CHAPTER 12

Understanding Bias in Nonlinear Panel Models: Some Recent Developments*

Manuel Arellano and Jinyong Hahn

1 INTRODUCTION

The purpose of this paper is to review recently developed bias-adjusted methods of estimation of nonlinear panel data models with fixed effects. Standard estimators such as maximum likelihood estimators are usually inconsistent if the number of individuals n goes to infinity while the number of time periods T is held fixed. For some models, like static linear and logit regressions, there exist fixed-T consistent estimators as $n \to \infty$ (see, e.g., Andersen, 1970). Fixed T consistency is a desirable property because for many panels T is much smaller than n. However, these type of estimators are not available in general, and when they are, their properties do not normally extend to estimates of average marginal effects, which are often parameters of interest. Moreover, without auxiliary assumptions, the common parameters of certain nonlinear fixed effects models are simply unidentified in a fixed T setting, so that fixed-T consistent point estimation is not possible (see, e.g., Chamberlain, 1992). In other cases, although identifiable, fixed-T consistent estimation at the standard root-n rate is impossible (see, e.g., Honoré and Kyriazidou, 2000; Hahn, 2001).

The number of periods available for many household, firm-level or country panels is such that it is not less natural to talk of time-series finite sample bias than of fixed-T inconsistency or underidentification. In this light, an alternative reaction to the fact that micro panels are short is to ask for approximately unbiased estimators as opposed to estimators with no bias at all. That is, estimators with biases of order $1/T^2$ as opposed to the standard magnitude of $1/T$. This alternative approach has the potential of overcoming some of the fixed-T identification difficulties and the advantage of generality.

The paper is organized as follows. Section 2 describes fixed effects estimators and the incidental parameters problem. Section 3 explains how to construct analytical bias correction of estimators. Section 4 describes bias correction of

* Prepared for the Econometric Society World Meetings, London, August 2005. We are grateful to Whitney Newey and Tiemen Woutersen for helpful comments on this and related work. The second author gratefully acknowledges financial support from NSF Grant SES-0313651.

the moment equation. Section 5 presents bias corrections for the concentrated likelihood. Section 6 discusses other approaches leading to bias correction, including Cox and Reid's and Lancaster's approaches based on orthogonality, and their extensions. Section 7 describes quasi maximum likelihood estimation for dynamic models. Section 8 considers the estimation of marginal effects. Section 9 discusses automatic methods based on simulation. Section 10 concludes.

2 INCIDENTAL PARAMETERS PROBLEM WITH LARGE T

We first describe fixed effects estimators. Let the data observations be denoted by $z_{it} = \left(y_{it}, x_{it}'\right)'$, $(t = 1, \ldots, T; i = 1, \ldots, n)$, where y_{it} denotes the "dependent" variable, and x_{it} denotes the strictly exogenous "explanatory" variable.[1] Let θ denote a parameter that is common to all i, α_i a scalar individual effect,[2] and $f(y_{i1}, \ldots, y_{iT} \mid \theta_0, \alpha_{i0})$

$$f(y_{i1}, \ldots, y_{iT} \mid \theta_0, \alpha_{i0}) = f(y_{i1}, \ldots, y_{iT} \mid x_{i1}, \ldots, x_{iT}, \theta_0, \alpha_{i0})$$

a density function of $y_{i1}, \ldots, y_{iT}$ conditional on the strictly exogenous explanatory variables $x_{i1}, \ldots, x_{iT}$. Assuming that y_{it} are independent across i and t, we obtain the log likelihood

$$\sum_{i=1}^{n} \sum_{t=1}^{T} \log f_{it}(y_{it} \mid \theta, \alpha_i),$$

where $f_{it}(y_{it} \mid \theta, \alpha_i)$ denotes the density of y_{it} conditional on $x_{i1}, \ldots, x_{iT}$. For notational simplicity, we will write f for f_{it} below. The fixed effects estimator is obtained by doing maximum likelihood treating each α_i as a parameter to be estimated. Concentrating out the α_i leads to the characterization

$$\widehat{\theta}_T \equiv \operatorname*{argmax}_{\theta} \sum_{i=1}^{n} \sum_{t=1}^{T} \log f(y_{it} \mid \theta, \widehat{\alpha}_i(\theta)),$$

$$\widehat{\alpha}_i(\theta) \equiv \operatorname*{argmax}_{\alpha} \sum_{t=1}^{T} \log f(y_{it} \mid \theta, \alpha).$$

Here the $\widehat{\alpha}_i(\theta)$ depends on the data only through the ith observation $z_{i1}, \ldots, z_{iT}$. Let

$$L(\theta) \equiv \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} E\left[\sum_{t=1}^{T} \log f(y_{it} \mid \theta, \widehat{\alpha}_i(\theta))\right].$$

It will follow from the usual extremum estimator properties (e.g., Amemiya, 1985) that as $n \to \infty$ with T fixed, $\widehat{\theta}_T = \theta_T + o_p(1)$, where $\theta_T \equiv \operatorname{argmax}_{\theta}$

[1] Throughout most of the paper except in Section 7, we will assume away dynamics or feedback.

[2] Our analysis extends easily, albeit with some notational complication, to the case where there are multiple fixed effects, that is, where α_i is a multidimensional vector.

$L(\theta)$. In general, $\theta_T \neq \theta_0$. This is the incidental parameters problem noted by Neyman and Scott (1948). The source of this problem is the estimation error of $\widehat{\alpha}_i(\theta)$. Because only a finite number T of observations are available to estimate each α_i, the estimation error of $\widehat{\alpha}_i(\theta)$ does not vanish as the sample size n grows, and this error contaminates the estimates of parameters of interest.

Example 1 *Consider a simple model where* $y_{it} \overset{i.i.d.}{\sim} \mathcal{N}(\alpha_{i0}, \sigma_0^2)$, $(t = 1, \ldots, T; i = 1, \ldots, n)$, *or*

$$\log f\left(y_{it}; \sigma^2, \alpha_i\right) = C - \frac{1}{2}\log\sigma^2 - \frac{(y_{it} - \alpha_i)^2}{2\sigma^2}.$$

This is a simpler version of the model considered by Chamberlain (1980). Here, we may write $\theta = \sigma^2$, *and the MLE is such that*

$$\widehat{\alpha}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it} \equiv \overline{y}_i, \quad \widehat{\theta} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(y_{it} - \overline{y}_i\right)^2.$$

It is straightforward to show that $\widehat{\theta} = \theta_0 - \frac{1}{T}\theta_0 + o_p(1)$ *as* $n \to \infty$ *with* T *fixed. In this example, the bias is easy to fix by equating the denominator with the correct degrees of freedom* $n(T-1)$.

Note that the bias should be small for large enough T, that is, $\lim_{T\to\infty}\theta_T = \theta_0$. Furthermore, for smooth likelihoods we usually have

$$\theta_T = \theta_0 + \frac{B}{T} + O\left(\frac{1}{T^2}\right) \tag{1}$$

for some B. In Example 1, $B = -\theta_0$. The fixed effects estimator $\widehat{\theta}$ will in general be asymptotically normal, although it will be centered at θ_T: as $n, T \to \infty$, $\sqrt{nT}(\widehat{\theta} - \theta_T) \xrightarrow{d} N(0, \Omega)$ for some Ω. Under these general conditions the fixed effects estimator is asymptotically biased even if T grows at the same rate as n. For $n/T \to \rho$, say,

$$\begin{aligned}\sqrt{nT}\left(\widehat{\theta} - \theta_0\right) &= \sqrt{nT}\left(\widehat{\theta} - \theta_T\right) + \sqrt{nT}\frac{B}{T} \\ &\quad + O\left(\sqrt{\frac{n}{T^3}}\right) \xrightarrow{d} N\left(B\sqrt{\rho}, \Omega\right).\end{aligned}$$

Thus, even when T grows as fast as n, asymptotic confidence intervals based on the fixed effects estimator will be incorrect, due to the limiting distribution of $\sqrt{nT}(\widehat{\theta} - \theta_0)$ not being centered at 0.

Similar to the bias of the fixed effects estimand $\theta_T - \theta_0$, the bias in the expected fixed effects score at θ_0 and the bias in the expected concentrated

likelihood at an arbitrary θ can also be expanded in orders of magnitude of T:

$$E\left[\frac{1}{T}\sum_{t=1}^{T}\frac{\partial}{\partial\theta}\log f\left(y_{it}\mid\theta_0,\widehat{\alpha}_i\left(\theta_0\right)\right)\right]=\frac{1}{T}b_i\left(\theta_0\right)+o\left(\frac{1}{T}\right) \tag{2}$$

and

$$\begin{aligned}E\left[\frac{1}{T}\sum_{t=1}^{T}\log f\left(y_{it}\mid\theta,\widehat{\alpha}_i\left(\theta\right)\right)-\frac{1}{T}\sum_{t=1}^{T}\log f\left(y_{it}\mid\theta,\overline{\alpha}_i\left(\theta\right)\right)\right]\\ =\frac{1}{T}\beta_i\left(\theta\right)+o\left(\frac{1}{T}\right)\end{aligned} \tag{3}$$

where $\overline{\alpha}_i(\theta)$ maximizes $\lim_{T\to\infty}E[T^{-1}\sum_{t=1}^{T}\log f(y_{it}\mid\theta,\alpha)]$. These expansions motivate alternative approaches to bias correction based on adjusting the estimator, the estimating equation, or the objective function. We next discuss these three approaches in turn. We shall refer to B/T, b_i/T, and β_i/T as the order $1/T$ biases of the fixed effects estimand, expected score, and expected concentrated likelihood, respectively.

3 BIAS CORRECTION OF THE ESTIMATOR

An analytical bias correction is to plug into the formula for B estimators of its unknown components to construct $\widehat{B}$, and then form a bias-corrected estimator

$$\widehat{\theta}^{1}\equiv\widehat{\theta}-\frac{\widehat{B}}{T}. \tag{4}$$

3.1 Formulae for the Order $1/T$ Bias

To implement this idea, we need to have an explicit formula for B. For this purpose, it is convenient to define

$$\begin{aligned}u_{it}\left(\theta,\alpha\right)&\equiv\frac{\partial}{\partial\theta}\log f\left(y_{it}|\theta,\alpha\right),\quad v_{it}\left(\theta,\alpha\right)\equiv\frac{\partial}{\partial\alpha_i}\log f\left(y_{it}|\theta,\alpha\right),\\ V_{2it}\left(\theta,\alpha\right)&\equiv v_{it}^{2}\left(\theta,\alpha\right)+\frac{\partial v_{it}\left(\theta,\alpha\right)}{\partial\alpha_i},\\ U_{it}\left(\theta,\alpha\right)&\equiv u_{it}\left(\theta,\alpha\right)-v_{it}\left(\theta,\alpha\right)E\left[v_{it}^{\alpha_i}\right]^{-1}E\left[u_{it}^{\alpha_i}\right],\\ \mathcal{I}_i&\equiv-E\left[\frac{\partial U_{it}\left(\theta_0,\alpha_{i0}\right)}{\partial\theta'}\right].\end{aligned}$$

Note that $E\left[U_{it}^{\alpha_i}\right]=0$, which in the MLE case implies that U_{it} and v_{it} are orthogonalized. We will denote the derivative with respect to θ or α_i by appropriate superscripts, for example, $U_{it}^{\alpha_i}\left(\theta,\alpha\right)\equiv\partial U_{it}\left(\theta,\alpha\right)/\partial\alpha_i$, $U_{it}^{\alpha_i\alpha_i}\left(\theta,\alpha\right)\equiv\partial^2U_{it}\left(\theta,\alpha\right)/\partial\alpha_i^2$. For notational convenience we suppress the arguments when

expressions are evaluated at the true values θ_0 and α_{i0}, for example $v_{it}^{\alpha_i} = \partial v_{it}(\theta_0, \alpha_{i0}) / \partial \alpha_i$.

It can be shown that

$$B = \left(\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathcal{I}_i \right)^{-1} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} b_i(\theta_0) \tag{5}$$

where $b_i(\theta_0)/T$ is the $1/T$ bias of the score function. It can also be shown that

$$b_i(\theta_0) = -\left(\frac{E\left[v_{it} U_{it}^{\alpha_i}\right]}{E\left[v_{it}^{\alpha_i}\right]} - \frac{E\left[U_{it}^{\alpha_i \alpha_i}\right] E\left[v_{it}^2\right]}{2\left(E\left[v_{it}^{\alpha_i}\right]\right)^2} \right). \tag{6}$$

or

$$b_i(\theta_0) = \left(\frac{-E\left[v_{it}^2\right]}{E\left[v_{it}^{\alpha_i}\right]} \right) \left[-\frac{1}{\left(-E\left[v_{it}^2\right]\right)} \left(E\left[v_{it} u_{it}^{\alpha_i}\right] - E\left[v_{it} v_{it}^{\alpha_i}\right] \frac{E\left[u_{it}^{\alpha_i}\right]}{E\left[v_{it}^{\alpha_i}\right]} \right) \right.$$
$$\left. -\frac{1}{2E\left[v_{it}^{\alpha_i}\right]} \left(E\left[u_{it}^{\alpha_i \alpha_i}\right] - E\left[v_{it}^{\alpha_i \alpha_i}\right] \frac{E\left[u_{it}^{\alpha_i}\right]}{E\left[v_{it}^{\alpha_i}\right]} \right) \right]. \tag{7}$$

Intuition on the derivation of the bias of the score function is provided in Section 4. See also Hahn and Newey (2004), for example. The bias correction formula (5) does not depend on the likelihood setting, and so would be valid for any fixed effects m-estimator.

However, in the likelihood setting because of the information identity $E[v_{it}^2] = -E[v_{it}^{\alpha_i}]$ and the Bartlett equality

$$E\left[v_{it} U_{it}^{\alpha_i}\right] + \frac{1}{2} E\left[U_{it}^{\alpha_i \alpha_i}\right] = -\frac{1}{2} E\left[V_{2it} U_{it}\right], \tag{8}$$

we can alternatively write

$$B = \frac{1}{2} \left(\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathcal{I}_i \right)^{-1} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{E\left[U_{it} V_{2it}\right]}{E\left[v_{it}^{\alpha_i}\right]}. \tag{9}$$

In Example 1 with $\theta = \sigma^2$, we can see that

$$u_{it} = -\frac{1}{2\theta_0} + \frac{(y_{it} - \alpha_i)^2}{2\theta_0^2}, \quad v_{it} = \frac{y_{it} - \alpha_{i0}}{\theta_0}, \quad E\left[v_{it}^{\alpha_i}\right] = -\frac{1}{\theta_0}$$
$$E\left[u_{it} v_{it}\right] = 0, \quad U_{it} = u_{it} = -\frac{1}{2\theta_0} + \frac{(y_{it} - \alpha_{i0})^2}{2\theta_0^2},$$
$$E\left[\mathcal{I}_i\right] = \frac{1}{2\theta_0^2}, \quad V_{2it} = \frac{(y_{it} - \alpha_{i0})^2}{\theta_0^2} - \frac{1}{\theta_0},$$
$$E\left[U_{it} V_{2it}\right] = \frac{1}{\theta_0^2}, \quad \frac{E\left[U_{it} V_{2it}\right]}{E\left[v_{it}^{\alpha_i}\right]} = -\frac{1}{\theta_0},$$
$$B = -\frac{1}{2} \left(\frac{1}{2\theta_0^2} \right)^{-1} \frac{1}{\theta_0} = -\theta_0,$$

and we obtain

$$\widehat{\theta}^{1} = \widehat{\theta} - \frac{\widehat{B}}{T} = \frac{T+1}{T}\widehat{\theta}.$$

Recall that $\widehat{\theta} = \theta_0 - \frac{1}{T}\theta_0 + o_p(1)$ as $n \to \infty$ with T fixed. It follows that

$$\widehat{\theta}^{1} = \theta_0 - \frac{1}{T^2}\theta_0 + o_p(1),$$

which shows that the bias of order T^{-1} is removed.

3.2 Estimators of the Bias

An estimator of the bias term can be formed using a sample counterpart of the previous formulae. One possibility is

$$\widehat{B}(\theta) = \left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\mathcal{I}}_i\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\widehat{b}_i(\theta) \tag{10}$$

where

$$\widehat{\mathcal{I}}_i = -\left(\widehat{E}_T\left[\widehat{u}_{it}^{\theta}\right] - \widehat{E}_T\left[\widehat{u}_{it}^{\alpha_i}\right]\widehat{E}_T\left[\widehat{v}_{it}^{\alpha_i}\right]^{-1}\widehat{E}_T\left[\widehat{u}_{it}^{\alpha_i\prime}\right]\right) \tag{11}$$

$$\begin{aligned}\widehat{b}_i(\theta) = &\left(\frac{-\widehat{E}_T\left[\widehat{v}_{it}^2\right]}{\widehat{E}_T\left[\widehat{v}_{it}^{\alpha_i}\right]}\right)\left[-\frac{1}{\left(-\widehat{E}_T\left[\widehat{v}_{it}^2\right]\right)}\left(\widehat{E}_T\left[\widehat{v}_{it}\widehat{u}_{it}^{\alpha_i}\right] - \widehat{E}_T\left[\widehat{v}_{it}\widehat{v}_{it}^{\alpha_i}\right]\frac{\widehat{E}_T\left[\widehat{u}_{it}^{\alpha_i}\right]}{\widehat{E}_T\left[\widehat{v}_{it}^{\alpha_i}\right]}\right)\right.\\ &\left.-\frac{1}{2\widehat{E}_T\left[\widehat{v}_{it}^{\alpha_i}\right]}\left(\widehat{E}_T\left[\widehat{u}_{it}^{\alpha_i\alpha_i}\right] - \widehat{E}_T\left[\widehat{v}_{it}^{\alpha_i\alpha_i}\right]\frac{\widehat{E}_T\left[\widehat{u}_{it}^{\alpha_i}\right]}{\widehat{E}_T\left[\widehat{v}_{it}^{\alpha_i}\right]}\right)\right]\end{aligned} \tag{12}$$

where $\widehat{E}_T(.) = \sum_{t=1}^{T}(.)/T$, $\widehat{u}_{it}^{\theta} = u_{it}^{\theta}(\theta, \widehat{\alpha}_i(\theta))$, $\widehat{u}_{it}^{\alpha_i} = u_{it}^{\alpha_i}(\theta, \widehat{\alpha}_i(\theta))$, etc. The bias corrected estimator can then be formed with $\widehat{B} = \widehat{B}\left(\widehat{\theta}_T\right)$.

The other possibility exploits the likelihood setting to replace some derivatives by outer product terms:

$$\widetilde{B}(\theta) = \left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathcal{I}}_i\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\widetilde{b}_i(\theta) \tag{13}$$

where

$$\begin{aligned}\widetilde{\mathcal{I}}_i &= -\left(\widehat{E}_T\left[\widehat{u}_{it}\widehat{u}_{it}'\right] - \widehat{E}_T\left[\widehat{u}_{it}\widehat{v}_{it}\right]\widehat{E}_T\left[\widehat{v}_{it}^2\right]^{-1}\widehat{E}_T\left[\widehat{v}_{it}\widehat{u}_{it}'\right]\right)\\ &= -\widehat{E}_T\left(\widehat{U}_{it}\widehat{U}_{it}'\right),\end{aligned} \tag{14}$$

$$\widetilde{b}_i(\theta) = \frac{\sum_{t=1}^{T}\widehat{U}_{it}(\theta, \widehat{\alpha}_i(\theta))V_{2it}(\theta, \widehat{\alpha}_i(\theta))}{2\sum_{t=1}^{T}v_{it}^{\alpha_i}(\theta, \widehat{\alpha}_i(\theta))}, \tag{15}$$

and

$$\widehat{U}_{it} \equiv \widehat{U}_{it}\left(\theta, \widehat{\alpha}_i(\theta)\right) = u_{it}\left(\theta, \widehat{\alpha}_i(\theta)\right) - \frac{\widehat{E}_T\left[\widehat{u}_{it}\widehat{v}_{it}\right]}{\widehat{E}_T\left[\widehat{v}_{it}^2\right]} v_{it}\left(\theta, \widehat{\alpha}_i(\theta)\right), \tag{16}$$

so that an alternative bias correction can be formed with $\widetilde{B} = \widetilde{B}\left(\widehat{\theta}_T\right)$.

3.3 Infinitely Iterated Analytic Bias Correction

If $\widehat{\theta}$ is heavily biased and it is used in the construction of $\widehat{B}$, it may adversely affect the properties of $\widehat{\theta}^1$. One way to deal with this problem is to use $\widehat{\theta}^1$ in the construction of another $\widehat{B}$, and then form a new bias corrected estimator as in equation (4). One could even iterate this procedure, updating $\widehat{B}$ several times using the previous estimator of $\widehat{\theta}$. To be precise, let $\overline{B}(\theta)$ denote an estimator of B depending on θ, and suppose that $\widehat{B} = \overline{B}(\widehat{\theta})$. Then $\widehat{\theta}^1 = \widehat{\theta} - \overline{B}(\widehat{\theta})/T$. Iterating gives $\widehat{\theta}^k = \widehat{\theta} - \overline{B}(\widehat{\theta}^{k-1})/T$, $(k = 2, 3, \ldots)$. If this estimator were iterated to convergence, it would give $\widehat{\theta}^\infty$ solving

$$\widehat{\theta}^\infty = \widehat{\theta} - \overline{B}\left(\widehat{\theta}^\infty\right)/T. \tag{17}$$

In general this estimator will not have improved asymptotic properties, but may have lower bias for small T. In Example 1 with $\theta_0 = \sigma_0^2$, we can see that

$$\widehat{\theta}^k = \frac{T^k + T^{k-1} + \ldots + 1}{T^k}\widehat{\theta} = \frac{T^{k+1} - 1}{T^k(T-1)}\widehat{\theta} \to \frac{T}{T-1}\widehat{\theta} = \widehat{\theta}^\infty$$

as $k \to \infty$, and the limit $\widehat{\theta}^\infty$ has zero bias.

4 BIAS CORRECTION OF THE MOMENT EQUATION

Another approach to bias correction for fixed effects is to construct the estimator as the solution to a bias-corrected version of the first-order conditions. Recall that the expected fixed effects score has the $1/T$ bias equal to $b_i(\theta_0)$ at the true value, as noted in (2). Let us consider $\widehat{S}(\theta) = \sum_{i=1}^n \sum_{t=1}^T u_{it}(\theta, \widehat{\alpha}_i(\theta))/(nT)$, so that the fixed effects estimator solves $\widehat{S}(\widehat{\theta}_T) = 0$, and let $\widehat{b}_i(\theta)/T$ be an estimator of the $1/T$ bias of the expected score at the true value. A score-corrected estimator is obtained by solving the modified moment equation

$$\widehat{S}(\theta) - \frac{1}{nT}\sum_{i=1}^n \widehat{b}_i(\theta) = 0. \tag{18}$$

To understand the idea of correcting the moment equation and its connection to estimating B, it is convenient to note that the MLE $\widehat{\theta}$ is a solution to

$$\sum_{i=1}^{n}\sum_{t=1}^{T} u_{it}\left(\widehat{\theta},\widehat{\alpha}_i\right)=0.$$

Consider an infeasible estimator $\overline{\theta}$ based on $\widehat{\alpha}_i(\theta_0)$ rather than $\widehat{\alpha}_i$, where $\overline{\theta}$ solves the first-order condition $0=\sum_{i=1}^{n}\sum_{t=1}^{T} U_{it}\left(\overline{\theta},\widehat{\alpha}_i(\theta_0)\right)$. Standard arguments suggest that

$$\sqrt{nT}\left(\overline{\theta}-\theta_0\right)\approx\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i\right)^{-1}\frac{1}{\sqrt{nT}}\sum_{i=1}^{n}\sum_{t=1}^{T} U_{it}\left(\theta_0,\widehat{\alpha}_i(\theta_0)\right).$$

Because $E\left[U_{it}\left(\theta_0,\widehat{\alpha}_i(\theta_0)\right)\right]\neq 0$, we cannot apply the central limit theorem to the numerator on the right side. We use a second-order Taylor series expansion to approximate $U_{it}\left(\theta_0,\widehat{\alpha}_i(\theta_0)\right)$ around α_{i0}:

$$\begin{aligned}\frac{1}{\sqrt{nT}}\sum_{i=1}^{n}\sum_{t=1}^{T} U_{it}\left(\theta_0,\widehat{\alpha}_i(\theta_0)\right)&\approx\frac{1}{\sqrt{nT}}\sum_{i=1}^{n}\sum_{t=1}^{T} U_{it}\\&+\frac{1}{\sqrt{nT}}\sum_{i=1}^{n}\sum_{t=1}^{T} U_{it}^{\alpha_i}\left(\widehat{\alpha}_i(\theta_0)-\alpha_{i0}\right)\\&+\frac{1}{2\sqrt{nT}}\sum_{i=1}^{n}\sum_{t=1}^{T} U_{it}^{\alpha_i\alpha_i}\left(\widehat{\alpha}_i(\theta_0)-\alpha_{i0}\right)^2.\end{aligned}$$

The first term on the right will follow a central limit theorem because $E[U_{it}]=0$. As for the second and third terms, we note that $\widehat{\alpha}_i(\theta_0)-\alpha_{i0}\approx -T^{-1}\sum_{t=1}^{T} v_{it}(E[v_{it}^{\alpha_i}])^{-1}$, and substituting for $\widehat{\alpha}_i(\theta_0)-\alpha_{i0}$ in the approximation for $U_{it}(\theta_0,\widehat{\alpha}_i(\theta_0))$ leads to

$$\begin{aligned}\sum_{i=1}^{n}\sum_{t=1}^{T} U_{it}\left(\theta_0,\widehat{\alpha}_i(\theta_0)\right)&\approx\sum_{i=1}^{n}\sum_{t=1}^{T} U_{it}\\&-\sum_{i=1}^{n}\left[\frac{\sum_{t=1}^{T} v_{it}}{\sqrt{T}E\left[v_{it}^{\alpha_i}\right]}\right]\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(U_{it}^{\alpha_i}-\frac{E\left[U_{it}^{\alpha_i\alpha_i}\right]}{2E\left[v_{it}^{\alpha_i}\right]}v_{it}\right)\right].\end{aligned}\tag{19}$$

Taking an expectation of the second term on the right and subtracting it from the LHS, we expect that

$$\begin{aligned}&\sum_{i=1}^{n}\sum_{t=1}^{T} U_{it}\left(\theta_0,\widehat{\alpha}_i(\theta_0)\right)+\sum_{i=1}^{n}\left(\frac{E\left[v_{it}U_{it}^{\alpha_i}\right]}{E\left[v_{it}^{\alpha_i}\right]}-\frac{E\left[U_{it}^{\alpha_i\alpha_i}\right]E\left[v_{it}^2\right]}{2\left(E\left[v_{it}^{\alpha_i}\right]\right)^2}\right)\\&\quad=\sum_{i=1}^{n}\sum_{t=1}^{T} U_{it}\left(\theta_0,\widehat{\alpha}_i(\theta_0)\right)-\sum_{i=1}^{n} b_i(\theta_0)\end{aligned}$$

is more centered at zero than $\sum_{i=1}^{n}\sum_{t=1}^{T} U_{it}(\theta_0,\widehat{\alpha}_i(\theta_0))$.

An estimator of the $1/T$ bias of the moment equation is given by $\widehat{b}_i(\theta)/T$ in (12). We then expect the solution to

$$\sum_{i=1}^{n}\left[\sum_{t=1}^{T} u_{it}\left(\theta, \widehat{\alpha}_i(\theta)\right) - \widehat{b}_i(\theta)\right] = 0 \tag{20}$$

to be less biased than the MLE $\widehat{\theta}_T$. Alternatively, the bias can be estimated using the estimator of the bias in (15) that exploits Bartlett identities, leading to the moment equation

$$\sum_{i=1}^{n}\left[\sum_{t=1}^{T} u_{it}\left(\theta, \widehat{\alpha}_i(\theta)\right) - \widetilde{b}_i(\theta)\right] = 0. \tag{21}$$

The first expression would be valid for any fixed effects m-estimator, whereas the second is appropriate in a likelihood setting. These two versions of bias-corrected moment equation are discussed in Hahn and Newey (2004).

In a likelihood setting it is also possible to form an estimate of $b_i(\theta)$ that uses expected rather than observed quantities, giving rise to alternative score-corrected estimators, such as those considered by Carro (2004) and Fernández-Val (2005) for binary choice models. To see a connection between bias correction of the moment equation and iterated bias correction of the estimator, it is useful to note that $\widehat{\theta}^{\infty}$ solves the equation $\widehat{\theta} - \theta = \overline{B}(\theta)/T$ or

$$\sum_{i=1}^{n}\left[\overline{\mathcal{I}}_i(\theta)\left(\widehat{\theta} - \theta\right) - \frac{1}{T}\overline{b}_i(\theta)\right] = 0 \tag{22}$$

where $\overline{B}(\theta)$ is as in (10) or (13). This equation can be regarded as an approximation to the previous corrected moment equations as long as $\overline{\mathcal{I}}_i(\theta)$ is an estimator of $\partial \widehat{E}_T\left[u_{it}\left(\theta, \widehat{\alpha}_i(\theta)\right)\right]/\partial\theta$ and $\overline{b}_i(\theta)/T$ is an estimator of the $1/T$ bias for $\widehat{E}_T\left[u_{it}\left(\theta, \widehat{\alpha}_i(\theta)\right)\right]$. Thus, the bias correction of the moment equation can be loosely understood to be an infinitely iterated bias correction of the estimator.

5 BIAS CORRECTION OF THE CONCENTRATED LIKELIHOOD

Because of the noise of estimating $\widehat{\alpha}_i(\theta)$, the expectation of the concentrated likelihood is not maximized at the true value of the parameter [see (3)]. In this section, we discuss how such problem can be avoided by correcting the concentrated likelihood.

Let $\ell_i(\theta, \alpha) = \sum_{t=1}^{T} \ell_{it}(\theta, \alpha)/T$ where $\ell_{it}(\theta, \alpha) = \log f(y_{it} \mid \theta, \alpha)$ denotes the log likelihood of one observation. Moreover, let $\overline{\alpha}_i(\theta) = \operatorname{argmax}_{\alpha} \operatorname{plim}_{T\to\infty} \ell_i(\theta, \alpha)$, so that under regularity conditions $\overline{\alpha}_i(\theta_0) = \alpha_{i0}$. Following Severini (2000) and Pace and Salvan (2006), the concentrated log likelihood for unit i

$$\widehat{\ell}_i(\theta) = \ell_i\left(\theta, \widehat{\alpha}_i(\theta)\right) \tag{23}$$

can be regarded as an estimate of the unfeasible concentrated log likelihood

$$\overline{\ell}_i(\theta) = \ell_i(\theta, \overline{\alpha}_i(\theta)). \tag{24}$$

The function $\overline{\ell}_i(\theta)$ is a proper log likelihood which assigns data a density of occurrence according to values of θ and values of the effects along the curve $\overline{\alpha}_i(\theta)$. It is a least-favorable target log likelihood in the sense that the expected information for θ calculated from $\overline{\ell}_i(\theta)$ coincides with the partial expected information for θ (c.f. Stein, 1956; Severini and Wong, 1992; and Newey, 1990, for related discussion on semiparametric bounds). $\overline{\ell}_i(\theta)$ has the usual log likelihood properties: it has zero mean expected score, it satisfies the information matrix identity, and is maximized at θ_0.

Now, define

$$H_i(\theta) = -E\left[\frac{\partial v_{it}(\theta, \overline{\alpha}_i(\theta))}{\partial \alpha}\right], \quad \Upsilon_i(\theta) = E\left\{[v_{it}(\theta, \overline{\alpha}_i(\theta))]^2\right\}.$$

A stochastic expansion for an arbitrary fixed θ gives

$$\widehat{\alpha}_i(\theta) - \overline{\alpha}_i(\theta) \approx H_i^{-1}(\theta)\, v_i(\theta, \overline{\alpha}_i(\theta)) \tag{25}$$

where $v_i(\theta, \alpha) = \sum_{t=1}^{T} v_{it}(\theta, \alpha)/T$. Next, expanding $\ell_i(\theta, \widehat{\alpha}_i(\theta))$ around $\overline{\alpha}_i(\theta)$ for fixed θ, we get

$$\begin{aligned}\ell_i(\theta, \widehat{\alpha}_i(\theta)) - \ell_i(\theta, \overline{\alpha}_i(\theta)) \approx\ & v_i(\theta, \overline{\alpha}_i(\theta))[\widehat{\alpha}_i(\theta) - \overline{\alpha}_i(\theta)] \\ & - \frac{1}{2} H_i(\theta)[\widehat{\alpha}_i(\theta) - \overline{\alpha}_i(\theta)]^2.\end{aligned} \tag{26}$$

Substituting (25) we get

$$\ell_i(\theta, \widehat{\alpha}_i(\theta)) - \ell_i(\theta, \overline{\alpha}_i(\theta)) \approx \frac{1}{2} H_i(\theta)[\widehat{\alpha}_i(\theta) - \overline{\alpha}_i(\theta)]^2. \tag{27}$$

Taking expectations, we obtain

$$E[\ell_i(\theta, \widehat{\alpha}_i(\theta)) - \ell_i(\theta, \overline{\alpha}_i(\theta))] \approx \frac{1}{2} H_i(\theta)\, Var[\widehat{\alpha}_i(\theta)] \approx \frac{\beta_i(\theta)}{T}$$

where

$$\beta_i(\theta) = \frac{1}{2} H_i(\theta)\, Var\left(\sqrt{T}[\widehat{\alpha}_i(\theta) - \overline{\alpha}_i(\theta)]\right) = \frac{1}{2} H_i^{-1}(\theta)\, \Upsilon_i(\theta). \tag{28}$$

Thus, we expect that

$$\sum_{i=1}^{n} \sum_{t=1}^{T} \ell_{it}(\theta, \widehat{\alpha}_i(\theta)) - \sum_{i=1}^{n} \beta_i(\theta)$$

is a closer approximation to the target log likelihood than $\sum_{i=1}^{n}\sum_{t=1}^{T}$ $\ell_{it}(\theta, \widehat{\alpha}_i(\theta))$. Letting $\widehat{\beta}_i(\theta)$ be an estimated bias,we then expect an estimator $\widetilde{\theta}$ that solves

$$\widetilde{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} \left[\sum_{t=1}^{T} \ell_{it}(\theta, \widehat{\alpha}_i(\theta)) - \widehat{\beta}_i(\theta) \right] \tag{29}$$

to be less biased than the MLE $\widehat{\theta}_T$.

We can consistently estimate $\beta_i(\theta)$ by

$$\widehat{\beta}_i(\theta) = \frac{1}{2} \left(-\frac{1}{T} \sum_{t=1}^{T} \frac{\partial v_{it}(\theta, \widehat{\alpha}_i(\theta))}{\partial \alpha} \right)^{-1} \frac{1}{T} \sum_{t=1}^{T} [v_{it}(\theta, \widehat{\alpha}_i(\theta))]^2 . \tag{30}$$

Using this form of $\widehat{\beta}_i(\theta)$ in (29), $\widetilde{\theta}$ solves the first-order conditions

$$\sum_{i=1}^{n}\sum_{t=1}^{T} u_{it}(\theta, \widehat{\alpha}_i(\theta)) - \sum_{i=1}^{n} \frac{\partial \widehat{\beta}_i(\theta)}{\partial \theta} = 0. \tag{31}$$

Because $\widehat{\alpha}_i(\theta)$ satisfies

$$0 = \sum_{t=1}^{T} v_{it}(\theta, \widehat{\alpha}_i(\theta)), \tag{32}$$

we can obtain

$$\frac{\partial \widehat{\alpha}_i(\theta)}{\partial \theta} = -\frac{\sum_{t=1}^{T} v_{it}^{\theta}(\theta, \widehat{\alpha}_i(\theta))}{\sum_{t=1}^{T} v_{it}^{\alpha_i}(\theta, \widehat{\alpha}_i(\theta))}. \tag{33}$$

Using this equation and the fact $v_{it}^{\theta} = u_{it}^{\alpha_i}$, it follows that

$$\frac{\partial \widehat{\beta}_i(\theta)}{\partial \theta} = \widehat{b}_i(\theta) \tag{34}$$

where $\widehat{b}_i(\theta)$ corresponds to the estimated score bias in (12). Therefore, the first-order conditions from (29) and the bias corrected moment (20) are identical.

Moreover, in the likelihood context, we can consider a local version of the estimated bias constructed as an expansion of $\widehat{\beta}_i(\theta)$ at θ_0 using that at the truth $H_i^{-1}(\theta_0)\Upsilon_i(\theta_0) = 1$ (Pace and Salvan, 2006):

$$\widehat{\beta}_i(\theta) = \widetilde{\beta}_i(\theta) + O\left(\frac{1}{T}\right) \tag{35}$$

where

$$\widetilde{\beta}_i(\theta) = -\frac{1}{2}\log\left(-\frac{1}{T}\sum_{t=1}^{T}\frac{\partial v_{it}(\theta, \widehat{\alpha}_i(\theta))}{\partial \alpha}\right) + \frac{1}{2}\log\left\{\frac{1}{T}\sum_{t=1}^{T}[v_{it}(\theta, \widehat{\alpha}_i(\theta))]^2\right\}. \tag{36}$$

This form of the estimated bias leads to the modified concentrated likelihood

$$\ell_i(\theta, \widehat{\alpha}_i(\theta)) + \frac{1}{2}\log\left\{-\frac{1}{T}\sum_{t=1}^{T}\left[\frac{\partial v_{it}(\theta, \widehat{\alpha}_i(\theta))}{\partial \alpha}\right]\right\} - \frac{1}{2}\log\left\{\frac{1}{T}\sum_{t=1}^{T}[v_{it}(\theta, \widehat{\alpha}_i(\theta))]^2\right\}. \tag{37}$$

This adjustment was considered by DiCiccio and Stern (1993) and DiCiccio et al. (1996). They showed that (37) reduces the bias of the concentrated score to $O(1/T)$ in the likelihood setting. In fact, it can be shown that (37) is maximized at $\frac{1}{n(T-1)}\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it} - \overline{y}_i)^2$ in Example 1.

It can be easily shown that

$$\frac{\partial \widetilde{\beta}_i(\theta)}{\partial \theta} = \frac{\widehat{E}_T\left[\widehat{v}_{it}^{\alpha_i}\right]}{\left(-\widehat{E}_T\left[\widehat{v}_{it}^2\right]\right)}\widehat{b}_i(\theta). \tag{38}$$

Therefore, the DiCiccio–Stern first-order condition is using a valid estimate of the concentrated score $1/T$ bias as long as the information identity holds, so that in general it will be appropriate in likelihood settings. Note that $\partial\widetilde{\beta}_i(\theta)/\partial\theta$ differs from $\widetilde{b}_i(\theta)$ in (15), which exploits Bartlett identities as well as the information equality.

In the likelihood setting it is also possible to form estimates of $H_i(\theta)$ and $\Upsilon_i(\theta)$ that use expected rather than observed quantities. An estimator of the bias of the form of (36) that uses the observed Hessian but an expectation-based estimate of the outer product term $\Upsilon_i(\theta)$ is closely related to Severini's (1998) approximation to the modified profile likelihood. Severini (2002) extends his earlier results to pseudo-ML estimation problems, and Sartori (2003) considers double asymptotic properties of modified concentrated likelihoods in the context of independent panel or stratified data with fixed effects.

6 OTHER APPROACHES LEADING TO BIAS CORRECTION

The incidental parameters problem in panel data models can be broadly viewed as a problem of inference in the presence of many nuisance parameters. The leading statistical approach under this circumstance has been to search for suitable modification of conditional or marginal likelihoods. The modified profile

likelihood of Barndorff-Nielsen (1983) and the approximate conditional likelihood of Cox and Reid (1987) belong to this category [see Reid (1995) for an overview]. However, the Barndorff-Nielsen formula is not generally operational, and the one in Cox and Reid requires the availability of an orthogonal effect.

We begin with discussion of Cox and Reid's (1987) adjustment to the concentrated likelihood followed by Lancaster's (2002) proposal.

6.1 Approaches Based on Orthogonality

6.1.1 Cox and Reid's Adjusted Profile Likelihood Approach

Cox and Reid (1987) considered the general problem of inference for a parameter of interest in the presence of nuisance parameters. They proposed a first-order adjustment to the concentrated likelihood to take account of the estimation of the nuisance parameters.

Their formulation required information orthogonality between the two types of parameters. That is, that the information matrix be block diagonal between the parameters of interest and the nuisance parameters. Suppose that the individual likelihood is given by $\prod_{t=1}^{T} f(y_{it} \mid \theta, \alpha_i)$. In general, the information matrix for (θ, α_i) will not be block-diagonal, although it may be possible to reparameterize α_i as a function of θ and some η_i such that the information matrix for (θ, η_i) is block-diagonal (Cox and Reid explained how to construct orthogonal parameters).

The discussion of orthogonality in the context of panel data models is due to Lancaster (2000, 2002), together with a Bayesian proposal that we consider below. The nature of the adjustment in a fixed effects model and some examples were also discussed in Cox and Reid (1992).

In the panel context, the Cox–Reid (1987) approach maximizes

$$\sum_{i=1}^{n}\sum_{t=1}^{T} \ell_{it}\left(y_{it}; \theta, \widehat{\alpha}_i(\theta)\right) - \frac{1}{2}\sum_{i=1}^{n} \log\left(-\sum_{t=1}^{T} \frac{\partial^2 \ell_{it}\left(y_{it}; \theta, \widehat{\alpha}_i(\theta)\right)}{\partial \alpha_i^2}\right). \tag{39}$$

The adjusted profile likelihood function (39) was derived by Cox and Reid as an approximation to the conditional likelihood given $\widehat{\alpha}_i(\theta)$. Their approach was motivated by the fact that in an exponential family model, it is optimal to condition on sufficient statistics for the nuisance parameters, and these can be regarded as the MLE of nuisance parameters chosen in a form to be orthogonal to the parameters of interest. For more general problems the idea was to derive a concentrated likelihood for θ conditioned on the MLE $\widehat{\alpha}_i(\theta)$, having ensured via orthogonality that $\widehat{\alpha}_i(\theta)$ changes slowly with θ.

6.1.1.1 Relation to Bias Correction of the Moment Equation. It is useful to spell out the first-order condition corresponding to the adjusted profile likelihood:

$$0 = \sum_{i=1}^{n} \left[\sum_{t=1}^{T} u_{it}\left(\theta, \widehat{\alpha}_i\left(\theta\right)\right) - \frac{1}{2} \frac{\sum_{t=1}^{T} u_{it}^{\alpha_i \alpha_i}\left(\theta, \widehat{\alpha}_i\left(\theta\right)\right)}{\sum_{t=1}^{T} v_{it}^{\alpha_i}\left(\theta, \widehat{\alpha}_i\left(\theta\right)\right)} - \frac{1}{2} \frac{\sum_{t=1}^{T} v_{it}^{\alpha_i \alpha_i}\left(\theta, \widehat{\alpha}_i\left(\theta\right)\right) \dfrac{\partial \widehat{\alpha}_i\left(\theta\right)}{\partial \theta}}{\sum_{t=1}^{T} v_{it}^{\alpha_i}\left(\theta, \widehat{\alpha}_i\left(\theta\right)\right)} \right] \tag{40}$$

where we used the fact $v_{it}^{\theta} = u_{it}^{\alpha_i}$. Moreover, using equations (32) and (33), we obtain that the moment equation of the adjusted profile likelihood is equal to

$$\sum_{i=1}^{n} \left[\sum_{t=1}^{T} u_{it}\left(\theta, \widehat{\alpha}_i\left(\theta\right)\right) - \widetilde{b}_i^{CR}\left(\theta\right) \right] = 0 \tag{41}$$

where

$$\widetilde{b}_i^{CR}\left(\theta\right) = \frac{1}{2} \frac{\widehat{E}_T\left[\widehat{u}^{\alpha_i \alpha_i}\right]}{\widehat{E}_T\left[\widehat{v}_{it}^{\alpha_i}\right]} - \frac{1}{2} \frac{\widehat{E}_T\left[\widehat{v}_{it}^{\alpha_i \alpha_i}\right] \widehat{E}_T\left[\widehat{u}_{it}^{\alpha_i}\right]}{\left(\widehat{E}_T\left[\widehat{v}_{it}^{\alpha_i}\right]\right)^2}. \tag{42}$$

Ferguson, Reid, and Cox (1991) showed that under orthogonality the expected moment equation has a bias of a smaller order of magnitude than the standard expected ML score.

Under information orthogonality $E\left[u_{it}^{\alpha_i}\right] = 0$ and $E\left[v_{it} u_{it}^{\alpha_i}\right] = -E\left[u_{it}^{\alpha_i \alpha_i}\right]$. Using these facts and the information identity, the bias formula (7) becomes

$$b_i\left(\theta_0\right) = \frac{1}{2} \frac{E\left[u_{it}^{\alpha_i \alpha_i}\right]}{E\left[v_{it}^{\alpha_i}\right]}. \tag{43}$$

Comparison with the Cox–Reid moment equation adjustment $\widetilde{b}_i^{CR}\left(\theta\right)$ reveals that the latter has an extra term whose population counterpart is equal to zero under orthogonality. It can in fact be shown that this term does not contribute anything to the asymptotic distribution of the resultant estimator under the large n large T asymptotics.

6.1.1.2 Relation to Bias Correction of the Concentrated Likelihood. To see the connection between the Cox–Reid's adjustment, which requires orthogonalization, and the one derived from the bias-reduction perspective in the previous section, which does not, note that (37) can be written as

$$\ell_i\left(\theta, \widehat{\alpha}_i\left(\theta\right)\right) - \frac{1}{2} \log\left\{ -\frac{1}{T} \sum_{t=1}^{T} \left[\frac{\partial v_{it}\left(\theta, \widehat{\alpha}_i\left(\theta\right)\right)}{\partial \alpha} \right] \right\} - \frac{1}{2} \log \widehat{\mathrm{Var}}\left(\sqrt{T}\left(\widehat{\alpha}_i\left(\theta\right) - \overline{\alpha}_i\left(\theta\right)\right) \right) \tag{44}$$

where

$$\widehat{\mathrm{Var}}\left(\sqrt{T}\left(\widehat{\alpha}_i(\theta)-\overline{\alpha}_i(\theta)\right)\right)=\frac{T\sum_{t=1}^{T}\left[v_{it}\left(\theta,\widehat{\alpha}_i(\theta)\right)\right]^2}{\left(\sum_{t=1}^{T}\left[v_{it}^{\alpha_i}\left(\theta,\widehat{\alpha}_i(\theta)\right)\right]\right)^2}. \tag{45}$$

Thus, a criterion of the form (44) can be regarded as a generalized Cox–Reid adjusted likelihood with an extra term given by an estimate of the variance of $\sqrt{T}\left(\widehat{\alpha}_i(\theta)-\overline{\alpha}_i(\theta)\right)$, which accounts for nonorthogonality (the discussion of this link is due to Pace and Salvan, 2006). Under orthogonality the extra term is irrelevant because the variance of $\widehat{\alpha}_i(\theta)$ does not change much with θ.

6.1.1.3 Other Features of Adjusted Likelihood Approach. We note that Cox and Reid's (1987) proposal and other methods in the same literature, were not developed to explicitly address the incidental parameter problem in the panel data context. Rather, they were concerned with inference in models with many nuisance parameters.

We also note that this class of approaches was not developed for the sole purpose of correcting for the bias of the resultant estimator. It was developed with the ambitious goal of making the modified concentrated likelihood behave like a proper likelihood, including the goal of stabilizing the behavior of the likelihood ratio statistic. We can see that it achieves some of these other goals at least in the context of Example 1, where it can be shown that

$$\widehat{\theta}=\frac{1}{n(T-1)}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(y_{it}-\overline{y}_i\right)^2$$

maximizes (39), and the second derivative of (39) delivers $\frac{2\theta^2}{n(T-1)}$ as the estimated variance of $\widehat{\theta}$. Because the actual variance of $\widehat{\theta}$ is equal to $\frac{2\theta^2}{n(T-1)}$, we can note that the Cox–Reid approach even takes care of the problem of correctly estimating the variance of the estimator. It is not clear whether such success is specific to the particular example, or not. More complete analysis of other aspects of inference such as variance estimation is beyond the scope of this survey.

6.1.2 Lancaster's (2002) Bayesian Inference

Lancaster (2002) proposed a method of Bayesian inference that is robust to the incidental parameters problem, which like Cox and Reid's method critically hinges on the availability of parameter orthogonality, which may not be feasible in many applications. Sweeting (1987) pointed out that such procedure is in fact approximately Bayesian. These approaches have been later generalized by Woutersen (2002) and Arellano (2003) to situations where orthogonality may not be available. Their generalizations are based on correcting the first-order condition of the adjusted profile likelihood estimator, and will be discussed in the next section.

In a Bayesian setting, fixed effects are integrated out of the likelihood with respect to the prior distribution conditional on the common parameters (and

covariates, if present) $\pi(\alpha \mid \theta)$. In this way, we get an integrated (or random effects) log likelihood of the form

$$\ell_i^I(\theta) = \log \int e^{T\ell_i(\theta,\alpha)} \pi(\alpha \mid \theta)\, d\alpha.$$

As is well known, the problem with inferences from $\ell_i^I(\theta)$ is that they depend on the choice of prior for the effects and are not in general consistent with T fixed. It can be shown that under regularity conditions the maximizer of $\sum_i \ell_i^I(\theta)$ has a bias of order $O(1/T)$ regardless of $\pi(\alpha \mid \theta)$. However, if α and θ are information orthogonal, the bias can be reduced to $O\left(1/T^2\right)$.

Lancaster (2002) proposes to integrate out the fixed effects η_i by using a noninformative prior, say a uniform prior, and use the posterior mode as an estimate of θ. The idea is to rely on prior independence between fixed effects and θ, having chosen an orthogonal reparameterization, say $\alpha_i = \alpha\left(\theta, \eta_i\right)$, that separates the common parameter θ from the fixed effects η_i in the information matrix sense. In other words, his estimator $\widehat{\theta}_L$ takes the form

$$\widehat{\theta}_L = \underset{\theta}{\operatorname{argmax}} \int \cdots \int \prod_{i=1}^{n} \prod_{t=1}^{T} f\left(y_{it} \mid \theta, \alpha\left(\theta, \eta_i\right)\right) d\eta_1 \cdots d\eta_n. \quad (46)$$

In Example 1 with $\theta = \sigma^2$, we have $E[u_{it} v_{it}] = 0$ so the reparameterization is unnecessary. Lancaster's estimator would therefore maximize

$$\int \cdots \int \prod_{i=1}^{n} \prod_{t=1}^{T} \frac{1}{\sqrt{\theta}} \exp\left(-\frac{(y_{it} - \alpha_i)^2}{2\theta}\right) d\alpha_1 \cdots d\alpha_n$$
$$\propto \frac{1}{\left(\sqrt{\theta}\right)^{T-1}} \exp\left(-\frac{\sum_{i=1}^{n} \sum_{t=1}^{T} \left(y_{it} - \overline{y}_i\right)^2}{2\theta}\right),$$

and

$$\widehat{\theta}_L = \frac{1}{n(T-1)} \sum_{i=1}^{n} \sum_{t=1}^{T} \left(y_{it} - \overline{y}_i\right)^2.$$

Note that $\widehat{\theta}_L$ has a zero bias.

Asymptotic properties of $\widehat{\theta}_L$ are not yet fully worked out except in a small number of specific examples. It is in general expected $\widehat{\theta}_L$ removes bias only up to $O(T^{-1})$, although we can find examples where $\widehat{\theta}_L$ eliminates bias of even higher order.

6.2 Overcoming Infeasibility of Orthogonalization

The Cox–Reid and Lancaster approaches are successful only when the parameter of interest can be orthogonalized with respect to the nuisance parameters. In general, such reparameterization requires solving some partial differential equations, and the solution may not exist. Because parameter orthogonalization

is not feasible in general, such approach cannot be implemented for arbitrary models. This problem can be overcome by adjusting the moment equation instead of the concentrated likelihood. We discuss two approaches in this regard, one introduced by Woutersen (2002) and the other by Arellano (2003). We will note that these two approaches result in identical estimators.

6.2.1 Woutersen's (2002) Approximation

Woutersen (2002) provided an insight on the role of Lancaster's posterior calculation in reducing the bias of the fixed effects. Assume for simplicity that the common parameter θ is orthogonal to α_i in the information sense, and no reparameterization is necessary to implement Lancaster's proposal. Given the posterior

$$\prod_{i=1}^{n}\left(\int\prod_{t=1}^{T}f\left(y_{it}|\,\theta,\alpha_i\right)d\alpha_i\right),$$

the first-order condition that characterize the posterior mode can be written as

$$0=\sum_{i=1}^{n}\frac{\int\left(\sum_{t=1}^{T}u_{it}\left(\theta,\alpha_i\right)\right)\prod_{t=1}^{T}f\left(y_{it}|\,\theta,\alpha_i\right)d\alpha_i}{\int\prod_{t=1}^{T}f\left(y_{it}|\,\theta,\alpha_i\right)d\alpha_i}. \tag{47}$$

Woutersen (2002) pointed out that the ith summand on the right can be approximated by

$$\begin{aligned}&\sum_{t=1}^{T}u_{it}\left(\theta,\widehat{\alpha}_i\left(\theta\right)\right)-\frac{1}{2}\frac{\sum_{t=1}^{T}u_{it}^{\alpha_i\alpha_i}\left(\theta,\widehat{\alpha}_i\left(\theta\right)\right)}{\sum_{t=1}^{T}v_{it}^{\alpha_i}\left(\theta,\widehat{\alpha}_i\left(\theta\right)\right)}\\&\quad+\frac{1}{2}\frac{\left(\sum_{t=1}^{T}v_{it}^{\alpha_i\alpha_i}\left(\theta,\widehat{\alpha}_i\left(\theta\right)\right)\right)\left(\sum_{t=1}^{T}u_{it}^{\alpha_i}\left(\theta,\widehat{\alpha}_i\left(\theta\right)\right)\right)}{\left(\sum_{t=1}^{T}v_{it}^{\alpha_i}\left(\theta,\widehat{\alpha}_i\left(\theta\right)\right)\right)^2},\end{aligned}$$

where $\widehat{\alpha}_i\left(\theta\right)$ is a solution to $\sum_{t=1}^{T}v_{it}\left(\theta,\widehat{\alpha}_i\left(\theta\right)\right)=0$. Therefore, Woutersen's estimator under parameter orthogonality is the solution to

$$0=\sum_{i=1}^{n}\left[\sum_{t=1}^{T}u_{it}\left(\theta,\widehat{\alpha}_i\left(\theta\right)\right)-\frac{1}{2}\frac{\widehat{E}_T\left[\widehat{u}^{\alpha_i\alpha_i}\right]}{\widehat{E}_T\left[\widehat{v}_{it}^{\alpha_i}\right]}+\frac{1}{2}\frac{\widehat{E}_T\left[\widehat{v}_{it}^{\alpha_i\alpha_i}\right]\widehat{E}_T\left[\widehat{u}_{it}^{\alpha_i}\right]}{\left(\widehat{E}_T\left[\widehat{v}_{it}^{\alpha_i}\right]\right)^2}\right]. \tag{48}$$

Note that this estimator solves the same moment equation as Cox and Reid's moment equation (41).

Woutersen pointed out that the moment function

$$\overline{u}_{it}\left(\theta,\alpha\right)\equiv u_{it}\left(\theta,\alpha\right)-\rho_i\left(\theta,\alpha\right)v_{it}\left(\theta,\alpha\right) \tag{49}$$

where

$$\rho_i(\theta,\alpha) \equiv \frac{\int u_i^{\alpha}(y;\theta,\alpha)\, f_i(y;\theta,\alpha)\, dy}{\int v_i^{\alpha}(y;\theta,\alpha)\, f_i(y;\theta,\alpha)\, dy} \tag{50}$$

would satisfy the orthogonality requirement in the sense that at true values

$$E\left[\overline{u}_{it}^{\alpha}(\theta_0,\alpha_{i0})\right] = 0.$$

Recall that $U_{it}(\theta,\alpha_i) \equiv u_{it} - v_{it}E[v_{it}^2]^{-1}E[v_{it}u_{it}]$ defined in Section 3 cannot be used as a basis of estimation because the ratio $E[v_{it}^2]^{-1}E[v_{it}u_{it}]$ is not known in general. It was used only as a theoretical device to understand the asymptotic property of various estimators. On the other hand, $\rho(\theta_0,\alpha_{i0}) = E[v_{it}^{\alpha}]^{-1}E[u_{it}^{\alpha}] = E[v_{it}^2]^{-1}E[v_{it}u_{it}]$, so we can consider $\overline{u}_{it}(\theta,\alpha_i)$ as a feasible version of $U_{it}(\theta,\alpha_i)$. Woutersen's moment equation when parameter orthogonality is unavailable is therefore obtained by replacing $u_{it}(\theta,\widehat{\alpha}_i(\theta))$ in (48) by $\overline{u}_{it}(\theta,\widehat{\alpha}_i(\theta))$.

6.2.2 Arellano's (2003) Proposal

An orthogonal transformation is a function $\eta_i = \eta_i(\theta,\alpha)$ such that

$$\frac{\eta_{\theta i}}{\eta_{\alpha i}} = \rho_i(\theta,\alpha)$$

where $\eta_{\theta i} = \partial\eta_i/\partial\theta$, $\eta_{\alpha i} = \partial\eta_i/\partial\alpha$, and $\rho_i(\theta,\alpha)$ is given in (50). Such a function may or may not exist, and if it does it need not be unique.

Arellano (2003) considers a Cox and Reid's (1987) objective function that is written for some transformation of the effects $\eta_i = \eta_i(\theta,\alpha)$ and he rewrites it in terms of the original parameterization. The resulting criterion is given by (39) with the addition of the Jacobian of the transformation:

$$\sum_{t=1}^{T}\ell_{it}(y_{it};\theta,\widehat{\alpha}_i(\theta)) - \frac{1}{2}\log\left(-\sum_{t=1}^{T}\frac{\partial^2\ell_{it}(y_{it};\theta,\widehat{\alpha}_i(\theta))}{\partial\alpha_i^2}\right) + \log\left(\widehat{\eta}_{\alpha i}\right)$$

where $\widehat{\eta}_{\alpha i} = \left(\eta_{\alpha i}\mid_{\alpha=\widehat{\alpha}_i(\theta)}\right)$. The corresponding moment equation is

$$\sum_{t=1}^{T}u_{it}(\theta,\widehat{\alpha}_i(\theta)) - \widetilde{b}_i^{CR}(\theta) + m_i(\theta)$$

where $\widetilde{b}_i^{CR}(\theta)$ is given in (42) and

$$\begin{aligned} m_i(\theta) &= \frac{\partial}{\partial\theta}\log\left(\widehat{\eta}_{\alpha i}\right) = \frac{\widehat{\eta}_{\alpha\theta i}}{\widehat{\eta}_{\alpha i}} + \frac{\widehat{\eta}_{\alpha\alpha i}}{\widehat{\eta}_{\alpha i}}\frac{\partial\widehat{\alpha}_i(\theta)}{\partial\theta} \\ &= \left(\frac{\partial}{\partial\alpha}\frac{\eta_{\theta i}}{\eta_{\alpha i}}\bigg|_{\alpha=\widehat{\alpha}_i(\theta)}\right) - \frac{\widehat{\eta}_{\alpha\alpha i}}{\widehat{\eta}_{\alpha i}}\left(\frac{\widehat{E}_T\left[\widehat{u}_{it}^{\alpha_i}\right]}{\widehat{E}_T\left[\widehat{v}_{it}^{\alpha_i}\right]} - \frac{\widehat{\eta}_{\theta i}}{\widehat{\eta}_{\alpha i}}\right). \end{aligned}$$

If $\eta_i(\theta, \alpha)$ is an orthogonal transformation

$$m_i(\theta) = \left.\frac{\partial \rho_i(\theta, \alpha)}{\partial \alpha}\right|_{\alpha=\widehat{\alpha}_i(\theta)} - \frac{\widehat{\eta}_{\alpha\alpha i}}{\widehat{\eta}_{\alpha i}}\left(\frac{\widehat{E}_T\left[\widehat{u}_{it}^{\alpha_i}\right]}{\widehat{E}_T\left[\widehat{v}_{it}^{\alpha_i}\right]} - \rho_i(\theta, \widehat{\alpha}_i(\theta))\right) \tag{51}$$

so that

$$m_i(\theta_0) = \left.\frac{\partial \rho_i(\theta_0, \alpha)}{\partial \alpha}\right|_{\alpha=\widehat{\alpha}_i(\theta_0)} + O\left(\frac{1}{T}\right).$$

Thus, regardless of the existence of an orthogonal transformation, it is always possible to obtain a locally orthogonal Cox and Reid moment equation. Arellano's moment equation is therefore obtained as

$$0 = \sum_{i=1}^{n}\left[\sum_{t=1}^{T} u_{it}(\theta, \widehat{\alpha}_i(\theta)) - \widetilde{b}_i^{CR}(\theta) + \left.\frac{\partial \rho_i(\theta, \alpha)}{\partial \alpha}\right|_{\alpha=\widehat{\alpha}_i(\theta)}\right], \tag{52}$$

after supressing the transformation-specific term in (51) that is irrelevant for the purpose of bias reduction. Indeed, Carro (2004) has shown that Arellano's moment equation reduces the order of the score bias regardless of the existence of an information orthogonal reparameterization.

It can be shown that this moment equation is identical to Woutersen's (2002) moment equation. This can be shown in the following way. Now note that Woutersen's (2002) moment equation is equal to

$$0 = \sum_{i=1}^{n}\left[\sum_{t=1}^{T}\overline{u}_{it}(\theta, \widehat{\alpha}_i(\theta)) - \frac{1}{2}\frac{\sum_{t=1}^{T}\overline{u}_{it}^{\alpha_i\alpha_i}(\theta, \widehat{\alpha}_i(\theta))}{\sum_{t=1}^{T} v_{it}^{\alpha_i}(\theta, \widehat{\alpha}_i(\theta))} + \frac{1}{2}\frac{\left(\sum_{t=1}^{T} v_{it}^{\alpha_i\alpha_i}(\theta, \widehat{\alpha}_i(\theta))\right)\left(\sum_{t=1}^{T}\overline{u}_{it}^{\alpha_i}(\theta, \widehat{\alpha}_i(\theta))\right)}{\left(\sum_{t=1}^{T} v_{it}^{\alpha_i}(\theta, \widehat{\alpha}_i(\theta))\right)^2}\right]. \tag{53}$$

Using (32), we can obtain:

$$\sum_{t=1}^{T}\overline{u}_{it}(\theta, \widehat{\alpha}_i(\theta)) = \sum_{t=1}^{T} u_{it}(\theta, \widehat{\alpha}_i(\theta)),$$

$$\sum_{t=1}^{T}\overline{u}_{it}^{\alpha_i}(\theta, \widehat{\alpha}_i(\theta)) = \sum_{t=1}^{T} u_{it}^{\alpha_i}(\theta, \widehat{\alpha}_i(\theta)) - \left(\sum_{t=1}^{T} v_{it}^{\alpha_i}(\theta, \widehat{\alpha}_i(\theta))\right)\rho_i(\theta, \alpha)\big|_{\alpha=\widehat{\alpha}_i(\theta)}$$

and

$$\sum_{t=1}^{T} \overline{u}_{it}^{\alpha_i \alpha_i}\left(\theta, \widehat{\alpha}_i\left(\theta\right)\right) = \sum_{t=1}^{T} u_{it}^{\alpha_i \alpha_i}\left(\theta, \widehat{\alpha}_i\left(\theta\right)\right) - \left(\sum_{t=1}^{T} v_{it}^{\alpha_i \alpha_i}\left(\theta, \widehat{\alpha}_i\left(\theta\right)\right)\right) \rho_i\left(\theta, \alpha\right)\Big|_{\alpha=\widehat{\alpha}_i(\theta)} - 2\left(\sum_{t=1}^{T} v_{it}^{\alpha_i}\left(\theta, \widehat{\alpha}_i\left(\theta\right)\right)\right) \frac{\partial \rho_i\left(\theta, \alpha\right)}{\partial \alpha}\Bigg|_{\alpha=\widehat{\alpha}_i(\theta)}.$$

Plugging these expressions to (53), we obtain after some simplification an alternative characterization of Woutersen's (2002) moment equation:

$$0 = \sum_{i=1}^{n}\left[\sum_{t=1}^{T} u_{it}\left(\theta, \widehat{\alpha}_i\left(\theta\right)\right) - \frac{1}{2}\frac{\sum_{t=1}^{T} u_{it}^{\alpha_i \alpha_i}\left(\theta, \widehat{\alpha}_i\left(\theta\right)\right)}{\sum_{t=1}^{T} v_{it}^{\alpha_i}\left(\theta, \widehat{\alpha}_i\left(\theta\right)\right)} + \frac{1}{2}\frac{\left(\sum_{t=1}^{T} v_{it}^{\alpha_i \alpha_i}\left(\theta, \widehat{\alpha}_i\left(\theta\right)\right)\right)\left(\sum_{t=1}^{T} u_{it}^{\alpha_i}\left(\theta, \widehat{\alpha}_i\left(\theta\right)\right)\right)}{\left(\sum_{t=1}^{T} v_{it}^{\alpha_i}\left(\theta, \widehat{\alpha}_i\left(\theta\right)\right)\right)^2} + \frac{\partial \rho_i\left(\theta, \alpha\right)}{\partial \alpha}\Bigg|_{\alpha=\widehat{\alpha}_i(\theta)}\right],$$

which can be seen to be identical to moment equation (52). We can therefore conclude that Woutesen's (2002) is identical to Arellano's (2003).

6.2.3 Relation to Bias Correction of the Moment Equation

The moment equation used by Woutersen, Arellano, and Carro can be written as

$$\sum_{i=1}^{n}\left[\sum_{t=1}^{T} u_{it}\left(\theta, \widehat{\alpha}_i\left(\theta\right)\right) - \widetilde{b}_i^{W}\left(\theta\right)\right] = 0 \tag{54}$$

where

$$\widetilde{b}_i^{W}\left(\theta\right) = \widetilde{b}_i^{CR}\left(\theta\right) - \frac{\partial \rho_i\left(\theta, \alpha\right)}{\partial \alpha}\Bigg|_{\alpha=\widehat{\alpha}_i(\theta)}, \tag{55}$$

$$\widetilde{b}_i^{CR}\left(\theta\right) = \frac{1}{2\widehat{E}_T\left[\widehat{v}_{it}^{\alpha_i}\right]}\left(\widehat{E}_T\left[\widehat{u}^{\alpha_i \alpha_i}\right] - \widehat{E}_T\left[\widehat{v}_{it}^{\alpha_i \alpha_i}\right]\frac{\widehat{E}_T\left[\widehat{u}_{it}^{\alpha_i}\right]}{\widehat{E}_T\left[\widehat{v}_{it}^{\alpha_i}\right]}\right),$$

and at true values

$$\frac{\partial \rho_i\left(\theta_0, \alpha_{i0}\right)}{\partial \alpha} = \frac{1}{E\left[v_{it}^{\alpha}\right]}\left(E\left[u_{it}^{\alpha\alpha}\right] - E\left[v_{it}^{\alpha\alpha}\right]\frac{E\left[u_{it}^{\alpha}\right]}{E\left[v_{it}^{\alpha}\right]}\right) + \frac{1}{E\left[v_{it}^{\alpha}\right]}\left(E\left[u_{it}^{\alpha} v_{it}\right] - E\left[v_{it}^{\alpha} v_{it}\right]\frac{E\left[u_{it}^{\alpha}\right]}{E\left[v_{it}^{\alpha}\right]}\right). \tag{56}$$

Comparing the resulting expression with the theoretical bias (7), we note that moment condition (54) is using a valid estimate of the concentrated score $1/T$ bias as long as the information identity holds, so that in general it will be appropriate in likelihood settings. The estimated bias $\widetilde{b}_i^W(\theta)$ uses a combination of observed and expected terms. Note that, contrary to the situation under orthogonality when the theoretical bias reduces to (43), there is no redundant term here.

The term $\partial \rho_i(\theta, \widehat{\alpha}_i(\theta)) / \partial \alpha$ in (52) can be interpreted as a measure of how much the variance of $\widehat{\alpha}_i(\theta)$ changes with θ. In this respect, note the equivalence between the derivative of the log variance of $\widehat{\alpha}_i(\theta)$ in (45) and a sample counterpart of (56):

$$
\begin{aligned}
&-\frac{\partial}{\partial \theta}\frac{1}{2}\log \widehat{\mathrm{Var}}\left(\sqrt{T}\left(\widehat{\alpha}_i(\theta)-\overline{\alpha}_i(\theta)\right)\right) \\
&\quad=\frac{1}{\widehat{E}_T\left[v_{it}^{\alpha_i}\right]}\left(\widehat{E}_T\left[\widehat{u}_{it}^{\alpha_i\alpha_i}\right]-\widehat{E}_T\left[\widehat{v}_{it}^{\alpha_i\alpha_i}\right]\frac{\widehat{E}_T\left[\widehat{u}_{it}^{\alpha_i}\right]}{\widehat{E}_T\left[\widehat{v}_{it}^{\alpha_i}\right]}\right) \\
&\qquad+\frac{1}{\left(-\widehat{E}_T\left[\widehat{v}_{it}^2\right]\right)}\left(\widehat{E}_T\left[\widehat{u}_{it}^{\alpha_i}\widehat{v}_{it}\right]-\widehat{E}_T\left[\widehat{v}_{it}^{\alpha_i}\widehat{v}_{it}\right]\frac{\widehat{E}_T\left[\widehat{u}_{it}^{\alpha_i}\right]}{\widehat{E}_T\left[\widehat{v}_{it}^{\alpha_i}\right]}\right). \qquad (57)
\end{aligned}
$$

7 QMLE FOR DYNAMIC MODELS

The starting point of our discussion so far has been the assumption that the fixed effects estimator actually maximizes the likelihood. When we defined $\widehat{\theta}_T$ to be a maximizer of

$$
\sum_{i=1}^{n}\sum_{t=1}^{T}\log f\left(y_{it}|\,\theta, \widehat{\alpha}_i(\theta)\right),
$$

we assumed that (i) xs are strictly exogenous, (ii) ys are independent over t given xs, and (iii) f is the correct (conditional) density of y given x. We noted that some of the bias correction methods did not depend on the likelihood setting, while others, that relied on the information or Bartlett identities, did. However, in all cases assumptions (i) and (ii) were maintained. For example, if the binary response model

$$
y_{it} = 1\left(x_{it}'\theta + \alpha_i + e_{it} > 0\right), \qquad (58)
$$

where the marginal distribution of e_{it} is $\mathcal{N}(0, 1)$, is such that e_{it} is independent over t, and if it is estimated by nonlinear least squares, our first bias formula is valid.

In the likelihood setting, assumption (ii) can be relaxed choosing estimates of bias corrections that use expected rather than observed quantities. This is possible because the likelihood fully specifies the dynamics, and it is simple if the required expected quantities have closed form expressions, as in the dynamic probit models in Carro (2004) and Fernández-Val (2005).

In a nonlikelihood setting, our analysis can be generalized to the case when the fixed effects estimator maximizes

$$\sum_{i=1}^{n}\sum_{t=1}^{T}\psi\left(z_{it};\theta,\widehat{\alpha}_i\left(\theta\right)\right)$$

for an arbitrary ψ under some regularity conditions, thereby relaxing assumptions (i) and (ii). For example, the binary response model (58) can still be analyzed by considering the fixed effects probit MLE even when e_{it} has an arbitrary unknown serial correlation.

The intuition for this more general model can still be obtained from the approximation of the moment equation as in (19), which can be corrected by calculating the approximate expectation of the correction term

$$\sum_{i=1}^{n}\left[\frac{\sum_{t=1}^{T}v_{it}}{\sqrt{T}E\left[v_{it}^{\alpha_i}\right]}\right]\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(U_{it}^{\alpha_i}-\frac{E\left[U_{it}^{\alpha_i\alpha_i}\right]}{2E\left[v_{it}^{\alpha_i}\right]}v_{it}\right)\right].$$

The analysis for this more general model gets to be more complicated because calculation of the expectation should incorporate the serial correlation in v_{it} and $U_{it}^{\alpha_i}$, which was a non-issue in the simpler context. Hahn and Kuersteiner (2004) provide an analysis that incorporate such complication.

8 ESTIMATION OF MARGINAL EFFECTS

It is sometimes of interest to estimate quantities such as

$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}m\left(z_{it};\theta,\alpha_i\right) \tag{59}$$

where $z_{it}=\left(y_{it},x_{it}'\right)'$. For example, it may be of interest to estimate the mean marginal effects

$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\phi\left(x_{it}'\theta+\alpha_i\right)\theta$$

for the binary response model (58), where ϕ denotes the density of $\mathcal{N}\left(0,1\right)$. It would be sensible to estimate such quantities by

$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}m\left(z_{it};\widetilde{\theta},\widehat{\alpha}_i\left(\widetilde{\theta}\right)\right)$$

where $\widetilde{\theta}$ denotes a bias-corrected version of $\widehat{\theta}$ computed by one of the methods discussed before, and $\widehat{\alpha}_i(\widetilde{\theta})$ denotes the estimate of α_i at $\widetilde{\theta}$. Hahn and Newey (2004), Carro (2004), and Fernandez-Val (2005) discuss estimation and bias correction of such quantity.

To relate our discussion with the bias-correction formula developed there, it is useful to think about the quantity (59) as a solution to the (infeasible) moment equation

$$\sum_{i=1}^{n}\sum_{t=1}^{T}\left(m\left(z_{it};\widehat{\alpha}_i\left(\theta_0\right)\right)-\widehat{\mu}\right)=0,\quad \sum_{t=1}^{T}v\left(z_{it};\widehat{\alpha}_i\left(\theta_0\right)\right)=0 \tag{60}$$

where, for simplicity of notation, we suppressed the dependence of m on θ. Let

$$M\left(z_{it};\alpha_i\right)=m\left(z_{it};\alpha_i\right)-v\left(z_{it};\alpha_i\right)\frac{E\left[m^{\alpha_i}\left(z_{it};\alpha_i\right)\right]}{E\left[v^{\alpha_i}\left(z_{it};\alpha_i\right)\right]}$$

and note that $\widehat{\mu}$ in (60) solves

$$0=\sum_{i=1}^{n}\sum_{t=1}^{T}\left(M\left(z_{it};\widehat{\alpha}_i\left(\theta_0\right)\right)-\widehat{\mu}\right). \tag{61}$$

Assuming that serial correlation can be ignored, we can bias-correct this moment equation using the same intuition as in Section 4. We then obtain a bias-corrected version of the moment equation

$$\begin{aligned}0=&\sum_{i=1}^{n}\sum_{t=1}^{T}\left(M\left(z_{it};\widehat{\alpha}_i\left(\theta_0\right)\right)-\widehat{\widehat{\mu}}\right)\\&+\sum_{i=1}^{n}\left(\frac{\sum_{t=1}^{T}v_{it}M_{it}^{\alpha_i}}{\sum_{t=1}^{T}v_{it}^{\alpha_i}}+\frac{\sum_{t=1}^{T}M_{it}^{\alpha_i\alpha_i}}{2\left(\sum_{t=1}^{T}v_{it}^{\alpha_i}\right)}\right)\end{aligned} \tag{62}$$

when the fixed effects estimator is based on a correctly specified likelihood, or

$$\begin{aligned}0=&\sum_{i=1}^{n}\sum_{t=1}^{T}\left(M\left(z_{it};\widehat{\alpha}_i\left(\theta_0\right)\right)-\widehat{\widehat{\mu}}\right)\\&+\sum_{i=1}^{n}\left(\frac{\sum_{t=1}^{T}v_{it}M_{it}^{\alpha_i}}{\sum_{t=1}^{T}v_{it}^{\alpha_i}}-\frac{\left(\sum_{t=1}^{T}v_{it}^{2}\right)\left(\sum_{t=1}^{T}M_{it}^{\alpha_i\alpha_i}\right)}{2\left(\sum_{t=1}^{T}v_{it}^{\alpha_i}\right)^{2}}\right)\end{aligned} \tag{63}$$

in general. Replacing $M\left(z_{it};\theta_0,\widehat{\alpha}_i\left(\theta_0\right)\right)$ in (62) by the feasible version

$$m\left(z_{it};\widetilde{\theta},\widehat{\alpha}_i\left(\widetilde{\theta}\right)\right)-v\left(z_{it};\widetilde{\theta},\widehat{\alpha}_i\left(\widetilde{\theta}\right)\right)\frac{\sum_{t=1}^{T}m^{\alpha_i}\left(z_{it};\widetilde{\theta},\widehat{\alpha}_i\left(\widetilde{\theta}\right)\right)}{\sum_{t=1}^{T}v^{\alpha_i}\left(z_{it};\widetilde{\theta},\widehat{\alpha}_i\left(\widetilde{\theta}\right)\right)},$$

we obtain the same bias-corrected estimator $\widehat{\widehat{\mu}}$ as in Hahn and Newey (2004), and Fernandez-Val (2005).

9 AUTOMATIC METHODS

We have so far discussed methods of bias correction based on some analytic formulae. Depending on applications, we may be able to by-pass such analysis, and rely on numerical methods. We discuss two such procedures here.

9.1 Panel Jackknife

The panel jackknife is an automatic method of bias correction. To describe it, let $\widehat{\theta}_{(t)}$ be the fixed effects estimator based on the subsample excluding the observations of the tth period. The jackknife estimator is

$$\widetilde{\theta} \equiv T\widehat{\theta} - (T-1)\sum_{t=1}^{T}\widehat{\theta}_{(t)}/T \tag{64}$$

or

$$\widetilde{\theta} \equiv \widehat{\theta} - \frac{\widetilde{B}}{T}, \qquad \frac{\widetilde{B}}{T} = (T-1)\left(\frac{1}{T}\sum_{t=1}^{T}\widehat{\theta}_{(t)} - \widehat{\theta}\right).$$

To explain the bias correction from this estimator it is helpful to consider a further expansion

$$\theta_T = \theta_0 + \frac{B}{T} + \frac{D}{T^2} + O\left(\frac{1}{T^3}\right). \tag{65}$$

The limit of $\widetilde{\theta}$ for fixed T and how it changes with T shows the effect of the bias correction. The estimator $\widetilde{\theta}$ will converge in probability to

$$\begin{aligned} T\theta_T - (T-1)\theta_{T-1} &= \theta_0 + \left(\frac{1}{T} - \frac{1}{T-1}\right)D + O\left(\frac{1}{T^2}\right) \\ &= \theta_0 + O\left(\frac{1}{T^2}\right) \end{aligned} \tag{66}$$

or

$$(T-1)(\theta_{T-1} - \theta_T) = \frac{B}{T} + O\left(\frac{1}{T^2}\right).$$

Thus, we see that the asymptotic bias of the jackknife corrected estimator is of order $1/T^2$. Consequently, this estimator will have an asymptotic distribution centered at 0 when $n/T \to \rho$. Hahn and Newey (2004) formally established that $\sqrt{nT}\left(\widetilde{\theta} - \theta_0\right)$ has the same asymptotic variance as $\sqrt{nT}\left(\widehat{\theta} - \theta_0\right)$ when $n/T \to \rho$. This implies that the bias reduction is achieved without any increase in the asymptotic variance. This suggests that, although there may be some small increase in variance as a result of bias reduction, the increase is so small that it is ignored when $n/T \to \rho$.

In Example 1, it is straightforward to show that

$$\widetilde{\theta} = \frac{1}{n(T-1)}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(y_{it} - \overline{y}_i\right)^2, \tag{67}$$

which is the estimator that takes care of the degrees of freedom problem. It is interesting to note that the jackknife bias correction completely removed bias in this example: $E(\widetilde{\theta}) = \theta$. This happens only because the $O(T^{-2})$ term is

identically equal to zero in this particular example, which is not expected to happen too often in practice.

It is natural to speculate that a higher-order version of the panel jackknife may correct even higher-order bias. For this purpose, assume that an expansion even higher than (65) is valid:

$$\theta_T = \theta_0 + \frac{B}{T} + \frac{D}{T^2} + \frac{F}{T^3} + \frac{G}{T^4} + O\left(\frac{1}{T^5}\right).$$

Because

$$\begin{aligned}
&\frac{1}{2}T^2\theta_T - (T-1)^2\theta_{T-1} + \frac{1}{2}(T-2)^2\theta_{T-2} \\
&\quad = \theta_0 + \frac{F}{T(T-1)(T-2)} + \frac{3T^2-6T+2}{T^2(T-1)^2(T-2)^2}G + O\left(\frac{1}{T^3}\right) \\
&\quad = \theta + O\left(\frac{1}{T^3}\right),
\end{aligned}$$

we can conjecture that an estimator of the form

$$\widetilde{\widetilde{\theta}} \equiv \frac{1}{2}T^2\widehat{\theta} - (T-1)^2\frac{\sum_{s=1}^{T}\widehat{\theta}_{(s)}}{T} + \frac{1}{2}(T-2)^2\frac{\sum_{s\neq s'}\widehat{\theta}_{(s,s')}}{T(T-1)},$$

where $\widehat{\theta}_{(s,s')}$ denotes the delete-2 estimator, will be centered at zero even at the asymptotics where $n = o\left(T^5\right)$.

The panel jackknife is easiest to understand when y_{it} is independent over time. When it is serially correlated, which is to be expected in many applications, it is not yet clear how it should be modified. To understand the gist of the problem, it is useful to investigate the role of $\sum_{t=1}^{T}\widehat{\theta}_{(t)}/T$ in (64). Note that it is the sample analog of θ_{T-1} in (66). When y_{it} is serially correlated, what should be used as the sample analog? One natural candidate is to use the same formula as in (64), with the understanding that $\widehat{\theta}_{(t)}$ should be the MLE maximizing the likelihood of $(y_{i1}, \ldots, y_{i,t-1}, y_{i,t+1}, \ldots, y_T)$ $i = 1, \ldots, n$. We are not aware of any formal result that establishes the asymptotic properties of the panel jackknife estimator, even in the simple dynamic panel model where $y_{it} = \alpha_i + \theta y_{i,t-1} + \varepsilon_{it}$ with $\varepsilon_{it} \sim \mathcal{N}\left(0, \sigma^2\right)$. Even if this approach is shown to have a desirable asymptotic property, we should bear in mind that such approach requires complete parametric specification of the distribution of $(y_{i1}, \ldots, y_{iT})$. In many applications, we do not have a complete specification of the likelihood.

Another possibility is to use $\widehat{\theta}_{(T)}$ as the sample analog of θ_{T-1}. Note that $\widehat{\theta}_{(T)}$ is the MLE based on the first $T-1$ observations. It turns out that such procedure will be accompanied by some large increase in variance. To understand this problem, it is useful to examine Example 1 again. It can be shown that

$$\widehat{\theta}_{(T-1)} = \frac{T}{T-1}\widehat{\theta} - \frac{T}{n(T-1)^2}\sum_{i=1}^{n}\left(\overline{y}_i - y_{iT}\right)^2$$

and therefore,

$$T\widehat{\theta} - (T-1)\widehat{\theta}_{(T-1)} = \frac{T}{n(T-1)} \sum_{i=1}^{n} \left(\overline{y}_i - y_{iT}\right)^2 .$$

We can write with some abuse of notation that $T\widehat{\theta} - (T-1)\widehat{\theta}_{(T-1)} \sim \frac{\theta_0}{n}\chi^2_n$, whereas $\widetilde{\theta}$ in (67) is distributed as $\frac{\theta_0}{n(T-1)}\chi^2_{n(T-1)}$. This implies that (i) $T\widehat{\theta} - (T-1)\widehat{\theta}_{(T-1)}$ is indeed bias free; and (ii) the variance of $T\widehat{\theta} - (T-1)\widehat{\theta}_{(T-1)}$ is $T-1$ times as large as that of as the jackknife estimator $\widetilde{\theta}$. When T is sufficiently large, this delete-last-observation approach will be unacceptable. We expect a similar problem when y_{it} is subject to serial correlation, and eliminate $T\widehat{\theta} - (T-1)\widehat{\theta}_{(T-1)}$ from our consideration.

We argued that the panel jackknife may not be attractive when serial correlation is suspected. The bootstrap is another way of reducing bias. A time series version of the bootstrap is block-bootstrap, which has been shown in many occasions to have desirable properties. We conjecture that some version of a bootstrap bias correction would also remove the asymptotic bias (e.g., with truncation as in Hahn, Kuersteiner, and Newey, 2002).

9.2 Bootstrap-Adjusted Concentrated Likelihood

Simulation methods can also be used for bias correction of moment equations and objective functions. Pace and Salvan (2006) have suggested a bootstrap approach to adjust the concentrated likelihood.

Consider generating parametric bootstrap samples $\{y_{i1}(r), \ldots, y_{iT}(r)\}_{i=1}^n$ $(r = 1, \ldots, R)$ from the models $\{\prod_{t=1}^T f(y_t \mid \widehat{\theta}, \widehat{\alpha}_i)\}_{i=1}^n$ to obtain $\widehat{\alpha}_i^{[r]}(\theta)$ as the solution to

$$\widehat{\alpha}_i^{[r]}(\theta) = \underset{\alpha}{\operatorname{argmax}} \sum_{t=1}^{T} \log f\left(y_{it}(r) \mid \theta, \alpha\right) \quad (r = 1, \ldots, R).$$

Pace and Salvan's (2006) simulation adjusted log-likelihood for the ith unit is

$$\overline{\ell}_i^S(\theta) = \frac{1}{R} \sum_{r=1}^{R} \sum_{t=1}^{T} \ell_{it}\left(\theta, \widehat{\alpha}_i^{[r]}(\theta)\right). \tag{68}$$

The criterion $\overline{\ell}_i^S(\theta)$ is invariant under one-to-one reparameterizations of α_i that leave θ fixed (invariant under "interest respecting reparameterizations").

Alternatively, Pace and Salvan consider the form in (30), using a bootstrap estimate of $V_i[\widehat{\alpha}_i(\theta)]$ given by

$$\widetilde{V}_i[\widehat{\alpha}_i(\theta)] = \frac{1}{R} \sum_{r=1}^{R} \left[\widehat{\alpha}_i^{[r]}(\theta) - \widehat{\alpha}_i(\theta)\right]^2, \tag{69}$$

which leads to

$$\overline{\ell}_i^{SA}(\theta) = \sum_{t=1}^{T} \ell_{it}\left(\theta, \widehat{\alpha}_i(\theta)\right) - \frac{1}{2}\left(-\frac{1}{T}\sum_{t=1}^{T} \frac{\partial v_{it}\left(\theta, \widehat{\alpha}_i(\theta)\right)}{\partial \alpha}\right) \widetilde{V}_i\left[\widehat{\alpha}_i(\theta)\right]. \quad (70)$$

10 CONCLUDING REMARKS

We discussed a variety of methods of estimation of nonlinear fixed effects panel data models with reduced bias properties. Alternative approaches to bias correction based on adjusting the estimator, the moment equation, and the criterion function have been considered. We have also discussed approaches relying on orthogonalization and automatic methods, as well as the connections among the various approaches.

All the approaches that we discuss in the paper are based on an asymptotic approximation where n and T grow to infinity at the same rate. Therefore, they are likely to be useful in applications in which the value of T is not negligible relative to n. Examples of this kind include data sets constructed from country or regional level macropanels, the balance-sheet-based company panels that are available in many countries, or the household incomes panel in the US (PSID). However, for n too large relative to T, the sampling distributions of the $1/T$ bias-corrected estimators will not provide accurate confidence intervals because their standard deviation will be small relative to bias. In those situations, an asymptotic approximation where n/T^3 converges to a constant may be called for, leading to $1/T^2$ bias-corrected estimators. A more general issue is how good are the n and T asymptotic approximations when the objective is to produce confidence intervals, or to test a statistical hypothesis. This is a question beyond the scope of this paper.

Next in the agenda, it is important to find out how well each of these bias correction methods work for specific models and data sets of interest in applied econometrics. In this regard, the Monte Carlo results and empirical estimates obtained by Carro (2004) and Fernández-Val (2005) for binary choice models are very encouraging. For a dynamic logit model, using the same simulation design as in Honoré and Kyriazidou (2000), they find that a score-corrected estimator and two one-step analytical bias-corrected estimators are broadly comparable to the Honoré–Kyriazidou estimator (which is consistent for fixed T) when $T = 8$ and $n = 250$. However, the finite sample properties of the bias correction seem to depend on how they are done. For dynamic logit, Carro's score-corrected estimator and Fernández-Val's bias-corrected estimator, which use expected quantities, are somewhat superior to a bias-corrected estimator using observed quantities, but more results are needed for other models and simulation designs.

We have focused on bias reduction, but other theoretical properties should play a role in narrowing the choice of bias-reducing estimation methods. In the likelihood context it is natural to seek an adjusted concentrated likelihood that behaves like a proper likelihood. In this respect, information bias reduction

and invariance to reparameterization are relevant properties in establishing the relative merits of different bias-reducing estimators.

References

AMEMIYA, T. (1985): *Advanced Econometrics*, Oxford: Basil Blackwell.

ANDERSEN, E. (1970): "Asymptotic Properties of Conditional Maximum Likelihood Estimators," *Journal of the Royal Statistical Society, Series B*, 32, 283–301.

ARELLANO, M. (2003): "Discrete Choices with Panel Data," *Investigaciones Económicas*, 27, 423–458.

BARNDORFF-NIELSEN, O. E. (1983): "On a Formula for the Distribution of the Maximum Likelihood Estimator," *Biometrika*, 70, 343–365.

CARRO, J. (2004): "Estimating Dynamic Panel Data Discrete Choice Models with Fixed Effects," Unpublished manuscript.

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

——— (1992): "Binary Response Models for Panel Data: Identification and Information," Unpublished manuscript.

COX, D. R. AND N. REID (1987): "Parameter Orthogonality and Approximate Conditional Inference" (with discussion), *Journal of the Royal Statistical Society, Series B*, 49, 1–39.

——— (1992): "A Note on the Difference Between Profile and Modified Profile Likelihood," *Biometrika*, 79, 408–411.

DICICCIO, T. J. AND S. E. STERN (1993): "An adjustment to Profile Likelihood Based on Observed Information," Technical Report, Department of Statistics, Stanford University.

DICICCIO, T. J., M. A. MARTIN, S. E. STERN, and G. A. YOUNG (1996): "Information Bias and Adjusted Profile Likelihoods," *Journal of the Royal Statistical Society, Series B*, 58, 189–203.

FERGUSON, H., N. REID, AND D. R. COX (1991): "Estimating Equations from Modified Profile Likelihood," in *Estimating Functions*, edited by V. P. Godambe, Oxford: Oxford University Press.

FERNÁNDEZ-VAL, I. (2005): "Estimation of Structural Parameters and Marginal Effects in Binary Choice Panel Data Models with Fixed Effects," Unpublished manuscript.

HAHN, J. (2001): "The Information Bound of a Dynamic Panel Logit Model with Fixed Effects," *Econometric Theory*, 17, 913–932.

HAHN, J. AND G. KUERSTEINER (2004): "Bias Reduction for Dynamic Nonlinear Panel Models with Fixed effects," Unpublished manuscript.

HAHN, J., G. KUERSTEINER, AND W. K. NEWEY (2002): "Higher Order Properties of Bootstrap and Jackknife Bias Corrected Maximum Likelihood Estimators," Unpublished manuscript.

HAHN, J. AND W. K. NEWEY (2004): "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models," *Econometrica*, 72, 1295–1319.

HONORÉ, B. E. AND E. KYRIAZIDOU (2000): "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica*, 68, 839–874.

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

——— (2002): "Orthogonal Parameters and Panel Data," *Review of Economic Studies*, 69, 647–666.

NEWEY, W. K. (1990): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135.

NEYMAN, J. AND E. L. SCOTT (1948): "Consistent Estimates Based on Partially Consistent Observations," *Econometrica*, 16, 1–32.

PACE, L. AND A. SALVAN (2006): "Adjustments of the Profile Likelihood from a New Perspective," *Journal of Statistical Planning and Inference*, 136, 3554–3564.

REID, N. (1995): "The Roles of Conditioning in Inference," *Statistical Science*, 10, 138–199.

SARTORI, N. (2003): "Modified Profile Likelihoods in Models with Stratum Nuisance Parameters," *Biometrika*, 90, 533–549.

SEVERINI, T. A. (1998): "An Approximation to the Modified Profile Likelihood Function," *Biometrika*, 85, 403–411.

——— (2000): *Likelihood Methods in Statistics*, Oxford: Oxford University Press.

——— (2002): "Modified Estimating Functions," *Biometrika*, 89, 333–343.

SEVERINI, T. A. AND W. H. WONG (1992): "Profile Likelihood and Conditionally Parametric Models," *The Annals of Statistics*, 20, 1768–1802.

STEIN, C. (1956): "Efficient Nonparametric Testing and Estimation," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, Vol. 1.

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

WOUTERSEN, T. (2002): "Robustness Against Incidental Parameters," Unpublished manuscript.