Robust Likelihood Estimation of Dynamic Panel Data Models[*]

Javier Alvarez[†]
Banco de España

Manuel Arellano[‡]
CEMFI, Madrid

This version: March 2021

[*] This article is dedicated to the memory of Gary Chamberlain. An earlier version appeared as CEMFI Working Paper No. 0421 in 2004. We are grateful to Olympia Bover, Gary Chamberlain, Tony Lancaster, Thierry Magnac, Whitney Newey, Enrique Sentana, and seminar audiences at Alicante, CEMFI, Princeton, University College London, and Yale for helpful comments; and to Laura Hospido and Gema Zamarro for very able research assistance. We extend our special thanks to Bryan Graham and Kei Hirano for their encouragement in revising the paper, and to an anonymous referee for helpful comments.

[†] Statistics Department, Banco de España, Alcalá 522, 28043 Madrid, Spain. *E-mail:* javier.alvarez@bde.es

[‡] Corresponding author. CEMFI, Casado del Alisal 5, 28014 Madrid, Spain. *E-mail:* arellano@cemfi.es

Abstract

We develop likelihood-based estimators for autoregressive panel data models that are consistent in the presence of time series heteroskedasticity. Bias-corrected conditional score estimators, random effects maximum likelihood in levels and first differences, and estimators that impose mean stationarity are considered for general autoregressive models with individual effects. We investigate identification under unit roots, and show that random effects estimation in levels may achieve substantial efficiency gains relative to estimation from data in differences. In an empirical application, we find evidence against unit roots in individual earnings processes from the Panel Study of Income Dynamics and the Spanish section of the European Community Household Panel.

Keywords: Autoregressive panel data models; time series heteroskedasticity; bias-corrected score; random effects; earnings process.

JEL: C23

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. Our methods are robust in the sense used in Gourieroux, Monfort, and Trognon (1984) of providing consistent estimates of the conditional mean parameters when the chosen likelihood function does not necessarily contain the true distribution.

The paper is organized as follows. Section 2 presents the model and a discussion of the assumptions. Section 3 explains how to obtain fixed-T consistent estimates of AR(p) coefficients from bias-corrected first-order conditions of a heteroskedastic within-group likelihood (bias-corrected score (BCS) estimation).

Section 4 presents ML estimates from a likelihood averaged with respect to normally distributed effects and initial observations (random effects maximum likelihood (RML) estimation). We show that such an averaging leads to a modified within-group criterion that balances off the within and between biases. The modification term, which depends on the data in levels, may lead to substantial efficiency gains relative to estimators from differenced data alone, and is crucial for identification in very short panels. Heteroskedastic RML is our recommended likelihood-based method. It is computationally straightforward and can be easily extended to unbalanced and multivariate panels. Moreover, work by Chamberlain and Moreira (2009) and Bai (2013) has established additional desirable properties of this type of estimator from a fixed-effects perspective in finite samples and in large T-and-N samples, respectively. We shall return to a discussion of these points in the concluding section.

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

Section 5 presents RML estimates from data in differences, and Section 6 discusses conditional and marginal ML estimation under stationarity in mean. Interestingly, we show that the random effects likelihood for the differenced data coincides with the likelihood conditioned on the estimated effects under mean stationarity, so that this restriction is immaterial to the data in differences when homoskedasticity is not imposed.

Section 7 discusses the possibility of identification failure for a first-order process with an unit root, in view that in a three-wave panel a random walk without heterogeneous drift is known to be underidentified. We show that in a four-wave panel there is local identification but not global identification under heteroskedasticity, and global identification but first-order underidentification under homoskedasticity. In panels with more than four waves, we find that the autoregressive coefficient is globally identified unless the error variances change with a constant rate of growth.

Section 8 reports numerical calculations of the asymptotic variances of BCS and RML estimators in differences relative to RML in levels, calculated under the assumption of normality. In Section 9 we present estimates of first- and second-order autoregressive equations for individual labour income using data from the Panel Study of Income Dynamics (PSID) and the Spanish section of the European Community Household Panel, and find evidence against unit roots in earnings. The PSID result differs greatly from the income processes that impose a unit root, often employed in the empirical literatures on consumption and labour supply (e.g. Hall and Mishkin, 1982; Abowd and Card, 1989, or Meghir and Pistaferri, 2004). However, it is not inconsistent with the findings of later studies that have explored more general models by allowing for richer forms of heterogeneity (Browning, Ejrnæs and Alvarez, 2010) or nonlinear dynamics (Arellano, Blundell and Bonhomme, 2017). Our result is unaffected by adding moving average components to the specification of the earnings process.

Finally, Section 10 contains further comments on the properties of our estimators taking into account results from the literature, and provides a summary of the major conclusions of the paper. Proofs and technical material are in the Appendix.

2 Model and Assumptions

We consider an autoregressive model for panel data given by

$$y_{it} = \alpha_1 y_{i(t-1)} + ... + \alpha_p y_{i(t-p)} + \eta_i + v_{it} \quad (t = 1, ..., T; i = 1, ..., N). \tag{1}$$

The variables $\left(y_{i(1-p)}, ..., y_{i0}, ..., y_{iT}\right)$ are observed but η_i is an unobservable individual effect. The $p \times 1$ vector of initial observations is denoted as $y_i^0 = \left(y_{i(1-p)}, ..., y_{i0}\right)'$.[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 $x_{it} = \left(y_{i(t-1)}, ..., y_{i(t-p)}\right)'$, $\alpha = (\alpha_1, ..., \alpha_p)'$, and write the model in the form:

$$y_i = X_i \alpha + \eta_i \iota + v_i \tag{2}$$

[2] We assume that y_i^0 is observed for notational convenience, so that the actual number of waves in the data is $T^o = T+p$.

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 : $\left\{\eta_i, y_i^0, y_{i1}, ..., y_{iT}\right\}_{i=1}^N$ is a random sample from a well defined joint distribution with finite fourth-order moments that satisfies

$$E\left(v_{it} \mid \eta_i, y_i^0, y_{i1}, ..., y_{i(t-1)}\right) = 0 \quad (t = 1, ..., T)\,. \tag{3}$$

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 nor conditional homoskedasticity are assumed.[3] That is, the unconditional variances of the errors, denoted as

$$E\left(v_{it}^2\right) = \sigma_t^2, \tag{4}$$

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

$$E\left(v_{it}^2 \mid \eta_i, y_i^0, y_{i1}, ..., y_{i(t-1)}\right).$$

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. See Arellano (2003, Section 6.4.3).

Also note that under stability of the process,[4] we do not assume stationarity in mean. Let the covariance matrix of $\left(\eta_i, y_i^0\right)$ be denoted as

$$Var\begin{pmatrix} \eta_i \\ y_i^0 \end{pmatrix} = \begin{pmatrix} \sigma_\eta^2 & \gamma_{\eta 0} \\ \gamma_{0\eta} & \Gamma_{00} \end{pmatrix}. \tag{5}$$

For example, when $p = 1$ (so that $\alpha = \alpha_1$, $y_i^0 = y_{i0}$, and $\Gamma_{00} = \gamma_{00}$) model (1) can be written as

$$y_{it} = \left(1 + \alpha + ... + \alpha^{t-1}\right)\eta_i + \alpha^t y_{i0} + \left(v_{it} + \alpha v_{i(t-1)} + ... + \alpha^{t-1} v_{i1}\right). \tag{6}$$

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

$$y_{i0} = \frac{\eta_i}{(1-\alpha)} + \sum_{j=0}^{\infty} \alpha^j v_{i(-j)}, \tag{7}$$

implying $\gamma_{\eta 0} = \sigma_\eta^2 / (1-\alpha)$ and $\gamma_{00} = \sigma_\eta^2 / (1-\alpha)^2 + \sum_{j=0}^{\infty} \alpha^{2j} \sigma_{-j}^2$.[5] 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

[3] Time series and conditional homoskedasticity assumptions are discussed in Arellano (2003, p. 82–83).

[4] That is, when the roots of the equation $z^p - \alpha_1 z^{p-1} - ... - \alpha_p = 0$ are inside the unit circle.

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

in first differences will generally depend on individual effects. Estimation under stationarity in mean is discussed in Section 6.

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, 2003a).

In the next two sections, we introduce the likelihood of the data given initial conditions and the two types of likelihood functions on which our estimates are based. Namely, these are a likelihood conditioned on the MLE of the effects and a ("random effects") likelihood in which the effects are averaged out using normal probability weights with a variance that is also estimated. Later, we examine the role of mean stationarity restrictions. Altogether, we consider four different estimators, which are displayed in Table A1 according to whether they use the data in levels or in differences, and whether they impose mean stationarity or not. Random effects ML in levels will emerge as the natural estimation approach unless the effects have a very large variance that cannot be estimated. As for mean stationarity restrictions, their justification is often weak, and in our applications they turn out not to be essential for precision nor are they supported by the data.

3 Bias-Corrected Conditional Score Estimation

3.1 Normal Likelihood Given Initial Observations and Effects

Under the normality assumption

$$y_{it} \mid y_i^0, ..., y_{i(t-1)}, \eta_i \sim \mathcal{N}\left(\alpha_1 y_{i(t-1)} + ... + \alpha_p y_{i(t-p)} + \eta_i, \sigma_t^2\right) \quad (t = 1, ..., T), \qquad (\textit{Assumption G}1)$$

the log density of y_i conditioned on $\left(y_i^0, \eta_i\right)$ is given by

$$\ln f\left(y_i \mid y_i^0, \eta_i\right) = -\frac{1}{2} \ln \det \Lambda - \frac{1}{2} v_i' \Lambda^{-1} v_i \tag{8}$$

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 (8) is

$$\widehat{\eta}_i = \overline{y}_i - \overline{x}_i' \alpha \tag{9}$$

where $\overline{y}_i$ and $\overline{x}_i$ denote weighted averages of the form $\overline{y}_i = \sum_{t=1}^T \varphi_t y_{it}$ with weights

$$\varphi_t = \frac{\sigma_t^{-2}}{\sigma_1^{-2} + ... + \sigma_T^{-2}}. \tag{10}$$

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

$$L^* = \frac{N}{2} \ln \det \Phi - \frac{NT}{2} \ln \omega_T - \frac{1}{2\omega_T} \sum_{i=1}^N v_i' \left(\Phi - \Phi \iota \iota' \Phi\right) v_i \tag{11}$$

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

$$\omega_T = Var\left(\overline{v}_i\right) = \frac{1}{\sigma_1^{-2} + ... + \sigma_T^{-2}}. \tag{12}$$

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

$$\begin{aligned} v_i' D' \left(D\Lambda D'\right)^{-1} D v_i &= \frac{1}{\omega_T} v_i' \left(\Phi - \Phi\iota\iota'\Phi\right) v_i = \sum_{t=1}^{T} \frac{\left(v_{it} - \overline{v}_i\right)^2}{\sigma_t^2} & (13) \\ \ln\det\left(D\Lambda D'\right) &= -\ln\det\Phi + (T-1)\ln\omega_T & (14) \end{aligned}$$

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.[6] Note that with $T = 3$ (i.e. $(3+p)$ 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

$$\sigma_t^2 = E\left[\left(v_{it} - \overline{v}_i\right)^2\right] + \omega_T \quad (t = 1, ...T). \tag{15}$$

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

$$\widehat{\alpha} = \left[\sum_{i=1}^{N}\sum_{t=1}^{T} \varphi_t \left(x_{it} - \overline{x}_i\right)\left(x_{it} - \overline{x}_i\right)'\right]^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T} \varphi_t \left(x_{it} - \overline{x}_i\right)\left(y_{it} - \overline{y}_i\right), \tag{16}$$

which in first differences can also be written as

$$\widehat{\alpha} = \left[\sum_{i=1}^{N} X_i' D' \left(D\Lambda D'\right)^{-1} D X_i\right]^{-1} \sum_{i=1}^{N} X_i' D' \left(D\Lambda D'\right)^{-1} D y_i. \tag{17}$$

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

$$\omega_T = E\left[\frac{1}{(T-1)} \sum_{t=1}^{T} \varphi_t \left(v_{it} - \overline{v}_i\right)^2\right]. \tag{18}$$

[6] According to (13), the weighted sum of squared errors in deviations is also a weighted sum of cross-products of the errors in first differences $\Delta v_{it} = \Delta y_{it} - \Delta x_{it}'\alpha$ contained in the vector Dv_i.

3.2 Likelihood Conditioned on the ML Estimates of the Effects

Provided $G1$ holds, the ML estimates of the effects $\widehat{\eta}_i = \eta_i + \overline{v}_i$ at the true values of the common parameters $\left(\alpha, \sigma_1^2, ..., \sigma_T^2\right)$ satisfy

$$\widehat{\eta}_i \mid y_i^0, \eta_i \sim \mathcal{N}\left(\eta_i, \omega_T\right). \tag{19}$$

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

$$\ln f\left(y_i \mid y_i^0, \eta_i, \widehat{\eta}_i\right) = -\frac{1}{2}\ln\det\left(D\Lambda D'\right) - \frac{1}{2}v_i'D'\left(D\Lambda D'\right)^{-1}Dv_i, \tag{20}$$

which is a within-group density that does not depend on η_i. This result follows from subtracting (8) and the normal density of $\widehat{\eta}_i \mid y_i^0, \eta_i$ while using the identities (13)-(14). Thus, (8) admits the decomposition

$$f\left(y_i \mid y_i^0, \eta_i\right) = f\left(y_i \mid y_i^0, \widehat{\eta}_i\right) f\left(\widehat{\eta}_i \mid y_i^0, \eta_i\right), \tag{21}$$

which confines the dependence on η_i to the conditional density of $\widehat{\eta}_i$. Similarly, any marginal density for $y_i \mid y_i^0$, which uses a prior distribution on the effects, can be written as

$$f\left(y_i \mid y_i^0\right) = f\left(y_i \mid y_i^0, \widehat{\eta}_i\right) f\left(\widehat{\eta}_i \mid y_i^0\right). \tag{22}$$

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

$$L_C = \frac{N}{2}\ln\det\Phi - \frac{N(T-1)}{2}\ln\omega_T - \frac{1}{2\omega_T}\sum_{i=1}^{N} v_i'\left(\Phi - \Phi\iota\iota'\Phi\right)v_i \tag{23}$$

or

$$L_C = -\frac{N}{2}\ln\det\left(D\Lambda D'\right) - \frac{1}{2}\sum_{i=1}^{N} v_i'D'\left(D\Lambda D'\right)^{-1}Dv_i, \tag{24}$$

which is similar to the concentrated likelihood (11) except that it incorporates a correction for degrees of freedom. In a model with strictly exogenous x_{it}, 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 (16) 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 the within-group estimator in the autoregressive model results from lack of orthogonality between the scores of α and the effects.

In the homoskedastic case with $p = 1$, 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.

3.2.1 First-Order Conditions

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

$$\frac{\partial L_C}{\partial \alpha} = \sum_{i=1}^{N} X_i' D' \left(D \Lambda D'\right)^{-1} D v_i. \tag{25}$$

$$\frac{\partial L_C}{\partial \theta} = \frac{1}{2} \sum_{i=1}^{N} K' \left(D \Lambda D' \otimes D \Lambda D'\right)^{-1} vec\left(D v_i v_i' D' - D \Lambda D'\right) \tag{26}$$

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

Thus, the conditional MLE of α and θ solve, respectively, (17) and

$$\widehat{\theta} = \left(K' \Upsilon^{-1} K\right)^{-1} K' \Upsilon^{-1} \frac{1}{N} \sum_{i=1}^{N} vec\left(D v_i v_i' D'\right). \tag{27}$$

where $\Upsilon = D \Lambda D' \otimes D \Lambda D'$.[7]

3.2.2 Bias-Corrected Conditional Scores

Under Assumption A the expected conditional ML scores are given by

$$E\left[X_i' D' \left(D \Lambda D'\right)^{-1} D v_i\right] = -h_T\left(\alpha, \varphi\right) \tag{28}$$

$$E\left[K' \left(D \Lambda D' \otimes D \Lambda D'\right)^{-1} vec\left(D v_i v_i' D' - D \Lambda D'\right)\right] = 0 \tag{29}$$

where

$$h_T\left(\alpha, \varphi\right) = \begin{pmatrix} \varphi' C_1 \iota \\ \vdots \\ \varphi' C_p \iota \end{pmatrix} \tag{30}$$

[7] Maximizing L_C with respect to ω_T and $\left(\varphi_1 ... \varphi_T\right)$ 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 (15) and (18) as shown in Appendix A.1.

with

$$C_j = \begin{pmatrix} 0 & 0 \\ B_{T-j}^{-1} & 0 \end{pmatrix} \tag{31}$$

and B_{T-j} is a $(T-j) \times (T-j)$ matrix such that

$$B_{T-j} = \begin{pmatrix} 1 & 0 & \dots & 0 & \dots & 0 & 0 \\ -\alpha_1 & 1 & & 0 & \dots & 0 & 0 \\ -\alpha_2 & -\alpha_1 & \ddots & 0 & \dots & 0 & 0 \\ & \ddots & \ddots & \ddots & & \vdots & \vdots \\ 0 & & & 0 & & 1 & 0 \\ 0 & 0 & \dots & -\alpha_p & \dots & -\alpha_1 & 1 \end{pmatrix}. \tag{32}$$

When $p = 1$, $h_T(\alpha, \varphi)$ is a scalar function given by

$$h_T(\alpha, \varphi) = \sum_{t=1}^{T-1} \left(1 + \alpha + ... + \alpha^{t-1}\right) \varphi_{t+1}. \tag{33}$$

Under homoskedasticity $\varphi_t = T^{-1}$ for all t, and the bias function (33) boils down to the expression in Nickell (1981) and Lancaster (2002), which for $|\alpha| < 1$ is[8]

$$h_T(\alpha) = \frac{1}{(1-\alpha)} \left[1 - \frac{1}{T}\left(\frac{1-\alpha^T}{1-\alpha}\right)\right]. \tag{34}$$

In view of (28)-(29), heteroskedasticity-consistent GMM estimators can be obtained as a solution to the nonlinear estimating equations

$$\begin{aligned} \sum_{i=1}^{N} X_i' D' \left(D \Lambda D'\right)^{-1} D v_i + N h_T(\alpha, \varphi) &= 0 & (35) \\ K' \left(D \Lambda D' \otimes D \Lambda D'\right)^{-1} vec \sum_{i=1}^{N} \left(D v_i v_i' D' - D \Lambda D'\right) &= 0. & (36) \end{aligned}$$

Consistency of the bias-corrected score estimator (BCS) that solves (35)-(36) does not depend on normality nor on conditional or time-series homoskedasticity.

BCS estimation is not possible from a $(2+p)$-wave panel (i.e. $T = 2$) because in that case α is not identified from the expected scores, which for $p = 1$ are given by

$$\begin{aligned} E\left[(y_{i1} - y_{i0})(v_{i2} - v_{i1})\right] &= -\sigma_1^2 & (37) \\ E\left[(v_{i2} - v_{i1})^2\right] &= \sigma_1^2 + \sigma_2^2. & (38) \end{aligned}$$

[8] Note that although the bias of the CML scores only depends on (α, φ), the asymptotic bias of the CML estimator of α as $N \to \infty$ also depends on the covariance matrix of $\left(\eta_i, y_i^0\right)$. Approximate bias formulae for homoskedastic WG were derived by Hahn and Kuersteiner (2002), and Alvarez and Arellano (2003). A bias-corrected estimator so constructed removes bias to order T^{-2} but is not fixed-T consistent.

This situation is in contrast with Lancaster's BCS estimator that enforces time series homoskedasticity (hence achieving identification from (37)-(38)), or the bias-corrected within-group estimator considered in Kiviet (1995).

The moment equations (28)-(29) are satisfied at the true values of (α, θ) but there may be other solutions. For example, in Section 7 we find that when $T = 3$, $p = 1$ and $\alpha = 1$ there are two observationally equivalent solutions under heteroskedasticity, so that α and θ are only locally identified. The solutions of the bias corrected scores for autoregressive models without heteroskedasticity have been studied by Dhaene and Jochmans (2016). A characterization of the solutions of bias corrected scores with time series heteroskedasticity remains an open question.

3.3 Modified Conditional Likelihood Interpretation

If the weights φ are known and $p = 1$, the method of moments estimators of α and ω_T based on the bias corrected scores

$$\begin{aligned} E\left[x_i' D' \left(D\Phi^{-1}D'\right)^{-1} Dv_i\right] &= -\omega_T h_T(\alpha, \varphi) \qquad (39) \\ E\left[v_i' D' \left(D\Phi^{-1}D'\right)^{-1} Dv_i\right] &= (T-1)\,\omega_T \qquad (40) \end{aligned}$$

can be regarded as the maximizers of the criterion function

$$L_{CR} = L_C + N b_T(\alpha, \varphi) \qquad (41)$$

where

$$b_T(\alpha, \varphi) = \sum_{t=1}^{T-1} \frac{(\varphi_{t+1} + \ldots + \varphi_T)}{t} \alpha^t, \qquad (42)$$

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, 2003a, p. 105)

$$L_{CR} = \sum_{i=1}^{N} \ln f\left(y_i \mid y_{i0}, \widehat{\lambda}_i\right), \qquad (43)$$

or as an integrated likelihood

$$L_{CR} = \sum_{i=1}^{N} \ln f(y_i \mid y_{i0}) = \sum_{i=1}^{N} \ln f(y_i \mid y_{i0}, \widehat{\eta}_i) + \sum_{i=1}^{N} \ln f(\widehat{\eta}_i \mid y_{i0}) \qquad (44)$$

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:

$$f(\widehat{\eta}_i \mid y_{i0}) = \kappa_i(\varphi)\, e^{b_T(\alpha, \varphi)} \qquad (45)$$

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 (21), whereas the second relies on factorization (22).

With unknown weights and $p > 1$ there is no orthogonal reparameterization, but for a heteroskedastic AR(p) model with unknown weights the BCS estimating equations coincide with the locally orthogonal Cox-Reid score function discussed in Woutersen (2002), Arellano (2003b), and Arellano and Hahn (2007), as we show in Appendix C. Thus, in our setting a first-order bias adjustment to the score is an exact correction that removes fully the bias, hence leading to fixed-T consistency.[9]

4 Random Effects Estimation

The analysis so far was conditional on y_i^0 and $\widehat{\eta}_i$. Conditioning on y_i^0 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_i^0, with linear mean and constant variance. A sufficient condition that we use for simplicity is:

Assumption G2: $\left(\eta_i, y_i^0\right)$ is jointly normally distributed with an unrestricted covariance matrix.

Normality of y_i^0 is inessential because its variance matrix is a free parameter, so the following analysis can be regarded as conditional on y_i^0. Clearly, assumptions $G1$ and $G2$ together imply that$\left(\eta_i, y_i^0, y_{i1}, ..., y_{iT}\right)$ are jointly normally distributed.

4.1 The Random Effects Log-Likelihood

Under $G2$, the MLE of η_i conditioned on y_i^0 is normally distributed as

$$\widehat{\eta}_i \mid y_i^0 \sim \mathcal{N}\left(\phi' y_i^0, \sigma_\varepsilon^2\right), \tag{46}$$

where $\phi = \Gamma_{00}^{-1}\gamma_{\eta 0}$ and $\sigma_\varepsilon^2 = \omega_T + \sigma_\eta^2 - \gamma_{\eta 0}'\Gamma_{00}^{-1}\gamma_{\eta 0}$. So, using factorization (22), the density of y_i conditioned on y_i^0 but marginal on η_i is:

$$\begin{aligned}\ln f\left(y_i \mid y_i^0\right) &= -\frac{1}{2}\ln\det\left(D\Lambda D'\right) - \frac{1}{2}v_i' D'\left(D\Lambda D'\right)^{-1} D v_i \\ &\quad -\frac{1}{2}\ln\sigma_\varepsilon^2 - \frac{1}{2\sigma_\varepsilon^2}\left(\overline{y}_i - \alpha'\overline{x}_i - \phi' y_i^0\right)^2.\end{aligned} \tag{47}$$

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

[9] Dhaene and Jochmans (2016, Lemma 2.2) show that in a homoskedastic AR(p) model there is an adjusted log-likelihood associated with the bias-corrected scores, despite the fact that orthogonalization is not possible when $p > 1$. They also show that there is a corresponding data-independent bias-reducing (bias-eliminating in this case) prior in the sense of Arellano and Bonhomme (2009).

given by

$$L_R = L_C - \frac{N}{2} \ln \sigma_\varepsilon^2 - \frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^{N} \left(\overline{u}_i - \phi' y_i^0 \right)^2, \tag{48}$$

with scores:

$$\frac{\partial L_R}{\partial \alpha} = \frac{\partial L_C}{\partial \alpha} + \frac{1}{\sigma_\varepsilon^2} \sum_{i=1}^{N} \overline{x}_i \left(\overline{u}_i - \phi' y_i^0 \right) \tag{49}$$

$$\frac{\partial L_R}{\partial \theta} = \frac{\partial L_C}{\partial \theta} + \frac{1}{\sigma_\varepsilon^2} \sum_{i=1}^{N} \Phi D' \left(D \Lambda D' \right)^{-1} D u_i \left(\overline{u}_i - \phi' y_i^0 \right) \tag{50}$$

$$\frac{\partial L_R}{\partial \phi} = \frac{1}{\sigma_\varepsilon^2} \sum_{i=1}^{N} y_i^0 \left(\overline{u}_i - \phi' y_i^0 \right) \tag{51}$$

$$\frac{\partial L_R}{\partial \sigma_\varepsilon^2} = \frac{1}{2\sigma_\varepsilon^4} \sum_{i=1}^{N} \left[\left(\overline{u}_i - \phi' y_i^0 \right)^2 - \sigma_\varepsilon^2 \right]. \tag{52}$$

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

$$E \left[\frac{1}{\sigma_\varepsilon^2} \overline{x}_i \left(\overline{u}_i - \phi' y_i^0 \right) \right] = h_T (\alpha, \varphi) \tag{53}$$

and

$$E \left[\frac{1}{\sigma_\varepsilon^2} \Phi D' \left(D \Lambda D' \right)^{-1} D v_i \left(\overline{u}_i - \phi' y_i^0 \right) \right] = 0. \tag{54}$$

Therefore, in view of (28) and (29), under Assumption A the expected scores evaluated at the true values of the parameters are equal to zero:

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

$$E \left[\frac{1}{2} K' \left(D \Lambda D' \otimes D \Lambda D' \right)^{-1} vec \left(D v_i v_i' D' - D \Lambda D' \right) \right.$$

$$\left. + \frac{1}{\sigma_\varepsilon^2} \Phi D' \left(D \Lambda D' \right)^{-1} D v_i \left(\overline{u}_i - \phi' y_i^0 \right) \right] = 0$$

$$E \left[y_i^0 \left(\overline{u}_i - \phi' y_i^0 \right) \right] = 0$$

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

The random effects maximum likelihood estimator (RML) solves the estimating equations (49)-(52) and is consistent and asymptotically normal under assumption A regardless of non-normality or conditional heteroskedasticity.

In a $(2+p)$-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 conditions

$$E\left[y_i^0\left(\Delta y_{i2}-\alpha_1 \Delta y_{i1}-\ldots-\alpha_p \Delta y_{i(2-p)}\right)\right]=0. \tag{55}$$

A random effects likelihood function for an autoregressive model with time series heteroskedasticity under the normality assumption $G2$ was first considered in Chamberlain (1980, p. 234-235). Similar likelihood functions for homoskedastic models have been considered in Blundell and Smith (1991), Sims (2000), and Alvarez and Arellano (2003). Correlated random effects approaches more generally are discussed in detail in Chamberlain (1984).

4.2 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 $2(T+p)$ moment conditions:

$$E\left[X_i' D'\left(D \Lambda D'\right)^{-1} D v_i\right]=-h_T(\alpha, \varphi) \tag{56}$$

$$E\left[K'\left(D \Lambda D' \otimes D \Lambda D'\right)^{-1} vec\left(D v_i v_i' D'-D \Lambda D'\right)\right]=0 \tag{57}$$

$$E\left[\left(D \Lambda D'\right)^{-1} D v_i\left(\overline{u}_i-\phi' y_i^0\right)\right]=0 \tag{58}$$

$$\begin{aligned} E\left[y_i^0\left(\overline{u}_i-\phi' y_i^0\right)\right] &= 0 && (59)\\ E\left[\left(\overline{u}_i-\phi' y_i^0\right)^2-\sigma_{\varepsilon}^2\right] &= 0. && (60) \end{aligned}$$

and a weight matrix calculated under the assumption of normality. Equation (53) gives the between-group covariance between the regressors and the error, in the same way as the BCS moments (56) specified the within-group covariance, but it is a redundant moment condition given (58), (59) and (60).[10]

BCS is based on moments (56) and (57), but RML is also using the information from the data in levels contained in (58). The $(T-1)$ overidentifying moments in (58) state the orthogonality between within-group and between-group errors (partialling out the initial observations). Finally, (59) and (60) 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 may not be ordered. Nevertheless, a GMM estimator based on (56)-(60) 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.

[10] Interestingly, $\left(\sigma_1^2, \sigma_2^2\right)$ are identified from the RML scores when $T=2$. In that case (57) determines $\left(\sigma_1^2+\sigma_2^2\right)$ and (58) determines φ_1. Note that when $T=2$ one of the two moments in (57) is redundant.

4.3 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:

$$L_R^* = L_C - \frac{N}{2} \ln \left[\left(\overline{y} - \alpha' \overline{x} \right)' S_0 \left(\overline{y} - \alpha' \overline{x} \right) \right] \tag{61}$$

where $S_0 = I_N - Y_0 \left(Y_0' Y_0 \right)^{-1} Y_0'$, and $Y_0 = \left(y_1^0, ..., y_N^0 \right)'$.

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

A simple OLS consistent estimator of the variance weights $\left(\varphi_1, ..., \varphi_{T-1} \right)$ for given α can be obtained from the fact that $E \left(\overline{u}_i \Delta v_{it} \right) = 0$. Such an estimator is useful for providing starting values for nonlinear likelihood-based estimation. The estimator is presented in Appendix A.4.

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_i^0 and η_i under the assumption that y_i^0 was observed but η_i was not. That is, a situation in which the data was a random sample of the vectors $\left(y_i^{0\prime}, y_{i1}, ..., y_{iT} \right)$. 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 $\left(\Delta y_{i(2-p)}, ..., \Delta y_{iT} \right)$. 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 $\left(\Delta y_{i(2-p)}, ..., \Delta y_{iT} \right)$ is equivalent to observing $\left(y_{i(2-p)}^\dagger, ..., y_{iT}^\dagger \right)' = \left(y_{i(2-p)} - y_{i(1-p)}, ..., y_{iT} - y_{i(1-p)} \right)'$, since the former results from multiplying the latter by the nonsingular transformation matrix of order $(T + p - 1)$:

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

with $\det \left(D^\dagger \right) = 1$. Also note that by construction $y_{i(1-p)}^\dagger = 0$. We shall use the notation $y_i^\dagger = y_i - y_{i(1-p)} \iota_T$ and $X_i^\dagger = X_i - y_{i(1-p)} \iota_T \iota_p'$. Similarly, $\overline{y}_i^\dagger = y_i^{\dagger\prime} \Phi \iota_T = \overline{y}_i - y_{i(1-p)}$, etc.[11]

[11] The following expression of $\overline{y}_i^\dagger$ makes explicit the connection to the data in differences:

$$\overline{y}_i^\dagger = \sum_{j=1}^{p} \Delta y_{i(1-p)+j} + \sum_{t=2}^{T} \left(\sum_{s=t}^{T} \varphi_s \right) \Delta y_{it}.$$

The original model can be written as

$$y_{i1}^{\dagger} = \alpha_1 y_{i0}^{\dagger} + ... + \alpha_{p-1} y_{i(2-p)}^{\dagger} + \eta_i^{\dagger} + v_{i1} \tag{62}$$

$$y_{it}^{\dagger} = \alpha_1 y_{i(t-1)}^{\dagger} + ... + \alpha_p y_{i(t-p)}^{\dagger} + \eta_i^{\dagger} + v_{it} \quad (t = 2, ..., T). \tag{63}$$

where

$$\eta_i^{\dagger} = \eta_i - (1 - \alpha_1 - ... - \alpha_p) y_{i(1-p)}. \tag{64}$$

Thus, the model for the deviations $y_{it}^{\dagger}$ can be regarded as a version of the original model in which $y_{i(1-p)}^{\dagger} = 0$ for all individuals and the effect is given by $\eta_i^{\dagger}$. From the point of view of this section, bundling together $y_{i(1-p)}$ 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 (62)-(63), applying (20) we have

$$\ln f\left(y_i^{\dagger} \mid y_i^{0\dagger}, \eta_i^{\dagger}, \widehat{\eta}_i^{\dagger}\right) = -\frac{1}{2} \ln \det \left(D\Lambda D'\right) - \frac{1}{2} v_i' D' \left(D\Lambda D'\right)^{-1} D v_i \tag{65}$$

where at true values

$$\widehat{\eta}_i^{\dagger} = \overline{y}_i^{\dagger} - \alpha' \overline{x}_i^{\dagger} = \eta_i^{\dagger} + \overline{v}_i = \overline{u}_i - \left(1 - \alpha' \iota_p\right) y_{i(1-p)}, \tag{66}$$

and following (19):

$$\widehat{\eta}_i^{\dagger} \mid y_i^{0\dagger}, \eta_i^{\dagger} \sim \mathcal{N}\left(\eta_i^{\dagger}, \omega_T\right). \tag{67}$$

Also, mimicking the marginal density decomposition in (22):

$$f\left(y_i^{\dagger} \mid y_i^{0\dagger}\right) = f\left(y_i^{\dagger} \mid y_i^{0\dagger}, \widehat{\eta}_i^{\dagger}\right) \int f\left(\widehat{\eta}_i^{\dagger} \mid y_i^{0\dagger}, \eta_i^{\dagger}\right) dG\left(\eta_i^{\dagger} \mid y_i^{0\dagger}\right). \tag{68}$$

Moreover, since $y_{i(1-p)}^{\dagger} = 0$ with probability one, for $p = 1$ densities conditioned on $y_i^{0\dagger}$ coincide with unconditional densities, and for $p > 1$ conditioning on $y_i^{0\dagger}$ is equivalent to conditioning on $\Delta y_i^0 = \left(\Delta y_{i(2-p)}, ..., \Delta y_{i0}\right)'$. Thus, for $p > 1$, $f\left(y_i^{\dagger} \mid \Delta y_i^0\right) = f\left(y_i^{\dagger} \mid y_i^{0\dagger}\right)$ and

$$\widehat{\eta}_i^{\dagger} \mid \Delta y_i^0, \eta_i^{\dagger} \sim \mathcal{N}\left(\eta_i^{\dagger}, \omega_T\right), \tag{69}$$

so that

$$f\left(y_i^{\dagger} \mid \Delta y_i^0\right) = f\left(y_i^{\dagger} \mid \Delta y_i^0, \widehat{\eta}_i^{\dagger}\right) f\left(\widehat{\eta}_i^{\dagger} \mid \Delta y_i^0\right). \tag{70}$$

Recall that the density $f\left(y_i^{\dagger} \mid \Delta y_i^0\right)$ is also the density of the first-differences of the data $(\Delta y_{i1}, ..., \Delta y_{iT})$ conditioned on Δy_i^0, which we are expressing as the product of the usual within-group conditional density and the density of $\widehat{\eta}_i^{\dagger}$ conditioned on Δy_i^0. Therefore, in the absence of steady state assumptions about initial conditions, the form of the density of panel AR(p) data in first differences depends on the distribution of the effects. In Section 6 we shall see that this dependence vanishes under the assumption of mean stationarity.

In sum, the BCS approach of Section 3 produces the same estimator regardless of whether one starts from the likelihood of the data in levels or in differences.

5.1 Random Effects Estimation in Differences

Next, we apply the random effects approach of Section 4 to the data in differences. This produces a different estimator (RML-dif), which as discussed below, has been shown to be more efficient than BCS but less efficient than RML under Assumption A (Bai, 2013). Interestingly, in Section 6 we will see that RML-dif can also be seen as a conditional ML estimator under mean stationarity.

Let $\left(\phi^{\dagger},\sigma_{\varepsilon}^{2\dagger}\right)$ denote the linear regression coefficients of $\widehat{\eta}_{i}^{\dagger}$ on Δy_{i}^{0}, so that $\sigma_{\varepsilon}^{2\dagger}$ satisfies

$$\sigma_{\varepsilon}^{2\dagger}=\sigma_{\eta^{\dagger}}^{2}+\omega_{T}-\phi^{\dagger\prime}Var\left(\Delta y_{i}^{0}\right)\phi^{\dagger}. \tag{71}$$

Under the normality assumption $G2$

$$\widehat{\eta}_{i}^{\dagger}\mid\Delta y_{i}^{0}\sim\mathcal{N}\left(\phi^{\dagger\prime}\Delta y_{i}^{0},\sigma_{\varepsilon}^{2\dagger}\right),$$

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

$$\begin{aligned}\ln f\left(\Delta y_{i1},...,\Delta y_{iT}\mid\Delta y_{i}^{0}\right) &= -\frac{1}{2}\ln\det\left(D\Lambda D'\right)-\frac{1}{2}v_{i}'D'\left(D\Lambda D'\right)^{-1}Dv_{i}\\ &\quad -\frac{1}{2}\ln\sigma_{\varepsilon}^{2\dagger}-\frac{1}{2\sigma_{\varepsilon}^{2\dagger}}\left(\overline{y}_{i}^{\dagger}-\overline{x}_{i}^{\dagger}\alpha-\phi^{\dagger\prime}\Delta y_{i}^{0}\right)^{2}\end{aligned} \tag{72}$$

Therefore, the random effects log-likelihood for the data in first-differences is a function of the parameter vector $\left(\alpha,\sigma_{1}^{2},...,\sigma_{T}^{2},\sigma_{\varepsilon}^{2\dagger},\phi^{\dagger}\right)$ given by

$$L_{RD}=L_{C}-\frac{N}{2}\ln\sigma_{\varepsilon}^{2\dagger}-\frac{1}{2\sigma_{\varepsilon}^{2\dagger}}\sum_{i=1}^{N}\left(\overline{y}_{i}^{\dagger}-\alpha'\overline{x}_{i}^{\dagger}-\phi^{\dagger\prime}\Delta y_{i}^{0}\right)^{2}. \tag{73}$$

Concentrating L_{RD} with respect to $\sigma_{\varepsilon}^{2\dagger}$ and $\phi^{\dagger}$, and letting $S_{\Delta}^{0}=I_{N}-Y_{\Delta}^{0}\left(Y_{\Delta}^{0\prime}Y_{\Delta}^{0}\right)^{-1}Y_{\Delta}^{0\prime}$ with $Y_{\Delta}^{0}=\left(\Delta y_{1}^{0},...,\Delta y_{N}^{0}\right)'$, we obtain the following criterion function that only depends on α and θ:

$$L_{RD}^{*}=L_{C}-\frac{N}{2}\ln\left(\overline{y}^{\dagger}-\alpha'\overline{x}^{\dagger}\right)'S_{\Delta}^{0}\left(\overline{y}^{\dagger}-\alpha'\overline{x}^{\dagger}\right), \tag{74}$$

which, in common with (61), 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.

In the $p=1$ case, the term Δy_{i}^{0} does not occur, so that (72) becomes a marginal density for the data in first differences and the log-likelihood is just a function of $\left(\alpha,\sigma_{1}^{2},...,\sigma_{T}^{2},\sigma_{\varepsilon}^{2\dagger}\right)$ given by

$$L_{RD}=L_{C}-\frac{N}{2}\ln\sigma_{\varepsilon}^{2\dagger}-\frac{1}{2\sigma_{\varepsilon}^{2\dagger}}\sum_{i=1}^{N}\left(\overline{y}_{i}^{\dagger}-\alpha\overline{x}_{i}^{\dagger}\right)^{2}. \tag{75}$$

5.2 Underidentification in a $(2+p)$-Wave Panel ($T=2$)

In common with BCS, RML-dif estimation is not possible from a $(2+p)$-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 illustrated by examining for $p=1$ the covariance matrix of the transformed series

$$Var\begin{pmatrix} y_{i0} \\ \Delta y_{i1} \\ \Delta y_{i2} \end{pmatrix} = \Omega^* = \begin{pmatrix} \gamma_{00} & \gamma_{0\Delta1} & \gamma_{0\Delta2} \\ \gamma_{0\Delta1} & & \\ \gamma_{0\Delta2} & & \Omega_\Delta \end{pmatrix},$$

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}, $\gamma_{0\Delta1}$ and $\gamma_{0\Delta2}$ unrestricted (Arellano, 2003a, p. 67). With $T=2$, the only identifying information about α is precisely the restriction $\gamma_{0\Delta2} = \alpha\gamma_{0\Delta1}$ 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.

5.3 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, using a factor analytical formulation, Bai (2013 Theorem S.4) shows that the RML (resp. RML-dif) estimator of α is fixed-T efficient in the sense of being asymptotically equivalent to the optimal GMM estimator that enforces the restrictions implied by our baseline assumption (Assumption A) on the second-order moments of the data in levels (resp. differences). Moreover, regardless of normality, under Assumption A estimates based on first-differences alone will never be more efficient than the 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 $\mu_i = \eta_i / (1 - \alpha' \iota_p)$. Specifically, we assume:

$$\gamma_{\eta 0} = \begin{pmatrix} Cov\left(\eta_i, y_{i(1-p)}\right) \\ \vdots \\ Cov\left(\eta_i, y_{i0}\right) \end{pmatrix} = \frac{\sigma_\eta^2}{(1 - \alpha' \iota_p)} \iota_p. \qquad (Assumption\ B)$$

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 a 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_i^0 given the effects:

$$y_i^0 \mid \mu_i \sim \mathcal{N}\left(\mu_i \iota_p, \Sigma_{00}\right) \qquad (Assumption\ G3)$$

where Σ_{00} satisfies $\Sigma_{00} = \Gamma_{00} - \iota_p \iota_p' \sigma_\eta^2 / (1 - \alpha' \iota_p)^2$.

Under assumptions $G1$ and $G3$

$$y_i^T \mid \mu_i \sim \mathcal{N}\left(\mu_i \bar{\iota}, V\right) \qquad (76)$$

where $y_i^T = \left(y_i^{0\prime}, y_{i1}, ..., y_{iT}\right)'$, $\bar{\iota}$ denotes a vector of ones of order $(T+p)$, and

$$V = \Gamma \Lambda^\dagger \Gamma' \qquad (77)$$

with

$$\Lambda^\dagger = \begin{pmatrix} \Sigma_{00} & 0 \\ 0 & \Lambda \end{pmatrix}, \qquad \Gamma = \begin{pmatrix} I_p & 0 \\ -B_T^{-1} B_{Tp} & B_T^{-1} \end{pmatrix} \qquad (78)$$

and

$$B_{Tp} = \begin{pmatrix} -\alpha_p & -\alpha_{p-1} & \dots & -\alpha_1 \\ 0 & -\alpha_p & \dots & -\alpha_2 \\ 0 & 0 & \ddots & \\ \vdots & \vdots & & \\ 0 & 0 & \dots & 0 \end{pmatrix}. \qquad (79)$$

Thus

$$\ln f\left(y_i^T \mid \mu_i\right) = -\frac{1}{2} \ln \det V - \frac{1}{2} \left(y_i^T - \mu_i \bar{\iota}\right)' V^{-1} \left(y_i^T - \mu_i \bar{\iota}\right). \qquad (80)$$

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

$$\widehat{\mu}_i = \left(\bar{\iota}' V^{-1} \bar{\iota}\right)^{-1} \bar{\iota}' V^{-1} y_i^T. \tag{81}$$

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

$$\mathcal{H} = \begin{pmatrix} \left(\bar{\iota}' V^{-1} \bar{\iota}\right)^{-1} \bar{\iota}' V^{-1} \\ \overline{D} \end{pmatrix}, \tag{82}$$

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

$$Var\left(\mathcal{H} y_i^T \mid \mu_i\right) = \begin{pmatrix} \left(\bar{\iota}' V^{-1} \bar{\iota}\right)^{-1} & 0 \\ 0 & \overline{D} V \overline{D}' \end{pmatrix} \tag{83}$$

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

$$f\left(y_i^T \mid \mu_i\right) = f\left(\mathcal{H} y_i^T \mid \mu_i\right) \left|\det \mathcal{H}\right| = f\left(\overline{D} y_i^T\right) f\left(\widehat{\mu}_i \mid \mu_i\right). \tag{84}$$

This is so because $\overline{D} y_i^T$ is independent of μ_i and the fact that $\left|\det \mathcal{H}\right| = 1$ (Arellano, 2003a, 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:

$$f\left(y_i^T \mid \widehat{\mu}_i, \mu_i\right) = \frac{f\left(y_i^T \mid \mu_i\right)}{f\left(\widehat{\mu}_i \mid \mu_i\right)} = f\left(\overline{D} y_i^T\right). \tag{85}$$

Thus, the log-likelihood conditioned on the ML estimates of the effects under mean stationarity is a function of $\left(\alpha, \sigma_1^2, ..., \sigma_T^2, vech\Sigma_{00}\right)$ given by

$$L_{CS} = -\frac{N}{2} \ln \det \left(\overline{D} V \overline{D}'\right) - \frac{1}{2} \sum_{i=1}^{N} y_i^{T\prime} \overline{D}' \left(\overline{D} V \overline{D}'\right)^{-1} \overline{D} y_i^T. \tag{86}$$

This result is similar to the one discussed by Lancaster (2002) for a homoskedastic stationary model with $p = 1$.

6.1.1 Comparison with the Marginal Likelihood for Differenced Data

Here we explain that L_{CS} in (86) is the same function as the random effects likelihood for differenced data in Section 5. The implication is that RML-dif without mean stationarity is the same estimator as conditional ML with mean stationarity.

In the previous section we obtained a random effects likelihood (73) for data in first-differences without assuming mean stationarity as a function of $\left(\alpha, \sigma_1^2, ..., \sigma_T^2, \sigma_\varepsilon^{2\dagger}, \phi^\dagger\right)$. This likelihood was conditioned on Δy_i^0 (unless $p = 1$), but adding to it the likelihood of Δy_i^0, we can write the likelihood of

$\overline{D}y_i^T$ in the absence of mean stationarity as a function of $\left(\alpha, \sigma_1^2, ..., \sigma_T^2, \sigma_\varepsilon^{2\dagger}, \phi^\dagger\right)$ and $\Sigma_\Delta = Var\left(\Delta y_i^0\right)$ given by[12]

$$L_{RDU} = L_{RD} - \frac{N}{2}\ln\det\Sigma_\Delta - \frac{1}{2}tr\left(\Sigma_\Delta^{-1}Y_\Delta^{0\prime}Y_\Delta^0\right). \tag{87}$$

If $p = 1$ the likelihood of $\overline{D}y_i^T$ in the absence of mean stationarity coincides with L_{RD} in (75).

In general, $\sigma_\varepsilon^{2\dagger}$ satisfies expression (71), which under mean stationarity becomes [13]

$$\sigma_\varepsilon^{2\dagger} = \left(1 - \alpha'\iota_p\right)^2\sigma_{00} + \omega_T - \sigma_{10}'D_p'\left(D_p\Sigma_{00}D_p'\right)^{-1}D_p\sigma_{10} \tag{88}$$

where we are using the partition of Σ_{00}

$$\Sigma_{00} = \begin{pmatrix} \sigma_{00} & \sigma_{10}' \\ \sigma_{10} & \Sigma_{11} \end{pmatrix}. \tag{89}$$

Similarly, under mean stationarity

$$\phi^\dagger = \left(D_p\Sigma_{00}D_p'\right)^{-1}D_p\sigma_{10}. \tag{90}$$

However, both $\sigma_\varepsilon^{2\dagger}$ and $\phi^\dagger$ remain free parameters because so is Σ_{00}.

Thus, the restriction of mean stationarity is immaterial to the data in first differences. L_{RDU} 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.[14] In particular the estimator that maximizes L_{CS} (or L_{RD}) will be consistent under Assumption A regardless of mean stationarity.[15]

Note that under homoskedasticity or covariance stationarity the situation is different because Σ_{00} is no longer a matrix of free parameters, but tied to α and the common variance σ^2.

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:

$$f\left(y_i^T\right) = \int f\left(y_i^T \mid \mu_i\right)dG\left(\mu_i\right) \tag{91}$$

[12] Note that $\Sigma_\Delta = D_p\Gamma_{00}D_p'$ where D_p is the first-difference operator of order $(p-1)\times p$.

[13] When $p = 1$ we just have $\sigma_{00} = \Sigma_{00}$ and $\sigma_\varepsilon^{2\dagger} = (1-\alpha)^2\sigma_{00} + \omega_T$.

[14] For further intuition, note that when $p = 1$, letting $\eta_i = (1-\alpha)\mu_i$ and $v_{i0} = y_{i0} - \mu_i$ we can write (6) as $y_{it} = \mu_i + \sum_{s=0}^t \alpha^s v_{i(t-s)}$ and in first-differences as $\Delta y_{it} = v_{it} - (1-\alpha)\sum_{s=1}^t \alpha^{s-1}v_{i(t-s)}$. Under mean stationarity μ_i and v_{i0} are uncorrelated, which constraints the data covariances in levels. However, the covariance restrictions for the differenced data remain the same regardless of mean stationarity; only the interpretation of the variance of v_{i0} will change.

[15] A conceptual difference is that since $\sigma_\varepsilon^{2\dagger}$ and $\phi^\dagger$ do not depend on σ_η^2 under mean stationarity, they would remain constant as $\sigma_\eta^2 \to \infty$.

whose log is given by

$$\ln f\left(y_i^T\right) = -\frac{1}{2}\ln\det\Omega - \frac{1}{2}y_i^T{}'\Omega^{-1}y_i^T \tag{92}$$

with

$$\Omega = \sigma_\mu^2\overline{\iota\iota}' + V. \tag{93}$$

Therefore, the random effects log-likelihood under mean stationarity is a function of the parameter vector $\left(\alpha, \sigma_1^2, ..., \sigma_T^2, vech\Sigma_{00}, \sigma_\eta^2\right)$ given by

$$L_{RS} = -\frac{N}{2}\ln\det\Omega - \frac{1}{2}\sum_{i=1}^{N}y_i^T{}'\Omega^{-1}y_i^T. \tag{94}$$

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 with $p = 1$ ($T = 2$), 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 $T + 1 + p\left(p+3\right)/2$ 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

6.2.1 Relation to RML without Mean Stationarity

Equation (48) in Section 4 gave the random effects log-likelihood conditioned on y_i^0. Adding to this expression the likelihood of y_i^0, we can write the likelihood of y_i^T in the absence of mean stationarity as a function of $\left(\alpha, \sigma_1^2, ..., \sigma_T^2, \phi, \sigma_\varepsilon^2, vech\Gamma_{00}\right)$ given by

$$L_{RU} = L_R - \frac{N}{2}\ln\det\Gamma_{00} - \frac{1}{2}tr\left(\Gamma_{00}^{-1}Y_0'Y_0\right). \tag{95}$$

If $p = 1$, in the parameterization of L_{RU}, mean stationarity can be expressed as the restriction

$$\sigma_\varepsilon^2 = \left(1-\alpha\right)\phi\left(1-\phi\right)\gamma_{00} + \omega_T. \tag{96}$$

Thus, RML-s can also be obtained maximizing L_{RU} subject to (96) in that case.

7 Unit Roots

In this section we discuss the possibility of identification failure when the autoregressive process has a unit root. We focus on the $p = 1$ case, so that the unit root model is

$$y_{it} = t\eta_i + v_{it} + v_{i(t-1)} + ... + v_{i1} + y_{i0}.$$

For this process, if $\sigma^2_\eta = 0$ the rank condition for GMM based on lagged levels as instruments for errors in differences fails, because changes in y_{it} are uncorrelated to lagged levels (e.g. Arellano and Honoré, 2001).[16] Thus, α would not be identified from RML in a three-wave panel ($T = 2$) when the true value is one, since in that case RML coincides with the IV estimator based on

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

Since the estimating criteria for the previous estimators depend on the data exclusively through second moments, it is useful to first look at the restrictions implied by the model on the data covariance matrix. Following Ahn and Schmidt (1995), for $T \geq 3$ these restrictions can be represented as

$$E\left[y_{is}\left(\Delta y_{it} - \alpha\Delta y_{i(t-1)}\right)\right] = 0 \quad (t = 2, ..., T; s = 0, ..., t-2) \tag{97}$$

$$E\left[\left(\Delta y_{i(t-1)} - \alpha\Delta y_{i(t-2)}\right)\left(y_{it} - \alpha y_{i(t-1)}\right)\right] = 0 \quad (t = 3, ..., T). \tag{98}$$

When $T = 3$ and the true values are $\overline{\alpha} = 1$ and $\overline{\sigma}^2_\eta = 0$, (98) consists of just one quadratic equation

$$a_1\alpha^2 + b_1\alpha + c_1 = 0 \tag{99}$$

with coefficients given by

$$\begin{aligned} a_1 &= E\left(y_{i2}\Delta y_{i1}\right) = \overline{\sigma}^2_1 \\ b_1 &= -E\left(y_{i2}\Delta y_{i2} + y_{i3}\Delta y_{i1}\right) = -\left(\overline{\sigma}^2_1 + \overline{\sigma}^2_2\right) \\ c_1 &= E\left(y_{i3}\Delta y_{i2}\right) = \overline{\sigma}^2_2 \end{aligned}$$

where $\overline{\sigma}^2_1$, $\overline{\sigma}^2_2$ and $\overline{\sigma}^2_3$ denote the true values of the error variances.

Equation (99) has two roots given by

$$\frac{\overline{\sigma}^2_1 + \overline{\sigma}^2_2 \pm \left(\overline{\sigma}^2_2 - \overline{\sigma}^2_1\right)}{2\overline{\sigma}^2_1} = \begin{cases} \alpha^*_1 = \overline{\sigma}^2_2/\overline{\sigma}^2_1 \\ \overline{\alpha} = 1 \end{cases} \tag{100}$$

Therefore, under time series heteroskedasticity there is local identification from (99) but not global identification. If $\overline{\sigma}^2_1 = \overline{\sigma}^2_2$ there is global identification but first-order underidentification, because the first derivative of (99)

$$2a_1\alpha + b_1 = 0 \tag{101}$$

[16] When $\alpha = 1$ and $\sigma^2_\eta = 0$, heterogeneity only plays a role in the determination of the initial observations of the process. In contrast, if $\sigma^2_\eta \neq 0$ the model is a random walk with heterogeneous linear growth.

vanishes at $\alpha = 1$. In that case there is second-order identification because $\alpha = 1$ is the only solution to equation (101) and the second derivative does not vanish (Sargan, 1983).[17]

In general, we get $T-2$ equations of the same form as (99), each one with a solution of the form $\alpha_t^* = \overline{\sigma}_{t+1}^2/\overline{\sigma}_t^2$, aside from unity. Thus, for $T > 3$ there is both first-order and global identification from (98) under heteroskedasticity, unless the unconditional variances change at a constant rate of growth (i.e. $\overline{\sigma}_{t+1}^2/\overline{\sigma}_t^2$ is constant for $t = 1, ..., T-2$).

7.1 Heteroskedastic BCS and Unit Roots

Next, we develop the local identification result for the bias-corrected CML scores when $T = 3$. The expected BCS equations are given by

$$E\left[x_i' D' \left(D\Lambda D'\right)^{-1} Dv_i\right] = -h_T\left(\alpha, \varphi\right) \tag{102}$$

$$E\left[K' \left(D\Lambda D' \otimes D\Lambda D'\right)^{-1} vec\left(Dv_i v_i' D' - D\Lambda D'\right)\right] = 0. \tag{103}$$

where

$$Dx_i = \begin{pmatrix} \Delta y_{i1} \\ \Delta y_{i2} \end{pmatrix}, \qquad Dy_i = \begin{pmatrix} \Delta y_{i2} \\ \Delta y_{i3} \end{pmatrix},$$

$$\left(D\Lambda D'\right)^{-1} = \frac{1}{\left(\sigma_1^2\sigma_2^2 + \sigma_1^2\sigma_3^2 + \sigma_2^2\sigma_3^2\right)} \begin{pmatrix} \sigma_2^2 + \sigma_3^2 & \sigma_2^2 \\ \sigma_2^2 & \sigma_1^2 + \sigma_2^2 \end{pmatrix},$$

and

$$h_T\left(\alpha, \varphi\right) = \varphi_2 + \left(1+\alpha\right)\varphi_3 = \frac{\sigma_1^2}{\left(\sigma_1^2\sigma_2^2 + \sigma_1^2\sigma_3^2 + \sigma_2^2\sigma_3^2\right)} \left[\sigma_3^2 + \left(1+\alpha\right)\sigma_2^2\right].$$

When the true values are $\overline{\alpha} = 1$ and $\overline{\sigma}_\eta^2 = 0$, the first score (102) can be written as

$$tr\left[\begin{pmatrix} \sigma_2^2 + \sigma_3^2 & \sigma_2^2 \\ \sigma_2^2 & \sigma_1^2 + \sigma_2^2 \end{pmatrix} \begin{pmatrix} \alpha\overline{\sigma}_1^2 & -\overline{\sigma}_2^2 \\ 0 & \alpha\overline{\sigma}_2^2 \end{pmatrix}\right] = \sigma_1^2\left[\left(\sigma_2^2 + \sigma_3^2\right) + \alpha\sigma_2^2\right] \tag{104}$$

Moreover,

$$E\left(Dv_i v_i' D'\right) = \begin{pmatrix} \overline{\sigma}_2^2 + \alpha^2\overline{\sigma}_1^2 & -\alpha\overline{\sigma}_2^2 \\ -\alpha\overline{\sigma}_2^2 & \overline{\sigma}_3^2 + \alpha^2\overline{\sigma}_2^2 \end{pmatrix}.$$

Hence, in view of the second block of scores (103), we have

$$\begin{aligned} \overline{\sigma}_2^2 + \alpha^2\overline{\sigma}_1^2 &= \sigma_1^2 + \sigma_2^2 \\ \alpha\overline{\sigma}_2^2 &= \sigma_2^2 \\ \overline{\sigma}_3^2 + \alpha^2\overline{\sigma}_2^2 &= \sigma_2^2 + \sigma_3^2 \end{aligned} \tag{105}$$

[17] A similar result under homoskedasticity was independently found by Ahn and Thomas (2006).

Now, substituting in (104)

$$tr\left[\left(\begin{array}{cc}\overline{\sigma}_3^2+\alpha^2\overline{\sigma}_2^2 & \alpha\overline{\sigma}_2^2\\ \alpha\overline{\sigma}_2^2 & \overline{\sigma}_2^2+\alpha^2\overline{\sigma}_1^2\end{array}\right)\left(\begin{array}{cc}\alpha\overline{\sigma}_1^2 & -\overline{\sigma}_2^2\\ 0 & \alpha\overline{\sigma}_2^2\end{array}\right)\right]=\left(\overline{\sigma}_2^2+\alpha^2\overline{\sigma}_1^2-\alpha\overline{\sigma}_2^2\right)\left(\overline{\sigma}_3^2+2\alpha^2\overline{\sigma}_2^2\right),$$

which can be rearranged as

$$(1-\alpha)\left(\overline{\sigma}_2^2-\overline{\sigma}_1^2\alpha\right)\left(\overline{\sigma}_3^2+2\overline{\sigma}_2^2\alpha^2\right)=0. \tag{106}$$

Thus, as before there are two real roots: $\overline{\alpha}=1$ and $\alpha^*=\overline{\sigma}_2^2/\overline{\sigma}_1^2$. Corresponding to $\alpha=1$ we have

$$\left(\begin{array}{c}\sigma_1^2\\ \sigma_2^2\\ \sigma_3^2\end{array}\right)=\left(\begin{array}{c}\overline{\sigma}_1^2\\ \overline{\sigma}_2^2\\ \overline{\sigma}_3^2\end{array}\right), \tag{107}$$

and corresponding to $\alpha=\overline{\sigma}_2^2/\overline{\sigma}_1^2$

$$\left(\begin{array}{c}\sigma_1^2\\ \sigma_2^2\\ \sigma_3^2\end{array}\right)=\left(\begin{array}{c}\sigma_1^{2*}\\ \sigma_2^{2*}\\ \sigma_3^{2*}\end{array}\right)\equiv\left(\begin{array}{c}\overline{\sigma}_2^2\\ \frac{\overline{\sigma}_2^4}{\overline{\sigma}_1^2}\\ \overline{\sigma}_3^2-\frac{\overline{\sigma}_2^4}{\overline{\sigma}_1^4}\left(\overline{\sigma}_1^2-\overline{\sigma}_2^2\right)\end{array}\right). \tag{108}$$

7.2 Expected RML Likelihood

Finally, we consider the expected random effects likelihood for one observation when $T=3$, $\overline{\alpha}=1$ and $\overline{\sigma}_\eta^2=0$. This is a function of $\left(\alpha,\sigma_1^2,\sigma_2^2,\sigma_3^2,\phi,\sigma_\varepsilon^2\right)$ given by

$$\begin{aligned}E\left(L_{Ri}\right) &= -\frac{1}{2}\ln\det\left(D\Lambda D'\right)-\frac{1}{2}tr\left[\left(D\Lambda D'\right)^{-1}E\left(Dv_iv_i'D'\right)\right]\\ &\quad -\frac{1}{2}\ln\sigma_\varepsilon^2-\frac{1}{2\sigma_\varepsilon^2}E\left[\left(\overline{u}_i-\phi y_{i0}\right)^2\right]\end{aligned} \tag{109}$$

Note that the true values of ϕ and σ_ε^2 are $\overline{\phi}=0$ and $\overline{\sigma}_\varepsilon^2=\overline{\omega}_T$.

Maximizing $E\left(L_{Ri}\right)$ with respect to $\phi,\sigma_\varepsilon^2$ for given $\left(\alpha,\sigma_1^2,\sigma_2^2,\sigma_3^2\right)$ we get

$$\phi=1-\alpha \tag{110}$$

$$\sigma_\varepsilon^2=E\left[\left(\overline{y}_i-\alpha\overline{x}_i-(1-\alpha)\,y_{i0}\right)^2\right]=E\left[\left(\overline{y}_i^\dagger-\alpha\overline{x}_i^\dagger\right)^2\right]. \tag{111}$$

Therefore, the concentrated expected likelihood for the data in levels and in differences coincide. An implication is that when $\overline{\alpha}=1$ and $\overline{\sigma}_\eta^2=0$, RML in levels and RML in differences are asymptotically equivalent.

Moreover, the maximum of $E\left(L_{Ri}\right)$ is attained at

$$\max E\left(L_{Ri}\right)=-\frac{1}{2}\ln\left(\overline{\sigma}_1^2\overline{\sigma}_2^2\overline{\sigma}_3^2\right)-\frac{3}{2} \tag{112}$$

by $\left(\overline{\alpha}, \overline{\sigma}_1^2, \overline{\sigma}_2^2, \overline{\sigma}_3^2, \overline{\phi}, \overline{\sigma}_\varepsilon^2\right)$ and $\left(\alpha^*, \sigma_1^{2*}, \sigma_2^{2*}, \sigma_3^{2*}, \phi^*, \sigma_\varepsilon^{2*}\right)$, where

$$\phi^* = 1 - \alpha^* = 1 - \frac{\overline{\sigma}_2^2}{\overline{\sigma}_1^2} \tag{113}$$

$$\sigma_\varepsilon^{2*} = \frac{\overline{\sigma}_1^2 \overline{\sigma}_2^2 \overline{\sigma}_3^2}{\overline{\sigma}_2^2 \overline{\sigma}_3^2 + \overline{\sigma}_2^4 \frac{\overline{\sigma}_3^2}{\overline{\sigma}_1^2} + \frac{\overline{\sigma}_2^{10}}{\overline{\sigma}_1^6}}, \tag{114}$$

which completes the characterization of the two observationally equivalent points.

8 Calculations of Relative Asymptotic Variances

We perform numerical calculations of the asymptotic variances for various estimators of the autoregressive coefficient. We report, for $p = 1$, the asymptotic variances of both homoskedastic and heteroskedastic BCS and RML-dif estimators, relative to the corresponding RML in levels, calculated under the assumption of normality. Formulae for the asymptotic variances are derived in Appendix B.

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 homoskedasticity with $\sigma^2 = 1$.[18]

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.

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.

[18] Because of stationarity $\gamma_{00} = \sigma^2 / \left(1 - \alpha^2\right)$, so that it increases with α.

Turning to heteroskedastic estimators, Figures 4 and 5 display relative inefficiency ratios for heteroskedastic BCS and RML-dif, similar to those in the previous figures. The calculations are under homoskedasticity and stationarity, for $\lambda = 0$ and 1, $T = 3$ and 9, and $\sigma^2 = 1$. As before, the inefficiencies of heteroskedastic BCS and RML-dif increase with α and decrease with λ and T, but they have larger magnitudes than those of their homoskedastic counterparts.

Table 1 illustrates the extent of these differences by showing the inefficiencies of homoskedastic and heteroskedastic estimators for selected values of the parameters. Some of the inefficiencies are quite large. For example, for the heteroskedastic estimators with $\alpha = 0.8$, $T^o = 4$ and $\lambda = 0$, the standard error of RML-dif is more than twice that of RML-lev, and the standard error of BCS is more than three times as large.

Finally, Figure 6 reports asymptotic standard deviations of BCS and RML when $\alpha = 1$ and $\lambda = 0$ (in this case RML-dif and RML-lev are asymptotically equivalent) for $T = 6$ and a single break in the error variance. Standard deviations (scaled by $\sqrt{900}$) are given as a function of the percentage change in variance and for two different locations of the variance break (which takes place either during the last 2 or the last 4 periods).[19] As expected, asymptotic standard deviations decrease with the strength of heteroskedasticity, and are smaller when the variance break is centrally located than when it only occurs during the last two periods.

9 Empirical Illustration: Individual Earnings Dynamics

In order to illustrate the properties of the previous methods, we estimate first- and second-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 Browning, Ejrnæs, and Alvarez (2010).

There are $N = 632$ individuals and $T^0 = 6$ waves in the Spanish data set, and $N = 792$ and $T^0 = 7$ 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 Browning, Ejrnæs, and Alvarez 2010, for further details on the PSID sample, and tables A2 and A3 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 AR(1) results for the Spanish data are reported in the first part of Table 2. Heteroskedastic

[19] When $\alpha = 1$, we considered choices of γ_{00} and $\gamma_{\eta 0}$ of the form $\gamma_{00} = \kappa\overline{\sigma}^2 + \kappa^2\sigma_\eta^2$ and $\gamma_{\eta 0} = \kappa\sigma_\eta^2$, where $\overline{\sigma}^2 = T^{-1}\sum_{t=1}^{T}\sigma_t^2$. But for the calculations in Figure 6, since $\sigma_\eta^2 = 0$ the results turn out to be invariant to the choice of κ.

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 AR(1) 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. The RMLr estimate subject to mean stationarity is smaller than system-GMM, but a Wald test of the mean stationarity restriction rejects (with a "t ratio" for σ_ε^2 of 2.54). Finally, the within-group (WG) estimate 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 AR(1) results for the PSID sample, reported in Table 3, also show a marked discrepancy between the likelihood-based estimates and GMM, and a similar rejection of mean stationarity from the incremental Sargan test, although not from RML estimation (the "t ratio" for σ_ε^2 is just 0.16). In the PSID data there is more state dependence than in the Spanish data, at least as measured by the first-order 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 AR(1) estimates reported in the tables, 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.

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, suggesting misspecification as the most likely reason for these discrepancies. This impression is confirmed by the AR(2) estimates and the simulation results reported below.

The upshot from the AR(2) estimates reported in the second parts of tables 2 and 3 is that there is a positive autoregressive root, in the $(0.4, 0.5)$ range for the Spanish panel and in the $(0.6, 0.7)$ range for PSID, and a negative root of around -0.2 in both datasets (so that an ARMA(1,1) model would provide a similar fit).

The AR(2) GMM estimates are still smaller than the likelihood-based estimates, and there is a discrepancy between BCS and RMLr (specially for PSID), all of which suggests that there may be some remaining misspecification.[20] This suggestion is reinforced by the robust GMM form of the estimates shown in tables A4 and A5, which provide evidence against the overidentifying restrictions in the PSID data. These GMM estimates use the $2(T+2)$ moments (56)-(60) with RML adding extra moment conditions to BCS. Moreover, mean stationarity is rejected in both datasets and, when

[20] We found that the BCS equations, in addition to the stable solution, had another solution with an explosive root.

enforced, leads to somewhat larger positive roots.

However, in contrast with other studies that either imposed or found a unit root in individual earnings (e.g. MaCurdy, 1982), we find no evidence of unit roots. The only way we managed to obtain a near-unit root is by imposing the restriction that the initial observations of earnings are orthogonal to the unobserved component (i.e. $\phi = 0$). Doing this led to an estimated positive root of 0.95 in both panels. Clearly, if only heterogeneity that is orthogonal to initial observations is allowed, any nonorthogonal heterogeneity will be captured by the autoregressive part of the model as spurious state dependence.[21]

9.1 Moving Average Errors

We checked whether this conclusion was affected by adding a moving average component to the specification of PSID earnings. In such a case the autoregressive coefficients can no longer be interpreted as a model for the conditional expectation of earnings given past observations, but an ARMA model might lead to a more parsimonious specification. Moreover, models of earnings that specify a measurement error component imply a reduced form with moving average errors. Appendix D describes our ARMA specification and the random effects ML estimators that we used.

Table 4 reports ARMA(1,1), ARMA(1,2), and ARMA(2,1) estimates from the PSID sample. As expected, the ARMA(1,1) estimates are similar to those obtained from the AR(2) specification. However, the ARMA(1,2) and the ARMA(2,1) estimates were very imprecise, suggesting that there is no enough variation in the data covariance matrix to support a three-parameter dynamic specification within this class of models.

9.2 Testing for Nonnormality

The distributions of the effects and the autoregressive errors are nonparametrically identified and can be estimated using deconvolution techniques as in Horowitz and Markatou (1996).

Horowitz and Markatou carried out graphical tests of normality of the distributions of errors and effects in a static earnings model using a two-wave panel from the CPS.[22] We used their diagnostics and found very similar results for PSID autoregressive models. A normal probability plot of residuals in first-differences (Figure 7) indicates that the tails of the distribution of errors are thicker than those of the normal distribution. However, a plot of the log empirical characteristic function of the effects

[21] Studies that have explored more general models of PSID earnings by allowing for richer forms of heterogeneity or nonlinear dynamics have found evidence of misspecification in conventional linear models. For example, Browning, Ejrnæs and Alvarez (2010) test the weaker hypothesis that some agents have a unit root and others a stable process; they reject the hypothesis that anyone has a unit root. Arellano, Blundell and Bonhomme (2017) develop a quantile-based framework to explore the nonlinear nature of income shocks; they find that the impact of past shocks can be altered by the size and sign of new shocks, so that the future persistence of a current shock is not fixed, as in a linear mean-reverting or unit-root model, but stochastic due to its dependence on future shocks.

[22] Figures 1 and 5 in their paper

against minus the square of its argument is almost a straight line, hence showing no deviation from normality (Figure 8).

9.3 Monte Carlo Simulations

To illustrate the properties of the estimators, we performed a small simulation exercise calibrated to the likelihood-based AR(1) estimates from 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 Table 5 we report means and standard deviations of the WG, GMM1, RML(nr), RML(r), and BCS estimators of the AR(1) model for $\alpha = 0.4$ and 0.8 (with $\overline{\sigma}_0^2 = 0.11$ and 0.28, respectively). The results show that both RML(r) and BCS are virtually unbiased. Those for $\alpha = 0.4$ nicely reproduce the WG downward bias and the RML(nr) upward bias that we found in the PSID sample. However, the results fail to explain the performance of GMM with the real data, which reinforces the evidence of misspecification in the AR(1) earnings models.

10 Concluding Remarks

In this paper we have considered likelihood-based estimation strategies of autoregressive panel models, which are consistent under the same baseline assumptions as Arellano-Bond and Ahn-Schmidt GMM estimators. The starting point is that to achieve this goal one has to allow for time-series heteroskedasticity.

Our leading method is a heteroskedastic correlated random-effects estimator (RML) that maximizes a marginal likelihood function where individual effects are normally distributed with a mean that depends linearly on the initial observations. The literature has uncovered some attractive properties of this estimator. Firstly, it is fixed-T efficient in the sense of being asymptotically equivalent to the optimal GMM estimator that enforces the restrictions implied by our baseline assumption (Assumption A) on the data second-order moments regardless of nonnormality (Bai 2013 Theorem S.4). Secondly, it does not lead to incidental-parameter bias when T and N are of comparable dimension (Bai 2013). Finally, under normality of the errors (Assumption G1) but not necessarily under normality of the effects (Assumption G2), RML is a finite-sample minimax optimal estimator for a suitable choice of prior distributions in the sense of Chamberlain and Moreira (2009).

For comparisons, we have considered two other likelihood-based estimators, which contrary to RML only depend on the data in first-differences. The first one (BCS) solves a bias-corrected score function of the heteroskedastic likelihood conditioned on the MLE of the incidental parameters. The other (RML-dif) maximizes a marginal likelihood function of the same form as RML but for the data in first-differences. The three estimators, RML, BCS and RML-dif, can be variously interpreted as fixed-effects, random effects, Bayesian, or method-of-moment estimators. For example, versions of BCS can be regarded as a random effects or Bayesian estimator that specifies a nonnormal prior for the effects with a very large variance (Chamberlain and Moreira 2009, p. 131; Dhaene and Jochmans

2016, p. 1184-1185). Moreover, the RML and RML-dif estimators coincide with the corresponding fixed-effects factor analytic estimators studied in Bai (2013), which estimate the sample variance of the fixed effects.

One major theme of this paper has been to highlight the advantages of heteroskedastic RML estimation relative to traditional GMM methods in finite samples and large T asymptotics. The other major theme has been to highlight the efficiency gains from using data in levels (as in RML) relative to only using data in differences (as in BCS or RML-dif). We have done so for our baseline model and for a more restrictive model that assumes stationarity in mean. In the latter case, the likelihood-based estimators that we discuss are consistent under the same assumptions as the Arellano-Bover and Blundell-Bond system GMM estimators.

Appendix

A Conditional Maximum Likelihood and Expected Scores

A.1 First-Order Conditions and Related Results

Equations (15), (18): Note that $\overline{v} = v'\Phi\iota$ and $\omega_T = Var\left(\overline{v}\right) = \left(\iota'\Lambda^{-1}\iota\right)^{-1}$, so that $\Lambda^{-1} = \left(1/\omega_T\right)\Phi$. Moreover, the equivalences in (14) also imply

$$\ln\det\Lambda = \ln\det\left(D\Lambda D'\right) + \ln\omega_T. \tag{A.1}$$

Clearly $0 \leq \varphi_t \leq 1$, $\sum_{t=1}^{T}\varphi_t = 1$, and under homoskedasticity $\varphi_t = 1/T$ for all t.

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 obtain expression (15), and also

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

Finally, equation (18) is easily verified from (15).

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

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

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

$$\frac{\partial\varphi}{\partial\theta'} = -\left(\Phi - \Phi\iota\iota'\Phi\right)\Lambda^{-1} = -D'\left(D\Lambda D'\right)^{-1}D\Phi. \tag{A.2}$$

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

$$\frac{\partial\omega_T}{\partial\sigma_s^2} = \frac{1/\sigma_s^4}{\left(\sigma_1^{-2} + ... + \sigma_T^{-2}\right)^2} = \varphi_s^2, \tag{A.3}$$

we get

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

First-Order Conditions for Variance Parameters when Maximizing L_C: The derivatives of L_C with respect to $\theta = \left(\sigma_1^2 ... \sigma_T^2\right)'$ given in (26) are

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

where K is a $(T-1)^2 \times T$ selection matrix such that $vec\,(D\Lambda D') = K\theta$. Let d_t and k_t be the t-th columns of D and K, respectively, so that $D\Lambda D' = \sum_{t=1}^{T}\sigma_t^2 d_t d_t'$, $K\theta = \sum_{t=1}^{T}\sigma_t^2 k_t$, and $k_t = d_t \otimes d_t$. Thus, also

$$\frac{\partial L_C}{\partial\sigma_t^2} = \frac{1}{2}\sum_{i=1}^{N} d_t' \left(D\Lambda D'\right)^{-1} \left(Dv_i v_i' D' - D\Lambda D'\right) \left(D\Lambda D'\right)^{-1} d_t \quad (t = 1, ..., T)\,. \tag{A.4}$$

Maximizing L_C in (23) 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 be written in a form analogous to (15) and (18) as

$$\sum_{i=1}^{N}\left[\frac{1}{(T-1)} v_i' \left(\Phi - \Phi\iota\iota'\Phi\right) v_i - \omega_T\right] = 0 \tag{A.5}$$

$$\sum_{i=1}^{N}\left[\left(v_{it} - \overline{v}_i\right)^2 - \left(v_{i(t-1)} - \overline{v}_i\right)^2 - \left(\sigma_t^2 - \sigma_{t-1}^2\right)\right] = 0 \quad (t = 2, ..., T)\,. \tag{A.6}$$

The details are as follows. For a matrix $A = (a_1, ..., a_n)'$, we use the notation $vec\,(A) = (a_1', ..., a_n')'$ and $A \otimes B = \{a_{jk}B\}$. The derivative of L_C with respect ω_T is

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

The concentrated likelihood with respect to ω_T is

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

and the Lagrangean

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

so that

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

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

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

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. This transformation is discussed next.

A.2 Heteroskedastic Orthogonal Deviations

The following equivalences also hold

$$v'D'\left(D\Lambda D'\right)^{-1}Dv=\sum_{t=1}^{T-1}\frac{\widetilde{v}_t^2}{\widetilde{\sigma}_t^2} \tag{A.7}$$

$$\ln\det\left(D\Lambda D'\right)=\sum_{t=1}^{T}\ln\sigma_t^2+\ln\left(\sigma_1^{-2}+...+\sigma_T^{-2}\right)=\sum_{t=1}^{T-1}\ln\widetilde{\sigma}_t^2 \tag{A.8}$$

where the heteroskedastic orthogonal deviations are given by

$$\widetilde{v}_t=\begin{cases} v_{T-1}-v_T & \text{for } t=T-1 \\ \\ v_t-\frac{\sigma_{t+1}^{-2}v_{t+1}+...+\sigma_T^{-2}v_T}{\sigma_{t+1}^{-2}+...+\sigma_T^{-2}} & \text{for } t=T-2,...,1\end{cases} \tag{A.9}$$

$$\widetilde{\sigma}_t^2=\begin{cases} \sigma_{T-1}^2+\sigma_T^2 & \text{for } t=T-1 \\ \\ \sigma_t^2+\frac{1}{\sigma_{t+1}^{-2}+...+\sigma_T^{-2}} & \text{for } t=T-2,...,1\end{cases}. \tag{A.10}$$

or

$$\widetilde{v}_t=\begin{cases} v_{T-1}-v_T & \text{for } t=T-1 \\ \\ (v_t-v_{t+1})+\lambda_{t+1}\widetilde{v}_{t+1} & \text{for } t=T-2,...,1\end{cases} \tag{A.11}$$

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

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

$$E\left[\left(v_{T-1} - v_T\right)\left(v_{T-2} - v_T\right)\right] = \sigma_T^2$$

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

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

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

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.

A.3 Score Bias Function

Proof of (28): We have

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

$$= -\omega_T E \begin{pmatrix} x_{1i}' \Lambda^{-1} \iota\iota' \Lambda^{-1} v_i \\ \vdots \\ x_{pi}' \Lambda^{-1} \iota\iota' \Lambda^{-1} v_i \end{pmatrix} = -\omega_T \begin{pmatrix} \iota' \Lambda^{-1} E\left(x_{1i} v_i'\right) \Lambda^{-1} \iota \\ \vdots \\ \iota' \Lambda^{-1} E\left(x_{pi} v_i'\right) \Lambda^{-1} \iota \end{pmatrix}$$

To obtain an expression for $E\left(x_{ji} v_i'\right)$ we need to develop a suitable notation. Let us write

$$\begin{pmatrix} I_p & 0 \\ B_{Tp} & B_T \end{pmatrix} \begin{pmatrix} y_i^0 \\ y_i \end{pmatrix} = \begin{pmatrix} y_i^0 \\ \eta_i \iota + v_i \end{pmatrix} \tag{A.12}$$

where

$$\begin{pmatrix} B_{Tp} & B_T \end{pmatrix} = \begin{pmatrix} -\alpha_p & -\alpha_{p-1} & \dots & -\alpha_1 & 1 & 0 & \dots & 0 & \dots & 0 & 0 \\ 0 & -\alpha_p & \dots & -\alpha_2 & -\alpha_1 & 1 & & 0 & \dots & 0 & 0 \\ 0 & 0 & \ddots & & -\alpha_2 & -\alpha_1 & \ddots & 0 & \dots & 0 & 0 \\ \vdots & \vdots & & & & \ddots & \ddots & \ddots & & \vdots & \vdots \\ 0 & 0 & & 0 & 0 & & & 0 & & 1 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & -\alpha_p & \dots & -\alpha_1 & 1 \end{pmatrix}.$$

Moreover,

$$\begin{pmatrix} y_i^0 \\ y_i \end{pmatrix} = \begin{pmatrix} I_p & 0 \\ \overline{C}_{Tp} & \overline{C}_T \end{pmatrix} \begin{pmatrix} y_i^0 \\ \eta_i \iota + v_i \end{pmatrix} \tag{A.13}$$

where $\overline{C}_T = B_T^{-1}$ and $\overline{C}_{Tp} = -B_T^{-1} B_{Tp}$, so that

$$y_i = \overline{C}_{Tp} y_i^0 + \eta_i \overline{C}_T \iota + \overline{C}_T v_i. \tag{A.14}$$

Thus,

$$E\left(y_i v_i'\right) = \overline{C}_{Tp} E\left(y_i^0 v_i'\right) + \overline{C}_T E\left(v_i v_i'\right) = \overline{C}_T \Lambda. \tag{A.15}$$

Let us consider now an expression for $x_{ji} = \left(y_{i(1-j)}, ..., y_{i0}, y_{i1}, ..., y_{i(T-j)}\right)'$. Since we have

$$\begin{pmatrix} y_{i(1-j)} \\ \vdots \\ y_{i0} \end{pmatrix} = \begin{pmatrix} 0 & I_j \end{pmatrix} y_i^0$$

and

$$\begin{pmatrix} y_{i1} \\ \vdots \\ y_{i(T-j)} \end{pmatrix} = \overline{C}_{(T-j)p} y_i^0 + \eta_i \overline{C}_{T-j} \iota_{T-j} + \overline{C}_{T-j} \begin{pmatrix} v_{i1} \\ \vdots \\ v_{i(T-j)} \end{pmatrix},$$

we can write x_{ji} as

$$x_{ji} = \begin{pmatrix} 0 & I_j \\ \overline{C}_{(T-j)p} & \end{pmatrix} y_i^0 + \eta_i \begin{pmatrix} 0 & 0 \\ \overline{C}_{T-j} & 0 \end{pmatrix} \begin{pmatrix} \iota_{T-j} \\ \iota_j \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ \overline{C}_{T-j} & 0 \end{pmatrix} \begin{bmatrix} \begin{pmatrix} v_{i1} \\ \vdots \\ v_{i(T-j)} \end{pmatrix} \\ \begin{pmatrix} v_{i(T-j+1)} \\ \vdots \\ v_{iT} \end{pmatrix} \end{bmatrix}$$

or

$$x_{ji} = C_{Tp}^j y_i^0 + \eta_i C_j \iota + C_j v_i \qquad (j = 1, ..., p) \tag{A.16}$$

where

$$C_j = \begin{pmatrix} 0 & 0 \\ \overline{C}_{T-j} & 0 \end{pmatrix} \qquad C_{Tp}^j = \begin{pmatrix} 0 & I_j \\ \overline{C}_{(T-j)p} & \end{pmatrix}.$$

Therefore,

$$E\left(x_{ji} v_i'\right) = C_j \Lambda, \tag{A.17}$$

and in view of the previous expression

$$E\left[X_i' D' \left(D \Lambda D'\right)^{-1} D v_i\right] = -\omega_T \begin{pmatrix} \iota' \Lambda^{-1} C_1 \iota \\ \vdots \\ \iota' \Lambda^{-1} C_p \iota \end{pmatrix} = -\begin{pmatrix} \varphi' C_1 \iota \\ \vdots \\ \varphi' C_p \iota \end{pmatrix}.$$

Moreover, note that weighted averages are given by

$$\overline{x}_{ji} = \varphi' x_{ji} = \eta_i \left(\varphi' C_j \iota\right) + \left(\varphi' C^j_{Tp}\right) y_i^0 + \varphi' C_j v_i \qquad (j = 1, ..., p) . \tag{A.18}$$

Also note that the variance of the average error can be eliminated to give rise to moment conditions that only depend on α and the weights.

Integral (42) of the Bias Function when $p = 1$:

To see that the integral of $h_T(\alpha, \varphi)$ when $p = 1$ is given by (42) 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

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

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 (A.2):

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

Proof of (53) and (54) for the Random Effects Scores:

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

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

Using this expression and (A.18) we have

$$\begin{aligned} \frac{1}{\sigma_\varepsilon^2} E\left[\overline{x}_i \left(\overline{u}_i - \phi' y_i^0\right)\right] &= \frac{1}{\sigma_\varepsilon^2} \left\{ E\left(\overline{x}_i \overline{v}_i\right) + E\left[\overline{x}_i \left(\eta_i - \phi' y_i^0\right)\right] \right\} \\ &= \frac{1}{\sigma_\varepsilon^2} \left[\omega_T^2 \Lambda^{-1} \iota' E\left(v_i X_i'\right) \Lambda^{-1} \iota + h_T(\alpha, \varphi) E\left(\eta_i \xi_i\right)\right] \\ &= \frac{1}{\sigma_\varepsilon^2} \left[\omega_T h_T(\alpha, \varphi) + h_T(\alpha, \varphi) Cov\left(\eta_i, \xi_i\right)\right] \\ &= h_T(\alpha, \varphi) \frac{1}{\sigma_\varepsilon^2} \left[Var\left(\overline{v}_i\right) + Var\left(\xi_i\right)\right] = h_T(\alpha, \varphi) . \end{aligned}$$

This proves result (53). Turning to (54), we have

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

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

A.4 A linear OLS Estimator of Variance Weights

A simple consistent estimator of the variance weights for given α can be obtained from the fact that $E\left(\overline{u}_i\Delta v_{it}\right)=0$. Such estimator may be useful for providing starting values for nonlinear likelihood-based estimation.

Enforcing the adding-up constraint, the average error can be written as

$$\begin{aligned}
\overline{u}_i &= \varphi_1u_{i1}+...+\varphi_{T-1}u_{i(T-1)}+\left(1-\varphi_1-...-\varphi_{T-1}\right)u_{iT}\\
&= u_{iT}-\varphi_1\left(u_{iT}-u_{i1}\right)-...-\varphi_{T-1}\left(u_{iT}-u_{i(T-1)}\right).
\end{aligned}\tag{A.19}$$

Letting $\varphi_o=\left(\varphi_1,...,\varphi_{T-1}\right)'$ and $w_i=\left[\left(u_{iT}-u_{i1}\right),...,\left(u_{iT}-u_{i(T-1)}\right)\right]'$, we have orthogonality between w_i and $\overline{u}_i$

$$E\left[w_i\left(u_{iT}-w_i'\varphi_o\right)\right]=0,\tag{A.20}$$

which suggests the following OLS estimator of φ_o for given α:

$$\widetilde{\varphi}_o=\left(\sum_{i=1}^N w_iw_i'\right)^{-1}\sum_{i=1}^N w_iu_{iT}.\tag{A.21}$$

This estimator satisfies the adding-up constraint, but not necessarily the non-negativity restrictions.

Given the $\widetilde{\varphi}_t$'s, estimates of ω_T and the σ_t^2's can be obtained from

$$\widetilde{\omega}_T = \frac{1}{(T-1)N}\sum_{i=1}^N\sum_{t=1}^T\widetilde{\varphi}_t\left(v_{it}-\overline{v}_i\right)^2\tag{A.22}$$

$$\widetilde{\sigma}_t^2 = \frac{\widetilde{\omega}_T}{\widetilde{\varphi}_t}.\tag{A.23}$$

B Asymptotic Variances of Estimators Under Normality

This Appendix presents the formulae for the asymptotic variances of RML and BCS estimators used for the inefficiency calculations reported in the main body of the paper. They are calculated under the assumption of normality for both homoskedastic and heteroskedastic estimators when $p = 1$. These formulas are not suggested for empirical standard error calculations (for which we use robust sample expressions that remain consistent under conditional heteroskedasticity and nonnormality), but in order to facilitate numerical comparisons of relative efficiency among alternative estimators.

B.1 Asymptotic Variance of the RML-dif Estimator

Letting $\eta_i^\dagger = \eta_i - (1-\alpha) y_{i0}$, the AR(1) 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,

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

where $\gamma = \left(\alpha, \sigma_1^2, ..., \sigma_T^2, \sigma^2_{\eta^\dagger}\right)'$.

Moreover, note that the heteroskedastic marginal MLE for the data in differences can be written as

$$\left(\widehat{\alpha}_D, \widehat{\sigma}_1^2 ..., \widehat{\sigma}_T^2, \widehat{\sigma}^2_{\eta^\dagger}\right) = \arg\min \left[\ln\det\Omega(\gamma) + \frac{1}{N} \sum_{i=1}^{N} (\Delta y_{i1}, ..., \Delta y_{iT}) \Omega^{-1}(\gamma) \begin{pmatrix} \Delta y_{i1} \\ \vdots \\ \Delta y_{iT} \end{pmatrix} \right].$$

Thus, under normality the asymptotic variance matrix of $\left(\widehat{\alpha}_D, \widehat{\sigma}_1^2 ..., \widehat{\sigma}_T^2, \widehat{\sigma}_{\eta\dagger}^2\right)$ is given by[23]

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

where

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

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

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

A similar expression is valid for the homoskedastic RML-dif estimator, except that in that case the parameter vector is redefined as $\gamma = \left(\alpha, \sigma^2, \sigma_{\eta\dagger}^2\right)'$.

B.2 Asymptotic Variance of the 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:

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

where $\gamma_{00} = Var\left(y_{i0}\right)$, $\gamma_{0\eta\dagger} = Cov\left(y_{i0}, \eta_i^\dagger\right)$, and $\gamma^* = \left(\alpha, \sigma_1^2, ..., \sigma_T^2, \sigma_{\eta\dagger}^2, \gamma_{0\eta\dagger}, \gamma_{00}\right)'$.

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

$$\left(\widehat{\alpha}_L, \widetilde{\sigma}_1^2, ..., \widetilde{\sigma}_T^2, \widetilde{\sigma}_{\eta\dagger}^2, \widetilde{\gamma}_{0\eta\dagger}, \widetilde{\gamma}_{00}\right) =$$

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

Thus, under normality the asymptotic variance matrix of $\left(\widehat{\alpha}_L, \widetilde{\sigma}_1^2, ..., \widetilde{\sigma}_T^2, \widetilde{\sigma}_{\eta\dagger}^2, \widetilde{\gamma}_{0\eta\dagger}, \widetilde{\gamma}_{00}\right)$ is given by

$$2\left\{H^*\left(\gamma^*\right)' \mathcal{D}^{*\prime} \left[\Omega^{*-1}\left(\gamma^*\right) \otimes \Omega^{*-1}\left(\gamma^*\right)\right] \mathcal{D}^* H^*\left(\gamma^*\right)\right\}^{-1} \tag{B.3}$$

where

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

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

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_η^2) given by

$$\gamma_{00} = \frac{\sigma_\eta^2}{\left(1-\alpha\right)^2} + \overline{\sigma}_0^2$$

and

$$\begin{aligned} \gamma_{0\eta\dagger} &\equiv Cov\left(y_{i0}, \eta_i^\dagger\right) = -\left(1-\alpha\right)\overline{\sigma}_0^2 \\ \sigma_{\eta\dagger}^2 &\equiv Var\left(\eta_i^\dagger\right) = \left(1-\alpha\right)^2 \overline{\sigma}_0^2, \end{aligned}$$

so that the restriction under mean stationarity is $\gamma_{0\eta\dagger}/\sigma_{\eta\dagger}^2 = -1/\left(1-\alpha\right)$. Homoskedasticity further restricts these coefficients to satisfy $\overline{\sigma}_0^2 = \sigma^2/\left(1-\alpha^2\right)$.

B.3 Asymptotic Variance of the Homoskedastic BCS 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 estimators that are consistent in view of the moment relationships:

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

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

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

where

$$\psi_i\left(\alpha\right) = x_i^{*\prime} v_i^* + v_i^{*\prime} v_i^* \frac{h_T\left(\alpha\right)}{\left(T-1\right)}. \tag{B.4}$$

Under suitable regularity conditions, if there is a consistent root of the equation $\sum_{i=1}^N \psi_i\left(a\right) = 0$,[24] its asymptotic variance is given by

$$v_\alpha = \frac{v}{d^2}. \tag{B.5}$$

where

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

and

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

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

Because of

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

we have

$$d = -E\left(x_i^{*\prime} x_i^*\right) + 2\sigma^2 \frac{h_T^2}{(T-1)} + \sigma^2 h_T' \tag{B.6}$$

where we are using h_T and h_T' for shortness.

Similarly,

$$v = E\left[\left(x_i^{*\prime} v_i^*\right)^2\right] + E\left[\left(v_i^{*\prime} v_i^*\right)^2\right] \frac{h_T^2}{(T-1)^2} + 2E\left[\left(x_i^{*\prime} v_i^*\right)\left(v_i^{*\prime} v_i^*\right)\right] \frac{h_T}{(T-1)}. \tag{B.7}$$

The availability of expression (B.1) allows us to calculate the term $E\left(x_i^{*\prime} x_i^*\right)$ that appears in (B.6) as follows

$$E\left(x_i^{*\prime} x_i^*\right) = E\left(x_i' D' \left(DD'\right)^{-1} D x_i\right) = tr\left[\left(DD'\right)^{-1} \Omega_{\Delta 11}\right] \tag{B.8}$$

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

Next, under normality and homoskedasticity we have

$$E\left[\left(x_i^{*\prime} v_i^*\right)^2\right] = \sigma^4 h_T^2 + \sigma^2 E\left(x_i^{*\prime} x_i^*\right) + \sigma^4 tr\left(Q C_T Q C_T\right) \tag{B.9}$$

$$E\left[\left(v_i^{*\prime} v_i^*\right)^2\right] = \sigma^4 (T+1)(T-1) \tag{B.10}$$

$$E\left[\left(x_i^{*\prime} v_i^*\right)\left(v_i^{*\prime} v_i^*\right)\right] = -\sigma^4 h_T (T+1) \tag{B.11}$$

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

Thus,

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

or

$$v = \sigma^2 E\left(x_i^{*\prime} x_i^*\right) + \sigma^4 tr\left(Q C_T Q C_T\right) - \frac{2}{(T-1)} \sigma^4 h_T^2. \tag{B.12}$$

To get the results (B.9)-(B.11) we have used the following intermediate formulae for moments of quadratic forms in normal variables:

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

B.4 Asymptotic Variance of the Heteroskedastic BCS Estimator

The i-th unit log-likelihood conditioned on the MLE of η_i and y_{i0} is given by

$$\ell_i = -\frac{1}{2}\ln\det\left(D\Lambda D'\right) - \frac{1}{2}v_i'D'\left(D\Lambda D'\right)^{-1}Dv_i$$

where D is the $(T-1)\times T$ first-difference matrix operator and $\Lambda = diag\left(\sigma_1^2, ..., \sigma_T^2\right)$. Also, let d_t be the t-th column of D, so that $D\Lambda D' = \sum_{t=1}^T \sigma_t^2 d_t d_t'$.

Using for shortness the notation $\Omega = D\Lambda D'$, the first and second derivatives of ℓ_i with respect to α and σ_t^2 are given by [25]

$$\frac{\partial \ell_i}{\partial \alpha} = x_i'D'\Omega^{-1}Dv_i$$

$$\frac{\partial \ell_i}{\partial \sigma_t^2} = \frac{1}{2}d_t'\Omega^{-1}\left(Dv_iv_i'D' - \Omega\right)\Omega^{-1}d_t \quad (t = 1, ..., T)$$

$$\frac{\partial^2 \ell_i}{\partial \alpha^2} = -x_i'D'\Omega^{-1}Dx_i$$

$$\frac{\partial^2 \ell_i}{\partial \sigma_t^2 \partial \alpha} = -d_t'\Omega^{-1}Dx_iv_i'D'\Omega^{-1}d_t \quad (t = 1, ..., T)$$

$$\frac{\partial^2 \ell_i}{\partial \sigma_t^2 \partial \sigma_s^2} = -\left(d_t'\Omega^{-1}d_s\right)\left(d_t'\Omega^{-1}Dv_iv_i'D'\Omega^{-1}d_s\right) + \frac{1}{2}\left(d_t'\Omega^{-1}d_s\right)^2.$$

Let $\ell_{1i} = \partial\ell_i/\partial\alpha$, $\ell_{2it} = \partial\ell_i/\partial\sigma_t^2$, $\ell_{11i} = \partial^2\ell_i/\partial\alpha^2$, etc., $\gamma = \left(\alpha, \sigma_1^2, ..., \sigma_T^2\right)'$, and $h = -E\left(\ell_{1i}\right)$, $h_1 = \partial h/\partial\alpha$, $h_{2t} = \partial h/\partial\sigma_t^2$. BCS is the GMM estimator based on the moments

$$\psi_i = \begin{pmatrix} \psi_{1i} \\ \psi_{2i} \end{pmatrix} = \begin{pmatrix} \ell_{1i} + h \\ \ell_{2i} \end{pmatrix}$$

whose asymptotic variance is

$$V_{BCS} = \left(D'V^{-1}D\right)^{-1}$$

where

$$D = E\left(\frac{\partial \psi_i}{\partial \gamma'}\right) = E\begin{pmatrix} \ell_{11i} & \ell_{12i} \\ \ell_{21i} & \ell_{22i} \end{pmatrix} + \begin{pmatrix} h_1 & h_2' \\ 0 & 0 \end{pmatrix}$$

and

$$V = E\left(\psi_i\psi_i'\right) = E\begin{pmatrix} \ell_{1i}^2 & \ell_{1i}\ell_{2i}' \\ \ell_{2i}\ell_{1i} & \ell_{2i}\ell_{2i}' \end{pmatrix} - \begin{pmatrix} h^2 & 0 \\ 0 & 0 \end{pmatrix}.$$

[25] Note that

$$\frac{\partial d_t'\Omega^{-1}d_t}{\partial \sigma_s^2} = -\left(d_t'\Omega^{-1}d_s\right)^2$$

and

$$\frac{\partial}{\partial \sigma_s^2}d_t'\Omega^{-1}\left(Dv_iv_i'D'\right)\Omega^{-1}d_t = -2\left(d_t'\Omega^{-1}d_s\right)\left(d_s'\Omega^{-1}Dv_iv_i'D'\Omega^{-1}d_t\right).$$

Letting $\Omega_{\Delta 11} = E\left(Dx_i x_i' D'\right)$, the expected second derivatives are

$$E\left(\ell_{11i}\right) \equiv E\left(\frac{\partial^2 \ell_i}{\partial \alpha^2}\right) = -tr\left(\Omega^{-1}\Omega_{\Delta 11}\right) \tag{B.13}$$

$$E\left(\ell_{21it}\right) \equiv E\left(\frac{\partial^2 \ell_i}{\partial \sigma_t^2 \partial \alpha}\right) = -d_t'\Omega^{-1}DC_1\Lambda D'\Omega^{-1}d_t \tag{B.14}$$

$$E\left(\ell_{22its}\right) \equiv E\left(\frac{\partial^2 \ell_i}{\partial \sigma_t^2 \partial \sigma_s^2}\right) = -\frac{1}{2}\left(d_t'\Omega^{-1}d_s\right)^2 \tag{B.15}$$

where $E\left(x_i v_i'\right) = C_1\Lambda$, and

$$C_1 = \begin{pmatrix} 0 & 0 \\ B_{T-1}^{-1} & 0 \end{pmatrix}.$$

Finally, the outer product terms are given by

$$\begin{aligned}
E\left(\ell_{1i}^2\right) &= E\left[\left(x_i'D'\Omega^{-1}Dv_i\right)^2\right] \\
E\left(\ell_{2it}\ell_{1i}\right) &= \frac{1}{2}E\left[\left(d_t'\Omega^{-1}Dv_i\right)^2\left(x_i'D'\Omega^{-1}Dv_i\right)\right] + \frac{1}{2}\left(d_t'\Omega^{-1}d_t\right)h \\
E\left(\ell_{2it}\ell_{2is}\right) &= \frac{1}{4}E\left[\left(d_t'\Omega^{-1}Dv_i\right)^2\left(d_s'\Omega^{-1}Dv_i\right)^2\right] - \frac{1}{4}\left(d_t'\Omega^{-1}d_t\right)\left(d_s'\Omega^{-1}d_s\right).
\end{aligned}$$

Under normality:

$$E\left(\ell_{1i}^2\right) = tr\left(\Omega^{-1}\Omega_{\Delta 11}\right) + tr\left(D'\Omega^{-1}DC_1\Lambda D'\Omega^{-1}DC_1\Lambda\right) + h^2 \tag{B.16}$$

$$E\left(\ell_{2it}\ell_{1i}\right) = d_t'\Omega^{-1}DC_1\Lambda D'\Omega^{-1}d_t \tag{B.17}$$

$$E\left(\ell_{2it}\ell_{2is}\right) = \frac{1}{2}\left(d_t'\Omega^{-1}d_s\right)^2 \tag{B.18}$$

Proof: Note that under normality:

$$\begin{aligned}
E\left[\left(d_t'\Omega^{-1}Dv_i\right)^2\left(d_s'\Omega^{-1}Dv_i\right)^2\right] &= E\left[\left(d_t'\Omega^{-1}Dv_i\right)^2\right]E\left[\left(d_s'\Omega^{-1}Dv_i\right)^2\right] \\
&\quad +2\left\{E\left[\left(d_t'\Omega^{-1}Dv_i\right)\left(d_s'\Omega^{-1}Dv_i\right)\right]\right\}^2 \\
&= \left(d_t'\Omega^{-1}d_t\right)\left(d_s'\Omega^{-1}d_s\right) + 2\left(d_t'\Omega^{-1}d_s\right)^2,
\end{aligned}$$

which proves (B.18) and also shows that $E\left(\ell_{2it}\ell_{2is}\right) = -E\left(\ell_{22its}\right)$.

To prove (B.16), let $v_i^* = \Omega^{-1/2}Dv_i$, $x_i^* = \Omega^{-1/2}Dx_i$ and note that

$$\begin{aligned}
E\left(\ell_{1i}^2\right) &= E\left[\left(x_i^{*\prime}v_i^*\right)^2\right] \\
&= \left[E\left(x_i^{*\prime}v_i^*\right)\right]^2 + tr\left[E\left(x_i^*x_i^{*\prime}\right)E\left(v_i^*v_i^{*\prime}\right)\right] + tr\left[E\left(x_i^*v_i^{*\prime}\right)E\left(x_i^*v_i^{*\prime}\right)\right] \\
&= h^2 + tr\left(\Omega^{-1}\Omega_{\Delta 11}\right) + tr\left(\Omega^{-1}DC_1\Lambda D'\Omega^{-1}DC_1\Lambda D'\right).
\end{aligned}$$

Finally, (B.17) can be proved as follows:

$$E\left[d_t'\Omega^{-1}Dv_iv_i'D'\Omega^{-1}d_t\left(x_i'D'\Omega^{-1}Dv_i\right)\right] = E\left[d_t'\Omega^{-1/2\prime}v_i^*v_i^{*\prime}\Omega^{-1/2}d_t\left(x_i^{*\prime}v_i^*\right)\right]$$

$$\begin{aligned}
&= E\left(d_t'\Omega^{-1/2\prime}v_i^*v_i^{*\prime}\Omega^{-1/2}d_t\right)E\left(x_i^{*\prime}v_i^*\right) + 2E\left(d_t'\Omega^{-1/2\prime}v_i^*x_i^{*\prime}\right)E\left(v_i^*v_i^{*\prime}\Omega^{-1/2}d_t\right) \\
&= -\left(d_t'\Omega^{-1}d_t\right)h + 2\left(d_t'\Omega^{-1}DC_1\Lambda D'\Omega^{-1}d_t\right)
\end{aligned}$$

To see this, letting $\widetilde{v}_{it} = d_t'\Omega^{-1/2\prime}v_i^*$, note that

$$E\left[d_t'\Omega^{-1/2\prime}v_i^*v_i^{*\prime}\Omega^{-1/2}d_t\left(x_i^{*\prime}v_i^*\right)\right] = \sum_s E\left(\widetilde{v}_{it}^2 x_{is}^* v_{is}^*\right)$$

$$\begin{aligned} &= \sum_s E\left(\widetilde{v}_{it}^2\right)E\left(x_{is}^*v_{is}^*\right) + 2\sum_s E\left(\widetilde{v}_{it}x_{is}^*\right)E\left(\widetilde{v}_{it}v_{is}^*\right) \\ &= E\left(d_t'\Omega^{-1/2\prime}v_i^*v_i^{*\prime}\Omega^{-1/2}d_t\right)E\left(x_i^{*\prime}v_i^*\right) + 2E\left(d_t'\Omega^{-1/2\prime}v_i^*x_i^{*\prime}\right)E\left(v_i^*v_i^{*\prime}\Omega^{-1/2}d_t\right). \end{aligned}$$

Thus, the information equality $E\left(\ell_{2it}\ell_{1i}\right) = -E\left(\ell_{21it}\right)$ also holds.

C Modified Conditional ML Score Interpretation of BCS

For the heteroskedastic AR(1) model we saw that BCS can be given a modified conditional likelihood interpretation when the weights φ are known. More generally, we show here that for a heteroskedastic AR(p) model with unknown weights, the BCS estimating equations coincide with the modified score vector discussed in Woutersen (2002), Arellano (2003b), and Arellano and Hahn (2007), which is first reviewed for convenience.

C.1 The Modified CML Score

Let $\ell_i(\beta, \eta_i)$ be an individual log-likelihood conditioned on z_i, and let $d_{\beta i}(\beta, \eta_i)$, $d_{\eta i}(\beta, \eta_i)$, $d_{\eta\eta i}(\beta, \eta_i)$ and $d_{\beta\eta i}(\beta, \eta_i)$ be first and second partial derivatives. The first argument is a vector common parameter β and η_i is a scalar individual effect. Let $\ell_i(\beta, \widehat{\eta}_i(\beta))$ be the concentrated log-likelihood, so that $d_{\beta i}(\beta, \widehat{\eta}_i(\beta))$ is the concentrated score.

The modified score discussed in Arellano (2003b) is given by

$$d_{Mi}(\beta) = d_{\beta i}(\beta, \widehat{\eta}_i(\beta)) - \frac{1}{2}\frac{\partial}{\partial\beta}\ln\left[-d_{\eta\eta i}(\beta, \widehat{\eta}_i(\beta))\right] + q_{\eta i}(\beta, \widehat{\eta}_i(\beta)) \tag{C.1}$$

where

$$q_{\eta i}(\beta, \eta_i) = \frac{\partial}{\partial\eta_i} q_i(\beta, \eta_i) \tag{C.2}$$

$$q_i(\beta, \eta_i) = \frac{\kappa_{\beta\eta i}(\beta, \eta_i)}{\kappa_{\eta\eta i}(\beta, \eta_i)} \tag{C.3}$$

and

$$\kappa_{\beta\eta i}(\beta_0, \eta_i) = E\left[\frac{1}{T} d_{\beta\eta i}(\beta_0, \eta_i) \mid x_i, \eta_i\right] \tag{C.4}$$

$$\kappa_{\eta\eta i}(\beta_0, \eta_i) = E\left[\frac{1}{T} d_{\eta\eta i}(\beta_0, \eta_i) \mid x_i, \eta_i\right]. \tag{C.5}$$

The first modification term provides a "degrees of freedom adjustment", whereas the second corrects for nonorthogonality between β and η_i. Note that if β and η_i are information orthogonal $\kappa_{\beta\eta i}(\beta, \eta_i) = 0$, so that $q_{\eta i}(\beta, \eta_i) = 0$ as well.

If there exists a scalar function $c_i(\beta, \eta_i)$ such that

$$\frac{\partial}{\partial\beta} c_i(\beta, \eta_i) = q_{\eta i}(\beta, \eta_i), \tag{C.6}$$

the modified score corresponds to the objective function

$$\ell_i(\beta, \widehat{\eta}_i(\beta)) - \frac{1}{2}\ln\left[-d_{\eta\eta i}(\beta, \widehat{\eta}_i(\beta))\right] + c_i(\beta, \eta_i), \tag{C.7}$$

which coincides with the Cox and Reid modified profile likelihood based on an orthogonal reparameterization of the effects. If $c_i(\beta, \eta_i)$ does not exist, there is no orthogonal reparameterization but the modified score $d_{Mi}(\beta)$ may still achieve bias reduction relative to $d_{\beta i}(\beta, \widehat{\eta}_i(\beta))$.

C.2 Application to AR(p) models

In the AR(p) model, $\beta = \left(\alpha', \theta'\right)'$, $z_i = y_i^0$, and

$$\ell_i\left(\beta,\eta_i\right) = -.5 \ln\det\Lambda - .5 v_i'\Lambda^{-1}v_i, \quad d_{\eta\eta i}\left(\beta,\eta_i\right) = -1/\omega_T,$$

$$d_{\beta\eta i}\left(\beta,\eta_i\right) = -\left(\frac{1}{\omega_T}\overline{x}_{1i},...,\frac{1}{\omega_T}\overline{x}_{pi},\frac{1}{\sigma_1^4}v_{i1},...,\frac{1}{\sigma_T^4}v_{iT}\right)'.$$

Thus, $\kappa_{\eta\eta i}\left(\beta,\eta_i\right) = -1/\left(T\omega_T\right)$,

$$\begin{aligned}\kappa_{\beta\eta i}\left(\beta_0,\eta_i\right) &= E\left[\frac{1}{T}d_{\beta\eta i}\left(\beta_0,\eta_i\right) \mid y_i^0,\eta_i\right] \\ &= -\frac{1}{T\omega_T}\left(E\left(\overline{x}_{1i} \mid y_i^0,\eta_i\right),...,E\left(\overline{x}_{pi} \mid y_i^0,\eta_i\right),0,...,0\right)',\end{aligned}$$

and

$$q_i\left(\beta,\eta_i\right) = \begin{pmatrix} \eta_i\left(\varphi' C_1\iota\right) + \left(\varphi' C_{Tp}^1\right)y_i^0 \\ \vdots \\ \eta_i\left(\varphi' C_p\iota\right) + \left(\varphi' C_{Tp}^p\right)y_i^0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad q_{\eta i}\left(\beta,\eta_i\right) = \begin{pmatrix} \varphi' C_1\iota \\ \vdots \\ \varphi' C_p\iota \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Therefore, the modified score vector is

$$d_{Mi}\left(\beta\right) = d_{\beta i}\left(\beta,\widehat{\eta}_i\left(\beta\right)\right) + \frac{1}{2}\frac{\partial}{\partial\beta}\ln\omega_T + \left(\varphi' C_1\iota,...,\varphi' C_p\iota,0,...,0\right)'$$

where

$$d_{\beta i}\left(\beta,\widehat{\eta}_i\left(\beta\right)\right) = \frac{\partial}{\partial\beta}\left[-\frac{1}{2}\ln\det\Phi - \frac{T}{2}\ln\omega_T - \frac{1}{2\omega_T}v_i'\left(\Phi - \Phi\iota\iota'\Phi\right)v_i\right],$$

which shows that BCS can be regarded as the solution to the estimating equations $\sum_{i=1}^N d_{Mi}\left(\beta\right) = 0$.

In this case it does not exist a function $c_i\left(\beta,\eta_i\right)$ such that

$$\left(\partial/\partial\beta\right)c_i\left(\beta,\eta_i\right) = \left(\varphi' C_1\iota,...,\varphi' C_p\iota,0,...,0\right)'.$$

This can be easily seen when $p = 1$. In that case $h_T\left(\alpha,\varphi\right) = \partial b_T\left(\alpha,\varphi\right)/\partial\alpha$ where $b_T\left(\alpha,\varphi\right) = \sum_{t=1}^{T-1}\left(\varphi_{t+1} + ... + \varphi_T\right)\alpha^t/t$, so that possible solutions for $c_i\left(\beta,\eta_i\right)$ would be of the form $b_T\left(\alpha,\varphi\right) + c\left(\theta\right)$. However, since $\partial b_T\left(\alpha,\varphi\right)/\partial\sigma_t^2$ depends on α and varies with t,[26] there is no $c\left(\theta\right)$ that can make $\partial c_i\left(\beta,\eta_i\right)/\partial\sigma_t^2$ equal to zero for any α and t as required.

Thus, in the heteroskedastic AR(p) setting, despite the lack of existence of an orthogonal transformation, a first-order bias adjustment to the score is an exact correction that removes fully the bias, hence leading to fixed-T consistent estimation.

[26] The expression is $\partial b_T\left(\alpha,\varphi\right)/\partial\sigma_t^2 = -\varphi_t^2\left[b_T\left(\alpha,\varphi\right) + \alpha + .. + \alpha^{t-1}\right]/\omega_T$.

D ARMA Models

Consider the model

$$y_{it} = \alpha_1 y_{i(t-1)} + ... + \alpha_p y_{i(t-p)} + \eta_i + v_{it} \quad (t = 1, ..., T; i = 1, ..., N) \tag{D.1}$$

where v_{it} is a moving average error of order q.

Following the notation introduced in (A.12), we can write

$$\begin{pmatrix} I_p & 0 \\ B_{Tp} & B_T \end{pmatrix} \begin{pmatrix} y_i^0 \\ y_i \end{pmatrix} = \begin{pmatrix} y_i^0 \\ \eta_i \iota + v_i \end{pmatrix}. \tag{D.2}$$

For an AR(p) process we have

$$Var \begin{pmatrix} y_i^0 \\ \eta_i \iota + v_i \end{pmatrix} = \begin{pmatrix} \Gamma_{00} & \gamma_{0\eta} \iota_T' \\ \iota_T \gamma_{0\eta}' & \sigma_\eta^2 \iota_T \iota_T' + \Lambda \end{pmatrix} \tag{D.3}$$

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

Similarly, for an ARMA(p, q) process

$$Var \begin{pmatrix} y_i^0 \\ \eta_i \iota + v_i \end{pmatrix} = \begin{pmatrix} \Gamma_{00} & \Upsilon_{pq} \quad \gamma_{0\eta} \iota_{T-q}' \\ \Upsilon_{pq}' & \\ \iota_{T-q} \gamma_{0\eta}' & \sigma_\eta^2 \iota_T \iota_T' + \Lambda_\psi \end{pmatrix}. \tag{D.4}$$

If $p \leq q$, the elements of Υ_{pq} are all unrestricted. However, if $p > q$ only the last q rows are unrestricted, and the $(p - q)$ first elements of the columns of Υ_{pq} coincide with those of $\gamma_{0\eta}$. Moreover, Λ_ψ is a moving average covariance matrix whose first q subdiagonals contain nonzero elements.

We adopt the following heteroskedastic moving-average specification for the errors in (D.1):

$$v_{it} = \sigma_t v_{it}^\dagger \tag{D.5}$$

$$v_{it}^\dagger = \zeta_{it} - \psi_1 \zeta_{i(t-1)} - ... - \psi_q \zeta_{i(t-q)} \tag{D.6}$$

where ζ_{it} is an $iid\,(0, 1)$ random error. In this way, we allow for arbitrary time series heteroskedasticity and at the same time specify a stationary serial correlation pattern for v_{it}. Thus,

$$v_i = \Lambda^{1/2} \Psi \left(\zeta_{i(1-q)}, ..., \zeta_{iT} \right)' \tag{D.7}$$

and

$$\Lambda_\psi = \Lambda^{1/2} \Psi \Psi' \Lambda^{1/2} \tag{D.8}$$

where Ψ is the $T \times (T + q)$ matrix

$$\Psi = \begin{pmatrix} -\psi_q & -\psi_{q-1} & \dots & -\psi_1 & 1 & 0 & \dots & 0 & \dots & 0 & 0 \\ 0 & -\psi_q & \dots & -\psi_2 & -\psi_1 & 1 & & 0 & \dots & 0 & 0 \\ \vdots & \vdots & & & & \ddots & \ddots & \ddots & & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & -\psi_q & \dots & -\psi_1 & 1 \end{pmatrix}.$$

Therefore, the covariance matrix of $y_i^T = \left(y_i^{0\prime}, y_i^{\prime}\right)^{\prime}$ is given by

$$\Omega\left(\gamma^*\right) = \left(\begin{array}{cc} I_p & 0 \\ B_{Tp} & B_T \end{array}\right)^{-1} \left(\begin{array}{ccc} \Gamma_{00} & \Upsilon_{pq} & \gamma_{0\eta}\iota_{T-q}^{\prime} \\ \Upsilon_{pq}^{\prime} & \sigma_\eta^2 \iota_T \iota_T^{\prime} + \Lambda_\psi & \\ \iota_{T-q}\gamma_{0\eta}^{\prime} & & \end{array}\right) \left(\begin{array}{cc} I_p & 0 \\ B_{Tp} & B_T \end{array}\right)^{-1\prime} \tag{D.9}$$

where the parameter vector γ^* consists of the autoregressive and moving average coefficients, $\gamma_{0\eta}, \sigma_\eta^2, \sigma_1^2, ..., \sigma_T^2$, and the unrestricted elements in Γ_{00} and Υ_{pq}.

The ARMA(p, q) log-likelihood is given by

$$L_{RS} = -\frac{N}{2} \ln\det\Omega\left(\gamma^*\right) - \frac{1}{2}\sum\nolimits_{i=1}^{N} y_i^{T\prime}\Omega\left(\gamma^*\right)^{-1} y_i^T. \tag{D.10}$$

Noting that

$$\det\left(\begin{array}{cc} I_p & 0 \\ B_{Tp} & B_T \end{array}\right) = 1,$$

and letting $u_i = \eta_i\iota + v_i$, $\Omega_{11} = \sigma_\eta^2\iota\iota^{\prime} + \Lambda_\psi$, $\Gamma_{01} = \left(\begin{array}{cc} \Upsilon_{pq} & \gamma_{0\eta}\iota_{T-q}^{\prime} \end{array}\right)$, and

$$\left(\begin{array}{cc} \Gamma_{00} & \Gamma_{01} \\ \Gamma_{01}^{\prime} & \Omega_{11} \end{array}\right)^{-1} = \left(\begin{array}{cc} \Gamma^{00} & \Gamma^{01} \\ \Gamma^{01\prime} & \Omega^{11} \end{array}\right),$$

where

$$\Omega_{11}^{-1} = \Omega^{11} - \Gamma^{01\prime}\left(\Gamma^{00}\right)^{-1}\Gamma^{01} \tag{D.11}$$

$$\det\Omega\left(\gamma^*\right) = \left(\det\Omega_{11}\right) / \left(\det\Gamma^{00}\right), \tag{D.12}$$

we have

$$\begin{aligned} y_i^{T\prime}\Omega\left(\gamma^*\right)^{-1} y_i^T &= \left(y_i^{0\prime}, u_i^{\prime}\right)\left(\begin{array}{cc} \Gamma^{00} & \Gamma^{01} \\ \Gamma^{01\prime} & \Omega^{11} \end{array}\right)\left(\begin{array}{c} y_i^0 \\ u_i \end{array}\right) \\ &= u_i^{\prime}\Omega_{11}^{-1}u_i + \left(y_i^0 + \left(\Gamma^{00}\right)^{-1}\Gamma^{01}u_i\right)^{\prime}\Gamma^{00}\left(y_i^0 + \left(\Gamma^{00}\right)^{-1}\Gamma^{01}u_i\right). \end{aligned} \tag{D.13}$$

Therefore, letting $\Psi_{00} = \left(\Gamma^{00}\right)^{-1}$ and $\Pi_{01} = -\left(\Gamma^{00}\right)^{-1}\Gamma^{01} = \Gamma_{01}\Omega_{11}^{-1}$, we obtain the following expression for L_{RS}:

$$\begin{aligned} L_{RS} &= -\frac{N}{2}\ln\det\Omega_{11} - \frac{1}{2}\sum_{i=1}^{N} u_i^{\prime}\Omega_{11}^{-1}u_i \\ &\quad -\frac{N}{2}\ln\det\Psi_{00} - \frac{1}{2}\sum_{i=1}^{N}\left(y_i^0 - \Pi_{01}u_i\right)^{\prime}\Psi_{00}^{-1}\left(y_i^0 - \Pi_{01}u_i\right). \end{aligned} \tag{D.14}$$

Concentrating the likelihood with respect to Ψ_{00} (which is unrestricted), we get

$$L_{RS}^* = -\frac{N}{2}\ln\det\Omega_{11} - \frac{1}{2}\sum_{i=1}^{N} u_i^{\prime}\Omega_{11}^{-1}u_i - \frac{N}{2}\ln\det\sum_{i=1}^{N}\left(y_i^0 - \Pi_{01}u_i\right)\left(y_i^0 - \Pi_{01}u_i\right)^{\prime}, \tag{D.15}$$

which we found computationally very useful.

References

Abowd, J.M. and D. Card (1989): “On the Covariance Structure of Earnings and Hours Changes”, *Econometrica*, 57, 411–445.

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

Ahn, S. C. and G. M. Thomas (2006): “Likelihood Based Inference for Dynamic Panel Data Models”. Unpublished manuscript.

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

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. and B. Honoré (2001): “Panel Data Models: Some Recent Developments”. In J. J. Heckman and E. Leamer (eds.): *Handbook of Econometrics*, Vol. 5, Chapter 53, North-Holland, 3229–3296.

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

Arellano, M. (2003b): “Discrete Choices with Panel Data”, *Investigaciones Económicas*, 27, 423–458.

Arellano, M. and J. Hahn (2007): “Understanding Bias in Nonlinear Panel Models: Some Recent Developments”. In: R. Blundell, W. Newey, and T. Persson (eds.): *Advances in Economics and Econometrics, Ninth World Congress*, Vol. III, Cambridge University Press, 381–409.

Arellano, M. and S. Bonhomme (2009): “Robust Priors in Nonlinear Panel Data Models”, *Econometrica*, 77, 489–536.

Arellano, M., R. Blundell, and S. Bonhomme (2017): “Earnings and Consumption Dynamics: A Nonlinear Panel Data Framework”, *Econometrica*, 85, 693–734.

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

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.

Browning, M., M. Ejrnæs, and J. Alvarez (2010): “Modelling Income Processes with Lots of Heterogeneity”, *Review of Economic Studies*, 77, 1353–1381.

Chamberlain, G. (1980): "Analysis of Covariance with Qualitative Data", *Review of Economic Studies*, 47, 225–238.

Chamberlain, G. (1984): "Panel Data". In Z. Griliches and M. D. Intriligator (eds.): *Handbook of Econometrics*, Vol. 2, Chapter 22, North Holland, 1247–1318.

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

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.

Dhaene, G. and K. Jochmans (2016): "Likelihood Inference in an Autoregression with Fixed Effects", *Econometric Theory*, 32, 1178–1215.

Gourieroux, C., A. Monfort, and A. Trognon (1984): "Pseudo Maximum Likelihood Methods: Theory", *Econometrica*, 52, 681–700.

Hahn, J. and G. Kuersteiner (2002): "Asymptotically Unbiased Inference for a Dynamic Panel Model with Fixed Effects When Both n and T are Large", *Econometrica*, 70, 1639–1657.

Hall, R. and F. Mishkin (1982): "The Sensitivity of Consumption to Transitory Income: Estimates from Panel Data on Households", *Econometrica*, 50, 461–481.

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.

Horowitz, J. L. and M. Markatou (1996): "Semiparametric Estimation of Regression Models for Panel Data", *Review of Economic Studies*, 63, 145–168.

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.

MaCurdy, T. E. (1982): "The Use of Time Series Processes to Model the Error Structure of Earnings in a Longitudinal Data Analysis", *Journal of Econometrics*, 18, 83–114.

Meghir, C. and L. Pistaferri (2004): "Income Variance Dynamics and Heterogeneity", *Econometrica*, 72, 1–32.

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

Sargan, J. D. (1983): "Identification and Lack of Identification", *Econometrica*, 51, 1605–1634.

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

Woutersen, T. (2002): "Robustness Against Incidental Parameters," Unpublished manuscript.

Table 1

Relative Inefficiency Ratios*

		Homosk.		Heterosk.	
		BCS	RMLdif	BCS	RMLdif
			$\alpha = 0.6$		
$T^o = 4$	$\lambda = 0$	1.45	1.33	2.21	1.59
	$\lambda = 1$	1.14	1.05	1.56	1.12
$T^o = 10$	$\lambda = 0$	1.06	1.04	1.07	1.05
	$\lambda = 1$	1.02	1.00	1.03	1.00
			$\alpha = 0.8$		
$T^o = 4$	$\lambda = 0$	1.93	1.70	3.16	2.15
	$\lambda = 1$	1.22	1.07	1.69	1.15
$T^o = 10$	$\lambda = 0$	1.22	1.13	1.28	1.15
	$\lambda = 1$	1.08	1.01	1.12	1.01

*Ratios of Asymptotic St.Deviations: Denominator is St.Dev. of RML-lev; T^o =no. of waves; $\lambda = \sigma_\eta^2/\sigma^2$.

Table 2

Autoregressive Model of Earnings

AR(1) Estimates for Spanish Data, 1994-1999

$N = 632, T^0 = 6$

	WG	GMM1	GMM2	System-GMM
α	-0.022	0.042	0.038	0.183
	(-0.95)	(0.93)	(0.87)	(7.00)
Sargan test (df)			6.11(9)	22.71(13)
$m1$		-9.67	-9.89	-13.73
$m2$		0.27	0.23	1.83

Likelihood-based Estimates

	BCS	RML(r)	RML(nr)	RML(r)	RML(r)
	(robust)	(robust)	(homosk.)	(mean stat.)	$(\phi = 0)$
α	0.218	0.200	0.207	0.164	0.926
	(7.04)	(7.07)	(3.83)	(5.32)	(87.05)
σ_1^2 (1995)	0.025	0.023	0.023	0.023	0.049
	(11.34)	(11.91)	(25.14)	(11.65)	(12.81)
σ_2^2 (1996)	0.022	0.021		0.021	0.042
	(8.55)	(9.28)		(9.04)	(14.40)
σ_3^2 (1997)	0.023	0.023		0.023	0.039
	(8.23)	(9.55)		(9.16)	(15.96)
σ_4^2 (1998)	0.023	0.023		0.022	0.039
	(10.26)	(10.60)		(10.47)	(14.74)
σ_5^2 (1999)	0.023	0.025		0.025	0.047
	(10.93)	(11.63)		(11.51)	(14.80)
ϕ		0.567	0.560	0.607	$0.^{\dagger}$
		(18.27)	(11.72)	(15.05)	
σ_ε^2		0.020	0.020	$0.024^{\dagger}$	0.003
		(10.37)	(7.53)		(9.77)
γ_{00}		0.111		0.100	
		(14.35)		(16.13)	

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 $(\overline{y} - \alpha\overline{y}_{-1})$ on y_0. †Implied by constraint.

Table 2 (continued)

Autoregressive Model of Earnings

AR(2) Estimates for Spanish Data, 1994-1999

$N = 632, T^0 = 6$

	WG	GMM1	GMM2	System-GMM
α_1	-0.131	0.112	0.138	0.311
	(5.06)	(1.20)	(1.58)	(7.91)
α_2	-0.118	0.051	0.070	0.176
	(3.78)	(0.93)	(1.41)	(4.87)
Sargan test (df)			4.21 (7)	16.02 (11)
$m1$		-6.41	-7.02	-11.56
$m2$		-0.75	-0.87	-1.55

Likelihood-based Estimates

	BCS	RML(r)	RML(nr)	RML(r)	RML(r)
	(robust)	(robust)	(homosk.)	(mean stat.)	($\phi = 0$)
α_1	0.218	0.201	0.210	0.300	0.600
	(4.47)	(4.89)	(2.73)	(4.69)	(25.40)
α_2	0.104	0.094	0.100	0.102	0.338
	(2.57)	(2.47)	(1.35)	(2.16)	(15.90)
σ_1^2 (1996)	0.022	0.022	0.023	0.026	0.037
	(7.93)	(8.69)	(25.14)	(7.17)	(11.90)
σ_2^2 (1997)	0.025	0.024		0.026	0.035
	(7.34)	(9.15)		(8.84)	(13.59)
σ_3^2 (1998)	0.023	0.023		0.024	0.033
	(8.85)	(9.87)		(10.10)	(12.85)
σ_4^2 (1999)	0.024	0.024		0.034	0.035
	(10.68)	(11.34)		(6.13)	(13.04)
ϕ_1		0.253	0.247		0.
		(5.39)	(5.50)		
ϕ_2		0.334	0.326		0.
		(6.51)	(6.12)		
σ_ε^2		0.016	0.015		0.005
		(8.24)	(7.57)		(12.62)
Root_1	0.450	0.424	0.437	0.503	0.954
Root_2	-0.232	-0.223	-0.228	-0.203	-0.354

Table 3

Autoregressive Model of Earnings

AR(1) Estimates 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)	(3.37)	(3.54)	(9.76)	
Sargan test (df)			15.61 (14)	46.59 (19)	
$m1$		-6.36	-6.40	-7.42	
$m2$		1.82	1.64	2.36	
			Likelihood-based Estimates		
	BCS	RML(r)	RML(nr)	RML(r)	RML(r)
	(robust)	(robust)	(homosk.)	(mean stat.)	($\phi = 0$)
α	0.387	0.367	0.416	0.366	0.902
	(9.64)	(10.09)	(8.27)	(10.04)	(43.93)
σ_1^2 (1978)	0.061	0.059	0.068	0.059	0.113
	(7.73)	(7.83)	(28.14)	(7.83)	(10.14)
σ_2^2 (1979)	0.062	0.058		0.058	0.085
	(6.10)	(6.08)		(6.07)	(8.73)
σ_3^2 (1980)	0.054	0.052		0.052	0.079
	(7.21)	(7.55)		(7.54)	(9.02)
σ_4^2 (1981)	0.046	0.046		0.046	0.080
	(6.62)	(7.41)		(7.40)	(8.79)
σ_5^2 (1982)	0.094	0.096		0.096	0.114
	(3.55)	(3.68)		(3.67)	(4.66)
σ_6^2 (1983)	0.086	0.091		0.091	0.132
	(5.34)	(5.31)		(5.31)	(6.97)
ϕ		0.385	0.352	0.384	0.†
		(11.84)	(8.35)	(11.75)	
σ_ε^2		0.045	0.042	0.046†	0.008
		(9.35)	(7.55)		(6.43)
γ_{00}		0.239		0.237	
		(12.92)		(13.34)	

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. †Value implied by constraint.

Table 3 (continued)

Autoregressive Model of Earnings

AR(2) Estimates for PSID Data, 1977-1983

$N = 792, T^0 = 7$

	WG	GMM1	GMM2	System-GMM
α_1	0.135	0.227	0.250	0.433
	(3.61)	(2.75)	(3.37)	(11.03)
α_2	−0.028	0.047	0.062	0.119
	(0.90)	(1.17)	(1.81)	(3.93)
Sargan test (df)			12.29 (12)	30.96 (17)
$m1$		−4.94	−5.47	−7.05
$m2$		2.19	1.79	1.45

Likelihood-based Estimates

	BCS	RML(r)	RML(nr)	RML(r)	RML(r)
	(robust)	(robust)	(homosk.)	(mean stat.)	($\phi = 0$)
α_1	0.473	0.419	0.496	0.518	0.673
	(5.29)	(8.32)	(5.49)	(8.79)	(18.30)
α_2	0.157	0.115	0.176	0.159	0.260
	(2.78)	(3.14)	(2.55)	(3.56)	(8.26)
σ_1^2 (1979)	0.070	0.064	0.076	0.071	0.082
	(4.84)	(6.19)	(28.14)	(7.00)	(8.52)
σ_2^2 (1980)	0.061	0.056		0.063	0.074
	(5.50)	(7.48)		(8.34)	(9.32)
σ_3^2 (1981)	0.057	0.051		0.059	0.072
	(5.21)	(7.01)		(7.88)	(8.36)
σ_4^2 (1982)	0.092	0.097		0.102	0.109
	(3.69)	(3.71)		(3.91)	(4.23)
σ_5^2 (1983)	0.091	0.090		0.096	0.108
	(4.88)	(5.28)		(5.68)	(6.47)
ϕ_1		0.096	0.065		0.
		(2.95)	(2.06)		
ϕ_2		0.262	0.174		0.
		(4.73)	(3.23)		
σ_ε^2		0.028	0.023		0.012
		(7.23)	(5.65)		(8.38)
Root_1	0.698	0.607	0.736	0.735	0.947
Root_2	−0.225	−0.189	−0.240	−0.217	−0.274

Table 4

ARMA Models of Earnings

RML Estimates for PSID Data, 1977-1983

$N = 792$, $T^0 = 7$

	ARMA$(1,1)$	ARMA$(1,2)$	ARMA$(2,1)$
α_1	0.655	0.336	0.210
	(3.34)	(1.74)	(0.32)
α_2			0.194
			(0.24)
ψ_1	0.205	-0.068	-0.175
	(1.69)	(0.41)	(0.18)
ψ_2		-0.139	
		(2.62)	
σ^2_{1978}	0.069	0.057	
	(4.74)	(6.63)	
σ^2_{1979}	0.063	0.062	0.065
	(6.69)	(6.11)	(0.48)
σ^2_{1980}	0.057	0.056	0.056
	(6.50)	(7.85)	(1.43)
σ^2_{1981}	0.055	0.049	0.048
	(4.80)	(6.43)	(1.52)
σ^2_{1982}	0.094	0.099	0.096
	(3.63)	(3.68)	(2.87)
σ^2_{1983}	0.093	0.092	0.089
	(5.32)	(5.36)	(4.68)
σ^2_η	0.021	0.073	0.064
	(0.90)	(1.75)	(1.20)

Table 5

Simulations for the First-Order Autoregressive Model

Means and standard deviations of the estimators

$N = 792, T^0 = 7$

	WG	GMM	RML(nr)	RML(r)	BCS
True values: $\alpha = 0.4$, $\overline{\sigma}_0^2 = 0.11$					
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)	(0.003)
σ_4^2				0.046	0.046
				(0.003)	(0.003)
σ_5^2				0.096	0.096
				(0.005)	(0.006)
σ_6^2				0.091	0.091
				(0.005)	(0.005)
True values: $\alpha = 0.8$, $\overline{\sigma}_0^2 = 0.28$					
α	0.488	0.772	0.882	0.804	0.804
	(0.016)	(0.076)	(0.028)	(0.037)	(0.040)
σ_1^2				0.059	0.059
				(0.004)	(0.004)
σ_2^2				0.058	0.058
				(0.004)	(0.004)
σ_3^2				0.052	0.052
				(0.004)	(0.004)
σ_4^2				0.046	0.046
				(0.003)	(0.003)
σ_5^2				0.096	0.096
				(0.005)	(0.006)
σ_6^2				0.091	0.091
				(0.005)	(0.005)

1000 replications. Variance values: $\sigma_1^2 = 0.059, \sigma_2^2 = 0.058$, $\sigma_3^2 = 0.052, \sigma_4^2 = 0.046, \sigma_5^2 = 0.096, \sigma_6^2 = 0.091, \sigma_\eta^2 = 0.07$.

Table A1

Heteroskedasticity-consistent likelihood-based estimators of Autoregressive Panel Models

	Data in levels	Data in differences
Unrestricted initial conditions	RML	BCS RML-dif
Imposing mean stationarity	RML-s	Conditional ML $\equiv$ RML-dif

RML: random effects maximum likelihood.

BCS: bias-corrected conditional score.

RML-dif: random effects ML in first-differences and conditional ML under mean stationarity.

RML-s: random effects ML under mean stationarity.

Table A2

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 A3

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

Table A4

Autoregressive Model of Earnings

AR(2) Estimates for Spanish Data, 1994-1999

$N = 632, T^0 = 6$

"Robust GMM" form of the estimates with RML

adding extra moment conditions to BCS

	GMM	RML	BCS
α_1	0.204	0.201	0.218
	(4.76)	(4.89)	(4.47)
α_2	0.098	0.094	0.104
	(2.55)	(2.47)	(2.57)
σ_1^2 (1996)	0.022	0.022	0.022
	(8.84)	(8.69)	(7.93)
σ_2^2 (1997)	0.024	0.024	0.025
	(9.39)	(9.15)	(7.34)
σ_3^2 (1998)	0.023	0.023	0.023
	(10.02)	(9.87)	(8.85)
σ_4^2 (1999)	0.024	0.024	0.024
	(11.27)	(11.34)	(10.68)
ϕ_1	0.251	0.253	
	(5.43)	(5.39)	
ϕ_2	0.329	0.334	
	(6.37)	(6.51)	
σ_ε^2	0.016	0.016	
	(8.24)	(8.24)	
Sargan test (p-value)	0.83 (0.84)		

GMM uses the $2(T+2)$ moments (56)-(60) in Section 4.

RML and BCS are as in Table 2 included here for convenience.

Table A5

Autoregressive Model of Earnings

AR(2) Estimates for PSID Data, 1977-1983

$N = 792, T^0 = 7$

"Robust GMM" form of the estimates with RML adding extra moment conditions to BCS

	GMM	RML	BCS
α_1	0.475	0.419	0.473
	(7.38)	(8.32)	(5.29)
α_2	0.134	0.115	0.157
	(3.34)	(3.14)	(2.78)
σ_1^2 (1979)	0.054	0.064	0.07
	(9.34)	(6.19)	(4.84)
σ_2^2 (1980)	0.052	0.056	0.061
	(8.84)	(7.48)	(5.50)
σ_3^2 (1981)	0.051	0.051	0.057
	(7.88)	(7.01)	(5.21)
σ_4^2 (1982)	0.075	0.097	0.092
	(4.47)	(3.71)	(3.69)
σ_5^2 (1983)	0.086	0.090	0.091
	(5.65)	(5.28)	(4.88)
ϕ_1	0.084	0.096	
	(2.61)	(2.95)	
ϕ_2	0.212	0.262	
	(3.25)	(4.73)	
σ_ε^2	0.023	0.028	
	(4.91)	(7.23)	
Sargan test (p-value)	10.78 (0.03)		

See notes to Table A3.

Figure 1
Relative Inefficiency Ratio ($\lambda = 0$)
Homoskedastic Estimators

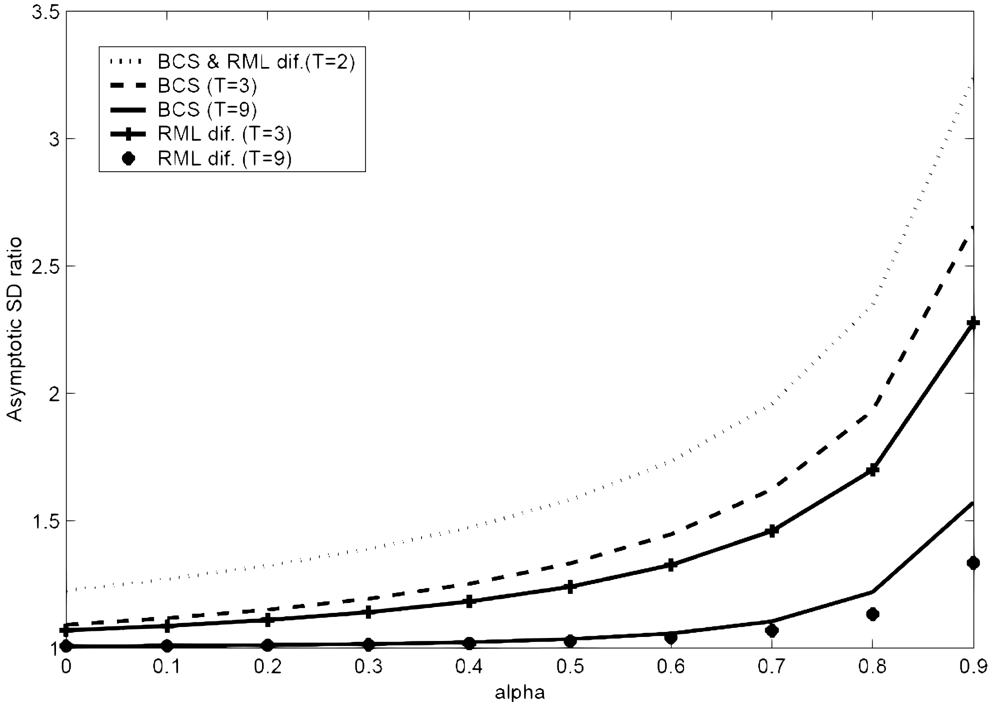


Figure 2
Relative Inefficiency Ratio ($\lambda = 1$)
Homoskedastic Estimators

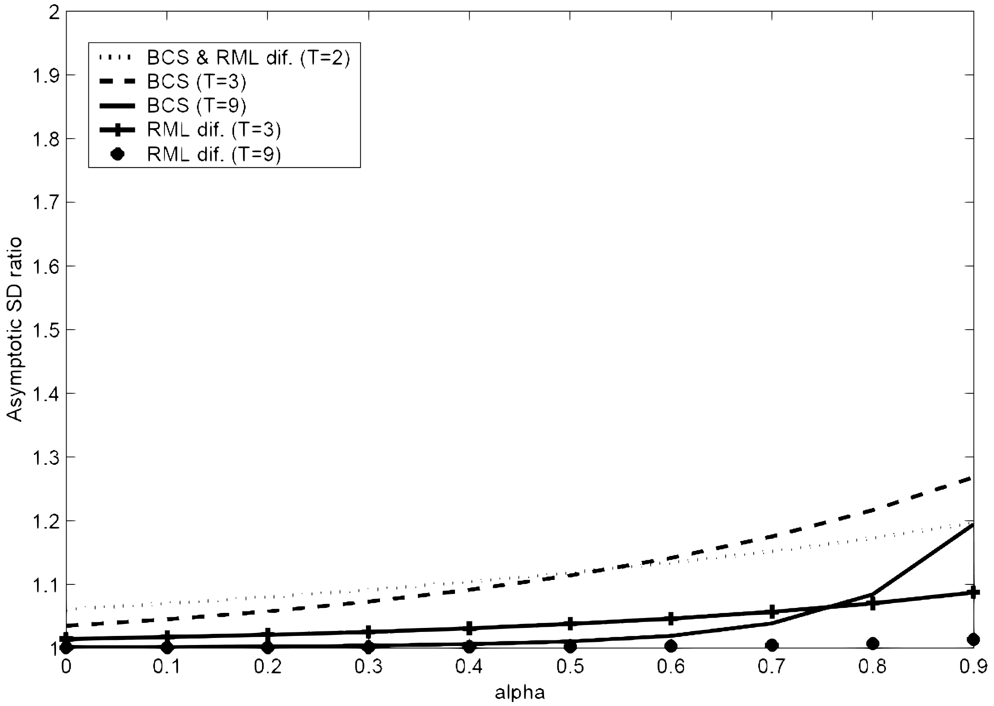


Figure 3
Relative Inefficiency Under Nonstationary Initial Variance ($T = 3, \alpha = 0.9, \lambda = 0$)
Homoskedastic Estimators

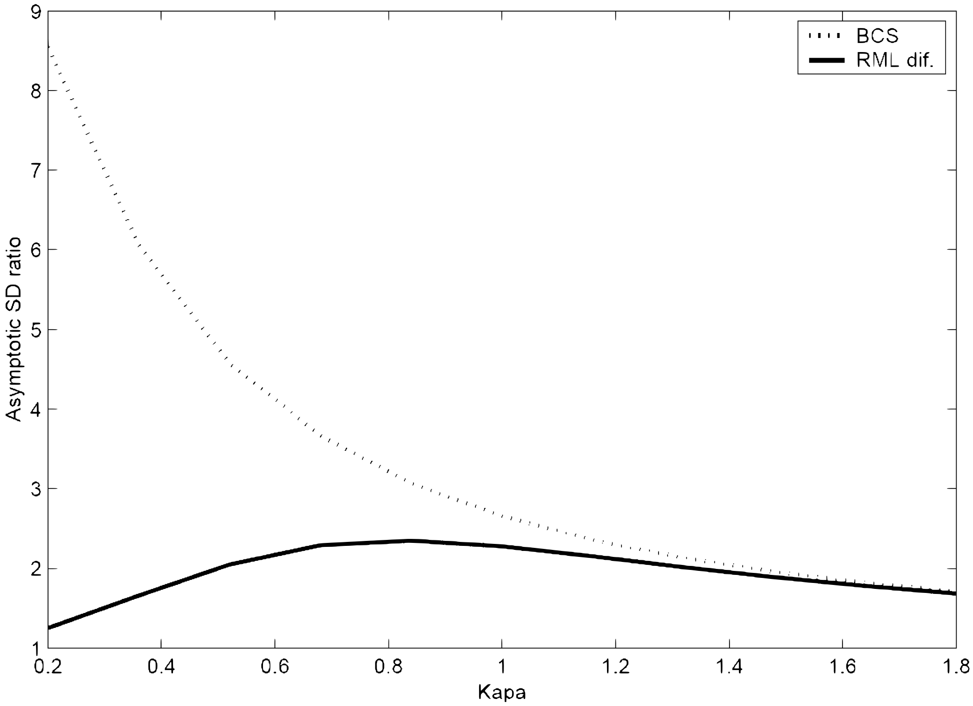


Figure 4
Relative Inefficiency Ratio ($\lambda = 0$)
Heteroskedastic Estimators

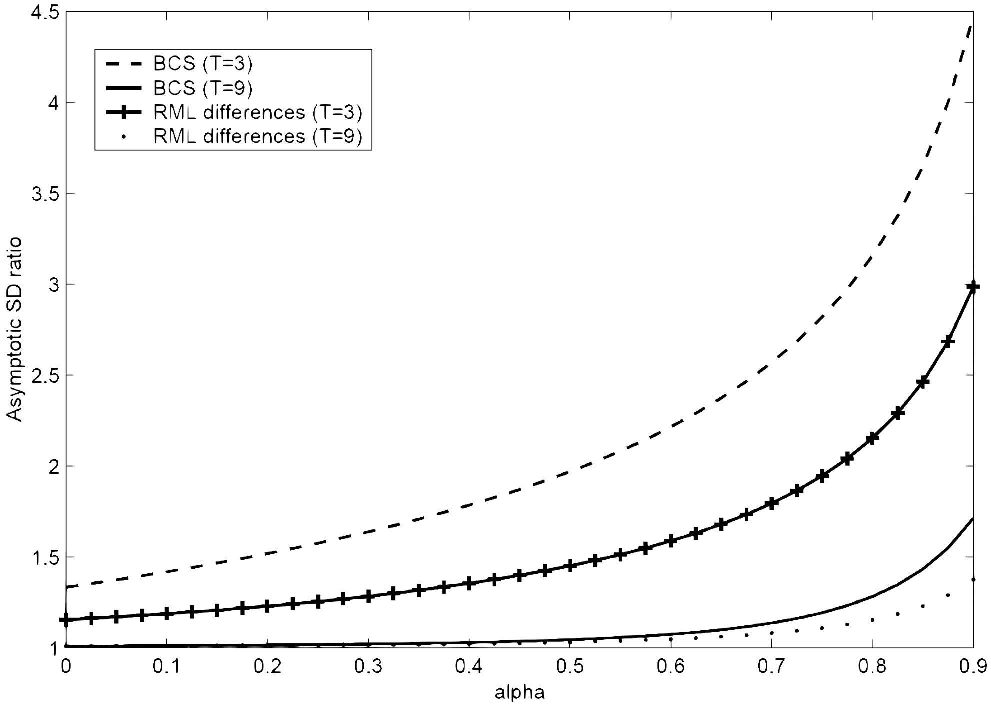


Figure 5
Relative Inefficiency Ratio ($\lambda = 1$)
Heteroskedastic Estimators

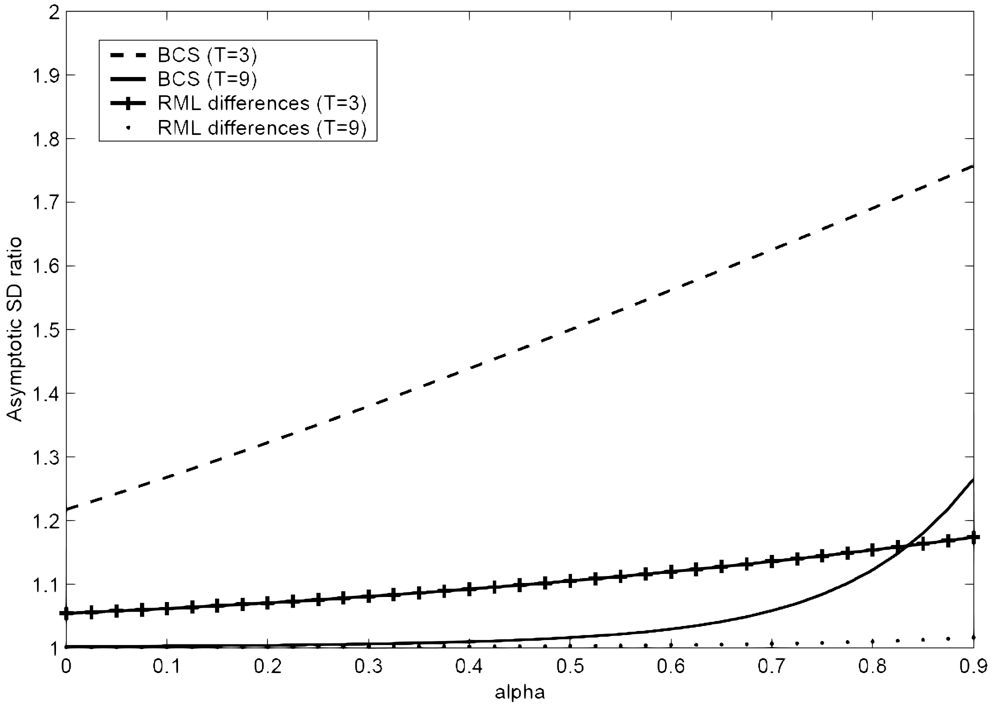


Figure 6
Asymptotic Standard Deviation Under Unit Root ($T = 6, \alpha = 1, \lambda = 0$)
By Location and Change in Single Break in Variances

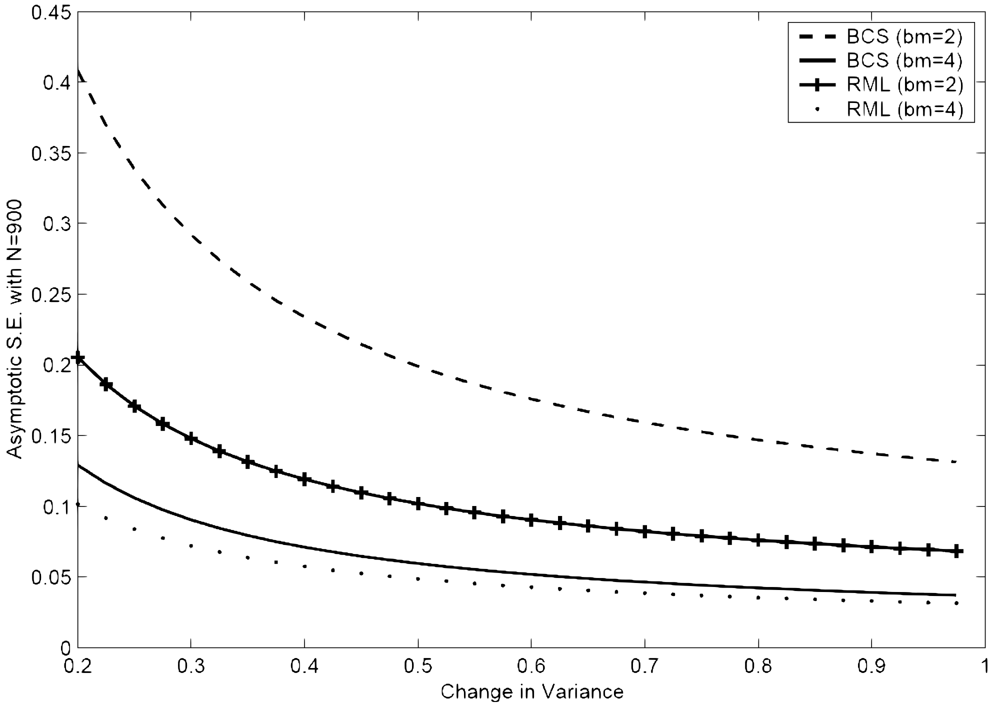


Figure 7
Distribution of Residuals in First Differences

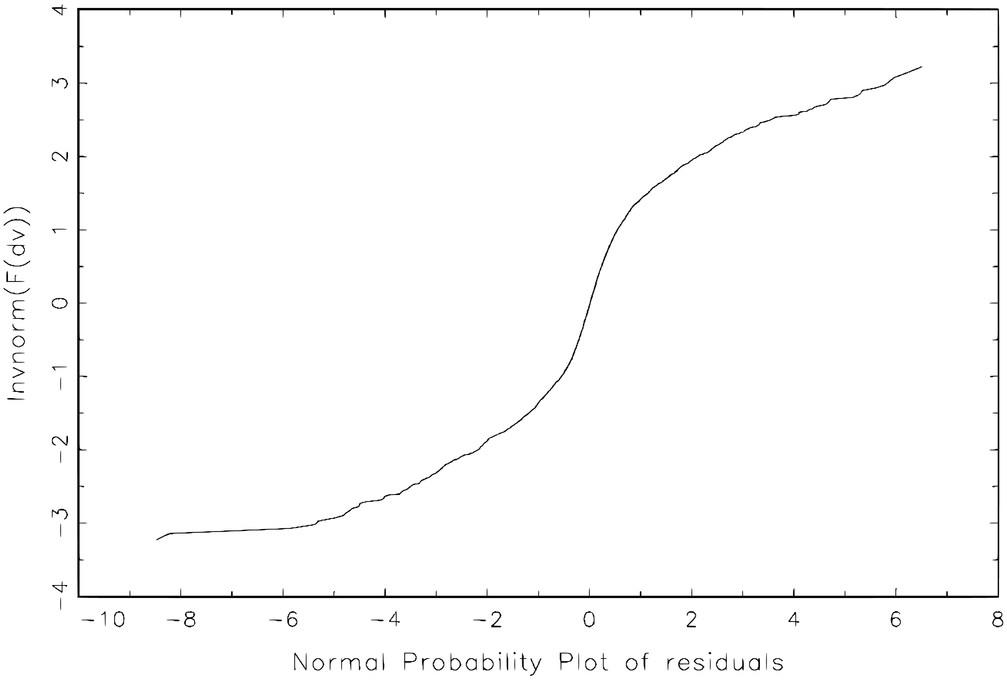
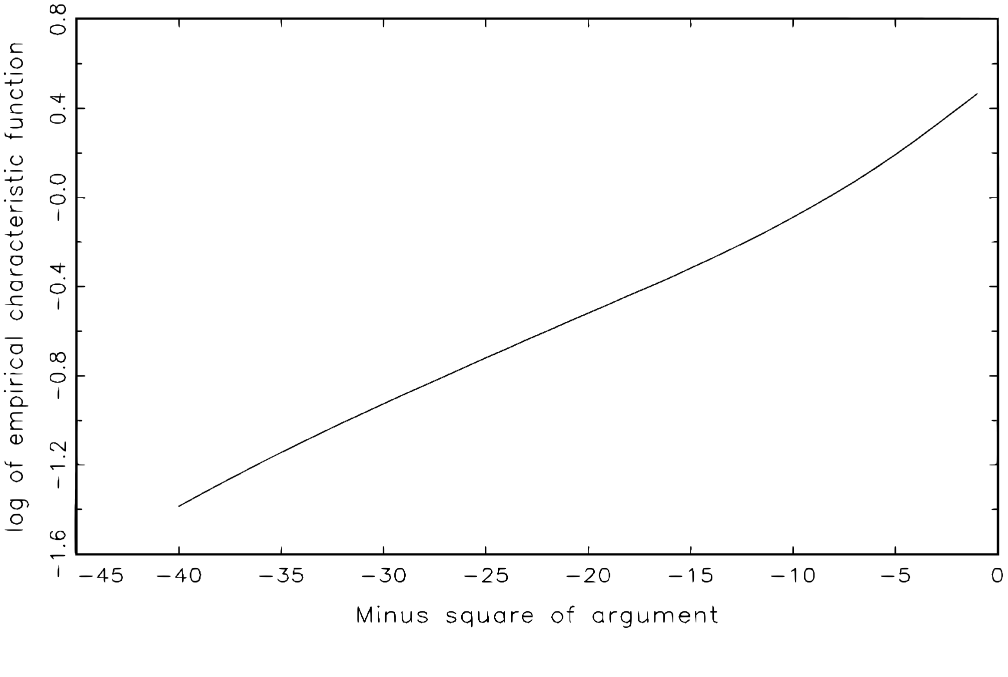


Figure 8
Graphical Test of Normality of Individual Effects


log of empirical characteristic function
Minus square of argument