Likelihood Inference in an Autoregression with Fixed Effects

Geert Dhaene, Koen Jochmans

Econometric Theory, **32**, 2016, 1178–1215.
doi:10.1017/S0266466615000146

LIKELIHOOD INFERENCE IN AN AUTOREGRESSION WITH FIXED EFFECTS

GEERT DHAENE
University of Leuven

KOEN JOCHMANS
Sciences Po

We calculate the bias of the profile score for the regression coefficients in a multistratum autoregressive model with stratum-specific intercepts. The bias is free of incidental parameters. Centering the profile score delivers an unbiased estimating equation and, upon integration, an adjusted profile likelihood. A variety of other approaches to constructing modified profile likelihoods are shown to yield equivalent results. However, the global maximizer of the adjusted likelihood lies at infinity for any sample size, and the adjusted profile score has multiple zeros. Consistent parameter estimates are obtained as local maximizers inside or on an ellipsoid centered at the maximum likelihood estimator.

1. INTRODUCTION

With nuisance parameters, inference based on the profile likelihood can be highly misleading. In an $N \times T$ data array setting with stratum nuisance parameters, the maximum likelihood estimator is often inconsistent as the number of strata, N, tends to infinity. This is the incidental-parameter problem (Neyman and Scott, 1948). It arises because profiling out the nuisance parameters from the likelihood introduces a nonnegligible bias into the (profile) score function. One possible solution is to calculate this bias and to subtract it from the profile score, as suggested by Neyman and Scott (1948) and McCullagh and Tibshirani (1990). When the bias is free of incidental parameters this yields a fully recentered score function which, in principle, paves the way for consistent estimation under Neyman–Scott asymptotics (Godambe and Thompson, 1974). This is the case in

We are grateful to Manuel Arellano, Stéphane Bonhomme, Gary Chamberlain, Jan Kiviet, Luc Lauwers, Thierry Magnac, Marcelo Moreira, Alessandra Salvan, Enrique Sentana, Thomas Severini, Frank Windmeijer, the co-editor Guido Kuersteiner, and four referees for comments and discussion. Research funding from the Flemish Science Foundation grants G.0505.11 and G.0628.07 is gratefully acknowledged. Address correspondence to Geert Dhaene, University of Leuven, Department of Economics, Naamsestraat 69, B-3000 Leuven, Belgium; e-mail: geert.dhaene@kuleuven.be; and Koen Jochmans, Sciences Po, Department of Economics, 28 rue des Saints-Pères, 75007 Paris, France; e-mail: koen.jochmans@sciencespo.fr.

the classic many-normal-means example, but little is known about this possibility in other situations.

In this paper we consider a time series extension of the classic example of Neyman and Scott (1948). The problem here is to estimate a pth order autoregressive model, possibly augmented with covariates, from data on N short time series of length T. The model has stratum-specific intercepts, the fixed effects. The distribution of the initial observations is left unrestricted and the p-vector of autoregressive parameters, ρ, may lie outside the stationary region. The incidental-parameter problem in this model is the subject of a substantial literature; see Arellano (2003b) for an overview and many references. The bias of the profile score is found to depend only on ρ and T. Hence, adjusting the profile score by subtracting its bias gives a fixed T unbiased estimating equation and, upon integration, an adjusted profile likelihood in the sense of Pace and Salvan (2006).

However, contrary to what standard maximum likelihood theory would suggest, the parameters of interest are *local* maximizers of the expected adjusted likelihood. The global maximum is reached at infinity. This phenomenon is not a small-sample problem or an artifact of an unbounded parameter space. The adjusted likelihood has its global maximum at infinity for any sample size, and may already be re-increasing in the stationary parameter region and reach its maximum at the boundary. Consistent estimation is achieved by locally maximizing the adjusted likelihood over a certain ellipsoid that is centered at the maximum likelihood estimator and is defined by the (unadjusted) likelihood function. The adjusted likelihood is re-increasing because the initial observations are unrestricted. This difficulty does not arise when stationarity of the initial observations is imposed, as in Cruddas, Reid, and Cox (1989). Further, when the data carry only little information, in a sense that we specify, the Hessian of the adjusted likelihood is singular, implying first-order underidentification (Sargan, 1983) and nonstandard asymptotic properties of the resulting point estimates, as derived in the case $p = 1$ by Kruiniger (2014).

These features are not unique to our approach. We show that several other routes to constructing modified objective functions for the dynamic linear fixed-effect model yield equivalent results. When $p = 1$, the adjusted profile likelihood coincides with the marginal posterior in Lancaster (2002), which, in the absence of covariates, is a Bayesian version of a Cox and Reid (1987) approximate conditional likelihood (see Sweeting, 1987). For general p, it is an integrated likelihood in the sense of Kalbfleish and Sprott (1970) and Arellano and Bonhomme (2009) where the fixed effects have been integrated out using a new data-independent bias-reducing prior. Such a prior was thought not to exist for this model. The adjusted likelihood can also be seen as a penalized likelihood as defined by Bester and Hansen (2009) (see DiCiccio, Martin, Stern, and Young, 1996, and Severini, 1998, for related approaches). The adjusted profile score equation, in turn, is a Woutersen (2002) integrated moment equation and a locally-orthogonal Cox and Reid (1987) moment equation, as defined in Arellano (2003a), and solving it is

equivalent to inverting the probability limit of the least-squares estimator, as proposed by Bun and Carree (2005) for the case $p = 1$.

The equivalence results allow to connect and complement various earlier least-squares and likelihood-based approaches. In particular, our analysis of the global properties of the modified objective function shows that it has to be maximized locally or, when solving the modified estimating equation, the appropriate solution has to be selected accordingly, an issue that has been overlooked. Corrected least-squares and likelihood-based methods have been proposed in this model as alternatives to the generalized method-of-moments estimators of Arellano and Bond (1991) and Ahn and Schmidt (1995). The latter estimators are well known to deliver biased point estimates and confidence regions with poor coverage when the data are persistent (Blundell and Bond, 1998) or when T is not negligible compared to N (Alvarez and Arellano, 2003). Inference based on the adjusted likelihood also becomes more fragile in the vicinity of a unit root, but does not deteriorate when T/N is nonnegligible. On a more general level, our findings highlight the difficulty of point identification under Neyman–Scott asymptotics and show that global maximization of a bias-adjusted profile likelihood, as if it were an ordinary likelihood, may fail badly.

Our focus is on short panels, that is, we treat T as fixed in the asymptotics. In related work, Kiviet (1995) and Hahn and Kuersteiner (2002) proposed methods to approximately bias-correct the within-group least-squares estimator. These approaches do not fully remove the bias but, rather, reduce its order from $O(T^{-1})$ down to $O(T^{-2})$. Hence, they are more suited for panels where T/N is nonnegligible. On the other hand, while a complete re-centering of the profile score equation is not generally possible in nonlinear fixed-effect models, approximate bias-corrected estimators have been derived under fairly general conditions; see, e.g., Hahn and Newey (2004), Arellano and Hahn (2006), Hahn and Kuersteiner (2011), and Dhaene and Jochmans (2015).

Sections 2 to 6 derive and study the adjusted profile likelihood. Proofs are given in the Appendix.

2. ADJUSTED PROFILE LIKELIHOOD

2.1. 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, for N strata i and T periods t. Consider the model

$$y_{it} = y_{it-}^\top \rho + x_{it}^\top \beta + \alpha_i + \varepsilon_{it}, \quad i = 1, \ldots, N;\ t = 1, \ldots, T, \tag{2.1}$$

where $y_{it-} = (y_{it-1}, \ldots, y_{it-p})^\top$, ρ and β are parameter vectors, α_i is a fixed effect, and ε_{it} is an error term. Let $z_{it} = (y_{it-}^\top, x_{it}^\top)^\top$ and $\theta = (\rho^\top, \beta^\top)^\top$. Further, let $y_i = (y_{i1}, \ldots, y_{iT})^\top$, $Y_{i-} = (y_{i1-}, \ldots, y_{iT-})^\top$, $X_i = (x_{i1}, \ldots, x_{iT})^\top$, $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iT})^\top$, and $Z_i = (Y_{i-}, X_i)$, so that $My_i = MZ_i\theta + M\varepsilon_i$ where $M = I_T - T^{-1}\iota\iota^\top$ and ι is a T-vector of ones. Also, let $y_i^0 = (y_{i(1-p)}, \ldots, y_{i0})^\top$ denote the initial values (which are observed). We make the following assumption.

Assumption 2.1. The variable y_{it} is generated by (2.1) and

(i) (Z_i, ε_i) and $(Z_{i'}, \varepsilon_{i'})$ are independent for all i and $i' \neq i$;

(ii) ε_{it} is i.i.d. for all i and t, is independent of X_i, has finite fourth moment, and satisfies

$$\mathbb{E}\left(\varepsilon_{it} | X_i, y_{it-1}, \ldots, y_{i1}, y_i^0\right) = 0,$$
$$\operatorname{Var}\left(\varepsilon_{it} | X_i, y_{it-1}, \ldots, y_{i1}, y_i^0\right) = \sigma^2 > 0;$$

(iii) $N^{-1}\sum_{i=1}^N Z_i^\top M Z_i$ and $\operatorname{plim}_{N\to\infty} N^{-1}\sum_{i=1}^N Z_i^\top M Z_i < \infty$ are nonsingular.

Thus, we assume cross-sectional independence, strict exogeneity, homoskedasticity, and no multicollinearity. On the other hand, we do not assume normality of ε_{it} and place no restrictions on how $\left(y_i^0, \alpha_i, X_i\right)$, $i = 1, \ldots, N$, are generated, thus allowing for nonstationarity across t. The unknown parameters are θ, σ^2, and $\alpha_1, \ldots, \alpha_N$. 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 require ρ_0 to lie in the stationary region of $\mathbb{R}^p$, i.e., we allow any $\rho_0 \in \mathbb{R}^p$.

We shall work with the Gaussian quasi-likelihood (i.e., acknowledging that ε_{it} may be nonnormal) but simply refer to it as the likelihood. Conditional on $y_1^0, \ldots, y_N^0$ and divided by NT, the log-likelihood is

$$-\frac{1}{2NT}\sum_{i=1}^N\sum_{t=1}^T\left(\log\sigma^2 + \frac{1}{\sigma^2}\left(y_{it} - z_{it}^\top\theta - \alpha_i\right)^2\right) + \mathrm{c},$$

where, here and later, c is a nonessential constant. Profiling out $\alpha_1, \ldots, \alpha_N$ and σ^2 gives the profile log-likelihood (divided by NT) for θ,

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

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

$$s_{\rho_j}(\theta) = \nabla_{\rho_j} l(\theta) = \frac{\sum_{i=1}^N (y_i - Z_i\theta)^\top M y_{i,-j}}{\sum_{i=1}^N (y_i - Z_i\theta)^\top M (y_i - Z_i\theta)}, \quad j = 1, \ldots, p,$$
$$s_{\beta_j}(\theta) = \nabla_{\beta_j} l(\theta) = \frac{\sum_{i=1}^N (y_i - Z_i\theta)^\top M x_{i,j}}{\sum_{i=1}^N (y_i - Z_i\theta)^\top M (y_i - Z_i\theta)}, \quad j = 1, \ldots, q,$$

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

For the analysis below, rewrite (2.1) as

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

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

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

Then

$$\begin{pmatrix} y_i^0 \\ y_i \end{pmatrix} = \xi_i + F\varepsilon_i, \tag{2.2}$$

where

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

and $y_{i,-j} = S_j(\xi_i + F\varepsilon_i)$ for selection matrices $S_j = (0_{T\times(p-j)}, I_T, 0_{T\times j})$.

2.2. Bias of the profile score

The profile score is asymptotically biased, that is, $\operatorname{plim}_{N\to\infty} s(\theta_0) \neq 0$. Therefore, the maximum likelihood estimator, solving $s(\theta) = 0$, is inconsistent. (Here and later, probability limits and expectations are taken conditionally on $\left(y_i^0, \alpha_i, X_i\right)$, $i = 1, \ldots, N$.) Lemma 2.1 below shows that the profile-score bias is a polynomial in the parameter ρ_0. For $k = (k_1, \ldots, k_p)^\top \in \mathbb{N}^p$, let $\rho^k = \prod_{j=1}^p \rho_j^{k_j}$. Also, let $\tau = (1, \ldots, p)^\top$,

$$\varphi_t = \sum_{\tau^\top k = t} \frac{(\iota^\top k)!}{k_1! \cdots k_p!} \rho^k, \quad t = 1, \ldots, T-1, \tag{2.3}$$

and set $\varphi_0 = 0$.

LEMMA 2.1. *Suppose Assumption 2.1 holds. Then, the asymptotic bias of the profile score is* $\operatorname{plim}_{N\to\infty} s(\theta_0) = b(\rho_0)$, *where* $b(\rho) = (b_1(\rho), \ldots, b_{p+q}(\rho))^\top$ *and*

$$\begin{aligned} b_j(\rho) &= -\textstyle\sum_{t=0}^{T-j-1} \frac{T-j-t}{T(T-1)} \varphi_t, \quad && j = 1, \ldots, p, \\ b_j(\rho) &= 0, && j = p+1, \ldots, p+q. \end{aligned}$$

The bias of the profile score depends only on ρ_0 and T. It does not depend on the distribution of ε_{it}. It is, furthermore, independent of the initial observations, the fixed effects, and the covariates. This is in sharp contrast with the bias of the maximum likelihood estimator, which was first derived by Nickell (1981) for the first-order autoregressive model under the assumption of stationarity of the initial observations. This bias depends on the initial observations, the fixed effects, and the covariate values.

2.3. Adjusted profile likelihood

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

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

is asymptotically unbiased, i.e., $\operatorname{plim}_{N\to\infty} s_{\mathrm{a}}(\theta_0) = 0$. Hence, $s_{\mathrm{a}}(\theta) = 0$ is a bias-adjusted estimating equation. The question arises whether there is a corresponding adjustment to the profile likelihood. This indeed turns out to be the case, as the following lemma shows.

LEMMA 2.2. *Let $b(\rho)$ be as defined in Lemma 2.1. Up to an arbitrary constant of integration, the solution to $\nabla_\theta a(\rho) = b(\rho)$ is given by*

$$a(\rho) = \sum_{S\in\mathcal{S}} a_S(\rho), \qquad a_S(\rho) = -\sum_{t=|S|}^{T-1} \frac{T-t}{T(T-1)} \sum_{k\in\mathcal{K}_S:\tau^\top k=t} \frac{(\iota^\top k-1)!}{k_1!\cdots k_p!}\rho_S^{k_S}, \quad \textbf{(2.4)}$$

where $\mathcal{S}$ is the collection of the nonempty subsets of $\{1,\dots,p\}$; $|S|$ is the sum of the elements of S; $\mathcal{K}_S = \{k\in\mathbb{N}^p | k_j > 0$ if and only if $j\in S\}$; 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.

It follows that $s_{\mathrm{a}}(\theta) = 0$ is an estimating equation associated with the function

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

which we call 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.

3. CONNECTIONS WITH THE LITERATURE

Before studying the adjusted profile likelihood as a tool for inference about θ_0 we show that it can also be obtained through various other routes that have been suggested in the literature.

Lancaster (2002) studied the first-order autoregressive model, with and without covariates, from a Bayesian perspective. With $p = 1$, we have $\varphi_t = \rho^t$ and

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

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

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

where $Q^2(\theta) = \sum_{i=1}^N (y_i - Z_i\theta)^\top M(y_i - Z_i\theta) \propto e^{-2l(\theta)}$. Integrating over σ^2 gives

$$f(\theta|\text{data}) \propto e^{-N(T-1)a(\rho)} \left(Q^2(\theta)\right)^{-N(T-1)/2}$$

and, hence,

$$f(\theta|\text{data}) \propto e^{N(T-1)l_{\text{a}}(\theta)}. \tag{3.1}$$

Thus, the posterior and the adjusted likelihood are equivalent. More generally, for any p and q, independent uniform priors on $\eta_1, \ldots, \eta_N, \theta, \log\sigma^2$, with $\eta_i = \alpha_i e^{-(T-1)a(\rho)}$ and $a(\rho)$ as in Lemma 2.2, yield a posterior $f(\theta|\text{data})$ that is related to $l_{\text{a}}(\theta)$ as in (3.1).

Lancaster's choice of a prior on the η_i that is independent of ϑ is motivated by a first-order autoregression without covariates. There, η_i is orthogonal to ϑ and the posterior $f(\theta|\text{data})$ (and, hence, also $e^{l_{\text{a}}(\theta)}$) has an interpretation as a Cox and Reid (1987) approximate conditional likelihood; see also Sweeting (1987). Orthogonalization to a multidimensional parameter is generally not possible (Severini, 2000, pp. 340–342). Here, orthogonalization is not possible when the model is augmented with covariates, as shown by Lancaster, or when the autoregressive order, p, is greater than one, as we show in the Appendix. From a bias correction perspective, however, orthogonality is not required. In the present model, for any p and q, $s_{\text{a}}(\theta) = 0$ is an unbiased estimating equation, and the bias calculation underlying it is immune to the nonexistence of orthogonalized fixed effects.

The approach of Arellano and Bonhomme (2009) shares the integration step with Lancaster (2002) but allows nonuniform priors on fixed effects or, equivalently, nonorthogonalized 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 present model (with general p and 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 log integrated likelihood (divided by NT) is

$$l_{\text{int}}(\vartheta) = \frac{1}{NT}\sum_{i=1}^N \log \int \sigma^{-T/2} e^{-\frac{1}{2\sigma^2}\sum_{t=1}^T \left(y_{it} - z_{it}^\top\theta - \alpha_i\right)^2} \pi_i(\alpha_i|\vartheta)\, \mathrm{d}\alpha_i + \text{c}.$$

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

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

Profiling out σ^2 gives $\sigma^2(\theta) = \arg\max_{\sigma^2} l_{\text{int}}(\vartheta) = Q^2(\theta)/(N(T-1))$, and so

$$l_{\text{int}}(\theta) = \max_{\sigma^2} l_{\text{int}}(\vartheta) = \frac{T-1}{T} l_{\text{a}}(\theta) + \text{c}.$$

Thus $l_{\text{int}}(\theta)$ and $l_{\text{a}}(\theta)$ 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. 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 setting a uniform prior on $\eta_i = \alpha_i e^{-(T-1)a(\rho)}$, as it leads to the same $l_{\text{int}}(\vartheta)$. Further, Arellano and Bonhomme (2009) give a necessary and sufficient condition for a uniform prior to be bias-reducing. With $\ell_i(\vartheta, \eta_i) = T^{-1} \sum_{t=1}^{T} \ell_{it}(\vartheta, \eta_i)$ denoting i's log-likelihood contribution (divided by T) in a parametrization η_i, the condition is that

$$\text{plim}_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \nabla_{\eta_i} \left(A_i^{-1} B_i \right) = o(1) \quad \text{as } T \to \infty, \tag{3.2}$$

where

$$A_i = A_i(\vartheta, \eta_i) = -\mathbb{E}_{\vartheta,\eta_i} \nabla_{\eta_i \eta_i} \ell_i(\vartheta, \eta_i), \qquad B_i = B_i(\vartheta, \eta_i) = \mathbb{E}_{\vartheta,\eta_i} \nabla_{\vartheta \eta_i} \ell_i(\vartheta, \eta_i),$$

and where $\nabla_{\eta_i} \left(A_i^{-1} B_i \right)$ is evaluated at the true parameter values. When η_i and ϑ are orthogonal, $B_i = 0$ and (3.2) holds. However, Condition (3.2) 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 (3.2) must hold for $\eta_i = \alpha_i e^{-(T-1)a(\rho)}$. Indeed, as we show in the Appendix,

$$\nabla_{\eta_i} \left(A_i^{-1} B_i \right) = 0 \tag{3.3}$$

because $A_i^{-1} B_i$ is free of η_i.

Woutersen (2002) derived a likelihood-based moment condition in which parameters of interest and fixed effects are orthogonal by construction even though orthogonality in the information matrix may not be possible. With $\ell_i = \ell_i(\vartheta, \alpha_i) = \sum_{t=1}^{T} \ell_{it}(\vartheta, \alpha_i)$ a generic log-likelihood for stratum i, let

$$g_i = g_i(\vartheta, \alpha_i) = \nabla_\vartheta \ell_i - \nabla_{\alpha_i} \ell_i \frac{\mathbb{E}_{\vartheta,\alpha_i} \nabla_{\alpha_i \vartheta} \ell_i}{\mathbb{E}_{\vartheta,\alpha_i} \nabla_{\alpha_i \alpha_i} \ell_i}. \tag{3.4}$$

Then $\mathbb{E}_{\vartheta,\alpha_i} g_i = 0$ and parameter orthogonality holds in the sense that $\mathbb{E}_{\vartheta,\alpha_i} \nabla_{\alpha_i} g_i = 0$ (under regularity conditions). The integrated moment estimator of ϑ minimizes $g_{\text{int}}^\top g_{\text{int}}$ where $g_{\text{int}} = (NT)^{-1} \sum_{i=1}^{N} g_{\text{int}i}$ and

$$g_{\text{int}i} = g_{\text{int}i}(\vartheta) = \left[g_i - \frac{1}{2} \frac{\nabla_{\alpha_i \alpha_i} g_i}{\nabla_{\alpha_i \alpha_i} \ell_i} + \frac{1}{2} \frac{\nabla_{\alpha_i \alpha_i \alpha_i} \ell_i}{\nabla_{\alpha_i \alpha_i} \ell_i} \nabla_{\alpha_i} g_i \right]_{\alpha_i = \widehat{\alpha}_i(\vartheta)},$$

with $\widehat{\alpha}_i(\vartheta) = \arg\max_{\alpha_i} \ell_i$. The function $g_{\text{int}i}$ is the Laplace approximation to $\int g_i e^{\ell_i} \mathrm{d}\alpha_i / \int e^{\ell_i} \mathrm{d}\alpha_i$, that is, to g_i with α_i integrated out using likelihood weights.[1] Arellano (2003b) obtained the same $g_{\text{int}i}$ as a locally orthogonal Cox and Reid (1987) moment function. Woutersen and Voia (2004) calculated g_{int} for the present model with $p = 1$. For any p and q, the integrated moment condition essentially coincides with the adjusted profile score. In the Appendix it is shown that

$$g_{\text{int}i}(\theta, \sigma^2) = \begin{pmatrix} \sigma^{-2} Z_i^\top M(y_i - Z_i\theta) - (T-1)\, b(\rho) \\ \sigma^{-4}(y_i - Z_i\theta)^\top M(y_i - Z_i\theta)/2 - \sigma^{-2}(T-1)/2 \end{pmatrix}. \tag{3.5}$$

On profiling out σ^2 from the minimand $g_{\text{int}}^\top g_{\text{int}}$ we obtain

$$g_{\text{int}}(\theta) = \frac{T-1}{T}\big(s(\theta) - b(\rho)\big) = \frac{T-1}{T} s_{\text{a}}(\theta).$$

Thus, the estimator of θ by Woutersen (2002) minimizes the norm of the adjusted profile score.

The adjusted likelihood can also be viewed as a penalized log-likelihood in the sense of Bester and Hansen (2009). With $\ell = \sum_{i=1}^N \sum_{t=1}^T \ell_{it}$, $\ell_{it} = \ell_{it}(\vartheta, \alpha_i)$, again denoting a generic log-likelihood, let $\pi_i = \pi_i(\vartheta, \alpha_i)$ be a function such that

$$\begin{aligned} \operatorname{plim}_{N\to\infty} \nabla_{\alpha_i} \pi_i = {} & \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^T \nabla_{\alpha_i\alpha_i} \ell_{it} \sum_{t=1}^T \psi_{it} \right] \\ & + \frac{1}{2} \mathbb{E}\big[\nabla_{\alpha_i\alpha_i\alpha_i} \ell_{it}\big] \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^T \psi_{it} \sum_{t=1}^T \psi_{it} \right], \\ \operatorname{plim}_{N\to\infty} \nabla_{\vartheta} \pi_i = {} & \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^T \nabla_{\alpha_i\vartheta} \ell_{it} \sum_{t=1}^T \psi_{it} \right] \\ & + \frac{1}{2} \mathbb{E}\big[\nabla_{\alpha_i\alpha_i\vartheta} \ell_{it}\big] \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^T \psi_{it} \sum_{t=1}^T \psi_{it} \right], \end{aligned}$$

where $\psi_{it} = -\mathbb{E}\big[\nabla_{\alpha_i\alpha_i} \ell_{it}\big]^{-1} \nabla_{\alpha_i} \ell_{it}$. Then $\ell_\pi = \ell - \sum_{i=1}^n \pi_i$ is a penalized log-likelihood. Bester and Hansen (2009) provide a function that satisfies these differential equations in a general class of fixed-effect models and show that it leads to ℓ_π whose first-order condition has bias $o(T^{-1})$. In the present model, the equations can be solved exactly, i.e., for finite T. With $\ell_{it} = -\frac{1}{2}\big[\log\sigma^2 + \big(y_{it} - z_{it}^\top\theta - \alpha_i\big)^2/\sigma^2\big] + \text{c}$, the relevant differential equations are

$$\nabla_{\alpha_i} \pi_i = 0, \qquad \nabla_\theta \pi_i = (T-1) b(\rho), \qquad \nabla_{\sigma^2} \pi_i = -\frac{1}{2\sigma^2},$$

which yields $\pi_i = -\frac{1}{2}\log\sigma^2 + (T-1)a(\rho) + \text{c}$. Therefore,

$$\ell_\pi = \ell + \frac{N}{2}\log\sigma^2 - N(T-1)a(\rho) + \text{c} \tag{3.6}$$

and $l_\pi(\theta) = \max_{\alpha_1,\ldots,\alpha_N,\sigma^2} \ell_\pi = N(T-1)l_{\mathrm{a}}(\theta) + \mathrm{c}$. Thus, the profile penalized log-likelihood and the adjusted log-likelihood are equivalent. Note that Bester and Hansen's approach is to adjust the likelihood *before* profiling out the incidental parameters, while we adjust it *after* doing so. In the present model, the two approaches coincide.

Finally, the adjusted profile score is also related to the approach of Bun and Carree (2005). Note that $s(\theta) = \sum_{i=1}^N Z_i^\top M(y_i - Z_i\theta)/Q^2(\theta)$ and $My_i = MZ_i\widehat{\theta} + M\widehat{\varepsilon}_i$ where $\widehat{\theta}$ is the maximum likelihood estimator, with residuals $\widehat{\varepsilon}_i$ satisfying $\sum_{i=1}^N Z_i^\top M\widehat{\varepsilon}_i = 0$. Therefore, solving $s_{\mathrm{a}}(\theta) = 0$ is equivalent to solving

$$\widehat{\theta} - \theta = \left(\sum_{i=1}^N Z_i^\top M Z_i\right)^{-1} b(\rho)Q^2(\theta). \tag{3.7}$$

When $p = 1$, solving (3.7) corresponds to the proposal by Bun and Carree (2005) for bias-correcting the maximum likelihood estimate.

4. GLOBAL PROPERTIES OF THE ADJUSTED PROFILE LIKELIHOOD

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

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

$$L(\theta) = -\frac{1}{2}\log\left(\mathrm{plim}_{N\to\infty}\frac{1}{N}\sum_{i=1}^N U_i(\theta)\right) + \mathrm{c}$$

for $U_i(\theta) = \varepsilon_i^\top M\varepsilon_i - 2(\theta - \theta_0)^\top Z_i^\top M\varepsilon_i + (\theta - \theta_0)^\top Z_i^\top MZ_i(\theta - \theta_0)$. Let $b_0 = b(\rho_0) = S(\theta_0)$ and note that

$$\mathrm{plim}_{N\to\infty}\frac{1}{N}\sum_{i=1}^N Z_i^\top M\varepsilon_i = \left(\mathrm{plim}_{N\to\infty}\frac{1}{N}\sum_{i=1}^N \varepsilon_i^\top M\varepsilon_i\right) b_0 = \sigma_0^2(T-1)\,b_0.$$

Hence, defining V_0 by

$$\mathrm{plim}_{N\to\infty}\frac{1}{N}\sum_{i=1}^N Z_i^\top MZ_i = \left(\mathrm{plim}_{N\to\infty}\frac{1}{N}\sum_{i=1}^N \varepsilon_i^\top M\varepsilon_i\right) V_0 = \sigma_0^2(T-1)\,V_0,$$

we can write

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

by absorbing the term $-\frac{1}{2}\log\left(\sigma_0^2(T-1)\right)$ into c. As $N \to \infty$, the maximum likelihood estimator of θ converges in probability to $\theta_{\mathrm{ml}} = \arg\max_\theta L(\theta) = \theta_0 + V_0^{-1}b_0$ and has asymptotic bias $V_0^{-1}b_0$. This expression generalizes the fixed T bias calculations in Nickell (1981) and Bun and Carree (2005). Note that $(\theta_0 - \theta_{\mathrm{ml}})^\top V_0(\theta_0 - \theta_{\mathrm{ml}}) = b_0^\top V_0^{-1} b_0$. Furthermore,

$$L(\theta) = -\frac{1}{2}\log\left(1 - b_0^\top V_0^{-1} b_0 + (\theta-\theta_{\mathrm{ml}})^\top V_0(\theta-\theta_{\mathrm{ml}})\right) + \mathrm{c},$$

$$S(\theta) = -\frac{V_0(\theta-\theta_{\mathrm{ml}})}{1 - b_0^\top V_0^{-1} b_0 + (\theta-\theta_{\mathrm{ml}})^\top V_0(\theta-\theta_{\mathrm{ml}})},$$

$$H(\theta) = -\frac{V_0}{1 - b_0^\top V_0^{-1} b_0 + (\theta-\theta_{\mathrm{ml}})^\top V_0(\theta-\theta_{\mathrm{ml}})} + 2S(\theta)S(\theta)^\top.$$

Note that $L(\cdot)$ and $H(\cdot)$ are even and $S(\cdot)$ is odd about θ_{ml} and that $H(\theta_0) = 2b_0b_0^\top - V_0$ and $H_{\mathrm{a}}(\theta_0) = 2b_0b_0^\top - V_0 - c_0$, 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 $\rho = \kappa r$ with r in the positive orthant of $\mathbb{R}^p$ and let $\kappa \to \infty$, the term $-a(\rho)$ dominates and $L_{\mathrm{a}}(\theta) \to \infty$.

It follows that $\theta_0 \neq \arg\max_\theta L_{\mathrm{a}}(\theta)$, and so θ_0 has to be identified as a functional of $L_{\mathrm{a}}(\theta)$ other than its global maximizer (as in standard maximum likelihood theory). 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, as shown below), $L_{\mathrm{a}}(\theta)$, being smooth and unbounded, must also have at least one local minimum. Because $l(\theta)$ is log-quadratic for any $N \geq 1$ and $a(\rho)$ does not depend on the data, $l_{\mathrm{a}}(\theta)$, too, is re-increasing, regardless of the sample size. Therefore, an estimation strategy based on solving $s_{\mathrm{a}}(\theta) = 0$ will generally lead to multiple solutions, from which the appropriate one has to be chosen.

4.1. First-order autoregression without covariates

Our focus in this and the following subsections is on how the parameter of interest, θ_0, is identified from the function $L_{\mathrm{a}}(\theta)$. We first examine the first-order autoregression without covariates, i.e., $p = 1$ and $q = 0$.

Letting $\zeta_0^2 = \left(V_0 - b_0^2\right)/V_0^2$, we have

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

$$S(\rho) = -\frac{\rho-\rho_{\mathrm{ml}}}{\zeta_0^2 + (\rho-\rho_{\mathrm{ml}})^2}, \qquad H(\rho) = -\frac{\zeta_0^2 - (\rho-\rho_{\mathrm{ml}})^2}{\left(\zeta_0^2 + (\rho-\rho_{\mathrm{ml}})^2\right)^2},$$

by absorbing $-\frac{1}{2}\log V_0$ into c. Note that $\zeta_0^2 = -1/H(\rho_{\mathrm{ml}})$, hence, ζ_0^2 is identified. Recall that $S(\rho)$ is odd about $\rho_{\mathrm{ml}} = \rho_0 + b_0/V_0$. The zeros of $H(\rho)$ are $\underline{\rho} = \rho_{\mathrm{ml}} - \zeta_0$ and $\overline{\rho} = \rho_{\mathrm{ml}} + \zeta_0$, so $S(\rho)$ decreases on $[\underline{\rho}, \overline{\rho}]$ and increases elsewhere. All of $\underline{\rho}$, $\overline{\rho}$, ρ_{ml}, and ζ_0 are identified by $S(\cdot)$, and 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 $[\underline{\rho}, \overline{\rho}]$.

There is a sharp lower bound on V_0. With ξ_{0i} and F_0 denoting ξ_i and F evaluated at ρ_0, we have $y_{i,-1} = S_1(\xi_{0i} + F_0\varepsilon_i)$. From the independence between ξ_{0i} and ε_i, we obtain

$$V_0 = \frac{\operatorname{plim}_{N\to\infty} \frac{1}{N}\sum_{i=1}^N y_{i,-1}^\top M y_{i,-1}}{\sigma_0^2 (T-1)} = V_0^{LB} + V_{\xi\xi},$$

where

$$V_0^{LB} = \frac{\operatorname{tr} F_0^\top S_1^\top M S_1 F_0}{T-1}, \qquad V_{\xi\xi} = \frac{\operatorname{plim}_{N\to\infty} \frac{1}{N}\sum_{i=1}^N \xi_{0i}^\top S_1^\top M S_1 \xi_{0i}}{\sigma_0^2 (T-1)}. \tag{4.1}$$

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 4.1. *Suppose Assumption 2.1 holds with $p = 1, q = 0$. Then, V_0^{LB}, as defined in (4.1), 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 $V_0^{LB} \geq 2b_0^2$ and $V_0^{LB} \geq 2b_0^2 - c_0$, each with equality if and only if $T = 2$ or $\rho_0 = 1$.

By Lemma 4.1, $H(\rho_0) = 2b_0^2 - V_0 \leq 0$ and, hence,

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

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, $S_{\mathrm{a}}(\rho)$ has finitely many zeros. Thus, because $S_{\mathrm{a}}(\rho_0) = 0$ and, by Lemma 4.1, $H_{\mathrm{a}}(\rho_0) = 2b_0^2 - V_0 - c_0 \leq 0$, it follows that $L_{\mathrm{a}}(\rho)$ has a local maximum or a flat inflection point at ρ_0. Our main result for the first-order autoregression without covariates is the uniqueness of such a point in $[\underline{\rho}, \overline{\rho}]$, thereby identifying ρ_0 as a functional of $L_{\mathrm{a}}(\rho)$. Equivalently, ρ_0 is the unique point in $[\underline{\rho}, \overline{\rho}]$ where $b(\rho)$ approaches $S(\rho)$ from below.

THEOREM 4.1. *Suppose Assumption 2.1 holds with $p = 1, q = 0$. Then, ρ_0 is the unique point in $[\underline{\rho}, \overline{\rho}]$ where $L_{\mathrm{a}}(\rho)$ has a local maximum or a flat inflection point.*

$L_{\mathrm{a}}(\rho)$ 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 $\rho_0^{j-1}y_i^0 + \alpha_i \sum_{k=1}^{j-1} \rho_0^{k-1}$, $j = 1, \ldots, T$, so $MS_1\xi_{0i} = 0$ if and only if $y_i^0(1 - \rho_0) = \alpha_i$. The following corollary has been independently obtained by Ahn and Thomas (2006).

COROLLARY 4.1. *Suppose Assumption 2.1 holds with $p = 1, q = 0$. Then, when $\rho_0 = 1$ and $\alpha_i = 0$, $L_{\mathrm{a}}(\rho)$ has a flat inflection point at ρ_0 for any T.*

This result is 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). It implies that inference based on the adjusted likelihood becomes nonstandard in the unit root case because $H_{\mathrm{a}}(\rho_0)$ vanishes.

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 plots 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 plots show $L(\rho)$ (solid line), $-a(\rho)$ (dashed line), and $L_{\mathrm{a}}(\rho) = L(\rho) - a(\rho)$ (thick line). In all the plots, 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 $[\underline{\rho}, \overline{\rho}]$.

The asymptotic bias of the maximum likelihood estimator has the same sign as b_0 because $\rho_{\mathrm{ml}} = \rho_0 + b_0/V_0$. The proof of Theorem 4.1, 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 $\rho_u - \rho_0$.

We note, finally, that $L_{\mathrm{a}}(\rho)$ may have more than one local maximum on $\mathbb{R}$. When $\rho_0 = -5$, $V_0 = V_0^{LB}$, and $T = 4$, $L_{\mathrm{a}}(\rho)$ has local maxima at -5 and -2.97, local minima at -3.78 and -0.27, and $[\underline{\rho}, \overline{\rho}] = [-5.08, -4.94]$. When $\rho_0 \in [-1, 1]$, however, we have not found a case where $L_{\mathrm{a}}(\rho)$ has multiple local

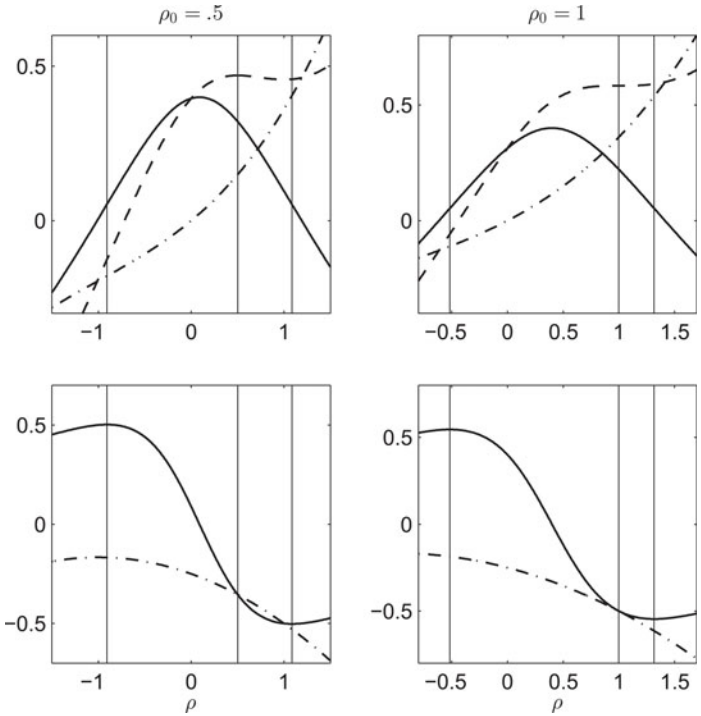

FIGURE 1. Identification in the first-order autoregression. Left: $T = 4$, $\rho_0 = .5$, $V_0 = V_0^{LB} + V_{\xi\xi}$, $V_{\xi\xi}$ corresponding to stationary initial observations. Right: $T = 4$, $\rho_0 = 1$, $V_0 = V_0^{LB}$. Top: $L(\rho)$ (solid), $-a(\rho)$ (dashed-dotted), $L_a(\rho)$ (dashed). Bottom: $S(\rho)$ (solid), $b(\rho)$ (dashed-dotted). Vertical lines at $\underline{\rho}$, ρ_0, and $\overline{\rho}$.

maxima on $\mathbb{R}$. This is in line with the result obtained by Kruiniger (2014, Lemmas 1 and 2) that, when $\rho_0 \geq -1$, $L_a(\rho)$ and $l_a(\rho)$ each have at most one local maximum on $[-1, \infty)$.

4.2. First-order autoregression with covariates

In the first-order autoregressive model with covariates ($p = 1, q \geq 1$), profiling out β yields a profile likelihood of ρ with essentially the same properties as in the model without covariates. Let $\beta(\rho) = \arg\max_\beta L_a(\rho, \beta) = \arg\max_\beta L(\rho, \beta) = \arg\min_\beta (\theta - \theta_{ml})^\top V_0 (\theta - \theta_{ml})$. Partition V_0, V_0^{-1}, and b_0 as

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

With

$$V_0^{\rho\rho} = \left(V_{0\rho\rho} - V_{0\rho\beta} V_{0\beta\beta}^{-1} V_{0\beta\rho} \right)^{-1}, \tag{4.2}$$

we have

$$V_{0\beta\beta}\,(\beta(\rho)-\beta_{\mathrm{ml}})=-V_{0\beta\rho}(\rho-\rho_{\mathrm{ml}}),$$
$$\min_{\beta}(\theta-\theta_{\mathrm{ml}})^{\top}V_0(\theta-\theta_{\mathrm{ml}})=(\rho-\rho_{\mathrm{ml}})^2/V_0^{\rho\rho},$$
$$1-b_0^{\top}V_0^{-1}b_0=1-b_{0\rho}^2V_0^{\rho\rho}.$$

The first of these equations, together with $V_0(\theta_0-\theta_{\mathrm{ml}})=-b_0$, 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

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

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

LEMMA 4.2. *Suppose Assumption 2.1 holds with $p=1, q\geq 1$. Let $V_0^{\rho\rho}$ be as defined in (4.2) and V_0^{LB} as defined in (4.1) and given in Lemma 4.1. Then, $(V_0^{\rho\rho})^{-1}\geq V_0^{LB}$.*

We can now invoke the result for the model without covariates. Let $\underline{\rho}=\rho_{\mathrm{ml}}-\zeta_0$ and $\overline{\rho}=\rho_{\mathrm{ml}}+\zeta_0$, with ζ_0 redefined as indicated.

THEOREM 4.2. *Suppose Assumption 2.1 holds with $p=1, q\geq 1$. 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 4.2, the conditions under which ρ_0 is a flat inflection point of $L_{\mathrm{a}}(\rho)$ are the same as before. The presence of covariates does not affect the sign of the asymptotic bias of the maximum likelihood estimator of ρ. It also follows from the proof of Lemma 4.2 that the inclusion of covariates in the model cannot increase $V_0^{\rho\rho}$, so the magnitude of $\rho_{\mathrm{ml}}-\rho_0=V_0^{\rho\rho}b_{0\rho}$ can only decrease relative to the model without covariates.

4.3. *p*th-order autoregression

Consider first an autoregression with $p>1$ and without covariates, i.e., $q=0$. Then

$$L(\rho)=-\frac{1}{2}\log\left(1+(\rho-\rho_{\mathrm{ml}})^{\top}W_0(\rho-\rho_{\mathrm{ml}})\right)+\mathrm{c},\quad W_0=\frac{V_0}{1-b_0^{\top}V_0^{-1}b_0},$$
$$S(\rho)=-\frac{W_0(\rho-\rho_{\mathrm{ml}})}{1+(\rho-\rho_{\mathrm{ml}})^{\top}W_0(\rho-\rho_{\mathrm{ml}})},$$
$$H(\rho)=-\frac{W_0}{1+(\rho-\rho_{\mathrm{ml}})^{\top}W_0(\rho-\rho_{\mathrm{ml}})}+2S(\rho)S(\rho)^{\top},$$

where $-\frac{1}{2}\log(1-b_0^\top V_0^{-1}b_0)$ is absorbed into c. Because $W_0=-H(\rho_{\mathrm{ml}})$, W_0 is identified by $L(\cdot)$.

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

$$V_0=\frac{\operatorname{plim}_{N\to\infty}\frac{1}{N}\sum_{i=1}^N Y_{i-}^\top M Y_{i-}}{\sigma_0^2(T-1)}=V_0^{LB}+V_{\xi\xi}$$

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

$$\left(V_0^{LB}\right)_{jk}=\frac{\operatorname{tr}F_0^\top S_j^\top M S_k F_0}{T-1},\qquad \left(V_{\xi\xi}\right)_{jk}=\frac{\operatorname{plim}_{N\to\infty}\frac{1}{N}\sum_{i=1}^N \xi_{0i}^\top S_j^\top M S_k \xi_{0i}}{\sigma_0^2(T-1)},$$

for $1\le j,k\le p$. Hence, $V_0-V_0^{LB}$ is positive semi-definite, which we write as $V_0\ge V_0^{LB}$. When $p\ge T$, while V_0 is nonsingular, $\operatorname{rank}(V_0^{LB})\le T-1$ because $S_jF_0=0$ for $j\ge T$, which implies that $\left(V_0^{LB}\right)_{jk}=0$ whenever $j\ge T$ or $k\ge T$. Thus, when $p\ge T$, although V_0 can be arbitrarily close to V_0^{LB}, $V_0\ne V_0^{LB}$. Further, when $p\ge T$, $b_j(\rho)=0$ for $j\ge T$ because the sum defining $b_j(\rho)$ is empty, and $c_{ij}(\rho)=0$ for $i+j\ge T$. Hence, when $p\ge T$, $V_0^{LB}-2b_0b_0^\top$ and $V_0^{LB}-2b_0b_0^\top+c_0$ have only zeros beyond their leading $(T-1)\times(T-1)$ blocks.

A proof of generalizations of Lemma 4.1 and Theorem 4.1 to the case $p>1$ would be desirable but is more difficult. 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 $p=2$. In comparison, $\varphi_t=\rho_1^t$ when $p=1$. We resorted to numerical computations, which suggest that

$$V_0^{LB}\ge 2b_0b_0^\top,\qquad V_0^{LB}\ge 2b_0b_0^\top-c_0, \tag{4.3}$$

$$\operatorname{rank}\left(V_0^{LB}-2b_0b_0^\top+c_0\right)=\begin{cases}\min(p,T-2) & \text{if } \sum_{j=1}^p\rho_{0j}\ne 1 \text{ or } T<p+2,\\ p-1 & \text{else.}\end{cases} \tag{4.4}$$

Specifically, we computed the eigenvalues of $V_0^{LB}-2b_0b_0^\top$ and $V_0^{LB}-2b_0b_0^\top+c_0$ for $p=2,3,4$; $T=2,\ldots,7$; 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

$$\begin{aligned}&-1\le\gamma_2\le 1, && \gamma_2-1\le\gamma_1\le 1-\gamma_2,\\ &-1\le\gamma_4\le 1, && \gamma_4-1\le\gamma_3\le 1-\gamma_4,\end{aligned} \tag{4.5}$$

the stationary region of the lag polynomial $\gamma(L)=\left(1-\gamma_1L-\gamma_2L^2\right)\left(1-\gamma_3L-\gamma_4L^2\right)$. 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 (4.5) 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 (4.5) 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 $V_0^{LB} - 2b_0b_0^\top$ and $V_0^{LB} - 2b_0b_0^\top + c_0$ are nonnegative and the rank of $V_0^{LB} - 2b_0b_0^\top + c_0$ is as given by (4.4). These findings, while obviously not a proof, support (4.3) and (4.4), and we shall proceed under the assumption that (4.3) and (4.4) hold.

Because $V_0 \geq V_0^{LB}$, (4.3) implies that $V_0 \geq 2b_0b_0^\top$ and that $H_a(\rho_0) = 2b_0b_0^\top - V_0 - c_0 \leq 0$. Pre- and postmultiplication of $V_0 \geq 2b_0b_0^\top$ by $b_0^\top V_0^{-1}$ and $V_0^{-1}b_0$ gives $b_0^\top V_0^{-1}b_0 \leq \frac{1}{2} \leq 1 - b_0^\top V_0^{-1}b_0$. Recalling that $(\rho_0 - \rho_{ml})^\top V_0(\rho_0 - \rho_{ml}) = b_0^\top V_0^{-1}b_0$, we have

$$(\rho_0 - \rho_{ml})^\top W_0(\rho_0 - \rho_{ml}) \leq 1.$$

Therefore, if (4.3) and (4.4) hold, ρ_0 is a point in the ellipsoidal disk $\mathcal{E} = \{\rho : (\rho - \rho_{ml})^\top W_0(\rho - \rho_{ml}) \leq 1\}$ where $L_a(\rho)$ 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 $L_a(\rho)$, starting at ρ_{ml} and using the Moore–Penrose inverse of $H_a(\rho)$ whenever $H_a(\rho)$ 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 $L_a(\rho)$ has a local maximum or a flat inflection point.

In the model with covariates, just as before, β can be profiled out of $L_a(\theta)$. Here, again, $\beta_0 = \beta(\rho_0)$. Lemma 4.2 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 $W_0 = (1 - b_{0\rho}^\top V_0^{\rho\rho} b_{0\rho})^{-1} V_{0\rho\rho}$, in obvious notation.

Figure 2 illustrates the identification for two cases with $p = 2$, $T = 4$, and without covariates. The plots on the left are for $\rho_0 = (.25, .25)'$ with $V_0 = V_0^{LB} + V_{\xi\xi}$ and $V_{\xi\xi}$ corresponding to stationary initial observations. Those on the right are for $\rho_0 = (.5, .5)'$ with $V_0 = V_0^{LB}$. In each figure, the ellipse $\mathcal{E}$, containing ρ_0, is drawn. The top figures, in addition, show contour plots of the adjusted log-likelihood, $L_a(\rho)$. (The unadjusted log-likelihood contours are not shown; they are elliptical and, like $\mathcal{E}$, centered at ρ_{ml}.) The figures at the bottom also plot the loci of the solution set of each of the adjusted profile score equations, $\nabla_{\rho_1} L_a(\rho) = 0$ and $\nabla_{\rho_2} L_a(\rho) = 0$, which intersect at ρ_0. In the case $\rho_0 = (.25, .25)'$, ρ_0 is the unique local maximizer of $L_a(\rho)$ in $\mathcal{E}$. There is also a second point in $\mathcal{E}$ where $S_a(\rho) = 0$, corresponding to a saddlepoint of $L_a(\rho)$. In the case $\rho_0 = (.5, .5)'$, ρ_0 is the unique point in $\mathcal{E}$ where $S_a(\rho) = 0$ and $H_a(\rho)$ is negative semidefinite. Here, ρ_0 is an isolated point of singularity of $H_a(\rho)$.

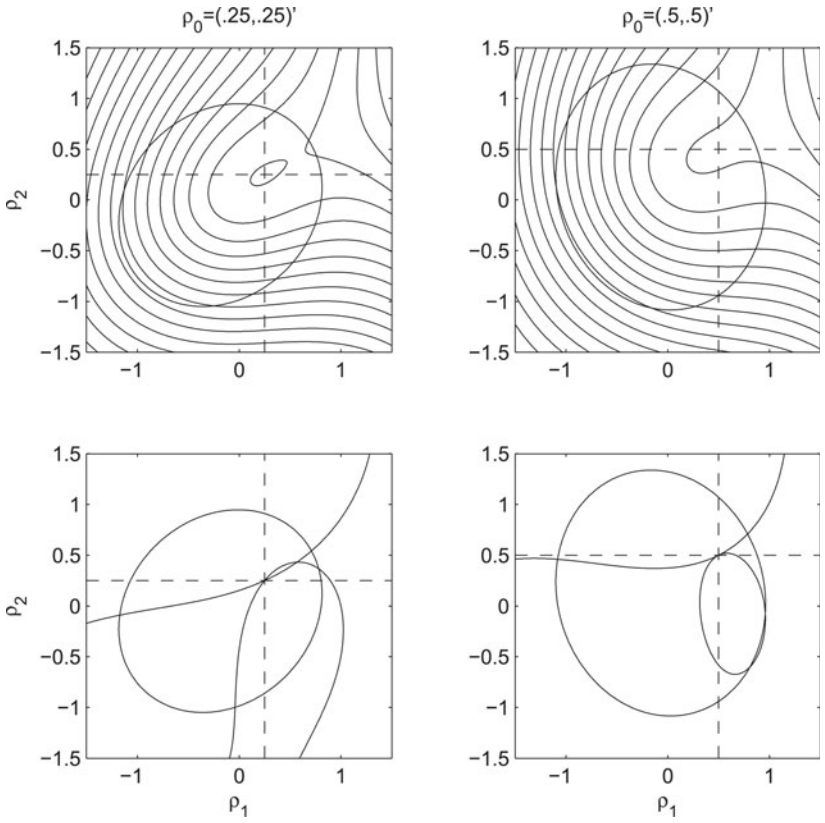

FIGURE 2. Identification in the second-order autoregression. Left: $T = 4$, $\rho_0 = (.25, .25)'$, $V_0 = V_0^{LB} + V_{\xi\xi}$, $V_{\xi\xi}$ corresponding to stationary initial observations . Right: $T = 4$, $\rho_0 = (.5, .5)'$, $V_0 = V_0^{LB}$. Top: contour plots of $L_a(\rho)$. Bottom: loci where $\nabla_{\rho_1} L_a(\rho) = 0$ (upper curve) and $\nabla_{\rho_2} L_a(\rho) = 0$ (lower curve). In all subplots: $\mathcal{E}$ (ellipse), dashed lines at ρ_0.

5. ESTIMATION AND INFERENCE

Let $\widehat{\beta}(\rho) = \arg\max_\beta l_a(\rho, \beta)$ for a given ρ. Note that

$$\widehat{\beta}(\rho) = \left(\sum_{i=1}^{N} X_i^\top M X_i\right)^{-1} \sum_{i=1}^{N} X_i^\top M (y_i - Y_{i-}\rho) = \arg\max_\beta l(\rho, \beta).$$

The unadjusted and adjusted profile log-likelihoods for ρ are $l(\rho) = l\big(\rho, \widehat{\beta}(\rho)\big)$ and $l_a(\rho) = l(\rho) - a(\rho)$. Let $s(\rho)$, $s_a(\rho)$, $h(\rho)$, and $h_a(\rho)$ be the corresponding profile scores and Hessians. Let $\widehat{W} = -h\big(\widehat{\rho}_{\text{ml}}\big)$, where $\widehat{\rho}_{\text{ml}}$ is the maximum likelihood estimator of ρ_0, and let $\widehat{\mathcal{E}} = \big\{\rho : \big(\rho - \widehat{\rho}_{\text{ml}}\big)^\top \widehat{W} \big(\rho - \widehat{\rho}_{\text{ml}}\big) \le 1\big\}$. We define

the adjusted likelihood estimator of ρ_0 as

$$\widehat{\rho}_{\mathrm{al}} = \arg\min_{\rho \in \widehat{\mathcal{E}}} s_{\mathrm{a}}^{\top}(\rho) s_{\mathrm{a}}(\rho) \quad \text{s.t.} \quad h_{\mathrm{a}}(\rho) \leq 0$$

and those of β_0 and θ_0 as $\widehat{\beta}_{\mathrm{al}} = \widehat{\beta}(\widehat{\rho}_{\mathrm{al}})$ and $\widehat{\theta}_{\mathrm{al}} = \left(\widehat{\rho}_{\mathrm{al}}^{\top}, \widehat{\beta}_{\mathrm{al}}^{\top}\right)^{\top}$. Some remarks and motivation are in order. In the most regular case, where $l_{\mathrm{a}}(\rho)$ has a unique strict local maximizer on the interior of $\widehat{\mathcal{E}}$, $\widehat{\rho}_{\mathrm{al}}$ coincides with the local maximizer of $l_{\mathrm{a}}(\rho)$. However, due to sampling variation with finite N, $l_{\mathrm{a}}(\rho)$ may have no local maximizer on $\widehat{\mathcal{E}}$, an event that occurs with positive probability. When this happens, $\widehat{\rho}_{\mathrm{al}}$ is defined as the point where $l_{\mathrm{a}}(\rho)$ varies the least, as measured by the norm of its gradient, $s_{\mathrm{a}}(\rho)$; in a sense this comes close to locally maximizing $l_{\mathrm{a}}(\rho)$. We also cannot exclude the possibility that $l_{\mathrm{a}}(\rho)$ has more than one local maximizer on the interior of $\widehat{\mathcal{E}}$ or that, when no local maximizer exists, the norm of $s_{\mathrm{a}}(\rho)$ has more than one minimizer. In this event, $\widehat{\rho}_{\mathrm{al}}$ becomes set-valued (if a single point estimate is required, one might choose the solution closest to $\widehat{\rho}_{\mathrm{ml}}$). However, in all cases where ρ_0 is identified as a unique local maximizer of $L_{\mathrm{a}}(\rho)$ on $\mathcal{E}$, the probability of nonexistence or nonuniqueness of a local maximizer of $l_{\mathrm{a}}(\rho)$ on $\widehat{\mathcal{E}}$ vanishes as $N \to \infty$.

The adjusted likelihood estimator is consistent and, when ρ_0 is identified as a local maximizer, asymptotically normal.

THEOREM 5.1. *Suppose Assumption 2.1 holds. Suppose also that ρ_0 is the unique point in $\mathcal{E}$ where $L_{\mathrm{a}}(\rho)$ has a local maximum or a flat inflection point. Then $\widehat{\theta}_{\mathrm{al}} \xrightarrow{p} \theta_0$ as $N \to \infty$.*

THEOREM 5.2. *Suppose Assumption 2.1 holds. Suppose also that ρ_0 is the unique point in $\mathcal{E}$ where $L_{\mathrm{a}}(\rho)$ has a local maximum and that $H_{\mathrm{a}}(\theta_0)$ is nonsingular. Then, as $N \to \infty$,*

$$\sqrt{N}\left(\widehat{\theta}_{\mathrm{al}} - \theta_0\right) \xrightarrow{d} \mathcal{N}(0, \Omega), \tag{5.1}$$

where $\Omega = H_{\mathrm{a}}(\theta_0)^{-1} \Sigma H_{\mathrm{a}}(\theta_0)^{-1}$, $\Sigma = \operatorname{plim}_{N\to\infty} N^{-1} \sum_{i=1}^{N} e_i e_i^{\top}$, *and* $e_i = \frac{1}{\sigma_0^2 (T-1)} \left(Z_i - \varepsilon_i b_0^{\top}\right)^{\top} M \varepsilon_i$.

We note that the information equality (i.e., the second Bartlett identity) does not hold for the adjusted likelihood, i.e., $-H_{\mathrm{a}}(\theta_0)$ differs from Σ, the asymptotic variance of $s_{\mathrm{a}}(\theta_0)$. One consequence is that the asymptotic variance of $\widehat{\theta}_{\mathrm{al}}$ is of the sandwich form. As $N \to \infty$, a consistent estimate of Ω is obtained by replacing $H_{\mathrm{a}}(\theta_0)$, b_0, ε_i, and σ_0^2 with $h\left(\widehat{\theta}_{\mathrm{al}}\right) - c\left(\widehat{\rho}_{\mathrm{al}}\right)$, $b\left(\widehat{\rho}_{\mathrm{al}}\right)$, $\widehat{\varepsilon}_i = y_i - Z_i \widehat{\theta}_{\mathrm{al}}$, and $(T-1)^{-1} N^{-1} \sum_{i=1}^{N} \widehat{\varepsilon}_i^{\top} M \widehat{\varepsilon}_i$, respectively. Wald tests and confidence ellipsoids then follow in the usual way, with correct asymptotic size and coverage probabilities, respectively. By contrast, Wald tests and confidence ellipsoids based on the unadjusted likelihood have asymptotic size equal to one and asymptotic coverage probabilities equal to zero. A further consequence of the failure of the second

Bartlett identity is that the adjusted likelihood ratio statistic (i.e., the LR statistic applied to the adjusted likelihood) is, under the null, asymptotically distributed as a weighted sum of χ^2_1 variates with the eigenvalues of $-\Sigma H_a(\theta_0)^{-1}$ as weights (instead of being χ^2_{p+q} asymptotically); see, e.g., Kent (1982), White (1982), and Vuong (1989). Although these weights can be estimated, the adjusted likelihood ratio statistic is unsuited for testing because it is ill-signed for large enough values of ρ. This is another consequence of the adjusted likelihood being re-increasing.

6. SIMULATIONS

In this section, we report simulation results for first- and second-order autoregressions without covariates, and a first-order autoregression with one stationary covariate. In all instances we chose ρ_0 in the interior of the stationary parameter region. We compare $\widehat{\theta}_{\mathrm{al}}$ with two other estimators: the one-step GMM estimator of Arellano and Bond (1991), $\widehat{\theta}_{\mathrm{ab}}$, and the estimator of Hahn and Kuersteiner (2002), $\widehat{\theta}_{\mathrm{hk}}$. The latter estimator is a large-T correction of the maximum likelihood estimator. In the first-order autoregression without covariates, $\widehat{\rho}_{\mathrm{hk}} = \widehat{\rho}_{\mathrm{ml}} + \left(1 + \widehat{\rho}_{\mathrm{ml}}\right)/T$ (Hahn and Kuersteiner, 2002), where one may view the bias correction term as resulting from $\rho_{\mathrm{ml}} = \rho_0 - (1+\rho_0)/T + o(1/T)$. In the second-order autoregression, the approach of Hahn and Kuersteiner (2002) gives $\widehat{\rho}_{\mathrm{hk}} = \widehat{\rho}_{\mathrm{ml}} + \iota_2\left(1 + \widehat{\rho}_{\mathrm{ml2}}\right)/T$, following from

$$\rho_{\mathrm{ml}} = \rho_0 - \iota_2(1+\rho_{02})/T + o(1/T), \tag{6.1}$$

as we show in the Appendix (with $\widehat{\rho}_{\mathrm{ml2}}$ and ρ_{02} denoting the second element of $\widehat{\rho}_{\mathrm{ml}}$ and ρ_0, respectively) (see Li, Lindsay, and Waterman, 2003, and Sartori, 2003). In the first-order autoregression with a covariate, $\widehat{\beta}_{\mathrm{hk}} = \left(\sum_{i=1}^N X_i^\top M X_i\right)^{-1} \sum_{i=1}^N X_i^\top M\left(y_i - y_{i,-1}\widehat{\rho}_{\mathrm{hk}}\right)$ and $\widehat{\rho}_{\mathrm{hk}} = \widehat{\rho}_{\mathrm{ml}} + \left(1 + \widehat{\rho}_{\mathrm{ml}}\right)/T$. While $\widehat{\theta}_{\mathrm{ab}}$ is fixed-T consistent, $\widehat{\theta}_{\mathrm{hk}}$ is consistent only as $T \to \infty$. On the other hand, under rectangular-array asymptotics (see (Li, Lindsay, and Waterman, 2003) and (Sartori, 2003)), $\widehat{\theta}_{\mathrm{ab}}$ is incorrectly centered, whereas $\widehat{\theta}_{\mathrm{hk}}$ is correctly centered. In line with the large-N, fixed-T approach in this paper, in the simulations presented here we set N relatively large compared to T ($N = 100, 500$ and $T = 2, 4, 8, 16$). We focus on small T because this setting is particularly relevant to micro-economic panel data applications, and it is also the more challenging setting. It should be noted that our setup is relatively unfavorable for $\widehat{\theta}_{\mathrm{hk}}$ in that its higher-order bias may show up and dominate in the distribution.

In all simulations, we generated ε_{it} and α_i as standard normal variates. We varied the informational content of the data by controlling 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 $\mu_i = \mathbb{E}\left(y_i^0|\alpha_i\right)$ and $\Sigma_i = \mathrm{Var}\left(y_i^0|\alpha_i\right)$. Let $G_i G_i^\top = \Sigma_i$ be the Cholesky factorization of Σ_i. We set $y_i^0 = \mu_i + \psi G_i \iota$ for a chosen scalar $\psi \geq 0$. This is a p-variate version of setting the initial observations ψ standard deviations away from their respective stationary means. 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. When $p = 1$ and $q = 0$, for example, $S_1\xi_{0i} = g_0(y_{i0} - \mu_i) + \iota\mu_i$ with $g_0 = \left(1, \rho_0, \ldots, \rho_0^{T-1}\right)^\top$, hence $V_{\xi\xi} = \frac{g_0^\top M g_0}{(1-\rho_0^2)(T-1)}\psi^2$. As $\rho_0 \uparrow 1$, $V_{\xi\xi} \to 0$ for any fixed ψ. We set $\psi = 0, 1, 2$ when $p = 1$ and $\psi = .3, 1, 2$ when $p = 2$. We do not consider $\psi = 0$ for $p = 2$ because the weight matrix of the Arellano and Bond (1991) estimator is singular for all T in this case.

Tables 1 to 3 present Monte Carlo estimates, based on 10,000 replications, of the bias and the standard deviation (std) of the estimators considered, as well as the coverage rates of the corresponding asymptotic 95% confidence intervals ($\text{ci}_{.95}$).

In the first-order autoregression with $\rho_0 = .5$ (first part of Table 1), both $\widehat{\rho}_{\text{al}}$ and $\widehat{\rho}_{\text{ab}}$ perform well. The adjusted likelihood estimator has smaller standard

TABLE 1. Simulation results for the first-order autoregression

				bias			std			$\text{ci}_{.95}$		
N	T	ψ	ρ_0	$\widehat{\rho}_{\text{al}}$	$\widehat{\rho}_{\text{ab}}$	$\widehat{\rho}_{\text{hk}}$	$\widehat{\rho}_{\text{al}}$	$\widehat{\rho}_{\text{ab}}$	$\widehat{\rho}_{\text{hk}}$	$\widehat{\rho}_{\text{al}}$	$\widehat{\rho}_{\text{ab}}$	$\widehat{\rho}_{\text{hk}}$
100	2	0	.5	−.142	—	−.747	.267	—	.153	.819	.921	.000
100	2	1	.5	.027	—	−.373	.266	—	.141	.903	.934	.090
100	2	2	.5	.019	—	.111	.166	—	.113	.946	.945	.880
100	4	0	.5	.008	−.039	−.295	.141	.148	.066	.924	.926	.004
100	4	1	.5	.016	−.053	−.139	.124	.164	.067	.945	.928	.327
100	4	2	.5	.001	−.016	.071	.064	.082	.056	.946	.936	.684
100	8	0	.5	.001	−.026	−.085	.056	.057	.042	.953	.918	.400
100	8	1	.5	−.001	−.040	−.045	.048	.070	.040	.943	.907	.730
100	8	2	.5	−.001	−.023	.028	.036	.051	.034	.946	.930	.812
100	16	0	.5	.000	−.019	−.021	.028	.030	.026	.944	.902	.841
100	16	1	.5	−.001	−.027	−.013	.027	.035	.025	.947	.879	.899
100	16	2	.5	−.001	−.023	.009	.023	.032	.023	.944	.893	.902
500	2	0	.5	−.106	—	−.750	.162	—	.067	.833	.927	.000
500	2	1	.5	.033	—	−.375	.168	—	.063	.931	.953	.000
500	2	2	.5	.004	—	.108	.067	—	.051	.952	.950	.400
500	4	0	.5	.012	−.008	−.295	.088	.069	.030	.946	.942	.000
500	4	1	.5	.003	−.010	−.139	.053	.076	.030	.958	.943	.002
500	4	2	.5	.000	−.003	.072	.028	.038	.025	.949	.946	.125
500	8	0	.5	.000	−.006	−.085	.025	.026	.019	.946	.941	.002
500	8	1	.5	.000	−.008	−.043	.021	.032	.018	.948	.939	.246
500	8	2	.5	.000	−.005	.029	.016	.023	.015	.951	.949	.431
500	16	0	.5	.000	−.004	−.021	.012	.014	.012	.951	.939	.509

Table continues on overleaf

TABLE 1. *continued*

				bias			std			$ci_{.95}$		
N	T	ψ	ρ_0	$\widehat{\rho}_{al}$	$\widehat{\rho}_{ab}$	$\widehat{\rho}_{hk}$	$\widehat{\rho}_{al}$	$\widehat{\rho}_{ab}$	$\widehat{\rho}_{hk}$	$\widehat{\rho}_{al}$	$\widehat{\rho}_{ab}$	$\widehat{\rho}_{hk}$
500	16	1	.5	.000	−.006	−.011	.012	.016	.011	.949	.936	.781
500	16	2	.5	.000	−.004	.009	.010	.015	.010	.948	.940	.807
100	2	0	.99	−.144	—	−.506	.265	—	.151	.821	.925	.034
100	2	1	.99	−.135	—	−.495	.267	—	.153	.821	.918	.043
100	2	2	.99	−.125	—	−.475	.266	—	.150	.827	.929	.053
100	4	0	.99	−.087	−.773	−.258	.123	.474	.073	.835	.651	.023
100	4	1	.99	−.082	−.771	−.249	.123	.475	.072	.839	.656	.027
100	4	2	.99	−.068	−.737	−.229	.123	.472	.072	.849	.675	.049
100	8	0	.99	−.046	−.472	−.132	.062	.198	.038	.847	.280	.022
100	8	1	.99	−.043	−.469	−.125	.061	.193	.038	.850	.281	.033
100	8	2	.99	−.028	−.434	−.104	.060	.198	.037	.881	.337	.098
100	16	0	.99	−.025	−.255	−.068	.031	.081	.020	.843	.034	.020
100	16	1	.99	−.020	−.254	−.061	.031	.080	.020	.867	.033	.045
100	16	2	.99	−.009	−.227	−.043	.030	.080	.019	.910	.057	.213
500	2	0	.99	−.107	—	−.505	.164	—	.067	.826	.931	.000
500	2	1	.99	−.102	—	−.497	.163	—	.067	.831	.928	.000
500	2	2	.99	−.090	—	−.476	.164	—	.067	.842	.932	.000
500	4	0	.99	−.056	−.748	−.256	.076	.474	.033	.839	.671	.000
500	4	1	.99	−.054	−.756	−.248	.076	.474	.032	.844	.681	.000
500	4	2	.99	−.039	−.640	−.226	.076	.489	.032	.864	.727	.000
500	8	0	.99	−.030	−.442	−.130	.038	.192	.017	.845	.310	.000
500	8	1	.99	−.025	−.459	−.123	.038	.194	.017	.860	.296	.000
500	8	2	.99	−.014	−.367	−.103	.038	.190	.016	.884	.425	.000
500	16	0	.99	−.015	−.218	−.067	.019	.077	.009	.852	.055	.000
500	16	1	.99	−.010	−.240	−.060	.019	.080	.009	.878	.040	.000
500	16	2	.99	−.002	−.180	−.042	.019	.074	.008	.924	.125	.000

Notes: Data generated as $y_{it} = \rho_0 y_{it-1} + \alpha_i + \varepsilon_{it}$ $(i = 1, \ldots, N; t = 1, \ldots T)$ with $\alpha_i \sim \mathcal{N}(0, 1)$, $\varepsilon_{it} \sim \mathcal{N}(0, 1)$, and ψ the degree of outlyingness of the initial observations y_{i0}. Entries: bias, standard deviation (std), and coverage rate of 95% confidence interval ($ci_{.95}$) of adjusted likelihood ($\widehat{\rho}_{al}$), Arellano-Bond ($\widehat{\rho}_{ab}$), and Hahn-Kuersteiner ($\widehat{\rho}_{hk}$) estimators; '—' indicates nonexistence of the moment; 10,000 Monte Carlo replications.

deviation and is virtually unbiased, except when $\psi = 0$ and $T = 2$. Both estimators also deliver 95% confidence intervals with broadly correct coverage and their biases decrease in N, as the theory predicts. The estimator of Hahn and Kuersteiner (2002) has substantial bias for small T and its performance is sensitive to ψ. In line with the theory, its bias decreases in T and is nearly constant in N.

When ρ_0 is increased to .99 (second part of Table 1), the performance of all estimators tends to worsen. $\widehat{\rho}_{ab}$ deteriorates the most, showing a substantial bias, large dispersion, and confidence intervals with lower coverage. $\widehat{\rho}_{al}$ continues to have little bias and provides confidence intervals with roughly correct coverage. When ρ_0 is large, the probability that the adjusted likelihood has

no interior local maximum in the relevant region is fairly large (up to the large-N theoretical maximum of 50%), in which case we set the confidence interval equal to $\mathbb{R}$. This probability decreases in N, T, and ψ. In most designs, $\widehat{\rho}_{\text{hk}}$ outperforms $\widehat{\rho}_{\text{ab}}$ in terms of bias and standard deviation. The associated confidence intervals, however, are not reliable.

In the second-order autoregression with $\rho_0 = (.6, .2)^\top$ (Table 2), both $\widehat{\rho}_{\text{al}}$ and $\widehat{\rho}_{\text{ab}}$ perform well in terms of bias. $\widehat{\rho}_{\text{al}}$ has the least bias, even though the bias is nonnegligible when $T = 2$ and also when $T = 4$ and the initial observations are strong inlyers. The comparison between $\widehat{\rho}_{\text{al}}$ and $\widehat{\rho}_{\text{ab}}$ in terms of dispersion shows no clear ordering; their standard errors tend to equalize as T or ψ grows. As before, $\widehat{\rho}_{\text{hk}}$ shows a substantial bias when T is very small. Together with its small standard deviation for most values of T, this again leads to confidence intervals being too narrow.

TABLE 2. Simulation results for the second-order autoregression

				bias			std			$\text{ci}_{.95}$		
N	T	ψ	ρ_0	$\widehat{\rho}_{\text{al}}$	$\widehat{\rho}_{\text{ab}}$	$\widehat{\rho}_{\text{hk}}$	$\widehat{\rho}_{\text{al}}$	$\widehat{\rho}_{\text{ab}}$	$\widehat{\rho}_{\text{hk}}$	$\widehat{\rho}_{\text{al}}$	$\widehat{\rho}_{\text{ab}}$	$\widehat{\rho}_{\text{hk}}$
100	2	.3	.6	−.143	—	−.848	.262	—	.304	.822	.927	.029
			.2	−.349	—	−.742	.665	—	.807	.956	.945	.479
100	2	1	.6	−.146	—	−.844	.267	—	.167	.811	.914	.000
			.2	−.173	—	−.730	.285	—	.273	.879	.918	.047
100	2	2	.6	−.143	—	−.846	.265	—	.152	.819	.923	.000
			.2	−.139	—	−.733	.244	—	.180	.838	.923	.001
100	4	.3	.6	−.069	−.282	−.327	.122	.283	.065	.860	.781	.000
			.2	−.030	−.123	−.121	.098	.135	.093	.930	.810	.528
100	4	1	.6	−.001	−.044	−.204	.122	.121	.061	.920	.919	.050
			.2	−.001	−.022	−.051	.095	.091	.083	.949	.927	.786
100	4	2	.6	.009	−.011	−.005	.082	.064	.048	.954	.941	.953
			.2	.005	−.005	.073	.073	.065	.066	.955	.941	.731
100	8	.3	.6	−.015	−.101	−.144	.064	.088	.043	.923	.778	.056
			.2	−.008	−.052	−.066	.055	.056	.046	.946	.843	.608
100	8	1	.6	.006	−.033	−.071	.063	.055	.039	.956	.895	.501
			.2	.003	−.015	−.003	.052	.045	.044	.962	.931	.915
100	8	2	.6	.000	−.012	.027	.038	.038	.031	.940	.930	.883
			.2	.001	−.002	.075	.036	.036	.038	.944	.947	.432
100	16	.3	.6	.003	−.041	−.050	.037	.036	.028	.963	.793	.512
			.2	.002	−.025	−.030	.035	.031	.028	.963	.868	.780
100	16	1	.6	.000	−.024	−.024	.031	.030	.027	.950	.871	.826
			.2	.000	−.011	.001	.029	.028	.027	.949	.926	.933
100	16	2	.6	.000	−.011	.015	.025	.026	.023	.945	.922	.905
			.2	.000	−.003	.041	.024	.024	.025	.946	.945	.576
500	2	.3	.6	−.104	—	−.845	.163	—	.134	.827	.928	.000
			.2	−.253	—	−.733	.331	—	.356	.952	.926	.142

Table continues on overleaf

TABLE 2. *continued*

				bias			std			$ci_{.95}$		
N	T	ψ	ρ_0	$\widehat{\rho}_{al}$	$\widehat{\rho}_{ab}$	$\widehat{\rho}_{hk}$	$\widehat{\rho}_{al}$	$\widehat{\rho}_{ab}$	$\widehat{\rho}_{hk}$	$\widehat{\rho}_{al}$	$\widehat{\rho}_{ab}$	$\widehat{\rho}_{hk}$
500	2	1	.6	−.105	—	−.843	.163	—	.075	.830	.925	.000
			.2	−.114	—	−.730	.160	—	.122	.864	.924	.000
500	2	2	.6	−.104	—	−.843	.162	—	.067	.834	.927	.000
			.2	−.098	—	−.729	.147	—	.080	.843	.927	.000
500	4	.3	.6	−.040	−.067	−.325	.076	.144	.030	.867	.900	.000
			.2	−.017	−.030	−.116	.051	.071	.042	.922	.909	.079
500	4	1	.6	.012	−.009	−.202	.077	.056	.027	.942	.938	.000
			.2	.007	−.003	−.049	.051	.041	.037	.962	.950	.566
500	4	2	.6	.001	−.002	−.006	.033	.029	.021	.956	.947	.951
			.2	.001	−.001	.074	.032	.029	.030	.946	.946	.206
500	8	.3	.6	−.003	−.023	−.143	.039	.041	.019	.927	.912	.000
			.2	−.002	−.012	−.065	.030	.028	.020	.948	.920	.076
500	8	1	.6	.003	−.007	−.070	.030	.025	.017	.964	.940	.014
			.2	.001	−.003	−.002	.024	.020	.020	.960	.946	.911
500	8	2	.6	.000	−.002	.026	.017	.017	.014	.949	.949	.568
			.2	.000	.000	.077	.016	.016	.017	.952	.952	.004
500	16	.3	.6	.001	−.010	−.050	.017	.017	.013	.962	.909	.016
			.2	.001	−.006	−.029	.016	.014	.012	.961	.932	.316
500	16	1	.6	.000	−.005	−.023	.013	.014	.012	.953	.933	.459
			.2	.000	−.002	.002	.013	.012	.012	.948	.943	.932
500	16	2	.6	.000	−.002	.015	.011	.011	.010	.948	.946	.701
			.2	.000	.000	.042	.011	.011	.011	.951	.949	.028

Notes: Data generated as $y_{it} = \rho_{01} y_{it-1} + \rho_{02} y_{it-2} + \alpha_i + \varepsilon_{it}$ $(i = 1, \ldots, N; t = 1, \ldots T)$ with $\alpha_i \sim \mathcal{N}(0, 1)$, $\varepsilon_{it} \sim \mathcal{N}(0, 1)$, and ψ the degree of outlyingness of the initial observations $(y_{i0}, y_{i,-1})$. Entries: bias, standard deviation (std), and coverage rate of 95% confidence interval ($ci_{.95}$) of adjusted likelihood ($\widehat{\rho}_{al}$), Arellano-Bond ($\widehat{\rho}_{ab}$), and Hahn-Kuersteiner ($\widehat{\rho}_{hk}$) estimators; '—' indicates nonexistence of the moment; 10,000 Monte Carlo replications.

Table 3 presents results for the first-order autoregression with a covariate, generated as $x_{it} = \delta\alpha_i + \gamma\, x_{it-1} + u_{it}$ with $u_{it} \sim \mathcal{N}(0, \sigma_u^2)$ and x_{i0} drawn from the stationary distribution. The mean and variance of the stationary distribution of y_{it} are

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

We set $\delta = \sigma_u = .5$ 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).

The first part of Table 3 corresponds to moderate persistence in y and x, with $\gamma = \rho_0 = .5$. In this case, $\widehat{\theta}_{al}$ and $\widehat{\theta}_{ab}$ perform very reasonably, both for ρ_0 and β_0. $\widehat{\theta}_{hk}$ performs well for β_0, except when $T = 2$ and ψ is 0 or 1, but not for ρ_0,

TABLE 3. Simulation results for the first-order autoregression with a covariate

					bias			std			$ci_{.95}$		
N	T	ψ	γ	θ_0	$\widehat{\rho}_{al}$	$\widehat{\rho}_{ab}$	$\widehat{\rho}_{hk}$	$\widehat{\rho}_{al}$	$\widehat{\rho}_{ab}$	$\widehat{\rho}_{hk}$	$\widehat{\rho}_{al}$	$\widehat{\rho}_{ab}$	$\widehat{\rho}_{hk}$
500	2	0	.5	.5	−.052	—	−.662	.166	—	.067	.867	.934	.000
				.5	−.012	—	−.165	.115	—	.085	.965	.954	.284
500	2	1	.5	.5	.031	—	−.417	.170	—	.064	.929	.947	.000
				.5	.000	—	.008	.112	—	.091	.967	.954	.831
500	2	2	.5	.5	.005	—	.075	.070	—	.052	.955	.947	.648
				.5	−.001	—	−.020	.112	—	.114	.952	.951	.837
500	4	0	.5	.5	.011	−.019	−.245	.080	.058	.030	.950	.928	.000
				.5	.002	−.002	−.029	.057	.056	.054	.955	.949	.847
500	4	1	.5	.5	.003	−.014	−.124	.050	.049	.030	.958	.936	.006
				.5	.000	.002	.016	.056	.056	.054	.950	.949	.896
500	4	2	.5	.5	.000	−.005	.078	.028	.028	.025	.945	.945	.085
				.5	.000	.002	−.030	.057	.057	.057	.949	.949	.872
500	8	0	.5	.5	.000	−.011	−.061	.022	.024	.018	.951	.926	.054
				.5	.000	.001	.003	.033	.033	.033	.947	.947	.929
500	8	1	.5	.5	.000	−.010	−.025	.020	.022	.017	.950	.927	.619
				.5	.000	.002	.005	.033	.033	.033	.948	.947	.932
500	8	2	.5	.5	.000	−.005	.042	.015	.016	.015	.949	.936	.138
				.5	.000	.002	−.016	.033	.033	.033	.947	.947	.908
500	16	0	.5	.5	.000	−.008	−.010	.012	.013	.011	.949	.905	.818
				.5	.000	.002	.002	.021	.021	.021	.950	.950	.943
500	16	1	.5	.5	.000	−.008	−.002	.011	.012	.011	.951	.909	.928
				.5	.000	.002	.001	.021	.022	.021	.950	.949	.944
500	16	2	.5	.5	.000	−.006	.017	.010	.011	.010	.950	.920	.503
				.5	.000	.002	−.006	.022	.022	.021	.950	.949	.935
500	2	0	.99	.99	−.104	—	−.503	.164	—	.067	.829	.827	.000
				.01	.001	—	−.002	.121	—	.100	.978	.983	.833
500	2	1	.99	.99	−.103	—	−.503	.164	—	.067	.830	.827	.000
				.01	.002	—	.004	.121	—	.100	.978	.981	.835
500	2	2	.99	.99	−.091	—	−.485	.164	—	.067	.840	.888	.000
				.01	.003	—	.009	.122	—	.101	.979	.971	.833
500	4	0	.99	.99	−.057	−.566	−.254	.077	.259	.032	.840	.331	.000
				.01	.000	−.003	−.002	.055	.055	.052	.976	.957	.906
500	4	1	.99	.99	−.056	−.575	−.253	.077	.259	.032	.837	.325	.000
				.01	.001	.008	.004	.055	.055	.052	.975	.957	.905
500	4	2	.99	.99	−.045	−.199	−.236	.077	.152	.032	.853	.706	.000
				.01	.001	.007	.008	.056	.054	.052	.974	.951	.904
500	8	0	.99	.99	−.029	−.315	−.129	.038	.101	.017	.852	.042	.000
				.01	.000	−.001	−.001	.028	.030	.027	.975	.952	.924
500	8	1	.99	.99	−.028	−.317	−.128	.038	.101	.017	.854	.040	.000
				.01	.001	.011	.004	.028	.030	.027	.975	.937	.922
500	8	2	.99	.99	−.018	−.119	−.112	.038	.056	.017	.883	.394	.000
				.01	.001	.009	.008	.028	.028	.027	.975	.938	.913

Table continues on overleaf

TABLE 3. *continued*

					bias			std			$ci_{.95}$		
N	T	ψ	γ	θ_0	$\widehat{\rho}_{al}$	$\widehat{\rho}_{ab}$	$\widehat{\rho}_{hk}$	$\widehat{\rho}_{al}$	$\widehat{\rho}_{ab}$	$\widehat{\rho}_{hk}$	$\widehat{\rho}_{al}$	$\widehat{\rho}_{ab}$	$\widehat{\rho}_{hk}$
500	16	0	.99	.99	−.014	−.163	−.066	.019	.038	.009	.854	.000	.000
				.01	.000	.000	.000	.014	.016	.014	.972	.950	.931
500	16	1	.99	.99	−.014	−.162	−.065	.019	.038	.009	.857	.000	.000
				.01	.001	.012	.005	.014	.016	.014	.971	.893	.914
500	16	2	.99	.99	−.005	−.065	−.050	.019	.021	.008	.900	.082	.000
				.01	.001	.010	.008	.014	.015	.014	.970	.897	.898

Notes: Data generated as $y_{it} = \theta_{01} y_{it-1} + \theta_{02} x_{it} + \alpha_i + \varepsilon_{it}$, $x_{it} = .5\alpha_i + \gamma x_{it-1} + u_{it}$ $(i = 1, \ldots, N; t = 1, \ldots T)$ with $\alpha_i \sim \mathcal{N}(0, 1)$, $\varepsilon_{it} \sim \mathcal{N}(0, 1)$, $u_{it} \sim \mathcal{N}(0, .25)$, ψ the degree of outlyingness of the initial observations y_{i0}, and x_{i0} drawn from the stationary distribution. Entries: bias, standard deviation (std), and coverage rate of 95% confidence interval ($ci_{.95}$) of adjusted likelihood ($\widehat{\theta}_{al}$), Arellano-Bond ($\widehat{\theta}_{ab}$), and Hahn-Kuersteiner ($\widehat{\theta}_{hk}$) estimators; '—' indicates nonexistence of the moment; 10,000 Monte Carlo replications.

except when T is sufficiently large. The second part of Table 3 shows a case where y and x are highly persistent, with $\gamma = \rho_0 = .99$. All estimators of β_0 tend to improve, while those of ρ_0 deteriorate. The latter results are in line with those for the first-order autoregression without covariates: $\widehat{\rho}_{ab}$ deteriorates the most, while $\widehat{\rho}_{al}$ is only moderately biased and the corresponding confidence intervals have reasonable coverage.

The results presented here are a subset of a larger set of simulations that we ran with T ranging from 2 tot 24 and N from 100 to 10,000. The complete set of results is available as supplementary material. The results are in line with the tendencies discussed here.

7. CONCLUDING REMARKS

We studied how the incidental parameter problem in a fixed-effect autoregression can be solved by adjusting the (quasi-)likelihood. Our approach, based on removing the bias of the profile score, turned out to be equivalent to several other recent likelihood-based proposals, offering a unifying perspective on these methods. Perhaps our main finding is that, even in regular cases, the parameters locally maximize the expected adjusted profile log-likelihood, not globally. Given that this difficulty arises here in a linear model, one may speculate that it will arise in other models as well.

The adjusted likelihood estimator, accordingly defined as a local maximizer of the adjusted likelihood (or the local minimizer of the norm of its score), is asymptotically normal in regular cases. We have not investigated its limit distribution in the case where the adjusted Hessian, $H_a(\theta_0)$, is singular (which includes all cases where ρ_0 is a flat inflection point of the adjusted profile likelihood limit). In this case, while the estimator is still consistent, its convergence rate is likely to be slower than $N^{-1/2}$. Presumably the limit distribution and convergence rate could be derived using arguments along the lines of Rotnitzky, Cox, Bottai, and

Robins (2000), extending the results of Kruiniger (2014) to the higher-order autoregressive case.

The construction of an adjusted profile score can be extended in several directions, for example to models featuring cross-sectional and time-series heteroskedasticity (see also Alvarez and Arellano, 2004), although we have not studied the global properties in this case. It may be noted that the adjustment term, $-b(\rho)$, is robust to cross-sectional heteroskedasticity (σ_i^2 instead of instead of σ^2). We investigated this in an earlier version of this paper (Dhaene and Jochmans, 2007), where we also derived the adjustment for the first-order autoregressive model with fixed effects and individual time trends. Juodis (2012) derived the adjustment in the first-order vector autoregressive model with fixed effects. In the latter two cases, the adjustment term is still a polynomial that depends on the autoregressive parameters only.

A limitation of our approach is the exogeneity assumption with regard to the covariates. It would be of interest to investigate if the approach could be extended to deal with feedback of lagged y on covariates.

NOTE

1. Several other authors also make use of the Laplace approximation in connection with panel models featuring incidental parameters. For example, Arellano and Bonhomme (2009) use it to approximate the bias of the integrated likelihood in general nonlinear models, and Gagliardini and Gouriéroux (2014) apply it to approximate the integrated likelihood of a default-risk factor model, where the integration is over the time path of the unobserved factors.

REFERENCES

Ahn, S.C. & P. Schmidt (1995) Efficient estimation of models for dynamic panel data. *Journal of Econometrics* 68, 5–27.

Ahn, S.C. & G.M. Thomas (2006) Likelihood based inference for dynamic panel data models. Mimeo, Arizona State University, W.P. Carey School of Business.

Alvarez, J. & M. Arellano (2003) The time series and cross-section asymptotics of dynamic panel data estimators. *Econometrica* 71, 1121–1159.

Alvarez, J. & M. Arellano (2004) Robust likelihood estimation of dynamic panel data models. Working Paper #0421, CEMFI.

Arellano, M. (2003a) Discrete choices with panel data. *Investigaciones Económicas* 27, 423–458.

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

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

Arellano, M. & J. Hahn (2006) A likelihood-based approximate solution to the incidental parameter problem in dynamic nonlinear models with multiple effects. Working Paper #0613, CEMFI.

Bester, C.A. & C. Hansen (2009) A penalty function approach to bias reduction in non-linear panel models with fixed effects. *Journal of Business and Economic Statistics* 27, 131–148.

Blundell, R.W. & S. Bond (1998) Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics* 87, 115–143.

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

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

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

Dhaene, G. & K. Jochmans (2007) An adjusted profile likelihood for non-stationary panel data models with incidental parameters. Mimeo, K.U. Leuven, Department of Economics.

Dhaene, G. & K. Jochmans (2015) Split-panel jackknife estimation of fixed-effect models. *Review of Economic Studies*. doi:10.1093/restud/rdv007.

DiCiccio, T.J., M.A. Martin, S.E. Stern, & A. Young (1996) Information bias and adjusted profile likelihoods. *Journal of the Royal Statistical Society, Series B* 58, 189–203.

Gagliardini, P. & C. Gouriéroux (2014) Efficiency in large dynamic panel models with common factors. *Econometric Theory* 30, 961–1020.

Godambe, V.P. & M.E. Thompson (1974) Estimating equations in the presence of a nuisance parameter. *Annals of Statistics* 2, 568–571.

Hahn, J. & G. Kuersteiner (2002) Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and T are large. *Econometrica* 70, 1639–1657.

Hahn, J. & G. Kuersteiner (2011) Bias reduction for dynamic nonlinear panel models with fixed effects. *Econometric Theory* 27, 1152–1191.

Hahn, J. & W.K. Newey (2004) Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica* 72, 1295–1319.

Juodis, A. (2012) On the relative merits of bias correction methods in panel var models. MPhil Thesis, Tinbergen Institute.

Kalbfleish, J.P. & D.A. Sprott (1970) Application of likelihood methods to models involving large numbers of parameters (with discussion). *Journal of the Royal Statistical Society, Series B* 32, 175–208.

Kent, J.T. (1982) Robust properties of likelihood ratio tests. *Biometrika* 69, 19–27.

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. (2014) A further look at modified ML estimation of the panel AR(1) model with fixed effects and arbitrary initial conditions. Available at SSRN: http://ssrn.com/abstract=2559784.

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

Li, H., B. Lindsay, & R. Waterman (2003) Efficiency of projected score methods in rectangular array asymptotics. *Journal of the Royal Statistical Society, Series B* 65, 191–208.

McCullagh, P. & R. Tibshirani (1990) A simple method for the adjustment of profile likelihoods. *Journal of the Royal Statistical Society, Series B* 52, 325–344.

Neyman, J. & 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.

Pace, L. & A. Salvan (2006) Adjustments of profile likelihood from a new perspective. *Journal of Statistical Planning and Inference* 136, 3554–3564.

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

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

Sartori, N. (2003) Modified profile likelihood in models with stratum nuisance parameters. *Biometrika* 90, 533–549.

Severini, T.A. (1998) An approximation to the modified profile likelihood function. *Biometrika* 85, 403–411.

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

Sweeting, T.J. (1987) Discussion of the paper by Professors Cox and Reid. *Journal of the Royal Statistical Society, Series B* 49, 20–21.

Vuong, Q.H. (1989) Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57, 307–333.

White, H. (1982) Maximum-likelihood estimation of misspecified models. *Econometrica* 50, 1–25.
Woutersen, T. (2002) Robustness against incidental parameters. Working Paper 2002_8, University of Westerm Ontario, Department of Economics.
Woutersen, T. & M. Voia (2004) Efficient estimation of the dynamic linear model with fixed effects and regressors. Working Paper 2002_10, University of Westerm Ontario, Department of Economics.

APPENDIX: Proofs

Proof of Lemma 2.1. Using (2.2),

$$\begin{aligned}\operatorname{plim}_{N\to\infty} s_{\rho_j}(\theta_0) &= \frac{\operatorname{plim}_{N\to\infty} N^{-1}\sum_{i=1}^N \varepsilon_i^\top M S_j(\xi_{0i}+F_0\varepsilon_i)}{\operatorname{plim}_{N\to\infty} N^{-1}\sum_{i=1}^N \varepsilon_i^\top M\varepsilon_i}\\ &= \frac{\mathbb{E}\left(\varepsilon_i^\top M S_j F_0\varepsilon_i\right)}{\mathbb{E}\left(\varepsilon_i^\top M\varepsilon_i\right)} = \frac{\operatorname{tr} M S_j F_0}{T-1},\end{aligned}$$

$$\operatorname{plim}_{N\to\infty} s_{\beta_j}(\theta_0) = 0,$$

where ξ_{0i} and F_0 are ξ_i and F, evaluated at θ_0. We now write $\operatorname{tr} M S_j F_0$ in terms of the φ_t. Note that

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

where D_j^{-1} is the leading $(T-j)\times(T-j)$ block of D^{-1}. For arbitrary $\rho_1,\ldots,\rho_{T-1}$, 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 $\phi_1,\ldots,\phi_{T-1}$ are recursively obtained as $\phi_1=\rho_1$ and $\phi_j=\rho_j+\sum_{k=1}^{j-1}\phi_k\rho_{j-k}$, $j=2,\ldots,T-1$. Recursive substitution gives

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

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

$$\frac{\operatorname{tr} M S_j F_0}{T-1} = -\frac{\iota^\top D_j^{-1}\iota}{T(T-1)} = -\sum_{t=0}^{T-j-1}\frac{T-j-t}{T(T-1)}\varphi_t, \qquad j=1,\ldots,p,$$

which equals $b_j(\rho)$. ∎

Proof of Lemma 2.2. For $j=1,\ldots,p$, 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

$$B_{j,S}(\rho) = -\sum_{t=0}^{T-j-1} \frac{T-j-t}{T(T-1)} \sum_{k \in \mathcal{K}_{j,S}: \tau^\top k = t} \frac{\left(\iota^\top k\right)!}{k_1! \cdots (k_j+1)! \cdots k_p!} \rho_j \rho_S^{k_S}$$

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

$$B_{j,S}(\rho) = -\sum_{t=|S|-j}^{T-j-1} \frac{T-j-t}{T(T-1)} \sum_{k \in \mathcal{K}_S: \tau^\top k = t+j} \frac{\left(\iota^\top k - 1\right)!}{k_1! \cdots k_p!} \rho_S^{k_S},$$

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 $\tau^\top k = t + j$. 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 (2.4). 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. ■

Proof of Equation (3.3). In the parameterization $\eta_i = \alpha_i e^{-(T-1)a(\rho)}$, we have

$$\ell_i(\vartheta, \eta_i) = -\frac{1}{2}\log\sigma^2 - \frac{1}{2T\sigma^2} \sum_{t=1}^{T} \left(y_{it} - z_{it}^\top \theta - \eta_i e^{(T-1)a(\rho)}\right)^2 + \mathrm{c},$$

$$\nabla_{\eta_i} \ell_i(\vartheta, \eta_i) = \frac{e^{(T-1)a(\rho)}}{T\sigma^2} \left(y_i - Z_i\theta - \eta_i e^{(T-1)a(\rho)} \iota\right)^\top \iota,$$

and

$$\mathbb{E}_{\vartheta,\eta_i} \nabla_{\eta_i\eta_i} \ell_i(\vartheta, \eta_i) = -\sigma^{-2} e^{2(T-1)a(\rho)},$$
$$\mathbb{E}_{\vartheta,\eta_i} \nabla_{\sigma^2\eta_i} \ell_i(\vartheta, \eta_i) = 0,$$
$$\mathbb{E}_{\vartheta,\eta_i} \nabla_{\theta\eta_i} \ell_i(\vartheta, \eta_i) = -\sigma^{-2} e^{(T-1)a(\rho)} \left(\eta_i (T-1) b(\rho) e^{(T-1)a(\rho)} + \frac{\mathbb{E}_{\vartheta,\eta_i} Z_i^\top \iota}{T}\right).$$

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

$$\mathbb{E}_{\vartheta,\eta_i} y_{i,-j}^\top \iota = \iota^\top S_j \xi_i = \iota^\top D_j^{-1} \iota \eta_i e^{(T-1)a(\rho)} + T m_j,$$

where

$$m_j = \iota^\top S_j \begin{pmatrix} y_i^0 \\ D^{-1}\left(C y_i^0 + X_i \beta\right) \end{pmatrix} / T.$$

Hence,

$$\mathbb{E}_{\vartheta,\eta_i} Z_{i-}^\top \iota / T = -\eta_i (T-1) b(\rho) e^{(T-1)a(\rho)} + m,$$

where $m = \left(m_1, \ldots, m_p, \iota^\top X_i / T\right)^\top$ is free of η_i. Consequently,

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

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

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

$$\mathbb{E}_{\vartheta,\alpha_i} \nabla_{\alpha_i \alpha_i} l_i(\vartheta, \alpha_i) = -\sigma^{-2}, \qquad \mathbb{E}_{\vartheta,\alpha_i} \nabla_{\sigma^2 \alpha_i} l_i(\vartheta, \alpha_i) = 0,$$
$$\mathbb{E}_{\vartheta,\alpha_i} \nabla_{\theta \alpha_i} l_i(\vartheta, \alpha_i) = -\sigma^{-2} \mathbb{E}_{\vartheta,\alpha_i} Z_i^\top \iota / T,$$

and so, by the preceding proof,

$$A_i^{-1} B_i = -\begin{pmatrix} \mathbb{E}_{\vartheta,\alpha_i} Z_i^\top \iota / T \\ 0 \end{pmatrix} = -\begin{pmatrix} -(T-1) b(\rho) \alpha_i + m \\ 0 \end{pmatrix}.$$

Suppose some reparameterized fixed effect, say ζ_i, is 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 nonzero, while (A.2) implies $\nabla_{\rho_j \beta_{j'}} \alpha_i = 0$, so the equations are inconsistent. Suppose $q = 0$. Then

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

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 $T m_p$ and $T m_{p-1}$ 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

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

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

$$\nabla_{\rho_{p-1}} \varphi_t = \sum_{\tau^\top k = t-p+1} \frac{\left(\iota^\top k + 1\right)!}{k_1! \cdots k_p!} \rho^k,$$

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. ■

Proof of Equation (3.5). By the preceding proof,

$$\frac{\mathbb{E}_{\vartheta,\alpha_i}\nabla_{\alpha_i\vartheta}\ell_i}{\mathbb{E}_{\vartheta,\alpha_i}\nabla_{\alpha_i\alpha_i}\ell_i} = \begin{pmatrix}\mathbb{E}_{\vartheta,\alpha_i} Z_i^\top \iota/T \\ 0\end{pmatrix}$$

and so

$$g_i = \begin{pmatrix}\sigma^{-2}\left(Z_i^\top - \mathbb{E}_{\vartheta,\alpha_i} Z_i^\top \iota\iota^\top/T\right)(y_i - Z_i\theta - \iota\alpha_i) \\ \sigma^{-4}(y_i - Z_i\theta - \iota\alpha_i)^\top(y_i - Z_i\theta - \iota\alpha_i)/2 - \sigma^{-2}T/2\end{pmatrix}.$$

Recalling $\mathbb{E}_{\vartheta,\alpha_i} Z_i^\top \iota/T = -(T-1)b(\rho)\alpha_i + m$, we have

$$\nabla_{\alpha_i\alpha_i} g_i = \begin{pmatrix}-2\sigma^{-2}T(T-1)b(\rho) \\ \sigma^{-4}T\end{pmatrix}, \qquad \nabla_{\alpha_i\alpha_i}\ell_i = -\sigma^{-2}T,$$

and therefore $g_i - \frac{1}{2}\frac{\nabla_{\alpha_i\alpha_i} g_i}{\nabla_{\alpha_i\alpha_i}\ell_i}$ equals

$$\begin{pmatrix}\sigma^{-2}\left(Z_i^\top - \mathbb{E}_{\vartheta,\alpha_i} Z_i^\top \iota\iota^\top/T\right)(y_i - Z_i\theta - \iota\alpha_i) - (T-1)b(\rho) \\ \sigma^{-4}(y_i - Z_i\theta - \iota\alpha_i)^\top(y_i - Z_i\theta - \iota\alpha_i)/2 - \sigma^{-2}(T-1)/2\end{pmatrix}.$$

Evaluating at $\alpha_i = \widehat{\alpha}_i(\vartheta) = \iota^\top(y_i - Z_i\theta)/T$ and noting that $\nabla_{\alpha_i\alpha_i}\ell_i = 0$ gives (3.5). ■

Proof of Lemma 4.1. Let $A = S_1 F_0$ and $B = \nabla_{\rho_0} A$. Then

$$b_0 = -\frac{\iota^\top A\iota}{T(T-1)}, \qquad c_0 = -\frac{\iota^\top B\iota}{T(T-1)},$$

and

$$V_0^{LB} = \frac{\mathrm{tr}A^\top MA}{T-1} = \frac{T\mathrm{tr}AA^\top - \iota^\top AA^\top\iota}{T(T-1)}.$$

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

$$T\mathrm{tr}AA^\top - \iota^\top AA^\top\iota - \frac{2\left(\iota^\top A\iota\right)^2}{T(T-1)} \geq 0, \tag{A.3}$$

$$T\mathrm{tr}AA^\top - \iota^\top AA^\top\iota - \frac{2\left(\iota^\top A\iota\right)^2}{T(T-1)} - \iota^\top B\iota \geq 0. \tag{A.4}$$

The matrix $A = A_T$ is

$$A = \begin{pmatrix}0 & 0 \\ 1 & 0\end{pmatrix}, \qquad T = 2,$$

$$A = \begin{pmatrix}A_{T-1} & 0 \\ a_T^\top & 0\end{pmatrix}, \qquad a_T = \left(\rho^{T-2}, \rho^{T-3}, \ldots, 1\right)^\top, \qquad T > 2,$$

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

$$\iota^\top A\iota = \sum_{j=0}^{T-2}(T-j-1)\rho^j, \qquad \iota^\top B\iota = \sum_{j=1}^{T-2} j\,(T-j-1)\rho^{j-1},$$

$$\mathrm{tr}AA^\top = \sum_{j=0}^{T-2}(T-j-1)\rho^{2j}, \qquad \iota^\top AA^\top\iota = \sum_{j=0}^{T-2}\left(\sum_{k=0}^{j}\rho^k\right)^2,$$

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

$$\begin{aligned}\Delta Q_T &= \Delta\left(T\mathrm{tr}AA^\top-\iota^\top AA^\top\iota-2\frac{\left(\iota^\top A\iota\right)^2}{T(T-1)}-\iota^\top B\iota\right)_T\\ &=\left\{\left(\mathrm{tr}AA^\top\right)_{T+1}-2\frac{\left(\iota^\top A\iota\right)^2_{T+1}}{T(T+1)}\right\}+\left\{2\frac{\left(\iota^\top A\iota\right)^2_T}{T(T-1)}-\Delta\left(\iota^\top B\iota\right)_T\right\}\\ &\quad+\left\{T\Delta\left(\mathrm{tr}AA^\top\right)_T-\Delta\left(\iota^\top AA^\top\iota\right)_T\right\}\end{aligned}$$

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

$$\begin{aligned}\tau_1 &= \sum_{j=0}^{T-1}(T-j)\rho^{2j}-\frac{2}{T(T+1)}\left(\sum_{j=0}^{T-1}(T-j)\rho^j\right)^2\\ &=\frac{2}{T(T+1)}\left(\sum_{i=0}^{T-1}\sum_{j=0}^{T-1}(T-i)(T-j)(\rho^{2j}-\rho^{i+j})\right)=\frac{2}{T(T+1)}u^\top Ru,\end{aligned}$$

where $u=(T,T-1,\ldots,1)^\top$ 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, rank $(R) \leq 2$. Therefore, R is positive semi-definite and $\tau_1 \geq 0$. Furthermore,

$$\tau_3 = T\sum_{j=0}^{T-1}\rho^{2j} - \left(\sum_{j=0}^{T-1}\rho^{j}\right)^2 = \sum_{i=0}^{T-1}\sum_{j=0}^{T-1}\rho^{2j} - \sum_{i=0}^{T-1}\sum_{j=0}^{T-1}\rho^{i+j} = \iota^\top R\iota \geq 0.$$

Use

$$\left(\iota^\top A\iota\right)_T^2 = \left(\sum_{j=0}^{T-2}(T-j-1)\rho^j\right)^2 = \sum_{j=0}^{2T-4} g_j\rho^j,$$

$$\Delta\left(\iota^\top B\iota\right)_T = \sum_{j=1}^{T-1} j\,(T-j)\,\rho^{j-1} - \sum_{j=1}^{T-2} j\,(T-j-1)\,\rho^{j-1} = \sum_{j=0}^{T-2} h_j\rho^j,$$

where

$$\begin{aligned} g_j &= \sum_{\substack{0\leq k,l\leq T-2\\ k+l=j}} (T-k-1)(T-l-1) \\ &= \begin{cases} (T-1)(T-j-1)(j+1)+j(j-1)(j+1)/6 & \text{if } 0\leq j\leq T-2, \\ (2T-j-1)(2T-j-2)(2T-j-3)/6 & \text{if } T-2<j\leq 2T-4, \end{cases} \\ h_j &= j+1, \qquad 0\leq j\leq T-2, \end{aligned}$$

to write τ_2 as a polynomial

$$\tau_2 = \sum_{j=0}^{2T-4} q_j\rho^j, \qquad q_j = \begin{cases} dg_j - h_j & \text{if } 0\leq j\leq T-2, \\ dg_j & \text{if } T-2<j\leq 2T-4, \end{cases}$$

where $d = \frac{2}{T(T-1)}$. When $T = 2$ or $\rho = 1$, $\tau_2 = 0$. For $T > 2$,

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

and, therefore, $\tau_2 = (1-\rho)^2 P(T,\rho)$, where $P(T,\rho) = \sum_{j=0}^{2T-6} p_j\rho^j$ is a polynomial of degree $2T-6$, with coefficients $p_j = \sum_{i=0}^{j}(j+1-i)\,q_i$ given by

$$p_j = \begin{cases} d\binom{j+3}{j-2} + d\binom{j+3}{j}(T-1)(T-j-2)/2 & \text{if } 0\leq j\leq T-2, \\ d\binom{2T-j-1}{2T-j-6} & \text{if } T-2<j\leq 2T-6. \end{cases}$$

Hence $\tau_2 \geq 0$ because $p_j > 0$. (An earlier version of this proof stated that p_j increases in j for $j \geq T-2$. This was incorrect, as noted by Kruiniger, 2014.) 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 4.1. $L_a(\rho)$ 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 $\left[\underline{\rho}, \overline{\rho}\right]$ at most once. From

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

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

$$b(\rho) = -\sum_{t=0}^{T-2} \frac{T-1-t}{T(T-1)} \rho^t,$$

$$c(\rho) = -\sum_{t=1}^{T-2} \frac{t(T-1-t)}{T(T-1)} \rho^{t-1}, \quad d(\rho) = -\sum_{t=2}^{T-2} \frac{t(t-1)(T-1-t)}{T(T-1)} \rho^{t-2}.$$

For $\rho \neq 1$,

$$b(\rho) = -\frac{T-1-T\rho+\rho^T}{T(T-1)(1-\rho)^2},$$

$$c(\rho) = -\frac{T-2-T\rho+T\rho^{T-1}-(T-2)\rho^T}{T(T-1)(1-\rho)^3},$$

$$d(\rho) = -\frac{2T-6-2T\rho+T(T-1)\rho^{T-2}-2T(T-3)\rho^{T-1}+(T-2)(T-3)\rho^T}{T(T-1)(1-\rho)^4}.$$

When $T \leq 3$, $b(\rho)$ is linear and so, given that $S(\rho)$ is concave-convex on $\left[\underline{\rho}, \overline{\rho}\right]$, $b(\rho) \uparrow S(\rho)$ on $\left[\underline{\rho}, \overline{\rho}\right]$ 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 $\left[\underline{\rho}, \overline{\rho}\right]$ can only be on $\left(\rho_{\mathrm{ml}}, \overline{\rho}\right]$, 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,

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

so $b(\rho)$ is strictly convex on $(-\infty, \rho_v]$ and strictly concave on $[\rho_v, \infty)$ for some $\rho_v \in \left(-1, -\frac{1}{2}\right)$ and decreases on $\mathbb{R}$. Define ρ_u by $b(\rho_u) = 0$, that is, by $T(1-\rho_u) = 1 - \rho_u^T$, $\rho_u \in \mathbb{R}_-$. Since $T \geq 5$, we have $-2 < \rho_u < -1$. Thus, $b(\rho)$ is negative and strictly convex on $\left(\rho_u, \rho_v\right]$, with $-2 < \rho_u < -1 < \rho_v < -\frac{1}{2}$. Let $R = \left[\rho_u, \rho_v\right] \cap \left[\rho_{\mathrm{ml}}, \overline{\rho}\right]$. If R is empty, then $\rho_v < \rho_{\mathrm{ml}}$ or $\overline{\rho} < \rho_u$; in either case, by the concavity-convexity of $S(\rho)$, $b(\rho) \uparrow S(\rho)$

on $\left[\underline{\rho},\overline{\rho}\right]$ at most once. If R is nonempty, to show that $b(\rho)\uparrow S(\rho)$ on $\left[\underline{\rho},\overline{\rho}\right]$ 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), $\rho_{\text{ml}}=\rho_0+b_0/V_0\geq\rho_0+2b_0>-\frac{1}{2}$ 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, $\rho_{\text{ml}}>0$ 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 $S(\rho_w)=b(\rho_v)$, $\rho_w\in[\rho_{\text{ml}},\overline{\rho}]$; and ρ'_w by $S(\rho'_w)=b(0)=-\frac{1}{T}$, $\rho'_w\in[\rho_{\text{ml}},\overline{\rho}]$. Then $\rho_w-\rho_{\text{ml}}<\rho'_w-\rho_{\text{ml}}=\frac{1}{2}\left(T-\sqrt{T^2-4\zeta_0^2}\right)$. 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 $[\rho_{\text{ml}},\overline{\rho}]$ and $H(\rho'_w)$ decreases in T and increases in ζ_0^2,

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

and so, $H(\rho)<-\frac{1}{2}$ for $\rho\in[\rho_{\text{ml}},\rho_w]$. 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 $\left[\underline{\rho},\overline{\rho}\right]$ 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

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

using $\left(\sum_{k=0}^j\rho^k\right)^2=\sum_{k=0}^j(k+1)\rho^k+\sum_{k=1}^j(j-k+1)\rho^{j+k}$. 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 $d_k=\sum_{j=0}^k\left(v_{2j}+v_{2j+1}\right)-\frac{T(T-1)}{2}$. 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, $f_0 = (T-2)(T-5) \geq 0$, $f_1 = \frac{1}{2}\left(3T^2 - 25T + 54\right) > 0$, 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. ■

Proof of Lemma 4.2. Use $y_{i,-1} = S_1(\xi_{0i} + F_0\varepsilon_i)$ to write $Z_i = (y_{i,-1}, X_i) = (S_1 F_0 \varepsilon_i, 0) + \Xi_i$, where $\Xi_i = (S_1\xi_{0i}, X_i)$ is independent of ε_i. Proceeding as above, we have

$$V_0 = \frac{\operatorname{plim}_{N\to\infty} \frac{1}{N}\sum_{i=1}^N Z_i^\top M Z_i}{\sigma_0^2(T-1)} = \begin{pmatrix} V_0^{LB} & 0 \\ 0 & 0 \end{pmatrix} + V_\Xi,$$

where

$$V_\Xi = \begin{pmatrix} V_{\xi\xi} & V_{\xi X} \\ V_{X\xi} & V_{XX} \end{pmatrix} = \frac{\operatorname{plim}_{N\to\infty} \frac{1}{N}\sum_{i=1}^N \Xi_i^\top M \Xi_i}{\sigma_0^2(T-1)}$$

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 $(V_0^{\rho\rho})^{-1} = V_0^{LB} + V_{\xi\xi} - V_{\xi X}V_{XX}^{-1}V_{X\xi} \geq V_0^{LB}$. ■

Proof of Theorem 5.1. As $N \to \infty$, $l_{\mathrm{a}}(\rho)$ 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, $\widehat{\rho}_{\mathrm{ml}} \xrightarrow{p} \rho_{\mathrm{ml}}$, $\widehat{W} \xrightarrow{p} -H(\rho_{\mathrm{ml}}) = W_0$, and $\widehat{\mathcal{E}} \xrightarrow{p} \mathcal{E}$ in the sense that $\Pr\left[\rho \in \widehat{\mathcal{E}}\right] \to 1_{\{\rho\in\mathcal{E}\}}$ for any ρ not on the boundary of $\mathcal{E}$. Therefore,

$$\widehat{\rho}_{\mathrm{al}} \xrightarrow{p} \left\{ \arg\min_{\rho\in\mathcal{E}} S_{\mathrm{a}}^\top(\rho) S_{\mathrm{a}}(\rho) \quad \text{s.t.} \quad H_{\mathrm{a}}(\rho) \leq 0 \right\} = \rho_0 \tag{A.5}$$

and $\widehat{\beta}_{\mathrm{al}} = \widehat{\beta}(\widehat{\rho}_{\mathrm{al}}) \xrightarrow{p} \beta_0$. ■

Proof of Theorem 5.2. The adjusted likelihood estimator satisfies $s_{\mathrm{a}}\left(\widehat{\theta}_{\mathrm{al}}\right) = 0$ with probability approaching 1 as $N \to \infty$. By a Taylor series expansion of $s_{\mathrm{a}}(\theta)$ around θ_0,

$$0 = s_{\mathrm{a}}(\theta_0) + \nabla_{\theta'} s_{\mathrm{a}}(\theta_0)\left(\widehat{\theta}_{\mathrm{al}} - \theta_0\right) + O_p\left(||\widehat{\theta}_{\mathrm{al}} - \theta_0||^2\right).$$

Rearringing and using $\nabla_{\theta'} s_{\mathrm{a}}(\theta_0) = H_{\mathrm{a}}(\theta_0) + o_p(1)$ gives

$$\sqrt{N}\left(\widehat{\theta}_{\mathrm{al}} - \theta_0\right) = -H_{\mathrm{a}}(\theta_0)^{-1}\sqrt{N} s_{\mathrm{a}}(\theta_0) + o_p(1).$$

Write $s_{\mathrm{a}}(\theta_0)$ as

$$s_{\mathrm{a}}(\theta_0) = \frac{N^{-1}\sum_{i=1}^N Z_i^\top M\varepsilon_i}{N^{-1}\sum_{i=1}^N \varepsilon_i^\top M\varepsilon_i} - b_0 = \frac{\sigma_0^2(T-1)}{N^{-1}\sum_{i=1}^N \varepsilon_i^\top M\varepsilon_i} N^{-1}\sum_{i=1}^N e_i.$$

Since $\mathbb{E}e_i = 0$ and $\operatorname{plim}_{N\to\infty} N^{-1}\sum_{i=1}^N \varepsilon_i^\top M\varepsilon_i = \sigma_0^2(T-1)$, we have $\sqrt{N}s_{\mathrm{a}}(\theta_0) \xrightarrow{d} \mathcal{N}(0,\Sigma)$ and the result follows. ■

Proof of Equation (6.1). Since $\rho_{\mathrm{ml}} = \rho_0 + V_0^{-1}b_0$ for every T, the result follows on calculating $\lim_{T\to\infty} TV_0^{-1}b_0$. With $\rho_0 = (\rho_{01},\rho_{02})^\top$ in the stationary region, y_{it} is eventually stationary as $t\to\infty$ and hence

$$\lim_{T\to\infty} V_0 = \lim_{T\to\infty} \frac{\mathbb{E}\left(Y_{i-}^\top M Y_{i-}\right)}{\sigma_0^2(T-1)} = \sigma_0^{-2}\lim_{t\to\infty} \operatorname{Var}\begin{pmatrix} y_{it} \\ y_{it-1}\end{pmatrix}$$
$$= a^{-1}\begin{pmatrix} 1-\rho_{02} & \rho_{01} \\ \rho_{01} & 1-\rho_{02}\end{pmatrix}, \qquad \textbf{(A.6)}$$

where $a = (1+\rho_{02})(1-\rho_{01}-\rho_{02})(1+\rho_{01}-\rho_{02})$. In the steady state, we have, by the proof of Lemma 2.1,

$$y_{it} = \alpha_i + \rho_{01}y_{it-1} + \rho_{02}y_{it-2} + \varepsilon_{it} = \sum_{j=0}^\infty \varphi_j\left(\alpha_i + \varepsilon_{it-j}\right).$$

Therefore, with $b_0 = (b_{01}, b_{02})^\top$,

$$\lim_{T\to\infty} Tb_{0j} = \lim_{T\to\infty} \frac{T\mathbb{E}\left(\varepsilon_i^\top M y_{i,-j}\right)}{\mathbb{E}(\varepsilon_i^\top M\varepsilon_i)}$$
$$= -\sigma_0^{-2}\lim_{T\to\infty}\mathbb{E}\left(T^{-1}\left(\sum_{t=1}^T \varepsilon_{it}\right)\left(\sum_{t=1}^T y_{it-j}\right)\right)$$
$$= -\sigma_0^{-2}\lim_{t\to\infty}\mathbb{E}\left(\varepsilon_{it}\left(\sum_{j=0}^\infty y_{it+j}\right)\right) = -\sum_{j=0}^\infty \varphi_j = -\frac{1}{1-\rho_{01}-\rho_{02}} \qquad \textbf{(A.7)}$$

for $j = 1, 2$. From (A.6)–(A.7), it follows that $\lim_{T\to\infty} TV_0^{-1}b_0 = -\iota_2(1+\rho_{02})$, which implies (6.1). ■