DECISION THEORY APPLIED TO A LINEAR PANEL DATA MODEL

Gary Chamberlain and Marcelo J. Moreira[1]

ABSTRACT

This paper applies some general concepts in decision theory to a linear panel data model. A simple version of the model is an autoregression with a separate intercept for each unit in the cross section, with errors that are independent and identically distributed with a normal distribution. There is a parameter of interest γ and a nuisance parameter τ, a $N \times K$ matrix, where N is the cross-section sample size. The focus is on dealing with the incidental parameters problem created by a potentially high-dimension nuisance parameter. We adopt a "fixed-effects" approach, that seeks to protect against any sequence of incidental parameters. We transform τ to (δ, ρ, ω), where δ is a $J \times K$ matrix of coefficients from the least squares projection of τ on a $N \times J$ matrix x of strictly exogenous variables, ρ is a $K \times K$ symmetric, positive semidefinite matrix obtained from the residual sums of squares and cross products in the projection of τ on x, and ω is a $(N-J)\times K$ matrix whose columns are orthogonal and have unit length. The model is invariant under the actions of a group on the sample space and the parameter space, and we find a maximal invariant statistic. The distribution of the maximal invariant statistic does not depend upon ω. There is a unique invariant distribution for ω. We use this invariant distribution as a prior distribution to obtain an integrated likelihood function. It depends upon the observation only through the maximal invariant statistic. We use the maximal invariant statistic to construct a marginal likelihood function. So we can eliminate ω by integration with respect to the invariant prior distribution, or by working with the marginal likelihood function. The two approaches coincide.

Decision rules based on the invariant distribution for ω have a minimax property. Given a loss function that does not depend upon ω, and given a prior distribution for (γ, δ, ρ), we show how to minimize the average—with respect to the prior distribution for (γ, δ, ρ)—of the maximum risk, where the maximum is with respect to ω.

There is a family of prior distributions for (δ, ρ) that leads to a simple closed form for the integrated likelihood function. This integrated likelihood function coincides with the likelihood function for a normal, correlated random effects model. Under random sampling, the corresponding quasi maximum likelihood estimator is consistent for γ as $N \to \infty$, with a standard limiting distribution. The limit results do not require normality or homoskedasticity (conditional on x) assumptions.

KEYWORDS: Autoregression, fixed effects, incidental parameters, invariance, minimax, correlated random effects

[1] We thank two referees and a co-editor for helpful comments. Financial support was provided by the National Science Foundation (SES-0819761).

June 2006
revised 17 July 2008

DECISION THEORY APPLIED TO A LINEAR PANEL DATA MODEL

1. INTRODUCTION

This paper applies some general concepts in decision theory to a linear panel data model. An example of the model is an autoregression with a separate intercept for each unit in the cross section, with errors that are independent and identically distributed with a normal distribution. There is a parameter of interest γ and a nuisance parameter τ, a $N \times K$ matrix, where N is the cross-section sample size. The focus is on dealing with the incidental parameters problem created by a potentially high-dimension nuisance parameter.

In our general model, the observation is the realized value of a $N \times M$ matrix Y of random variables. We shall be conditioning on the value of a $N \times J$ matrix x, which is observed and has rank J. Our model specifies a conditional distribution for Y given x, as a function of the parameter of interest γ and the nuisance parameter τ:

$$Y \mid x \overset{d}{=} xa(\gamma) + \tau b(\gamma) + Wc(\gamma), \tag{1}$$

where τ is $N \times K$, W is $N \times p$, and $J + K \leq N$, $J + M \leq N$, $M \leq p$. The components of W, conditional on x, are independently and identically distributed $\mathcal{N}(0, 1)$, which we shall denote by

$$\mathcal{L}(W) = \mathcal{N}(0, I_N \otimes I_p).$$

The functions a, b, and c are given. (For a random matrix V, the notation $\mathcal{L}(V) = \mathcal{N}(\mu, \Lambda)$ indicates that the vector formed by joining the rows of V has a multivariate normal distribution with covariance matrix Λ and mean vector formed by joining the rows of the matrix μ.) All distributions throughout the paper are conditional on x.

A simple version of our model arises from the reduced form of the following autoregression:

$$Y_{it} = \psi Y_{i,t-1} + \alpha_i + U_{it} \qquad (i = 1, \dots, N;\ t = 1, \dots, \bar{T}),$$

where the U_{it} are independent and identically distributed $\mathcal{N}(0, \sigma^2)$. We observe the realized value of the random variable Y_{it} for $i = 1, \dots, N$ and $t = 1, \dots, \bar{T}$. We do not observe Y_{i0}. The reduced form is

$$Y_{i1} = \psi Y_{i0} + \alpha_i + U_{i1},$$
$$Y_{it} = \psi^t Y_{i0} + (1 + \psi + \cdots + \psi^{t-1})\alpha_i + U_{it} + \psi U_{i,t-1} + \cdots + \psi^{t-1} U_{i1} \qquad (t = 2, \dots, \bar{T}).$$

Conditional on $Y_{i0} = y_{i0}$, we can write this as

$$Y = \tau b(\gamma) + W c(\gamma), \tag{2}$$

where $\gamma = (\psi, \sigma)$,

$$Y = \begin{pmatrix} Y_{11} & \dots & Y_{1\bar{T}} \\ \vdots & & \vdots \\ Y_{N1} & \dots & Y_{N\bar{T}} \end{pmatrix}, \quad \tau = \begin{pmatrix} y_{10} & \alpha_1 \\ \vdots & \vdots \\ y_{N0} & \alpha_N \end{pmatrix}, \quad W = \begin{pmatrix} W_{11} & \dots & W_{1\bar{T}} \\ \vdots & & \vdots \\ W_{N1} & \dots & W_{N\bar{T}} \end{pmatrix}, \tag{3}$$

$$b(\gamma) = \begin{pmatrix} \psi & \psi^2 & \dots & \psi^{\bar{T}} \\ 1 & (1+\psi) & \dots & (1 + \psi + \cdots + \psi^{\bar{T}-1}) \end{pmatrix}, \quad c(\gamma) = \sigma \begin{pmatrix} 1 & \psi & \dots & \psi^{\bar{T}-1} \\ 0 & 1 & \dots & \psi^{\bar{T}-2} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}, \tag{4}$$

and the W_{it} are independent and identically distributed $\mathcal{N}(0, 1)$.

The observation is the realized value of Y. The parameters are γ and τ. We shall focus on inference for γ, and treat the initial conditions and individual effects in τ as nuisance parameters. We shall try to deal with the large number of incidental parameters in τ that arises when N is large. We shall adopt a "fixed-effects" approach, that seeks to protect against any sequence of incidental parameters in τ. There are recent discussions of incidental parameters and panel data in Lancaster (2000, 2002) and Arellano (2003).

An alternative analysis could be based on the distribution of $(Y_{i2}, \dots, Y_{i\bar{T}})$ conditional on the observed value $Y_{i1} = y_{i1}$. This can fit into our framework by removing the first column of Y and including y_{i1} in the i^{th} row of x. We prefer to work with the full distribution of the observed Y in order to avoid possible loss of information from conditioning.

Now consider a second-order autoregression with time-varying coefficients on the individual effect (a factor model), and time-varying variances for the innovations:

$$Y_{it} = \psi_1 Y_{i,t-1} + \psi_2 Y_{i,t-2} + \alpha_i \zeta_t + U_{it} \qquad (t = 1, \ldots, \bar{T}),$$

where $Y_{i0} = y_{i0}$ and $Y_{i,-1} = y_{i,-1}$ are not observed, and the U_{it} are mutually independent with $U_{it} \sim \mathcal{N}(0, \sigma_t^2)$. With Y and W defined as above, we can write this as

$$Y d(\psi) = \tau \tilde{b}(\psi, \zeta) + W \tilde{c}(\sigma),$$

where

$$\tau = \begin{pmatrix} y_{10} & y_{1,-1} & \alpha_1 \\ \vdots & \vdots & \vdots \\ y_{N0} & y_{N,-1} & \alpha_N \end{pmatrix}, \quad d(\psi) = \begin{pmatrix} 1 & -\psi_1 & -\psi_2 & \ldots & 0 \\ 0 & 1 & -\psi_1 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \ldots & -\psi_2 \\ 0 & 0 & 0 & \ldots & -\psi_1 \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix},$$

$$\tilde{b}(\psi, \zeta) = \begin{pmatrix} \psi_1 & \psi_2 & 0 & \ldots & 0 \\ \psi_2 & 0 & 0 & \ldots & 0 \\ \zeta_1 & \zeta_2 & \zeta_3 & \ldots & \zeta_{\bar{T}} \end{pmatrix}, \quad \tilde{c}(\sigma) = \mathrm{diag}(\sigma_1, \ldots, \sigma_{\bar{T}}).$$

We can impose a normalization such as $\sum_{t=1}^{\bar{T}} \zeta_t^2 = 1$. The reduced form of the model is

$$Y = \tau b(\gamma) + W c(\gamma),$$

with $\gamma = (\psi, \zeta, \sigma)$ and

$$b(\gamma) = \tilde{b}(\psi, \zeta) d(\psi)^{-1}, \quad c(\gamma) = \tilde{c}(\sigma) d(\psi)^{-1}.$$

We can include strictly exogenous variables x_{it}:

$$Y_{it} = x_{it}'\xi + \psi_1 Y_{i,t-1} + \psi_2 Y_{i,t-2} + \alpha_i \zeta_t + U_{it},$$

where x_{it} and ξ are $L \times 1$ matrices,

$$x = \begin{pmatrix} x_{11}' & \ldots & x_{1\bar{T}}' \\ \vdots & & \vdots \\ x_{N1}' & \ldots & x_{N\bar{T}}' \end{pmatrix},$$

and conditional on x, the U_{it} are mutually independent with $U_{it} \sim \mathcal{N}(0, \sigma_t^2)$. The reduced form of this model is

$$Y = xa(\gamma) + \tau b(\gamma) + Wc(\gamma),$$

with $\gamma = (\xi, \psi, \zeta, \sigma)$, $\tilde{a}(\xi) = I_{\bar{T}} \otimes \xi$, and

$$a(\gamma) = \tilde{a}(\xi)d(\psi)^{-1}, \quad b(\gamma) = \tilde{b}(\psi, \zeta)d(\psi)^{-1}, \quad c(\gamma) = \tilde{c}(\sigma)d(\psi)^{-1}.$$

Note that if ψ_1 or ψ_2 is not equal to zero, then the reduced form has a distributed lag: the conditional expectation of Y_{it} given x depends upon $x_{i1}, \ldots, x_{it}$. An alternative model has

$$Y_{it} = x_{it}'\xi + \alpha_i \zeta_t + U_{it},$$

where, conditional on x, the vector $(U_{i1}, \ldots, U_{i\bar{T}})$ is independent and identically distributed with a multivariate normal distribution:

$$(U_{i1}, \ldots, U_{i\bar{T}}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Lambda(\chi)).$$

The function Λ is given and specifies the variances and serial correlations of the errors U_{it} as a function of the parameter vector χ with fixed dimension. We can write this as

$$Y = x\tilde{a}(\xi) + \tau\tilde{b}(\psi, \zeta) + W\tilde{c}(\chi),$$

where $\tilde{c}(\chi)$ is the symmetric square root of $\Lambda(\chi)$: $\tilde{c}(\chi)^2 = \Lambda(\chi)$.

In our general model, the observation is the realized value of a $N \times M$ matrix Y of random variables. For example, in a vector autoregression involving the variables $Y^{(1)}, \ldots, Y^{(k)}$, the i^{th} row of Y could be

$$(Y_{i1}^{(1)}, \ldots, Y_{i1}^{(k)}, \ldots, Y_{i\bar{T}}^{(1)}, \ldots, Y_{i\bar{T}}^{(k)}),$$

so that $M = k\bar{T}$. We only consider linear, complete systems whose reduced forms match equation (1). See Arellano's (2003, p. 144) discussion of incomplete systems with unspecified feedback processes.

The next section shows that the model is invariant under the actions of a group. This group is isomorphic to $O(N-J)$, the group of orthogonal matrices with $N-J$ rows and columns. This isomorphism suggests a canonical form for the model, based on a one-to-one transformation, which simplifies the subsequent analysis. We transform τ to (δ, ρ, ω), where δ is a $J \times K$ matrix of coefficients from the least squares projection of τ on x, ρ is a $K \times K$ symmetric, positive semidefinite matrix obtained from the residual sums of squares and cross products in the projection of τ on x, and ω is a $(N-J) \times K$ matrix whose columns are orthogonal and have unit length. Only ω has a dimension that increases with N. Section 3 finds a maximal invariant statistic. The distribution of the maximal invariant statistic does not depend upon ω. Section 4 obtains the unique, invariant distribution for ω. We use this invariant distribution as a prior distribution to obtain an integrated likelihood function. It depends upon the observation only through the maximal invariant statistic. We use the maximal invariant statistic to construct a marginal likelihood function. So we can eliminate ω by integration with respect to the invariant prior distribution, or by working with the marginal likelihood function. The two approaches coincide.

Section 5 shows that decision rules based on the invariant distribution for ω have a minimax property. Given a loss function that does not depend upon ω, and given a prior distribution for (γ, δ, ρ), we show how to minimize the average—with respect to the prior distribution for (γ, δ, ρ)—of the maximum risk, where the maximum is with respect to ω.

Section 6 shows that there is a family of prior distributions for (δ, ρ) that leads to a simple closed form for the integrated likelihood function. This integrated likelihood function coincides with the likelihood function for a normal, correlated random effects model.

Section 7 develops the example of a simple autoregression and relates our results to the literature. Under random sampling, the quasi maximum likelihood estimator for the correlated random effects model is consistent for γ as $N \to \infty$, with a standard limiting distribution. The limit results do not require normality or homoskedasticity (conditional on x) assumptions.

2. MODEL INVARIANCE AND CANONICAL FORM

This section shows that the model is invariant under the actions of a group. This group invariance implies a maximal invariant statistic and an invariant prior distribution. Their derivation is simplified by working with a canonical form for the model, based on a one-to-one transformation. We shall briefly describe invariance in the original form of the model, and then provide more detail in the canonical form, where most of the subsequent analysis takes place.

Let $O(N)$ denote the group of $N \times N$ orthogonal matrices ($gg' = g'g = I_N$). The group $\tilde{G}$ is the subgroup of $O(N)$ that preserves x:

$$\tilde{G} = \{\tilde{g} \in O(N) : \tilde{g}x = x\}.$$

The action of $\tilde{G}$ on the sample space maps y to $\tilde{g}y$. Note that

$$\tilde{g}Y \mid x \stackrel{d}{=} xa(\gamma) + (\tilde{g}\tau)b(\gamma) + Wc(\gamma)$$

(because $\mathcal{L}(\tilde{g}W) = \mathcal{L}(W)$). The action of $\tilde{G}$ on the parameter space maps (γ, τ) to $(\gamma, \tilde{g}\tau)$. So the model is invariant under the actions of $\tilde{G}$ on the sample space and the parameter space.

The canonical form follows from recognizing that the group $\tilde{G}$ is in fact isomorphic to $O(N-J)$. To see this, use the polar decomposition of x to obtain

$$x = q \begin{pmatrix} s \\ 0 \end{pmatrix},$$

where $q \in O(N)$ and s is the unique symmetric, positive semidefinite square root of $x'x$: $s = (x'x)^{1/2}$, with $ss = x'x$. (See Golub and Van Loan (1996, p. 149).) The $J \times J$ matrix s is positive definite because x has full column rank J. Then $\tilde{g}x = x$ is equivalent to

$$(q'\tilde{g}q) \begin{pmatrix} s \\ 0 \end{pmatrix} = \begin{pmatrix} s \\ 0 \end{pmatrix}.$$

Let

$$q'\tilde{g}q = \begin{pmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{pmatrix}.$$

Then we have

$$h_{11}s = s \quad \Rightarrow \quad h_{11} = I_J,$$

$$h_{21}s = 0 \quad \Rightarrow \quad h_{21} = 0,$$

and so $q'\tilde{g}q \in O(N)$ implies that

$$q'\tilde{g}q = \begin{pmatrix} I_J & 0 \\ 0 & g \end{pmatrix}, \tag{5}$$

where $g \in O(N-J)$. Define $G = O(N-J)$ and define the map $\iota\colon \tilde{G} \to G$ by using (5) to map $\tilde{g}$ to g. It is straightforward to check that this map is bijective. In addition,

$$\begin{aligned} q'(\tilde{g}_1\tilde{g}_2)q &= (q'\tilde{g}_1q)(q'\tilde{g}_2q) \\ &= \begin{pmatrix} I_J & 0 \\ 0 & g_1 \end{pmatrix}\begin{pmatrix} I_J & 0 \\ 0 & g_2 \end{pmatrix} = \begin{pmatrix} I_J & 0 \\ 0 & g_1g_2 \end{pmatrix}, \end{aligned}$$

so that ι is a group homomorphism:

$$\iota(\tilde{g}_1\tilde{g}_2) = g_1g_2 = \iota(\tilde{g}_1)\iota(\tilde{g}_2).$$

Since ι is a bijective homomorphism, the groups $\tilde{G}$ and G are isomorphic.

Because of this isomorphism, the action of $\tilde{G}$ on the sample space implies an action of G on the sample space:

$$\tilde{g}y = q(q'\tilde{g}q)q'y = q\begin{pmatrix} I_J & 0 \\ 0 & g \end{pmatrix}\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = q\begin{pmatrix} z_1 \\ gz_2 \end{pmatrix},$$

where $z = q'y$. This applies the orthogonal transformation q' to y, multiplies the last $N-J$ rows by the orthogonal matrix g, and then uses q to transform back. So we can simplify notation by working with the one-to-one transformation $Z \equiv q'Y$, with

$$Z \,|\, x \stackrel{d}{=} \begin{pmatrix} s \\ 0 \end{pmatrix} a(\gamma) + \tilde{\tau}b(\gamma) + Wc(\gamma)$$

and $\tilde{\tau} = q'\tau$. Let $\mathcal{Z} = \mathcal{R}^{N\times M}$ denote the sample space and partition a point $z \in \mathcal{Z}$ as

$$z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix},$$

where z_1 is $J \times M$ and z_2 is $(N-J) \times M$. Then the action of G on the sample space is given by

$$m_1\colon G \times Z \to Z, \quad m_1(g, z) = \begin{pmatrix} I_J & 0 \\ 0 & g \end{pmatrix} z = \begin{pmatrix} z_1 \\ g z_2 \end{pmatrix}.$$

We shall abbreviate $m_1(g, z) = g \cdot z$. This defines a group action because for all $g_1, g_2 \in G$ and $z \in \mathcal{Z}$, we have $e \cdot z = z$ and $(g_1 g_2) \cdot z = g_1 \cdot (g_2 \cdot z)$, where $e = I_{N-J}$ is the identity element in G.

Partition $\tilde{\tau}$ as

$$\tilde{\tau} = \begin{pmatrix} \tilde{\tau}_1 \\ \tilde{\tau}_2 \end{pmatrix},$$

where $\tilde{\tau}_1$ is $J \times K$ and $\tilde{\tau}_2$ is $(N-J) \times K$. Note that

$$g \cdot Z \,|\, x \stackrel{d}{=} \begin{pmatrix} s \\ 0 \end{pmatrix} a(\gamma) + \begin{pmatrix} \tilde{\tau}_1 \\ g\tilde{\tau}_2 \end{pmatrix} b(\gamma) + W c(\gamma) \tag{6}$$

(because $\mathcal{L}(g \cdot W) = \mathcal{L}(W)$). It is convenient to define $\delta = s^{-1}\tilde{\tau}_1$ and write (6) as

$$g \cdot Z \,|\, x \stackrel{d}{=} \begin{pmatrix} s \\ 0 \end{pmatrix} \pi(\gamma, \delta) + \begin{pmatrix} 0 \\ g\tilde{\tau}_2 \end{pmatrix} b(\gamma) + W c(\gamma),$$

where

$$\pi(\gamma, \delta) = a(\gamma) + \delta b(\gamma).$$

Note that the least-squares projection coefficient of τ on x is

$$(x'x)^{-1} x'\tau = s^{-1}\tilde{\tau}_1 = \delta.$$

So δ captures a linear relationship between the individual effects τ and x.

Let $\mathcal{F}_{K,N-J}$ denote the set of $(N-J) \times K$ matrices whose columns are orthogonal and have unit length:

$$\mathcal{F}_{K,N-J} = \{d \in \mathcal{R}^{(N-J)\times K} : d'd = I_K\}.$$

($\mathcal{F}_{K,N-J}$ is the Stiefel manifold of ordered sets of K orthonormal vectors in $\mathcal{R}^{N-J}$; see Bishop and Crittenden (1964, p. 137).) The matrix $\tilde{\tau}_2$ has polar decomposition

$$\tilde{\tau}_2 = \omega\rho, \quad \omega \in \mathcal{F}_{K,N-J}, \quad \rho = (\tilde{\tau}_2'\tilde{\tau}_2)^{1/2},$$

where ρ is the unique, symmetric positive semidefinite square root of $\tilde{\tau}_2'\tilde{\tau}_2$. Partition $q = (\, q_1 \quad q_2 \,)$, where q_1 is $N \times J$ and q_2 is $N \times (N - J)$. Note that $x = q_1 s$ implies that

$$x(x'x)^{-1}x' = q_1 q_1',$$

and $qq' = I_N$ implies that

$$\begin{aligned} q_2 q_2' &= I_N - x(x'x)^{-1}x', \\ \rho^2 = \tilde{\tau}_2'\tilde{\tau}_2 &= \tau' q_2 q_2' \tau = \tau'(I_N - x(x'x)^{-1}x')\tau. \end{aligned}$$

So ρ^2 is formed from the residual sums of squares and cross products in the least squares projection of τ on x.

Then we can write the model in (1) as

$$Z \,|\, x \overset{d}{=} \begin{pmatrix} s \\ 0 \end{pmatrix} \pi(\gamma, \delta) + \begin{pmatrix} 0 \\ \omega \end{pmatrix} \rho b(\gamma) + W c(\gamma), \quad \mathcal{L}(W) = \mathcal{N}(0, I_N \otimes I_p). \tag{7}$$

Let $\theta = (\beta, \omega)$ denote the parameter, with $\beta = (\gamma, \delta, \rho)$. The parameter space is

$$\Theta = \Theta_1 \times \Theta_2 \quad \text{with} \quad \Theta_2 = \mathcal{F}_{K,N-J}$$

(and Θ_1 is a subset of some Euclidean space). We shall let P_θ denote the distribution of Z (conditional on x) when the parameter takes on the value θ: $\mathcal{L}(Z) = P_\theta$.

The action of the group G on the parameter space is given by

$$m_2 \colon G \times \Theta \to \Theta, \quad m_2(g, \theta) = m_2(g, (\beta, \omega)) = (\beta, g\omega).$$

We shall abbreviate $m_2(g, \theta) = g \cdot \theta$. This defines a group action because for all $g_1, g_2 \in G$ and $\theta \in \Theta$, we have $e \cdot \theta = \theta$ and $(g_1 g_2) \cdot \theta = g_1 \cdot (g_2 \cdot \theta)$. Then

$$g \cdot Z \,|\, x \overset{d}{=} \begin{pmatrix} s \\ 0 \end{pmatrix} \pi(\gamma, \delta) + \begin{pmatrix} 0 \\ g\omega \end{pmatrix} \rho b(\gamma) + W c(\gamma), \quad \mathcal{L}(W) = \mathcal{N}(0, I_N \otimes I_p),$$

and so

$$\mathcal{L}(Z) = P_\theta \quad \text{implies} \quad \mathcal{L}(g \cdot Z) = P_{g \cdot \theta},$$

and the model is invariant under the actions of G on the sample space and the parameter space.

3. MAXIMAL INVARIANT STATISTIC

A statistic S is a (measurable) function defined on the sample space. S is invariant if $S(g \cdot z) = S(z)$ for all $g \in G$ and $z \in \mathcal{Z}$. Let P_θ^S denote the distribution of $S(Z)$ when $\mathcal{L}(Z) = P_\theta$. If S is an invariant statistic, then for all $g \in G$ and $\theta \in \Theta$:

$$P_\theta^S = \mathcal{L}(S(Z)) = \mathcal{L}(S(g \cdot Z)) = P_{g \cdot \theta}^S.$$

The orbit of a point $\theta \in \Theta$ under the action of G is the set $\{g \cdot \theta : g \in G\}$. Note that for any $\omega_1, \omega_2 \in \Theta_2$, there exists a $g \in G$ such that $g\omega_1 = \omega_2$, and hence for any $\beta \in \Theta_1$, the points (β, ω) are in the same orbit for all $\omega \in \Theta_2$. (The action of G on Θ_2, defined by $m(g, \omega) = g\omega$, is transitive.) So the distribution of an invariant statistic does not depend upon ω.

Let $T(z) = (T_1(z), T_2(z)) = (z_1, z_2' z_2)$. Then

$$T(g \cdot (z_1, z_2)) = T(z_1, g z_2) = (z_1, z_2' g' g z_2) = (z_1, z_2' z_2),$$

and so T is an invariant statistic. We shall show that T is a maximal invariant statistic: if S is an invariant statistic, then for any $z, \tilde{z} \in \mathcal{Z}$, $T(z) = T(\tilde{z})$ implies that $S(z) = S(\tilde{z})$. This result is a consequence of the following proposition:

Proposition 1. If $T(z) = t = (t_1, t_2)$, then there exists a $g_z \in G$ such that $z = g_z \cdot r(t)$, where

$$r(t) = \begin{pmatrix} t_1 \\ \begin{pmatrix} t_2^{1/2} \\ 0 \end{pmatrix} \end{pmatrix} \in \mathcal{Z}.$$

Proof. The matrix z_2 can be decomposed as

$$z_2 = h \begin{pmatrix} (z_2' z_2)^{1/2} \\ 0 \end{pmatrix} \quad \text{where} \quad h \in O(N - J).$$

Set $g_z = h$. Then

$$g_z^{-1} \cdot z = \begin{pmatrix} I_J & 0 \\ 0 & g_z^{-1} \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} t_1 \\ \begin{pmatrix} t_2^{1/2} \\ 0 \end{pmatrix} \end{pmatrix} = r(t). \quad \diamond$$

Corollary. T is a maximal invariant statistic.

Proof. Suppose that S is an invariant statistic. If $T(z) = T(\tilde{z}) = t$, then Proposition 1 implies that

$$g_z^{-1} \cdot z = g_{\tilde{z}}^{-1} \cdot \tilde{z} = r(t) \quad \text{with} \quad g_z,\, g_{\tilde{z}} \in G.$$

Hence z and $\tilde{z}$ are in the same orbit:

$$g \cdot z = \tilde{z} \quad \text{for} \quad g = g_{\tilde{z}} g_z^{-1} \in G.$$

So

$$S(z) = S(g \cdot z) = S(\tilde{z}). \quad \diamond$$

The orbit of a point $z \in \mathcal{Z}$ under the action of G is the set $\{g \cdot z : g \in G\}$. The maximal invariant T indexes the orbits in the sample space: if $T(z_1) = T(z_2) = t$, then z_1 and z_2 are in the orbit of $r(t)$. The set $\{r(T(z)) : z \in \mathcal{Z}\}$ contains one point from each orbit. It is a measurable cross section; see Eaton (1989, p. 58). In the parameter space, for any point $\beta \in \Theta_1$, the points (β, ω) are in the same orbit for all $\omega \in \Theta_2$. So we can fix some point $\omega_0 \in \Theta_2$, and then the set $\{(\beta, \omega_0) : \beta \in \Theta_1\}$ contains one point from each orbit in the parameter space. It is a measurable cross section in the parameter space.

4. INVARIANT PRIOR DISTRIBUTION

Since G is a compact group, there is a unique invariant distribution μ on G: Haar measure normalized so that $\mu(G) = 1$. Let U denote a random variable taking on values in G. The invariance property is that

$$\mathcal{L}(U) = \mu \quad \text{implies} \quad \mathcal{L}(gU) = \mathcal{L}(Ug) = \mu \quad \text{for all} \quad g \in G.$$

We shall refer to the invariant distribution μ as the uniform distribution on G. This invariant distribution on G implies a unique invariant distribution λ on the compact set $\Theta_2 = \mathcal{F}_{K,N-J}$; see Eaton (1989, example 2.10, p. 27). This distribution can be obtained from μ by fixing some point

$\omega_0 \in \Theta_2$ and setting $\lambda = \mathcal{L}(U\omega_0)$, where $\mathcal{L}(U) = \mu$. The distribution λ does not depend upon the point ω_0, since if ω_1 is some other point in Θ_2, with $\omega_1 = g\omega_0$ for some $g \in G$, then

$$\mathcal{L}(U\omega_1) = \mathcal{L}(U(g\omega_0)) = \mathcal{L}((Ug)\omega_0) = \mathcal{L}(U\omega_0) = \lambda.$$

Let V be a random variable taking on values in Θ_2. Then the invariance property of λ is that

$$\mathcal{L}(V) = \lambda \quad \text{implies} \quad \mathcal{L}(gV) = \mathcal{L}(g(U\omega_0)) = \mathcal{L}((gU)\omega_0) = \mathcal{L}(U\omega_0) = \lambda$$

for all $g \in G$. We shall refer to the invariant distribution λ as the uniform distribution on Θ_2.

Define

$$\Omega(\gamma) = c(\gamma)'c(\gamma),$$

and assume that $c(\gamma)$ has full column rank for all $\beta = (\gamma, \delta, \rho) \in \Theta_1$. Let $f(z \,|\, \beta, \omega)$ denote the likelihood function:

$$f(z \,|\, \beta, \omega) = (2\pi)^{-NM/2} \det(\Omega(\gamma))^{-N/2} \exp\Big(-\frac{1}{2}\text{trace}[\Omega(\gamma)^{-1} k(z, \beta, \omega)' k(z, \beta, \omega)]\Big),$$

where

$$k(z, \beta, \omega) = z - \begin{pmatrix} s \\ 0 \end{pmatrix} \pi(\gamma, \delta) - \begin{pmatrix} 0 \\ \omega \end{pmatrix} \rho b(\gamma).$$

We can use the uniform distribution on Θ_2 as a prior distribution to obtain an integrated likelihood function:

$$f_\lambda(z \,|\, \beta) = \int_{\Theta_2} f(z \,|\, \beta, \omega)\, \lambda(d\omega).$$

The next proposition shows that this integrated likelihood function depends upon z only through the maximal invariant $T(z)$.

Proposition 2. For all $z \in \mathcal{Z}$ and $\beta \in \Theta_1$, $f_\lambda(z \,|\, \beta) = f_\lambda(r(T(z)) \,|\, \beta)$.

Proof. Note that for any $g \in G$,

$$k(g^{-1} \cdot z, \beta, \omega) = \begin{pmatrix} I_J & 0 \\ 0 & g^{-1} \end{pmatrix} [z - \begin{pmatrix} s \\ 0 \end{pmatrix} \pi(\gamma, \delta) - \begin{pmatrix} 0 \\ g\omega \end{pmatrix} \rho b(\gamma)] = \begin{pmatrix} I_J & 0 \\ 0 & g^{-1} \end{pmatrix} k(z, \beta, g\omega),$$

and so, for all $z \in \mathcal{Z}$ and $(\beta, \omega) \in \Theta$,

$$f(g^{-1} \cdot z \,|\, \beta, \omega) = f(z \,|\, \beta, g\omega).$$

(See Eaton (1989, p. 44) for a general discussion of this point.) As in Proposition 1, $z = g_z \cdot r(T(z))$. So

$$\begin{aligned}
\int_{\Theta_2} f(z \,|\, \beta, \omega)\, \lambda(d\omega) &= \int_{\Theta_2} f(g_z \cdot r(T(z)) \,|\, \beta, \omega)\, \lambda(d\omega) \\
&= \int_{\Theta_2} f(r(T(z)) \,|\, \beta, g_z^{-1}\omega)\, \lambda(d\omega) \\
&= \int_{\Theta_2} f(r(T(z)) \,|\, \beta, \omega)\, \lambda(d\omega). \quad \diamond
\end{aligned}$$

We can use the maximal invariant statistic T to construct a marginal likelihood function, based on a density for the distribution of T. The next proposition uses Proposition 2 to show that this marginal likelihood function can be obtained from the integrated likelihood function. Let P_β^T denote the distribution of $T(Z)$ when $\mathcal{L}(Z) = P_{(\beta,\omega)}$; the value of ω does not matter since T is an invariant statistic. Let ζ denote Lebesgue measure on $\mathcal{R}^N \times \mathcal{R}^M$, and let $\nu = \zeta T^{-1}$ denote the following measure:

$$\nu(A) = \zeta(T^{-1}(A))$$

for (measurable) sets A in a Euclidean space containing $T(\mathcal{Z})$. Define

$$f^T(t \,|\, \beta) = f_\lambda(r(t) \,|\, \beta) \quad \text{for} \quad t \in T(\mathcal{Z}),\, \beta \in \Theta_1.$$

Proposition 3 shows that $f^T(t \,|\, \beta)$ provides a density function for P_β^T:

$$P_\beta^T(A) = \int_A f^T(t \,|\, \beta)\, \nu(dt).$$

Proposition 3. $f_\lambda(r(t) \,|\, \beta)$ is a density for P_β^T with respect to the measure ν.

Proof. For all $\omega \in \Theta_2$,

$$P_\beta^T(A) = P_{(\beta,\omega)}(T^{-1}(A)),$$

and so

$$
\begin{aligned}
P_\beta^T(A) &= \int_{\Theta_2} P_{(\beta,\omega)}(T^{-1}(A))\,\lambda(d\omega) \\
&= \int_{\Theta_2} \left[\int_{T^{-1}(A)} f(z\,|\,\beta,\omega)\,\zeta(dz)\right] \lambda(d\omega) \\
&= \int_{T^{-1}(A)} \left[\int_{\Theta_2} f(z\,|\,\beta,\omega)\,\lambda(d\omega)\right] \zeta(dz) \\
&= \int_{T^{-1}(A)} f_\lambda(r(T(z))\,|\,\beta)\,\zeta(dz) \\
&= \int_A f_\lambda(r(t)\,|\,\beta)\,\zeta T^{-1}(dt). \quad \diamond
\end{aligned}
$$

We can eliminate the parameter ω by integration with respect to the invariant prior distribution, to obtain the integrated likelihood function $f_\lambda(z\,|\,\beta)$. Or we can eliminate ω by working with the marginal likelihood function $f^T(t\,|\,\beta)$, based on the maximal invariant statistic T. Propositions 2 and 3 show that these likelihood functions coincide:

$$f_\lambda(z\,|\,\beta) = f_\lambda(r(T(z))\,|\,\beta) = f^T(T(z)\,|\,\beta).$$

Having eliminated ω, we can ask whether γ is identified in these likelihood functions. This will depend on the particular specifications for $a(\gamma)$, $b(\gamma)$, $c(\gamma)$, and one can examine the following moment conditions based on $T(Z)$:

$$
\begin{aligned}
&E[(x'x)^{-1}x'Y] = s^{-1}E(Z_1) = \pi(\gamma,\delta), \\
&\qquad E(Y'Y) = E(Z_1'Z_1 + Z_2'Z_2) = \pi(\gamma,\delta)'x'x\pi(\gamma,\delta) + b(\gamma)'\rho^2 b(\gamma) + Nc(\gamma)'c(\gamma).
\end{aligned}
$$

5. OPTIMALITY

Using the likelihood function of an invariant statistic has the advantage of eliminating dependence on the parameter ω. The concern is that, even using the maximal invariant statistic, we are not using all of the data. This concern can be addressed in our case, since the marginal likelihood function based on T coincides with the integrated likelihood function when we use the invariant prior distribution for ω.

Suppose the loss function does not depend upon ω:

$$L : \Theta_1 \times \mathcal{A} \to \mathcal{R},$$

where $\mathcal{A}$ is the action space. The corresponding risk function is

$$R((\beta, \omega), d) = \int_{\mathcal{Z}} L(\beta, d(z)) f(z \,|\, \beta, \omega)\, \zeta(dz),$$

where $d : \mathcal{Z} \to \mathcal{A}$ is in the set $\mathcal{D}$ of feasible decision rules; $\mathcal{D}$ is unrestricted except for regularity conditions. Let η be some prior distribution on Θ_1, and consider the average risk with respect to the prior distribution $\eta \times \lambda$ on Θ:

$$\begin{aligned} R^*(\eta \times \lambda, d) &= \int_{\Theta_1} \int_{\Theta_2} R((\beta, \omega), d)\, \lambda(d\omega)\, \eta(d\beta) \\ &= \int_{\Theta_1} \int_{\mathcal{Z}} L(\beta, d(z)) f_\lambda(z \,|\, \beta)\, \zeta(dz)\, \eta(d\beta). \end{aligned}$$

So choosing d to minimize this average risk function can be based on the integrated likelihood function. Under regularity conditions, we have the standard result that the optimal d is obtained by minimizing posterior expected loss:

$$\begin{aligned} d(z) &= \arg\min_{a \in \mathcal{A}} \int_{\Theta_1} L(\beta, a) f_\lambda(z \,|\, \beta)\, \eta(d\beta) \\ &= \arg\min_{a \in \mathcal{A}} \int_{\Theta_1} L(\beta, a) f^T(T(z) \,|\, \beta)\, \eta(d\beta). \end{aligned}$$

So we can obtain an optimal decision rule using the marginal likelihood function. This optimal decision rule is a function of the maximal invariant statistic—it depends upon z only through $T(z)$—but this was not imposed as a constraint on $\mathcal{D}$ in the optimization. See Eaton (1989, Chapter 6) for a general discussion of invariant decision rules.

Suppose that $d_{\eta \times \lambda}$ minimizes average risk:

$$d_{\eta \times \lambda} = \arg\min_{d \in \mathcal{D}} R^*(\eta \times \lambda, d)$$

and depends upon z only through $T(z)$:

$$d_{\eta\times\lambda}(z) = \tilde{d}_{\eta\times\lambda}(T(z)).$$

The next proposition establishes a minimax property for this decision rule. The argument is based on Chamberlain (2007, Theorem 6.1).

Proposition 4. $d_{\eta\times\lambda}$ solves the following problem, which combines the average risk and maximum risk criteria:

$$d_{\eta\times\lambda} = \arg\min_{d\in\mathcal{D}} \int_{\Theta_1} [\sup_{\omega\in\Theta_2} R((\beta,\omega),d)]\,\eta(d\beta).$$

Proof. Let $\mathcal{L}(Z) = P_{(\beta,\omega)}$. Then

$$R((\beta,\omega),d_{\eta\times\lambda}) = E[L(\beta,\tilde{d}_{\eta\times\lambda}(T(Z)))],$$

which does not depend upon ω since T is an invariant statistic. So we can fix a point $\omega_0 \in \Theta_2$, define $\tilde{R}(\beta,d_{\eta\times\lambda}) = R((\beta,\omega_0),d_{\eta\times\lambda})$, and then, for all $\beta\in\Theta_1$ and $\omega\in\Theta_2$, we have $R((\beta,\omega),d_{\eta\times\lambda})$ $= \tilde{R}(\beta,d_{\eta\times\lambda})$. For any $d\in\mathcal{D}$,

$$\begin{aligned}
&\int_{\Theta_1} [\sup_{\omega\in\Theta_2} R((\beta,\omega),d)]\,\eta(d\beta) \\
&\qquad \geq \int_{\Theta_1} \left[\int_{\Theta_2} R((\beta,\omega),d)\,\lambda(d\omega)\right]\eta(d\beta) \\
&\qquad \geq \int_{\Theta_1} \left[\int_{\Theta_2} R((\beta,\omega),d_{\eta\times\lambda})\,\lambda(d\omega)\right]\eta(d\beta) \\
&\qquad = \int_{\Theta_1} \tilde{R}(\beta,d_{\eta\times\lambda})\,\eta(d\beta) \\
&\qquad = \int_{\Theta_1} [\sup_{\omega\in\Theta_2} R((\beta,\omega),d_{\eta\times\lambda})]\,\eta(d\beta). \quad \diamond
\end{aligned}$$

The use of minimax here does not eliminate the choice of a prior distribution; the average risk criteria on the parameter space Θ_1 for β requires that we specify a prior distribution η. But we can replace the choice of a prior distribution on the parameter space $\mathcal{F}_{K,N-J}$ for ω by the maximum risk criterion. The solution to the minimax problem calls for a particular, least favorable, distribution on $\mathcal{F}_{K,N-J}$: the uniform distribution λ. This minimax treatment of the incidental parameters can be obtained using the marginal likelihood function $f^T(\cdot \,|\, \beta)$ based on the maximal invariant statistic.

Recall from Section 2 that

$$\delta = (x'x)^{-1}x'\tau, \quad \rho^2 = \tau'(I_N - x(x'x)^{-1}x')\tau.$$

Define the set-valued function $B(\cdot\,,\cdot)$ by

$$B(\delta,\rho) = \{\tau \in \mathcal{R}^{N\times K} : (x'x)^{-1}x'\tau = \delta,\ \tau'(I_N - x(x'x)^{-1}x')\tau = \rho^2\}.$$

The minimax result in Proposition 4 can be related to the original parametrization by replacing the sup over ω in Θ_2 by the sup over τ in $B(\delta,\rho)$.

6. A CLOSED FORM INTEGRATED LIKELIHOOD AND CORRELATED RANDOM EFFECTS

Our finite sample optimality result uses a prior distribution η for β, where $\beta = (\gamma,\delta,\rho)$. This section develops a family of prior distributions for (δ,ρ) that leads to a simple, explicit form for the integrated likelihood. The basic idea is that the uniform distribution λ on $\mathcal{F}_{K,N-J}$ can be combined with a central Wishart distribution to obtain a multivariate normal distribution.

We start with a family of prior distributions for ρ that is indexed by a parameter Φ, which is a $K\times K$ symmetric, positive semidefinite matrix. Let

$$\mathcal{L}(Q) = \mathcal{N}(0, I_{N-J} \otimes \Phi).$$

Let

$$\kappa_\Phi = \mathcal{L}((Q'Q)^{1/2})$$

be the prior distribution for ρ with parameter Φ. Then the corresponding integrated likelihood function is

$$\begin{aligned} f_{\lambda,\kappa}(z \,|\, \gamma,\delta,\Phi) &= \int f_\lambda(z \,|\, (\gamma,\delta,\rho))\, \kappa_\Phi(d\rho) \\ &= \int\!\!\int f(z \,|\, (\gamma,\delta,\rho),\omega)\, \lambda(d\omega)\, \kappa_\Phi(d\rho). \end{aligned}$$

Suppose that V is independent of $Q'Q$, with $\mathcal{L}(V) = \lambda$. Then

$$\mathcal{L}(Q) = \mathcal{L}(V(Q'Q)^{1/2});$$

see Eaton (1989, Example 4.4, p. 61). So the distribution for $\tilde{\tau}_2 = \omega\rho$ implied by $\lambda \times \kappa_\Phi$ is $\mathcal{N}(0, I_{N-J} \otimes \Phi)$. This implies that the log of the integrated likelihood function is

$$\begin{aligned}
\log[f_{\lambda,\kappa}(z \,|\, \gamma, \delta, \Phi)] = &-\frac{NM}{2}\log(2\pi) - \frac{J}{2}\log[\det(\Omega(\gamma))] \\
&- \frac{N-J}{2}\log\big[\det[b(\gamma)'\Phi b(\gamma) + \Omega(\gamma)]\big] \\
&- \frac{1}{2}\text{trace}[\Omega(\gamma)^{-1}(z_1 - s\pi(\gamma,\delta))'(z_1 - s\pi(\gamma,\delta))] \\
&- \frac{1}{2}\text{trace}\big[[b(\gamma)'\Phi b(\gamma) + \Omega(\gamma)]^{-1} z_2' z_2\big].
\end{aligned}$$

Fix a value for the parameter in (7):

$$\beta^* = (\gamma^*, \delta^*, \rho^*) \in \Theta_1, \quad \omega^* \in \Theta_2.$$

Let $\mathcal{L}(Z) = P_{(\beta^*,\omega^*)}$ and define

$$l(\gamma, \delta, \Phi) = E\big[\log[f_{\lambda,\kappa}(Z \,|\, \gamma, \delta, \Phi)]\big].$$

Note that this expectation does not depend upon the normality assumption for W in (7); only the first and second moments of Z are used, and so $l(\gamma, \delta, \Phi)$ depends upon $\mathcal{L}(W)$ only through its first and second moments. Evaluating $E(Z_1)$, $E(Z_1'Z_1)$, and $E(Z_2'Z_2)$ gives

$$\begin{aligned}
l(\gamma, \delta, \Phi) = &-\frac{NM}{2}\log(2\pi) - \frac{J}{2}\log[\det(\Omega(\gamma))] \\
&- \frac{N-J}{2}\log\big[\det[b(\gamma)'\Phi b(\gamma) + \Omega(\gamma)]\big] \\
&- \frac{1}{2}\text{trace}\big[\Omega(\gamma)^{-1}[(\pi(\gamma^*,\delta^*) - \pi(\gamma,\delta))'x'x(\pi(\gamma^*,\delta^*) - \pi(\gamma,\delta))] + J\Omega(\gamma)^{-1}\Omega(\gamma^*)\big] \\
&- \frac{1}{2}\text{trace}\big[[b(\gamma)'\Phi b(\gamma) + \Omega(\gamma)]^{-1}[b(\gamma^*)'\rho^{*2} b(\gamma^*) + (N-J)\Omega(\gamma^*)]\big].
\end{aligned}$$

The maximum of $l(\gamma, \delta, \Phi)$ is attained at

$$\gamma = \gamma^*, \quad \delta = \delta^*, \quad \Phi = {\rho^*}^2/(N-J).$$

This result is useful for obtaining asymptotic properties of the estimator that maximizes the integrated (quasi) log-likelihood function.

In order to make connections between a correlated random effects model and our fixed-effects approach, it is convenient to introduce a prior distribution for δ, in addition to the prior distribution for ρ that was chosen to obtain a closed form for the integrated likelihood. The family of prior distributions for (δ, ρ) is indexed by the parameter (ι, Φ), where ι is a $J \times K$ matrix and Φ is a $K \times K$ symmetric, positive semidefinite matrix. Let

$$\mathcal{L}(\begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix}) = \mathcal{N}(0, I_N \otimes \Phi),$$

where Q_1 is $J \times K$ and Q_2 is $(N-J) \times K$. Let

$$\kappa_{\iota,\Phi} = \mathcal{L}(\iota + s^{-1}Q_1, (Q_2'Q_2)^{1/2})$$

be the prior distribution for (δ, ρ). The distribution for $\omega\rho$ implied by $\lambda \times \kappa_{\iota,\Phi}$ is $\mathcal{N}(0, I_{N-J} \otimes \Phi)$ (as above), and the distribution for $(s\pi(\gamma, \delta), \omega\rho)$ is

$$\mathcal{N}(s\pi(\gamma, \iota), I_J \otimes b(\gamma)'\Phi b(\gamma)) \times \mathcal{N}(0, I_{N-J} \otimes \Phi).$$

The corresponding integrated likelihood function is

$$\bar{f}_{\lambda,\kappa}(z \,|\, \gamma, \iota, \Phi) = \int f_\lambda(z \,|\, (\gamma, \delta, \rho))\, \kappa_{\iota,\Phi}(d\delta, d\rho).$$

Evaluating the log of this integrated likelihood function gives

$$\begin{aligned} \log[\bar{f}_{\lambda,\kappa}(z \,|\, \gamma, \iota, \Phi)] = &-\frac{NM}{2}\log(2\pi) - \frac{N}{2}\log\big[\det[b(\gamma)'\Phi b(\gamma) + \Omega(\gamma)]\big] \\ &- \frac{1}{2}\mathrm{trace}\big[[b(\gamma)'\Phi b(\gamma) + \Omega(\gamma)]^{-1}[(z_1 - s\pi(\gamma,\iota))'(z_1 - s\pi(\gamma,\iota)) + z_2'z_2]\big]. \end{aligned} \tag{8}$$

As above, fix a value (β^*, ω^*) for the parameter, let $\mathcal{L}(Z) = P_{(\beta^*,\omega^*)}$, and define

$$\bar{l}(\gamma, \iota, \Phi) = E\big[\log[\bar{f}_{\lambda,\kappa}(Z \mid \gamma, \iota, \Phi)]\big].$$

As before, this expectation does not depend upon the normality assumption for W in (7); only the first and second moments of Z are used. Evaluating $E(Z_1)$, $E(Z_1'Z_1)$, and $E(Z_2'Z_2)$ gives

$$\begin{aligned}\bar{l}(\gamma, \iota, \Phi) = &-\frac{NM}{2}\log(2\pi) - \frac{N}{2}\log\big[\det[b(\gamma)'\Phi b(\gamma) + \Omega(\gamma)]\big] \\ &- \frac{1}{2}\text{trace}\big[[b(\gamma)'\Phi b(\gamma) + \Omega(\gamma)]^{-1}[(\pi(\gamma^*, \delta^*) - \pi(\gamma, \iota))'x'x(\pi(\gamma^*, \delta^*) - \pi(\gamma, \iota)) \\ &+ b(\gamma^*)'\rho^{*2}b(\gamma^*) + N\Omega(\gamma^*)]\big].\end{aligned} \tag{9}$$

The maximum of $\bar{l}(\gamma, \iota, \Phi)$ is attained at

$$\gamma = \gamma^*, \quad \iota = \delta^*, \quad \Phi = \rho^{*2}/N.$$

Consider the following correlated random effects specification for the incidental parameters:

$$\tau \mid x \stackrel{d}{=} \mathcal{N}(x\iota, I_N \otimes \Phi). \tag{10}$$

Combining this with the model in (1), the implied distribution for the observation is

$$Y \mid x \stackrel{d}{=} \mathcal{N}(x\pi(\gamma, \iota), I_N \otimes [b(\gamma)'\Phi b(\gamma) + \Omega(\gamma)]), \tag{11}$$

and the log-likelihood function is

$$\begin{aligned}\log[f^{re}(y \mid \gamma, \iota, \Phi)] = &-\frac{NM}{2}\log(2\pi) - \frac{N}{2}\log\big[\det[b(\gamma)'\Phi b(\gamma) + \Omega(\gamma)]\big] \\ &- \frac{1}{2}\text{trace}\big[[b(\gamma)'\Phi b(\gamma) + \Omega(\gamma)]^{-1}[(y - x\pi(\gamma, \iota))'(y - x\pi(\gamma, \iota))]\big].\end{aligned} \tag{12}$$

We shall refer to this as the normal, correlated random effects model. As in Section 2, let

$$x = q\begin{pmatrix} s \\ 0 \end{pmatrix}, \quad z = q'y,$$

where q is a $N \times N$ orthogonal matrix; then we can write the log-likelihood function as

$$\log[f^{re}(qz \mid \gamma, \iota, \Phi)] = -\frac{NM}{2}\log(2\pi) - \frac{N}{2}\log\big[\det[b(\gamma)'\Phi b(\gamma) + \Omega(\gamma)]\big] \tag{13}$$
$$- \frac{1}{2}\text{trace}\big[[b(\gamma)'\Phi b(\gamma) + \Omega(\gamma)]^{-1}[(z_1 - s\pi(\gamma,\iota))'(z_1 - s\pi(\gamma,\iota)) + z_2'z_2]\big]$$
$$= \log[\bar{f}_{\lambda,\kappa}(z \mid \gamma, \iota, \Phi)].$$

So the log of the normal, correlated random effects likelihood function coincides with the log of the integrated likelihood function in equation (8). This connection with a correlated random effects model helps to relate our results to the literature on panel data.

7. CONNECTIONS WITH THE LITERATURE

A simple version of our model arises from the reduced form of the following autoregression:

$$Y_{it} = \psi Y_{i,t-1} + \alpha_i + U_{it} \qquad (i = 1, \ldots, N;\ t = 1, \ldots, \bar{T}), \tag{14}$$

where the U_{it} are independent and identically distributed $\mathcal{N}(0, \sigma^2)$. We observe the realized value of the random variable Y_{it} for $i = 1, \ldots, N$ and $t = 1, \ldots, \bar{T}$. We do not observe Y_{i0}. This specification implies a likelihood function for $\{Y_{i1}, \ldots, Y_{i\bar{T}}\}_{i=1}^N$, conditional on $\{y_{i0}, \alpha_i\}_{i=1}^N$. Our framework allows for conditioning on time-varying covariates x_{it}, but in this simple version we shall just use $x = 1_N$, where 1_N denotes a $N \times 1$ matrix of ones. To obtain our canonical form, use an $N \times N$ orthogonal matrix q whose first column is proportional to 1_N: $q = \begin{pmatrix} 1_N/\sqrt{N} & q_2 \end{pmatrix}$, so that

$$x = q\begin{pmatrix} \sqrt{N} \\ 0 \end{pmatrix}.$$

Note that $qq' = I_N$ implies that $q_2 q_2' = I_N - 1_N 1_N'/N$. Then our transformation of the $N \times \bar{T}$ matrix Y is

$$Z = q'Y = \begin{pmatrix} \sqrt{N}\bar{Y} \\ q_2'Y \end{pmatrix}, \quad \text{where} \quad \bar{Y} = \begin{pmatrix} \sum_{i=1}^N Y_{i1}/N & \ldots & \sum_{i=1}^N Y_{i\bar{T}}/N \end{pmatrix}.$$

Our transformation of the parameters uses

$$\tilde{\tau} = q'\tau = \begin{pmatrix} \sqrt{N}\bar{\tau} \\ \tilde{\tau}_2 \end{pmatrix}$$

with $\tilde{\tau}_2 = q_2'\tau$ and

$$\begin{aligned}
\tilde{\tau}_2'\tilde{\tau}_2 &= \tau' q_2 q_2' \tau = (\tau - 1_N\bar{\tau})'(\tau - 1_N\bar{\tau}) \\
&= \begin{pmatrix} \sum_{i=1}^N (y_{i0} - \bar{y}_0)^2 & \sum_{i=1}^N (y_{i0} - \bar{y}_0)(\alpha_i - \bar{\alpha}) \\ \sum_{i=1}^N (y_{i0} - \bar{y}_0)(\alpha_i - \bar{\alpha}) & \sum_{i=1}^N (\alpha_i - \bar{\alpha})^2 \end{pmatrix}.
\end{aligned} \tag{15}$$

Then

$$\delta = \bar{\tau} = \begin{pmatrix} \sum_{i=1}^N y_{i0}/N & \sum_{i=1}^N \alpha_i/N \end{pmatrix}, \quad \rho = (\tilde{\tau}_2'\tilde{\tau}_2)^{1/2}.$$

The maximal invariant statistic $T(Z) = (T_1(Z), T_2(Z))$ has

$$T_1(Z) = \sqrt{N}\bar{Y}, \quad T_2(Z) = (Y - 1_N\bar{Y})'(Y - 1_N\bar{Y}).$$

The distribution of this statistic depends only upon $(\psi, \sigma, \delta, \rho)$, which has dimension seven; the distribution does not depend upon ω, which has dimension $2N-5$. Let $\gamma = (\psi, \sigma)$. The distribution of the maximal invariant statistic has density $f^T(\cdot \,|\, \gamma, \delta, \rho)$, which is based on a normal distribution for $\bar{Y}$ and an independent noncentral Wishart distribution for $(Y - 1_N\bar{Y})'(Y - 1_N\bar{Y})$.

Our optimality result requires a prior distribution for $(\psi, \sigma, \delta, \rho)$. The dimension reduction shows up in not requiring a prior distribution for ω—that is where the minimax result is used.

There is a particular family of prior distributions for (δ, ρ) that connects to the literature on random effects models. The family is indexed by the parameter (ι, Φ), where ι is 1×2 and Φ is a 2×2 symmetric, positive semidefinite matrix. Let

$$\mathcal{L}\left(\begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix}\right) = \mathcal{N}(0, I_N \otimes \Phi),$$

where Q_1 is 1×2 and Q_2 is $(N-1) \times 2$. The prior for (δ, ρ) is

$$\kappa_{\iota,\Phi} = \mathcal{L}(\iota + N^{-1/2}Q_1, (Q_2'Q_2)^{1/2}).$$

Combining $f^T(T(z) \,|\, \gamma, \delta, \rho)$ with this family of prior distributions for (δ, ρ) gives the integrated likelihood $\bar{f}_{\lambda,\kappa}(z \,|\, \gamma, \iota, \Phi)$, as in equation (8) in Section 6.

Now consider the following normal random effects model: the specification in (14) plus

$$(y_{i0}, \alpha_i) \overset{\text{i.i.d.}}{\sim} \mathcal{N}((\iota_1, \iota_2), \Phi) \qquad (i = 1, \ldots, N). \tag{16}$$

The likelihood function is $f^{re}(y \,|\, \gamma, \iota, \Phi)$, as in equation (12) in Section 6. Using the transformation $z = q'y$, our result in (13) shows that

$$f^{re}(qz \,|\, \gamma, \iota, \Phi) = \bar{f}_{\lambda,\kappa}(z \,|\, \gamma, \iota, \Phi).$$

A prior distribution for $(\psi, \sigma, \iota, \Phi)$, of dimension seven, is needed to obtain minimum average risk in finite samples. We do not have a specific recommendation for this prior. Our point is that the normal random effects likelihood function can be obtained from the likelihood function for the maximal invariant, in which the dimension of the parameter space has already been reduced to a number (seven) that does not depend upon N. In this sense, the incidental parameter problem has been dealt with in the original fixed-effects model in (14), which conditions on $\{y_{i0}, \alpha_i\}_{i=1}^N$, without relying on the specification of a random-effects distribution in (16).

Our paper has a finite sample perspective, but in connecting with the literature we shall consider limits as $N \to \infty$ in the context of our general model. In equation (9) of Section 6, we fix a "true value" $\theta^* = (\gamma^*, \delta^*, \rho^*, \omega^*)$ for the parameter in (7), and evaluate

$$\bar{l}(\gamma, \iota, \Phi) = E\big[\log[\bar{f}_{\lambda,\kappa}(Z \,|\, \gamma, \iota, \Phi)]\big],$$

with Z distributed according to P_{θ^*}. This expectation does not depend upon the normality assumption for W in (7), and the maximum of $\bar{l}(\gamma, \iota, \Phi)$ is attained at $\gamma = \gamma^*$, $\iota = \delta^*$, $\Phi = \rho^{*2}/N$. If there is a unique maximizing value for γ, it should be feasible to go from here to a consistency result for γ^* (as $N \to \infty$), without the normality assumption for W.

In fact, if we add an assumption of random sampling over the cross-section dimension i, then the asymptotics are straightforward and well known. Let $Y_{(i)}$, $x_{(i)}$, and $\tau_{(i)}$ denote the i^{th} rows of Y, x, and τ. Assume that

$$(Y_{(i)}, x_{(i)}, \tau_{(i)}) \qquad (i = 1, \ldots, N)$$

are independent and identically distributed from a joint distribution F, and let E_F denote expectation with respect to this distribution. We shall assume that the (unconditional) second moments of $(Y_{(i)}, x_{(i)})$ correspond to the normal, correlated random effects model, but we shall not make normality or homoskedasticity (conditional on x) assumptions in obtaining the limit distribution of the estimator. We shall refer to this semiparametric model simply as the correlated random effects model. Assume that

$$E_F[x'_{(i)}(Y_{(i)} - x_{(i)}\pi(\gamma^*, \iota^*))] = 0,$$

so that $x_{(i)}\pi(\gamma^*, \iota^*)$ is the minimum mean-square-error linear predictor of $Y_{(i)}$ given $x_{(i)}$. Define

$$\epsilon_{(i)} = Y_{(i)} - x_{(i)}\pi(\gamma^*, \iota^*)$$

and assume that

$$E_F(\epsilon'_{(i)}\epsilon_{(i)}) = b(\gamma^*)'\Phi^* b(\gamma^*) + \Omega(\gamma^*).$$

Let υ denote the column vector formed from γ, ι, and the lower triangle of Φ, and let

$$h(Y_{(i)}, x_{(i)}, \upsilon) = -\frac{1}{2}\log\big[\det[b(\gamma)'\Phi b(\gamma) + \Omega(\gamma)]\big]$$
$$-\frac{1}{2}\mathrm{trace}\big[[b(\gamma)'\Phi b(\gamma) + \Omega(\gamma)]^{-1}[(Y_{(i)} - x_{(i)}\pi(\gamma,\iota))'(Y_{(i)} - x_{(i)}\pi(\gamma,\iota))]\big].$$

Then it is straightforward to show that

$$\max_{\upsilon} E_F[h(Y_{(1)}, x_{(1)}, \upsilon)]$$

is attained at υ^*, which is formed from the distinct elements of $(\gamma^*, \iota^*, \Phi^*)$. The quasi-ML estimator is

$$\hat{\upsilon}_N = \arg\max_{\upsilon} \frac{1}{N}\sum_{i=1}^{N} h(Y_{(i)}, x_{(i)}, \upsilon).$$

Standard method-of-moments arguments, as in Hansen (1982), MaCurdy (1982), and White (1982), provide regularity conditions under which $\hat{\upsilon}_N$ has a limiting normal distribution as $N \to \infty$ (with J, K, and M fixed):

$$\sqrt{N}(\hat{\upsilon}_N - \upsilon^*) \xrightarrow{d} \mathcal{N}(0, \Lambda^*),$$

where

$$\Lambda^* = [E_F \frac{\partial^2 h(Y_{(1)}, x_{(1)}, \upsilon^*)}{\partial \upsilon \partial \upsilon'}]^{-1} [E_F \frac{\partial h(Y_{(1)}, x_{(1)}, \upsilon^*)}{\partial \upsilon} \frac{\partial h(Y_{(1)}, x_{(1)}, \upsilon^*)}{\partial \upsilon'}][E_F \frac{\partial^2 h(Y_{(1)}, x_{(1)}, \upsilon^*)}{\partial \upsilon \partial \upsilon'}]^{-1}.$$

Since $\bar{f}_{\lambda,\kappa}(q'y \mid \gamma, \iota, \Phi)$ equals $f^{re}(y \mid \gamma, \iota, \Phi)$, this limit distribution result applies to a quasi maximum likelihood estimator based on the integrated likelihood $\bar{f}_{\lambda,\kappa}$ from Section 6.

The quasi maximum likelihood estimator is asymptotically equivalent to a minimum distance estimator that imposes the restrictions on the second moments. An optimal minimum distance estimator uses a weight matrix based on the covariance matrix of the sample second moments. The minimum distance estimator corresponding to quasi-ML uses a weight matrix that would be optimal under normality but not in general. See Chamberlain (1984, section 4.4) and Arellano (2003, sections 5.4.3 and 7.4.2).

Returning to the example in (14), we can use the reduced form from equations (2)–(4) in Section 1 to calculate moment conditions based on $T(Z)$, conditional on $\{y_{i0}, \alpha_i\}_{i=1}^N$. This gives

$$E(\bar{Y}) = N^{-1/2} E(Z_1) = \bar{\tau} b(\gamma),$$
$$E(Y'Y) = E(Z_1'Z_1 + Z_2'Z_2) = b(\gamma)'(N\bar{\tau}'\bar{\tau} + \tilde{\tau}_2'\tilde{\tau}_2) b(\gamma) + Nc(\gamma)'c(\gamma),$$

where $\gamma = (\psi, \sigma)$, $\tilde{\tau}'\tilde{\tau} = \rho^2$ is displayed in (15), and $b(\gamma)$ and $c(\gamma)$ are displayed in (4). These moments can be used to examine the identification of ψ and σ, treating $\bar{\tau}$ and $\tilde{\tau}_2'\tilde{\tau}_2$ as unrestricted.

Now add the normal random effects specification in (16), and calculate moment conditions without conditioning on $\{y_{i0}, \alpha_i\}_{i=1}^N$:

$$E(\bar{Y}) = \iota b(\gamma),$$
$$E(Y'Y) = b(\gamma)'(N\iota'\iota + N\Phi) b(\gamma) + Nc(\gamma)'c(\gamma).$$

Working with these moments leads to the same identification analysis for γ, because ι and Φ are unrestricted.

Bhargava and Sargan (1983) consider maximum likelihood estimation in a model with lagged dependent variables and strictly exogenous variables. They use a normal, correlated random effects model for the initial conditions. Their model is discussed in Arellano (2003, sections 7.4.1 and 7.4.2). Arellano (2003, section 7.4.3) considers a normal, correlated random effects specification for the individual effects in the Bhargava-Sargan model. Chamberlain (1980, p. 234–235) and Blundell and Smith (1991) consider maximum likelihood estimation, conditional on the first observation, in normal, correlated random effects models. Alvarez and Arellano (2003, section 3.5) obtain limiting results for inference in these models as N and $\bar{T}$ tend to infinity.[2] Alvarez and Arellano (2004) consider quasi maximum likelihood estimators in correlated random effects models, with a stress on allowing for time-series heteroskedasticity.

Lancaster (2002) deals with incidental parameters by first reparametrizing so that the information matrix is block diagonal, with the common parameters in one block and the incidental parameters in the other. In his application to a nonstationary dynamic regression model (p. 653), the parameter space for the reparametrized incidental parameters is $\mathcal{R}^N$. Then he forms an integrated likelihood function, integrating with respect to Lebesgue measure on $\mathcal{R}^N$. He shows that maximizing this integrated likelihood function provides a consistent estimator of the common parameters. Note that the information matrix block diagonality would be preserved by a smooth bijective transformation of the incidental parameters, so the use of Lebesgue measure does not by itself provide a unique prior measure. Our approach is similar in that it uses an integrated likelihood function. The prior measure, however, is different. Our reparametrization is motivated by the invariance of the model under the actions of the orthogonal group, and this determines a unique invariant distribution for ω on the compact space $\mathcal{F}_{K,N-J}$. This distribution is least favorable in our minimax optimality result. Another difference is that the use of Lebesgue measure on $\mathcal{R}^N$ for a prior measure does not correspond to the normal, correlated random effects model. It amounts to specifying that the (reparametrized) individual effects have very large variances, instead of treating the individual effects as draws from a distribution whose variance is a parameter to be estimated.

Sims (2000) uses a likelihood perspective in his analysis of dynamic panel data models. He deals

[2] Hahn and Kuersteiner (2002) consider maximum likelihood estimation in fixed effects models and obtain bias corrections as N and $\bar{T}$ tend to infinity.

with incidental parameters by treating the individual effects and initial conditions as draws from a bivariate normal distribution (p. 454). Our approach has a different starting point, since our model treats the individual effects and initial conditions as parameters (fixed effects). But our minimax optimality argument calls for a particular least favorable distribution for ω. We have seen that this unique distribution can be combined with a particular family of prior distributions for (δ, ρ) to obtain a normal, correlated random effects model, which corresponds to Sims's specification.

8. CONCLUSION

We started with a fixed-effects model. After reparametrizing, only the parameter ω has dimension depending on the cross-section sample size N. The model is invariant under the actions of the orthogonal group, and we obtained a maximal invariant statistic, T, whose distribution does not depend upon ω. So we can solve the incidental parameters problem by working with a marginal likelihood, based on the sampling distribution of T. This approach has a finite sample, minimax optimality. The argument is based on expressing the marginal likelihood as an integrated likelihood for a particular prior distribution for ω. The prior distribution is the unique, invariant distribution under the group action on that part of the parameter space.

In addition to ω, the nuisance parameter consists of (δ, ρ), whose dimension does not depend upon N. A convenient way to implement our approach is to use a particular family of prior distributions for (δ, ρ), indexed by the parameter (ι, Φ). This leads to an integrated likelihood function with a closed form expression. It is a function of (γ, ι, Φ), where γ is the original parameter of interest, which is not affected by the reparametrization. It turns out that this integrated likelihood function coincides with the likelihood function for a normal, correlated random effects model.

So our finite sample optimality arguments take us from the initial fixed-effects model to a normal, correlated random effects model. The normal distribution for the effects is not part of our model in equations (1) and (7); the model only specifies a normal distribution for the errors. The normal distribution for the effects arises from two sources: the unique uniform distribution for ω on the compact manifold $\mathcal{F}_{K,N-J}$, whose dimension depends upon N, and the convenient choice of prior distribution for (δ, ρ), whose dimension does not depend upon N. The first source is motivated by our invariance and minimax arguments. The second source lacks this motivation, but since the dimension of (δ, ρ) does not depend upon N, the particular choice made here may not be so important when N is large. In fact, using the integrated likelihood function as a quasi-likelihood, the large N asymptotics of the quasi-ML estimator are covered by standard arguments, under random sampling. These large N arguments do not require the assumption of normal errors in (1) and (7).

So one way to view our finite sample results is that, starting with a fixed-effects model, they provide motivation for a normal, correlated random effects model. At that point, robustness concerns can lead to dropping the normality assumption. Our quasi-ML estimator can still provide the basis for large N inference, but it would not be (semiparametric) efficient. So one may prefer to use a different weighting scheme for the moment restrictions implied by the correlated random effects model. This leads to standard optimal minimum distance and generalized method of moments estimators for the correlated random effects model.

Dept. of Economics, Harvard University, Cambridge, MA 02138, U.S.A.;
gary_chamberlain@harvard.edu

and

Dept. of Economics, Harvard University, Cambridge, MA 02138, U.S.A. and FGV/EPGE,
Rio de Janeiro, RJ 22250, Brazil; mjmoreira@columbia.edu

REFERENCES

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

Alvarez, J., and M. Arellano (2004): "Robust Likelihood Estimation of Dynamic Panel Data Models," Unpublished Manuscript, CEMFI.

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

Bhargava, A., and J.D. Sargan (1983): "Estimating Dynamic Random Effects Models from Panel Data Covering Short Time Periods," *Econometrica*, 51, 1635–1659.

Bishop, R., and R. Crittenden (1964): *Geometry of Manifolds*. Originally published: New York: Academic Press, 1964. Reprinted with corrections by the American Mathematical Society, 2001. Providence: AMS Chelsea Publishing.

Blundell, R., and R. Smith (1991): "Initial Conditions and Efficient Estimation in Dynamic Panel Data Models: An Application to Company Investment Behaviour," *Annales d'Économie et de Statistique*, 20/21, 109–123.

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

Chamberlain, G. (1984): "Panel Data," in *Handbook of Econometrics*, Volume II, ed. by Z. Griliches and M. Intriligator, 1247–1318. Amsterdam: North-Holland.

Chamberlain, G. (2007): "Decision Theory Applied to an Instrumental Variables Model," *Econometrica*, 75, 609–652.

Eaton, M. (1989): *Group Invariance Applications in Statistics*, Regional Conference Series in Probability and Statistics, Volume 1. Hayward, CA: Institute of Mathematical Statistics.

Golub, G. and C. Van Loan (1996): *Matrix Computations*, Third Edition. Baltimore: The Johns Hopkins University Press.

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

Hansen, L.P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

Lancaster, T. (2000): "The Incidental Parameter Problem Since 1948," *Journal of Econometrics*, 95, 391–413.

Lancaster, T. (2002): “Orthogonal Parameters and Panel Data,” *Review of Economic Studies*, 69, 647–666.

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

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

White, H. (1982): “Maximum Likelihood Estimation of Misspecified Models,” *Econometrica*, 50, 1–25.