

An Adjusted Profile Likelihood for Non-Stationary Panel Data Models with Incidental Parameters

Geert Dhaene* Koen Jochmans
K.U.Leuven K.U.Leuven

15 March 2007

Abstract

We calculate the exact bias of the profile score for the first-order autoregressive parameter, ρ , in a Gaussian $N \times T$ panel data model with arbitrary initial conditions and arbitrary heterogeneity in intercepts, trends and error variances. The bias is a polynomial in ρ and does not depend on the initial values or the incidental parameters. Subtracting its integral from the profile loglikelihood leads to an adjusted profile likelihood which, in the case without incidental trends and error variances, coincides with Lancaster's (2002) marginal posterior density for ρ . We show, largely by simulation, that the expected adjusted profile loglikelihood (and hence the expected marginal posterior log-density), in addition to attaining a local maximum on $[-1, 1]$ at the true value of ρ , may attain a global maximum at 1. The latter occurs when the initial values are strong inliers relative to the stationary distribution, which leads to weakly informative data when the autoregressive parameter is moderate to large, even with very large N .

JEL: C13, C22

Keywords: Fixed effects, incidental parameters, dynamic panel, heterogeneity, panel unit root, bias correction, profile score

*Corresponding author. Address: K.U.Leuven, Department of Economics, Naamsestraat 69, B-3000 Leuven, Belgium. Email: geert.dhaene@econ.kuleuven.be.

1 Introduction

The dominant mode of inference in dynamic panel models with fixed effects is moment-based. GMM estimators defined by appropriate moment conditions (Anderson and Hsiao, 1981, 1982; Arellano and Bond, 1991; Ahn and Schmidt, 1995) are fixed- T consistent, regardless of the way the initial observations are generated. Unlike GMM, the maximum likelihood estimator in general (Neyman and Scott, 1948), and the least-squares estimator in particular (Nickell, 1981), are biased and inconsistent as $N \rightarrow \infty$ with T fixed. The bias depends on the relationship between the initial observations and the fixed effects. Assumptions regarding this relationship, such as stationarity of the initial observations, are often delicate and may seriously bias the results if incorrectly imposed. Therefore, it is on the safe side to conduct inference conditional on the initial observations. With short panels this implies allowing for non-stationarity and precludes, for example, simple patches of the least-squares estimator based on the inversion of Nickell's (1981) bias formula. For these reasons, GMM is currently the most widely applied method of inference in dynamic panels with fixed effects.

In an interesting paper, Lancaster (2002) argued in favour of likelihood-based analysis of fixed effects models. His (Bayesian) resolution of the incidental parameters problem in the non-stationary AR(1) model consists of two steps. First, he reparameterises the fixed effects so as to make them orthogonal (in the information sense) to the common parameters. Second, he integrates the orthogonalised fixed effects from the (Gaussian) likelihood, using a uniform prior. The first-order conditions for the mode of the resulting marginal posterior are unbiased estimating equations for the common parameters. Lancaster notes that Cox and Reid's (1987) approximate conditional likelihood coincides with the marginal posterior (see also Arellano, 2003). Alvarez and Arellano (2004) also adopt a likelihood perspective in a fixed effects setting. They follow Cox and Reid to derive bias-corrected score equations in a model featuring time series heteroskedasticity, with qualitatively similar results as Lancaster, including fixed- T consistency.

In this paper, we give a third interpretation to Lancaster’s result. We arrive at it through a route that, unlike the other approaches, does not involve parameter orthogonalisation or a conditioning argument as in the Cox and Reid approach. Motivated by a desire to bias-correct the score associated with the profile (or concentrated) loglikelihood, we calculate the bias of the profile score for the autoregressive parameter, ρ . The (finite N, T) bias turns out to be a polynomial of degree $T - 2$ in ρ . Surprisingly, the bias is independent of the initial observations (and also of the fixed effects); the coefficients in the polynomial depend only on T . Subtracting the bias from the profile score provides an unbiased estimating equation that coincides with Lancaster’s equation locating the marginal posterior mode for ρ . Subtracting the integral of the bias from the profile loglikelihood yields an adjusted profile loglikelihood which then, of course, coincides with Lancaster’s marginal posterior log-density for ρ . Our approach to adjust the profile loglikelihood follows, in part, McCullagh and Tibshirani (1990) who, in addition to re-centring the profile score, rescale it to make it information unbiased. We do not carry out the latter step, because it is essentially equivalent to using a sandwich form for the variance of the estimator. Our result coincides with Alvarez and Arellano’s (2004) in the case of homoskedasticity and follows, as in their analysis, from a bias calculation. The two approaches differ in that we calculate the bias of the profile score for ρ (only), while they calculate the bias of the conditional score for all common parameters, given maximum likelihood estimates of the fixed effects. The results coincide for Gaussian likelihoods, but not in general.

We also derive the proposed profile likelihood adjustments for models with incidental trends and/or error variances. Incidental trends alter the adjustment, but the essential features remain. The bias of the profile score is still a polynomial of degree $T - 2$ in ρ , with coefficients depending only on T . Incidental error variances leave the adjustments unchanged although, of course, they alter the profile likelihood. In particular, the (unadjusted and adjusted) profile likelihoods now feature a polynomial of degree $2N$, which makes it somewhat more difficult to locate the maximum. Presumably the

adjusted profile likelihoods in these models can also be obtained via Lancaster’s Bayesian approach and via Cox and Reid’s approximate conditional likelihood approach.

We subsequently return to the model without incidental trends and error variances. In view of the identical results obtained from three different angles, it is tempting to speculate that the adjusted profile loglikelihood (or, equivalently, the marginal posterior log-density and the approximate conditional loglikelihood) is a genuine pseudo-loglikelihood, in the sense that its expectation attains a unique global maximum at the true parameter value, ρ_0 . We show, by simulation, that this is not true in general. It appears that the expected adjusted profile loglikelihood can only have two local maxima on $[-1, 1]$. There is, as formally shown by Lancaster, always a local maximum at $\rho = \rho_0$. In addition, there may be a second local maximum at $\rho = 1$, which may even be the global maximum. This suggests that ρ_0 is identified from the expected adjusted profile loglikelihood curve, but *not necessarily as its global maximiser*. If correct, this is a rather puzzling state of affairs, which demands theoretical explanation. Moreover, it presents numerical challenges as one may need to look for a *local* optimum of the adjusted profile loglikelihood rather than the global one. At present we don’t have a theoretical answer to offer, but make some observations as to why the problem of weak identification emerges.

The model is presented in Section 2. In its most general form, it features fixed effects, incidental trends and incidental error variances. Submodels are considered, always including fixed effects and possibly in combination with incidental trends and/or incidental error variances. The adjustment to the profile likelihood, which is based on a finite N, T bias calculation of the profile score for ρ , is developed in Section 3. Section 4 addresses the large N global properties of the adjusted profile likelihood. These properties lead to the definition of the estimator for ρ_0 . An appendix contains some algebra and figures.

2 The model

Suppose we observe a scalar variable y_{it} ($i = 1, \dots, N$; $t = 0, \dots, T$) which, given y_{i0} and $\lambda_i = (\alpha_i, \beta_i, \sigma_i^2)$, is generated by

$$y_{it} = \alpha_i + \beta_i t + \rho y_{it-1} + \varepsilon_{it}, \quad \varepsilon_{it} \sim N(0, \sigma_i^2),$$

where the errors, ε_{it} , are independent across i and t . Our interest lies in estimating the autoregressive parameter, whose true value will be denoted by ρ_0 . We assume $-1 \leq \rho_0 \leq 1$ and allow for arbitrary initial observations. Specifically, the pairs (y_{i0}, λ_i) are assumed to be drawn independently across i from a distribution G which is essentially unconstrained. We distinguish between four cases: incidental trends (where $\beta_i = 0$ if $\rho_0 = 1$) or no incidental trends (where $\beta_i = 0$ and, if $\rho_0 = 1$, $\alpha_i = 0$); and incidental error variances (where $\sigma_i^2 > 0$) or no incidental error variances (where $\sigma_i^2 = \sigma^2 > 0$). All cases allow for arbitrary dependence between y_{i0} and λ_i . For example, y_{i0} given λ_i may be degenerate.

We shall use the following notation: $y_i = (y_{i1}, \dots, y_{iT})'$, $y_{i-} = (y_{i0}, \dots, y_{iT-1})'$, $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{iT})'$ and

$$M = \begin{cases} I_T - \iota_T(\iota_T' \iota_T)^{-1} \iota_T' & \text{in cases without incidental trends,} \\ I_T - J_T(J_T' J_T)^{-1} J_T' & \text{in cases with incidental trends,} \end{cases} \quad (1)$$

where $J_T = [\iota_T : \tau_T]$, $\iota_T = (1, \dots, 1)'$ and $\tau_T = (1, \dots, T)'$. This allows us to write $M y_i = \rho M y_{i-} + M \varepsilon_i$ in any case. The conditional distribution of y_i , given y_{i0}, λ_i , will be denoted as H_0 . Expectations, $E_0(\cdot)$, are taken with respect to $H_0 \times G$.

3 Adjustment of the profile likelihood

Let $l(\rho)$ and $s(\rho) = dl(\rho)/d\rho$ be the profile loglikelihood function and the profile score function, normalised by the number of observations. The presence of the incidental parameters leads, in this model, to an asymptotically¹ biased profile score equation, $s(\rho) = 0$. That is, $\lim_{N \rightarrow \infty} E_0[s(\rho_0)] \neq 0$. Hence

¹Throughout the paper, the asymptotics are large N , fixed T asymptotics.

$\rho_0 \neq \arg \max_{\rho \in [-1,1]} \lim_{N \rightarrow \infty} E_0[l(\rho)]$ and the maximum likelihood estimate, solving $s(\rho) = 0$, is inconsistent. This difficulty with likelihood-based inference in short dynamic panels remained unresolved until Lancaster (2002).

Our approach to solving the incidental parameters problem in this model starts by calculating the finite N, T bias of the profile score,

$$B(\rho_0) = E_0[s(\rho_0)],$$

in order to re-centre (or adjust) the profile score function, through

$$s_A(\rho) = s(\rho) - B(\rho).$$

By construction, the re-centred profile score function delivers an unbiased estimating equation, $s_A(\rho) = 0$. That is, $E_0[s_A(\rho_0)] = 0$. This estimating equation, moreover, turns out to be conditionally unbiased, in the sense that $E_0[s_A(\rho_0) | \{y_{i0}, \lambda_i\}_{i=1}^N] = 0$. Subsequently, we integrate the re-centred profile score function to obtain an adjusted profile loglikelihood function,

$$l_A(\rho) = \int s_A(\rho) d\rho = l(\rho) - \int B(\rho) d\rho.$$

Estimation of ρ will be considered in the next section, on inspecting the properties of the functions $s_A(\rho)$ and $l_A(\rho)$. We now derive these functions for the model defined in the previous section, starting with the case considered by Lancaster.

3.1 No incidental trends, no incidental error variances

The loglikelihood function, conditional on $\{y_{i0}\}_{i=1}^N$ and normalised by the number of observations, is

$$-\frac{1}{2NT} \sum_{i=1}^N \sum_{t=1}^T \left[\log \sigma^2 + \frac{(y_{it} - \alpha_i - \rho y_{it-1})^2}{\sigma^2} \right] + c$$

where, here and later, c is an inessential constant. Replacing $\alpha_1, \dots, \alpha_N$ and σ^2 with their maximum likelihood estimates for any fixed value of ρ gives the

(normalised) profile loglikelihood function,

$$l(\rho) = -\frac{1}{2} \log \left(\sum_{i=1}^N (y_i - \rho y_{i-})' M (y_i - \rho y_{i-}) \right) + c. \quad (2)$$

The profile score function is

$$s(\rho) = \frac{dl(\rho)}{d\rho} = \frac{\sum_{i=1}^N (y_i - \rho y_{i-})' M y_{i-}}{\sum_{i=1}^N (y_i - \rho y_{i-})' M (y_i - \rho y_{i-})}. \quad (3)$$

We now calculate the bias of the profile score, $B(\rho_0) = E_0[s(\rho_0)]$. Let $\gamma = (1, \rho_0, \dots, \rho_0^{T-1})'$ and

$$\Gamma = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ \rho_0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_0^{T-3} & \rho_0^{T-4} & \rho_0^{T-5} & \cdots & 0 & 0 \\ \rho_0^{T-2} & \rho_0^{T-3} & \rho_0^{T-4} & \cdots & 1 & 0 \end{pmatrix}.$$

Then $y_{i-} = \Gamma \varepsilon_i + \zeta_{i0}$, where $\zeta_{i0} = \Gamma \iota_T \alpha_i + \gamma y_{i0}$. Hence we can write

$$\begin{aligned} s(\rho_0) &= \frac{\sum_{i=1}^N \varepsilon_i' M y_{i-}}{\sum_{i=1}^N \varepsilon_i' M \varepsilon_i} \\ &= \frac{\sum_{i=1}^N \varepsilon_i' M \Gamma M \varepsilon_i}{\sum_{i=1}^N \varepsilon_i' M \varepsilon_i} + \frac{\sum_{i=1}^N \varepsilon_i' M \Gamma (I_T - M) \varepsilon_i}{\sum_{i=1}^N \varepsilon_i' M \varepsilon_i} + \frac{\sum_{i=1}^N \varepsilon_i' M \zeta_{i0}}{\sum_{i=1}^N \varepsilon_i' M \varepsilon_i}. \end{aligned}$$

The last two terms of $s(\rho_0)$ have zero expectation, because $(I_T - M)\varepsilon_i$ and $M\varepsilon_i$ are independent, ε_i and ζ_{i0} are independent, and $\varepsilon_i' M / \sum_{i=1}^N \varepsilon_i' M \varepsilon_i$ is symmetrically distributed around zero. With regard to the first term of $s(\rho_0)$, using an argument that goes back to Fisher (1930) and Geary (1933), we show that

$$E \left(\frac{\sum_{i=1}^N \varepsilon_i' M \Gamma M \varepsilon_i}{\sum_{i=1}^N \varepsilon_i' M \varepsilon_i} \right) = \frac{E(\varepsilon_i' M \Gamma M \varepsilon_i)}{E(\varepsilon_i' M \varepsilon_i)}.$$

Transform the vector ε_i into its average, $T^{-1} \iota_T' \varepsilon_i$, and the radius, $(\varepsilon_i' M \varepsilon_i)^{1/2}$, and $T - 2$ polar angles of $M \varepsilon_i$. All of these—the mean, the radius, and the

polar angles—are independent. Hence, the ratio $\varepsilon'_i M \Gamma M \varepsilon_i / \varepsilon'_i M \varepsilon_i$, being a function of the polar angles only (regardless of Γ), is independent of $\varepsilon'_i M \varepsilon_i$. Therefore,

$$E(\varepsilon'_i M \Gamma M \varepsilon_i) = E\left(\frac{\varepsilon'_i M \Gamma M \varepsilon_i}{\varepsilon'_i M \varepsilon_i} \varepsilon'_i M \varepsilon_i\right) = E\left(\frac{\varepsilon'_i M \Gamma M \varepsilon_i}{\varepsilon'_i M \varepsilon_i}\right) E(\varepsilon'_i M \varepsilon_i)$$

and so

$$E\left(\frac{\varepsilon'_i M \Gamma M \varepsilon_i}{\varepsilon'_i M \varepsilon_i}\right) = \frac{E(\varepsilon'_i M \Gamma M \varepsilon_i)}{E(\varepsilon'_i M \varepsilon_i)}.$$

In fact, the last equality holds generally when $\varepsilon_i \sim N(0, \sigma^2)$ and M is symmetric and idempotent.² Therefore, letting $\varepsilon = (\varepsilon'_1, \dots, \varepsilon'_N)'$,

$$\begin{aligned} E\left(\frac{\sum_{i=1}^N \varepsilon'_i M \Gamma M \varepsilon_i}{\sum_{i=1}^N \varepsilon'_i M \varepsilon_i}\right) &= E\left(\frac{\varepsilon'(I_T \otimes M \Gamma M)\varepsilon}{\varepsilon'(I_T \otimes M)\varepsilon}\right) = \frac{E(\varepsilon'(I_T \otimes M \Gamma M)\varepsilon)}{E(\varepsilon'(I_T \otimes M)\varepsilon)} \\ &= \frac{E\left(\sum_{i=1}^N \varepsilon'_i M \Gamma M \varepsilon_i\right)}{E\left(\sum_{i=1}^N \varepsilon'_i M \varepsilon_i\right)} = \frac{E(\varepsilon'_i M \Gamma M \varepsilon_i)}{E(\varepsilon'_i M \varepsilon_i)}. \end{aligned}$$

Hence the bias of the profile score is

$$B(\rho_0) = \frac{\text{tr}(\Gamma M)}{\text{tr}(M)}. \quad (4)$$

It will turn out that this expression holds for all cases considered. In the current case, $M = I_T - \nu_T(\nu'_T \nu_T)^{-1} \nu'_T$ and

$$B(\rho_0) = -\frac{1}{T(T-1)} \sum_{t=1}^{T-1} (T-t) \rho_0^{t-1}. \quad (5)$$

The bias of the profile score is related to the bias of the maximum likelihood (here, least-squares) estimator, calculated by Nickell (1981). On rewriting $B(\rho_0)$ as

$$B(\rho_0) = -\frac{1}{(T-1)(1-\rho_0)} \left[1 - \frac{1}{T} \left(\frac{1-\rho_0^T}{1-\rho_0} \right) \right] \quad \text{if } -1 \leq \rho_0 < 1,$$

²For an excellent discussion and historical perspective on this device, see Conniffe and Spencer (2001).

we recognise the numerator of the Nickell bias as $(T - 1)B(\rho_0) = \text{tr}(\Gamma M)$. There are two key differences between $B(\rho_0)$ and the Nickell bias. First, unlike the Nickell bias, which is derived under stationarity of the initial observations, $B(\rho_0)$ is independent of the relationship between the initial values and the fixed effects. Put differently, because $B(\rho_0)$ does not depend on G and G is allowed to be degenerate, the conditional bias of the profile score, given *any* initial values and fixed effects, is also $B(\rho_0)$. That is, $E_0[s(\rho_0) | \{y_{i0}, \lambda_i\}_{i=1}^N] = B(\rho_0)$. Second, the Nickell bias concerns a probability limit as $N \rightarrow \infty$, whereas (5) is a finite sample result. It holds for any $T \geq 2$ and $N \geq 1$, and hence may be of independent interest in a time series setting.

The re-centred profile score function and the adjusted loglikelihood function are

$$s_A(\rho) = \frac{\sum_{i=1}^N (y_i - \rho y_{i-})' M y_{i-}}{\sum_{i=1}^N (y_i - \rho y_{i-})' M (y_i - \rho y_{i-})} + \frac{1}{T(T-1)} \sum_{t=1}^{T-1} (T-t) \rho^{t-1},$$

$$l_A(\rho) = -\frac{1}{2} \log \left(\sum_{i=1}^N (y_i - \rho y_{i-})' M (y_i - \rho y_{i-}) \right) + \frac{1}{T(T-1)} \sum_{t=1}^{T-1} \left(\frac{T-t}{t} \right) \rho^t + c.$$

$l_A(\rho)$ is identical to Lancaster's marginal posterior log-density for ρ , and $s_A(\rho) = 0$ is the first-order condition for its posterior mode. Notice that $s_A(\rho) = 0$ is a conditionally unbiased estimating equation, given $\{y_{i0}, \lambda_i\}_{i=1}^N$. Lancaster shows that $E_0[l_A(\rho)]$ attains a strict local maximum at ρ_0 .

3.2 Incidental trends, no incidental error variances

In the case with incidental trends, the profile loglikelihood and score functions are still given by (2) and (3), but with M redefined according to (1). The calculation of the bias of the profile score proceeds along the same lines as above. In particular, (4) still holds. After some algebra, summarised in the

Appendix, we find

$$\begin{aligned} B(\rho_0) &= -\frac{\text{tr}(J_T' \Gamma J_T (J_T' J_T)^{-1})}{T-2} \\ &= -\frac{2}{T(T-2)} \sum_{j=1}^{T-1} \left(T-j - j \frac{T^2-j^2}{T^2-1} \right) \rho_0^{j-1}. \end{aligned} \quad (6)$$

Phillips and Sul (2007) calculate the bias of the least-squares estimator for this case, assuming stationary initial observations. On rewriting $B(\rho_0)$ as

$$B(\rho_0) = -\frac{2}{(T-2)(1-\rho_0)} \left[1 - \frac{2}{T-1} \left(\frac{C}{1-\rho_0} \right) \right] \quad \text{if } -1 \leq \rho_0 < 1,$$

where

$$C = 1 - \frac{1}{T+1} \left(1 + \frac{1}{T} \frac{1-\rho_0^3}{(1-\rho_0)^3} \right) + \left(\frac{1}{2} + \frac{1}{T+1} \left(\frac{1+2\rho_0}{1-\rho_0} + \frac{1}{T} \frac{1-\rho_0^3}{(1-\rho_0)^3} \right) \right) \rho_0^T,$$

we see that the numerator of Philips and Sul's formula is $(T-2)B(\rho_0) = \text{tr}(\Gamma M)$. Again, $B(\rho_0)$ is the conditional bias of the profile score, given $\{y_{i0}, \lambda_i\}_{i=1}^N$, a result that holds for all $T \geq 3$ and $N \geq 1$.

The re-centred profile score function and the adjusted loglikelihood function are

$$\begin{aligned} s_A(\rho) &= \frac{\sum_{i=1}^N (y_i - \rho y_{i-})' M y_{i-}}{\sum_{i=1}^N (y_i - \rho y_{i-})' M (y_i - \rho y_{i-})} \\ &\quad + \frac{2}{T(T-2)} \sum_{j=1}^{T-1} \left(T-j - j \frac{T^2-j^2}{T^2-1} \right) \rho^{j-1}, \\ l_A(\rho) &= -\frac{1}{2} \log \left(\sum_{i=1}^N (y_i - \rho y_{i-})' M (y_i - \rho y_{i-}) \right) \\ &\quad + \frac{2}{T(T-2)} \sum_{j=1}^{T-1} \left(\frac{T-j}{j} - \frac{T^2-j^2}{T^2-1} \right) \rho^j + c. \end{aligned}$$

As before, $s_A(\rho) = 0$ is a conditionally unbiased estimating equation, given $\{y_{i0}, \lambda_i\}_{i=1}^N$.

3.3 Incidental error variances

In the cases with incidental error variances, the loglikelihood function is

$$-\frac{1}{2NT} \sum_{i=1}^N \sum_{t=1}^T \left[\log(\sigma_i^2) + \frac{(y_{it} - \alpha_i - \beta_i t - \rho_0 y_{it-1})^2}{\sigma_i^2} \right] + c,$$

with the understanding that $\beta_i = 0$ when there are no incidental trends. The profile loglikelihood and profile score functions are

$$l(\rho) = -\frac{1}{2N} \sum_{i=1}^N \log [(y_i - \rho y_{i-})' M (y_i - \rho y_{i-})] + c,$$

$$s(\rho) = \frac{1}{N} \sum_{i=1}^N \frac{(y_i - \rho y_{i-})' M y_{i-}}{(y_i - \rho y_{i-})' M (y_i - \rho y_{i-})}.$$

It is easily seen that the bias of the profile score is still given by (4). Hence (5) and (6) remain valid in the presence of incidental error variances. As before, $B(\rho_0)$ equals the conditional bias of the profile score, given $\{y_{i0}, \lambda_i\}_{i=1}^N$, for all $T \geq 2$ (without incidental trends) or $T \geq 3$ (with incidental trends) and $N \geq 1$. The re-centred profile score function, $s_A(\rho)$, and the adjusted profile loglikelihood function, $l_A(\rho)$, follow immediately. The estimating equation $s_A(\rho) = 0$ is conditionally unbiased, given $\{y_{i0}, \lambda_i\}_{i=1}^N$.

It should be remarked that the presence of incidental error variances causes $l_A(\rho)$ and its derivatives to be less smooth when T is very small. In particular, $l_A(\rho)$ may be unbounded or have multiple local maxima. The problem is not due to the adjustment of the profile loglikelihood, as it occurs to $l(\rho)$ alike and the adjustment term is smooth. To understand what happens, notice that $l(\rho)$ depends on ρ through $P(\rho) = \prod_{i=1}^N (y_i - \rho y_{i-})' M (y_i - \rho y_{i-})$, a polynomial of degree $2N$ in ρ . When $T = 2$ (in the case without incidental trends) or $T = 3$ (in the case with incidental trends), this polynomial is zero at all points $y'_{i-} M y_i / y'_{i-} M y_{i-}$, resulting in an unbounded likelihood at those points. When $T = 3$ or $T = 4$ (without and with incidental trends, resp.), the probability of an unbounded likelihood is zero, but $P(\rho)$ often has multiple local minima, implying $l(\rho)$ has multiple local maxima, a feature that carries

over to $l_A(\rho)$. Simulations with very large N indicate that the multiplicity of local maxima does not disappear asymptotically. Noting that the estimation problem is, in essence, one of estimating a location parameter from normal data with one degree of freedom, there appears to be a connection with the fact that the asymptotic number of local but not global maxima of a Cauchy likelihood for a location parameter is a Poisson variate (Reeds, 1985). This needs to be investigated more thoroughly, but is confirmed by the fact that in simulations the problem of multiple local maxima disappears as soon as $T > 3$ or $T > 4$ (without and with incidental trends, resp.). The multiplicity of local maxima of $l_A(\rho)$ is, in itself, not a fundamental problem, although it may well pose numerical difficulties. A further consequence is that it requires $T \geq 4$ or $T \geq 5$ (without and with incidental trends, resp.) to use the local curvature of $l_A(\rho)$ in a variance formula for the maximiser of $l_A(\rho)$.

Figures 1–3 in the Appendix illustrate the numerical properties of $l(\rho)$ and $l_A(\rho)$ just described, and those implied for $s_A(\rho)$. Each figure shows six curves, all computed from the same simulated data set. Different figures use different data sets, with $T = 2, 3, 4$ in Figures 1, 2, 3, respectively. All three data sets were generated with $N = 1000$, $\rho_0 = \alpha_i = 0$, $\sigma_i = 1$, $y_{i0} = 2$ ($i = 1, \dots, N$) and no incidental error variances. The three curves on the left of each figure are, from top to bottom, $l(\rho)$, $l_A(\rho)$ and $s_A(\rho)$ for the model without incidental trends and error variances; the curves on the right are $l(\rho)$, $l_A(\rho)$ and $s_A(\rho)$ for the model without incidental trends, but with incidental error variances. Both models are correctly specified, but the former is more parsimonious, as it correctly imposes error variance homogeneity. The absolute levels of $l(\rho)$ and $l_A(\rho)$ cannot be compared across models (due to the use of different constants) but their range can be compared, as an informal measure of information content. From the Bayesian point of view and using Lancaster’s (uninformative) priors, $l_A(\rho)$ is the marginal posterior log-density, normalised by NT . For both models and for all three data sets, the maximiser of $l(\rho)$ shows a downward bias, while the maximiser of $l_A(\rho)$ is close to $\rho_0 = 0$. Recall, however, that $l(\rho)$ and $l_A(\rho)$ are unbounded (at $N = 1000$ points) in the model with incidental error

variances and $T = 2$. This is disguised in the top-right curves in Figure 1, due to the grid (2000 points only) over which the functions were computed, to the log-transformation involved and to the averaging over N loglikelihood contributions. Lack of smoothness is most clearly seen through $s_A(\rho)$. The multiplicity of local maxima when $T = 3$ is well illustrated in the bottom-right curve of Figure 2. Apart from the lack of smoothness for very small T in the model with incidental error variances, $l(\rho)$ and $l_A(\rho)$ agree well across the two models. We conjecture that, when the two models are correctly specified (which is the case for the data used for Figures 1–3), they share the *same* curves $\lim_{N \rightarrow \infty} E_0[l(\rho)]$ and $\lim_{N \rightarrow \infty} E_0[l_A(\rho)]$. If this is correct, there is no information loss in allowing error variances to differ across i .

To appreciate the difference between the (adjusted) profile loglikelihoods with and without incidental error variances, we repeated the above exercise with mild error variance heterogeneity and smaller N . We set $N = 100$, $\sigma_i = 1$ for $i = 1, \dots, 95$ and $\sigma_i = 5$ for $i = 96, \dots, 100$. In other respects the design was as above. Figures 4–6 in the Appendix present the results. For all three data sets, the maximiser of the heteroskedastic version of $l_A(\rho)$ (or a smoothed variant thereof, for $T = 2$) is much closer to $\rho_0 = 0$ than the maximiser of the homoskedastic version. Further, when $T = 4$, the heteroskedastic version of $s_A(\rho)$ is steeper around $\rho_0 = 0$ than the homoskedastic version, confirming the intuition that the heteroskedastic (adjusted) loglikelihood extracts the information from the data more efficiently than the (mis-specified) homoskedastic one. The same holds when $T = 3$, on smoothing the heteroskedastic version of $s_A(\rho)$.

4 Global properties of the adjusted profile likelihood

The question arises whether the curve $\lim_{N \rightarrow \infty} E_0[l_A(\rho)]$ identifies ρ_0 , regardless of G . In particular: is ρ_0 the unique maximiser of $\lim_{N \rightarrow \infty} E_0[l_A(\rho)]$? Is the solution to $\lim_{N \rightarrow \infty} E_0[s_A(\rho)] = 0$ unique (hence equal to ρ_0)? We have no definitive results regarding these questions, but the subsequent analysis

indicates that the answer is no to both questions. Yet, the analysis also indicates that ρ_0 is identified as the *interior* maximiser of $\lim_{N \rightarrow \infty} E_0[l_A(\rho)]$ if such a maximiser exists, and otherwise as the global maximiser on $[-1, 1]$. We limit the discussion to the case without incidental trends and error variances. The cases with incidental trends and/or incidental error variances are expected to yield the same conclusions.

Let $T \geq 2$. Let $\mathcal{F}(T, \rho_0)$ be the family of curves $\lim_{N \rightarrow \infty} E_0[l_A(\rho)]$, $\rho \in [-1, 1]$, obtained by varying (G, σ^2) . This family coincides with the family of curves $\lim_{N \rightarrow \infty} E_0[\int s_A(\rho) d\rho]$, $\rho \in [-1, 1]$, obtained by varying (G, σ^2) . Define the scalar random variable

$$z_{i0} = \begin{cases} \left(y_{i0} - \frac{\alpha_i}{1 - \rho_0} \right) / \sqrt{\frac{\sigma^2}{1 - \rho_0^2}} & \text{if } -1 < \rho_0 < 1, \\ \left(y_{i0} - \frac{\alpha_i}{2} \right) / \sigma & \text{if } \rho_0 = \pm 1. \end{cases}$$

Because $s(\rho)$ is invariant under (individual-specific) location and (common) scale transformations of the data, $s(\rho)$ depends on $\{y_{i0}, \lambda_i\}_{i=1}^N$ only through $\{z_{i0}\}_{i=1}^N$ (and, when $\rho_0 = 1$, $s(\rho)$ does not depend on $\{y_{i0}, \lambda_i\}_{i=1}^N$ at all). Hence $\lim_{N \rightarrow \infty} E_0[s_A(\rho)]$ depends on (G, σ^2) only through the distribution that (G, σ^2) induces on z_{i0} . Therefore, the family of curves $\lim_{N \rightarrow \infty} E_0[s_A(\rho)]$ obtained by varying (G, σ^2) is equivalently obtained by varying the distribution of z_{i0} . Integrating those curves, in turn, generates $\mathcal{F}(T, \rho_0)$, including level shifts appearing as a constant of integration. So, without loss of generality, we can put $\sigma = 1$, $\alpha_i = 0$ and generate $\mathcal{F}(T, \rho_0)$ by varying the distribution of z_{i0} or, equivalently, by varying the distribution of y_{i0} . Now,

$$\begin{aligned} \lim_{N \rightarrow \infty} E_0[s(\rho)] &= \frac{p \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N (y_i - \rho y_{i-})' M y_{i-}}{p \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N (y_i - \rho y_{i-})' M (y_i - \rho y_{i-})} \\ &= \frac{E_0 [(y_i - \rho y_{i-})' M y_{i-}]}{E_0 [(y_i - \rho y_{i-})' M (y_i - \rho y_{i-})]}. \end{aligned}$$

Recalling that $y_{i-} = \Gamma \varepsilon_i + \Gamma \iota_T \alpha_i + \gamma y_{i0} = \Gamma \varepsilon_i + \gamma y_{i0}$, we find

$$\begin{aligned} E_0 [(y_i - \rho y_{i-})' M y_{i-}] &= \text{tr}(M\Gamma) + (\rho_0 - \rho)A(\omega^2), \\ E_0 [(y_i - \rho y_{i-})' M (y_i - \rho y_{i-})] &= \text{tr}(M) + 2(\rho_0 - \rho)\text{tr}(M\Gamma) + (\rho_0 - \rho)^2 A(\omega^2), \end{aligned}$$

where

$$A(\omega^2) = \text{tr}(\Gamma' M \Gamma) + \gamma' M \gamma \omega^2, \quad \omega^2 = E_0(y_{i0}^2).$$

Hence

$$\begin{aligned} \lim_{N \rightarrow \infty} E_0[s(\rho)] &= \frac{\text{tr}(M \Gamma) + (\rho_0 - \rho) A(\omega^2)}{\text{tr}(M) + 2(\rho_0 - \rho) \text{tr}(M \Gamma) + (\rho_0 - \rho)^2 A(\omega^2)} \\ &= h(\rho_0, \rho, \omega^2), \end{aligned}$$

say. Therefore, $\mathcal{F}(T, \rho_0)$ is generated by the integrals

$$H(\rho_0, \rho, \omega^2) = \int (h(\rho_0, \rho, \omega^2) - B(\rho)) d\rho,$$

with ω^2 varying over the positive half-line. Note that

$$H(\rho_0, \rho, \omega^2) = E_0[l_A(\rho) | \{y_{i0} = \omega\}_{i=1}^N].$$

It appears that, when $-1 < \rho_0 < 1$, $H(\rho_0, \rho, \omega^2)$ always has a unique interior strict local maximum, at $\rho = \rho_0$, for any $\omega^2 > 0$.³ When $\rho_0 = \pm 1$, there is no interior local maximum, and the global maximum is reached at $\rho = \rho_0$. However, when $-1 < \rho_0 < 1$ and ω^2 is small enough (given T and ρ_0), $H(\rho_0, \rho, \omega^2)$ may reach its global maximum at $\rho = 1$. The precise conditions under which this happens are not fully clear, but the intuition why it happens is as follows. Observe that, when $-1 < \rho_0 < 1$, $\omega^2 = E(z_{i0}^2)(1 - \rho_0^2)^{-1}$ and $E(z_{i0}^2) = 1$ can be interpreted as initial values being drawn from the stationary distribution. Values of $E(z_{i0}^2) > 1$ can be interpreted as initial observations that are outlying relative to the stationary distribution; values of $E(z_{i0}^2) < 1$ imply inlying initial observations. Very small values of ω^2 correspond to strong inlying initial observations. When the initial observations are inlying, the data carry less information about ρ_0 . In the extreme case where $\omega^2 = 0$, the initial pairs of observations, (y_{i0}, y_{i1}) , carry no information at all about ρ_0 . Moreover, the effect of strong inlying initial observations on

³We computed $H(\rho_0, \rho, \omega^2)$ numerically, as $l_A(\rho)$, from a data set generated with $N = 10^6$, for any desired T , ρ_0 and ω^2 , where we set $\sigma = 1$, $\alpha_i = 0$ and $y_{i0} = \omega = z_{i0}(1 - \rho_0^2)^{-1/2}$ ($i = 1, \dots, N$).

the information in the data is larger when T is small (because it takes time to revert to the stationary distribution) and when $|\rho_0|$ is large (because it takes more time to revert to the stationary distribution; e.g. when $\rho_0 = 0$, y_{i1} is drawn from the stationarity distribution, regardless of y_{i0}). In our computations, with z_{i0} fixed at small values (say, between 0 and 0.5), small T and a whole range of values of ρ_0 , $H(\rho_0, \rho, \omega^2)$ starts to re-increase in ρ somewhere between ρ_0 and 1, and has its global maximum at 1.⁴ Our computations also indicate that that $H(\rho_0, \rho, \omega^2)$ is never re-increasing when $z_{i0} \geq 1$.

Figure 7 in the Appendix illustrates the fact that, when the initial observations are inlying, $\lim_{N \rightarrow \infty} E_0[l_A(\rho)]$ may reach its global maximum at 1 even though $\rho_0 \neq 1$. The top curve gives $\lim_{N \rightarrow \infty} E_0[l_A(\rho)]$ when $T = 2$, $\rho_0 = -0.2$ and $z_{i0} = 0.5$ ($i = 1, \dots, N$). The bottom curve gives $\lim_{N \rightarrow \infty} E_0[l_A(\rho)]$ when $T = 5$, $\rho_0 = 0.6$ and $z_{i0} = 0$ ($i = 1, \dots, N$). In the latter case, $\lim_{N \rightarrow \infty} E_0[l_A(\rho)]$ is essentially flat between 0.5 and 1, suggesting ρ_0 is nearly unidentified by the curve $\lim_{N \rightarrow \infty} E_0[l_A(\rho)]$.

Based on the global properties of $\lim_{N \rightarrow \infty} E_0[l_A(\rho)]$, we suggest to estimate ρ_0 by the interior strict local maximiser of $l_A(\rho)$ (or a smoothed variant thereof, in the case of incidental error variances with small enough T)—if it exists, and by the global maximiser otherwise.

⁴Lancaster (2002) show graphs of the marginal posterior for ρ in a model with covariates. In some of these graphs, in particular those where N is small, the marginal posterior is also re-increasing. Our findings suggest that this need not only arise due to small sample variation, but may reflect a more fundamental property of the marginal posterior or, equivalently, of the adjusted profile likelihood.

Appendix

Derivation of (6). A tedious but straightforward calculation shows that, if $-1 \leq \rho_0 < 1$,

$$\begin{aligned} \iota'_T \Gamma \iota_T &= \frac{T}{1 - \rho_0} - \frac{1 - \rho_0^T}{(1 - \rho_0)^2}, \\ \iota'_T \Gamma \tau_T &= \frac{(T+1)(T-2)}{2(1 - \rho_0)} + \frac{1 - T\rho_0}{(1 - \rho_0)^2} + \frac{\rho_0^2(1 - \rho_0^{T-1})}{(1 - \rho_0)^3}, \\ \tau'_T \Gamma \iota_T &= \frac{(T+1)(T+2)}{2(1 - \rho_0)} - \frac{2 + T(1 - \rho_0 - \rho_0^T)}{(1 - \rho_0)^2} - \frac{\rho_0^2(1 - \rho_0^{T-2})}{(1 - \rho_0)^3}, \\ \tau'_T \Gamma \tau_T &= \frac{(T-1)(5T + 2T^2 + 6)}{6(1 - \rho_0)} - \frac{(T+2)(T-1) + 2\rho_0}{2(1 - \rho_0)^2} \\ &\quad - \frac{T\rho_0^{T+1}}{(1 - \rho_0)^3} + \frac{\rho_0(1 - \rho_0^T)}{(1 - \rho_0)^4}, \end{aligned}$$

and, if $\rho_0 = 1$,

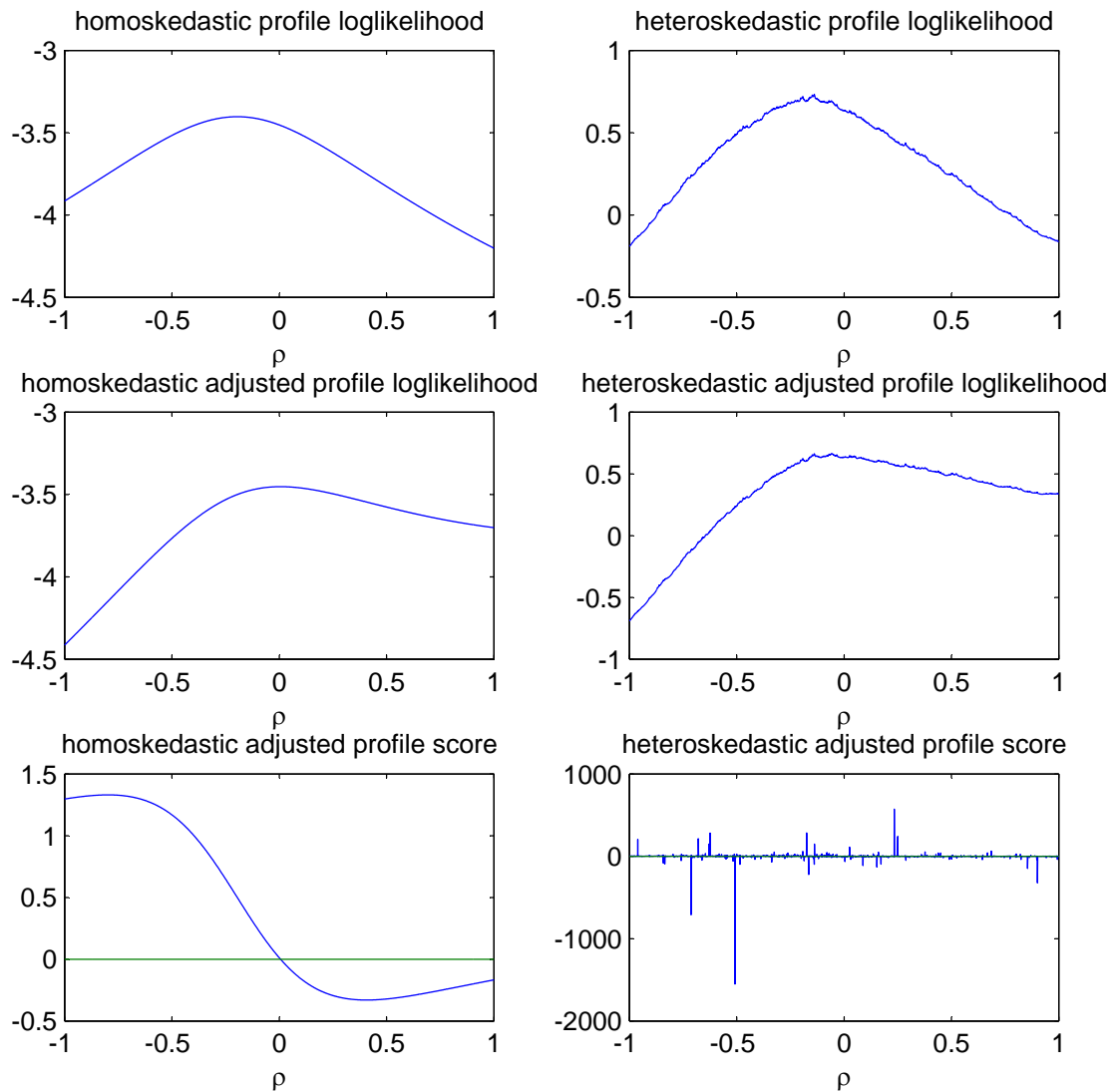
$$\begin{aligned} \iota'_T \Gamma \iota_T &= \frac{1}{2}T(T-1), \\ \iota'_T \Gamma \tau_T &= \frac{1}{6}T(T-1)(T+1), \\ \tau'_T \Gamma \iota_T &= \frac{1}{3}T(T-1)(T+1), \\ \tau'_T \Gamma \tau_T &= \frac{1}{24}T(T-1)(3T+2)(T+1). \end{aligned}$$

Using these expressions, together with

$$(J'_T J_T)^{-1} = \frac{2}{T(T-1)} \begin{pmatrix} 2T+1 & -3 \\ -3 & 6(T+1)^{-1} \end{pmatrix},$$

(6) follows on rearranging. \square

Figure 1: Profile loglikelihood and score functions (without incidental trends), $T = 2$



Note: Data were generated with $T = 2$, $N = 1000$, $\rho_0 = \alpha_i = \beta_i = 0$, $\sigma_i = 1$, $y_{i0} = 2$ ($i = 1, \dots, N$).

Figure 7: Re-increasing adjusted profile loglikelihood when initial observations are inlying

Notes: Top curve generated with $N = 10^6$, $T = 2$, $\rho_0 = -0.2$, $\alpha_i = \beta_i = 0$, $\sigma_i = 1$, $y_{i0} = z_{i0}(1 - \rho_0^2)^{-1/2}$, $z_{i0} = 0.5$ ($i = 1, \dots, N$). Bottom curve generated with $N = 10^6$, $T = 5$, $\rho_0 = 0.6$, $\alpha_i = \beta_i = 0$, $\sigma_i = 1$, $y_{i0} = z_{i0}(1 - \rho_0^2)^{-1/2}$, $z_{i0} = 0$ ($i = 1, \dots, N$).

References

- [1] Ahn, S. and P. Schmidt (1995). Efficient estimation of models for dynamic panel data. *Journal of Econometrics*, 68, 5—27.
- [2] Alvarez, J. and M. Arellano (2004), Robust likelihood estimation of dynamic panel data models, unpublished manuscript.
- [3] Anderson, T.W. and C. Hsiao (1981), Estimation of dynamic models with error components, *Journal of the American Statistical Association*, 76, 598–606.
- [4] Anderson, T.W. and C. Hsiao (1982), Formulation and estimation of dynamic panel data models using panel data, *Journal of Econometrics*, 18, 47–82.
- [5] Arellano, M. (2003), *Panel Data Econometrics*, Oxford University Press, Oxford.
- [6] Arellano, M. and S. Bond (1991), Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations, *Review of Economic Studies*, 58, 277–297.
- [7] Conniffe, D. and J.E. Spencer (2001), When moments of ratios are ratios of moments, *The Statistician*, 50, 161–168.
- [8] Cox, D.R. and N. Reid (1987), Parameter orthogonality and approximate conditional inference, *Journal of the Royal Statistical Society, Series B*, 49, 1–39.
- [9] Fisher, R.A. (1930), The moments of the distribution for normal samples of measures of departure from normality, *Proceedings of the Royal Society, A*, 130, 16–28.
- [10] Geary, R.C. (1933), A general expression for the moments of certain symmetrical functions of normal samples, *Biometrika*, 25, 184–186.

- [11] Lancaster, T. (2002), Orthogonal parameters and panel data, *Review of Economic Studies*, 69, 647–666.
- [12] McCullagh, P. and R. Tibshirani (1990), A simple method for the adjustment of profile likelihoods, *Journal of the Royal Statistical Society, Series B*, 52, 325–344.
- [13] Neyman, J. and E. L. Scott (1948), Consistent estimates based on partially consistent observations, *Econometrica*, 16, 1–32.
- [14] Nickell, S. (1981), Biases in dynamic models with fixed effects, *Econometrica*, 49, 1417–1426.
- [15] Phillips, P.C.B. and D. Sul (2007), Bias in dynamic panel estimation with fixed effects, incidental trends and cross section dependence, *Journal of Econometrics*, 137, 162–188.
- [16] Reeds, J.A. (1985), Asymptotic number of roots of Cauchy location likelihood equations, *The Annals of Statistics*, 13, 775–784.