæs, 2001, for further details on the PSID sample, and tables A1 and A2 for the Spanish sample). Log earnings have a much higher variance in the PSID sample than in the Spanish one. Moreover, the PSID data show a sharp rise in the variance of earnings in 1982 (a widely documented fact), whereas there is no appreciable change in the variance in the Spanish sample during the (different) years that we observe.

The results for the Spanish data are reported in Table 1. Heteroskedastic bias-corrected score (BCS) and random effects (RMLr) estimates of the autoregressive coefficient are very similar. They are also very close to the homoskedastic random effects estimate (RMLnr), which is not surprising given the absence of change in the period-specific variance estimates reported in the table. By comparison, the GMM estimates (one- and two-step) are very small, given that GMM, BCS, and RMLr are all consistent under similar assumptions. The system GMM estimator, that relies on mean stationarity, is more in line with the likelihood-based estimates, although probably for the wrong reasons, given the rejection of mean stationarity that is apparent from the Sargan test. Finally, within-groups (WG) and the random effects estimate that rules out correlation between the effects and initial observations (RML, $\phi = 0$) exhibit, respectively, the downward and upward biases that would be predicted from theory.

The results for the PSID sample, reported in Table 2, also show a marked discrepancy between the likelihood-based estimates and GMM, and a similar rejection of mean stationarity from the incremental Sargan test. In the PSID data there is more state dependence than in the Spanish data, at least as measured by the autoregressive coefficient. There is also more variation in the errors and substantial time series heteroskedasticity. The latter translate into a small but noticeable upward bias in the RML estimate calculated under the assumption of homoskedasticity.

Given the estimates reported in tables 1 and 2, the variance of the effects can be recovered from $\sigma_\eta^2 = \sigma_\varepsilon^2 + \phi^2 \gamma_{00} - \omega_T$ (as explained in Section 4), which gives $\widehat{\sigma}_\eta^2 = 0.05$ for the Spanish data, and $\widehat{\sigma}_\eta^2 = 0.07$ for the PSID.

The regression coefficient of η on y_0 under mean stationarity is

$$\phi^* = \frac{1}{(1-\alpha)} \frac{\sigma_\eta^2}{\gamma_{00}}.$$

The implied RML estimates of this quantity for the Spanish data ($\widehat{\phi}^* = 0.563$) is very similar to the unrestricted estimate in Table 1, whereas the estimate from the PSID sample ($\widehat{\phi}^* = 0.464$) is somewhat larger than the corresponding unrestricted estimate in Table 2.

GMM estimates are known to be downward biased in finite samples, specially when the number of moments is large and the instruments are weak. However, the magnitude of the bias in our application (relative to the likelihood estimates) is puzzling for the values of α and T/N that we have, unless related to other aspects of the specification of the models. In particular, it may be useful to estimate a second-order autoregressive model. The evidence that we have against second-order autocorrelation in the first differenced errors from the $m2$ test statistics (Arellano and Bond, 1991) is conclusive for the Spanish panel but marginal for the PSID.

Monte Carlo Simulations In order to illustrate the properties of the estimators, we performed a small simulation exercise calibrated to the likelihood-based estimates from the PSID data. We generated 1000 replications with $N = 792$, $T^o = 7$, $\eta_i \sim \mathcal{N}\left(0, \sigma_\eta^2\right)$, $v_{it} \sim \mathcal{N}\left(0, \sigma_t^2\right)$, $\sigma_\eta^2 = 0.07$, and mean stationarity.

In Tables 3 and 4 we report means and standard deviations of the WG, GMM1, RML(nr), RML(r), and BCS estimators for $\alpha = 0.4$ and 0.8, respectively (with $\overline{\sigma}_0^2 = 0.11$ and 0.28). The results show that both RML(r) and BCS are virtually unbiased. Those in Table 3 nicely reproduce the WG downward bias and the RML(nr) upward bias that we found in the PSID sample. However, further investigation under alternative specifications of the data generating process is required, because so far the results fail to explain the performance of GMM with the real data.

9. Concluding Remarks

From a GMM perspective, a motivation for considering likelihood based estimators is to reduce the number of moments available for estimation. The number of orthogonality conditions of optimal GMM estimators in autoregressive panel models grows at a rate of $T(T-1)/2$, whereas the number of score equations for the heteroskedastic likelihood estimators grows at a rate of T. An interesting question is to characterize the potential incidental parameter problem that occurs for these estimators as T tends to infinity.

From ongoing work by the authors, we conjecture that in a double asymptotic setup where T/N tends to a finite constant, the estimators with unrestricted time series variances remain consistent and asymptotically normal, but have a bias term in the asymptotic distribution when the data are not symmetrically distributed.

Table 1
Autoregressive Model of Earnings
Results for Spanish Data, 1994-1999
$N = 632, T^0 = 6$

	WG	GMM1	GMM2	System-GMM
α	-0.022	0.041	0.035	0.183
	(-0.95)	(1.12)	(1.87)	(7.00)
Sargan test (df)			11.99(9)	22.71(13)
$m1$		-9.66	-9.89	
$m2$		-0.27	-0.23	
	Likelihood-based Estimates			
	BCS	RML(r)	RML(nr)	RML
	(robust)	(robust)	(homosk.)	($\phi = 0$)
α	0.210	0.200	0.207	0.926
	(8.03)	(9.14)	(4.39)	(131.9)
σ_1^2 (1995)	0.025	0.023		
	(11.26)	(18.04)		
σ_2^2 (1996)	0.022	0.021		
	(8.46)	(20.45)		
σ_3^2 (1997)	0.022	0.023		
	(6.47)	(21.31)		
σ_4^2 (1998)	0.025	0.023		
	(10.00)	(20.25)		
σ_5^2 (1999)	0.020	0.025		
	(3.77)	(19.36)		
ϕ		0.567		
		(24.67)		
σ_ε^2		0.020		
		(23.52)		
γ_{00}		0.111		

Data are log earnings residuals from a regression on age, education and year dummies. γ_{00} is the sample variance of y_0.
$t-$ratios robust to conditional heteroskedasticity.
$m1$ and $m2$ are serial correlation tests for differenced errors.
$(\phi, \sigma_\varepsilon^2)$ are regression coeffs. of $\left(\overline{y} - \alpha\overline{y}_{-1}\right)$ on y_0.

Table 2
Autoregressive Model of Earnings
Results for PSID Data, 1977-1983
$N = 792, T^0 = 7$

	WG	GMM1	GMM2	System-GMM
α	0.184	0.171	0.157	0.311
	(6.08)	(3.34)	(3.53)	(9.75)
Sargan test (df)			15.61 (14)	46.59 (19)
$m1$		-6.36	-6.39	
$m2$		1.82	1.65	
	Likelihood-based Estimates			
	BCS (robust)	RML(r) (robust)	RML(nr) (homosk.)	RML ($\phi = 0$)
α	0.385	0.367	0.416	0.902
	(9.55)	(31.82)	(18.03)	(118.2)
σ_1^2 (1978)	0.063	0.059		
	(7.74)	(34.76)		
σ_2^2 (1979)	0.065	0.058		
	(5.95)	(39.63)		
σ_3^2 (1980)	0.066	0.052		
	(6.03)	(29.23)		
σ_4^2 (1981)	0.032	0.046		
	(4.16)	(30.56)		
σ_5^2 (1982)	0.102	0.096		
	(3.93)	(84.79)		
σ_6^2 (1983)	0.071	0.091		
	(2.23)	(58.04)		
ϕ		0.384		
		(24.76)		
σ_ε^2		0.045		
		(28.09)		
γ_{00}		0.239		

Data are log earnings residuals from a regression on age, education and year dummies. γ_{00} is the sample variance of y_0.
*See notes to Table 1.

Table 3
Simulations for the Autoregressive Model
Means and standard deviations of the estimators
$N = 792, T^0 = 7$

	WG	GMM	RML(nr)	RML(r)	BCS
α	0.178	0.396	0.430	0.400	0.400
	(0.015)	(0.035)	(0.021)	(0.020)	(0.021)
σ_1^2				0.059	0.059
				(0.003)	(0.004)
σ_2^2				0.058	0.058
				(0.003)	(0.004)
σ_3^2				0.052	0.052
				(0.003)	(0.004)
σ_4^2				0.046	0.046
				(0.003)	(0.005)
σ_5^2				0.096	0.096
				(0.005)	(0.007)
σ_6^2				0.091	0.091
				(0.005)	(0.006)

1000 replications. Parameter values: $\alpha = 0.4$, $\sigma_1^2 = 0.059$, $\sigma_2^2 = 0.058, \sigma_3^2 = 0.052, \sigma_4^2 = 0.046, \sigma_5^2 = 0.096, \sigma_6^2 = 0.091$ $\sigma_\eta^2 = 0.07$, $\overline{\sigma_0^2} = 0.11$.

Table 4
Simulations for the Autoregressive Model
Means and standard deviations of the estimators
$N = 792, T^0 = 7$

	WG	GMM	RML(nr)	RML(r)	BCS
α	0.488 (0.016)	0.772 (0.076)	0.882 (0.028)	0.804 (0.037)	0.805 (0.038)
σ_1^2				0.059 (0.004)	0.059 (0.004)
σ_2^2				0.058 (0.004)	0.058 (0.005)
σ_3^2				0.052 (0.004)	0.052 (0.005)
σ_4^2				0.046 (0.003)	0.046 (0.005)
σ_5^2				0.096 (0.005)	0.096 (0.007)
σ_6^2				0.091 (0.005)	0.091 (0.006)

1000 replications. Parameter values: $\alpha = 0.8$, $\sigma_1^2 = 0.059$, $\sigma_2^2 = 0.058, \sigma_3^2 = 0.052, \sigma_4^2 = 0.046, \sigma_5^2 = 0.096, \sigma_6^2 = 0.091$ $\sigma_\eta^2 = 0.07$, $\overline{\sigma_0^2} = 0.28$.

Table A1
Sample characteristics: Spanish Data, 1994-1999
$N = 632, T^0 = 6$

	Mean	Min	Max
age	43.5	23	65
tenure (years of exp in the job)	13.4	0	20
real labor income (euros)	13296.8	3529.1	72825.8
real capital income (euros)	276.6	0	27761.8
% less than sec educ	28.3		
% secondary educ	46.3		
% university educ	25.4		
% industry	37.0		
% service	63.0		
% private sector	65.0		

Table A2
Regression results first-step
Dependent variable: log of real labor income
Spanish Data, 1994-1999

	Coefficient	t-ratio
constant	7.269	54.98
age	0.076	12.79
age2	-0.001	-11.47
sec educ	0.267	19.98
univ educ	0.717	46.48
private sector	0.073	5.73
services	-0.006	-0.50
d94	-0.040	-2.15
d95	-0.051	-2.79
d96	-0.054	-2.95
d97	-0.049	-2.68
d98	-0.027	-1.50

References

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

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

Alvarez, J., M. Browning, and M. Ejrnæs (2001): "Modelling Income Processes with Lots of Heterogeneity", University of Copenhagen, unpublished.

Anderson, T. W. and C. Hsiao (1981): "Estimation of Dynamic Models with Error Components", *Journal of the American Statistical Association*, 76, 598-606.

Arellano, M. and S. R. 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.

Arellano, M. and O. Bover (1995): "Another Look at the Instrumental-Variable Estimation of Error-Components Models", *Journal of Econometrics*, 68, 29-51.

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

Barro, R. J. and X. Sala-i-Martin (1995): *Economic Growth*, McGraw-Hill, New York.

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), "Initial Conditions and Efficient Estimation in Dynamic Panel Data Models", *Annales d'Economie et de Statistique*, 20/21, 109-123.

Cox, D. R. and N. Reid (1987): "Parameter Orthogonality and Approximate Conditional Inference" (with discussion), *Journal of the Royal Statistical Society*, Series B, 49, 1-39.

Hause, J. C. (1980): ""The Fine Structure of Earnings and the On-the-Job Training Hypothesis", *Econometrica*, 48, 1013-1029.

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

Hsiao, C., M. H. Pesaran, and A. K. Tahmiscioglu (2002): "Maximum Likelihood Estimation of Fixed Effects Dynamic Panel Data Models Covering Short Time Periods", *Journal of Econometrics*, 109, 107–150.

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

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

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

Sims, C. A. (2000): "Using a Likelihood Perspective to Sharpen Econometric Discourse: Three Examples", *Journal of Econometrics*, 95, 443-462.

Appendix

A. Asymptotic Variances of Estimators Under Normality

A.1. Asymptotic Variance of the Homoskedastic RML-dif Estimator

Letting $\eta_i^{\dagger} = \eta_i - (1-\alpha)\, y_{i0}$, the model can be written as

$$\begin{aligned} \Delta y_{i1} &= \eta_i^{\dagger} + v_{i1} \\ \Delta y_{it} &= \alpha \Delta y_{i(t-1)} + \Delta v_{it} \quad (t = 2, ..., T) \end{aligned}$$

or in vector notation

$$B \begin{pmatrix} \Delta y_{i1} \\ \vdots \\ \Delta y_{iT} \end{pmatrix} = D^{\dagger} \begin{pmatrix} \eta_i^{\dagger} + v_{i1} \\ \vdots \\ \eta_i^{\dagger} + v_{iT} \end{pmatrix} \equiv D^{\dagger} u_i^{\dagger}$$

where B and $D^{\dagger}$ are $T \times T$ matrices of the form

$$B = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ -\alpha & 1 & \dots & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \dots & -\alpha & 1 \end{pmatrix}, \quad D^{\dagger} = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ & & D & & \end{pmatrix}.$$

Moreover,

$$Var\left(D^{\dagger} u_i^{\dagger}\right) = D^{\dagger}\left(\sigma^2_{\eta\dagger}\iota\iota' + \Lambda\right) D^{\dagger\prime}$$

where $\sigma^2_{\eta\dagger} = Var\left(\eta_i^{\dagger}\right)$ and under homoskedascity $\Lambda = \sigma^2 I_T$.

Therefore,

$$Var \begin{pmatrix} \Delta y_{i1} \\ \vdots \\ \Delta y_{iT} \end{pmatrix} = B^{-1} D^{\dagger}\left(\sigma^2_{\eta\dagger}\iota\iota' + \Lambda\right) D^{\dagger\prime} B^{-1\prime} \equiv \Omega(\gamma) \qquad \text{(A.1)}$$

where $\gamma = \left(\alpha, \sigma^2, \sigma^2_{\eta\dagger}\right)'$.[4]

[4] The same expression is valid for the heteroskedastic RML-dif except that in that case $\gamma = \left(\alpha, \sigma_1^2, ..., \sigma_T^2, \sigma^2_{\eta\dagger}\right)'$.

Moreover, note that the homoskedastic marginal MLE for the data in differences in (B.14) can be written as

$$\left(\widehat{\alpha}_D, \widehat{\sigma}^2, \widehat{\sigma}^2_{\eta\dagger}\right) = \arg\min \left[\ln\det\Omega\left(\gamma\right) + \frac{1}{N}\sum_{i=1}^{N}\left(\Delta y_{i1}, ..., \Delta y_{iT}\right)\Omega^{-1}\left(\gamma\right)\left(\begin{array}{c} \Delta y_{i1} \\ \vdots \\ \Delta y_{iT} \end{array}\right)\right].$$

Thus, under normality the asymptotic variance matrix of $\left(\widehat{\alpha}_D, \widehat{\sigma}^2, \widehat{\sigma}^2_{\eta\dagger}\right)$ is given by[5]

$$2\left\{H\left(\gamma\right)'\mathcal{D}'\left[\Omega^{-1}\left(\gamma\right)\otimes\Omega^{-1}\left(\gamma\right)\right]\mathcal{D}H\left(\gamma\right)\right\}^{-1}$$

where

$$H\left(\gamma\right) = \frac{\partial vech\left[\Omega\left(\gamma\right)\right]}{\partial\gamma'}$$

and $\mathcal{D}$ is the selection matrix

$$\mathcal{D} = \frac{\partial vec\Omega}{\partial\left(vech\Omega\right)'}.$$

[5] See for example Arellano (2003, p. 72).

A.2. Asymptotic Variance of the Homoskedastic RML-lev Estimator

In order to exploit the previous result for the differences, we express the covariance structure corresponding to the levels using the transformation:

$$Var\begin{pmatrix} y_{i0} \\ \Delta y_{i1} \\ \vdots \\ \Delta y_{iT} \end{pmatrix} = \begin{pmatrix} \gamma_{00} & \gamma_{0\eta\dagger} & \alpha\gamma_{0\eta\dagger} & \ldots & \alpha^{T-1}\gamma_{0\eta\dagger} \\ \gamma_{0\eta\dagger} & & & & \\ \alpha\gamma_{0\eta\dagger} & & \Omega(\gamma) & & \\ \vdots & & & & \\ \alpha^{T-1}\gamma_{0\eta\dagger} & & & & \end{pmatrix} = \Omega^*(\gamma^*)$$

where $\gamma_{00} = Var(y_{i0})$, $\gamma_{0\eta\dagger} = Cov\left(y_{i0}, \eta_i^\dagger\right)$, and $\gamma^* = \left(\alpha, \sigma^2, \sigma^2_{\eta\dagger}, \gamma_{0\eta\dagger}, \gamma_{00}\right)'$.

Arguing as in the previous case, the homoskedastic marginal MLE for the data in levels can be written as

$$\left(\widehat{\alpha}_L, \widetilde{\sigma}^2, \widetilde{\sigma}^2_{\eta\dagger}, \widetilde{\gamma}_{0\eta\dagger}, \widetilde{\gamma}_{00}\right) =$$

$$\arg\min\left[\ln\det\Omega^*(\gamma^*) + \frac{1}{N}\sum_{i=1}^{N}(y_{i0}, \Delta y_{i1}, ..., \Delta y_{iT})\,\Omega^{*-1}(\gamma^*)\begin{pmatrix} y_{i0} \\ \Delta y_{i1} \\ \vdots \\ \Delta y_{iT} \end{pmatrix}\right].$$

Thus, under normality the asymptotic variance matrix of $\left(\widehat{\alpha}_L, \widetilde{\sigma}^2, \widetilde{\sigma}^2_{\eta\dagger}, \widetilde{\gamma}_{0\eta\dagger}, \widetilde{\gamma}_{00}\right)$ is given by

$$2\left\{H^*(\gamma^*)'\mathcal{D}^{*\prime}\left[\Omega^{*-1}(\gamma^*)\otimes\Omega^{*-1}(\gamma^*)\right]\mathcal{D}^*H^*(\gamma^*)\right\}^{-1}$$

where

$$H^*(\gamma^*) = \frac{\partial vech\left[\Omega^*(\gamma^*)\right]}{\partial\gamma^{*\prime}}$$

and $\mathcal{D}^*$ is the selection matrix

$$\mathcal{D}^* = \frac{\partial vec\Omega^*}{\partial\left(vech\Omega^*\right)'}.$$

Note that in this parameterization, under stationary initial conditions, γ_{00} remains a free parameter (which determines σ^2_η) and

$$\begin{aligned} \gamma_{0\eta\dagger} &\equiv Cov\left(y_{i0}, \eta_i^\dagger\right) = -\frac{\sigma^2}{(1+\alpha)} \\ \sigma^2_{\eta\dagger} &\equiv Var\left(\eta_i^\dagger\right) = \left(\frac{1-\alpha}{1+\alpha}\right)\sigma^2. \end{aligned}$$

A.3. Asymptotic Variance of the Homoskedastic Lancaster Estimator

Because of the incidental parameters problem, the ML estimates of α and σ^2 estimated jointly with the effects are inconsistent for fixed T. However, as noted by Lancaster (2002), we can obtain score adjusted (or "degrees of freedom" adjusted) estimators that are consistent in view of the moment relationships:

$$\begin{aligned} E\left(x_i^{*\prime} v_i^*\right) &= -\sigma^2 h_T(\alpha) \\ E\left(v_i^{*\prime} v_i^*\right) &= (T-1)\,\sigma^2 \end{aligned}$$

where x_i^* and v_i^* denote orthogonal deviations of the original variables.

By substituting the second equation we can eliminate σ^2 and get

$$E\left[\psi_i(\alpha)\right] = 0$$

where

$$\psi_i(\alpha) = x_i^{*\prime} v_i^* + v_i^{*\prime} v_i^* \frac{h_T(\alpha)}{(T-1)}. \tag{A.2}$$

Under suitable regularity conditions, if there is a consistent root of the equation $\sum_{i=1}^{N} \psi_i(a) = 0$,[6] its asymptotic variance is given by

$$v_\alpha = \frac{v}{d^2}. \tag{A.3}$$

where

$$v = E\left[\psi_i^2(\alpha)\right]$$

and

$$d = E\left[\frac{\partial \psi_i(\alpha)}{\partial \alpha}\right].$$

Because of

$$\frac{\partial \psi_i(\alpha)}{\partial \alpha} = -x_i^{*\prime} x_i^* - 2 x_i^{*\prime} v_i^* \frac{h_T(\alpha)}{(T-1)} + \frac{v_i^{*\prime} v_i^*}{(T-1)} h_T'(\alpha),$$

we have

$$d = -E\left(x_i^{*\prime} x_i^*\right) + 2\sigma^2 \frac{h_T^2}{(T-1)} + \sigma^2 h_T' \tag{A.4}$$

where we are using h_T and h_T' for shortness.

[6] A formal proof of consistency is given in Lancaster (2002), Theorem A1.

Similarly,

$$v = E\left[\left(x_i^{*\prime}v_i^*\right)^2\right] + E\left[\left(v_i^{*\prime}v_i^*\right)^2\right]\frac{h_T^2}{(T-1)^2} + 2E\left[\left(x_i^{*\prime}v_i^*\right)\left(v_i^{*\prime}v_i^*\right)\right]\frac{h_T}{(T-1)}. \quad (A.5)$$

The availability of expression (A.1) allows us to calculate the term $E\left(x_i^{*\prime}x_i^*\right)$ that appears in (A.4) as follows

$$E\left(x_i^{*\prime}x_i^*\right) = E\left(x_i'D'\left(DD'\right)^{-1}Dx_i\right) = tr\left[\left(DD'\right)^{-1}\Omega_{11}\right] \quad (A.6)$$

where $\Omega_{11} = E\left(Dx_ix_i'D'\right)$ is the $(T-1)\times(T-1)$ north-west submatrix of $\Omega(\gamma)$.

Next, under normality and homoskedasticity we have

$$E\left[\left(x_i^{*\prime}v_i^*\right)^2\right] = \sigma^4h_T^2 + \sigma^2E\left(x_i^{*\prime}x_i^*\right) + \sigma^4 tr\left(QC_TQC_T\right) \quad (A.7)$$

$$E\left[\left(v_i^{*\prime}v_i^*\right)^2\right] = \sigma^4(T+1)(T-1) \quad (A.8)$$

$$E\left[\left(x_i^{*\prime}v_i^*\right)\left(v_i^{*\prime}v_i^*\right)\right] = -\sigma^4h_T(T+1) \quad (A.9)$$

where $Q = I_T - \iota\iota'/T$ and C_T is such that $E\left(x_iv_i'\right) = \sigma^2C_T$.

Thus,

$$v = \sigma^4h_T^2 + \sigma^2E\left(x_i^{*\prime}x_i^*\right) + \sigma^4tr\left(QC_TQC_T\right) - \sigma^4h_T^2\left(\frac{T+1}{T-1}\right)$$

or

$$v = \sigma^2E\left(x_i^{*\prime}x_i^*\right) + \sigma^4tr\left(QC_TQC_T\right) - \frac{2}{(T-1)}\sigma^4h_T^2. \quad (A.10)$$

To get the results (A.7)-(A.9) we have used the following intermediate formulae for moments of quadratic forms in normal variables:

$$\begin{aligned}
E\left[\left(x_i^{*\prime}v_i^*\right)^2\right] &= \left[E\left(x_i^{*\prime}v_i^*\right)\right]^2 + tr\left[E\left(x_i^*x_i^{*\prime}\right)E\left(v_i^*v_i^{*\prime}\right)\right] + tr\left[E\left(x_i^*v_i^{*\prime}\right)E\left(x_i^*v_i^{*\prime}\right)\right] \\
E\left[\left(v_i^{*\prime}v_i^*\right)^2\right] &= tr^2\left[E\left(v_i^*v_i^{*\prime}\right)\right] + 2tr\left[E\left(v_i^*v_i^{*\prime}\right)E\left(v_i^*v_i^{*\prime}\right)\right] = (T-1)^2\sigma^4 + 2\sigma^4(T-1) \\
E\left[\left(x_i^{*\prime}v_i^*\right)\left(v_i^{*\prime}v_i^*\right)\right] &= E\left(x_i^{*\prime}v_i^*\right)E\left(v_i^{*\prime}v_i^*\right) + 2tr\left[E\left(x_i^*v_i^{*\prime}\right)E\left(v_i^*v_i^{*\prime}\right)\right] \\
&= -\sigma^4h_T(T-1) - 2\sigma^4h_T.
\end{aligned}$$

B. Proofs and Intermediate Results

B.1. Bias-Corrected Score Estimation

B.1.1. Heteroskedastic First-Differences, Within-Groups, and Orthogonal Deviations

Let D be the $(T-1)\times T$ first-difference matrix operator. For any $(\sigma_1^2,...,\sigma_T^2)$ and $v=(v_1,...,v_T)'$ the following equivalences hold:

$$\begin{aligned} v'D'\left(D\Lambda D'\right)^{-1}Dv &= \sum_{t=1}^{T}\frac{\left(v_t-\overline{v}\right)^2}{\sigma_t^2}=\frac{1}{\omega_T}\sum_{t=1}^{T}\varphi_t\left(v_t-\overline{v}\right)^2 \\ &= \frac{1}{\omega_T}v'\left(\Phi-\Phi\iota\iota'\Phi\right)v \end{aligned} \quad \text{(B.1)}$$

where $\Lambda = diag\left(\sigma_1^2,...,\sigma_T^2\right)$,

$$\overline{v}=\sum_{s=1}^{T}\varphi_s v_s\equiv v'\Phi\iota$$

$$\varphi_s=\frac{\sigma_s^{-2}}{\sigma_1^{-2}+...+\sigma_T^{-2}},$$

also $\Phi = diag\left(\varphi_1,...,\varphi_T\right)$, and

$$\omega_T=Var\left(\overline{v}\right)=\sum_{s=1}^{T}\varphi_s^2\sigma_s^2=\left(\sigma_1^{-2}+...+\sigma_T^{-2}\right)^{-1}=\left(\iota'\Lambda^{-1}\iota\right)^{-1}$$

so that $\Lambda^{-1}=\left(1/\omega_T\right)\Phi$.

It is also true that

$$\omega_T=E\left[\frac{1}{(T-1)}\sum_{t=1}^{T}\varphi_t\left(v_t-\overline{v}\right)^2\right], \quad \text{(B.2)}$$

so that for fixed T a degrees of freedom correction is needed.

Regarding period-specific variances, taking into account that:

$$E\left[\left(v_t-\overline{v}\right)^2\right]=\sigma_t^2+\omega_T-2E\left(v_t\overline{v}\right)=\sigma_t^2+\omega_T-2\varphi_t\sigma_t^2=\sigma_t^2+\omega_T-2\omega_T$$

we have the following expressions:

$$\sigma_t^2 = E\left[(v_t - \overline{v})^2\right] + \omega_T \tag{B.3}$$

or

$$\sigma_t^2 - \sigma_{t-1}^2 = E\left[(v_t - \overline{v})^2\right] - E\left[(v_{t-1} - \overline{v})^2\right] \quad (t = 2, ...T).$$

Note that $0 \leq \varphi_t \leq 1$, $\sum_{t=1}^{T} \varphi_t = 1$, and that under homoskedasticity $\varphi_t = 1/T$ for all t. Also note that (B.3) can be used to easily verify (B.2).

Moreover,

$$\begin{aligned}
\ln\det(D\Lambda D') &= \sum_{t=1}^{T} \ln\sigma_t^2 + \ln\left(\sigma_1^{-2} + ... + \sigma_T^{-2}\right) = \sum_{t=1}^{T-1} \ln\widetilde{\sigma}_t^2 \\
&= -\sum_{t=1}^{T} \ln\varphi_t - (T-1)\ln\left(\sigma_1^{-2} + ... + \sigma_T^{-2}\right) \\
&= -\ln\det\Phi + (T-1)\ln\omega_T
\end{aligned} \tag{B.4}$$

Note that these equivalences also imply

$$\ln\det\Lambda = \ln\det(D\Lambda D') + \ln\omega_T. \tag{B.5}$$

Heteroskedastic Orthogonal Deviations The following equivalence also holds

$$v'D'(D\Lambda D')^{-1} Dv = \sum_{t=1}^{T-1} \frac{\widetilde{v}_t^2}{\widetilde{\sigma}_t^2} \tag{B.6}$$

where the heteroskedastic orthogonal deviations are given by

$$\widetilde{v}_t = \begin{cases} v_{T-1} - v_T & \text{for } t = T-1 \\ v_t - \frac{\sigma_{t+1}^{-2} v_{t+1} + ... + \sigma_T^{-2} v_T}{\sigma_{t+1}^{-2} + ... + \sigma_T^{-2}} & \text{for } t = T-2, ..., 1 \end{cases} \tag{B.7}$$

$$\widetilde{\sigma}_t^2 = \begin{cases} \sigma_{T-1}^2 + \sigma_T^2 & \text{for } t = T-1 \\ \sigma_t^2 + \frac{1}{\sigma_{t+1}^{-2} + ... + \sigma_T^{-2}} & \text{for } t = T-2, ..., 1 \end{cases}. \tag{B.8}$$

or

$$\widetilde{v}_t = \begin{cases} v_{T-1} - v_T & \text{for } t = T-1 \\ (v_t - v_{t+1}) + \lambda_{t+1}\widetilde{v}_{t+1} & \text{for } t = T-2, ..., 1 \end{cases} \tag{B.9}$$

where $\lambda_t = \sigma_t^2/\widetilde{\sigma}_t^2$, $(t = T-1, ..., 1)$.

To clarify the mapping between $(\sigma_1^2, ..., \sigma_T^2)$ and $\left(\widetilde{\sigma}_1^2, ..., \widetilde{\sigma}_{T-1}^2\right)$ note that

$$E\left[(v_{T-1} - v_T)(v_{T-2} - v_T)\right] = \sigma_T^2$$
$$E\left(\widetilde{v}_t\right) = \widetilde{\sigma}_t^2 \quad (t = T-1, ..., 1).$$

So we identify σ_T^2 as a covariance between $(v_{T-1} - v_T)$ and $(v_{T-2} - v_T)$, and $\widetilde{\sigma}_{T-1}^2$ as the variance of $\widetilde{v}_{T-1} = (v_{T-1} - v_T)$, so that $\sigma_{T-1}^2 = \widetilde{\sigma}_{T-1}^2 - \sigma_T^2$. We can get

$$\lambda_{T-1} = \frac{\sigma_{T-1}^2}{\widetilde{\sigma}_{T-1}^2} = \frac{\sigma_{T-1}^2}{\sigma_{T-1}^2 + \sigma_T^2}$$

and use it to form

$$\widetilde{v}_{T-2} = (v_{T-2} - v_{T-1}) + \lambda_{T-1}\widetilde{v}_{T-1},$$

which allows us to get $\widetilde{\sigma}_{T-2}^2$. Now we can get $\sigma_{T-2}^2 = \widetilde{\sigma}_{T-2}^2 - 1/\left(\sigma_{T-1}^{-2} + \sigma_T^{-2}\right)$, $\lambda_{T-2} = \sigma_{T-2}^2/\widetilde{\sigma}_{T-2}^2$, and proceed recursively to obtain the remaining terms. Note that the $\widetilde{\sigma}_t^2$ will be nonnegative by construction, so that the non-negativity problem is confined to σ_T^2.

Idempotent Matrices Concerning relevant idempotent matrices, letting $Q = \Phi - \Phi\iota\iota'\Phi$, note that the matrix $Q^\dagger = I - \Phi^{1/2}\iota\iota'\Phi^{1/2}$ is idempotent, and that $Q = \Phi^{1/2}Q^\dagger\Phi^{1/2}$. Also

$$Q^\dagger = \Lambda^{1/2}D'\left(D\Lambda D'\right)^{-1}D\Lambda^{1/2} = I - \omega_T\Lambda^{-1/2}\iota\iota'\Lambda^{-1/2}$$

and $D'\left(D\Lambda D'\right)^{-1}D = \Lambda^{-1/2}Q^\dagger\Lambda^{-1/2}$. So that

$$D'\left(D\Lambda D'\right)^{-1}D = \Lambda^{-1} - \omega_T\Lambda^{-1}\iota\iota'\Lambda^{-1} = \frac{1}{\omega_T}Q. \tag{B.10}$$

Derivatives Letting $\varphi = (\varphi_1, ..., \varphi_T)' = \Phi\iota$, we have the following result:

$$\frac{\partial\varphi}{\partial\theta'} = -\left(\Phi - \Phi\iota\iota'\Phi\right)\Lambda^{-1} = -D'\left(D\Lambda D'\right)^{-1}D\Phi. \tag{B.11}$$

To see this recall that $\varphi_s = \omega_T/\sigma_T^2$ and consider

$$d\varphi = \omega_T\frac{\partial}{\partial\theta'}\begin{pmatrix} 1/\sigma_1^2 \\ \vdots \\ 1/\sigma_T^2 \end{pmatrix}d\theta + \begin{pmatrix} 1/\sigma_1^2 \\ \vdots \\ 1/\sigma_T^2 \end{pmatrix}\frac{\partial\omega_T}{\partial\theta'}d\theta.$$

Also using

$$\frac{\partial \omega_T}{\partial \sigma_s^2} = \frac{1/\sigma_s^4}{\left(\sigma_1^{-2} + ... + \sigma_T^{-2}\right)^2} = \varphi_s^2, \tag{B.12}$$

we get

$$\begin{aligned}
\frac{\partial \varphi}{\partial \theta'} &= -\omega_T \begin{pmatrix} 1/\sigma_1^4 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1/\sigma_T^4 \end{pmatrix} + \begin{pmatrix} 1/\sigma_1^2 \\ \vdots \\ 1/\sigma_T^2 \end{pmatrix} \begin{pmatrix} \varphi_1^2 & \dots & \varphi_T^2 \end{pmatrix} \\
&= -\frac{1}{\omega_T} \begin{pmatrix} \varphi_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \varphi_T^2 \end{pmatrix} + \frac{1}{\omega_T} \begin{pmatrix} \varphi_1 \\ \vdots \\ \varphi_T \end{pmatrix} \begin{pmatrix} \varphi_1^2 & \dots & \varphi_T^2 \end{pmatrix} \\
&= -\frac{1}{\omega_T} \Phi\Phi - \frac{1}{\omega_T} \begin{pmatrix} \varphi_1 \\ \vdots \\ \varphi_T \end{pmatrix} \begin{pmatrix} \varphi_1 & \dots & \varphi_T \end{pmatrix} \Phi = -\frac{1}{\omega_T} \left(\Phi - \Phi \iota \iota' \Phi\right) \Phi.
\end{aligned}$$

B.1.2. Joint MLE of Common Parameters and Effects

Under Assumption $G1$ the joint log likelihood of common parameters and effects is

$$L\left(\alpha, \sigma_1^2, ..., \sigma_T^2, \eta_1, ..., \eta_N\right) = -\frac{N}{2} \sum_{t=1}^{T} \ln \sigma_t^2 - \frac{1}{2} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\frac{v_{it}^2}{\sigma_t^2}\right)$$

The MLE of η_i for given α and $(\sigma_1^2, ..., \sigma_T^2)$ is:

$$\widehat{\eta}_i = \sum_{t=1}^{T} \varphi_t \left(y_{it} - \alpha y_{i(t-1)}\right).$$

Thus the likelihood concentrated with respect to $\eta_1, ..., \eta_N$ is

$$L^*\left(\alpha, \sigma_1^2, ..., \sigma_T^2\right) = -\frac{N}{2} \sum_{t=1}^{T} \ln \sigma_t^2 - \frac{1}{2} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\frac{v_{it} - \overline{v}_t}{\sigma_t}\right)^2.$$

The MLE of α for given $(\sigma_1^2, ..., \sigma_T^2)$ is:

$$\widehat{\alpha} = \left[\sum_{i=1}^{N} \sum_{t=1}^{T} \varphi_t \left(y_{i(t-1)} - \overline{y}_{(-1)}\right)\right]^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \varphi_t \left(y_{i(t-1)} - \overline{y}_{i(-1)}\right) \left(y_{it} - \overline{y}_i\right)$$

The MLE of $(\sigma_1^2, ..., \sigma_T^2)$ for given α and $(\eta_1, ..., \eta_N)$ is[7]

$$\widehat{\sigma}_t^2 = \frac{1}{N} \sum_{i=1}^{N} \left(y_{it} - \alpha y_{i(t-1)} - \eta_i \right)^2 .$$

Therefore, $\widehat{\sigma}_t^2$ solves

$$\widehat{\sigma}_t^2 = \frac{1}{N} \sum_{i=1}^{N} \left[(y_{it} - \overline{y}_i) - \widehat{\alpha} \left(y_{i(t-1)} - \overline{y}_{i(-1)} \right) \right]^2 \quad (t = 1, ..., T)$$

where $\overline{y}_i = \sum_{s=1}^{T} \widehat{\varphi}_s y_{is}$, $\widehat{\varphi}_t = \widehat{\sigma}_t^{-2} / \sum_{s=1}^{T} \widehat{\sigma}_s^{-2}$, and $\widehat{\alpha}$ is as above but with weights evaluated at their ML estimates.

Finally, expressing L^* in terms of φ_s and ω_T

$$L^* \left(\alpha, \sigma_1^2, ..., \sigma_T^2 \right) = \frac{N}{2} \sum_{t=1}^{T} \ln \varphi_t - \frac{NT}{2} \ln \omega_T - \frac{1}{2\omega_T} \sum_{i=1}^{N} \sum_{t=1}^{T} \varphi_t \left(v_{it} - \overline{v}_t \right)^2 ,$$

we can see that the MLE of ω_T for given α and $\left(\varphi_1, ..., \varphi_{T-1} \right)$ is

$$\widehat{\omega}_T = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \varphi_t \left(v_{it} - \overline{v}_t \right)^2 ,$$

which is inconsistent for fixed T because of the absence of degrees of freedom adjustment.

B.1.3. First-Order Conditions for Conditional ML Estimation

We are using

$$A = \begin{pmatrix} a_1' \\ \vdots \\ a_n' \end{pmatrix} \quad vec\,(A) = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$$

$$A \otimes B = \{ a_{jk} B \}$$

[7]Note that

$$\frac{\partial L}{\partial \sigma_t^2} = \frac{N}{2\sigma_t^4} \left(\frac{1}{N} \sum_{i=1}^{N} v_{it}^2 - \sigma_t^2 \right).$$

$$
\begin{aligned}
vec\,(ABC) &= (A \otimes C')\, vec\,(B) \\
tr\,(A'B) &= [vec\,(A)]'\, vec\,(B) \\
d \ln \det (\Omega) &= tr\left(\Omega^{-1} d\Omega\right) \\
d\left(\Omega^{-1}\right) &= -\Omega^{-1}\,(d\Omega)\,\Omega^{-1}.
\end{aligned}
$$

The derivative of L_C with respect ω_T is

$$
\begin{aligned}
\frac{\partial L_C}{\partial \omega_T} &= \frac{N\,(T-1)}{\omega_T^2}\left[\frac{1}{N\,(T-1)}\sum_{i=1}^{N} v_i'\,(\Phi - \Phi\iota\iota'\Phi)\, v_i - \omega_T\right] \\
&= \frac{1}{\omega_T^2}\sum_{i=1}^{N}\left[v_i'\,(\Phi - \Phi\iota\iota'\Phi)\, v_i - (T-1)\,\omega_T\right].
\end{aligned}
$$

The concentrated likelihood with respect to ω_T is

$$
L_C^* = \frac{N}{2}\sum_{t=1}^{T} \ln \varphi_t - \frac{N\,(T-1)}{2} \ln \sum_{i=1}^{N}\sum_{t=1}^{T} \varphi_t\,(v_{it} - \overline{v}_i)^2,
$$

and the Lagrangean

$$
\mathcal{L} = L_C^* + \lambda\left(1 - \sum_{t=1}^{T}\varphi_t\right),
$$

so that

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \varphi_t} &= \frac{N}{2}\frac{1}{\varphi_t} - \frac{1}{2\widehat{\omega}_T}\sum_{i=1}^{N}\left[(v_{it} - \overline{v}_i)^2 - 2 v_{it}\overline{v}_i\left(1 - \sum_{s=1}^{T}\varphi_s\right)\right] - \lambda \\
\frac{\partial \mathcal{L}}{\partial \lambda} &= 1 - \sum_{t=1}^{T}\varphi_t.
\end{aligned}
$$

Inserting the restriction, the first-order conditions for the weights are

$$
\frac{1}{\varphi_t} = \frac{1}{\widehat{\omega}_T}\frac{1}{N}\sum_{i=1}^{N}(v_{it} - \overline{v}_i)^2 + \lambda,
$$

and taking first-differences to eliminate the Lagrange multiplier

$$
\frac{1}{\varphi_t} - \frac{1}{\varphi_{t-1}} = \frac{1}{\widehat{\omega}_T}\frac{1}{N}\sum_{i=1}^{N}(v_{it} - \overline{v}_i)^2 - \frac{1}{\widehat{\omega}_T}\frac{1}{N}\sum_{i=1}^{N}\left(v_{i(t-1)} - \overline{v}_i\right)^2
$$

or

$$\frac{\widehat{\omega}_T}{\varphi_t} - \frac{\widehat{\omega}_T}{\varphi_{t-1}} = \frac{1}{N}\sum_{i=1}^{N}\left[(v_{it} - \overline{v}_i)^2 - \left(v_{i(t-1)} - \overline{v}_i\right)^2\right].$$

Nonnegativity constraints The nonnegativity constraints $\sigma_t^2 > 0$ may be enforced through the parameterization $(\omega_T, \varphi_1, ..., \varphi_T)$ imposing adding-up and non-negativity restrictions to the weights. Alternatively, transformed variances for errors in orthogonal deviations can be used, which confine nonnegativity restrictions to σ_T^2.

B.1.4. Bias corrected conditional ML scores

Under Assumption A

$$E\left[x_i' D' \left(D\Lambda D'\right)^{-1} D v_i\right] = -h_T(\alpha, \varphi)$$

where

$$h_T(\alpha, \varphi) = \sum_{t=1}^{T-1}\left(\frac{1-\alpha^t}{1-\alpha}\right)\varphi_{t+1}.$$

An alternative expression is

$$h_T(\alpha, \varphi) = \frac{1}{(1-\alpha)}\left[1 - \varphi_1 - \sum_{t=2}^{T}\varphi_t \alpha^{t-1}\right]$$

so that

$$h_T(\alpha, \varphi) = \begin{cases} \frac{1}{(1-\alpha)}\left[1 - \sum_{t=1}^{T}\varphi_t \alpha^{t-1}\right] & if\ \alpha \neq 0 \\ (1-\varphi_1) & if\ \alpha = 0. \end{cases}$$

Proof. Using (B.10) we have

$$\begin{aligned} E\left[x_i' D' \left(D\Lambda D'\right)^{-1} D v_i\right] &= E\left(x_i'\Lambda^{-1}v_i\right) - \omega_T E\left(x_i'\Lambda^{-1}\iota\iota'\Lambda^{-1}v_i\right) \\ &= -\omega_T \iota'\Lambda^{-1}E\left(v_i x_i'\right)\Lambda^{-1}\iota \end{aligned}$$

where $E(v_i x_i') = \Lambda C_T$. The (t, s)th element of C_T is $\alpha^{(s-t-1)}$ for $t < s$, and zero otherwise. Thus,

$$E\left[x_i' D' \left(D\Lambda D'\right)^{-1} D v_i\right] = -\omega_T \iota' C_T \Lambda^{-1}\iota = -\iota' C_T \Phi \iota$$

$$
\begin{aligned}
&= -\left(0,1,1+\alpha,1+\alpha+\alpha^2,...,1+\alpha+...+\alpha^{T-2}\right)\begin{pmatrix}\varphi_1\\ \vdots \\ \varphi_T\end{pmatrix}\\
&= -\left[\left(\frac{1-\alpha}{1-\alpha}\right)\varphi_2+\left(\frac{1-\alpha^2}{1-\alpha}\right)\varphi_3+...+\left(\frac{1-\alpha^{T-1}}{1-\alpha}\right)\varphi_T\right]\\
&= -\sum_{t=1}^{T-1}\left(\frac{1-\alpha^t}{1-\alpha}\right)\varphi_{t+1}.
\end{aligned}
$$

■

The variance of the average error can be eliminated to give rise to moment conditions that only depend on α and the weights as follows:

$$
\begin{aligned}
E\left[x_i'\left(\Phi-\Phi\iota\iota'\Phi\right)v_i\right] &= -\omega_T\sum_{t=1}^{T-1}\left(\frac{1-\alpha^t}{1-\alpha}\right)\varphi_{t+1}\\
\omega_T &= E\left[\frac{1}{(T-1)}v_i'\left(\Phi-\Phi\iota\iota'\Phi\right)v_i\right]
\end{aligned}
$$

$$
E\left[x_i'\left(\Phi-\Phi\iota\iota'\Phi\right)v_i-\frac{1}{(T-1)}v_i'\left(\Phi-\Phi\iota\iota'\Phi\right)v_i\sum_{t=1}^{T-1}\left(\frac{1-\alpha^t}{1-\alpha}\right)\varphi_{t+1}\right]=0.
$$

Integral of the Heteroskedastic Bias Function The integral of $h_T(\alpha,\varphi)$ is given by

$$
b_T(\alpha,\varphi)=\sum_{t=1}^{T-1}\frac{\left(\varphi_{t+1}+...+\varphi_T\right)}{t}\alpha^t.
$$

To see this note that using

$$
\begin{aligned}
h_T(\alpha,\varphi) &= \sum_{t=1}^{T-1}\left(1+\alpha+..+\alpha^{t-1}\right)\varphi_{t+1}\\
&= \sum_{t=1}^{T-1}\varphi_{t+1}+\alpha\sum_{t=2}^{T-1}\varphi_{t+1}+\alpha^2\sum_{t=3}^{T-1}\varphi_{t+1}+...+\alpha^{T-2}\varphi_T,
\end{aligned}
$$

we can write

$$
b_T(\alpha,\varphi) = \alpha\sum_{s=1}^{T-1}\varphi_{s+1}+\frac{\alpha^2}{2}\sum_{s=2}^{T-1}\varphi_{s+1}+\frac{\alpha^3}{3}\sum_{s=3}^{T-1}\varphi_{s+1}+..+\frac{\alpha^{T-1}}{T-1}\varphi_T
$$

$$= \sum_{t=1}^{T-1} \frac{\left(\varphi_{t+1} + ... + \varphi_T\right)}{t} \alpha^t.$$

Derivatives of $b_T(\alpha, \varphi)$ with respect to φ_t are:

$$\frac{\partial b_T(\alpha, \varphi)}{\partial \varphi_t} = \begin{cases} 0 & for\ t = 1 \\ \sum_{s=1}^{t-1} \frac{\alpha^s}{s} & for\ t > 1 \end{cases}$$

and in view of (B.11):

$$\frac{\partial b_T(\alpha, \varphi)}{\partial \theta} = \left(\frac{\partial \varphi}{\partial \theta'}\right)' \frac{\partial b_T(\alpha, \varphi)}{\partial \varphi} = -\Phi D' \left(D \Lambda D'\right)^{-1} D \begin{pmatrix} 0 \\ \alpha \\ \alpha + \frac{\alpha^2}{2} \\ \alpha + \frac{\alpha^2}{2} + \frac{\alpha^3}{3} \\ \vdots \end{pmatrix}.$$

B.2. Random Effects Estimation

Proof of Results for the Random Effects Scores We show that under Assumption A:

$$\frac{1}{\sigma_\varepsilon^2} E\left[\overline{x}_i \left(\overline{u}_i - \phi y_{i0}\right)\right] = h_T(\alpha, \varphi)$$

and

$$E\left[\frac{1}{\sigma_\varepsilon^2} \Phi D' \left(D \Lambda D'\right)^{-1} D v_i \left(\overline{u}_i - \phi y_{i0}\right)\right] = 0$$

First, using (2.6) and (3.25), the lagged average $\overline{x}_i$ can be written as

$$\overline{x}_i = h_T(\alpha, \varphi) \eta_i + \left(\varphi_1 + \sum_{t=1}^{T-1} \varphi_{t+1} \alpha^t\right) y_{i0} + \sum_{t=1}^{T-1} \varphi_{t+1} \left(v_{it} + \alpha v_{i(t-1)} + ... + \alpha^{t-1} v_{i1}\right). \tag{B.13}$$

To see this, note that

$$\begin{aligned} \overline{x}_i &= \varphi_1 y_{i0} + \sum_{t=1}^{T-1} \varphi_{t+1} y_{it} \\ &= \varphi_1 y_{i0} + \sum_{t=1}^{T-1} \varphi_{t+1} \left[\left(\frac{1 - \alpha^t}{1 - \alpha}\right) \eta_i + \alpha^t y_{i0} + \left(v_{it} + \alpha v_{i(t-1)} + ... + \alpha^{t-1} v_{i1}\right)\right] \end{aligned}$$

$$= \left[\sum_{t=1}^{T-1} \varphi_{t+1} \left(\frac{1-\alpha^t}{1-\alpha}\right)\right] \eta_i + \left(\varphi_1 + \sum_{t=1}^{T-1} \varphi_{t+1} \alpha^t\right) y_{i0}$$
$$+ \sum_{t=1}^{T-1} \varphi_{t+1} \left(v_{it} + \alpha v_{i(t-1)} + ... + \alpha^{t-1} v_{i1}\right).$$

Moreover, letting $\xi_i = \eta_i - \phi y_{i0}$, so that

$$\sigma_\varepsilon^2 = Var\left(\overline{v}_i\right) + Var\left(\xi_i\right).$$

Using these two expressions we have

$$\begin{aligned}
\frac{1}{\sigma_\varepsilon^2} E\left[\overline{x}_i \left(\overline{u}_i - \phi y_{i0}\right)\right] &= \frac{1}{\sigma_\varepsilon^2} \left\{E\left(\overline{x}_i \overline{v}_i\right) + E\left[\overline{x}_i \left(\eta_i - \phi y_{i0}\right)\right]\right\} \\
&= \frac{1}{\sigma_\varepsilon^2} \left[\omega_T^2 \Lambda^{-1} \iota' E\left(v_i x_i'\right) \Lambda^{-1} \iota + h_T\left(\alpha, \varphi\right) E\left(\eta_i \xi_i\right)\right] \\
&= \frac{1}{\sigma_\varepsilon^2} \left[\omega_T h_T\left(\alpha, \varphi\right) + h_T\left(\alpha, \varphi\right) Cov\left(\eta_i, \xi_i\right)\right] \\
&= h_T\left(\alpha, \varphi\right) \frac{1}{\sigma_\varepsilon^2} \left[Var\left(\overline{v}_i\right) + Var\left(\xi_i\right)\right] = h_T\left(\alpha, \varphi\right).
\end{aligned}$$

This proves the first result. Turning to the second result, we have

$$E\left[\frac{1}{\sigma_\varepsilon^2} \Phi D' \left(D \Lambda D'\right)^{-1} D v_i \left(\overline{u}_i - \phi y_{i0}\right)\right] =$$

$$\begin{aligned}
&= \frac{1}{\sigma_\varepsilon^2} \Phi D' \left(D \Lambda D'\right)^{-1} E\left(D v_i \overline{v}_i\right) \\
&= \frac{1}{\sigma_\varepsilon^2} \Phi D' \left(D \Lambda D'\right)^{-1} D E\left(v_i v_i'\right) \Phi \iota \\
&= \frac{1}{\sigma_\varepsilon^2} \Phi D' \left(D \Lambda D'\right)^{-1} D \Lambda \Phi \iota \\
&= \frac{\omega_T}{\sigma_\varepsilon^2} \Phi D' \left(D \Lambda D'\right)^{-1} D \Lambda \Lambda^{-1} \iota = \frac{\omega_T}{\sigma_\varepsilon^2} \Phi D' \left(D \Lambda D'\right)^{-1} D \iota = 0.
\end{aligned}$$

B.3. Estimation from the Data in Differences

Average of deviations from y_{i0} in terms of first-differences We use the notation $x_i^\dagger = \left(y_{i0}^\dagger, ..., y_{i(T-1)}^\dagger\right)'$, so that $y_i^\dagger = y_i - y_{i0}\iota$ and $x_i^\dagger = x_i - y_{i0}\iota$. Similarly, $\overline{y}_i^\dagger = y_i^{\dagger\prime}\Phi\iota = \overline{y}_i - y_{i0}$, etc. The following expression makes explicit that $\overline{y}_i^\dagger$ only depends on the data in differences:

$$\begin{aligned}
\overline{y}_i^\dagger &= \iota'\Phi y_i^\dagger = \iota'\Phi D^{\dagger-1}\left(\Delta y_{i1}, ..., \Delta y_{iT}\right)' \\
&= \left(\varphi_1, ..., \varphi_T\right)\begin{pmatrix} \Delta y_{i1} \\ \Delta y_{i1} + \Delta y_{i2} \\ \vdots \\ \Delta y_{i1} + ... + \Delta y_{iT} \end{pmatrix} = \sum_{t=1}^{T}\left(\sum_{s=t}^{T}\varphi_s\right)\Delta y_{it}.
\end{aligned}$$

Under homoskedasticity $\varphi_t = 1/T$ and the previous expression reduces to

$$\overline{y}_i^\dagger = \Delta y_{i1} + \frac{T-1}{T}\Delta y_{i2} + ... + \frac{1}{T}\Delta y_{iT} = \sum_{t=1}^{T}\left(\frac{T-t+1}{T}\right)\Delta y_{it}.$$

Alternative Expressions for densities and criteria for first-differences We have the following "random effects" marginal log density for the data in first differences

$$\begin{aligned}
\ln f\left(\Delta y_{i1}, ..., \Delta y_{iT}\right) &= -\frac{1}{2}\ln\det\left(D\Lambda D'\right) - \frac{1}{2}v_i'D'\left(D\Lambda D'\right)^{-1}Dv_i \\
&\quad -\frac{1}{2}\ln\sigma_\varepsilon^{2\dagger} - \frac{1}{2\sigma_\varepsilon^{2\dagger}}\left(\overline{y}_i^\dagger - \overline{x}_i^\dagger\alpha\right)^2
\end{aligned}$$

or

$$\begin{aligned}
\ln f\left(\Delta y_{i1}, ..., \Delta y_{iT}\right) &= \frac{1}{2}\ln\det\Phi - \frac{(T-1)}{2}\ln\omega_T - \frac{1}{2\omega_T}v_i'D'\left(D\Phi^{-1}D'\right)^{-1}Dv_i \\
&\quad -\frac{1}{2}\ln\sigma_\varepsilon^{2\dagger} - \frac{1}{2\sigma_\varepsilon^{2\dagger}}\left[\sum_{t=1}^{T}\left(\sum_{s=t}^{T}\varphi_s\right)\Delta v_{it}\right]^2.
\end{aligned}$$

Therefore, the random effects log likelihood for the data in first-differences is

$$\begin{aligned}
L_{RD} &= \frac{N}{2}\sum_{t=1}^{T}\ln\varphi_t - \frac{(T-1)N}{2}\ln\omega_T - \frac{1}{2\omega_T}\sum_{i=1}^{N}v_i'D'\left(D\Phi^{-1}D'\right)^{-1}Dv_i \\
&\quad -\frac{N}{2}\ln\sigma_\varepsilon^{2\dagger} - \frac{1}{2\sigma_\varepsilon^{2\dagger}}\sum_{i=1}^{N}\left(\overline{y}_i^\dagger - \overline{x}_i^\dagger\alpha\right)^2.
\end{aligned}$$

Concentrating L_{RD} with respect to ω_T and $\sigma_\varepsilon^{2\dagger}$ we can define the

$$\begin{aligned}(\widehat{\alpha}_{DH}, \widehat{\varphi}_{DH}) &= \arg\min \left\{ \ln \left(\sum_{i=1}^{N} v_i' D' \left(D \Phi^{-1} D' \right)^{-1} D v_i \right) - \frac{1}{(T-1)} \sum_{t=1}^{T} \ln \varphi_t \right. \\ & \left. + \frac{1}{(T-1)} \ln \left(\sum_{i=1}^{N} \overline{u}_i^{\dagger 2} \right) \right\}\end{aligned}$$

subject to $0 < \varphi_t < 1$ and $\sum_{t=1}^{T} \varphi_t = 1$.

Finally, the corresponding homoskedastic estimator and log likelihood are

$$\widehat{\alpha}_D = \arg\min \left\{ \ln \left(\sum_{i=1}^{N} v_i' D' \left(D D' \right)^{-1} D v_i \right) + \frac{1}{(T-1)} \ln \sum_{i=1}^{N} \left[\overline{u}_i - (1-\alpha) y_{i0} \right]^2 \right\}. \tag{B.14}$$

and

$$\begin{aligned}\ln f(\Delta y_{i1}, ..., \Delta y_{iT}) &\propto -\frac{(T-1)}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} v_i^{*\prime} v_i^* \\ & -\frac{1}{2} \ln \sigma_\varepsilon^{2\dagger} - \frac{1}{2\sigma_\varepsilon^{2\dagger}} \left[\overline{y}_i - \alpha \overline{x}_i - (1-\alpha) y_{i0} \right]^2.\end{aligned}$$

The first-order conditions for α in the homoskedastic case are

$$\frac{\partial \ln f(\Delta y_{i1}, ..., \Delta y_{iT})}{\partial \alpha} = \frac{1}{\sigma^2} x_i^{*\prime} v_i^* + \frac{1}{\sigma_\varepsilon^{2\dagger}} \left[\overline{u}_i - (1-\alpha) y_{i0} \right] (\overline{x}_i - y_{i0}).$$

Figure 1
Relative Inefficiency Ratios (lamda=0)

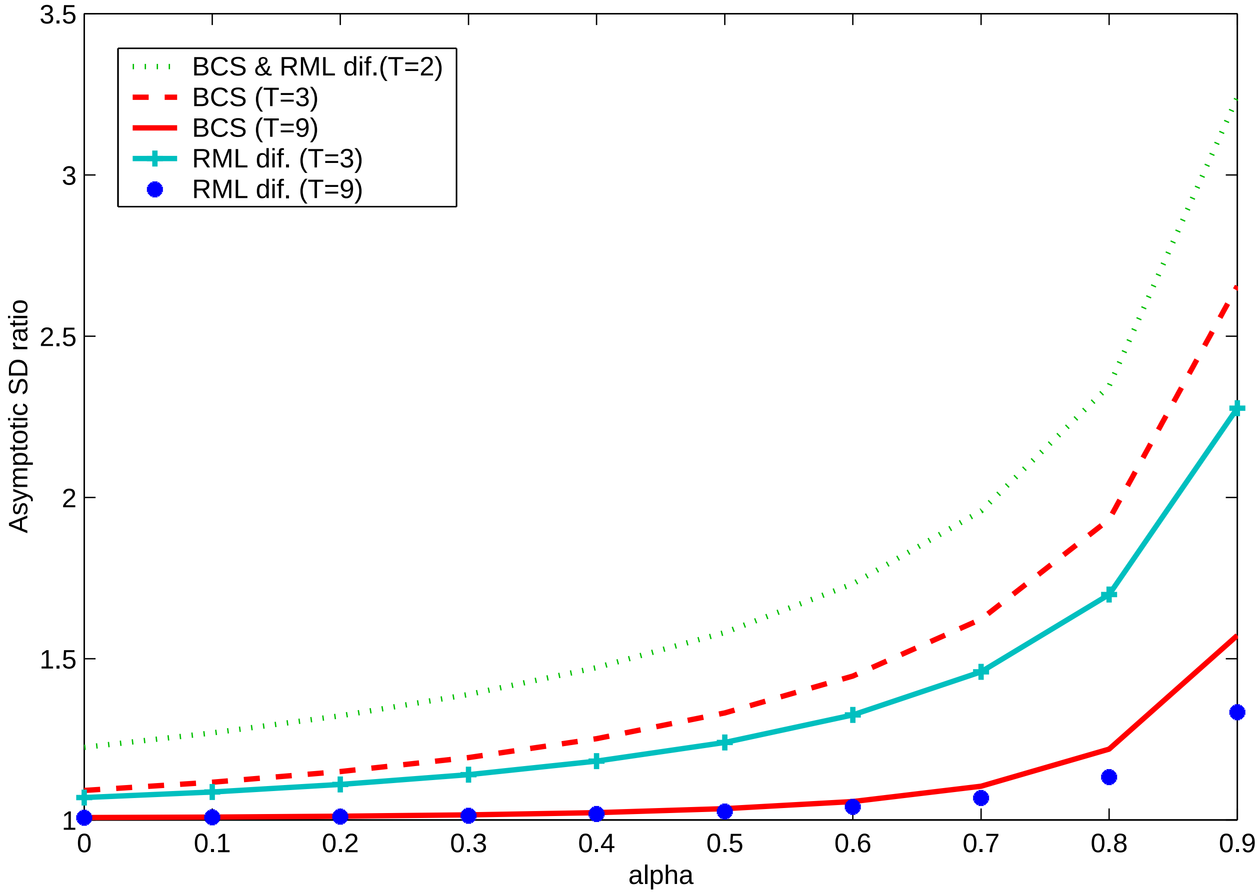

Figure 2

Relative Inefficiency Ratios (lamda=1)

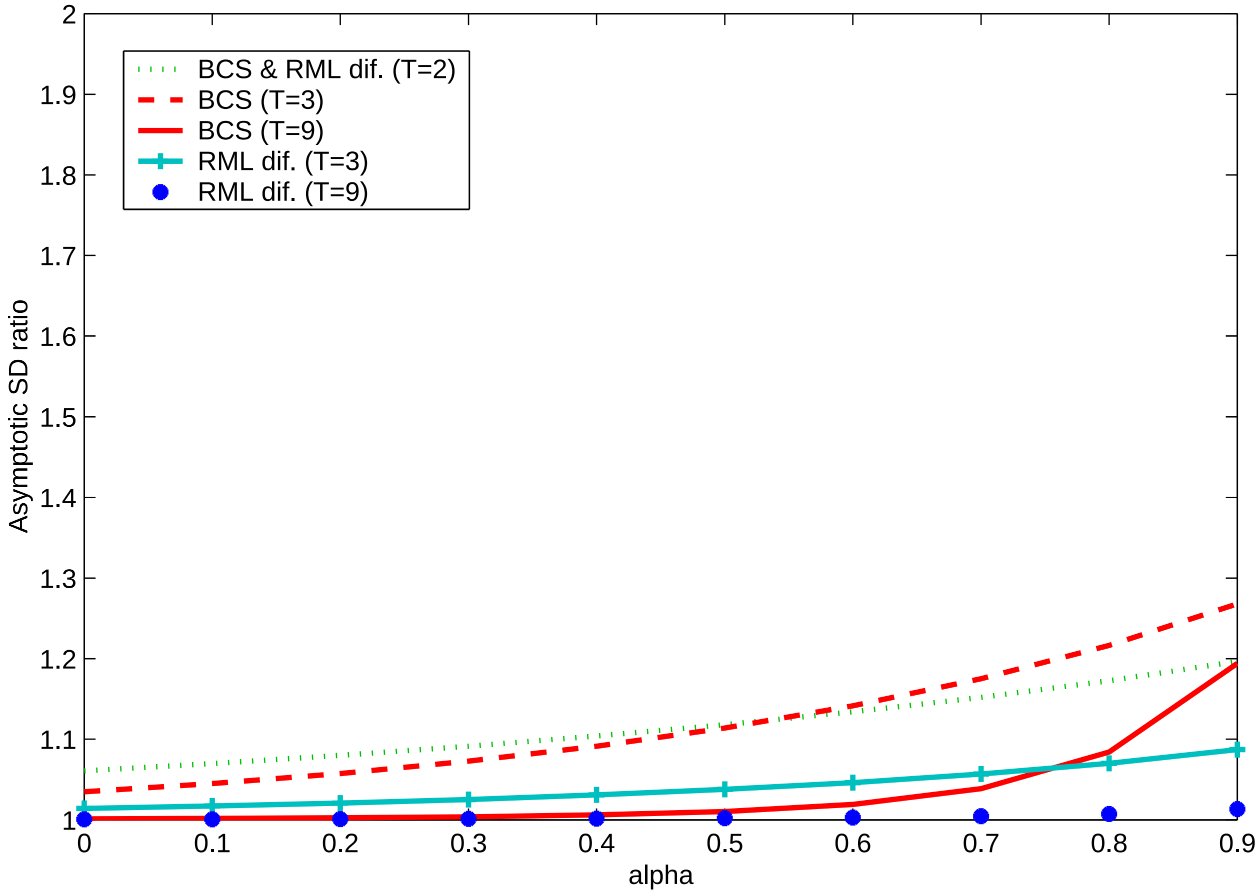

Figure 3
Relative Inefficiency Under Nonstationary Initial Variance
(T=3, alpha=0.9, lamda=0)