Review of Economic Studies (2002) **69**, 647–666 0034-6527/02/00250647$02.00
© 2002 The Review of Economic Studies Limited

Orthogonal Parameters and Panel Data

TONY LANCASTER
Brown University

First version received 8 *March* 2000; *final version accepted* 22 *October* 2001 (*Eds.*)

This paper describes a class of consistent estimators for short panels with fixed effects. The method is to find an orthogonal reparametrization of the fixed effects and then to integrate the new effects from the likelihood with respect to an appropriately chosen prior density. The resulting marginal posterior densities of the common parameters have modes that are shown to be consistent in the models examined here. The main result concerns the first-order autoregressive model with agent specific intercepts where the likelihood is conditional on the set of initial observations. This paper provides a consistent likelihood-based estimator for this model. Some numerical illustrations are given. The first-order conditions for the posterior mode can also be thought of as new moment conditions for GMM estimation.

1. INTRODUCTION

The general failure of maximum likelihood in short panels with agent specific fixed effects has been known for 50 years, since the seminal paper by Neyman and Scott (1948). The principal reaction among econometricians has been the widespread abandonment of likelihood based methods of inference with such data. Instead, researchers look for valid instrumental variables and orthogonality conditions which are then combined in a generalized method of moments calculation.[1]

In this paper I propose a new class of procedures for consistent estimation in short panels with fixed effects. Unlike most recent work the method is likelihood based, but it is *not* maximum likelihood—it proceeds by integrating the likelihood, not maximizing it and therefore is essentially Bayesian in spirit. This approach makes it possible to secure (large N) consistent, likelihood-based, estimators of common parameters and, moreover, to make inferences that are not only consistent but also, unlike GMM procedures, exact for any size of panel. I shall present this argument in the same way that Neyman and Scott demonstrated the failure of maximum likelihood, by example. Specifically, I shall exhibit exact and consistent likelihood-based inference in the dynamic linear model with additive fixed effects. The approach I shall describe is applicable, in principle, to a large class of nonlinear models.

The methods depend on two ideas. The first is a reparametrization of the fixed effect such that the score for the new fixed effect is uncorrelated with that for the common parameter. To put it another way, the information matrix is block diagonal in the new parametrization. This orthogonal transformation achieves an exact or approximate likelihood separation of the fixed effect and the common parameter and paves the way for consistent inference about the latter even though, of course, consistent inference about the former is impossible in a panel of fixed length. The key features of a fixed effect are that it be agent specific and time invariant. In most

1. See, for example, Arellano and Bond (1991), Ahn and Schmidt (1995) and Arellano and Bover (1995). Such methods can produce estimators for the parameters that are common to all agents and which are consistent as the number of agents, N, becomes large.

applications any parametrization that respects these properties is as good as any other so there can be no economic objection to working with a redefined effect.

The second idea is uniform integration of the orthogonal fixed effect with respect to a prior exhibiting independence of the two sets of parameters. This has two aspects; the first is forming a prior in which (orthogonal) fixed effects and common parameters are independent. This independence is motivated by the fact that, by construction, these parameters are separate (in a sense to be described precisely) in the likelihood. The second aspect is integration of the orthogonal fixed effect with respect to a suitable uniform prior. The idea here is to be uninformative about the orthogonal fixed effects. "Uninformativeness" is a notoriously slippery notion and in the present applications we explore what seem to be the obvious ways of constructing such a prior, looking for a version that gives sensible results. Thus the choice among arguably uninformative priors for the new fixed effects is essentially pragmatic. We do not have to look far since the natural choices lead immediately to marginal posterior distributions for the common parameter from which exact inferences can be made, and whose modes are large N consistent for the parameter in the models considered here.

Two types of orthogonal reparametrization are defined in Section 2 and the approach advocated in this paper is then applied to three simple examples. The first is the panel Poisson count model which has been widely, but mistakenly, believed to suffer from an incurable incidental parameter problem. The second is the linear model with exogenous covariates and additive fixed effects. The third is the stationary first-order autoregressive linear panel data model with exogenous regressors and additive fixed effects. In Section 3 I move on to the main part of the paper and deal with the possibly non-stationary dynamic linear model with additive fixed effects in which the likelihood is conditioned on the initial observation.

In all these models I shall find an orthogonal reparametrization of the fixed effects whose uniform integration then gives a marginal posterior density for the common parameters whose mode is large N consistent. The new results in these sections include

(1) An exactly orthogonal fixed effect for the dynamic regression model in which the likelihood is conditioned on the initial observations.
(2) Consistent, likelihood-based estimators of the autoregressive coefficient in the dynamic regression model.
(3) New moment conditions for the dynamic model with additive fixed effects.
(4) Methods for exact, finite sample, inference in several panel data models.

In Section 4 I give a short account of previous work which exploits orthogonal reparametrization, in particular the important work of Cox and Reid (1987) who use orthogonalization as the basis for a frequentist, approximate conditional likelihood, approach for estimation of common parameters. Section 5 concludes.

2. ORTHOGONALIZATION AND THREE SIMPLE EXAMPLES

In this section I shall define and illustrate two definitions of parameter orthogonality and apply the Bayesian approach of this paper to three simple examples.

The setting is panel data models with N agents observed for T periods. The models have agent specific, time invariant, fixed effects, f_i, together with a parameter θ of fixed and finite dimension common to all agents. Conditional on parameters and observed covariates, observations for different agents are stochastically independent. The total number of parameters is N plus the dimension of the common parameter vector, say K. Direct application of maximum likelihood to estimate all parameters can yield inconsistent ($N \to \infty$) estimators for the

common parameters. This is the *incidental parameter problem* on which Neyman and Scott (1948) wrote the seminal paper. With a total of $N + K$ parameters standard theorems on the consistency of maximum likelihood, which require a parameter of fixed dimension, fail.

It has been less noticed[2] that incidental parameters raise an analogous problem for Bayesian inference about the common parameters. In the Bayesian approach the fixed effects would be integrated out of the likelihood with respect to a prior distribution conditional on the common parameters and covariates. Inference about the common parameters would then be based on their marginal posterior distribution. These inferences will depend, even asymptotically, on the choice of prior for the fixed effects and will not, in general, be consistent as $N \to \infty$ with T fixed.[3]

The objective is to provide consistent Bayesian inference about θ as $N \to \infty$ with T fixed. We shall denote the likelihood function for the data provided by a single agent as $\ell_i(f_i, \theta)$.

Suppose that ℓ_i factors as

$$\ell_i(f_i, \theta) = \ell_{i1}(f_i)\ell_{i2}(\theta) \tag{2.1}$$

where ℓ_{i1}, ℓ_{i2} are themselves likelihood functions. If a likelihood factors as (2.1) and the parameters f, θ are variation independent, they are *orthogonal* (Jeffreys (1960), Anscombe (1964)). By inspection of (2.1), the maximum likelihood estimator of θ is quite independent of the value taken by f_i and, as long as f_i and θ are independent *a priori*, Bayesian inference about θ does not depend at all on prior or posterior opinions about f_i. In these cases, so far as θ is concerned, we may take the likelihood to be (the product of terms like) ℓ_{i2}. Typically we can then show that Bayes or maximum likelihood estimators based on this product are consistent.[4]

Now whether a likelihood factors as in (2.1) depends on the parametrization. A model may factor in one parametrization but not in others. But, by assumption, we are interested in estimating θ, and the fixed effects are nuisance parameters whose definition we can change at will as long as they remain time invariant and agent specific. Thus we are free to seek a reparametrization of the fixed effects in such a way that they *are orthogonal to* θ. If we can find such a reparametrization there are incidental parameters but there is, in general, no incidental parameter problem. Notice that a redefinition of the fixed effects does not change the ml estimator of θ—this is the invariance property of maximum likelihood estimators. The redefinition of the fixed effects to put the likelihood in the form (2.1) merely reveals the absence of an incidental parameter problem.

2.1. *Example* 1

We can illustrate orthogonality using a simple but important econometric panel data model, that for a sequence of Poisson counts. Let the data consist of T Poisson counts y_{it} with means $E(y_{it}) = f_i \exp\{x_{it}\theta\}$, $i = 1, \ldots, N$, $t = 1, \ldots, T$ for each of N agents. Conditional on the unknown parameters $f = (f_i, \ldots, f_N)$ and θ, and upon the known covariates x_{it}, the y_{it} are stochastically independent. Therefore, the likelihood contribution of individual i is

$$\ell_i(f_i, \theta) \propto e^{-f_i \Sigma_t \exp\{x_{it}\theta\}} f_i^{\Sigma_t y_{it}} e^{\theta' \Sigma_t y_{it} x_{it}}, \tag{2.2}$$

2. Diaconis and Freedman (1986) develop examples of inconsistent Bayes inference in the presence of infinitely many incidental parameters. Wasserman (1998) provides a recent discussion of Bayesian inconsistency in the presence of infinite-dimensional parameter spaces.

3. This point has come up rather indirectly in the literature in the following way. From a formal point of view, integrating out fixed effects is equivalent to a "random effects" model with the prior distribution playing the role of a mixing distribution. It is well known, particularly in the duration literature, that inferences can be sensitive to the choice of mixing distribution.

4. Consistency will be immediate if l_{i2} is itself a probability density function.

(Here and throughout the paper I omit irrelevant constants of proportionality from probability distributions.) The expression (2.2) is not in the orthogonal form (2.1) but, by inspection, we can see that if we define a new fixed effect as

$$g_i = f_i \Sigma_t e^{\theta' x_{it}} \tag{2.3}$$

then the likelihood does take a product form. Under this reparametrization the likelihood contribution of agent i is

$$\ell_i(g_i, \theta) \propto e^{-g_i} g_i^{\Sigma_t y_{it}} \times \frac{e^{\theta' \Sigma_t y_{it} x_{it}}}{(\Sigma_t e^{\theta' x_{it}})^{\Sigma_t y_{it}}}. \tag{2.4}$$

This is precisely in the form (2.1) with fixed effects g_i and common parameter θ. Consequently, orthogonality in the sense of (2.1) can be achieved in this model. The term corresponding to ℓ_{i1} in (2.1) is a Poisson likelihood with parameter g_i; and the term corresponding to ℓ_{i2} is a multinomial likelihood with T cells, and cell probabilities proportional to $e^{\theta' x_{it}}$, conditional on a total number of events given by $\Sigma_t y_{it}$.

The ml estimator of θ maximizes the product over agents of these conditional multinomials. It is consistent for θ under conditions provided in Andersen (1970). Hausman *et al*. (1984), who used the Poisson model to study the dependence of patent filings on R&D expenditures, wrote "But we cannot simply estimate separate f_i parameters (in the Poisson model) because for θ held fixed and N large we have the incidental parameter problem and maximum likelihood need not be consistent (see Neyman and Scott, . . .). Instead we use the conditional maximum likelihood approach of Andersen and condition on the sum of patents $\Sigma_t\ Y_{it}$". The above analysis shows that conditional maximum likelihood is not an alternative to ml in this model, it is identical to it. And even though it is true that we cannot (consistently) estimate separate f_i parameters, the ml estimator of θ is consistent.[5] There are incidental parameters but there is no incidental parameter problem.

Exact Bayesian inference is straightforward in this model and as long as the prior for g, θ exhibits independence of these parameters the posterior mode for θ will be consistent as $N \to \infty$. ‖

The set of models whose likelihood can put be in the form (2.1) by reparametrization of the fixed effects appears to be small. But a weaker type of orthogonality can be achieved in many econometric models. A consequence of (2.1) is that

$$\frac{\partial^2 L_i(f_i, \theta)}{\partial f_i \partial \theta} = 0 \tag{2.5}$$

where $L_i = \log \ell_i$. If we cannot choose the parametrization such that (2.5) is identically true, but we can choose it such that (2.5) is true on average, we have *information orthogonality*. Specifically, if

$$E\left(\frac{\partial^2 L_i(f_i, \theta)}{\partial f_i \partial \theta}\right) = 0 \tag{2.6}$$

and f, θ are variation independent we have an information orthogonal parametrization of the fixed effects.[6] Equation (2.6) states that the information matrix is block diagonal.

An information orthogonal reparametrization, if it exists, may be found by the following argument. Let f be the original effect and g the proposed orthogonal effect (dropping the i

5. Frank Windmeijer in Blundell *et al*. (1997) independently noted the consistency of maximum likelihood in the panel Poisson count model.

6. Cox and Reid (1987) is the principal modern reference on information orthogonality.

subscripts for simplicity). Write $f = f(\theta, g)$ where θ is the K-dimensional common parameter. If $L(\theta, g)$ is the log likelihood for agent i and superscripts denote derivatives we must have

$$L^g = \frac{\partial f}{\partial g} L^f,$$
$$E(L^{g\theta}) = \frac{\partial f}{\partial g}\left[\frac{\partial f}{\partial \theta} E(L^{ff}) + E(L^{f\theta})\right]$$

where the second equation follows from the fact that $E(L^f) = 0$. Since we require the function $f(\cdot)$ to be such that $E(L^{g\theta}) = 0$ we see that it must solve the system of differential equations

$$\frac{\partial f}{\partial \theta} = -[E(L^{ff})]^{-1} E(L^{f\theta}). \tag{2.7}$$

The orthogonal parameter g may then be introduced as a constant of integration in the solution of this first-order differential equation.

We may illustrate this procedure using the Poisson model of Example 1.

$$L = -f\Sigma e^{\theta x_t} + \log f \Sigma_t y_t + \theta \Sigma y_t x_t; \qquad L^f = -\Sigma_t e^{\theta x_t} + \Sigma_t y_t / f;$$
$$L^{ff} = -\Sigma_t y_t / f^2; \qquad E(L^{ff}) = -\Sigma_t e^{\theta x_t}/f; \qquad L^{f\theta} = -\Sigma_t x_t e^{\theta x_t} = E(L^{f\theta}).$$

Thus the differential equation is

$$\frac{1}{f}\frac{\partial f}{\partial \theta} = -\frac{\Sigma_t x_t e^{\theta x_t}}{\Sigma_t e^{\theta x_t}},$$

with general solution

$$\log f = -\log \Sigma_t e^{\theta x_t} + \log g,$$

where the arbitrary constant of integration has been written as $\log g$. This immediately implies the orthogonal parameter

$$g = f\Sigma e^{\theta x_t}$$

given earlier in (2.3). Note that since we could have written the constant of integration in many alternative ways, orthogonal parameters are not unique.

An information orthogonal parameterization may not exist since the system (2.7) may have no solution. Non-existence can be shown by examining the cross-partial derivatives implied by the system. If these are unequal the equations are inconsistent. We shall note an instance of this in Section 3.

2.2. *Example* 2

For a model in which orthogonality cannot be achieved but information orthogonality can, consider the panel linear model with exogenous covariates. Let $y_{it} = f_i + x_{it}\beta + u_{it}$, where the u_{it} are independently $n(0, \sigma^2)$ conditional on the regressor sequence, f_i, and $\theta = (\beta, \sigma^2)$. In vector terms

$$y_i = \jmath f_i + x_i\beta + u_i; \qquad u_i | x_i, f_i, \theta \sim n(0, \sigma^2 I_T), \qquad i = 1, \ldots, N, \tag{2.8}$$

where $\jmath$ contains T ones. The likelihood for the information provided by agent i is

$$\ell_i \propto \frac{1}{\sigma^T} \exp\{-(1/2\sigma^2)(y_i - f_i\jmath - x_i\beta)'(y_i - f_i\jmath - x_i\beta)\}.$$

It is apparent that ℓ will not factor but an information orthogonal reparametrization of f_i is easily found to be $g_i = f_i + \bar{x}_i\beta$ where $\bar{x}_i$ contains the time means of the K regressor vectors for agent i.

This familiar transformation is such that the information matrix for g_i, β, σ^2 is block diagonal as between g_i and β, σ^2 as can be easily verified.

The second part of the proposed program of inference is the uniform integration of the orthogonal fixed effect. To see how this works write the model in terms of g_i, β, σ^2,

$$y_i = \jmath g_i + \tilde{x}_i\beta + u_i, \qquad u_i|x_i, g_i, \beta, \sigma^2 \sim n(0, \sigma^2 I_T)$$

where $\tilde{x}_i = x_i - \bar{x}_i \jmath$. Now y_i is equivalent to $\bar{y}_i$ and Dy_i where D is the $T-1 \times T$ first differencing matrix with typical row $(0, \ldots, 0, 1, -1, 0, \ldots, 0)$. Moreover $\bar{y}_i \sim n(g_i, \sigma^2/T)$ and $Dy_i \sim n(Dx_i\beta, \sigma^2 DD')$, and $\bar{y}$, Dy_i are independent. Thus the likelihood for agent i is

$$\ell_i(f_i, \theta) \propto \frac{1}{\sigma}\exp\{-(T/2\sigma^2)(\bar{y}_i - g_i)^2\} \times \frac{1}{\sigma^{T-1}}\exp\{-(1/2\sigma^2)(y_i - x_i\beta)'D'(DD')^{-1}D(y_i - x_i\beta)\}.$$

Taking the prior for g, β, σ^2 as $\propto \pi(\beta, \sigma^2)$ so that g_i are independently uniformly distributed on the real line we see that g_i are distributed independently as $n(\bar{y}_i, \sigma^2/T)$ given σ^2, *a posteriori*. Carrying out the integration of the g_i for each i and collecting terms gives the marginal posterior of β, σ^2 as

$$p(\beta, \sigma^2) \propto \frac{1}{\sigma^{N(T-1)}}\exp\{-(1/2\sigma^2)\Sigma_i(y_i - x_i\beta)'D'(DD')^{-1}D(y_i - x_i\beta)\}\pi(\beta, \sigma^2).$$

This is, of course, just the posterior density based on the first differenced data. Under a standard flat prior for β and $\log\sigma$, the β mode is just the "within" estimator; and that of σ^2 is the residual sum of squares divided by the "correct" degrees of freedom, $N(T-1)$, apart from a negligible term of $O(1/NT)$. These estimators are consistent for β, σ^2. In contrast, as is well known since this is essentially Neyman and Scott's Example 1, the maximum likelihood estimator of σ^2 is inconsistent as $N \to \infty$.[7]

It is critical to the method we are illustrating here that there are *two* steps to carry out. The first is development of an orthogonal fixed effect; the second is integration of that fixed effect with respect to a prior distribution. There is absolutely no point whatever in defining an orthogonal fixed effect and then maximizing the likelihood with respect to it and the common parameter θ since, by invariance, the ml estimator of θ will be independent of the parametrization of the fixed effect. ‖

2.3. *Example* 3

As a final introductory example consider a stationary dynamic version of the last example in which the scalar error covariance matrix is replaced by that of the stationary first-order autoregressive model:

$$y_i = \jmath f_i + x_i\beta + u_i, \qquad u_i|x_i,\ f_i, \theta \sim n(0, \sigma^2 V), \tag{2.9}$$

where

$$V = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & \rho & \ldots & \rho^{T-1} \\ \rho & 1 & \ldots & \cdot \\ \cdot & \cdot & \ldots & \cdot \\ \rho^{T-1} & \cdot & \ldots & 1 \end{pmatrix} \qquad |\rho| < 1,$$

7. The ml estimator of each f_i is $\bar{y}_i$ so the concentrated likelihood is $\sigma^{-NT}\exp\{-(1/2\sigma^2)\Sigma_i u_i' D'(DD')^{-1}u_i\}$ whose mode is the within residual sum of squares divided by NT. This converges to $\sigma^2(T-1)/T$.

where $\theta = (\beta, \rho, \sigma^2)$. We seek a transformation of f_i such that it is information orthogonal to θ. By the solution of the differential equation (2.7) or otherwise we find that

$$g_i = f_i + (\jmath' V^{-1} \jmath)^{-1} \jmath' V^{-1} x_i \beta \tag{2.10}$$

is information orthogonal to θ. Rewriting the model in the new parametrization we have

$$y_i = \jmath g_i + \widetilde{x}_i \beta + u_i, \qquad u_i | x_i, g_i, \theta \sim n(0, \sigma^2 I_T)$$

where $\widetilde{x}_i = x_i - \jmath (\jmath' V^{-1} \jmath)^{-1} \jmath' V^{-1} x_i$.

As in Example 2 we consider the likelihood for $\overline{y}$, Dy which, exploiting $D\jmath = 0$, satisfy $Dy_i = Dx_i \beta + Du_i$ and $\overline{y}_i = g_i \overline{u}_i$ with error covariance matrix

$$V \begin{pmatrix} Du_i \\ \overline{u}_i \end{pmatrix} | x_i, f_i, \theta = \sigma^2 \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} = \sigma^2 \begin{pmatrix} DVD' & DV\jmath/T \\ \jmath' V D'/T & \jmath' V \jmath / T^2 \end{pmatrix}.$$

Note that, unlike Example 2, Dy and $\overline{y}$ are dependent, but we may still consider their joint distribution, as the product of the marginal of Dy and the conditional of $\overline{y}$ given Dy. To write these distributions concisely set $\Omega = DVD'$ and $\omega = V_{22} - V_{21} V_{11}^{-1} V_{12}$. These are the variance of Du and the conditional variance of $\overline{u}$ given Du respectively. Then we have

$$\begin{aligned} p(Dy_i) &\propto \sigma^{-(T-1)} |\Omega|^{-1/2} \exp\{-(1/2\sigma^2) u_i' D' \Omega^{-1} Du_i\}; \\ p(\overline{y}_i | Dy_i) &\propto \sigma^{-1} \omega^{-1/2} \exp\{-(1/2\omega\sigma^2)(\overline{u}_i - V_{21} V_{11}^{-1} Du_i)^2\}; \\ \text{where } \overline{u}_i &= \overline{y}_i - g_i; \qquad Du_i = Dy_i - Dx_i \beta. \end{aligned}$$

As before, we take the prior of g, β, ρ, σ^2 to be of the form $\pi(\beta, \rho, \sigma^2)$ so that g_i are independently uniformly distributed on the real line. Then integration of g_i which are independently normally distributed *a posteriori* leads to the marginal posterior density

$$\begin{aligned} p(\beta, \rho, \sigma^2 | data) \propto{}& \sigma^{-N(T-1)} |\Omega|^{-N/2} \\ & \times \exp\{-(1/2\sigma^2)(y_i - x_i \beta)' D' \Omega^{-1} D(y_i - x_i \beta)\} \pi(\beta, \rho, \sigma^2). \end{aligned} \tag{2.11}$$

This is the posterior density that would be formed solely from the likelihood of the first differenced data. Its mean for β given ρ is the generalized least squares estimator based on those data. ‖

The second and third examples show that uniformly integrating an orthogonally parameterized fixed effect leads to inference based on the first differenced data.[8] Indeed, this is a Bayesian interpretation of first differencing in linear models. As soon as one moves outside the simple linear model framework, however, we find that uniform integration of an orthogonal fixed effect does **not** lead to inference based on the likelihood for the first differenced data.

3. THE NONSTATIONARY DYNAMIC REGRESSION MODEL

The dynamic model that argues conditionally on observed initial conditions and permits the autoregressive parameter to equal or exceed unity appears to be the dominant one in the literature. This is a radically different model from that of Example 3 and its large N consistent inference has provoked the production of a large amount of (frequentist) literature. In this section we analyze Bayesian inference in this model using the program of uniform integration of an orthogonal fixed effect. We show that there exists an orthogonal reparametrization of the fixed effect in this model

8. Example 3 may be generalized to the case of an arbitrary $T \times T$ error covariance matrix. This generalization gives a Bayesian interpretation of first differencing in panel data models and a Bayesian interpretation of the approach of Kiefer (1980). See Hansen (2000).

and that uniform integration of such an effect leads to a marginal posterior distribution for the autoregressive parameter whose mode, if it exists, is large N consistent. A by-product of the calculation is a new set of moment conditions for the model.

The model for agent $i = 1, 2, \ldots, N$ is

$$y_i = f_i \iota + \rho y_{i-} + x_i \beta + u_i, \qquad u_i | f_i, \theta, x_i, y_{i0} \sim n(0, \sigma^2 I_T). \tag{3.1}$$

Here y_{i-} is the T dimensional column vector $(y_{i0}, y_{i1}, \ldots, y_{i,T-1})$ and $\theta = (\rho, \beta, \sigma^2)$. We shall impose the condition that the columns of the $T \times K$ matrix x_i are measured from their agent specific means, which is an initial redefinition of the fixed effects. The parameter space is such that β and $\log \sigma^2$ are unrestricted and

$$-\rho_l \leq \rho \leq \rho_u, \qquad \rho_l, \rho_u \geq 0.$$

In particular, ρ is allowed to exceed or equal 1 in modulus. Note that the likelihood is conditioned on the set of N initial values y_{i0} which are not modelled. They are N parameters whose values are revealed when the data are observed. Their prior distribution is irrelevant as long as it is not dogmatic.

A simplification of subsequent analysis arises when we can take the origin of each of the processes to be zero. This may be achieved by working with the likelihood for $y_{it} - y_{i0}$, a sequence satisfying

$$y_{it} - y_{i0} = [f_i - y_{i0}(1 - \rho)]\iota + \rho(y_{it-1} - y_{i0}) + x_{it}\beta + u_{it}$$

under the same stochastic assumptions as (3.1). Henceforth we shall interpret y_{it} as $y_{it} - y_{i0}$ and f_i as $f_i - y_{i0}(1 - \rho)$ and take the origin for every agent as zero.

To find the orthogonal reparametrization we form the differential equation (2.7) and for this we require, first of all, the elements of the information matrix corresponding to f_i, θ. The log likelihood (for agent i) is

$$L_i(f_i, \theta) = -\frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y_i - f_i\iota - \rho y_{i-} - x_i\beta)'(y_i - f_i\iota - \rho y_{i-} - x_i\beta). \tag{3.2}$$

Its derivative with respect to f_i is

$$\frac{\partial L_i}{\partial f_i} = \frac{1}{\sigma^2}(y_i - f_i\iota - \rho y_{i-} - x_i\beta)'\iota = \frac{1}{\sigma^2}(y_i - f_i\iota - \rho y_{i-})'\iota \tag{3.3}$$

since the columns of x_i have zero sum. Consequently, the elements of the f_i, θ part of the information matrix are

$$E\left(\frac{\partial^2 L_i}{\partial f_i \partial \beta}\right) = 0, \tag{3.4}$$

$$E\left(\frac{\partial^2 L_i}{\partial f_i \partial \sigma^2}\right) = -\frac{1}{\sigma^4}E(y_i - f_i\iota - \rho y_{i-})'\iota = 0, \tag{3.5}$$

$$E\left(\frac{\partial^2 L_i}{\partial f_i \partial \rho}\right) = -\frac{TE(\overline{y}_{i-})}{\sigma^2} \neq 0, \tag{3.6}$$

where $\overline{y}_{i-} = (y_{i0} + y_{i1} + \ldots y_{i,T-1})/T$.

The vector of lagged values has the representation

$$y_{i-} = f_i c + Cx_i\beta + Cu_i \tag{3.7}$$

where C is a $T \times T$ matrix and c is a $T \times 1$ vector whose elements are polynomials in ρ:

$$c = \begin{pmatrix} 0 \\ 1 \\ 1+\rho \\ \cdot \\ 1+\rho+\rho^2+\ldots+\rho^{T-2} \end{pmatrix}; \qquad C = \begin{pmatrix} 0 & 0 & 0 & \ldots & 0 \\ 1 & 0 & 0 & \ldots & 0 \\ \rho & 1 & 0 & \ldots & 0 \\ \cdot & \cdot & \cdot & \ldots & \cdot \\ \rho^{T-2} & \rho^{T-3} & \ldots & 1 & 0 \end{pmatrix}.$$

3.1. *No covariates*

We first take the case in which $\beta = 0$ so we have a pure first-order autoregression.

The value of $TE(\bar{y}_{i-})$ is then $f_i \jmath' c$ where $\jmath' c = \sum_{t=1}^{T-1}(T-t)\rho^{t-1}$. Define the function

$$b(\rho) = \frac{1}{T}\sum\nolimits_{t=1}^{T-1}\frac{T-t}{t}\rho^t, \tag{3.8}$$

whose derivative is $\jmath' c/T$. Then, in the absence of covariates,

$$E(\bar{y}_{i-}) = f_i b'(\rho)$$

so the required expression for the f_i, ρ element of the expected Hessian is

$$E\left(\frac{\partial^2 L_i}{\partial f_i \partial \rho}\right) = -\frac{TE(\bar{y}_{i-})}{\sigma^2} = -\frac{f_i T b'(\rho)}{\sigma^2}. \tag{3.9}$$

This last result tells us that in this model the fixed effects are *not* information orthogonal to the common parameters. So let us find an information orthogonal reparametrization $f_i = f_i(g_i, \rho)$ such that

$$E\left(\frac{\partial^2 L_i}{\partial g_i \partial \rho}\right) = 0$$

where the new fixed effects are also information orthogonal to β, σ^2.

To find an orthogonal fixed effect requires solving (2.7) which, using (3.4)–(3.6), (3.9) and $E(L^{ff}) = -T/\sigma^2$, is

$$\begin{pmatrix} \frac{\partial f_i}{\partial \rho} \\ \frac{\partial f_i}{\partial \sigma^2} \end{pmatrix} = \frac{\sigma^2}{T}\begin{pmatrix} -\frac{T f_i b'(\rho)}{\sigma^2} \\ 0 \end{pmatrix} = \begin{pmatrix} -f_i b'(\rho) \\ 0 \end{pmatrix}. \tag{3.10}$$

These equations are consistent and they reduce to

$$\frac{\partial f_i}{\partial \rho} = -f_i b'(\rho) \tag{3.11}$$

which has the general solution

$$f_i = g_i e^{-b(\rho)}. \tag{3.12}$$

The parameter g_i is a version of the arbitrary constant of integration and will act as (a version of) the orthogonalized fixed effect.

To interpret g_i consider the behavior of the function $b(\rho)$. For all $\rho \geq -1$ we find that

$$\lim_{T\to\infty} e^{-b(\rho)} = \begin{cases} 1-\rho & -1 \leq \rho \leq 1 \\ 0 & \rho > 1 \end{cases}$$

while there is no limit for $\rho < -1$. Thus for any $-1 \leq \rho \leq 1$ and for large T we have that

$$f_i = g_i(1-\rho). \tag{3.13}$$

Since the mean of y_{it} for large T is $f_i/(1-\rho)$ if $|\rho| < 1$ we see that the orthogonal fixed effect approaches the long run mean of the process in the stationary case.

The second step in the agenda is to form a prior on $\{g_i\}$, σ^2 and ρ and find the marginal posterior density of ρ. Motivated by the information orthogonality of θ we take these parameters to be independent and we assign independent uniform priors to the g_i. This leads to the following calculation. Let $\theta = (\rho, \sigma^2)$ and, for notational simplicity, let $w_i = y_i - \rho y_{i-}$. Then

$$p(\theta|data) \propto \sigma^{-NT} \Pi_{i=1}^{N} \int_{-\infty}^{\infty} \exp\{-(1/2\sigma^2)\Sigma_i (w_i - f_{iJ})'(w_i - f_{iJ})\} dg_i \pi(\theta), \tag{3.14}$$

$$= \sigma^{-NT} e^{Nb(\rho)} \Pi_{i=1}^{N} \int_{-\infty}^{\infty} \exp\{-(1/2\sigma^2)\Sigma_i (w_i - f_{iJ})'(w_i - f_{iJ})\} df_i \pi(\theta), \tag{3.15}$$

$$= \sigma^{-N(T-1)} e^{Nb(\rho)} \times \exp\{-(1/2\sigma^2)\Sigma_i (y_i - \rho y_{i-})' H (y_i - \rho y_{i-})\} \pi(\theta), \tag{3.16}$$

where H is a $T \times T$ matrix that subtracts the mean from a T vector.

The move from (3.14) to (3.15) is the change of variable from g to f with Jacobian $e^{b(\rho)}$, this change being performed N times. Notice that in order to perform this calculation we do not need to know the orthogonal parameter and so we do not need the complete solution of the differential equation (2.7). All that is required is the Jacobian from g to f. This is significant since in some cases this Jacobian is easier to obtain than is a full solution for the orthogonal parameter.

The form of the joint posterior, (3.16), is interesting. The expression

$$\sigma^{-NT} \exp\{-(1/2\sigma^2)\Sigma_i (y_i - \rho y_{i-})' H (y_i - \rho y_{i-})\} \tag{3.17}$$

is the likelihood *concentrated* with respect to f so that the maximum likelihood estimator of ρ and σ maximizes (3.17). This is the estimator you would get by treating the model as if it were a linear model with strictly exogenous covariates and then measuring data from their agent specific means to eliminate f_i. The expression

$$\sigma^{-N(T-1)} \exp\{-(1/2\sigma^2)\Sigma_i (y_i - \rho y_{i-})' H (y_i - \rho y_{i-})\} \tag{3.18}$$

is the likelihood *integrated uniformly* with respect to f. This is the naive Bayes posterior that assigns a uniform prior to the f_i, a procedure that works (in the sense of producing large N consistent estimates of ρ) in the linear model with exogenous covariates but not here. Finally, (3.16) is the posterior in which the uniform prior is assigned to g and not to f. It can be thought of as adding a correction term $N \log \sigma + Nb(\rho)$ to the concentrated log likelihood. Note that for large T this correction is approximately $N \log \sigma - N \log(1 - \rho)$ for $-1 \leq \rho < 1$.

With $P = \log p(\theta|data)$ the first-order condition for the mode of the ρ, σ^2 joint posterior is

$$\frac{\partial P}{\partial \rho} = Nb'(\rho) + (1/\sigma^2)\Sigma_i (y_i - \rho y_{i-})' H y_{i-} = 0 \tag{3.19}$$

$$\frac{\partial P}{\partial \sigma^2} = -\frac{N(T-1)}{2\sigma^2} + \frac{1}{2\sigma^4}\Sigma_i (y_i - \rho y_{i-})' H (y_i - \rho y_{i-}) = 0. \tag{3.20}$$

We show in the appendix that the expressions (3.19) and (3.20) have expectation zero when evaluated at the true value of ρ, and that when N is large this point locates a maximum of the joint posterior density.

It is the marginal posterior density for ρ that is the natural basis for inference from a Bayesian perspective and, under a conventional prior $\propto 1/\sigma$ for σ, and a uniform prior for ρ

over $\rho_l \leq \rho \leq \rho_u$ this density is

$$p(\rho|data) \propto e^{Nb(\rho)}[\Sigma_i(y_i - \rho y_{i-})' H(y_i - \rho y_{i-})]^{-\frac{N(T-1)}{2}}. \tag{3.21}$$

Under rather unrestrictive conditions the logarithm of this function will converge uniformly in ρ to a function that is maximized at the true value of this parameter. Both the prior for σ and that for ρ have an effect on the shape of the marginal posterior density that is $O(1/N)$ and thus they have negligible effect asymptotically in N.

The first-order condition for a mode of (3.21) is

$$b'(\rho) + (T-1)\frac{\Sigma_i(y_i - \rho y_{i-})' H y_{i-}}{\Sigma_i(y_i - \rho y_{i-})' H(y_i - \rho y_{i-})} = 0. \tag{3.22}$$

3.2. *Including covariates*

We return to the model with covariates whose log likelihood is (3.2). The interest here, apart from the model's practical relevance, is that we unearth (and resolve) a problem with the orthogonalization approach, namely the non-existence of a transformation that will diagonalize the information matrix. We first look for a transformation of the form $f_i = f_i(g_i, \rho, \beta)$ such that g_i is information orthogonal to both ρ and β. Using earlier results the relevant differential equation to determine this function is

$$\begin{pmatrix} \frac{\partial f_i}{\partial \rho} \\ \frac{\partial f_i}{\partial \beta} \\ \frac{\partial f_i}{\partial \sigma^2} \end{pmatrix} = \begin{pmatrix} -\frac{E(J' y_{i-})}{T} \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} -f_i J' c - J' C x_i \beta \\ 0 \\ 0 \end{pmatrix}. \tag{3.23}$$

But unfortunately these equations are inconsistent since if we calculate $\partial^2 f_i / \partial\rho\partial\beta$ from the first of these equations, remembering that both c and C are functions only of ρ, we find that it is $-J' C x_i \beta$ but from the second equation it is zero. This is a contradiction, so the equations have no solution. There is no parametrization in which $f_i = f_i(g_i, \rho, \beta)$ such that g_i is information orthogonal to ρ and β.

Fortunately, for consistent inference such a complete orthogonalization turns out to be unnecessary, since inspection of the first-order conditions for the maximization of the joint log posterior density of $\theta = (\rho, \beta, \sigma^2)$ when g_i is defined, as before, by (3.12) reveal that they provide a large N consistent estimator of this parameter. To see this return to (3.14) and interpret w_i now as $y_i - \rho y_{i-} - x_i\beta$. The joint posterior density then is

$$p(\theta|data) \propto \sigma^{-N(T-1)} \exp\{Nb(\rho) - (1/2\sigma^2)\textstyle\sum_i (y_i - \rho y_{i-} - x_i\beta)' H(y_i - \rho y_{i-} - x_i\beta)\}. \tag{3.24}$$

The corresponding marginal posterior density for ρ under a flat prior for β and $\log\sigma^2$ is

$$p(\rho|data) \propto e^{Nb(\rho)}[\Sigma_i w_i' H w_i - \Sigma_i w_i' H x_i (\Sigma_i x_i' H x_i)^{-1} \Sigma_i x_i' H w_i]^{-[N(T-1)-K]/2} \tag{3.25}$$

for $w_i = y_i - \rho y_{i-}$. The complete Bayesian analysis would proceed by sampling from the univariate density (3.25) and then using the realizations to form and sample from the normal/gamma conditional joint posterior density of $\beta, 1/\sigma^2$. Sampling from (3.25) is most easily carried out by discretizing the density at, say, 1000 points and then sampling each discrete value according to its probability.

The first-order conditions for a maximum of the joint posterior, with $P = \log p$, are

$$\frac{\partial P}{\partial \rho} = Nb'(\rho) + \textstyle\sum_i y_{i-}' H(y_i - \rho y_{i-} - x_i\beta) = 0$$

$$\frac{\partial P}{\partial \beta} = \sum\nolimits_i x_i' H(y_i - \rho y_{i-} - x_i \beta) = 0$$
$$\frac{\partial P}{\partial \sigma^2} = -\frac{N(T-1)}{2} \log \sigma^2 + \frac{1}{2\sigma^4} \sum\nolimits_i (w_i - x_i \beta)' H(w_i - x_i \beta) = 0$$

and these may be shown to have mean zero when evaluated at the true parameter values.

Again, as in the case without covariates, these equations provide consistent inference as $N \to \infty$ for any $T \geq 2$ although a solution providing a maximum may not exist when N and/or T are small.

3.3. *Some illustrative calculations*

In this subsection we report the behavior of the marginal posterior density of ρ, (3.25), using some artificially generated panel data.[9] The marginal posterior density of the parameter of interest is the natural outcome of a Bayesian study. The marginal and joint densities of the other parameters, β and σ^2, conditional on ρ are of standard normal/inverted gamma form so the joint posterior density of all parameters is both well behaved and readily sampled if that of ρ is. The results of this section do not amount to a systematic numerical study and the graphs are intended mainly to give the flavor of what can be found.

3.3.1. Data generation. The data were generated using the model (3.1) with covariates which are $T - 1$ period-specific dummy variables. The errors were iid normal with mean zero and variance one in the first three figures and variance three in the fourth figure. The fixed effects, $\{f_i\}$, were produced as normal variates, independent across agents, with mean and variance both equal to one.

3.3.2. Construction of the figures. Each of the four figures contains nine graphs. Each of the nine graphs within a figure was constructed from data generated using the same values of N, T and σ^2. But each graph within a figure uses different, randomly generated values for the $T - 1$ values of β, the $T \times N$ values of $\{f_i\}$ and y. For each graph in each figure the N values of $\{y_{i0}\}$ are set to zero.

The plots show the marginal posterior density of ρ under a flat prior for $\{g_i\}$ and uniform priors for β, $\log \sigma^2$ and ρ; this is (3.25) after division by its integral (computed numerically) over the set of ρ values indicated by the scale on the horizontal axis. For example, in the graphs of Figure 1 the values for ρ_l and ρ_u were taken to be 0·6 and 1·2 respectively. Choosing a lower limit smaller than 0·6 would only include further regions of zero density, while the upper limit was chosen by a guess as to what might be the largest thinkable value in economic applications. In practice, the investigator can always explore what the density looks like at larger and larger values of ρ. The upper limit on the vertical scale was equal to the largest ordinate of the two plotted densities and so it varies from graph to graph.

This marginal posterior density for ρ is shown as a **solid line**. It is the density that, we argue, provides the basis for large N consistent inference in this model. The value of the parameter ρ used to generate the data was, in all cases, 0·9 and a vertical line is drawn on each graph to indicate this value.

Also shown, as a **dotted curve**, is the marginal posterior density of ρ for a prior which is flat in $\{f_i\}$, β, $\log \sigma^2$ and ρ—it is the marginal of ρ corresponding to (3.18). The mode of this density is a Bayesian version of the "within estimator" that is known to be large N inconsistent.

9. Some results using real (cross-country, macro) data are reported in Lancaster and Aiyar (2000).

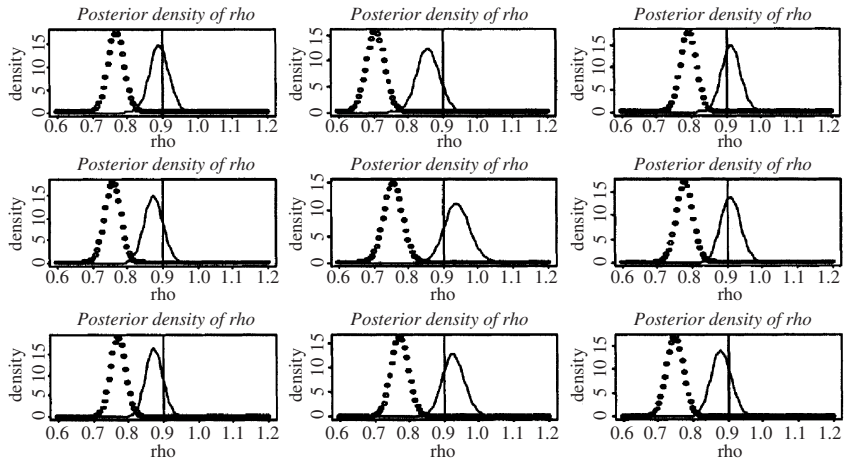

FIGURE 1
$N = 100$; $T = 6$.

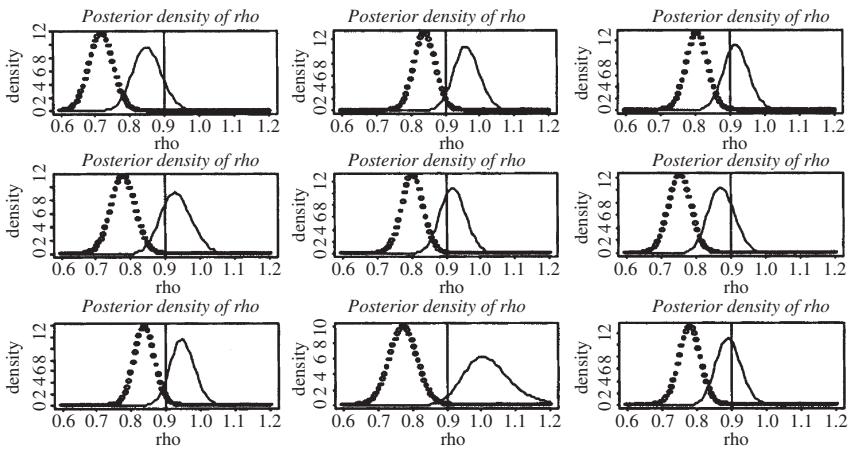

FIGURE 2
$N = 50$; $T = 6$.

3.3.3. The figures. Figure 1 suggests that data from a model with $N = 100$ and $T = 6$ with $\sigma^2 = 1$ and values of f that are rather highly dispersed (with a coefficient of variation of 1) provide posterior densities that are centered, on average, about the value of ρ that provided the data. In contrast, the posterior density under a flat prior on the non-orthogonalized fixed effect is bell shaped and rather tightly centered on a value of ρ that is too low. With this posterior the correct value of ρ is never a plausible value.

In Figure 2 the value of N has been halved to 50 with other parameters remaining the same as in Figure 1. The general effect of this reduction in data is to raise the dispersion of both densities.

In Figure 3 N is now reduced to 40 and T to 3 with other parameters the same. Now, most densities are roughly bell shaped but the ninth graph (bottom right) starts to show the effect of the prior *on the shape*, as well as the location, of the distribution. The term coming from the uniform

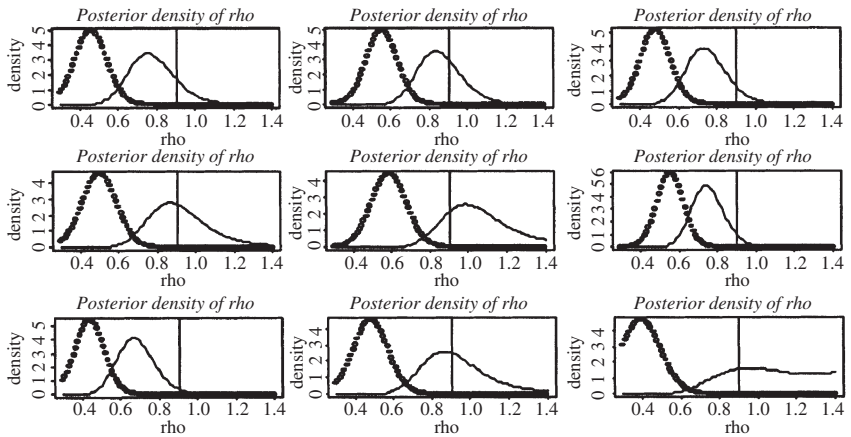

FIGURE 3
$N = 40$; $T = 3$.

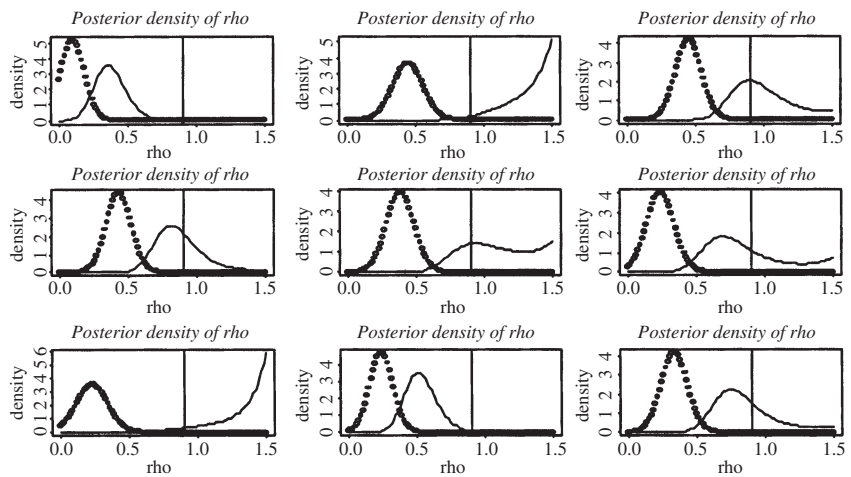

FIGURE 4
$N = 40$; $T = 3$; $\sigma^2 = 3$.

prior on g has the form $e^{Nb(\rho)}$ where, for $T = 3$, $b(\rho)$ is quadratic and it is the effect of this that is shown in the ninth graph. Nonetheless each graph has a clear mode within the "relevant" region (which has now been extended up to 1·4.) However, in the last of these graphs it is clear that the width of highest posterior density regions for ρ will be strongly affected by the choice of ρ_u.

The final figure shows densities from data produced under the least informative of the four data generating processes—here we have tripled the error variance while retaining $N = 40$, $T = 3$. Now the effect of the prior on the shape of the posterior density is even more apparent than in Figure 3. Two of the graphs have no mode within the relevant interval and in the fifth, which has an interior mode, the width of highest posterior density intervals will again depend on the fairly arbitrary choice of ρ_u.

3.4. *Conclusions from Section 3*

We conclude this section with a brief discussion and summary. The marginal posterior distribution of ρ derived by uniform integration of orthogonal fixed effects provides, in its mode, a large N consistent estimator of the autoregressive parameter. This is true even though there exists no complete orthogonalization of the fixed effects, as when covariates enter the model. Judging by the figures given earlier one could presumably prove asymptotic normality of the marginal posterior under suitable further regularity conditions. It is in this sense that we can, as claimed, find likelihood-based consistent estimates for the panel first-order autoregression. The method is likelihood based but not, of course, maximum likelihood.

However, when the total information in the panel is small relative to the number of quantities to be estimated, both real data—Lancaster and Aiyar (2000)—and simulated data (as shown above) can produce marginal posterior densities that are hard to interpret and which do not point conclusively to any one value of ρ.

A number of theoretical questions remain. Foremost of these is how to interpret situations like those in the second and seventh graphs of Figure 4. Is it the case that these data sets carry virtually no information about ρ? Would it then be reasonable to use a more informative prior such as that suggested by Sims (2000)? What would be the consequence of using a prior other than the uniform on the $\{g_i\}$? How would efficient GMM procedures work with such data? Would "lack of information" show up in the standard errors of GMM estimates or would reliance on asymptotic approximations to exact sampling distributions conceal this situation?

A related issue is the behavior of the posterior of ρ under a flat prior on $\{f_i\}$—this is the dotted curve which is systematically well behaved and wrong! Presumably such a prior is far from uninformative. What is the misinformation that this prior conveys?

Another set of issues concerns extensions of the method to more general error distributions, possibly heteroscedastic, and to allow the covariate sequences to be only weakly exogenous. From another point of view, how robust is the method to failure of normality, homoscedasticity, strict exogeneity? Consistency does not depend on the normality assumption nor on that of homoscedasticity across agents but would presumably fail if the covariates were not strictly exogenous.

A final, and harder question is how to extend the separation approach to situations in which there exists no full information orthogonality.

The implication of this work for practitioners is as follows. If you have a (first-order) autoregressive panel model and you are unwilling to formulate a marginal distribution for the initial observations, $\{y_{i0}\}$, so that you plan to work conditionally on their values, then draw the marginal posterior density of ρ, (3.25), in the spirit of "what if". It requires only a few lines of code and will tell you what you would now think, after seeing the data, *if you had believed that the covariates were strictly exogenous and the errors homoscedastic and normal*. If the density you draw has a mode you will be looking at a large N consistent estimate of ρ under these conditions. Such a drawing is clearly informative unless the researcher dogmatically believes that the covariates are not strictly exogenous and the errors definitely not homoscedastic or normal, and the drawing is almost costless.

4. RELATED DISCUSSIONS OF ORTHOGONALITY

The idea of orthogonalizing parameters is not new in the statistical literature. Jeffreys in the second (1948) and subsequent editions of his "Theory of Probability" discussed orthogonalization in his chapter on approximate methods (including maximum likelihood). He argued that information or local orthogonality simplified the computation of ml estimators.

Anscombe (1964) made a similar argument. More recently Cox and Reid (1987) explored orthogonalization from the point of view of approximate frequentist inference.[10] The proposal was to base inference about a parameter of interest on the likelihood conditional on the maximum likelihood estimate of an orthogonalized nuisance parameter **when the common parameter is taken as given**. They also gave a tractable formula for this conditional likelihood using a saddle point approximation to the sampling distribution of the maximum likelihood estimator of the nuisance parameters. In the present context this method would mean orthogonalizing the fixed effects, as we have done; finding the ml estimator of the new fixed effects with θ known; and then working from the distribution of the data given this quantity. The emphasis in Cox and Reid was on testing hypotheses about the common parameter. Liang (1987) took the same approach but emphasized point estimation and the construction of orthogonality conditions[11] based on the likelihood conditioned on ml estimates of fixed effects.[12] If the orthogonalized fixed effect has a maximum likelihood estimate which is free of θ this leads to standard conditional likelihood panel data procedures. For example, in panel Poisson counts it leads to the multinomial conditional likelihood procedure which, as I showed in Section 2, is identical to the standard ml estimator. In panel logit models, where the orthogonalized fixed effect is $g_i = \Sigma_t Pr(Y_{it} = 1|x_i, f_i, \beta)$ with ml estimator $\Sigma_t Y_{it}$ it leads to the usual conditional logit estimator. In Example 3, the stationary autoregressive panel model, the orthogonal fixed effect is 2.9 and conditioning on its maximum likelihood estimator leads to the likelihood for the first differenced data.[13] Thus both Bayesian and approximate conditional frequentist inference lead to the same criterion functions in these models.

Interestingly enough the Bayes procedure for the non-stationary model of Section 3 also has an approximate conditional likelihood interpretation. The orthogonal fixed effect is $g_i = f_i e^{-b(\rho)}$ and when ρ, σ^2 are given the maximum likelihood estimator of g_i is $(\overline{y}_i - \rho\overline{y}_{i-})e^{-b(\rho)}$. This is normally distributed with mean g_i and standard deviation $e^{b(\rho)}\sigma/\sqrt{T}$. Dividing the product of N such terms into the likelihood again leads to the Bayes marginal posterior under a uniform prior for $\{g_i\}$.

5. CONCLUSIONS

The inconsistency of both maximum likelihood and Bayes estimators in models with infinitely many incidental parameters can be avoided in some models of econometric interest by redefining the fixed effects to be information orthogonal to the parameter of interest. I have shown how this can be done in two dynamic regression models. The method is to integrate out orthogonalized fixed effects with respect to an uninformative prior distribution and to base inferences on the resulting marginal posterior distribution for the common parameters. I have also shown that the same criterion functions can be deduced from Cox and Reid's approximate conditional likelihood approach to inference.[14] This approach provides, depending on ones point of view, likelihood

10. Sweeting (1987) gave a Bayesian interpretation of Cox and Reid's approach.

11. Such GMM procedures, called generalized estimating equations, have been used in the biostatistics literature since Godambe (1960).

12. An enlightening recent discussion of frequentist conditional inference with comments on its connection with Bayesian procedures is to be found in two papers by Reid (1995) and Liang (1995).

13. Cruddas *et al*. (1989) studied the stationary model of Example 3 but without covariates. They noted that $\{f_i\}$ were orthogonal to ρ and concentrated on finding a transformation of σ^2 that would orthogonalize it to ρ. They did not remark on the result (2.11).

14. This parameter (information) orthogonalization is also possible in panel Weibull duration data where again the marginal posterior distribution provides consistent inferences for common parameters. Orthogonalization is also readily achieved in all panel binary response models in which the (non-orthogonalized) fixed effect appears as the intercept in a latent linear regression model. Lancaster (2000*b*) applies orthogonalization methods to autoregressive duration data with fixed effects.

based, and in particular Bayesian, methods of inference in dynamic regression models, or new consistent GMM procedures.[15]

The point of orthogonalizing is to separate the problem of inference about the common parameters from that of inference about the fixed effects, about which consistent inferences cannot be made. Information orthogonality achieves an approximate separation and this is enough to permit consistent inference in the dynamic regression models studied here. In other models information orthogonality of the fixed effects does not lead to consistent marginal posterior inference. We should also note that there may be no orthogonal reparametrization, as for example, appears to be true for autoregressive models of order greater than one.

I have argued in this paper for a return to likelihood-based analyses of fixed effects panels. In so far as the reason for their abandonment by the profession has been the perceived failure of maximum likelihood—the incidental parameter problem—I would argue that reason no longer exists.[16] To be sure, maximum likelihood is inappropriate, but maximizing is not what should be done.

APPENDIX

In this appendix we shall prove consistency of the joint posterior mode by showing that the conditions of the consistency theorem for extremum estimators, Theorem 4.1.2. of Amemiya (1985) is satisfied. First we recall the model; then we state the assumptions and theorem adapted to the present problem; finally, we show that the verifiable assumptions hold.

A1. *The model*

The model is

$$y_i = f_i \jmath + \rho y_{i-} + x_i\beta + u_i, \qquad u_i \,|\, f_i, x_i, \beta, \sigma \sim n(0, \sigma^2 I_T) \tag{A1}$$

$$y_{i-} = f_i c + Cx_i\beta + Cu_i, \tag{A2}$$

$$c = \begin{pmatrix} 0 \\ 1 \\ 1+\rho \\ \cdot \\ 1+\rho+\rho^2+\cdots+\rho^{T-2} \end{pmatrix}; \qquad C = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ \rho & 1 & 0 & \dots & 0 \\ \cdot & \cdot & \cdot & \dots & \cdot \\ \rho^{T-2} & \rho^{T-3} & \dots & 1 & 0 \end{pmatrix}$$

for $T \geq 3$ and $N \geq 1$. The $T \times T$ matrix $H = I_T - (1/T)\jmath\jmath'$ which subtracts the mean from a T vector. Observations are stochastically independent over i given the values of $\{f_i\}, \{x_i\}, \beta, \sigma^2$. The vector β is of length K. The function $b(\cdot)$ is

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

The matrix C and the function $b(\cdot)$ satisfy

$$trHC = trC'H = -b'(\rho);$$

since $trHC = trC - (1/T)\jmath'C\jmath = 0 - b'(\rho)$.

We focus on the mode of the joint posterior density, the arguments for consistency of modes of the marginal posterior densities are similar. The logarithm of the joint posterior density divided by N is

$$Q_N(r, b, s^2) = -\left(\frac{1}{2s^2N}\right)\sum_i (y_i - ry_{i-} - x_ib)'H(y_i - ry_{i-} - x_ib) - \frac{T-1}{2}\log s^2 + b(r),$$

15. It is worth noting that consistency in the non-stationary model does not depend on normality of the errors.

16. Lancaster (2000*a*), gives a brief history of the incidental parameter problem through to recent developments in orthogonalization.

and using the expression (A1) for y_i we have

$$\begin{aligned} Q_N(r,b,s^2) = &-(1/2s^2N)[(\rho-r)^2\Sigma_i y'_{i-}Hy_{i-} + (\beta-b)'\Sigma_i x'_i Hx_i(\beta-b) \\ &+ \Sigma_i u'_i Hu_i + 2(\rho-r)\Sigma_i y'_{i-}Hu_i + 2(\rho-r)\Sigma_i y'_{i-}Hx_i(\beta-b) \\ &+ 2\Sigma_i u'_i Hx_i(\beta-b)] - (T-1)/2)\log s^2 + b(r)\cdot \end{aligned}$$

Now using the definition, A2, of y_{i-}, and writing

$$w_i = f_i c_2, \qquad z_i = Cx_i, \qquad v_i = w_i + z_i\beta,$$

we have

$$\begin{aligned} (1/N)E(\Sigma_i y'_{i-}Hy_{i-}) &= (1/N)\Sigma_i(w'_i Hw_i + \beta' z'_i Hz_i\beta + \sigma^2 N tr C'HC + 2w'_i Hz_i\beta) \\ &= (1/N)\Sigma_i v'_i Hv_i + \sigma^2 tr C'HC \\ (1/N)E(\Sigma_i y'_{i-}Hu_i) &= \sigma^2 tr C'H \\ (1/N)E(\Sigma_i y'_{i-}Hx_i) &= (1/N)\Sigma_i(w'_i Hx_i + \beta' z'_i Hx_i) = (1/N)\Sigma_i v'_i Hx_i \end{aligned}$$

A2. *Assumptions and theorem*

One way of doing the asymptotics is to consider a sequence $\{f_i\}, \{x_i\}$ consisting of independent and identically distributed realizations of a $K+1$ dimensional random variable with finite first two moments. We then let various sample moments converge in probability to their expectations under this distribution.

Let

(1) $\{u_i\}$ be iid realizations of $n(0, \sigma^2 I_T)$ given $x_i, f_i, \beta, \rho, \sigma^2$.
(2) $\{f_i, x_i\}$be iid realizations from a distribution with finite second moments.
(3) $\text{plim}_{N\to\infty}(1/N)\Sigma_i v'_i Hv_i, = E(v'Hv) > 0$,
(4) $\text{plim}_{N\to\infty}(1/N)\Sigma_i v'_i Hx_i = E(v'Hx)$,
(5) $\lim_{N\to\infty}(1/N)\Sigma_i x'_i Hx_i = E(x'Hx)$, positive definite.

For notational simplicity we shall denote the expectations of $x'Hx, v'Hx, v'Hv$ by these expressions without the expectation operator. Note that there is no requirement that f_i and x_i are independent.

Further, let $\theta = (r, b, s^2) \in \Theta$ be an open subset of Euclidean $K+2$ space with $\theta_0 = (\rho, \beta, \sigma^2)$.

Theorem A1. *Let Θ_N be the set of roots of the equation*

$$\frac{\partial Q_N}{\partial \theta} = 0$$

corresponding to the local maxima. If that set is empty set Θ_N equal to zero. Then, for any $\epsilon > 0$,

$$\lim_{N\to\infty} P[\inf_{\theta\in\Theta_N}(\theta-\theta_0)'(\theta-\theta_0) > \epsilon] = 0\cdot$$

Proof. Since Q_N is polynomial in (r, b) and differentiable in s^2 conditions 1, 2 and 3 are sufficient to show that Q_N converges in probability uniformly in θ in any open neighbourhood of θ_0 to the function $Q(\theta)$ which is

$$\begin{aligned} Q(\theta) = &-(1/2s^2)\{(\rho-r)^2[v'Hv + \sigma^2 tr C'HC] \\ &+ (\beta-b)'x'Hx(\beta-b) + \sigma^2(T-1) + 2(\rho-r)\sigma^2 tr C'H \\ &+ 2(\rho-r)v'Hx(\beta-b)\} \\ &- \frac{T-1}{2}\log s^2 + b(r)\cdot \end{aligned}$$

It suffices to show that $Q(\theta)$ has a strict local maximum at θ_0. To show this consider $\partial Q/\partial\theta$:

$$\begin{aligned} \frac{\partial Q}{\partial r} &= (1/s^2)[(\rho-r)\{v'Hv + \sigma^2 tr C'HC\} \\ &\quad + \sigma^2 tr C'H + v'Hx(\beta-b)] + b'(r), \\ \frac{\partial Q}{\partial b} &= (1/s^2)[x'Hx(\beta-b) + (\rho-r)v'Hx] \end{aligned}$$

$$\frac{\partial Q}{\partial s^2} = (1/2s^4)\{(\rho - r)^2[v'Hv + \sigma^2 trC'HC]$$
$$(\beta - b)'x'Hx(\beta - b) + \sigma^2(T-1) + 2(\rho - r)\sigma^2 trC'HC$$
$$2(\rho - r)v'Hx(\beta - b)\} - (T-1)/2s^2.$$

In view of $trC'H = -b'(\rho)$, the equation $\partial Q/\partial\theta = 0$ has a solution at θ_0. To determine that this solution provides a maximum consider the second derivatives at $\theta_0 = (\rho, \beta, \sigma^2)$. These are

$$\frac{\partial^2 Q}{\partial r\,\partial r} = -\frac{v'Hv}{\sigma^2} - trC'HC + b''(\rho),$$
$$\frac{\partial^2 Q}{\partial r\,\partial b} = \frac{-v'Hx}{\sigma^2},$$
$$\frac{\partial^2 Q}{\partial r\,\partial s^2} = -\frac{trC'H}{\sigma^2} = \frac{b'(\rho)}{\sigma^2},$$
$$\frac{\partial^2 Q}{\partial b\,\partial b'} = -\frac{x'Hx}{\sigma^2},$$
$$\frac{\partial^2 Q}{\partial b\,\partial s^2} = 0,$$
$$\frac{\partial^2 Q}{\partial s^2\,\partial s^2} = -\frac{T-1}{2\sigma^4}.$$

To show that θ_0 locates a maximum we must show that $-H$ is positive definite. The principal minors are

(1) $trC'HC - b''(\rho) + v'Hv/\sigma^2$. Since $v'Hv$ is positive, this minor is certainly positive if $g_1(\rho) = trC'HC - b''(\rho)$ is. This is a polynomial of degree $T-1$ in ρ. In the case $T = 3$ for example, it takes the form

$$g_1(\rho) = \frac{3 - 2\rho + 2\rho^2}{3}$$

which is positive everwhere on the real line. Direct calculation for larger values of T establishes $g(\rho)$ is in fact always positive.

(2) $\frac{T-1}{2\sigma^4}[trC'HC - b''(\rho) + \frac{v'Hv}{\sigma^2}] - \frac{(b'(\rho))^2}{\sigma^4}$. Since $v'Hv$ is positive this minor will certainly be positive if $g_2(\rho) = \frac{T-1}{2}[trC'HC - b''(\rho)] - (b'(\rho))^2$ is positive. Direct calculation shows $g_2(\rho)$ is positive except when $\rho = 1$ when it equals zero. In this case the condition $v'Hv > 0$ establishes positivity.

(3) The final minor is the sum of two terms. The first is equal to $x'Hx/\sigma^2$ times $g_2(\rho)$ and is thus positive. The second is $(T-1)/2$ times $v'Hv \cdot x'Hx - (v'Hx)^2$ which is non-negative by the Cauchy–Schwarz inequality. (This presumes $K = 1$, a similar argument applies when $K > 1$·)

Thus, the Hessian of $Q(\theta)$ is negative definite at θ_0 which point provides a strict local maximum. ‖

Essentially the same argument can be used to show the consistency of the marginal posterior modes.

Acknowledgements. Bernard Lindenhovius participated in the initial work on model Example 3 of Section 2. I would also like to acknowledge helpful comments on earlier versions of this paper by Shekhar Aiyar, Moshe Buchinsky, Gary Chamberlain, Xiaohong Chen, Jim Feyrer, Jerry Hausman, Karsten Hansen, Hide Ichimura, Robin Lumsdaine, Jim Powell, Chris Sims, Tom Wansbeek, Tiemen Woutersen, Arnold Zellner and seminar participants at Harvard/MIT, Berkeley, Stanford, UCSD, USC/UCLA, Bristol and University College London. I would also like to acknowledge the stimulating, though foreshortened, hospitality of the Department of Economics at UCSD in the spring of 1997, where the first draft of this paper was written.

REFERENCES

AHN, S. and SCHMIDT, P. (1995), "Efficient Estimation of Models for Dynamic Panel Data", *Journal of Econometrics*, **68**, 5–27.
AMEMIYA, T. (1985), *Advanced Econometrics* (Cambridge, MA: Harvard University Press).
ANDERSEN, E. B. (1970), "Asymptotic Properties of Conditional Maximum Likelihood Estimators", *Journal of the Royal Statistical Society, B*, **32**, 283–301.
ANSCOMBE, F. J. (1964), "Normal Likelihoods", *Annals of the Institute of Statistical Mathematics*, **26**, 1–19.

ARELLANO, M. and BOND, S. (1991), "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations", *Review of Economic Studies*, **58**, 227–297.
ARELLANO, M. and BOVER, O. (1995), "Another Look at the Instrumental Variable Estimation of Error Components Models", *Journal of Econometrics*, **68**, 29–51.
BLUNDELL, R., GRIFFITH, R. and WINDMEIJER, F. (1997), *Individual Effects and Dynamics in Count Data Models*, typescript (London: University College).
COX, D. R. and REID, N. (1987), "Parameter Orthogonality and Approximate Conditional Inference (with Discussion)", *Journal of the Royal Statistical Society, B*, **49** (1), 1–39.
CRUDDAS, A. M., COX, D. R. and REID, N. (1989), "A Time Series Illustration of Approximate Conditional Likelihood", *Biometrika*, **76** (2), 231–237.
DIACONIS, P. and FREEDMAN, D. (1986), "On the Consistency of Bayes Estimates", *Annals of Statistics*, **14**, 1–67.
GODAMBE, V. P. (1960), "An Optimum Property of Regular Maximum Likelihood Estimation", *Annals of Mathematical Statistics*, **31**, 1208–1211.
HANSEN, K. (2000), Linear Panel Models: When Should Bayesians Difference? (Mimeo, Department of Economics, Brown University).
HAUSMAN, J., HALL, B. H. and GRILICHES, Z. (1984), "Econometric Models for Count Data with an Application to the Patents – R & D Relationship", *Econometrica*, **52** (4), 909–938.
JEFFREYS, H. (1960), *Theory of Probability*, 3rd edition (Oxford University Press).
KIEFER, N. M. (1980), "Estimation of Fixed Effects Models for Time Series of Cross-sections with Arbitrary Intertemporal Covariance", *Journal of Econometrics*, **14**, 195–202.
LANCASTER, T. (2000*a*), "The Incidental Parameter Problem Since 1948", *Journal of Econometrics*, **95**, 391–413.
LANCASTER, T. (2000*b*), "Some Econometrics of Scarring", in C. Hsiao, K. Morimune and J. L. Powell (eds.) *Nonlinear Statistical Modelling* (Cambridge: Cambridge University Press).
LANCASTER, T. and AIYAR, S. (2000), "Econometric Analysis of Dynamic Models: A Growth Theory Example", in H. Bunzel, P. Jensen, N. M. Kiefer and D. T. Mortensen (eds.) *Panel Data and Structural Labor Market Models* (Amsterdam: North Holland).
LIANG, K.-Y. (1987), "Estimating Functions and Approximate Conditional Likelihood", *Biometrika*, **74** (4), 695–702.
LIANG, K.-Y (1995), "Inference Based on Estimating Functions in the Presence of Nuisance Parameters", *Statistical Science*, **10** (2), 158–172.
NEYMAN, J. and SCOTT, E. L. (1948), "Consistent Estimation from Partially Consistent Observations", *Econometrica*, **16**, 1–32.
REID, N. (1995), "The Roles of Conditioning in Inference", *Statistical Science*, **10** (2), 138–157.
SIMS, C. A. (2000), "Using a Likelihood Perspective to Sharpen Econometric Discourse: Three Examples", *Journal of Econometrics*, **95** (2), 443–462.
SWEETING, T. (1987), Discussion of Cox and Reid (1987).
WASSERMAN, L. (1998), "Asymptotic Properties of Bayesian Nonparametric Procedures", in Dipak Dey and Debajyoti Sinha (eds.) *Practical nonparametric and Semiparametric Bayesian Statistics*, Chapter 16 (Springer).