Nonlinear Panel Data Analysis[*]

Manuel Arellano   Stéphane Bonhomme
CEMFI, Madrid   CEMFI, Madrid

Revised version: January 2011

Abstract

Nonlinear panel data models arise naturally in economic applications, yet their analysis is challenging. Here we provide a progress report on some recent advances in the area. We start by reviewing the properties of random-effects likelihood approaches. We emphasize a link with Bayesian computation and Markov Chain Monte Carlo, which provides a convenient approach to estimation and inference. Relaxing parametric assumptions on the distribution of individual effects raises serious identification problems. In discrete choice models, common parameters and average marginal effects are generally set-identified. The availability of continuous outcomes, however, provides opportunities for point-identification. We end the paper by reviewing recent progress on non fixed-T approaches. In panel applications where the time dimension is not negligible relative to the size of the cross-section, it makes sense to view the estimation problem as a time-series finite sample bias. Several perspectives to bias reduction are now available. We review their properties, with a special emphasis on random-effects methods.

JEL codes: C23.

Keywords: Panel data, incidental parameters.

1   Introduction

The great advantage of panel data is that they allow identification of models that would not be identified on single-outcome data. This is due to observing repeated choices or outcomes from the same economic units over time. Included in this category are models with heterogeneous preferences and/or heterogeneous constraints, as well as models with state dependent choices and/or state dependent shocks. Typically, the identification gains from panel data happen in combination with additional assumptions, which place some form of stability in the time pattern of choices and requirements on the number of outcomes per unit.

The worthiness of panel data has by now been well established in such diverse areas of empirical research as household-level demand and labor supply decisions, workers' wage processes, firm-level productivity, or cross-country determinants of economic growth.[1] However, the empirical success of panel data is mostly confined to linear models and special nonlinearities, for which a more or less complete understanding of identification and inference is available.[2]

For many other nonlinear models of interest in economics the situation is very different. We have a collection of point-identification (and more recently set-identification) results concerning particular model quantities under certain assumptions. Associated to these results, there are some innovative estimation methods, but often little is known about their statistical and numerical properties in practice. We think it is fair to say that we are still short of answers for panel versions of many commonly used models in applied work.

In this survey we attempt to provide a progress report from traditional and more recent perspectives on the current state of the econometrics of nonlinear panel data models. We hope the paper will appeal to both econometricians and empirical economists interested in the use of panel data.

Outline of the paper. In an attempt to provide applied motivation for our review, we start in Section 2 with a list of linear and nonlinear example panel data models of applied interest. In many economic applications, nonlinearity arises naturally when flexible response functions or preferences are sought.

Section 3 introduces a general estimation approach: the so-called "random-effects" (or "correlated random-effects") perspective. We emphasize the link with the Bayesian approach, which provides alternative methods to compute the random-effects estimates and their confidence intervals. The random-effects approach, however, relies on parametric assumptions on the joint distribution of individual effects and exogenous covariates. When these assumptions are relaxed, the estimates are subject to an incidental parameter problem, just as in standard fixed-effects maximum likelihood (Neyman and Scott, 1948). As a result, random-effects estimators are generally inconsistent for a fixed number of time periods T. Moreover, point-identification itself becomes problematic.

In Section 4, we discuss in some detail the identification problem when T is fixed and the distribution of individual effects is left unrestricted. In discrete choice panel models, structural parameters are typically set-identified, unless the model belongs to a very specific parametric class (logistic). We review various approaches to construct population identified sets, and discuss ways of conducting estimation and inference.

Next, we argue that panel data offer opportunities for point-identification. One reason for this is that, even in discrete choice models, *some* quantities of interest such as average marginal effects may be point-identified, although others are not. Also, in panel data models with continuous outcomes, the

---

[1]Arellano (2003) provides examples of panel data applications in these areas.
[2]See the survey by Arellano and Honoré (2001).

availability of repeated outcomes for a time-invariant structure of heterogeneity is a powerful source of identification. Lastly, restricting the conditional distribution of individual effects given exogenous covariates may be another useful source of (point-) identification.

In many applications, the time-series dimension T of the panel is not negligible relative to its cross-sectional dimension N. In such cases, it makes sense to view the incidental parameter problem as time-series finite-sample bias. In Section 5, we review some recently proposed bias reduction methods. We also study the properties of random-effects estimators in this context. In general, random-effects estimates are consistent as T increases but suffer from finite-sample bias. We discuss estimation of average marginal effects in this context.

Lastly, Section 6 concludes.

2 A menu of panel models

The overriding characteristic of panel data is the presence of time. Variation over time and timing considerations provide a host of opportunities for addressing richer economic questions and conducting more sophisticated empirical analyses than those affordable from purely cross-sectional data. These opportunities include individual-specific effects, time-dependent patterns of endogeneity and exogeneity, and dynamic relationships.

Individual effects. Firstly, individual specific effects are a way of allowing for fixed-effects endogeneity and heterogeneous responses. The most standard setup is a linear model with an additive time-invariant effect:

$$y_{it} = x'_{it}\beta + \alpha_i + \sigma v_{it},$$

together with an assumption of some form of independence between v_{it} and $(x_{i1}, ..., x_{iT}, \alpha_i)$. These variables may represent, for example, (transformations of) individual consumer or leisure demand (y_{it}), prices (x_{it}), the marginal utility of initial wealth (α_i), and a preference shifter (v_{it}). The model allows for fixed-effects endogeneity (cross-sectional correlation between prices and α_i), but rules out any dependence between prices and preference shifters (strict exogeneity of x_{it}). An alternative version of the model is to transfer the strict exogeneity assumption to an external instrumental variable z_{it} (e.g. a tax component of the price), thereby allowing x_{it} to be strictly endogenous.

The motives for using the previous model quickly lead to nonlinearities if more flexible response functions or preferences are sought. A simple specification with interactions between observables and unobservables is a location-scale model with heterogeneous volatility:

$$y_{it} = \left(x'_{it}\beta + \alpha_{0i}\right) + \sigma\left(x'_{it}\gamma + \alpha_{1i}\right)v_{it}.$$

A semiparametric generalization of the above is the quantile model

$$y_{it} = x'_{it}\beta\left(u_{it}\right) + \alpha_i\gamma\left(u_{it}\right)$$

3

where u_{it} is the rank of the error v_{it}, so that

$$u_{it} \mid x_{i1}, ..., x_{iT}, \alpha_i \quad \sim \quad \mathcal{U}(0, 1), \tag{1}$$

and $\beta(u)$ and $\gamma(u)$ are nonparametric functions. While the econometrics of the linear model is standard for both least-squares and instrumental-variable versions, the one of the location-scale model is not, and the quantile panel model is a topic of ongoing research.[3] Other forms of interaction between observables and unobservables arise in Chamberlain (1992)'s linear random coefficient model:[4]

$$y_{it} = g_0(x_{it}, \theta) + g_1(x_{it}, \theta)' \alpha_i + v_{it},$$

which has been recently re-examined in work by Graham and Powell (2010) and Arellano and Bonhomme (2010).

Non-additive unobservables arise naturally in the context of discrete choice models, such as corner-solution models of leisure demand, of the form

$$y_{it} = \mathbf{1}\left\{ F\left(x_{it}'\beta + \alpha_i \right) \geq u_{it} \right\},$$

where y_{it} is a $0-1$ indicator of participation, $1(.)$ is the indicator function, u_{it} is a rank variable as before, and $F(.)$ is a cumulative distribution function. However, non-additive fixed effects may also arise in continuous response functions. An example is the following heterogeneous constant elasticity of substitution (CES) production function:

$$\log y_{it} = \lambda \log \ell_{it} + (1 - \lambda) \log \left[\gamma x_{it}^{\sigma_i} + (1 - \gamma) z_{it}^{\sigma_i} \right]^{1/\sigma_i} + \alpha_i + v_{it}, \tag{2}$$

which allows for different degrees of complementarity between low-skill labor (ℓ_{it}), high-skill labor (x_{it}), and capital equipment (z_{it}). More generally, equilibrium conditions suggest a broad class of GMM estimation problems for instrumental variable models à la Hansen and Singleton (1982) with fixed effects:

$$\mathbb{E}\left[z_i \otimes g\left(y_{it}, x_{it}, \theta, \alpha_i \right) \right] = 0.$$

Time patterns. The previous examples include fixed effects but do not allow for time patterns in the dependence between observables and time-varying unobservables. However, the availability of a time dimension makes it conceptually possible to go beyond the cross-sectional notions of strict exogeneity and strict endogeneity, whereby the full time series of a regressor is either fully independent or fully dependent of the full time series of error terms. Thus, x may depend on past v's but not on future v's (predeterminedness), or on v's that are close in time but not on v's from distant periods.

---

[3]See Koenker (2004) and Abrevaya and Dahl (2008), for example.
[4]See also Wooldridge (1997, 1999).

For example, a discrete choice model or a quantile model with general predetermined variables would replace the strict exogeneity assumption (1) with the sequential conditioning assumption

$$u_{it} \mid x_{i1}, ..., x_{it}, \alpha_i \quad \sim \quad \mathcal{U}(0,1). \qquad (3)$$

Time patterns of dependence arise naturally in the context of dynamic models. These are models that consider the effects of lagged outcomes and/or lagged and current interventions on current outcomes, and also models of the transition between states. In this context the gap between what is well known for the econometrics of linear and nonlinear models is particularly large, despite the existence of a broad array of nonlinear situations of applied interest.[5] Examples include mixed discrete/continuous VAR models of transmission of shocks

$$\begin{aligned}
y_{it} &= \left(\rho y_{it-1} + x'_{it}\beta + \alpha_{1i} + v_{it} \right) d_{it} \\
d_{it} &= \mathbf{1} \left\{ \gamma d_{it-1} + z'_{it}\delta + \alpha_{2i} + \phi v_{it} + \varepsilon_{it} \geq 0 \right\},
\end{aligned}$$

or state-dependent discrete choice responses

$$y_{it} = \mathbf{1} \left\{ F \left(\gamma y_{it-1} + x'_{it}\beta + \alpha_i \right) \geq u_{it} \right\},$$

where $u_{it} \mid x_i^t, y_i^{t-1}, \alpha_i \sim \mathcal{U}(0,1)$, and the feedback process $f\left(x_{it} \mid x_i^{t-1}, y_i^{t-1}, \alpha_i \right)$ and the probability distribution of initial conditions $f(y_{i1}, x_{i1}, \alpha_i)$ are unrestricted, where x_i^t denotes the sequence $(x_{i1}, ..., x_{it})$.

3 Random effects and incidental parameters

Random-effects approaches provide a general solution to the estimation of panel data models. We argue that these approaches are conveniently interpreted from a Bayesian perspective, and that Markov Chain Monte Carlo (MCMC) methods are useful tools for estimation and inference in this context. When the parametric assumptions on the distribution of individual effects are relaxed, however, random-effects estimators suffer from an incidental parameter problem, just as fixed-effects maximum likelihood.

3.1 Average likelihood

We start by presenting the general setup. Let $y_i = (y_{i1}, ..., y_{iT})$ denote the full sequence of outcomes, and let $x_i = (x_{i1}, ..., x_{iT})$ denote a sequence of strictly exogenous covariates. The likelihood of y_i conditioned on x_i and the vector of individual effects α_i is assumed to belong to a parametric family $f_{y|x,\alpha}(y_i|x_i, \alpha_i; \theta)$ characterized by the parameter θ.

---

[5]See for example Altonji, Smith and Vidangos (2009), Arellano and Carrasco (2003), Honoré and Kyriazidou (2000), and Wooldridge (2005).

This framework is not limited to static models. One can include a finite number of lagged outcomes, as:

$$f_{y|x,\alpha}(y_i|x_i,\alpha_i;\theta) = \prod_{t=1}^{T} f_{y_t|y^{t-1},x,\alpha}(y_{it}|y_i^{t-1},x_i,\alpha_i;\theta),$$

where $y_i^t = (y_{it}, y_{i,t-1}, ...)$, in which case x_i contains strictly exogenous regressors and initial conditions.

As a simple (linear) example, let us consider the dynamic Gaussian autoregressive model:

$$y_{it} = \rho y_{i,t-1} + x_{it}'\beta + \alpha_i + v_{it}, \tag{4}$$

where v_{it} is i.i.d. across individuals and time, and follows a normal distribution with zero mean and variance σ^2. In this case, the conditional likelihood function is fully characterized by the parameter $\theta = (\rho, \beta, \sigma^2)$, and is given by:

$$f_{y|x,\alpha}(y_i|x_i,y_{i0},\alpha_i;\theta) = \frac{1}{\sigma^T} \prod_{t=1}^{T} \phi\left(\frac{y_{it} - \rho y_{i,t-1} - x_{it}'\beta - \alpha_i}{\sigma}\right),$$

where ϕ denotes the standard normal density, and y_{i0} is the initial observation of the y_i process.

Additionally, when y_{it} is a vector of random variables (e.g., containing outcomes *per se* as well as non-exogenous regressors), this representation allows for general feedback effects, as long as the researcher is willing to parametrically specify the feedback process– that is, the conditional distribution of x_{it} given x_i^{t-1}, y_i^{t-1}, and α_i. General predetermined regressors, in the sense of variables associated with unrestricted feedback processes, would give rise to a semiparametric likelihood, and are therefore not covered in this discussion.

In a random-effects fashion, the researcher will complete the model by specifying a parametric distribution for the individual effects, conditional on exogenous covariates and initial conditions. Let $f_{\alpha|x}(\alpha_i|x_i;\xi)$ denote that distribution, which is fully characterized by the parameter ξ.

A popular example is to specify α_i to be Gaussian with a mean that is a linear combination of exogenous covariates, and a constant variance, yielding:

$$f_{\alpha|x}(\alpha_i|x_i;\xi) = \frac{1}{\nu}\phi\left(\frac{\alpha_i - x_i'\mu}{\nu}\right),$$

where $\xi = (\mu, \nu)$. Chamberlain (1984) introduces this specification in the static probit model. Alvarez and Arellano (2003) use a similar specification for the distribution of individual effects of an autoregressive model where the conditional mean of α_i is linear in the initial condition of the process.

Once a distribution has been postulated for the individual effects, the researcher will base inference on the average (or integrated) likelihood:

$$f_{y|x}(y_i|x_i;\theta,\xi) = \int f_{y|x,\alpha}(y_i|x_i,\alpha;\theta) f_{\alpha|x}(\alpha|x_i;\xi)\,d\alpha, \tag{5}$$

where the integral is taken over the support of the distribution of individual effects (typically the real line when α_i is scalar). Note that the integrated likelihood function is fully characterized by the parameter (θ, ξ) so that, under correct specification, a parametric approach can be used for estimation and inference.

Bayesian interpretation. The average likelihood function given by (5) is also appealing from a Bayesian perspective. A Bayesian researcher would start by specifying a joint prior distribution for $(\alpha_1, ..., \alpha_N, \theta)$. Viewing $\alpha_1, ..., \alpha_N$ as an i.i.d. sample of missing data, it is natural to assume prior conditional independence of $\alpha_1, ..., \alpha_N$ given θ. Under this assumption, the joint prior conditioned on covariates can be decomposed as:

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

In this case the posterior distribution for θ is proportional to:

$$p(\theta|y_1, ..., y_N, x_1, ..., x_N) \quad \propto \quad \pi(\theta) \int f_{y|x,\alpha}(y_1|x_1, \alpha_1; \theta)\pi_1(\alpha_1|\theta)\, d\alpha_1 \times ...$$
$$... \times \int f_{y|x,\alpha}(y_N|x_N, \alpha_N; \theta)\pi_N(\alpha_N|\theta)\, d\alpha_N. \qquad (6)$$

Therefore, the random-effects integrated likelihood (5) can be interpreted in a Bayesian perspective as a marginal likelihood, where the (hierarchical) prior specification on individual effects is given by:

$$\pi_i(\alpha_i|\theta; \xi) = f_{\alpha|x}(\alpha_i|x_i; \xi).$$

In words, random-effects specifications are a special case of hierarchical Bayesian approaches, where the prior distribution of individual effects is assumed independent of common parameters.

3.2 Integration versus simulation

Suppose first that interest centers on structural parameters θ only. For example, in the above example of the CES production function given by (2), the researcher may be interested in the relative elasticity λ of low-skill labor. A classical approach to estimation is to maximize the log-average likelihood, and to estimate θ as:

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmax}} \left(\underset{\xi}{\operatorname{argmax}} \sum_{i=1}^{N} \log \int f_{y|x,\alpha}(y_i|x_i, \alpha; \theta)f_{\alpha|x}(\alpha|x_i; \xi)\, d\alpha \right), \qquad (7)$$

where the integral is taken on the support of individual effects.

Solving (7) requires computing integrals with respect to α. In the linear autoregressive model with a Gaussian specification for α_i, the average likelihood function is analytical (e.g., Alvarez and Arellano, 2003). However, in nonlinear panel models the integrals are generally not available in closed form and must be approximated numerically. Quadrature methods (Butler and Moffitt, 1982), and simulation-based approaches such as importance sampling (Geweke, 1989) may be used for this purpose.[6]

The connection with Bayesian approaches, which we have emphasized above, suggests another way to estimate θ. Indeed, from (6) the classical random-effects estimator $\widehat{\theta}$ in (7) coincides with the

---

[6]Judd (1998) is a useful reference on numerical integration techniques.

posterior mode of θ, where the prior for α_i is $f_{\alpha|x}(.|x_i;\xi)$, and where θ and the hyperparameter ξ have been endowed with independent flat (improper) priors. So, an alternative approach to estimation is to generate a Markov chain of parameter draws, in a purely Bayesian fashion, using these prior specifications. This approach may be interpreted as a computationally convenient way of calculating the random-effects maximum likelihood estimate $\widehat{\theta}$.

It is well-known that the statistical equivalence between Bayesian and classical approaches is not limited to the case of the posterior mode with flat priors. Using any non-dogmatic priors on θ and ξ instead will result in asymptotically equivalent estimates as N tends to infinity. Also, using posterior mean instead of posterior mode will have a negligible effect on the asymptotic distribution of the estimate.[7]

As a result of the increase in computation power over the last decades, Bayesian estimation approaches have become increasingly attractive from a practical perspective. This is leading in turn to a pragmatic synthesis of Bayesian and frequentist approaches, as MCMC methods can be viewed as a way of computing estimators that are justified from a frequentist point of view. As we will see below, Bayesian techniques are also useful devices to compute frequentist confidence intervals for the parameters of interest.

The principle of Markov Chain Monte Carlo (MCMC) methods is to generate a sequence of draws from the posterior distribution of the model's parameters. The draws are generated in a recursive manner, starting with initial parameter values. The posterior distribution corresponds to the equilibrium distribution of the Markov chain, which is usually reached after a sufficiently large number of steps. The output of the chain– that is, the sequence of parameter values– is then interpreted as a sequence of draws from the posterior distribution of the parameter, and features of that distribution (such as mean, mode, or quantiles) can be directly computed.

In a panel data context, it is often convenient to introduce $\alpha_1, ..., \alpha_N$ as additional parameters that we will draw jointly with θ and the hyperparameters ξ. The sth step of the Markov chain then typically takes the following form, reminiscent of the Gibbs sampler:

- Update $\xi^{(s)}$ given $\alpha_1^{(s-1)}, ..., \alpha_N^{(s-1)}$.

  This step treats the draws of individual effects obtained in the previous step as observations.

- For each $i = 1, ..., N$, update $\alpha_i^{(s)}$ given y_i, x_i, $\theta^{(s-1)}$, and $\xi^{(s)}$.

  For example, when α_i is scalar and enters y_{it} additively, this second step requires to draw from the posterior distribution of a mean of T observations (N times).

- Update $\theta^{(s)}$ given $y_1, ..., y_N$, $x_1, ..., x_N$, and $\alpha_1^{(s)}, ..., \alpha_N^{(s)}$.

---

[7]However, the priors on $\alpha_1, ..., \alpha_N$ are very informative when T is small. See the discussion of misspecification in Subsection 3.3 below.

To draw θ, the researcher proceeds as if the individual effects were observed. Metropolis methods are typically used here.

Parametric Bayesian approaches have been used in many economic applications, some good examples being Rossi, McCulloch and Allenby (1995) for demand analysis in marketing, and Geweke and Keane (2000) for modelling earnings dynamics. See also Hajivassiliou and McFadden (1998). The textbook of Tony Lancaster (2004) devotes a chapter to the application of Bayesian techniques to panel data models.

An appealing feature of Bayesian techniques is that the output of the Markov chain does not only provide estimates of θ and ξ, but also asymptotically valid frequentist confidence intervals. Under quite general conditions (see van der Vaart, 2007, p. 141), the Bernstein-Von Mises theorem guarantees that 5%-95% quantiles of the posterior distribution of θ are equivalent (in a first-order sense) to the endpoints of a frequentist 90%-confidence interval of $\widehat{\theta}$ as N tends to infinity.

Average marginal effects. So far, we have discussed estimation and inference for the vector of structural parameters θ. In many applications, the researcher is also interested in averages of individual responses taken over the distribution of individual effects. These policy parameters, or average marginal effects, take the general form:

$$M = \mathbb{E}\left[m\left(x_i, \alpha_i; \theta\right)\right],$$

where $m()$ is a known function, and where the expectation is taken with respect to the joint distribution of x_i and α_i.[8]

For example, in the CES production function (2), one may be interested in the average (semi-) elasticity of high-skill labor. That is, including all common parameters in θ:

$$\mathbb{E}\left[\frac{\partial \mathbb{E}\left(\log y_{it}|x_i, \ell_i, z_i, \alpha_i, \sigma_i; \theta\right)}{\partial x_{it}}\right] = \mathbb{E}\left[\frac{\gamma(1-\lambda)x_{it}^{\sigma_i-1}}{\gamma x_{it}^{\sigma_i} + (1-\gamma)z_{it}^{\sigma_i}}\right]. \tag{8}$$

Note that this quantity involves an integral over the distribution of individual effects.

A first approach to estimate the average elasticity in (8) is to replace the distribution of individual effects by its random-effects estimate, and to substitute the latter in the expectation. This results in the following estimate:

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\int \frac{\widehat{\gamma}(1-\widehat{\lambda})x_{it}^{\sigma-1}}{\widehat{\gamma}x_{it}^{\sigma} + (1-\widehat{\gamma})z_{it}^{\sigma}} f_{\sigma|x}\left(\sigma|x_i; \widehat{\xi}\right) d\sigma, \tag{9}$$

where $f_{\sigma|x}$ denotes the postulated random-effects distribution of σ_i. Under correct specification, the classical estimate (9) is typically root-N consistent for the average elasticity, and asymptotic standard

---

[8]One may also consider expectations with respect to other distributions. See for example Altonji and Matzkin (2005), and Blundell and Powell (2004).

errors can be obtained *via* the delta-method. Note that evaluating (9) requires to compute an integral numerically.

Here also, an alternative estimate may be computed from the outcome of a Markov chain. To see this, let us denote as:

$$M\left(\gamma, \lambda, \sigma_1, ..., \sigma_N\right) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{\gamma(1-\lambda)x_{it}^{\sigma_i-1}}{\gamma x_{it}^{\sigma_i} + (1-\gamma)z_{it}^{\sigma_i}}. \tag{10}$$

MCMC techniques will deliver a sequence of draws of the model parameters γ, λ, and the individual effects $\sigma_1, .., \sigma_N$, from which it is easy to construct a sequence of draws from the posterior distribution of the average marginal effect $M\left(\gamma, \lambda, \sigma_1, ..., \sigma_N\right)$, simply using (10). A natural estimate is then the posterior mode, or mean, of the effect (as proposed in Arellano and Bonhomme, 2009).

Under correct specification of the distribution of individual effects, the advantage of using the posterior mode or mean of $M\left(\gamma, \lambda, \sigma_1, ..., \sigma_N\right)$ instead of the classical random-effects estimate (8) is computational, as no extra calculation (i.e, numerical integration) is needed once the Markov chain is available. Moreover, as in the case of common parameters, frequentist confidence intervals can be read on the posterior distribution of the average marginal effect. In Section 5, we will see that when the distribution of individual effects is misspecified, there is an additional reason for using the posterior mean or mode to estimate average marginal effects, as the latter is consistent as T tends to infinity while the classical random-effects estimator is not. This is due to the fact that as the time-series information accumulates, the impact of the prior distribution of the effects on the posterior of the policy parameter tends to disappear.

3.3   Properties under misspecification

Panel data researchers have long been concerned with the possibility that the functional form of the distribution of α_i might be misspecified (e.g., Chamberlain, 1980). The standard approach in the literature is to maintain parametric assumptions on the conditional distribution of y_i given x_i and α_i, while at the same time leaving the distribution of α_i given x_i unrestricted. The asymmetry between the two parts of the average likelihood function is usually motivated by the desire to treat α_i as missing data given the availability of repeated outcomes for each unit.

Formally, the researcher will thus interpret (5) as an average *pseudo*-likelihood, possibly misspecified if the population cross-sectional distribution of individual effects does not belong to the parametric family $f_{\alpha|x}\left(.|x_i; \xi\right)$. Robustness of estimation methods to possible misspecification is often an important issue in empirical applications. One source of misspecification may be that the marginal distribution of α_i is incorrect (for example, the true distribution of α_i is not Gaussian). Another, empirically relevant, source is the incorrect conditioning with respect to exogenous covariates. This second problem is especially severe in dynamic models, where one needs to properly control for initial

conditions.[9]

In linear models, simple random-effects specifications may lead to consistent estimates, even though they are misspecified. This situation arises in the linear autoregressive model when one postulates a Gaussian distribution for the individual effects, where the conditional mean of α_i is linear in the initial condition y_{i0}. In this model, correlated Gaussian random-effects remains consistent even if the population distribution of individual effects is not Gaussian. One simply needs to adjust the computation of standard errors (Alvarez and Arellano, 2003).

Incidental parameters. In nonlinear models, however, we are not aware of similar robustness results for random-effects specifications. Random-effects maximum likelihood falls into the general class of pseudo-likelihood approaches. As a consequence, it will asymptotically deliver pseudo-true parameter values, which minimize the (Kullback-Leibler) distance between the postulated parametric family of distributions and the population distribution of the data (White, 1982). In nonlinear models, there does not seem to be any special reason to expect pseudo-true parameter values to coincide with true parameter values in general. Random-effects maximum likelihood will thus yield inconsistent estimates of the parameters.

A notable exception is given by models where the conditional likelihood factors as:

$$f_{y|x,\alpha}(y_i|x_i,\alpha;\theta) = g\left(y_i, x_i; \theta\right) h\left(y_i, x_i, \alpha\right), \tag{11}$$

as in the panel Poisson counts model (Lancaster, 2002, Blundell *et al.*, 2002). The reason is that, when (11) holds, θ satisfies moment conditions that do not depend on the postulated distribution of the effects.[10] However, this likelihood property is very rarely exactly satisfied in nonlinear models.

Interestingly, as T increases random-effects estimators become consistent, irrespective of the form of the postulated distribution of individual effects (Arellano and Bonhomme, 2009). The reason is that:

$$\log f_{y|x,\alpha}(y_i|x_i,\alpha;\theta) = \sum_{t=1}^{T} \log f_{y_t|y^{t-1},x,\alpha}(y_{it}|y_i^{t-1}, x_i, \alpha; \theta)$$

is a sum of T time-series observations, so the effect of the prior distribution $f_{\alpha|x}$ becomes negligible compared to that of the likelihood as the number of time periods increases. For small T, however, the estimator suffers from a bias of order $1/T$ under standard regularity conditions.

The properties of misspecified random-effects maximum likelihood are thus similar to those of a "fixed-effects" maximum likelihood (ML) approach that treats the individual effects as parameters to be estimated jointly with θ. In this case, the ML estimates of α_i suffer from an estimation bias. As a consequence, in a nonlinear setup the ML estimate of θ is generally inconsistent for fixed T.[11] As

---

[9]Heckman (1981) and Wooldridge (2005) describe different approaches to the treatment of initial conditions in dynamic panel data models.

[10]Specifically, $\mathbb{E}\left[\partial \log g\left(y_i, x_i; \theta\right) / \partial \theta\right] = 0$.

[11]Here also, an exception is given by models where the conditional likelihood factors as in (11).

the precision of the α_i estimates increases as T increases, the ML estimate of θ is consistent in the limit. However, it suffers from a $1/T$ bias. The fixed-T inconsistency of fixed-effects and random-effects maximum likelihood approaches may thus be viewed as a manifestation of the same *incidental parameter* problem.

Flexible random-effects. In contexts where the parameter θ is point-identified, a general solution is to postulate a flexible parametric model for the individual effects, and to let the dimension of the parameter ξ increase with the sample size. Sieve– random-effects– maximum likelihood (Shen, 1997, Chen, 2007) has been advocated in the context of nonlinear models with latent variables such as panel data models (Hu and Schennach, 2008, Bester and Hansen, 2007). A related approach is given by nonparametric Bayesian methods based on Dirichlet process priors (West, Müller and Escobar, 1994). Hirano (2002) applies these methods to an autoregressive panel data model. A challenge in panel data applications is the conditioning on (time-series sequences of) covariates and initial conditions, which raises an issue of curse of dimensionality.

Recent work, however, has emphasized the possibility that common parameters– and average marginal effects– may fail to be point-identified. As we will see in the next section, lack of identification may arise in simple semiparametric models such as probit, where the only nonstandard feature of the model is due to the presence of an unrestricted distribution for the individual effects. When point-identification fails, "flexible" random-effects estimation approaches may provide misleading answers in short panels. For this reason, the study of identification is an important part of nonlinear panel data analysis.

4   Fixed-T, fixed-effects (set)-identification

Identification requires that the parameter values that generate the data be uniquely defined. In panel data analysis, the model's parameters include the unknown distribution of individual effects, so identification is a non-trivial issue. Recent work in panel data emphasizes the possibility that the parameters of interest be set-identified in some of the most widely used models. Here we first review part of this work. Then, we discuss some situations where quantities of interest are point-identified for fixed T.

4.1   Widespread identification failure...

When panel data outcomes are discrete, serious identification issues arise. This point was made clear in an early paper by Gary Chamberlain (recently published in 2010), and has recently been re-emphasized in the literature.

As before, we will work with a parametric conditional model for y_i given x_i and α_i characterized by a parameter θ. Assuming that outcomes have discrete support, the average likelihood function of

an individual observation is given by:

$$\Pr\left(y_i|x_i;\theta\right) = \int \Pr\left(y_i|x_i,\alpha;\theta\right) f_{\alpha|x}\left(\alpha|x_i\right) d\alpha, \tag{12}$$

As an example, one may consider a static binary choice model of the form:

$$y_{it} = \mathbf{1}\left\{x_{it}'\theta + \alpha_i \geq v_{it}\right\}, \tag{13}$$

where v_{it} are i.i.d. draws from a known distribution F (e.g. normal, logistic), independent from the sequence of x's and the individual effects. In this case the conditional outcome probabilities are given by:

$$\Pr\left(y_i|x_i,\alpha;\theta\right) = \prod_{t=1}^{T} F\left(x_{it}'\theta + \alpha\right)^{y_{it}} \left[1 - F\left(x_{it}'\theta + \alpha\right)\right]^{1-y_{it}}.$$

The identification problem comes from the fact that the individual effects are unobserved to the econometrician, so the conditional probabilities $P\left(y_i|x_i,\alpha;\theta\right)$ have no direct counterpart in the data. Here we leave the conditional distribution of individual effects unrestricted, consistently with a "fixed-effects" perspective. Thus, *via* (12), the observed data frequencies involve an unknown mixing distribution.

Chamberlain's underidentification argument. Suppose that the researcher is interested in estimating the structural parameter vector θ. The identification question is the following: is there a unique value of θ such that (12) is satisfied for *some* conditional distribution of individual effects $f_{\alpha|x}$? In the static binary choice model (13), with $T = 2$ and exogenous covariates with bounded support, Chamberlain (2010) finds a surprisingly simple answer to the identification question: θ is not point-identified, unless F is the logistic distribution.

Chamberlain's proof has two steps. In a first step, it is shown that for θ to be identified, there must exist a linear combination of the conditional probabilities that is equal to zero. In other words, there must exist some non-trivial ψ_j's, possibly dependent on x, such that:

$$\psi_1 \Pr\left(0,0|x,\alpha;\theta\right) + \psi_2 \Pr\left(1,0|x,\alpha;\theta\right) + \psi_3 \Pr\left(0,1|x,\alpha;\theta\right) + \psi_4 \Pr\left(1,1|x,\alpha;\theta\right) = 0, \quad \text{for all } x,\alpha.$$
$$\tag{14}$$

The second step in the proof is to show that the logistic distribution is the *only* continuous distribution for which (14) is satisfied. To see this, note that taking $\alpha \to \pm\infty$ in (14) yields $\psi_1 = \psi_4 = 0$, provided α_i is supported on the full real line given x. It thus follows that:

$$\psi_2 F\left(x_1'\theta + \alpha\right)\left(1 - F\left(x_2'\theta + \alpha\right)\right) = \psi_3 F\left(x_2'\theta + \alpha\right)\left(1 - F\left(x_1'\theta + \alpha\right)\right).$$

This is an equation of the form:

$$G(u+d) = a(d)G(u),$$

with $d = (x_2 - x_1)'\theta$ and $G(u) = F(u)/(1 - F(u))$, the solution of which is $G(u) = e^{a+bu}$. This solution for G yields the logistic form for F: $F(u) = e^{a+bu}/\left(1 + e^{a+bu}\right)$.

Restricting the support of individual effects. To provide some intuition about Chamberlain's underidentification result, it is useful to work under the assumption that the distribution of individual effects has known discrete support. We will denote its points of support as $\underline{\alpha}_k$, $k = 1, ..., K$, with K possibly very large. Then, (12) takes the form of a simple linear system of equations:

$$\Pr(y_i|x_i; \theta) = \sum_{k=1}^{K} P_k(y_i|x_i; \theta) \pi_k(x_i), \tag{15}$$

where $\pi_k(x_i)$, $k = 1, ..., K$, denote the conditional probabilities of individual effects given covariates, and where:

$$P_k(y_i|x_i; \theta) = \Pr(y_i|x_i, \alpha_i = \underline{\alpha}_k; \theta), \quad k = 1, ...K.$$

Bonhomme (2010) notes that, when the number of points of support of y_i is large enough relative to K, there exists a simple way to obtain restrictions on θ that do not involve the individual effects. To see this, let $P_x(\theta)$ denote the $2^T \times K$ matrix of conditional probabilities $P_k(y|x; \theta)$ for a given value $x_i = x$ of the exogenous covariates, and denote as $P_{y|x}$ the $2^T \times 1$ vector of data frequencies, and as π_x the $K \times 1$ vector of probabilities of individual effects. Writing (15) in matrix form yields:

$$P_{y|x} = P_x(\theta)\pi_x, \quad \text{for all } x.$$

So, assuming that $P_x(\theta)$ has independent columns,[12] we obtain the following restrictions on θ alone:

$$\left[I - P_x(\theta) \left(P_x(\theta)' P_x(\theta) \right)^{-1} P_x(\theta)' \right] P_{y|x} = 0, \tag{16}$$

where I is the identity matrix of size 2^T.

This "functional differencing" approach differences out the probability distribution of the individual effects. A differencing strategy is feasible even though the panel data model is nonlinear, as the system (15) that relates outcome probabilities to the probabilities of individual effects is linear. Moreover, as the projection matrix in (16) multiplies the vector of outcome probabilities, this approach delivers conditional moment restrictions– given covariates x_i– on common parameters θ.

In discrete choice models with large K, the moment restrictions (16) will often be noninformative about θ. Indeed, the projection matrix on the left-hand side of (16) is zero when the rows of $P_x(\theta)$ are linearly independent. When $T = 2$, the dimensions of the matrix $P_x(\theta)$ are $4 \times K$. So, when $K \geq 4$ the rows of $P_x(\theta)$ are independent in general. An important exception is obtained for the logit model. To see the link with Chamberlain (2010)'s argument, note that (14) simply means that the rows of the $P_x(\theta)$ are *dependent*. In the logit model, even when K is large relative to 2^T, the functional differencing approach delivers informative moment restrictions on θ, which actually coincide with the first-order conditions of the conditional maximum likelihood estimator of Andersen (1970).

---

[12]This assumption may be easily relaxed, using a generalized matrix pseudo-inverse (e.g., Moore-Penrose).

In situations where the support of individual effects is less rich than the support of outcomes, this discussion also highlights identification possibilities.[13] Indeed, when K is smaller than the number of points of support of y_i (that is, 2^T) the rows of $P_x(\theta)$ are necessarily linearly dependent. As a result, (16) is generally informative about θ. This suggests that, when the support of outcomes is richer than the support of individual effects, panel data offer the possibility to derive restrictions on θ that do not depend on α_i. This intuition will be important when discussing the identification of panel data models with continuous outcomes.

Partial identification. In many instances, however, the researcher is unwilling to limit *a priori* the number of values that α_i can take. When α_i can take an arbitrarily large number of values, Chamberlain (2010)'s result suggests that most discrete outcomes panel data models with discrete regressors will not be point-identified. Recently, several authors have acknowledged the widespread identification failure of fixed-effects discrete choice panel models with discrete x's, and proposed methods that deal with the lack of point-identification.

When parameters fail to be point-identified, one may still be able to characterize the region where the true parameter belongs. This is the *identified set* of the parameter θ, which is given by:

$$\Theta^I = \left\{ \theta\text{'s, there exists a set of probabilities } \pi_k(x) \text{ such that:} \right.$$

$$\left. \Pr(y_i = y | x_i = x) = \sum_{k=1}^{K} P_k(y|x;\theta) \pi_k(x) \text{ for all values of } y, x. \right\}$$

One method to construct Θ^I is given in Honoré and Tamer (2006). They start by fixing a value θ of common parameters. Then, they note that θ belongs to the identified set if and only if a system of linear equality and inequality restrictions is satisfied.[14] So, checking that $\theta \in \Theta^I$ may be done in the same way as one verifies the feasibility of a linear program, for which there exist fast and widely available algorithms. Chernozhukov, Hahn and Newey (2005) and Honoré and Tamer (2006) show that a linear programming approach can also be used to compute bounds on average marginal effects.

Another approach to construct θ is given in Chernozhukov *et al.* (2007). They note that true values of θ can be characterized as the minimand of the (weighted) Euclidean distance between the data frequencies $\Pr(y_i = y | x_i = x)$ and the model predictions $\sum_{k=1}^{K} P_k(y|x;\theta) \pi_k(x)$, across outcome and regressor values, where the minimization is with respect to the vector of probabilities of individual effects $\pi_k(x)$. Hence, an alternative approach to compute the identified set Θ^I is to solve this quadratic programming problem, where the probabilities $\pi_k(x)$ are constrained to belong to the unit simplex.

There exists interesting evidence on the size of identified sets in some simple models. In a dynamic probit model without regressors, Honoré and Tamer (2006) find that the identified region for the

---

[13]See the recent paper by Browning and Carro (2009) for a related approach, in the context of dynamic binary outcome models.

[14]The inequality restrictions come from the fact that $\pi_k(x)$– which is a probability– must be nonnegative.

autoregressive parameter θ is indistinguishable from a singleton when $T \geq 4$. This means that θ is "nearly" point-identified in this model. In the static probit model with an i.i.d. binary regressor, the numerical calculations in Chernozhukov *et al.* (2007) suggest that the identified set for the slope coefficient shrinks very fast as T increases. However, they also emphasize that the convergence rate of the identified set to a singleton as T increases depends on the properties of the joint distribution or regressors and individual effects.

The main advantage of the approach in Chernozhukov *et al.* relative to Honoré and Tamer's is that it can be used for estimation and inference on the identified set. Indeed, part of the linear equality constraints that define the identified set in the Honoré and Tamer approach involve data frequencies. When population quantities are replaced by empirical counterparts, these constraints will not hold in general. In contrast, the quadratic program of Chernozhukov *et al.* can be applied to empirical frequencies. Building on the recent literature on estimation and inference of partially identified models (i.e., Chernozhukov, Hong and Tamer, 2007), Chernozhukov *et al.* (2007) propose to estimate the identified region as the set of θ's for which the quadratic objective function is small enough. They provide a consistency result for their estimator and show how to construct asymptotically valid confidence regions for the identified set Θ^I.

So far, these approaches are limited to simple setups with discrete outcomes and discrete regressors. At the same time, the econometric literature on partial identification and inference is rapidly developing (e.g., Tamer, 2010). Thus, we expect progress to be done in this direction, which could be part of a future survey on nonlinear panel data analysis.

A numerical illustration. As an illustration, we compute the population identified set in a simple static probit model:

$$y_{it} = \mathbf{1}\left\{\theta(t-1) + \alpha_i \geq v_{it}\right\},$$

where v_{it} is i.i.d. standard normal. We vary the number of points of support K of α_i, that we suppose uniformly distributed on the interval $(-3, 3)$. In addition, to generate the data we take $\theta = .7$, and we choose the probabilities π_k to closely resemble those of a normal distribution, as in Honoré and Tamer (2006):
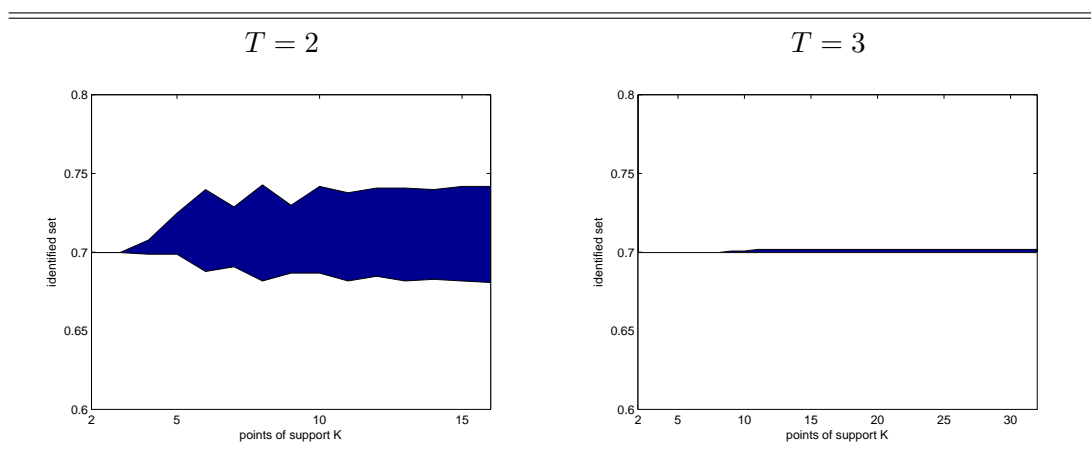
$$\pi_k = \Phi\left(\frac{\alpha_k + \alpha_{k+1}}{2}\right) - \Phi\left(\frac{\alpha_k + \alpha_{k-1}}{2}\right), \quad k = 2, ..., K-1,$$

where Φ is the cumulative distribution function of the standard normal distribution. In addition, $\pi_1 = \Phi\left(\frac{\alpha_1 + \alpha_2}{2}\right)$, and $\pi_K = 1 - \Phi\left(\frac{\alpha_K + \alpha_{K-1}}{2}\right)$.

Figure 1 shows the results, for $T = 2$ and $T = 3$, and for various values of K. We have used a quadratic programming method to compute the identified sets. We see that, when the support of α_i is known and smaller than $2^T - 1$ (that is, $K = 3$ for $T = 2$, and $K = 7$ for $T = 3$), the structural parameter θ is point-identified. This suggests that excess support in outcomes relative to individual

effects may lead to point-identification. When K is larger, the parameter is partially identified. For $T = 3$, the identified set is already very small in this model.

Figure 1: Identified sets in a simple static logit model with time trend

Note: The true θ is .7, and α_i has K points of support, where K is shown on the x-axis. The y-axis features the identified set for θ.

4.2   ... but there is room for point-identification

The conclusion of the previous subsection may seem pessimistic: in many panel data models, point-identification is rather the exception than the rule. So one needs to resort to (yet) non-standard set estimation and inference approaches, which are (so far) limited to simple setups. Here we argue, however, that panel data, thanks to the presence of time variation, offer opportunities for point-identification.

4.2.1   Point-identified objects of interest

Binary choice.   In various panel models, *some* objects of interest are point-identified, although others are not. For illustration, let us start by considering a simple static probit model:

$$y_{it} = \mathbf{1}\left\{\theta x_{it} + \alpha_i \geq v_{it}\right\}, \tag{17}$$

where v_{it} are i.i.d. standard normal, and where x_{it} is a sequence of binary exogenous regressors. We have seen above that θ is not point-identified in this model. However, in an application, the researcher

17

may not be interested in θ *per se*, but rather in the average marginal effect of an increase in x_{it} from 0 to 1, that is:

$$\begin{aligned} \Delta &= \mathbb{E}\left[\mathbb{E}\left(y_{it}|x_{it}=1,\alpha_i\right) - \mathbb{E}\left(y_{it}|x_{it}=0,\alpha_i\right)\right] \\ &= \mathbb{E}\left[\Phi\left(\theta+\alpha_i\right) - \Phi\left(\alpha_i\right)\right], \end{aligned}$$

where Φ is the cumulative distribution function of the standard normal distribution.

In this model, although the overall average Δ is *not* point-identified for fixed T, the average effect on the subpopulation of units whose x's change over time is point-identified (Chamberlain, 1982, Hahn, 2001). To see this, let us take $T=2$. We have:

$$\begin{aligned} \Delta_{10} &= \mathbb{E}\left[\mathbb{E}\left(y_{i1}|x_{i1}=1,\alpha_i\right) - \mathbb{E}\left(y_{i1}|x_{i1}=0,\alpha_i\right)\Big|x_{i1}=1,x_{i2}=0\right] \\ &= \mathbb{E}\left[\mathbb{E}\left(y_{i1}|x_{i1}=1,x_{i2}=0,\alpha_i\right) - \mathbb{E}\left(y_{i2}|x_{i2}=0,\alpha_i\right)\Big|x_{i1}=1,x_{i2}=0\right] \\ &= \mathbb{E}\left[\mathbb{E}\left(y_{i1}|x_{i1}=1,x_{i2}=0,\alpha_i\right) - \mathbb{E}\left(y_{i2}|x_{i1}=1,x_{i2}=0,\alpha_i\right)\Big|x_{i1}=1,x_{i2}=0\right] \\ &= \mathbb{E}\left[y_{i1}|x_{i1}=1,x_{i2}=0\right] - \mathbb{E}\left[y_{i2}|x_{i1}=1,x_{i2}=0\right]. \end{aligned}$$

To derive this series of equalities, we have used two types of assumptions. Note that the normality assumption is not needed. The first assumption we have used is the strict exogeneity of x_{it}, which ensures that $\mathbb{E}\left(y_{i1}|x_{i1},x_{i2},\alpha_i\right)$ and $\mathbb{E}\left(y_{i1}|x_{i1},\alpha_i\right)$ coincide. The second is a marginal stationarity assumption, which implies that the conditional expectation $\mathbb{E}\left(y_{it}|x_{it},\alpha_i\right)$ does not depend on t. These two types of assumptions have been used in other contexts to derive point- and set-identified effects of interest (Chernozhukov *et al.*, 2009, Hoderlein and White, 2010).

It thus follows that:

$$\Delta_{10} = \mathbb{E}\left[y_{i1}-y_{i2}|x_{i1}=1,x_{i2}=0\right],$$

so that Δ_{10} is point-identified from the data. A similar result holds for the average effect in the subpopulation of units for which $x_{i1}=0$ and $x_{i2}=1$. However, in this model, the two remaining conditional averages– for $(x_{i1},x_{i2})=(0,0)$ or $(1,1)$– are not point-identified. The approach in Chernozhukov *et al.* (2007) delivers bounds on these terms, hence on the unconditional effect Δ.

Random coefficients.   As another example where point-identified average marginal effects are available, consider the following random coefficient model:

$$y_{it} = z_{it}'\delta + \alpha_i + \beta_i x_{it} + v_{it}, \tag{18}$$

where $x_{it}\in\{0,1\}$ and z_{it} are strictly exogenous. In this model, identification of δ is obtained *via* quasi-differencing, provided $T\geq 3$ (Chamberlain, 1992).

Here, the marginal effect of x_{it} for individual i is:

$$\mathbb{E}\left(y_{it}|x_{it}=1,\alpha_i,\beta_i\right) - \mathbb{E}\left(y_{it}|x_{it}=0,\alpha_i,\beta_i\right) = \beta_i.$$

The researcher may be interested in estimating average marginal effects, i.e. means of β_i in specific subpopulations. Arellano and Bonhomme (2010) use this framework to model the effect that a mother smokes during pregnancy ($x_{it} = 1$) on the weight of the child at birth (y_{it}). The "panel" component of the data comes from the fact that mothers have multiple children. They are interested in estimating the mean effect of smoking across mothers, while acknowledging that this effect may be heterogeneous.

Note that (18) implies that:

$$\mathbb{E}\left(\beta_i \Delta x_{it} | x_i\right) = \mathbb{E}\left[\Delta\left(y_{it} - z_{it}'\delta\right)\Big| x_i\right]$$

is point-identified, where $\Delta w_{it} = w_{it} - w_{i,t-1}$ is the first-difference operator, and x_i denotes the full sequence of regressors x_{it}. Therefore, similarly to the above binary choice model, and taking $T = 2$ for simplicity, the mean of β_i is point-identified in the subpopulation of individuals whose x's vary over time. In the smoking example, this means that the average smoking effect is identified in the subpopulation of mothers whose smoking status changed between pregnancies. However, unlike the binary choice example, bounds on the smoking effect for mothers whose smoking status did not change will not be informative here, as the range of the outcome variable is too wide. This means that there will be considerable uncertainty about the overall average smoking effect in the total population.

When x_{it} is continuous instead of binary and the subpopulation of units whose x's change over time coincides with the total population, the researcher will be able to identify the overall average effect. The analysis in Graham and Powell (2010), however, shows that one must proceed with caution in estimation. To see why, let us write the population average of β_i as:

$$\mathbb{E}\left(\beta_i\right) = \mathbb{E}\left[\frac{\Delta\left(y_{it} - z_{it}'\delta\right)}{\Delta x_{it}}\right]. \tag{19}$$

The estimation problem comes from the presence of the denominator in (19): when x_{it} is a persistent process, Δx_{it} will be close to zero with non-negligible probability, so a naive empirical counterpart of (19) will be very imprecise. Graham and Powell solve this problem by using a trimming strategy, dropping a small percentage of problematic observations that goes to zero as the sample size increases.

Predetermined regressors. Lastly, although we have focused on static models, it may also be that interesting average effects are point-identified in models with dynamics. To give a simple example, let us consider again model (18), but now allowing that x_{it} be predetermined, as opposed to strictly exogenous. Namely, x_{it} is allowed to be correlated to past shocks $v_{i,t-1}$, $v_{i,t-2}$, ..., though not to current and future ones. Predeterminedness seems more appealing than strict exogeneity in the context of smoking behavior, as one could expect that a mother whose first child had a low birthweight (i.e., low v_{i1}) could react by quitting smoking before her second pregnancy ($x_{i2} = 0$).

Let us start by assuming that $\delta = 0$. It turns out that interesting average effects remain point-

identified when regressors are predetermined. To see why, note that

$$
\begin{aligned}
\mathbb{E}\left(\Delta y_{it}|x_{i,t-1}=1\right) &= \mathbb{E}\left(\beta_i \Delta x_{it}|x_{i,t-1}=1\right) + \mathbb{E}\left(\Delta v_{it}|x_{i,t-1}=1\right) \\
&= \mathbb{E}\left(\beta_i \Delta x_{it}|x_{i,t-1}=1\right),
\end{aligned}
$$

where we have used that both v_{it} and $v_{i,t-1}$ are mean independent of $x_{i,t-1}$. Moreover, using that x_{it} can take only two values:

$$
\mathbb{E}\left(\beta_i \Delta x_{it}|x_{i,t-1}=1\right) = -\Pr\left(x_{it}=0|x_{i,t-1}=1\right)\mathbb{E}\left(\beta_i|x_{it}=0, x_{i,t-1}=1\right).
$$

A similar argument shows that the mean of β_i on the subpopulation for which $x_{it}=1$ and $x_{i,t-1}=0$ is also identified. This shows that, when $T=2$, the mean of β_i in the subpopulation of mothers whose smoking status changes over time remains point-identified when smoking behavior is predetermined.

Note, however, that δ has been assumed known (and normalized to zero) in the discussion. In this framework, when δ is unknown and in the absence of external sources of identification (for example, an instrument), the above average effects are not point-identified. See Chamberlain (1993) for an illustration of this remark. Finding interesting point-identified effects in dynamic panel data models seems an interesting research question.

4.2.2 Exploiting excess support

Our analysis of discrete choice panel models suggests that the identification failure is due to the lack of support in the outcome variables, when the researcher wants to allow for a rich specification of unobserved heterogeneity (i.e., a continuously distributed vector of individual effects). In panel data models with continuously distributed outcomes, however, the situation is very different, and time variation offers opportunities for point-identification.

Kotlarski's result. As a first example, let us consider the simple linear model:

$$
y_{it} = x_{it}'\theta + \alpha_i + v_{it}, \quad t = 1,2. \tag{20}
$$

We assume that x_{it} is strictly exogenous, so that $\mathbb{E}\left(v_{it}|x_{i1}, x_{i2}, \alpha_i\right) = 0$. Then, θ and the mean of α_i are point-identified. When the dependence structure of (v_{i1}, v_{i2}) is unrestricted, however, the variance of α_i is fundamentally unidentified.

In many applications, it makes sense to restrict the dynamics of time-varying errors, which the researcher interprets as transitory, though persistent, shocks. Suppose in the simple linear model that v_{i1} and v_{i2} are assumed statistically independent. In addition, let us assume that the errors are statistically independent of the individual effects α_i. Then, a remarkable result due to Kotlarski

(1967) shows that, under weak technical conditions, the three distributions of α_i, v_{i1} and v_{i2} are nonparametrically point-identified.[15]

Kotlarski's result emphasizes the identification power of having repeated continuous outcomes and time-invariant individual effects. Here, the data provide the researcher with a bivariate continuous distribution– that of (y_{i1}, y_{i2})– while the three unknown distributions are univariate. Excess support in outcome variables is thus at the heart of the result.

Extensions. Kotlarski's insight has been generalized in several ways in relation to panel data modelling. Arellano and Bonhomme (2010) provide nonparametric identification results for the joint distribution of (α_i, β_i) and for the distribution of v_{it} in the random coefficients model (18). Their analysis requires $T \geq 3$, in order to recover a bivariate distribution of individual effects. Moreover, they show that the assumption of statistically independent errors which underlies Kotlarski's result can be relaxed when $T > 3$, and that it is possible to allow for moving average or autoregressive error structures with independent underlying disturbances.

Evdokimov (2010) uses Kotlarski's intuition in a different model. He considers a nonlinear regression model of the form:

$$y_{it} = g\left(x_{it}, \alpha_i\right) + v_{it}, \quad t = 1, 2, \tag{21}$$

where α_i is a scalar individual effects, v_{i1} and v_{i2} are independent, and the function $g()$ is unknown, weakly increasing in α_i. His analysis proceeds in three steps. First, Kotlarski's argument is applied conditionally on $x_{i1} = x_{i2}$ to recover the distribution of v_{it}. In a second step, a similar ("deconvolution") argument is applied given $x_{it} = x$ to recover the conditional distribution of $g\left(x_{it}, \alpha_i\right)$. Lastly, the function $g()$ and the distribution of α_i given covariates are identified using the monotonicity property of $g()$. The intuition behind this result is similar to Kotlarski: from the knowledge of a bivariate distribution of outcomes, it is possible to recover three univariate distributions (of α_i, v_{i1} and v_{i2}) together with the univariate function $\alpha \mapsto g(\alpha, x)$. Evdokimov's result thus also emphasizes the identification power of repeated continuous outcomes.

Functional differencing. Bonhomme (2010) considers a general parametric conditional model of outcomes given individual effects with parameter θ, as in Section 3. When outcomes are continuously distributed, the matrix $P_x(\theta)$ of conditional probabilities becomes a linear mapping, or *operator*, which maps function of q-dimensional vectors α, where q is the dimension of the vector of individual effects (e.g., $q = 1$ when α_i is scalar), to functions of T-dimensional vectors y. The image of a function $g(\alpha)$ by this operator is given by a function $L_{\theta,x} g$ of y such that:

$$\left[L_{\theta,x} g\right](y) = \int f_{y|x,\alpha}\left(y|x, \alpha; \theta\right) g(\alpha) d\alpha, \quad \text{for all } y. \tag{22}$$

---

[15]Independently, Horowitz and Markatou (1996) showed identification of these distributions under symmetry of transitory errors.

Bonhomme shows that a similar projection ("functional differencing") approach as in the discrete case can be applied in the continuous case, yielding a system of equations similar to (16). This approach provides conditional moment restrictions on θ that do not involve α_i. For these restrictions to be informative, it is necessary that the operator $L_{\theta,x}$ be *non-surjective*. In other words, its image must not span the whole space of functions of y. In the discrete case, this condition requires that the rows of the matrix $P_x(\theta)$ be linearly dependent, and is automatically satisfied provided the number of points of support of y_i exceeds that of α_i. In the continuous case, primitive conditions for non-surjectivity are given in the censored panel data model with random coefficients and normal disturbances, and a nonlinear regression model with independent errors. These examples lead to the conjecture that $L_{\theta,x}$ should be generally non-surjective provided $T > q$, i.e. provided the support of outcomes be richer than the support of individual effects.

Conditional independence restrictions. The analysis of Bonhomme (2010) is limited to setups where the distribution of y_i given x_i and α_i is parametric. In more general models where that distribution is not assumed to belong to a parametric family, conditional independence restrictions may be a powerful source of identification. In the related context of nonlinear measurement error models, general nonparametric identification results have been derived recently (Hu, 2008, Hu and Schennach, 2008, Hu and Shum, 2009). Nonlinear panel models often satisfy the type of conditional independence restrictions that these papers assume. For example, Hu and Shum (2009) exploit the identifying power of Markovian restrictions. An interesting aspect of their work is that they allow for general time-varying unobservables. These identification results, however, depend on high-level assumptions (such as operator injectivity). Providing primitive conditions in the context of specific panel data models seems of interest.

4.2.3 The role of covariates

In a fixed-effects setting, the conditional distribution of individual effects is left completely unspecified. This contrasts with the strong (e.g., parametric) assumptions that are often made on the conditional distribution of outcomes given individual effects and covariates. In empirical applications, researchers may be willing to impose some assumptions on the distribution of α_i given x_i, in order to gain some flexibility on the other part of the model.

Independent random-effects. A first example where restricting the distribution of α_i given x_i may be useful is given by discrete choice models. Assuming that some continuously distributed covariates are statistically independent of individual effects may be enough to ensure identification. In effect, this type of assumption allows to compensate for the lack of variation in the outcomes by using continuous variation in the regressors that is unrelated to the individual effects.

To see this, let us consider the binary choice model (13), where for simplicity we assume that individual effects have finite support. Assuming that x_i is statistically independent of α_i yields, using (15):

$$\Pr\left(y_i|x_i;\theta\right) = \sum_{k=1}^{K} P_k\left(y_i|x_i;\theta\right)\pi_k, \tag{23}$$

where now π_k does not depend on x_i. On the left-hand side of (23) there are as many frequencies as points in the *joint* support of (y_i, x_i), while there are $K + \dim\theta$ unknown parameters. Therefore, there will be room for point-identification of θ when K is small relative to the number of points of support of (y_i, x_i). Even though y_{it} is binary and T is small, θ may thus be point-identified in the presence of a rich distribution of individual effects (i.e, a large K), provided the support of x_i be sufficiently rich.[16]

Under this type of independence restrictions, panel data are actually not needed for identification, and a cross-section of data is enough. Beran and Hall (1992) and Hoderlein *et al.* (2010) apply this idea to a cross-sectional linear model with random coefficients, while Gautier and Kitamura (2009) treat the case of a binary choice model. Honoré and Lewbel (2002) use a similar strategy to deal with a panel data binary choice model with predetermined regressors, where one of the regressors is assumed independent of α_i given the other covariates. They discuss some empirical applications where this assumption makes sense. The plausibility of this type of excluding restrictions should be argued on a case-by-case basis, just as the validity of instrumental variables in general.

Exchangeability. Finally, the identification strategy in Altonji and Matzkin (2005) also exploits restrictions on the conditional distribution of individual effects. Their approach relies on an assumption of *exchangeability*, according to which the conditional distribution of unobservables (including the individual effects) given the sequence of regressors $x_{i1}, ..., x_{iT}$ does not depend on the order in which the x_{it} appear. Exchangeability is natural in sibling applications, less so in proper panel data applications where t indexes time and dependence of x_{it} over time may be expected.

To mention a simple case where exchangeability holds, let us consider a general response model:

$$y_{it} = g\left(x_{it}, \alpha_i, v_{it}\right), \quad t = 1, 2, \tag{24}$$

where we suppose that the α_i and v_{it} are independent of x_{it} given the mean covariates $\overline{x}_i$, and where $g()$ is an unspecified function. For simplicity, we also suppose that x_{it} is binary, although the approach can be generalized to continuous x's. In this case, the average marginal effect of an increase of x_{i1}

---

[16]See Kasahara and Shimotsu (2009) for related work concerning structural dynamic discrete choice models with heterogeneity.

from 0 to 1 is point-identified as, for example:

$$
\begin{aligned}
\Delta_1 &= \mathbb{E}\left[g\left(1, \alpha_i, v_{i1}\right) - g\left(0, \alpha_i, v_{i1}\right) | x_{i1} = 1\right] \\
&= \mathbb{E}\left[y_{i1} | x_{i1} = 1\right] - \mathbb{E}\left[\mathbb{E}\left(g\left(0, \alpha_i, v_{i1}\right) | x_{i1} = 1, \overline{x}_i\right) | x_{i1} = 1\right] \\
&= \mathbb{E}\left[y_{i1} | x_{i1} = 1\right] - \mathbb{E}\left[\mathbb{E}\left(g\left(0, \alpha_i, v_{i1}\right) | x_{i1} = 0, \overline{x}_i\right) | x_{i1} = 1\right] \\
&= \mathbb{E}\left[y_{i1} | x_{i1} = 1\right] - \mathbb{E}\left[\mathbb{E}\left(y_{i1} | x_{i1} = 0, \overline{x}_i\right) | x_{i1} = 1\right],
\end{aligned}
\tag{25}
$$

where in the third equality we have used the conditional independence assumption:

$$
\mathbb{E}\left[g\left(0, \alpha_i, v_{i1}\right) | x_{i1} = 1, \overline{x}_i\right] = \mathbb{E}\left[g\left(0, \alpha_i, v_{i1}\right) | x_{i1} = 0, \overline{x}_i\right].
$$

Similarly, the average effect given $x_{i1} = 0$ is also-point identified, and so is the unconditional average.

Note that this situation contrasts with the binary choice model (17) that we analyzed above, where the unconditional average is not point-identified. Exchangeability (in this case, conditional independence given the mean covariates), when justified, thus appears as a useful source of identification.

Note also that repeated outcomes are not needed for the Altonji-Matzkin strategy to work. Data on the first period outcomes and regressors and the mean regressor $\overline{x}_i$ will be sufficient to estimate an empirical counterpart of (25). This makes this approach very different from the other approaches to identification that we mentioned in this subsection, where repeated observations on y_{it} is often essential.

5   When the choice of population framework matters

Due to the incidental parameter problem, many natural estimation approaches– such as maximum likelihood– lead to inconsistent estimators in the presence of fixed effects, when the number of time periods T is fixed. In this perspective, point-identification itself is sometimes problematic. In many applications, however, the time dimension T of the panel is not negligible relative to its cross-sectional dimension N. Following this insight, it has been recently argued that the incidental parameter problem may be viewed as time-series finite sample bias. This change of population framework yields a very different view of panel data estimators, which we now review.

5.1   Non fixed-T asymptotic properties

Let us consider the same setup as in Section 3, based on a parametric likelihood function for y_i given x_i and α_i, characterized by a parameter θ, and an unrestricted conditional distribution for α_i given x_i. Following Arellano and Bonhomme (2009), we note that many estimation approaches to θ are based on an average likelihood that assigns weights to different values of α_i:

$$
f^a(y_i | x_i; \theta) = \int f_{y|x,\alpha}(y_i | x_i, \alpha; \theta)\omega_i\left(\alpha\right) d\alpha,
\tag{26}
$$

where the weights $\omega_i(\alpha)$ may depend on θ and exogenous covariates x_i.

The random-effects integrated likelihood (5) of Section 3 is a first example. In this case, $\omega_i(\alpha)$ depends on hyperparameters ξ, and possibly on exogenous covariates or initial conditions. However, it is independent of θ. In a Bayesian approach, (26) will be the marginal likelihood on which to base inference, provided the researcher has assumed prior conditional independence of individual effects given θ. In this situation, $\omega_i(\alpha)$ is understood as the prior distribution of α_i given θ.

A fixed-effects approach that estimates $\alpha_1, ..., \alpha_N$ jointly with θ by maximum likelihood is another special case of the average likelihood representation (26). To see this, note that the fixed-effects maximum likelihood estimator is given by:

$$
\begin{aligned}
\widehat{\theta}_{ML} &= \operatorname*{argmax}_{\theta, \alpha_1, ..., \alpha_N} \sum_{i=1}^{N} \log f_{y|x,\alpha}(y_i|x_i, \alpha_i; \theta) \\
&= \operatorname*{argmax}_{\theta} \sum_{i=1}^{N} \log f_{y|x,\alpha}(y_i|x_i, \widehat{\alpha}_i(\theta); \theta),
\end{aligned}
$$

where $\widehat{\alpha}_i(\theta) = \operatorname{argmax}_{\alpha_i} \log f_{y|x,\alpha}(y_i|x_i, \alpha_i; \theta)$. It thus follows that fixed-effects maximum likelihood may be interpreted as an average likelihood approach, where the weight $\omega_i(\alpha)$ assigns all mass to the fixed-effects estimate $\widehat{\alpha}_i(\theta)$.

When T is fixed and N tends to infinity, all these average likelihood approaches will generally be inconsistent, as a consequence of the incidental parameter problem. This inconsistency reflects a poor finite-sample performance when T is very small relative to N. However, in many panel data applications, the ratio T/N is not negligible. For example, in panel growth applications (e.g., Caselli et al., 1996) N and T are of similar orders of magnitude. In microeconometric applications, the widely used PSID dataset now has a total of $T \approx 30$ years for a few thousands individuals. In these applications, it makes sense to consider an alternative asymptotic experiment where N and T tend jointly to infinity.

In this perspective, let us consider the properties of a generic average likelihood estimator:

$$
\widehat{\theta} = \operatorname*{argmax}_{\theta} \sum_{i=1}^{N} \log f^a(y_i|x_i; \theta), \tag{27}
$$

where $f^a(y_i|x_i; \theta)$ is given by (26) for some weights $\omega_i(\alpha)$. Under standard regularity conditions, the probability limit θ_T of $\widehat{\theta}$ as N tends to infinity may be expanded as follows:

$$
\theta_T = \theta + \frac{B}{T} + O\left(\frac{1}{T^2}\right). \tag{28}
$$

In general, θ_T differs from the true value θ of the parameter, reflecting the fact that $\widehat{\theta}$ is inconsistent for fixed T. Moreover, the first-order bias B/T is generally non-zero. The recent panel data literature on non fixed-T asymptotics emphasizes the possibility to remove that term, so that the resulting bias-reduced estimator has a bias of order $1/T^2$, as opposed to $1/T$. When T is moderately large, bias reduction may improve the finite-sample properties of $\widehat{\theta}$ substantially.

Asymptotic distribution. An interesting property of panel data estimators is that bias reduction happens with no increase in the asymptotic variance as T/N tends to a constant. When conditions are met that ensure that the average likelihood estimator is asymptotically normal, we will have:

$$\sqrt{NT}\left(\widehat{\theta} - \theta_T\right) \overset{d}{\to} \mathcal{N}\left(0, V\right), \tag{29}$$

where (under general conditions) V is the large-T asymptotic variance of the maximum likelihood estimator. A bias-reduced estimator $\widehat{\theta}^R$ will satisfy:

$$\widehat{\theta}^R = \widehat{\theta} - \frac{B}{T} + o_p\left(1\right).$$

So:

$$\begin{aligned} \widehat{\theta}^R - \theta &= \widehat{\theta} - \theta - \frac{B}{T} + o_p\left(1\right) \\ &= \widehat{\theta} - \theta_T + o_p\left(1\right). \end{aligned}$$

It thus follows that, as T and N tend to infinitely simultaneously such that T/N tends to a non-zero constant:

$$\sqrt{NT}\left(\widehat{\theta}^R - \theta\right) \overset{d}{\to} \mathcal{N}\left(0, V\right),$$

where V is the same asymptotic variance as in (29). This situation contrasts with the pure time-series case, where bias reduction is usually associated with variance inflation.

Note that if N/T tends to 0, then:

$$\sqrt{NT}\left(\widehat{\theta} - \theta\right) \overset{d}{\to} \mathcal{N}\left(0, V\right),$$

whereas if $N/T^3 \to 0$

$$\sqrt{NT}\left(\widehat{\theta} - \theta - \frac{B}{T}\right) \overset{d}{\to} \mathcal{N}\left(0, V\right).$$

To obtain sufficiently accurate confidence intervals from this type of asymptotic approximation, the bias should be small relative to the standard deviation. For first-order bias corrected estimators, this requires that N be small relative to T^3 (for example, N small relative to $1,000$ or to $8,000$ for $T = 10$ or 20, respectively).

Non-standard situations. Before reviewing the various approaches to bias reduction used in the literature, it is interesting to mention two cases where the asymptotic expansion (28), on which standard bias reduction techniques are based, is invalid. The first case corresponds to non-regular models, where the objective function is not smooth. An example is the panel data quantile regression estimator (Koenker, 2004). Galvao *et al.* (2010) have recently argued that quantile estimates satisfy non-standard asymptotic expansions, so that usual bias reduction approaches cannot be applied in this context.

The second case arises in the presence of time effects: unlike the fixed-T case, allowing for time dummies is difficult here, as there is a "double" incidental parameter problem when N (the number of individual effects) and T (the number of time effects) grow simultaneously. This will have an effect on the asymptotic expansion of average likelihood estimates: in addition to a $O(1/T)$ term that reflects the estimation of individual effects, the expansion also involves a term of the order $1/N$ which is due to the estimation of the time effects (Fernández-Val and Weidner, 2010).

5.2   Varieties of bias reduction

The first class of approaches to reduce the bias of average likelihood estimators is based on analytical methods. Suppose, indeed, that we have constructed a consistent estimator of B in (28).[17] Then, the following estimator:

$$\widehat{\theta}^R = \widehat{\theta} - \frac{\widehat{B}}{T}$$

has a small bias by construction.

Analytical calculations of B are available (Hahn and Newey, 2004), and may be used to construct an empirical counterpart $\widehat{B}$. Moreover, a similar approach can be used to reduce the bias of the estimating equations– as opposed to reducing the bias of the estimator– and the bias of the objective function. Arellano and Hahn (2007) provide a thorough review of the literature based on analytical bias corrections.

Bias-reducing priors.   A different approach to bias reduction is introduced in Arellano and Bonhomme (2009). Focusing on estimators that maximize an average likelihood with weights $\omega_i(\alpha)$, they ask the following question: how should one choose the weights so that the average likelihood estimator be unbiased to first-order, so that the B term in (28) be zero? A feature of their analysis is that it covers fixed-effects, random-effects and Bayesian approaches in a single framework.

As shown by Sweeting (1987) in his discussion of Cox and Reid (1987)'s classic paper, the answer is simple when θ and α_i are *information orthogonal*, i.e. when the likelihood function satisfies:

$$\mathbb{E}\left(\frac{\partial^2 \log f_{y|x,\alpha}\left(y_i|x_i,\alpha_i;\theta\right)}{\partial\theta\partial\alpha_i'}\right) = 0. \tag{30}$$

When parameters are information orthogonal, choosing uniform weights $\omega_i(\alpha) = 1$ will lead to first-order unbiasedness. More generally, taking as weight– or prior– for α_i any distribution that is independent of θ will also be bias-reducing for the average likelihood estimator.

Lancaster (2002) noted that a similar approach can be applied in situations where there exists a reparameterization of the effects where information orthogonality (30) is satisfied. Orthogonal reparametrizations exist in first-order linear autoregressive models and in static binary choice, for

---

[17]Here, "consistency" is understood as N and T tend to infinity.

example. In these models, using as weight (or prior) the Jacobian of the reparameterization will lead to bias-reduction.

However, orthogonal reparameterizations do not always exist. For example, in most of the models of applied interest reviewed in Section 2, Lancaster (2002)'s approach will not be applicable. Arellano and Bonhomme (2009) provide general conditions for weights to lead to biased-reduced estimates. In the absence of orthogonal reparameterizations, the bias-reducing weights– or "priors"– depend on the distribution of the data. In particular, they show that using as weights the normal approximation to the sampling distribution of the estimated individual effects given θ:

$$\omega_i(\alpha) = \frac{1}{\sqrt{\widehat{\mathrm{Var}}\left[\widehat{\alpha}_i(\theta)\right]}} \phi\left(\frac{\alpha - \widehat{\alpha}_i(\theta)}{\sqrt{\widehat{\mathrm{Var}}\left[\widehat{\alpha}_i(\theta)\right]}}\right) \tag{31}$$

leads to bias-reduction. Using (31) is intuitive: when individual effects are precisely estimated given θ, then the weights are tightly concentrated around the maximum likelihood estimate, while when $\widehat{\alpha}_i(\theta)$ is imprecise, the attached weight or prior on α_i is very diffuse. Moreover, this choice of weights represents a general solution to bias reduction that does not rely on parameter orthogonality.

The average likelihood approach based on bias-reducing weights such as (31) is computationally attractive, as it relies on simulation rather than integration. Just as in the random-effects approach reviewed in Section 3, one may generate a Markov chain of parameter draws and compute the estimate as the posterior mode or mean of the chain. In addition, frequentist asymptotic confidence intervals can be directly read on the posterior distribution.[18]

Automatic approaches. In addition to analytical approaches and weighted likelihood approaches based on suitable weights, the recent literature has emphasized automatic approaches to bias reduction. In static panel data models, Hahn and Newey (2004) propose to use the delete-one jackknife. The split-panel jackknife method of Dhaene and Jochmans (2010), which we now review, allows for dynamics. The idea is to estimate the average likelihood estimator $\widehat{\theta}$ on the two subsamples $[1, T/2]$ and $[T/2+1, T]$ (assuming T even for simplicity). Let $\widehat{\theta}_1$ and $\widehat{\theta}_2$ denote the two estimates, and let $\widehat{\theta}$ denote the estimate based on the full sample. The first-order bias term of $\widehat{\theta}_1$ is $B/(T/2) = 2B/T$, while that of $\widehat{\theta}$ is B/T. It thus follows that:

$$\widehat{\theta}^R = 2\widehat{\theta} - \frac{\widehat{\theta}_1 + \widehat{\theta}_2}{2} \tag{32}$$

is unbiased to first order.

The available evidence on the finite-sample performance of the various approaches to bias reduction is encouraging. In static and dynamic settings (e.g., Carro, 2007), these techniques tend to remove at least half of the bias, while keeping the variance virtually unchanged. An issue concerns the possibility

---

[18]The large-T validity of this inference method relies on the properties of the pseudo-Bayesian approach of Chernozhukov and Hong (2003). See Arellano and Bonhomme (2009).

to reduce the bias further. Second-order bias reduction can be simply implemented using a variant of the split-panel jackknife approach. However, the Monte Carlo evidence presented in Dhaene *et al.* (2010) suggests that higher-order bias reduction may be associated with increased variance.

To conclude this brief review, there is so far too little comparison of the various bias reduction approaches on simulated data. Moreover, although panel data bias reduction has been used in some empirical applications (e.g., Hospido, 2010, Fernández-Val and Vella, 2009), more applications are needed.

5.3 The bias of random-effects estimators

Random-effects specifications are an important special case of average likelihood approaches. It is thus interesting to study their properties when N and T tend to infinity simultaneously.

Arellano and Bonhomme (2009) provide conditions for the first-order bias term of random-effects maximum likelihood estimates to be zero. The conditions they find are quite restrictive. They are satisfied by Gaussian random-effects when the model is linear. However, in nonlinear models, usual random-effects specifications (e.g. Gaussian, Gamma) lead very generally to the presence of a first-order bias.

In addition, they show that the bias term B/T in (28) is a function of the distance (in a Kullback-Leibler sense) between the population cross-sectional density of the individual effects and its best approximation in the parametric family $f_{\alpha|x}(.|x_i; \xi)$. As a special case, the bias is zero when the population density belongs to the random-effects family. Indeed, in this case the random-effects maximum likelihood estimator is fixed-T consistent.

This characterization suggests that one may achieve bias reduction by letting the parametric distribution of individual effects become increasingly flexible as N increases. In a model without covariates, Arellano and Bonhomme (2009) model the distribution of α_i as a mixture of normals with an increasing number of components, and show that the resulting sieve random-effects estimator of θ is unbiased to first order. In the presence of covariates, however, achieving the required level of "flexibility" so as to remove the first-order bias on the parameter of interest seems challenging.

Average marginal effects. In applications, the researcher is usually not interested only in common parameters θ, but also in average marginal effects. In Section 3, we have mentioned two different ways to estimate average marginal effects. When N and T grow simultaneously, and when the population distribution of individual effects does not necessarily belong to the postulated random-effects family, these two estimation approaches have strikingly different properties.

The classical estimate relies on the postulated distribution $f_{\alpha|x}(.|x_i; \xi)$, see equation (9). When that distribution is misspecified, the estimate is generally inconsistent as T tends to infinity. To give a simple example, suppose that the prior specification imposes that α_i and x_i are independent. Then a

classical estimate of the correlation between α_i and any component of x_i will be zero by construction. This will be true even if α_i and x_i are not independent in the population, and no matter how large T is.

In contrast, the posterior mean (or mode) of the average marginal effect uses the random-effects model as a prior specification that is updated using data information. As a result, the posterior mean will be consistent as T tends to infinity. In the previous example, as T grows, the posterior mean of the correlation between α_i and x_i will increasingly reflect the degree of correlation that exists in the population.[19]

This discussion highlights an interesting contrast. Random-effects estimates of structural parameters θ are always consistent as T tends to infinity, even under incorrect specification. This is due to the fact that the "prior" information embodied in the random-effects distribution vanishes as T increases. However, only posterior means (or modes) of average marginal effects remain consistent for large T under misspecification. This is because, unlike the classical estimates, the posterior mean updates the prior information by using the data.

Lastly, although they are consistent as T tends to infinity, posterior mean estimates of average marginal effects suffer from a first-order bias in general. Arellano and Bonhomme (2009) derive the expression of bias-reducing weights in this case. As in the analytical approach of Hahn and Newey (2004), the bias-reducing weights depend on the form of the marginal effect. Here also, in the absence of covariates, a sieve random-effects approach will remove the first-order bias.

6 Conclusion

Linearity and homogeneity assumptions are rarely justified by the economics of a problem. The availability of panel data makes it conceptually possible to tackle the issues of nonlinearity and unobserved heterogeneity simultaneously. However, the analysis of nonlinear panel data models remains a challenge for econometricians. This survey presents some recent advances in this area.

A general approach is to postulate a parametric model, which includes in particular a model for the distribution of the unobserved individual effects. We have emphasized the relationship between classical "random-effects" approaches and Bayesian computation techniques, as we think that Markov Chain Monte Carlo methods are convenient tools for estimation and inference in this context.

Random-effects methods are parametric in nature. Relaxing this assumption and leaving the conditional distribution of individual effects unrestricted raises an identification challenge. There is growing evidence that discrete-choice panel data models are partially identified in general. In this

---

[19]Consider the model $y_{it} = \rho y_{i,t-1} + \alpha_i + v_{it}$, where v_{it} is i.i.d. Gaussian $(0, \sigma^2)$, and ρ and σ^2 are known. Assuming a random-effects specification according to which α_i is Gaussian $(0, \sigma_\alpha^2)$ independent of y_{i0}, it can be shown that the posterior mean of $M = \mathbb{E}[\alpha_i y_{i0}]$ is $M/(1 + \frac{\sigma^2}{T\sigma_\alpha^2})$. In contrast, the classical random-effects estimate of M is zero, therefore inconsistent as T tends to infinity.

context, estimation and inference methods are needed. In models with continuous outcomes, however, panel data offer opportunities for point-identification that, for a large part, remain to be explored.

References

[1] Abrevaya, J., and C. M. Dahl (2008): "The Effects of Birth Inputs on Birthweight", *Journal of Business & Economic Statistics*, 26, 379–397.

[2] Altonji, J. G., A. Smith, and I. Vidangos (2009): "Modeling Earnings Dynamics", NBER Working Paper 14743.

[3] Altonji, J. G., and R. L. Matzkin (2005): "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors," *Econometrica*, 73(4), 1053–1102.

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

[5] Andersen, E.B. (1970): "Asymptotic Properties of Conditional Maximum Likelihood Estimators," *Journal of the Royal Statistical Society B*, 32, 283-301.

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

[7] Arellano, M., and S. Bonhomme (2009): "Robust Priors in Nonlinear Panel Data Models", *Econometrica*, 77, 489–536.

[8] Arellano, M., and S. Bonhomme (2010): "Identifying Distributional Characteristics in Random Coefficients Panel Data Models", unpublished manuscript.

[9] Arellano, M., and R. Carrasco (2003): "Binary Choice Panel Data Models with Predetermined Variables," *Journal of Econometrics*, 115(1), 125–157.

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

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

[12] Bester, A., and C. Hansen (2007): "Flexible Correlated Random Effects Estimation in Panel Models with Unobserved Heterogeneity," unpublished manuscript.

[13] Beran, R., and P. Hall (1992): "Estimating Coefficient Distributions in Random Coefficient Regressions," *Annals of Statistics*, 20(4), 1970–1984.

[14] Blundell, R., R. Griffith, and F. Windmeijer (2002): "Individual Effects and Dynamics in Count Data Models," *Journal of Econometrics*, 108(1), 113–131.

[15] Blundell, R., and J. L. Powell (2004): "Endogeneity in Semiparametric Regression Models", *Review of Economic Studies*, 71(3), 581–913.

[16] Bonhomme, S. (2010): "Functional Differencing," unpublished manuscript.

[17] Browning, M., and J. Carro (2009): "Dynamic Binary Outcome Models with Maximal Heterogeneity", unpublished manuscript.

[18] Butler, J. S., and R. Moffitt (1982): "A Computationally Efficient Quadrature Procedure for the One-Factor Multinomial Probit Model," *Econometrica*, 50(3), 761–64.

[19] Carro, J. (2007): "Estimating Dynamic Panel Data Discrete Choice Models with Fixed Effects", *Journal of Econometrics*, 140, 503–528.

[20] Caselli, F., G. Esquivel, and F. Lefort (1996): "Reopening the Convergence Debate: A New Look at Cross-Country Growth Empirics", *Journal of Economic Growth*, 1, 363–389.

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

[22] Chamberlain, G. (1982): "Multivariate Regression Models for Panel Data ," *Journal of Econometrics*, 18(1), 5–46.

[23] Chamberlain, G. (1992): "Efficiency Bounds for Semiparametric Regression", *Econometrica*, 60, 567–596.

[24] Chamberlain, G. (1993): "Feedback in Panel Data Models", unpublished manuscript, Department of Economics, Harvard University.

[25] Chamberlain, G. (2010): "Binary Response Models for Panel Data: Identification and Information", *Econometrica*, 78, 159–168.

[26] Chen, X. (2007): "Large Sample Sieve Estimation of Semi-nonparametric Models," chapter 76 in *Handbook of Econometrics*, vol. 6B, 2007, eds. J.J. Heckman and E.E. Lleamer, North-Holland.

[27] Chernozhukov, V., I. Fernández-Val, and W. Newey (2009): "Quantile and Average Effects in Nonseparable Panel Models," unpublished manuscript.

[28] Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2007): "Identification and Estimation of Marginal Effects in Nonlinear Panel Models," unpublished manuscript.

[29] Chernozhukov, V., J. Hahn, and W. Newey (2005): "Bound Analysis in Panel Models with Correlated Random Effects," unpublished manuscript.

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

[31] Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models ," *Econometrica*, 75(5), 1243–1284.

[32] 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.

[33] Dhaene, G., and K. Jochmans (2010): "Split-Panel Jackknife Estimation of Fixed-Effect Models,"' unpublished manuscript.

[34] Evdokimov, K. (2010): "Identification and Estimation of a Nonparametric Panel Data Model with Unobserved Heterogeneity," unpublished manuscript.

[35] Fernández-Val, I., and F. Vella (2009): "Bias Corrections for Two-Step Fixed Effects Panel Data Estimators," unpublished manuscript.

[36] Fernández-Val, I., and M. Weidner (2010): "Individual and Time Effects in Nonlinear Panel Data Models with Large N,T," unpublished manuscript.

[37] Galvao, A., K. Kato, and G. Montes-Rojas (2010): "Asymptotics and Bootstrap Inference for Panel Quantile Regression Models with Fixed Effects," unpublished manuscript.

[38] Gautier, E., and Y. Kitamura (2009): "Nonparametric Estimation in Random Coefficients Binary Choice Models," unpublished manuscript.

[39] Geweke, J. (1989): "Bayesian Inference in Econometric Models Using Monte Carlo Integration", *Econometrica*, 57, 1317–1339.

[40] Geweke, J., and M. Keane (2000): "An Empirical Analysis of Earnings Dynamics Among Men in the PSID: 1968-1989," *Journal of Econometrics*, 96(2), 293–356.

[41] Graham, B.S., and J.L. Powell (2008): "Identification and Estimation of Irregular Correlated Random Coefficient Models," unpublished manuscript.

[42] Hahn, J. (2001): "Comment: Binary Regressors in Nonlinear Panel-Data Models with Fixed Effects", *Journal of Business & Economic Statistics*, 19(1), 16–17.

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

[44] Hajivassiliou, V. A., and D. L. McFadden (1998): "The Method of Simulated Scores for the Estimation of LDV Models", *Econometrica*, 66(4), 863–896.

[45] Hansen, L. P., and K. J. Singleton (1982): "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models", *Econometrica*, 50(5), 1269–86.

[46] Heckman, J. J. (1981): "The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time-Discrete Data Stochastic Process and Some Monte Carlo Evidence," in *Structural Analysis of Discrete Data*, D. McFadden and C. Manski eds. MIT Press. Cambridge.

[47] Hirano, K. (2002): "Semiparametric Bayesian Inference in Autoregressive Panel Data Models," *Econometrica*, 70(2), 781–799.

[48] Hoderlein, S., Klemelä, J., and E. Mammen (2010): "Analyzing the Random Coefficient Model Nonparametrically", *Econometric Theory*, 26(3), 804–837.

[49] Hoderlein, S., and H. White (2010): "Nonparametric Identification in Nonseparable Panel Data Models with Generalized Fixed Effects," unpublished manuscript.

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

[51] Honoré, B., and A. Lewbel (2002): "Semiparametric Binary Choice Panel Data Models without Strictly Exogenous Regressors," *Econometrica*, 70, 2053–2063.

[52] Honoré, B., and E. Tamer (2006): "Bounds on Parameters in Panel Dynamic Discrete Choice Models," *Econometrica*, 74, 611–629.

[53] Horowitz, J. L., and M. Markatou (1996): "Semiparametric Estimation of Regression Models for Panel Data", *Review of Economic Studies*, 63, 145–168.

[54] Hospido, L. (2010): "Modelling Heterogeneity and Dynamics in the Volatility of Individual Wages", *Journal of Applied Econometrics*, forthcoming.

[55] Hu, Y. (2008): "Identification and Estimation of Nonlinear Models with Misclassification Error Using Instrumental Variables: A General Solution," *Journal of Econometrics*, 144(1), 27–61.

[56] Hu, Y., and S. M. Schennach (2008): "Instrumental Variable Treatment of Nonclassical Measurement Error Models," *Econometrica*, 76(1), 195-216.

[57] Hu, Y., and M. Shum (2009): "Nonparametric Identification of Dynamic Models with Unobserved State Variables," unpublished manuscript.

[58] Judd, K. (1998): *Numerical Methods in Economics*, MIT Press. Cambridge, London.

[59] Kasahara, H., and K. Shimotsu (2009): "Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices," *Econometrica*, 77(1), 135–175.

[60] Koenker, R. (2004): "Quantile Regression for Longitudinal Data," *Journal of Multivariate Analysis*, 91(1), 74–89.

[61] Kotlarski, I. (1967): "On Characterizing the Gamma and Normal Distribution," *Pacific Journal of Mathematics*, 20, 69–76.

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

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

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

[65] Rossi, P., R. E. McCulloch, and G. M. Allenby (1995): "Hierarchical Modelling of Consumer Heterogeneity: An Application to Target Marketing," in Kass and Singpurwalla: *Case Studies in Bayesian Statistics*. New York: Springer Verlag, 323–50.

[66] Shen, X. (1997): "On Methods of Sieves and Penalization," *The Annals of Statistics*, 25, 2555–2591.

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

[68] Tamer, E. (2010): "Partial Identification in Econometrics," *Annual Review of Economics*, 2, 167–195.

[69] Van der Vaart, A. W. (2007): *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics.

[70] West, M., Müller, P., and M. D. Escobar (1994): *Hierarchical priors and mixture models, with application in regression and density estimation,* in Aspects of Uncertainty: A Tribute to D. V. Lindley, A.F.M. Smith and P. Freeman eds. Wiley.

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

[72] Wooldridge, J. (1997): "Multiplicative Panel Data Models without the Strict Exogeneity Assumption", *Econometric Theory* 13, 667–678.

[73] Wooldridge, J. (1999): "Distribution-Free Estimation of Some Nonlinear Panel Data Models", *Journal of Econometrics*, 90, 77–97.

[74] Wooldridge, J. (2005): "Simple Solutions to the Initial Conditions Problem in Dynamic, Nonlinear Panel Data Models with Unobserved Heterogeneity," *Journal of Applied Econometrics*, 20(1), 39–54.