ROBUST PRIORS IN NONLINEAR PANEL DATA MODELS

By Manuel Arellano and Stéphane Bonhomme[1]

Many approaches to estimation of panel models are based on an average or integrated likelihood that assigns weights to different values of the individual effects. Fixed effects, random effects, and Bayesian approaches all fall in this category. We provide a characterization of the class of weights (or priors) that produce estimators that are first-order unbiased. We show that such bias reducing weights will depend on the data in general unless an orthogonal reparameterization or an essentially equivalent condition is available. Two intuitively appealing weighting schemes are discussed. We argue that asymptotically valid confidence intervals can be read from the posterior distribution of the common parameters when N and T grow at the same rate. Next, we show that random effects estimators are not bias reducing in general and discuss important exceptions. Moreover, the bias depends on the Kullback-Leibler distance between the population distribution of the effects and its best approximation in the random effects family. Finally, we show that in general standard random effects estimation of marginal effects is inconsistent for large T, whereas the posterior mean of the marginal effect is large-T consistent, and we provide conditions for bias reduction. Some examples and Monte Carlo experiments illustrate the results.

Keywords: Panel data, incidental parameters, bias reduction, integrated likelihood, priors.

[1] We thank Victor Aguirregabiria, Josh Angrist, Gary Chamberlain, Gabriele Fiorentini, Jean-Pierre Florens, Jinyong Hahn, Laura Hospido, Guido Imbens, Thierry Magnac, Jean-Marc Robin, Enrique Sentana, and two anonymous referees for useful comments. All remaining errors are our own. Research funding from the Spanish MEC, Grant SEJ2005-08880 is gratefully acknowledged.

1 Introduction

In a panel model the likelihood of the data y_i for a given unit is typically a function $f\left(y_i, \theta, \alpha_i\right) = f_i\left(\theta, \alpha_i\right)$ of common and individual specific parameters θ and α_i, respectively. Interest centers in the estimation of θ or other common policy parameters constructed as summary measures of the two types of parameters and data. The central feature of this estimation problem is the presence of many nuisance parameters (the individual effects) when the cross-sectional dimension is large relative to the number of time series observations.

Many approaches to estimation of θ in this context are based on an average likelihood that assigns weights to different values of α_i:

$$f_i^a\left(\theta\right) = \int f_i\left(\theta, \alpha_i\right) w_i\left(\alpha_i\right) d\alpha_i \tag{1}$$

where $w_i\left(\alpha_i\right)$ is a possibly θ-specific weight, related to a discrete or continuous measure. An estimate of θ is then usually chosen to maximize the average likelihood of the sample under cross-sectional independence: $\sum_{i=1}^{N} \ln f_i^a\left(\theta\right)$.

A fixed effects approach that estimates θ jointly with the individual effects by maximum likelihood (ML) falls in this category with weights assigning all mass to $\alpha_i = \widehat{\alpha}_i\left(\theta\right)$, where $\widehat{\alpha}_i\left(\theta\right)$ is the maximum likelihood estimator of α_i for given θ. That is,

$$w_i\left(\alpha_i\right) = \delta\left(\alpha_i - \widehat{\alpha}_i\left(\theta\right)\right) \tag{2}$$

where $\delta(.)$ denotes Dirac's delta function. The resulting average likelihood in this case is just the concentrated likelihood $f_i\left(\theta, \widehat{\alpha}_i\left(\theta\right)\right)$.

A random effects approach is also based on an average likelihood in which the weights are chosen as a model for the distribution of individual effects in the population given covariates and initial observations. In this case $w_i\left(\alpha_i\right)$ is a parametric or semiparametric density or probability mass function which does not depend on θ, but includes additional unknown coefficients:

$$w_i\left(\alpha_i\right) = \pi_i\left(\alpha_i; \xi\right).$$

Finally, in a Bayesian approach, beginning with a joint prior for common and individual parameters $\pi\left(\theta, \alpha_1, ..., \alpha_N\right)$, an average likelihood is also constructed. In this case, weights are chosen as a formulation of the prior probability distribution of α_i given θ, covariates and initial observations, under the assumption of prior conditional independence of $\alpha_1, ..., \alpha_N$

given θ:

$$w_i(\alpha_i) = \pi_i(\alpha_i|\theta),$$

such that

$$\pi(\theta, \alpha_1...\alpha_N) = \pi_1(\alpha_1|\theta)...\pi_N(\alpha_N|\theta)\pi(\theta). \tag{3}$$

However, α_i and θ need not be independent, so that the weights assigned to different values of α_i may depend on the value of θ.

All these approaches, in general, lead to estimators of θ that are not consistent as N tends to infinity for fixed T, but have large-N biases of order $1/T$. This situation, known as the "incidental parameter problem", is of particular concern when T is small relative to N (a common situation in applications), and has become one of the main challenges in modern econometrics.[2]

The traditional reaction to this problem has been to look for estimators yielding fixed-T consistency as N goes to infinity.[3] One drawback of these methods is that they are somewhat limited to linear models and certain nonlinear models, often due to the fact that fixed-T identification itself is problematic. Other considerations are that their properties may deteriorate as T increases, and that there may be superior methods that are not fixed-T consistent.[4]

More recently, it has been argued that the incidental parameter problem can be viewed as time-series finite-sample bias when T tends to infinity. Following this perspective, several approaches have been proposed to correct for the time-series bias. These methods include bias-correction of the ML estimator of the common parameters (Hahn and Newey 2004, Hahn and Kuersteiner 2004, Dhaene *et al.*, 2006), of the moment equation (Woutersen 2002, Arellano 2003, Carro 2007) or of the objective function (Arellano and Hahn 2006, 2007, Bester and Hansen 2005a, Hospido 2006), each of them based on analytical or simulation-based approximations.

The aim in this literature has been to obtain estimators of θ with biases of order $1/T^2$ (as opposed to $1/T$) and similar large-sample dispersion as the corresponding uncorrected methods when T/N tends to a constant. This is done in the hope that the reduction in the order of magnitude of the bias will essentially eliminate the incidental parameter

[2] The classic reference on the incidental parameter problem is Neyman and Scott (1948). Lancaster (2000) reviews the history of the problem since then.

[3] See Arellano and Honoré (2001) for a review.

[4] Alvarez and Arellano (2003) showed that standard panel GMM estimators of linear dynamic models are asymptotically biased as T and N increase at the same rate.

problem, even in panels where T is much smaller than N, as long as individual time series are statistically informative.

In this paper, we consider estimators that maximize an average likelihood such as (1) and provide a characterization of the class of weights that produce estimators that are first-order unbiased. Specifically, we consider $\widehat{\theta} = \arg\max_{\theta} \sum_{i=1}^{N} \ln f_i^a(\theta)$ for general weight functions, or priors, $w_i(\alpha_i)$.[5] For fixed T, we can define the pseudo true value $\theta_T = \text{plim}_{N\to\infty}\widehat{\theta}$. In general, $\theta_T \neq \theta_0$. However, expanding in powers of T:

$$\theta_T = \theta_0 + \frac{B}{T} + o\left(\frac{1}{T}\right).$$

We look for priors that yield $B = 0$.

Our results suggest new bias reducing estimators with attractive computational properties, as well as a natural way of obtaining asymptotic confidence intervals. They also provide important insights into the properties of fixed effects, random effects, and Bayesian nonlinear panel estimators in a unified framework.

The approach we follow was first considered in the panel data context by Lancaster (2002) from a Bayesian perspective, in situations where common parameters and fixed effects can be made information orthogonal by reparameterization.[6] Indeed, it can be shown that under information orthogonality taking a uniform prior for the effects reduces the bias on the parameter of interest. In this paper we generalize this approach to situations where orthogonal reparameterizations do not exist.

We start with a characterization of bias reducing priors. For a given weight function or prior, we derive the expression of the $1/T$ term of the bias of the average likelihood relative to an infeasible average likelihood without uncertainty about pseudo true values of the effects for given values of θ. We use this finding to show that there always exist bias reducing weights. This result provides a generalization of Lancaster's approach to a much wider class of models. We also find an expression for the bias of the score of the average or integrated likelihood, which allows us to make the link with information orthogonality. Namely we show that, when (generalized) orthogonal reparameterizations of the fixed effects are not available, bias reducing priors will in general depend on the data.

We discuss two specific data dependent bias reducing priors. The first one, that we call

[5] We shall indistinctly use the terms "weights" and "priors", since in this paper we treat priors as automatic weighting schemes.

[6] The classic paper on information orthogonality is Cox and Reid (1987), and its discussion by Sweeting (1987) makes the connection between orthogonality and inference from the integrated likelihood.

the "robust" prior, can be written as a combination of a Hessian and an outer product of score term. As such it is related to, but different from, the non-subjective prior introduced by Harold Jeffreys. The second bias reducing prior is just the normal approximation to the sampling distribution of the estimated effects for *given* θ:

$$w_i(\alpha_i) \sim \mathcal{N}\left(\widehat{\alpha}_i(\theta), \widehat{\mathrm{Var}}\left[\widehat{\alpha}_i(\theta)\right]\right).$$

The bias reduction property comes from the fact that, contrary to (2), the variability of the fixed effects estimates and its dependence on θ are taken into account.

Given a bias reducing prior, estimation of the common parameters can be performed by integration methods, as well as using Bayesian simulation techniques such as Markov Chain Monte Carlo. The possibility of using computationally efficient techniques for estimation is an appealing feature of the method we propose. In addition, simulation methods can also be useful to compute confidence intervals. Building on Chernozhukov and Hong (2003), we argue that asymptotically valid confidence intervals of the parameter estimates can be read from the quantiles of the posterior distribution of θ when N and T grow at the same rate.

Next we study random effects estimation, which we see as a particular case of the previous analysis when the priors on the individual effects are independent of the common parameters. We find that, in the absence of prior knowledge on the distribution of the individual effects in the population, it is not possible in general to correct for first-order bias. For a given random effects specification, we characterize the set of models for which random effects maximum likelihood (REML) is robust. As an important special case, we derive a necessary and sufficient condition for the Gaussian REML estimator to be bias reducing, which includes the class of linear autoregressive models. In more general nonlinear models, however, the use of Gaussian REML has no bias reducing asymptotic justification.

In contrast, if the random effects family approximates the population distribution of individual effects well, the properties of REML improve. Specifically, we show that the first-order bias of the REML estimator depends on the distance between the distribution of individual effects and its best approximation, in a Kullback-Leibler sense, in the random effects family. This suggests that using a flexible distribution for the effects may reduce the bias on the parameter of interest. As an example, we consider the case of a normal mixture with a number of components that grows with N, and obtain first-order bias reduction of the REML estimator in a model without covariates.

Finally, we study the estimation of averages over individual effects, such as average marginal effects. We compare two estimators. Firstly, the standard random effects estimator, which is inconsistent for large T unless the population distribution of the effects belongs to the chosen family of priors. Secondly, the Bayesian fixed effects (BFE) estimator, defined as the posterior mean of the marginal effect, which is large-T consistent. Thus, in the presence of misspecification, by updating the prior given the data, the bias of marginal effects is reduced by an order of magnitude.

We compute the first-order bias term of BFE estimators of marginal effects. Priors that are bias reducing for the common parameters do not lead in general to bias reduction of marginal effects, and bias reducing priors for marginal effects are specific to the effect considered. The BFE first-order bias depends on the distance between the population distribution of the effects and its best fitting approximation in the chosen family of priors. So, while updating lowers the bias on the marginal effects by an order of magnitude, the bias can be further reduced either by using a bias-reducing prior or a sufficiently close approximating family to the distribution of the effects.

The related literature includes Woutersen (2002), which obtained the first-order bias of the integrated likelihood estimator in the case where parameters are information orthogonal, and proposed a modification of the score when there is no orthogonality. In a contribution closely related to ours, Severini (1999) studies the conditions under which a classical pseudo-likelihood is asymptotically equivalent to some integrated likelihood, corresponding to a given prior distribution for the effects. The conditions he finds can be seen as a special case of our results when parameters are information orthogonal. Some of the results of this paper have been independently obtained by Bester and Hansen (2005b). They consider the form of bias reducing priors for general parametric likelihood models, and provide a data dependent prior, which coincides with one of our proposals, but their focus is not on panel data, and they do not discuss the duality between existence of orthogonal reparameterizations and non-data dependent bias reducing priors. Other important differences are that we provide a formal justification for bias reduction in the panel context, and that we are also concerned with developing a framework where we can study the bias reducing properties of random effects estimators.

The plan of the paper is as follows. In section 2, we derive the expression of the bias of the average likelihood and make the link with information orthogonality. In section 3, we obtain analytical expressions of two special bias reducing weight functions and discuss inference issues. Section 4 focuses on the bias reducing properties of random effects estimators. In section 5 we study the properties of marginal effects. Section 6 illustrates the results by means of two examples: the dynamic AR(p) model and the static logit model with fixed effects. In section 7, we report a small Monte-Carlo simulation to study the finite-sample behavior of the proposed estimators. Lastly, section 8 concludes. The appendix contains proofs of results from sections 2-3 and subsections 4.1-4.2. Proofs of the remaining results, which are of a more technical nature, are in an online supplementary appendix on the journal's website.

2 Biases of the integrated likelihood and score

In this section, we derive the expression of the first-order bias of the integrated likelihood with respect to an arbitrary prior distribution for the individual effects. We start by setting the notation.

2.1 Notation

Let $(y_{it}, x_{it}')'$, $i = 1, ..., N$ and $t = 0, 1, ..., T$ be the set of observations on the endogenous variable y_{it} and a vector of strictly exogenous variables x_{it}, that we assume i.i.d. across individuals. The density of y_{it} conditioned on $(x_{i1}, ..., x_{iT})$ and lagged $y's$ is given by:

$$f_{it}(y_{it}|\theta_0, \alpha_{i0}) \equiv f(y_{it}|x_{it}, y_{i(t-1)}; \theta_0, \alpha_{i0}),$$

which leads to the expression for the scaled individual log-likelihood conditioned on exogenous covariates and initial observations:

$$\ell_i(\theta, \alpha_i) = \frac{1}{T}\sum_{t=1}^{T} \ln f_{it}(y_{it}|\theta, \alpha_i).$$

The likelihood is assumed to depend on a vector of common parameters θ and scalar individual fixed effects $\alpha_1 ... \alpha_N$.[7] Then, let $\pi_i(\alpha_i|\theta)$ be a conditional prior distribution on the individual fixed effect given θ. The conditioning on θ follows from our treatment of α_i as nuisance parameters, while θ are the parameters of interest. Moreover, the subindex i in π_i refers to possible conditioning on strictly exogenous regressors and initial conditions.

Throughout the paper, we will assume that standard regularity conditions are satisfied (e.g., Severini, 1999). In particular, all likelihood and pseudo-likelihood functions as well

[7] Considering further lags and multiple fixed effects would complicate the notation, but leave the essence of what follows unaltered.

as all priors will be three-times differentiable. We will also assume that the prior is not dogmatic in the following sense.

Assumption 1 *The support of $\pi_i(\alpha_i|\theta)$ contains an open neighborhood of the true parameters (α_{i0}, θ_0).*

The prior will generally depend on T. We assume that the order of magnitude of the logarithm of the prior is bounded as T increases:

Assumption 2 *When T tends to infinity we have, for all θ and α_i:*

$$\ln \pi_i(\alpha_i|\theta) = O(1), \quad \textit{uniformly over } i.^{8}$$

Concentrated likelihood. Our analysis makes use of three different objective functions at the individual level. The first one is the concentrated or profile likelihood. It is defined as $\ell_i^c(\theta) = \ell_i(\theta, \widehat{\alpha}_i(\theta))$, where the fixed effects estimates solve $\widehat{\alpha}_i(\theta) = \text{argmax}_{\alpha_i} \ell_i(\theta, \alpha_i)$. Thus, the ML estimator solves $\widehat{\theta}_{ML} = \text{argmax}_{\theta} \sum_{i=1}^{N} \ell_i^c(\theta)$. As is well-known, $\widehat{\theta}_{ML}$ is in general inconsistent for fixed T as $N \to \infty$.

Integrated likelihood. Bias-corrected estimators for θ based on the concentrated likelihood have been recently studied in the statistical and econometric literatures (Arellano and Hahn, 2007). In this paper, we study the behavior of the integrated likelihood with respect to a given prior $\pi_i(\alpha_i|\theta)$. The individual log integrated likelihood is given by:

$$\ell_i^I(\theta) = \frac{1}{T} \ln \int \exp\left[T \ell_i(\theta, \alpha_i)\right] \pi_i(\alpha_i|\theta) d\alpha_i.$$

As noted by Berger *et al.* (1999), this likelihood would be acceptable to a subjective Bayesian whose joint prior is separable in the individual effects, see (3). From this perspective, in this paper we implicitly assume a uniform prior on θ: $\pi(\theta) \propto 1$.[9] Allowing for any non dogmatic prior on θ does not affect the analysis.

Target likelihood. We shall compute the first-order bias of the integrated likelihood relative to a target likelihood without uncertainty about the value of the effects for given θ. Let the target likelihood be $\overline{\ell}_i(\theta) = \ell_i(\theta, \overline{\alpha}_i(\theta))$, where $\overline{\alpha}_i(\theta) = \text{argmax}_{\alpha_i} \text{plim}_{T\to\infty} \ell_i(\theta, \alpha_i)$. This

[8] In what follows, uniformity is implicitly assumed everywhere.

[9] We write $a \propto b$ to denote that a and b are equal up to a multiplicative constant.

function possesses many properties of a proper likelihood. In particular, it is maximized at θ_0 and satisfies Bartlett identities (Severini, 2000). Note that the effects $\overline{\alpha}_i(\theta)$– and as such the likelihood $\overline{\ell}_i(\theta)$– are infeasible. The target likelihood provides a useful theoretical benchmark to compute first-order biases. It is a “least favorable” target likelihood in the sense that the expected information for θ calculated from $\overline{\ell}_i(\theta)$ coincides with the partial expected information.

The concentrated and target likelihood functions can be regarded as integrated likelihood functions with respect to the priors

$$\overline{\pi}_i(\alpha_i|\theta) = \delta\left(\alpha_i - \overline{\alpha}_i(\theta)\right), \text{ and } {\pi_i}^c(\alpha_i|\theta) = \delta\left(\alpha_i - \widehat{\alpha}_i(\theta)\right),$$

respectively. In this perspective, π_i^c can be interpreted as a sample counterpart of $\overline{\pi}_i$. Below, we investigate the existence of non-degenerate feasible counterparts of $\overline{\pi}_i$ that, unlike π_i^c, reduce first-order bias.

Lastly, we denote the observed score with respect to the fixed effect as

$$v_i(\theta,\alpha_i) = \frac{\partial \ell_i(\theta,\alpha_i)}{\partial \alpha_i},$$

and its derivatives as

$$v_i^{\alpha_i}(\theta,\alpha_i) = \frac{\partial v_i(\theta,\alpha_i)}{\partial \alpha_i}, \quad v_i^{\theta}(\theta,\alpha_i) = \frac{\partial v_i(\theta,\alpha_i)}{\partial \theta}, \quad v_i^{\alpha_i\alpha_i}(\theta,\alpha_i) = \frac{\partial^2 v_i(\theta,\alpha_i)}{\partial \alpha_i^2}, \quad \text{etc.}$$

2.2 Bias of the integrated likelihood

We now derive the expression of the first-order bias of the individual integrated likelihood relative to the target likelihood:

$$\mathbb{E}_{\theta_0,\alpha_{i0}}\left[\ell_i^I(\theta) - \overline{\ell}_i(\theta)\right] = C^{st} + \frac{\beta_i(\theta)}{T} + O\left(\frac{1}{T^2}\right),$$

for a given prior $\pi_i(\alpha_i|\theta)$.[10] The expectation is taken with respect to $\exp\left[T\ell_i\left(\theta_0,\alpha_{i0}\right)\right]$, so that a quantity like $\mathbb{E}_{\theta_0,\alpha_{i0}}\left[\ell_i^I(\theta)\right]$ will depend on θ, θ_0 and α_{i0}. We shall proceed in two steps.

In a first step, we use a Laplace approximation (e.g., Tierney *et al.*, 1989) to link the integrated and the concentrated likelihood functions. The result is contained in the following lemma.

[10] Throughout the paper, we use C^{st} to denote any constant term, which depending on the context may be scalar or vector-valued, and stochastic or nonstochastic.

Lemma 1 *Let Assumptions 1 and 2 hold. Then:*

$$\mathbb{E}_{\theta_0,\alpha_{i0}}\left[\ell_i^I(\theta)-\ell_i^c(\theta)\right]=C^{st}-\frac{1}{2T}\ln\mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\alpha_i}(\theta,\overline{\alpha}_i(\theta))\right]+\frac{1}{T}\ln\pi_i(\overline{\alpha}_i(\theta)|\theta)+O\left(\frac{1}{T^2}\right). \quad (4)$$

In a second step we use the formula that gives the first-order bias of the concentrated likelihood (e.g., Arellano and Hahn, 2006, 2007):

$$\mathbb{E}_{\theta_0,\alpha_{i0}}\left[\ell_i^c(\theta)-\overline{\ell}_i(\theta)\right]=\frac{1}{2T}\left\{\mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\alpha_i}(\theta,\overline{\alpha}_i(\theta))\right]\right\}^{-1}\mathbb{E}_{\theta_0,\alpha_{i0}}\left[Tv_i^2(\theta,\overline{\alpha}_i(\theta))\right]+O\left(\frac{1}{T^2}\right). \quad (5)$$

The expression of the first-order bias of the integrated likelihood then follows directly.

Theorem 1 *Let Assumptions 1 and 2 hold. Then:*

$$\mathbb{E}_{\theta_0,\alpha_{i0}}\left[\ell_i^I(\theta)-\overline{\ell}_i(\theta)\right]=C^{st}+\frac{\beta_i(\theta)}{T}+O\left(\frac{1}{T^2}\right)$$

where

$$\begin{aligned}\beta_i(\theta) &= \frac{1}{2}\left\{\mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\alpha_i}(\theta,\overline{\alpha}_i(\theta))\right]\right\}^{-1}\mathbb{E}_{\theta_0,\alpha_{i0}}\left[Tv_i^2(\theta,\overline{\alpha}_i(\theta))\right]\\ &\qquad -\frac{1}{2}\ln\mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\alpha_i}(\theta,\overline{\alpha}_i(\theta))\right]+\ln\pi_i(\overline{\alpha}_i(\theta)|\theta). \quad (6)\end{aligned}$$

As the right-hand side of (6) is $O(1)$, Theorem 1 shows that the effect of the prior vanishes as the amount of data increases. When T goes to infinity, the bias of the integrated likelihood goes to zero irrespective of the prior, provided that the latter is non-dogmatic. In section 4, we will see that this property is shared by random effects panel data models. However, it turns out that the prior has an effect on the first-order bias of the integrated likelihood as, in general, $\beta_i(\theta)$ is not locally constant around θ_0.

2.3 Bias of the integrated score

We start with a definition of robust priors.

Definition 1 *Let $b_i(\theta_0)=\frac{\partial}{\partial\theta}\big|_{\theta_0}\beta_i(\theta)$ be the first-order bias of the integrated score evaluated at the true value. A prior family is said to be bias reducing, or robust, if and only if*

$$b_\infty(\theta_0)\equiv\operatorname*{plim}_{N\to\infty}\ \frac{1}{N}\sum_{i=1}^N b_i(\theta_0)=o\left(1\right).$$

Bias reduction of the moment equation implies bias reduction of the estimator (e.g., Arellano and Hahn, 2006). So, for a robust prior family the mode of the integrated likelihood:

$$\widehat{\theta}_{IML} = \underset{\theta}{\text{argmax}} \sum_{i=1}^{N} \ell_i^I(\theta)$$

has zero first-order bias; that is:

$$\underset{N\to\infty}{\text{plim}} \quad \widehat{\theta}_{IML} = \theta_0 + o\left(\frac{1}{T}\right).$$

We now use the results of the previous subsection to characterize robust priors. From Theorem 1 we can obtain the expression of the bias of the integrated score evaluated at the true value, $b_i(\theta_0)$. It is convenient, in the likelihood context, to use a simplification proposed by Pace and Salvan (2006). At the true value θ_0, where the information matrix equality is satisfied, we have:

$$\begin{aligned} \left.\frac{\partial}{\partial\theta}\right|_{\theta_0} & \left(\left\{\mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\alpha_i}(\theta,\overline{\alpha}_i(\theta))\right]\right\}^{-1} \mathbb{E}_{\theta_0,\alpha_{i0}}\left[Tv_i^2(\theta,\overline{\alpha}_i(\theta))\right] \right) = \\ & \left.\frac{\partial}{\partial\theta}\right|_{\theta_0} \ln\left(\left\{\mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\alpha_i}(\theta,\overline{\alpha}_i(\theta))\right]\right\}^{-1} \mathbb{E}_{\theta_0,\alpha_{i0}}\left[Tv_i^2(\theta,\overline{\alpha}_i(\theta))\right]\right). \end{aligned} \tag{7}$$

The bias of the integrated score is thus given by:

$$b_i(\theta_0) = \left.\frac{\partial}{\partial\theta}\right|_{\theta_0} \ln\pi_i(\overline{\alpha}_i(\theta)|\theta) - \left.\frac{\partial}{\partial\theta}\right|_{\theta_0} \ln\left(\mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\alpha_i}(\theta,\overline{\alpha}_i(\theta))\right]\left\{\mathbb{E}_{\theta_0,\alpha_{i0}}\left[Tv_i^2(\theta,\overline{\alpha}_i(\theta))\right]\right\}^{-1/2}\right). \tag{8}$$

Hence the following result:

Theorem 2 *A prior π_i is bias reducing if:*

$$\left.\frac{\partial}{\partial\theta}\right|_{\theta_0} \ln\pi_i(\overline{\alpha}_i(\theta)|\theta) = \left.\frac{\partial}{\partial\theta}\right|_{\theta_0} \ln\left(\mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\alpha_i}(\theta,\overline{\alpha}_i(\theta))\right]\left\{\mathbb{E}_{\theta_0,\alpha_{i0}}\left[Tv_i^2(\theta,\overline{\alpha}_i(\theta))\right]\right\}^{-1/2}\right) + O\left(\frac{1}{T}\right).$$

Theorem 2 gives a sufficient condition for bias reduction. The reason why the condition is not always necessary is that bias reduction might happen because of cross-sectional averaging, i.e. $b_\infty(\theta_0)$ could be $O(1/T)$ even if some of the $b_i(\theta_0)$, $i = 1...N$, are not. However, the bias reducing priors that we discuss in the next section will satisfy $b_i(\theta_0) = O(1/T)$ for all i.

2.4 Non-distribution dependent bias reducing priors and orthogonality

We turn to consider the role of information orthogonality. The next proposition shows the link between the ability of a prior to reduce bias and information orthogonality.

Proposition 1 *The following equality holds:*

$$b_i(\theta_0) = \left.\frac{\partial}{\partial \theta}\right|_{\theta_0} \ln \pi_i(\overline{\alpha}_i(\theta)|\theta) + \left.\frac{\partial}{\partial \alpha_i}\right|_{\alpha_{i0}} \rho_i(\theta_0, \alpha_i) \tag{9}$$

where

$$\rho_i(\theta, \alpha_i) \equiv \left\{\mathbb{E}_{\theta,\alpha_i}\left[-v_i^{\alpha_i}(\theta, \alpha_i)\right]\right\}^{-1} \mathbb{E}_{\theta,\alpha_i}\left[v_i^{\theta}(\theta, \alpha_i)\right].$$

Proposition 1 shows that the quantity $\rho_i(\theta, \alpha_i)$, the projection coefficient in the efficient score for θ, is key in the ability of a given prior to reduce bias. A particular case is the one of information orthogonality studied by Cox and Reid (1987) and Lancaster (2002). In that case the information matrix is block diagonal so that $\mathbb{E}_{\theta,\alpha_i}\left[v_i^{\theta}(\theta, \alpha_i)\right]$ is identically zero. It follows from Proposition 1 that the uniform prior $\pi_i(\alpha_i|\theta) \propto 1$ is bias reducing. The same is true of all priors that are independent of θ in light of Proposition 1 and the fact that

$$\left.\frac{\partial \overline{\alpha}_i(\theta)}{\partial \theta}\right|_{\theta_0} = \rho_i(\theta_0, \alpha_{i0}). \tag{10}$$

Conversely, Proposition 1 implies that the uniform prior reduces bias if and only if:

$$\operatorname*{plim}_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \left.\frac{\partial}{\partial \alpha_i}\right|_{\alpha_{i0}} \rho_i(\theta_0, \alpha_i) = o(1). \tag{11}$$

Condition (11) is slightly more general than information orthogonality. For it to be satisfied, it suffices that $\rho_i(\theta, \alpha_i)$ is a function of θ only.

The uniform prior does not depend on the distribution of the data. That is, it is independent of the true parameters $\theta_0, \alpha_{10}, ..., \alpha_{N0}$. We shall refer to the (infeasible) weighting schemes that depend on the true values of the parameters as distribution dependent. In particular, the uniform prior is non-distribution dependent.

Other non-distribution dependent priors are given by orthogonal reparameterizations of the fixed effects, when available. Let $\psi_i = \psi_i(\alpha_i, \theta)$ be a reparameterization of the individual effects. To every prior $\widetilde{\pi}_i(\psi_i|\theta)$ on ψ_i we can associate the transformed prior in the original parameterization:

$$\pi_i(\alpha_i|\theta) = \widetilde{\pi}_i(\psi_i(\alpha_i, \theta)|\theta) \left|\frac{\partial \psi_i(\alpha_i, \theta)}{\partial \alpha_i}\right|.$$

The following result shows that the bias reducing properties of a prior are not affected by a reparameterization of the effects.

Proposition 2 $\widetilde{\pi}_i$ *is bias reducing in the transformed parameterization* ψ_i *if and only if* π_i *is bias reducing in the original parameterization* α_i.

We now apply Proposition 2 to a reparameterization $\psi_i = \psi_i(\alpha_i, \theta)$ such that ψ_i and θ are information orthogonal in the sense of equation (11). In this case the uniform prior on ψ_i is bias reducing. Hence, using Proposition 2, the transformed prior on α_i:

$$\pi_i(\alpha_i|\theta) = \left|\frac{\partial \psi_i(\alpha_i, \theta)}{\partial \alpha_i}\right|$$

is also bias reducing. Remark that this prior is the Jacobian of the transformation which maps (α_i, θ) onto (ψ_i, θ). Conversely, any non-distribution dependent bias reducing prior $\pi_i(\alpha_i|\theta)$ can be associated an orthogonal reparameterization in the sense of equation (11). It suffices to take $\psi_i = \psi_i(\alpha_i, \theta)$, where:

$$\psi_i(\alpha_i, \theta) = \int_{-\infty}^{\alpha_i} \pi_i(\alpha|\theta) d\alpha.$$

This discussion shows that there exists a mapping between non-distribution dependent bias reducing priors and orthogonal reparameterizations in the sense of (11). Now, such reparameterizations do not always exist. In the multiparameter case (when θ is a vector) one ends up with a partial differential equation which has no solution in general, in close analogy with the case of strict information orthogonality (Cox and Reid, 1987). Hence, to deal with the case where orthogonal reparameterizations are not available, it is in general necessary to search for robust priors that depend on the distribution of the data. We address this task in the next section.

3 Constructive bias reducing priors

In this section we discuss two specific data dependent priors that are bias reducing independently of the possibility of orthogonalization.

3.1 A robust prior

Theorem 2 shows that the following prior is bias reducing:

$$\pi_i^R(\alpha_i|\theta) \propto \widehat{\mathbb{E}}\left[-v_i^{\alpha_i}(\theta, \alpha_i)\right] \left\{\widehat{\mathbb{E}}\left[v_i^2(\theta, \alpha_i)\right]\right\}^{-1/2} \qquad (12)$$

where $\widehat{\mathbb{E}}\left[-v_i^{\alpha_i}(\theta, \alpha_i)\right]$ and $\widehat{\mathbb{E}}\left[v_i^2(\theta, \alpha_i)\right]$ are consistent estimates of $\mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\alpha_i}(\theta, \alpha_i)\right]$ and $\mathbb{E}_{\theta_0,\alpha_{i0}}\left[v_i^2(\theta, \alpha_i)\right]$, respectively, when T tends to infinity. Remark that replacing the expectations by large-T consistent estimates in the condition of Theorem 2 does not affect the result.[11]

The bias reducing prior (12), which we call the "robust" prior, depends on the data. The discussion in the previous section has shown that non-data dependent priors are generally not robust in cases when orthogonal reparameterizations of the fixed effects are not available.[12]

Moreover, π_i^R is the combination of a Hessian term ($\widehat{\mathbb{E}}\left[-v_i^{\alpha_i}(\theta,\alpha_i)\right]$) and a outer product term ($\widehat{\mathbb{E}}\left[v_i^2(\theta,\alpha_i)\right]$). A closely related expression appears in Jeffreys' automatic prior when θ is kept fixed, the expression of which is:

$$\pi_i^J(\alpha_i|\theta) \propto \left\{\mathbb{E}_{\theta,\alpha_i}\left[-v_i^{\alpha_i}(\theta,\alpha_i)\right]\right\}^{1/2}. \tag{13}$$

A crucial difference between $\pi_i^R(\alpha_i|\theta)$ and $\pi_i^J(\alpha_i|\theta)$ is that Jeffreys' prior does not depend on the data. In fact, Jeffreys' prior (13) is generally not bias reducing (see Hahn, 2004).

Before ending this discussion, note that we have assumed a likelihood set-up, as opposed to a pseudo-likelihood set-up. The likelihood assumption is required to obtain equation (7), which uses the information identity at true parameter values. In the pseudo-likelihood case, however, it is still possible to use Theorem 1 to obtain a robust weighting scheme for an integrated objective function. In effect, using the expression of the bias of the integrated likelihood (6), it is straightforward to show that the following prior is bias reducing in both likelihood and pseudo-likelihood settings:

$$\left\{\widehat{\mathbb{E}}\left[-v_i^{\alpha_i}(\theta,\alpha_i)\right]\right\}^{1/2}\exp\left(-\frac{T}{2}\left\{\widehat{\mathbb{E}}\left[-v_i^{\alpha_i}(\theta,\alpha_i)\right]\right\}^{-1}\widehat{\mathbb{E}}\left[v_i^2(\theta,\alpha_i)\right]\right). \tag{14}$$

Coming back to the likelihood set-up, note that Proposition 1 shows that many other priors are robust. In particular, the two priors given by (12) and (14) are bias reducing. Using (14) instead of (12) for estimation can make a difference in finite samples. The Monte Carlo simulations reported below will illustrate this remark.

3.2 Robust reparameterizations

The following result provides an additional characterization of the robust prior.

[11] Thus, the problem of computing bias reducing priors is analogous to the problem of estimating an additive bias correction to the concentrated likelihood. See for example Hahn and Kuersteiner (2004), Arellano and Hahn (2006, 2007), and Pace and Salvan (2006).

[12] This result is in a similar spirit to one in Wasserman (2000), which shows that for certain mixture models data dependent priors are the only priors that produce intervals with second-order frequentist coverage.

Proposition 3 *We have:*

$$\pi_i^R(\widehat{\alpha}_i(\theta)|\theta) \propto \frac{1}{\sqrt{\widehat{\mathrm{Var}}\left(\widehat{\alpha}_i(\theta)\right)}}\left(1+O_p\left(\frac{1}{T}\right)\right) \tag{15}$$

where $T\widehat{\mathrm{Var}}\left(\widehat{\alpha}_i(\theta)\right)$ is a consistent estimate of the asymptotic variance of $\sqrt{T}\left(\widehat{\alpha}_i(\theta)-\overline{\alpha}_i(\theta)\right)$ when T tends to infinity. In addition, every non-dogmatic prior satisfying (15) is bias reducing.

Proposition (3) sheds some light on the properties of the robust prior. To see why, let us consider the reparameterization:

$$\psi_i(\alpha_i,\theta) = \frac{\alpha_i - \widehat{\alpha}_i(\theta)}{\sqrt{\widehat{\mathrm{Var}}\left(\widehat{\alpha}_i(\theta)\right)}}. \tag{16}$$

Reparameterizing the individual effects as in (16) amounts to rescaling the effects, weighting them in inverse proportion to the standard deviation of the fixed effects MLE.

Specifically, let us consider a prior on ψ_i that is independent of θ, with *pdf* f. In terms of the original parameterization, the prior is:[13]

$$\widetilde{\pi}_i^R(\alpha_i|\theta) = \frac{1}{\sqrt{\widehat{\mathrm{Var}}\left(\widehat{\alpha}_i(\theta)\right)}} f\left(\frac{\alpha_i - \widehat{\alpha}_i(\theta)}{\sqrt{\widehat{\mathrm{Var}}\left(\widehat{\alpha}_i(\theta)\right)}}\right).$$

Then, clearly:

$$\widetilde{\pi}_i^R(\widehat{\alpha}_i(\theta)|\theta) \propto \frac{1}{\sqrt{\widehat{\mathrm{Var}}\left(\widehat{\alpha}_i(\theta)\right)}}.$$

It thus follows from Proposition 3 that $\widetilde{\pi}_i^R$ is bias reducing.

For the particular choice of $\psi_i \sim \mathcal{N}(0,1)$, we obtain the result that the normal approximation to the sampling distribution of the MLE $\widehat{\alpha}_i(\theta)$ is a bias reducing weighting scheme for α_i:

$$\alpha_i|\theta \sim \mathcal{N}(\widehat{\alpha}_i(\theta), \widehat{\mathrm{Var}}\left(\widehat{\alpha}_i(\theta)\right)). \tag{17}$$

Specifying a prior distribution on the fixed effects as in (17) is intuitively appealing from the point of view of bias reduction. First, unlike the robust prior (π_i^R), this prior is proper, so that it will unambiguously lead to a proper posterior. Second, it can be seen as a feasible counterpart of the (degenerate) prior associated to the target likelihood ($\overline{\pi}_i$). Unlike the prior

[13] Note that $\widetilde{\pi}_i^R$ does not satisfy Assumption 2. This does not matter for the present discussion, however, as shown by the proof of Proposition 3.

associated with the concentrated likelihood (π_i^c), it takes into account the way the precision of $\widehat{\alpha}_i(\theta)$ varies with θ. When $\mathrm{Var}\left(\widehat{\alpha}_i(\theta)\right)$ varies slowly with θ, the uniform prior on the original effects is bias reducing. This happens when parameters are information orthogonal.

3.3 Asymptotic distribution and inference

Here we derive the asymptotic distribution of the integrated likelihood estimator, and discuss how to perform inference from the posterior distribution of θ.

Let $\ell_i^I(\theta)$ be associated with a bias reducing prior. Let $\widehat{\theta}_{IML} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^N \ell_i^I(\theta)$ be the mode of the integrated likelihood. We are interested in the asymptotic distribution of $\widehat{\theta}_{IML}$ when N and T tend simultaneously to infinity at the same rate: $T/N \to C^{st} > 0$.

Let $\overline{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^N \overline{\ell}_i(\theta)$ be the (infeasible) mode of the target likelihood. Because the prior is bias reducing, we have:

$$\widehat{\theta}_{IML} = \overline{\theta} + o_p\left(\frac{1}{T}\right).$$

So, when N and T tend to infinity at the same rate:

$$\sqrt{NT}\left(\widehat{\theta}_{IML} - \overline{\theta}\right) = o_p(1).$$

The mode of the integrated likelihood and the mode of the target likelihood are thus asymptotically equivalent. In particular, the asymptotic variance of $\sqrt{NT}\left(\widehat{\theta}_{IML} - \theta_0\right)$ is equal to that of $\sqrt{NT}\left(\overline{\theta} - \theta_0\right)$. Now, $\overline{\theta}$ has the same asymptotic dispersion as the maximum likelihood estimator $\widehat{\theta}_{ML}$. So, as in the case of the additive approaches to bias reduction (Hahn and Newey, 2004), bias reduction occurs with no increase in the asymptotic variance relative to fixed effects maximum likelihood.

Given a robust weighting scheme, estimation based on the integrated likelihood can be performed using classical or Bayesian techniques. For this purpose, one can use integration routines (quadrature, Monte Carlo) to compute the integrated likelihood, and then maximize the latter using optimization algorithms. This is the approach we have adopted in the Monte Carlo experiments reported below. However, in highly nonlinear models with possibly many parameters, this approach can be problematic. Our connection to Bayesian statistics makes it possible to use Bayesian techniques, such as Markov Chain Monte Carlo, to perform the estimation.

Moreover, an additional appealing feature of the simulation approach is the ability to read confidence intervals directly from the posterior distribution. Following Chernozhukov and Hong (2003), it can be shown that, in a double asymptotics perspective when N and T tend to infinity at the same rate, the quantiles of the posterior distribution of θ provide asymptotically valid confidence intervals for θ_0. Indeed, the marginal posterior of θ can be interpreted as a pseudo-posterior calculated from the integrated likelihood. Moreover, this objective function satisfies a generalized information equality in a double asymptotic sense.

4 Random effects and bias reduction

In this section, we study the first-order bias properties of random effects maximum likelihood (REML) estimators.

4.1 The random effects model

We assume that α_{i0}, $i = 1...N$, are drawn from a distribution with density π_0 conditioned on covariates and initial observations. The marginal density of an observation is thus given by

$$f_i(y_{i1}, ..., y_{iT}|y_{i0}, \theta_0, \pi_0) = \int \prod_{t=1}^{T} f(y_{it}|x_{it}, y_{i(t-1)}; \theta_0, \alpha_i)\pi_0(\alpha_i)d\alpha_i.$$

This model is very common in the panel data literature. Often, π_0 is supposed to belong to a known parametric family such as the normal or a multinomial distribution with a finite number of mass points, possibly independent of covariates. In contrast, here we make no assumption about the functional form of π_0.

Let ξ be a parameter and $\pi_i(\alpha_i; \xi)$ be a family of prior distributions indexed by ξ. A typical example is when $\pi(\alpha_i; \xi)$ is a normal distribution with unknown mean and variance, $\xi = (m, s^2)$. Importantly, $\pi_i(\alpha_i; \xi)$ does not depend directly on the common parameter θ, nor on the *cdf* of the distribution of the data (that is, on the true parameters θ_0, α_{i0}). Nevertheless, we do allow π_i to depend on conditioning covariates and/or initial conditions. For example, the mean and variance of the normal m and s^2 may be functions of covariates and/or initial conditions as in Chamberlain (1984).

The function $\pi_i(\alpha_i; \xi)$ has two possible interpretations. It can be regarded as a model for the population distribution of α_{i0}; this is the "random effects" perspective. In a Bayesian perspective, it can also be seen as a hierarchical prior assuming independence between α_i and θ. In both approaches, we are interested in the random effects pseudo-likelihood:

$$\ell_i^{RE}(\theta; \xi) = \frac{1}{T} \ln \int \exp\left[T\ell_i(\theta, \alpha_i)\right] \pi_i(\alpha_i; \xi)d\alpha_i,$$

which is the integrated likelihood with respect to the prior $\pi_i(\alpha_i;\xi)$.

4.2 Robust random effects

Here we study the existence of random effects specifications that are bias reducing for any population distribution of the individual effects π_0.[14]

It is convenient to start by concentrating the likelihood with respect to ξ. Let:

$$\widehat{\xi}(\theta)=\underset{\xi}{\operatorname{argmax}}\sum_{i=1}^{N}\ell_i^{RE}(\theta;\xi).$$

The score of the concentrated random effects likelihood is given by:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\theta}\bigg|_{\theta_0}\ell_i^{RE}(\theta;\widehat{\xi}(\theta))=\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\theta}\bigg|_{\theta_0}\ell_i^{RE}(\theta;\widehat{\xi}(\theta_0)),$$

where the equality comes from the envelope theorem.

The bias of the score of the concentrated random effects likelihood is thus:

$$\begin{aligned}b_\infty(\theta_0)&=\underset{N\to\infty}{\operatorname{plim}}\ \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\theta}\bigg|_{\theta_0}\ell_i^{RE}(\theta;\widehat{\xi}(\theta_0))\\&=\underset{N\to\infty}{\operatorname{plim}}\ \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\pi_0}\left(\frac{\partial}{\partial\theta}\bigg|_{\theta_0}\ell_i^{RE}(\theta;\overline{\xi}(\theta_0))\right),\end{aligned}\tag{18}$$

where: $\overline{\xi}(\theta)=\underset{N\to\infty}{\operatorname{plim}}\left(\widehat{\xi}(\theta)\right)$. The following result helps to interpret the pseudo true value $\overline{\xi}(\theta_0)$.

Lemma 2 *For all θ, we have:*

$$\underset{N\to\infty}{\operatorname{plim}}\ \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\pi_0}\left(\frac{\partial\ln\pi_i(\overline{\alpha}_i(\theta);\overline{\xi}(\theta))}{\partial\xi}\right)=O\left(\frac{1}{T}\right).\tag{19}$$

Lemma 2 provides a heuristic interpretation of $\overline{\xi}(\theta)$, up to a $O(1/T)$ term, as the pseudo true value of ξ for the model $\pi_i(.;\xi)$ and the "data" $\overline{\alpha}_1(\theta),...,\overline{\alpha}_N(\theta)$. Evaluated at $\theta=\theta_0$, equation (19) shows that $\pi_i\left(.;\overline{\xi}(\theta_0)\right)$ is the best approximation to π_0, in a Kullback-Leibler sense, in the family $\pi_i(.;\xi)$. In the next subsection, we will see that the distance between π_0 and its best approximation also matters for bias reduction.

[14] In general, π_0 is conditional on covariates and initial conditions, but for simplicity our notation does not make explicit that π_0 may be unit-specific.

Equation (18) shows that the first-order bias properties of the random effects likelihood are the same as the ones of an integrated likelihood with prior $\pi_i(\alpha_i;\overline{\xi}(\theta_0))$. In particular, using Proposition 1 we obtain:

$$b_\infty(\theta_0) \;=\; \operatorname*{plim}_{N\to\infty} \;\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\pi_0}\left(\left.\frac{\partial}{\partial\theta}\right|_{\theta_0}\ln\pi_i\left(\overline{\alpha}_i(\theta);\overline{\xi}(\theta_0)\right)+\left.\frac{\partial}{\partial\alpha_i}\right|_{\alpha_{i0}}\rho_i(\theta_0,\alpha_i)\right). \tag{20}$$

So, using (20) together with equation (10) and rearranging, we find that REML is first-order bias reducing if and only if:

$$\operatorname*{plim}_{N\to\infty}\;\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\pi_0}\left(\frac{1}{\pi_i\left(\alpha_{i0};\overline{\xi}(\theta_0)\right)}\left.\frac{\partial}{\partial\alpha_i}\right|_{\alpha_{i0}}\pi_i(\alpha_i;\overline{\xi}(\theta_0))\rho_i(\theta_0,\alpha_i)\right)=o\left(1\right). \tag{21}$$

A first implication of (21) is that, if the common parameters and the individual effects are information orthogonal, then every REML estimator is bias reducing. This is because in this case $\rho_i(\theta,\alpha)=0$ is identically zero.

Another case where REML is bias reducing is when π_0 belongs to the parametric family $\pi_i(.;\xi)$. Then the random effects model is correctly specified. So, under standard identification conditions, the REML estimator is fixed-T consistent, hence bias reducing.

Moreover, equation (21) allows to characterize the set of models for which a given random effects specification is bias reducing, as shown by the following theorem.

Theorem 3 *Let $\pi_i(.;\xi)$ be a random effects specification depending on a q-dimensional vector of hyperparameters ξ. Then REML is bias reducing for all π_0 and covariate distributions if and only if there exists a constant* $\dim(\theta)\times q$ *matrix $\Gamma(\theta)$ such that:*

$$\left.\frac{\partial}{\partial\alpha}\right|_{\alpha_i}\rho_i\left(\theta,\alpha\right)\pi_i\left(\alpha;\overline{\xi}\left(\theta\right)\right)=\Gamma(\theta)\left.\frac{\partial}{\partial\xi}\right|_{\overline{\xi}(\theta)}\pi_i\left(\alpha_i;\xi\right)+o\left(1\right). \tag{22}$$

Theorem 3 shows that, for a given random effects family, the set of models where there is bias reduction is limited: it corresponds to ρ_i being a linear combination of q functions, where q is the number of hyperparameters. As an important special case, we mention the following corollary.

Corollary 1 ***(uncorrelated random effects)*** *REML based on a location-scale family reduces first-order bias for all π_0 and covariate distributions if and only if there exist $\gamma_1(\theta)$ and $\gamma_2(\theta)$ such that:*

$$\rho_i(\theta,\alpha_i)=\gamma_1(\theta)+\gamma_2(\theta)\alpha_i+o\left(1\right). \tag{23}$$

Corollary 1 gives a necessary and sufficient condition for REML based on a location-scale family to reduce bias. In the corollary, the mean and variance hyperparameters are independent of x_i. We also have the following result, where we let the mean depend linearly on x_i (correlated random effects).[15]

Corollary 2 ***(correlated random effects)*** *REML based on a location-scale family with mean depending linearly on x_i reduces first-order bias for all π_0 and covariate distributions if and only if there exist $\gamma_1(\theta)$ and $\gamma_2(\theta)$ such that:*

$$\rho_i(\theta, \alpha_i) = \gamma_1(\theta) x_i + \gamma_2(\theta)\alpha_i + o(1). \tag{24}$$

In particular, these results apply to Gaussian REML. Section 6 will give examples of models that satisfy conditions (23) or (24), such as dynamic AR(p) models with or without strictly exogenous regressors. In these models, the bias of REML based on the Gaussian family is of order $1/T^2$. Still, most models do not satisfy conditions (23) or (24). In those cases, the bias of the Gaussian REML estimator is of order $1/T$.

Corollaries 1 and 2 are interestingly related to the minimax finite sample result obtained by Chamberlain and Moreira (2008). Using a very different perspective, our results also emphasize the importance of the model's linearity in order for Gaussian REML to have good properties.

4.3 Flexible random effects

In the previous subsection we asked the question: Given a random effects family of priors, what is the set of models in which REML is robust for any population distribution of the individual effects? In particular, we required bias reduction to hold even if the population distribution π_0 was very poorly approximated by the parametric family of prior distributions $\pi_i(.;\xi)$. In contrast, here we ask: Is it possible to reduce the bias on θ by choosing a family of priors that approximates π_0 "sufficiently well"? Our motivation comes from the fact that, in the absence of misspecification, that is when π_0 belongs to the chosen family of prior distributions, the bias is zero.

To answer this question, it is convenient to define the following objects:

$$\overline{\xi}_0 = \underset{\xi}{\text{argmax}} \quad \mathbb{E}_{\pi_0}\left(\ln \pi\left(\alpha_{i0};\xi\right)\right), \qquad \text{and:} \qquad \widetilde{\pi}_0 \equiv \pi\left(.;\overline{\xi}_0\right).$$

[15] As in Chamberlain's (1984) random effects probit, for example.

$\overline{\xi}_0$ is the infeasible ML estimand of ξ, for the "data" $\alpha_{10}, ..., \alpha_{N0}$. So $\widetilde{\pi}_0$ is the best approximation to π_0, in a Kullback-Leibler sense, in the family $\pi(.;\xi)$. Note that both $\overline{\xi}_0$ and $\widetilde{\pi}_0$ are theoretical objects.[16] Remark also that we have assumed for expositional simplicity that $\pi_i(.;\xi) \equiv \pi(.;\xi)$ does not depend on covariates. We come back to this point at the end of this subsection.

It is also convenient to define, for a density p:

$$\mathcal{K}(\pi_0, p) = \left[\mathbb{E}_{\pi_0}\left(\ln \frac{p(\alpha_{i0})}{\pi_0(\alpha_{i0})}\right)^2\right]^{1/2}.$$

$\mathcal{K}(\pi_0, p)$ is the L^2 Kullback-Leibler loss. We will use it to measure how close the true π_0 and its best parametric approximation $\widetilde{\pi}_0$ are.

Let

$$\widehat{\theta}_{REML} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \ell_i^{RE}\left(\theta, \widehat{\xi}(\theta)\right)$$

be the REML estimator, and let $\overline{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \overline{\ell}_i(\theta)$ be the infeasible mode of the target likelihood. Unlike that of $\overline{\theta}$, the asymptotic distribution of $\widehat{\theta}_{REML}$ is generally not centered at zero. The following theorem shows that the bias in the asymptotic distribution of $\widehat{\theta}_{REML}$ depends on the discrepancy between the true density π_0 and its best fitting approximation $\widetilde{\pi}_0$, as measured by the L^2 Kullback-Leibler loss. The theorem requires some conditions on the tails of π_0 that we detail in the supplementary appendix, together with its proof.

Theorem 4 *Let N and T tend to infinity such that $N/T \to C^{st}$. Under suitable regularity conditions:*

$$\sqrt{NT}\left(\widehat{\theta}_{REML} - \theta_0\right) \quad = \quad \sqrt{NT}\left(\overline{\theta} - \theta_0\right) + O\left(\mathcal{K}(\pi_0, \widetilde{\pi}_0)\right) + o_p\left(1\right).$$

Theorem 4 shows that if the distance between π_0 and its best parametric approximation $\widetilde{\pi}_0$ is $o(1)$, then the REML estimator is first-order unbiased and has the same asymptotic variance as the fixed effects estimator.

As a special case, Theorem 4 implies that $\widehat{\theta}_{REML}$ and $\overline{\theta}$ are asymptotically equivalent if the model is correctly specified and π_0 belongs to the parametric family $\pi(.;\xi)$.

More interestingly, the result in Theorem 4 also suggests that, for a flexible choice of $\pi(.;\xi)$, one should be able to obtain asymptotically unbiased inference on θ. The following

[16] Note also that $\overline{\xi}_0$ does not coincide with $\overline{\xi}(\theta_0)$, although due to (19) their difference is $O(1/T)$.

result formalizes this intuition in the case of normal mixtures. For this purpose, we adopt the set-up in Ghosal and Van der Vaart (2001).

Corollary 3 *Assume that π_0 can be expressed as a mixture of normals of the form:*

$$\pi_0(\alpha) = \int \frac{1}{\sigma}\varphi\left(\frac{\alpha-\mu}{\sigma}\right) dH_0(\mu,\sigma)$$

where $\sigma \in [\underline{\sigma}, \overline{\sigma}]$ belongs to a compact interval. Let π be the pdf of a finite mixture of K normal components:

$$\pi(\alpha) = \sum_{k=1}^{K} p_k \frac{1}{\sigma_k}\varphi\left(\frac{\alpha-\mu_k}{\sigma_k}\right)$$

where $p_k \geq 0$, $\sum_{k=1}^{K} p_k = 1$ and $\mu_k \in [-A, A]$, with $A = O\left((\ln N)^{\nu}\right)$ for some $\nu > 0$. Assume also that there exists $\delta \in]0,1]$ such that:[17]

$$\int_{\pi_0(\alpha)/\widetilde{\pi}_0(\alpha)\geq e^{1/\delta}} \left(\frac{\pi_0(\alpha)}{\widetilde{\pi}_0(\alpha)}\right)^{\delta} \pi_0(\alpha)d\alpha < \infty. \qquad (25)$$

Then, for $K \geq C \ln N$ with C large enough

$$\mathcal{K}(\pi_0, \widetilde{\pi}_0) = O\left(N^{-\frac{1}{2}+\gamma}\right), \qquad \textit{for any } \gamma > 0. \qquad (26)$$

So, when N, T tend to infinity such that $N/T \to C^{st}$:

$$\sqrt{NT}\left(\widehat{\theta}_{REML} - \theta_0\right) \quad = \quad \sqrt{NT}\left(\overline{\theta} - \theta_0\right) + o_p(1). \qquad (27)$$

Corollary 3 shows that, in the case where π_0 is a mixture of normals, the rate of convergence of the discrete sieve MLE is almost root-N in (26). As noted by Ghosal and Van der Vaart (2007) this near-parametric rate is driven by the assumptions on π_0. Working under much weaker assumptions, Ghosal and Van der Vaart (2007) find convergence rates of sieve MLEs that are close to the rate of nonparametric kernel estimators $O(N^{-2/5})$. Applied to the case of finite mixtures of normals, their results imply that (27) holds for REML based on a normal mixture with a sufficiently large number of components, under much weaker assumptions on π_0. Indeed, for (27) to hold we only need that $\mathcal{K}(\pi_0, \widetilde{\pi}_0) = o(1)$, and do not require a specific convergence rate.

[17] Condition (25) imposes that the tails of $\widetilde{\pi}_0$ are not too thin relative to that of π_0. We need this condition because Ghosal and Van der Vaart (2001) bound the Hellinger distance between the two distributions (i.e., the L^2 distance of square roots), while we need to bound the L^2 Kullback-Leibler loss. A useful inequality between the two distances is given in Wong and Shen (1995). Also remark that (25) is clearly satisfied if π_0 is compactly supported.

Importantly, all results in this section are stated under the assumption that π and π_0 do not depend on covariates, or that covariates are discrete and the analysis is conducted for specific values. If π_0 depends on more general x's, then the statements of the theorem and corollary will still hold, provided that we let $\pi_i(.;\xi)$ depend in an unrestricted way on x_i.

5 Policy parameters: marginal effects

5.1 Estimating marginal effects

In this section we study the bias properties of some estimators of averages over individual effects, such as average marginal effects. We consider quantities of the form:

$$M = \frac{1}{N}\sum_{i=1}^{N} m_i(\theta_0, \alpha_{i0}).$$

A first example is the marginal effect of a covariate in a probit or logit model, e.g. for probit: $m_i(\theta,\alpha_i) = \theta_k \frac{1}{T}\sum_{t=1}^{T}\varphi\left(x_{it}'\theta + \alpha_i\right)$, where φ is the $\mathcal{N}(0,1)$ density. Other examples are moments of the distribution of individual effects: $m_i(\theta,\alpha_i) = \alpha_i^k$.

A standard fixed effects estimator of M is given by

$$\widehat{M}_{FE} = \frac{1}{N}\sum_{i=1}^{N} m_i\left(\widehat{\theta}, \widehat{\alpha}_i\left(\widehat{\theta}\right)\right)$$

where $\widehat{\alpha}_i(\theta)$ is the MLE of α_i given θ, and $\widehat{\theta}$ is a possibly bias reducing estimator of θ. This estimator was studied by Hahn and Newey (2004). Whether $\widehat{\theta}$ is bias corrected or not $\widehat{M}_{FE}$ has generally a non-zero first-order bias term. Hahn and Newey suggest an approach to bias correct the marginal effects also, and obtain a bias of order $1/T^2$.

We consider two other estimators of M. In a random effects framework with family $\pi_i(.;\xi)$ we may consider the standard random effects estimator given by:

$$\widehat{M}_{RE} = \frac{1}{N}\sum_{i=1}^{N}\int m_i(\widehat{\theta},\alpha_i)\pi_i\left(\alpha_i;\widehat{\xi}\left(\widehat{\theta}\right)\right)d\alpha_i$$

where $\widehat{\theta}$ is a large-T consistent estimator of θ, for example the REML estimator, and $\widehat{\xi}(\theta)$ is the MLE of ξ given θ.

More generally, assuming a family of prior distributions $\pi_i(\alpha_i|\theta)$ we can consider a Bayesian fixed effects (BFE) estimator of M as:

$$\widehat{M}_{BFE} \quad = \quad \int ... \int \left(\frac{1}{N}\sum_{i=1}^{N} m_i(\theta,\alpha_i)\right) p(\alpha_1,...,\alpha_N,\theta|y,x)d\alpha_1...d\alpha_N d\theta$$

where p is the posterior distribution of the model's parameters given the data. $\widehat{M}_{BFE}$ is the posterior mean of $\frac{1}{N}\sum_{i=1}^{N} m_i(\theta,\alpha_i)$. One could as well consider the posterior mode. As before, assuming a non-flat prior on θ does not affect the large-T bias or the asymptotic distribution of the estimator.[18]

5.2 Bayesian fixed effects estimation

The following theorem gives the large-T bias of the BFE estimator $\widehat{M}_{BFE}$.

Theorem 5 *When T tends to infinity:*

$$\operatorname*{plim}_{N\to\infty} \left(\widehat{M}_{BFE} - M\right) = \frac{B_M}{T} + o\left(\frac{1}{T}\right),$$

where

$$\begin{aligned} B_M &= \left[\operatorname*{plim}_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial\theta}\bigg|_{\theta_0} m_i(\theta,\alpha_{i0})\right] B \\ &\quad + \operatorname*{plim}_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\pi_i(\alpha_{i0}|\theta_0)} \frac{\partial}{\partial\alpha}\bigg|_{\alpha_{i0}} \left[\mathbb{E}_{\theta_0,\alpha_{i0}}\left(-v_i^{\alpha_i}(\theta_0,\alpha)\right)\right]^{-1} \pi_i(\alpha|\theta_0) m_i^{\alpha_i}(\theta_0,\alpha), \end{aligned}$$

and B is the first-order bias of the mode of the integrated likelihood (or, equivalently, of the posterior mean of θ).

Theorem 5 shows that the BFE estimator of M is large-T consistent, independently of π_i, and gives an expression of the first-order bias. It follows that taking a robust prior on α_i leads to first-order unbiasedness for θ ($B = 0$), but not for M in general ($B_M \neq 0$). An exception where the two bias terms are zero occurs when $M = m(\theta_0)$ does not depend on the individual effects. So the properties of the BFE estimator are similar to those of the standard fixed effects estimator.

As in the case of common parameters θ, one may look for priors on α_i that yield $B_M = 0$. If parameters are information orthogonal the uniform prior is not bias reducing for M, if the marginal effect depends on individual effects. Instead, one may consider:

$$\pi_i^m(\alpha_i) = \frac{1}{m_i^{\alpha_i}(\theta_0,\alpha_i)} \mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\alpha_i}(\theta_0,\alpha_i)\right]. \qquad (28)$$

[18] In a random effects model, we could also consider another estimator, that one could refer to as "Bayesian random effects", namely the posterior mean or mode of $\frac{1}{N}\sum_{i=1}^{N}\int m_i(\theta,\alpha_i)\pi_i(\alpha_i;\xi)\,d\alpha_i$. Using a Laplace approximation, it is easy to show that this estimator is asymptotically equivalent to $\widehat{M}_{RE}$ when N and T tend to infinity at the same rate.

Under information orthogonality, π_i^m is both bias reducing for θ and M. In the general case, one can verify that the following prior is robust for θ and M simultaneously:

$$\pi_i^{R,m}(\alpha_i|\theta) = \frac{m_i^{\alpha_i}(\theta_0, \overline{\alpha}_i(\theta))}{m_i^{\alpha_i}(\theta_0, \alpha_i)} \mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\alpha_i}(\theta_0, \alpha_i)\right] \left\{\mathbb{E}_{\theta_0,\alpha_{i0}}\left[v_i^2(\theta_0, \overline{\alpha}_i(\theta))\right]\right\}^{-\frac{1}{2}}. \tag{29}$$

As the robust priors considered in section 3, $\pi_i^{R,m}$ depends on the distribution of the data.[19] However, $\pi_i^{R,m}$ also depends on m_i, and, although it is not unique, there does not seem to be a way of finding priors that are bias reducing for any marginal effect considered. So, in practice one would need to estimate the model with different priors on α_i for the various marginal effects that one would consider.

In keeping with the discussion in section 4, we now look for a flexible specification for π_i that is bias reducing, independently of the marginal effect considered. For this purpose we use the set-up of subsection 4.3, and denote the population distribution of individual effects as π_0, the parametric random effects family as $\pi_i(.;\xi)$, and the best fitting approximation as $\widetilde{\pi}_0$. Then we have the following corollary to Theorem 5.

Corollary 4 *Under suitable regularity conditions given in the supplementary appendix:*

$$\operatorname*{plim}_{N\to\infty} \left(\widehat{M}_{BFE} - M\right) = O\left(\frac{\mathcal{K}(\pi_0, \widetilde{\pi}_0)}{T}\right) + o\left(\frac{1}{T}\right).$$

Corollary 4 shows that the first-order bias of the BFE estimator of M depends on the distance between the true π_0 and its parametric approximation $\widetilde{\pi}_0$, as measured by the L^2 Kullback-Leibler loss. As in the case of Corollary 3, in order to eliminate first-order bias, one could choose $\pi_i(.;\xi)$ to be the *pdf* of a finite normal mixture with a sufficiently large number of components.

Finally, let us discuss inference when N and T tend to infinity at the same rate. Provided that one uses either a robust prior for M or a flexible random effects specification, the asymptotic distribution of $\sqrt{NT}\left(\widehat{M}_{BFE} - M\right)$ is normal with zero mean and variance given by the large-T inverse information matrix.[20] In addition, asymptotically valid confidence intervals can be read from the posterior distribution of the marginal effects, as in the case of common parameters.

[19] As such, $\pi_i^{R,m}$ and π_i^m are infeasible. Feasible counterparts could be constructed as explained in section 3.

[20] Remark that, if instead we are interested in inference about the plim of M, then (unless m_i is independent of α_i) the confidence intervals will be of order $1/\sqrt{N}$ as opposed to $1/\sqrt{NT}$. This is because, when N and T grow at the same rate, the sampling error due to the averaging over cross-sectional units dominates.

5.3 Random effects estimation

Let us now turn to random effects estimation of marginal effects. The following theorem shows that $\widehat{M}_{RE}$ is generally inconsistent when N and T tend to infinity.

Theorem 6 *When T tends to infinity:*

$$\underset{N\rightarrow\infty}{\text{plim}} \left(\widehat{M}_{RE} - M\right) = \underset{N\rightarrow\infty}{\text{plim}} \frac{1}{N}\sum_{i=1}^{N}\int m_i(\theta_0, \alpha_i)\left(\widetilde{\pi}_0\left(\alpha_i\right) - \pi_0\left(\alpha_i\right)\right) d\alpha_i + O\left(\frac{1}{T}\right).$$

In a random effects framework one can use either $\widehat{M}_{RE}$ or $\widehat{M}_{BFE}$ to estimate M. Theorem 5 showed that the BFE estimator of M is large-T consistent, independently of the priors postulated on the individual effects. In sharp contrast with this result, Theorem 6 shows that standard random effects estimators of M are inconsistent in general. This happens because, in the estimation of M, $\widehat{M}_{BFE}$ updates the prior knowledge on the distribution of the fixed effects using the data while $\widehat{M}_{RE}$ does not.[21]

To summarize the results in this section, the comparison of Bayesian fixed effects and random effects estimators of marginal effects shows the benefits of updating by relying on the posterior distribution, as this reduces bias by an order of magnitude, from $O(1)$ to $O(1/T)$. Moreover, the magnitude of the bias of the Bayesian fixed effects estimator depends on how well the parametric distribution of priors approximates the population distribution of individual effects.

6 Examples

In this section and the next we consider two specific examples: a dynamic AR(p) model, and a static logit model. Derivations and an additional example concerning a Poisson counts model are available in section S2 of the supplementary appendix.

6.1 Dynamic AR(p)

The model we consider is given by:

$$y_{it} = \mu_{10}y_{i,t-1} + ... + \mu_{p0}y_{i,t-p} + \alpha_{i0} + \varepsilon_{it}, \quad i = 1...N, \quad t = 1...T.$$

[21] Under suitable tail assumptions it can be shown that the bias of $\widehat{M}_{RE}$ is $O\left(\mathcal{K}\left(\pi_0, \widetilde{\pi}_0\right)\right)$. However, using a flexible parametric family to reduce the bias would increase the asymptotic variance of the estimator, because $\widetilde{\pi}_0$ appears in the first term of the expansion of $\widehat{M}_{RE}$.

Let $y_i^0 = (y_{i,1-p}, ..., y_{i0})'$ be the vector of initial conditions, that we assume observed. Observations are iid across i. Moreover, it is assumed that:

$$(\varepsilon_{i1}, ..., \varepsilon_{iT})' | \alpha_{i0}, y_i^0 \sim \mathcal{N}\left(0, \sigma_0^2 I_T\right),$$

where I_T is the identity matrix of order T.

For this model there exist likelihood-based fixed-T consistent estimators (see for example Alvarez and Arellano, 2004), which can provide a useful benchmark for the application of our general methods. Another interesting aspect of this illustration is that, as we argue below, an orthogonal reparameterization is available for the first-order process but not for models with $p > 1$.

The individual log likelihood is given by:

$$\ell_i(\mu, \sigma^2, \alpha_i) = \frac{1}{T} \ln f(y_i | y_i^0, \alpha_i; \mu, \sigma^2) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2) - \frac{1}{2T}\sum_{t=1}^{T} \frac{(y_{it} - x_{it}'\mu - \alpha_i)^2}{\sigma^2}$$

where $x_{it} = (y_{i,t-1}, ..., y_{i,t-p})'$ and $\mu = \left(\mu_1, ..., \mu_p\right)'$.

We show in the supplementary appendix that a robust prior can be chosen as a large-T consistent estimate of the following infeasible quantity:

$$\pi_i^{IR}\left(\alpha_i | \mu, \sigma^2\right) \propto \Big(1 + a(\mu - \mu_0) + b_i(\mu - \mu_0, \alpha_i - \alpha_{i0})\Big)^{-1/2},$$

where $a(.)$ and $b_i(.,.)$ are linear and quadratic functions, respectively, the coefficients of which depend on true parameter values and initial conditions. More precisely, $a \equiv a(\mu_0)$ is a function of μ_0 only, while $b_i \equiv b(\mu_0, \alpha_{i0}, y_{i0})$ depends on true values and initial conditions.

The quadratic term $b_i(\mu - \mu_0, \alpha_i - \alpha_{i0})$ has no effect on the bias. Indeed, it could be replaced by any other quadratic function in differences $\mu - \mu_0$ and $\alpha_i - \alpha_{i0}$. Removing the quadratic terms we may consider:

$$\widetilde{\pi}^{IR}\left(\alpha_i | \mu, \sigma^2\right) \propto \{1 + a(\mu - \mu_0)\}^{-1/2}. \tag{30}$$

The prior $\widetilde{\pi}^{IR}$ is also bias reducing. Note that, as $a(\mu - \mu_0)$ is linear, the function $\widetilde{\pi}^{IR}\left(\alpha_i | \mu, \sigma^2\right)$ is degenerate for some values of μ. When estimating the prior in practice, this degeneracy can be a problem. It can then make sense to use the alternative expression (14) for the robust prior and consider instead:

$$\widetilde{\pi}^{IR}\left(\alpha_i | \mu, \sigma^2\right) \propto \exp\left(-\frac{1}{2} a(\mu - \mu_0)\right). \tag{31}$$

Now, the priors given by (30) and (31), are distribution dependent because a depends on μ_0. Looking for a non-distribution dependent prior requires solving:

$$\left.\frac{\partial}{\partial \mu}\right|_{\mu_0,\sigma_0^2} \ln \pi\left(\overline{\alpha}_i\left(\mu,\sigma^2\right)|\mu,\sigma^2\right) \propto \left.\frac{\partial}{\partial \mu}\right|_{\mu_0} \ln\left(\left\{1+a(\mu-\mu_0)\right\}^{-1/2}\right), \tag{32}$$

for some function π independent of $(\mu_0, \sigma_0^2, \alpha_{i0})$.

In the AR(1) case, we show in the supplementary appendix that

$$\left.\frac{\partial}{\partial \mu}\right|_{\mu_0} \ln\left(\left\{1+a(\mu-\mu_0)\right\}^{-1/2}\right) = \frac{1}{T}\sum_{t=1}^{T-1}(T-t)\mu_{10}^{t-1}.$$

In this case, equation (32) admits solutions independent of true parameter values. For example, the following choice works:

$$\pi\left(\alpha_i|\mu,\sigma^2\right) = \exp\left(\frac{1}{T}\sum_{t=1}^{T-1}\frac{T-t}{t}\mu^t\right). \tag{33}$$

This is the prior found by Lancaster (2002) in terms of the original (non information orthogonal) parameterization. Note that this property is specific to the AR(1) case. In the AR(p) model, $p > 1$, there generally does not exist a non-data dependent bias reducing prior. At the end of this section we discuss the existence of bias reducing data dependent priors for the AR(p) model that are independent of the common parameters, in the context of random effects estimation.

6.2 Static logit

We now consider the model:

$$y_{it} = \mathbf{1}\left\{x_{it}'\theta_0 + \alpha_{i0} + \varepsilon_{it} > 0\right\}, \quad i = 1...N, \quad t = 1...T$$

where the x's are known, and ε_{it} are i.i.d. and drawn from the logistic distribution with *cdf* Λ.

The individual log-likelihood is given by:

$$\ell_i(\theta,\alpha_i) = \frac{1}{T}\sum_{t=1}^{T}\left\{y_{it}\ln\Lambda(x_{it}'\theta+\alpha_i) + (1-y_{it})\ln\left[1-\Lambda(x_{it}'\theta+\alpha_i)\right]\right\}.$$

In the supplementary appendix we derive the expression of a robust prior as a consistent estimate of:

$$\pi_i^{IR}(\alpha_i|\theta) \propto \left(\sum_{t=1}^{T}\mathbb{E}_{\theta_0,\alpha_{i0}}\left(\left[y_{it}-\Lambda(x_{it}'\theta+\alpha_i)\right]^2\right)\right)^{-1/2}\sum_{t=1}^{T}\Lambda(x_{it}'\theta+\alpha_i)\left[1-\Lambda(x_{it}'\theta+\alpha_i)\right]. \tag{34}$$

As shown in Lancaster (2000), there also exists an orthogonal reparameterization in this model. Let:

$$\psi_i = \sum_{t=1}^{T} \Lambda(x_{it}'\theta + \alpha_i).$$

Then ψ_i and θ are information orthogonal.

The uniform prior on ψ_i is thus bias reducing. The corresponding prior on the original individual effects is:

$$\pi_i(\alpha_i|\theta) \propto \sum_{t=1}^{T} \Lambda(x_{it}'\theta + \alpha_i)\left[1 - \Lambda(x_{it}'\theta + \alpha_i)\right]. \tag{35}$$

Note that in this case, Jeffreys' prior is given by $\pi_i^J(\alpha_i|\theta) \propto \{\pi_i(\alpha_i|\theta)\}^{1/2}$. It is readily verified that π_i^J is not bias reducing. On the other hand, both π_i^{IR} and π_i reduce bias.

In practice, one can thus compute the following robust prior:

$$\pi_i^R(\alpha_i|\theta) \propto \left\{\sum_{t=1}^{T}\left((y_{it} - \Lambda(x_{it}'\theta + \alpha_i))^2\right)\right\}^{-1/2} \sum_{t=1}^{T} \Lambda(x_{it}'\theta + \alpha_i)\left[1 - \Lambda(x_{it}'\theta + \alpha_i)\right]. \tag{36}$$

One can also use expected quantities and compute:

$$\begin{aligned}\pi_i^R(\alpha_i|\theta) \quad \propto \quad & \left\{\sum_{t=1}^{T} \Lambda(x_{it}'\widehat{\theta} + \widehat{\alpha}_i)\left[1 - 2\Lambda(x_{it}'\theta + \alpha_i)\right] + \left[\Lambda(x_{it}'\theta + \alpha_i)\right]^2\right\}^{-1/2} \\ & \times \sum_{t=1}^{T} \Lambda(x_{it}'\theta + \alpha_i)\left[1 - \Lambda(x_{it}'\theta + \alpha_i)\right], \end{aligned} \tag{37}$$

where $\widehat{\theta}$ and $\widehat{\alpha}_i$ are consistent estimates of the true parameters when T tends to infinity (for example maximum likelihood estimates).

6.3 Random effects

We study the properties of random effects maximum likelihood (REML) estimators in the previous examples.

Dynamic AR(p). We start with the dynamic AR(p) model of subsection 6.1. We show in the supplementary appendix that, for this model:

$$\rho_i(\mu, \sigma^2, \alpha_i) = a_0(\mu) y_i^0 + a_1(\mu)\alpha_i,$$

where y_i^0 is the vector of initial conditions, and $a_0(\mu)$ and $a_1(\mu)$ are matrices. Moreover, if the process is stationary then $a_0(\mu) = O(1/T)$. Hence, it follows from Corollary 1 that uncorrelated Gaussian REML is bias reducing for this model. This result was proven by Cho, Hahn and Kuersteiner (2004) in the case $p = 1$. If strictly exogenous covariates are included in the model then it is easy to check that correlated Gaussian REML is robust, while uncorrelated REML is not in general.

Linear model with one endogenous regressor and many instruments. A closely related example is the following linear model with one endogenous regressor in a panel context:[22]

$$\begin{aligned} y_{it} &= \theta\alpha_i + u_{it}, \\ x_{it} &= \alpha_i + v_{it}, \end{aligned}$$

where errors are i.i.d. and:

$$\begin{pmatrix} u_{it} \\ v_{it} \end{pmatrix} \sim \mathcal{N}(0, \Omega).$$

In the following we assume that covariance matrix Ω is given. We let

$$\Omega^{-1} = \begin{pmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{pmatrix}.$$

In this example there is an analogy between having a large number of individual effects and a large number of instruments in a simultaneous equations perspective (see Hahn, 2000).

We show in the supplementary appendix that:

$$\rho_i(\theta, \alpha_i) = \alpha_i \frac{-\omega_{11}\theta - \omega_{12}}{\omega_{11}\theta^2 + 2\omega_{12}\theta + \omega_{22}}.$$

We are thus in the case of Corollary 1, and Gaussian REML is bias reducing. A related situation arises in Chamberlain and Imbens' (2004) use of REQML under Bekker's (1994) asymptotics. Our treatment of this example shows that the linearity of the model is crucial for the success of random effects methods.

Static logit. In the case of the static logit model, we have that:

$$\rho_i(\theta, \alpha_i) = -\frac{\sum_{t=1}^T \Lambda(x_{it}'\theta + \alpha_i)(1 - \Lambda(x_{it}'\theta + \alpha_i))x_{it}}{\sum_{t=1}^T \Lambda(x_{it}'\theta + \alpha_i)(1 - \Lambda(x_{it}'\theta + \alpha_i))}.$$

This is a highly nonlinear expression in α_i, θ and $x_i = (x_{i1}...x_{iT})'$. Thus, usual REML estimators are not bias reducing. For example, Corollary 1 shows that uncorrelated Gaussian REML is not robust.

[22] We are grateful to Jinyong Hahn for this suggestion.

Note that this lack of unbiasedness is not corrected for by allowing the prior to depend on covariates x_{it}, as in Chamberlain's (1984) probit model. In that case, it is still impossible to correct for the first-order bias without permitting the prior to depend on the common parameters θ. In nonlinear models, thus, the success of random effects likelihood inference depends critically on prior knowledge about the form of the fixed effects.

7 Monte Carlo simulation

In this section, we provide some Monte Carlo evidence on the finite sample behavior of integrated likelihood estimators.

7.1 Static logit

We first focus on the static logit model:

$$y_{it} = \mathbf{1}\left\{x_{it}'\theta_0 + \alpha_{i0} + \varepsilon_{it} > 0\right\}, \quad i = 1...N, \quad t = 1...T. \tag{38}$$

The x_{it} are constant across simulations and drawn from a $\mathcal{N}(0,1)$ distribution. The individual effects are drawn in each simulation from $\mathcal{N}(\overline{x}_i, 1)$, where $\overline{x}_i = \frac{1}{T}\sum_{t=1}^{T} x_{it}$. Lastly, ε_{it} are i.i.d. draws from the logistic *cdf*, and θ_0 is set to one. In all the experiments N is 100.

Table I shows some statistics of the empirical distribution of 100 draws of $\widehat{\theta}$, where $\widehat{\theta}$ can be one of the following estimators: "uncorrected" refers to the MLE, and "corrected" to the corrected MLE, obtained using the DiCiccio and Stern (1993) adjustment based on equation (5), see Arellano and Hahn (2007, p.392); "uniform" is the integrated likelihood estimator with uniform prior $\pi_i \propto 1$; "Lancaster" is the integrated likelihood with the uniform prior on the orthogonal parameters written in terms of the original effects, see equation (35); "robust, observed" refers to the integrated likelihood with the robust prior constructed from observed quantities, see (36), while "robust, infeasible" refers to the integrated likelihood with the robust prior estimated using expected quantities where the true parameter θ_0 is assumed known, see (37); "robust, iterated 1" refers to the same estimator, but when the expectation in (37) is evaluated at $\widehat{\theta}$, the "robust" integrated likelihood estimator; then, "robust, iterated ∞" is obtained iterating this procedure until convergence; "random effects" is the Gaussian random effects estimator; lastly, "conditional logit" is Chamberlain's (1980) conditional logit.[23]

[23] Both the random effects and conditional logit estimators were computed using the STATA *xtlogit* and *clogit* commands, respectively. The other estimators were computed using GAUSS.

Table I shows that the bias of the MLE can be large: it is equal to 33% for $T = 5$ and still 6% for $T = 20$. The corrections based on the concentrated likelihood and the various integrated likelihoods give roughly the same results. In all cases considered, using one of these corrections reduces the bias by a factor between 2 and 3. The best performance, in terms of bias, mean squared error (MSE) and mean absolute error (MAE), is achieved by Lancaster's (1998) integrated likelihood given by equation (35). Note that the infeasible estimator based on (37) and the iterated corrections do not give better results than the correction based on observed quantities.

The Gaussian random effects MLE gives rather good results. Our experiments (not reported) showed that the relative performance of REML worsens when the correlation between α_{i0} and x_i increases, and when the sampling distribution of the individual effects departs from the normal. Lastly, the conditional logit estimator is consistent for fixed T. Still, note that several corrected/integrated estimators yield MSE and MAE comparable to– or lower than– the ones of conditional logit for $T = 10$ and $T = 20$. This suggests that, for intermediate values of T, it may not be obvious to choose a fixed-T consistent estimator rather than bias-corrected alternatives. Hahn, Kuersteiner and Newey (2004) show that bias-corrected estimators are second-order efficient. Clearly, under suitable regularity conditions our robust integrated likelihood estimator falls into the class considered by these authors.[24] In contrast, there is a potential efficiency loss in conditioning on the sufficient statistic in the conditional logit model.

Finally, in Figure 1 we draw the likelihood function of the static logit model (thin line). The thick line and the dashed line show the bias-corrected likelihood function (using the DiCiccio and Stern formula) and the robust integrated likelihood. The two pseudo-likelihoods are concave. Moreover, it is clear on the figure that they both correct bias with respect to the MLE.

7.2 Dynamic AR(1)

Next, we consider the dynamic AR(1) model:

$$y_{it} = \mu_{10} y_{it-1} + \alpha_{i0} + \varepsilon_{it}, \quad i = 1...N, \quad t = 1...T. \tag{39}$$

[24] A second-order Laplace approximation of the integrated likelihood (as in Tierney *et al.*, 1989) is necessary to prove this result formally.

Individual effects are drawn in each simulation from a standard distribution. Moreover, the initial condition y_{i0} is drawn in the stationary distribution of y_{it} for fixed i. Lastly, ε_{it} are i.i.d. standard normal draws, and μ_{10} is set to .5. As before, N is 100. The standard deviation of errors, set to one, is treated as known.

With non i.i.d. data, the choice of local approximation of the formulas for prior distributions may be important, as illustrated in Figure 2. The left panel in Figure 2 shows the likelihood function of the dynamic AR(1) model (thin line). The thick line shows the integrated likelihood with prior given by the formula (30), obtained using expected quantities. The function is degenerate around $\mu_1 = .8$. Moreover, a close look at the Figure shows two local extrema. The local maximum corresponds to μ_1 around .5, which means that inference from this local maximum is bias reducing. Still, the flatness of the curve suggests that one might have trouble trying to find this maximum using standard maximization algorithms. This problem is likely to be worse in situations with more parameters to consider. The right panel on the same figure shows the integrated likelihood for the prior (31). The situation there is strikingly different, as the pseudo-likelihood is nicely concave. Moreover, its maximum is still much closer to the truth than the MLE. In the rest of this section, we use the prior (31) to estimate common parameters.

Table II shows some statistics of the empirical distributions of some estimators for $T = 10$: the MLE ("observed"), and diverse corrections based on various degrees of trimming (from $q = 1$ to $q = 3$); Then, the integrated likelihood based on the uniform prior ("uniform") and on the Lancaster prior ("Lancaster") given by (33); the "robust" expression of the prior is based on (14) where the outer product is estimated using observed quantities with various degrees of trimming; the "expected" prior is the one given by (31), and plugged-in the "robust, $q = 2$" result to start the iterations in "iterated"; "GMM" refers to the estimator discussed in Arellano and Bond (1991); lastly, "random effects (uncorr.)" and "random effects (corr.)" refer to the Gaussian random effects estimators assuming that the individual effects are independent of initial conditions, or allowing that the mean depends linearly on the initial condition.[25]

We find a large bias of the MLE (30%) that is corrected for by almost one half by both the corrections of the concentrated likelihood and the robust integrated likelihood. In both cases the preferred degree of trimming is 2. The uniform prior yields no bias reduction at

[25] We computed the GMM estimator using the STATA command *xtabond2*, with the option *noleveleq*. The other estimators were programmed in GAUSS.

all, and the Lancaster prior based on the available orthogonalization gives almost no bias. Interestingly, the infeasible robust prior based on expected quantities and the true value of μ_{10} gives even better results, in terms of bias, MSE and MAE. Moreover, the iterated estimators have also very good finite sample properties. In our simulations, we found that two iterations were enough to get very close to the infinitely iterated estimator. As the formulas of these priors are not based on parameter orthogonalization, these results suggest that iteration of the analytical expressions of the prior such as (14) can be useful in order to deal with non i.i.d. data. Lastly, remark that the GMM estimator suffers from a small bias, which disappears when N grows (recall that $N = 100$ in the experiments). Moreover, it has larger variance than all the other estimators. The result is that the integrated likelihood functions with priors based on analytical calculations (infeasible and iterated) compare favorably with the fixed-T consistent GMM estimator in terms of MSE and MAE.

The last two rows of Table II show the behavior of random effects estimators. In the dynamic AR(1) model, Alvarez and Arellano (2003) showed that the Gaussian RE pseudo-likelihood based on $\alpha_i \sim \mathcal{N}(m_1 + m_2 y_{i0}, s^2)$ reduces bias. Then, Cho *et al.* (2004) showed that this is also the case of the RE specification $\alpha_i \sim \mathcal{N}(m, s^2)$, where the mean of α_i is misspecified to be independent of the initial observation y_{i0}. We have shown that this result generalizes to dynamic AR(p) models without exogenous covariates. The numbers reported show that, in spite of the theoretical result, the uncorrelated REML estimator is substantially biased compared to its correlated counterpart. Thus, in dynamic linear models, it may be important to allow (even parametrically) for correlation between the individual effects and the initial conditions in the estimation. Lastly, note that the correlated random effects estimator compares favorably to all other estimators studied, except the infeasible and infinitely iterated robust integrated likelihood estimators.

7.3 Dynamic AR(2)

We end this simulation section by considering the dynamic AR(2) model

$$y_{it} = \mu_{10} y_{it-1} + \mu_{20} y_{it-2} + \alpha_{i0} + \varepsilon_{it}, \quad i = 1...N, \quad t = 1...T. \tag{40}$$

As before, the individual effects are drawn in each simulation from a standard distribution and the initial conditions $y_{i,-1}$ and y_{i0} are drawn in the stationary distribution of (y_{it}, y_{it+1}) for fixed i. Then, ε_{it} are i.i.d. standard normal draws, μ_{10} is set to .5 and μ_{20} to 0. Lastly, N is 100, and the standard deviation of errors, set to one, is treated as known.

To estimate the priors, we use the robust formula given in (14). Analytical expressions are given in the supplementary appendix. Table III presents the results for $T = 10$. We find that the MLE is biased. A difference with the AR(1) case is that if the corrected concentrated likelihood and the robust integrated likelihood estimated using observed quantities reduce bias, they do so only for the first autoregressive parameter. In that case, only the analytical correction ("infeasible") reduces both biases. Interestingly, as before only one or two iterations starting with the "robust" estimate get close to these infeasible estimates. Moreover, as in the AR(1) case, the iterated analytical corrections compare favorably with the GMM estimator. Note that in the AR(2) case no orthogonal reparameterization is available. The results obtained for the iterated estimators thus seem remarkable, both in terms of bias and mean squared error.

8 Conclusion

Many approaches to the estimation of panel data models rely on an average likelihood that assigns weights to different values of the individual effects. In this paper, we study under which conditions such weighting schemes are robust, in that they yield biases of order $1/T^2$ as opposed to $1/T$.

We find that robust weights, or priors, will in general satisfy two conditions. First, they depend on the data, unless an orthogonal reparameterization is available. Second, they do not impose prior independence between the common parameters and the individual effects, as we show that random effects specifications are not bias reducing in general.

We propose two bias reducing priors, which deal with the incidental parameter problem by taking into account the uncertainty about the individual effects. Our approach, based on prior distributions and integration, has a natural connection with simulation-based estimation techniques, such as MCMC. In addition, we argue that asymptotically valid confidence intervals can be read from the quantiles of the posterior distribution.

We show that in general standard random effects estimation of policy parameters is inconsistent for large T, whereas the posterior mean is large-T consistent, and we provide conditions for bias reduction. Priors that are bias reducing for the common parameters do not lead to bias reduction of marginal effects, and bias reducing priors for marginal effects are specific to the effect considered.

We also show that in random effects models, both the estimators of common parameters and the posterior means of marginal effects have first-order biases that depend on the Kullback-Leibler distance between the population distribution of the effects and its best approximation in the random effects family. So, while updating the prior given the data lowers the bias on the marginal effects by an order of magnitude, the bias can be further reduced by using either a bias-reducing prior or a sufficiently close approximating family to the distribution of the effects.

The Monte Carlo evidence suggests rather good finite sample properties of integrated likelihood estimates based on robust priors. It seems very interesting to investigate the behavior of our method as the complexity of the model increases. If what we propose turns out to be feasible and satisfying, then structural microeconometric models would be a natural field of application.

CEMFI; Casado del Alisal, 5, 28014 Madrid, Spain; arellano@cemfi.es

and

CEMFI; Casado del Alisal, 5, 28014 Madrid, Spain; bonhomme@cemfi.es.

APPENDIX: Proofs

This appendix provides proofs of the results in sections 2-3 and subsections 4.1-4.2. Proofs of the results from subsection 4.3 are in the supplementary appendix.

Proof of Lemma 1: Let us fix i, and denote

$$L_i^I(\theta)=\int \exp\left[T\ell_i(\theta,\alpha_i)\right]\pi_i(\alpha_i|\theta)d\alpha_i.$$

Assuming that $\ell_i(\theta,\alpha_i)$ has a unique maximum $\widehat{\alpha}_i(\theta)$ and using a Laplace approximation as in Tierney *et al.* (1989) we obtain:

$$\begin{aligned}
L_i^I(\theta) &= \pi_i(\widehat{\alpha}_i(\theta)|\theta)\int \exp\left(T\ell_i(\theta,\widehat{\alpha}_i(\theta))+\frac{T}{2}v_i^{\alpha_i}(\theta,\widehat{\alpha}_i(\theta))\left(\alpha_i-\widehat{\alpha}_i(\theta)\right)^2\right)d\alpha_i\left(1+O_p\left(\frac{1}{T}\right)\right)\\
&= \pi_i(\widehat{\alpha}_i(\theta)|\theta)\exp\left[T\ell_i(\theta,\widehat{\alpha}_i(\theta))\right]\int \exp\left(\frac{T}{2}v_i^{\alpha_i}(\theta,\widehat{\alpha}_i(\theta))\left(\alpha_i-\widehat{\alpha}_i(\theta)\right)^2\right)d\alpha_i\left(1+O_p\left(\frac{1}{T}\right)\right),\\
&= \pi_i(\widehat{\alpha}_i(\theta)|\theta)\sqrt{2\pi}\left\{-Tv_i^{\alpha_i}(\theta,\widehat{\alpha}_i(\theta))\right\}^{-1/2}\exp\left[T\ell_i(\theta,\widehat{\alpha}_i(\theta))\right]\left(1+O_p\left(\frac{1}{T}\right)\right).
\end{aligned}$$

It thus follows that:

$$\ell_i^I(\theta)-\ell_i^c(\theta)=\frac{1}{2T}\ln\left(\frac{2\pi}{T}\right)-\frac{1}{2T}\ln\left(-v_i^{\alpha_i}(\theta,\widehat{\alpha}_i(\theta))\right)+\frac{1}{T}\ln\pi_i(\widehat{\alpha}_i(\theta)|\theta)+O_p\left(\frac{1}{T^2}\right), \qquad \text{(A1)}$$

where Assumption 1 allows us to take logs.

Now by expanding the sample moment condition $v_i(\theta, \widehat{\alpha}_i(\theta)) = 0$ around $\overline{\alpha}_i(\theta)$ we immediately find that

$$\widehat{\alpha}_i(\theta) - \overline{\alpha}_i(\theta) = \frac{A}{\sqrt{T}} + O_p\left(\frac{1}{T}\right),$$

where $A = O_p(1)$ and $\mathbb{E}_{\theta_0,\alpha_{i0}}[A] = 0$. This implies that:

$$v_i^{\alpha_i}(\theta, \widehat{\alpha}_i(\theta)) = v_i^{\alpha_i}(\theta, \overline{\alpha}_i(\theta)) + \frac{B}{\sqrt{T}} + O_p\left(\frac{1}{T}\right) = \mathbb{E}_{\theta_0,\alpha_{i0}}\left[v_i^{\alpha_i}(\theta, \overline{\alpha}_i(\theta))\right] + \frac{C}{\sqrt{T}} + O_p\left(\frac{1}{T}\right),$$

where B and C are $O_p(1)$ with zero mean. Expanding the log yields:

$$\mathbb{E}_{\theta_0,\alpha_{i0}} \ln\left(-v_i^{\alpha_i}(\theta, \widehat{\alpha}_i(\theta))\right) = \ln \mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\alpha_i}(\theta, \overline{\alpha}_i(\theta))\right] + O\left(\frac{1}{T}\right). \tag{A2}$$

Likewise, using Assumption 2 we obtain:

$$\mathbb{E}_{\theta_0,\alpha_{i0}} \ln \pi_i(\widehat{\alpha}_i(\theta)|\theta) = \ln \pi_i(\overline{\alpha}_i(\theta)|\theta) + O\left(\frac{1}{T}\right). \tag{A3}$$

Taking expectations in (A1) and combining the result with (A2) and (A3) yields:

$$\mathbb{E}_{\theta_0,\alpha_{i0}}\left[\ell_i^I(\theta) - \ell_i^c(\theta)\right] = \frac{1}{2T}\ln\left(\frac{2\pi}{T}\right) - \frac{1}{2T}\ln \mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\alpha_i}(\theta, \overline{\alpha}_i(\theta))\right] + \frac{1}{T}\ln \pi_i(\overline{\alpha}_i(\theta)|\theta) + O\left(\frac{1}{T^2}\right).$$

Q.E.D.

Proof of Theorem 1: Immediate from (4) and (5). *Q.E.D.*

Proof of Theorem 2: Immediate using (8). *Q.E.D.*

In preparation for the proof of Proposition 1 we state the following lemma:

Lemma A1

$$\left.\frac{\partial}{\partial\theta}\right|_{\theta_0} \overline{\alpha}_i(\theta) = \left\{\mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\alpha_i}(\theta_0, \alpha_{i0})\right]\right\}^{-1} \mathbb{E}_{\theta_0,\alpha_{i0}}\left[v_i^{\theta}(\theta_0, \alpha_{i0})\right] \equiv \rho_i(\theta_0, \alpha_{i0}). \tag{A4}$$

Proof of Lemma A1: By differentiating the moment condition solved by $\overline{\alpha}_i(\theta)$ with respect to θ:

$$\mathbb{E}_{\theta_0,\alpha_{i0}}\left[v_i(\theta, \overline{\alpha}_i(\theta))\right] = 0.$$

Q.E.D.

Proof of Proposition 1: The bias of the integrated score is:

$$b_i(\theta_0) = \left.\frac{\partial}{\partial\theta}\right|_{\theta_0} \ln \pi_i(\overline{\alpha}_i(\theta)|\theta) - \underbrace{\left.\frac{\partial}{\partial\theta}\right|_{\theta_0}\left(\ln\left(\mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\alpha_i}(\theta, \overline{\alpha}_i(\theta))\right]\left\{\mathbb{E}_{\theta_0,\alpha_{i0}}\left[v_i^2(\theta, \overline{\alpha}_i(\theta))\right]\right\}^{-1/2}\right)\right)}_{A}.$$

In addition to Lemma A1, we need the information matrix equality at true values:

$$\mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\alpha_i}(\theta_0, \alpha_{i0})\right] = T\mathbb{E}_{\theta_0,\alpha_{i0}}\left[v_i^2(\theta_0, \alpha_{i0})\right]. \tag{A5}$$

In order to simplify the notation, we drop the arguments inside the expectation terms when they are evaluated at true values. We obtain:

$$\begin{aligned}
A &= \frac{\mathbb{E}(v_i^{\alpha_i\theta}) + \rho_i \mathbb{E}(v_i^{\alpha_i\alpha_i})}{\mathbb{E}(v_i^{\alpha_i})} - \frac{1}{2}\cdot\frac{2\mathbb{E}(v_i^{\theta}v_i) + 2\rho_i\mathbb{E}(v_i^{\alpha_i}v_i)}{\mathbb{E}(v_i^2)} \\
&= \frac{-1}{\mathbb{E}(-v_i^{\alpha_i})}\left\{\mathbb{E}(v_i^{\alpha_i\theta}) + T\mathbb{E}(v_i^{\theta}v_i) + \rho_i\left[\mathbb{E}(v_i^{\alpha_i\alpha_i}) + T\mathbb{E}(v_i^{\alpha_i}v_i)\right]\right\} \\
&= \frac{-1}{\mathbb{E}(-v_i^{\alpha_i})^2}\left\{\mathbb{E}(-v_i^{\alpha_i})\left(\mathbb{E}(v_i^{\alpha_i\theta}) + T\mathbb{E}(v_i^{\theta}v_i)\right) + \mathbb{E}(v_i^{\theta})\left(\mathbb{E}(v_i^{\alpha_i\alpha_i}) + T\mathbb{E}(v_i^{\alpha}v_i)\right)\right\} \\
&= \frac{-1}{\mathbb{E}(-v_i^{\alpha_i})^2}\left\{\mathbb{E}(-v_i^{\alpha_i})\frac{\partial}{\partial\alpha_i}\bigg|_{\theta_0,\alpha_{i0}}\mathbb{E}_{\theta,\alpha_i}(v_i^{\theta}(\theta,\alpha_i)) - \mathbb{E}(v_i^{\theta})\frac{\partial}{\partial\alpha_i}\bigg|_{\theta_0,\alpha_{i0}}\mathbb{E}_{\theta,\alpha_i}(-v_i^{\alpha_i}(\theta,\alpha_i))\right\},
\end{aligned}$$

where

$$\mathbb{E}_{\theta,\alpha_i}(v_i^{\theta}(\theta,\alpha_i)) = \int v_i^{\theta}(\theta,\alpha_i)f_i(y;\theta,\alpha_i)dy; \text{ and: } \mathbb{E}_{\theta,\alpha_i}(v_i^{\alpha_i}(\theta,\alpha_i)) = \int v_i^{\alpha_i}((\theta,\alpha_i)f_i(y;\theta,\alpha_i)dy.$$

It follows that

$$A = -\frac{\partial}{\partial\alpha_i}\bigg|_{\theta_0,\alpha_{i0}}\left(\left\{\mathbb{E}_{\theta,\alpha_i}\left[-v_i^{\alpha_i}(\theta,\alpha_i)\right]\right\}^{-1}\mathbb{E}_{\theta,\alpha_i}\left[v_i^{\theta}(\theta,\alpha_i)\right]\right),$$

and the proposition is proved. *Q.E.D.*

PROOF OF PROPOSITION 2: We have:

$$b_i(\theta_0) = \frac{\partial}{\partial\theta}\bigg|_{\theta_0}\ln\pi_i(\overline{\alpha}_i(\theta)|\theta) - \frac{\partial}{\partial\theta}\bigg|_{\theta_0}\ln\left(\mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\alpha_i}(\theta,\overline{\alpha}_i(\theta))\right]\left\{\mathbb{E}_{\theta_0,\alpha_{i0}}\left[Tv_i^2(\theta,\overline{\alpha}_i(\theta))\right]\right\}^{-1/2}\right).$$

Note that it follows from the invariance property of ML that

$$\overline{\psi}_i(\theta) = \psi_i(\overline{\alpha}_i(\theta),\theta).$$

Moreover it is easily verified that:

$$\mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\alpha_i}(\theta,\alpha_i)\right] = \left(\frac{\partial\psi_i(\alpha_i,\theta)}{\partial\alpha_i}\right)^2\mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\psi_i}(\theta,\psi_i(\alpha_i,\theta))\right] - \frac{\partial^2\psi_i(\alpha_i,\theta)}{\partial\alpha_i^2}\mathbb{E}_{\theta_0,\alpha_{i0}}\left[v_i(\theta,\psi_i(\alpha_i,\theta))\right],$$

and:

$$\mathbb{E}_{\theta_0,\alpha_{i0}}\left[v_i^2(\theta,\alpha_i)\right] = \left(\frac{\partial\psi_i(\alpha_i,\theta)}{\partial\alpha_i}\right)^2\mathbb{E}_{\theta_0,\alpha_{i0}}\left[v_i^2(\theta,\psi_i(\alpha_i,\theta))\right],$$

where with some abuse of notation we have written $v_i(\theta,\psi_i)$ for the score of the reparameterized likelihood with respect to the new fixed effects. Evaluating these two equalities at $(\theta,\overline{\alpha}_i(\theta))$ and using that $E_{\theta_0,\alpha_{i0}}\left[v_i(\theta,\overline{\psi}_i(\theta))\right] = 0$ yields:

$$\mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\alpha_i}(\theta,\overline{\alpha}_i(\theta))\right] = \left(\frac{\partial\psi_i(\overline{\alpha}_i(\theta),\theta)}{\partial\alpha_i}\right)^2\mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\psi_i}(\theta,\overline{\psi}_i(\theta))\right],$$

and:

$$\mathbb{E}_{\theta_0,\alpha_{i0}}\left[v_i^2(\theta,\overline{\alpha}_i(\theta))\right] = \left(\frac{\partial\psi_i(\overline{\alpha}_i(\theta),\theta)}{\partial\alpha_i}\right)^2\mathbb{E}_{\theta_0,\alpha_{i0}}\left[v_i^2(\theta,\overline{\psi}_i(\theta))\right].$$

Hence:

$$\begin{aligned}
b_i(\theta_0) &= \frac{\partial}{\partial\theta}\bigg|_{\theta_0}\ln\pi_i(\overline{\alpha}_i(\theta)|\theta) - \frac{\partial}{\partial\theta}\bigg|_{\theta_0}\ln\left(\mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\psi_i}(\theta,\overline{\psi}_i(\theta))\right]\left\{\mathbb{E}_{\theta_0,\alpha_{i0}}\left[Tv_i^2(\theta,\overline{\psi}_i(\theta))\right]\right\}^{-1/2}\right) \\
&\qquad - \frac{\partial}{\partial\theta}\bigg|_{\theta_0}\ln\left|\frac{\partial\psi_i(\overline{\alpha}_i(\theta),\theta)}{\partial\alpha_i}\right| \\
&= \frac{\partial}{\partial\theta}\bigg|_{\theta_0}\ln\widetilde{\pi}_i(\overline{\psi}_i(\theta)|\theta) - \frac{\partial}{\partial\theta}\bigg|_{\theta_0}\ln\left(\mathbb{E}_{\theta_0,\psi_{i0}}\left[-v_i^{\psi_i}(\theta,\overline{\psi}_i(\theta))\right]\left\{\mathbb{E}_{\theta_0,\psi_{i0}}\left[Tv_i^2(\theta,\overline{\psi}_i(\theta))\right]\right\}^{-1/2}\right).
\end{aligned}$$

The proposition follows. *Q.E.D.*

Proof of Proposition 3: A stochastic expansion of $v_i(\theta,\widehat{\alpha}_i(\theta))$ in the neighborhood of $(\theta,\overline{\alpha}_i(\theta))$ yields:

$$\widehat{\alpha}_i(\theta)-\overline{\alpha}_i(\theta)=\left\{\mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\alpha_i}(\theta,\overline{\alpha}_i(\theta))\right]\right\}^{-1}v_i(\theta,\overline{\alpha}_i(\theta))+O_p\left(\frac{1}{T}\right).$$

This yields:

$$\mathbb{E}_{\theta_0,\alpha_{i0}}\left(\widehat{\alpha}_i(\theta)-\overline{\alpha}_i(\theta)\right)=O\left(\frac{1}{T}\right),$$

and:

$$\mathbb{E}_{\theta_0,\alpha_{i0}}\left[\left(\widehat{\alpha}_i(\theta)-\overline{\alpha}_i(\theta)\right)^2\right]=\left\{\mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\alpha_i}(\theta,\overline{\alpha}_i(\theta))\right]\right\}^{-2}\mathbb{E}_{\theta_0,\alpha_{i0}}\left[v_i^2(\theta,\overline{\alpha}_i(\theta))\right]+O\left(\frac{1}{T^2}\right).$$

Hence:

$$\widehat{\text{Var}}\left(\widehat{\alpha}_i(\theta)\right)=\left[\pi_i^R\left(\overline{\alpha}_i(\theta)|\theta\right)\right]^{-2}+O_p\left(\frac{1}{T^2}\right).$$

Thus, as $\widehat{\text{Var}}\left(\widehat{\alpha}_i(\theta)\right)=O_p(1/T)$ we have:

$$\pi_i^R(\overline{\alpha}_i(\theta)|\theta)\propto\frac{1}{\sqrt{\widehat{\text{Var}}\left(\widehat{\alpha}_i(\theta)\right)}}\left(1+O_p\left(\frac{1}{T}\right)\right).$$

Equation (15) follows by remarking that

$$\pi_i^R(\widehat{\alpha}_i(\theta)|\theta)=\pi_i^R(\overline{\alpha}_i(\theta)|\theta)\left(1+O_p\left(\frac{1}{T}\right)\right),$$

by the same arguments as in the proof of Lemma 1.

To show the second part of the Proposition, let π_i be a non-dogmatic prior satisfying:

$$\pi_i(\widehat{\alpha}_i(\theta)|\theta)\propto\frac{1}{\sqrt{\widehat{\text{Var}}\left(\widehat{\alpha}_i(\theta)\right)}}\left(1+O_p\left(\frac{1}{T}\right)\right).$$

Then the proof of Lemma 1 shows that the only quantity that matters for bias reduction is $\ln\pi_i(\widehat{\alpha}_i(\theta)|\theta)$. This result comes directly from the Laplace approximation to the integrated likelihood, and does not require Assumption 2 to hold. As

$$\ln\pi_i(\widehat{\alpha}_i(\theta)|\theta)=\ln\pi_i^R(\widehat{\alpha}_i(\theta)|\theta)+O_p\left(\frac{1}{T}\right),$$

and as π_i^R is robust, it follows that π_i is also bias reducing. *Q.E.D.*

Proof of Lemma 2: The first-order conditions of the maximization imply that:

$$0=\sum_{i=1}^N\frac{\partial\ell_i^{RE}(\theta;\widehat{\xi}(\theta))}{\partial\xi}=\sum_{i=1}^N\frac{1}{T}\frac{\int\exp\left[T\ell_i(\theta,\alpha_i)\right]\left\{\partial\pi_i(\alpha_i;\widehat{\xi}(\theta))/\partial\xi\right\}d\alpha_i}{\int\exp\left[T\ell_i(\theta,\alpha_i)\right]\pi_i(\alpha_i;\widehat{\xi}(\theta))d\alpha_i}.$$

A Laplace approximation of the two integrals yields, as in the proof of Lemma 1:

$$\begin{aligned}
\int \exp\left(T\ell_i(\theta,\alpha_i)\right)\frac{\partial \pi_i(\alpha_i;\widehat{\xi}(\theta))}{\partial \xi}d\alpha_i &= \sqrt{2\pi}\left(-Tv_i^{\alpha_i}(\theta,\widehat{\alpha}_i(\theta))\right)^{-1/2}\exp\left[T\ell_i(\theta,\widehat{\alpha}_i(\theta))\right] \\
&\quad \times \frac{\partial \pi_i(\widehat{\alpha}_i(\theta);\widehat{\xi}(\theta))}{\partial \xi}\left(1+O_p\left(\frac{1}{T}\right)\right), \\
\int \exp\left(T\ell_i(\theta,\alpha_i)\right)\pi_i(\alpha_i;\widehat{\xi}(\theta))d\alpha_i &= \sqrt{2\pi}\left(-Tv_i^{\alpha_i}(\theta,\widehat{\alpha}_i(\theta))\right)^{-1/2}\exp\left[T\ell_i(\theta,\widehat{\alpha}_i(\theta))\right] \\
&\quad \times \pi_i(\widehat{\alpha}_i(\theta);\widehat{\xi}(\theta))\left(1+O_p\left(\frac{1}{T}\right)\right).
\end{aligned}$$

Hence we obtain:

$$\frac{1}{N}\sum_{i=1}^N \frac{\partial \ln \pi_i(\widehat{\alpha}_i(\theta);\widehat{\xi}(\theta))}{\partial \xi}\left(1+O_p\left(\frac{1}{T}\right)\right)=0.$$

Then taking the probability limit we have:

$$\operatorname*{plim}_{N\to\infty}\ \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\pi_0}\left(\mathbb{E}_{\theta_0,\alpha_{i0}}\frac{\partial \ln \pi_i(\widehat{\alpha}_i(\theta);\overline{\xi}(\theta))}{\partial \xi}\right)=O\left(\frac{1}{T}\right).$$

Lastly, using that $\mathbb{E}_{\theta_0,\alpha_{i0}}(\widehat{\alpha}_i(\theta)-\overline{\alpha}_i(\theta))=O(1/T)$ we obtain:

$$\operatorname*{plim}_{N\to\infty}\ \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\pi_0}\left(\frac{\partial \ln \pi_i(\overline{\alpha}_i(\theta);\overline{\xi}(\theta))}{\partial \xi}\right)=O\left(\frac{1}{T}\right).$$

Q.E.D.

Proof of Theorem 3: Let $\pi_i(\alpha_i,\xi)$ be a class of random effects distributions indexed by ξ. Also, let π_{0G} be a population joint density of individual effects and exogenous covariates. Lemma 2 implies that the pseudo-true value $\overline{\xi}(\theta_0)$ satisfies:

$$\mathbb{E}_{\pi_{0G}}\left(\frac{\partial \ln \pi_i\left(\alpha_{i0};\overline{\xi}(\theta_0)\right)}{\partial \xi}\right)=O\left(\frac{1}{T}\right). \tag{A6}$$

Note that $\overline{\xi}(\theta_0)$ is population specific. Moreover, it follows from the analysis in section 4 that $\pi_i(\alpha_i,\xi)$ is bias reducing if and only if $\pi_i\left(\alpha_i,\overline{\xi}(\theta_0)\right)$ is bias reducing, that is:

$$\mathbb{E}_{\pi_{0G}}\left(\left.\frac{\partial}{\partial \alpha_i}\right|_{\alpha_{i0}}\rho_i(\theta_0,\alpha_i)+\rho_i(\theta_0,\alpha_{i0})\frac{\partial \ln \pi_i\left(\alpha_{i0};\overline{\xi}(\theta_0)\right)}{\partial \alpha_i}\right)=o(1). \tag{A7}$$

Here we ask the question: in which case is $\pi_i(\alpha_i,\xi)$ bias reducing *for all* π_{0G}? Clearly, this will hold if and only if (A7) holds for all π_{0G} such that (A6) is satisfied. We now provide a linear algebra interpretation of this statement, which leads to an explicit solution.

Let us consider the Hilbert space L^2, endowed with the inner product

$$<\varphi,\psi>=\int \varphi(\alpha)\psi(\alpha)d\alpha, \quad (\varphi,\psi)\in L^2\times L^2.$$

We have, for any function ψ:

$$\mathbb{E}_{\pi_{0G}}\left(\psi(\alpha_{i0})\right)=<\pi_{0G},\psi>.$$

So (A6) is equivalent to

$$< \frac{\partial \ln \pi_i \left(.; \overline{\xi}\left(\theta_0\right)\right)}{\partial \xi} - A_T, \pi_{0G} >= 0 \tag{A8}$$

and (A7) is equivalent to

$$< \frac{\partial \rho_i\left(\theta_0, .\right)}{\partial \alpha_i} + \rho_i\left(\theta_0, .\right) \frac{\partial \ln \pi_i \left(.; \overline{\xi}\left(\theta_0\right)\right)}{\partial \alpha_i} - B_T, \pi_{0G} >= 0 \tag{A9}$$

where $A_T = O\left(\frac{1}{T}\right)$ and $B_T = o(1)$.

So, $\pi_i\left(\alpha_i, \xi\right)$ is bias reducing for all π_{0G} if and only if, for all $\pi_{0G} \in L^2$ such that (A8) holds, (A9) holds also.[26]

Let $A^{\perp}$ denote the orthogonal complement of $A \subset L^2$. $\pi_i\left(\alpha_i, \xi\right)$ is thus bias reducing for all π_{0G} if and only if

$$\frac{\partial \rho_i\left(\theta_0, .\right)}{\partial \alpha_i} + \rho_i\left(\theta_0, .\right) \frac{\partial \ln \pi_i \left(.; \overline{\xi}\left(\theta_0\right)\right)}{\partial \alpha_i} - B_T \in \left[\left(\frac{\partial \ln \pi_i \left(.; \overline{\xi}\left(\theta_0\right)\right)}{\partial \xi} - A_T\right)^{\perp}\right]^{\perp}.$$

Now, as there is a finite number of first-order conditions in (A6), the vector space spanned by $\frac{\partial \ln \pi_i\left(.; \overline{\xi}\left(\theta_0\right)\right)}{\partial \xi} - A_T$ is finite dimensional. So (e.g., Griffel, 1989, p.66):

$$\left[\left(\frac{\partial \ln \pi_i \left(.; \overline{\xi}\left(\theta_0\right)\right)}{\partial \xi} - A_T\right)^{\perp}\right]^{\perp} = \text{Vect}\left(\frac{\partial \ln \pi_i \left(.; \overline{\xi}\left(\theta_0\right)\right)}{\partial \xi} - A_T\right),$$

where Vect(V) denotes the vector space spanned by V.

So (A7) and (A6) hold for all π_{0G} if and only if there exists a matrix $\Gamma(\theta_0)$, with as many columns as the number of hyperparameters ξ, such that:

$$\begin{aligned}\frac{\partial}{\partial \alpha_i}\bigg|_{\alpha_{i0}} \rho_i\left(\theta_0, \alpha_i\right) + \rho_i\left(\theta_0, \alpha_{i0}\right) \frac{\partial \ln \pi_i \left(\alpha_{i0}; \overline{\xi}\left(\theta_0\right)\right)}{\partial \alpha_i} - B_T \\ -\Gamma(\theta_0)\left(\frac{\partial \ln \pi_i \left(\alpha_{i0}; \overline{\xi}\left(\theta_0\right)\right)}{\partial \xi} - A_T\right) = 0,\end{aligned}$$

or, equivalently:

$$\frac{\partial}{\partial \alpha_i}\bigg|_{\alpha_{i0}} \rho_i\left(\theta_0, \alpha_i\right) + \rho_i\left(\theta_0, \alpha_{i0}\right) \frac{\partial \ln \pi_i \left(\alpha_{i0}; \overline{\xi}\left(\theta_0\right)\right)}{\partial \alpha_i} - \Gamma(\theta_0) \frac{\partial \ln \pi_i \left(\alpha_{i0}; \overline{\xi}\left(\theta_0\right)\right)}{\partial \xi} = o\left(1\right),$$

that is:

$$\frac{\partial}{\partial \alpha_i}\bigg|_{\alpha_{i0}} \left(\rho_i\left(\theta_0, \alpha_i\right) \pi_i\left(\alpha_i; \overline{\xi}\left(\theta_0\right)\right)\right) - \Gamma(\theta_0) \frac{\partial \pi_i \left(\alpha_{i0}; \overline{\xi}\left(\theta_0\right)\right)}{\partial \xi} = o\left(1\right).$$

This ends the proof. *Q.E.D.*

Proof of Corollary 1: In the location-scale case we have $\pi\left(\alpha_i\right) = \frac{1}{\sigma} f\left(\frac{\alpha_i - \mu}{\sigma}\right)$, where f is a known *pdf*, and μ and σ^2 are hyperparameters. Then (22) yields:[27]

$$\rho_i\left(\theta, \alpha_i\right) = \Gamma_1(\theta) + \Gamma_2(\theta)\left(\frac{\alpha_i - \overline{\mu}(\theta)}{\overline{\sigma}(\theta)}\right) + o\left(1\right).$$

[26] Strictly speaking, bias reduction holds for any *density* π_{0G}, so equations (A8) and (A9) hold only for all $\pi_{0G} \in L^2$ that are nonnegative and integrate to one. However, this does not matter for the argument.

[27] We are only looking at the solutions that satisfy: $\lim_{\alpha \to \pm\infty} \rho_i(\theta_0, \alpha)\pi_i\left(\alpha; \overline{\xi}(\theta_0)\right) = 0$.

The corollary follows. *Q.E.D.*

Proof of Corollary 2: In that case $\pi_i\left(\alpha_i\right)=\frac{1}{\sigma}f\left(\frac{\alpha_i-x_i'\mu}{\sigma}\right)$, and (22) yields:

$$\rho_i\left(\theta,\alpha_i\right)=\Gamma_1(\theta)x_i+\Gamma_2(\theta)\left(\frac{\alpha_i-x_i'\overline{\mu}(\theta)}{\overline{\sigma}(\theta)}\right)+o\left(1\right).$$

Q.E.D.

References

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

[2] Alvarez, J. and M. Arellano (2004): "Robust Likelihood Estimation of Dynamic Panel Data Models", unpublished manuscript.

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

[4] Arellano, M. and S. R. Bond (1991): "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations", *Review of Economic Studies*, 58, 277-297.

[5] Arellano, M. and B. Honoré (2001): "Panel Data Models: Some Recent Developments", in J. Heckman and E. Leamer (eds.), *Handbook of Econometrics*, vol. 5, North Holland, Amsterdam.

[6] Arellano, M., and J. Hahn (2006): "A Likelihood-based Approximate Solution to the Incidental Parameter Problem in Dynamic Nonlinear Models with Multiple Effects", unpublished manuscript.

[7] Arellano, M., and J. Hahn (2007): "Understanding Bias in Nonlinear Panel Models: Some Recent Developments,". In: R. Blundell, W. Newey, and T. Persson (eds.): *Advances in Economics and Econometrics, Ninth World Congress*, vol. 3, Cambridge University Press.

[8] Bekker, P.A. (1994): "Alternative Approximations to the Distributions of Instrumental Variable Estimators", *Econometrica*, 62, 657-681.

[9] Berger, J., B. Liseo, and R.L. Wolpert (1999): "Integrated Likelihood Methods for Eliminating Nuisance Parameters", *Statistical Science*, 14, 1–22.

[10] Bester, C. A. and C. Hansen (2005a): "A Penalty Function Approach to Bias Reduction in Non-linear Panel Models with Fixed Effects", unpublished manuscript.

[11] Bester, C. A. and C. Hansen (2005b): "Bias Reduction for Bayesian and Frequentist Estimators", unpublished manuscript.

[12] Carro, J. (2007): "Estimating Dynamic Panel Data Discrete Choice Models with Fixed Effects", *Journal of Econometrics*, 127, 503-528.

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

[14] Chamberlain, G. (1984): "Panel Data", in Z. Griliches and M. D. Intriligator (eds.), *Handbook of Econometrics*, Vol. 2, Elsevier Science.

[15] Chamberlain, G. and G. Imbens (2004): "Random Effects Estimators with many Instrumental Variables", *Econometrica*, 72, 295-306.

[16] Chamberlain, G. and M. Moreira (2008): "Decision Theory Applied to a Linear Panel Data Model", *Econometrica*, forthcoming.

[17] Chernozhukov, V. and H. Hong (2003): "An MCMC Approach to Classical Estimation", *Journal of Econometrics*, 115, 293–346.

[18] Cho, M.H., J. Hahn, and G. Kuersteiner (2004): "Asymptotic Distribution of Misspecified Random Effects Estimator for a Dynamic Panel Model with Fixed Effects When Both n and T are large", *Economics Letters*, 84, 117–125.

[19] 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.

[20] Dhaene, G., K. Jochmans, and B. Thuysbaert (2006): "Split-Panel Jacknife Estimation of Fixed Effects Models", unpublished manuscript.

[21] DiCiccio, T. J. and S. E. Stern (1993): "An adjustment to Profile Likelihood Based on Observed Information", Technical Report, Department of Statistics, Stanford University.

[22] Ghosal, S., and A. W. Van der Vaart (2001): "Rates of Convergence for Bayes and Maximum Likelihood Estimation for Mixture of Normal Densities", *Annals of Statistics*, 29, 1233–1263.

[23] Ghosal, S., and A. W. Van der Vaart (2007): "Posterior Convergence Rates of Dirichlet Mixtures of Normal Distributions at Smooth Densities", *Annals of Statistics*, 35, 697–723.

[24] Griffel, D.H. (1989): *Linear Algebra and its Applications. Volume 2: More Advanced.* New York: Ellis Horwood.

[25] Hahn, J. (2000): "Parameter orthogonalization and Bayesian inference", unpublished manuscript.

[26] Hahn, J. (2004): "Does Jeffrey's Prior Alleviate the Incidental Parameter Problem?", *Economics Letters*, 82, 135–138.

[27] Hahn, J., and W.K. Newey (2004): "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models", *Econometrica*, 72, 1295–1319.

[28] Hahn, J., and G. Kuersteiner (2004): "Bias Reduction for Dynamic Nonlinear Panel Models with Fixed Effects", unpublished manuscript.

[29] Hahn, J., G. Kuersteiner, and W. Newey (2004): "Higher Order Efficiency of Bias Corrections", unpublished manuscript.

[30] Hospido, L. (2006): "Modelling Heterogeneity and Dynamics in the Volatility of Individual Wages", unpublished manuscript.

[31] Lancaster, T. (1998): "Panel Binary Choice with Fixed Effects", unpublished manuscript.

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

[33] Lancaster, T. (2002): "Orthogonal Parameters and Panel Data", *Review of Economic Studies*, 69, 647–666.

[34] Neyman, J. and E. L. Scott (1948): "Consistent Estimates Based on Partially Consistent Observations", *Econometrica*, 16, 1–32.

[35] Pace, L. and A. Salvan (2006): "Adjustments of the Profile Likelihood from a New Perspective", *Journal of Statistical Planning and Inference*, 136, 3554–3564.

[36] Severini, T. A. (1999): "On the Relationship Between Bayesian and Non-Bayesian Elimination of Nuisance Parameters", *Statistica Sinica*, 9, 713–724.

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

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

[39] Tierney, L., R.E. Kass and J.B. Kadane (1989): "Fully Exponential Laplace Approximations to Expectations and Variances of Nonpositive Functions", *J. Am. Stat. Ass.,* 84, 710–716.

[40] Wasserman, L. (2000): "Asymptotic Inference for Mixture Models Using Data-Dependent Priors", *Journal of the Royal Statistical Society*, Series B, 62, 159–180.

[41] Wong, W.H., and X. Shen (1995): "Probability Inequalities for Likelihood Ratios and Convergence Rates of Sieve MLEs", *Annals of Statistics*, 23, 339–362.

[42] Woutersen, T. (2002): "Robustness against Incidental Parameters", unpublished manuscript.

TABLE I

Various Estimators of θ in the Static Logit Model[a]

$T = 5$

	Mean	Median	STD	$\widehat{p}, .05$	$\widehat{p}, .10$	MSE	MAE
uncorrected	1.33	1.30	.235	.929	1.08	.163	.335
corrected	1.12	1.08	.188	.838	.868	.0489	.170
uniform	1.61	1.62	.260	1.22	1.29	.442	.613
Lancaster	1.06	1.05	.150	.800	.843	.0260	.126
robust, observed	1.11	1.09	.199	.821	.867	.0523	.176
robust, infeasible	1.18	1.17	.146	.950	.963	.0530	.193
robust, iterated 1	1.13	1.14	.184	.878	.914	.0504	.172
robust, iterated ∞	1.23	1.22	.195	1.01	1.03	.0907	.236
random effects	1.14	1.13	.163	.854	.905	.0418	.178
conditional logit	.997	.983	.172	.749	.793	.0283	.138

$T = 10$

	Mean	Median	STD	$\widehat{p}, .05$	$\widehat{p}, .10$	MSE	MAE
uncorrected	1.13	1.13	.117	.950	.994	.0296	.140
corrected	1.06	1.05	.0975	.902	.927	.0136	.0943
uniform	1.26	1.26	.147	1.05	1.06	.0893	.263
Lancaster	1.02	1.03	.0911	.880	.899	.00880	.0790
robust, observed	1.05	1.05	.109	.884	.909	.0145	.0974
robust, infeasible	1.07	1.06	.100	.895	.933	.0142	.0946
robust, iterated 1	1.04	1.04	.0892	.918	.932	.00976	.0785
robust, iterated ∞	1.08	1.06	.0896	.939	.970	.0139	.0938
random effects	1.03	1.03	.0986	.865	.906	.00848	.0832
conditional logit	.997	.998	.0961	.859	.884	.0105	.0754

$T = 20$

	Mean	Median	STD	$\widehat{p}, .05$	$\widehat{p}, .10$	MSE	MAE
uncorrected	1.06	1.06	.0683	.947	.971	.00826	.0757
corrected	1.02	1.03	.0606	.912	.946	.00424	.0530
uniform	1.12	1.11	.0683	.990	1.03	.0184	.119
Lancaster	.997	.997	.0548	.900	.921	.00298	.0429
robust, observed	1.01	1.00	.0702	.905	.929	.00500	.0527
robust, infeasible	1.04	1.04	.0613	.923	.955	.00558	.0629
robust, iterated 1	1.01	1.00	.0673	.885	.934	.00459	.0536
robust, iterated ∞	1.02	1.02	.0688	.893	.948	.00525	.0567
random effects	1.02	1.01	.0664	.920	.940	.00579	.0523
conditional logit	1.01	.995	.0682	.905	.920	.00492	.0535

[a] Estimates of θ in model (38). $N = 100$, 100 simulations. $\theta_0 = 1$.

TABLE II

Various Estimators of μ_1 in the Dynamic AR(1) Model[a]

	$T = 10$						
	Mean	Median	STD	$\widehat{p}, .05$	$\widehat{p}, .10$	MSE	MAE
uncorrected	.333	.328	.0320	.288	.300	.0290	.167
corrected, $q = 1$	.391	.390	.0341	.336	.342	.0131	.109
corrected, $q = 2$	.402	.402	.0327	.348	.359	.0107	.0984
corrected, $q = 3$	.384	.384	.0343	.328	.340	.0145	.116
uniform	.336	.335	.0330	.277	.296	.0281	.164
Lancaster	.504	.506	.0374	.435	.455	.00140	.0302
robust, observed $q = 1$	.393	.394	.0296	.335	.352	.0123	.107
robust, observed $q = 2$	.409	.413	.0304	.356	.368	.00920	.0910
robust, observed $q = 3$	.394	.395	.0345	.332	.342	.0125	.106
robust, infeasible	.500	.502	.0302	.449	.455	.000903	.0240
robust, iterated 1	.479	.477	.0299	.429	.436	.00133	.0299
robust, iterated ∞	.499	.497	.0323	.445	.455	.00104	.0264
GMM	.455	.459	.0608	.340	.373	.00567	.0602
random effects (uncorr.)	.562	.560	.0501	.448	.498	.00629	.0663
random effects (corr.)	.500	.498	.0348	.435	.461	.00120	.0274

[a] Estimates of μ_1 in model (39). $N = 100$, 100 simulations. $\mu_{10} = .5$.

TABLE III

Various Estimators of (μ_1, μ_2) in the Dynamic AR(2) Model[a]

	$T = 10$			
	Mean $\widehat{\mu}_1$	MSE $\widehat{\mu}_1$	Mean $\widehat{\mu}_2$	MSE $\widehat{\mu}_2$
uncorrected	.385	.0146	-.0774	.00700
corrected, $q = 1$	.419	.00808	-.101	.0111
corrected, $q = 2$	.423	.00734	-.0780	.00715
uniform	.369	.0189	-.104	.0119
robust, observed $q = 1$	.451	.00371	-.137	.0198
robust, observed $q = 2$	.435	.00602	-.0873	.00868
robust, infeasible	.451	.00352	-.00801	.00117
robust, iterated 1	.441	.00455	-.0262	.00203
robust, iterated ∞	.446	.00405	-.0187	.00175
GMM	.440	.00739	-.0278	.00297

[a] Estimates of μ_1 and μ_2 in model (39). $N = 100$, 100 simulations. $\mu_{10} = .5$, $\mu_{20} = 0$.

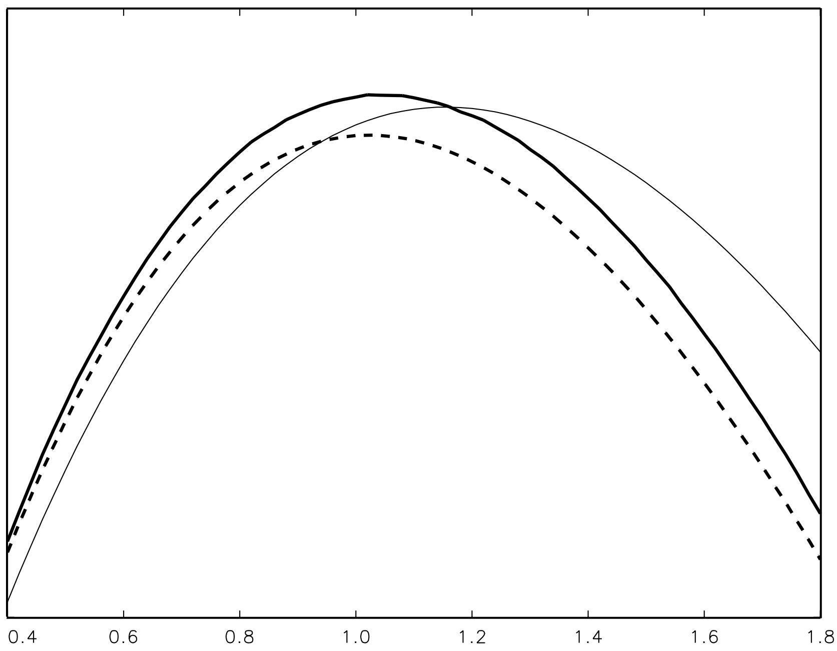

FIGURE 1. Likelihood functions in the static logit model ($T = 10$, $N = 100$, $\theta_0 = 1$). The thin line represents the likelihood function, the thick line the bias-corrected likelihood using DiCiccio and Stern (1993), and the dashed line represents the robust integrated likelihood.

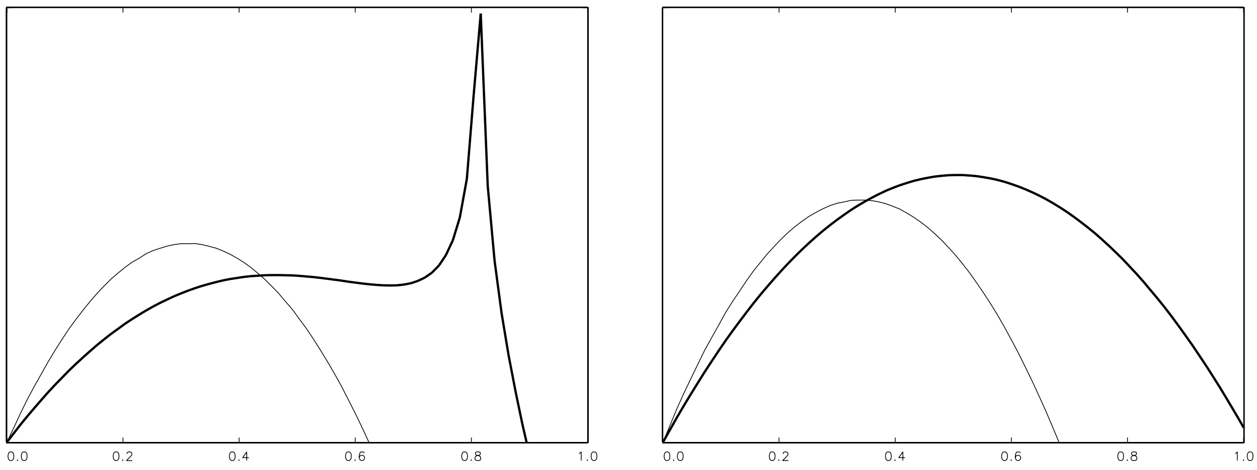

FIGURE 2. Likelihood functions in the dynamic AR(1) model (one simulation, $T = 10$, $N = 100$, $\mu_{10} = .5$). The thin line represents the likelihood function, and the thick line represents the robust integrated likelihood. Left: prior based on equation (12). Right: prior based on equation (14).