ROBUSTNESS AGAINST INCIDENTAL PARAMETERS AND MIXING DISTRIBUTIONS*

By Tiemen Woutersen†

Abstract

Neyman and Scott (1948) define the incidental parameter problem. In panel data with T observations per individual and unobservable individual-specific effects, the inconsistency of the maximum likelihood estimator of the common parameters is in general $O(T^{-1})$. This paper considers the integrated likelihood estimator and develops the integrated moment estimator. It shows that the inconsistency of the integrated likelihood estimator reduces from $O(T^{-1})$ to $O(T^{-2})$ if an information orthogonal parametrization is used. It derives information orthogonal moment functions for the general linear model and the index model with weakly exogenous regressors and thereby offers an approximate solution for the incidental parameter problem for a wide range of models. It argues that reparametrizations are easier in a Bayesian framework and shows how to use the $O(T^{-2})$-result to increase the robustness against the choice of mixing distribution. The integrated likelihood estimator is consistent and adaptive for asymptotics in which $T \propto N^{\alpha}$ where $\alpha > \frac{1}{3}$. The paper also shows that likelihood methods that use sufficient statistics for the individual-specific effects can be viewed as a special case of the integrated likelihood estimator.

Keywords: Incidental parameters, predetermined variables, panel data

*I gratefully acknowlegde stimulating suggestions from Tony Lancaster, Bo Honoré, Wolfgang Polasek, Nancy Reid, Geert Ridder, and seminar participants at the Departments of Economics at Stanford University, University College London, and University of Western Ontario. The CIBC Human Capital and Productivity Program provided financial support. All errors are mine.

†Mailing Address: University of Western Ontario, Department of Economics, Social Science Center, London, Ontario, N6A 5C2, Canada. Email: twouters@uwo.ca.

1 Introduction

One way to control for the heterogeneity in panel data is to allow for time-invariant, individual specific parameters. This fixed effect approach introduces many parameters into the model which causes the 'incidental parameter problem' of Neyman and Scott (1948): the maximum likelihood estimator is in general inconsistent. It follows from Liang (1987) that this inconsistency is $O(T^{-1})$ where T is the number of periods for which we observe an individual.

Cox and Reid (1987) propose using a parametrization of the likelihood that is information-orthogonal and then applying their conditional profile likelihood method. Lancaster (1997 and 2000) applies this orthogonality idea to panel data and develops an 'integrated likelihood estimator'. However, Lancaster does not present general results. This paper shows that the inconsistency or bias of the integrated likelihood estimator is in general $O(T^{-1})$ and that information-orthogonality of the likelihood reduces this inconsistency to $O(T^{-2})$. We derive this result using a Laplace approximation for the ratio of integrals, as derived by Kass, Tierney, and Kadane (1990). We show how to attain an information-orthogonal parametrization of the likelihood for the general nonlinear model and for index models with lagged dependent and exogenous variables. We develop the integrated moment estimator for models with general predetermined or weakly exogenous regressors and fixed effects. Moreover, we extend the integrated likelihood approach by allowing for implicitly defined likelihoods. We thereby solve the incidental parameter problem up to $O(T^{-2})$ for a wide range of models, including the dynamic linear, logit and probit model with fixed effects and predetermined variables.

Cox and Reid (1987 and 1993) only consider models in which the likelihood can be written as an analytical function of the parameters of interest and the information-orthogonal nuisance parameters. This requires that a particular differential equation can be solved analytically. The integrated likelihood does not require an analytical solution which implies that the integrated likelihood can be used much more generally than the conditional profile likelihood method. This is an argument in favor of Bayesian analysis. Berger et al. (1999) give an overview of integrated likelihoods methods but only mention 'generality' and 'simplicity'

and 'accounts for parameter uncertainty in special cases' as advantages over profile likelihood methods. We argue that the argument of implicit versus explicit reparametrization is a more tangible advantage of integrated likelihoods. We show the information-orthogonal parametrizations of the models of the handbook chapters covering panel data by Chamberlain (1984) and Honoré and Arellano (2001).

Alvarez and Arellano (1998) develop an alternative asymptotic where T increases at the same rate as the number of individuals, N. We show that the integrated likelihood estimator is asymptotically unbiased under this alternative asymptotic. Given that T is smaller than N in most panel data, we argue that it might be more interesting to let T increase at a slower rate than N; we also show that the integrated likelihood estimator is asymptotically unbiased and adaptive as long as $T \propto N^{\alpha}$ where $\alpha > \frac{1}{3}$. This means that the asymptotic variance of the integrated likelihood estimator equals the asymptotic variance of the infeasible maximum likelihood estimator that assumes the values of the nuisance parameters to be known. This implies efficiency of the integrated likelihood if $T \propto N^{\alpha}$ where $\alpha > \frac{1}{3}$.

We also consider fixed T asymptotics and show how the choice of a prior may induce the mode of the posterior to be a consistent estimator for $N \to \infty$. We call such a prior a frequentist prior and derive sufficient conditions for its existence. An interesting application of the frequentist prior is the dynamic linear model with fixed effects. Blundell et al. (2000) and Alvarez and Arellano (1998) give recent reviews of moment estimators for this model. The integrated likelihood estimator is consistent for fixed T, adaptive under stationairity and $T \propto N^{\alpha}$ where $\alpha > \frac{1}{3}$, and superconsistent for the non-stationary case.

For a couple of models, a sufficient statistic for the fixed effect has been found. These models are the panel version of the logit, the Poisson, the Weibull models and the linear model with known variance, see Chamberlain (1984 and 1985) for an overview. We show that these likelihood methods that use a sufficient statistic for the fixed effect can be viewed as a special case of the integrated likelihood estimator. Honoré and Kyriazidou (2000) develop an estimator for the dynamic binary model that requires a continuous distribution of all exogenous regressors. They use semiparametric matching in a space with dimension K^L

where K is the number of exogenous regressors and L the number of lags of y_{it}. Semiparametric matching decreases the rate of convergence of the estimator and Honoré and Kyriazidou do not consider cases where $L > 2$. The integrated likelihood estimator with an explicit parametrization reproduces the estimator by Honoré and Kyriazidou. If we use an implicit parametrization, however, we do not need matching so that the rate of convergence increases and adaptiveness can be achieved if $T \propto N^{\alpha}$ where $\alpha > \frac{1}{3}$.

From a formal point of view, integrating out fixed effects is equivalent to a 'random effects' model with a prior distribution playing the role of a mixing distribution. By specifying the mixing distribution over an information-orthogonal parametrization, the $O(T^{-2})$ and adaptiveness results apply to these models as well. We thus generalize Mundlak's (1978) linear random effects model to nonlinear models.

The argument against fixed effect models is usually that the set of models that can be estimated is so small. The reason for aversion against random effects models is usually based on the sensitivity to the choice of mixing distribution, see Nerlove (2000) and Trognon (2000) for a recent exposition of these arguments. This paper gives an approximate solution for the incidental parameter problem and thereby allows a fixed effect estimation for a much wider class of models. Moreover, the same algebra can be used to increase the robustness of mixing distributions by changing the interpretation of the prior.

This paper is organized as follows. Section 2 shows that the inconsistency of the integrating likelihood estimator is, in general, $O(T^{-1})$. Section 3 shows that information-orthogonality reduces this inconsistency to $O(T^{-2})$ and gives the asymptotics for which the integrated likelihood estimator is adaptive. Section 4 deals with random effects models and shows that specifying the random distribution over information-orthogonal parameters gives the same adaptiveness and $O(T^{-2})$-result. Section 5 gives examples. Section 6 shows that 'differencing out' and 'conditioning on a sufficient statistic' are special cases of the integrating out approach with an information-orthogonal parametrization. Section 7 derives the integrated moment estimator for nonlinear models with predetermined variables and section 8 concludes.

2 An expression for the inconsistency

Suppose we observe N individuals for T periods. Let the log likelihood contribution of the t^{th} spell of individual i be denoted by L^{it}. Summing over the contributions of individual i yields the log likelihood contribution,

$$L^i(\beta, \lambda_i) = \sum_t L^{it}(\beta, \lambda_i),$$

where β is the common parameter and λ_i is the individual specific effect. Suppose that the parameter β is of interest and that the fixed effect λ_i is a nuisance parameter that controls for heterogeneity. This paper considers elimination of nuisance parameters by integration. This Bayesian treatment of nuisance parameters is straightforward: Formulate a prior on all the nuisance parameters and then integrate the likelihood with respect to that prior distribution of the nuisance parameters, see Gelman et al. (1995) for an overview. For a panel data model with fixed effects this means that we have to specify priors on the common parameters and all the fixed effects. Chamberlain (1984) describes the problem of this method for panel data:

> "In a Bayesian framework, β and λ_i would be treated symmetrically, with a prior distribution for both. Since I [Chamberlain] am only going to use asymptotic results on inference, however, a "gentle" prior distribution for β will be dominated. That this need not be true for λ_i is one of the interesting aspects of our problem".

Later in his handbook chapter, Chamberlain gives an example of an estimator for which the inconsistency is $O(T^{-1})$. One could try to find a prior that ensures the mode of the posterior to be a consistent estimator for β. In particular, one could structure this search by interpreting the prior as a Jacobian. Finding a favorable prior is then equivalent to finding a particular parametrization where the reparametrization generates the Jacobian or prior. We discuss the integrated likelihood in this section and explore favorable reparametrizations thereafter. Berger et al. (1999) review integrated likelihood methods in which flat priors are used for both the parameter of interest and the nuisance parameters. The individual

specific nuisance parameters are then eliminated by integration. We denote the logarithm of the integrated likelihood contribution by $L^{i,I}(\beta)$, i.e.

$$L^{i,I}(\beta) = \ln \int e^{L^i(\beta,\lambda)} d\lambda_i.$$

Summing over i yields the logarithm of the integrated likelihood,

$$L^I(\beta) = \sum_i L^{i,I}(\beta) = \sum_i \ln \int e^{L^i(\beta,\lambda)} d\lambda_i.$$

After integrating out the fixed effects, the mode of the integrated likelihood can be used as an estimator.[1] We thus define the *integrated likelihood estimator* $\hat{\beta}$ to be the mode of $L^I(\beta)$.

$$\hat{\beta} = \arg\max_\beta L^I(\beta).$$

The remainder of this paper studies properties of $\hat{\beta}$ and shows how to choose a parametrization that minimizes its mean squared error. At the mode, $\hat{\beta}$, we have

$$L^I_\beta(\hat{\beta}) = \frac{\partial L^I(\beta)}{\partial \beta}|_{\beta=\hat{\beta}} = 0.$$

The delta method gives

$$L^I_\beta(\hat{\beta}) = L^I_\beta(\beta_0) + (\hat{\beta} - \beta_0) L^I_{\beta\beta}(\bar{\beta})$$

where $\bar{\beta}$ is a mean value on the line joining $\hat{\beta}$ and β_0 and $L^I_{\beta\beta}(\bar{\beta})$ is the matrix of second derivatives. We assume that $L^I_{\beta\beta}(\bar{\beta})$ has full rank and omit the argument of expressions when we evaluate them at the truth. This yields

$$\begin{aligned} (\hat{\beta} - \beta_0) &= -[\frac{L^I_{\beta\beta}(\bar{\beta})}{NT}]^{-1} \frac{L^I_\beta}{NT} \\ &= -[\frac{L^I_{\beta\beta}(\bar{\beta})}{NT}]^{-1} \frac{EL^I_\beta}{NT} - [\frac{L^I_{\beta\beta}(\bar{\beta})}{NT}]^{-1} \frac{L^I_\beta - EL^I_\beta}{NT}, \end{aligned} \tag{1}$$

where $E()$ denotes the expectation over the dependent variable and the regressors. The second argument in equation (1), $[\frac{L^I_{\beta\beta}(\bar{\beta})}{NT}]^{-1} \frac{L^I_\beta - EL^I_\beta}{NT}$, converges in probability to zero if $N \to \infty$. The first expression, however, depends on $\frac{EL^I_\beta}{NT} = \frac{EL^{i,I}_\beta}{T}$, which is a function of T but not of N. If EL^I_β is nonzero, then $(\hat{\beta} - \beta_0)$ will be nonzero for any N and $\hat{\beta}$ does not converge in probability to β_0 for $N \to \infty$. This potential inconsistency is the integrated likelihood analogue

of the incidental parameter problem of Neyman and Scott (1948). Neyman and Scott (1948) give some examples of fixed effect or incidental parameter models in which the maximum likelihood estimator fails to be consistent for $N \to \infty$. An intuition for their examples is that the marginal likelihood of the incidental parameters is not sufficiently concentrated.[2] The same intuition applies here. We will now study the properties of $L_\beta^{i,I}$ in order to derive an approximate solution to the incidental parameter problem. With some abuse of the notation, the vector $L_\beta^{i,I}$ can be written as follows,

$$L_\beta^{i,I} = \frac{\int L_\beta e^L d\lambda}{\int e^L d\lambda}. \tag{2}$$

Kass et al. (1990, theorem 7) give a Laplace approximation[3] for ratios of integrals.

$$L_\beta^{i,I} = \frac{\int L_\beta^i e^{L^i} d\lambda}{\int e^{L^i} d\lambda} = L_\beta^i(\hat{\lambda}) - \frac{1}{2}\frac{L_{\beta\lambda\lambda}^i(\hat{\lambda})}{L_{\lambda\lambda}^i(\hat{\lambda})} + \frac{1}{2}\frac{L_{\lambda\lambda\lambda}^i(\hat{\lambda}) L_{\beta\lambda}^i(\hat{\lambda})}{\{L_{\lambda\lambda}^i(\hat{\lambda})\}^2} + O_p(T^{-1}). \tag{3}$$

The expansion is valid for well behaved likelihoods. The following assumption states the regularity conditions as well as the substantive condition that the likelihood has a dominant peak at $\hat{\lambda}$.

Assumption 1: (i) With probability one, $L^i(\beta, \lambda)$ is six times continuously differentiable with respect to λ; (ii) with probability one, $L_\beta^i(\beta, \lambda)$ is four times continuously differentiable with respect to λ; (iii) there exist $\varepsilon > 0$ such that for all $\tilde{\lambda} \in [\lambda_0 - \varepsilon, \lambda_0 + \varepsilon]$, $lim_{T\to\infty}\, L_{\lambda\lambda}^i(\tilde{\lambda}) > -\infty$ and $lim_{T\to\infty}\, L_{\lambda\lambda}^i(\tilde{\lambda}) < 0$; (iv) either $\{\beta_0, \lambda_0\}$ is an element of the interior of a convex set Θ and $L^i(\beta, \lambda)$ is concave for all i or $\{\beta_0, \lambda_0\} \in \Theta$ which is compact; (v) $\{\beta, \lambda\} \neq \{\beta_0, \lambda_0\}$ and $\{\beta, \lambda\} \in \Theta$ implies $L(\beta, \lambda) \neq L(\beta_0, \lambda_0)$ (vi) $E(|\ln L(\beta, \lambda)| < \infty)$ for all $\{\beta, \lambda\} \in \Theta$ (vii) the dependent variable and the regressors are ergodic and strictly stationary with marginal density $p(x, y|\beta, \lambda)$.

This assumption implies that $L_{\lambda\lambda}^i$ is proportional to T. In general, $L_{\lambda\lambda\lambda}^i(\hat{\lambda})$ and $L_{\beta\lambda}^i(\hat{\lambda})$ are $O_p(T)$ so that the second and third term of equation (3) are $O_p(1)$. Thus, $\frac{1}{T}L_\beta^{i,I}$ is $O_p\left(T^{-1}\right)$. Averaging over individuals and taking expectations yields the following lemma.

Lemma 1 Let Assumption 1 hold. Then $\frac{1}{NT}EL_\beta^I$ is $O\left(T^{-1}\right)$.

Proof: See appendix 1.

Lemma 1 states an order result for the 'score' of the integrated likelihood. To derive a theorem about the order of the asymptotic bias, we consider the second derivative of the integrated likelihood, $\frac{1}{NT}L^I_{\beta\beta}(\bar{\beta})$, in equation (1). This matrix converges to its expectation which yields the following order result for the asymptotic bias.

Theorem 1

Let assumptions 1 hold. Then $E(\hat{\beta}-\beta_0)$ *is* $O\left(T^{-1}\right)$.

Proof: See appendix 2.

A slight adjustment of the proof yields that the asymptotic bias of the maximum likelihood is of the same order, $O\left(T^{-1}\right)$. Theorem 1 is of particular interest if the order of the squared bias is equal or higher then the order of the variance. The variance of $\hat{\beta}$ is $O((TN)^{-1})$ where NT is the number of observations. The squared bias is the dominant term of the mean squared error of $\hat{\beta}$ under the following assumption.

Assumption 2: $T \propto N^{\alpha}$ *where* $\alpha \leq 1$.

Assumption 2 states that T increases at the same or slower rate than N. Combining the expression of $(\hat{\beta}-\beta_0)$ in equation (1) with theorem 1 and assumption 3 yields the following theorem.

Theorem 2

Let assumptions 1-2 hold. Then $\hat{\beta}-\beta_0$ *is* $O_p\left(T^{-1}\right)$.

Proof: See appendix 3.

Theorem 2 states that the integrated likelihood is consistent with $O_p\left(T^{-1}\right)$, and that the rate of convergence does not depend on how fast N increases as long as $\alpha \leq 1$. It follows from theorem 1 that this slow rate of convergence is the result of the asymptotic bias. In other words, theorem 1 gives the degree of inconsistency of the integrated likelihood estimator, $O\left(T^{-1}\right)$, and thereby states an order result for the incidental parameter problem. It follows

from Liang (1987) and Ferguson (1991) that the degree of inconsistency of the maximum likelihood estimator is of the same order, $O\left(T^{-1}\right)$. The "invariance result" of the maximum likelihood estimator implies that reparametrizations do not change the estimates. In particular, an information-orthogonal parametrization would yield the same estimates for β as a parametrization that is not information-orthogonal. However, the integrating out method does not have this invariance property and in the next section we show that information-orthogonality can reduce the inconsistency to $O(T^{-2})$ for the integrated likelihood estimator.

3 Orthogonality reduces the inconsistency to $O(T^{-2})$

In this section, we show that information-orthogonality reduces the inconsistency of the integrated likelihood estimator from $O(T^{-1})$ to $O(T^{-2})$. A parametrization of the likelihood is information-orthogonal if the information matrix is block diagonal. That is

$$\begin{aligned} EL_{\beta\lambda}(\beta_0, \lambda_0) &= 0 \\ & \quad \text{i.e.} \\ \int_{y_{\min}}^{y_{\max}} L_{\beta\lambda}(\beta_0, \lambda_0) e^{L(\beta_0, \lambda_0)} dy &= 0, \end{aligned}$$

where y denotes the dependent variable, $y \in [y_{\min}, y_{\max}]$ and $\{\beta_0, \lambda_0\}$ denote the true value of the parameters. Cox and Reid (1987) and Jeffreys (1961) use this concept and refer to it as 'orthogonality'. We prefer the term information-orthogonality to distinguish it from the other orthogonality concepts and to stress that it is defined in terms of the properties of the information matrix. See Tibshirani and Wasserman (1994) and Woutersen (2000) for an overview of orthogonality concepts.

Chamberlain (1984), and Arellano and Honoré (2001) review panel data econometrics in their handbook chapters. All but two of their models can be written in information-orthogonal form.[4] Information orthogonality can require to trim the distribution of the error term or to normalize the regressors. For example, the linear model with exogenous regressors is information-orthogonal if we normalize the regressors to have mean zero, $\sum_t x_{it} = 0$, for all i. Normalizing regressors can be viewed as a reparametrization of the likelihood, see appendix

4 for details. In general, let the individual nuisance parameter that is *not* information-orthogonal be denoted by f. We can interpret f as a function of β and information-orthogonal λ, $f(\beta,\lambda)$, and write the log likelihood as $L(\beta, f(\beta,\lambda))$. Differentiating $L(\beta, f(\beta,\lambda))$ with respect to β and λ yields

$$\begin{aligned}\frac{\partial L(\beta, f(\beta,\lambda))}{\partial \beta} &= L_\beta + L_f \frac{\partial f}{\partial \beta} \\ \frac{\partial^2 L(\beta, f(\beta,\lambda))}{\partial \lambda \partial \beta} &= L_{f\beta}\frac{\partial f}{\partial \lambda} + L_{ff}\frac{\partial f}{\partial \lambda}\frac{\partial f}{\partial \beta} + L_f \frac{\partial^2 f}{\partial \lambda \partial \beta}\end{aligned}$$

where L_f is a score and therefore $EL_f = 0$. Information orthogonality requires the cross-derivative $\frac{\partial^2 L(\beta, f(\beta,\lambda))}{\partial \lambda \partial \beta}$ to be zero in expectation, i.e.

$$EL_{\beta\lambda} = EL_{f\beta}\frac{\partial f}{\partial \lambda} + EL_{ff}\frac{\partial f}{\partial \lambda}\frac{\partial f}{\partial \beta} = 0.$$

This implies the following differential equation

$$EL_{f\beta} + EL_{ff}\frac{\partial f}{\partial \beta} = 0. \tag{4}$$

If equation (4) has an analytical solution then $L(\beta, f(\beta,\lambda))$ is an explicit function of $\{\beta,\lambda\}$ and we refer to such a parametrization as an *explicit parametrization.* In most cases, however, equation (4) has an implicit solution and we have to recover the Jacobian $\frac{\partial \lambda}{\partial f}$ from this implicit solution. In this case, $L(\beta,\lambda)$ has an *implicit parametrization.* The general nonlinear model and the single index model have an information-orthogonal parametrization that is implicit, as shown in appendix 5. The conditional likelihood approach of Cox and Reid (1987) involves maximizing the likelihood with respect to its arguments and Cox and Reid (1987 and 1993) as well as Lancaster (2000) only consider explicit parametrizations. An implicit likelihood calls for a Bayesian framework in which we integrate out the implicitly defined nuisance parameters. We thus extend the integrated likelihood approach to implicitly defined likelihoods. For the remainder of the paper, we assume information-orthogonality.

Assumption 3: $EL_{\beta\lambda}(\beta_0,\lambda_0) = 0$.

Using the Laplace approximation of equation (3) and taking expectations gives

$$EL_\beta^{i,I} = EL_\beta^i(\hat{\lambda}) - \frac{1}{2}E\frac{L^i_{\beta\lambda\lambda}(\hat{\lambda})}{L^i_{\lambda\lambda}(\hat{\lambda})} + \frac{1}{2}E\frac{L^i_{\lambda\lambda\lambda}(\hat{\lambda}) * L^i_{\beta\lambda}(\hat{\lambda})}{\{L^i_{\lambda\lambda}(\hat{\lambda})\}^2} + O(T^{-1}). \tag{5}$$

Given that information-orthogonality is defined at the true value, λ_0, we use a Taylor expansion to write the approximation of $EL_\beta^{i,I}$ as a function of λ_0. We omit the argument if $\lambda = \lambda_0$ and ignore terms that are $O(T^{-1})$. This yields

$$EL_\beta^{i,I} = E[L_\beta^i - \{\frac{L_{\beta\lambda}^i L_\lambda^i + L_{\beta\lambda\lambda}^i}{L_{\lambda\lambda}^i}\} + \frac{1}{2}\{\frac{L_{\lambda\lambda}^i - (L_\lambda^i)^2}{L_{\lambda\lambda}^i}\}\{\frac{L_{\lambda\lambda\lambda}^i L_{\beta\lambda}^i}{(L_{\lambda\lambda}^i)^2}\}] + O(T^{-1}), \tag{6}$$

see appendix 6 for details. The zero score property implies that $EL_\beta^i = 0$. The next two lemmas show that information-orthogonality implies that the last two terms of equation (6) are both $O(T^{-1})$. The intuition for equation (6) being $O(T^{-1})$ is that the last two terms would vanish if $L_{\beta\lambda}(\lambda) = 0$ for all λ. That is, $L_{\beta\lambda}(\lambda) = 0$ for all λ implies $L_{\beta\lambda\lambda} = 0$ and $L_{\beta\lambda} = 0$. We only assume that $EL_{\beta\lambda} = 0$ but this is enough to ensure that EL_β^I is $O(T^{-1})$. We give this result in two lemma's.

Lemma 2 If $EL_{\beta\lambda}^i = 0$ *then* $E\{\frac{L_{\beta\lambda}^i L_\lambda^i + L_{\beta\lambda\lambda}^i}{L_{\lambda\lambda}^i}\}$ is $O(T^{-1})$.
Proof: Differentiating both sides of the equation $EL_{\beta\lambda}^i = 0$ with respect to λ gives $E\{L_{\beta\lambda}^i L_\lambda^i + L_{\beta\lambda\lambda}^i\} = 0$; see appendix 7 for details.

Lemma 3 If $EL_{\beta\lambda}^i = 0$ *then* $E[\{\frac{L_{\lambda\lambda}^i - (L_\lambda^i)^2}{L_{\lambda\lambda}^i}\}\{\frac{L_{\lambda\lambda\lambda}^i L_{\beta\lambda}^i}{(L_{\lambda\lambda}^i)^2}\}]$ *is* $O(T^{-1})$.
Proof: $EL_{\beta\lambda}^i = 0$ implies that $EL_{\beta\lambda}^i = 0$ is $O(\sqrt{T})$; see appendix 8 for details.

Combining lemma 2 and 3 with the expansions of the last section gives a theorem about the reduced asymptotic bias of the integrated likelihood estimator.

Theorem 3
Let assumptions 1-3 hold. Then $E\hat{\beta} - \beta_0$ *is* $O\left(T^{-2}\right)$.
Proof: See appendix 9.

By reducing the order of the bias, theorem 3 also limits the region in the space of asymptotics for which the bias is the dominant term of the mean squared error. This is reflected in the fact that the following assumption is stronger than assumption 3.

Assumption 4: $T \propto N^\alpha$ *where* $\alpha \leq \frac{1}{3}$.

Theorem 4

Let assumptions 1, 3 and 4 hold. Then $\hat{\beta} - \beta_0$ *is* $O_p\left(T^{-2}\right)$.

Proof: See appendix 10.

By reducing the degree of inconsistency to $O\left(T^{-2}\right)$ and increasing the rate of convergence to T^2, theorem 3 and 4 give an approximate solution to the incidental parameter problem of Neyman and Scott (1948). An attractive feature of this 'solution' is that it is likelihood based and gives exact inference in small samples. In contrast, bias-correction and most GMM methods do not have this exact inference feature and can be sensitive to the choice of asymptotics. The author views the integrated likelihood as a convenient way to derive moments that can be robust against misspecification of the parametric error term. In particular, the parametric assumptions on the error term are irrelevant for the models with additive error terms that are discussed in Arellano and Honoré (2001). Given that most panel datasets have more individuals than time periods, we argue in the next subsection that the relevant limiting distribution has T increasing at a slower rate than N. We show that the integrated likelihood estimator is adaptive if $T \propto N^{\alpha}$ where $\alpha > \frac{1}{3}$. In section 4, we discuss priors on λ that are not flat. We can interpret these priors as mixing distributions and discuss how theorems 3 and 4 can be used to increase robustness against the choice of mixing distribution.

3.1 Adaptive Estimation

Alvarez and Arellano (1998) develop an alternative asymptotic where T and N increase at the same rate. Given that T is smaller than N in most panel data, we prefer an asymptotic that includes cases in which T increases at a slower rate than N. Thus, we therefore assume the following throughout this subsection

Assumption 5: $T \propto N^{\alpha}$ *where* $\alpha > \frac{1}{3}$ *and* $EL_{\beta\lambda} = 0$.

In this asymptotic, the integrated likelihood estimator is asymptotically unbiased, normally distributed and adaptive. Theorem 3 states that $(E\hat{\beta} - \beta_0)$ is $O(T^{-2})$. Thus, under

assumption 1, 2 and 5 we have

$$\sqrt{NT}(E\hat{\beta} - \beta_0) = O(\sqrt{\frac{N}{T^3}}) = o(1) \text{ for } N, T \to \infty. \tag{7}$$

Equation (7) shows that the integrated likelihood estimator is asymptotically unbiased. Using this unbiasedness result, it can be easily shown that the integrated likelihood estimator $\hat{\beta}$ has the following asymptotic distribution.

$$\sqrt{NT}(\hat{\beta} - \beta_0) \to N(0, \Psi)$$

where

$$\Psi = \lim_{N,T\to\infty} [\frac{1}{NT} L^I_{\beta\beta}]^{-1} [\frac{1}{NT} (L^I_\beta)(L^I_\beta)'] [\frac{1}{NT} L^I_{\beta\beta}]^{-1}. \tag{8}$$

In the remainder of this subsection, we show that Ψ equals the variance-covariance matrix of the maximum likelihood estimator for known values of the nuisance parameters. Efficiency of the infeasible maximum likelihood estimator implies adaptiveness of the integrated likelihood estimator. We first show that the variance-covariance matrix of equation (8) can be simplified to $\Psi = [\frac{1}{NT} E\{(L^I_\beta)(L^I_\beta)'\}]^{-1}$. The Hessian of the integrated likelihood estimator has the following form.

$$\begin{aligned} L^{i,I}_{\beta\beta} &= \frac{\partial}{\partial\beta} \sum_i \frac{\int L^i_\beta e^{L^i} d\lambda}{\int e^{L^i} d\lambda} \\ &= \sum_i \frac{\int (L^i_{\beta\beta} + L^i_\beta L^{i\prime}_\beta) e^{L^i} d\lambda}{\int e^{L^i} d\lambda} - \sum_i (L^{i,I}_\beta)(L^{i,I}_\beta)'. \end{aligned}$$

A Laplace approximation and the information equality yields that $\frac{1}{NT} \sum_i \{\frac{\int (L^i_{\beta\beta} + L^i_\beta L^{i\prime}_\beta) e^{L^i} d\lambda}{\int e^{L^i} d\lambda}\}$ is $o_p(T^{-1/2})$, see appendix 11 for details. This yields

$$\frac{1}{NT} E L^I_{\beta\beta} = -\frac{1}{NT} E L^I_\beta L^{I\prime}_\beta + o(T^{-1/2}). \tag{9}$$

Thus, the variance-covariance matrix of equation (8) simplifies in our asymptotics[5] in which $T \propto N^\alpha$ where $\alpha > \frac{1}{3}$,

$$\Psi = \lim_{N,T\to\infty} \frac{1}{NT} E[\{(L^I_\beta)(L^I_\beta)'\}]^{-1}.$$

The Laplace approximation gives

$$L_\beta^{i,I} = L_\beta^i + R^i$$

where appendix 1 shows that R^i is $O_p(1)$ and lemma 2 and 3 imply that ER^i is $O(T^{-1})$. Thus

$$\begin{aligned} \frac{L_\beta^I}{\sqrt{NT}} &= \frac{\sum_i L_\beta^i}{\sqrt{NT}} + O_p(T^{-1/2}) + O(\sqrt{\frac{N}{T^3}}) \\ &= \frac{L_\beta}{\sqrt{NT}} + o_p(1). \end{aligned}$$

Therefore,

$$\frac{1}{NT} E\{(L_\beta^I)(L_\beta^I)'\} = \frac{1}{NT} E\{L_\beta L_\beta'\} + o(1).$$

This last equation states that the asymptotic variance of the integrated likelihood estimator equals the asymptotic variance of the maximum likelihood estimator for λ_0 known. We summarize the findings of this subsection in the following theorem.

Theorem 5

Let assumptions 1, and 5 hold. Let the asymptotic variance of $\hat{\beta}_{ML} = \arg\max_\beta L(\beta, \lambda_0)$ *equal* $\Psi = \frac{1}{NT} E\{L_\beta L_\beta'\}$. *Then the integrated likelihood estimator* $\hat{\beta}$ *is an adaptive estimator and*

$$\sqrt{NT}(\hat{\beta} - \beta_0) \to_d N(0, \Psi).$$

Proof: See appendix *12*.

Theorem 5 states that knowledge of λ_0, the true value of the nuisance parameter, does *not* change the asymptotic variance. This implies that the integrated likelihood estimator is adaptive and is therefore efficient. See Bickel (1982), Newey (1990) and Bickel, Klaassen, Ritov and Wellner (1993) for an overview of adaptive estimation. Note that the nuisance parameters are not identified in the sense that $\sqrt{NT}(\hat{\lambda} - \lambda_0) = O_p(\sqrt{N})$ and therefore increases with N. Nevertheless, the assumed asymptotics and $EL_{\beta\lambda} = 0$ ensure that the integrated likelihood estimator is adaptive. Theorem 5 excludes superconsistent estimators by assuming that $\hat{\beta}_{ML}$ converges at the rate of $\sqrt{NT}$. Superconsistency usually implies that

L_β is $O_p(T)$ and $[L_{\beta\beta}]^{-1}$ is $O_p(T^2)$. Using the approximations of equation (3) and (9) shows that L^I_β is $O_p(T)$ and $[L^I_{\beta\beta}]^{-1}$ is $O_p(T^2)$ under regularity conditions. This implies that the integrated likelihood is superconsistent for those cases. Unit roots are usually studied in linear models so we conclude this section by considering the dynamic linear model with fixed effects:

$$y_{it} = y_{i,t-1}\beta + f_i + \varepsilon_{it} \text{ where } E\varepsilon_{it} = 0, E\varepsilon_{it}^2 < \infty \text{ for } E\varepsilon_{is}\varepsilon_{it} = 0 \text{ for } s \neq t \text{ and } t = 1, ..., T.$$

Lancaster (2000) conditions on y_{i0} and suggests the following information-orthogonal parametrization,[6]

$$f_i = y_{i0}(1-\beta) + \lambda_i e^{-b(\beta)} \text{ where } b(\beta) = \frac{1}{T}\sum_{t=1}^{T}\frac{T-t}{t}\beta^t.$$

Analogue to the quasi-maximum likelihood estimator of White (1982), we assume normality of the error terms in order to derive the integrated likelihood estimator. The estimator, however, depends only on the first two moments of y_{it} and is superconsistent in the sense that $T\sqrt{N}(\hat{\beta} - \beta) = O_p(1)$, see appendix 13 for details. In the next section, we show that the dynamic linear model belongs to a class of models for which the integrated likelihood estimator is consistent for fixed T and $N \to \infty$.

3.2 Fixed T Asymptotics

Suppose that T is rather small so that an asymptotics in which $T \propto N^\alpha$ where $\alpha > \frac{1}{3}$ is not satisfactory. Above we showed how to reduce the inconsistency from $O_p(T^{-1})$ to $O_p(T^{-2})$ and in this section we discuss a class of models for which we can derive consistent estimators for T being constant and $N \to \infty$. Note that we can specify an independent[7] prior on the information-orthogonal λ without changing the results of the last sections: $E\frac{\partial^2 L}{\partial\beta\partial\lambda} = 0$ implies $E\frac{\partial^2\{L+\ln\pi(\lambda)\}}{\partial\beta\partial\lambda} = 0$ and the earlier theorems still hold. However, the integrated likelihood contribution changes slightly:

$$L^{i,I}(\beta) = \ln\int \mathfrak{L}^i\pi(\lambda)d\lambda.$$

We define a prior $\pi(\lambda)$ to be a *frequentist prior* if $E\frac{\int L_\beta e^L\pi(\lambda)d\lambda}{\int e^L\pi(\lambda)d\lambda} = 0$. A frequentist prior can sometimes be found by a reparametrization. Suppose we have two information-orthogonal

parametrizations, $\{\beta, \lambda\}$ and $\{\beta, \lambda^*\}$. Consider the following change of variable

$$\begin{aligned} L^I(\beta) &= \ln \int \mathfrak{L}^i d\lambda \\ &= \ln \int \mathfrak{L}^i \frac{\partial \lambda}{\partial \lambda^*} d\lambda^*. \end{aligned}$$

The Jacobian $\frac{\partial \lambda}{\partial \lambda^*}$ can be interpreted as a prior on λ^* in the sense that assuming the prior $\frac{\partial \lambda}{\partial \lambda^*}$ is equivalent to a reparametrization. Sufficient conditions for the existence of a frequentist prior can be stated in terms of $\frac{\partial \lambda}{\partial \lambda^*}$ or in terms of a characterization of $\{\beta, \lambda\}$. The next theorem states a primitive condition for consistency on the parametrization $\{\beta, \lambda\}$.

Theorem 6

Suppose $EL_{\beta\lambda} = 0$, $L_{\beta\lambda\lambda} = 0$, *and* $EL_{\beta\lambda}\tilde{\lambda} = 0$ *where* $\tilde{\lambda}$ *denotes the posterior mean. Let assumption 1 hold and the solution to* $EL^I_\beta(\beta) = 0$ *be unique. Assume that* $N \to \infty$ *and that* T *is fixed. Then*

$$\sqrt{N}(\hat{\beta} - \beta_0) \to N(0, \Psi)$$

where

$$\Psi = [\frac{1}{NT} EL^I_{\beta\beta}]^{-1} [\frac{1}{NT} E\{(L^I_\beta)(L^I_\beta)'\}] [\frac{1}{NT} EL^I_{\beta\beta}]^{-1}.$$

Proof: See appendix 14

Examples of models where the assumption of theorem 6 hold are the exponential model with hazard $f_i e^{x_{it}\beta}$, the Poisson model, and the dynamic linear model with fixed effects that we discussed in the last section. It can be easily shown that the shape of the frequentist prior depends on the parametrization. Using a frequentist prior yields a consistent estimator for $N \to \infty$ and an adaptive estimator for $T \propto N^\alpha$ where $\alpha > \frac{1}{3}$. We return to fixed T asymptotics in section 6.2 in which we show that the integrated likelihood estimator can be viewed as a generalization of the sufficiency principle.

4 Priors and Mixing Distributions

In fixed effect models, the analysis is conditional on the fixed effects. Therefore, the distribution of the fixed effects is not specified[8] and does not need to be estimated. An alternative

estimation strategy is to estimate the parameters of interest and the distribution of the heterogeneity simultaneously. Such models are usually referred to as 'random effect' or 'mixing' models, see Hsiao (1986) for an overview. Obviously, the fact that the mixing distribution is unobserved complicates its estimation. Hsiao (1986), Lancaster (1990) and Van den Berg (2000) argue that the estimates of the parameter of interest are sensitive to the choice of mixing distribution. So there is arguably a need for some robustness against a wrong choice of mixing distribution or its imprecise estimation. Integrating out fixed effects is formally equivalent to a 'random effects' model with a prior distribution playing the role of a mixing distribution and we use this fact extensively in this section. We show that the adaptiveness and $O_p(T^{-2})$-result of the previous section also hold for the random effects model if one specifies the mixing distribution as a function of the orthogonal nuisance parameter.

Let γ be the parameter vector describing the mixing distribution and let the logarithm of the mixing distribution be denoted by $M(\gamma, \lambda)$ where $M(\gamma, \lambda)$ is bounded and γ does *not* contain elements of the common parameter β. If we interpret $M(\gamma, \lambda)$ as a prior then it is a function of λ only. Integrating out the mixing distribution or prior gives an 'integrated likelihood' as a function of the common parameters. Analogue to the last section, we have

$$L_\beta^{i,I} = \frac{\int L_\beta^i e^{L^i + M} d\lambda}{\int e^{L^i + M} d\lambda}.$$

To determine the order of $EL_\beta^{i,I}$, we use the Laplace approximation of Kass et al. (1990) and Tierney et al. (1989),

$$L_\beta^{i,I} = L_\beta^i(\hat{\lambda}) - \frac{M_\lambda(\hat{\lambda}) L_\lambda^i(\hat{\lambda})}{L_{\lambda\lambda}^i(\hat{\lambda})} - \frac{1}{2} \frac{L_{\beta\lambda\lambda}^i(\hat{\lambda})}{L_{\lambda\lambda}^i(\hat{\lambda}) + M_{\lambda\lambda}(\hat{\lambda})} + \frac{1}{2} \frac{L_{\beta\lambda}^i(\hat{\lambda}) \{L_{\lambda\lambda\lambda}^i(\hat{\lambda}) + M_{\lambda\lambda\lambda}(\hat{\lambda})\}}{\{L_{\lambda\lambda}^i(\hat{\lambda}) + M_{\lambda\lambda}(\hat{\lambda})\}^2} + O_p(T^{-1}).$$

Note that $M_{\lambda\lambda}$ is $O_p(1)$ and therefore

$$\begin{aligned} \frac{1}{L_{\lambda\lambda}^i(\hat{\lambda}) + M_{\lambda\lambda}(\hat{\lambda})} &= \frac{1}{L_{\lambda\lambda}^i} + O_p(T^{-2}) \\ \frac{L_{\lambda\lambda\lambda}^i(\hat{\lambda}) + M_{\lambda\lambda\lambda}(\hat{\lambda})}{\{L_{\lambda\lambda}^i(\hat{\lambda}) + M_{\lambda\lambda}(\hat{\lambda})\}^2} &= \frac{L_{\lambda\lambda\lambda}^i}{(L_{\lambda\lambda}^i)^2} + O_p(T^{-2}), \end{aligned}$$

see appendix 15 for details. This yields

$$L_\beta^{i,I} = L_\beta^i(\hat{\lambda}) - \frac{M_\lambda(\hat{\lambda}) L_\lambda^i(\hat{\lambda})}{L_{\lambda\lambda}^i(\hat{\lambda})} - \frac{1}{2} \frac{L_{\beta\lambda\lambda}^i(\hat{\lambda})}{L_{\lambda\lambda}^i(\hat{\lambda})} + \frac{1}{2} \frac{L_{\beta\lambda}^i(\hat{\lambda}) L_{\lambda\lambda\lambda}^i(\hat{\lambda})}{\{L_{\lambda\lambda}^i(\hat{\lambda})\}^2} + O_p(T^{-1}). \tag{10}$$

Under certain regularity conditions given below, $\frac{M_\lambda(\hat{\lambda})L^i_\lambda(\hat{\lambda})}{L^i_{\lambda\lambda}(\hat{\lambda})}$ is $O(T^{-1/2})$ and $E(\frac{M_\lambda(\hat{\lambda})L^i_\lambda(\hat{\lambda})}{L^i_{\lambda\lambda}(\hat{\lambda})})$ is $O(T^{-1})$. Thus, ignoring the mixing distribution gives the same Laplace approximation as in previous sections. After integrating out the mixing distribution or prior we have a function, L^I, that depends only on the common parameters. This distribution concentrates and therefore its mode and marginal posterior give, asymptotically, the same inferences for β. The following theorem states that using a non-flat prior or mixing distribution M does not effect the results of theorem 4 and 5 in the sense that the rate of convergence is T^2 if $T \propto N^\alpha$ where $\alpha \leq \frac{1}{3}$ and the estimator is adaptive if $\alpha > \frac{1}{3}$.

Assumption 6: *The mixing distribution M is a function of γ and λ where $\{\beta \cup \lambda\} = \emptyset$. If M is interpreted as the logarithm of a prior then M is a function of λ only. With probability one, M is two times continuously differentiable with respect to λ. Let $0 \leq e^M < \infty$ be bounded over the whole domain of γ and λ* and let $e^{M(\gamma_0,\lambda_0)} > 0$.

Theorem 7

Let assumptions 1, 3, 4 and 6 hold. Let

$$\{\hat{\beta}, \hat{\gamma}\} = \arg\max_{\beta,\gamma} \frac{1}{NT} \sum_i \ln \int e^{L^i + M} d\lambda.$$

Then $\hat{\beta} - \beta_0$ is $O_p\left(T^{-2}\right)$.

Let assumptions 1, 5 and 6 hold. Let the asymptotic variance of $\hat{\beta}_{ML} = \arg\max_\beta L(\beta, \lambda_0)$ equal $\Psi = \frac{1}{NT} E\{L_\beta L_\beta{}'\}$. Then the integrated likelihood estimator $\hat{\beta}$ is adaptive and

$$\sqrt{NT}(\hat{\beta} - \beta_0) \rightarrow_d N(0, \Psi).$$

Proof: Assumption 1 ensures identification and assumption 6 implies that the derivative of the logarithm can be approximated by the Laplace approximation of equation (3). The previous theorems use the same approximation and the result follows immediately.

An intuition for theorem 7 is that β is 'information-orthogonal' to all other parameters in the sense that

$$\frac{\partial^2 L(\beta,\lambda)+M(\gamma,\lambda)}{\partial\beta\partial\gamma}=0 \quad \text{and} \quad E\frac{\partial^2 L(\beta,\lambda)+M(\gamma,\lambda)}{\partial\beta\partial\lambda}=0.$$

In this interpretation, γ can be viewed as another nuisance parameter that is information-orthogonal to β. Consequently, the inconsistency of the estimate of $\hat{\beta}$ must be $O\left(T^{-2}\right)$.

Mundlak (1978) increases the robustness of a random effects estimator by allowing for correlation between the random effect λ and the mean of the exogenous regressors, $\bar{x}_i = \frac{1}{T}\sum x_{it}$. Mundlak considers only the linear model and writes the random effect λ as a linear function of $\bar{x}_i$ plus random noise. Chamberlain (1980 and 1982) gives reasons why the fixed effect might depend on x_i other than through $\bar{x}_i$. Consider writing the fixed effect as $\lambda_i = v(x_i,\theta)+\eta$ where $v(x_i,\theta)$ is a continuous function in θ and η has a distribution that does not depend on θ or x_i. Suppose we want to estimate the following model

$$y_{it} = x_{it}\beta + \lambda_i + \varepsilon_{it} \quad \text{where } \varepsilon_{it}\sim N(0,\sigma^2) \text{ and } \lambda_i = v(x_i,\theta)+\eta,$$

where x_{it} is a vector of exogenous regressors. The distributional assumptions are not restrictive since the integrated likelihood method yields a projection estimator for the linear model with exogenous regressors. By writing the likelihood as a function of $\tilde{x}_{it} = x_{it} - \bar{x}_i$ we ensure orthogonality between β and λ.

$$\frac{\partial^2 L}{\partial\beta\partial\lambda_i} = \sum_t \tilde{x}_{it} = 0.$$

We can write the estimation problem in the form

$$\max_{\beta,\theta,\gamma}\int e^{L(\beta,\lambda)+M(\theta,\gamma,\lambda)}d\lambda, \tag{11}$$

where $M(\theta,\gamma,\lambda)$ is implied by $v(x_i,\theta)$ and the distribution of η. The requirements of theorem 7 are satisfied and β can be estimated up to $O_p(T^{-2})$ or adaptively. This approximation result holds for all models for which we can find an explicit parametrization of the nuisance parameter. Therefore, equation (11) with an information-orthogonal parametrization is a generalization of Mundlak's (1978) framework to nonlinear models. Thus, specifying the prior or mixing distribution in terms of an informational orthogonal nuisance parameter yields a robust and potentially adaptive estimator.

5 Examples

In this section we discuss two examples. The example of the gamma distribution illustrates that the inconsistency of the integrated likelihood estimator is $O_p(T^{-2})$. The second example is an 'inference problem' by Neyman and Scott (1948) and we show that the integrated likelihood provides a natural solution to this problem.

5.1 Gamma distribution with individual parameters

Consider the gamma distribution and assume that we observe T observations per individual for N individuals. The observations are independent of each other and have the following density function

$$y_{it} \sim Gamma(\alpha, f_i)$$

where α is a common parameter and f_i is a person specific nuisance parameter. We are interested in estimating α, the relative variance of this gamma distribution. As shown in appendix 16, f_i is *not* informational-orthogonal to α but an information-orthogonal parameterization can be derived. In particular, defining $\lambda_i = \frac{f_i}{\alpha}$ as the new nuisance parameter yields an information-orthogonal parametrization. In this parametrization, we have

$$y_{it} \sim Gamma(\alpha, \alpha\lambda_i).$$

Thus, we changed the *parametrization* of the model but none of the assumptions. In the new parametrization, the likelihood contribution of individual i has the following form

$$\mathfrak{L}^i(\alpha, \lambda_i) = \prod_t \frac{(\alpha\lambda_i)^\alpha y_{it}^{\alpha-1} e^{-\alpha\lambda_i y_{it}}}{\Gamma(\alpha)} \text{ for } i = 1, ..., N. \tag{12}$$

Integrating with respect to λ_i yields the integrated likelihood contribution,.

$$\begin{aligned}
\mathfrak{L}^{i,I}(\alpha, \lambda_i) &= \int \mathfrak{L}^i d\lambda_i = \int \frac{(\alpha\lambda_i)^{T\alpha} e^{-\alpha\lambda_i \sum_t y_{it}} \prod_t (y_{it}^{\alpha-1})}{\Gamma(\alpha)^T} d\lambda_i \\
&= \frac{\alpha^{T\alpha} \prod_t (y_{it}^{\alpha-1})}{\Gamma(\alpha)^T} \int {\lambda_i}^{T\alpha} e^{-\alpha\lambda_i \sum_t y_{it}} d\lambda_i \\
&= \frac{\prod_t (y_{it}^{\alpha-1}) * \Gamma(T\alpha + 1)}{\alpha\Gamma(\alpha)^T (\sum_t y_{it})^{T\alpha+1}}.
\end{aligned}$$

To derive the score of the integrated likelihood, we take the logarithm of $\mathfrak{L}^I$,

$$L^{i,I} = \ln \mathfrak{L}^{i,I} = (\alpha - 1) \sum_t \ln y_{it} + \ln \Gamma(T\alpha + 1) - \ln(\alpha) - T \ln \Gamma(\alpha) - (T\alpha + 1) \ln(\sum_t y_{it}).$$

Differentiating with respect to α gives

$$L^I_\alpha = \sum_t \ln y_{it} + T\psi(T\alpha + 1) - \frac{1}{\alpha} - T\psi(\alpha) - T \ln(\sum_t y_{it}).$$

The equation $EL^I_\alpha = 0$ is uniquely solved for $\alpha = \alpha_0$. Appendix 17 shows that the integrated likelihood estimator is consistent for either T or N going to infinity where the inconsistency of the maximum likelihood is $O(T^{-1})$.

Suppose we want to use a nonflat prior for λ, e.g. $\pi(\lambda) = \lambda$. This prior is neither proper nor bounded. Integrating the likelihood with respect to the prior and taking logarithms yields

$$L^{i,I} = (\alpha - 1) \sum_t \ln y_{it} + \ln \Gamma(T\alpha + 2) - 2 \ln(\alpha) - T \ln \Gamma(\alpha) - (T\alpha + 2) \ln(\sum_t y_{it}).$$

Differentiating with respect to α gives

$$\begin{aligned} L^I_\alpha &= \sum_t \ln y_{it} + T\psi(T\alpha + 2) - \frac{2}{\alpha} - T\psi(\alpha) - T \ln(\sum_t y_{it}) \\ EL^I_\alpha &= T\psi(T\alpha + 2) - \frac{2}{\alpha} - T\psi(T\alpha) = \frac{1}{\alpha(T\alpha + 1)} \text{ is } O(T^{-1}). \end{aligned}$$

Note that

$$L^I_{\alpha\alpha} = T^2 \psi'(T\alpha + 2) + \frac{2}{\alpha^2} - T\psi'(\alpha)$$

so that $\frac{1}{L^I_{\alpha\alpha}}$ is $O(T^{-1})$. Thus, the inconsistency ($N \to \infty$) of the integrated likelihood estimator, $\frac{EL^I_\alpha}{EL^I_{\alpha\alpha}}$, is $O(T^{-2})$ in this example. Note that using the prior $\pi(\lambda) = \lambda$ in Gamma$(\alpha, \alpha\lambda)$ is equivalent to the parametrization Gamma$(\alpha, \frac{\alpha\lambda^2}{2})$ since $\frac{\partial \lambda^2}{\partial \lambda} = \lambda$. For this reason, we use the term 'integrated likelihood' for all priors that correspond to an alternative parametrization of the likelihood.

[graph "bias" about here]

The graph shows the performance of the estimators for small T and $N \to \infty$. The absolute bias is shown as a function of T. We choose the true value of the parameter of interest, $\alpha_0 = 1$. A nice feature of this example is that the distribution of f does not effect the small T performance of the estimators and that the variance of the estimators is the same. Using a flat prior, the integrated hazard estimator asymptotically unbiased. If we use the prior $\pi(\lambda) = \lambda$ then the bias order result seems to be relevant for small T since the absolute bias is approximately $\frac{1}{T^2}$. In contrast, the absolute bias of the maximum likelihood estimator decreases at the rate $\frac{1}{T}$.

5.2 Neyman and Scott (1948), example 2

Neyman and Scott (1948) consider several examples in which the maximum likelihood estimator fails to be consistent. Their example 2 (page 4) assumes a model where the dependent variable y_{it} has a common variance but allows for individual means.

$$y_{it} \sim N(\lambda_i, \sigma^2).$$

Neyman and Scott note that the inconsistency of the maximum likelihood estimator for σ^2 is $O(T^{-1})$. We show that the integrated likelihood is unbiased and a consistent estimator. Obviously, the log likelihood contribution of individual i is

$$L^i = -\frac{\log \sigma^2}{2} - \frac{\sum_t (y_{it} - \lambda_i)^2}{2T\sigma^2}.$$

Note that σ^2 and λ_i are information-orthogonal. Let the likelihood contribution of individual i be denoted by $\mathfrak{L}^i$,

$$\mathfrak{L}^i \propto \sigma^{-T} \exp\{-\frac{1}{2}(\frac{\sum_t (y_{it} - \bar{y}_i)^2}{\sigma^2} + \frac{T(\bar{y}_i - \lambda_i)^2}{\sigma^2})\}$$

where $\bar{y}_i = \sum_t y_{it}/T$. We integrate with respect to the nuisance parameter λ_i to derive the integrated likelihood contribution, $\mathfrak{L}^{i,I}$.

$$\begin{aligned} \mathfrak{L}^{i,I} &= \int \mathfrak{L}^i \, d\lambda_i \propto \int \sigma^{-T} \exp\{-\frac{1}{2}(\frac{\sum_t (y_{it}^2 - \bar{y}_i^2)}{\sigma^2} + \frac{T(\bar{y}_i - \lambda_i)^2}{\sigma^2})\} d\lambda_i \\ &\propto \sigma^{-(T-1)} \exp(-\frac{\sum_t (y_{it} - \bar{y}_i)^2}{\sigma^2}), \end{aligned}$$

see appendix 18 for details. This yields

$$L^{i,I} = -(\frac{T-1}{2})\log\sigma^2 - \frac{\sum_t (y_{it}-\bar{y}_i)^2}{\sigma^2}.$$

Differentiating with respect to the parameter of interest, σ^2, gives

$$L^{i,I}_{\sigma^2} = -(\frac{T-1}{2})\frac{1}{\sigma^2} + \frac{\sum_t (y_{it}-\bar{y}_i)^2}{\sigma^4}.$$

Thus, the integrated likelihood estimator for σ^2 is

$$\widehat{\sigma^2} = \frac{1}{N}\sum_i \frac{\sum_t (y_{it}-\bar{y}_i)^2}{T-1}.$$

Note that $E\widehat{\sigma^2} = \widehat{\sigma_0^2}$ and that the integrated likelihood estimator is consistent and unbiased for fixed T. This estimator of σ^2 is usually obtained with an 'ad hoc' solution: replacing T by $(T-1)$ in the denominator. With the integrated likelihood estimator, no 'ad hoc' adjustment is required.

6 'Differencing out' and 'Sufficient Statistic' as special cases

6.1 'Differencing out'

For a couple of models we can difference out the fixed effect and then derive a consistent estimator for the common parameters. This class of models seems to be limited to:

(i). The Weibull and exponential hazard model with exogenous regressors. The likelihood of the differenced logarithms of the durations does not depend on the fixed effects.

(ii). The linear regressor model with exogenous regressors: $y_{it} = x_{it}\beta + \lambda_i + \varepsilon_{it}$ where $E\varepsilon_{it} = 0$ and $E\varepsilon_{it}^2 < \infty$. Regressing $y_{it} - y_{i,t-1}$ on $x_{it} - x_{i,t-1}$ yields a consistent estimator.

For both models, we can attain an information-orthogonal parametrization.

ad. (i). For the Weibull model, Cox and Reid (1987) show that the orthogonal fixed effect can be written as $f_i = e^{\alpha\lambda_i + \psi(2)}$ where α is the Weibull parameter and $\sum_t x_{it} = 0$. Integrating the likelihood with respect to λ_i yields the likelihood for the first differenced data, see Lancaster (2000).

ad. (ii). If we require $\sum_t x_{it} = 0$ in the linear model then β and λ_i are information-orthogonal.[9] Assuming normality of ε_{it} and integrating out the fixed effects yields the usual GLS difference estimator. This estimator requires the first two moments of the error term to be finite but does not require normality.

6.2 'Sufficient Statistic'

For a small class of models, we can eliminate the nuisance parameters by conditioning on a sufficient statistic. This class seems to be limited to the following fixed effects models of the exponential family: the Poisson, logit, Weibull, and linear model with known variance. Lancaster (2000) shows that the integrating out method yields the usual moment functions for the Poisson model. The informative observations in a logit model are of those individuals that have both possible outcomes. The likelihood of these observations does not depend on the individual effect so there is no need for integration. Conditioning on a sufficient statistic in the Weibull or linear model with known variance yields the 'difference' estimator. We discussed in the last subsection that these 'difference estimators' are special cases of the more general integrated likelihood estimator with information-orthogonal fixed effects. Cox and Reid (1987, page 8) present their conditional profile likelihood method as a generalization of eliminating nuisance parameters by sufficient statistics. By conditioning on maximum likelihood estimates of the nuisance parameters given β_0, Cox and Reid derive the following objective function.[10]

$$L^{CR}(\beta) = L(\beta, \hat{\lambda}) - \frac{1}{2}\ln|L_{\lambda\lambda}(\beta, \hat{\lambda})|. \tag{13}$$

Differencing with respect to β gives

$$L^{CR}_{\beta}(\beta) = L_{\beta}(\beta, \hat{\lambda}) + \frac{1}{2}\frac{L_{\lambda\lambda\beta}(\beta, \hat{\lambda})}{L_{\lambda\lambda}(\beta, \hat{\lambda})} + \frac{1}{2}\frac{L_{\lambda\lambda\lambda}(\beta, \hat{\lambda})}{L_{\lambda\lambda}(\beta, \hat{\lambda})}\frac{\partial\hat{\lambda}}{\partial\beta}$$

where $\frac{\partial\hat{\lambda}}{\partial\beta} = -\frac{L_{\beta\lambda}}{L_{\lambda\lambda}} + O_p(T^{-1})$. Thus

$$L^{CR}_{\beta}(\beta) = L_{\beta}(\beta, \hat{\lambda}) - \frac{1}{2}\frac{L_{\lambda\lambda\beta}(\beta, \hat{\lambda})}{L_{\lambda\lambda}(\beta, \hat{\lambda})} - \frac{1}{2}\frac{L_{\lambda\lambda\lambda}(\beta, \hat{\lambda})L_{\beta\lambda}(\beta, \hat{\lambda})}{L^2_{\lambda\lambda}(\beta, \hat{\lambda})} + O_p(T^{-1}).$$

The last expression is the approximation of L_β^I as was given in equation (3). Thus, the conditional profile likelihood can be viewed as an approximation of the integrated likelihood method. Although both methods use information-orthogonal parametrizations, the conditional profile likelihood requires an explicit parametrization of the likelihood. As we discussed above, an implicit parametrization suffices for the integrated likelihood method. This advantage of Bayesian analysis is not confined to panel data. If we can write f as an analytical function of β and λ, i.e. $\lambda = g(\beta, \lambda)$, then we can obviously write the likelihood function in terms of β and λ. The question is, however, whether there exist an analytical solution to the differential equation of equation (4). In his review of Cox and Reid (1987), Critchley (1987) wonders "How often are the differential equations[11] soluble analytically?". An analytical solution is a necessary condition for writing the likelihood as an analytical function of β and λ. An example of Hills (1987), however, shows that being able to write λ as a function of β and f does not imply that f can be written as an explicit function of β and λ: "the inverse of this transformation is not explicit and therefore it is not possible to write the likelihood function in the form $L(\beta, \lambda)$". A reparametrization is relatively easy in a Bayesian framework since we only need the Jacobian $\frac{\partial \lambda}{\partial f}$.

$$\int e^L d\lambda = \int e^L \frac{\partial \lambda}{\partial f} df.$$

Thus, the information-orthogonality can be used for a wider class of models than considered by Cox and Reid (1987 and 1993). Berger et al. (1999) consider inference in the presence of nuisance parameters and mention generality and simplicity as arguments in favor of the integrated likelihood method (over the profile likelihood). To the author, however, implicit versus explicit parametrization is a more tangible advantage of the Bayesian approach. We therefore formulate all theorems in terms of the integrated likelihood estimator. Cox and Reid (1987 and 1993) and Ferguson, Cox and Reid (1991) do not consider panel data. Given that all the theorems are based on the approximation of equation (3), they also hold for the conditional profile likelihood. The following theorem follows directly from theorem 4 and 5. The conditional profile likelihood estimator is denoted by $\hat{\beta}_{CPL}$ and maximizes the objective function of equation (13).

Theorem 8

Let assumptions 1, 3 and 4 hold. Then $\hat{\beta}_{CPL} - \beta_0$ *is* $O_p\left(T^{-2}\right)$.

Let assumptions 1 and 5 hold and let the asymptotic variance of $\hat{\beta}_{ML} = \arg\max_\beta L(\beta, \lambda_0)$ *be* $\Psi = \frac{1}{NT} E\{L_\beta L_\beta{}'\}$. *Then* $\hat{\beta}_{CPL}$ *is an adaptive estimator and*

$$\sqrt{NT}(\hat{\beta}_{CPL} - \beta_0) \to_d N(0, \Psi).$$

Proof: Assumption 1 implies identification. The difference between the scores of the conditional profile likelihood and the integrated likelihood is $O(T^{-1})$ and the result follows immediately.

The beauty of the integrating out approach is its combination of simplicity and generality: It is as easy or easier to compute than competing methods and 'differencing out' or 'conditioning methods' can be viewed as special cases.

7 Predetermined Variables

Neyman and Scott (1948) describe the incidental parameter problem by showing that the maximum likelihood estimator fails to be consistent in a couple of examples. The regressors of these examples are all exogenous but the incidental parameter problem obviously remains when the assumption of exogeneity is relaxed. In the previous sections, we showed that the incidental parameter problem can be solved by using an information-orthogonal parametrization of the likelihood. This framework allows for predetermined regressors that are lagged dependent variables. Quite often, however, one is not willing to specify the stochastic process of such general predetermined variables. In their handbook chapter, Arellano and Honoré (2001) note, "almost nothing is known about nonlinear models with general predetermined variables". This section derives new estimators for single index models with general predetermined variables and fixed effects. The following definition generalizes Chamberlain's (1984) concept of conditional strict exogeneity to predetermined or weakly exogenous variables. A

regressor x is *conditionally weakly exogenous* if

$$P(y_{it}|x_{i1},...,x_{it},...,x_{iT},\lambda_i) = P(y_{it}|x_{i1},...,x_{it},\lambda_i) \text{ for all } i, \tag{14}$$

where x_{it} can include lagged values of y_{it}. For models with weakly exogenous regressors, the technique to condition on a sufficient statistic cannot work. A sufficient statistic for the incidental parameter λ_i would be a function of $y_{i1},...,y_{iT}$. Conditioning requires the distribution of the sufficient statistic conditional on the predetermined regressors of all periods. This distribution is not specified since the regressors are only required to be predetermined. The integrated likelihood approach allows for some misspecification of the likelihood but requires that $EL_\beta(\beta_0,\lambda_0) = 0$ and $EL_{\beta\lambda}(\beta_0,\lambda_0) = 0$. For predetermined variables, we propose using a moment function that has zero expectation at the true values for β and λ and whose derivative with respect to λ is zero in expectation. Thus, the moment function $Q_\beta(\beta,\lambda)$ is an *information-orthogonal moment function* if (i) $EQ_\beta(\beta_0,\lambda_0) = 0$ and (ii) $EQ_{\beta\lambda}(\beta_0,\lambda_0) = 0$ where $\{\beta_0,\lambda_0\}$ denotes the true values of the parameters of interest and the vector of nuisance parameters. The function $Q_\beta(\beta,\lambda)$ plays the same role as the score function L_β of the previous section. Its dependence on λ is reduced by a particular parametrization and the next step is to integrate out λ with respect to the likelihood.

Consider the binairy choice model with predetermined variables, $Pr(Y_{it} = 1) = f(\mu_{it})$ where $\mu_{it} = x_{it}\beta + \lambda_i$ and $f()$ denotes a density function. The logit, probit and all other parametric binairy choice models fall in this class. Let $L_{\mu_{it}}$ denotes the log likelihood diffentiated with respect to μ_{it} and $L_{\mu_{it}\mu_{it}}$ denotes the second derivative. We propose the following moment function for this class of models.

$$\begin{aligned} Q^I_\beta(\beta) &= \sum_i \frac{\int Q^i_\beta(\beta,\lambda)e^{L^i}d\lambda}{\int e^{L^i}d\lambda} \text{ where} \\ Q^i_\beta(\beta,\lambda) &= \sum_{t=1}^{T-1} x_{it}\{L_{\mu_{it}} - (EL_{\mu_{it}\mu_{it}}|x^t_{i,},\beta,\lambda_i)(EL_{\mu_{i,t+1}\mu_{i,t+1}}|x^{t+1}_i,\beta,\lambda_i)^{-1}L_{\mu_{i,t+1}}\}, \end{aligned} \tag{15}$$

where x_{it} denotes a vector of predetermined variables and $x^t_i = \{x_{i1},...,x_{it}\}$. Note that $Q^i_\beta(\beta,\lambda)$ is information-orthogonal since $\frac{\partial\mu_{it}}{\partial\lambda} = 1$ and

$$EQ^i_{\beta\lambda}(\beta) = E\sum_{t=1}^{T} x_{it}\{L_{\mu_{it}\mu_{it}} - E(L_{\mu_{it}\mu_{it}}|x^t_{i,},\beta,\lambda_i)\} = 0.$$

The same vector moment can be used for the General Linear Model. Suppose $y_{it} = G(\mu_{it}) + \varepsilon_{it}$ where $\mu_{it} = X_{it}\beta + \lambda_i$, $G()$ is a known function, $E(\varepsilon_{it}|x_i^t) = 0$ and $E(\varepsilon_{it}^2|x_i^t) < \infty$. Assuming normality and homoscedasticity of the error term yields a quasi likelihood. Applying the moment function of equation (15) to the linear model with predetermined variables yields a familiar[12] moment function:

$$Q_\beta^I(\beta) = \sum_{i=1}^{N}\sum_{t=1}^{T-1} x_{it}(\varepsilon_{it} - \varepsilon_{i,t+1}).$$

Identification conditions are given in appendix 19 and Woutersen (2000) gives a general discussion how orthogonality concepts simplify identification proofs. The moment function of equation (15) can be adapted to deal with endogenous regressors,

$$Q_\beta^i(\beta) = \sum_{t=1}^{T-1} x_{i,t-1}\{L_{\mu_{it}} - (EL_{\mu_{it}\mu_{it}}|x_{i,},\lambda_i)(EL_{\mu_{i,t+1}\mu_{i,t+1}}|x_{i,t+1},\lambda_i)^{-1}L_{\mu_{t+1}}\}.$$

Note that $x_{i,t-1}$ can be replaced by an instrument z_{it} that is independent of $L_{\mu_{it}}$ and $L_{\mu_{i,t+1}}$ for all i. In both cases, $Q_\beta^i(\beta)$ is information-orthogonal. Analogue to the previous theorems about score functions, we now derive a theorem for information-orthogonal moment functions.

Theorem 9

Suppose $\hat{\beta} = \arg\min_\beta\{Q_\beta^I(\beta)'Q_\beta^I(\beta)\}$, $EQ_\beta = 0$, $EQ_{\beta\lambda} = 0$ *and* $T \propto N^\alpha$ *where* $\alpha > \frac{1}{3}$. *Let assumption 1 hold and the solution to* $EQ_\beta^I(\beta) = 0$ *be unique. Then*

$$\sqrt{NT}(\hat{\beta} - \beta_0) \to N(0, \Psi)$$

where

$$\Psi = [\frac{1}{NT}E(Q_\beta L_\beta{}')]^{-1}[\frac{1}{NT}E(Q_\beta Q_\beta{}')][\frac{1}{NT}E(Q_\beta L_\beta{}')]^{-1}.$$

Proof: See appendix 20.

8 Conclusion

This paper develops the integrated moment estimator and extends the integrated likelihood estimator to implicitly defined likelihoods. It shows that the integrated likelihood method yields an approximate solution to the incidental parameter problem of Neyman and Scott

(1948) for information-orthogonal likelihoods. A nice feature of the integrated likelihood method is its generality: 'differencing' and 'sufficient statistic' estimators were shown to be special cases. In a Bayesian framework, reparametrization of a nuisance parameter only requires an expression of the Jacobian. Berger et al. (1999) ignore this advantage but, to the author, this is the most tangible argument in favor of Bayesian analysis. All but two models considered by Chamberlain (1984) and Arellano and Honoré (2001) are information-orthogonal with an explicit parametrization. Using an implicit parametrization, we derived new estimators for the single index models with lagged dependent variables that converge at a faster rate than existing ones.

In their conclusion, Arellano and Honoré (2001) note, "almost nothing is known about nonlinear models with general predetermined variables." Using information-orthogonality, we derive new estimators for the general linear model and single index model with fixed effects and predetermined variables. It thus seems that implicit parametrizations and informational moment functions are very promising approaches to study models with general predetermined variables and many nuisance parameters.

It was also shown that the prior could be interpreted as a mixing distribution so that the robustness of the random effects model could be increased. If it is considered to be important that empirical results can be replicated by people with different ideas about the prior or intestable aspects of the mixing distribution, then it is important to limit the influence of priors and mixing distributions. This paper makes a contribution to attaining that goal.

9 Appendices

Appendix 1. Lemma 1: *Let assumption 1 hold. Then* $\frac{1}{NT}EL_\beta^I$ is $O\left(T^{-1}\right)$.

Proof: Equation (3) in the text states that

$$L_\beta^{i,I} = \frac{\int L_\beta^i e^{L^i} d\lambda}{\int e^{L^i} d\lambda} = L_\beta^i(\hat{\lambda}) - \frac{1}{2}\frac{L_{\beta\lambda\lambda}^i(\hat{\lambda})}{L_{\lambda\lambda}^i(\hat{\lambda})} + \frac{1}{2}\frac{L_{\lambda\lambda\lambda}^i(\hat{\lambda})L_{\beta\lambda}^i(\hat{\lambda})}{\{L_{\lambda\lambda}^i(\hat{\lambda})\}^2} + O_p(T^{-1})$$

where the second and third term are in general $O(1)$. To ensure that these terms do not cancel we derive a Taylor approximation around λ.

$$L_\beta^{i,I} = L_\beta^i + L_{\beta\lambda}^i(\hat{\lambda}-\lambda) + \frac{1}{2}L_{\beta\lambda\lambda}^i(\hat{\lambda}-\lambda)^2 - \frac{1}{2}\frac{L_{\beta\lambda\lambda}^i}{L_{\lambda\lambda}^i} + \frac{1}{2}\frac{L_{\lambda\lambda\lambda}^i L_{\beta\lambda}^i}{\{L_{\lambda\lambda}^i\}^2} + O_p(T^{-1/2}).$$

The definition of $\hat{\lambda}$ implies that

$$L_\lambda^i(\hat{\lambda}) = L_\lambda^i + (\hat{\lambda}-\lambda_0)L_{\lambda\lambda}^i(\bar{\lambda}) = 0, \text{ and}$$

$$(\hat{\lambda}-\lambda_0) = -\frac{L_\lambda^i}{L_{\lambda\lambda}^i(\bar{\lambda})} \text{ is } O_p(T^{-1}).$$

Similarly,

$$L_\lambda^i(\hat{\lambda}) = L_\lambda^i + (\hat{\lambda}-\lambda_0)L_{\lambda\lambda}^i + \frac{1}{2}(\hat{\lambda}-\lambda_0)^2 L_{\lambda\lambda\lambda}^i + \frac{1}{6}(\hat{\lambda}-\lambda_0)^3 L_{\lambda\lambda\lambda\lambda}^i(\bar{\lambda}) = 0.$$

This yields

$$(\hat{\lambda}-\lambda_0) = -\frac{L_\lambda^i}{L_{\lambda\lambda}^i} + O_p(T^{-1}),$$

and

$$\begin{aligned}
(\hat{\lambda}-\lambda_0) &= -\frac{L_\lambda^i}{L_{\lambda\lambda}^i} - \frac{1}{2}\frac{L_{\lambda\lambda\lambda}^i}{L_{\lambda\lambda}^i}(\frac{L_\lambda^i}{L_{\lambda\lambda}^i})^2 + O_p(T^{-3/2}) \\
(\hat{\lambda}-\lambda_0)^2 &= (\frac{L_\lambda^i}{L_{\lambda\lambda}^i})^2 + O_p(T^{-3/2}) \\
\frac{1}{L_{\lambda\lambda}^i(\hat{\lambda})} &= \frac{1}{L_{\lambda\lambda}^i} + O_p(T^{-3/2}) = \frac{1}{EL_{\lambda\lambda}^i} + O_p(T^{-3/2}).
\end{aligned}$$

Using these terms in the Taylor expansion gives the following

$$\begin{aligned}
L_\beta^{i,I} &= L_\beta^i + L_{\beta\lambda}^i(\hat{\lambda}-\lambda_0) + \frac{1}{2}L_{\beta\lambda\lambda}^i(\hat{\lambda}-\lambda_0)^2 - \frac{1}{2}\frac{L_{\beta\lambda\lambda}^i}{L_{\lambda\lambda}^i} + \frac{1}{2}\frac{L_{\lambda\lambda\lambda}^i L_{\beta\lambda}^i}{\{L_{\lambda\lambda}^i\}^2} + O_p(T^{-1/2}) \\
&= L_\beta^i - \frac{L_\lambda^i L_{\beta\lambda}^i}{L_{\lambda\lambda}^i} - \frac{1}{2}\frac{L_{\lambda\lambda\lambda}^i L_{\beta\lambda}^i}{L_{\lambda\lambda}^i}(\frac{L_\lambda^i}{L_{\lambda\lambda}^i})^2 + \frac{1}{2}(\frac{L_\lambda^i}{L_{\lambda\lambda}^i})^2 L_{\beta\lambda\lambda}^i - \frac{1}{2}\frac{L_{\beta\lambda\lambda}^i}{L_{\lambda\lambda}^i} + \frac{1}{2}\frac{L_{\lambda\lambda\lambda}^i L_{\beta\lambda}^i}{\{L_{\lambda\lambda}^i\}^2} + O_p(T^{-1/2}) \\
&= L_\beta^i - \{\frac{L_{\beta\lambda}^i L_\lambda^i + L_{\beta\lambda\lambda}^i}{L_{\lambda\lambda}^i}\} + \frac{1}{2}\{\frac{L_{\lambda\lambda}^i + (L_\lambda^i)^2}{L_{\lambda\lambda}^i}\}\frac{L_{\beta\lambda}^i}{L_{\lambda\lambda}^i} + \frac{1}{2}\{\frac{L_{\lambda\lambda}^i - (L_\lambda^i)^2}{L_{\lambda\lambda}^i}\}\{\frac{L_{\lambda\lambda\lambda}^i L_{\beta\lambda}^i}{(L_{\lambda\lambda}^i)^2}\} + O_p(T^{-1/2}).
\end{aligned}$$

Thus

$$L_\beta^{i,I} = L_\beta^i - \frac{L_\lambda^i L_{\beta\lambda}^i}{L_{\lambda\lambda}^i} + O_p(1)$$

Note that the information equality implies that $E\{L_{\lambda\lambda}^i + (L_\lambda^i)^2\} = 0$ and that $L_{\beta\lambda\lambda}^i - EL_{\beta\lambda\lambda}^i$ is $O(\sqrt{T})$. Thus,

$$E[\{\frac{L_{\lambda\lambda}^i + (L_\lambda^i)^2}{L_{\lambda\lambda}^i}\}\frac{L_{\beta\lambda\lambda}^i}{L_{\lambda\lambda}^i}] = E[\{\frac{L_{\lambda\lambda}^i + (L_\lambda^i)^2}{EL_{\lambda\lambda}^i}\}\frac{L_{\beta\lambda\lambda}^i - EL_{\beta\lambda\lambda}^i}{EL_{\lambda\lambda}^i}] \text{ is } O(T^{-1/2}).$$

This gives

$$EL_\beta^{i,I} = -E[\{\frac{L_{\beta\lambda}^i L_\lambda^i + L_{\beta\lambda\lambda}^i}{L_{\lambda\lambda}^i}\} + \frac{1}{2}\{\frac{L_{\lambda\lambda}^i - (L_\lambda^i)^2}{L_{\lambda\lambda}^i}\}\{\frac{L_{\lambda\lambda\lambda}^i L_{\beta\lambda}^i}{(L_{\lambda\lambda}^i)^2}\}] + O_p(T^{-1/2})$$

which is $O(1)$. Therefore, $\frac{1}{NT}EL_\beta^I$ is $O\left(T^{-1}\right)$. *Q.E.D.*

Appendix 2. Theorem 1: *Let assumption 1 hold. Then* $E(\hat{\beta} - \beta_0)$ *is* $O\left(T^{-1}\right)$.

Proof: Equation (1) in the text states that

$$(\hat{\beta} - \beta_0) = -[\frac{L_{\beta\beta}^I(\bar{\beta})}{NT}]^{-1}\frac{EL_\beta^I}{NT} - [\frac{L_{\beta\beta}^I(\bar{\beta})}{NT}]^{-1}\frac{L_\beta^I - EL_\beta^I}{NT}.$$

The elements of the matrix $\frac{1}{NT}L_{\beta\beta}^I(\bar{\beta})$ are just the averages of the second derivatives,

$$\begin{aligned} L_{\beta\beta}^{i,I} &= \frac{\partial}{\partial\beta}\sum_i \frac{\int L_\beta^i e^{L^i} d\lambda}{\int e^{L^i} d\lambda} \\ &= \sum_i \frac{\int \{L_{\beta\beta}^i - (L_\beta^i)^2\} e^{L^i} d\lambda}{\int e^{L^i} d\lambda} + \sum_i \{\frac{\int L_\beta^i e^{L^i} d\lambda}{\int e^{L^i} d\lambda}\}\{\frac{\int L_\beta^i e^{L^i} d\lambda}{\int e^{L^i} d\lambda}\}'. \end{aligned} \tag{16}$$

The elements of $\sum_i [\{\frac{\int L_\beta^i e^{L^i} d\lambda}{\int e^{L^i} d\lambda}\}\{\frac{\int L_\beta^i e^{L^i} d\lambda}{\int e^{L^i} d\lambda}\}']$ are proportional to NT so that $[\frac{L_{\beta\beta}^I(\bar{\beta})}{NT}]^{-1}$ is $O(1)$. Lemma 1 states that $\frac{EL_\beta^I}{NT}$ is $O(T^{-1})$ so that $-[\frac{L_{\beta\beta}^I(\bar{\beta})}{NT}]^{-1}\frac{EL_\beta^I}{NT}$ is $O(T^{-1})$. We now show that $E([\frac{L_{\beta\beta}^I(\bar{\beta})}{NT}]^{-1}\frac{L_\beta^I - EL_\beta^I}{NT})$ is $O(1/(NT))$. Note that $\frac{L_\beta^I - EL_\beta^I}{NT}$ has expectation zero so that

$$E([\frac{L_{\beta\beta}^I(\bar{\beta})}{NT}]^{-1}\frac{L_\beta^I - EL_\beta^I}{NT}) = E\{([\frac{L_{\beta\beta}^I(\bar{\beta})}{NT}]^{-1} - E[\frac{L_{\beta\beta}^I(\bar{\beta})}{NT}]^{-1})\frac{L_\beta^I - EL_\beta^I}{NT}\}$$

where $([\frac{L_{\beta\beta}^I(\bar{\beta})}{NT}]^{-1} - E[\frac{L_{\beta\beta}^I(\bar{\beta})}{NT}]^{-1})$ and $\frac{L_\beta^I - EL_\beta^I}{NT}$ are $O(1/\sqrt{NT})$. Thus, $E(\hat{\beta} - \beta_0)$ is $O(T^{-1})$ for N being constant or increasing. *Q.E.D.*

Note that the intermediate value $\bar{\beta} \in [\hat{\beta}, \beta_0]$ converges to β_0 as $\hat{\beta}$ converges to β_0. Thus $\frac{1}{NT}L_{\beta\beta}^I(\bar{\beta})$ converges in probability to the Hessian, $\frac{1}{NT}EL_{\beta\beta}^I$.

Appendix 3. Theorem 2: *Let assumptions 1-2 hold. Then* $\hat{\beta} - \beta_0$ *is* $O_p\left(T^{-1}\right)$.

Proof: The proof is trivial for constant T. For increasing T, we prove the theorem in three steps.

Identification: Consider the following objective function $Q_0(\beta, \lambda) = EL(\beta, \lambda)$. The assumptions *(v),* $\{\beta, \lambda\} \neq \{\beta_0, \lambda_0\}$ *and* $\{\beta, \lambda\} \in \Theta$ *implies* $L(\beta, \lambda) \neq L(\beta_0, \lambda_0)$, and *(vi),* $E(|\ln L(\beta, \lambda)| < \infty)$, imply that $Q_0(\beta, \lambda)$ is uniquely maximized at $\{\beta_0, \lambda_0\}$.

Consistency: We first prove consistency under the assumption that the parameter space is compact: Let $\{\beta_0, \lambda_0\} \in \Theta$ which is compact. A Laplace approximation of L^I yields the following

$$\begin{aligned} L^{i,I}(\beta) &= L^i(\beta, \hat{\lambda}) + \ln(|L^i_{\lambda\lambda}(\hat{\lambda})|) + O_p(T^{-1}) \\ &= L^i(\beta, \hat{\lambda}) + o_p(T^{\gamma}) \text{ where } \gamma > 0 \end{aligned}$$

where γ is arbitrarily close to zero. Thus, the difference between $\frac{L^{i,I}(\beta)}{NT}$ and $\frac{L^i(\beta,\hat{\lambda})}{NT}$ is $o(T^{\gamma-1})$. The function $\frac{L^I(\beta)}{NT}$ is bounded and continuous over a compact set. Assumption *(vii)* implies that $\frac{L^i(\beta,\lambda)}{NT}$ is ergodic and stationair so that the assumptions of Newey and McFadden (1994, Lemma 2.4) are satisfied. Therefore, $\frac{L^I(\beta)}{NT}$ converges uniformly to $Q_0(\beta, \lambda) = EL(\beta, \lambda)$ and consistency follows from of Newey and McFadden (1994, Theorem 2.1).

Instead of assuming that the parameter space is compact we now assume that $\{\beta_0, \lambda_0\}$ is an element of the interior of a convex set Θ and $L^i(\beta, \lambda)$ is concave for all i. All the requirements of Newey and McFadden (1994, Theorem 2.7) are satisfied and consistency of the integrated likelihood estimator follows.

Note that the Laplace approximation of this proof differs from the Laplace approximations of $L^{i,I}_{\beta}(\beta)$ that are offered in the text. This reflects the fact that interpreting the integrated likelihood estimator as a solution to its first order conditions is more restrictive than necessary. The rates of convergence, however, are easier found by analyzing the first order conditions, $\frac{\sum_i L^{i,I}_{\beta}(\hat{\beta})}{NT} = 0$. Newey and McFadden (1994) maximize the objective functions to prove consistency and then analyze the first order conditions to derive the rate of convergence of the maximum likelihood and other estimator. A similar approach is followed here.

Rate of Convergence: Equation (1) gives an expression for $(\hat{\beta} - \beta_0)$. Assumption 2 ensures that the term that causes the bias in equation (1), $\frac{EL^I_\beta}{NT}$, is of the same or higher order than the term that induces the variance, $\frac{L^I_\beta - EL^I_\beta}{NT}$. Appendix 2 shows that $[\frac{L^I_{\beta\beta}(\bar{\beta})}{NT}]^{-1}$ is $O(1)$. Thus, $[\frac{L^I_{\beta\beta}(\bar{\beta})}{NT}]^{-1}\frac{EL^I_\beta}{NT}$ is $O_p(T^{-1})$. *Q.E.D.*

Appendix 4. Handbook Chapters

For linear models, we assume normality of the error terms in order to derive a likelihood. The resulting moment estimators are a function of the first two moments of the error term. That is, the only substantial restriction is that the first two moments are finite. In this methodology, robustness is proven by showing that the moment estimator is a consistent estimator for a class of models or likelihoods. Trimming of the error is applied in order to make the error normality distribution hold. In particular, trimming is applied to ensure that the zero mean condition holds. In this case, symmetry of the error distribution is a substantial condition along with the first two moments being finite. This yields the estimator of Honoré (1992) as well as subsequent papers that apply trimming, e.g. Kyriazidou (1997), Honoré and Kyriazidou (1999).

The models that are considered by Chamberlain(1984) are information-orthogonal or can be reparametrized such that they are information-orthogonal. The linear model and exponential model with exogenous regressors are information-orthogonal if we normalize the regressors to have mean zero, $\sum_t x_{it} = 0$, for all i. The information-orthogonal parametrization of the dynamic linear model with fixed effects is discussed in section 3. Conditioning on sufficient statistics is a special case of the integrated likelihood and discussed in section 6.2.

Arellano and Honoré (2001) review recent development in panel data econometrics. Besides the models mentioned above they also discuss models with a multiplicative nuisance parameter and the dynamic logit model. The multiplicative models are information-orthogonal after the transformations that are discussed in Arellano and Honoré (2001) section 2.4. and after normality of the error term is assumed. This yields a quasi integrated likelihood estimator that coincides with the moment estimators discussed in the handbook chapter. Honoré and Kyriazidou (2000) derive a very creative estimator for the logit model with exogenous re-

gressors and one lagged dependent variable. The parametrization is explicit and information-orthogonal if their matching technique is used. Appendix 5 gives an implicit parametrization for this model as well as other single index models. As we discuss in section 7, choosing a different $\frac{\partial f_i}{\partial \beta}$ for each period yields a common estimator for the linear model with predetermined variables.

Appendix 5. Information Orthogonality in the general nonlinear model

Consider the following classes of panel data models:

(i). Let the observations for a single agent be stochastically independent and dependent on unknowns only through $\mu_{it} = v_i + x_{it}\beta$, where v_i denotes an individual specific fixed effect and x_{it} a vector of exogenous regressors that can include lagged dependent variables. The panel Poisson, logit and probit models are of this single index form.

(ii). Let $y_{it} = G(x_{it}\beta + v_i) + \varepsilon_{it}$, where $\varepsilon_{it} \sim N(0, \sigma^2)$; $G()$ is unrestricted and v_i is a fixed effect and x_{it} a vector of exogenous regressors that can include lagged dependent variables.

(iii) As model (ii) but now we require $E\varepsilon_{it} = 0$, $E\varepsilon_{it}^2 < \infty$ and the distribution of ε_{it} to be known. This includes moving the average for ε_{it}, $t = 1, ..., T$. Given that some parameters of the error distribution are usually unknown, this class is merely of theoretical interest.

In this appendix, we show how information-orthogonality can be obtained for these classes of panel data models and discuss some generalizations of model (ii). The parameters β and λ_i are information-orthogonal if the following condition is satisfied:

$$E\frac{\partial^2 L(\beta, v(\beta, \lambda)}{\partial \lambda_i \partial \beta} = 0 \text{ at } \{\beta_0, \lambda_0\}$$

where L denotes the conditional log likelihood function (conditional on x) and λ_i the individual parameter in information-orthogonal reparametrization. The information matrix is evaluated at the true value; therefore, $E\frac{\partial^2 L}{\partial \lambda_i \partial \beta}$ has to hold at $\{\beta_0, \lambda_0\}$. We can rewrite $L(\beta, v_i)$ as $L(\mu_{i1}, .., \mu_{iT})$, where $\mu_{it} = x_{it}\beta + v_i$; then (we omit the subscript i) :

$$\frac{\partial L}{\partial \lambda} = \frac{\partial v}{\partial \lambda}\frac{\partial L}{\partial v} = \frac{\partial v}{\partial \lambda}\sum_t \frac{\partial L}{\partial \mu_t}\frac{\partial \mu_t}{\partial v} = \frac{\partial v}{\partial \lambda}\sum_t \frac{\partial L}{\partial \mu_t}.$$

We can rephrase the information-orthogonality condition as:

$$\begin{aligned} E\frac{\partial^2 L}{\partial\beta\partial\lambda} &= \frac{\partial v}{\partial\lambda}E\frac{\partial^2 L}{\partial\beta\partial v} = \frac{\partial v}{\partial\lambda}E\frac{\partial\{\sum_t \frac{\partial L}{\partial\mu_t}\}}{\partial\beta} \\ &= \frac{\partial v}{\partial\lambda}E\sum_t \frac{\partial\{\frac{\partial L}{\partial\mu_t}\}}{\partial\mu_t}\frac{\partial\mu_t}{\partial\beta} = \frac{\partial v}{\partial\lambda}E\{\sum_t L_{\mu_t\mu_t}\frac{\partial\mu_t}{\partial\beta}\} \\ &= 0. \end{aligned}$$

This gives

$$E\{\sum_t L_{\mu_t\mu_t}\frac{\partial\mu_t}{\partial\beta}\} = E\{\sum_t L_{\mu_t\mu_t}(x_t + \frac{\partial v}{\partial\beta})\} = 0,$$

which gives, omitting subscripts for μ :

$$\frac{\partial v}{\partial\beta} = -\frac{E\sum_t L_{\mu\mu}x_{it}}{E\sum_t L_{\mu\mu}} = -\frac{\sum_t x_{it}EL_{\mu\mu}}{\sum_t EL_{\mu\mu}}.$$

The solution for this differential equation is:

$$\lambda = \sum_t \int_{-\infty}^{v+x\beta} EL_{\mu\mu}d\mu.$$

This solution can be easily checked by total differentiation. We calculate (numerically) the integrated log likelihood:

$$L^I = \sum_i L_i^I = \sum_i \ln\int_0^{\bar\lambda} e^L d\lambda = \sum_i \ln\int_{-\infty}^{\infty} e^L\frac{\partial\lambda}{\partial v}dv,$$

where $\bar\lambda = \sum_t \int_{-\infty}^{\infty} EL_{\mu\mu}d\mu$.

If the regressor x_{it} is a $(K\times 1)$ vector then we want to (information) orthogonalize the fixed effects (incidental parameters) to all common parameters.

This gives us the following differential equations:

$$E\frac{\partial^2 L}{\partial\lambda_i\partial\beta_j} = 0 \text{ for } j = 1, ...K.$$

And the solution is similar to the above (but now x_{it} is a vector):

$$\lambda = \sum_t \int_{-\infty}^{v_i+x_{it}\beta} EL_{\mu\mu}d\mu,$$

and

$$\frac{\partial\lambda}{\partial v_i} = \sum_t EL_{\mu\mu}(v_i + x_{it}\beta).$$

For model (ii), we also need that σ^2 is information-orthogonal to λ :

$$\begin{aligned} L^{it}(\beta,\lambda,\sigma^2) &= -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y-g(\mu_{it}))^2 \\ \frac{\partial L^{it}(\beta,\lambda,\sigma^2)}{\partial\sigma^2} &= -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(y-g(\mu_{it}))^2 \\ \frac{\partial^2 L^{it}(\beta,\lambda,\sigma^2)}{\partial\sigma^2\partial\mu_{it}} &= \frac{1}{\sigma^2}\frac{\partial L^{it}(\beta,\lambda,\sigma^2)}{\partial\mu_{it}}. \end{aligned}$$

The score has expectation zero, i.e. $E\frac{\partial L^{it}(\beta,\lambda,\sigma^2)}{\partial\mu_{it}} = 0$. Therefore

$$E\frac{\partial^2 L^{it}(\beta,\lambda,\sigma^2)}{\partial\sigma^2\partial\mu_{it}} = \frac{1}{\sigma^2}E\frac{\partial^2 L^{it}(\beta,\lambda,\sigma^2)}{\partial\sigma^2\partial\mu_{it}} = 0.$$

The linear model with individual effects is a degenerated case of model (ii) and (iii):

$$y_{it} = \mu_{it} + \varepsilon_{it} = x_{it}\beta + v_i + \varepsilon_{it}$$

where x_{it} can include lagged values of y_{it}. Two remarks about the linear model:

(i) The error term enters additively in model II and III and therefore the log likelihood contribution of the s^{th} spell of individual i can be written in terms of ε_{it},i.e. $L^{it}(y_{it}-\mu_{it}) = L(\varepsilon_{it})$. The linearity assumption, $y_{it} = \mu_{it} + \varepsilon_{it}$ implies that all the derivatives with respect to μ can also be written as functions ε_{it}. This implies that $EL_{\mu\mu}$ is not a function of x_{it} and therefore $\frac{\partial f}{\partial\lambda} = EL_{\mu\mu} = c$ for any density function.

(ii) If the regressors x_{it} are exogenous then we have to deal with a somewhat strange modelling strategy: The assumptions of the model imply that changes of y_{it} are modelled as changes in x_{it} and, therefore, the change or differences regression should be applied without the introduction of individual effects. As shown in section 6, the change regression a special case of the integrated likelihood method.

Appendix 6. From $\hat{\lambda}$ to λ_0

Proof: Equation (3) in the text states that

$$L_\beta^{i,I} = \frac{\int L_\beta^i e^{L^i} d\lambda}{\int e^{L^i} d\lambda} = L_\beta^i(\hat{\lambda}) - \frac{1}{2}\frac{L_{\beta\lambda\lambda}^i(\hat{\lambda})}{L_{\lambda\lambda}^i(\hat{\lambda})} + \frac{1}{2}\frac{L_{\lambda\lambda\lambda}^i(\hat{\lambda})L_{\beta\lambda}^i(\hat{\lambda})}{\{L_{\lambda\lambda}^i(\hat{\lambda})\}^2} + O_p(T^{-1}).$$

We use the Laplace and Taylor expansions given in appendix 1 in order to derive an approximation whose remainder term is $O(T^{-1})$.

$$EL_\beta^{i,I} = E[L_\beta^i + L_{\beta\lambda}^i(\hat{\lambda}-\lambda) + \frac{1}{2}L_{\beta\lambda\lambda}^i(\hat{\lambda}-\lambda)^2 - \frac{1}{2}\frac{L_{\beta\lambda\lambda}^i}{L_{\lambda\lambda}^i} + \frac{1}{2}\frac{L_{\lambda\lambda\lambda}^i L_{\beta\lambda}^i}{\{L_{\lambda\lambda}^i\}^2}] + O(T^{-1}).$$

Using the expansions of appendix 1 yields

$$L_\lambda^i(\hat{\lambda}) = L_\lambda^i + (\hat{\lambda}-\lambda_0)L_{\lambda\lambda}^i + \frac{1}{2}(\hat{\lambda}-\lambda_0)^2 L_{\lambda\lambda\lambda}^i + \frac{1}{6}(\hat{\lambda}-\lambda_0)^3 L_{\lambda\lambda\lambda\lambda}^i + \frac{1}{12}(\hat{\lambda}-\lambda_0)^4 L_{\lambda\lambda\lambda\lambda\lambda}^i(\bar{\lambda}) = 0.$$

The following relations are helpful for approximating $EL_\beta^{i,I}$.

$$\begin{aligned}
(\hat{\lambda}-\lambda_0) &= -\frac{L_\lambda^i}{L_{\lambda\lambda}^i} - \frac{1}{2}\frac{L_{\lambda\lambda\lambda}^i}{L_{\lambda\lambda}^i}\{(\frac{L_\lambda^i}{L_{\lambda\lambda}^i})^2 + \frac{1}{2}\frac{L_{\lambda\lambda\lambda}^i}{L_{\lambda\lambda}^i}(\frac{L_\lambda^i}{L_{\lambda\lambda}^i})^3\} - \frac{1}{6}(\frac{L_\lambda^i}{L_{\lambda\lambda}^i})^3 L_{\lambda\lambda\lambda\lambda}^i + O_p(T^{-2}) \\
(\hat{\lambda}-\lambda_0)^2 &= (\frac{L_\lambda^i}{L_{\lambda\lambda}^i})^2 + \frac{L_{\lambda\lambda\lambda}^i}{L_{\lambda\lambda}^i}(\frac{L_\lambda^i}{L_{\lambda\lambda}^i})^3 + O_p(T^{-2}) \\
\frac{1}{L_{\lambda\lambda}^i} &= \frac{1}{EL_{\lambda\lambda}^i} + O_p(T^{-3/2}) \\
\frac{1}{(L_{\lambda\lambda}^i)^2} &= \frac{1}{(EL_{\lambda\lambda}^i)^2} + 2\frac{EL_{\lambda\lambda}^i - L_{\lambda\lambda}^i}{L_{\lambda\lambda}^i(EL_{\lambda\lambda}^i)^2} + O(T^{-3}) = \frac{1}{(EL_{\lambda\lambda}^i)^2} + O(T^{-3}) \\
\frac{1}{(L_{\lambda\lambda}^i)^3} &= \frac{1}{(EL_{\lambda\lambda}^i)^3} + O(T^{-4}).
\end{aligned}$$

Using these terms in the Taylor expansion for $L_\beta^{i,I}$ gives the following

$$\begin{aligned}
EL_\beta^{i,I} &= E[L_\beta^i - \frac{L_\lambda^i L_{\beta\lambda}^i}{L_{\lambda\lambda}^i} - \frac{1}{2}\frac{L_{\lambda\lambda\lambda}^i L_{\beta\lambda}^i}{L_{\lambda\lambda}^i}(\frac{L_\lambda^i}{L_{\lambda\lambda}^i})^2 + \frac{1}{2}(\frac{L_\lambda^i}{L_{\lambda\lambda}^i})^2 L_{\beta\lambda\lambda}^i - \frac{1}{2}\frac{L_{\beta\lambda\lambda}^i}{L_{\lambda\lambda}^i} + \frac{1}{2}\frac{L_{\lambda\lambda\lambda}^i L_{\beta\lambda}^i}{\{L_{\lambda\lambda}^i\}^2}] + O(T^{-1}) \\
&= E[L_\beta^i - \{\frac{L_{\beta\lambda}^i L_\lambda^i + L_{\beta\lambda\lambda}^i}{L_{\lambda\lambda}^i}\} + \frac{1}{2}\{\frac{L_{\lambda\lambda}^i + (L_\lambda^i)^2}{L_{\lambda\lambda}^i}\}\frac{L_{\beta\lambda\lambda}^i}{L_{\lambda\lambda}^i} + \frac{1}{2}\{\frac{L_{\lambda\lambda}^i - (L_\lambda^i)^2}{L_{\lambda\lambda}^i}\}\{\frac{L_{\lambda\lambda\lambda}^i L_{\beta\lambda}^i}{(L_{\lambda\lambda}^i)^2}\} + O(T^{-1}).
\end{aligned}$$

We complete the proof by showing that

$$E[\{\frac{L_{\lambda\lambda}^i + (L_\lambda^i)^2}{L_{\lambda\lambda}^i}\}\frac{L_{\beta\lambda\lambda}^i}{L_{\lambda\lambda}^i}] \text{ is } O(T^{-1}).$$

The proof of this last statement is slightly more general then necessary so that will be helpful for proving Lemma 2. Consider

$$E[\{\frac{L_{\lambda\lambda}^i + (L_\lambda^i)^2}{L_{\lambda\lambda}^i}\}\frac{L_{\beta\lambda\lambda}^i}{L_{\lambda\lambda}^i}] = E[\{\frac{L_{\lambda\lambda}^i + (L_\lambda^i)^2}{EL_{\lambda\lambda}^i}\}\{\frac{L_{\beta\lambda\lambda}^i - EL_{\beta\lambda\lambda}^i}{EL_{\lambda\lambda}^i}\}] + O(T^{-1})$$

since $E[\{L_{\lambda\lambda}^i + (L_\lambda^i)^2\} = 0$. Using $E\{L_{\beta\lambda\lambda}^i - EL_{\beta\lambda\lambda}^i\} = 0$ and $\frac{1}{L_{\lambda\lambda}^i} = \frac{1}{EL_{\lambda\lambda}^i} + O_p(T^{-3/2})$ gives

$$E[\{\frac{L_{\lambda\lambda}^i + (L_\lambda^i)^2}{L_{\lambda\lambda}^i}\}\frac{L_{\beta\lambda\lambda}^i}{L_{\lambda\lambda}^i}] = \frac{1}{\sqrt{T}}E[(\frac{T}{EL_{\lambda\lambda}^i})^2\frac{L_\lambda^i}{\sqrt{T}}\frac{L_\lambda^i}{\sqrt{T}}\frac{(L_{\beta\lambda\lambda}^i - EL_{\beta\lambda\lambda}^i)}{\sqrt{T}}] + O(T^{-1}).$$

The terms $\frac{L^i_\lambda}{\sqrt{T}}$ and $\frac{(L^i_{\beta\lambda}-EL^i_{\beta\lambda})}{\sqrt{T}}$ have expectation zero and their asymptotic distribution is normal with mean zero. It is trivial to show that the product of three normally distributed stochasts with mean zero has expectation zero. Consider $\eta \sim N(0,\Sigma)$ where Σ is a 3×3 matrix. Write η_1 as $\eta_1 = \rho_{12}\eta_2+\rho_{13}\eta_3+\varepsilon_1$ where ε_1 is uncorrelated with η_2 and η_3. Similarly, $\eta_2 = \rho_{23}\eta_3+\varepsilon_2$ where ε_2 is uncorrelated with η_3.This gives

$$\begin{aligned} E(\eta_1\eta_2\eta_3) &= E\{(\rho_{12}\eta_2+\rho_{13}\eta_3+\varepsilon_1)\eta_2\eta_3\} \\ &= E\{(\rho_{12}\eta_2+\rho_{13}\eta_3)\eta_2\eta_3\} \\ &= E\{(\rho_{12}\rho_{23}\eta_3+\rho_{13}\eta_3)\rho_{23}\eta_3\eta_3\} = 0. \end{aligned}$$

Thus,

$$\begin{aligned} \frac{(L^i_{\beta\lambda\lambda}-EL^i_{\beta\lambda\lambda})}{\sqrt{T}} &= \eta_1 + O_p(T^{-1/2}) \\ \frac{L^i_\lambda}{\sqrt{T}} &= \eta_2 + O_p(T^{-1/2}) = \eta_3 + O_p(T^{-1/2}) \end{aligned}$$

and the expectation of the product of these three terms is $O(T^{-1/2})$. We use a similar proof Lemma 2 and therefore ignore the fact that $\eta_2 = \eta_3$. Thus, $E[\{\frac{L^i_{\lambda\lambda}+(L^i_\lambda)^2}{EL^i_{\lambda\lambda}}\}\{\frac{L^i_{\beta\lambda\lambda}-EL^i_{\beta\lambda\lambda}}{EL^i_{\lambda\lambda}}\}]$ is $O(T^{-1})$. *Q.E.D.*

Appendix 7. Lemma 2

To be shown: *If* $EL^i_{\beta\lambda}=0$ *then* $E\{\frac{L^i_{\beta\lambda}L^i_\lambda+L^i_{\beta\lambda\lambda}}{L^i_{\lambda\lambda}}\}$ is $O(T^{-1})$.

Differentiating $EL^i_{\beta\lambda}=0$ with respect to λ gives $EL^i_{\beta\lambda\lambda}+EL^i_{\beta\lambda}L^i_\lambda=0$. Thus, $E\{\frac{L^i_{\beta\lambda}L^i_\lambda+L^i_{\beta\lambda\lambda}}{EL^i_{\lambda\lambda}}\} = 0$ and

$$\begin{aligned} E\{\frac{L^i_{\beta\lambda}L^i_\lambda+L^i_{\beta\lambda\lambda}}{L^i_{\lambda\lambda}}\} &= E\{\frac{L^i_{\beta\lambda}L^i_\lambda+L^i_{\beta\lambda\lambda}}{L^i_{\lambda\lambda}}\} - E\{\frac{L^i_{\beta\lambda}L^i_\lambda+L^i_{\beta\lambda\lambda}}{EL^i_{\lambda\lambda}}\} \\ &= E\{\frac{(L^i_{\beta\lambda}L^i_\lambda+L^i_{\beta\lambda\lambda})(EL^i_{\lambda\lambda}-L^i_{\lambda\lambda})}{L^i_{\lambda\lambda}EL^i_{\lambda\lambda}}\} \\ &= E\{\frac{(L^i_{\beta\lambda}L^i_\lambda+L^i_{\beta\lambda\lambda})(EL^i_{\lambda\lambda}-L^i_{\lambda\lambda})}{(EL^i_{\lambda\lambda})^2}\} \end{aligned}$$

since $\frac{1}{L^i_{\lambda\lambda}} = \frac{1}{EL^i_{\lambda\lambda}} + \frac{EL^i_{\lambda\lambda} - L^i_{\lambda\lambda}}{L^i_{\lambda\lambda} EL^i_{\lambda\lambda}} = \frac{1}{EL^i_{\lambda\lambda}} + O(T^{-3/2})$. This gives

$$\begin{aligned} E\{\frac{L^i_{\beta\lambda}L^i_{\lambda} + L^i_{\beta\lambda\lambda}}{L^i_{\lambda\lambda}}\} &= E\{\frac{(L^i_{\beta\lambda}L^i_{\lambda})(EL^i_{\lambda\lambda} - L^i_{\lambda\lambda})}{(EL^i_{\lambda\lambda})^2}\} + E\{\frac{(L^i_{\beta\lambda\lambda} - EL^i_{\beta\lambda\lambda})(EL^i_{\lambda\lambda} - L^i_{\lambda\lambda})}{(EL^i_{\lambda\lambda})^2}\} \\ &= \frac{1}{\sqrt{T}} E\{\frac{T}{EL^i_{\lambda\lambda}}\frac{T}{EL^i_{\lambda\lambda}}\frac{L^i_{\beta\lambda}}{\sqrt{T}}\frac{L^i_{\lambda}}{\sqrt{T}}\frac{EL^i_{\lambda\lambda} - L^i_{\lambda\lambda}}{\sqrt{T}}\} + O(T^{-1}). \end{aligned}$$

The terms $\frac{L^i_{\beta\lambda}}{\sqrt{T}}$, $\frac{L^i_{\lambda}}{\sqrt{T}}$, and $\frac{(EL^i_{\lambda\lambda} - L^i_{\lambda\lambda})}{\sqrt{T}}$ have expectation zero and their asymptotic distribution is normal with mean zero. As shown above, the product of three normally distributed stochasts with mean zero has expectation zero. *Q.E.D.*

Appendix 8. Lemma 3

To be shown: *If* $EL^i_{\beta\lambda} = 0$ *then* $E[\{\frac{L^i_{\lambda\lambda} - (L^i_{\lambda})^2}{L^i_{\lambda\lambda}}\}\{\frac{L^i_{\lambda\lambda\lambda}L^i_{\beta\lambda}}{(L^i_{\lambda\lambda})^2}\}]$ *is* $O(T^{-1})$.

Obviously,

$$\begin{aligned} E[\{\frac{L^i_{\lambda\lambda} - (L^i_{\lambda})^2}{L^i_{\lambda\lambda}}\}\{\frac{L^i_{\lambda\lambda\lambda}L^i_{\beta\lambda}}{(L^i_{\lambda\lambda})^2}\}] &= E[\{\frac{L^i_{\lambda\lambda\lambda}L^i_{\beta\lambda}}{(EL^i_{\lambda\lambda})^2}\}] - E[\{\frac{(L^i_{\lambda})^2}{EL^i_{\lambda\lambda}}\}\{\frac{L^i_{\lambda\lambda\lambda}L^i_{\beta\lambda}}{(EL^i_{\lambda\lambda})^2}\}] + O(T^{-1}) \\ &= E[\{\frac{L^i_{\beta\lambda}(L^i_{\lambda\lambda\lambda} - EL^i_{\lambda\lambda\lambda})}{(EL^i_{\lambda\lambda})^2}\}] - E[\{\frac{(L^i_{\lambda})^2}{EL^i_{\lambda\lambda}}\}\{\frac{L^i_{\lambda\lambda\lambda}L^i_{\beta\lambda}}{(EL^i_{\lambda\lambda})^2}\}] + O(T^{-1}) \\ &= -\frac{1}{\sqrt{T}} E\{(\frac{T}{EL^i_{\lambda\lambda}})^3 \frac{L^i_{\lambda}}{\sqrt{T}}\frac{L^i_{\lambda}}{\sqrt{T}}\frac{L^i_{\beta\lambda}}{\sqrt{T}}\frac{EL^i_{\lambda\lambda\lambda}}{T}\} + O(T^{-1}) \end{aligned}$$

since $L^i_{\beta\lambda}$ and $(L^i_{\lambda\lambda\lambda} - EL^i_{\lambda\lambda\lambda})$ are $O(\sqrt{T})$, and $\frac{L^i_{\lambda}}{\sqrt{T}}$ and $\frac{L^i_{\beta\lambda}}{\sqrt{T}}$ have an asymptotic distribution that is normal with mean zero. *Q.E.D.*

Appendix 9. Theorem 3

Let assumptions 1-3 hold. Then $E\hat{\beta} - \beta_0$ *is* $O\left(T^{-2}\right)$.

Proof: The proof follows the proof of theorem 1 in appendix 2. Equation (1) in the text states that

$$(\hat{\beta} - \beta_0) = -[\frac{L^I_{\beta\beta}(\bar{\beta})}{NT}]^{-1}\frac{EL^I_{\beta}}{NT} - [\frac{L^I_{\beta\beta}(\bar{\beta})}{NT}]^{-1}\frac{L^I_{\beta} - EL^I_{\beta}}{NT}.$$

It follows from Lemma 3 and 4 that $\frac{1}{NT}EL^I_{\beta}$ is $O\left(T^{-2}\right)$. Thus, $E[\frac{L^I_{\beta\beta}(\bar{\beta})}{NT}]^{-1}\frac{EL^I_{\beta}}{NT}]$ is $O\left(T^{-2}\right)$.

$$E([\frac{L^I_{\beta\beta}(\bar{\beta})}{NT}]^{-1}\frac{L^I_{\beta} - EL^I_{\beta}}{NT}) = E\{([\frac{L^I_{\beta\beta}(\bar{\beta})}{NT}]^{-1} - E[\frac{L^I_{\beta\beta}(\bar{\beta})}{NT}]^{-1})\frac{L^I_{\beta} - EL^I_{\beta}}{NT}\}$$

where $(\frac{L^I_{\beta\beta}(\bar{\beta})}{NT}]^{-1} - E[\frac{L^I_{\beta\beta}(\bar{\beta})}{NT}]^{-1})$ and $\frac{L^I_\beta - EL^I_\beta}{NT}$ are $O(1/\sqrt{NT})$ so that $E([\frac{L^I_{\beta\beta}(\bar{\beta})}{NT}]^{-1}\frac{L^I_\beta - EL^I_\beta}{NT})$ is $O(1/(NT))$. Assumption 3 ensures $O(1/(NT))$ is $O(T^{-2})$ and the result follows. *Q.E.D.*

Appendix 10. Theorem 4

Let assumptions 1, 3 and 4 hold. Then $\hat{\beta} - \beta_0$ *is* $O_p\left(T^{-2}\right)$.

Proof: The proof closely follows the proof of theorem 2 in appendix 3, the only difference being that the bias, $E\hat{\beta} - \beta_0$, is of a lower order. The order of the variance is still the inverse of the number of observations. Thus.

$$\hat{\beta} - \beta_0 = \{\hat{\beta} - E(\hat{\beta})\} + \{E\hat{\beta} - \beta_0\} = O_p(1/\sqrt{NT}) + O\left(T^{-2}\right).$$

Assumption 4 states that $T \propto N^\alpha$ *where* $\alpha \leq \frac{1}{3}$. So that $\hat{\beta} - \beta_0$ *is* $O_p\left(T^{-2}\right)$. *Q.E.D.*

Appendix 11.

To be shown

$$\frac{1}{NT}EL^I_{\beta\beta} = -\frac{1}{NT}EL^I_\beta L^{I\prime}_\beta + o(1/\sqrt{T}).$$

By definition,

$$\begin{aligned} L^{i,I}_{\beta\beta} &= \frac{\partial}{\partial\beta}\sum_i \frac{\int L^i_\beta e^{L^i} d\lambda}{\int e^{L^i} d\lambda} \\ &= \sum_i \frac{\int\{L^i_{\beta\beta} + L^i_\beta L^{i\prime}_\beta\} e^{L^i} d\lambda}{\int e^{L^i} d\lambda} - \sum_i (L^{i,I}_\beta)(L^{i,I}_\beta)'. \end{aligned}$$

A Laplace approximation gives

$$\begin{aligned} \sum_i \{\frac{\int\{L^i_{\beta\beta} + L^i_\beta L^{i\prime}_\beta\} e^{L^i} d\lambda}{\int e^{L^i} d\lambda}\} &= \sum_i \{L^i_{\beta\beta}(\hat{\lambda}) + L^i_\beta(\hat{\lambda})^2 + O_p(1)\} \\ &= \sum_i \{L^i_{\beta\beta} + L^i_\beta L^{i\prime}_\beta\} + O_p(N) = L_{\beta\beta} + L^2_\beta + O_p(N). \end{aligned}$$

The information inequality states that $E(L_{\beta\beta} + L^2_\beta) = 0$. Therefore, $(L_{\beta\beta} + L^2_\beta)$ is $O_p(\sqrt{NT})$. This yields

$$\frac{1}{NT}EL^I_{\beta\beta} = -\frac{1}{NT}EL^I_\beta L^{I\prime}_\beta + o_p(T^{-1/2}).$$

Appendix 12.

Using a central limit theorem and the delta method yields that $\sqrt{NT}(\hat{\beta}-\beta_0)$ is asymptotically normally distributed,

$$\sqrt{NT}(\hat{\beta}-\beta_0) \rightarrow N(0,\Psi).$$

As shown in the text and appendix 11,

$$\Psi = \frac{1}{NT} E\{L_\beta L_\beta{}'\}.$$

Q.E.D.

Appendix 13. Unit Root

To be shown: $T\sqrt{N}(\hat{\beta}-\beta) = O_p(1)$ if $\beta = 1$.

$$L^i = -\frac{1}{2}\sum_t (\tilde{y}_t - \tilde{y}_{t-1}\beta - \lambda e^{-b(\beta)})^2 \text{ where } \tilde{y}_t = y_t - y_0.$$

Integrating the likelihood contribution with respect to λ gives the integrated likelihood contribution. Note that $\frac{\partial \lambda}{\partial f} = e^{b(\beta)}$ does *not* depend on f.

$$\begin{aligned}
e^{L^{i,I}} &= \int e^{-\frac{1}{2}\sum_t (\tilde{y}_t - \tilde{y}_{t-1}\beta - \lambda e^{-b(\beta)})^2} d\lambda \\
&= e^{b(\beta)} \int e^{-\frac{1}{2}\sum_t (y_t - y_{t-1}\beta - f)^2} df \\
&= e^{b(\beta) - \frac{1}{2}\sum_t (y_t - y_{t-1}\beta)^2} \int e^{-\frac{T}{2}\{f^2 - 2f(\overline{y_t - y_{t-1}\beta})\}} df \\
&\propto e^{b(\beta) - \frac{1}{2}\sum_t (y_t - y_{t-1}\beta)^2 + \frac{T}{2}(\overline{y_t - y_{t-1}\beta})^2}.
\end{aligned}$$

where we omit the subscript i. Differentiating with respect to β yields

$$\begin{aligned}
L_\beta^{i,I} &= b'(\beta) + \sum_t (y_t - y_{t-1}\beta) y_{t-1} - T(\overline{y_t - y_{t-1}\beta})\overline{y_{t-1}} \\
L_{\beta\beta}^{i,I} &= b''(\beta) - \sum_t y_{t-1}^2 + T\,\overline{y_{t-1}}^2
\end{aligned}$$

where $b(\beta) = \frac{1}{T}\sum_{t=1}^T \frac{T-t}{t}\beta^t$, $b(\beta)' = \frac{1}{T}\sum_{t=1}^T (T-t)\beta^{t-1}$, $b(\beta)'' = \frac{1}{T}\sum_{t=1}^T (T-t)(t-1)\beta^{t-2}$. In particular, if $\beta = 1$ then

$$b'(\beta = 1) = \frac{1}{T}\sum_{t=1}^T (T-t) = T - \frac{1}{T}\sum_{t=1}^T t = \frac{T-1}{2}.$$

Thus $L^{i,I}_{\beta}$ is $O_p(T)$ and L^{I}_{β} is $O_p(T\sqrt{N})$. The second derivative $L^{i,I}_{\beta\beta}$ is $O_p(T^2)$ and $L^{I}_{\beta\beta}$ is $O_p(T^2N)$ and the result follows.

Appendix 14. Theorem 6

To be shown

$$\begin{aligned}\sqrt{N}(\hat{\beta}-\beta_0) &\rightarrow N(0,\Psi) \text{ where} \\ \Psi &= [\frac{1}{NT}EL^I_{\beta\beta}]^{-1}[\frac{1}{NT}E\{(L^I_\beta)(L^I_\beta)'\}][\frac{1}{NT}EL^I_{\beta\beta}]^{-1}.\end{aligned}$$

Proof: The assumption $L_{\beta\lambda\lambda}=0$ implies that $L_{\beta\lambda}$ is not a function of λ so that the Taylor expansion around λ_0 has the following form:

$$L^I_\beta = \frac{\int L_\beta e^L d\lambda}{\int e^L d\lambda} = E\frac{\int\{L_\beta(\lambda_0)+(\lambda-\lambda_0)L_{\beta\lambda}\}e^L d\lambda}{\int e^L d\lambda}.$$

This yields

$$EL^I_\beta(\beta_0) = EL_\beta(\beta_0,\lambda_0) + EL_{\beta\lambda}(\beta_0)\frac{\int(\lambda-\lambda_0)e^L d\lambda}{\int e^L d\lambda} = 0$$

since the posterior mean $\tilde{\lambda} = \frac{\int \lambda e^L d\lambda}{\int e^L d\lambda}$ is uncorrelated with $L_{\beta\lambda}$. It was assumed that the solution to $EL^I_\beta(\beta)=0$ be unique. Thus, all the conditions for the method of moment estimator are satisfied and consistency follows from Newey and McFadden (1994, Theorem 2.6). The variance-covariance matrix and the asymptotic normality follow from Newey and McFadden (1994, Theorem 3.4). *Q.E.D.* Note that condition (vii) of assumption 1 is not neccesary for theorem 6.

Appendix 15. Mixing Distributions

The Laplace approximation in the text states that

$$L^{i,I}_\beta = L^i_\beta(\hat{\lambda}) - \frac{M_\lambda(\hat{\lambda})L^i_\lambda(\hat{\lambda})}{L^i_{\lambda\lambda}(\hat{\lambda})} - \frac{1}{2}\frac{L^i_{\beta\lambda\lambda}(\hat{\lambda})}{L^i_{\lambda\lambda}(\hat{\lambda})+M_{\lambda\lambda}(\hat{\lambda})} + \frac{1}{2}\frac{L^i_{\beta\lambda}(\hat{\lambda})\{L^i_{\lambda\lambda\lambda}(\hat{\lambda})+M_{\lambda\lambda\lambda}(\hat{\lambda})\}}{\{L^i_{\lambda\lambda}(\hat{\lambda})+M_{\lambda\lambda}(\hat{\lambda})\}^2} + O_p(T^{-1}).$$

By assumption, the prior and mixing distribution cannot be a function of β. If this dependence would not have been ruled out then the Laplace approximation would have one more term, $M_\beta(\hat{\lambda})$. This term is $O(1)$ so that the Laplace approximation would *not* be the same, up to

$O_p(T^{-1})$, as in the previous section. Note that $M_{\lambda\lambda}$ is $O_p(1)$ and therefore

$$\begin{aligned}\frac{1}{L^i_{\lambda\lambda}(\hat{\lambda})+M_{\lambda\lambda}(\hat{\lambda})} &= \frac{1}{L^i_{\lambda\lambda}(\hat{\lambda})}-\frac{M_{\lambda\lambda}}{L^i_{\lambda\lambda}(\hat{\lambda})(L^i_{\lambda\lambda}(\hat{\lambda})+M_{\lambda\lambda}(\hat{\lambda}))}=\frac{1}{L^i_{\lambda\lambda}}+O_p(T^{-2})\\ \frac{L^i_{\lambda\lambda\lambda}(\hat{\lambda})+M_{\lambda\lambda\lambda}(\hat{\lambda})}{\{L^i_{\lambda\lambda}(\hat{\lambda})+M_{\lambda\lambda}(\hat{\lambda})\}^2} &= \frac{L^i_{\lambda\lambda\lambda}}{{L^i_{\lambda\lambda}}^2}+O_p(T^{-2}).\end{aligned}$$

Thus,

$$L^{i,I}_\beta = L^i_\beta(\hat{\lambda})-\frac{M_\lambda(\hat{\lambda})L^i_\lambda(\hat{\lambda})}{L^i_{\lambda\lambda}(\hat{\lambda})}-\frac{1}{2}\frac{L^i_{\beta\lambda\lambda}(\hat{\lambda})}{L^i_{\lambda\lambda}(\hat{\lambda})+M_{\lambda\lambda}(\hat{\lambda})}+\frac{1}{2}\frac{L^i_{\beta\lambda}(\hat{\lambda})\{L^i_{\lambda\lambda\lambda}(\hat{\lambda})+M_{\lambda\lambda\lambda}(\hat{\lambda})\}}{\{L^i_{\lambda\lambda}(\hat{\lambda})+M_{\lambda\lambda}(\hat{\lambda})\}^2}+O_p(T^{-1}).$$

Appendix 16. Gamma Distribution

The density has the following form

$$f(y_{i1},...y_{iT}|\alpha,f_i)=\prod_t \frac{f_i^\alpha y_{it}^{\alpha-1}e^{-f_i y_{it}}}{\Gamma(\alpha)} \text{ for } i=1,...,N.$$

This yields,

$$\begin{aligned}L &= T\alpha\ln f_i+(\alpha-1)\sum_t\ln(y_{it})-f_i\sum_t y_{it}+T\ln\Gamma(\alpha),\\ L_f &= \frac{T\alpha}{f_i}-\sum_t y_{it},\ L_{ff}=-\frac{T\alpha}{f_i^2},\text{ and } L_{f\alpha}=\frac{T}{f_i}.\end{aligned}$$

Thus, $EL_{f\alpha}=\frac{T}{f_i}\neq 0$. We can interpret f_i as a function of α and the information-orthogonal nuisance parameter λ_i. Equation (4) states the following differential equation.

$$EL_{f\beta}+EL_{ff}\frac{\partial f}{\partial\beta}=0.$$

In this case,

$$\begin{aligned}\frac{T}{f_i}-\frac{T\alpha}{f_i^2}\frac{\partial f}{\partial\alpha} &= 0,\\ \frac{\partial f_i}{\partial\alpha} &= \frac{f_i}{\alpha}.\end{aligned}$$

A solution is $f_i(\alpha,\lambda)=a\lambda$. This implies the following log likelihood,

$$L^i(\alpha,\lambda_i)=T\alpha\ln\alpha+T\alpha\ln\lambda_i+(\alpha-1)\sum_t\ln y_{it}-\alpha\lambda_i\sum_t y_{it}-T\ln\Gamma(\alpha).$$

Differentiating with respect to α and λ_i gives

$$L^i_{\alpha\lambda_i}(\alpha,\lambda_i) = \frac{T}{\lambda_i} - \sum_t y_{it},$$

and the parametrization is information-orthogonal,

$$EL^i_{\alpha\lambda_i}(\alpha_0,\lambda_{0,i}) = \frac{T}{\lambda_{0,i}} - \sum_t Et_{it} = 0.$$

Appendix 17.

To be shown

$$\alpha \to_p \alpha \text{ for } N \to \infty \text{ or } T \to \infty.$$

The moment function has the following expectation,

$$EL^I_\alpha = \sum_t E\ln y_{it} + T\psi(T\alpha+1) - \frac{1}{\alpha} - T\psi(\alpha) - TE\ln(\sum_t y_{it}).$$

Note that

$$\begin{aligned} y_{it} &\sim Gamma(\alpha,\alpha\lambda_i) \text{ and therefore} \\ E\ln y_{it} &= -\ln(\alpha) - \ln\lambda_i + \psi(\alpha), \end{aligned}$$

see Lancaster (1990), appendix 1. Similarly,

$$\begin{aligned} \sum_t y_{it} &\sim Gamma(T\alpha,\alpha\lambda_i) \text{ and therefore} \\ E\ln\sum_t y_{it} &= -\ln(\alpha) - \ln\lambda_i + \psi(T\alpha). \end{aligned}$$

Thus,

$$E\sum_t \ln y_{it} - ET\ln\sum_t y_{it} = T\psi(\alpha_0) - T\psi(T\alpha_0).$$

This gives

$$\begin{aligned} EL^I_\alpha(\alpha) &= E\sum_t\{\ln y_{it}\} - TE\{\ln(\sum_t y_{it})\} + T\psi(T\alpha+1) - \frac{1}{\alpha} - T\psi(\alpha) \\ &= T\psi(\alpha_0) - T\psi(T\alpha_0) - T\psi(\alpha) + T\psi(T\alpha+1) - \frac{1}{\alpha}. \end{aligned}$$

Note that $\psi(T\alpha+1)=\psi(T\alpha)+\frac{1}{T\alpha}$ and therefore

$$\begin{aligned} EL^I_\alpha(\alpha) &= T\psi(\alpha_0)-T\psi(\alpha)-T\psi(T\alpha_0)+T\psi(T\alpha) \\ EL^I_{\alpha\alpha}(\alpha) &= -T\psi'(\alpha)+T^2\psi'(T\alpha), \end{aligned}$$

which is negative, see Lancaster (1990). Thus $EL^I_\alpha(\alpha)=0$ is uniquely solved for $\alpha=\alpha_0$. Consistency follows from Newey and McFadden (1994, Theorem 2.6).

We concentrate the likelihood by replacing λ_i by its maximum likelihood estimator.

$$L^i_\lambda(\alpha,\lambda_i)=\frac{T\alpha}{\lambda_i}-\alpha\sum_t y_{it},.$$

Thus

$$\hat\lambda_i=\frac{T}{\sum_t y_{it}}=\frac{1}{\bar t_i}.$$

Differentiating the concentrated log likelihood with respect to α gives

$$\begin{aligned} L^i_\alpha(\alpha,\hat\lambda_i) &= T+T\ln\alpha+T\ln\hat\lambda_i+\sum_t \ln y_{it}-\hat\lambda_i\sum_t y_{it}-T\psi(\alpha) \\ &= T\ln\alpha-T\ln\sum_t y_{it}+T\ln T+\sum_t \ln y_{it}-T\psi(\alpha). \end{aligned}$$

Note that

$$E\sum_t \ln y_{it}-ET\ln\sum_t y_{it}=T\psi(\alpha)-T\psi(T\alpha),$$

see Lancaster (1990, appendix 1). This gives

$$EL^i_\alpha(\alpha,\hat\lambda_i)=T\ln\alpha+T\ln T-T\psi(\alpha)$$

which is $O(1)$. The second derivative has the following form,

$$L^i_{\alpha\alpha}(\alpha,\lambda_i)=\frac{T}{\alpha}-T\psi'(\alpha).$$

Besides the incidental parameter bias, this maximum likelihood estimator also has the small sample bias of $O((NT)^{-1})$. We assume that N is increasing so that we are left with the incidental parameter bias,

$$\begin{aligned} E\hat\alpha-\alpha &= \frac{EL^i_\alpha(\alpha,\hat\lambda_i)}{\frac{T}{\alpha}-T\psi'(\alpha)}+o(T^{-1}) \\ &= \frac{T\ln\alpha+T\ln T-T\psi(\alpha)}{\frac{T}{\alpha}-T\psi'(\alpha)} \text{ is } O(T^{-1}). \end{aligned}$$

Appendix 18. Neyman and Scott Example:

$$\mathfrak{L}^{i,I} = \int \mathfrak{L}^i \, d\lambda_i \propto \int \sigma^{-T} \exp\{-\frac{1}{2}(\frac{\sum_t (y_{it}-\bar{y}_i)^2}{\sigma^2} + \frac{T(\bar{y}_i-\lambda_i)^2}{\sigma^2})\} d\lambda_i.$$

Note that $\frac{\sum_t (y_{it}-\bar{y}_i)^2}{\sigma^2}$ does not depend on λ_i and therefore

$$\mathfrak{L}^I \propto \sigma^{-T} \exp(-\frac{\sum_t (y_{it}-\bar{y}_i)^2}{\sigma^2}) * \sigma/\sqrt{T} \int \frac{1}{\sigma/\sqrt{T}} \exp\{-\frac{1}{2}\frac{T(\bar{y}_i-\lambda_i)^2}{\sigma^2}\} d\lambda_i.$$

Note that $\int \frac{1}{\sigma/\sqrt{T}} \exp\{-\frac{1}{2}\frac{T(\bar{y}_i-\lambda_i)^2}{\sigma^2}\} d\lambda_i = 1$ and therefore

$$\mathfrak{L}^I \propto \sigma^{-(T-1)} \exp(-\frac{\sum_t (y_{it}-\bar{y}_i)^2}{\sigma^2}).$$

Appendix 19.

Local Identification. Consider the derivative of the moment function $Q^{i,I}_\beta(\beta)$ at $\beta = \beta_0$,

$$\begin{aligned} Q^{i,I}_{\beta\beta} &= \frac{\partial}{\partial\beta} \sum_i \frac{\int Q^i_\beta e^{L^i} d\lambda}{\int e^{L^i} d\lambda} \\ &= \sum_i \frac{\int \{Q^i_{\beta\beta} + Q^i_\beta L^{i\,\prime}_\beta\} e^{L^i} d\lambda}{\int e^{L^i} d\lambda} + \sum_i (Q^{i,I}_\beta)(L^{i,I}_\beta)'. \end{aligned}$$

A Laplace approximation and the following equation

$$EQ^i_{\beta\beta}(\beta) = \sum_{t=1}^{T-1} \{x^2_{it} EL_{\mu_{it}\mu_{it}} - x_{it}x_{i,t+1} EL_{\mu_{i,t}\mu_{i,t}}\} = -EQ^i_\beta L^i_\beta$$

imply that $\sum_i \{\frac{\int \{L^i_{\beta\beta} + L^i_\beta L^{i\,\prime}_\beta\} e^{L^i} d\lambda}{\int e^{L^i} d\lambda}\}$ is $o(NT)$. This yields

$$\begin{aligned} \frac{1}{NT} EQ^I_{\beta\beta} &= -\frac{1}{NT} EQ^I_\beta L^{I\,\prime}_\beta + o(1) \\ &= -\frac{1}{NT} EQ_\beta L_\beta{}' + o(1) \end{aligned}$$

where we assume that $T \propto N^\alpha$ where $\alpha > \frac{1}{3}$. Suppose x_{is} is a vector, then

$$\begin{aligned} Q^i_\beta(\beta) &= \sum_{t=1}^{T-1} \{x_{it}\{L_{\mu_{it}} - (EL_{\mu_{it}\mu_{it}}|x_{i,},\beta,\lambda_i)(EL_{\mu_{i,t+1}\mu_{i,t+1}}|x_{i,t+1},\beta,\lambda_i)^{-1} L_{\mu_{t+1}}\} \\ EQ_\beta L_\beta{}' &= -Q^i_{\beta\beta}(\beta) \\ &= -\sum_{t=1}^{T-1} \{x_{it}x'_{it} EL_{\mu_{it}\mu_{it}} - x_{it}x'_{i,t+1} EL_{\mu_{t+1}\mu_{t+1}}\} \\ &= -\sum_{t=2}^{T-1} (x_{it}x'_{it} - x_{i,t-1}x'_{it}) EL_{\mu_{it}\mu_{it}} - x_{i1}x'_{i1} EL_{\mu_{i1}\mu_{i1}} + x_{i,T-1}x'_{iT} EL_{\mu_{iT}\mu_{iT}}. \end{aligned}$$

Thus, local identification is ensured when $\frac{1}{NT}EQ^I_{\beta\beta}$ or the asymptotically equivalent matrix $EQ_\beta L_\beta{}'$ is positive definite.

Global Identification.

The following Laplace approximation shows that, for identification, we only need to evaluate $Q^i_\beta(\beta,\hat\lambda)$ where $\hat\lambda$ is the maximum likelihood estimate of λ for given β.

$$\begin{aligned} Q^{i,I}_\beta(\beta) &= \frac{\int Q^i_\beta(\beta,\lambda)e^{L^i(\beta,\lambda)}d\lambda}{\int e^{L^i(\beta,\lambda)}d\lambda} = Q^i_\beta(\beta,\hat\lambda)+\frac{1}{2}\frac{Q^i_{\beta\lambda\lambda}(\beta,\hat\lambda)}{L^i_{\lambda\lambda}(\beta,\hat\lambda)}+\frac{1}{2}\frac{L^i_{\lambda\lambda\lambda}(\beta,\hat\lambda)Q^i_{\beta\lambda}(\beta,\hat\lambda)}{\{L^i_{\lambda\lambda}(\beta,\hat\lambda)\}^2}+O_p(T^{-1}) \\ &= Q^i_\beta(\beta,\hat\lambda)+O_p(1). \end{aligned}$$

Honoré and Hu (1999) ensure global identification by using moments that converge at a rate slower than the inverse of root of the number of observations. These relatively slow converging moments shrink the parameter space that needs to be considered so that only local identification needs to be proven. One can apply the same technique here. In particular, the maximum likelihood estimator converges at a slow rate, $\hat\beta_{ML}-\beta=O_p((NT)^{-1/2})+O(T^{-1})$, where assuming 1 ensures global identification. Analogue to Honoré and Hu (1999), one can require the integrated moment estimator $\hat\beta$ to be in a shrinking neighborhood of $\hat\beta_{ML}$. This ensures global identification under the condition of local identification of the integrated moment estimator and global identification of the maximum likelihood estimator.

Appendix 20. Predetermined Variables

To be shown

$$\begin{aligned} \sqrt{NT}(\hat\beta-\beta_0) &\to N(0,\Psi) \text{ where} \\ \hat\beta &= \arg\min_\beta\{Q^I_\beta(\beta)'Q^I_\beta(\beta)\}. \end{aligned}$$

Information orthogonality of the moment function implies that $Q^i_{\beta\lambda}$ is $O_p(\sqrt{T})$ and $EQ^i_{\beta\lambda}+EQ^i_{\beta\lambda\lambda}=0$. To evaluate the integral $Q^I_\beta=\sum_i\frac{\int Q^i_\beta(\beta_0,\lambda)e^{L^i(\beta_0,\lambda)}d\lambda}{\int e^{L^i(\beta_0,\lambda)}d\lambda}$ we use the same Taylor approximation as for the integrated likelihood. This gives

$$Q^{i,I}_\beta=Q^i_\beta+O_p(1).$$

Analogue to Lemma 2 and 3, information-orthogonality implies that

$$E\{\frac{Q^i_{\beta\lambda}L^i_\lambda + Q^i_{\beta\lambda\lambda}}{L^i_{\lambda\lambda}}\} \text{ is } O(T^{-1}) \text{ and}$$
$$E[\{\frac{L^i_{\lambda\lambda} - (L^i_\lambda)^2}{L^i_{\lambda\lambda}}\}\{\frac{L^i_{\lambda\lambda\lambda}Q^i_{\beta\lambda}}{(L^i_{\lambda\lambda})^2}\}] \text{ is } O(T^{-1}).$$

This gives

$$\begin{aligned} EQ^{i,I}_\beta &= O(T^{-1}), \\ Q^{i,I}_\beta &= Q^i_\beta + O_p(1) \text{ and} \\ \frac{Q^I_\beta}{\sqrt{NT}} &= \frac{Q_\beta}{\sqrt{NT}} + O(\sqrt{\frac{N}{T^3}}) + O_p(T^{-1/2}). \end{aligned}$$

Thus

$$\begin{aligned} \frac{1}{NT}(Q^I_\beta)(Q^I_\beta)' &= \frac{1}{NT}(Q_\beta)(Q_\beta)' + o(1) \\ &= \frac{1}{NT}E\{(Q_\beta)(Q_\beta)'\} + o(1). \end{aligned}$$

Consider the derivative of the moment function $Q^{i,I}_\beta(\beta)$.

$$\begin{aligned} Q^{i,I}_{\beta\beta} &= \frac{\partial}{\partial\beta}\sum_i \frac{\int Q^i_\beta e^{L^i}d\lambda}{\int e^{L^i}d\lambda} \\ &= \sum_i \frac{\int\{Q^i_{\beta\beta} + Q^i_\beta L^{i\prime}_\beta\}e^{L^i}d\lambda}{\int e^{L^i}d\lambda} + \sum_i (Q^{i,I}_\beta)(L^{i,I}_\beta)'. \end{aligned}$$

A Laplace approximation and the following equation

$$EQ^i_{\beta\beta}(\beta) = \sum_{t=1}^{T-1}\{x^2_{it}EL_{\mu_{it}\mu_{it}} - x_{it}x_{i,t+1}EL_{\mu_{i,t}\mu_{i,t}}\} = -EQ^i_\beta L^i_\beta$$

imply that $\sum_i\{\frac{\int\{L^i_{\beta\beta}+L^i_\beta L^{i\prime}_\beta\}e^{L^i}d\lambda}{\int e^{L^i}d\lambda}\}$ is $o(NT)$. This yields

$$\begin{aligned} \frac{1}{NT}EQ^I_{\beta\beta} &= -\frac{1}{NT}EQ^I_\beta L^{I\prime}_\beta + o(1) \\ &= -\frac{1}{NT}EQ_\beta L_\beta{}' + o(1) \end{aligned}$$

where our assumption $T \propto N^{\alpha}$ where $\alpha > \frac{1}{3}$ is crucial for the last step. The regularity conditions of assumption 1 and the algebra above implies that

$$\sqrt{NT}(\hat{\beta} - \beta_0) \rightarrow N(0, \Psi)$$

where

$$\Psi = [\frac{1}{NT} E Q_\beta L_\beta]^{-1} [\frac{1}{NT} E Q_\beta Q_\beta{}'] [\frac{1}{NT} E Q_\beta L_\beta]^{-1}.$$

Q.E.D.

10 Reference

Abrevaya, J. (1998): "Leapfrog Estimation of a Fixed-Effects Model with Unknown Transformation of the Dependent Variable," unpublished manuscript, Graduate School of Business, University of Chicago.

Alvarez, J. and M. Arellano (1998): "The Time Series and Cross-Section Asymptotics of Dynamic Panel Data Estimators," Working Paper 9808, CEMFI, Madrid.

Anderson, T. W. and C. Hsiao (1981): "Estimation of Dynamic Models with Error Components," *Journal of the American Statistical Society*, 76, 598-606.

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.

Arellano, M., and B. E. Honoré, (2001): "Panel Data Models: Some Recent Developments," in *Handbook of Econometrics,* Vol. 5, ed. by J. Heckman and E. Leamer, Amsterdam: North-Holland.

Baltagi, B. H. (1995): *Econometric Analysis of Panel Data*, New York: John Wiley and Sons, New York.

Berger, J. O., B. Liseo, and R. L. Wolpert (1999): "Integrated Likelihood Methods for Eliminating Nuisance Parameters," *Statistical Science*, 14, 1-28.

Bickel, P. J. (1982), "On Adaptive estimation," *Annals of Statistics,* 10, 647 - 671.

Bickel P. J., C. A. J. Klaassen, Y. Ritov and J. A. Wellner (1993), *Efficient and Adaptive*

Estimation for Semiparametric Models. Johns Hopkins series in mathematical science.

Blundell, R., S. Bond, and F. Windmeijer (2000): "Estimation in Dynamic Panel Data Models: Improving on the Performance of the Standard GMM Estimators", Working Paper 00/12, Institute for Fiscal Studies, London.

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

——— (1982): "Multivariate Regression for Panel Data," *Journal of Econometrics*, 18, 5-46.

——— (1984): "Panel Data," in *Handbook of Econometrics,* Vol. 2, ed. by Z. Griliches and M. D. Intriligator. Amsterdam: North-Holland.

——— (1985): "Heterogeneity, Omitted Variable Bias, and Duration Dependence," in *Longitudinal Analysis of Labor Market Data*, ed. by J. J. Heckman and B. Singer. Cambridge: Cambridge University Press.

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.

——— (1993): "A Note on the Calculation of Adjusted Profile Likelihood," *Journal of the Royal Statistical Society*, Series B, 45, 467-471.

Critchley, F. (1987): "Discussion of Parameter Orthogonality and Approximate Conditional Inference (by Cox, D. R., and N. Reid)," *Journal of the Royal Statistical Society*, Series B, 49, 25-26.

Ferguson, H. (1991): "Asymptotic Properties of a Conditional Maximum-likelihood Estimator," *The Canadian Journal of Statistics*, 20, 63-75.

Ferguson, H., N. Reid and D. R. Cox (1989): "Estimating Equations from Modified Profile Likelihood," in *Estimating Functions*, ed. by V. P. Godambe, Oxford: Oxford University Press.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Rubin (1995): *Bayesian Data Analysis.* New York: Chapman and Hall.

Griliches, Z. and J. A. Hausman (1986): "Errors in Variables in Panel Data", *Journal of*

Econometrics, 31, 93-118.

Heckman, J., H. Ichimura, J. Smith and P. Todd (1998): "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017-1098.

Hsiao, C. (1986): *"Analysis of Panel Data"*. New York: Cambridge University Press.

Hills, S. E. (1987): "Discussion of Parameter Orthogonality and Approximate Conditional Inference (by Cox, D. R., and N. Reid)," *Journal of the Royal Statistical Society*, Series B, 49, 23-24.

Holtz-Eakin, D., W. Newey, and H. Rosen (1988): "Estimating Vector Autoregressions with Panel Data", *Econometrica*, 56, 1371-1395.

Honoré, B. E. (1992): "Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects," *Econometrica,* 60, 533-565.

Honoré, B. E. and L. Hu (1999): "Estimation of Censored Regression Models with Endogeneity," unpublished manuscript, Department of Economics, Princeton University.

Honoré, B. E. and E. Kyriazidou (2000): "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica*, 68, 839-874.

Jeffreys, H. (1961): *Theory of Probability*, 3rd ed. Oxford: Clarendon Press.

Kass, R. E., L. Tierney, and J. B. Kadane (1990): "The Validity of Posterior Expansions Based on Laplace's Method," in *Essays in Honor of George Barnard*, ed. by S. Geiser, S. J. Press, and A. Zellner. Amsterdam: North-Holland.

Kiviet (1995): "On Bias, Inconsistency and Efficiency of Various Estimators in Dynamic Panel Data Models," *Journal of Econometrics*, 68, 53-78.

Lancaster, T. (1997): "Orthogonal Parameters and Panel Data," Working paper nr 97-32, Department of Economics, Brown University.

——— (2000): "The Incidental Parameters since 1948," *Journal of Econometrics,* 95, 391-413.

Liang, K. (1987): "Estimating Functions and Approximate Conditional Likelihood", *Biometrika*, 74, 695-702.

Mundlak, Y. (1961): "Empirical Production Function Free of Management Bias," *Journal*

of Farm Economics, 43, 44-56.

Nerlove, M. (2000): "The Future of Panel Data Econometrics," Working Paper, Department of Agriculture and Resource Economics, University of Maryland.

Neyman, J., and E. L. Scott (1948): "Consistent Estimation from Partially Consistent Observations," *Econometrica*, 16, 1-32.

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

Newey, W. K., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics,* Vol. 4, ed. by R. F. Engle and D. MacFadden. Amsterdam: North-Holland.

Reid, N. (1995): "The Roles of Conditioning in Inference," *Statistical Science,* 10, 138-157.

Tibshirani, R., and L. Wasserman (1994): "Some Aspects of Reparametrization of Statistical Models," *The Canadian Journal of Statistics,* Vol. 22, 163-173.

Tierney, L., R. E. Kass and J. B. Kadane (1989): "Fully Exponential Laplace Approximations to Expectations and Variances of Nonpositive Functions," *Journal of the American Statistical Society*, 84, 710-716.

Trognon, A. (2000): "Panel Data Econometrics: A Successful Past and a Promising Future," Working Paper, Genes (INSEE).

Van den Berg, G. J. (2001): "Duration Models: Specification, Identification, and Multiple Duration," in *Handbook of Econometrics,* Vol. 5, ed. by J. Heckman and E. Leamer. Amsterdam: North-Holland.

Van der Vaart, A. W. (1998): *Asymptotic Statistics,* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

White, H. (1982): "Maximum Likelihood Estimation of Misspecified Models", *Econometrica*, 50, 1-25.

Woutersen, T. M. (2000): "Consistent Estimation and Orthogonality," Working paper, Department of Economics, University of Western Ontario.

——— (2001) "Estimating the Hand of the Past: New Estimators for Hazard Models with

Endogenous Regressors and Endogenous Censoring," Working paper, Department of Economics, University of Western Ontario.

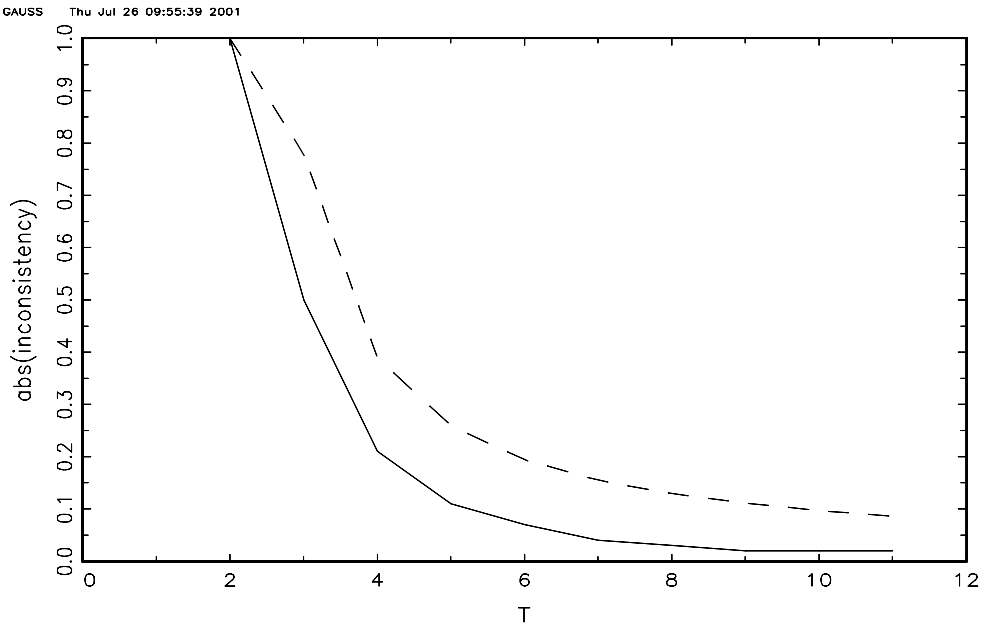

Figure 1: The Absolute Inconsistency of $\hat{\beta}_{ML}$, $\hat{\beta}_I$ with $\pi(\lambda) = \lambda$, and $\hat{\beta}_I$ with $\pi(\lambda) = 1$.

$\lim_{N\to\infty}|\hat{\beta}_{ML} - \beta_0|$: Dashed line

$\lim_{N\to\infty}|\hat{\beta}_I - \beta_0|$ where $\pi(\lambda) = \lambda$: Solid line

$\lim_{N\to\infty}|\hat{\beta}_I - \beta_0|$ where $\pi(\lambda) = 1$: T-ax.

Notes

[1]As $N \to \infty$, using the marginal posteriors is asymptotically equivalent. Considering the mode of the posterior, however, simplifies the algebra.

[2]Lancaster (2000) gives a historical overview of the incidental parameter problem and notes that Neyman and Scott (1948) was cited 207 times in 1997 alone (science citation index).

[3]An earlier version of this paper used a Taylor expansion around λ_0 and than an usual Laplace approximation in the second step. Using Kass et al. (1990) reduces the algebra without changing any of the theorems.

[4]The transformation model of Abrevaya (1998) and one discrete choice model by Honoré and Kyriazidou (2000) are not information-orthogonal. Both models require infinite support for the regressor, can be estimated using a sign function and will be discussed in a separate paper that deals with 'information-orthogonality' of sign functions.

[5]Equation (9) and (**??**) are true for any asymptotics that have T increasing; we maintain "$T \propto N^{\alpha}$ where $\alpha > \frac{1}{3}$" to ensure asymptotic unbiasedness.

[6]In appendix 5 we derive information-orthogonal parametrizations for linear models with more than one autoregressive term.

[7]independent of β.

[8]For this reason, Lancaster (2000) calls it a 'semiparametric model'.

[9]In fact, even exactly orthogonal in the sense that $L_{\beta\lambda} = 0$.

[10]Cox and Reid (1987), equation 10.

[11]Equation (4) in our notation with $N = 1$.

[12]This moment function is mentioned by Anderson and Hsiao (1981), Griliches and Hausman (1986), Holtz-Eakin, Newey, and Rosen (1988), and Arellano and Bond (1991) amongst others.