Robust Priors in Nonlinear Panel Data Models*

Manuel Arellano
CEMFI, Madrid

Stéphane Bonhomme
CEMFI, Madrid

First draft: June 2006
This version: December 2006

Abstract

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 must depend on the data 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. Finally, we show that random effects estimators are not bias reducing in general and discuss important exceptions. Three examples and some Monte Carlo experiments illustrate the results.

JEL CODE: C23.
KEYWORDS: Panel data, incidental parameters, bias reduction, integrated likelihood, priors.

*We thank Victor Aguirregabiria, Gabriele Fiorentini, Jean-Pierre Florens, Jinyong Hahn, Laura Hospido, Thierry Magnac, Jean-Marc Robin, Enrique Sentana, and seminar audiences at CEMFI, Toulouse, and University College London, for useful comments.

1 Introduction

In a panel model the likelihood of the data for a given unit is typically a function $f_i(\theta, \alpha_i)$ 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:

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

where $w_i(\alpha_i)$ is a possibly θ-specific weight and $d\alpha_i$ is 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: $\sum_{i=1}^{N} \ln f_i^a(\theta)$.

A fixed effects approach that estimates θ jointly with the individual effects by maximum likelihood (ML) falls in this category with weights

$$w_i(\alpha_i) = \begin{cases} 1 & \text{if } \alpha_i = \widehat{\alpha}_i(\theta) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $\widehat{\alpha}_i(\theta)$ is the maximum likelihood estimator of α_i for given θ. The resulting average likelihood in this case is just the concentrated likelihood $f_i(\theta, \widehat{\alpha}_i(\theta))$.

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 $w_i(\alpha_i)$ is a parametric or semiparametric density or probability mass function which does not depend on θ, but includes additional unknown coefficients:

$$w_i(\alpha_i) = g_i(\alpha_i; \xi).$$

Finally, in a Bayesian approach, an average likelihood is also constructed, choosing as weights 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. 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.[1]

The traditional reaction to this problem has been to look for estimators yielding fixed-T consistency as N goes to infinity.[2] 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.[3]

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), of the moment equation (Woutersen 2002, Arellano 2003, Carro 2006) or of the objective function (Arellano and Hahn 2006a,b, 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 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, $w_i(\alpha_i)$.[4] For fixed T, we can define the pseudo true value $\theta_T = \text{plim}_{N\to\infty} \widehat{\theta}$. In

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

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

[3] 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.

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

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.[5] 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 make four contributions. First, 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. This allows us to make the link with information orthogonality. Namely, we show that information orthogonality or an essentially equivalent condition is both necessary and sufficient for the uniform prior on the fixed effects to be bias reducing.

Moreover, when (generalized) orthogonal reparameterizations of the fixed effects are not available, every bias reducing prior has to be data dependent. We denote as "data dependent" a theoretical weight function which depends on the true values θ_0 and α_{i0}:

$$w_i(\alpha_i) = \pi_i(\alpha_i \mid \theta; \theta_0, \alpha_{i0}),$$

so that a feasible counterpart will depend on the data in general.

In a second contribution, we discuss two special bias reducing priors. The first one, that we call the "robust" prior, can be written as a combination of a Hessian and an outer product

[5] 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.

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:

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

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. Both robust weighting schemes are functions of the data.

The third contribution concerns estimation and inference from the integrated likelihood. As the expression of the robust priors is close to additive corrections of the bias of the concentrated likelihood (e.g. Di Ciccio and Stern, 1993), one can choose among several already available methods to find a feasible counterpart for the weight function. Then, 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. Simulation methods can also be useful to compute confidence intervals. Building on the results in Chernozhukov and Hong (2003), we show that asymptotically valid confidence intervals of the parameter estimates can be read from the quantiles of the pseudo-posterior distribution when N and T grow at the same rate.

Finally, we study the existence of bias reducing priors on the individual effects that are independent of the common parameters, as is the case in the context of random-effects models, which are very popular in applied work. 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. In particular, we derive a necessary and sufficient condition for the Gaussian random effects maximum likelihood (REML) estimator to be bias reducing. An important special case is the class of linear autoregressive models. In more general nonlinear models, however, the use of Gaussian REML has no bias-reducing asymptotic justification.

The related literature includes Woutersen (2002), which obtained the first-order bias of the integrated likelihood 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. In Section 4, we illustrate these results by means of three examples: the dynamic AR(p) model, the Poisson counts model and the static logit model with fixed effects. In Section 5, we discuss issues of estimation and inference. Section 6 focuses on the bias reducing properties of random effects estimators. In Section 7, we report a small Monte-Carlo simulation to study the finite-sample behavior of the proposed estimators. Lastly, Section 8 concludes.

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 likelihood conditioned on 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 $\alpha_1 ... \alpha_N$.[6] 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 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 will assume that the order of magnitude of the logarithm of the prior is bounded when T increases:

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

$$\ln \pi_i(\alpha_i|\theta) = O(1).$$

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

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

Hahn, 2006a). 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:

$$\pi(\theta, \alpha_1 ... \alpha_N) = \pi(\theta)\pi_1(\alpha_1|\theta) ... \pi_N(\alpha_N|\theta).$$

From this perspective, in this paper we will assume a uniform prior on θ: $\pi(\theta) \propto 1$.[7] 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 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 will provide 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 likelihoods can be regarded as integrated likelihoods with respect to the priors

$$\overline{\pi}_i(\alpha_i|\theta) = \mathbf{1}\{\alpha_i = \overline{\alpha}_i(\theta)\}, \text{ and } \pi_i{}^c(\alpha_i|\theta) = \mathbf{1}\{\alpha_i = \widehat{\alpha}_i(\theta)\},$$

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:

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

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

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:

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

for a given prior $\pi_i(\alpha_i|\theta)$.[8] 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 likelihoods. The result is contained in the following lemma.

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

$$\mathbb{E}_{\theta_0,\alpha_{i0}}\left[\ell_i^I(\theta)-\ell_i^c(\theta)\right]=C^{st}-\frac{1}{2T}\ln\mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\alpha_i}(\theta,\overline{\alpha}_i(\theta))\right]+\frac{1}{T}\ln\pi_i(\overline{\alpha}_i(\theta)|\theta)+O\left(\frac{1}{T^2}\right). \quad (3)$$

Proof. See Appendix. ■

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

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

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

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

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

where

$$\begin{aligned}B_i(\theta) &= \frac{1}{2}\left\{\mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\alpha_i}(\theta,\overline{\alpha}_i(\theta))\right]\right\}^{-1}\mathbb{E}_{\theta_0,\alpha_{i0}}\left[Tv_i^2(\theta,\overline{\alpha}_i(\theta))\right]\\ &\quad -\frac{1}{2}\ln\mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\alpha_i}(\theta,\overline{\alpha}_i(\theta))\right]+\ln\pi_i(\overline{\alpha}_i(\theta)|\theta). \quad (5)\end{aligned}$$

[8] 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.

Proof. Combining (3) and (4). ■

As the right-hand side of (5) is $O(1)$, Theorem 1 illustrates the "dominance" argument (e.g. Lancaster, 2004) 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 6, 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, $B_i(\theta)$ is not locally constant around θ_0.

2.3 Bias of the integrated score

From Theorem 1 we can obtain the expression of the bias of the integrated score evaluated at the true value θ_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:

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

The bias of the integrated score is thus given by:

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

Hence the following characterization of bias reducing priors:

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

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

Proof. The condition is an immediate application of (7). Then, lack of first-order bias of the estimator follows from lack of first-order bias in the score or estimating equation. For a theory for general bias corrected estimating equations, see Arellano and Hahn (2006b), for example. ■

2.4 Non-data 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:*

$$\frac{\partial}{\partial\theta}\Big|_{\theta_0} B_i(\theta) = \frac{\partial}{\partial\theta}\Big|_{\theta_0} \ln \pi_i(\overline{\alpha}_i(\theta)|\theta) + \frac{\partial}{\partial\alpha_i}\Big|_{\alpha_{i0}} \rho_i(\theta_0, \alpha_i) \tag{8}$$

where

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

Proof. See Appendix. ■

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

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

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

$$\frac{\partial}{\partial\alpha_i}\Big|_{\alpha_{i0}} \rho_i(\theta_0, \alpha_i) = O\left(\frac{1}{T}\right). \tag{9}$$

Condition (9) 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 $\theta_0, \alpha_{10}, ..., \alpha_{N0}$. Other non-data dependent priors are given by orthogonal reparameterizations of the fixed effects, when available. Let $\psi_i = \psi_i(\alpha_i, \theta)$ be a reparameterization such that ψ_i and θ are information orthogonal in the sense of equation (9). In that case equation (9) shows that the uniform prior on ψ_i is bias-reducing. Hence 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, as this prior is the Jacobian of the transformation which maps (α_i, θ) onto (ψ_i, θ). Conversely, any non-data dependent bias-reducing prior $\pi_i(\alpha_i|\theta)$ can be associated an orthogonal reparameterization in the sense of equation (9). 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-data dependent bias reducing priors and orthogonal reparameterizations in the sense of (9). 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). Appendix B makes this statement more precise. Hence, to deal with the general case where orthogonal reparameterizations are not available, it is necessary to search for robust priors that depend on the data. We address this task in the next section.

Note also that to every reparameterization of the fixed effects $\psi_i(\alpha_i, \theta)$, and every prior $\widetilde{\pi}_i(\psi_i|\theta)$ on ψ_i we can associate the transformed prior in the original parameterization:

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

Then we show the following result in Appendix, which is a corollary of Theorem 2.

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.*

Proof. See Appendix ■

Proposition 2 shows that the bias reducing properties of a prior are not affected by a reparameterization of the effects.

3 Two bias reducing priors

3.1 Robust prior

Theorem 1, together with equation (6), show that the following prior is robust, in the sense that it yields first-order unbiasedness:

$$\pi_i^R(\alpha_i|\theta) \propto \mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\alpha_i}(\theta, \alpha_i)\right] \left\{\mathbb{E}_{\theta_0,\alpha_{i0}}\left[v_i^2(\theta, \alpha_i)\right]\right\}^{-1/2}. \tag{10}$$

This bias-reducing prior (10), which we will call the "robust" prior, is data dependent, as both expectation terms depend on the true parameters θ_0 and α_{i0}.[9] In particular, different robust priors are associated with different individual units. The discussion in the previous section has shown that non-data dependent priors cannot be robust in cases when orthogonal reparameterizations of the fixed effects are not available.

Moreover, π_i^R involves a Hessian term ($\mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\alpha_i}(\theta,\alpha_i)\right]$) and an outer product term ($\mathbb{E}_{\theta_0,\alpha_{i0}}\left[v_i^2(\theta,\alpha_i)\right]$). A closely related expression appears in Jeffreys' automatic prior when θ is kept fixed, the expression of which is:

$$\pi_i^J(\alpha_i|\theta) \propto \left\{\mathbb{E}_{\theta,\alpha_i}\left[-v_i^{\alpha_i}(\theta,\alpha_i)\right]\right\}^{1/2}. \tag{11}$$

A crucial difference between $\pi_i^R(\alpha_i|\theta)$ and $\pi_i^J(\alpha_i|\theta)$ is that in the latter the expectation is taken with respect to $\exp\left[T\ell_i\left(\theta,\alpha_i\right)\right]$ as opposed to $\exp\left[T\ell_i\left(\theta_0,\alpha_{i0}\right)\right]$. Thus, in particular Jeffreys' prior does not depend on the data. Evaluated at true values, the robust prior π^R boils down to Jeffreys'. However, the distinction between arbitrary parameter values and true values appears only in (10), and is critical in ensuring bias reduction. In fact, Jeffreys' prior (11) 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 (6), 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 (5), it is straightforward to show that the following prior is bias reducing in both likelihood and pseudo-likelihood settings:

$$\left\{\mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\alpha_i}(\theta,\alpha_i)\right]\right\}^{1/2}\exp\left(-\frac{T}{2}\left\{\mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\alpha_i}(\theta,\alpha_i)\right]\right\}^{-1}\mathbb{E}_{\theta_0,\alpha_{i0}}\left[v_i^2(\theta,\alpha_i)\right]\right). \tag{12}$$

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 (10) and (12) are bias reducing. Using (12) instead of (10) 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.

[9] Thus $\pi_i^R(\alpha_i|\theta)$ should be regarded as a shorthand for $\pi_i^R(\alpha_i|\theta;\theta_0,\alpha_{i0})$.

Proposition 3

$$\pi_i^R(\widehat{\alpha}_i(\theta)|\theta) \propto \frac{1}{\sqrt{\text{Var}\left(\widehat{\alpha}_i(\theta)\right)}}\left(1 + O_p\left(\frac{1}{T}\right)\right). \tag{13}$$

In addition, every non-dogmatic prior satisfying (13) is bias reducing.

Proof. See Appendix. ■

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

$$\psi_i(\alpha_i, \theta) = \frac{\alpha_i - \widehat{\alpha}_i(\theta)}{\sqrt{\text{Var}\left(\widehat{\alpha}_i(\theta)\right)}}. \tag{14}$$

Reparameterizing the individual effects as in (14) 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 *cdf* F and *pdf* f. In terms of the original parameterization, the prior is:[10]

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

Then, clearly:

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

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 (large-T) asymptotic sampling distribution of the MLE $\widehat{\alpha}_i(\theta)$ is a bias reducing prior for α_i:

$$\alpha_i|\theta \sim \quad \mathcal{N}(\widehat{\alpha}_i(\theta), \text{Var}\left(\widehat{\alpha}_i(\theta)\right)). \tag{15}$$

Specifying the *a priori* distribution of the fixed effects as in (15) is intuitively appealing. 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 associated with the concentrated likelihood (π_i^c), it takes into account the way the precision of $\widehat{\alpha}_i(\theta)$ varies with θ. In the limit, if $\text{Var}\left(\widehat{\alpha}_i(\theta)\right)$ varies slowly with θ then we obtain the uniform prior on the original effects. This happens when parameters are information orthogonal.

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

4 Examples

We turn to consider three specific examples: the dynamic AR(p) model, the Poisson counts model, and the static logit model.

4.1 Dynamic AR(p)

The model we consider is given by:

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

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 later, 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 Appendix C that the robust prior can be written as:

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

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:

$$\widetilde{\pi}^R\left(\alpha_i | \mu, \sigma^2\right) \propto \{1 + a(\mu - \mu_0)\}^{-1/2}. \tag{16}$$

The prior $\widetilde{\pi}^R$ is also bias-reducing. Note that, as $a(\mu-\mu_0)$ is linear, the function $\widetilde{\pi}^R\left(\alpha_i|\mu,\sigma^2\right)$ 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 (12) for the robust prior and consider instead:

$$\widetilde{\pi}^R\left(\alpha_i|\mu,\sigma^2\right) \propto \exp\left(-\frac{1}{2}a(\mu-\mu_0)\right). \tag{17}$$

Now, the priors given by (16) and (17), are data dependent because a depends on μ_0. Looking for a non-data dependent prior requires solving:

$$\frac{\partial}{\partial\mu}\Big|_{\mu_0,\sigma_0^2} \ln\pi\left(\overline{\alpha}_i\left(\mu,\sigma^2\right)|\mu,\sigma^2\right) \propto \frac{\partial}{\partial\mu}\Big|_{\mu_0} \ln\left(\left\{1+a(\mu-\mu_0)\right\}^{-1/2}\right), \tag{18}$$

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

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

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

In this case, equation (18) admits solutions independent of true parameter values. For example, the following choice works:

$$\pi\left(\alpha_i|\mu,\sigma^2\right) = \exp\left(\frac{1}{T}\sum_{t=1}^{T-1}\frac{T-t}{t}\mu^t\right). \tag{19}$$

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. In Section 6 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.

4.2 Poisson counts

Let the data consist of T Poisson counts y_{it} with individual means:

$$\mathbb{E}_{\theta_0,\alpha_{i0}}(y_{it}) = \alpha_{i0}\exp(x_{it}'\theta_0), \quad i=1...N, \quad t=1...T,$$

where x_{it} are known covariates. The individual log-likelihood is given by:

$$\ell_i(\theta,\alpha_i) \propto -\alpha_i\frac{1}{T}\sum_{t=1}^{T}\exp(x_{it}'\theta) + \frac{1}{T}\sum_{t=1}^{T}y_{it}\ln(\alpha_i) + \frac{1}{T}\sum_{t=1}^{T}y_{it}x_{it}'\theta.$$

We show in Appendix C that the robust prior is given by:

$$\pi_i^R(\alpha_i|\theta) \propto \frac{1}{\alpha_i}\left(\sum_{t=1}^{T}\alpha_{i0}\exp(x_{it}'\theta_0) + \left[\alpha_{i0}\exp(x_{it}'\theta_0) - \alpha_i\exp(x_{it}'\theta)\right]^2\right)^{-1/2}. \tag{20}$$

Then, by Proposition 1 one can add a quadratic adjustment in $(\theta-\theta_0)$ and $(\alpha_i-\alpha_{i0})$ to the logarithm of π_i^R without altering its bias properties. It follows that:

$$\widetilde{\pi}(\alpha_i|\theta) \propto \frac{1}{\alpha_i} \tag{21}$$

is also bias-reducing. Note that π_i^R is proper, while $\widetilde{\pi}$ is not.

As in Lancaster (2002), let us consider the reparameterization: $\psi_i = \alpha_i\sum_{t=1}^{T}\exp(x_{it}'\theta)$. Then it is straightforward to show that: $\frac{\partial^2\ell_i(\theta,\psi_i)}{\partial\theta\partial\psi_i} = 0$. In this reparameterized model, parameters are fully orthogonal, not just information orthogonal. In particular, the uniform prior is bias-reducing. Therefore, in terms of the original reparameterization, the following prior reduces bias:

$$\pi_i\left(\alpha_i|\theta\right) \propto \left|\frac{\partial\psi_i(\alpha_i,\theta)}{\partial\alpha_i}\right| = \sum_{t=1}^{T}\exp(x_{it}'\theta).$$

Interestingly, the robust prior and Lancaster's prior are directly related, as:[11]

$$\pi_i{}^R(\overline{\alpha}_i(\theta)|\theta) \propto \widetilde{\pi}(\overline{\alpha}_i(\theta)|\theta) = \sum_{t=1}^{T}\exp(x_{it}'\theta) = \pi_i(\alpha_i|\theta).$$

4.3 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 Appendix C we derive the expression of the robust prior:

$$\pi_i^R(\alpha_i|\theta) \propto \left(\sum_{t=1}^{T}\mathbb{E}_{\theta_0,\alpha_{i0}}\left(\left[y_{it}-\Lambda(x_{it}'\theta+\alpha_i)\right]^2\right)\right)^{-1/2}\sum_{t=1}^{T}\Lambda(x_{it}'\theta+\alpha_i)\left[1-\Lambda(x_{it}'\theta+\alpha_i)\right]. \tag{22}$$

[11] This result follows directly from the expression of $\overline{\alpha}_i(\theta)$ given in the appendix.

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:

$$\pi_i(\alpha_i|\theta) \propto \sum_{t=1}^{T} \Lambda(x_{it}'\theta + \alpha_i)\left[1 - \Lambda(x_{it}'\theta + \alpha_i)\right]. \tag{23}$$

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^R and π_i reduce bias.

5 Estimation and inference

The previous analysis has shown that, absent the possibility of orthogonalization, the only priors that lead to bias reduction are data-dependent priors.[12] We here explain how to find feasible counterparts for the robust priors, and we consider methods to perform the estimation of θ. We then discuss inference issues.

5.1 Estimation

Prior. The expression for the robust prior (10) is very similar to the expression for the bias of the concentrated likelihood given by equation (4). It also involves the Hessian term $\mathbb{E}_{\theta_0,\alpha_{i0}}\left(-v_i^{\alpha_i}(\theta,\alpha_i)\right)$, as well as the outer product term $\mathbb{E}_{\theta_0,\alpha_{i0}}\left(v_i^2(\theta,\alpha_i)\right)$. For this reason, the problem of finding a feasible counterpart for the robust prior is analogous to the problem of estimating an additive bias correction for the concentrated likelihood.

The Hessian term can be consistently estimated by the observed Hessian:

$$-\frac{1}{T}\sum_{t=1}^{T} v_{it}^{\alpha_i}(\theta,\alpha_i) = \frac{1}{T}\sum_{t=1}^{T} \frac{\partial^2 \ell_{it}(\theta,\alpha_i)}{\partial \alpha_i^2}.$$

Moreover, in the case of independent observations the outer product term can be consistently estimated by

$$\frac{1}{T^2}\sum_{t=1}^{T} v_{it}^2(\theta,\alpha_i) = \frac{1}{T^2}\sum_{t=1}^{T}\left(\frac{\partial \ell_{it}(\theta,\alpha_i)}{\partial \alpha_i}\right)^2.$$

[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.

However, when observations are not independent the same strategy cannot be applied because of the dynamic dependence of the score. Indeed, as $v_i(\theta, \widehat{\alpha}_i(\theta)) = \frac{1}{T}\sum_{t=1}^{T} v_{it}(\theta, \widehat{\alpha}_i(\theta)) = 0$, it follows that the expression: $-v_i^{\alpha_i}(\theta, \alpha_i)\left\{v_i^2(\theta, \alpha_i)\right\}^{-1/2}$ is degenerate at $(\theta, \widehat{\alpha}_i(\theta))$.

One possibility to estimate the outer product term consistently is to use expected quantities. Note that estimation of the expectation requires to plug-in consistent estimates of the true parameters (θ_0, α_{i0}). Another possibility is to use a trimmed version of the empirical mean, as in Hahn and Kuersteiner (2004) or Arellano and Hahn (2006b). Lastly, one can make use of the identity (13) and estimate the variance of $\widehat{\alpha}_i(\theta)$ by parametric bootstrap. This last idea was proposed by Pace and Salvan (2006) in the context of bias correction of the concentrated likelihood.

Example. In the static logit example (see 4.3), one can use observed quantities and compute:

$$\widehat{\pi}_i^R(\alpha_i|\theta) \propto \left\{\sum_{t=1}^{T}\left((y_{it} - \Lambda(x_{it}'\theta + \alpha_i))^2\right)\right\}^{-1/2} \sum_{t=1}^{T}\Lambda(x_{it}'\theta + \alpha_i)\left[1 - \Lambda(x_{it}'\theta + \alpha_i)\right]. \quad (24)$$

One can also use expected quantities as:

$$\begin{aligned}\widehat{\widehat{\pi}}_i^R(\alpha_i|\theta) \quad \propto \quad & \left\{\sum_{t=1}^{T}\Lambda(x_{it}'\widehat{\theta} + \widehat{\alpha}_i)\left[1 - 2\Lambda(x_{it}'\theta + \alpha_i)\right] + \left[\Lambda(x_{it}'\theta + \alpha_i)\right]^2\right\}^{-1/2} \times \\ & \sum_{t=1}^{T}\Lambda(x_{it}'\theta + \alpha_i)\left[1 - \Lambda(x_{it}'\theta + \alpha_i)\right], \end{aligned} \quad (25)$$

where $\widehat{\theta}$ and $\widehat{\alpha}_i$ are consistent estimates of the true parameters when T tends to infinity. Maximum Likelihood estimates are natural candidates.

Estimation of common parameters. Once a feasible robust weighting scheme is available, 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, as explained in the next subsection.

Freedom to choose. Lastly, it is worth reiterating that the robust prior can be modified in a way that does not create first-order bias, but can make a difference in finite samples. This gives the researcher some degree of freedom in her choice of prior, even if this choice is constrained by the fact that the bias of the score of the integrated likelihood has to be (asymptotically close to) zero. In the case of the dynamic AR(p) model studied in the previous section, arbitrary quadratic terms in $\mu - \mu_0$ and $\alpha_i - \alpha_{i0}$ can be added to the prior while keeping the bias-reduction property. However, linear terms cannot be changed without creating bias. As showed by the Poisson counts and static logit examples, this property is not limited to linear models. The choice of bias-reducing priors in practice remains an open area for study.

5.2 Inference

Let ℓ_i^I be associated with a bias reducing prior. We define $\widehat{\theta}$ as:

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

The analysis below still applies if instead of the mode of the integrated likelihood one considers its mean or its median, provided that these quantities exist. Throughout this section, we assume that the integrated likelihood is proper:

$$\int \left\{ \prod_{i=1}^{N} \exp\left(T\ell_i^I(\theta)\right) \right\} d\theta < \infty.$$

In this section, we are concerned with computing confidence intervals for $\widehat{\theta}$. For this we need some additional notation. Let

$$H_{iT} = \mathbb{E}_{\theta_0,\alpha_{i0}}\left(-\frac{\partial^2 \ell_i^I(\theta_0)}{\partial\theta\partial\theta'}\right), \quad \text{and} \quad \Omega_{iT} = \mathbb{E}_{\theta_0,\alpha_{i0}}\left(\frac{\partial \ell_i^I(\theta_0)}{\partial\theta}\frac{\partial \ell_i^I(\theta_0)}{\partial\theta'}\right)$$

be the Hessian and the outer product of the individual integrated likelihood at the truth, and let

$$\overline{H}_{iT} = \mathbb{E}_{\theta_0,\alpha_{i0}}\left(-\frac{\partial^2 \overline{\ell}_i(\theta_0)}{\partial\theta\partial\theta'}\right), \quad \text{and} \quad \overline{\Omega}_{iT} = \mathbb{E}_{\theta_0,\alpha_{i0}}\left(\frac{\partial \overline{\ell}_i(\theta_0)}{\partial\theta}\frac{\partial \overline{\ell}_i(\theta_0)}{\partial\theta'}\right)$$

be the same quantities associated with the target likelihood. Then we have the following lemma.

Lemma 2 *Let $\ell_i^I(\theta)$ be an integrated likelihood associated with a prior such that:*

$$\frac{\partial}{\partial\theta}\Big|_{\theta_0} B_i(\theta) = O\left(\frac{1}{T}\right); \quad \frac{\partial^2}{\partial\theta\partial\theta'}\Big|_{\theta_0} B_i(\theta) = O\left(\frac{1}{T}\right), \tag{26}$$

where $B_i(\theta)$ is given by equation (5). Then:

$$H_{iT} = \overline{H}_{iT} + O\left(\frac{1}{T^2}\right), \tag{27}$$

$$\Omega_{iT} = \overline{\Omega}_{iT} + \Xi_{iT} + O\left(\frac{1}{T^2}\right), \tag{28}$$

where Ξ_{iT} is a term of order 1/T that does not depend on the expression of the prior.

Proof. See Appendix. ■

Conditions (26) are bias reduction conditions that are satisfied by all the robust priors derived in the previous sections, as well as by their robust approximations near the true parameter values.

To understand the lemma, one has to note that H_{iT} is $O(1)$ while Ω_{iT} is $O(1/T)$, as the prior is bias reducing. Lemma 2 thus shows that the Hessian of the integrated likelihood and that of the target are equal up to a small $1/\mathrm{T}^2$ term. However, the outer product terms of the integrated and the target likelihoods need not coincide to a 1/T order of magnitude, as in general the term Ξ_{iT} is not zero.

A first application of Lemma 2 is that the information bias, defined as:

$$\Delta_{iT} = T\Omega_{iT} + H_{iT},$$

is independent of the form of the robust prior used for estimation. In general, it is $O(1)$ as the target likelihood has no information bias. Di Ciccio *et al.* (1996) use a multiplicative correction on the score of the corrected concentrated likelihood that reduces the bias to an order $O(1/T)$. Our result shows that, in general, no prior reduces both the bias of the integrated likelihood and the information bias. The intuition behind this result is that, as the use of Laplace approximations makes clear, the prior behaves asymptotically as an additive correction to the concentrated likelihood.

We now turn to the computation of confidence intervals for θ. Let us assume to start with that T is fixed. Let us define the pseudo true value associated with the problem of maximizing the integrated likelihood:

$$\theta_T = \underset{\theta}{\operatorname{argmax}} \quad \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\theta_0,\alpha_{i0}}\left(\ell_i^I(\theta)\right).$$

Under fixed-T asymptotics, when N tends to infinity, one has:

$$\sqrt{NT}(\widehat{\theta} - \theta_T) \underset{d}{\to} \mathcal{N}(0, V_T), \tag{29}$$

and the asymptotic variance is given by the "sandwich" formula:

$$V_T = H_T^{-1} \Omega_T H_T^{-1}, \tag{30}$$

where

$$H_T = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} H_{iT}; \qquad \text{and} \qquad \Omega_T = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \Omega_{iT}.$$

As the information bias is not zero, the "sandwich" formula does not simplify. This is due to the fact that the integrated likelihood is not a proper likelihood. In particular, it does not satisfy Bartlett identities.

Let then

$$p(\theta) = \frac{\prod_{i=1}^{N} \exp\left[T\ell_i^I(\theta)\right]}{\int \left\{\prod_{i=1}^{N} \exp\left[T\ell_i^I(\theta)\right]\right\} d\theta} \tag{31}$$

be the pseudo-posterior distribution associated with the integrated likelihood and a uniform prior for θ.[13] As the integrated likelihood is assumed proper, the denominator exists.

Theorem 4 in Chernozhukov and Hong (2003) shows that, under suitable regularity conditions, p is asymptotically (when N tends to infinity for fixed T) equivalent to: $\mathcal{N}(\theta_T, H_T^{-1})$.

Now, (27) shows that H_T is equal to the Hessian of the target likelihood, up to second-order terms. Moreover, (28) makes clear that the outer products of the target and the integrated likelihoods are generally different. Hence, unlike the fixed-T variance V_T, H_T^{-1} does not take into account the variability of the fixed effects estimates in the calculation of the asymptotic distribution. This variability has only second-order effects on the confidence intervals. As a consequence, the quantiles of the pseudo-posterior distribution are generally not valid confidence intervals for θ_T under fixed-T asymptotics.

Nonetheless, the pseudo-posterior distribution of θ is a valid guide for making inference about θ_0 when N and T tend simultaneously to infinity at the same rate. Indeed, in that case the information matrix equality is satisfied at the limit and both V_T and H_T^{-1} tend to $V_\infty = H_\infty^{-1}$. It follows that:

$$\sqrt{NT}(\widehat{\theta} - \theta_0) \underset{d}{\to} \mathcal{N}(0, V_\infty), \qquad \text{when } N \to \infty, \qquad \text{and} \qquad \frac{T}{N} \to c > 0. \tag{32}$$

[13] Note that p is a *pseudo*-posterior distribution for θ, since ℓ^I is a pseudo-likelihood.

Therefore, in a double asymptotics perspective, V_T and H_T^{-1} have the same justification. However, a Bayesian derivation points to a justification of H_T^{-1} as providing confidence intervals that can be read directly from the quantiles of the pseudo-posterior distribution. In practice, the quantiles can be computed directly from the empirical distribution of $\widehat{\theta}$, e.g. given by a Markov chain.

6 Random-effects and bias reduction

In this section, we study the first-order bias properties of random-effects maximum likelihood (REML) estimators. We first focus on the case of an integrated likelihood with prior independence between α_i and θ. We then show that random-effects ML estimators can, to first-order, be embedded into this framework.

6.1 The random-effects model

In this section, 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. Importantly, $\pi_i(\alpha_i; \xi)$ does not depend directly on the common parameter θ, nor on the *cdf* 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.

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:

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

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

6.2 Random-effects without hyperparameters

For expositional simplicity, we start with the case where there are no hyperparameters ξ and the prior is given by $\pi_i(\alpha_i)$, independent of θ and independent of the data.

Let the bias of the score of the random-effects likelihood at the truth be:

$$\frac{B_\infty(\theta_0)}{T} \equiv \operatorname*{plim}_{N\to\infty} \quad \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \ell_i^{RE}(\theta_0)}{\partial \theta}.$$

Using Proposition 1 we obtain:

$$\begin{aligned} B_\infty(\theta_0) &= \operatorname*{plim}_{N\to\infty} \quad \frac{1}{N}\sum_{i=1}^{N}\left\{ \frac{\partial}{\partial\theta}\Big|_{\theta_0} \ln \pi_i(\overline{\alpha}_i(\theta)) + \frac{\partial}{\partial\alpha_i}\Big|_{\alpha_{i0}} \rho_i(\theta_0,\alpha_i)\right\} + O\left(\frac{1}{T}\right) \\ &= \lim_{N\to\infty} \quad \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\pi_0}\left(\frac{\partial}{\partial\theta}\Big|_{\theta_0} \ln \pi_i(\overline{\alpha}_i(\theta)) + \frac{\partial}{\partial\alpha_i}\Big|_{\alpha_{i0}} \rho_i(\theta_0,\alpha_i)\right) + O\left(\frac{1}{T}\right), \end{aligned} \tag{33}$$

where π_0 is the population distribution of the individual effects.[14] At this stage, it is useful to recall that:

$$\frac{\partial}{\partial\theta}\Big|_{\theta_0} \overline{\alpha}_i(\theta) = \rho_i(\theta_0, \alpha_{i0}). \tag{34}$$

Using (33) and (34), we find that π_i is first-order bias reducing if and only if:

$$\lim_{N\to\infty} \quad \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\pi_0}\left(\frac{1}{\pi_i(\alpha_{i0})}\frac{\partial}{\partial\alpha_i}\Big|_{\alpha_{i0}} \pi_i(\alpha_i)\rho_i(\theta_0,\alpha_i)\right) = O\left(\frac{1}{T}\right). \tag{35}$$

In the particular case where $\pi_i = \pi_0$ is the population density from which the fixed effects are drawn, there is no bias. To see why, remark that, in this case:

$$\mathbb{E}_{\pi_0}\left(\frac{1}{\pi_i(\alpha_{i0})}\frac{\partial}{\partial\alpha_i}\Big|_{\alpha_{i0}} \pi_i(\alpha_i)\rho_i(\theta_0,\alpha_i)\right) = \int \left(\frac{\partial}{\partial\alpha_i}\Big|_{\alpha_i} \pi_0(\alpha_i)\rho_i(\theta_0,\alpha_i)\right) d\alpha_i. \tag{36}$$

To make the argument formally, we use the following assumption:

Assumption 3

$$\lim_{\alpha_i\to\pm\infty} \quad \pi_0(\alpha_i)\rho_i(\theta_0,\alpha_i) \quad = 0.$$

If Assumption 3 holds, then the right-hand side of (36) is zero. Hence, if π_i is the population density of the individual effects, then the random-effects likelihood has no first-order bias.

[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.

Moreover, under Assumption 3 it can easily be checked that the bias of the score of the random-effects likelihood admits the following alternative expression:

$$B_\infty(\theta_0) \;=\; \operatorname*{plim}_{N\to\infty} \;\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\pi_0}\left(\rho_i(\theta_0,\alpha_{i0})\frac{\partial}{\partial\alpha_i}\Big|_{\alpha_{i0}}\ln\frac{\pi_i(\alpha_i)}{\pi_0(\alpha_i)}\right). \tag{37}$$

Equation (37) suggests that the bias of the random-effects has two sources: (i) the non-orthogonality of the effects, i.e. the presence of the ρ term; and (ii) the distance between the population density of the effects π_0 and the postulated density π_i.

6.3 Random-effects with hyperparameters

We now turn to Random-Effects Maximum Likelihood (REML) estimation. Let ξ be a set of parameters and $\pi_i(\alpha_i;\xi)$ be a family of prior distributions indexed by ξ. As in the previous paragraph, $\pi_i(\alpha_i;\xi)$ does not depend directly on the common parameter θ, nor does it directly depend on the data through the true parameter values. We are interested in the asymptotic properties of the estimator that maximizes the random-effects pseudo-likelihood with respect to θ and ξ. A typical example is when $\pi(\alpha_i;\xi)$ is a normal distribution with unknown mean and variance, $\xi=(m,s^2)$. In another example, the parameters m and s^2 may be functions of covariates and/or initial conditions as in Chamberlain (1984).

To study the bias properties of the REML estimator, it is convenient to start by concentrating the likelihood with respect to ξ. Let:

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

Then the score of the concentrated random-effects likelihood is given by:

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

where the equality comes from the envelope theorem. The bias of the score of the concentrated random-effects likelihood is thus:

$$\begin{aligned}\frac{B_\infty(\theta_0)}{T} &= \operatorname*{plim}_{N\to\infty}\;\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\theta}\Big|_{\theta_0}\ell_i^{RE}(\theta;\widehat{\xi}(\theta_0))\\ &= \lim_{N\to\infty}\;\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\pi_0}\left(\frac{\partial}{\partial\theta}\Big|_{\theta_0}\ell_i^{RE}(\theta;\overline{\xi}(\theta_0))\right),\end{aligned} \tag{38}$$

where:

$$\overline{\xi}(\theta)=\operatorname*{plim}_{N\to\infty}\;\widehat{\xi}(\theta).$$

Equation (38) shows that the first-order bias properties of the random effects likelihood are the same as the ones of an integrated likelihood with prior $\pi_i(\alpha_i; \overline{\xi}(\theta_0))$. The analysis of the previous subsection is thus easily extended to the cases where hyperparameters are present. In particular, using equation (35) we see that REML is bias reducing if and only if

$$\lim_{N\to\infty} \quad \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\pi_0}\left(\frac{1}{\pi_i(\alpha_{i0};\overline{\xi}(\theta_0))}\frac{\partial}{\partial\alpha_i}\Big|_{\alpha_{i0}} \pi_i(\alpha_i;\overline{\xi}(\theta_0))\rho_i(\theta_0,\alpha_i)\right) = O\left(\frac{1}{T}\right). \tag{39}$$

In addition, we show the following result in the Appendix, which helps to interpret the pseudo true value $\overline{\xi}(\theta_0)$.

Lemma 3 *For all θ, we have:*

$$\lim_{N\to\infty} \quad \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\pi_0}\left(\frac{\partial \ln \pi_i(\overline{\alpha}_i(\theta);\overline{\xi}(\theta))}{\partial\xi}\right) = O\left(\frac{1}{T}\right). \tag{40}$$

Proof. See Appendix. ■

Lemma 3 provides a heuristic interpretation of $\overline{\xi}(\theta)$, 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 (40) shows that $\overline{\xi}(\theta_0)$ provides the best approximation, in the Kullback-Leibler sense, to the distribution π_0 on the basis of the family π_i.

We can now state two sufficient conditions for bias reduction:

Proposition 4 *(i) If the common parameters and the individual effects are information orthogonal, then every REML estimator is bias reducing.*

(ii) If π_0 belongs to the parametric family $\pi_i(.;\xi)$, and if Assumption 3 holds, then REML is bias reducing.

Proof. Part (i) comes from the fact that, if parameters are information orthogonal, then $\rho_i(\theta,\alpha) = 0$ for all (θ,α). This implies that (39) is satisfied.

To show part (ii), assume that there exists a ξ_0 such that $\pi_0 = \pi_i(.;\xi_0)$. Under standard identification conditions in parametric models, equation (40) yields that $\overline{\xi}(\theta_0) = \xi_0 + O(1/T)$. The same argument as at the end of the previous subsection follows. ■

The sufficient conditions stated in Proposition 4 are restrictive. In general, REML based on a given parametric family of priors does not reduce bias. We now discuss an important special case and study when Gaussian REML is robust. We prove the following theorem in the Appendix.

Theorem 3 *Gaussian REML reduces first-order bias if and only if there exist $a_{i0}(\theta)$ and $a_{i1}(\theta)$, possibly dependent on exogenous covariates and/or initial conditions, such that:*

$$\rho_i(\theta, \alpha_i) = a_{i0}(\theta) + a_{i1}(\theta)\alpha_i + O\left(\frac{1}{T}\right). \tag{41}$$

Proof. See Appendix. ■

Theorem 3 gives a necessary and sufficient condition for Gaussian REML to reduce bias. The next subsection gives examples of models that satisfy condition (41), such as the dynamic AR(p) model. In these models, the bias of REML based on the Gaussian family is of order $1/T^2$. Still, most models do not satisfy condition (41). In those cases, the bias of the Gaussian REML estimator is of order $1/T$.

6.4 Examples

We turn to reexamine the three examples of Section 4. We first study linear dynamic autoregressive models, and show that the Gaussian REML estimator is first-order bias reducing, irrespective of the form of the individual effects. We also provide a connection to Gaussian random-effects estimation in a linear model with one endogenous regressor and many instruments. Next, in the Poisson counts case, we find that there exists an improper robust prior independent of the common parameters. Moreover, usual RE specifications lead to bias reduction. Lastly, in the static logit case we find that no REML estimator reduces bias. In nonlinear models, thus, the success of random-effects likelihood inference depends critically on prior knowledge about the form of the fixed effects.

Dynamic AR(p). We start with the dynamic AR(p) model of Section 4. We show in Appendix C 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. Hence, it follows from Theorem 3 that Gaussian REML is bias reducing for this model. This result was proven by Cho, Hahn and Kuersteiner (2004) in the case $p = 1$. Moreover, it is easy to check that it still holds if strictly exogenous covariates are included.

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:[15]

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

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 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 Theorem 3, 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.

Poisson counts. For the Poisson counts model, we have:

$$\rho_i(\theta, \alpha_i) = -\alpha_i h(x_i, \theta),$$

where:

$$h(x_i, \theta) = \frac{\sum_{t=1}^{T} \exp(x_{it}'\theta) x_{it}}{\sum_{t=1}^{T} \exp(x_{it}'\theta)}.$$

It follows that Gaussian REML is also bias reducing in this model.

In addition, remark that the local approximation to the robust prior:

$$\widetilde{\pi}(\alpha_i|\theta) = \frac{1}{\alpha_i}$$

is a bias reducing prior that is independent of θ. However, $\widetilde{\pi}$ is an improper prior which does not correspond to a random-effects specification.

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

Assume now that π belongs to the $\Gamma(p,r)$ family, for some $p > 0$, $r > 0$. This family has been widely used to estimate θ by REML in order to correct for overdispersion (see e.g. Gouriéroux *et al.*, 1984). Here we study the bias reduction properties of this specification. We have:

$$\pi(\alpha_i; p, r) = \frac{p^r \alpha_i^{r-1} \exp(-p\alpha_i)}{\Gamma(r)}.$$

It is straightforward to check that the left-hand side in equation (39) is equal to:

$$\lim_{N\to\infty} \quad -\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\pi_0}\left(\overline{r}(\theta_0) - \overline{p}(\theta_0)\alpha_{i0}\right) h(x_i, \theta_0). \tag{42}$$

Now, Lemma 3 implies that:

$$\mathbb{E}_{\pi_0}\left(\frac{\partial \ln \pi(\alpha_{i0}; \overline{p}(\theta_0), \overline{r}(\theta_0))}{\partial p}\right) = O\left(\frac{1}{T}\right).$$

This implies:

$$\frac{1}{\overline{p}(\theta_0)} \mathbb{E}_{\pi_0}\left(\overline{r}(\theta_0) - \overline{p}(\theta_0)\alpha_{i0}\right) = O\left(\frac{1}{T}\right),$$

which in turn implies that (42) is O$(1/T)$. As a consequence, Gamma REML is also bias reducing in the Poisson model.

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, it is very likely that no prior independent of θ will be bias reducing. For example, Theorem 3 shows that Gaussian REML is not robust. This will be the case of virtually all REML estimators of the static logit model.

Note that this lack of unbiasedness is not corrected for by allowing the prior to depend on covariates x_{it}, as in Chamberlain (1984)'s 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.

7 Monte Carlo simulation

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

Static logit model. We first focus on the static logit model:

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

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.

Tables 1 and 2 show 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 Di Ciccio and Stern (1993) adjustment based on equation (4), see Arellano and Hahn (2006a, p.13-14); "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 (23); "robust, observed" refers to the integrated likelihood with the robust prior constructed from observed quantities, see (24), 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 (25); "robust, iterated 1" refers to the same estimator, but when the expectation in (25) 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.[16]

Tables 1 and 2 show 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 (1998)'s integrated likelihood given by equation (23). Note that the infeasible estimator based on (25) 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 the RMLE worsens when the correlation between α_{i0} and x_i increases, and when the sampling distribution of the individual effects

[16] 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.

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.[17] 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 Di Ciccio 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.

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

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

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. Panel a) of 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 (16), 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. Panel b)

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

Table 1: Various estimators of θ in the static logit model, $T = 5$ and $T = 10$

$T = 5$							
	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

Table 2: Various estimators of θ in the static logit model, $T = 20$ and $T = 100$

$T = 20$							
	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
$T = 100$							
	Mean	Median	STD	$\widehat{p}, .05$	$\widehat{p}, .10$	MSE	MAE
uncorrected	1.01	1.01	.0326	.948	.961	.00113	.0275
corrected	.999	.998	.0303	.949	.958	.000910	.0233
uniform	1.02	1.03	.0249	.981	.993	.00119	.0288
Lancaster	1.00	1.01	.0293	.955	.967	.000869	.0234
robust, observed	.989	.988	.0275	.941	.953	.000863	.0237
robust, infeasible	1.00	1.01	.0280	.954	.962	.000789	.0223
robust, iterated 1	.998	1.00	.0282	.949	.961	.000790	.0227
robust, iterated ∞	1.00	1.00	.0283	.953	.964	.000792	.0229
random effects	1.00	1.00	.0278	.953	.975	.000821	.0202
conditional logit	1.00	1.00	.0264	.957	.969	.000764	.0227

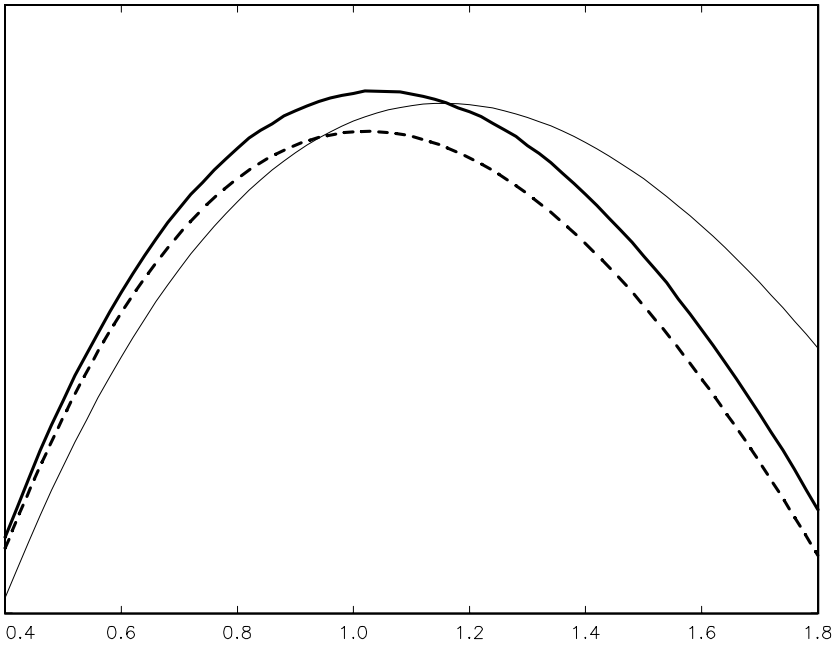
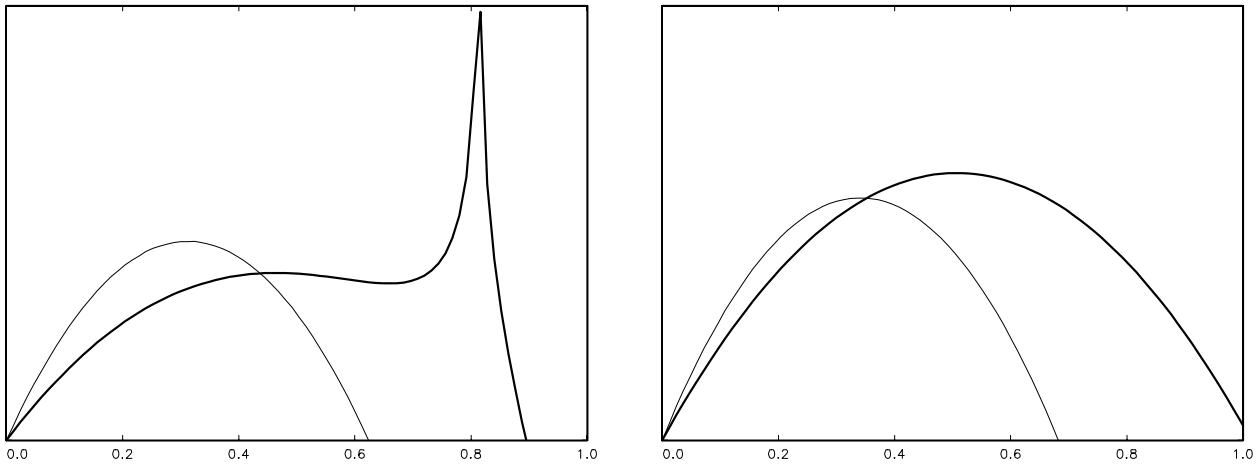


Figure 1: The uncorrected (thin line), bias-corrected (thick line) and integrated (dashed line) likelihoods, one simulation, $T = 10$ (Logit model)



a) Prior based on equation (10)

b) Prior based on equation (12)

Figure 2: The uncorrected (thin line) and integrated (thick line) likelihoods, one simulation, $T = 10$ (AR(1) model)

on the same figure shows the integrated likelihood for the prior (17). 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 (17) to estimate common parameters.

Table 3 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 (19); the "robust" expression of the prior is based on (12) where the outer product is estimated using observed quantities with various degrees of trimming; the "expected" prior is the one given by (17), and plugged-in the "robust, $q = 2$" result to start the iterations in "iterated"; lastly, "GMM" refers to the estimator discussed in Arellano and Bond (1991).

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 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 (12) 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.

We then look at the behavior of random-effects estimators in the dynamic AR(1) model. In this setting, 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

Table 3: Various estimators of μ_1 in the dynamic AR(1) model

$T = 10$							
	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	.468	.470	.0667	.349	.375	.00545	.0583

to dynamic AR(p) models with exogenous covariates. Here we investigate the finite-sample behavior of the two estimators ("correlated" and "independent", respectively) for various values of T. Table 4 shows that, in spite of the theoretical result, the "independent" REML estimator is substantially biased for T as large as 20, compared to its "correlated" counterpart (which is also fixed-T consistent). 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.

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

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

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 (12). Analytical expressions are given in the Appendix. Table 5 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

Table 4: Gaussian random-effect ML estimators of θ in the dynamic AR(1) model

	Mean	Median	STD	$\widehat{p}, .05$	$\widehat{p}, .10$	MSE	MAE
$T = 5$							
independent	.872	.873	.0222	.830	.840	.143	.372
correlated	.620	.639	.0984	.440	.469	.0263	.134
$T = 10$							
independent	.872	.871	.0171	.842	.845	.140	.372
correlated	.519	.506	.0713	.430	.459	.00624	.0497
$T = 20$							
independent	.860	.863	.0248	.814	.823	.130	.360
correlated	.502	.503	.0233	.464	.478	.000399	.0183
$T = 100$							
independent	.500	.500	.00882	.487	.489	.0000771	.00669
correlated	.501	.502	.0101	.485	.488	.0000865	.00828

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, must in general satisfy two conditions. First, they have to depend on the data, in the case where orthogonal reparameterizations do not exist.

Table 5: Various estimators of (μ_1, μ_2) in the dynamic AR(2) model

$T = 10$				
	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	.452	.00680	-.0285	.00304

Second, they must 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, that 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 show that asymptotically valid confidence intervals can be read from the quantiles of the pseudo-posterior distribution.

Our Monte Carlo evidence suggests rather good finite sample properties. 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 could be a natural field of application.

APPENDIX

A Proofs

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

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

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:

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

It thus follows that:

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

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:

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

Likewise, using Assumption 2 we obtain:

$$\mathbb{E}_{\theta_0,\alpha_{i0}}\ln\pi_i(\widehat{\alpha}_i(\theta)|\theta) = \ln\pi_i(\overline{\alpha}_i(\theta)|\theta) + O\left(\frac{1}{T}\right). \qquad (3)$$

Taking expectations in (1) and combining the result with (2) and (3) yields:

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

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

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

We first need the following Lemma:

Lemma 4

$$\frac{\partial}{\partial\theta}\Big|_{\theta_0} \overline{\alpha}_i(\theta) = \left\{\mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\alpha_i}(\theta_0,\alpha_{i0})\right]\right\}^{-1} \mathbb{E}_{\theta_0,\alpha_{i0}}\left[v_i^{\theta}(\theta_0,\alpha_{i0})\right] \equiv \rho_i(\theta_0,\alpha_{i0}). \tag{4}$$

Proof. Straightforward, by differentiating the moment condition solved by $\overline{\alpha}_i(\theta)$ with respect to θ:

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

■

We also need the information matrix equality at true values:

$$\mathbb{E}_{\theta_0,\alpha_{i0}}\left[-v_i^{\alpha_i}(\theta_0,\alpha_{i0})\right] = T\mathbb{E}_{\theta_0,\alpha_{i0}}\left[v_i^2(\theta_0,\alpha_{i0})\right]. \tag{5}$$

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

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

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

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

and the proposition is proved.

Proof of Proposition 2. Assume that π_i is bias reducing. Then Theorem 2 implies:

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

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:

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

Hence one implication. The other implication follows by symmetry.

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:

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

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:

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

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

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

Equation (13) 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:

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

Then the proof of Lemma 1 shows that the only quantity that matters in 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.

Proof of Lemma 2. We have, for all θ:

$$\ell_i^c(\theta) - \overline{\ell}_i(\theta) \;=\; \mathbb{E}_{\theta_0,\alpha_{i0}}\left(\ell_i^c(\theta) - \overline{\ell}_i(\theta)\right) + \widehat{A}_i(\theta),$$

where $\widehat{A}_i(\theta) = O_p(1/\sqrt{T})$, and $\mathbb{E}_{\theta_0,\alpha_{i0}}\left(\widehat{A}_i(\theta)\right) = 0$. Hence:

$$\begin{aligned}
\ell_i^I(\theta) - \overline{\ell}_i(\theta) &= \frac{1}{2T}\ln\left(\frac{2\pi}{T}\right) - \frac{1}{2T}\ln\left(-v_i^{\alpha_i}(\theta, \widehat{\alpha}_i(\theta))\right) + \frac{1}{T}\ln \pi_i(\widehat{\alpha}_i(\theta)|\theta) \\
&\quad + \mathbb{E}_{\theta_0,\alpha_{i0}}\left(\ell_i^c(\theta) - \overline{\ell}_i(\theta)\right) + \widehat{A}_i(\theta) + O_p\left(\frac{1}{T^2}\right), \\
&= C^{st} + \widehat{A}_i(\theta) + \frac{\widehat{B}_i(\theta)}{T} + O_p\left(\frac{1}{T^2}\right),
\end{aligned}$$

where $\mathbb{E}_{\theta_0,\alpha_{i0}}[\widehat{B}_i(\theta)/T] = B_i(\theta)/T$ is the bias of the integrated likelihood, given by (5).

As $\mathbb{E}_{\theta_0,\alpha_{i0}}(\widehat{A}_i(\theta)) = 0$ for all θ, we have:

$$\mathbb{E}_{\theta_0,\alpha_{i0}}\left(\widehat{A}_i(\theta_0)\right) = 0; \quad \mathbb{E}_{\theta_0,\alpha_{i0}}\left(\frac{\partial}{\partial\theta}\Big|_{\theta_0}\widehat{A}_i(\theta)\right) = 0; \quad \mathbb{E}_{\theta_0,\alpha_{i0}}\left(\frac{\partial^2}{\partial\theta\partial\theta'}\Big|_{\theta_0}\widehat{A}_i(\theta)\right) = 0.$$

Moreover it follows from (26) that:

$$\mathbb{E}_{\theta_0,\alpha_{i0}}\left(\frac{\partial}{\partial\theta}\Big|_{\theta_0}\widehat{B}_i(\theta)\right) = 0; \quad \mathbb{E}_{\theta_0,\alpha_{i0}}\left(\frac{\partial^2}{\partial\theta\partial\theta'}\Big|_{\theta_0}\widehat{B}_i(\theta)\right) = 0.$$

The lemma immediately follows:

$$\begin{aligned}
\mathbb{E}_{\theta_0,\alpha_{i0}}\left(\frac{\partial^2 \ell_i^I(\theta_0)}{\partial\theta\partial\theta'}\right) &= \mathbb{E}_{\theta_0,\alpha_{i0}}\left(\frac{\partial^2 \overline{\ell}_i(\theta_0)}{\partial\theta\partial\theta'}\right) + O\left(\frac{1}{T^2}\right), \\
\mathbb{E}_{\theta_0,\alpha_{i0}}\left(\frac{\partial \ell_i^I(\theta_0)}{\partial\theta}\frac{\partial \ell_i^I(\theta_0)}{\partial\theta'}\right) &= \mathbb{E}_{\theta_0,\alpha_{i0}}\left(\frac{\partial \overline{\ell}_i(\theta_0)}{\partial\theta}\frac{\partial \overline{\ell}_i(\theta_0)}{\partial\theta'}\right) + \mathbb{E}_{\theta_0,\alpha_{i0}}\left(\frac{\partial \overline{\ell}_i(\theta_0)}{\partial\theta}\frac{\partial \widehat{A}_i(\theta_0)}{\partial\theta'}\right) \\
&\quad + \mathbb{E}_{\theta_0,\alpha_{i0}}\left(\frac{\partial \widehat{A}_i(\theta_0)}{\partial\theta}\frac{\partial \overline{\ell}_i(\theta_0)}{\partial\theta'}\right) + \mathbb{E}_{\theta_0,\alpha_{i0}}\left(\frac{\partial \widehat{A}_i(\theta_0)}{\partial\theta}\frac{\partial \widehat{A}_i(\theta_0)}{\partial\theta'}\right) + O\left(\frac{1}{T^2}\right).
\end{aligned}$$

Note that

$$\Xi_{iT} \equiv \mathbb{E}_{\theta_0,\alpha_{i0}}\left(\frac{\partial \overline{\ell}_i(\theta_0)}{\partial\theta}\frac{\partial \widehat{A}_i(\theta_0)}{\partial\theta'}\right) + \mathbb{E}_{\theta_0,\alpha_{i0}}\left(\frac{\partial \widehat{A}_i(\theta_0)}{\partial\theta}\frac{\partial \overline{\ell}_i(\theta_0)}{\partial\theta'}\right) + \mathbb{E}_{\theta_0,\alpha_{i0}}\left(\frac{\partial \widehat{A}_i(\theta_0)}{\partial\theta}\frac{\partial \widehat{A}_i(\theta_0)}{\partial\theta'}\right)$$

need not be zero in general. Note also that, as: $\mathbb{E}_{\theta_0,\alpha_{i0}}\left(\frac{\partial}{\partial\theta}\Big|_{\theta_0}\overline{\ell}_i(\theta)\right) = 0$, all the terms in (28) are $O(1/T)$.

Proof of Lemma 3 The first-order conditions of the maximization imply that:

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

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:

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

where we have denoted by π_ξ the derivative of π with respect to ξ. Then, taking the plim we have:

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

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

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

Proof of Theorem 3 In this proof, we assume away individual covariates. Including them complicates the notation, but the essence of the proof remains the same.

The Gaussian prior satisfies:

$$\ln \pi(\alpha_i; \mu, \sigma^2) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\alpha_i - \mu)^2.$$

Then let $\overline{\mu}(\theta_0) = \underset{N\to\infty}{\text{plim}} \quad \widehat{\mu}(\theta_0)$. We verify easily that:

$$\overline{\mu}(\theta_0) = \mathbb{E}_{\pi_0}(\alpha_{i0}) + O(\frac{1}{T}); \qquad \overline{\sigma}^2(\theta_0) = \mathbb{E}_{\pi_0}(\alpha_{i0}^2) - \{\mathbb{E}_{\pi_0}(\alpha_{i0})\}^2 + O\left(\frac{1}{T}\right).$$

Let us assume that π is bias reducing. As there are no individual covariates, the bias of the integrated score writes:

$$\mathbb{E}_{\pi_0}\left(\frac{\partial}{\partial\alpha_i}\Big|_{\alpha_{i0}} \rho_i(\theta_0, \alpha_i)\right) - \mathbb{E}_{\pi_0}\left(\rho_i(\theta_0, \alpha_{i0})\left(\frac{\alpha_{i0} - \overline{\mu}(\theta_0)}{\overline{\sigma}^2(\theta_0)}\right)\right) = O\left(\frac{1}{T}\right).$$

Assuming that $\rho_i(\theta, \alpha_i)$ is continuous in α_i and θ we can suppose that:

$$\rho_i(\theta, \alpha_i) = \sum_{k=0}^{\infty} a_k(\theta)\alpha_i^k,$$

where the a_k functions possibly depend on covariates x_i. We then have:

$$\sum_{k=0}^{\infty} a_k(\theta_0)\left(k\mathbb{E}_{\pi_0}\left(\alpha_{i0}^{k-1}\right) - \mathbb{E}_{\pi_0}\left(\alpha_{i0}^k\left(\frac{\alpha_{i0} - \overline{\mu}(\theta_0)}{\overline{\sigma}^2(\theta_0)}\right)\right)\right) = O\left(\frac{1}{T}\right).$$

The two first terms in this sum are zero. We thus have:

$$\sum_{k=2}^{\infty} a_k(\theta_0)\left(k\mathbb{E}_{\pi_0}\left(\alpha_{i0}^{k-1}\right) - \mathbb{E}_{\pi_0}\left(\alpha_{i0}^k\left(\frac{\alpha_{i0} - \mathbb{E}_{\pi_0}(\alpha_{i0})}{\mathbb{E}_{\pi_0}(\alpha_{i0}^2) - \mathbb{E}_{\pi_0}(\alpha_{i0})^2}\right)\right)\right) = O\left(\frac{1}{T}\right).$$

Hence:

$$\sum_{k=2}^{\infty} a_k(\theta_0)\left(\mathbb{E}_{\pi_0}\left(\alpha_{i0}^{k+1}\right) - \mathbb{E}_{\pi_0}(\alpha_{i0})\mathbb{E}_{\pi_0}(\alpha_{i0}^k) - k\left(\mathbb{E}_{\pi_0}(\alpha_{i0}^2) - \mathbb{E}_{\pi_0}(\alpha_{i0})^2\right)\mathbb{E}_{\pi_0}\left(\alpha_{i0}^{k-1}\right)\right) = O\left(\frac{1}{T}\right).$$

This equality has to be satisfied for all distribution of fixed effects α_{i0}, hence for each set of moments. Taking a distribution such that $\mathbb{E}_{\pi_0}(\alpha_{i0}) = 0$ and $\mathbb{E}_{\pi_0}(\alpha_{i0}^2) - \mathbb{E}_{\pi_0}(\alpha_{i0})^2 = 1$ yields the following simplification:

$$\sum_{k=2}^{\infty} a_k(\theta_0)\left(\mathbb{E}_{\pi_0}\left(\alpha_{i0}^{k+1}\right) - k\mathbb{E}_{\pi_0}\left(\alpha_{i0}^{k-1}\right)\right) = O\left(\frac{1}{T}\right).$$

Then it can be argued, by induction, that $a_k(\theta_0) = O(1/T)$ for all $k \geq 2$. Hence one implication. The other implication is straightforward.

B Existence of non-data dependent bias-reducing priors

In this section of the Appendix, we show that the existence of orthogonal reparameterizations in the sense of equation (9) is not guaranteed in general. To proceed, remark that, by Proposition 1, a fixed prior is bias reducing if and only if the following equation holds:

$$\rho_i(\theta_0, \alpha_{i0}) \frac{\partial \ln \pi_i(\alpha_{i0}|\theta_0)}{\partial \alpha} + \frac{\partial \ln \pi_i(\alpha_{i0}|\theta_0)}{\partial \theta} + \frac{\partial}{\partial \alpha_i}\Big|_{\alpha_{i0}} \rho_i(\theta_0, \alpha_i) = O\left(\frac{1}{T}\right).$$

Up to a term in $O(1/T)$, this is a first-order partial differential equation of the form:

$$\frac{\partial g}{\partial \alpha} f + \frac{\partial g}{\partial \theta} + h = 0, \tag{6}$$

where f and h are known vector functions, and g is an unknown scalar function.

One can solve for g in (6) equation-by-equation. Let

$$\frac{\partial g}{\partial \alpha} f_k + \frac{\partial g}{\partial \theta_k} + h_k$$

be the kth component of the Left-Hand Side in (6). Let also $\alpha = \alpha^{(k)}(\psi^{(k)}, \theta)$ be a reparameterization such that $\partial \alpha^{(k)} / \partial \theta_k = f_k$. We suppose that we have chosen one possible reparameterization among the possible ones, and we denote also $\psi^{(k)}$ the inverse transformation of $\alpha^{(k)}$. Lastly, let $g^{(k)}(\psi^{(k)}, \theta) = g(\alpha^{(k)}(\psi^{(k)}, \theta), \theta)$. One has $\partial g^{(k)} / \partial \theta_k + h_k = 0$, which can be solved as:

$$g^{(k)}(\psi^{(k)}, \theta) = -\int_{-\infty}^{\theta_k} h_k(\alpha^{(k)}(\psi^{(k)}, \widetilde{\theta}_k, \theta_{-k}), \widetilde{\theta}_k, \theta_{-k}) d\widetilde{\theta}_k + \varphi^{(k)}(\psi^{(k)}, \theta_{-k}). \tag{7}$$

In this equation, θ_{-k} denotes vector θ without its kth component, and $\varphi^{(k)}$ is an arbitrary function of $\psi^{(k)}$ and θ_{-k}.

Now, $g^{(k)}(\psi^{(k)}, \theta) = g(\alpha^{(k)}(\psi^{(k)}, \theta), \theta)$ for all k. Equation (7), for all k, thus defines a set of restrictions that g has to satisfy simultaneously.

These restrictions are generally incompatible, as the following argument shows. Let us take $k \neq k'$. Then:

$$\begin{aligned} \varphi^{(k)}(\psi^{(k)}(\alpha, \theta), \theta_{-k}) - \varphi^{(k')}(\psi^{(k')}(\alpha, \theta), \theta_{-k'}) \quad = \quad & \int_{-\infty}^{\theta_k} h_k(\alpha^{(k)}(\psi^{(k)}(\alpha, \theta), \widetilde{\theta}_k, \theta_{-k}), \widetilde{\theta}_k, \theta_{-k}) d\widetilde{\theta}_k \\ & - \int_{-\infty}^{\theta_{k'}} h_{k'}(\alpha^{(k')}(\psi^{(k')}(\alpha, \theta), \widetilde{\theta}_{k'}, \theta_{-k'}), \widetilde{\theta}_{k'}, \theta_{-k'}) d\widetilde{\theta}_{k'} \end{aligned} \tag{8}$$

does not depend on g (that is, on the prior). Now, the left-hand side in (8) generates a (continuous) manifold of dimension at most K is the space $\mathbb{R}^{K+1}$ spanned by (α, θ). Equation (8) thus forms a non trivial set of restrictions. There is no general guarantee that these restrictions are satisfied.

C Derivations for the three examples

For notational simplicity we drop the indices of the expectation terms when they are evaluated at true parameter values.

C.1 Dynamic AR(p)

Let $y_i^0 = (y_{i,1-p}, ..., y_{i0})'$ be the vector of initial conditions, that we assume observed. In matrix form, we have:

$$y_i = X_i\mu_0 + \alpha_{i0}\iota + \varepsilon_i,$$

where the tth row of X_i is $x_{it}' = (y_{i,t-p}, ..., y_{i,t-1})$, $\mu_0 = \left(\mu_{10}...\mu_{p0}\right)'$, and ι is a $T \times 1$ vector of ones. The scaled individual log-likelihood is given by:

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

We thus have:

$$v_i(\mu, \sigma^2, \alpha_i) = \frac{1}{T}\sum_{t=1}^{T}\frac{(y_{it} - x_{it}'\mu - \alpha_i)}{\sigma^2},$$

and hence:

$$\mathbb{E}\left[-v_i^{\alpha_i}(\mu, \sigma^2, \alpha_i)\right] = \frac{1}{\sigma^2}.$$

Moreover:

$$\begin{aligned}\mathbb{E}\left[v_i^2(\mu, \sigma^2, \alpha_i)\right] &= \frac{1}{T^2\sigma^4}\iota'\mathbb{E}\left((y_i - X_i\mu - \alpha_i\iota)(y_i - X_i\mu - \alpha_i\iota)'\right)\iota, \\ &= \frac{1}{T^2\sigma^4}\iota'\mathbb{E}\left((X_i(\mu_0 - \mu) + (\alpha_{i0} - \alpha_i)\iota + \varepsilon_i)(X_i(\mu_0 - \mu) + (\alpha_{i0} - \alpha_i)\iota + \varepsilon_i)'\right)\iota.\end{aligned}$$

Note that this expectation depends on the true values of the parameters. Note also that the expectation is taken for i fixed. The same will be true of the variances and covariances that we will consider in this section of the Appendix.

Computation of $\mathbb{E}\left[v_i^2(\mu, \sigma^2, \alpha_i)\right]$. One has:

$$\mathrm{Var}\left(\varepsilon_i + X_i(\mu_0 - \mu)\right) = \mathrm{Var}\left(\varepsilon_i + \left[(\mu_0 - \mu)' \otimes I_T\right]\mathrm{vec}\, X_i\right).$$

Let $B(\mu_0, \mu) = (\mu_0 - \mu)' \otimes I_T$. Then:

$$\begin{aligned}\mathrm{Var}\left(\varepsilon_i + X_i(\mu_0 - \mu)\right) &= \sigma^2 I_T + \mathbb{E}\left(\varepsilon_i\left(\mathrm{vec}\, X_i\right)'\right)B(\mu_0, \mu)' + B(\mu_0, \mu)\mathbb{E}\left(\varepsilon_i\left(\mathrm{vec}\, X_i\right)'\right)' \\ &\quad + B(\mu_0, \mu)\,\mathrm{Var}\left(\mathrm{vec}\, X_i\right)B(\mu_0, \mu)'.\end{aligned}$$

To compute these expressions, we shall write the model as (see Alvarez and Arellano, 2004, appendix A.3):

$$\begin{pmatrix} I_p & 0 \\ B_{Tp} & B_T \end{pmatrix}\begin{pmatrix} y_i^0 \\ y_i \end{pmatrix} = \begin{pmatrix} y_i^0 \\ \alpha_i\iota + \varepsilon_i \end{pmatrix},$$

where

$$\begin{pmatrix} B_{Tp} & B_T \end{pmatrix} = \begin{pmatrix} -\mu_{p0} & -\mu_{(p-1)0} & ... & -\mu_{10} & 1 & 0 & 0 & ... & 0 & 0 \\ 0 & -\mu_{p0} & ... & -\mu_{20} & -\mu_{10} & 1 & 0 & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & ... & 0 & 0 & 0 & 0 & ... & -\mu_{10} & 1 \end{pmatrix}.$$

Inverting the system yields:

$$y_i = \overline{C}_{Tp} y_i^0 + \alpha_i \overline{C}_T \iota + \overline{C}_T \varepsilon_i,$$

where $\overline{C}_T = B_T^{-1}$ and $\overline{C}_{Tp} = -B_T^{-1} B_{Tp}$.

At this stage, it is convenient to introduce the $(T+p) \times (Tp)$ selection matrix such that

$$\text{vec}(X_i) = P' \begin{pmatrix} y_i^0 \\ y_i \end{pmatrix}.$$

Moreover, the matrix $B(\mu_0, \mu)P'$ reads:

$$\begin{pmatrix} \mu_{10} - \mu_1 & \mu_{20} - \mu_2 & ... & \mu_{p0} - \mu_p & 0 & 0 & ... & 0 & 0 \\ 0 & \mu_{10} - \mu_1 & \mu_{20} - \mu_2 & ... & \mu_{p0} - \mu_p & 0 & ... & 0 & 0 \\ 0 & 0 & \mu_{10} - \mu_1 & \mu_{20} - \mu_2 & ... & \mu_{p0} - \mu_p & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... & 0 \\ 0 & 0 & 0 & 0 & ... & 0 & ... & 0 & 0 \\ 0 & 0 & 0 & 0 & ... & 0 & ... & \mu_{p0} - \mu_p & 0 \end{pmatrix}.$$

We shall write:

$$B(\mu_0, \mu)P' = \begin{pmatrix} \overline{A}(\mu_0, \mu) & \overline{B}(\mu_0, \mu) \end{pmatrix},$$

where $\overline{A}(\mu_0, \mu)$ is $T \times p$ and $\overline{B}(\mu_0, \mu)$ is $T \times T$.

Now:

$$\text{vec}(X_i) = P' \begin{pmatrix} y_i^0 \\ y_i \end{pmatrix} = P' \begin{pmatrix} I_p \\ \overline{C}_{Tp} \end{pmatrix} y_i^0 + \alpha_i P' \begin{pmatrix} 0 \\ \overline{C}_T \iota \end{pmatrix} + P' \begin{pmatrix} 0 \\ \overline{C}_T \varepsilon_i \end{pmatrix}. \tag{9}$$

It thus follows that

$$\begin{aligned} \mathbb{E}\left[\varepsilon_i \left(\text{vec}\, X_i\right)'\right] B(\mu_0, \mu)' &= \sigma_0^2 \begin{pmatrix} 0_p & \overline{C}_T' \end{pmatrix} P B(\mu_0, \mu)' \\ &= \sigma_0^2 \overline{C}_T' \overline{B}(\mu_0, \mu)'. \end{aligned}$$

Then:

$$\begin{aligned} B(\mu_0, \mu)\, \text{Var}\left(\text{vec}\, X_i\right) B(\mu_0, \mu)' &= \sigma_0^2 B(\mu_0, \mu) P' \begin{pmatrix} 0 & 0 \\ 0 & \overline{C}_T \overline{C}_T' \end{pmatrix} P B(\mu_0, \mu)' \\ &= \sigma_0^2 \overline{B}(\mu_0, \mu) \overline{C}_T \overline{C}_T' \overline{B}(\mu_0, \mu)'. \end{aligned}$$

Hence:

$$\begin{aligned} \text{Var}\left(\varepsilon_i + X_i(\mu_0 - \mu)\right) &= \sigma_0^2 I_T + \sigma_0^2 \overline{C}_T' \overline{B}(\mu_0, \mu)' + \sigma_0^2 \overline{B}(\mu_0, \mu) \overline{C}_T \\ &\quad + \sigma_0^2 \overline{B}(\mu_0, \mu) \overline{C}_T \overline{C}_T' \overline{B}(\mu_0, \mu)'. \end{aligned}$$

Now:

$$\mathbb{E}\left(\left(X_i(\mu_0-\mu)+(\alpha_{i0}-\alpha_i)\iota+\varepsilon_i\right)\left(X_i(\mu_0-\mu)+(\alpha_{i0}-\alpha_i)\iota+\varepsilon_i\right)'\right)$$
$$=\mathrm{Var}\left(\varepsilon_i+X_i(\mu_0-\mu)\right)+\mathbb{E}\left(X_i(\mu_0-\mu)+(\alpha_{i0}-\alpha_i)\iota+\varepsilon_i\right)\mathbb{E}\left(X_i(\mu_0-\mu)+(\alpha_{i0}-\alpha_i)\iota+\varepsilon_i\right)'.$$

Since:

$$\mathrm{vec}(X_i)=P'\left(\begin{array}{c} I_p \\ \overline{C}_{Tp}\end{array}\right)y_i^0+\alpha_{i0}P'\left(\begin{array}{c} 0 \\ \overline{C}_T\iota\end{array}\right)+P'\left(\begin{array}{c} 0 \\ \overline{C}_T\varepsilon_i\end{array}\right),$$

it follows that

$$\begin{aligned}\mathbb{E}\left[X_i(\mu_0-\mu)\right] &= B(\mu_0,\mu)\mathbb{E}\left[\mathrm{vec}(X_i)\right]\\ &= \left(\overline{A}(\mu_0,\mu)+\overline{B}(\mu_0,\mu)\overline{C}_{Tp}\right)y_i^0+\alpha_{i0}\overline{B}(\mu_0,\mu)\overline{C}_T\iota.\end{aligned}$$

The previous results yield:

$$\begin{aligned}\mathbb{E}\left[v_i^2(\mu,\sigma^2,\alpha_i)\right] &= \frac{1}{T^2\sigma^4}\iota'\Big\{\sigma_0^2I_T+\sigma_0^2\overline{C}_T'\overline{B}(\mu_0,\mu)'+\sigma_0^2\overline{B}(\mu_0,\mu)\overline{C}_T\\ &\quad+\sigma_0^2\overline{B}(\mu_0,\mu)\overline{C}_T\overline{C}_T'\overline{B}(\mu_0,\mu)'\\ &\quad+\left[\left(\overline{A}(\mu_0,\mu)+\overline{B}(\mu_0,\mu)\overline{C}_{Tp}\right)y_i^0+\alpha_{i0}\overline{B}(\mu_0,\mu)\overline{C}_T\iota+(\alpha_{i0}-\alpha_i)\iota\right]\times\\ &\quad\left[\left(\overline{A}(\mu_0,\mu)+\overline{B}(\mu_0,\mu)\overline{C}_{Tp}\right)y_i^0+\alpha_{i0}\overline{B}(\mu_0,\mu)\overline{C}_T\iota+(\alpha_{i0}-\alpha_i)\iota\right]'\Big\}\iota.\end{aligned}$$

The robust prior is thus given by:

$$\begin{aligned}\pi_i^R\left(\alpha_i|\mu,\sigma^2\right) &\propto \Big(\iota'\Big\{\sigma_0^2I_T+\sigma_0^2\overline{C}_T'\overline{B}(\mu_0,\mu)'+\sigma_0^2\overline{B}(\mu_0,\mu)\overline{C}_T\\ &\quad+\sigma_0^2\overline{B}(\mu_0,\mu)\overline{C}_T\overline{C}_T'\overline{B}(\mu_0,\mu)'\\ &\quad+\left[\left(\overline{A}(\mu_0,\mu)+\overline{B}(\mu_0,\mu)\overline{C}_{Tp}\right)y_i^0+\alpha_{i0}\overline{B}(\mu_0,\mu)\overline{C}_T\iota+(\alpha_{i0}-\alpha_i)\iota\right]\times\\ &\quad\left[\left(\overline{A}(\mu_0,\mu)+\overline{B}(\mu_0,\mu)\overline{C}_{Tp}\right)y_i^0+\alpha_{i0}\overline{B}(\mu_0,\mu)\overline{C}_T\iota+(\alpha_{i0}-\alpha_i)\iota\right]'\Big\}\iota\Big)^{-1/2},\\ &\propto \Big(1+a(\mu-\mu_0)+b(\mu-\mu_0,\alpha_i-\alpha_{i0})\Big)^{-1/2},\end{aligned}$$

where

$$a(\mu-\mu_0)=\frac{1}{T}\iota'\Big\{\overline{C}_T'\overline{B}(\mu_0,\mu)'+\overline{B}(\mu_0,\mu)\overline{C}_T\Big\}\iota \tag{10}$$

is a linear function of $\mu-\mu_0$, and

$$\begin{aligned}b(\mu-\mu_0,\alpha_i-\alpha_{i0}) &= \frac{1}{T\sigma_0^2}\iota'\Big\{\sigma_0^2\overline{B}(\mu_0,\mu)\overline{C}_T\overline{C}_T'\overline{B}(\mu_0,\mu)'\\ &\quad+\left[\left(\overline{A}(\mu_0,\mu)+\overline{B}(\mu_0,\mu)\overline{C}_{Tp}\right)y_i^0+\alpha_{i0}\overline{B}(\mu_0,\mu)\overline{C}_T\iota+(\alpha_{i0}-\alpha_i)\iota\right]\times\\ &\quad\left[\left(\overline{A}(\mu_0,\mu)+\overline{B}(\mu_0,\mu)\overline{C}_{Tp}\right)y_i^0+\alpha_{i0}\overline{B}(\mu_0,\mu)\overline{C}_T\iota+(\alpha_{i0}-\alpha_i)\iota\right]'\Big\}\iota\end{aligned} \tag{11}$$

is a quadratic function of $\mu-\mu_0$ and $\alpha_i-\alpha_{i0}$.

The AR(1) case. Let us assume that $p = 1$. Then:

$$\overline{C}_T = \begin{pmatrix} 1 & 0 & ... & 0 \\ \mu_{10} & 1 & ... & 0 \\ ... & ... & ... & ... \\ \mu_{10}^{T-1} & \mu_{10}^{T-2} & ... & 1 \end{pmatrix},$$

so that:

$$\overline{C}_{T}\iota = \frac{1}{1-\mu_1^0} \begin{pmatrix} 1-\mu_{10} \\ 1-\mu_{10}^2 \\ ... \\ 1-\mu_{10}^T \end{pmatrix}.$$

Moreover:

$$\overline{C}_{Tp} = \begin{pmatrix} \mu_{10} \\ \mu_{10}^2 \\ ... \\ \mu_{10}^T \end{pmatrix},$$

and

$$\overline{A}(\mu_0, \mu) = \begin{pmatrix} \mu_{10} - \mu_1 \\ 0 \\ ... \\ 0 \end{pmatrix}, \qquad \overline{B}(\mu_0, \mu) = \begin{pmatrix} 0 & 0 & ... & 0 & 0 \\ \mu_{10} - \mu_1 & 0 & ... & 0 & 0 \\ 0 & \mu_{10} - \mu_1 & ... & 0 & 0 \\ 0 & 0 & ... & \mu_{10} - \mu_1 & 0 \end{pmatrix}.$$

Hence $\pi_i^R\left(\mu_i|\mu, \sigma^2\right)$ is proportional to

$$\begin{aligned} &\left\{\sigma_0^2 T + 2\sigma_0^2 \frac{\mu_{10} - \mu_1}{1-\mu_{10}} \cdot \sum_{t=1}^{T-1}\left(1-\mu_{10}^t\right) + \sigma_0^2 \left(\frac{\mu_{10} - \mu_1}{1-\mu_{10}}\right)^2 \cdot \sum_{t=1}^{T-1}\left(1-\mu_{10}^t\right)^2 + \right. \\ &\left[\left((\mu_{10} - \mu_1)\frac{1-\mu_{10}^T}{1-\mu_{10}}\right) y_i^0 + \alpha_{i0}\frac{\mu_{10} - \mu_1}{1-\mu_{10}} \cdot \sum_{t=1}^{T-1}\left(1-\mu_{10}^t\right) + (\alpha_{i0} - \alpha_i)T\right] \times \\ &\left.\left[\left((\mu_{10} - \mu_1)\frac{1-\mu_{10}^T}{1-\mu_{10}}\right) y_i^0 + \alpha_{i0}\frac{\mu_{10} - \mu_1}{1-\mu_{10}} \cdot \sum_{t=1}^{T-1}\left(1-\mu_{10}^t\right) + (\alpha_{i0} - \alpha_i)T\right]'\right\}^{-1/2}. \end{aligned}$$

We thus obtain:

$$\begin{aligned} \pi_i^R\left(\overline{\alpha}_i(\mu, \sigma^2)|\mu, \sigma^2\right) \quad \propto \quad &\left\{T + 2\frac{\mu_{10} - \mu_1}{1-\mu_{10}} \cdot \sum_{t=1}^{T-1}\left(1-\mu_{10}^t\right)\right. \\ &\left. + \left(\frac{\mu_{10} - \mu_1}{1-\mu_{10}}\right)^2 \cdot \sum_{t=1}^{T-1}\left(1-\mu_{10}^t\right)^2\right\}^{-1/2}. \end{aligned}$$

Hence, for π to reduce bias we need that:

$$\frac{\partial \ln \pi\left(\overline{\alpha}_i(\mu, \sigma^2)|\mu, \sigma^2\right)}{\partial \mu}\Big|_{\mu_{10}, \sigma_0^2, \alpha_{i0}} = \frac{1}{T(1-\mu_{10})} \cdot \sum_{t=1}^{T-1}\left(1-\mu_{10}^t\right) = \frac{1}{T}\sum_{t=1}^{T-1}(T-t)\mu_{10}^{t-1}.$$

Gaussian REML. We have:

$$v_i(\mu, \sigma^2, \alpha_i) = \frac{1}{T}\sum_{t=1}^{T} \frac{(y_{it} - x_{it}'\mu - \alpha_i)}{\sigma^2},$$

and hence:

$$\mathbb{E}\left(-v_i^{\alpha_i}(\mu, \sigma^2, \alpha_i)\right) = \frac{1}{\sigma^2}; \qquad \mathbb{E}\left(-v_i^{\mu}(\mu, \sigma^2, \alpha_i)\right) = -\frac{1}{T\sigma^2}\sum_{t=1}^{T} x_{it}; \qquad \mathbb{E}\left(-v_i^{\sigma^2}(\mu, \sigma^2, \alpha_i)\right) = 0.$$

Dropping for simplicity the derivative with respect to σ^2 we obtain:

$$\rho_i(\mu, \alpha_i) = -\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}(x_{it}).$$

Let us define the following $p \times (T + p)$ matrix:

$$Q = \begin{pmatrix} I_p & ... & I_p \end{pmatrix} P'.$$

Then as

$$\sum_{t=1}^{T} x_{it} = \begin{pmatrix} I_p & ... & I_p \end{pmatrix} \mathrm{vec}(X_i),$$

we obtain, using (9):

$$\rho_i(\mu, \alpha_i) = -\frac{1}{T}\left(Q\begin{pmatrix} I_p \\ \overline{C}_{Tp} \end{pmatrix} y_i^0 + \alpha_i Q \begin{pmatrix} 0 \\ \overline{C}_{T\iota} \end{pmatrix}\right),$$

where $\overline{C}_{Tp}$ and $\overline{C}_T$ are functions of μ.

C.2 Linear model with one endogenous regressor

The individual log-likelihood is given by (see e.g. Hahn, 2000):

$$\ell_i(\theta, \alpha_i) = -\frac{1}{2}\ln|\Omega| - \frac{1}{2T}\omega_{11}\sum_{t=1}^{T}(y_{it} - \theta\alpha_i)^2 - \frac{1}{T}\omega_{12}\sum_{t=1}^{T}(y_{it} - \theta\alpha_i)(x_{it} - \alpha_i) - \frac{1}{2T}\omega_{22}\sum_{t=1}^{T}(x_{it} - \alpha_i)^2.$$

We thus have:

$$v_i(\theta, \alpha_i) = \frac{1}{T}\omega_{11}\theta\sum_{t=1}^{T}(y_{it} - \theta\alpha_i) + \frac{1}{T}\omega_{12}\sum_{t=1}^{T}(y_{it} - 2\theta\alpha_i + \theta x_{it}) + \frac{1}{T}\omega_{22}\sum_{t=1}^{T}(x_{it} - \alpha_i).$$

Then:

$$\mathbb{E}\left(-v_i^{\alpha_i}(\theta, \alpha_i)\right) = \omega_{11}\theta^2 + 2\omega_{12}\theta + \omega_{22},$$

and:

$$v_i^{\theta}(\theta, \alpha_i) = \frac{1}{T}\omega_{11}\sum_{t=1}^{T}(y_{it} - 2\theta\alpha_i) + \frac{1}{T}\omega_{12}\sum_{t=1}^{T}(-2\alpha_i + x_{it}).$$

Hence, at true values:

$$\mathbb{E}_{\theta_0, \alpha_{i0}}\left(v_i^{\theta}(\theta_0, \alpha_{i0})\right) = -\omega_{11}\theta_0\alpha_{i0} - \omega_{12}\alpha_{i0}.$$

We obtain that:

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

C.3 Poisson counts

We have:

$$v_i(\theta, \alpha_i) = \frac{1}{T\alpha_i} \sum_{t=1}^{T} \left(y_{it} - \alpha_i \exp(x_{it}'\theta) \right).$$

Note that it follows that:

$$\overline{\alpha}_i(\theta) = \alpha_{i0} \frac{\sum_{t=1}^{T} \exp(x_{it}'\theta_0)}{\sum_{t=1}^{T} \exp(x_{it}'\theta)}. \tag{12}$$

Moreover:

$$\mathbb{E}\left(-v_i^{\alpha_i}(\theta, \alpha_i)\right) = \frac{1}{T\alpha_i^2} \sum_{t=1}^{T} \alpha_{i0} \exp(x_{it}'\theta_0),$$

and:

$$\begin{aligned}
\mathbb{E}\left(v_i^2(\theta, \alpha_i)\right) &= \frac{1}{T^2\alpha_i^2} \sum_{t=1}^{T} \mathbb{E}\left((y_{it} - \alpha_i \exp(x_{it}'\theta))^2 \right), \\
&= \frac{1}{T^2\alpha_i^2} \sum_{t=1}^{T} \left(\mathbb{E}\left((y_{it} - \mathbb{E}(y_{it}))^2 \right) + \left(\mathbb{E}(y_{it}) - \alpha_i \exp(x_{it}'\theta) \right)^2 \right), \\
&= \frac{1}{T^2\alpha_i^2} \sum_{t=1}^{T} \alpha_{i0} \exp(x_{it}'\theta_0) + \left(\alpha_{i0} \exp(x_{it}'\theta_0) - \alpha_i \exp(x_{it}'\theta) \right)^2,
\end{aligned}$$

where we have used that $\mathrm{Var}(y_{it}) = \mathbb{E}(y_{it}) = \alpha_{i0} \exp(x_{it}'\theta_0)$. Hence:

$$\pi_i^R(\alpha_i|\theta) \propto \frac{1}{\alpha_i} \left(\sum_{t=1}^{T} \alpha_{i0} \exp(x_{it}'\theta_0) + \left(\alpha_{i0} \exp(x_{it}'\theta_0) - \alpha_i \exp(x_{it}'\theta) \right)^2 \right)^{-1/2}.$$

C.4 Static logit

We have:

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

It follows that:

$$\mathbb{E}\left[-v_i^{\alpha_i}(\theta, \alpha_i)\right] = \frac{1}{T} \sum_{t=1}^{T} \Lambda(x_{it}'\theta + \alpha_i)(1 - \Lambda(x_{it}'\theta + \alpha_i)), \tag{13}$$

and:

$$\begin{aligned}
\mathbb{E}\left[v_i^2(\theta, \alpha_i)\right] &= \mathbb{E}\left(\frac{1}{T} \sum_{t=1}^{T} \left(y_{it} - \Lambda(x_{it}'\theta + \alpha_i) \right) \right)^2 \\
&= \frac{1}{T^2} \sum_{t=1}^{T} \mathbb{E}\left(\left(y_{it} - \Lambda(x_{it}'\theta + \alpha_i) \right)^2 \right), \qquad (14)
\end{aligned}$$

where we have used the fact that observations are i.i.d. across T.

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 (2006a): "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*, Cambridge University Press, forthcoming.

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

[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. (2006): "Estimating Dynamic Panel Data Discrete Choice Models with Fixed Effects", *Journal of Econometrics*, forthcoming.

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

[17] 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.

[18] 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.

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

[20] DiCiccio, T. J., M. A. Martin, S. E. Stern, and G. A. Young (1996): "Information Bias and Adjusted Profile Likelihoods", *Journal of the Royal Statistical Society*, Series B, 58, 189–203.

[21] Gourieroux, C., A. Montfort and A. Trognon (1984). "Pseudo Maximum Likelihood Methods: Applications to. Poisson Models", *Econometrica*, 52, 701–720.

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

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

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

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

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

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

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

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

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

[31] Lancaster, T. (2004): *An Introduction to Modern Bayesian Econometrics*, Blackwell.

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

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

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

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

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

[37] 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.

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

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