An Adjusted profile likelihood for non-stationary panel data models with fixed effects

Geert Dhaene, Koen Jochmans

AN ADJUSTED PROFILE LIKELIHOOD FOR NON-STATIONARY PANEL DATA MODELS WITH FIXED EFFECTS

Geert Dhaene* Koen Jochmans†

K.U. Leuven CORE

First version: March 2007; This version: March 22, 2011

Abstract: We calculate the bias of the profile score for the autoregressive parameters ρ and covariate slopes β in the linear model for $N \times T$ panel data with p lags of the dependent variable, exogenous covariates, fixed effects, and unrestricted initial observations. The bias is a vector of multivariate polynomials in ρ with coefficients that depend only on T. We center the profile score and, on integration, obtain an adjusted profile likelihood. When $p = 1$, the adjusted profile likelihood coincides with Lancaster's (2002) marginal posterior. More generally, it is an integrated likelihood, in the sense of Arellano and Bonhomme (2009), with fixed effects integrated out using a new data-independent prior. It appears that ρ and β are identified as the unique point where the large N adjusted profile likelihood reaches a *local* maximum (or a flat inflection point, as a limiting case) inside or on an ellipsoid centered at the maximum likelihood estimator. We prove this when $p = 1$ and report numerical calculations that support it when $p > 1$. The global maximum of the adjusted profile likelihood lies at infinity for any N.

JEL classification: C13, C22, C23
Keywords: adjusted likelihood, bias correction, dynamic panel data, fixed effects, non-stationarity.

Introduction

Maximum likelihood estimates of dynamic linear fixed effect models from short panels are known to be severely biased and inconsistent because of the incidental parameter problem first described by Neyman and Scott (1948); see Nerlove (1967, 1971) and Nickell (1981), for example. The failure of maximum likelihood has shifted interest towards GMM-based estimation in these models. An overview and discussion is provided by Arellano (2003, Chapter 6).

In an interesting paper, Lancaster (2002) argued for a return to likelihood-based inference and suggested a Bayesian resolution of the incidental parameter problem in the non-stationary AR(1) model.[1] The fixed effects are first orthogonalized (in the information sense) to the

*Address: K.U. Leuven, Department of Economics, Naamsestraat 69, B-3000 Leuven, Belgium. Tel. +32 16 326798; Fax +32 16 326796; E-mail: `geert.dhaene@econ.kuleuven.be`.

†Address: U.C. Louvain, Center for Operations Research and Econometrics, Voie du Roman Pays 34, B-1348 Louvain-la-Neuve, Belgium. Tel. +32 10 474329; E-mail: `koen.jochmans@uclouvain.be`.

[1]Throughout, non-stationarity will refer to initial observations being left unrestricted (and being conditioned upon) and the possibility that the autoregressive parameter vector lies outside the stationary region. Likelihood-based estimation of AR(1) models under restrictions on the initial observations has been considered by Kiefer (1980), Cruddas, Reid, and Cox (1989), Hsiao, Pesaran, and Tahmiscioglu (2002), and Kruiniger (2008).

common parameters and then integrated from the likelihood using uniform priors. The first order condition for the posterior mode is unbiased. Unfortunately, orthogonalization is not possible when the AR(1) model is augmented with covariates, as Lancaster showed, or when the autoregressive order is greater than one, as we show. Lancaster noted, however, that incomplete orthogonalization suffices in the AR(1) model with covariates. Arellano and Bonhomme (2009) considered integrating possibly non-orthogonalized fixed effects from the likelihood using more general priors, generalizing Lancaster (2002). Their characterization of bias-reducing priors in a general class of models (including nonlinear ones) further shows that orthogonality is sufficient but not necessary for being able to improve on maximum likelihood.

In this paper, we deal with the incidental parameter problem in the non-stationary AR(p) model with covariates by adjusting the profile likelihood. Our approach is frequentist and does not require orthogonalization or finding suitable priors. The adjustment is based on a bias calculation, and subsequent centering, of the profile score, similar in spirit to McCullagh and Tibshirani (1990).[2] An intermediate result of our analysis is a simple unbiased estimating equation for the AR(p) model. The adjustment term to the profile log-likelihood that results from it is a multivariate polynomial in the autoregressive parameters with coefficients depending only on the number of time periods. Thus, the adjustment to the profile likelihood is independent of the data and the true parameter values.

When $p = 1$, the adjusted profile likelihood coincides with the marginal posterior obtained by Lancaster (2002). Similarly, when $p > 1$, it coincides with the marginal posterior obtained by integrating out suitably reparameterized fixed effects using a uniform prior. This view relates the adjusted likelihood to Lancaster's approach, although the transformed fixed effects are no longer information orthogonal when $p > 1$ or when covariates are present. We show that the adjusted profile likelihood is also an integrated likelihood in the sense of Arellano and Bonhomme (2009). That is, for a specific choice of prior, the integrated likelihood coincides with the adjusted profile likelihood. From the integrated likelihood point of view, what we find is a data-independent prior for the AR(p) model that fully eliminates the bias from the first order condition. Such a prior was thought not to exist when $p > 1$. We also show that centering the profile score is equivalent to inverting a plug-in estimate of the probability limit of the within-group estimator, complementing and generalizing Bun and Carree (2005).

Recently, Chamberlain and Moreira (2009) and Moreira (2009) suggested the use of invariance arguments to eliminate the fixed effects from the likelihood. It would be interesting if a connection could be established between their invariant likelihood and the adjusted profile likelihood, but we doubt that this is possible in the present model. In the simplest case, the

[2] Our bias calculation is related to Alvarez and Arellano's (2004) in the case of time series homoskedasticity. We calculate the bias of the profile score for the regression coefficients, while they calculated the bias of the conditional score for all common parameters given maximum likelihood estimates of the fixed effects.

AR(1) model without covariates, Moreira's (2009) maximal invariant likelihood differs from the adjusted profile likelihood.

Perhaps our main (and unexpected) finding is that one should look for a *local* maximizer of the adjusted likelihood/marginal posterior because the global maximum is reached at infinity. The reason for this is that the profile log-likelihood is log-quadratic in the parameters while the adjustment term is a polynomial, with positive coefficients, in the autoregressive parameters. Thus, for large enough autoregressive parameter values, the adjustment term dominates, and the adjusted profile log-likelihood must be re-increasing. This finding complements observations by Lancaster (2002), who noticed in Monte Carlo experiments that his marginal posterior was sometimes re-increasing. This phenomenon is not just a small sample problem or an artifact of an unbounded parameter space. The adjusted likelihood/marginal posterior has its global maximum at infinity for *any* sample size, and it may already be re-increasing in the stationary parameter region and reach its maximum at the boundary. Consequently, identification is not achieved by global maximization, and estimation based solely on the first order condition is incomplete.

To examine identification, we profile out the covariate slopes, which are unique given the autoregressive parameters, so the analysis reduces to that of the model without covariates.

In the AR(1) model, we show that the autoregressive parameter is identified as the unique local maximizer—or flat inflection point, as a limiting case—of the (large sample) adjusted profile likelihood in the interval where the *un*adjusted profile score is downward sloping. The (large sample) maximum likelihood estimate is the midpoint of this interval. Identification as a flat inflection point is an instance of first order lack of identification, in Sargan's (1983) terminology. The leading case of interest is a unit root without deterministic trends, which is identified as a flat inflection point regardless of the number of time periods, in line with the rank deficiency of the expected Jacobian associated with the moment conditions of Ahn and Schmidt (1995); see Alvarez and Arellano (2004). For the unit root case the flat inflection point property has been noted independently by Ahn and Thomas (2006).

Identification in the AR(p) model via the adjusted likelihood is a harder problem. It appears that the autoregressive parameter vector is identified as the unique local maximizer or flat inflection point of the adjusted profile likelihood on a region bounded by a well-identified ellipsoid that is centered at the maximum likelihood estimate. We have no proof of this conjecture, but we report numerical calculations that support it.

The model parameters can be estimated by the local maximizer of the adjusted likelihood over a plug-in estimate of the ellipsoid (or, when no local maximizer exists, by the minimizer of the norm of the adjusted profile score). When the parameters are identifed as local maximizers, the estimator converges at the parametric rate, and its limit distribution is normal and correctly centered.

In Section 1, we present the model and derive the bias of the profile score. The adjusted likelihood is given in Section 2, where we also connect with Lancaster (2002), Arellano and Bonhomme (2009), and Bun and Carree (2005). Global properties of the adjusted likelihood are discussed in Section 3. In Section 4, we define the adjusted likelihood estimator and derive its main asymptotic properties. Section 5 discusses some simulation results. Technical proofs are collected in an appendix.

1 Bias of the profile score

1.1 Model and profile likelihood

Suppose we observe a scalar variable y, the first $p \geq 1$ lags of y, and a q-vector of covariates x (which may include lags), for N units i and T periods t. Assume that y_{it} is generated by

$$y_{it} = y_{it-}'\rho + x_{it}'\beta + \alpha_i + \varepsilon_{it}, \qquad i = 1, ..., N; t = 1, ..., T, \tag{1.1}$$

where $y_{it-} = (y_{it-1}, ..., y_{it-p})'$ and the ε_{it} are identically distributed with mean 0 and variance σ^2 and are independent across i and t and also of $x_{i't'}$ for all i' and t'. Let $y_i^0 = (y_{i(1-p)}, ..., y_{i0})'$, $X_i = (x_{i1}, ..., x_{iT})'$, and $\varepsilon_i = (\varepsilon_{i1}, ..., \varepsilon_{iT})'$. We make no assumptions about how (y_i^0, α_i, X_i), $i = 1, ..., N$, are generated, except for the assumption that the covariates are not linearly dependent. The unknown parameters are $\theta = (\rho', \beta')'$, σ^2, and $\alpha_1, ..., \alpha_N$. Let θ_0 and σ_0^2 be the true values of θ and σ^2. Our interest lies in consistently estimating θ_0 under large N and fixed T asymptotics. We do not impose the stationarity condition on ρ_0, i.e., we allow any $\rho_0 \in \mathbb{R}^p$.

Let $z_{it} = (y_{it-}', x_{it}')'$, $Y_{i-} = (y_{i1-}, ..., y_{iT-})'$, $Z_i = (Y_{i-}, X_i)$, and $y_i = (y_{i1}, ..., y_{iT})'$, so that $My_i = MZ_i\theta + M\varepsilon_i$ where $M = I_T - T^{-1}\iota\iota'$ and ι is a conformable vector of ones. We assume that $N^{-1}\sum_{i=1}^N Z_i'MZ_i$ and its probability limit as $N \to \infty$ are nonsingular. The Gaussian quasi-log-likelihood, conditional on $y_1^0, ..., y_N^0$ and normalized by the number of observations, is given by

$$-\frac{1}{2NT}\sum_{i=1}^N\sum_{t=1}^T\left(\log\sigma^2 + \frac{1}{\sigma^2}(y_{it} - z_{it}'\theta - \alpha_i)^2\right) + \mathrm{c},$$

where, here and later, c is a non-essential constant. Replacing the nuisance parameters $\alpha_1, ..., \alpha_N$ and σ^2 with their MLE for given θ gives the (normalized) profile log-likelihood,

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

The profile score, $s(\theta) = \nabla_\theta l(\theta)$, has elements

$$\begin{aligned} s_{\rho_j}(\theta) &= \frac{\sum_{i=1}^N (y_i - Z_i\theta)' M y_{i,-j}}{\sum_{i=1}^N (y_i - Z_i\theta)' M (y_i - Z_i\theta)}, \qquad j = 1, ..., p, \\ s_{\beta_j}(\theta) &= \frac{\sum_{i=1}^N (y_i - Z_i\theta)' M x_{i,j}}{\sum_{i=1}^N (y_i - Z_i\theta)' M (y_i - Z_i\theta)}, \qquad j = 1, ..., q, \end{aligned}$$

where $y_{i,-j}$ is the jth column of Y_{i-} and $x_{i,j}$ is the jth column of X_i.

1.2 Bias of the profile score

It is well known that the profile score is asymptotically biased, that is,

$$\text{plim}_{N\to\infty} s(\theta_0) \neq 0,$$

where, here and later, probability limits and expectations are taken conditionally, given (y_i^0, α_i, X_i), $i = 1, ..., N$. Hence, $\theta_0 \neq \arg\max_\theta \text{plim}_{N\to\infty} l(\theta)$ and the MLE, solving $s(\theta) = 0$, is inconsistent.

The following calculation yields the bias of the profile score, either asymptotically or, when $\varepsilon_{it} \sim \mathcal{N}(0, \sigma^2)$, for fixed N. Rewrite (1.1) as

$$Dy_i = Cy_i^0 + X_i\beta + \iota\alpha_i + \varepsilon_i, \qquad i = 1, ..., N,$$

where $D = D(\rho)$ and $C = C(\rho)$ are the $T \times T$ and $T \times p$ matrices

$$D = \begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ -\rho_1 & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ -\rho_p & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -\rho_p & \cdots & -\rho_1 & 1 \end{pmatrix}, \quad C = \begin{pmatrix} \rho_p & \cdots & \cdots & \cdots & \rho_1 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & \rho_p \\ 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 \end{pmatrix}.$$

Then

$$\begin{pmatrix} y_i^0 \\ y_i \end{pmatrix} = \xi_i + F\varepsilon_i, \qquad \xi_i = \begin{pmatrix} y_i^0 \\ D^{-1}\left(Cy_i^0 + X_i\beta + \iota\alpha_i\right) \end{pmatrix}, \qquad F = \begin{pmatrix} 0 \\ D^{-1} \end{pmatrix},$$

and $y_{i,-j} = S_j(\xi_i + F\varepsilon_i)$, where $S_j = (0_{T\times(p-j)} : I_T : 0_{T\times j})$, a selection matrix. Therefore,

$$\begin{aligned} \text{plim}_{N\to\infty} s_{\rho_j}(\theta_0) &= \frac{\text{plim}_{N\to\infty} N^{-1} \sum_{i=1}^N \varepsilon_i' M S_j(\xi_{0i} + F_0\varepsilon_i)}{\text{plim}_{N\to\infty} N^{-1} \sum_{i=1}^N \varepsilon_i' M \varepsilon_i} = \frac{\mathbb{E}(\varepsilon_i' M S_j F_0 \varepsilon_i)}{\mathbb{E}(\varepsilon_i' M \varepsilon_i)} \\ &= \frac{\text{tr} M S_j F_0}{T - 1}, \\ \text{plim}_{N\to\infty} s_{\beta_j}(\theta_0) &= 0, \end{aligned}$$

where ξ_{0i} and F_0 are ξ_i and F, evaluated at θ_0. If, in addition, $\varepsilon_{it} \sim \mathcal{N}(0, \sigma^2)$, then

$$\begin{aligned}
\mathbb{E}[s_{\rho_j}(\theta_0)] &= \mathbb{E}\left(\frac{\sum_{i=1}^N \varepsilon_i' M S_j(\xi_{0i} + F_0\varepsilon_i)}{\sum_{i=1}^N \varepsilon_i' M \varepsilon_i}\right) = \mathbb{E}\left(\frac{\sum_{i=1}^N \varepsilon_i' M S_j F_0 M \varepsilon_i}{\sum_{i=1}^N \varepsilon_i' M \varepsilon_i}\right) \\
&= \frac{\mathbb{E}(\varepsilon_i' M S_j F_0 M \varepsilon_i)}{\mathbb{E}(\varepsilon_i' M \varepsilon_i)} = \frac{\operatorname{tr} M S_j F_0}{T-1}, \qquad (1.2) \\
\mathbb{E}[s_{\beta_j}(\theta_0)] &= 0,
\end{aligned}$$

by well-known properties of the normal distribution and the following geometric argument, which goes back to Fisher (1930) and Geary (1933). Let $v \sim \mathcal{N}(0, \sigma^2 I_g)$ and let Q be a $g \times h$ matrix such that $Q'Q = I_h$, so QQ' is idempotent. Write $I_g - QQ'$ as PP', where $P'P = I_{g-h}$. Transform v into $m = P'v$, the radius $r = (v'QQ'v)^{1/2}$, and the $h-1$ polar angles a of $Q'v$. Then the elements of $(m', r, a')'$ are independent. Therefore, for any $g \times g$ matrix W, if $A = v'QQ'WQQ'v$ and $B = v'QQ'v$, then the ratio A/B depends on v only through a and hence is independent of B, which implies that $\mathbb{E}(A) = \mathbb{E}(A/B)\mathbb{E}(B)$ and $\mathbb{E}(A/B) = \mathbb{E}(A)/\mathbb{E}(B)$.[3] The transition to (1.2) now follows from applying this property to the ratio

$$\frac{\sum_{i=1}^N \varepsilon_i' M S_j F_0 M \varepsilon_i}{\sum_{i=1}^N \varepsilon_i' M \varepsilon_i} = \frac{\varepsilon'(I_N \otimes M)(I_N \otimes S_j F_0)(I_N \otimes M)\varepsilon}{\varepsilon'(I_N \otimes M)\varepsilon}$$

with $v = \varepsilon = (\varepsilon_1', \dots, \varepsilon_N')'$, $QQ' = I_N \otimes M$, and $W = I_N \otimes S_j F_0$.

To summarize the results so far regarding the bias of the profile score, let

$$b(\rho) = (b_1(\rho), \dots, b_{p+q}(\rho))', \qquad b_j(\rho) = \begin{cases} \dfrac{\operatorname{tr} M S_j F}{T-1}, & j = 1, \dots, p; \\ 0, & j = p+1, \dots, p+q. \end{cases}$$

Then $\operatorname{plim}_{N\to\infty} s(\theta_0) = b(\rho_0)$; if, in addition, $\varepsilon_{it} \sim \mathcal{N}(0, \sigma_0^2)$, then $\mathbb{E}[s(\theta_0)] = b(\rho_0)$.

Two remarks are worth making. First, the bias of the profile score, $b(\rho_0)$, depends only on ρ_0 and T. It is independent of the initial observations, the fixed effects, and the covariates (recall that the expectations are conditioning on these). This is in sharp contrast with the bias of the ML (the within-group) estimator, which was first derived by Nickell (1981) for the first order autoregressive model under the assumption of stationarity of the initial observations. The bias of the MLE depends on the initial observations, the fixed effects, and the covariate values; it is smaller when the initial observations are more outlying relative to the stationary distributions. Second, the Nickell bias concerns a probability limit as $N \to \infty$ whereas here, when the errors are normal, $\mathbb{E}[s(\theta_0)] = b(\rho_0)$ is a finite sample result holding for fixed N and T and may therefore be of independent interest in a time series setting.

[3] For a discussion and historical perspective on this device, see Conniffe and Spencer (2001).

For $k = (k_1, ..., k_p)' \in \mathbb{N}^p$, let $\rho^k = \prod_{j=1}^p \rho_j^{k_j}$, and let $\tau = (1, ..., p)'$. It is shown in the Appendix that

$$D^{-1} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \varphi_1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \varphi_{T-1} & \cdots & \varphi_1 & 1 \end{pmatrix}, \qquad \varphi_t = \sum_{\tau' k = t} \frac{(\iota' k)!}{k_1! \cdots k_p!} \rho^k, \tag{1.3}$$

where the summation is over all $k \in \mathbb{N}^p$ satisfying $\tau' k = t$. Therefore, since

$$S_j F = \begin{pmatrix} 0 & 0 \\ D_j^{-1} & 0 \end{pmatrix}$$

where D_j^{-1} is the leading $(T-j) \times (T-j)$ block of D^{-1}, we obtain

$$b_j(\rho) = -\frac{\iota' D_j^{-1} \iota}{T(T-1)} = -\sum_{t=0}^{T-j-1} \frac{T-j-t}{T\,(T-1)} \varphi_t, \qquad j = 1, ..., p,$$

where $\varphi_0 = 1$. Hence, for $j = 1, ..., p$, $b_j(\rho)$ is a multivariate polynomial of degree $T - j - 1$ in $\rho_1, ..., \rho_p$ with non-positive coefficients.

2 Adjusted profile likelihood

2.1 Centered profile score and adjusted profile likelihood

By construction, the centered (or adjusted) profile score,

$$s_{\mathrm{A}}(\theta) = s(\theta) - b(\rho),$$

is asymptotically unbiased, i.e., $\mathrm{plim}_{N \to \infty} s_{\mathrm{A}}(\theta_0) = 0$. In the Appendix, we show that the differential equation $\nabla_\theta a(\rho) = b(\rho)$ has a solution which, up to an arbitrary constant of integration, is

$$a(\rho) = \sum_{S \in \mathcal{S}} a_S(\rho), \qquad a_S(\rho) = -\sum_{t=|S|}^{T-1} \frac{T-t}{T(T-1)} \sum_{k \in \mathcal{K}_S : \tau' k = t} \frac{(\iota' k - 1)!}{k_1! \cdots k_p!} \rho_S^{k_S}, \tag{2.1}$$

where $\mathcal{S}$ is the collection of the non-empty subsets of $\{1, ..., p\}$; $|S|$ is the sum of the elements of S; $\mathcal{K}_S = \{k \in \mathbb{N}^p | k_j > 0 \text{ iff } j \in S\}$; and $\rho_S = (\rho_j)_{j \in S}$ and $k_S = (k_j)_{j \in S}$ are subvectors of ρ and k determined by S.

We call the function

$$l_{\mathrm{A}}(\theta) = l(\theta) - a(\rho),$$

which has derivative $\nabla_\theta l_{\mathrm{A}}(\theta) = s_{\mathrm{A}}(\theta)$, an adjusted profile log-likelihood. Every subvector ρ_S of ρ contributes to $l_{\mathrm{A}}(\theta)$ an adjustment term, $-a_S(\rho)$, which takes the form of a multivariate polynomial in ρ_j, $j \in S$, with positive coefficients that are independent of p.[4,5]

2.2 Connections with the literature

Lancaster (2002) studied the AR(1) model, with and without covariates, from a Bayesian perspective. It turns out that $e^{l_{\mathrm{A}}(\theta)}$ coincides, up to a non-essential normalization, with Lancaster's posterior density for θ, thereby giving it a frequentist interpretation. With $p = 1$, we have $\varphi_t = \rho^t$ and

$$b_1(\rho) = -\sum_{t=1}^{T-1} \frac{T-t}{T(T-1)}\rho^{t-1}, \qquad a(\rho) = -\sum_{t=1}^{T-1} \frac{T-t}{T(T-1)t}\rho^{t}.$$

With independent uniform priors on the reparameterized effects $\eta_i = \alpha_i e^{-(T-1)a(\rho)}$ and on θ and $\log\sigma^2$, Lancaster's posterior density for $\vartheta = (\theta', \sigma^2)'$ is

$$f(\vartheta|data) \propto \sigma^{-N(T-1)-2} \exp\left(-N(T-1)a(\rho) - Q^2(\theta)\sigma^{-2}/2\right),$$

where $Q^2(\theta) = \sum_{i=1}^N (y_i - Z_i\theta)' M (y_i - Z_i\theta) \propto e^{-2l(\theta)}$. Integrating over σ^2 gives $f(\theta|data) \propto e^{-N(T-1)a(\rho)}(Q^2(\theta))^{-N(T-1)/2}$ and, hence,

$$f(\theta|data) \propto e^{N(T-1)l_{\mathrm{A}}(\theta)}. \tag{2.2}$$

Thus, the posterior density and the adjusted likelihood for θ are equivalent as a basis for inference. More generally, for any p and q, independent uniform priors on $\eta_1, ..., \eta_N, \theta, \log\sigma^2$, with $\eta_i = \alpha_i e^{-(T-1)a(\rho)}$ and $a(\rho)$ as in (2.1), yield a posterior $f(\theta|data)$ that is related to $l_{\mathrm{A}}(\theta)$ as in (2.2).

Lancaster's choice of independent uniform priors on η_i has its foundation in the AR(1) model without covariates, where η_i is information orthogonal to ϑ and the posterior density (hence also $e^{l_{\mathrm{A}}(\theta)}$) has an interpretation as a Cox and Reid (1987) approximate conditional likelihood (see also Arellano, 2003, pp. 103–106). Information orthogonalization to a multidimensional parameter is generally not possible (see, e.g., Severini, 2000, pp. 340–342). Here, orthogonalization is not possible when the AR(1) model is augmented with covariates, as shown by Lancaster, or when the autoregressive order, p, is greater than one, as shown in the Appendix. From a bias correction perspective, however, information orthogonality is sufficient but not necessary. For any p and q, $s_{\mathrm{A}}(\theta) = 0$ is an unbiased estimating equation, and the bias calculation underlying it is immune to the non-existence of orthogonalized fixed effects.

[4] The adjustment, $-a(\rho)$, is robust to cross-sectional heteroskedasticity (σ_i^2 instead of σ^2, say). We investigated this in an older version of this paper (Dhaene and Jochmans, 2007), where we also derived the adjustment for the AR(1) model when there are incidental time trends. In the latter case, the adjustment is still a polynomial.

[5] Unbalanced panel data can be accommodated by writing $l(\theta)$ as a weighted average of K balanced subpanel profile log-likelihoods, say $l_k(\theta)$, and adjusting each $l_k(\theta)$. Missing data, with observations missing completely at random, fit into this scheme by forming complete subseries from each incomplete time series.

Arellano and Bonhomme's (2009) approach shares the integration step with Lancaster (2002) but allows non-uniform priors on fixed effects or, equivalently, non-orthogonalized fixed effects. Of interest are bias-reducing priors, i.e., weighting schemes that deliver an integrated likelihood whose score equation has bias $o(T^{-1})$ as opposed to the standard $O(T^{-1})$. The linear dynamic model (with general p, q) illustrates an interesting result of Arellano and Bonhomme that generalizes the scope of uniform integration to situations where orthogonalization is impossible. For a given prior $\pi_i(\alpha_i|\vartheta)$, the (normalized) log integrated likelihood is

$$l^I(\vartheta) = \frac{1}{NT}\sum_{i=1}^{N} \log \int \sigma^{-T/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_{it} - z_{it}'\theta - \alpha_i)^2\right) \pi_i(\alpha_i|\vartheta)\, \mathrm{d}\alpha_i + \mathrm{c}.$$

Choosing $\pi_i(\alpha_i|\vartheta) \propto e^{-(T-1)a(\rho)}$ yields

$$l^I(\vartheta) = -\frac{T-1}{2T}\log\sigma^2 - \frac{T-1}{T}a(\rho) - \frac{Q^2(\theta)}{2NT\sigma^2} + \mathrm{c}.$$

Profiling out σ^2 gives $\sigma^2(\theta) = \arg\max_{\sigma^2} l^I(\vartheta) = Q^2(\theta)/(N(T-1))$ and

$$l^I(\theta) = \max_{\sigma^2} l^I(\vartheta) = \frac{T-1}{T} l_{\mathrm{A}}(\theta) + \mathrm{c},$$

so $l^I(\theta)$ and $l_{\mathrm{A}}(\theta)$ are equivalent. Because $a(\rho)$ does not depend on true parameter values, $\pi_i(\alpha_i|\vartheta) \propto e^{-(T-1)a(\rho)}$ is a data-independent bias-reducing (in fact, bias-eliminating) prior in the sense of Arellano and Bonhomme.[6] Now, $\pi_i(\alpha_i|\vartheta) \propto e^{-(T-1)a(\rho)}$ is equivalent to $\pi_i(\eta_i|\vartheta) \propto 1$, i.e., to a uniform prior on $\eta_i = \alpha_i e^{-(T-1)a(\rho)}$, leading to the same $l^I(\vartheta)$. Arellano and Bonhomme (2009, Eq. (11)) give a necessary and sufficient condition for a uniform prior to be bias-reducing. With $\ell_i(\vartheta, \eta_i)$ denoting i's log-likelihood contribution in a parametrization η_i, the condition is that

$$\mathrm{plim}_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \nabla_{\eta_i}(A_i^{-1}B_i) = o(1) \qquad \text{as } T \to \infty, \tag{2.3}$$

where $A_i = A_i(\vartheta, \eta_i) = -\mathbb{E}_{\vartheta,\eta_i}\nabla_{\eta_i\eta_i}\ell_i(\vartheta, \eta_i)$, $B_i = B_i(\vartheta, \eta_i) = \mathbb{E}_{\vartheta,\eta_i}\nabla_{\vartheta\eta_i}\ell_i(\vartheta, \eta_i)$, and $\nabla_{\eta_i}(A_i^{-1}B_i)$ is evaluated at the true parameter values. When η_i and ϑ are information orthogonal, $B_i = 0$ and (2.3) holds. However, condition (2.3) is considerably weaker than parameter orthogonality. In the present model, when $p > 1$ or $q > 0$, and thus no orthogonalization is possible, it follows from our analysis and Arellano and Bonhomme (2009) that (2.3) must hold for $\eta_i = \alpha_i e^{-(T-1)a(\rho)}$. Indeed, as we show in the Appendix,

$$\nabla_{\eta_i}(A_i^{-1}B_i) = 0 \tag{2.4}$$

[6] Our focus is on estimating θ_0 but, noting the degrees-of-freedom correction in $\sigma^2(\theta)$, the estimating equation derived from $l^I(\vartheta)$ is also unbiased for σ_0^2.

because $A_i^{-1}B_i$ is free of η_i.

Finally, the centered profile score relates directly to Bun and Carree (2005). Observe that $s(\theta) = \sum_{i=1}^{N} Z_i' M(y_i - Z_i\theta)/Q^2(\theta)$. Further, $My_i = MZ_i\widehat{\theta} + M\widehat{\varepsilon}_i$ where $\widehat{\theta}$ is the within-group estimator, with residuals $\widehat{\varepsilon}_i$ satisfying $\sum_{i=1}^{N} Z_i' M\widehat{\varepsilon}_i = 0$. Therefore, solving $s_{\mathrm{A}}(\theta) = 0$ is equivalent to solving

$$\widehat{\theta} - \theta = \left(\sum_{i=1}^{N} Z_i' M Z_i\right)^{-1} b(\rho)Q^2(\theta). \tag{2.5}$$

In the AR(1) model, (2.5) corresponds to Bun and Carree's (2005) proposal for bias-correcting the within-group estimator.

3 Global properties of the adjusted profile likelihood when N is large

At this point it is tempting to anticipate that θ_0 maximizes $\mathrm{plim}_{N\to\infty} l_{\mathrm{A}}(\theta)$. However, as the following analysis shows, $-a(\rho)$ dominates $\mathrm{plim}_{N\to\infty} l(\theta)$ as $\|\rho\| \to \infty$ in almost all directions and, as a result, $\mathrm{plim}_{N\to\infty} l_{\mathrm{A}}(\theta)$ is unbounded from above.

Let $h(\theta) = \nabla_{\theta'} s(\theta)$, $c(\rho) = \nabla_{\theta'} b(\rho)$, and

$$\begin{aligned} L_{\mathrm{A}}(\theta) &= L(\theta) - a(\rho), & L(\theta) &= \mathrm{plim}_{N\to\infty} l(\theta), \\ S_{\mathrm{A}}(\theta) &= S(\theta) - b(\rho), & S(\theta) &= \mathrm{plim}_{N\to\infty} s(\theta), \\ H_{\mathrm{A}}(\theta) &= H(\theta) - c(\rho), & H(\theta) &= \mathrm{plim}_{N\to\infty} h(\theta). \end{aligned}$$

Using $M(y_i - Z_i\theta) = -MZ_i(\theta - \theta_0) + M\varepsilon_i$, we have

$$L(\theta) = -\frac{1}{2}\log\left(\mathrm{plim}_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\left(\varepsilon_i' M\varepsilon_i - 2(\theta-\theta_0)' Z_i' M\varepsilon_i + (\theta-\theta_0)' Z_i' M Z_i(\theta-\theta_0)\right)\right) + \mathrm{c}.$$

Let $b_0 = b(\rho_0) = S(\theta_0)$ and note that

$$\mathrm{plim}_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} Z_i' M\varepsilon_i = \left(\mathrm{plim}_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i' M\varepsilon_i\right) b_0 = \sigma_0^2\,(T-1)\, b_0.$$

Hence, defining $V_0 = V(\theta_0)$ by

$$\mathrm{plim}_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} Z_i' M Z_i = \left(\mathrm{plim}_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i' M\varepsilon_i\right) V_0 = \sigma_0^2\,(T-1)\, V_0,$$

we can write

$$L(\theta) = -\frac{1}{2}\log\left(1 - 2(\theta-\theta_0)' b_0 + (\theta-\theta_0)' V_0(\theta-\theta_0)\right) + \mathrm{c}$$

by absorbing the term $-\frac{1}{2}\log\left(\sigma_0^2\,(T-1)\right)$ into c. As $N \to \infty$, the ML estimator of θ converges in probability to $\theta_{ML} = \arg\max_\theta L(\theta) = \theta_0 + V_0^{-1} b_0$ and has asymptotic bias $V_0^{-1} b_0$. This expression generalizes the bias calculations in Nickell (1981). Note that $(\theta_0-\theta_{ML})'V_0(\theta_0-\theta_{ML}) = b_0'V_0^{-1}b_0$. Furthermore,

$$
\begin{aligned}
L(\theta) &= -\frac{1}{2}\log\left(1-b_0'V_0^{-1}b_0+(\theta-\theta_{ML})'V_0(\theta-\theta_{ML})\right)+\mathrm{c},\\
S(\theta) &= -\frac{V_0(\theta-\theta_{ML})}{1-b_0'V_0^{-1}b_0+(\theta-\theta_{ML})'V_0(\theta-\theta_{ML})},\\
H(\theta) &= -\frac{V_0}{1-b_0'V_0^{-1}b_0+(\theta-\theta_{ML})'V_0(\theta-\theta_{ML})}+2S(\theta)S(\theta)'.
\end{aligned}
$$

Note that $L(\cdot)$ and $H(\cdot)$ are even and $S(\cdot)$ is odd about θ_{ML} and that $H(\theta_0) = 2b_0b_0' - V_0$ and $H_{\mathrm{A}}(\theta_0) = 2b_0b_0' - V_0 - c_0$, where $c_0 = c(\rho_0)$. Since $L(\theta)$ is log-quadratic in θ and $a(\rho)$ is a multivariate polynomial with negative coefficients, $L_{\mathrm{A}}(\theta) = L(\theta) - a(\rho)$ is unbounded from above. For example, if we put $\rho = kr$ with r in the positive orthant of $\mathbb{R}^p$ and let $k\to\infty$, then the term $-a(\rho)$ dominates and $L_{\mathrm{A}}(\theta)\to\infty$.[7] Hence, $\theta_0 \neq \arg\max_\theta L_{\mathrm{A}}(\theta)$ and θ_0 has to be identified as a functional of $L_{\mathrm{A}}(\theta)$ other than its maximizer. Because $S_{\mathrm{A}}(\theta_0) = 0$, we need to select θ_0 from the set of stationary points of $L_{\mathrm{A}}(\theta)$, that is, from the set of zeros of $S_{\mathrm{A}}(\theta)$. In general, this set is not a singleton. Indeed, whenever θ_0 is a local maximizer of $L_{\mathrm{A}}(\theta)$ (which will often be the case), $L_{\mathrm{A}}(\theta)$, being smooth and unbounded, must also have at least one local minimum. A moment-based estimation strategy based solely on solving $s_{\mathrm{A}}(\theta) = 0$ is, therefore, incomplete.

3.1 AR(1) without covariates

Consider the AR(1) model without covariates, i.e., $p = 1$ and $q = 0$. Let $\zeta_0^2 = (V_0 - b_0^2)/V_0^2$. Then,

$$
\begin{aligned}
L(\rho) &= -\frac{1}{2}\log\left(\zeta_0^2+(\rho-\rho_{ML})^2\right)+\mathrm{c},\\
S(\rho) &= -\frac{\rho-\rho_{ML}}{\zeta_0^2+(\rho-\rho_{ML})^2}, \qquad H(\rho) = -\frac{\zeta_0^2-(\rho-\rho_{ML})^2}{\left(\zeta_0^2+(\rho-\rho_{ML})^2\right)^2},
\end{aligned}
$$

by absorbing $-\frac{1}{2}\log V_0$ into c. Note that $\zeta_0^2 = -1/H(\rho_{ML})$. Recall that $S(\rho)$ is odd about $\rho_{ML} = \rho_0 + b_0/V_0$. The zeros of $H(\rho)$ are $\underline{\rho} = \rho_{ML} - \zeta_0$ and $\overline{\rho} = \rho_{ML} + \zeta_0$, so $S(\rho)$ decreases on $[\underline{\rho}, \overline{\rho}]$ and increases elsewhere. Note that all of $\underline{\rho}$, $\overline{\rho}$, ρ_{ML}, and ζ_0 are identified by $S(\cdot)$. Further, ρ_{ML} and ζ_0 act as location and scale parameters of $S(\cdot)$. For any given ρ_0, ρ_{ML} and ζ_0 are determined by V_0. As V_0 increases, $|b_0/V_0|$ and ζ_0 decrease, that is, the bias of ρ_{ML} decreases in absolute value, the length of $[\underline{\rho}, \overline{\rho}]$ shrinks, and $S(\rho)$ becomes steeper on $(\underline{\rho}, \overline{\rho})$.

[7] It has already been observed that, in finite samples, a bias-adjusted likelihood may be re-increasing. Moreover, Lancaster (2002) presented graphs of a re-increasing marginal posterior log-density for ρ in an AR(1) model with time dummies. Arellano and Bonhomme (2009), using a slightly different prior, presented a graph of a re-increasing integrated likelihood for an AR(1) model. Because $l(\theta)$ is log-quadratic for any $N \geq 1$ and $a(\rho)$ does not depend on the data, $l_{\mathrm{A}}(\theta)$ is re-increasing regardless of the sample size.

There is a sharp lower bound on V_0. From $y_{i,-1} = S_1(\xi_{0i} + F_0\varepsilon_i)$ and the independence between ξ_{0i} and ε_i, we obtain

$$V_0 = \frac{\text{plim}_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} y_{i,-1}' M y_{i,-1}}{\sigma_0^2 (T-1)} = V_0^{LB} + V_{\xi\xi},$$

where

$$V_0^{LB} = \frac{\text{tr}F_0' S_1' M S_1 F_0}{T-1}, \qquad V_{\xi\xi} = \frac{\text{plim}_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} \xi_{0i}' S_1' M S_1 \xi_{0i}}{\sigma_0^2 (T-1)}.$$

So $V_0 \geq V_0^{LB}$ and this lower bound implies an upper bound on $|b_0/V_0|$ and on the length of $[\underline{\rho}, \overline{\rho}]$, and a lower bound on the steepness of $S(\rho)$ on $(\underline{\rho}, \overline{\rho})$.

Lemma 1. *V_0^{LB} is given by*

$$V_0^{LB} = \frac{1}{T-1}\left(\sum_{j=0}^{T-2}(T-j-1)\,\rho_0^{2j} - \frac{1}{T}\sum_{j=0}^{T-2}\left(\sum_{k=0}^{j}\rho_0^k\right)^2\right)$$

and satisfies (i) $V_0^{LB} \geq 2b_0^2$; (ii) $V_0^{LB} \geq 2b_0^2 - c_0$ with equality if and only if $T = 2$ or $\rho_0 = 1$.

By Lemma 1, $H(\rho_0) = 2b_0^2 - V_0 \leq 0$ and, hence,

$$(\overline{\rho} - \rho_{ML})^2 = \frac{V_0 - b_0^2}{V_0^2} \geq \frac{b_0^2}{V_0^2} = (\rho_0 - \rho_{ML})^2.$$

Therefore, $\rho_0 \in [\underline{\rho}, \overline{\rho}]$. Since $S(\rho)$ is a rational function that vanishes at $\pm\infty$ and $b(\rho)$ is a polynomial, $S_{\text{A}}(\rho)$ has finitely many zeros. Thus, because $S_{\text{A}}(\rho_0) = 0$ and, by Lemma 1, $H_{\text{A}}(\rho_0) = 2b_0^2 - V_0 - c_0 \leq 0$, it follows that $L_{\text{A}}(\rho)$ has a local maximum or a flat inflection point at ρ_0. Our main result for the AR(1) model without covariates is the uniqueness of such a point in $[\underline{\rho}, \overline{\rho}]$, thereby identifying ρ_0 as a functional of $L_{\text{A}}(\rho)$.

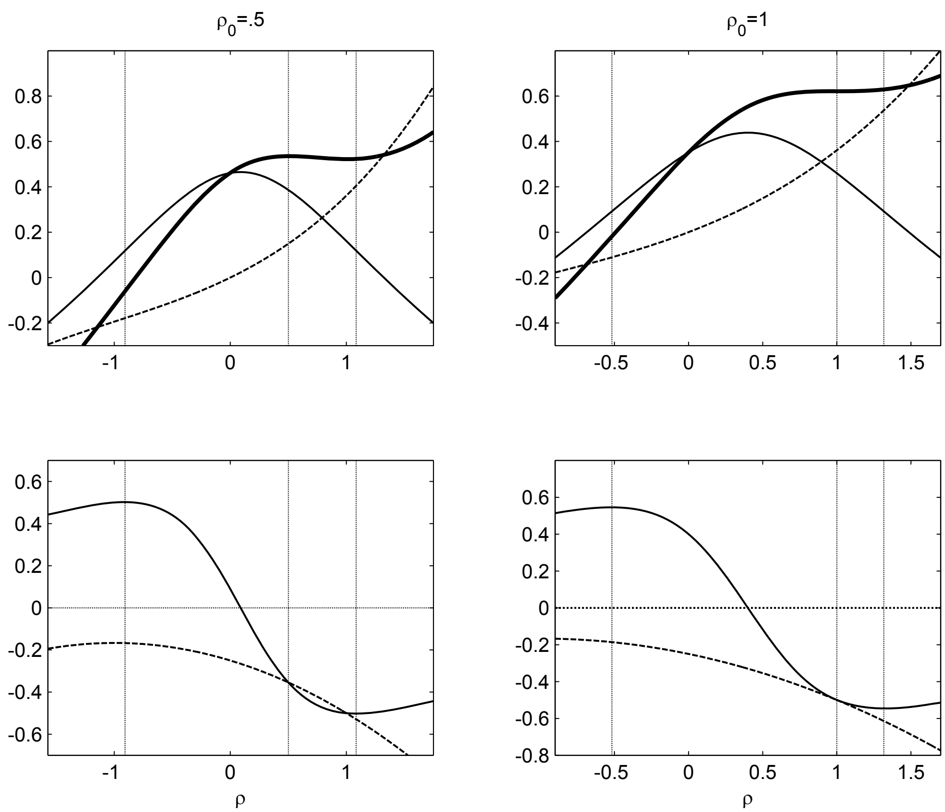
Theorem 1. *ρ_0 is the unique point in $[\underline{\rho}, \overline{\rho}]$ where $L_{\text{A}}(\rho)$ has a local maximum or a flat inflection point.*

Equivalently, ρ_0 is the unique point in $[\underline{\rho}, \overline{\rho}]$ where $b(\rho)$ approaches $S(\rho)$ from below.

$L_{\text{A}}(\rho)$ has a flat inflection point at ρ_0 if and only if $V_0 = V_0^{LB} = 2b_0^2 - c_0$. The latter equality holds if and only if $T = 2$ or $\rho_0 = 1$. The former holds if and only if $V_{\xi\xi} = 0$, which requires $MS_1\xi_{0i}$ to be negligibly small for almost all i. The elements of $S_1\xi_{0i}$ are $\rho_0^{j-1}y_i^0 + \alpha_i\sum_{k=1}^{j-1}\rho_0^{k-1}$, $j = 1, ..., T$, so $MS_1\xi_{0i} = 0$ if and only if $y_i^0(1-\rho_0) = \alpha_i$. Thus, when $\rho_0 = 1$ and there are no deterministic trends, i.e., $\alpha_i = 0$, $L_{\text{A}}(\rho)$ has a flat inflection point at ρ_0 for any T. When $\rho_0 \neq 1$, $V_0 = V_0^{LB} = 2b_0^2 - c_0$ only when $T = 2$ and a very strong condition holds on the initial observations and the fixed effects, which is unlikely to hold in situations where a fixed effect modeling approach is called for. Thus, when $\rho_0 \neq 1$, except in quite special circumstances, ρ_0 is the unique point in $[\underline{\rho}, \overline{\rho}]$ where $L_{\mathrm{A}}(\rho)$ attains a strict local maximum. Note that, when ρ_0 is a local maximizer of $L_{\mathrm{A}}(\rho)$, it need not be the global maximizer on $[\underline{\rho}, \overline{\rho}]$, which may instead be $\overline{\rho}$. To see why this may happen, interpret the situation where $L_{\mathrm{A}}(\rho)$ has a flat inflection point at ρ_0 as a limiting case of the property that $L_{\mathrm{A}}(\rho)$ is re-increasing.

Figure 1 illustrates how ρ_0 is identified by $L_{\mathrm{A}}(\rho)$ for two cases, each with $T = 4$. The figures on the left correspond to the case $\rho_0 = .5$ with $V_0 = V_0^{LB} + V_{\xi\xi}$ and $V_{\xi\xi}$ corresponding to stationary initial observations. Those on the right correspond to the unit root case without deterministic trends, i.e., $\rho_0 = 1$ and $V_0 = V_0^{LB}$. In each case, the bottom figures show $S(\rho)$ (solid line) and $b(\rho)$ (dashed line); the top figures show $L(\rho)$ (solid line), $-a(\rho)$ (dashed line), and $L_{\mathrm{A}}(\rho) = L(\rho) - a(\rho)$ (thick line). In all the figures, vertical lines indicate $\underline{\rho}$, ρ_0, and $\overline{\rho}$, from left to right. In the case of $\rho_0 = .5$, ρ_0 is the unique local maximizer of $L_{\mathrm{A}}(\rho)$ on $[\underline{\rho}, \overline{\rho}]$. Note that there is a second solution of $S_{\mathrm{A}}(\rho) = 0$ on $[\underline{\rho}, \overline{\rho}]$, which corresponds to a local minimum of $L_{\mathrm{A}}(\rho)$. In the unit root case, ρ_0 is the unique flat inflection point of $L_{\mathrm{A}}(\rho)$ on $[\underline{\rho}, \overline{\rho}]$.

Figure 1: Identification



Left: $\rho_0 = 0.5$. Right: $\rho_0 = 1$. Bottom: $S(\rho)$ (solid), $b(\rho)$ (dashed). Top: $L(\rho)$ (solid), $-a(\rho)$ (dashed), $L_{\mathrm{A}}(\rho)$ (thick). Vertical lines at $\underline{\rho}$, ρ_0, and $\overline{\rho}$.

Note, finally, that the asymptotic bias of the MLE has the same sign as b_0 because $\rho_{ML} =$

$\rho_0 + b_0/V_0$. The proof of Theorem 1, as a by-product, shows that if T is even, then $b_0 < 0$; and, if T is odd, then $b(\rho)$ decreases and has a unique zero at some point $\rho_u \in [-2, -1)$, so b_0 has the same sign as $\rho_u - \rho_0$.

3.2 AR(1) with covariates

In the AR(1) model with covariates, profiling out β yields a profile likelihood of ρ with essentially the same properties as in the AR(1) model without covariates. Let $\beta(\rho) = \arg\max_\beta L_{\mathrm{A}}(\rho, \beta)$. Clearly, $\beta(\rho) = \arg\max_\beta L(\rho, \beta) = \arg\min_\beta (\theta - \theta_{ML})' V_0 (\theta - \theta_{ML})$. Partition V_0, V_0^{-1}, and b_0 as

$$V_0 = \begin{pmatrix} V_{0\rho\rho} & V_{0\rho\beta} \\ V_{0\beta\rho} & V_{0\beta\beta} \end{pmatrix}, \qquad V_0^{-1} = \begin{pmatrix} V_0^{\rho\rho} & V_0^{\rho\beta} \\ V_0^{\beta\rho} & V_0^{\beta\beta} \end{pmatrix}, \qquad b_0 = \begin{pmatrix} b_{0\rho} \\ 0 \end{pmatrix},$$

where $V_{0\rho\rho}$, $V_0^{\rho\rho}$, and $b_{0\rho}$ are scalars. Using $V_0^{\rho\rho} = (V_{0\rho\rho} - V_{0\rho\beta} V_{0\beta\beta}^{-1} V_{0\beta\rho})^{-1}$, we have

$$\begin{aligned} V_{0\beta\beta}\,(\beta(\rho) - \beta_{ML}) &= -V_{0\beta\rho}(\rho - \rho_{ML}), \\ \min_\beta (\theta - \theta_{ML})' V_0 (\theta - \theta_{ML}) &= (\rho - \rho_{ML})^2 / V_0^{\rho\rho}, \\ 1 - b_0' V_0^{-1} b_0 &= 1 - b_{0\rho}^2 V_0^{\rho\rho}. \end{aligned}$$

The first of these equations, together with $V_0(\theta_0 - \theta_{ML}) = -b_0$, yields $\beta(\rho_0) = \beta_0$, so β_0 is identified whenever ρ_0 is. Profiling out β from $L(\rho, \beta)$ gives the limiting profile log-likelihood of ρ as

$$L(\rho) = L(\rho, \beta(\rho)) = -\frac{1}{2} \log\left(\zeta_0^2 + (\rho - \rho_{ML})^2\right) + \mathrm{c}$$

(slightly abusing notation), where ζ_0^2 is redefined as $\zeta_0^2 = \left(1 - b_{0\rho}^2 V_0^{\rho\rho}\right) V_0^{\rho\rho}$ and $\frac{1}{2}\log V_0^{\rho\rho}$ is absorbed into c.

Lemma 2. *$(V_0^{\rho\rho})^{-1} \geq V_0^{LB}$, with V_0^{LB} as defined earlier and given in Lemma 1.*

In view of Lemma 2, we can invoke the result for the AR(1) model without covariates. If $\underline{\rho} = \rho_{ML} - \zeta_0$ and $\overline{\rho} = \rho_{ML} + \zeta_0$ with ζ_0 redefined as indicated, then ρ_0 is the unique point in $[\underline{\rho}, \overline{\rho}]$ where $L_{\mathrm{A}}(\rho) = L(\rho) - a(\rho)$ has a local maximum or a flat inflection point. By the proof of Lemma 2, the conditions under which ρ_0 is a flat inflection point of $L_{\mathrm{A}}(\rho)$ are the same as in the AR(1) model without covariates. Note, also, that the presence of covariates does not affect the sign of the asymptotic bias of the MLE of ρ. Furthermore, it follows from the proof of Lemma 2 that the inclusion of covariates in the model cannot increase $V_0^{\rho\rho}$, so the magnitude of $\rho_{ML} - \rho_0 = V_0^{\rho\rho} b_{0\rho}$ can only decrease relative to the model without covariates. This generalizes an observation of Phillips and Sul (2007) in the stationary AR(1) model, and contrasts with Nickell (1981, p. 1424).

3.3 AR(p)

Consider first the AR(p) model with $p > 1$ and without covariates, i.e., $q = 0$. Then

$$\begin{aligned}
L(\rho) &= -\frac{1}{2}\log\left(1 + (\rho - \rho_{ML})' W_0 (\rho - \rho_{ML})\right) + \mathrm{c}, \qquad W_0 = \frac{V_0}{1 - b_0' V_0^{-1} b_0}, \\
S(\rho) &= -\frac{W_0(\rho - \rho_{ML})}{1 + (\rho - \rho_{ML})' W_0 (\rho - \rho_{ML})}, \\
H(\rho) &= -\frac{W_0}{1 + (\rho - \rho_{ML})' W_0 (\rho - \rho_{ML})} + 2S(\rho)S(\rho)',
\end{aligned}$$

where $-\frac{1}{2}\log(1 - b_0' V_0^{-1} b_0)$ is absorbed into c. Note that $W_0 = -H(\rho_{ML})$, so W_0 is identified by $L(\cdot)$.

As in the AR(1) case, there is a lower bound on V_0. Recalling that $Y_{i-} = (y_{i,-1}, ..., y_{i,-p})$ and $y_{i,-j} = S_j(\xi_{0i} + F_0 \varepsilon_i)$, where ξ_{0i} and ε_i are independent, we have

$$V_0 = \frac{\operatorname{plim}_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} Y_{i-}' M Y_{i-}}{\sigma_0^2 (T-1)} = V_0^{LB} + V_{\xi\xi}$$

where V_0^{LB} and $V_{\xi\xi}$ have elements

$$\left(V_0^{LB}\right)_{jk} = \frac{\mathrm{tr} F_0' S_j' M S_k F_0}{T-1}, \qquad \left(V_{\xi\xi}\right)_{jk} = \frac{\operatorname{plim}_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \xi_{0i}' S_j' M S_k \xi_{0i}}{\sigma_0^2 (T-1)},$$

for $1 \leq j, k \leq p$. Hence, $V_0 - V_0^{LB}$ is positive semi-definite, which we write as $V_0 \geq V_0^{LB}$. When $p \geq T$, while V_0 is nonsingular by assumption, $\mathrm{rank}(V_0^{LB}) \leq T - 1$ because $S_j F_0 = 0$ for $j \geq T$, which implies that $\left(V_0^{LB}\right)_{jk} = 0$ whenever $j \geq T$ or $k \geq T$. Thus, when $p \geq T$, although V_0 can be arbitrarily close to V_0^{LB}, $V_0 \neq V_0^{LB}$. Further, when $p \geq T$, $b_j(\rho) = 0$ for $j \geq T$ because the sum defining $b_j(\rho)$ is empty, and $c_{ij}(\rho) = 0$ for $i + j \geq T$. Hence, when $p \geq T$, $V_0^{LB} - 2b_0 b_0'$ and $V_0^{LB} - 2b_0 b_0' + c_0$ have only zeros beyond their leading $(T-1) \times (T-1)$ blocks.

A proof of generalizations of (i)–(ii) of Lemma 1 and Theorem 1 to the AR(p) model would be desirable but is more difficult than in the AR(1) case and our attempts failed.[8] Numerical computations, however, suggest that

$$V_0^{LB} \geq 2b_0 b_0', \qquad V_0^{LB} \geq 2b_0 b_0' - c_0, \tag{3.1}$$

$$\mathrm{rank}(V_0^{LB} - 2b_0 b_0' + c_0) = \begin{cases} \min(p, T-2) & \text{if } \sum_{j=1}^{p} \rho_{0j} \neq 1 \text{ or } T < p + 2, \\ p - 1 & \text{else.} \end{cases} \tag{3.2}$$

[8] A major difficulty is the rapidly increasing complexity of φ_t as p increases. For example, $\varphi_t = \sum_{k=0}^{\lfloor t/2 \rfloor} \frac{(t-k)!}{(t-2k)!k!} \rho_1^{t-2k} \rho_2^k$ when $p = 2$. Compare this with $\varphi_t = \rho_1^t$ when $p = 1$ and note that V_0^{LB}, b_0, and c_0 depend in non-trivial ways on the φ_t's.

Specifically, we computed the eigenvalues of $V_0^{LB} - 2b_0b_0'$ and $V_0^{LB} - 2b_0b_0' + c_0$ for $p = 2, 3, 4$; $T = 2, ..., 7$; and all ρ_0 in a subset of $\mathbb{R}^p$ chosen as follows. For $p = 4$, we put a square grid on the Cartesian product of the two triangles defined by

$$\begin{array}{ll} -1 \leq \gamma_2 \leq 1, & \gamma_2 - 1 \leq \gamma_1 \leq 1 - \gamma_2, \\ -1 \leq \gamma_4 \leq 1, & \gamma_4 - 1 \leq \gamma_3 \leq 1 - \gamma_4, \end{array} \tag{3.3}$$

which is the stationary region of the lag polynomial $\gamma(L) = (1 - \gamma_1 L - \gamma_2 L^2)(1 - \gamma_3 L - \gamma_4 L^2)$. For each point on this grid and for each of the values $m = 1, 2, 4$, ρ_0 was calculated by equating the coefficients on both sides of $m - \rho_{01}L - \rho_{02}L^2 - \rho_{03}L^3 - \rho_{04}L^4 = m\gamma(L)$. For $m = 1$, the stationary region is covered, while for larger m a larger region is covered, though less densely. In addition to (3.3) we set $\gamma_4 = 0$ for $p = 3$, and $\gamma_3 = \gamma_4 = 0$ for $p = 2$. The grid points on the region defined by (3.3) were spaced at intervals of .002 when $p = 2$, .02 when $p = 3$, and .1 when $p = 4$. We found that, uniformly over this numerical design, the eigenvalues of $V_0^{LB} - 2b_0b_0'$ and $V_0^{LB} - 2b_0b_0' + c_0$ are non-negative and the rank of $V_0^{LB} - 2b_0b_0' + c_0$ is as given by (3.2). These findings, while obviously not a proof, support (3.1) and (3.2), and we shall proceed under the assumption that (3.1) and (3.2) hold.[9]

Because $V_0 \geq V_0^{LB}$, (3.1) implies that $V_0 \geq 2b_0b_0'$ and that $H_{\mathrm{A}}(\rho_0) = 2b_0b_0' - V_0 - c_0 \leq 0$. Pre- and postmultiplication of $V_0 \geq 2b_0b_0'$ by $b_0'V_0^{-1}$ and $V_0^{-1}b_0$ gives $b_0'V_0^{-1}b_0 \leq \frac{1}{2} \leq 1 - b_0'V_0^{-1}b_0$. Recalling that $(\rho_0 - \rho_{ML})'V_0(\rho_0 - \rho_{ML}) = b_0'V_0^{-1}b_0$, we see that

$$(\rho_0 - \rho_{ML})'W_0(\rho_0 - \rho_{ML}) \leq 1.$$

Therefore, if (3.1) and (3.2) hold, ρ_0 is a point in the ellipsoidal disk $\mathcal{E} = \{\rho : (\rho - \rho_{ML})'W_0(\rho - \rho_{ML}) \leq 1\}$ where $L_{\mathrm{A}}(\rho)$ has a local maximum or a flat inflection point. We approached the question of uniqueness of such a point numerically. For the same numerical design as above and with $V_0 = V_0^{LB}$, we applied the Newton-Raphson algorithm to find a stationary point of $L_{\mathrm{A}}(\rho)$, starting at ρ_{ML} and using the Moore-Penrose inverse of $H_{\mathrm{A}}(\rho)$ whenever $H_{\mathrm{A}}(\rho)$ is singular. Uniformly over this design, the algorithm was found to converge to ρ_0, thus supporting the conjecture that ρ_0 is the unique point in $\mathcal{E}$ where $L_{\mathrm{A}}(\rho)$ has a local maximum or a flat inflection point.[10]

In the AR(p) model with covariates, β can be profiled out of $L_{\mathrm{A}}(\theta)$ just as in the AR(1) model. Here, again, $\beta_0 = \beta(\rho_0)$. Lemma 2 continues to hold for $p > 1$. Hence, if ρ_0 is identified in the model without covariates in the way we suggested, then it is identified in the model with covariates in exactly the same way, now with $\mathcal{E}$ defined through $W_0 = (1 - b_{0\rho}'V_0^{\rho\rho}b_{0\rho})^{-1}V_{0\rho\rho}$, in obvious notation.

[9] The same computations but with $T = 8, 9, 10$ further supported the conclusions. Here, however, when $m = 4$ and $p = 3, 4$ the computations are numerically less stable because the polynomial terms may be extremely large and their sum numerically imprecise. It is natural to expect, however, that (3.1) and (3.2) hold for all T if they hold for the small values of T considered.

[10] Computations with $T = 8, 9, 10$ gave the same results except in certain cases with $m = 4$ and $p = 3, 4$ where the algorithm failed to converge because the rank of $H_{\mathrm{A}}(\rho)$ was underestimated.

4 Estimation

For a given ρ, we define

$$\widehat{\beta}_{\mathrm{AL}}(\rho) = \arg\max_{\beta} l_{\mathrm{A}}(\rho, \beta) = \left(\sum_{i=1}^{N} X_i' M X_i\right)^{-1} \sum_{i=1}^{N} X_i' M \left(y_i - Y_{i-}\rho\right)$$

as the adjusted likelihood (AL) estimator of β_0. (Note that $\widehat{\beta}_{\mathrm{AL}}(\rho)$ coincides with the MLE of β_0 for a given ρ.) The unadjusted and adjusted profile log-likelihoods for ρ are $l(\rho) = l(\rho, \widehat{\beta}_{\mathrm{AL}}(\rho))$ and $l_{\mathrm{A}}(\rho) = l(\rho) - a(\rho)$. Let $s(\rho)$, $s_{\mathrm{A}}(\rho)$, $h(\rho)$, and $h_{\mathrm{A}}(\rho)$ be the corresponding profile scores and Hessians. Let $\widehat{W} = -h(\widehat{\rho}_{ML})$, where $\widehat{\rho}_{ML}$ is the MLE of ρ_0, and let $\widehat{\mathcal{E}} = \{\rho : (\rho - \widehat{\rho}_{ML})'\widehat{W}(\rho - \widehat{\rho}_{ML}) \leq 1\}$. We define the AL estimator of ρ_0 as

$$\widehat{\rho}_{\mathrm{AL}} = \arg\min_{\rho \in \widehat{\mathcal{E}}} s_{\mathrm{A}}'(\rho) s_{\mathrm{A}}(\rho) \quad \text{s.t.} \quad h_{\mathrm{A}}(\rho) \leq 0,$$

that is, as the strict local maximizer of $l_{\mathrm{A}}(\rho)$ on the interior of $\widehat{\mathcal{E}}$ if such a maximizer exists and otherwise as the minimizer of the norm of $s_{\mathrm{A}}(\rho)$ on $\widehat{\mathcal{E}}$. The AL estimator of β_0, then, is $\widehat{\beta}_{\mathrm{AL}} = \widehat{\beta}_{\mathrm{AL}}(\widehat{\rho}_{\mathrm{AL}})$.

Let $N \to \infty$. Then, $l_{\mathrm{A}}(\rho)$ converges to $L_{\mathrm{A}}(\rho)$ uniformly in ρ since $-a(\rho)$ is nonstochastic and $\sup_{\rho} |l(\rho) - L(\rho)| = o_p(1)$. Further, $\widehat{\rho}_{ML} \xrightarrow{p} \rho_{ML}$, $\widehat{W} \xrightarrow{p} -H(\rho_{ML}) = W_0$, and $\widehat{\mathcal{E}} \xrightarrow{p} \mathcal{E}$ in the sense that $\Pr[\rho \in \widehat{\mathcal{E}}] \to 1_{\{\rho \in \mathcal{E}\}}$ for any ρ not on the boundary of $\mathcal{E}$. It follows that $\widehat{\theta}_{\mathrm{AL}} = (\widehat{\rho}_{\mathrm{AL}}', \widehat{\beta}_{\mathrm{AL}}')' \xrightarrow{p} \theta_0$. When $H_{\mathrm{A}}(\theta_0)$ is nonsingular, by the usual argument involving a Taylor series expansion of $s_{\mathrm{A}}(\theta)$ around θ_0,

$$\sqrt{N}(\widehat{\theta}_{\mathrm{AL}} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Omega), \qquad \Omega = H_{\mathrm{A}}(\theta_0)^{-1}(V_0 - b_0 b_0') H_{\mathrm{A}}(\theta_0)^{-1}. \tag{4.1}$$

The asymptotic variance can be estimated in the usual way. We have not investigated the limit distribution of $\widehat{\theta}_{\mathrm{AL}}$ in the situation where $H_{\mathrm{A}}(\theta_0)$ is singular. Presumably this could be done along the lines of Rotnitzky, Cox, Bottai, and Robins (2000).

5 Simulations

We used simulations to examine the finite sample properties of the AL estimator in AR(1) and AR(2) models without covariates and in an AR(1) model with a single stationary and exogenous regressor. We compared the AL with the one-step GMM estimator of Arellano and Bond (1991), which leaves the initial observations unrestricted.[11] In all the designs, we set $N = 100$, generated

[11] For the GMM estimator we used all the moments that follow from the exogeneity of x_{it}, i.e. $\mathbb{E}\left[x_{ij}(\varepsilon_{it} - \varepsilon_{it-1})\right] = 0, j = 1, ..., T; t = 2, ..., T$; and $\mathbb{E}\left[y_{it-j}(\varepsilon_{it} - \varepsilon_{it-1})\right] = 0, j = 2, ..., t; t = 2, ..., T$. It should be noted that the GMM estimator remains consistent under heteroskedasticity while the bias calculation underlying the AL estimator requires time series homoskedasticity.

ε_{it} and α_i as $\mathcal{N}(0,1)$ variates,[12] and chose ρ_0 in the interior of the stationary region, which implies that y_{it} is eventually stationary as $t \to \infty$. We varied the information content of the data through the initial observations. Let $\mu_i = \lim_{t\to\infty} \mathbb{E}(y_{it-}|\alpha_i)$ and $\Sigma_i = \lim_{t\to\infty} \mathrm{Var}(y_{it-}|\alpha_i)$, so, if y_i^0 was drawn from the stationary distribution, we would just have $\mu_i = \mathbb{E}(y_i^0|\alpha_i)$ and $\Sigma_i = \mathrm{Var}(y_i^0|\alpha_i)$. Let $G_i G_i' = \Sigma_i$ be the Cholesky factorization of Σ_i. We set $y_i^0 = \mu_i + \psi G_i \iota$ for some chosen scalar $\psi \geq 0$, which is a p-variate version of setting the initial observations ψ standard deviations away from the stationary mean. So ψ controls the outlyingness of the initial observations relative to the stationary distributions. All else being equal, V_0 increases in ψ and $V_0 \to V_0^{LB}$ as $\psi \to 0$, so the data carry less information as ψ gets smaller. The effect of strong inlying observations (small ψ) on the informativeness of the data is stronger when T is small because it takes time to revert to the stationary distribution. The effect of ψ is vanishingly small as ρ_0 moves to the boundary of the stationary region.[13] We set $\psi = 0, 1, 2$ in the AR(1) models and $\psi = .3, 1, 2$ in the AR(2) model.[14]

In the models without a covariate, μ_i and Σ_i follow immediately from α_i and ρ_0. In the AR(1) model with a covariate, x_{it} was generated by a stationary AR(1), $x_{it} = \delta\alpha_i + \gamma x_{it-1} + u_{it}$ with $u_{it} \sim \mathcal{N}(0, \sigma_u^2)$ and x_{i0} drawn from the stationary distribution. Here,

$$\mu_i = \frac{\alpha_i}{1-\rho_0}\left(1 + \frac{\delta\beta_0}{1-\gamma}\right), \qquad \Sigma_i = \frac{1}{1-\rho_0^2}\left(1 + \frac{\beta_0^2}{1-\gamma^2}\left(\frac{1+\gamma\rho_0}{1-\gamma\rho_0}\right)\sigma_u^2\right).$$

We set $\delta = \gamma = \sigma_u = .5$ and $\beta_0 = 1 - \rho_0$, inducing dependence between the covariate and the fixed effect, and keeping the long-run multiplier of x on y constant at unity across designs, as in Kiviet (1995).

We varied ψ, T, and ρ_0. Tables 1–3 present Monte Carlo estimates, based on 10,000 replications at each design point, of the bias and standard deviation (STD) of the AL and GMM estimators and of the coverage rates of the asymptotic and bootstrap 95% confidence intervals ($\mathrm{CI}^A_{.95}$ and $\mathrm{CI}^B_{.95}$).[15]

In the AR(1) model (Table 1), for all designs $\widehat{\rho}_{\mathrm{AL}}$ has less bias and less standard deviation than $\widehat{\rho}_{\mathrm{GMM}}$. When $\rho_0 = .5$, the AL estimator is virtually unbiased except when $\psi = 0$ and $T = 2$, in which case ρ_0 is weakly identified.[16] Both estimators deliver 95% confidence intervals with broadly correct coverage, although the coverage errors are somewhat larger for $\widehat{\rho}_{\mathrm{GMM}}$,

[12] The AL estimator is invariant with respect to the fixed effects, but the GMM estimator is not.

[13] In the AR(1) model without covariates, for example, $S_1\xi_{0i} = g_0(y_{i0} - \mu_i) + \iota\mu_i$ with $g_0 = (1, \rho_0, ..., \rho_0^{T-1})'$, hence $V_{\xi\xi} = \frac{g_0' M g_0}{(1-\rho_0^2)(T-1)}\psi^2$. As $\rho_0 \uparrow 1$, $V_{\xi\xi} \to 0$ for any fixed ψ.

[14] In the AR(2) model, when $\psi = 0$ the weighting matrix used by GMM is singular for all T and the likelihood function does not depend on ρ_2 for $T = 2$. Therefore, we set $\psi > 0$ although it is possible to work with $\psi = 0$ for $T \geq 3$ by dropping the earliest observation from the set of instruments.

[15] To compute $\mathrm{CI}^B_{.95}$, we used the bootstrap percentile method with 39 bootstrap samples formed by randomly drawing N units i with replacement from $\{1, ..., N\}$. A small number of bootstrap draws suffices for studying coverage rates via simulation.

[16] When $\psi = 0$ and $T = 2$, $l(\rho)$ depends on ρ_0 and ρ only through $\rho_0 - \rho$, so the distribution of $\widehat{\rho}_{\mathrm{AL}}$ is

Table 1: Simulations for the AR(1) model, $N = 100$

			bias		STD		$\text{CI}^A_{.95}$		$\text{CI}^B_{.95}$	
ψ	T	ρ_0	$\widehat{\rho}_{\text{AL}}$	$\widehat{\rho}_{\text{GMM}}$	$\widehat{\rho}_{\text{AL}}$	$\widehat{\rho}_{\text{GMM}}$	$\widehat{\rho}_{\text{AL}}$	$\widehat{\rho}_{\text{GMM}}$	$\widehat{\rho}_{\text{AL}}$	$\widehat{\rho}_{\text{GMM}}$
0	2	.5	−.146	–	.267	–	.880	.916	.924	.976
1	2	.5	.032	–	.269	–	.944	.934	.944	.965
2	2	.5	.029	–	.173	–	.967	.941	.942	.944
0	4	.5	.006	−.043	.142	.150	.954	.921	.947	.911
1	4	.5	.014	−.057	.124	.163	.965	.927	.944	.906
2	4	.5	.002	−.015	.064	.083	.968	.940	.943	.934
0	8	.5	.000	−.028	.056	.058	.966	.914	.946	.860
1	8	.5	.000	−.039	.048	.070	.960	.906	.944	.842
2	8	.5	−.001	−.022	.036	.052	.953	.925	.946	.887
0	24	.5	.000	−.018	.021	.022	.944	.871	.944	.726
1	24	.5	−.001	−.022	.020	.025	.946	.859	.945	.700
2	24	.5	.000	−.020	.018	.023	.950	.868	.947	.721
0	2	.95	−.144	–	.266	–	.883	.919	.927	.978
1	2	.95	−.118	–	.268	–	.888	.924	.933	.979
2	2	.95	−.055	–	.267	–	.913	.928	.942	.979
0	4	.95	−.087	−.680	.124	.465	.894	.667	.903	.615
1	4	.95	−.063	−.696	.124	.464	.907	.688	.933	.627
2	4	.95	−.016	−.389	.123	.426	.936	.815	.946	.756
0	8	.95	−.043	−.345	.064	.176	.888	.438	.902	.155
1	8	.95	−.025	−.398	.063	.192	.914	.403	.942	.141
2	8	.95	.003	−.225	.063	.156	.952	.655	.948	.350
0	24	.95	−.006	−.090	.024	.033	.923	.157	.940	.006
1	24	.95	.000	−.127	.024	.042	.941	.063	.940	.004
2	24	.95	.001	−.091	.018	.038	.961	.166	.940	.022

‘–’ indicates non-existence of the corresponding population moment.

Table 2: Simulations for the AR(2) model, $N = 100$

			bias		STD		$\mathrm{CI}^A_{.95}$		$\mathrm{CI}^B_{.95}$	
ψ	T	ρ_0	$\widehat{\rho}_{\mathrm{AL}}$	$\widehat{\rho}_{\mathrm{GMM}}$	$\widehat{\rho}_{\mathrm{AL}}$	$\widehat{\rho}_{\mathrm{GMM}}$	$\widehat{\rho}_{\mathrm{AL}}$	$\widehat{\rho}_{\mathrm{GMM}}$	$\widehat{\rho}_{\mathrm{AL}}$	$\widehat{\rho}_{\mathrm{GMM}}$
.3	2	.6	−.141	–	.267	–	.882	.920	.930	.977
		.2	−.129	–	.748	–	.959	.945	.940	.983
1	2	.6	−.141	–	.267	–	.881	.916	.930	.977
		.2	−.124	–	.316	–	.909	.920	.920	.977
2	2	.6	−.147	–	.266	–	.880	.916	.927	.977
		.2	−.130	–	.258	–	.890	.918	.922	.978
.3	4	.6	−.071	−.279	.123	.280	.900	.782	.914	.698
		.2	−.033	−.125	.098	.133	.936	.808	.932	.763
1	4	.6	−.001	−.043	.123	.120	.945	.921	.945	.895
		.2	−.002	−.023	.094	.089	.959	.931	.950	.925
2	4	.6	.008	−.011	.082	.064	.967	.939	.946	.933
		.2	.003	−.007	.072	.064	.965	.943	.949	.942
.3	8	.6	−.016	−.102	.065	.088	.941	.773	.941	.536
		.2	−.009	−.053	.056	.056	.949	.833	.943	.695
1	8	.6	.006	−.033	.063	.055	.965	.897	.948	.816
		.2	.002	−.015	.052	.045	.963	.929	.947	.910
2	8	.6	.000	−.011	.037	.037	.953	.932	.947	.913
		.2	.000	−.003	.037	.036	.944	.943	.945	.942
.3	24	.6	.000	−.028	.024	.025	.956	.795	.949	.607
		.2	.000	−.018	.024	.023	.950	.869	.945	.755
1	24	.6	.000	−.020	.022	.023	.951	.859	.950	.732
		.2	.000	−.011	.022	.022	.945	.914	.946	.865
2	24	.6	.000	−.011	.020	.020	.950	.913	.951	.856
		.2	.000	−.003	.019	.019	.946	.944	.948	.939
.3	2	1	−.144	–	.265	–	.882	.918	.934	.977
		−.2	−.083	–	.640	–	.969	.968	.945	.992
1	2	1	−.145	–	.265	–	.880	.913	.927	.979
		−.2	−.073	–	.234	–	.928	.921	.931	.978
2	2	1	−.147	–	.266	–	.881	.918	.930	.977
		−.2	−.078	–	.165	–	.901	.918	.919	.979
.3	4	1	−.070	−.235	.118	.256	.902	.808	.906	.721
		−.2	.004	−.021	.087	.084	.966	.934	.945	.937
1	4	1	.009	−.030	.115	.098	.951	.930	.947	.908
		−.2	.006	−.006	.081	.078	.965	.939	.944	.943
2	4	1	.003	−.008	.059	.051	.966	.940	.945	.936
		−.2	.001	−.003	.060	.059	.949	.941	.945	.943
.3	8	1	−.009	−.068	.059	.072	.953	.830	.945	.653
		−.2	.002	−.015	.051	.045	.965	.926	.945	.911
1	8	1	.002	−.024	.051	.048	.966	.916	.947	.856
		−.2	.001	−.001	.042	.041	.950	.945	.944	.946
2	8	1	−.001	−.009	.033	.034	.948	.936	.946	.922
		−.2	.001	.002	.033	.034	.944	.945	.947	.947
.3	24	1	−.001	−.018	.022	.023	.950	.872	.950	.770
		−.2	.000	−.008	.022	.022	.947	.927	.946	.906
1	24	1	−.001	−.015	.021	.022	.948	.898	.948	.818
		−.2	.000	−.004	.021	.021	.947	.940	.947	.935
2	24	1	−.001	−.010	.019	.020	.950	.920	.947	.879
		−.2	.000	.000	.019	.019	.946	.947	.948	.949

'–' indicates non-existence of the corresponding population moment.

Table 3: Simulations for the AR(1) model with a covariate, $N = 100$

			bias		STD		$\mathrm{CI}^A_{.95}$		$\mathrm{CI}^B_{.95}$	
ψ	T	θ_0	$\widehat{\theta}_{\mathrm{AL}}$	$\widehat{\theta}_{\mathrm{GMM}}$	$\widehat{\theta}_{\mathrm{AL}}$	$\widehat{\theta}_{\mathrm{GMM}}$	$\widehat{\theta}_{\mathrm{AL}}$	$\widehat{\theta}_{\mathrm{GMM}}$	$\widehat{\theta}_{\mathrm{AL}}$	$\widehat{\theta}_{\mathrm{GMM}}$
0	2	.5	−.092	−.051	.266	–	.900	.882	.943	.945
		.5	−.027	−.015	.249	–	.965	.943	.945	.962
1	2	.5	.019	.001	.269	–	.939	.910	.947	.945
		.5	−.001	.000	.258	–	.968	.950	.939	.949
2	2	.5	.030	.002	.182	–	.966	.940	.949	.945
		.5	−.010	−.002	.261	–	.949	.941	.942	.941
0	4	.5	.014	−.093	.141	.119	.957	.845	.945	.737
		.5	.004	−.008	.127	.124	.959	.941	.947	.946
1	4	.5	.012	−.069	.119	.102	.968	.883	.948	.802
		.5	−.001	.010	.126	.123	.951	.943	.949	.947
2	4	.5	.001	−.024	.063	.062	.968	.921	.944	.890
		.5	.000	.009	.129	.128	.944	.940	.941	.941
0	8	.5	−.001	−.053	.051	.050	.965	.813	.948	.582
		.5	.001	.004	.074	.075	.943	.943	.946	.945
1	8	.5	−.001	−.047	.045	.047	.959	.837	.945	.627
		.5	.001	.011	.075	.076	.940	.938	.943	.937
2	8	.5	−.001	−.026	.035	.036	.950	.884	.939	.754
		.5	.001	.011	.075	.075	.948	.945	.949	.943
0	24	.5	−.001	−.032	.020	.020	.948	.646	.943	.323
		.5	.000	.007	.038	.038	.950	.945	.950	.941
1	24	.5	.000	−.031	.019	.020	.946	.652	.945	.343
		.5	.000	.009	.038	.038	.950	.942	.952	.937
2	24	.5	.000	−.026	.017	.018	.947	.700	.948	.416
		.5	.000	.009	.038	.038	.949	.943	.952	.935
0	2	.95	−.145	−.904	.266	–	.879	.809	.930	.926
		.05	−.007	−.030	.234	–	.974	.977	.947	.991
1	2	.95	−.124	−.975	.266	–	.891	.811	.937	.923
		.05	.004	.046	.240	–	.974	.977	.941	.989
2	2	.95	−.068	−.534	.264	–	.910	.842	.948	.936
		.05	.007	.063	.246	–	.972	.968	.943	.980
0	4	.95	−.086	−.641	.124	.253	.892	.267	.906	.123
		.05	−.002	−.013	.121	.115	.973	.944	.948	.955
1	4	.95	−.067	−.595	.123	.252	.906	.312	.930	.148
		.05	.004	.041	.122	.114	.973	.938	.948	.943
2	4	.95	−.020	−.338	.122	.207	.939	.561	.950	.346
		.05	.003	.056	.128	.120	.971	.916	.945	.917
0	8	.95	−.043	−.347	.064	.100	.888	.036	.903	.001
		.05	.000	−.006	.073	.075	.970	.943	.948	.946
1	8	.95	−.026	−.306	.064	.098	.911	.067	.942	.004
		.05	.003	.028	.075	.074	.972	.926	.944	.933
2	8	.95	.003	−.155	.064	.068	.947	.317	.939	.056
		.05	.000	.032	.075	.072	.967	.928	.951	.918
0	24	.95	−.007	−.116	.024	.022	.920	.000	.943	.000
		.05	.000	−.001	.038	.039	.970	.946	.950	.949
1	24	.95	.000	−.101	.024	.021	.942	.000	.944	.000
		.05	.000	.009	.038	.039	.964	.942	.950	.942
2	24	.95	.001	−.055	.017	.015	.963	.024	.946	.000
		.05	.000	.010	.038	.038	.950	.941	.951	.937

'–' indicates non-existence of the corresponding population moment.

where they increase in T. When ρ_0 is increased to .95, $\widehat{\rho}_{\mathrm{GMM}}$ has a large negative bias and its standard deviation increases considerably. The AL estimator also deteriorates, but much less. The confidence intervals based on $\widehat{\rho}_{\mathrm{AL}}$ continue to provide good coverage, with the bootstrap being slightly superior, while those based on $\widehat{\rho}_{\mathrm{GMM}}$ have large coverage errors for all ψ, especially when $T \geq 8$.

Turning to the AR(2) model (Table 2), the AL and GMM estimators perform well in terms of bias (for the chosen values of ρ_0), although there is a non-negligible bias when $T = 2$ and also when $T = 4$ and the initial observations are strong inlyers. As $N = 100$, the probability that the adjusted likelihood has no local maximum in the relevant region is fairly large when both T and ψ are small. In most designs, $\widehat{\rho}_{\mathrm{AL}}$ has smaller standard deviation than $\widehat{\rho}_{\mathrm{GMM}}$, with the difference decreasing in T and ψ. For both estimators, the confidence intervals have very reasonable coverage, with those based on $\widehat{\rho}_{\mathrm{AL}}$ being slightly more accurate in most designs.

The results for the AR(1) model with a covariate (Table 3) follow a similar pattern as in the AR(1) model without covariates. The coefficient on the exogenous variable, β_0, is generally estimated with small bias by both estimators. When $(\rho_0, \beta_0) = (.5, .5)$, the biases and standard deviations are fairly small and, here, the AL estimator does not uniformly dominate GMM. When $T = 2$, $\widehat{\theta}_{\mathrm{GMM}}$ is less biased than $\widehat{\theta}_{\mathrm{AL}}$ while the opposite holds when $T \geq 4$. When (ρ_0, β_0) is changed to $(.95, .05)$, the bias of $\widehat{\beta}_{\mathrm{GMM}}$ increases only moderately, but $\widehat{\rho}_{\mathrm{GMM}}$ is now heavily biased. As a result, GMM-based confidence intervals for ρ_0 tend to have poor coverage while those for β_0 have broadly correct coverage. Compared to GMM, the AL estimator is much less biased and yields confidence intervals with approximately correct coverage, especially when computed by bootstrapping.

Appendix

Proof of equations (1.3). For arbitrary $\rho_1, ..., \rho_{T-1}$, D and its inverse are

$$D = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -\rho_1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ -\rho_{T-1} & \cdots & -\rho_1 & 1 \end{pmatrix}, \qquad D^{-1} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \phi_1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \phi_{T-1} & \cdots & \phi_1 & 1 \end{pmatrix},$$

where $\phi_1, ..., \phi_{T-1}$ are recursively obtained as $\phi_1 = \rho_1$ and $\phi_j = \rho_j + \sum_{k=1}^{j-1} \phi_k \rho_{j-k}$, $j = 2, ..., T-1$. Recursive substitution gives

$$\phi_j = \sum_{k_1 + 2k_2 + \cdots + jk_j = j} \frac{(k_1 + ... + k_p)!}{k_1! \cdots k_p!} \rho_1^{k_1} \rho_2^{k_2} \cdots \rho_j^{k_j}.$$

equivariant under translations of ρ_0 and, hence, the bias, standard deviation, and coverage rates are independent of ρ_0.

Putting $\rho_{p+1} = \cdots = \rho_{T-1} = 0$ gives $\phi_j = \varphi_j$. $\square$

Proof of equations (2.1). For $j = 1, ..., p$, let $\mathcal{S}_j = \{S \in \mathcal{S} | j \in S\}$. Group terms by $S \in \mathcal{S}_j$ to write

$$\int b_j(\rho) d\rho_j = \sum_{S \in \mathcal{S}_j} B_{j,S}(\rho) + \mathrm{c},$$

where

$$B_{j,S}(\rho) = -\sum_{t=0}^{T-j-1} \frac{T-j-t}{T(T-1)} \sum_{k \in \mathcal{K}_{j,S} : \tau' k = t} \frac{(\iota' k)!}{k_1! \cdots (k_j+1)! \cdots k_p!} \rho_j \rho_S^{k_S}$$

and $\mathcal{K}_{j,S} = \{k \in \mathbb{N}^p | \text{ for all } j' \neq j,\, k_{j'} > 0 \text{ iff } j' \in S\} \supset \mathcal{K}_S$. A change of variable from $k_j + 1$ to k_j gives

$$B_{j,S}(\rho) = -\sum_{t=|S|-j}^{T-j-1} \frac{T-j-t}{T(T-1)} \sum_{k \in \mathcal{K}_S : \tau' k = t+j} \frac{(\iota' k - 1)!}{k_1! \cdots k_p!} \rho_S^{k_S},$$

where the lower limit in the first sum changed from 0 to $|S| - j$ because, when $t < |S| - j$, no $k \in \mathcal{K}_S$ satisfies $\tau' k = t + j$. A further change of variable from $t + j$ to t gives $B_{j,S}(\rho) = a_S(\rho)$, with $a_S(\rho)$ as defined in (2.1). Therefore,

$$b_j(\rho) = \nabla_{\rho_j} \sum_{S \in \mathcal{S}_j} a_S(\rho) = \nabla_{\rho_j} \sum_{S \in \mathcal{S}} a_S(\rho) = \nabla_{\rho_j} a(\rho)$$

which completes the proof. $\square$

Proof of (2.4). In the parameterization $\eta_i = \alpha_i e^{-(T-1)a(\rho)}$, we have

$$\begin{aligned} \ell_i(\vartheta, \eta_i) &= -\frac{1}{2} \log \sigma^2 - \frac{1}{2T\sigma^2} \sum_{t=1}^{T} (y_{it} - z_{it}' \theta - \eta_i e^{(T-1)a(\rho)})^2 + \mathrm{c}, \\ \nabla_{\eta_i} \ell_i(\vartheta, \eta_i) &= \frac{e^{(T-1)a(\rho)}}{T\sigma^2} (y_i - Z_i \theta - \eta_i e^{(T-1)a(\rho)} \iota)' \iota, \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}_{\vartheta,\eta_i} \nabla_{\eta_i \eta_i} \ell_i(\vartheta, \eta_i) &= -\sigma^{-2} e^{2(T-1)a(\rho)}, \qquad \mathbb{E}_{\vartheta,\eta_i} \nabla_{\sigma^2 \eta_i} \ell_i(\vartheta, \eta_i) = 0, \\ \mathbb{E}_{\vartheta,\eta_i} \nabla_{\theta \eta_i} \ell_i(\vartheta, \eta_i) &= -\sigma^{-2} e^{(T-1)a(\rho)} \Big(\eta_i (T-1) b(\rho) e^{(T-1)a(\rho)} + \mathbb{E}_{\vartheta,\eta_i} Z_i' \iota / T \Big). \end{aligned}$$

The jth column of Y_{i-} is $y_{i,-j} = S_j(\xi_i + F\varepsilon_i)$, so the jth element of $\mathbb{E}_{\vartheta,\eta_i} Y_{i-}' \iota$ is

$$\mathbb{E}_{\vartheta,\eta_i} y_{i,-j}' \iota = \iota' S_j \xi_i = \iota' D_j^{-1} \iota \eta_i e^{(T-1)a(\rho)} + m_j, \qquad m_j = \iota' S_j \begin{pmatrix} y_i^0 \\ D^{-1}(C y_i^0 + X_i \beta) \end{pmatrix}.$$

Hence,

$$\mathbb{E}_{\vartheta,\eta_i} Z_{i-}' \iota / T = -\eta_i (T-1) b(\rho) e^{(T-1)a(\rho)} + m,$$

where $m = (m_1, ..., m_p, \iota' X_i/T)'$ is free of η_i. Consequently,

$$A_i^{-1} B_i = -e^{-(T-1)a(\rho)} \begin{pmatrix} m \\ 0 \end{pmatrix}$$

and $\nabla_{\eta_i}(A_i^{-1} B_i) = 0$. $\square$

Proof that no orthogonalization exists when $p > 1$. In the original parameterization, if $l_i(\vartheta, \alpha_i)$ is i's log-likelihood contribution, we have

$$\begin{aligned} \mathbb{E}_{\vartheta,\alpha_i} \nabla_{\alpha_i \alpha_i} l_i(\vartheta, \alpha_i) &= -\sigma^{-2}, \qquad\qquad \mathbb{E}_{\vartheta,\alpha_i} \nabla_{\sigma^2 \alpha_i} l_i(\vartheta, \alpha_i) = 0, \\ \mathbb{E}_{\vartheta,\alpha_i} \nabla_{\theta \alpha_i} l_i(\vartheta, \alpha_i) &= -\sigma^{-2} \mathbb{E}_{\vartheta,\alpha_i} Z_i' \iota / T, \end{aligned}$$

and so, by the preceeding proof,

$$A_i^{-1} B_i = -\begin{pmatrix} \mathbb{E}_{\vartheta,\alpha_i} Z_i' \iota / T \\ 0 \end{pmatrix} = -\begin{pmatrix} -(T-1) b(\rho) \alpha_i + m \\ 0 \end{pmatrix}.$$

Suppose some reparameterized fixed effect, say ζ_i, is information orthogonal to ϑ. Then $\alpha_i = \alpha_i(\vartheta, \zeta_i)$ must satisfy the differential equation $\nabla_\vartheta \alpha_i = A_i^{-1} B_i$, that is,

$$\nabla_{\rho_j} \alpha_i = (T-1) b_j(\rho) \alpha_i - m_j, \qquad j = 1, \ldots, p, \tag{A.1}$$

$$\nabla_{\beta_j} \alpha_i = -m_{p+j}, \qquad j = 1, \ldots, q, \tag{A.2}$$

and $\nabla_{\sigma^2} \alpha_i = 0$. We show that these equations are inconsistent. Suppose $q > 0$. Then (A.1) implies $\nabla_{\rho_j \beta_{j'}} \alpha_i = -\nabla_{\beta_{j'}} m_j$, which is generally non-zero, while (A.2) implies $\nabla_{\rho_j \beta_{j'}} \alpha_i = 0$, so the equations are inconsistent. Suppose $q = 0$. Then

$$m_j = \iota' S_j \begin{pmatrix} I_p \\ D^{-1} C \end{pmatrix} y_i^0, \qquad j = 1, \ldots, p,$$

and, because $\nabla_{\rho_{j'}} b_j(\rho) = \nabla_{\rho_{j'} \rho_j} a(\rho) = \nabla_{\rho_j} b_{j'}(\rho)$, (A.1) will be inconsistent if $\nabla_{\rho_{j'}} m_j \neq \nabla_{\rho_j} m_{j'}$ for some j, j'. Take $j = p$ and $j' = p-1$. The first element of y_i^0 appears in m_p and m_{p-1} with coefficients $\gamma_p = 1 + \rho_p \sum_{t=0}^{T-p-1} \varphi_t$ and $\gamma_{p-1} = \rho_p \sum_{t=0}^{T-p} \varphi_t$, respectively. Differentiating gives

$$\nabla_{\rho_{p-1}} \gamma_p = \rho_p \sum_{t=0}^{T-p-1} \nabla_{\rho_{p-1}} \varphi_t = \rho_p \sum_{t=1}^{T-p} \nabla_{\rho_p} \varphi_t, \qquad \nabla_{\rho_p} \gamma_{p-1} = \rho_p \sum_{t=1}^{T-p} \nabla_{\rho_p} \varphi_t + \sum_{k=0}^{T-p} \varphi_t,$$

using $\varphi_0 = 1$ and $\nabla_{\rho_{p-1}} \varphi_t = \nabla_{\rho_p} \varphi_{t+1}$. The latter follows from differentiating φ_t and a change of variable from $k_{p-1} - 1$ to k_{p-1}, giving

$$\nabla_{\rho_{p-1}} \varphi_t = \sum_{\tau' k = t-p+1} \frac{(\iota' k + 1)!}{k_1! \cdots k_p!} \rho^k,$$

which is invariant under a unit shift of p and t. Therefore, $\nabla_{\rho_{p-1}} \gamma_p \neq \nabla_{\rho_p} \gamma_{p-1}$, and (A.1) is inconsistent. $\square$

Proof of Lemma 1. Let $A = S_1 F_0$ and $B = \nabla_{\rho_0} A$. Then

$$b_0 = -\frac{\iota' A \iota}{T(T-1)}, \qquad c_0 = -\frac{\iota' B \iota}{T(T-1)}, \qquad V_0^{LB} = \frac{\mathrm{tr} A' M A}{T-1} = \frac{T \mathrm{tr} A A' - \iota' A A' \iota}{T(T-1)}.$$

Hence, $V_0^{LB} \geq 2b_0^2$ and $V_0^{LB} \geq 2b_0^2 - c_0$ if and only if

$$T \mathrm{tr} A A' - \iota' A A' \iota - \frac{2(\iota' A \iota)^2}{T(T-1)} \;\geq\; 0, \tag{A.3}$$

$$T \mathrm{tr} A A' - \iota' A A' \iota - \frac{2(\iota' A \iota)^2}{T(T-1)} - \iota' B \iota \;\geq\; 0. \tag{A.4}$$

The matrix $A = A_T$ is

$$\begin{aligned} A &= \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \qquad T = 2, \\ A &= \begin{pmatrix} A_{T-1} & 0 \\ a_T' & 0 \end{pmatrix}, \qquad a_T = (\rho^{T-2}, \rho^{T-3}, ..., 1)', \qquad T > 2, \end{aligned}$$

where the subscript on ρ is omitted. By recursion, it can be deduced that

$$\begin{aligned} \iota' A \iota &= \sum_{j=0}^{T-2} (T-j-1)\rho^j, \qquad \iota' B \iota = \sum_{j=1}^{T-2} j(T-j-1)\rho^{j-1}, \\ \mathrm{tr} A A' &= \sum_{j=0}^{T-2} (T-j-1)\rho^{2j}, \qquad \iota' A A' \iota = \sum_{j=0}^{T-2} \left(\sum_{k=0}^{j} \rho^k \right)^2, \end{aligned}$$

yielding V_0^{LB} as stated in the lemma. Now let $r > 0$ and use the equalities just obtained to see that if (A.4) holds for $\rho = r$, then (A.3) holds for $\rho = r$ and (A.3) and (A.4) hold for $\rho = -r$, with strict inequalities for $T \geq 3$. Hence, we only need to show that (A.4) holds for $\rho \geq 0$, with equality if and only if $T = 2$ or $\rho = 1$. Write (A.4) as $Q_T \geq 0$. Because $Q_2 = 0$, to show that (A.4) holds, it suffices to show that $\Delta Q_T \geq 0$ for $T \geq 2$, where $\Delta(\cdot)_T = (\cdot)_{T+1} - (\cdot)_T$. Write ΔQ_T as

$$\begin{aligned} \Delta Q_T &= \Delta \left(T \mathrm{tr} A A' - \iota' A A' \iota - 2\frac{(\iota' A \iota)^2}{T(T-1)} - \iota' B \iota \right)_T \\ &= \left\{ (\mathrm{tr} A A')_{T+1} - 2\frac{(\iota' A \iota)^2_{T+1}}{T(T+1)} \right\} + \left\{ 2\frac{(\iota' A \iota)^2_T}{T(T-1)} - \Delta(\iota' B \iota)_T \right\} \\ &\quad + \left\{ T\Delta(\mathrm{tr} A A')_T - \Delta(\iota' A A' \iota)_T \right\} \end{aligned}$$

and denote the quantities in braces as τ_1, τ_2, and τ_3. Using $T(T+1)/2 = \sum_{i=0}^{T-1}(T-i)$, we have

$$\begin{aligned}
\tau_1 &= \sum_{j=0}^{T-1}(T-j)\rho^{2j} - \frac{2}{T(T+1)}\left(\sum_{j=0}^{T-1}(T-j)\rho^j\right)^2 \\
&= \frac{2}{T(T+1)}\left(\sum_{i=0}^{T-1}\sum_{j=0}^{T-1}(T-i)(T-j)\rho^{2j} - \sum_{i=0}^{T-1}\sum_{j=0}^{T-1}(T-i)(T-j)\rho^{i+j}\right) \\
&= \frac{2}{T(T+1)}u'Ru,
\end{aligned}$$

where $u = (T, T-1, ..., 1)'$ and

$$R = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \rho^2 & \rho^2 & \cdots & \rho^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{2T-2} & \rho^{2T-2} & \cdots & \rho^{2T-2} \end{pmatrix} - \begin{pmatrix} 1 & \rho & \cdots & \rho^{T-1} \\ \rho & \rho^2 & \cdots & \rho^T \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \rho^T & \cdots & \rho^{2T-2} \end{pmatrix}.$$

Consider the principal minors of R. Those of order 1 are 0; those of order 2 are

$$\det\begin{pmatrix} 0 & \rho^{2i} - \rho^{i+j} \\ \rho^{2j} - \rho^{i+j} & 0 \end{pmatrix} = \rho^{i+j}\left(\rho^j - \rho^i\right)^2 \geq 0, \qquad 0 < i < j < T,$$

given $\rho \geq 0$; and those of order greater than 2 are 0 because R is the sum of two matrices of rank 1 and, hence, $\operatorname{rank}(R) \leq 2$. Therefore, R is positive semi-definite and $\tau_1 \geq 0$. Furthermore,

$$\tau_3 = T\sum_{j=0}^{T-1}\rho^{2j} - \left(\sum_{j=0}^{T-1}\rho^j\right)^2 = \sum_{i=0}^{T-1}\sum_{j=0}^{T-1}\rho^{2j} - \sum_{i=0}^{T-1}\sum_{j=0}^{T-1}\rho^{i+j} = \iota' R\iota \geq 0.$$

Use

$$\begin{aligned}
\Delta\left(\iota' B\iota\right)_T &= \sum_{j=1}^{T-1} j(T-j)\rho^{j-1} - \sum_{j=1}^{T-2} j(T-j-1)\rho^{j-1} = \sum_{j=0}^{T-2}(j+1)\rho^j \\
&= \frac{2}{T(T-1)}\sum_{i=0}^{T-2}\sum_{j=0}^{T-2}(T-i-1)(j+1)\rho^j
\end{aligned}$$

to write

$$\tau_2 = d\left(\sum_{i=0}^{T-2}\sum_{j=0}^{T-2}(T-i-1)(T-j-1)\rho^{i+j} - \sum_{i=0}^{T-2}\sum_{j=0}^{T-2}(T-i-1)(j+1)\rho^j\right),$$

where $d = \frac{2}{T(T-1)}$. Note that τ_2 is a polynomial of degree $2T-4$ in ρ. When $T = 2$ or $\rho = 1$, $\tau_2 = 0$. When $\rho \neq 1$,

$$\left(\iota' A\iota\right)_T = \frac{T - 1 - T\rho + \rho^T}{(1-\rho)^2}, \qquad \Delta\left(\iota' B\iota\right)_T = \frac{1 - \rho^T - T\rho^{T-1} + T\rho^T}{(1-\rho)^2},$$

and so

$$\tau_2 = \frac{d\left(T-1-T\rho+\rho^T\right)^2}{\left(1-\rho\right)^4} - \frac{1-\rho^T-T\rho^{T-1}+T\rho^T}{\left(1-\rho\right)^2}.$$

For $T>2$,

$$\lim_{\rho\to 1}\tau_2\left(1-\rho\right)^{-2} = \frac{1}{72}T\left(T-1\right)\left(T-2\right)\left(T+1\right) > 0$$

and, therefore, $\tau_2 = \left(1-\rho\right)^2 P\left(T,\rho\right)$, where $P\left(T,\rho\right)$ is a polynomial of degree $2T-6$. If all coefficients of $P\left(T,\rho\right) = \sum_{j=0}^{2T-6} p_j\rho^j$ are positive, we conclude that $\tau_2 \geq 0$. Write $\tau_2 = \sum_{j=0}^{2T-4} q_j\rho^j$, where q_j is found as

$$q_j = \begin{cases} \frac{d}{6}\left\{\left(j+1\right)\left(j\left(j-1\right)+6\left(T-1\right)\left(T-j-1\right)\right)-3jT(T-1)\right\}, & j \leq T-2, \\ \frac{d}{6}\left(2T-j-1\right)\left(2T-j-2\right)\left(2T-j-3\right), & T-1 \leq j. \end{cases}$$

Equating the coefficients of τ_2 and $\left(1-\rho\right)^2 P\left(T,\rho\right)$ gives $p_k = \sum_{j=0}^{k}\left(k+1-j\right)q_j$. To show that $p_k > 0$ for $0 \leq k \leq 2T-6$, we only need to show that $p_k > 0$ for k up to $T-2$ because for larger k, $q_k > 0$ and so p_k increases in k. For k up to $\min(T-2, 2T-6)$, we obtain

$$p_k = \frac{d}{12}\left(k+1\right)\left(k+2\right)\left(k+3\right)\left(\left(T-1\right)\left(2T-k-2\right)+\frac{k}{10}\left(k-1\right)-T\left(T-1\right)\right)$$

and, hence, $p_k > 0$ because either $k < T-2$, implying $2T-k-2 > T$, or $k = T-2 \leq 2T-6$, implying $T \geq 4$ and $k \geq 2$. Therefore, $\tau_2 \geq 0$. This establishes $Q_T \geq 0$, that is, (A.4). Recall that $Q_2 = 0$ and note that $\rho = 1$ implies $\tau_1 = \tau_2 = \tau_3 = 0$ and, hence, $Q_T = 0$. Therefore, $Q_T = 0$ if $T = 2$ or $\rho = 1$. If $T \geq 2$ and $\rho \neq 1$, then $\Delta Q_T > 0$ because $\tau_3 > 0$ when $T = 2$ and $\tau_2 > 0$ when $T > 2$. Therefore, $Q_T = 0$ only if $T = 2$ or $\rho = 1$. □

Proof of Theorem 1. $L_{\mathrm{A}}(\rho)$ having a local maximum or a flat inflection point at ρ_0 is equivalent to $b(\rho)$ approaching $S(\rho)$ from below as ρ approaches ρ_0 from the left. We will write this as $b(\rho) \uparrow S(\rho)$ at ρ_0, and show that $b(\rho) \uparrow S(\rho)$ on $[\underline{\rho}, \overline{\rho}]$ at most once. From

$$\nabla_\rho H(\rho) = \frac{2(\rho-\rho_{ML})\left(3\zeta_0^2-(\rho-\rho_{ML})^2\right)}{\left(\zeta_0^2+(\rho-\rho_{ML})^2\right)^3}$$

it follows that $S(\rho)$ is strictly concave on $[\underline{\rho}, \rho_{ML}]$ and strictly convex on $[\rho_{ML}, \overline{\rho}]$. Because $\varphi_t = \rho^t$, $b(\rho)$ and its first two derivatives are

$$\begin{aligned} b(\rho) &= -\sum_{t=0}^{T-2}\frac{T-1-t}{T\left(T-1\right)}\rho^t, \\ c(\rho) &= -\sum_{t=1}^{T-2}\frac{t\left(T-1-t\right)}{T\left(T-1\right)}\rho^{t-1}, \qquad d(\rho) = -\sum_{t=2}^{T-2}\frac{t\left(t-1\right)\left(T-1-t\right)}{T\left(T-1\right)}\rho^{t-2}. \end{aligned}$$

For $\rho \neq 1$,

$$
\begin{aligned}
b(\rho) &= -\frac{T-1-T\rho+\rho^T}{T\left(T-1\right)\left(1-\rho\right)^2}, \qquad c(\rho) = -\frac{T-2-T\rho+T\rho^{T-1}-\left(T-2\right)\rho^T}{T\left(T-1\right)\left(1-\rho\right)^3}, \\
d(\rho) &= -\frac{2T-6-2T\rho+T\left(T-1\right)\rho^{T-2}-2T\left(T-3\right)\rho^{T-1}+\left(T-2\right)\left(T-3\right)\rho^T}{T\left(T-1\right)\left(1-\rho\right)^4}.
\end{aligned}
$$

When $T \leq 3$, $b(\rho)$ is linear and so, given that $S(\rho)$ is concave-convex on $[\underline{\rho}, \overline{\rho}]$, $b(\rho) \uparrow S(\rho)$ on $[\underline{\rho}, \overline{\rho}]$ at most once. Suppose $T \geq 4$. Then, $b(\rho)$ is a polynomial of degree 2 or higher with negative coefficients, so $b(\rho)$ is negative, decreasing, and strictly concave, on $\mathbb{R}_+$. Further, by Descartes' rule of signs, $c(\rho)$ has one zero on $\mathbb{R}_-$ when T is even and none when T is odd, and $d(\rho)$ has no zeros on $\mathbb{R}_-$ when T is even and one when T is odd. Suppose T is even. Then $c(-1) = 0$ and $b(-1) = -\frac{1}{2(T-1)} < 0$, so $b(\rho)$ is negative and strictly concave on $\mathbb{R}$, and, hence, its intersection with $S(\rho)$ on $[\underline{\rho}, \overline{\rho}]$ can only be on $(\rho_{ML}, \overline{\rho}]$, where $S(\rho)$ is strictly convex and is approached from below by $b(\rho)$ at most once. Now suppose T is odd and $T \geq 5$. Then,

$$
d(-1) = \frac{T-3}{4T} > 0, \qquad d(-\tfrac{1}{2}) = -\frac{2^{4-T}\left(T-2\right)\left(2^T-3T+1\right)}{27T\left(T-1\right)} < 0,
$$

so $b(\rho)$ is strictly convex on $(-\infty, \rho_v]$ and strictly concave on $[\rho_v, \infty)$ for some $\rho_v \in (-1, -\frac{1}{2})$ and decreases on $\mathbb{R}$. Define ρ_u by $b(\rho_u) = 0$, that is, by $T\left(1-\rho_u\right) = 1-\rho_u^T$, $\rho_u \in \mathbb{R}_-$. Since $T \geq 5$, we have $-2 < \rho_u < -1$. Thus, $b(\rho)$ is negative and strictly convex on $(\rho_u, \rho_v]$, with $-2 < \rho_u < -1 < \rho_v < -\frac{1}{2}$. Let $R = [\rho_u, \rho_v] \cap [\rho_{ML}, \overline{\rho}]$. If R is empty, then $\rho_v < \rho_{ML}$ or $\overline{\rho} < \rho_u$; in either case, by the concavity-convexity of $S(\rho)$, $b(\rho) \uparrow S(\rho)$ on $[\underline{\rho}, \overline{\rho}]$ at most once. If R is non-empty, to show that $b(\rho) \uparrow S(\rho)$ on $[\underline{\rho}, \overline{\rho}]$ at most once, it suffices to show that $S(\rho)$ decreases faster than $b(\rho)$ on R, i.e., $H(\rho) < c(\rho)$ for $\rho \in R$. We will show below that (i) $V_0^{LB} \geq \frac{T-1}{T}$ if $\rho_0 \leq 0$; (ii) $V_0^{LB} \geq \frac{1}{2}$ if $\rho_0 > 0$. By (ii), $\rho_{ML} = \rho_0 + b_0/V_0 \geq \rho_0 + 2b_0 > -\frac{1}{2}$ if $0 < \rho_0 \leq 1$ because $b(0) = -\frac{1}{T}$, $b(1) = -\frac{1}{2}$, and $b(\rho)$ is concave on $[0, 1]$. Further, $\rho_{ML} > 0$ if $\rho_0 > 1$ because, then, $\frac{b_0}{V_0} > \frac{1}{2b_0} > -1$. Hence, R is empty if $\rho_0 > 0$. Now suppose $\rho_0 \leq 0$. Define ρ_w by $S(\rho_w) = b(\rho_v), \rho_w \in [\rho_{ML}, \overline{\rho}]$; and ρ_w' by $S(\rho_w') = b(0) = -\frac{1}{T}, \rho_w' \in [\rho_{ML}, \overline{\rho}]$. Then $\rho_w - \rho_{ML} < \rho_w' - \rho_{ML} = \frac{1}{2}(T - \sqrt{T^2 - 4\zeta_0^2})$. By (i), $\zeta_0^2 = \frac{V_0 - b_0^2}{V_0^2} \leq \frac{1}{V_0} \leq \frac{T}{T-1} \leq \frac{5}{4}$. Since $H(\rho)$ increases on $[\rho_{ML}, \overline{\rho}]$ and $H(\rho_w')$ decreases in T and increases in ζ_0^2,

$$
\begin{aligned}
H(\rho_w) &= -\frac{\zeta_0^2}{\left(\zeta_0^2+\left(\rho_w-\rho_{ML}\right)^2\right)^2} + 2S^2(\rho_w) < -\frac{\zeta_0^2}{\left(\zeta_0^2+\left(\rho_w'-\rho_{ML}\right)^2\right)^2} + \frac{2}{T^2} \\
&\leq -\frac{5/4}{\left(\frac{5}{4}+\frac{1}{4}\left(5-\sqrt{20}\right)^2\right)^2} + \frac{2}{25} < -\frac{1}{2}
\end{aligned}
$$

and so, $H(\rho) < -\frac{1}{2}$ for $\rho \in [\rho_{ML}, \rho_w]$. On the other hand, $T(1-\rho_u) = 1 - \rho_u^T$ implies $\frac{1-\rho_u}{\rho_u} = \frac{1-\rho_u^{T-1}}{T-1}$ and, therefore,

$$c(\rho_u) = -\frac{-T + T\rho_u^{T-1}}{T(T-1)(1-\rho_u)^2} = \frac{1}{\rho_u(1-\rho_u)} > -\frac{1}{2}.$$

So, $c(\rho) > -\frac{1}{2}$ for $\rho \in [\rho_u, \rho_v]$ and $H(\rho) < c(\rho)$ for $\rho \in R$. We conclude that $b(\rho) \uparrow S(\rho)$ on $[\underline{\rho}, \overline{\rho}]$ at most once, provided (i) and (ii) hold, which we now show. Write $V_0^{LB} = \frac{1}{T(T-1)} \sum_{j=0}^{2T-4} v_j \rho_0^j$, where

$$\begin{aligned} v_{2j} &= T(T-j-1) \\ &\quad - \left\{(2j+1)(T-j-1) - j(j+1) + (2j-T+1)(2j-T+2)\,1_{\{2j \geq T\}}\right\}, \\ v_{2j+1} &= -\left\{(2j+2)(T-j-2) - j(j+1) + (2j-T+2)(2j-T+3)\,1_{\{2j+1 \geq T\}}\right\}, \end{aligned}$$

using $(\sum_{k=0}^{j} \rho^k)^2 = \sum_{k=0}^{j}(k+1)\rho^k + \sum_{k=1}^{j}(j-k+1)\rho^{j+k}$. Clearly, $v_{2j+1} < 0$. Further, $v_{2j} > 0$ because

$$v_{2j} = \begin{cases} (T-2j-1)(T-j-1) + j(j+1) & \text{if } 0 \leq 2j < T, \\ (T-j-1)(j+1) & \text{if } T \leq 2j \leq 2T-4. \end{cases}$$

Hence, V_0^{LB} decreases in ρ_0 on $\mathbb{R}_-$ and (i) follows because $V_0^{LB} = \frac{T-1}{T}$ when $\rho_0 = 0$. When $0 < \rho_0 < 1$, a sufficient condition for $V_0^{LB} \geq \frac{1}{2}$ is that $d_k \geq 0$ for $0 \leq k \leq T-2$, where $d_k = \sum_{j=0}^{k}(v_{2j} + v_{2j+1}) - \frac{T(T-1)}{2}$. We have

$$v_{2j} + v_{2j+1} = \begin{cases} (T-2j-1)(T-j-1) - (T-2j-2)(2j+2) & \text{if } 2j+1 < T, \\ (2j-T+3)(T-j-1) & \text{if } 2j+1 \geq T. \end{cases}$$

Only when $2j+1 < T$ is it possible that $v_{2j} + v_{2j+1} < 0$, so it suffices to show that $d_k \geq 0$ for $2k+1 < T$. We obtain, for $2k+1 < T$,

$$d_k = \frac{1}{2}(k+1)\left(2T^2 - 5Tk + 4k^2 - 8T + 13k + 10\right) - \frac{T(T-1)}{2}.$$

Define f_k by $d_k = \frac{1}{2}(k+1) f_k$. Then, $f_0 = (T-2)(T-5) \geq 0$, $f_1 = \frac{1}{2}(3T^2 - 25T + 54) > 0$, and, for $k \geq 2$,

$$\begin{aligned} f_k &> \frac{5}{3}T^2 - 5Tk + 4k^2 - 8T + 13k + 10 \\ &= \frac{1}{3}\left((T-2k-2)(5T-6k-16) + k(T-5) + 2(T-1)\right) > 0. \end{aligned}$$

Hence, $V_0^{LB} \geq \frac{1}{2}$ when $0 < \rho_0 < 1$. When $\rho_0 \geq 1$, it also holds that $V_0^{LB} \geq \frac{1}{2}$ because then $b_0 < b(1) = -\frac{1}{2}$ and $V_0^{LB} \geq 2b_0^2$. Therefore, (ii) holds. $\square$

Proof of Lemma 2. Use $y_{i,-1} = S_1(\xi_{0i} + F_0\varepsilon_i)$ to write $Z_i = (y_{i,-1}, X_i) = (S_1F_0\varepsilon_i, 0) + \Xi_i$, where $\Xi_i = (S_1\xi_{0i}, X_i)$ is independent of ε_i. Proceeding as above, we have

$$V_0 = \frac{\text{plim}_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} Z_i' M Z_i}{\sigma_0^2\,(T-1)} = \begin{pmatrix} V_0^{LB} & 0 \\ 0 & 0 \end{pmatrix} + V_\Xi,$$

where

$$V_\Xi = \begin{pmatrix} V_{\xi\xi} & V_{\xi X} \\ V_{X\xi} & V_{XX} \end{pmatrix} = \frac{\text{plim}_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} \Xi_i' M \Xi_i}{\sigma_0^2\,(T-1)}$$

is positive semi-definite and V_{XX} is positive definite by assumption. Therefore, $V_{\xi\xi} - V_{\xi X}V_{XX}^{-1}V_{X\xi} \geq 0$ and $(V_0^{\rho\rho})^{-1} = V_0^{LB} + V_{\xi\xi} - V_{\xi X}V_{XX}^{-1}V_{X\xi} \geq V_0^{LB}$. □

References

Ahn, S. C. 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. (2004). Robust likelihood estimation of dynamic panel data models. Unpublished manuscript.

Arellano, M. (2003). *Panel Data Econometrics*. Advanced Texts in Econometrics. Oxford University Press.

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.

Arellano, M. and S. Bonhomme. (2009) Robust priors in nonlinear panel data models. *Econometrica*, 77:489–536.

Bun, M. J. and M. A. Carree. (2005). Bias-corrected estimation in dynamic panel data models. *Journal of Business and Economic Statistics*, 23:200–210.

Chamberlain, G. and M. J. Moreira. (2009). Decision theory applied to a linear panel data model. *Econometrica*, 77:107–133.

Conniffe, D. and J. E. Spencer. (2001). When moments of ratios are ratios of moments. *The Statistician*, 50:161–168.

Cox, D. R. and N. Reid. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society, Series B*, 49:1–39.

Cruddas, A., Reid, N. and D. Cox. (1989). A time series illustration of approximate conditional likelihood. *Biometrika*, 76:231–237.

Dhaene, G. and K. Jochmans. (2007). An adjusted profile likelihood for non-stationary panel data models with incidental parameters. Unpublished manuscript.

Fisher, R. (1930). The moments of the distribution for normal samples of measures of departure from normality. *Proceedings of the Royal Society of London, Series A*, 130:16–28.

Geary, R. (1933). A general expression for the moments of certain symmetrical functions of normal samples. *Biometrika*, 25:184–186.

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.

Kiefer, N. M. (1980). Estimation of fixed effects models for time series of cross-sections with arbitrary intertemporal covariance. *Journal of Econometrics*, 14:195–202.

Kiviet, J. F. (1995). On bias, inconsistency, and efficiency of various estimators in dynamic panel data models. *Journal of Econometrics*, 68:53–78.

Kruiniger, H. (2008). Maximum likelihood estimation and inference methods for the covariance stationary panel AR(1)/unit root model. *Journal of Econometrics*, 144:447–464.

Lancaster, T. (2002). Orthogonal parameters and panel data. *Review of Economic Studies*, 69:647–666.

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.

Moreira, M. J. (2009). A maximum likelihood method for the incidental parameter problem. *Annals of Statistics*, 37:3660–3696.

Nerlove, M. (1967). Experimental evidence on the estimation of dynamic economic relations from a time series of cross-sections. *Economic Studies Quarterly*, 18:42–74.

Nerlove, M. (1971). Further evidence on the estimation of dynamic economic relations from a time series of cross-sections. *Econometrica*, 39:359–382.

Neyman, J. and E. L. Scott. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16:1–32.

Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica*, 49:1417–1426.

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.

Rotnitzky, A., D. R. Cox, M. Bottai, and J. Robins. (2000). Likelihood-based inference with singular information matrix. *Bernoulli*, 6:243–284.

Sargan, J. D. (1983). Identification and lack of identification. *Econometrica*, 51:1605–1633.

Severini, T. A. (2000). *Likelihood Methods in Statistics.* Oxford Statistical Science Series. Oxford University Press.