Responses to the Incidental Parameter Problem

Dit proefschrift is tot stand gekomen in het kader van EDE-EM (European Doctorate in Economics – Erasmus Mundus), met als doel het behalen van een gezamenlijk doctoraat. Het proefschrift is voorbereid aan de Faculteit Economie en Bedrijfskunde van de Universiteit van Amsterdam en aan de Center for Operations Research and Econometrics van de Université Catholique de Louvain.

La thèse a été préparée dans le cadre du programme doctoral européen EDE-EM (European Doctorate in Economics – Erasmus Mundus). Cette thèse a été préparée conjointment au Faculteit Economie en Bedrijfskunde, Universiteit van Amsterdam et au Center for Operations Research and Econometrics, Université Catholique de Louvain.

This thesis has been written within the framework of the EDE-EM (European Doctorate in Economics – Erasmus Mundus), with the purpose of obtaining a joint doctorate degree. The thesis was prepared in the Faculty of Economics and Business at the University of Amsterdam and in the Center for Operations Research and Econometrics at the Université Catholique de Louvain.

Layout and cover design by Andrew Adrian Yu Pua

ISBN 978-94-91030-84-0
NUR 916

RESPONSES TO THE INCIDENTAL PARAMETER PROBLEM

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor

aan de Universiteit van Amsterdam

op gezag van de Rector Magnificus

prof. dr. D. C. van den Boom

ten overstaan van een door het College voor Promoties ingestelde

commissie, in het openbaar te verdedigen in de Agnietenkapel

op donderdag 10 maart 2016, te 14:00 uur

door

Andrew Adrian Yu Pua

geboren te Manilla, Filipijnen

Promotiecommissie:

Promotor:	Prof. dr. H. P. Boswijk	Universiteit van Amsterdam
	Prof. dr. S. van Bellegem	Université Catholique de Louvain
Copromotor:	Dr. M. J. G. Bun	Universiteit van Amsterdam
Overige leden:	Prof. dr. G. Dhaene	Katholieke Universiteit Leuven
	Dr. K. J. van Garderen	Universiteit van Amsterdam
	Dr. N. P. A. van Giersbergen	Universiteit van Amsterdam
	Prof. dr. C. M. Hafner	Université Catholique de Louvain
	Prof. dr. S. Khan	Duke University
	Prof. dr. F. R. Kleibergen	Universiteit van Amsterdam

Faculteit: Economie en Bedrijfskunde

Acknowledgements

I acknowledge the funding and support of the Education, Audiovisual and Culture Executive Agency (EACEA) of the European Union during my stay in Europe from September 2009 to August 2014. The agency financed both my scholarship for the Erasmus Mundus Master Course QEM and my fellowship for the Erasmus Mundus Joint Doctorate EDEEM. I also thank my promotor Peter Boswijk for offering a teaching gig that allowed me to stay at the University of Amsterdam until 1 February 2016.

I would like to thank six sets of people: my family, my friends, my colleagues, the participants at talks, the support staff, and the nameless future reader.

First, I spent most of my time with colleagues at the University of Amsterdam (UvA) and at the Center for Operations Research and Econometrics (CORE). I thank my promotors, Peter Boswijk and Sébastien van Bellegem, for all the talks, discussions, and the candidness. I also thank Maurice Bun for his patience in going through the manuscript. They have decided to trust me and I hope I was able to deliver. I also thank my doctoral committee for taking the time to read my manuscript. Their comments have been useful in rethinking about the approaches I considered in the thesis. Let me also single out members of my doctoral committee – Geert Dhaene, Shakeeb Khan, and Frank Kleibergen, for their support in my job search.

Second, I thank all the people who have attended my talks or listened to my ideas (either forced or of their own volition). Let me single out people who have offered some perspective through their comments – Luc Bauwens, Stéphane Bonhomme, Simon Broda, Martin Carree, Pavel Čížek, Geert Dhaene, Firmin Doko Tchatoka, Jianqing Fan, Kees Jan van Garderen, Noud van Giersbergen, Refet Gürkaynak, Christian Hafner, Harry Haupt, Artūras Juodis, Shakeeb Khan, Jan Kiviet, Frank Kleibergen, Thierry Magnac, Michael Massmann, Salvador Navarro, Serena Ng, Cavit Pakel, Dale Poirier, Renata Rabovič, Douglas Steigerwald, Martin Weidner, Frank Windmeijer, and Jeffrey Wooldridge. I also thank Roy van der Weide for sharing the data used in Chapter 5.

Third, I thank all my friends for their support, even if I am usually not around. Most of my friends are back home in the Philippines and I thank them for making my return home so much fun. I also thank the EDEEM cohort for their help in administrative matters.

Fourth, the support staff at UvA and CORE have made smooth transitions possible. Arnold van Meteren was one of my earliest contacts at UvA. He was responsible for facilitating my long-stay visa application in the Netherlands. José Kiss was very helpful in facilitating accommodation in Amsterdam and registration at the UvA. Kees Nieuwland made office life smoother by being there for computer-related issues. Jolanda Vroons also took his place as IT liaison and was very quick to respond. Evelien Brink, Ana Colic, Wilma de Kruijf, and Robert Helmink are always there to

help whenever I would need assistance. Marc van Steekelenburg has been helpful in dealing with renewing my residence permit. Catherine Germain is possibly one of the best multi-taskers I have ever seen in action. She helped in smoothing out my move to Belgium, dealing with French-speaking authorities, and expediting the final activities of the dissertation defense phase. Marie-Hélène Chassagne has also been very helpful with these final activities as well. Raphaël Tursis was one of the nicer IT guys I have met. I also thank Caroline Dutry, the only support staff at the coordinating institution of the doctoral programme, for dealing with both administrative and finance-related issues. The support staff is really the heart of any institution!

Fifth, I thank the reader of this thesis. I hope you enjoy reading this work just as I have enjoyed (though not without heartbreak) working on it. In case you did not notice, the last few pages of the dissertation are blanks meant for notes.

Finally, I thank my mother for understanding the nature of what I have been doing for the past years, despite her initial hesitations. I thank my brother and sister for being there with my mother in my absence. Although infuriating at times, I would like to thank the cats and our lone dog back in our house, as they have stabilized the household. I thank my better half Stephanie for being one of the constants in my life.

Contents

Chapter 1

Introduction

1.1 The promise of panel data

In this chapter, I show through a series of examples that panel data offer researchers three broad but sometimes competing advantages – estimating structural or common parameters more precisely, allowing for dynamics and feedback, and control of time-invariant unobserved heterogeneity. I am working within usual panel data context where the cross-sectional units i are independently sampled.

Let $y_i^t = (y_{i1}, \ldots, y_{it})$ and $x_i^t = (x_{i1}, \ldots, x_{it})$ for $i = 1, \ldots, n$ and $t = 1, \ldots, T$. The variable y_{it} is the outcome of interest and x_{it} is a vector of regressors – both of which are observable. We are interested in the conditional distribution of the observables y_i^T given x_i^T, which is indexed by a finite-dimensional parameter θ. Unfortunately, the presence of the unobservable α_i, which is an individual-specific effect capturing time-invariant unobserved heterogeneity potentially correlated with the regressors, obscures our ability to estimate and make inferences about θ. To see this, consider a prototypical panel data model where the previously mentioned elements can be found in the following integral equation, i.e.,

$$f_{y|x}\left(y_i^T|x_i^T;\theta\right) = \int f_{y|x,\alpha}\left(y_i^T|x_i^T,\alpha_i;\theta\right) f_{\alpha|x}\left(\alpha_i|x_i^T\right) d\alpha_i,$$

where $f_{y|x,\alpha}\left(y_i^T|x_i^T,\alpha_i;\theta\right)$ is a conditional model and $f_{\alpha|x}\left(\alpha_i|x_i^T\right)$ is the distribution of time-invariant unobserved heterogeneity.

The integral equation can be modified can allow for x to be strictly exogenous and for y to have dynamics (where y_{i1} plays the role of the initial condition), i.e.,

$$f\left(y_{iT},\ldots,y_{i2}|y_{i1},x_i^T;\theta\right) = \int f_1\left(y_{iT},\ldots,y_{i2}|y_{i1},x_i^T,\alpha_i;\theta\right) f_2\left(\alpha_i|y_{i1},x_i^T\right) d\alpha_i,$$

where the integrand is given by

$$f_1\left(\cdot|y_{i1}, x_i^T, \alpha_i; \theta\right) = g_T\left(y_{iT}|x_i^T, y_i^{T-1}, \alpha_i\right) \times \ldots \times g_2\left(y_{i2}|x_i^T, y_{i1}, \alpha_i\right).$$

The integral equation can also be modified to allow for x to have feedback, i.e.,

$$\begin{aligned} & f\left(y_{iT}, \ldots, y_{i2}, x_{iT}, \ldots, x_{i2}|y_{i1}, x_{i1}; \theta\right) \\ = & \int f_1\left(y_{iT}, \ldots, y_{i2}, x_{iT}, \ldots, x_{i2}|y_{i1}, x_{i1}, \alpha_i; \theta\right) f_2\left(\alpha_i|y_{i1}, x_{i1}\right) d\alpha_i, \end{aligned}$$

where the integrand is given by

$$\begin{aligned} f_1\left(\cdot|y_{i1}, x_{i1}, \alpha_i; \theta\right) = & \; g_T\left(y_{iT}|x_i^T, y_i^{T-1}, \alpha_i\right) h_T\left(x_{iT}|y_i^{T-1}, x_i^{T-1}, \alpha_i\right) \times \ldots \\ & \times g_2\left(y_{i2}|x_i^2, y_{i1}, \alpha_i\right) h_2\left(x_{i2}|y_{i1}, x_{i1}, \alpha_i\right). \end{aligned}$$

It is certainly possible for each of the terms of the above expression to be indexed by some finite-dimensional parameter θ. Furthermore, it is also possible to have a multi-dimensional fixed effect α_i. Note that having the time series dimension provides more degrees of freedom for which to estimate θ but these degrees of freedom may get consumed by considering more and more complex models, even if we retain fully parametric specifications.

A large part of research in panel data econometrics adopts a fully parametric specification for $f_{y|x,\alpha}$ while leaving $f_{\alpha|x}$ unspecified (see the surveys by Chamberlain (1984), Arellano and Honoré (2001), and Arellano and Bonhomme (2011)). Leaving $f_{\alpha|x}$ unspecified is at the core of the fixed-effects approach because one has to account for sources of heterogeneity not always observed by the econometrician. Since there is scarce guidance from economic theory as to the nature of heterogeneity observed units should possess, we start with a widely used notion of heterogeneity – that any differences among observed units are relatively stable over time but are allowed to be correlated with the included regressors. Unfortunately, the presence of individual-specific effects complicates the estimation of common parameters in dynamic nonlinear fixed effects panel data models, as we shall see in the examples in the next section. Alternatively, correlated random effects approaches, where some aspects of the distribution $f_{\alpha|x}$ are specified, can be beneficial as discussed in Example 1.2.6. In practice, either we impose assumptions on the first and second moments of $f_{\alpha|x}$ for linear models or we impose fully parametric assumptions on $f_{\alpha|x}$ for nonlinear models.

The conditional model with $f_{y|x,\alpha}$ fully specified can also be used as a starting point while treating the α_i's as parameters to be estimated. In this case, Neyman and Scott (1948) call θ the structural parameter and α_i the incidental parameter. The distinguishing feature of parametric statistical models with incidental parameters is the presence of a parameter α_i that appears in only a finite number of proba-

bility distributions (in particular, that of ith cross-sectional unit). Neyman and Scott (1948) have shown that the maximum likelihood estimator (MLE) of θ may not be consistent in this case.[1] This unfortunate consequence of using ML have henceforth been referred to as the incidental parameter problem (see Lancaster (2000), Arellano and Honoré (2001), and Arellano and Bonhomme (2011) for surveys of some recent developments).[2]

More formally, this incidental parameter problem arises because the MLE $\widehat{\theta}$ has the following property for fixed T:

$$\begin{aligned}
\widehat{\theta} &= \underset{\theta}{\operatorname{argmax}} \frac{1}{n}\sum_{i=1}^{n} \log f_{y|x,\alpha}\left(y_i^T | x_i^T, \widehat{\alpha_i}(\theta); \theta\right) && (1.1.1)\\
&\xrightarrow{p} \underset{\theta}{\operatorname{argmax}} \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\log f_{y|x,\alpha}\left(y_i^T | x_i^T, \widehat{\alpha_i}(\theta); \theta\right)\right] \\
&\neq \underset{\theta}{\operatorname{argmax}} \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\log f_{y|x,\alpha}\left(y_i^T | x_i^T, \alpha_i\right)\right] && (1.1.2)
\end{aligned}$$

Note that in (1.1.1), we have substituted an estimator of α_i. Hence, the right hand side of (1.1.1) is called the profile or concentrated likelihood. Plugging in an estimator for a finite-dimensional nuisance parameter usually has an asymptotically negligible effect on the estimator for the parameter of interest. In contrast, when we substitute an estimator $\widehat{\alpha_i}(\theta)$ for α_i in (1.1.1), there is an asymptotically nonnegligible effect. The inconsistency of $\widehat{\theta}$ can be traced to four interrelated reasons: (a) the parameter space grows with n, (b) the finite sample bias of $\widehat{\theta}$ that does not disappear in the limit as seen in (1.1.2), (c) the profile or concentrated likelihood does not correspond to a joint density of the observables, and (d) the profile score, which is the derivative of the profile log-likelihood with respect to θ, is not necessarily an unbiased estimating equation. Since these reasons are interrelated, general purpose solutions (some of which are surveyed from an econometrics perspective by Arellano and Hahn (2007) along with its references and from the statistics perspective by Reid

[1] They also show using the example of estimating a normal mean with variances as incidental parameters that sometimes the MLE can be consistent but is no longer asymptotically efficient. They also propose a bias-adjustment method in the spirit of a profile score adjustment. Finally, they sketch the efficiency losses resulting from the incidental parameter problem.

[2] It would seem that treating α_i's as random variables (or random effects) and treating α_i's as parameters are not different from each other. The former subsumes the usual random effects specification where $f_{\alpha|x} = f_\alpha$. Leaving $f_{\alpha|x}$ unspecified is sometimes called the fixed-effects approach. These two models generate estimators that actually have different distribution theories. Sims (2000) argues that "there is a random effects distribution theory for the fixed effects estimator and vice versa." The measurement error literature has been much more explicit about this distinction with respect to its treatment of the latent variable representing the true value of the measurement. The two models are called structural and functional, respectively. See Moran (1971) for more details. Semiparametric estimation and efficiency theory has also been explicit with respect to the distinction. See Moran (1971), Bickel and Klaassen (1986), Bhanja and Ghosh (1992a; 1992b; 1992c), Bickel, Klaassen, et al. (1993), and Pfanzagl (1993) for more details.

(2013), which contain some of the different likelihoods available in the literature) will tend to focus on directly addressing one of these four reasons.

Because the incidental parameter problem is difficult to handle for many nonlinear panel models, some approaches that weaken the fixed-effects approach have been proposed. Typically, the search for consistent estimators of common parameters depends on a set of auxiliary assumptions. Assumptions include, but are not limited to, correlated random effects strategies where the α_i's are drawn from a known $f_{\alpha|x}$ (a particular approach involving sparsity is explored in Chapter 5), fixed-T or large-T bias corrections that exploit full specification of $f_{y|x,\alpha}$ (some of which are explored further in Chapter 4), and approaches invoking discrete support for $f_{\alpha|x}$ (explored further in a simultaneous equations context in Chapter 3). The next four chapters of this dissertation provide specific theoretical or empirical situations for which these auxiliary assumptions may be appropriate (or inappropriate as will be seen in Chapter 2). Before discussing the rest of the thesis, I first discuss the incidental parameter problem in more detail using seven examples.

1.2 Sketching some of the arguments

In this section, I consider some examples that demonstrate the theoretical and practical relevance of the incidental parameter problem along with some proposed solutions. Example 1.2.1 is the many normal means problem posed in Neyman and Scott (1948) where the parameter of interest is the common variance of the observations. The MLE in this example is inconsistent and model-specific solutions are proposed to remedy this inconsistency.

Example 1.2.2 reconsiders the solutions in Example 1.2.1 when both $n, T \to \infty$. Next, Example 1.2.3 is an illustration of the more general case where the $O\left(T^{-1}\right)$ incidental parameter bias is characterized so that we can pursue a general purpose solution. The model-specific nature of fixed-T solutions is further explored in Examples 1.2.4 and 1.2.5. Sometimes these structural parameters are not of main interest and we want to determine how to recover average marginal effects. Example 1.2.6 contains a discussion of how this can be accomplished in fixed-T and large-T situations. Finally, I consider situations where $f_{\alpha|x}$ has discrete support in Example 1.2.7.

Example 1.2.1. (Neyman and Scott (1948), Waterman (1993), and Hahn and Newey (2004)) Let y_{it} be iid draws from a $N\left(\alpha_{i0}, \sigma_0^2\right)$ distribution for $i = 1, \ldots, n$ and $t = 1, \ldots, T$. The parameter of interest in this classic example is the variance parameter σ_0^2. The model allows for one individual-specific effect and does not contain any time-varying regressors. The log-likelihood for one observation is given by

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

The MLE satisfies the following first order conditions obtained by taking the derivative of the log-likelihood with respect to σ^2 and α_i:

$$\begin{aligned} \sum_i \sum_t \left[-\frac{1}{2\sigma^2} + \frac{(y_{it} - \alpha_i)^2}{2\sigma^4} \right] &= 0, \\ \sum_t \left(\frac{y_{it} - \alpha_i}{\sigma^2} \right) &= 0. \end{aligned} \tag{1.2.1}$$

Profiling out the α_i's using the second equation above gives

$$\widehat{\alpha_i}\left(\sigma^2\right) = \frac{1}{T}\sum_t y_{it} = \bar{y}_i. \tag{1.2.2}$$

Note that (1.2.2) is written as a function of σ^2 even though σ^2 does not explicitly appear in the expression for this simple setup. In general, however, the profiled α_i is going to depend on the structural parameter. Substituting this into (1.2.1) and solving for σ^2 gives

$$\widehat{\sigma^2} = \frac{1}{nT}\sum_i \sum_t (y_{it} - \bar{y}_i)^2 . \tag{1.2.3}$$

Note that (1.2.2) does not depend on σ_0^2 and both (1.2.2) and (1.2.3) are available in closed form. The normality and independence assumptions imply that

$$\widehat{\alpha_i}\left(\sigma^2\right) = \bar{y}_i \sim N\left(\alpha_{i0}, \sigma_0^2/T\right).$$

Results from normal theory (applied to time series observations for the ith cross sectional unit) allow us to conclude that $\sum_t (y_{it} - \bar{y}_i)^2 \sim \sigma_0^2 \chi^2_{T-1}$ for every i. Since we have independence across i, we can write

$$\sum_i \sum_t (y_{it} - \bar{y}_i)^2 \sim \sigma_0^2 \chi^2_{n(T-1)}.$$

Furthermore, taking the expectation of $\widehat{\sigma^2}$ gives

$$\mathbb{E}\widehat{\sigma^2} = \frac{1}{nT}\mathbb{E}\left[\sigma_0^2 \chi^2_{n(T-1)}\right] = \sigma_0^2\left(1 - \frac{1}{T}\right). \tag{1.2.4}$$

As a consequence, $\widehat{\sigma^2}$ is not an unbiased estimator of σ_0^2 in finite samples.

If we want to determine if this finite sample bias disappears in large samples, we have to think of the dimensions in which sample sizes could grow, i.e., the consistency of $\widehat{\sigma^2}$ will depend on the asymptotic embedding. When $T \to \infty$ and n is fixed, $\widehat{\sigma^2}$ is consistent for σ_0^2. When $n \to \infty$ and T is fixed, however, $\widehat{\sigma^2}$ is inconsistent for σ_0^2 because of (1.2.4). As a result, the finite sample bias does not disappear even if $n \to \infty$. We can correct the finite-sample bias directly by using the bias-corrected

estimator $\widehat{\sigma_c^2} = \frac{T}{T-1}\widehat{\sigma^2}$. The degrees of freedom correction produces an unbiased and consistent estimator in this case. ■

The previous example is practically relevant because it is a restricted version of a static linear panel data model with strictly exogenous covariates. In particular, setting $\beta_0 = 0$ in the model where $y_{it}|x_i^T \overset{iid}{\sim} N\left(\alpha_{i0} + \beta_0 x_{it}, \sigma_0^2\right)$ produces the previous example.

Note that the bias in (1.2.4) arises from the finite T setting. One can argue that we can view this bias as finite sample bias in the time series dimension brought about by our inability to consistently estimate α_i . Letting $T \to \infty$ while fixing n is a solution for panel data typically encountered in financial (and sometimes macroeconomic) situations. In contrast, many existing datasets derived from surveys have a large-n dimension with a relatively small T. Therefore, a slight change in the asymptotic scheme may be fruitful.

Example 1.2.2. (Continuation of Example 1.2.1) Let us return to the earlier example. When both $n, T \to \infty$ at some unspecified rate, $\widehat{\sigma^2}$ will be consistent for σ_0^2. Unfortunately, the limiting distribution of $\widehat{\sigma^2}$ may be incorrectly centered. Consider the limiting distribution of $\sqrt{nT}\left(\widehat{\sigma^2} - \sigma_0^2\right)$. We have

$$\begin{aligned}
\sqrt{nT}\left(\widehat{\sigma^2} - \sigma_0^2\right) &= \sqrt{nT}\left(\frac{1}{nT}\sum_i\sum_t (y_{it} - \bar{y}_i)^2 - \sigma_0^2\right) \\
&= \sqrt{nT}\left(\frac{1}{nT}\sum_i\sum_t (y_{it} - \alpha_{i0} + \alpha_{i0} - \bar{y}_i)^2 - \sigma_0^2\right) \\
&= \underbrace{\sqrt{nT}\left(\frac{1}{nT}\sum_i\sum_t (y_{it} - \alpha_{i0})^2 - \sigma_0^2\right)}_{Z_1} \\
&\quad \underbrace{-\sqrt{nT}\left(\frac{1}{n}\sum_i (\bar{y}_i - \alpha_{i0})^2\right)}_{Z_2}
\end{aligned}$$

where $Z_1 \xrightarrow{d} N\left(0, 2\sigma_0^4\right)$ as $n, T \to \infty$ and

$$Z_2 = \sqrt{\frac{n}{T}}\sigma_0^2\left(\frac{1}{n}\sum_i\left(\frac{\bar{y}_i - \alpha_{i0}}{\sigma_0/\sqrt{T}}\right)^2\right) = \sqrt{\frac{n}{T}}\sigma_0^2\left(\frac{1}{n}\sum_i \chi_1^2\right) \xrightarrow{p} \kappa\sigma_0^2$$

as $n, T \to \infty$ while $n/T \to \kappa^2$ for some finite constant $\kappa > 0$.[3] As a consequence, we have

$$\sqrt{nT}\left(\widehat{\sigma^2} - \sigma_0^2\right) \xrightarrow{d} N\left(-\kappa\sigma_0^2, 2\sigma_0^4\right)$$

[3]The result depends on sequential asymptotics. Here, we have $T \to \infty$ first then $n \to \infty$.

This example shows that the relative growth rates of the two dimensions influence the magnitude of the nonzero center $-\kappa\sigma_0^2$. This nonzero center disappears when $n/T \to 0$. Otherwise, we can remove the nonzero center as follows:

$$\sqrt{nT}\left(\widehat{\sigma^2} - \sigma_0^2\right) + Z_2 = \sqrt{nT}\left(\widehat{\sigma^2} - \sigma_0^2 + \frac{\sigma_0^2}{T}\right) \xrightarrow{d} N\left(0, 2\sigma_0^4\right).$$

By plugging in a consistent estimator for σ_0^2/T under this asymptotic scheme, we are able to bias-correct $\widehat{\sigma^2}$. The bias-corrected estimator $\widehat{\sigma_c^2} = \widehat{\sigma^2} + \widehat{\sigma^2}/T$ will have a limiting distribution that is centered at zero. Interestingly, the asymptotic variance of $\widehat{\sigma_c^2}$ coincides with the asymptotic variance of $\widehat{\sigma^2}$. Finally, note that $\mathbb{E}\widehat{\sigma_c^2} = \sigma_0^2\left(1 - 1/T^2\right)$. As a result, this corrected estimator is different from the degrees of freedom correction considered in Example 1.2.1 because this corrected estimator is biased for fixed T but it no longer has the $O\left(T^{-1}\right)$ bias. ■

The discussion in Examples 1.2.1 and 1.2.2 provides ways in which we can achieve either consistency for fixed T or a correctly centered asymptotic distribution when both $n, T \to \infty$ at rate $n/T \to \kappa^2$. First, we have a closed form solution (1.2.3) for the MLE of the structural parameter and a complete specification of the density of the data. Thus, we can derive the finite-sample distribution of (1.2.3). Second, the bias of the MLE in (1.2.3) also has a closed form and does not depend on α_i (see (1.2.4)). In general, these conditions rarely arise so a general characterization of the nonzero center is needed, as seen in the next example.

Example 1.2.3. (Hahn and Newey (2004), Arellano and Hahn (2007), and Hahn and Kuersteiner (2011)) In the previous example, we have seen an indication that the bias in the estimator for the parameter of interest in a model with incidental parameters is of order $O\left(T^{-1}\right)$. We can think of this bias as time series finite sample bias and consider again the asymptotic setting where both $n, T \to \infty$ and $n/T \to \kappa^2$. This asymptotic setting will allow us to more generally approximate the asymptotic bias in the estimator and then reduce its impact. Assume that $\widehat{\theta}$ is a consistent estimator under this asymptotic setting, i.e. $\lim_{T\to\infty} \theta_T = \theta_0$, where θ_T is the large-n, fixed-T limit of some extremum estimator. Further assume that $\sqrt{nT}\left(\widehat{\theta} - \theta_T\right) \xrightarrow{d} N(0, \Omega)$. Under these assumptions along with a stochastic expansion of θ_T, i.e., $\theta_T = \theta_0 + B/T + O\left(T^{-2}\right)$, we can write

$$\begin{aligned}
\sqrt{nT}\left(\widehat{\theta} - \theta_0\right) &= \sqrt{nT}\left(\widehat{\theta} - \theta_T\right) + \sqrt{nT}\left(\theta_T - \theta_0\right) \\
&= \sqrt{nT}\left(\widehat{\theta} - \theta_T\right) + \sqrt{nT}\frac{B}{T} + \sqrt{nT}O\left(T^{-2}\right) \\
&= \sqrt{nT}\left(\widehat{\theta} - \theta_T\right) + \sqrt{\frac{n}{T}}B + O\left(\sqrt{\frac{n}{T^3}}\right) \\
&\xrightarrow{d} N(B\kappa, \Omega). \qquad (1.2.5)
\end{aligned}$$

Note that (1.2.5) is not centered at 0. In the previous example, we were able to derive that $B = -\sigma_0^2$. To remove the nonzero center in (1.2.5), we need to characterize B and its components of this term because a characterization is essential for the practical purpose of bias reduction and for the theoretical purpose of understanding the sources of incidental parameter bias.

Hahn and Newey (2004) study the case of static panel data models with strictly exogenous regressors. In this example, I highlight the general setting considered by Hahn and Kuersteiner (2011). They show that in panel data models with fully-specified dynamics, the bias term is given by

$$B = -\mathscr{I}^{-1}\left(\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{f_i^{VU^\alpha}}{\mathbb{E}\left(V_{it}^{\alpha_i}\right)} - \frac{1}{2}\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{E}\left(U_{it}^{\alpha_i\alpha_i}\right)f_i^{VV}}{\left(\mathbb{E}\left(V_{it}^{\alpha_i}\right)\right)^2}\right),$$

where the components of B involve (a) the information matrix $\mathscr{I}^{-1}$, (b) the cross-covariances of the α_i-score V_{it} and the α_i-derivative of the θ-score U_{it}:[4]

$$f_i^{VU^\alpha} = \sum_{l=-\infty}^{\infty}\mathrm{Cov}\left(V_{it}, U_{i,t-l}^{\alpha_i}\right),$$

(c) the autocovariances of the α_i-score

$$f_i^{VV} = \sum_{l=-\infty}^{\infty}\mathrm{Cov}\left(V_{it}, V_{i,t-l}\right),$$

and (d) the expectation of the second α_i-derivative matrix of the θ-score, denoted by $\mathbb{E}\left(U_{it}^{\alpha_i\alpha_i}\right)$.[5] The other remaining component of B is the α_i-derivative of the α_i-score V_{it}, denoted by $V_{it}^{\alpha_i}$.

The characterization of the nonzero center allows us to develop a bias correction under large-n, large-T asymptotics. Observe that a feasible version of the correction requires us to specify a trimming parameter (called bandwidth) for the infinite sums that form B.

Unfortunately, there are negative results with respect to the point identification of common parameters in fixed-T settings (see Chamberlain (2010)). Honoré and Tamer (2006) show that the common parameters of panel data dynamic discrete choice models are only partially identified. Furthermore, bias correction may fail to provide improvements in fixed-T settings. Given that the MLE is heavily biased without bias correction (as documented by numerous Monte Carlo experiments in the literature), it seems advisable to apply these corrections. In general, it is likely that bias-corrected estimators of the common parameters will be found inside the

[4] In the linear model with strictly exogenous regressors, this cross-covariance is zero. Once dynamics are allowed, this cross-covariance is not necessarily zero.

[5] In the linear model, this expectation is zero regardless of whether the regressors are strictly exogenous or not.

identified set. Although no proof of the previous claim exists, we obtain point identification anyway once T becomes very large. ■

Observe that the examples so far apply to panel data models with strictly exogenous regressors and variables with fully-specified feedback mechanisms. On the other hand, GMM based estimation of linear dynamic panel data methods in the spirit of Arellano and Bond (1991) can in principle allow for regressors whose dynamics are not fully modeled. Unfortunately, these GMM estimators also have an asymptotic distribution with a nonzero center under large-n, large-T asymptotics (see Alvarez and Arellano (2003)). Furthermore, these GMM estimators have been documented to have poor finite sample performance and are susceptible to weak instruments (see Bun and Sarafidis (2015) and its references).

It should not be surprising that there is no uniformly good solution to the incidental parameter problem that would apply to every theoretical or empirical situation. As a result, it helps to look for solutions on a case-by-case basis. One possible approach is to exploit the properties of the chosen parametric family to develop a bias-correction. For instance, in Example 1.2.1, consider transforming the data y_{it} into $y_{it} - \bar{y}_i$. The transformation allows us to eliminate the α_i's because the distribution of the transformed data only depends on σ_0^2. As a result, the likelihood function formed from the transformed data can be used to conduct estimation and inference for σ_0^2. The resulting likelihood is called a marginal likelihood in the statistical literature.[6] Yet, it may be very difficult to find transformations or even subsets of the data that will allow us to construct a marginal likelihood. Despite this, there are successful applications of this idea even outside the likelihood setting as the following example illustrates.

Example 1.2.4. (Honoré, 1992) Consider a linear panel data regression model where $y_{it}^* = \alpha_i + \beta x_{it} + \varepsilon_{it}$ for $i = 1, \ldots, n$ and $t = 1, 2$. For simplicity, assume that x_{it} is scalar. Assume that $\left\{\left(y_{i1}^*, x_{i1}, y_{i2}^*, x_{i2}\right) : i = 1, \ldots, n\right\}$ form a random sample but we only get to observe data on both y and x when $y_{i1}^* > 0$ and $y_{i2}^* > 0$. Further assume that ε_{i1} and ε_{i2} are independent, identically and continuously distributed conditional on $(x_{i1}, x_{i2}, \alpha_i)$ for all i.

Honoré (1992) develops a semiparametric approach in the spirit of a marginal likelihood calculation. The idea is to look for a subset of

$$\left\{\left(y_{i1}^*, y_{i2}^*\right) : y_{i1}^* \in \mathbb{R}, y_{i2}^* \in \mathbb{R}\right\}$$

[6] Some authors call the likelihoods obtained after integrating out the nuisance parameters as marginal likelihoods. See Chamberlain (1980) for an example. To avoid confusion, I will call them integrated likelihoods instead. In contrast, we obtain profile likelihoods by maximizing out the nuisance parameters. These two likelihoods represent different ways of eliminating nuisance parameters (see Basu (1977) and Berger, Liseo, and Wolpert (1999) for more details). The meaning of marginal likelihood I use fits with the notion of marginal inference. See Kalbfleisch and Sprott (1970) and Christensen and Kiefer (2000) for more details. A more recent discussion on the types of likelihood functions can be found in Reid (2013).

that is unaffected by truncation. Observe that such a subset allows us to eliminate α_i by differencing. In other words, we have $y_{i1}^* = y_{i1}$, $y_{i2}^* = y_{i2}$ and both time series observations obey $y_{it} = \alpha_i + \beta x_{it} + \varepsilon_{it}$. Notice that this differencing strategy is exactly the same strategy applied to a linear panel data model (as in Example 1.2.1).

Define $\Delta y_i = y_{i1} - y_{i2}$, $\Delta x_i = x_{i1} - x_{i2}$, and $\Delta \varepsilon_i = \varepsilon_{i1} - \varepsilon_{i2}$. Assume that $\beta \Delta x_i > 0$. Consider the following sets

$$\begin{aligned} A &= \left\{\left(y_{i1}^*, y_{i2}^*\right) : y_{i1}^* > \beta \Delta x_i, y_{i2}^* > y_{i1}^* - \beta \Delta x_i\right\}, \\ B &= \left\{\left(y_{i1}^*, y_{i2}^*\right) : y_{i1}^* > \beta \Delta x_i, 0 < y_{i2}^* < y_{i1}^* - \beta \Delta x_i\right\}. \end{aligned}$$

Notice that whenever $y_{i1}^* > \beta \Delta x_i$, we must have $y_{i2}^* > 0$. Observe that

$$\begin{aligned} & \Pr\left(\left(y_{i1}^*, y_{i2}^*\right) \in A | x_{i1}, x_{i2}, \alpha_i\right) \\ = \; & \Pr\left(y_{i2}^* - y_{i1}^* > -\beta \Delta x_i, y_{i2}^* + y_{i1}^* > \beta \Delta x_i | x_{i1}, x_{i2}, \alpha_i\right) \\ = \; & \Pr\left(\varepsilon_{i2} - \varepsilon_{i1} > 0, \varepsilon_{i2} + \varepsilon_{i1} > -2\alpha_i - 2\beta x_{i2} | x_{i1}, x_{i2}, \alpha_i\right) \\ = \; & \Pr\left(\Delta \varepsilon_i < 0 | x_{i1}, x_{i2}, \alpha_i, \underbrace{\varepsilon_{i2} + \varepsilon_{i1} > -2\alpha_i - 2\beta x_{i2}}_{D_i}\right) \\ & \times \Pr\left(D_i | x_{i1}, x_{i2}, \alpha_i\right). \end{aligned}$$

Similarly, we can write

$$\begin{aligned} & \Pr\left(\left(y_{i1}^*, y_{i2}^*\right) \in B | x_{i1}, x_{i2}, \alpha_i\right) \\ = \; & \Pr\left(y_{i2}^* - y_{i1}^* < -\beta \Delta x_i, y_{i2}^* + y_{i1}^* > \beta \Delta x_i | x_{i1}, x_{i2}, \alpha_i\right) \\ = \; & \Pr\left(\varepsilon_{i2} - \varepsilon_{i1} < 0, \varepsilon_{i2} + \varepsilon_{i1} > -2\alpha_i - 2\beta x_{i2} | x_{i1}, x_{i2}, \alpha_i\right) \\ = \; & \Pr\left(\Delta \varepsilon_i > 0 | x_{i1}, x_{i2}, \alpha_i, D_i\right) \Pr\left(D_i | x_{i1}, x_{i2}, \alpha_i\right) \end{aligned}$$

Under the assumption that the distribution of $\Delta \varepsilon_i$ conditional on $\varepsilon_{i1} + \varepsilon_{i2}$ and on $(x_{i1}, x_{i2}, \alpha_i)$ is symmetric and unimodal around zero,[7] we can then conclude that

$$\Pr\left(\left(y_{i1}^*, y_{i2}^*\right) \in A | x_{i1}, x_{i2}, \alpha_i\right) = \Pr\left(\left(y_{i1}^*, y_{i2}^*\right) \in B | x_{i1}, x_{i2}, \alpha_i\right).$$

Furthermore, these two sets are unaffected by truncation and will be observable (since these sets satisfy $y_{i1}^* > \beta \Delta x_i > 0$ and $y_{i2}^* > 0$). As a result,

$$\Pr\left(\left(y_{i1}, y_{i2}\right) \in A | x_{i1}, x_{i2}\right) = \Pr\left(\left(y_{i1}, y_{i2}\right) \in B | x_{i1}, x_{i2}\right).$$

Therefore, the union of these two sets

$$A \cup B = \left\{\left(y_{i1}, y_{i2}\right) : y_{i1} > \beta \Delta x_i, y_{i2} > 0\right\}$$

[7] See Honoré (1992) for a sufficient condition.

is the basis for constructing a moment condition that only involves the observables but not the fixed effect α_i. Observe further that

$$\begin{aligned}
& \mathbb{E}[\mathbf{1}\{(y_{i1}, y_{i2}) \in A\}\, \Delta\varepsilon_i | x_{i1}, x_{i2}, \alpha_i] \\
= & \int_{\mathbf{1}\{(y_{i1}, y_{i2}) \in A\}} u f_{\Delta\varepsilon|x_1, x_2, \alpha, D}(u)\, du \\
= & \frac{\displaystyle\int_{-\infty}^{0} u f_{\Delta\varepsilon|x_1, x_2, \alpha, D}(u)\, du}{F_{\Delta\varepsilon|x_1, x_2, \alpha, D}(0) \Pr(D_i | x_{i1}, x_{i2}, \alpha_i)} \\
= & \frac{\displaystyle\int_{-\infty}^{0} u f_{\Delta\varepsilon|x_1, x_2, \alpha}(-u)\, du}{\big(1 - F_{\Delta\varepsilon|x_1, x_2, \alpha, D}(0)\big) \Pr(D_i | x_{i1}, x_{i2}, \alpha_i)} \\
= & \frac{\displaystyle -\int_{0}^{\infty} v f_{\Delta\varepsilon|x_1, x_2, \alpha}(v)\, dv}{\big(1 - F_{\Delta\varepsilon|x_1, x_2, \alpha, D}(0)\big) \Pr(D_i | x_{i1}, x_{i2}, \alpha_i)} \\
= & -\mathbb{E}[\mathbf{1}\{(y_{i1}, y_{i2}) \in B\}\, \Delta\varepsilon_i | x_{i1}, x_{i2}, \alpha_i]. \qquad (1.2.6)
\end{aligned}$$

The previous derivation involves the expectation of a truncated random variable and the i.i.d. assumption on the errors. We use the symmetry assumption to obtain the third equality. Using (1.2.6), we are able to show that the moment condition

$$\begin{aligned}
& \mathbb{E}[\mathbf{1}\{(y_{i1}, y_{i2}) \in A \cup B\}\, (\Delta y_i - \beta \Delta x_i)\, \Delta x_i] \\
= & \mathbb{E}[\mathbf{1}\{(y_{i1}, y_{i2}) \in A\}\, \Delta\varepsilon_i \Delta x_i] + \mathbb{E}[\mathbf{1}\{(y_{i1}, y_{i2}) \in B\}\, \Delta\varepsilon_i \Delta x_i] \\
= & \mathbb{E}[\mathbb{E}[\mathbf{1}\{(y_{i1}, y_{i2}) \in A\}\, \Delta\varepsilon_i \Delta x_i | x_{i1}, x_{i2}, \alpha_i]] \\
& + \mathbb{E}[\mathbb{E}[\mathbf{1}\{(y_{i1}, y_{i2}) \in B\}\, \Delta\varepsilon_i \Delta x_i | x_{i1}, x_{i2}, \alpha_i]] \qquad (1.2.7) \\
= & 0
\end{aligned}$$

is satisfied, where $\mathbf{1}(\cdot)$ is the indicator function. The case where $\beta \Delta x_i \leq 0$ is analogous and will be part of the criterion function for estimating β. Notice that without the indicator function in (1.2.7), we have the moment condition for β in the static linear panel data model with strictly exogenous covariates. A least squares objective function can be formed where the resulting first-order condition is exactly the sample analog of (1.2.7).[8] ■

Searching for a suitable subset of the data is what makes marginal likelihood approaches (or any other approach in the same spirit) highly model-specific. Furthermore, assumptions have to be changed in very specific ways to accommodate slight changes in the model. Extensions of the previous example to allow for lagged dependent variables can be found in Honoré (1993) but require a modification of the

[8] See Honoré (1992) for more details.

argument along with the assumptions in the previous example. Abrevaya (1999) proposes an estimator for fixed effects models with an unknown transformation of the dependent variable that also has the flavor of a marginal likelihood approach. Strictly speaking, the estimators discussed here are semiparametric in nature but the common feature is the search for subsets of the data from which to construct moment conditions or likelihoods that do not depend on α_i, but are informative about the structural parameters. Bonhomme (2012) provides a theory that allows any user of a likelihood-based panel data model with strictly exogenous regressors to construct moment conditions that are free of the fixed effects. One can think of the theory as a general treatment of the marginal likelihood approach. Unfortunately, it is possible that certain panel models will not possess moment conditions that are informative of the structural parameters. Aspects of this theory will be discussed further in Example 1.2.7.

Yet another approach is to find appropriate conditioning sets so that a conditional likelihood that does not depend on α_i can be constructed. As a result, the score of the conditional likelihood is itself a moment condition that is free of α_i. The following example illustrates this approach in a dynamic logit model.

Example 1.2.5. (Chamberlain, 1985; Maddala, 1987; Honoré and Kyriazidou, 2000) Consider a dynamic panel logit model with one strictly exogenous regressor. In particular, we have for $i = 1, \ldots, n$ and $t = 1, \ldots, T$:

$$\Pr\left(y_{it} = 1 | x_{i1}, \ldots, x_{iT}, y_{i0}, \ldots, y_{i,T-1}, \alpha_i\right) = \frac{\exp\left(\beta x_{it} + \gamma y_{i,t-1} + \alpha_i\right)}{1 + \exp\left(\beta x_{it} + \gamma y_{i,t-1} + \alpha_i\right)}. \quad (1.2.8)$$

This means that

$$\Pr\left(y_{it} = 0 | x_{i1}, \ldots, x_{iT}, y_{i0}, \ldots, y_{i,T-1}, \alpha_i\right) = \frac{1}{1 + \exp\left(\beta x_{it} + \gamma y_{i,t-1} + \alpha_i\right)}.$$

Assume that y_{i0} is observed and $T = 3$. Hence, we have a total of four observations. Define the sets

$$\begin{aligned} A &= \{y_{i1} = 0, y_{i2} = 1, y_{i3} = d_3\}, \\ B &= \{y_{i1} = 1, y_{i2} = 0, y_{i3} = d_3\}, \end{aligned}$$

where $d_3 \in \{0, 1\}$. Let $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})$ and $d_0 \in \{0, 1\}$. We can calculate the following conditional probabilities:

$$\begin{aligned} & \Pr(A | \mathbf{x}_i, y_{i0} = d_0, \alpha_i) \\ = \; & \Pr(y_{i3} = d_3 | \mathbf{x}_i, y_{i0} = d_0, y_{i1} = 0, y_{i2} = 1) \\ & \times \Pr(y_{i2} = 1 | \mathbf{x}_i, y_{i0} = d_0, y_{i1} = 0) \times \Pr(y_{i1} = 0 | \mathbf{x}_i, y_{i0} = d_0) \end{aligned}$$

$$
\begin{aligned}
= & \frac{\exp(d_3(\beta x_{i3}+\gamma+\alpha_i))}{1+\exp(\beta x_{i3}+\gamma+\alpha_i)} \times \frac{\exp(\beta x_{i2}+\alpha_i)}{1+\exp(\beta x_{i2}+\alpha_i)} \\
& \times \frac{1}{1+\exp(\beta x_{i1}+\gamma d_0+\alpha_i)}, \\
& \Pr(B|\mathbf{x}_i, y_{i0}=d_0, \alpha_i) \\
= & \Pr(y_{i3}=d_3|\mathbf{x}_i, y_{i0}=d_0, y_{i1}=1, y_{i2}=0) \\
& \times \Pr(y_{i2}=0|\mathbf{x}_i, y_{i0}=d_0, y_{i1}=1) \times \Pr(y_{i1}=1|\mathbf{x}_i, y_{i0}=d_0) \\
= & \frac{\exp(d_3(\beta x_{i3}+\alpha_i))}{1+\exp(\beta x_{i3}+\alpha_i)} \times \frac{1}{1+\exp(\beta x_{i2}+\gamma+\alpha_i)} \\
& \times \frac{\exp(\beta x_{i1}+\gamma d_0+\alpha_i)}{1+\exp(\beta x_{i1}+\gamma d_0+\alpha_i)}.
\end{aligned}
$$

Choosing $A \cup B$ as a conditioning set and noting that A and B are disjoint sets, the definition of conditional probability allows us to write

$$
\begin{aligned}
& \Pr(A|y_{i0}=d_0, A\cup B, \alpha_i) \\
= & \frac{\Pr(A|\mathbf{x}_i, y_{i0}=d_0, \alpha_i)}{\Pr(A\cup B|\mathbf{x}_i, y_{i0}=d_0, \alpha_i)} \\
= & \frac{\Pr(A|\mathbf{x}_i, y_{i0}=d_0, \alpha_i)}{\Pr(A|\mathbf{x}_i, y_{i0}=d_0, \alpha_i)+\Pr(B|\mathbf{x}_i, y_{i0}=d_0, \alpha_i)}, \qquad (1.2.9) \\
& \Pr(B|y_{i0}=d_0, A\cup B, \alpha_i) \\
= & 1-\Pr(A|\mathbf{x}_i, y_{i0}=d_0, A\cup B, \alpha_i). \qquad (1.2.10)
\end{aligned}
$$

Both probabilities in (1.2.9) and (1.2.10) still depend on α_i.

Consider first the case where $\beta = 0$. Observe that

$$
\begin{aligned}
& \Pr(A\cup B|\mathbf{x}_i, y_{i0}=d_0, \alpha_i) \\
= & \frac{\exp(d_3(\gamma+\alpha_i))}{1+\exp(\gamma+\alpha_i)} \times \frac{\exp(\alpha_i)}{1+\exp(\alpha_i)} \times \frac{1}{1+\exp(\gamma d_0+\alpha_i)} \\
& + \frac{\exp(d_3\alpha_i)}{1+\exp(\alpha_i)} \times \frac{1}{1+\exp(\gamma+\alpha_i)} \times \frac{\exp(\gamma d_0+\alpha_i)}{1+\exp(\gamma d_0+\alpha_i)} \\
= & \frac{\exp(d_3\alpha_i)\exp(\alpha_i)[\exp(d_3\gamma)+\exp(\gamma d_0)]}{[1+\exp(\alpha_i)][1+\exp(\gamma+\alpha_i)][1+\exp(\gamma d_0+\alpha_i)]}.
\end{aligned}
$$

Therefore, we can write (1.2.9) as

$$
\begin{aligned}
& \Pr(A|y_{i0}=d_0, A\cup B, \alpha_i) \\
= & \frac{\dfrac{\exp(d_3(\gamma+\alpha_i))}{1+\exp(\gamma+\alpha_i)} \times \dfrac{\exp(\alpha_i)}{1+\exp(\alpha_i)} \times \dfrac{1}{1+\exp(\gamma d_0+\alpha_i)}}{\dfrac{\exp(d_3\alpha_i)\exp(\alpha_i)[\exp(d_3\gamma)+\exp(d_0\gamma)]}{[1+\exp(\alpha_i)][1+\exp(\gamma+\alpha_i)][1+\exp(\gamma d_0+\alpha_i)]}}
\end{aligned}
$$

$$= \frac{\exp(d_3\gamma)}{\exp(d_3\gamma)+\exp(d_0\gamma)}.$$

Similarly, (1.2.10) can be written as

$$\Pr(B|y_{i0}=d_0, A\cup B, \alpha_i) = \frac{\exp(d_0\gamma)}{\exp(d_3\gamma)+\exp(d_0\gamma)}.$$

Both these conditional probabilities do not depend on α_i and can be used to form a conditional likelihood depending only on γ.

Now, consider the case where $\beta \neq 0$. Honoré and Kyriazidou (2000) show that by further conditioning on the event $\{x_{i2}=x_{i3}\}$, assumed to have positive probability, we can eliminate the dependence of (1.2.9) and (1.2.10) on α_i. In particular, we have

$$\begin{aligned}
\Pr(A|\mathbf{x}_i, y_{i0}=d_0, A\cup B, x_{i2}=x_{i3}, \alpha_i) &= \frac{1}{1+\exp(\beta(x_{i1}-x_{i2})+\gamma(d_0-d_3))}, \\
\Pr(B|\mathbf{x}_i, y_{i0}=d_0, A\cup B, x_{i2}=x_{i3}, \alpha_i) &= \frac{\exp(\beta(x_{i1}-x_{i2})+\gamma(d_0-d_3))}{1+\exp(\beta(x_{i1}-x_{i2})+\gamma(d_0-d_3))}
\end{aligned}$$

and a conditional likelihood will be formed from observations where $x_{i2}=x_{i3}$ and $y_{i1}+y_{i2}=1$, i.e., a conditional MLE can be computed from the following optimization problem:

$$\max_{\beta,\gamma} \sum_{i=1}^{N} \mathbf{1}\{y_{i1}+y_{i2}=1\}\,\mathbf{1}\{x_{i2}=x_{i3}\} \log\left(\frac{[\exp(\beta(x_{i1}-x_{i2})+\gamma(d_0-d_3))]^{y_{i1}}}{1+\exp(\beta(x_{i1}-x_{i2})+\gamma(d_0-d_3))}\right).$$

The condition $x_{i2}=x_{i3}$ is unlikely to be satisfied, so a kernel function replaces the indicator function above. Because we introduce a kernel function, the estimators for the structural parameters converge at a rate slower than the usual parametric rate. We also cannot allow for time dummies because they never satisfy $x_{i2}=x_{i3}$ by definition. Extensions to a semiparametric specification of the probability function (1.2.8) in the spirit of Manski (1987b), the multinomial logit case, and more than four observations for every i are available in Honoré and Kyriazidou (2000). ■

Even though the approaches in Examples 1.2.4 and 1.2.5 are both appealing and insightful, the search for appropriate transformations of the data or appropriate conditioning sets will become cumbersome when T is a bit larger or when we make slight changes to the model.

Some authors like Wooldridge (2005b) and Arellano and Bonhomme (2011) argue that the structural parameters may not be of primary interest especially for policy. Policy parameters are usually of the form $\mathbb{E}\left[m\left(x_i^T, \alpha_i\right)\right]$, where m is some function of the regressors and unobserved heterogeneity. These policy parameters have been called many names depending on the form of m, such as the average structural func-

tion (Blundell and Powell, 2004), quantile structural function (Chernozhukov et al., 2013), average index function (Lewbel, Dong, and Yang, 2012), average marginal effect (Wooldridge, 2005b), and local average response (Altonji and Matzkin, 2005). These policy parameters represent summary measures that describe outcomes of certain thought experiments. One such thought experiment involves a prediction of what m will be when we set x_i^T at some fixed value $\bar{x}$ while holding unobserved heterogeneity constant. Another thought experiment would involve predictions as to how m changes when we change the value $\bar{x}$ while holding unobserved heterogeneity constant. Unfortunately, these policy parameters are hard to identify and require understanding the tradeoffs among competing assumptions as seen in the next example.

Example 1.2.6. (Hoderlein and White (2012)) Consider the following nonseparable model where $Y_{it} = g(X_{it}, \alpha_i, \varepsilon_{it})$ for $i = 1, \ldots, n$ and $t = 1, 2$. An object of interest for policy is how $E(Y_{it}|X_{i1} = x_1, X_{i2} = x_2)$ changes with x_1 or x_2, holding the source of unobserved heterogeneity constant. In other words, the policy parameters of interest or average marginal effects at x_1 and x_2, are given by

$$\begin{aligned} ME_1(x_1, x_2) &= \int\int \frac{\partial g(x_1, a, e)}{\partial x_1} f_{\alpha_i, \varepsilon_{i1}|X_i}(a, e|x_1, x_2)\, da\, de, \\ ME_2(x_1, x_2) &= \int\int \frac{\partial g(x_2, a, e)}{\partial x_2} f_{\alpha_i, \varepsilon_{i2}|X_i}(a, e|x_1, x_2)\, da\, de. \end{aligned}$$

Had we known what $f_{\alpha_i, \varepsilon_{it}|X_i}$ is, then everything becomes straightforward and calculating $ME_1(x)$ and $ME_2(x)$ can be done directly. This situation is really the idea behind the calculation of average marginal effects from fully parametric models with correlated random effects proposed by Chamberlain (1984) and Wooldridge (2005b). If we do not know $f_{\alpha_i, \varepsilon_{it}|X_i}$, we have to indirectly recover $ME_1(x)$ and $ME_2(x)$ somehow. In particular, we have the following

$$\begin{aligned} \mathbb{E}(Y_{i1}|X_{i1} = x_1, X_{i2} = x_2) &= \int\int g(x_1, a, e) f_{\alpha_i, \varepsilon_{i1}|X_i}(a, e|x)\, da\, de, \\ \mathbb{E}(Y_{i2}|X_{i1} = x_1, X_{i2} = x_2) &= \int\int g(x_2, a, e) f_{\alpha_i, \varepsilon_{i2}|X_i}(a, e|x)\, da\, de, \end{aligned}$$

with four derivatives given by

$$\begin{aligned} \frac{\partial \mathbb{E}(Y_{i1}|X_{i1} = x_1, X_{i2} = x_2)}{\partial x_1} &= ME_1(x_1, x_2) \\ &\quad + \int\int g(x_1, a, e) \frac{\partial f_{\alpha_i, \varepsilon_{i1}|X_i}(a, e|x)}{\partial x_1}\, da\, de \\ \frac{\partial \mathbb{E}(Y_{i1}|X_{i1} = x_1, X_{i2} = x_2)}{\partial x_2} &= \int\int g(x_1, \alpha, \varepsilon) \frac{\partial f_{\alpha_i, \varepsilon_{i1}|X_i}(a, e|x)}{\partial x_2}\, da\, de \end{aligned}$$

$$\begin{aligned}
\frac{\partial \mathbb{E}(Y_{i2}|X_{i1}=x_1,X_{i2}=x_2)}{\partial x_1} &= \int\int g(x_2,a,e)\frac{\partial f_{\alpha_i,\varepsilon_{i2}|X_i}(a,e|x)}{\partial x_1}\,da\,de \\
\frac{\partial \mathbb{E}(Y_{i2}|X_{i1}=x_1,X_{i2}=x_2)}{\partial x_2} &= ME_2(x_1,x_2) \\
&\quad + \int\int g(x_2,a,e)\frac{\partial f_{\alpha_i,\varepsilon_{i2}|X_i}(a,e|x)}{\partial x_2}\,da\,de
\end{aligned}$$

The left hand side of the above derivatives are observable from the data. In contrast, the right hand side involves objects that are unknown to the econometrician, specifically the distribution of the errors $f_{\alpha_i,\varepsilon_{it}|X_i}$ and their associated derivatives $\partial f_{\alpha_i,\varepsilon_{it}|X_i}/\partial x$. To recover $ME_1(x_1,x_2)$ and $ME_2(x_1,x_2)$ from the four preceding equations, we have to make further assumptions since there are more unknowns than the number of equations. It is not enough that we assume a form of time homogeneity (which ensures that a repeated measurement will be beneficial with respect to controlling for α_i), i.e.

$$f_{\alpha_i,\varepsilon_{i1}|X_i} = f_{\alpha_i,\varepsilon_{i2}|X_i}$$

because we are still unable to completely remove the distortion caused by the effect of changing x_1 or x_2 on the distribution of the errors. In addition, we have to condition on the set where $X_{i1}=X_{i2}=x$ to completely remove this distortion. As a result, we are able to identify the marginal effects by conditioning on an appropriate set under no assumptions about the nonseparable model and the distribution of the errors (aside from time homogeneity):[9]

$$\begin{aligned}
ME_1(x) &= \frac{\partial \mathbb{E}(Y_{i1}|X_{i1}=X_{i2}=x)}{\partial x_1} - \frac{\partial \mathbb{E}(Y_{i2}|X_{i1}=X_{i2}=x)}{\partial x_1}, \\
ME_2(x) &= \frac{\partial \mathbb{E}(Y_{i2}|X_{i1}=X_{i2}=x)}{\partial x_2} - \frac{\partial \mathbb{E}(Y_{i1}|X_{i1}=X_{i2}=x)}{\partial x_2}.
\end{aligned}$$

Without conditioning on $X_{i1}=X_{i2}=x$, there are multiple avenues to recover the average marginal effect. In general, we can only partially identify the average marginal effect when there are bounds on $g(X_{it},\alpha_i,\varepsilon_{it})$ (see Chernozhukov et al. (2013)).

To avoid conditioning on $X_{i1}=X_{i2}=x$, we may consider correlated random effects strategies that use exchangeability (see Altonji and Matzkin (2005) for more) and dimension reduction to construct "instruments" that allow us to nullify the distortions brought about by the effect of changing x on the distribution of the errors. Bester and Hansen (2009b) show that if there exists a sufficient statistic that could reduce the dimension of the conditioning set $\{X_{i1}=x_1,X_{i2}=x_2,\ldots,X_{iT}=x_T\}$, then it is possible to recover the average marginal effect if $T\geq 3$. Bester and Hansen (2009b) are actually able to weaken the assumption of time homogeneity in this

[9] Imposing further restrictions may help in trading off some assumptions for others. The gains will have to be explored on a case-by-case basis.

case. Testing some of these assumptions is the subject of Ghanem (2015).

Note that the discussion so far focuses on strictly exogenous regressors. Extending the ideas to dynamic models are not very straightforward under the conditions maintained in the earlier discussion. Bounds for the dynamic model are also available in Chernozhukov et al. (2013). Parametric approaches that fully specify the distribution of the errors are available in Wooldridge (2005b). Large-T bias corrections of marginal effects obtained from parametric fixed-effects models can be found in Hahn and Newey (2004), Bester and Hansen (2009a), and Fernandez-Val (2009). ■

In the final example, I show that reducing the support of the distribution of the fixed effects may be helpful in identification of structural parameters. The lack of point identification of structural parameters in nonlinear panel data models has been documented by Honoré and Tamer (2006) and Chamberlain (2010) if we leave the distribution of the fixed effects unspecified. This lack of point identification can also be illustrated in the next example.

Example 1.2.7. (Bajari et al. (2011) and Bonhomme (2012)) Consider the following panel binary choice model with strictly exogenous regressors x_{it}:

$$\Pr(y_{it} = 1|x_i, \alpha_i) = H(\alpha_i + \beta x_{it}),\ i = 1, \ldots, n;\ t = 1, \ldots, T \tag{1.2.11}$$

where the distribution of the individual-specific fixed effect α_i given $x_i = (x_{i1}, \ldots, x_{iT})$ has finite and discrete support, i.e.

$$\Pr(\alpha_i = \alpha_k|x_i = x) = \pi_{x,k},\ k = 1, \ldots, K.$$

A fixed-effects setup means that we leave the $\pi_{x,k}$'s unspecified and possibly dependent on x. Assume further that the inverse link function H is specified in advance. Since the α_i's are unobservable, we have to look at the full conditional distribution of $y_i = (y_{i1}, \ldots, y_{iT})$ given $x_i = x$ alone. As a consequence of the law of total probability, this full conditional distribution can be written as

$$\Pr(y_i = y|x_i = x) = \sum_{k=1}^{K} \Pr(y_i = y|x_i = x, \alpha_i = \alpha_k; \beta) \Pr(\alpha_i = \alpha_k|x_i = x) \tag{1.2.12}$$

for some binary sequence y . The left hand side of (1.2.12) can be recovered from the data on frequencies of each of the 2^T possible binary sequences. We can collect every (1.2.12) for each possible binary sequence so that we have a matrix equation

$$P_{y|x} = P_x(\beta)\, \pi_x,$$

where $\pi_x = \left(\pi_{x,1}, \ldots, \pi_{x,K}\right)^{\mathrm{T}}$ is a $K \times 1$ vector, $P_x(\beta)$ is a $2^T \times K$ matrix based on the

specification (1.2.11), and $P_{y|x}$ is a $2^T \times 1$ vector of conditional probabilities observed from the data.

Instead of differencing out every α_i which is not generalizable outside linear models, we difference out π_x by annihilating the matrix $P_x(\beta)$, i.e.

$$\left[I - P_x(\beta) P_x(\beta)^-\right] P_{y|x} = \left[I - P_x(\beta) P_x(\beta)^-\right] P_x(\beta) \pi_x = 0. \tag{1.2.13}$$

Note that $P_x(\beta)^-$ is the Moore-Penrose inverse of $P_x(\beta)$. The main message behind (1.2.13) is not that it is possible to construct moment conditions that do not depend on α_i but that the rank of the matrix $P_x(\beta)$ matters. If we know that $K \geq 2^T$, then (1.2.13) is not informative of β at all. On the other hand, considering models for the fixed effects for which $K < 2^T$ may be useful. We can interpret $K < 2^T$ as the support of the fixed effects being less rich than the support of outcomes. In general, we will not know whether $K < 2^T$ or otherwise. There are empirical situations, such as the game-theoretic model estimated by Hahn and Moon (2010) and the one discussed in Chapter 3, where we would know the value of K.

In Hahn and Moon (2010), the reduced support of the fixed effects arises because the fixed effects represent which of the two pure strategy equilibria is selected by players and maintained over time. Further work that allow for time-varying fixed effects with limited support has been studied by Bonhomme and Manresa (2015). The latter paper and Hahn and Moon (2010) have shown that bias correction in a large-n, large-T context like we have seen in Example 1.2.3 is not needed at all. ■

1.3 How should we respond?

The discussion in the previous section comes from a perspective which emphasizes either the elimination of nuisance parameters or the robustification of estimation and inference methods in the presence of nuisance parameters. Furthermore, the distribution of the fixed effects is left unspecified as seen in Examples 1.2.1 to 1.2.5. As models become more complicated, this emphasis may become increasingly untenable, especially when there is meaning to be attached to nuisance parameters or when interest centers on functions of interest and nuisance parameters as seen in Examples 1.2.6 and 1.2.7. Empirically relevant models also have to allow for dynamics and predetermined regressors. Therefore, we need to search for methods that work in slightly complicated settings at the cost of making assumptions that may nevertheless be motivated theoretically or empirically. I now describe the ideas pursued in the succeeding chapters.

Many empirical situations (see Chapter 2 for examples) call for the estimation of a dynamic binary choice model with fixed effects. In Chapter 2, I demonstrate that it is inappropriate to estimate such a model by applying IV to a dynamic linear probability model. Motivations behind the use of a dynamic linear probability model

include the ability to directly recover average marginal effects and the availability of software without additional programming. IV or GMM based estimators of the dynamic linear probability model can also allow for predetermined regressors. We saw the difficulties in recovering average marginal effects and allowing for dynamics in Example 1.2.6. The main results of the chapter actually suggest that IV estimators of the linear probability model converge to an average marginal effect with incorrect weighting. Furthermore, this large-n limit might not even be found inside the large-n limit of the bounds proposed by Chernozhukov et al. (2013). In addition, these IV estimators do not converge to the true average marginal effect even as $T \to \infty$. As a result, this chapter gives an example for which dealing with the incidental parameter problem using IV may not be a good response.

Another empirical situation of interest involves the estimation of simultaneously determined discrete outcomes. Allowing for fixed effects in these models has not been explored fully, since most research has focused on either cross-sectional models, continuous outcomes, or random effects (see for example, the research by Cornwell, Schmidt, and Wyhowski (1992), Leon-Gonzalez (2003), Matzkin (2008), Matzkin (2012), and Masten (2015)). Parameter identification in these models is further complicated by the nonexistence of a unique reduced form. One way of partially resolving the identification problem is to introduce coherency conditions. Unfortunately, the coherency condition needs to be imposed a priori. In Chapter 3, I propose using panel data to estimate such models by allowing the data to determine how the coherency condition will hold. The manner in which the coherency condition holds can be represented as an incidental parameter that has finite support, in the spirit of what we have seen in Example 1.2.7.

The discussion in Example 1.2.3 is an estimator-based bias correction. One will observe that papers proposing an analytical correction of the estimator typically motivate the correction using the score. In Chapter 4, I develop a score-based correction involving projections. This approach is a useful and intuitive alternative when constructing estimating equations for the structural parameters that are relatively insensitive to inconsistent plug-ins. I show that the method can produce familiar estimators in special cases. I also show that projection exploits correct specification to reap the gains from bias reduction especially when T is very small.

Although the notion of time-invariant heterogeneity is hardly unique, a large gap exists between specifications where we allow for full heterogeneity (i.e., acknowledging that all units are different from each other) and full homogeneity (i.e., acknowledging that all units are the same). This gap enables us to explore different notions of partial pooling. Researchers acknowledge that units might be different from one another yet they may believe that some units are more alike than others. Despite this, they might be unwilling to specify which units are different from each other and which units are similar to one another. I formalize the preceding intuition by allowing some incidental parameters to take on the same value, namely zero.

In Chapter 5, I demonstrate that some notion of sparsity of the incidental parameters may be useful in constructing fixed-T consistent estimators that converge at the root-n rate. In particular, I tune the lasso (see Tibshirani (1996; 2011) and Chapter 5 for more) so that it will be able to detect the non-zero incidental parameters. A subsample for which the incidental parameters are set to zero can then be used for estimation and inference. This is in contrast to the machine learning and big data literature where the main developments have concentrated on uncovering non-zero effects in a sea of zero effects.

The four essays included in this thesis demonstrate several different ways to cope with the incidental parameter problem. None of these essays offer a general solution. Instead, these essays provide situations for which the incidental parameter problem may not be a serious impediment to theoretical and empirical work. However, I restrict myself to parametric situations and leave the nonparametric situations to future research.

Chapter 2

On IV estimation of a dynamic linear probability model with fixed effects

2.1 Introduction

Many researchers still use the dynamic linear probability model (LPM) with fixed effects when analyzing a panel of binary choices. Several applications of the dynamic LPM with fixed effects can be found in papers published in top journals. Applications include assessing the magnitude of state dependence in female labor force participation (Hyslop, 1999), examining the factors that affect exporting decisions (Bernard and Jensen, 2004), determining the effect of income on transitions in and out of democracy (Acemoglu et al., 2009), and determining how overnight rates affect a bank's decision to provide loans (Jiménez et al., 2014). A more suitable approach, however, is to use limited dependent variable (LDV) models when analyzing discrete choice. Unfortunately, the inclusion of fixed effects creates an incidental parameter problem that complicates the estimation of average marginal effects, especially when the time dimension is small (see the survey by Arellano and Bonhomme (2011)). Resorting to a random effects or correlated random effects approach may require specifying the full distribution of the fixed effects and initial conditions[1] – something that researchers may be unwilling to do because of the lack of specific subject

[1] Typically, only the first two moments of the full distribution are required in the case of linear models. In contrast, nonlinear models would typically require the full distribution because we use this distribution to integrate the fixed effects out of the distribution. There are some approaches that can be thought of as being in the middle of correlated random effects approaches and fixed-effects approaches. A prominent example is using a special regressor to consistently estimate common parameters without imposing a parametric assumption on the distribution of the fixed effects and initial conditions, as proposed by Honoré and Lewbel (2002).

matter knowledge to construct such a distribution. Linear dynamic panel data methods present an alternative that allows for fixed effects, dynamics, predetermined regressors, fewer functional form restrictions, and even allow for heteroscedasticity. Therefore, using methods intended for linear dynamic panel data models seems to be an attractive alternative in this setting.

In contrast, my results provide arguments against a commonly held sentiment among researchers expressed quite forcefully in Angrist and Pischke (2009, p.107):

> The upshot of this discussion is that while a nonlinear model may fit the CEF for LDVs more closely than a linear model, when it comes to marginal effects, this probably matters a little. This optimistic conclusion is not a theorem, but, as in the empirical example here, it seems to be fairly robustly true.
>
> Why, then, should we bother with nonlinear models and marginal effects? One answer is that the marginal effects are easy enough to compute now that they are automated in packages like Stata. But there are a number of decisions to make along the way (e.g., the weighting scheme, derivatives versus finite differences), while OLS is standardized. Nonlinear life also gets considerably more complicated when we work with instrumental variables and panel data. Finally, extra complexity comes into the inference step as well, since we need standard errors for marginal effects.

In this paper, I explain why usual dynamic panel data methods, specifically instrumental variable (IV) estimation, are inappropriate for estimating average marginal effects if the conditional expectation function (CEF) is truly nonlinear. In particular, I show the large-n limit of the Anderson-Hsiao (1981; 1982) IV estimator (henceforth AH) is an average marginal effect but subject to incorrect weighting. Given that the AH estimator is a special case of GMM, estimators in the spirit of Arellano and Bond (1991) may be subject to the same problem. I also show that the effect of this incorrect weighting does not disappear even when T is large. Furthermore, I give examples to show that there are certain parameter configurations and fixed effect distributions for which the large-n limit of the AH estimator is outside the nonparametric bounds derived by Chernozhukov et al. (2013).

Much research has been done on whether using the LPM is suitable. A particularly eye-catching example was provided by Lewbel, Dong, and Yang (2012). They show, in a toy example, that OLS applied to the LPM cannot even get the correct sign of the treatment effect even in the situation where there is just a binary exogenous regressor and a high signal-to-noise ratio. Horrace and Oaxaca (2006) show that the linear predictor for the probability of success should be in $[0,1]$ for all observations for the OLS estimator to be consistent for the regression coefficients because the zero conditional mean assumption does not hold when there are observations (whether in the sample or in the population) that produce success probabilities outside $[0,1]$. On the other hand, Wooldridge (2010) argues that "the case for the LPM is even stronger if most the regressors are discrete and take on only a few values". Problem 15.1 of his book asks the reader to show that we need not worry about success probabilities being outside $[0,1]$ in a saturated model. If we specialize the results in Wooldridge

(2005a) and Murtazashvili and Wooldridge (2008) to the LPM, then they show that fixed-effects estimation applied to the LPM with strictly exogenous regressors can be used to consistently estimate average marginal effects under a specific correlated random coefficients condition.

I organize the rest of the chapter as follows. In Section 2.2, I present an example to show that it is possible to use the LPM to recover an average treatment effect under very special assumptions that researchers are unwilling to make. In Section 2.3, I derive analytically the consequences of not meeting these special assumptions when interest centers on the average marginal effect of state dependence for the cases of $T = 3$ and $T \to \infty$. Next, I examine the practical implications of these results using a numerical example and an empirical application on female labor force participation and fertility in Section 2.4. The last section contains concluding remarks followed by a technical appendix.

2.2 A situation where the LPM is a good idea

Suppose we have a two-period panel binary choice model with a strictly exogenous binary regressor:

$$\Pr(y_{it} = 1|x_i, \alpha_i) = \Pr(y_{it} = 1|x_{it}, \alpha_i) = H(\alpha_i + \beta x_{it}), \tag{2.2.1}$$

where $H : [0,1] \to \mathbb{R}$ is some inverse link function that is increasing, $y_{it} \in \{0,1\}$ and $x_i = (x_{i1}, x_{i2}) = (0,1)$ for all $i = 1, \ldots, n$ and $t = 1, 2$. Assume that for all i, we have $y_{i1} \perp y_{i2} | x_i, \alpha_i$.

The regressor x is a strictly exogenous treatment indicator such that all individuals are treated in the second period but not in the first period. In other words, specification (2.2.1) is basically a before-and-after analysis. In this setting, α_i is an individual-specific fixed effect drawn from some unspecified density $g(\alpha)$.

Suppose one ignores the binary nature of the outcome variable y_{it} and one starts with an LPM with fixed effects, i.e., $y_{it} = \alpha_i + \beta x_{it} + \varepsilon_{it}$ instead. The within estimator for β, which is equivalent to the first-difference estimator for $T = 2$, is then given by

$$\widehat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_{i2} - y_{i1})(x_{i2} - x_{i1})}{\frac{1}{n}\sum_{i=1}^{n}(x_{i2} - x_{i1})^2} = \frac{1}{n}\sum_{i=1}^{n}(y_{i2} - y_{i1})\,\mathbf{1}(y_{i1} + y_{i2} = 1) = \frac{1}{n}(n_{01} - n_{10}),$$

where $\mathbf{1}(\cdot)$ is the indicator function. The second equality follows from the definition of x and the implication that $y_{i2} - y_{i1} = 0$ for all i such that $y_{i1} = y_{i2}$. The third equality follows from defining $n_{ab} = \sum_{i=1}^{n} \mathbf{1}(y_{i1} = a, y_{i2} = b)$ as the number

of observations for which we observe the sequence ab. Thus, only those i for which $y_{i1} \neq y_{i2}$ enter into the calculation of $\widehat{\beta}$.

When we calculate the large-n limit of the within estimator, we have

$$\begin{aligned}
\widehat{\beta} \xrightarrow{p} & \int \Pr(y_{i2}=1|x_i,\alpha)\Pr(y_{i1}=0|x_i,\alpha)\,g(\alpha)\,d\alpha \\
& -\int \Pr(y_{i2}=0|x_i,\alpha)\Pr(y_{i1}=1|x_i,\alpha)\,g(\alpha)\,d\alpha \\
= & \int [(1-H(\alpha))H(\alpha+\beta)-H(\alpha)(1-H(\alpha+\beta))]\,g(\alpha)\,d\alpha \\
= & \int [H(\alpha+\beta)-H(\alpha)]\,g(\alpha)\,d\alpha
\end{aligned}$$

In the situation I have described, the average marginal effect $\Delta = \mathbb{E}[y_{i2}-y_{i1}|x_i=(0,1)]$ can be written as:

$$\begin{aligned}
\Delta &= \mathbb{E}[\mathbb{E}(y_{i2}|x_i=(0,1),\alpha)-\mathbb{E}(y_{i1}|x_i=(0,1),\alpha)] \\
&= \mathbb{E}[\mathbb{E}(y_{i2}|x_{i2}=1,\alpha)-\mathbb{E}(y_{i1}|x_{i1}=0,\alpha)] && (2.2.2) \\
&= \mathbb{E}[H(\alpha+\beta)-H(\alpha)] && (2.2.3)
\end{aligned}$$

Despite the inability of the within estimator to consistently estimate β,[2] the within estimator does coincide with Δ even if the true model is nonlinear. In addition, the sample analog of Δ is exactly the within estimator.

Notice that the result arises because of a lucky coincidence of factors – (a) the strict exogeneity of x (allowing us to obtain (2.2.2)), (b) the independence of α_i and x_i (allowing us to obtain (2.2.3)), and (c) the time homogeneity assumption because H does not depend on time (which follows from (2.2.1)). Despite starting from a fixed effects treatment of α_i, one has no choice but to assume independence of α_i and x_i in order to obtain (2.2.3). This already violates the need to allow for arbitrary correlation between α_i and x_i. It is as if an omniscient Nature did not use the knowledge of α_i to assign a corresponding treatment vector x_i to every unit.

Hahn's (2001) discussion of Angrist (2001) has already pointed out the special conditions under which the within estimator is able to estimate an average treatment effect. In addition, he emphasizes that the simple strategies suggested by Angrist (2001) require knowledge of the "structure of treatment assignment and careful re-expression of the new target parameter". Chernozhukov et al. (2013) also make the same point and further show that the within estimator converges to some weighted average of individual difference of means for a specific subset of the data. They also show that this weighted average is not the average marginal effect of interest.

[2] Incidentally, Chamberlain (2010) shows that β is not even point identified in this example unless H is logistic. The result of Manski (1987a) does not apply here. He shows that β is identified up to scale when one of the strictly exogenous regressors has unbounded support.

Despite all these concerns, researchers still insist on estimating LPMs with fixed effects. One may argue that the example above does not really arise in empirical applications but the example already gives an indication that complicated binary choice models estimated through an LPM are unlikely to produce intended results. In particular, the lucky coincidence of factors mentioned earlier does not hold at all for the dynamic LPM which I discuss next.

2.3 Main results

2.3.1 The case of three time periods

Consider the following specification of a dynamic discrete choice model with fixed effects and no additional regressors:

$$\begin{aligned} \Pr\left(y_{it}=1|y_i^{t-1},\alpha_i\right) &= \Pr\left(y_{it}=1|y_{i,t-1},\alpha_i,y_{i0}\right) \\ &= H\left(\alpha_i+\rho y_{i,t-1}\right),\ i=1,\ldots,n,\ t=1,2,3, \end{aligned} \tag{2.3.1}$$

where y_i^{t-1} is the past history of y, α_i is an individual-specific fixed effect, y_{i0} is an observable initial condition, and $H:\mathbb{R}\to[0,1]$ is some inverse link function. Assume that $(y_{i0},y_{i1},y_{i2},y_{i3},\alpha_i)$ are independently drawn from their joint distribution for all i. I leave the joint density of (α_i,y_{i0}), denoted by f, unspecified. This data generating process satisfies Assumptions 1, 3, 5, and 6 of Chernozhukov et al. (2013).

If H is the logistic function, then ρ can be estimated consistently using conditional logit (Chamberlain, 1985). If H happens to be the standard normal cdf, then ρ is not even point-identified (Honoré and Tamer, 2006).[3] In both these cases, we also cannot point-identify the average marginal effect Δ:

$$\Delta=\int\left[\Pr\left(y_{it}=1|y_{i,t-1}=1,\alpha,y_0\right)-\Pr\left(y_{it}=1|y_{i,t-1}=0,\alpha,y_0\right)\right]f\left(\alpha,y_0\right)\,d\alpha dy_0 \tag{2.3.2}$$

even if we know H but leave the density of (y_{i0},α_i) unspecified. This average marginal effect is of practical interest because it measures the effect of state dependence in the presence of individual-specific unobserved heterogeneity.

Despite these negative results, researchers still insist on using a dynamic LPM on the grounds that linearity still provides a good approximation even if the true H is nonlinear.[4] I use this as a starting point and determine the large-n limit of IV estimators for the dynamic LPM. The linear model researchers have in mind can be expressed as:

$$y_{it}=\alpha_i+\rho y_{i,t-1}+\epsilon_{it},\ i=1,\ldots,n,\ t=1,2,3,$$

[3] Honoré and Tamer (2006) actually show that the sign of ρ is identified for any strictly increasing cdf H and unrestricted distribution of (y_{i0},α_i).

[4] The dynamic LPM is really a special case of (2.3.1), where H is the identity function.

where $\epsilon_{it} = y_{it} - E\left(y_{it}|y_i^{t-1}, \alpha_i\right)$. We now take first-differences to eliminate α_i:

$$\Delta y_{it} = \rho \Delta y_{i,t-1} + \Delta \epsilon_{it},\ i = 1, \ldots, n,\ t = 2, 3.$$

Because the differenced regressor $\Delta y_{i,t-1}$ is correlated with the differenced error $\Delta \epsilon_{it}$, IV or GMM estimators have been used to estimate ρ. Using lagged differences as instruments, the AH estimator can be written as

$$\widehat{\rho}_{AHd} = \frac{\sum_{i=1}^{n} \Delta y_{i1} \Delta y_{i3}}{\sum_{i=1}^{n} \Delta y_{i1} \Delta y_{i2}}.$$

Because of the binary nature of the sequences $\{(y_{i0}, y_{i1}, y_{i2}, y_{i3}) : i = 1, \ldots, n\}$, it is certainly possible for some of the first differences to be equal to zero. Therefore, there are only certain types of sequences that enter into the expression above. If we enumerate all these 16 possible sequences, we can rewrite the estimator as

$$\widehat{\rho}_{AHd} = \frac{n_{0110} + n_{1001} - n_{1010} - n_{0101}}{n_{0100} + n_{1010} + n_{0101} + n_{1011}},$$

where $n_{abcd} = \sum_{i=1}^{n} \mathbf{1}(y_{i0} = a, y_{i1} = b, y_{i2} = c, y_{i3} = d)$ denotes the number of observations in the data for which we observe the sequence $abcd$.[5]

It can be shown[6] that the large-n limit of $\widehat{\rho}_{AHd}$ is

$$\begin{aligned} \widehat{\rho}_{AHd} \quad &\xrightarrow{p} \quad \frac{\int H(\alpha)(1 - H(\alpha + \rho))(H(\alpha + \rho) - H(\alpha))\, g(\alpha)\, d\alpha}{\int H(\alpha)(1 - H(\alpha + \rho))\, g(\alpha)\, d\alpha} \\ &= \quad \int w_d(\alpha, \rho)(H(\alpha + \rho) - H(\alpha))\, d\alpha \\ &= \quad \int \int w_d(\alpha, \rho)\left[\Pr\left(y_{it} = 1 | y_{i,t-1} = 1, \alpha, y_0\right) - \Pr\left(y_{it} = 1 | y_{i,t-1} = 0, \alpha, y_0\right)\right] d\alpha\, dy_0 \end{aligned} \tag{2.3.3}$$

where

$$w_d(\alpha, \rho) = \frac{H(\alpha)(1 - H(\alpha + \rho))\, g(\alpha)}{\int H(\alpha)(1 - H(\alpha + \rho))\, g(\alpha)\, d\alpha}.$$

Note that the weighting function $w_d(\alpha, \rho)$ depends on the true value of ρ and the

[5] Note that we cannot just drop those sequences for which $y_{i1} + y_{i2} \neq 1$, like in conditional logit. If we do this, the resulting AH estimator becomes

$$\widetilde{\rho}_{AHd} = \frac{-n_{1010} - n_{0101}}{n_{0100} + n_{1010} + n_{0101} + n_{1011}},$$

which is always negative regardless of the sign of ρ or Δ. Observe that identification arguments based on the conditional logit do not necessarily translate to other inverse link functions, including that of the identity function.

[6] A part of the derivation can be found in the appendix.

marginal distribution of the fixed effects $g(\alpha)$. The correct weighting function should have been the joint density of (y_0, α) as in (2.3.2). Therefore, $\widehat{\rho}_{AHd}$ is inconsistent for Δ because of the incorrect weighting of the individual marginal dynamic effect $H(\alpha+\rho)-H(\alpha)$.

It is difficult to give a general indication of whether we overestimate or underestimate Δ, because the results depend on the joint distribution of (y_0, α). If it happens that $\rho = 0$ (so that $\Delta = 0$), then $\widehat{\rho}_{AHd}$ is consistent for Δ.

The analysis above can be extended to the AH estimator which uses levels as the instrument set. It can be shown that this AH estimator has the following form:

$$\begin{aligned}\widehat{\rho}_{AHl} &= \frac{\sum_{i=1}^{n}\sum_{t=2}^{3} y_{i,t-2}\Delta y_{it}}{\sum_{i=1}^{n}\sum_{t=2}^{3} y_{i,t-2}\Delta y_{i,t-1}} \\ &= \frac{n_{0110}-n_{0101}+n_{1110}-n_{1010}+n_{1100}-n_{1011}}{n_{1010}+n_{1000}+n_{1001}+n_{1011}+n_{0100}+n_{1100}+n_{0101}+n_{1101}}.\end{aligned}$$

Calculations similar to (2.3.3) allow us to derive the large-n limit of $\widehat{\rho}_{AHl}$, i.e.

$$\begin{aligned}\widehat{\rho}_{AHl} \xrightarrow{p} & \frac{\int (1-H(\alpha+\rho))(1+H(\alpha+\rho))(H(\alpha+\rho)-H(\alpha))f(\alpha,1)\,d\alpha}{\int [(1-H(\alpha+\rho))(1+H(\alpha+\rho))f(\alpha,1)+(1-H(\alpha+\rho))H(\alpha)f(\alpha,0)]\,d\alpha} \\ & + \frac{\int (1-H(\alpha+\rho))H(\alpha)(H(\alpha+\rho)-H(\alpha))f(\alpha,0)\,d\alpha}{\int [(1-H(\alpha+\rho))(1+H(\alpha+\rho))f(\alpha,1)+(1-H(\alpha+\rho))H(\alpha)f(\alpha,0)]\,d\alpha} \\ = & \int\int w_l(\alpha,\rho,y_0)(H(\alpha+\rho)-H(\alpha))\,dy_0\,d\alpha \\ = & \int\int w_l(\alpha,\rho,y_0)\big[\Pr\big(y_{it}=1|y_{i,t-1}=1,\alpha,y_0\big)-\Pr\big(y_{it}=1|y_{i,t-1}=0,\alpha,y_0\big)\big]\,d\alpha\,dy_0,\end{aligned}$$

where

$$\begin{aligned}w_l(\alpha,\rho,0) &= \frac{(1-H(\alpha+\rho))H(\alpha)f(\alpha,0)}{\int [(1-H(\alpha+\rho))(1+H(\alpha+\rho))f(\alpha,1)+(1-H(\alpha+\rho))H(\alpha)f(\alpha,0)]\,d\alpha}, \\ w_l(\alpha,\rho,1) &= \frac{(1-H(\alpha+\rho))(1+H(\alpha+\rho))f(\alpha,1)}{\int [(1-H(\alpha+\rho))(1+H(\alpha+\rho))f(\alpha,1)+(1-H(\alpha+\rho))H(\alpha)f(\alpha,0)]\,d\alpha},\end{aligned}$$

I denote $f(\alpha,0) = \Pr(y_0=0|\alpha)\,g(\alpha)$ and $f(\alpha,1) = \Pr(y_0=1|\alpha)\,g(\alpha)$. Note that the weighting function $w_l(\alpha,\rho,y_0)$ depends on the true value of ρ and the joint distribution of (y_0,α). Once again, we have an incorrect weighting function $w_l(\alpha,\rho,y_0)$ instead of the joint distribution of (y_0,α). As a result, $\widehat{\rho}_{AHl}$ is inconsistent for Δ.[7]

[7] For the case where we have one less time period, i.e. we observe sequences of the form $\{(y_{i0}, y_{i1}, y_{i2}) : i = 1,\ldots,n\}$, the large-$n$ limit of $\widehat{\rho}_{AHl}$ depends only on $f(\alpha,1)$.

I was able to obtain neat analytical expressions because there are no other regressors aside from the lagged dependent variable. However, the results above can be extended to the case where we have a predetermined binary regressor (at the cost of more complicated notation). Furthermore, the results can also be extended to regressors with richer support. But these exercises will also point to the same inconsistency of IV estimators for the average marginal effect.

2.3.2 Large-T case

A natural question to ask is whether the inconsistency results extend to the case where the number of time periods T is large. An intuitive response would be to say that as $T \to \infty$, the fixed effects α_i can be estimated consistently. Therefore, we should be able to estimate average marginal effects consistently. Unfortunately, this intuition may be mistaken.

To address this issue, I use sequential asymptotics where I let $T \to \infty$ and then $n \to \infty$ (see Phillips and Moon (1999)). I first derive the large-T limits of the two AH estimators ($\widehat{\rho}_{AHd}$ and $\widehat{\rho}_{AHl}$) and the first-difference OLS estimator $\widehat{\rho}_{FD}$ for the dynamic LPM:

$$y_{it} = \alpha_i + \rho y_{i,t-1} + \epsilon_{it},\ i = 1,\dots,n,\ t = 1,\dots,T.$$

Recall that these estimators are given by the following expressions:

$$\widehat{\rho}_{AHd} = \frac{\sum_{i=1}^{n}\sum_{t=3}^{T}\Delta y_{i,t-2}\Delta y_{it}}{\sum_{i=1}^{n}\sum_{t=3}^{T}\Delta y_{i,t-2}\Delta y_{i,t-1}}, \widehat{\rho}_{AHl} = \frac{\sum_{i=1}^{n}\sum_{t=2}^{T} y_{i,t-2}\Delta y_{it}}{\sum_{i=1}^{n}\sum_{t=2}^{T} y_{i,t-2}\Delta y_{i,t-1}}, \widehat{\rho}_{FD} = \frac{\sum_{i=1}^{n}\sum_{t=2}^{T}\Delta y_{it}\Delta y_{i,t-1}}{\sum_{i=1}^{n}\sum_{t=2}^{T}\left(\Delta y_{i,t-1}\right)^2}.$$

It can be shown that as $T \to \infty$,[8]

$$\begin{aligned}
\frac{1}{T}\sum_{t=3}^{T}\Delta y_{i,t-2}\Delta y_{it} &\xrightarrow{p} -\int (1-H(\alpha+\rho))H(\alpha)(H(\alpha+\rho)-H(\alpha))\,g(\alpha)\,d\alpha,\\
\frac{1}{T}\sum_{t=3}^{T}\Delta y_{i,t-2}\Delta y_{i,t-1} &\xrightarrow{p} -\int (1-H(\alpha+\rho))H(\alpha)\,g(\alpha)\,d\alpha,\\
\frac{1}{T}\sum_{t=2}^{T} y_{i,t-2}\Delta y_{it} &\xrightarrow{p} -\int \frac{(1-H(\alpha+\rho))H(\alpha)}{1-H(\alpha+\rho)+H(\alpha)}(H(\alpha+\rho)-H(\alpha))\,g(\alpha)\,d\alpha,\\
\frac{1}{T}\sum_{t=2}^{T} y_{i,t-2}\Delta y_{i,t-1} &\xrightarrow{p} -\int \frac{(1-H(\alpha+\rho))H(\alpha)}{1-H(\alpha+\rho)+H(\alpha)}g(\alpha)\,d\alpha,\\
\frac{1}{T}\sum_{t=2}^{T}\Delta y_{it}\Delta y_{i,t-1} &\xrightarrow{p} -\int (1-H(\alpha+\rho))H(\alpha)\,g(\alpha)\,d\alpha,
\end{aligned}$$

[8] Some of the calculations can be found in the Appendix. Note that even with fixed n, the inconsistency is still present.

$$\frac{1}{T}\sum_{t=2}^{T}\left(\Delta y_{i,t-1}\right)^2 \quad \xrightarrow{p} \quad -2\int \frac{(1-H(\alpha+\rho))H(\alpha)}{1-H(\alpha+\rho)+H(\alpha)} g(\alpha)\, d\alpha.$$

Notice that the limiting quantities above do not depend on i. Therefore, as $n \to \infty$, we must have

$$\widehat{\rho}_{AHd} \quad \xrightarrow{p} \quad \int w_d(\alpha,\rho)(H(\alpha+\rho)-H(\alpha))\, d\alpha, \tag{2.3.4}$$

$$\widehat{\rho}_{AHl} \quad \xrightarrow{p} \quad \int w_l(\alpha,\rho)(H(\alpha+\rho)-H(\alpha))\, d\alpha, \tag{2.3.5}$$

$$\widehat{\rho}_{FD} \quad \xrightarrow{p} \quad \frac{1}{2}\left[1-\int w_l(\alpha,\rho)(H(\alpha+\rho)-H(\alpha))\, d\alpha\right], \tag{2.3.6}$$

where the weighting functions are given by

$$\begin{aligned} w_d(\alpha,\rho) &= \frac{H(\alpha)(1-H(\alpha+\rho))\, g(\alpha)}{\int H(\alpha)(1-H(\alpha+\rho))\, g(\alpha)\, d\alpha}, \\ w_l(\alpha,\rho) &= \frac{\frac{(1-H(\alpha+\rho))H(\alpha)}{1-H(\alpha+\rho)+H(\alpha)} g(\alpha)}{\int \frac{(1-H(\alpha+\rho))H(\alpha)}{1-H(\alpha+\rho)+H(\alpha)} g(\alpha)\, d\alpha}. \end{aligned}$$

As for the behavior of the fixed effects (FE) estimator in the large-T case, I rely on Proposition 3.1 of Galvao and Kato (2014). In the context I consider, the linear probability model is misspecified and the true model is the nonlinear model (2.3.1). As a result, the conditional mean $\mathbb{E}\left(y_{it}|y_{i,t-1},\alpha_i\right)$ is misspecified as additive and linear when in fact it is nonlinear. Under their assumptions A1 to A3, they show that the FE estimator converges to the following pseudo-true parameter:

$$\beta_0 = \frac{\mathbb{E}\left(\widetilde{y}_{it}\widetilde{y}_{i,t-1}\right)}{\mathbb{E}\left(\widetilde{y}_{i,t-1}^2\right)},$$

where $\widetilde{y}_{it} = y_{it} - \mathbb{E}\left(y_{it}|y_{i,t-1},\alpha_i\right)$. Assumption A1 of their paper require that the marginal distribution of $\left(\alpha_i, y_{it}, y_{i,t-1}\right)$ is invariant with respect to (i,t). As a result, the initial condition is drawn from the stationary distribution conditional on α_i. Notice that I did not impose this assumption. In the appendix, I show that this pseudo-true parameter is given by

$$\beta_0 = \frac{\mathbb{E}\left[(H(\alpha+\rho)-H(\alpha))\Pr\left(y_{i,t-1}=1|\alpha\right)\left(1-\Pr\left(y_{i,t-1}=1|\alpha\right)\right)\right]}{\mathbb{E}\left[\Pr\left(y_{i,t-1}=1|\alpha\right)\left(1-\Pr\left(y_{i,t-1}=1|\alpha\right)\right)\right]}, \tag{2.3.7}$$

where the expectations are calculated with respect to the marginal distribution of α.

Clearly, the FE estimator does not converge to the correct average marginal effect and the weighting function is given by

$$w_{FE}(\alpha,\rho)=\frac{\Pr\left(y_{i,t-1}=1|\alpha\right)\left(1-\Pr\left(y_{i,t-1}=1|\alpha\right)\right)}{\mathbb{E}\left[\Pr\left(y_{i,t-1}=1|\alpha\right)\left(1-\Pr\left(y_{i,t-1}=1|\alpha\right)\right)\right]}.$$

The result (2.3.6) is very troubling. When $\rho = 0$ (so that the true average marginal effect is 0), $\widehat{\rho}_{FD}$ converges to 0.5, grossly overstating the true Δ. In contrast, the other two AH estimators and the FE estimator are available to consistently estimate Δ when $\rho = 0$ (so that $\Delta = 0$). Unfortunately, for all other values of ρ, these two AH estimators and the FE estimator still cannot consistently estimate the correct Δ because of incorrect weighting in (2.3.4), (2.3.5), and (2.3.7). The appropriate weighting function is now the marginal distribution of the fixed effects $g(\alpha)$, because the effect of the initial condition disappears as $T \to \infty$. Moreover, just as in the fixed-T case considered earlier, it is still not possible to determine the direction of inconsistency. Finally, Chernozhukov et al. (2013) show in their Theorem 4 that the identified set for Δ shrinks to a singleton as $T \to \infty$. Thus, it becomes more likely that the large-T limits in (2.3.4), (2.3.5), and (2.3.6) are outside the identified set.

2.4 Practical implications

Based on the results of the previous section, we should not be using IV estimators for the dynamic LPM. Despite these negative results, the IV estimators are able to estimate a zero average marginal effect, if it were the truth. This observation may allow us to construct a test of the hypothesis that $\Delta = 0$. Unfortunately, this may not be so straightforward since the appropriate standard errors for the AH estimators still depend on the unknown joint distribution of (y_0, α). Although of practical interest, testing the hypothesis $\Delta = 0$ may be infeasible.

To further persuade researchers not to use IV for the dynamic LPM, I adopt the example in Chernozhukov et al. (2013) to show that, even in the simplest of cases, we cannot ignore the distortion brought about by the incorrect weighting function. Chernozhukov et al. (2013) consider a data generating process where H is the standard normal cdf, $y_{i0} \perp \alpha_i$, $\Pr(y_{i0} = 1) = 0.5$, and $T = 3$.

I use four distributions for the fixed effects, as described in Table 2.4.1. The first is the standard normal distribution which is a usual choice in Monte Carlo simulations and serves as a benchmark. The second is a mixture of a standard normal and a normal distribution with mean 2 and variance 0.5^2. This mixture makes it more likely for cross-sectional units to have $y_{it} = 1$ across time. The third is a distribution which favors the LPM because the support of α_i is on a bounded set $(0, 1)$. Finally,

the fourth is a mixture of two normals with negative means. This mixture achieves the opposite effect compared to the second mixture.

Table 2.4.1: Distribution of fixed effects for computations

	$N(0,1)$	$0.5N(0,1)+$ $0.5N(2,0.5)$	$Beta(4,2)$	$0.5N(-2,0.1)+$ $0.5N(-1,1)$
Mean	0	1	0.667	-1.5
Variance	1	1.625	0.032	0.755
Skewness	0	-0.543	-0.468	1.132
Kurtosis	3	2.402	2.625	4.070
Multimodal?	Unimodal	Bimodal	Unimodal	Bimodal

In Figure 2.4.1, I calculate[9] the large-n limits of the AH estimators (in blue for $\widehat{\rho}_{AHd}$ and green for $\widehat{\rho}_{AHl}$) and large-n limits of the nonparametric bounds proposed by Chernozhukov et al. (2013) (in red for the lower bound $\hat{\Delta}_l$ and orange for the upper bound $\hat{\Delta}_u$) evaluated at different values of $\rho \in [-2,2]$. I also calculate the true Δ (in black) using the true distribution of (y_0, α).

Even in the benchmark case where $\alpha_i \sim N(0,1)$, both the large-n limits of the AH estimators are larger than Δ when $\rho > 0$. Further note that when $\rho < -0.5$[10], both these large-n limits are outside the identified set. For $\alpha_i \sim 0.5N(0,1)+0.5N(2,0.5)$, both the large-$n$ limits of the AH estimators nearly coincide and are much larger than Δ even for less persistent state dependence. For $\alpha_i \sim Beta(4,2)$, the large-n limits of the AH estimators are practically the same as Δ and both can be found in the identified set. The key seems to be that the bounded support for the fixed effect, which is $(0,1)$. Finally, the large-n limit of the AH estimator using levels as the instrument set is smaller than Δ for $\alpha_i \sim 0.5N(-2,0.1)+0.5N(-1,1)$.

Although I do not have analytical results for GMM applied to the dynamic LPM, I illustrate why GMM may not be a good idea using the empirical application by Chernozhukov et al. (2013) on female labor force participation and fertility. They estimate the following model using complete longitudinal data on 1587 married women selected from the National Longitudinal Survey of Youth 1979 and observed for three years – 1990, 1992, and 1994:

$$LFP_{it} = \mathbf{1}(\beta \cdot kids_{it} + \alpha_i \geq \epsilon_{it}).$$

The parameter of interest is the average marginal effect of fertility on female labor force participation. The dependent variable is a labor force participation indicator,

[9] A Mathematica notebook containing the calculations is available at `http://andrew-pua.ghost.io`.

[10] Negative state dependence has been found in the literature on scarring effects (see references in Torgovitsky (2015)).

the regressor is a fertility indicator that takes the value 1 if the woman has a child less than 3 years old, and α_i is the individual-specific fixed effect.

Chernozhukov et al. (2013) compute nonparametric bounds for the average marginal effect under the assumption that the fertility indicator is strictly exogenous (called static bounds) and that the average marginal effect is decreasing[11] in the fertility indicator. I replicate their bounds and they can be found in row (2) of Table 2.4.3. I also include static bounds without monotonicity for comparison in row (1).[12] In addition, I compute two other nonparametric bounds with and without the monotonicity assumption under the assumption that the fertility indicator is predetermined (called dynamic bounds) in rows (3) and (4).

Table 2.4.3: Female LFP and fertility ($n = 1587, T = 3$)

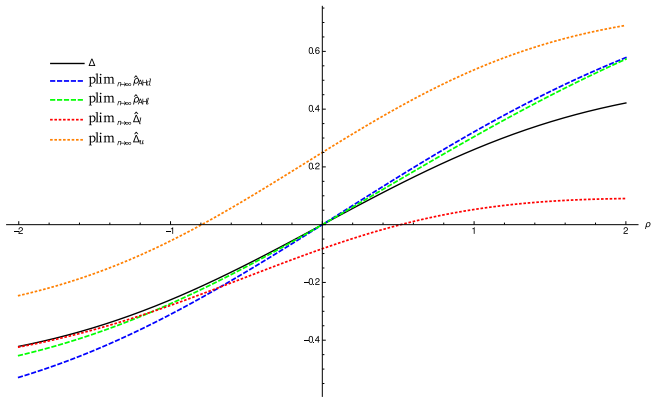
		Avg. Marginal Effect	95% CI
(1)	Static NP bounds	$[-0.40, 0.09]$	
(2)	(1) under monotonicity	$[-0.40, -0.04]$	
(3)	Dynamic NP bounds	$[-0.39, 0.11]$	
(4)	(3) under monotonicity	$[-0.39, -0.19]$	
(5)	Random effects probit	-0.11	$[-0.13, -0.08]$
(6)	Fixed effects OLS	-0.08	$[-0.11, -0.06]$
(7)	First-difference OLS	-0.08	$[-0.09, -0.04]$
(8)	AH (differences)	-0.01	$[-0.14, 0.13]$
(9)	AH (levels)	-0.02	$[-0.07, 0.03]$
(10)	Arellano-Bond	-0.02	$[-0.07, 0.03]$

I also report estimates based on the linear probability model along with the usual 95% asymptotic confidence intervals. Both the fixed effects and first-differenced estimates (rows (6) and (7)) can be found inside the static bounds. This is no longer the case when static bounds are computed under the monotonicity assumption. In contrast, the estimated average marginal effect from the random effects probit (row (5)) is inside the static bounds with or without monotonicity, despite the very incredible assumption where the fixed effects are independent of the fertility indicator. Finally, note that the AH and Arellano-Bond estimates (rows (8) to (10)), which actually assume predeterminedness, are outside the dynamic bounds under monotonicity.

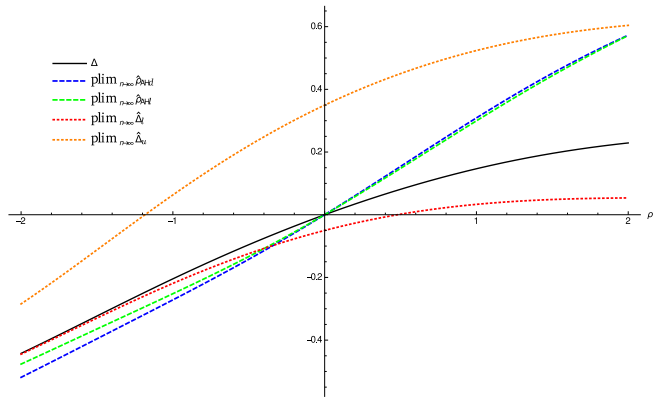
[11] Details as to how to construct the bounds under monotonicity can be found in the Supplemental Material to Chernozhukov et al. (2013).

[12] A Stata do-file is available for replication at `http://andrew-pua.ghost.io`.

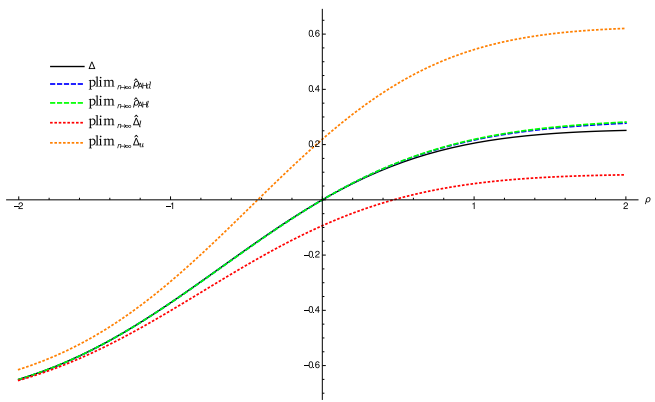
Figure 2.4.1: Large-n limits of the AH estimators under different distributions for the fixed effects



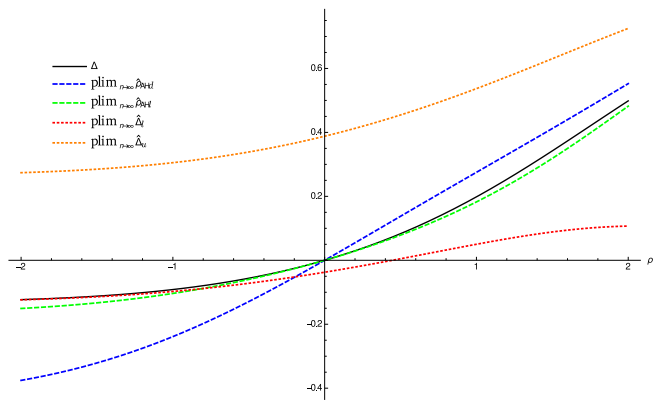
$\alpha_i \sim N(0,1)$



$\alpha_i \sim 0.5N(0,1) + 0.5N(2,0.5)$



$\alpha_i \sim Beta(4,2)$



$\alpha_i \sim 0.5N(-2,0.1) + 0.5N(-1,1)$

2.5 Concluding remarks

I show that using IV methods to estimate the dynamic LPM with fixed effects is inappropriate even in large samples (whether n or T diverge). The analytical results indicate that incorrect weighting of the individual treatment effect is the source of the problem. The numerical results indicate that the estimators may be outside the identified set in both finite and large samples. Therefore, it is more appropriate to use the nonparametric bounds proposed by Chernozhukov et al. (2013), especially if one is unwilling to specify the form for the inverse link function and the joint distribution of the initial conditions and the fixed effects.

The large-n, large-T results I obtain are based on sequential asymptotics. I conjecture that we should obtain similar inconsistency results based on joint asymptotics. It is also unclear whether bias corrections that are derived under large-n, large-T asymptotics can be directly applied to the dynamic LPM with fixed effects. The results in the paper point out that the direction of the asymptotic bias of the estimator for the average marginal effect cannot be obtained. This in stark contrast with the direction of the asymptotic bias derived by Nickell (1981). Although the Monte Carlo experiments of Fernandez-Val (2009) indicate good finite sample performance when we apply the large-T bias corrections, future work should study what exactly these corrections are doing.

It would also be interesting to derive similar analytical results for correlated random effects models so that the results in Wooldridge (2005a) and Murtazashvili and Wooldridge (2008) can be extended to the dynamic case. In the empirical application, I find that the average marginal effect from the usual random effects probit under strict exogeneity can be found in the static nonparametric bounds. Respecting the inherent nonlinearity of a discrete choice model may be responsible for this finding. Future work on this will be of practical interest.

2.6 Appendix

Some calculations for (2.3.3)

We calculate $\mathbb{E}[\mathbf{1}(y_{i0}=0,y_{i1}=1,y_{i2}=1,y_{i3}=0)]$ in detail since the other expressions follow similarly. This expression is equal to

$$\begin{aligned}
& \Pr(y_{i0}=0,y_{i1}=1,y_{i2}=1,y_{i3}=0) \\
= & \int \Pr(y_{i0}=0,y_{i1}=1,y_{i2}=1,y_{i3}=0|\alpha)\, g(\alpha)\, d\alpha \\
= & \int \Pr(y_{i3}=0|y_{i0}=0,y_{i1}=1,y_{i2}=1,\alpha)\Pr(y_{i2}=1|y_{i0}=0,y_{i1}=1,\alpha)\times \\
& \Pr(y_{i1}=1|y_{i0}=0,\alpha)\Pr(y_{i0}=0|\alpha)\, g(\alpha)\, d\alpha
\end{aligned}$$

$$
\begin{aligned}
&= \int \Pr(y_{i3}=0|y_{i2}=1,\alpha)\Pr(y_{i2}=1|y_{i1}=1,\alpha) \\
&\quad \times \Pr(y_{i1}=1|y_{i0}=0,\alpha) f(\alpha,0)\,d\alpha \\
&= \int (1-H(\alpha+\rho))H(\alpha+\rho)H(\alpha)f(\alpha,0)\,d\alpha, \qquad (2.6.1)
\end{aligned}
$$

where f is the joint density of (α, y_0). Similarly, we have the following:

$$
\begin{aligned}
\mathbb{E}[\mathbf{1}(y_{i0}=1,y_{i1}=0,y_{i2}=0,y_{i3}=1)] &= \int H(\alpha)(1-H(\alpha))(1-H(\alpha+\rho))f(\alpha,1)\,d\alpha \\
\mathbb{E}[\mathbf{1}(y_{i0}=1,y_{i1}=0,y_{i2}=1,y_{i3}=0)] &= \int (1-H(\alpha+\rho))H(\alpha)(1-H(\alpha+\rho))f(\alpha,1)\,d\alpha \\
\mathbb{E}[\mathbf{1}(y_{i0}=0,y_{i1}=1,y_{i2}=0,y_{i3}=1)] &= \int H(\alpha)(1-H(\alpha+\rho))H(\alpha)f(\alpha,0)\,d\alpha \\
\mathbb{E}[\mathbf{1}(y_{i0}=0,y_{i1}=1,y_{i2}=0,y_{i3}=0)] &= \int (1-H(\alpha))(1-H(\alpha+\rho))H(\alpha)f(\alpha,0)\,d\alpha \\
\mathbb{E}[\mathbf{1}(y_{i0}=1,y_{i1}=0,y_{i2}=1,y_{i3}=1)] &= \int H(\alpha+\rho)H(\alpha)(1-H(\alpha+\rho))f(\alpha,1)\,d\alpha
\end{aligned}
$$

Assembling these expressions together in the expression for the large-sample limit of $\widehat{\rho}_{AHd}$ gives (2.3.3).

Some calculations for the large-T case

Note that $\Delta y_{i,t-2}\Delta y_{it} = y_{i,t-2}y_{it} - y_{i,t-3}y_{it} - y_{i,t-2}y_{i,t-1} + y_{i,t-3}y_{i,t-1}$. Observe that the binary nature of y allows us to write

$$
\frac{1}{T}\sum_{t=3}^{T} y_{i,t-2}y_{it} \xrightarrow{p} \lim_{T\to\infty} \frac{1}{T}\sum_{t=3}^{T} \Pr\big(y_{it}=1, y_{i,t-2}=1\big).
$$

By the law of total probability, the definition of conditional probability, and calculations similar to (2.6.1), we are able to express $\Pr\big(y_{it}=1, y_{i,t-2}=1\big)$ as

$$
\begin{aligned}
& \Pr\big(y_{it}=1, y_{i,t-2}=1\big) \\
= \; & \Pr\big(y_{it}=1, y_{i,t-1}=0, y_{i,t-2}=1\big) + \Pr\big(y_{it}=1, y_{i,t-1}=1, y_{i,t-2}=1\big) \\
= \; & \int H(\alpha)(1-H(\alpha+\rho))\Pr\big(y_{i,t-2}=1|\alpha\big)g(\alpha)\,d\alpha \\
& + \int H(\alpha+\rho)^2 \Pr\big(y_{i,t-2}=1|\alpha\big)g(\alpha)\,d\alpha.
\end{aligned}
$$

As a result, we have

$$
\frac{1}{T}\sum_{t=3}^{T} y_{i,t-2}y_{it} \xrightarrow{p} \int \big[H(\alpha+\rho)^2 + H(\alpha)(1-H(\alpha+\rho))\big]\left[\lim_{T\to\infty}\frac{1}{T}\sum_{t=3}^{T}\Pr\big(y_{i,t-2}=1|\alpha\big)\right] g(\alpha)\,d\alpha.
$$

Finally observe that $\Pr\left(y_{i,t-2}=1|\alpha\right)$ obeys a first-order nonhomogeneous difference equation. In particular, note that

$$\begin{aligned}
\Pr(y_{i1}=1|\alpha) &= \Pr(y_{i1}=1|y_{i0}=1,\alpha)\Pr(y_{i0}=1|\alpha) \\
&\quad +\Pr(y_{i1}=1|y_{i0}=0,\alpha)\Pr(y_{i0}=0|\alpha) \\
&= [H(\alpha+\rho)-H(\alpha)]\Pr(y_{i0}=1|\alpha)+H(\alpha) \\
\Pr(y_{i2}=1|\alpha) &= \Pr(y_{i2}=1|y_{i1}=1,\alpha)\Pr(y_{i1}=1|\alpha) \\
&\quad +\Pr(y_{i2}=1|y_{i1}=0,\alpha)\Pr(y_{i1}=0|\alpha) \\
&= [H(\alpha+\rho)-H(\alpha)]\Pr(y_{i1}=1|\alpha)+H(\alpha) \\
&\vdots \\
\Pr(y_{it}=1|\alpha) &= [H(\alpha+\rho)-H(\alpha)]\Pr\left(y_{i,t-1}=1|\alpha\right)+H(\alpha)
\end{aligned}$$

The solution to the above difference equation can be written as

$$\Pr(y_{it}=1|\alpha)=[H(\alpha+\rho)-H(\alpha)]^t\Pr(y_{i0}=1|\alpha)+\sum_{s=0}^{t-1}[H(\alpha+\rho)-H(\alpha)]^s H(\alpha).$$

Note that $|H(\alpha+\rho)-H(\alpha)|<1$. As a result, the effect of the initial condition disappears as $T\to\infty$:

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=3}^{T}\Pr\left(y_{i,t-2}=1|\alpha\right)=\frac{H(\alpha)}{1-H(\alpha+\rho)+H(\alpha)}.$$

Thus, we have

$$\frac{1}{T}\sum_{t=3}^{T}y_{i,t-2}y_{it}\xrightarrow{p}\int\left[H(\alpha+\rho)^2+H(\alpha)(1-H(\alpha+\rho))\right]\left[\frac{H(\alpha)}{1-H(\alpha+\rho)+H(\alpha)}\right]g(\alpha)\,d\alpha.$$

Following similar calculations, we can derive the large-T limits of the other components. In particular,

$$\begin{aligned}
\frac{1}{T}\sum_{t=3}^{T}y_{i,t-2}y_{i,t-1} &\xrightarrow{p} \lim_{T\to\infty}\frac{1}{T}\sum_{t=3}^{T}\Pr\left(y_{i,t-1}=1,y_{i,t-2}=1\right) \\
&= \int H(\alpha+\rho)\left[\frac{H(\alpha)}{1-H(\alpha+\rho)+H(\alpha)}\right]g(\alpha)\,d\alpha. \\
\frac{1}{T}\sum_{t=3}^{T}y_{i,t-3}y_{it} &\xrightarrow{p} \lim_{T\to\infty}\frac{1}{T}\sum_{t=3}^{T}\Pr\left(y_{it}=1,y_{i,t-3}=1\right) \\
&= \int H(\alpha+\rho)^3\left[\frac{H(\alpha)}{1-H(\alpha+\rho)+H(\alpha)}\right]g(\alpha)\,d\alpha \\
&\quad +\int 2H(\alpha+\rho)H(\alpha)(1-H(\alpha+\rho))\left[\frac{H(\alpha)}{1-H(\alpha+\rho)+H(\alpha)}\right]g(\alpha)\,d\alpha
\end{aligned}$$

$$+\int H(\alpha)(1-H(\alpha))(1-H(\alpha+\rho))\left[\frac{H(\alpha)}{1-H(\alpha+\rho)+H(\alpha)}\right]g(\alpha)\,d\alpha.$$

Observe that the last term

$$\frac{1}{T}\sum_{t=3}^{T} y_{i,t-3}y_{i,t-1}$$

has the same probability limit as

$$\frac{1}{T}\sum_{t=3}^{T} y_{i,t-2}y_{i,t}$$

as $T\to\infty$. Assembling all the results together, we have as $T\to\infty$,

$$\frac{1}{T}\sum_{t=3}^{T}\Delta y_{i,t-2}\Delta y_{it} \xrightarrow{p} -\int (1-H(\alpha+\rho))H(\alpha)(H(\alpha+\rho)-H(\alpha))\,g(\alpha)\,d\alpha.$$

The other large-T results now follow similar computations.

Derivation of the large-n, large-T limit of the fixed effects estimator

Galvao and Kato (2014) impose assumptions A1 to A3 to derive the large-n, large-T limit of the fixed effects estimator. Assumption A1 is about independence across cross-sectional units and a mild form of time series dependence conditional on α_i. For my case, I needed to impose the assumption that the initial condition is drawn from its stationary distribution conditional on α_i, unlike the derivations for the AH estimators.

Let $\widetilde{y}_{it} = y_{it} - \mathbb{E}(y_{it}=1|\alpha_i) = y_{it} - \Pr(y_{it}=1|\alpha_i)$ for $t=1,\ldots,T$. Assumption A2 is about the existence and boundedness of the moments of $\widetilde{y}_{it}$. These moments are guaranteed to exist and be bounded because $\widetilde{y}_{it}$ has a Bernoulli distribution with probability $\Pr(y_{it}=1|\alpha_i)\in(0,1)$. They show that the fixed effects estimator converges to the following pseudo-true parameter:

$$\beta_0 = \frac{\mathbb{E}\left(\widetilde{y}_{it}\widetilde{y}_{i,t-1}\right)}{\mathbb{E}\left(\widetilde{y}_{i,t-1}^2\right)}.$$

I now calculate the denominator explicitly. First, note that

$$\begin{aligned}\widetilde{y}_{i,t-1}^2 &= y_{i,t-1}^2 - 2y_{i,t-1}\Pr\left(y_{i,t-1}=1|\alpha_i\right) + \left(\Pr\left(y_{i,t-1}=1|\alpha_i\right)\right)^2 \\ &= y_{i,t-1} - 2y_{i,t-1}\Pr\left(y_{i,t-1}=1|\alpha_i\right) + \left(\Pr\left(y_{i,t-1}=1|\alpha_i\right)\right)^2.\end{aligned}$$

Taking expectations, we have

$$\begin{aligned}
\mathbb{E}\left(\widetilde{y}_{i,t-1}^2\right) &= \mathbb{E}\Big[y_{i,t-1} - 2y_{i,t-1}\Pr\left(y_{i,t-1}=1|\alpha_i\right) + \left(\Pr\left(y_{i,t-1}=1|\alpha_i\right)\right)^2\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\left(y_{i,t-1}|\alpha_i\right) - 2\mathbb{E}\left(y_{i,t-1}|\alpha_i\right)\Pr\left(y_{i,t-1}=1|\alpha_i\right) + \left(\Pr\left(y_{i,t-1}=1|\alpha_i\right)\right)^2\Big] \\
&= \mathbb{E}\Big[\Pr\left(y_{i,t-1}=1|\alpha_i\right) - \left(\Pr\left(y_{i,t-1}=1|\alpha_i\right)\right)^2\Big] \\
&= \mathbb{E}\Big[\Pr\left(y_{i,t-1}=1|\alpha_i\right)\left(1-\Pr\left(y_{i,t-1}=1|\alpha_i\right)\right)\Big].
\end{aligned}$$

Note that $\mathbb{E}\left(\widetilde{y}_{i,t-1}^2\right) > 0$ and satisfies assumption A3 of Galvao and Kato (2014). As for the numerator, note that

$$\begin{aligned}
\widetilde{y}_{it}\widetilde{y}_{i,t-1} &= y_{it}y_{i,t-1} - y_{it}\Pr\left(y_{i,t-1}=1|\alpha_i\right) - y_{i,t-1}\Pr\left(y_{it}=1|\alpha_i\right) \\
&\quad + \Pr\left(y_{it}=1|\alpha_i\right)\Pr\left(y_{i,t-1}=1|\alpha_i\right). \qquad (2.6.2)
\end{aligned}$$

Take the first two terms of the right hand side of (2.6.2). Applying law of iterated expectations and $\mathbb{E}\left(y_{it}|y_{i,t-1},\alpha_i\right) = \Pr\left(y_{it}=1|y_{i,t-1},\alpha_i\right)$ gives

$$\begin{aligned}
&\mathbb{E}\left(\left(y_{i,t-1} - \Pr\left(y_{i,t-1}=1|\alpha_i\right)\right)y_{it}\right) \\
=\ &\mathbb{E}\Big[\mathbb{E}\left(\mathbb{E}\left(\left(y_{i,t-1} - \Pr\left(y_{i,t-1}=1|\alpha_i\right)\right)y_{it}|y_{i,t-1},\alpha_i\right)|\alpha_i\right)\Big] \\
=\ &\mathbb{E}\Big[\mathbb{E}\left(\left(y_{i,t-1} - \Pr\left(y_{i,t-1}=1|\alpha_i\right)\right)\mathbb{E}\left(y_{it}|y_{i,t-1},\alpha_i\right)|\alpha_i\right)\Big] \\
=\ &\mathbb{E}\Big[\mathbb{E}\left(\left(y_{i,t-1} - \Pr\left(y_{i,t-1}=1|\alpha_i\right)\right)H\left(\alpha_i+\rho y_{i,t-1}\right)|\alpha_i\right)\Big] \\
=\ &\mathbb{E}\Big[\left(1-\Pr\left(y_{i,t-1}=1|\alpha_i\right)\right)H\left(\alpha_i+\rho\right)\Pr\left(y_{i,t-1}=1|\alpha_i\right)\Big] \\
&-\mathbb{E}\Big[\Pr\left(y_{i,t-1}=1|\alpha_i\right)H\left(\alpha_i\right)\left(1-\Pr\left(y_{i,t-1}=1|\alpha_i\right)\right)\Big].
\end{aligned}$$

The last two terms of the right hand side of (2.6.2) is equal to zero. As a result, we obtain

$$\mathbb{E}\left(\widetilde{y}_{it}\widetilde{y}_{i,t-1}\right) = \mathbb{E}\Big[\left(H\left(\alpha_i+\rho\right)-H\left(\alpha_i\right)\right)\Pr\left(y_{i,t-1}=1|\alpha_i\right)\left(1-\Pr\left(y_{i,t-1}=1|\alpha_i\right)\right)\Big].$$

Combining all these findings give us the final form for the pseudo-true parameter:

$$\beta_0 = \frac{\mathbb{E}\Big[\left(H\left(\alpha_i+\rho\right)-H\left(\alpha_i\right)\right)\Pr\left(y_{i,t-1}=1|\alpha_i\right)\left(1-\Pr\left(y_{i,t-1}=1|\alpha_i\right)\right)\Big]}{\mathbb{E}\Big[\Pr\left(y_{i,t-1}=1|\alpha_i\right)\left(1-\Pr\left(y_{i,t-1}=1|\alpha_i\right)\right)\Big]}.$$

Chapter 3

Simultaneous equations models for discrete outcomes: Coherence and completeness using panel data

3.1 Introduction

In this chapter, I show how to estimate a dynamic simultaneous equations panel data model with discrete outcomes. There are two main issues involved in this endeavor, namely, the manner in which time-invariant unobserved heterogeneity is introduced and the manner in which the nonexistence of a unique reduced form is addressed. Both these issues have implications for how the parameters of the simultaneous equations model are going to be identified and estimated. I show how both these issues can be tackled at the same time.

Researchers who want to estimate a dynamic simultaneous equations model with discrete outcomes using panel data would have to introduce an additive individual-specific fixed effect into the latent index. Unfortunately, time-invariant unobserved heterogeneity cannot be left unrestricted in dynamic nonlinear panel data models, especially when the number of time periods T is small (see Section 4 of Arellano and Bonhomme (2011)). Although we have bias reduction procedures for parameters of interest, they are motivated from a large-T perspective. Results of existing Monte Carlo simulations for dynamic nonlinear panel data models indicate that T has to be much larger than 10 in order to reap the gains from bias reduction (see Bester and Hansen (2009a), Carro (2007), Fernandez-Val (2009), and Hahn and Kuersteiner

(2011)). Furthermore, the fixed-T solution proposed by Bonhomme (2012) only applies to models without dynamics. As a compromise, we have to restrict some features of the distribution of time-invariant unobserved heterogeneity. Allowing for fixed effects in these models has not been explored fully, since most research has focused on either cross-sectional models, continuous outcomes (or the latent outcomes themselves), or random effects (for examples, see the work by Cornwell, Schmidt, and Wyhowski (1992), Leon-Gonzalez (2003), Matzkin (2008), Matzkin (2012), and Masten (2015)).

Even if T is large, coherency conditions would still have to be imposed. Coherency conditions effectively convert a model where the endogenous variables are jointly determined into a model which is triangular or recursive. A triangular model restricts the direction in which an endogenous variable affects other endogenous variables (for all observations). Triangularity implies that there are either zero or inequality restrictions on the parameters or functions of the parameters. For example, when there are two binary endogenous variables y_1 and y_2 that are jointly determined, y_2 should not enter the equation for y_1 or vice-versa. As a result, we have to choose beforehand how the coherency conditions should be imposed.

The literature on coherency conditions started with research aiming to extend the simultaneous equations approach of the Cowles Commission to endogenous variables that are subject to censoring or truncation. Some representative papers in this area include Maddala and Lee (1976), Heckman (1978), Gourieroux, Laffont, and Monfort (1980), and Schmidt (1981). Blundell and Smith (1993) summarize this strand of the literature. They have all shown that parameter restrictions are typically required to ensure the existence and uniqueness of the reduced form. As a result, ensuring coherency is a first step prior to discussing identification. Later research has focused more on how to avoid imposing the coherency conditions (see the contribution of Tamer (2003)).

In order to avoid imposing these coherency conditions, a separate strand of the literature has emphasized that the uniqueness of equilibrium in game-theoretic models has parallels with the problems involving uniqueness of the reduced form for simultaneous equation models with discrete outcomes. Early work in the estimation of game-theoretic models such as Bjorn and Vuong (1984), Bresnahan and Reiss (1991), and Kooreman (1994) attempt to overcome the possibility of multiple equilibria either by introducing a selection mechanism which assumes that players choose one of the equilibria at random or by fusing multiple equilibria as one outcome. Tamer (2003) shows that point identification and consistent estimation is still possible without imposing a set of auxiliary assumptions that resolve the underlying multiplicity. All that is needed is the presence of a regressor with large support. He suggests a semiparametric ML estimator that is more efficient than the ML estimator that fuses multiple equilibria as one outcome. In fact, Tamer (2003) introduces new terminology to differentiate models whose reduced form is nonexistent and models

whose reduced form is nonunique. He calls these models incoherent and incomplete, respectively. Note that these cited papers focus on the incompleteness aspect because the signs of some parameters of game-theoretic models may be known a priori. These sign restrictions effectively rule out potential incoherence of the model.

Some like Dagenais (1999), Massacci (2010), and Hajivassiliou and Savignac (2011) have attempted to resolve both incompleteness and incoherence by imposing error-support restrictions. They all offer estimation methods that involve reweighting the likelihood contributions to reflect the restrictions on the error supports. In the most recent work by Chesher and Rosen (2012), they show how identified sets can be constructed without resorting to the restriction of error supports and a priori sign restrictions at all. In the process of constructing these identified sets, they were able to unify the different approaches that are available in the literature in the most general way possible. With the exception of Hajivassiliou and Savignac (2011), the preceding papers focus on either cross-sectional or time-series settings. On the other hand, Hajivassiliou and Savignac (2011) use panel data to estimate a model of the joint determination of a firm's decision to innovate and a firm's exposure to higher credit constraints but do not allow for fixed effects.

My proposal approaches the problem of incompleteness and incoherence from a different perspective. I exploit Lewbel's (2007) characterization of coherence and completeness when one of the endogenous variables is binary. He shows that a possible characterization involves indexing the direction of causality by a dummy variable which may be observable or modeled. In contrast, I do not restrict the dependence on observables but assume that this dependence is individual-specific. As a result, I allow for the estimation of a panel data simultaneous equations model with discrete outcomes where the individual-specific fixed effect can be interpreted as the manner in which the coherency condition holds or the direction in which causality flows from one discrete endogenous variable to another.

The paper is organized as follows. In Section 3.2, I provide a motivating example to demonstrate my proposal. In Section 3.3, I discuss how identification, estimation and inference may proceed in the model considered by Hajivassiliou and Ioannides (2007) (henceforth, HI). In Section 3.4, I revisit the empirical application of HI and cast doubt on the coherency conditions they have imposed. I conclude and suggest avenues for further work in Section 3.5.

3.2 A stylized example

3.2.1 Coherence and completeness

I start by introducing some terminology in the context of a simple example. Consider the situation where two dummy variables are jointly determined. This situation typically arises in many empirical applications, such as determining whether binary

choices are substitutes or complements (Lewbel, 2007), estimating game-theoretic models with discrete actions (Bjorn and Vuong, 1984; Bresnahan and Reiss, 1991; Kooreman, 1994; Tamer, 2003; Hahn and Moon, 2010), modelling vote trading among congressmen for agricultural issues (Stratmann, 1992; Dagenais, 1999), and modelling fertility decisions among couples (Sobel and Arminger, 1992), to name a few.

Let (y_1, y_2) be two dummy endogenous variables jointly determined by the system

$$y_1^* = y_2\alpha_1 + \epsilon_1,\ y_1 = 1\left\{y_1^* \geq 0\right\}, \tag{3.2.1}$$
$$y_2^* = y_1\alpha_2 + \epsilon_2,\ y_2 = 1\left\{y_2^* \geq 0\right\}, \tag{3.2.2}$$

where and (ϵ_1, ϵ_2) are the error terms. Only the signs of y_1^* and y_2^* are observable.[1] There are four possible outcomes for (y_1, y_2) and they arise according to the following rule:

$$(y_1, y_2) = \begin{cases} (1,1) & \text{if } \epsilon_1 > -\alpha_1,\ \epsilon_2 > -\alpha_2 \\ (1,0) & \text{if } \epsilon_1 > 0,\ \epsilon_2 \leq -\alpha_2 \\ (0,1) & \text{if } \epsilon_1 \leq -\alpha_1,\ \epsilon_2 > 0 \\ (0,0) & \text{if } \epsilon_1 \leq 0,\ \epsilon_2 \leq 0 \end{cases}.$$

Geometrically, the inequalities define regions in (ϵ_1, ϵ_2)-space. These regions will overlap when $\alpha_2\alpha_1 > 0$. As a result, y_1 may be assigned the value 0 or 1 in the overlapping region. This non-uniqueness of y_1 is called incompleteness. The model is indeterminate for y_1 for some (ϵ_1, ϵ_2).

On the other hand, the inequalities may lead to an empty region when $\alpha_2\alpha_1 < 0$. As a result, the model is unable to definitively assign a value for y_1 in the empty region. This nonexistence is called incoherence. Unfortunately, the data do not allow us to distinguish between the two cases unless we have prior information about the sign of $\alpha_2\alpha_1$ or we have a way of resolving how Nature (or perhaps the observed units) would assign (or choose) values for y_1 in those regions. We can overcome this by assuming $\alpha_2\alpha_1 = 0$. This restriction, called the coherency condition, assumes away the simultaneity initially posited for the endogenous variables.[2] However, even

[1] Blundell and Smith (1994) call the model above Type II because it is the observed indicators (y_1, y_2) that enter as right-hand side variables. In contrast, Type I models are models in which the latent variables $\left(y_1^*, y_2^*\right)$ enter as right-hand side variables. In the latter case, standard simultaneous equation methods can be applied. Matzkin (2012) considers panel data versions of Type I models.

[2] Another example is where $y_2^* = y_2$ is fully observable, as opposed to just the sign of y_2 being observable. Solving for y_1^* gives us

$$y_1^* = y_1\alpha_2\alpha_1 + \alpha_1\epsilon_2 + \epsilon_1.$$

There are only two possible observable values for y_1:

$$y_1 = \begin{cases} 0 & \text{if } \epsilon_1 + \alpha_1\epsilon_2 \leq 0 \\ 1 & \text{if } \epsilon_1 + \alpha_1\epsilon_2 > -\alpha_2\alpha_1 \end{cases}.$$

if we take the required condition that $\alpha_2\alpha_1 = 0$ seriously, it is not clear whether we should proceed with identification and estimation under $\alpha_1 = 0$ or $\alpha_2 = 0$. Because the coherency condition has to be imposed, most empirical applications would proceed by (a) producing two sets of results (depending on whether $\alpha_1 = 0$ or $\alpha_2 = 0$) (b) choosing to start with a triangular model from the onset (c) introducing the latent variables y_1^* and y_2^* instead of y_1 and y_2 in Equations (3.2.1) and (3.2.2).

Lewbel (2007) shows that it is possible to avoid setting either $\alpha_1 = 0$ or $\alpha_2 = 0$ by choosing a coherent and complete representation. In a nonseparable simultaneous equations model where one of the endogenous variables is binary, he shows that coherence and completeness restrict the manner in which some of the endogenous variables enter into the structural equations:

Theorem 3.2.1. *Assume that $y_1 \in \{0,1\}$, $y_2 \in \Psi$, and $w \in \Omega$ for some support sets Ψ and Ω.*[3] *The system*

$$\begin{aligned} y_1 &= H_1(y_1, y_2, w), \\ y_2 &= H_2(y_1, y_2, w) \end{aligned}$$

is coherent and complete if and only if there exists a function $g : \{0,1\} \times \Omega \to \Psi$ such that for all $w \in \Omega$, we have

$$\begin{aligned} H_1(0, g(0,w), w) &= H_1(1, g(1,w), w), \\ y_2 &= g(y_1, w). \end{aligned}$$

The proof of this theorem can be found in Lewbel (2007). As a result, the function H_1 should not depend on y_1. More importantly, the theorem allows us to choose g to ensure coherence and completeness without imposing sign restrictions or imposing error support restrictions that may be data-dependent. In the context of the model in (3.2.1) and (3.2.2), he shows that a coherent and complete representation can be chosen by defining a dummy function $d : \Omega \to \{0,1\}$ such that

$$\begin{aligned} y_1 &= 1\{(1 - d(w)) y_2\alpha_1 + \epsilon_1 \geq 0\}, \\ y_2 &= 1\{d(w) y_1\alpha_2 + \epsilon_2 \geq 0\}. \end{aligned}$$

The inclusion of d has intuitive appeal because some units may have $d = 0$ or $d = 1$ depending on the values of the observables and unobservables. As a result, y_1

Once again, we must have $\alpha_2\alpha_1 = 0$ so that a unique reduced form would exist. More empirical applications belong to this category of models. Examples include modelling the export-productivity relationship (Clerides, Lach, and Tybout, 1998), the household's bundle of appliance holdings and electricity consumption (Dubin and McFadden, 1984), and the measurement of recovery via output growth from crises (Cerra and Saxena, 2008).

[3] These support sets may contain either a finite or an infinite number of elements. The only requirement is that y_1 is binary.

depends on y_2 when $d = 0$ and y_2 depends on y_1 when $d = 1$. Unfortunately, Lewbel (2007) assumes that d is observable or could be modeled in some way.[4]

Note that Lewbel's (2007) approach gives a new interpretation for simultaneity when compared to its traditional usage in econometrics. Instead of simultaneity being a purely structural aspect of the model, simultaneity arises because of the econometrician's inability to distinguish the direction of dependence (from y_1 to y_2 or vice versa). Intuitively, information from repeated measurements can be useful in overcoming this inability without making parametric assumptions on the dummy function d.

In contrast to linear simultaneous equations models where there is a form of "bidirectional" causality (not in the usual Granger-causality sense) between two continuously distributed variables, the models I consider only allow for one-directional causality from y_1 to y_2 for a subset of observations and one-directional causality from y_2 to y_1 for the remaining set of observations. This is more credible than imposing the coherency conditions which will ultimately result in either one-directional causality from y_1 to y_2 for *all* observations or one-directional causality from y_2 to y_1 for *all* observations, but not both.

The model in (3.2.1) and (3.2.2) can be thought of as a static discrete game of complete information. In contrast to game-theoretic models, the models I consider allow for both incompleteness and incoherence. Incompleteness arising from multiple equilibria is a common feature in the estimation of game-theoretic models. Hahn and Moon (2010) estimate (3.2.1) and (3.2.2) using panel data of pairs of agents using Nash play. They assume that the model is incomplete by imposing sign restrictions $\alpha_1 < 0$ and $\alpha_2 < 0$ and that pairs of agents choose one of the two equilibria and stick to that same choice throughout time.

Since I am working with a complete and coherent representation by introducing d, it may be useful to motivate the representation in game-theoretic terms. One can view the representation as arising from the inability of the econometrician to observe how multiple equilibria or absence of equilibrium was resolved, possibly through an unmodelled communication or coordination device. Alternatively, the econometrician may also have neglected to model the sequential nature of the game and was unable to observe which player moved first (but the agents are aware of the sequential nature of the game).

[4]Introducing this dummy function is just one way to obtain a coherent and complete representation whenever one of the endogenous variables is binary. One could argue that this device may be useful in a linear simultaneous equations setting. A step in this direction is the introduction of random coefficients in a linear simultaneous equations model (see Masten (2015) for more details).

3.2.2 Why a cross section is not enough

Suppose we treat d as unobservable. For the moment, assume we only have a cross-section of observations which form a random sample. Consider the model

$$y_{1i} = 1((1-d_i)\,y_{2i}\alpha_1 + \epsilon_{1i} \geq 0), \tag{3.2.3}$$
$$y_{2i} = 1(d_i y_{1i}\alpha_2 + \epsilon_{2i} \geq 0), \tag{3.2.4}$$

for $i = 1,\ldots,n$. Let $\Pr(d_i = 1) = p_i$ and $\Pr(d_i = 0) = 1-p_i$, where $p_i \in (0,1)$. I assume that d_i, ϵ_{1i} and ϵ_{2i} are i.i.d. draws from their joint distribution. There are four joint probabilities of the form $\Pr(y_{1i} = j, y_{2i} = k)$, where $(j,k) \in \{0,1\}\times\{0,1\}$, that are observable from the data but only three of these provide restrictions on the parameters of the model (since all these four probabilities should sum up to one). As a result, it is not possible to have point identification for all parameters, even if we set $p_i = p$ for all i. The main reason is that there are four parameters $(\alpha_1, \alpha_2, \rho, p)$ to identify given the three joint probabilities obtained from the data.

To illustrate and simplify things further, assume normality[5][6] for the errors given d_i for all i, i.e.

$$\left.\begin{pmatrix} \epsilon_{1i} \\ \epsilon_{2i} \end{pmatrix}\right| d_i \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right). \tag{3.2.5}$$

As a result, we have

$$\left.\begin{pmatrix} \epsilon_{1i} \\ \epsilon_{2i} \end{pmatrix}\right| d_i = 0 \sim \left.\begin{pmatrix} \epsilon_{1i} \\ \epsilon_{2i} \end{pmatrix}\right| d_i = 1,$$

i.e., $(\epsilon_{1i}, \epsilon_{2i})$ is independent of d_i. The parameters of interest are α_1, α_2, and ρ. Note that for all $(j,k) \in \{0,1\}\times\{0,1\}$, we have

$$\begin{aligned}\Pr(y_{1i} = j, y_{2i} = k) = {} & \Pr(y_{1i} = j, y_{2i} = k | d_i = 1)\, p_i \\ & + \Pr(y_{1i} = j, y_{2i} = k | d_i = 0)(1-p_i).\end{aligned}$$

Specifically, we have

$$\Pr(y_{1i} = 0, y_{2i} = 0) = \Pr(\epsilon_{1i} \leq 0, \epsilon_{2i} \leq 0;\, \rho), \tag{3.2.6}$$

$$\Pr(y_{1i} = 1, y_{2i} = 0) = \Pr(\epsilon_{1i} \geq 0, \epsilon_{2i} \leq -\alpha_2;\, \rho)\, p_i + \Pr(\epsilon_{1i} \geq 0, \epsilon_{2i} \leq 0;\, \rho)(1-p_i), \tag{3.2.7}$$

$$\Pr(y_{1i} = 0, y_{2i} = 1) = \Pr(\epsilon_{1i} \leq -\alpha_1, \epsilon_{2i} \geq 0;\, \rho)(1-p_i) + \Pr(\epsilon_{1i} \leq 0, \epsilon_{2i} \geq 0;\, \rho)\, p_i. \tag{3.2.8}$$

[5] Normality is not really required here. All that is required is a bivariate cdf that is strictly monotonic in ρ. In cases where ρ is not a correlation coefficient but some one-dimensional parameter that indexes dependence modelled via a copula. See Section 3.3 for more details.

[6] Bivariate normality of the errors can be thought of as a simple factor structure. In particular, ϵ_{1i} can be written as $\epsilon_{1i} = \rho\epsilon_{2i} + v_i$ where $v_i \sim N\left(0, 1-\rho^2\right)$, $\epsilon_{2i} \sim N(0,1)$, and ϵ_{2i} is independent of v_i.

The left hand sides of (3.2.6), (3.2.7), and (3.2.8) are observable from the data. We are unable to observe the mixing probability p_i. As long as we observe outcomes of the form $(y_{1i} = 0, y_{2i} = 0)$ from the data,[7] (3.2.6) can be used to identify ρ because the bivariate normal cdf is strictly monotonic in ρ (hence, the bivariate normal cdf is invertible with respect to ρ). This is in contrast with identification problems associated with ρ as documented by Freedman and Sekhon (2010) and Meango and Mourifie (2013) in the context of triangular models with a dummy endogenous regressor.

Since ρ is identified, we can treat it as known for the next step. In particular, we can now use (3.2.7) and (3.2.8) to identify whether (i) $\alpha_1 \leq 0$ or $\alpha_1 \geq 0$ and (ii) $\alpha_2 \leq 0$ or $\alpha_2 \geq 0$. Whenever the cross-sectional frequency of $(y_{1i} = 1, y_{2i} = 0)$ is less than or equal to $\Pr(\epsilon_{1i} \geq 0, \epsilon_{2i} \leq 0; \rho)$, we must have $\alpha_2 \geq 0$. Similarly, showing that the cross-sectional frequency of $(y_{1i} = 0, y_{2i} = 1)$ is less than or equal to $\Pr(\epsilon_{1i} \leq 0, \epsilon_{2i} \geq 0; \rho)$ allows us to conclude that $\alpha_1 \geq 0$. The other cases follow analogously.

Unfortunately, we need external information to determine the grouping of every ith unit. One possible route is to impose $p_i = \Pr(d_i = 1) = \Pr(d = 1) = p$ for all i. Knowing the signs of α_1, α_2 allows us to determine the group to which all units belong. There are four cases to consider as shown in Table 3.2.1.

Table 3.2.1: Group assignment rules for model (3.2.3) and (3.2.4)

Condition	Rule
$\alpha_1 \geq 0$, $\alpha_2 \geq 0$	$d = 0$ iff frequency of $y_{1i} = 0$ less than $\Pr(\epsilon_{1i} \leq 0)$ $d = 1$ iff frequency of $y_{2i} = 0$ less than $\Pr(\epsilon_{2i} \leq 0)$
$\alpha_1 \leq 0$, $\alpha_2 \leq 0$	$d = 0$ iff frequency of $y_{1i} = 0$ greater than $\Pr(\epsilon_{1i} \leq 0)$ $d = 1$ iff frequency of $y_{2i} = 0$ greater than $\Pr(\epsilon_{2i} \leq 0)$
$\alpha_1 \geq 0$, $\alpha_2 \leq 0$	$d = 0$ iff frequency of $y_{1i} = 0$ less than $\Pr(\epsilon_{1i} \leq 0)$ $d = 1$ iff frequency of $y_{2i} = 0$ greater than $\Pr(\epsilon_{2i} \leq 0)$
$\alpha_1 \leq 0$, $\alpha_2 \geq 0$	$d = 0$ iff frequency of $y_{1i} = 0$ greater than $\Pr(\epsilon_{1i} \leq 0)$ $d = 1$ iff frequency of $y_{2i} = 0$ less than $\Pr(\epsilon_{2i} \leq 0)$

Once we know whether $d = 0$ or $d = 1$ for all i, we are able to only point-identify either α_1 or α_2 but not both. Suppose $d = 0$ for the moment. Since $\Pr(y_{1i} = 0, y_{2i} = 1|d = 0) = \Pr(\epsilon_{1i} \leq -\alpha_1, \epsilon_{2i} \geq 0; \rho)$ and $\Pr(\epsilon_{1i} \leq -\alpha_1, \epsilon_{2i} \geq 0; \rho)$ is strictly decreasing in α_1 for fixed ρ, we are able to point-identify α_1. In contrast, α_2 is set-identified because $d = 0$ for all observations and we know the sign

[7] The extreme case where we do not observe any other outcome aside from $(y_{1i} = 0, y_{2i} = 0)$ is ruled out.

of α_2. Similarly, we are able to point-identify α_2 from $\Pr(y_{1i}=1, y_{2i}=0|d=1) = \Pr(\epsilon_{1i} \geq 0, \epsilon_{2i} \leq -\alpha_2; \rho)$ but only set-identify α_1.

3.2.3 Why panel data may be useful

At this point, panel data may be useful for point identification of all the parameters. We can introduce individual-specific effects through d to allow for unrestricted dependence on the observables in a time-invariant manner. Note that I do not introduce individual-specific effects as intercepts in the linear predictors. As a result, time-invariant variables or even variables that do not have too much variation over time may be included in the model. In contrast, the usual way of introducing fixed effects additively precludes the inclusion of time-invariant variables of interest.

If we had panel data and imposed the assumption that p_i does not vary over time, we will be able to achieve point identification. Consider once again the model but this time adapted to panel data:

$$\begin{aligned} y_{1it} &= 1((1-d_i)\, y_{2it}\alpha_1 + \epsilon_{1it} \geq 0), \\ y_{2it} &= 1(d_i y_{1it}\alpha_2 + \epsilon_{2it} \geq 0), \end{aligned}$$

for $i = 1, \ldots, n$ and $t = 1, \ldots, T$. Let $\Pr(d_i = 1) = p_i$ and $\Pr(d_i = 0) = 1 - p_i$, where $p_i \in (0,1)$. I assume that d_i, ϵ_{1it} and ϵ_{2it} are i.i.d. draws from their joint distribution. Assume bivariate normality once again as in (3.2.5).[8] Analogously, we have

$$\Pr(y_{1it}=0, y_{2it}=0) = \Pr(\epsilon_{1it} \leq 0, \epsilon_{2it} \leq 0; \rho), \tag{3.2.9}$$

$$\Pr(y_{1it}=1, y_{2it}=0) = \Pr(\epsilon_{1it} \geq 0, \epsilon_{2it} \leq -\alpha_2; \rho)\, p_i + \Pr(\epsilon_{1it} \geq 0, \epsilon_{2it} \leq 0; \rho)(1-p_i), \tag{3.2.10}$$

$$\Pr(y_{1it}=0, y_{2it}=1) = \Pr(\epsilon_{1it} \leq -\alpha_1, \epsilon_{2it} \geq 0; \rho)(1-p_i) + \Pr(\epsilon_{1it} \leq 0, \epsilon_{2it} \geq 0; \rho)\, p_i. \tag{3.2.11}$$

We now follow the same steps in the identification argument of the previous subsection. Data on the observed frequencies of $(y_{1it}=0, y_{2it}=0)$ for all i and t point identify ρ from (3.2.9). After plugging in the value of ρ from the previous step, (3.2.10) and (3.2.11) for all i and t can be used to identify the signs of α_1, α_2 as before. Once the signs are identified, we can modify the group assignment rules in Table 3.2.1. Instead of using cross-sectional variation, we use time series variation of

[8] In footnote 6, bivariate normality can be rewritten as a factor structure. In the panel data case, the possibilities are richer. In particular, we may have $\epsilon_{1it} = \lambda_1\theta_i + \nu_{1it}$ and $\epsilon_{2it} = \lambda_2\theta_i + \nu_{2it}$, where θ_i is an individual-specific factor. A similar idea appears in Cameron and Taber (2004). They impose a random effects assumption on θ_i. It may be possible to modify the identification argument I present to allow for this. Unfortunately, the right hand side of Equation 3.2.9 will become an integral that depends on the distribution of θ_i. An identification argument based on the large-sample limit of the likelihood function for observations where $(y_{1it}=0, y_{2it}=0)$ may be used. It is unclear whether we can avoid the distributional assumption on θ_i. Preliminary research by Khan, Maurel, and Zhang (2015) points to the identifying power of factor structures in triangular discrete response models.

every unit to decide whether $d_i = 0$ or $d_i = 1$. Notice that d is now allowed to vary across cross-sectional units. Data on the observed frequencies of $(y_{1it} = 0, y_{2it} = 1)$ for all t and i such that $d_i = 0$ point-identify α_1. Similarly, data on the observed frequencies of $(y_{1it} = 1, y_{2it} = 0)$ for all t and i such that $d_i = 1$ point-identify α_2.

What I have shown is that point identification may be possible in a model where both (y_1, y_2) are dummy endogenous variables jointly determined by the system (3.2.1) and (3.2.2) without imposing sign restrictions, as long as we have access to panel data. The approach considered here is slightly different from the entry-exit game estimated by Hahn and Moon (2010) using panel data. They impose sign restrictions (motivated by economic theory) producing an incomplete game-theoretic model. In their model, the econometrician is unable to observe which equilibrium was selected by the players but whichever equilibrium is selected becomes fixed across time. The approach considered here is also different from Hajivassiliou and Savignac (2011). They do not have fixed effects and they restrict error supports conditional on the restriction that both α_1 and α_2 must not have the same sign.

The intuition behind the identification argument in both the cross-section and panel data case is to find a subset of the data unaffected by the mixing probability p_i. I use this subset of the data to identify the common parameters except for α_1 and α_2. Deciding whether $d_i = 0$ or $d_i = 1$ depends on time series variation. After deciding the grouping, α_1 and α_2 can now be point identified. It is important to note that the number of groups is known in advance. Lewbel's (2007) characterization of complete and coherent two-equation systems ensure that the number of groups is fixed at 2.

There are extensions of situations that follow essentially the identification argument above. For instance, the identification argument can be extended to the case where we have an intercept, strictly exogenous regressors, and lagged dependent variables. Another possibility is for the error terms to have other known marginal distributions linked by a parametric copula as in Han and Vytlacil (2015) but extended to the panel data case. Finally, we can accommodate other discrete choice models such as the multinomial logit/probit and ordered logit/probit.

3.3 The model

3.3.1 Background

I now describe how to identify and estimate the parameters of the model to be used in the empirical application. HI (2007) construct a model of a household head living in finite time who chooses consumption and hours worked subject to a liquidity constraint and a quantity constraint on labor supply. As a consequence of liquidity constraints, household heads cannot hold negative wealth at any time over the life cycle. Furthermore, they may be subject to involuntary unemployment / underem-

ployment, voluntary employment, or involuntary overemployment. As a result, they can be in one of the following situations – (a) they are able to work but are unable to reach their desired number of hours, (b) they are able to work at their desired number of hours, or (c) they are working beyond their desired level of hours. HI (2007) derive the solutions to the optimization problem faced by a representative household. The solutions to the optimization problem represent the optimal path of assets and hours worked over the life cycle of the household.

The econometrician is able to observe whether or not the household head is liquidity constrained, i.e., the liquidity constraint indicator S_{it} takes on the value 1 or 0, respectively. Household heads could be involuntarily overemployed ($E_{it} = -1$), voluntarily employed ($E_{it} = 0$), or involuntarily unemployed / underemployed ($E_{it} = 1$). The authors describe how these indicators were constructed in their earlier paper (see HI (1995)).

Since these optimal paths of assets and hours worked are determined jointly over the household's life cycle, the econometric treatment would have to acknowledge the underlying simultaneity. HI (2007) argue that one can either model the employment constraint indicator conditionally on the liquidity constraint indicator or vice versa. They further point out that this is consistent with the intertemporal two-stage budgeting of households described in Blundell and Walker (1986). Clearly, Lewbel's (2007) characterization can be exploited without us choosing the causal direction in advance or by presenting two sets of results.

HI (2007) specify their dynamic simultaneous equations model as follows:

$$\begin{aligned}
S_{it}^{*} &= \gamma_{11} S_{i,t-1} + \gamma_{12} S_{i,t-2} + \delta_0 E_{it} + \delta_1 E_{i,t-1} + \delta_2 E_{i,t-2} \\
&\quad + X_{1it}\beta^{bp} + \epsilon_{it}^{bp}, \qquad (3.3.1) \\
E_{it}^{*} &= \gamma_{21} E_{i,t-1} + \gamma_{22} E_{i,t-2} + \kappa_0 S_{it} + \kappa_1 S_{i,t-1} + \kappa_2 S_{i,t-2} \\
&\quad + X_{2it}\beta^{op} + \epsilon_{it}^{op}, \qquad (3.3.2) \\
S_{it} &= 1\left\{S_{it}^{*} \geq 0\right\}, \\
E_{it} &= -1\left\{E_{it}^{*} < \theta^{-}\right\} + 1\left\{E_{it}^{*} > \theta^{+}\right\}
\end{aligned}$$

where θ^{-} and θ^{+} are lower and upper thresholds. Just as in HI (2007), I also normalize $\theta^{+} = 0$. Observe that all the employment status indicators should really enter as two dummies because there are three categories. For example, $\delta_0 E_{it}$ can be decomposed into $\delta_{01} 1\{E_{it} = -1\} + \delta_{02} 1\{E_{it} = 1\}$. HI (1995; 2007) show that the coherency conditions are $(\delta_{01} + \delta_{02})\kappa_0 = 0$ and $\delta_{01}\delta_{02}\kappa_0 = 0$. As a result, we either have $\kappa_0 = 0$ or $\delta_{01} = \delta_{02} = 0$. If we exploit Lewbel's (2007) result, we need not impose these coherency conditions at all. Next, I show how to identify the parameters of their model without imposing coherency conditions.

3.3.2 Identification

To discuss the identification argument for the parameters in the model (3.3.1) and (3.3.2). I make the following assumptions:

A1 (Data generating process) For all i and t, S_{it} and E_{it} are generated by the model

$$\begin{aligned}
S_{it}^* &= \gamma_{11}S_{i,t-1} + \gamma_{12}S_{i,t-2} + d_i\delta_{01}1\{E_{it}=-1\} + d_i\delta_{02}1\{E_{it}=1\} \\
&\quad +\delta_{11}1\left\{E_{i,t-1}=-1\right\} + \delta_{12}1\left\{E_{i,t-1}=1\right\} + \delta_{21}1\left\{E_{i,t-2}=-1\right\} \\
&\quad +\delta_{22}1\left\{E_{i,t-2}=1\right\} + X_{1it}\beta^{bp} + \epsilon_{it}^{bp}, \qquad (3.3.3) \\
E_{it}^* &= \gamma_{211}1\left\{E_{i,t-1}=-1\right\} + \gamma_{212}1\left\{E_{i,t-1}=1\right\} + \gamma_{221}1\left\{E_{i,t-2}=-1\right\} \\
&\quad +\gamma_{222}1\left\{E_{i,t-2}=1\right\} + (1-d_i)\kappa_0 S_{it} + \kappa_1 S_{i,t-1} + \kappa_2 S_{i,t-2} \\
&\quad +X_{2it}\beta^{op} + \epsilon_{it}^{op} \qquad (3.3.4) \\
S_{it} &= 1\left\{S_{it}^* \geq 0\right\}, \\
E_{it} &= -1\left\{E_{it}^* < \theta^-\right\} + 1\left\{E_{it}^* > 0\right\},
\end{aligned}$$

where S_{it}^* and E_{it}^* are latent variables. The parameters representing the simultaneous effects δ_{01}, δ_{02}, and κ_0 cannot all be jointly equal to zero.

The representation of the model in A1 is a result of applying Lewbel's (2007) characterization of a coherent and complete representation. Note that $X_{1it} \in \mathbb{R}^{p_1}$ and $X_{2it} \in \mathbb{R}^{p_2}$ may have common elements. The equation for S_{it}^* is a binary choice model while the equation for E_{it}^* is an ordered choice model. The superscripts bp and op refer to binary probability model and ordered probability model, respectively.

A2 (Exogeneity restrictions) Let

$$\begin{aligned}
Z_i^t &= \left(S_{i,t-1},\ldots,S_{i0},S_{i,-1},E_{i,t-1},\ldots,E_{i0},E_{i,-1}\right), \\
Z_{it} &= \left(S_{i,t-1},S_{i,t-2},E_{i,t-1},E_{i,t-2}\right),
\end{aligned}$$

$X_{1i}^T = \left(X_{1i,-t},X_{1it}\right)$, and $X_{2i}^T = \left(X_{2i,-t},X_{2it}\right)$. For all i and t, the error terms satisfy

$$\left(\begin{array}{c}\epsilon_{it}^{bp}\\ \epsilon_{it}^{op}\end{array}\right)\Bigg| Z_i^t, X_{1i}^T, X_{2i}^T \sim \left(\begin{array}{c}\epsilon_{it}^{bp}\\ \epsilon_{it}^{op}\end{array}\right)\Bigg| Z_{it}, X_{1it}, X_{2it}.$$

Assumption A2 establishes some notation adapted from the dynamic panel data and game theory literatures. The notation for X_{1i}^T splits $(X_{1i1},X_{1i2},\ldots,X_{1it},\ldots,X_{1iT})$ into a period t component X_{1it} and a component

$$X_{1i,-t} = \left(X_{1i1},X_{1i2},\ldots,X_{1i,t-1},X_{1i,t+1}\ldots,X_{1iT}\right)$$

representing all the other time periods except period t. I use the same notation for X_{2i}^T. Assumption A2 establishes that Z_{it} represents the predetermined regressors and X_{1it}, X_{2it} represent the strictly exogenous regressors.

A3 (Error distribution) The error terms are i.i.d. draws from the conditional distribution

$$\left(\begin{array}{c} \epsilon_{it}^{bp} \\ \epsilon_{it}^{op} \end{array} \right) \Bigg| Z_{it}, X_{1it}, X_{2it} \sim C\left(F_{\epsilon^{bp}}\left(\epsilon^{bp}\right), F_{\epsilon^{op}}\left(\epsilon^{op}\right); \rho\right)$$

where $C(\cdot,\cdot;\rho)$ is a copula known up to a scalar parameter $\rho \in \Omega$ such that $C : (0,1)\times(0,1) \to (0,1)$ and Ω is an open subset of $\mathbb{R}$. The copula $C(u_1,u_2;\rho)$ is continuously differentiable everywhere in its domain $(u_1,u_2,\rho) \in (0,1) \times (0,1) \times \Omega$. $F_{\epsilon^{bp}}$ and $F_{\epsilon^{op}}$ are known marginal distribution functions for ϵ^{bp} and ϵ^{op}, respectively, that are strictly increasing, are absolutely continuous with respect to Lebesgue measure, and such that $\mathbb{E}\left(\epsilon^{bp}\right) = \mathbb{E}(\epsilon^{op}) = 0$ and $Var\left(\epsilon^{bp}\right) = Var\left(\epsilon^{op}\right) = 1$.

In contrast to the previous section where I imposed bivariate normality, I allow for a larger class of parametric models in A3. Furthermore, there is a large selection of copulas that are available (see the survey by Trivedi and Zimmer (2007), a textbook treatment by Nelsen (2006), and an application by Winkelmann (2012)). In contrast to Han and Vytlacil (2015), I do not impose any stochastic dominance assumptions on the selected copula. The assumptions on the marginal distributions $F_{\epsilon^{bp}}$ and $F_{\epsilon^{op}}$ are needed to ensure smoothness and invertibility. The restrictions on the moments of the error terms are typical normalizations in the discrete choice literature since the parameters are identified up to scale.

A4 (Finite support of fixed effects) The fixed effects d_i have known finite support $\{0,1\}$ for all i and are conditionally independent draws from some unknown distribution. Furthermore, $d_i \perp \left(\epsilon_{it}^{bp}, \epsilon_{it}^{op}\right) | S_{i0}, S_{i,-1}, E_{i0}, E_{i,-1}, X_{1i}^T, X_{2i}^T$ for all i and t.

Lewbel's (2007) characterization ensures that the support of the fixed effects is finite and has cardinality equal to 2. Assumption A4 is an assumption in the spirit of fixed-effects models. The independence assumption, however, is much stronger than the zero correlation between the fixed effect and the idiosyncratic error one usually encounters in linear panel data models.

A5 (Support and rank conditions) For all i and t, there exists some regressor (say the kth regressor) X_{1itk} with $\beta_k^{bp} \neq 0$ such that the distribution of $X_{1itk} | X_{1it,-k}$ has an everywhere positive Lebesgue density where

$$X_{1it,-k} = \left(X_{1it1}, \ldots, X_{1it,k-1}, X_{1it,k+1}, \ldots, X_{1itp_1}\right).$$

For all i and t, the regressors X_{1it} and X_{2it} have full column rank. Furthermore, for all i and t, we have

$$\begin{aligned}\Pr\left(\text{supp}\left(X_{1it}\beta^{bp}\right)\cap\text{supp}\left(X_{1it}\beta^{bp}+s_1\right)\right) &> 0,\\ \Pr\left(\text{supp}\left(X_{2it}\beta^{op}\right)\cap\text{supp}\left(X_{2it}\beta^{op}+s_2\right)\right) &> 0,\end{aligned}$$

where $s_1\in\{\gamma_{11},\gamma_{12},\delta_{11},\delta_{12},\delta_{21},\delta_{22}\}$ and $s_2\in\{\gamma_{211},\gamma_{212},\gamma_{221},\gamma_{222},\kappa_1,\kappa_2\}$.

Assumption A5 imposes full rank on the regressors. It also assumes the existence of a regressor with large support only in the binary choice model. As a result, this strictly exogenous regressor X_{1itk} has to be continuous. In contrast, Tamer (2003) requires the existence of a regressor with large support in any of the two equations. Finally, the last set of conditions in A5 ensures that we can identify the coefficients of the predetermined regressors .

Let us now summarize the steps made for the identification argument. Note that the probability that $S_{it}=0, E_{it}=0$ is unaffected by the presence of the fixed effect d_i, just like the stylized example in the previous section. Although the stylized example considers the case where both endogenous variables are binary, the intuition underlying the identification argument remains the same. To see this, first let W_{it}^{bp} and W_{it}^{op} be the value of the linear predictor excluding the contemporaneous endogenous variables in (3.3.3) and (3.3.4), respectively, i.e.,

$$\begin{aligned}W_{it}^{bp} &= \gamma_{11}S_{i,t-1}+\gamma_{12}S_{i,t-2}+\delta_{11}1\left\{E_{i,t-1}=-1\right\}+\delta_{12}1\left\{E_{i,t-1}=1\right\}\\ &\quad+\delta_{21}1\left\{E_{i,t-2}=-1\right\}+\delta_{22}1\left\{E_{i,t-2}=1\right\}+X_{1it}\beta^{bp},\\ W_{it}^{op} &= \gamma_{211}1\left\{E_{i,t-1}=-1\right\}+\gamma_{212}1\left\{E_{i,t-1}=1\right\}+\gamma_{221}1\left\{E_{i,t-2}=-1\right\}\\ &\quad+\gamma_{222}1\left\{E_{i,t-2}=1\right\}+\kappa_1S_{i,t-1}+\kappa_2S_{i,t-2}+X_{2it}\beta^{op}.\end{aligned}$$

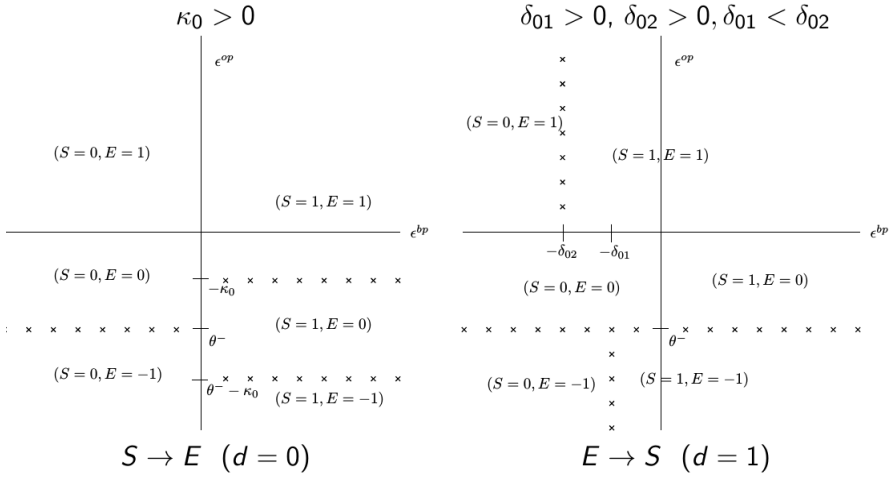
Next, we compute the probability that $S_{it}=0, E_{it}=0$ given the strictly exogenous regressors and predetermined regressors as follows:

$$\begin{aligned}&\Pr\left(S_{it}=0,E_{it}=0|Z_i^t,X_{1i}^T,X_{2i}^T\right)\\ &\overset{\text{A1}}{=}\Pr\left(S_{it}^*\le 0,\theta^-\le E_{it}^*\le 0|Z_i^t,X_{1i}^T,X_{2i}^T\right)\\ &\overset{\text{A1,A2,A4}}{=}\Pr\left(\epsilon_{it}^{bp}\le -W_{it}^{bp},\theta^- -W_{it}^{op}\le\epsilon_{it}^{op}\le -W_{it}^{op}\right)\\ &\overset{\text{A3}}{=}\Pr\left(F_{\epsilon^{bp}}\left(\epsilon_{it}^{bp}\right)\le F_{\epsilon^{bp}}\left(-W_{it}^{bp}\right),F_{\epsilon^{op}}\left(\theta^- -W_{it}^{op}\right)\le F_{\epsilon^{op}}\left(\epsilon_{it}^{op}\right)\le F_{\epsilon^{op}}\left(-W_{it}^{op}\right)\right)\\ &\overset{\text{A3}}{=}C\left(F_{\epsilon^{bp}}\left(-W_{it}^{bp}\right),F_{\epsilon^{op}}\left(-W_{it}^{op}\right);\rho\right)-C\left(F_{\epsilon^{bp}}\left(-W_{it}^{bp}\right),F_{\epsilon^{op}}\left(\theta^- -W_{it}^{op}\right);\rho\right).\end{aligned}\tag{3.3.5}$$

The probability computed in (3.3.5) is always positive since $\theta^-<0$. Figure 3.3.1 confirms the calculation made in (3.3.5). In the figure, think of the origin as the ordered pair of linear predictors $\left(W_{it}^{bp},W_{it}^{op}\right)$. The probability mass over the region

where we have $S_{it} = 0, E_{it} = 0$ is unaffected by the presence of the fixed effect given the assumptions I have imposed. Further observe that (3.3.5) can be thought of as a binary choice model where the outcomes are either the event $S_{it} = 0, E_{it} = 0$ or the event where $(S_{it} = 0, E_{it} = -1)$, $(S_{it} = 0, E_{it} = 1)$, $(S_{it} = 1, E_{it} = -1)$, or $(S_{it} = 1, E_{it} = 1)$, i.e. all the other configurations of (S_{it}, E_{it}).

Figure 3.3.1: An illustration of a case of (3.3.3) and (3.3.4)



The steps below summarize the identification argument. Steps 3 to 7 follow an argument similar to the stylized example in the previous section. The only steps that are new are the first two steps which account for how we will identify the coefficients of the lagged dependent variables and the coefficients of the strictly exogenous variables. The full details of the argument can be found in the Appendix.

Step 1. This step is the nonconstructive part of the identification argument. Take two points $\left(x_1^-, x_1\right) \in \text{supp}\left(X_{1i}^T\right)$, $\left(x_2^-, x_2\right) \in \text{supp}\left(X_{2i}^T\right)$, and $z \in \text{supp}\left(Z_i^{t-1}\right)$. Collect the observed frequencies of $S_{it} = 0$, $E_{it} = 0$ conditional on $Z_i^{t-1} = z$, $Z_{it} = (0,0,0,0)$, $X_{1i}^T = \left(x_1^-, x_1\right)$, $X_{2i}^T = \left(x_2^-, x_2\right)$. Use an identification at infinity argument like the one used by Tamer (2003) to identify β^{bp} and β^{op}. We also identify θ^- in this step.

Step 2. Take another point $\tilde{x}_1 \in \text{supp}\left(X_{1it}\right)$. We use Manski's (1985; 1988) argument to identify (γ_{11}, κ_1) using the observed frequencies of $S_{it} = 0$, $E_{it} = 0$ conditional on $Z_{it} = (1,0,0,0)$, $X_{1it} = \tilde{x}_1, X_{2it} = \tilde{x}_2$ and the observed frequencies of $S_{it} = 0$, $E_{it} = 0$ conditional on $Z_{it} = (0,0,0,0)$, $X_{1it} = x_1, X_{2it} = x_2$. Repeat the argument to identify the coefficients of the other lagged dependent variables using the appropriate Z_{it}, i.e. (γ_{12}, κ_2), $(\delta_{11}, \gamma_{211})$, $(\delta_{12}, \gamma_{212})$, $(\delta_{21}, \gamma_{221})$, and $(\delta_{22}, \gamma_{222})$. For instance, we should set $Z_{it} = (0,0,1,0)$ to identify $(\delta_{12}, \gamma_{212})$.

Step 3. Since (3.3.5) has the form of a fully parametric binary choice model as discussed earlier, the copula dependence parameter ρ can be identified immediately since the values of the parameters in Steps 1 and 2 are identified and can be taken as known.

Step 4. The signs of δ_{01}, δ_{02}, and κ_0 can now be identified.

Step 5. All the information from the previous steps can now be used to determine whether $d_i = 0$ or $d_i = 1$.

Step 6. Since the groupings are now identified, we can recover the values of δ_{01}, δ_{02}, and κ_0.

Steps 1 to 3 of the preceding argument can also be replaced by an alternative argument where we exploit the form of (3.3.5). (3.3.5) is a fully parametric binary choice model and a likelihood function formed from pooling all the cross-sectional and time series information can be used to identify all the parameters mentioned in Steps 1 to 3. This alternative avoids the rather nonconstructive nature of Step 1. Presenting the argument in these two ways is to set the stage for future work on weakening some of the parametric assumptions in Assumption A3. Furthermore, these two arguments may have different implications for estimation and inference. What I have shown is that point identification of all the common parameters is possible for (3.3.3) and (3.3.4).

3.3.3 Estimation and inference

Even though there would be incidental parameter bias when T is fixed and small, Hahn and Moon (2010) show that the incidental parameter bias disappears at a much faster rate. Although the context they have in mind is estimating a game-theoretic model where the fixed effect represents the equilibrium chosen by players (when there are multiple equilibria), the idea that the fixed effect takes only a finite number of values applies to my proposal.

In particular, they show that, under certain regularity conditions, the reduction in support is automatically bias-reducing under an asymptotic scheme where $n, T \to \infty$ with n typically growing as an exponential function of T. Since the asymptotic distribution of the MLE no longer has a noncentrality parameter (as opposed to the usual case where individual-specific effects are allowed to have full support over the real line; see Hahn and Kuersteiner (2011)), inferences can be justified without resorting to bias-reduction procedures.

In this subsection, we impose the Gaussian copula for C and standard normal cumulative distribution functions for the margins $F_{\epsilon^{bp}}$ and $F_{\epsilon^{op}}$ in Assumption A3, just as in HI (2007). As a result, the dependence parameter $\rho \in (-1, 1)$ coincides with the usual correlation coefficient of the bivariate normal distribution. I still impose all

the assumptions required for identification here in this subsection. I use maximum likelihood for estimation and inference. Collect all the common parameters into a vector $\lambda \in \Lambda$ and treat $d_i \in \{0,1\}$ as a parameter to be estimated. As a result, the log-likelihood for an arbitrary i and t is given by

$$l_{it}(\lambda, d_i) = \sum_{j\in\{0,1\}} \sum_{k\in\{-1,0,1\}} \mathbf{1}(S_{it} = j, E_{it} = k) \log \Pr\left(S_{it} = j, E_{it} = k | Z_i^t, X_{1i}^T, X_{2i}^T; \lambda, d_i\right).$$

Aggregating over time for a fixed cross-sectional unit gives us the log-likelihood for the ith cross-sectional unit:

$$l_i(\lambda, d_i) \quad = \quad \sum_{t=1}^{T} l_{it}(\lambda, d_i).$$

Next, I impose the following additional assumptions:

E1 The parameters representing the simultaneous effects δ_{01}, δ_{02}, and κ_0 cannot all be jointly equal to zero.

E2 Let $y_{it} = \left(S_{it}, E_{it}, S_{i,t-1}, E_{i,t-1}, S_{i,t-2}, E_{i,t-2}, X_{1it}, X_{2it}\right)$ be the data for the ith unit and tth time period and $d_{i0} \in \{0,1\}$ be the true value of d_i. For each i, $\{y_{it} : t = 1, 2, \ldots\}$ is strictly stationary. The differences of the joint distribution of $\{y_{i1}, y_{i2}, \ldots\}$ across i is completely characterized by d_{i0}.

E3 Let

$$\varepsilon^* = \inf_i \left[G_{(i)}(\lambda_0, d_{i0}) - \sup_{\{d_i \neq d_{i0}\}} G_{(i)}(\lambda, d_i) \right] > 0,$$

where

$$G_{(i)}(\lambda, d) = \mathbb{E}_{(\theta_0, d_{i0})}[l_i(\lambda, d)].$$

For all $\eta > 0$,

$$\inf_i \left[G_{(i)}(\lambda_0, d_{i0}) - \sup_{\{|\lambda - \lambda_0| > \eta, d\}} G_{(i)}(\lambda, d) \right] > 0.$$

The parameter space Λ is compact. There exists some $M(y_{it})$ such that

$$\sup_{\lambda, d} \left| \frac{\partial^k l_{it}(\lambda, d_i)}{\partial \lambda^k} \right| \leq M(y_{it})$$

for $k = 0, 1$ and $\max_i \mathbb{E}[M(y_{it})]^2 < \infty$.

E4 Let $\varepsilon > 0$, $\eta > 0$ and θ be given. There exists some $h(T)$ strictly increasing in T, such that, for all $\left(d_i, d_i'\right)$ combinations, we have

$$\begin{aligned}
\Pr\left[\frac{1}{T}\left|\sum_{t=1}^{T}\left(l_{it}(\lambda, d_i) - \mathbb{E}\left[l_{it}(\lambda, d_i)\right]\right)\right| > \frac{\eta}{3}\right] &= o\left(\frac{1}{h(T)}\right), \\
\Pr\left[\frac{1}{T}\left|\sum_{t=1}^{T}\left(M(y_{it}) - \mathbb{E}\left[M(y_{it})\right]\right)\right| > \frac{\eta}{3\varepsilon}\right] &= o\left(\frac{1}{h(T)}\right),
\end{aligned}$$

where the probability and the expectation are calculated with respect to the density of $(y_{i1}, \ldots, y_{iT})$ indexed by $\left(\lambda_0, d_i'\right)$.

Note that the individual-specific likelihood function under $d_i = 0$ becomes automatically distinguishable from the one under $d_i = 1$ provided that assumption E1 holds. If all these parameters representing simultaneous effects are jointly equal to zero, there is no way to use time series variation to differentiate between $d_i = 0$ and $d_i = 1$. This can be seen easily from Figure (3.3.1). Assumption E3 also holds because of the previous statements along with the point-identification result in the previous subsection. The compactness of Λ and the boundedness conditions on the likelihood and its score are standard regularity conditions imposed in maximum likelihood estimation. Note that the log-likelihood I consider are continuously differentiable over the compact parameter space. Furthermore, the parametric forms and time homogeneity assumed for the model in (3.3.3) and (3.3.4) ensures that the data for every cross-sectional unit are strictly stationary which satisfies Assumption E2. Finally, Assumption E4 is a technical condition required to identify the correct group assignment. This assumption has been used in the literature on discrete parameter models (refer to Choirat and Seri (2012) and its references). Hahn and Moon (2010) and Choirat and Seri (2012) show that $h(T)$ is typically an exponential function of T.

Since Assumptions E2 to E4 are the same conditions used by Hahn and Moon (2010), adapting their Theorem 1 gives us:

Theorem 3.3.1. *Let*

$$\begin{aligned}
\widehat{d_i}(\lambda) &= arg\max\left\{l_i(\lambda, d_i = 1), l_i(\lambda, d_i = 0)\right\}, \\
\widehat{\lambda} &= arg\max_{\lambda}\sum_{i=1}^{n}\sum_{t=1}^{T} l_{it}\left(\lambda, \widehat{d_i}(\lambda)\right), \\
\widetilde{\lambda} &= arg\max_{\lambda}\sum_{i=1}^{n}\sum_{t=1}^{T} l_{it}(\lambda, d_{i0}).
\end{aligned}$$

Suppose that $\sqrt{nT}\left(\widetilde{\lambda} - \lambda\right) \xrightarrow{d} N(0, \Sigma)$ *for some* Σ. *Under Assumptions E1 to E4, we have* $\sqrt{nT}\left(\widehat{\lambda} - \lambda\right) \xrightarrow{d} N(0, \Sigma)$ *if* $n \to \infty$ *and* $T \to \infty$ *such that* $n = O(h(T))$.

The theorem states that the substitution of a plug-in $\widehat{d_i}(\lambda)$ for d_i is asymptotically negligible. The covariance matrix Σ can either be the inverse of the Hessian or the covariance matrix based on the sandwich formula. To estimate all the common parameters, I use the following iterative approach:[9]

1. Set $s = 0$. Fix starting points for λ at $\lambda^{(0)}$.
2. Let $l_i(\lambda, d_i)$ be the log-likelihood for the ith unit. If $l_i\left(\lambda^{(s)}, 1\right) > l_i\left(\lambda^{(s)}, 0\right)$, then we set $\widehat{d_i^{(s)}} = 1$. Otherwise, $\widehat{d_i^{(s)}} = 0$.
3. Find the maximizer of $\sum_i l_i\left(\lambda^{(s)}, \widehat{d_i^{(s)}}\right)$ and call it $\theta^{(s+1)}$.
4. Set s to be $s + 1$. Repeat Steps 2 and 3 until convergence.

Note that Step 2 corresponds to the profiling out of the fixed effects and that Step 3 corresponds to finding the maximizer of the profile likelihood. The zigzag method proposed is slightly slow in the application because I have to estimate around 54 to 75 parameters. However, Step 2 is likely to be faster than the case where the fixed effect could take on any value.

3.4 Revisiting the results of HI (1995; 2007)

3.4.1 Similarities and differences

Using PSID data[10] from Waves 1 to 20, the authors estimate an econometric model based on the simultaneous determination of (S_{it}, E_{it}) as seen in (3.3.1) and (3.3.2). They estimate both a binary probit and an ordered probit model where both indicators are jointly determined. Both their 1995 and 2007 papers impose the coherency conditions $\kappa_0 = 0$ or $\delta_{01} = \delta_{02} = 0$. Therefore, they will have two sets of results – a set of results based on $\delta_{01} = \delta_{02} = 0$ and another based on $\kappa_0 = 0$. In contrast, I jointly estimate (3.3.3) and (3.3.4) without imposing the coherency conditions.

They incorporate dynamic effects in the model by introducing lagged values of the corresponding indicators. The other regressors are variables that represent characteristics of the household head and the labor market to which the household head was exposed. The list of regressors used in both the 1995 and 2007 papers can be found in Table 3.6.2 found in the Appendix. I exclude the cube of age in the list of regressors because the resulting Hessian was singular.[11] The 1995 paper makes use of exclusion restrictions when estimating (3.3.1) and (3.3.2). On the other hand,

[9] The algorithm is not exactly an application of the EM algorithm. The fixed effects d_i I introduce into the model are not just labels. The values that d_i take have a direct interpretation.

[10] I use data made available by the authors in the Journal of Applied Econometrics data archive.

[11] Freedman and Sekhon (2010) document some of the numerical issues involved in estimating systems of equations with discrete endogenous variables even with very few regressors.

the 2007 paper does not have any exclusion restrictions at all. For instance, the age of the household head may influence both S_{it} and E_{it}, but being a union member may influence E_{it} but not S_{it}. Meango and Mourifie (2013) show that some parameters are only partially identified in two-equation probit models with a dummy endogenous regressor when there are no exclusion restrictions. In contrast, I apply the exclusion restrictions in HI (1995).[12]

In addition to (3.3.1) and (3.3.2), they assume that the error terms $\left(\epsilon_{it}^{bp}, \epsilon_{it}^{op}\right)$ have an AR(1) structure:

$$\epsilon_{it}^{bp} = \eta_i^{bp} + \varsigma_{it}^{bp},\ \varsigma_{it}^{bp} = \rho^{bp}\varsigma_{i,t-1}^{bp} + \xi_{it}^{bp},\ \left|\rho^{bp}\right| < 1, \tag{3.4.1}$$

$$\epsilon_{it}^{op} = \eta_i^{op} + \varsigma_{it}^{op},\ \varsigma_{it}^{op} = \rho^{op}\varsigma_{i,t-1}^{op} + \xi_{it}^{op},\ |\rho^{op}| < 1, \tag{3.4.2}$$

where $\left(\eta_i^{bp}, \eta_i^{op}\right)$ represent time-invariant unobserved heterogeneity and $\left(\xi_{it}^{bp}, \xi_{it}^{op}\right)$ are both i.i.d. Gaussian random variables with mean zero, variance 1 and have nonzero correlation ρ conditional on the strictly exogenous regressors $X_i = (X_{i1}, \ldots, X_{iT})$. They use the Mundlak-Chamberlain device to model $\left(\eta_i^{bp}, \eta_i^{op}\right)$. In particular, they assume that $\eta_i^{bp}|X_i \sim N\left(\bar{X}_{i.}\theta^{bp}, \sigma_{bp}^2\right)$ and $\eta_i^{op}|X_i \sim N\left(\bar{X}_{i.}\theta^{op}, \sigma_{op}^2\right)$. They also model the initial conditions using an analogous assumption. In contrast, I use d_i to represent time-invariant unobserved heterogeneity that may be arbitrarily correlated with X_i. All my results are conditional on the initial observation. I allowed for a similar AR(1) structure but results were not forthcoming as will be discussed below.

Strictly speaking, the model I consider neither encompasses nor generalizes the model that HI (1995; 2007) consider. HI (1995; 2007) use large-n, fixed-T asymptotics to justify their results. In contrast, I use large-n, large-T asymptotics to justify my results. Introducing additive fixed effects and modelling these effects using the Mundlak-Chamberlain device may require substantial changes to the identification argument and the justification of the estimation procedure. Even without resorting to the Mundlak-Chamberlain device, it is not clear how to justify existing bias reduction procedures meant for panel data models with fixed effects that have full support. Despite these concerns, the results I present point to the conclusion that imposing coherency conditions may be inappropriate.

3.4.2 Results

There are a total of 32408 observations on 2410 male household heads observed for an average number of 14 periods. Complete spells accounted for 528 out of the 2410 male household heads. I exclude all household heads with spells of length 1 in

[12] Nevertheless, I estimate the model without the exclusion restrictions because recent work by Han and Vytlacil (2015) point to the possibility of point identification even if there are common exogenous regressors and there are no exclusion restrictions. For now, the identification argument in Section 3.3 require exclusion restrictions. The regressors with large support in the binary choice equation that HI (1995; 2007) use are food needs (`fneed`) and the growth of food needs (`gfneed`).

the analysis. Table 3.6.1 in the Appendix has the distribution of spell length for the household heads in the sample.

I compute all the results in this section using R (R Core Team, 2014). I use the `optimx` package, the accompanying BFGS algorithm, and the programmed tests for the Karush-Kuhn-Tucker optimality conditions (see Nash and Varadhan (2011) and Nash (2014) for more details). All results presented below have passed these tests. I use estimates found in HI (1995; 2007) as possible starting points for the algorithm.[13]

Given the discussion in the previous subsection, I estimate 3 different specifications:

1. Specification A uses the list of regressors in HI (1995) but only allows for own state dependence, i.e. $\delta_{11} = \delta_{12} = \delta_{21} = \delta_{22} = 0$ in (3.3.3) and $\kappa_1 = \kappa_2 = 0$ in (3.3.4). There are 56 parameters to be estimated in this case.

2. Specification B uses the list of regressors in HI (1995) but removes the restrictions just mentioned. As a result, I account for some form of spillover effect. There are 62 parameters to be estimated in this case.

3. Specification C uses the same set of regressors in both (3.3.3) and (3.3.4). The restrictions just mentioned are also removed. There are 73 parameters to be estimated in this case.

Furthermore, I consider two samples – Sample 1 consists of observations from Waves 1 to 14 while Sample 2 consists of observations from Waves 1 to 20. We present the coefficient estimate and their corresponding standard errors immediately below. Coefficients that are statistically significant at the 1% level are in **bold**.

Table 3.4.1: Main Results

	Specification A		Specification B		Specification C	
	Sample 1	Sample 2	Sample 1	Sample 2	Sample 1	Sample 2
δ_{01}	**0.693**	**0.293**	**0.489**	**0.283**	**0.489**	**0.309**
	(0.084)	(0.057)	(0.078)	(0.058)	(0.078)	(0.058)
δ_{02}	0.042	**−0.402**	**−0.721**	**−0.612**	**−0.770**	**−0.601**
	(0.053)	(0.036)	(0.054)	(0.041)	(0.055)	(0.040)
κ_0	**1.326**	**1.128**	**1.429**	**1.282**	**1.413**	**1.281**
	(0.034)	(0.025)	(0.036)	(0.028)	(0.036)	(0.028)
θ^-	**−5.697**	**−5.645**	**−5.714**	**−5.668**	**−5.719**	**−5.680**
	(0.008)	(0.007)	(0.008)	(0.007)	(0.008)	(0.007)
ρ	−0.007	0.001	0.000	−0.001	0.001	−0.000
	(0.005)	(0.004)	(0.004)	(0.004)	(0.005)	(0.004)

[13] These starting points are based on Tables 8 and 9 of HI (1995) and Tables VI and VII of HI (2007). I find that the results are not sensitive to these starting points.

I impose the AR(1) error structure used by the authors. However, the estimated first-order autocorrelation coefficients are extremely small (with sizes around 10^{-6}).[14] In contrast, the authors found first-order autocorrelation coefficients around the range of 0.40 to 0.68 and are significantly different from zero. Therefore, I set aside the AR(1) structure for the rest of the calculations.

The results in Table 3.4.1 indicate that imposing the coherency condition may not be appropriate. Note that δ_{01}, δ_{02}, and κ_0 are significantly different from zero across all specifications and samples (except for the one found in Specification A, Sample 1). Furthermore, the signs are very different from their results. For instance, the immediate effect of being involuntarily unemployed/underemployed on the probability of being liquidity constrained is negative while they estimate it as positive. The absolute values of the coefficients are much larger compared to the results by the authors. For instance, their estimates of κ_0 range from 0.12 to 0.13. There are some differences in the estimates for δ_{01} and δ_{02} in Specification A relative to the other specifications because Specification A does not include lagged spillover effects. There might also be indications of parameter nonconstancy as one moves from Sample 1 to Sample 2. Nevertheless, the results are qualitatively unchanged.

Table 3.4.2: Results on the effects of state dependence for Specification A

	Equation for S_{it}		Equation for E_{it}	
	Sample 1	Sample 2	Sample 1	Sample 2
$S_{i,t-1}$	**1.512**	**1.535**		
	(0.031)	(0.025)		
$S_{i,t-2}$	**0.313**	**0.459**		
	(0.031)	(0.025)		
$\mathbf{1}\{E_{i,t-1}=-1\}$			**−1.868**	**−1.885**
			(0.051)	(0.040)
$\mathbf{1}\{E_{i,t-2}=-1\}$			**−0.811**	**−0.883**
			(0.053)	(0.042)
$\mathbf{1}\{E_{i,t-1}=1\}$			**0.923**	**0.983**
			(0.028)	(0.021)
$\mathbf{1}\{E_{i,t-2}=1\}$			**0.523**	**0.539**
			(0.028)	(0.022)

It is clear from Table 3.4.1 that the coherency conditions imposed by HI (1995; 2007) are unlikely to be true for all household heads. The estimated lower threshold associated with involuntary unemployment/underemployment relative to voluntary is twice the value estimated by the authors. This means that the lower threshold is not as tight as HI (1995; 2007) estimate. The estimated correlation ρ between

[14] Even with different starting values, as noted in the preceding footnote, the estimates for the first-order autocorrelation coefficients are also very near zero.

the error terms $\left(\epsilon_{it}^{bp}, \epsilon_{it}^{op}\right)$ is not significantly different from zero, while the authors estimate this correlation at around 0.34 to 0.43 and are significantly different from zero. It is may be possible that the nonzero correlation of the error terms estimated by HI (1995; 2007) is an artifact of imposing the coherency conditions.

The results in Tables 3.4.2, 3.4.3, and 3.4.4 indicate that there are statistically significant effects of state dependence. In particular, the existence of own state dependence is a major feature common to the three tables. As a result, household heads that were liquidity constrained in the previous periods are more likely to be liquidity constrained now. I find a similar result for employment status. In particular, household heads who were overemployed in previous periods are more likely to be overemployed now.

Tables 3.4.3 and 3.4.4 indicate the possibility of lagged spillover effects, especially for employment status.[15] In particular, household heads that were liquidity constrained in previous periods are more likely to be overemployed now. The results also indicate that past employment status (except for household heads who were involuntarily unemployed/underemployed one period ago) may not be a significant indicator for determining whether household heads are more or less likely to be liquidity constrained now. Since being liquidity constrained or being involuntarily overemployed or unemployed for two periods in the past still has a significant effect on the current state of the household head, even after controlling for contemporaneous effects, the results paint quite a negative picture of the lasting effects of liquidity and employment constraints.

Table 3.4.3: Results on the effects of state dependence for Specification B

	Equation for S_{it}		Equation for E_{it}	
	Sample 1	Sample 2	Sample 1	Sample 2
$S_{i,t-1}$	**1.496**	**1.519**	**−0.385**	**−0.385**
	(0.031)	(0.025)	(0.038)	(0.030)
$S_{i,t-2}$	**0.297**	**0.444**	**−0.128**	**−0.115**
	(0.031)	(0.025)	(0.035)	(0.028)
$\mathbf{1}\{E_{i,t-1}=-1\}$	−0.018	−0.013	**−1.856**	**−1.876**
	(0.059)	(0.046)	(0.051)	(0.040)
$\mathbf{1}\{E_{i,t-2}=-1\}$	−0.097	−0.042	**−0.821**	**−0.882**
	(0.058)	(0.046)	(0.053)	(0.042)
$\mathbf{1}\{E_{i,t-1}=1\}$	**0.124**	**0.123**	**0.941**	**1.008**
	(0.032)	(0.025)	(0.028)	(0.021)
$\mathbf{1}\{E_{i,t-2}=1\}$	0.046	0.059	**0.529**	**0.561**
	(0.032)	(0.025)	(0.029)	(0.022)

[15] Note that these lagged spillover effects do not exactly represent the absence of Granger non-causality since the model includes contemporaneous terms.

Table 3.4.4: Results on the effects of state dependence for Specification C

	Equation for S_{it}		Equation for E_{it}	
	Sample 1	Sample 2	Sample 1	Sample 2
$S_{i,t-1}$	**1.480**	**1.507**	**−0.382**	**−0.359**
	(0.031)	(0.025)	(0.038)	(0.030)
$S_{i,t-2}$	**0.291**	**0.441**	**−0.134**	**−0.112**
	(0.031)	(0.025)	(0.035)	(0.028)
$\mathbf{1}\{E_{i,t-1}=-1\}$	−0.012	−0.010	**−1.856**	**−1.870**
	(0.059)	(0.047)	(0.051)	(0.040)
$\mathbf{1}\{E_{i,t-2}=-1\}$	−0.095	−0.041	**−0.814**	**−0.868**
	(0.058)	(0.046)	(0.053)	(0.042)
$\mathbf{1}\{E_{i,t-1}=1\}$	**0.109**	**0.101**	**0.939**	**1.009**
	(0.033)	(0.026)	(0.028)	(0.021)
$\mathbf{1}\{E_{i,t-2}=1\}$	0.039	0.051	**0.526**	**0.572**
	(0.033)	(0.025)	(0.029)	(0.022)

Note that the estimates in the preceding tables are not directly interpretable. Marginal effects would have to be computed and I leave this to future work. I conjecture that no bias correction would be required when estimating marginal effects unlike the case where the fixed effects have full support (for example, see Bester and Hansen (2009a)). Furthermore, the estimators for these marginal effects may be much slower when the fixed effects have full support, as documented by Fernandez-Val and Weidner (2013). It is unclear whether this will be the case when the fixed effects have finite support.

An alternative to estimating marginal effects is to estimate the ratio of the coefficient estimates. The ratio of the coefficient estimates is usually the ratio of marginal effects. Stewart (2004) show in the context of an ordered probit model that the ratios of the coefficient estimates can be interpreted as slopes of indifference curves. If we apply the idea to my context, the slope of this indifference curve represents the required tradeoff in one regressor so that a change in a different regressor will not alter the state of the household head. Unfortunately, these ratios cannot be obtained from the ratios of marginal effects because the probability in (3.3.5) is a joint probability and involves a difference of two probabilities.

The estimated fixed effects $\widehat{d_i}$ can also be obtained and be used to describe which of the household heads have a particular direction of causality. I calculate the estimated distribution of the fixed effects in Table 3.4.5. Compared to Hajivassiliou and Ioannides (1995; 2007), either all 2410 males have only a single direction of causality (say from S_{it} to E_{it}) or all of them have the other direction. Since we allow for the direction of causality to vary across males, we are able to count how many of these household heads have a pattern where S_{it} affects E_{it} and vice-versa. I find

that around half of the 2410 males have a pattern where S_{it} affects E_{it} across specifications and across different samples. I also find that some of these males change patterns from Sample 1 to Sample 2. In particular, around 12% of the males from Sample 1 change patterns once we observed them for more time periods.

Table 3.4.5: Estimated distribution of the fixed effects $\widehat{d_i}$

	Specification A		Specification B		Specification C	
	Sample 1	Sample 2	Sample 1	Sample 2	Sample 1	Sample 2
$S_{it} \to E_{it}$	1101	1214	1100	1221	1103	1209
$E_{it} \to S_{it}$	1001	1186	1002	1179	999	1191
Total	2102	2400	2102	2400	2102	2400

Finally, the tables in the appendix contain results for the coefficients of the strictly exogenous regressors across different specifications and samples. Apart from a few coefficients changing signs across specifications and samples (in particular, household age and some of the dummies representing residence, ethnicity, and religion), the results are quite similar to one another and are consistent with expectations. However, most of the positive coefficients in Sample 1 are larger than those in Sample 2. Similarly, most of the negative coefficients in Sample 1 are larger in absolute value than those in Sample 2.

3.5 Concluding remarks

In this chapter, I have developed a route toward identification, estimation, and inference in dynamic simultaneous equations models with discrete outcomes when panel data is available. These models are subject to the incidental parameter problem when individual-specific fixed effects are included and are also subject to incoherence and incompleteness. I introduce a specific type of individual-specific fixed effect so that the coherency condition need not be imposed across all observations. This proposal allows us to avoid imposing sign restrictions or to avoid restricting error supports.

Specifically, I use a subset of the observables unaffected by the individual-specific fixed effect to identify the common parameters of the model. I then use time series variation to identify the individual-specific fixed effect. This fixed effect represents the direction of causality from one endogenous variable to another. Knowing the direction allows us to identify the coefficients of the endogenous variables. Consistent estimation and correct inference without any need for bias reduction follows from the large-n, large-T asymptotic theory. I revisit the empirical application of Hajivassiliou and Ioannides (1995; 2007) and find strikingly different results with respect to

the contemporaneous interaction and dynamic structure of employment status and liquidity constraints.

Future work may consider the computation of certain types of marginal effects defined in Lewbel, Dong, and Yang (2012). Future work in this area may also allow the specification of individual-specific effects to be time-varying as well, just as Bonhomme and Manresa (2015) do for groupings in the linear model. There seems to be some slight evidence in the empirical application that cast some doubt on the assumption that the direction of causality is time-invariant. However, it may well be the case that we have restricted time-invariant unobserved heterogeneity too much. Introducing another fixed effect in the linear predictor may be fruitful but is beyond the scope of this chapter. Although the approach would seem fruitful, bias-reduction procedures have to be adapted for the case I considered. A natural alternative would be to use each cross-section to set-identify the parameters of the model (as seen in Section 2) and find methods to combine these set estimates across different time periods.

3.6 Appendix

Details of identification argument in Section 3.3.2

Let us now examine the details behind each step of the identification argument. In Step 1, we need to calculate the probability that $S_{it} = 0$ and $E_{it} = 0$ conditional on $Z_{it} = (0,0,0,0), X_{1i}^T = \left(x_1^-, x_1\right), X_{2i}^T = \left(x_2^-, x_2\right)$:

$$
\begin{aligned}
&\Pr\left(S_{it}^* \leq 0, \theta^- \leq E_{it}^* \leq 0 | Z_i^{t-1} = z, Z_{it} = (0,0,0,0), X_{1i}^T = \left(x_1^-, x_1\right), X_{2i}^T = \left(x_2^-, x_2\right)\right) \\
&\overset{\text{A2}}{=} \Pr\left(S_{it}^* \leq 0, \theta^- \leq E_{it}^* \leq 0 | Z_{it} = (0,0,0,0), X_{1it} = x_1, X_{2it} = x_2\right) \\
&\overset{\text{A1}}{=} \Pr\left(\epsilon_{it}^{bp} \leq -x_1\beta^{bp}, \theta^- - x_2\beta^{op} \leq \epsilon_{it}^{op} \leq -x_2\beta^{op}\right)
\end{aligned}
$$

Let $\left(\tilde{\beta}^{bp}, \tilde{\beta}^{op}\right)$ be such that $\left(\beta^{bp}, \beta^{op}\right) \neq \left(\tilde{\beta}^{bp}, \tilde{\beta}^{op}\right)$. Without loss of generality, let x_{1k} be the kth regressor in x_1 and $\beta_k^{bp}, \tilde{\beta}_k^{bp} > 0$ be the associated coefficient of this regressor. As $x_{1k} \to -\infty$ given the other regressors in x_1, we have $-x_{1k}\beta_k^{bp}, -x_{1k}\tilde{\beta}_k^{bp} \to \infty$. Since x_2 has full rank by A5, we have x_2 such that $x_2\beta^{op} \neq x_2\tilde{\beta}^{op}$. We now have

$$
\begin{aligned}
&\Pr\left(\epsilon_{it}^{bp} \leq -x_1\beta^{bp}, \theta^- - x_2\beta^{op} \leq \epsilon_{it}^{op} \leq -x_2\beta^{op}\right) \\
&\approx \Pr\left(\theta^- - x_2\beta^{op} \leq \epsilon_{it}^{op} \leq -x_2\beta^{op}\right) \\
&\neq \Pr\left(\theta^- - x_2\tilde{\beta}^{op} \leq \epsilon_{it}^{op} \leq -x_2\tilde{\beta}^{op}\right) \\
&\approx \Pr\left(\epsilon_{it}^{bp} \leq -x_1\beta^{bp}, \theta^- - x_2\tilde{\beta}^{op} \leq \epsilon_{it}^{op} \leq -x_2\tilde{\beta}^{op}\right).
\end{aligned}
$$

As a result, β^{op} is identified. Since x_1 has full rank by A5, we have x_1 such that $x_1\beta^{bp} \neq x_1\tilde{\beta}^{bp}$. Following the same argument as before, we have

$$\begin{aligned}&\Pr\left(\epsilon_{it}^{bp} \leq -x_1\beta^{bp}, \theta^- - x_2\beta^{op} \leq \epsilon_{it}^{op} \leq -x_2\beta^{op}\right)\\ &\neq \Pr\left(\epsilon_{it}^{bp} \leq -x_1\tilde{\beta}^{bp}, \theta^- - x_2\beta^{op} \leq \epsilon_{it}^{op} \leq -x_2\beta^{op}\right).\end{aligned}$$

As a result, β^{bp} is identified. For the case where $\tilde{\beta}_k^{bp} < 0$, we have $-x_{1k}\beta_k^{bp} \to \infty$ but $-x_{1k}\tilde{\beta}_k^{bp} \to -\infty$. Following the same argument as before, we can identify both β^{bp} and β^{op}. Note that the constant terms in β^{bp} and β^{op} are also identified.

Now, we identify θ^-. Without loss of generality, let $\tilde{\theta}^- < \theta^- < 0$. Since β^{bp} and β^{op} are both identified, we take them as fixed in this step. Recall that we have

$$\begin{aligned}&\Pr\left(\epsilon_{it}^{bp} \leq -x_1\beta^{bp}, \theta^- - x_2\beta^{op} \leq \epsilon_{it}^{op} \leq -x_2\beta^{op}\right)\\ &\neq \Pr\left(\epsilon_{it}^{bp} \leq -x_1\beta^{bp}, \tilde{\theta}^- - x_2\beta^{op} \leq \epsilon_{it}^{op} \leq -x_2\beta^{op}\right).\end{aligned}$$

As a result, θ^- is identified.

Step 2 uses Manski's (1985; 1988) identification argument to identify the coefficients of the lagged dependent variables. To illustrate, consider the following probabilities:

$$\begin{aligned}&\Pr\left(S_{it}=0, E_{it}=0 | Z_i^{t-1}=z, Z_{it}=(0,0,0,0), X_{1i}^T=\left(x_1^-, x_1\right), X_{2i}^T=\left(x_2^-, x_2\right)\right)\\ &= \Pr\left(\epsilon_{it}^{bp} \leq -x_1\beta^{bp}, \theta^- - x_2\beta^{op} \leq \epsilon_{it}^{op} \leq -x_2\beta^{op}\right),\end{aligned} \tag{3.6.1}$$

and

$$\begin{aligned}&\Pr\left(S_{it}=0, E_{it}=0 | Z_i^{t-1}=z, Z_{it}=(1,0,0,0), X_{1i}^T=\left(x_1^-, \widetilde{x_1}\right), X_{2i}^T=\left(x_2^-, \widetilde{x_2}\right)\right)\\ &= \Pr\left(\epsilon_{it}^{bp} \leq -\tilde{x}_1\beta^{bp} - \gamma_{11}, \theta^- - \tilde{x}_2\beta^{op} - \kappa_1 \leq \epsilon_{it}^{op} \leq -\tilde{x}_2\beta^{op} - \kappa_1\right).\end{aligned} \tag{3.6.2}$$

The expressions (3.6.1) and (3.6.2) will only be equal if and only if

$$\begin{aligned}-x_1\beta^{bp} &= -\widetilde{x_1}\beta^{bp} - \gamma_{11},\\ -x_2\beta^{op} &= -\widetilde{x_2}\beta^{op} - \kappa_1.\end{aligned}$$

Therefore, both γ_{11} and κ_1 are identified because $\gamma_{11} = (x_1 - \widetilde{x_1})\beta^{bp}$ and $\kappa_1 = (x_2 - \widetilde{x_2})\beta^{op}$ under the support condition in assumption A5. Similar arguments can be used to identify the coefficients of the other lagged dependent variables.

Step 3 follows from recognizing that we have a fully parametric binary choice model in (3.3.5) with only one copula dependence parameter left to identify.

In Step 4, we identify the signs of δ_{01}, δ_{02}, and κ_0. Note that we have

$$\begin{aligned}
&\Pr\left(S_{it}=0, E_{it}=-1 | Z_i^{t-1}=z,\, Z_{it}=(0,0,0,0),\, X_{1i}^T=\left(x_1^-, x_1\right), X_{2i}^T=\left(x_2^-, x_2\right)\right) \\
=&\Pr\left(\epsilon_{it}^{bp} \leq -x_1\beta^{bp} - \delta_{01}, \epsilon_{it}^{op} \leq \theta^- - x_2\beta^{op}\right)\Pr(d_i=1) \\
&+\Pr\left(\epsilon_{it}^{bp} \leq -x_1\beta^{bp}, \epsilon_{it}^{op} \leq \theta^- - x_2\beta^{op}\right)\Pr(d_i=0).
\end{aligned}$$

Showing that this conditional probability is greater than

$$\Pr\left(\epsilon_{it}^{bp} \leq -x_1\beta^{bp}, \epsilon_{it}^{op} \leq \theta^- - x_2\beta^{op}\right)$$

allows us to conclude that $\delta_{01} < 0$. The other cases follow analogously. Note that to avoid cumbersome notation, I omit the conditioning set in $\Pr(d_i=1)$ and $\Pr(d_i=0)$.

For Step 5, there are eight cases to consider. The resulting group assignment rules follows the same intuition as Table 3.2.1 and by sketching figures like Figure (3.3.1). One of the cases is that once we know that $\delta_{01} > 0$, $\delta_{02} > 0$, and $\kappa_0 > 0$, we must assign $d_i = 0$ if and only if

$$\begin{aligned}
&\Pr\left(E_{it}=0 | Z_i^{t-1}=z,\, Z_{it}=(0,0,0,0),\, X_{1i}^T=\left(x_1^-, x_1\right), X_{2i}^T=\left(x_2^-, x_2\right)\right) \\
>\;& \Pr\left(\epsilon_{it}^{op} \geq -x_2\beta^{op}\right)
\end{aligned}$$

or assign $d_i = 1$ if and only if

$$\begin{aligned}
&\Pr\left(S_{it}=0 | Z_i^{t-1}=z,\, Z_{it}=(0,0,0,0),\, X_{1i}^T=\left(x_1^-, x_1\right), X_{2i}^T=\left(x_2^-, x_2\right)\right) \\
>\;& \Pr\left(\epsilon_{it}^{bp} \leq -x_1\beta^{bp}\right)
\end{aligned}$$

Of course, these assignment rules can be altered by changing the conditioning sets. The other cases follow similarly.

In Step 6, we can now point-identify δ_{01}, δ_{02}, and κ_0. One route is to look at the conditional probability of $(S_{it}=1, E_{it}=1)$ given $Z_i^{t-1}=z$, $Z_{it}=(0,0,0,0)$, $X_{1i}^T=\left(x_1^-, x_1\right)$, $X_{2i}^T=\left(x_2^-, x_2\right)$, $d_i=0$. This conditional probability is now a function of κ_0 and can be used to point-identify κ_0. The conditional probability of $(S_{it}=0, E_{it}=-1)$ given $Z_i^{t-1}=z$, $Z_{it}=(0,0,0,0)$, $X_{1i}^T=\left(x_1^-, x_1\right)$, $X_{2i}^T=\left(x_2^-, x_2\right)$, $d_i=1$ can be used to point-identify δ_{01}. Finally, the conditional probability of $(S_{it}=0, E_{it}=1)$ given $Z_i^{t-1}=z$, $Z_{it}=(0,0,0,0)$, $X_{1i}^T=\left(x_1^-, x_1\right)$, $X_{2i}^T=\left(x_2^-, x_2\right)$, $d_i=1$ can be used to point-identify δ_{02}. Alternative routes include changing the vector Z_{it} or using other regions found in Figure 3.3.1.

Empirical Results

There are five tables in this Appendix. Table 3.6.1 contain the distribution of spell lengths in the data. Table 3.6.2 contains a description of the regressors used in the empirical application. Tables 3.6.3, 3.6.4, and 3.6.5 contain the estimation results for Specifications A, B, and C, respectively, across Samples 1 and 2.

Table 3.6.1: Length of spells observed in the data

Number of periods	1	2	3	4	5	6	7	8	9	10
Number of males	10	13	23	30	130	131	93	132	121	116
Number of periods	11	12	13	14	15	16	17	18	19	20
Number of males	103	121	138	124	124	118	125	127	103	528

Table 3.6.2: List of variables in Hajivassiliou and Ioannides (1995)

Model for	Regressors in X_{it}
S_{it}	educational category of head (`edycat`)
	dummies for 1976-79 and 1980-83 periods (`era7679`, `era8083`)
	food needs, growth of food needs (`fneed`, `gfneed`)
	age, age squared, age cubed (`hage`)
	live in north/central, south, west, other regions (`liveinnc`, `liveinso`, `liveinwe`, `liveinot`)
	married, race is black or other (`msm`, `raceb`, `raceo`)
	religion is Christian, Jewish, or Protestant (`religceo`, `religjsh`, `religpro`)
	real rate of interest (`rri`)
E_{it}	county unemployment rate (`cunemp`)
	head disabled, educational category of head (`disab`, `edycat`)
	dummies for 1976-79 and 1980-83 periods (`era7679`, `era8083`)
	food needs, growth of food needs (`fneed`, `gfneed`)
	age, age squared, age cubed (`hage`)
	tenure, tenure squared (`htenure`)
	unemployment insurance received by head (`hunemins`)
	imputed wage (`impwage`)[a]
	tightness of labor market conditions (labmkt)
	live in north/central, south, west, other regions (`liveinnc`, `liveinso`, `liveinwe`, `liveinot`)
	married, number of children between 0-5 (`msm`, `numch05`)
	occupational unemployment rate (`occunemp`)
	race is black or other (`raceb`, `raceo`)
	religion is Christian, Jewish, or Protestant (`religceo`, `religjsh`, `religpro`)
	real rate of interest (`rri`)
	head is union member (`unionmem`)

[a] The authors include a variable representing some measure of the imputed wage (`impwage`). Unfortunately, the JAE data archive did not include this variable.

Table 3.6.3: Results for Specification A

Sample 1						Sample 2					
Equation for S_{it}			Equation for E_{it}			Equation for S_{it}			Equation for E_{it}		
Variable	Coef	SE	Variable	Coef	SE	Variable	Coef	SE	Variable	Coef	SE
Regressors common to both equations											
Intercept	1.270	0.151	Intercept	-2.366	0.166	Intercept	0.939	0.122	Intercept	-1.572	0.129
era7679	0.141	0.033	era7679	0.002	0.035	era7679	0.134	0.030	era7679	0.035	0.030
era8083	-0.311	0.054	era8083	-0.460	0.054	era8083	-0.091	0.031	era8083	0.099	0.030
edycat	-0.043	0.008	edycat	-0.027	0.008	edycat	-0.054	0.006	edycat	-0.071	0.006
hage	-11.077	0.775	hage	1.925	0.721	hage	-9.399	0.616	hage	-1.118	0.552
hagesq	9.207	0.944	hagesq	-4.754	0.872	hagesq	7.221	0.746	hagesq	-0.471	0.660
liveinnc	-0.071	0.037	liveinnc	-0.086	0.036	liveinnc	-0.068	0.030	liveinnc	-0.108	0.028
liveinot	0.502	0.151	liveinot	0.228	0.161	liveinot	0.365	0.109	liveinot	0.123	0.111
liveinso	0.073	0.038	liveinso	0.123	0.039	liveinso	0.064	0.030	liveinso	-0.046	0.029
liveinwe	0.018	0.043	liveinwe	-0.436	0.042	liveinwe	0.050	0.033	liveinwe	-0.400	0.033
mss	0.577	0.043	mss	-0.209	0.042	mss	0.547	0.034	mss	-0.214	0.031
raceb	0.380	0.056	raceb	0.208	0.051	raceb	0.390	0.045	raceb	0.275	0.040
raceo	-0.372	0.054	raceo	0.140	0.049	raceo	-0.347	0.049	raceo	0.222	0.043
religceo	0.092	0.044	religceo	0.273	0.042	religceo	0.125	0.034	religceo	0.235	0.032
religjsh	0.201	0.081	religjsh	0.198	0.081	religjsh	0.197	0.069	religjsh	0.165	0.067
religpro	0.060	0.033	religpro	0.006	0.033	religpro	0.157	0.025	religpro	0.070	0.025
rri	10.017	1.115	rri	13.997	1.141	rri	6.891	0.616	rri	4.193	0.632
Regressors that are included in one of the equations but excluded in the other											
fneed	0.254	0.374	cunemp	0.874	0.596	fneed	0.306	0.317	cunemp	0.679	0.391
gfneed	-0.400	0.047	disab	0.270	0.044	gfneed	-0.508	0.039	disab	0.136	0.034
			htenure	-3.542	0.412				htenure	-3.938	0.325
			htenursq	8.650	1.519				htenursq	8.051	1.183
			hunemins	0.557	0.032				hunemins	0.354	0.016
			labmkt	0.032	0.014				labmkt	0.055	0.011
			numch05	0.007	0.022				numch05	0.018	0.016
			occunemp	5.865	0.518				occunemp	3.647	0.319
			unionmem	0.215	0.029				unionmem	0.213	0.023

Table 3.6.4: Results for Specification B

Sample 1						Sample 2					
Equation for S_{it}			Equation for E_{it}			Equation for S_{it}			Equation for E_{it}		
Variable	Coef	SE	Variable	Coef	SE	Variable	Coef	SE	Variable	Coef	SE
Regressors common to both equations											
Intercept	1.347	0.153	Intercept	-2.106	0.169	Intercept	0.896	0.123	Intercept	-1.391	0.131
era7679	0.141	0.033	era7679	0.011	0.034	era7679	0.124	0.030	era7679	0.038	0.030
era8083	-0.320	0.054	era8083	-0.454	0.054	era8083	-0.081	0.031	era8083	0.088	0.030
edycat	-0.047	0.008	edycat	-0.034	0.008	edycat	-0.050	0.006	edycat	-0.074	0.006
hage	-11.074	0.781	hage	1.150	0.725	hage	-9.344	0.618	hage	-1.348	0.556
hagesq	9.133	0.950	hagesq	-4.131	0.873	hagesq	7.227	0.748	hagesq	-0.502	0.663
liveinnc	-0.068	0.037	liveinnc	-0.048	0.036	liveinnc	-0.065	0.030	liveinnc	-0.111	0.028
liveinot	0.505	0.153	liveinot	0.327	0.162	liveinot	0.388	0.109	liveinot	0.138	0.112
liveinso	0.074	0.038	liveinso	0.128	0.039	liveinso	0.069	0.030	liveinso	-0.036	0.029
liveinwe	0.012	0.043	liveinwe	-0.401	0.042	liveinwe	0.055	0.034	liveinwe	-0.390	0.032
mss	0.569	0.044	mss	-0.146	0.042	mss	0.542	0.034	mss	-0.185	0.031
raceb	0.405	0.057	raceb	0.213	0.052	raceb	0.381	0.045	raceb	0.280	0.040
raceo	-0.399	0.055	raceo	0.154	0.050	raceo	-0.355	0.049	raceo	0.219	0.043
religceo	0.097	0.044	religceo	0.247	0.042	religceo	0.131	0.034	religceo	0.235	0.032
religjsh	0.209	0.081	religjsh	0.174	0.082	religjsh	0.199	0.070	religjsh	0.137	0.067
religpro	0.053	0.034	religpro	-0.027	0.033	religpro	0.160	0.025	religpro	0.061	0.025
rri	10.481	1.125	rri	13.672	1.137	rri	6.476	0.622	rri	4.247	0.632
Regressors that are included in one of the equations but excluded in the other											
fneed	0.101	0.379	cunemp	1.233	0.597	fneed	0.189	0.319	cunemp	0.338	0.391
gfneed	-0.395	0.048	disab	0.280	0.044	gfneed	-0.512	0.039	disab	0.135	0.034
			htenure	-3.695	0.413				htenure	-4.087	0.329
			htenursq	9.044	1.528				htenursq	8.060	1.208
			hunemins	0.537	0.032				hunemins	0.356	0.016
			labmkt	0.037	0.014				labmkt	0.060	0.010
			numch05	0.013	0.022				numch05	0.019	0.016
			occunemp	5.889	0.519				occunemp	3.717	0.318
			unionmem	0.184	0.029				unionmem	0.204	0.023

Table 3.6.5: Results for Specification C

	Sample 1				Sample 2			
	Equation for S_{it}		Equation for E_{it}		Equation for S_{it}		Equation for E_{it}	
Variable	Coef	SE	Coef	SE	Coef	SE	Coef	SE
Intercept	1.124	0.170	-1.869	0.171	0.759	0.136	-1.403	0.133
era7679	0.083	0.035	0.007	0.034	0.062	0.032	-0.038	0.031
era8083	-0.369	0.055	-0.426	0.054	-0.136	0.042	-0.146	0.040
era8487					-0.037	0.036	-0.282	0.034
edycat	-0.032	0.008	-0.028	0.008	-0.037	0.007	-0.050	0.006
hage	-10.555	0.801	-1.679	0.793	-8.819	0.642	-2.965	0.613
hagesq	8.625	0.966	-0.819	0.950	6.766	0.767	1.550	0.725
liveinnc	-0.093	0.039	-0.046	0.036	-0.074	0.030	-0.099	0.028
liveinot	0.492	0.152	0.379	0.163	0.386	0.109	0.197	0.112
liveinso	0.046	0.040	0.136	0.039	0.058	0.031	-0.014	0.029
liveinwe	-0.010	0.043	-0.393	0.042	0.043	0.034	-0.372	0.033
mss	0.570	0.045	-0.030	0.046	0.541	0.035	-0.060	0.035
raceb	0.408	0.057	0.194	0.052	0.393	0.045	0.237	0.040
raceo	-0.376	0.056	0.159	0.050	-0.344	0.050	0.166	0.043
religceo	0.087	0.044	0.218	0.042	0.123	0.034	0.203	0.032
religjsh	0.197	0.082	0.144	0.082	0.184	0.070	0.093	0.068
religpro	0.038	0.034	-0.036	0.034	0.150	0.026	0.033	0.025
rri	8.812	1.168	13.064	1.139	5.351	0.780	7.374	0.733
fneed	0.071	0.383	2.637	0.344	0.308	0.325	2.657	0.291
gfneed	-0.390	0.048	-0.088	0.050	-0.511	0.039	-0.121	0.039
cunemp	-1.819	0.625	1.295	0.597	-1.180	0.427	0.369	0.391
disab	0.069	0.048	0.277	0.044	0.022	0.038	0.136	0.034
htenure	-1.168	0.429	-3.921	0.414	-1.514	0.348	-4.804	0.326
htenursq	2.596	1.624	9.529	1.528	2.586	1.322	10.308	1.180
hunemins	0.085	0.034	0.542	0.032	0.044	0.017	0.365	0.016
labmkt	0.031	0.015	0.038	0.014	0.017	0.011	0.048	0.011
numch05	-0.012	0.022	0.000	0.021	-0.021	0.016	0.010	0.016
occunemp	3.381	0.537	5.730	0.519	2.029	0.343	4.045	0.321
unionmem	-0.015	0.032	0.183	0.029	-0.017	0.026	0.198	0.023

Chapter 4

Estimation and inference in dynamic nonlinear fixed effects panel data models by projection

4.1 Introduction

Neyman and Scott (1948) show that the method of maximum likelihood may fail to produce consistent and asymptotically efficient estimators when there are incidental parameters. Lancaster (2000) documents some of the developments after the publication of their paper. Roughly, these developments can be classified into two classes of solutions to the incidental parameter problem: solutions that exploit the structure of the model and solutions that involve orthogonal reparametrization. The latter has been explored more fully in Lancaster (2002) and Woutersen (2003; 2011). Most of the solutions that have been documented are called fixed-T solutions. If one would choose to use an asymptotic scheme where the number of cross-sectional units n grow large, leaving the number of time periods T fixed, then one has to use choose procedures that ensure that the estimating function is both functionally and stochastically independent of the incidental parameters.

Since incidental parameters in panel data models are represented as time-invariant parameters that appear in only a finite number of probability distributions, estimating these parameters induces finite sample bias in the time series dimension. This phenomenon allows us to reconsider the choice of asymptotic scheme. Research by Waterman (1993), Li, Lindsay, and Waterman (2003), and Hahn and Newey (2004)

has paved the way for these large-T bias corrections. Arellano and Hahn (2007) primarily survey these developments for static panel data models with strictly exogenous regressors. They also document the three related ways of constructing these corrections – correcting the objective function, the moment equation, or the estimator itself. Although one can find consistent estimators of the common parameters, their asymptotic distributions are incorrectly centered. Under this asymptotic scheme, the nonzero center can be estimated when both the number of cross-sectional units and time periods grow at a particular rate (say $n/T \to c \in (0, \infty)$). As a result, one can construct an estimator with a correctly centered asymptotic distribution.

In this paper, I adjust the score or some suitably chosen moment function for the common parameter so that a consistent root of the adjusted score has a correctly centered asymptotic distribution. Furthermore, there are cases for which the adjustment produces a fixed-T consistent estimator. The score or some moment function is the most natural object to adjust because they are the starting points for proofs of consistency and asymptotic normality under regularity conditions. Depending on how one sees the multiple root problem, an issue with score-based adjustments is root selection.[1] In addition, when the common parameter is vector-valued, reconstructing a corrected objective function from the adjusted score or adjusted moment function may no longer be possible.[2] Despite these issues, I discuss some of the advantages of using this score-based adjustment.

First, the computation of the large-T bias-corrected estimator typically requires the user to select an integer bandwidth whenever a model with some dynamics is being considered. This is true even for the case of a model with lagged dependent variables and strictly exogenous regressors (see for example, Bester and Hansen (2009a) and Hahn and Kuersteiner (2011)) or a static binary choice model with predetermined regressors (see Fernandez-Val (2009)). Arellano and Hahn (2006) modify the objective function which also requires bandwidth selection. The proposed adjustment would not require bandwidth selection just like other score-based corrections (see for example, Woutersen (2003), Carro (2007), and Dhaene and Jochmans (2015b)). One can consider this as an improvement because score-based adjustments exploit the model structure fully in order to create the correction. As a result, finite sample performance may improve, especially in short panels.[3]

[1] Small, Wang, and Yang (2000) survey some existing methods for dealing with multiple roots.

[2] One can only recover a quasi-likelihood function from a quasi-score function if the quasi-score is a conservative vector field (see Sections 6.4 and 6.5 of McLeish and Small (1994) for more details). The integration required to go from quasi-score to quasi-likelihood may be path dependent leading to nonuniqueness. The main requirement for a conservative vector field is the symmetry of the derivative matrix of the score. Examples where the latter is not satisfied is in the modelling of covariance matrices in longitudinal data (see Firth and Harris (1991)). It turns out that the symmetry is also required In the context of deriving an information-orthogonal reparameterization. See Section 3.2 of Lancaster (2002).

[3] The score-based adjustment to be discussed later requires the calculation of expectations based on the assumed parametric model. One can avoid the calculation of these expectations by using sample

Second, the approach can accommodate multiple *individual-specific* fixed effects. Multiple fixed effects may arise when the thresholds in ordered choice models are individual-specific in addition to accounting for individual-specific effects in the linear predictor (see Bester and Hansen (2009a) and Carro and Traferri (2012)). They also arise when a model explicitly allows for a vector of individual-specific effects. For example, Hausman and Pinkovskiy (2013) approximate a dynamic nonlinear model with general predetermined regressors and a scalar individual-specific effect by a Taylor series expansion around an estimator for the scalar individual-specific effect. They show that the transformed model is an affine function of a vector of fixed effects. The elements of this vector are the positive integer powers of the deviation of the scalar individual-specific effect from its estimator. Multiple fixed effects also arise when a model contains time dummies. I do not consider this case but Fernandez-Val and Weidner (2013) have recently proposed and justified the large-T bias corrections in this context.

Third, the approach can accommodate predetermined regressors aside from lagged dependent variables. The approach considered in this paper can accommodate predetermined regressors provided that the feedback process is specified to some degree. The feedback process can either be structural or be some flexible reduced form in the spirit of the Mundlak-Chamberlain device. The specification of the feedback process is partly a matter of interpretation. The Mundlak-Chamberlain device is a correlated random effects approach where the individual-specific fixed effect is usually expressed as a linear projection of the individual-specific fixed effect on the observable characteristics of the cross-sectional unit and a residual (see Mundlak (1978) and Chamberlain (1984)). As proposed by Wooldridge (2000) and applied by Moral-Benito (2013; 2014), the Mundlak-Chamberlain device can be used to flexibly specify the feedback process. In contrast to Wooldridge (2000), we do not specify reduced forms for the individual-specific fixed effect. Corrections that allow for general predetermined regressors without resorting to the device include work by Woutersen (2003), Fernandez-Val (2009), and Fernandez-Val and Weidner (2013).

I give details on the projection approach and its properties in Section 4.2. I also discuss some examples where analytical results are available. In Section 4.3, I present the results of two small-scale Monte Carlo simulations where I compare the projected score to the corrections proposed by Woutersen (2003), Carro (2007), Fernandez-Val (2009), and Hahn and Kuersteiner (2011). Other corrections that were not implemented include the corrections based on (i) modifying the likelihood (see Arellano and Hahn (2006) and Bartolucci et al. (2014)) or integrating the likelihood (see Arellano and Bonhomme (2009) and De Bin, Sartori, and Severini (2015)) and (ii) simulation (see Kim and Sun (2009) and Dhaene and Jochmans (2015b)). I

equivalents of these expectations. In this sense, one is able to "loosen" the use of the model structure as T becomes large.

conclude in Section 4.4 and include a technical appendix for some of the calculations and proofs.

4.2 The projection approach

4.2.1 Concept

Suppose we draw a random sample $\{y_i = (y_{i1}, \ldots, y_{iT}) : i = 1, \ldots, n\}$ from some known density $f(y_i; \theta_0, \alpha_{i0})$, where θ_0 is the true value for the common parameter and α_{i0} is the true value for the incidental parameter. Note that these parameters may be vector-valued but I assume that these are scalars for the purposes of illustration. Denote $\mathbb{E}[\cdot; \theta_0, \alpha_{i0}]$ to be the expectation at the true values of the parameters. Denote $\partial^k_{\alpha_i}$ to be the kth order partial derivative with respect to α_i.

To construct consistent estimators for θ_0 in the presence of unknown α_{i0} that have to be estimated, we need a concept that will quantify reduced sensitivity to perturbations of the true value of the incidental parameter, denoted by α_i', holding θ_0 fixed. This means that aside from searching for unbiased estimating functions $g(\theta, \alpha_i; y_i)$ that have zero expectation at the true value, i.e.

$$\mathbb{E}[g(\theta_0, \alpha_{i0}; y_i); \theta_0, \alpha_{i0}] = 0,$$

we have to further narrow the search to classes of estimating functions that satisfy either of the following conditions:

1. Global ancillarity, where the expectation of the estimating function does not depend on the perturbed value α_i':

$$\mathbb{E}\left[g(\theta_0, \alpha_{i0}; y_i); \theta_0, \alpha_i'\right] = 0, \ \forall \alpha_i' \neq \alpha_{i0}, \tag{4.2.1}$$

2. rth-order local $\mathbb{E}$-ancillarity, where the expectation of the estimating function does not depend on the perturbed value α_i' within some neighborhood of α_{i0}:

$$\partial^k_{\alpha_i'} \mathbb{E}\left[g(\theta_0, \alpha_{i0}; y_i); \theta_0, \alpha_i'\right]\Big|_{\alpha_i' = \alpha_{i0}} = 0, \quad \text{for } k = 1, \ldots, r \tag{4.2.2}$$

Moment functions satisfying (4.2.1) are difficult to construct. Bonhomme (2012) provides a theory that characterizes such moment functions using functional differencing, which is motivated by the theory of orthogonal projections. He also shows that fixed-T consistent estimation is possible in fully parametric and static and some dynamic panel data settings under some conditions on the distribution of the incidental parameters. Global ancillarity is also equivalent to what Cox and Reid (1987)

call global orthogonality. Tibshirani and Wasserman (1994) call this exact orthogonality in expectation. Woutersen (2011) calls this a zero-score property which holds not just at the true value α_{i0}. Therefore, a sample analog of the score will produce a consistent root regardless of the value plugged in for the incidental parameter.

A more attainable goal is to consider (4.2.2) so that (4.2.1) holds in a smaller region of the parameter space. To further motivate this condition, I expand, up to the second order, the density f in the left hand side of (4.2.1), i.e.,

$$\begin{aligned}
& \mathbb{E}\left[g\left(\theta_0,\alpha_{i0};y_i\right);\theta_0,\alpha_i'\right] \\
= & \int g\left(\theta_0,\alpha_{i0};y_i\right) f\left(y_i;\theta_0,\alpha_i'\right)\, dy_i \\
= & \int g\left(\theta_0,\alpha_{i0};y_i\right) f\left(y_i;\theta_0,\alpha_{i0}\right)\, dy_i \\
& + \int g\left(\theta_0,\alpha_{i0};y_i\right) \partial_{\alpha_i'} f\left(y_i;\theta_0,\alpha_i'\right)\Big|_{\alpha_i'=\alpha_{i0}} \left(\alpha_i'-\alpha_{i0}\right)\, dy_i \\
& + \frac{1}{2}\int g\left(\theta_0,\alpha_{i0};y_i\right) \partial_{\alpha_i'}^2 f\left(y_i;\theta_0,\alpha_i'\right)\Big|_{\alpha_i'=\alpha_{i0}} \left(\alpha_i'-\alpha_{i0}\right)^2\, dy_i \\
= & \underbrace{\mathbb{E}\left[g\left(\theta_0,\alpha_{i0};y_i\right);\theta_0,\alpha_{i0}\right]}_{(a)} + \underbrace{\partial_{\alpha_i'}\mathbb{E}\left[g\left(\theta_0,\alpha_{i0};y_i\right);\theta_0,\alpha_i'\right]\Big|_{\alpha_i'=\alpha_{i0}}}_{(b)} \left(\alpha_i'-\alpha_{i0}\right) \\
& + \frac{1}{2}\partial_{\alpha_i'}^2\mathbb{E}\left[g\left(\theta_0,\alpha_{i0};y_i\right);\theta_0,\alpha_i'\right]\Big|_{\alpha_i'=\widetilde{\alpha}_i} \left(\alpha_i'-\alpha_{i0}\right)^2,
\end{aligned}$$

where $\widetilde{\alpha}_i$ is in between α_i' and α_{i0}. Since g is an unbiased estimating function, the term (a) in the preceding derivation is equal to zero. Under first-order local $\mathbb{E}$-ancillarity, the term (b) is also equal to zero. As a result, we have

$$\mathbb{E}\left[g\left(\theta_0,\alpha_{i0};y_i\right);\theta_0,\alpha_i'\right] = o\left(\alpha_i'-\alpha_{i0}\right).$$

Obviously, the extension to rth-order local $\mathbb{E}$-ancillarity will allow us to conclude that

$$\mathbb{E}\left[g\left(\theta_0,\alpha_{i0};y_i\right);\theta_0,\alpha_i'\right] = o\left(\alpha_i'-\alpha_{i0}\right)^r.$$

Notice that more and more smoothness would be required as one increases r.[4]

First-order local $\mathbb{E}$-ancillarity is what Cox and Reid (1987) call information orthogonality or local orthogonality when applied to the likelihood setting. They suggest finding a reparametrization so that θ and α_i are information orthogonal. They call the required transformation an orthogonal reparametrization, which means that,

[4]Nonsmooth objective functions, especially those that arise in quantile regressions, are not covered by these ancillarity conditions. It is unclear how smoothing these objective functions will affect the bias-reducing properties of these ancillarity conditions.

up to a certain order, estimating α_i will have minimal impact on consistently estimating θ. Lancaster (2002) and Woutersen (2011) derive orthogonal reparametrizations for common panel data models such as the static single index model with strictly exogenous regressors and the linear AR(1) dynamic panel data model. Unfortunately, finding an orthogonal reparametrization requires finding a solution (which may not exist) to a system of partial differential equations.

4.2.2 Implications

Instead of finding solutions to the system of partial differential equations and applying the reparametrization, we can determine how g should be restricted so that g will satisfy (4.2.2). Notice that rth-order local $\mathbb{E}$-ancillarity is equivalent to searching for g such that the following set of moment conditions will hold:

$$\mathbb{E}\Big[g\left(\theta_0, \alpha_{i0}; y_i\right) V_i^{(k)}\left(\theta_0, \alpha_{i0}\right)\Big] = 0, \text{ for } k = 1, \ldots, r, \tag{4.2.3}$$

where

$$V_i^{(k)}\left(\theta_0, \alpha_{i0}\right) = \frac{\partial_{\alpha_i}^k f\left(y_i; \theta_0, \alpha_{i0}\right)}{f\left(y_i; \theta_0, \alpha_{i0}\right)} \tag{4.2.4}$$

is the kth element of the so-called Bhattacharyya basis (see the pioneering works by Bhattacharyya (1946; 1947; 1948)).[5] To show the equivalence, write the left hand side of (4.2.3) as

$$\begin{aligned}
\mathbb{E}\Big[g\left(\theta_0, \alpha_{i0}; y_i\right) V_i^{(k)}\left(\theta_0, \alpha_{i0}\right)\Big] &= \int g\left(\theta_0, \alpha_{i0}; y_i\right) \partial_{\alpha_i'}^k f\left(y_i; \theta_0, \alpha_i'\right)\Big|_{\alpha_i' = \alpha_{i0}} dy_i \\
&= \partial_{\alpha_i'}^k \int g\left(\theta_0, \alpha_{i0}; y_i\right) f\left(y_i; \theta_0, \alpha_i'\right) dy_i \Big|_{\alpha_i' = \alpha_{i0}} \\
&= \partial_{\alpha_i'}^k \mathbb{E}\Big[g\left(\theta_0, \alpha_{i0}; y_i\right); \theta_0, \alpha_i'\Big]\Big|_{\alpha_i' = \alpha_{i0}},
\end{aligned}$$

[5] The Bhattacharyya basis is a natural basis to use when studying the effects of fluctuations of the incidental parameters (around the true value) on the density $f\left(y_i; \theta_0, \alpha_{i0}\right)$. Consider a perturbation in the incidental parameter from α_{i0} to α_i'. A Taylor series expansion of f about α_{i0} can be written as the following infinite sum

$$f\left(y_i; \theta_0, \alpha_i'\right) = f\left(y_i; \theta_0, \alpha_{i0}\right) + \partial_{\alpha_i} f\left(y_i; \theta_0, \alpha_{i0}\right)\left(\alpha_i' - \alpha_{i0}\right) + \partial_{\alpha_i}^2 f\left(y_i; \theta_0, \alpha_{i0}\right)\left(\alpha_i' - \alpha_{i0}\right)^2/2 + \cdots.$$

The likelihood ratio obtained from comparing the perturbed model to the true model can be written as

$$\begin{aligned}
\frac{f\left(y_i; \theta_0, \alpha_i'\right)}{f\left(y_i; \theta_0, \alpha_{i0}\right)} &= 1 + \frac{\partial_{\alpha_i} f\left(y_i; \theta_0, \alpha_{i0}\right)}{f\left(y_i; \theta_0, \alpha_{i0}\right)}\left(\alpha_i' - \alpha_{i0}\right) + \frac{1}{2}\frac{\partial_{\alpha_i}^2 f\left(y_i; \theta_0, \alpha_{i0}\right)}{f\left(y_i; \theta_0, \alpha_{i0}\right)}\left(\alpha_i' - \alpha_{i0}\right)^2 + \cdots \\
&= 1 + V_i^{(1)}\left(\theta_0, \alpha_{i0}\right)\left(\alpha_i' - \alpha_{i0}\right) + \frac{1}{2} V_i^{(2)}\left(\theta_0, \alpha_{i0}\right)\left(\alpha_i' - \alpha_{i0}\right)^2 + \cdots.
\end{aligned}$$

Relative to the true model, the perturbed model can be "summarized" in terms of an infinite number of basis elements of the form $V_i^{(k)}\left(\theta_0, \alpha_{i0}\right)$.

where the last expression is equal to zero by (4.2.2). Note that whenever an estimating function g satisfies rth-order local $\mathbb{E}$-ancillarity, it also satisfies kth-order local $\mathbb{E}$-ancillarity for all $k = 1, \ldots, r-1$.

At this point, I will reduce notation by suppressing the arguments (θ_0, α_{i0}). I now show some of the consequences of (4.2.3) when $r = 2$. First, note that

$$\mathbb{E}\left[\partial_{\alpha_{i0}} g\right] = \partial_{\alpha_{i0}} \mathbb{E}[g] - \mathbb{E}\left[g V_i^{(1)}\right] = 0, \tag{4.2.5}$$

which follows from the requirement that g be an unbiased estimating function and (4.2.3) when $r = 1$. Furthermore, another consequence of (4.2.3) when $r = 2$ is

$$Cov\left(V_i^{(1)}, \partial_{\alpha_{i0}} g\right) = \mathbb{E}\left[V_i^{(1)} \partial_{\alpha_{i0}} g\right] - \mathbb{E}\left[V_i^{(1)}\right] \mathbb{E}\left[\partial_{\alpha_{i0}} g\right] = \mathbb{E}\left[V_i^{(1)} \partial_{\alpha_{i0}} g\right] = 0. \tag{4.2.6}$$

This zero covariance property follows from calculating the derivative of (4.2.3) with respect to α_{i0}:

$$\partial_{\alpha_{i0}} \mathbb{E}\left[g V_i^{(1)}\right] = \mathbb{E}\left[g V_i^{(2)}\right] - \mathbb{E}\left[V_i^{(1)} \partial_{\alpha_{i0}} g\right]. \tag{4.2.7}$$

Since g satisfies first-order local $\mathbb{E}$-ancillarity, the expression $\mathbb{E}\left[g V_i^{(1)}\right]$ on the left hand side is equal to zero. Since g satisfies second-order local $\mathbb{E}$-ancillarity, the first term in the right hand side of (4.2.7) is equal to zero. As a result, the covariance between $V_i^{(1)}$ and $\partial_{\alpha_{i0}} g$ is zero whenever g satisfies second-order local $\mathbb{E}$-ancillarity. Finally,

$$\mathbb{E}\left[\partial_{\alpha_{i0}}^2 g\right] = \partial_{\alpha_{i0}} \mathbb{E}\left[\partial_{\alpha_{i0}} g\right] - \mathbb{E}\left[V_i^{(1)} \partial_{\alpha_{i0}} g\right] = 0, \tag{4.2.8}$$

which follows from (4.2.5) and (4.2.6).

It is exactly this zero covariance property (4.2.6), along with the consequences of second-order local $\mathbb{E}$-ancillarity (4.2.5) and (4.2.8), that mimics the bias reduction that has already been developed in the literature. Estimator-based corrections in the spirit of Hahn and Newey (2004) and Hahn and Kuersteiner (2011) trace the source of the bias in the estimator to the $O\left(T^{-1}\right)$ bias in the unadjusted score or moment function. To illustrate how their work relates to the projected score, I reproduce their calculation of the bias of some moment function u_{it} for the common parameter θ. Note that v_{it} is the moment function for the incidental parameter α_i. In the context of a static panel data model, the bias of u_{it} is given by

$$\mathbb{E}[u_{it}(\theta, \widehat{\alpha}_i)] = \frac{1}{T}\left\{\mathbb{E}\left[\partial_{\alpha_i} u_{it}\right]\beta_i + \mathbb{E}\left[\psi_{it} \partial_{\alpha_i} u_{it}\right] + \frac{1}{2}\mathbb{E}\left[\partial_{\alpha_i}^2 u_{it}\right]\mathbb{E}\left[\psi_{it}^2\right]\right\} + o\left(T^{-1}\right),$$

where ψ_{it} and β_i are components of the higher-order asymptotic expansion for $\widehat{\alpha}_i$, i.e.,

$$\psi_{it} = -\mathbb{E}\left[\partial_{\alpha_i} v_{it}\right]^{-1} v_{it},\ \beta_i = -\mathbb{E}\left[\partial_{\alpha_i} v_{it}\right]^{-1}\left\{\mathbb{E}\left[\psi_{it} \partial_{\alpha_i} v_{it}\right] + \frac{1}{2}\mathbb{E}\left[\partial_{\alpha_i}^2 v_{it}\right]\mathbb{E}\left[\psi_{it}^2\right]\right\}$$

Notice that if we chose a moment function u_{it} such that

$$\mathbb{E}\left[\partial_{\alpha_i} u_{it}\right] = 0,\ \mathbb{E}\left[\psi_{it}\partial_{\alpha_i} u_{it}\right] = 0,\ \mathbb{E}\left[\partial^2_{\alpha_i} u_{it}\right] = 0,$$

the $O\left(T^{-1}\right)$ bias disappears. These three equations are exactly the consequences of first-order local $\mathbb{E}$-ancillarity, the zero-covariance property in (4.2.6), and second-order local $\mathbb{E}$-ancillarity, respectively. It is in this sense that starting from local $\mathbb{E}$-ancillarity may be more transparent and intuitive when considering bias corrections.

Let us now consider the case of dynamic nonlinear panel data models. In their motivation for their bias correction procedure, Hahn and Kuersteiner (2011) show that the nonzero center of the asymptotic distribution of the uncorrected MLE is

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\mathbb{E}\left[\partial_{\alpha_i} v_{it}\right]}\mathbb{E}\left[\frac{1}{T}\left(\sum_{t=1}^{T} v_{it}\right)\left(\sum_{t=1}^{T}\partial_{\alpha_i} u_{it}\right)\right] - \frac{\mathbb{E}\left[\partial^2_{\alpha_i} u_{it}\right]}{2\left(\mathbb{E}\left[\partial_{\alpha_i} v_{it}\right]\right)^2}\mathbb{E}\left[\frac{1}{T}\left(\sum_{t=1}^{T} v_{it}\right)^2\right].$$

Once again notice that if we choose a moment function for the common parameters u_{it} that satisfies second-order local $\mathbb{E}$-ancillarity, this nonzero center disappears.

In addition to the preceding discussion, the criterion of second-order local $\mathbb{E}$-ancillarity is also constructive because we can interpret (4.2.3) in Hilbert space terms, where the expectation operator is the inner product. We can think of (4.2.3) as finding g that is orthogonal to a linear subspace spanned by $\left(V_i^{(1)}, \ldots, V_i^{(r)}\right)$. This linear subspace represents local effects of incidental parameter fluctuations. An analogous idea appears in linear regression settings so that we can interpret the desired estimating function g as a residual orthogonal to the explanatory variables $\left(V_i^{(1)}, \ldots, V_i^{(r)}\right)$. This residual is called the rth-order projected score. In principle, one can construct the rth-order projected score but a lot of the benefits in terms of bias correction can already be reaped at the second order as seen in the preceding discussions.

4.2.3 Computation

Let us consider the situation where one has a complete specification of a likelihood for the data. For every $i = 1, \ldots, n$, let $z_i = (y_{i0}, y_{i1}, \ldots, y_{iT}, x_{i1}, \ldots, x_{iT})$ be the data for the ith unit and $z = (z_1, \ldots, z_n)$ be the full data. Let $f(z_{it}; \theta, \alpha_i)$ be the density of the data where $\theta \in \mathbb{R}^p$ and $\alpha_i \in \mathbb{R}^q$. Assume the cross-sectional units are independent of each other. The joint density of the observables is given by

$$f(z; \theta, \alpha) = \prod_{i=1}^{n} f(z_i; \theta, \alpha_i).$$

Note that the density $f(z_i; \theta, \alpha_i)$ is specified such that predetermined regressors can be accommodated. For example, if we let $x_i^t = (x_{i1}, \ldots, x_{it})$ and $y_i^t = (y_{i0}, y_{i1}, \ldots, y_{it})$,

we can write $f(z_i;\theta,\alpha_i)$ as

$$\begin{aligned} f(z_i;\theta,\alpha_i) &= f(y_{iT}|x_i^T,y_i^{T-1};\theta,\alpha_i)\times f(x_{iT}|y_i^{T-1},x_i^{T-1})\times\ldots\times f(y_{i2}|x_i^2,y_i^1;\theta,\alpha_i) \\ &\quad \times f(x_{i2}|y_i^1,x_{i1})\times f(y_{i1}|x_{i1},y_{i0};\theta,\alpha_i)\times f(y_{i0},x_{i1}) \end{aligned}$$

We usually specify parametric models for $f\left(y_{it}|x_i^t,y_i^{t-1};\theta,\alpha_i\right)$ and treat these models as structural. Flexible reduced forms can then be used to specify the feedback processes $f\left(x_{it}|y_i^{t-1},x_i^{t-1}\right)$. These flexible reduced forms can introduce further individual-specific fixed effects different from α_i. Examples can be found in Moral-Benito (2013; 2014). Note that the distribution of the initial values $f(y_{i0},x_{i1})$ can be specified or be left unspecified. If left unspecified, I condition on initial values.

The θ-score and α_i-score be can be written as

$$\begin{aligned} U_{i,0}(\theta,\alpha_i;z_i) &= \partial_\theta \log f(z_i;\theta,\alpha_i), \\ V_i(\theta,\alpha_i;z_i) &= \partial_{\alpha_i}\log f(z_i;\theta,\alpha_i). \end{aligned}$$

Observe that the α_i-score only uses the time-series observations for the ith cross-sectional unit and is a function of α_i and not of α_j for $j\neq i$.

When we set $k=1$ in (4.2.4), $V_i^{(1)}$ coincides with the α_i-score so that $V_i^{(1)}=V_i$. The second-order terms $V_i^{(2)}$ can be written as

$$V_i^{(2)} = \partial_{\alpha_i^{\mathrm{T}}} V_i + V_i V_i^{\mathrm{T}}. \tag{4.2.9}$$

The preceding recurrence relation, which can be generalized to the rth order, is a consequence of

$$\partial_{\alpha_i^{\mathrm{T}}} V_i = \partial_{\alpha_i^{\mathrm{T}}}\left(\frac{\partial_{\alpha_i} f}{f}\right) = \frac{f\times\partial_{\alpha_i,\alpha_i^{\mathrm{T}}} f - \partial_{\alpha_i} f\times\partial_{\alpha_i^{\mathrm{T}}} f}{f^2} = \frac{\partial_{\alpha_i,\alpha_i^{\mathrm{T}}} f}{f} - \frac{\partial_{\alpha_i} f}{f}\frac{\partial_{\alpha_i^{\mathrm{T}}} f}{f} = V_i^{(2)} - V_i V_i^{\mathrm{T}},$$

which follows from the quotient rule for derivatives. Note that (4.2.9) is a recurrence relation because one can generate $V_i^{(r)}$ from $V_i^{(r-1)}$. Define the second-order extended information matrix as

$$M_{i,2} = \mathbb{E}\left[\begin{pmatrix} U_{i,0} \\ V_i \\ vec\left[V_i^{(2)}\right] \end{pmatrix}\begin{pmatrix} U_{i,0}^{\mathrm{T}} & V_i^{\mathrm{T}} & vec\left[V_i^{(2)}\right]^{\mathrm{T}} \end{pmatrix}\right] = \begin{pmatrix} M_{11,i} & M_{12,i} \\ M_{21,i} & M_{22,i} \end{pmatrix},$$

where the submatrices are defined as follows:

$$\begin{aligned} M_{11,i} &= \mathbb{E}\left[U_{i,0}U_{i,0}^{\mathrm{T}}\right], \\ M_{12,i} &= M_{21,i}^{\mathrm{T}} = \mathbb{E}\left[\begin{pmatrix} U_{i,0}V_i^{\mathrm{T}} & U_{i,0}vec\left[V_i^{(2)}\right]^{\mathrm{T}} \end{pmatrix}\right], \end{aligned}$$

$$M_{22,i} = \mathbb{E}\left[\begin{pmatrix} V_i \\ vec\left[V_i^{(2)}\right] \end{pmatrix}\begin{pmatrix} V_i^{\mathrm{T}} & vec\left[V_i^{(2)}\right]^{\mathrm{T}} \end{pmatrix}\right].$$

The second-order projected score and its information matrix for the ith unit could be expressed as

$$U_{i,2} = U_{i,0} - M_{12,i}\left(M_{22,i}\right)^{-}\begin{pmatrix} V_i \\ vec\left[V_i^{(2)}\right] \end{pmatrix}, \tag{4.2.10}$$

$$I_{i,2} = M_{11,i} - M_{12,i}\left(M_{22,i}\right)^{-} M_{21,i}. \tag{4.2.11}$$

where $\left(M_{22,i}\right)^{-}$ is the Moore-Penrose inverse of $M_{22,i}$. As discussed in the previous subsection, the second-order projected score is really the residual orthogonal to the linear subspace spanned by $\left(V_i^{(1)}, V_i^{(2)}\right)$. Thus, the second-order projected score $U_{i,2}$ makes the θ-score $U_{i,0}$ less sensitive to the presence of the incidental parameters. The second-order projected score and its associated information matrix for the full data can then be computed by summing up n components of the form (4.2.10) and (4.2.11).

As a result of all the preceding discussions, I present the following lemma and a more formal proof in the appendix.

Lemma 4.2.1. The second-order projected score $U_{i,2}$ is an unbiased estimating equation that satisfies second-order local $\mathbb{E}$-ancillarity (4.2.2).

In general, the projected score may depend on both θ and α_i. Thus, we have to substitute an estimator for α_i to form a plug-in projected score. The first-order projected score for the ith unit can be written as

$$U_{i,1} = U_{i,0} - \mathbb{E}\left(U_{i,0}V_i^{\mathrm{T}}\right)\left[\mathbb{E}\left(V_iV_i^{\mathrm{T}}\right)\right]^{-} V_i.$$

Solving $V_i = 0$ gives an estimator for α_i given θ, denoted by $\widehat{\alpha_i}(\theta)$. The plug-in first-order projected score $\sum_i \widehat{U_{i,1}}$ coincides with the profile score for θ. Dhaene and Jochmans (2015b) show that the panel Poisson model and panel exponential duration model have profile scores that have zero expectation. Therefore, the plug-in first-order projected score mimics the behavior of the profile score when applied to these models.

On the other hand, the second-order projected score is given by

$$\begin{aligned} U_{i,2} = {} & U_{i,0} - \mathbb{E}\left(U_{i,0}V_i^{\mathrm{T}}\right)\left[\mathbb{E}\left(V_iV_i^{\mathrm{T}}\right)\right]^{-} V_i \\ & - \mathbb{E}\left(U_{i,0}vec\left[V_i^{(2)}\right]^{\mathrm{T}}\right)\left[\mathbb{E}\left(vec\left[V_i^{(2)}\right]vec\left[V_i^{(2)}\right]^{\mathrm{T}}\right)\right]^{-} vec\left[V_i^{(2)}\right]. \end{aligned} \tag{4.2.12}$$

The next two propositions show that the plug-in second-order projected score matches the properties of existing bias corrections.

Proposition 4.2.2. *Assume that the conditions for (4.5.2), the conditions of Lemma 4.2.1, and the central limit theorems for V_i , V_i^2, $\partial_{\alpha_i} U_{i,2}$, and $\partial_{\alpha_i}^2 U_{i,2}$ hold. Then,*

$$\mathbb{E}\left(\widehat{U_{i,2}}(\theta_0) - U_{i,2}\right) = \mathbb{E}\left(\widehat{U_{i,2}}(\theta_0) - U_{i,0}\right) = O\left(T^{-1}\right). \tag{4.2.13}$$

Since the second-order projected score $U_{i,2}$ satisfies second-order local $\mathbb{E}$-ancillarity by Lemma 4.2.1, we have (a) $\mathbb{E}\left[\partial_{\alpha_i} U_{i,2}\right] = 0$, (b) zero covariance between $\partial_{\alpha_i} U_{i,2}$ and V_i, and (c) $\mathbb{E}\left[\partial_{\alpha_i}^2 U_{i,2}\right] = 0$. These three implications of second-order local $\mathbb{E}$-ancillarity are the most crucial reasons why $U_{i,2}$ already provides much of the bias reduction that existing methods aim to provide, just as sketched in the preceding subsection.

Assume that the system of equations implied by the plug-in second-order projected score has a solution in some neighborhood of the true value θ_0. We denote this solution by $\widehat{\theta^c}$ and it satisfies $\sum_i \widehat{U_{i,2}}\left(\widehat{\theta^c}\right) = 0$. This solution has an asymptotic distribution that is exactly the asymptotic distribution of the MLE.

Proposition 4.2.3. *Under the asymptotic scheme where $n, T \to \infty$, $n/T \to c \in (0, \infty)$, and $n/T^3 \to 0$, we have*

$$\sqrt{nT}\left(\widehat{\theta^c} - \theta_0\right) \xrightarrow{d} N\left(0, \left(\lim_{n,T\to\infty} \frac{1}{nT} \sum_{i=1}^{n} \mathbb{E}\left[U_{i,0} U_{i,0}^{\mathrm{T}}\right]\right)^{-1}\right). \tag{4.2.14}$$

4.2.4 Examples

Consider the following examples to demonstrate the calculations and some of the complications (and virtues) that may arise for the plug-in second-order projected score.

Example 4.2.4. (Linear AR(1) dynamic panel data model) Let $y_{it} = \alpha_i + \rho y_{i,t-1} + \varepsilon_{it}$ where $\varepsilon_{it} \sim \text{iid}\, N\left(0, \sigma^2\right)$ for all $i = 1, \ldots, N$ and $t = 1, 2$. Note that I do not restrict ρ so that y_{it} will be stationary. I condition on y_{i0} and assume that it is uncorrelated with future realizations of ε_{it}. The MLE for α_i given ρ and σ^2 is $\widehat{\alpha_i}\left(\rho, \sigma^2\right) = \bar{y}_i - \rho \bar{y}_{i,-1}$, where $\bar{y}_i = \left(y_{i1} + y_{i2}\right)/2$ and $\bar{y}_{i,-1} = \left(y_{i0} + y_{i1}\right)/2$. After calculating the second-order projected score for this case,[6] we substitute the MLE for $\widehat{\alpha_i}\left(\rho, \sigma^2\right)$ and obtain the following system of equations:

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \frac{\sigma^2 + \left(y_{i1} - y_{i0}\right)\left(y_{i2} - y_{i1} - \rho\left(y_{i1} - y_{i0}\right)\right)}{2\sigma^2} &= 0, \\ \frac{1}{n}\sum_{i=1}^{n} \frac{-2\sigma^2 + \left(y_{i2} - y_{i1} - \rho\left(y_{i1} - y_{i0}\right)\right)^2}{4\sigma^4} &= 0. \end{aligned}$$

[6] Explicit calculations can be found in the appendix.

Eliminating σ^2 from the preceding system gives

$$\frac{2}{n}\sum_{i=1}^{n}(y_{i1}-y_{i0})(y_{i2}-y_{i1}-\rho(y_{i1}-y_{i0}))+\frac{1}{n}\sum_{i=1}^{n}(y_{i2}-y_{i1}-\rho((y_{i1}-y_{i0})))^2=0.$$

Simplifying the equation above gives a quadratic equation in ρ of the form $A_n\rho^2+B_n\rho+C_n=0$ where

$$\begin{aligned}
A_n &= \frac{1}{n}\sum_{i=1}^{n}(y_{i1}-y_{i0})^2,\\
B_n &= -\frac{2}{n}\sum_{i=1}^{n}\left[(y_{i2}-y_{i1})(y_{i1}-y_{i0})+(y_{i1}-y_{i0})^2\right],\\
C_n &= \frac{1}{n}\sum_{i=1}^{n}(y_{i2}-y_{i1})^2+\frac{2}{n}\sum_{i=1}^{n}(y_{i1}-y_{i0})(y_{i2}-y_{i1}).
\end{aligned}$$

I now show consistency of one of the roots of the quadratic equation. First, assume that $A_n \xrightarrow{p} A \neq 0$. Since $Cov(\epsilon_{i2}-\epsilon_{i1}, y_{i1}-y_{i0})=-\sigma^2$, we must have $B_n \xrightarrow{p} -2\rho A+2\sigma^2-2A$ and $C_n \xrightarrow{p} \rho^2 A-2\rho\sigma^2+2\rho A$. By Slutsky's lemma, we also have $B_n^2-4A_nC_n \xrightarrow{p} 4\left(\sigma^2-A\right)^2$. This means that the quadratic equation will always have real roots. As a result, we have

$$\widehat{\rho_n}=\frac{-B_n\pm\sqrt{B_n^2-4A_nC_n}}{2A_n}\xrightarrow{p}\rho-\left(\frac{\sigma^2}{A}-1\right)\pm\left|\frac{\sigma^2}{A}-1\right|,$$

where we either have $\widehat{\rho_n}\xrightarrow{p}\rho$ or $\widehat{\rho_n}\xrightarrow{p}\rho-2\left(\sigma^2/A-1\right)$. The estimator for $\widehat{\sigma_n^2}$ is given by

$$\widehat{\sigma_n^2}=-\frac{1}{n}\sum_{i=1}^{n}(y_{i1}-y_{i0})(y_{i2}-y_{i1}-\widehat{\rho_n}(y_{i1}-y_{i0})),$$

and will only be consistent if $\widehat{\rho_n}$ is consistent. Notice that the roots were obtained without resorting to an iterative procedure unlike the bias correction proposal by Bun and Carree (2005).

Which of the two roots should be chosen? To illustrate, consider the case where we have stationarity. Assume that y_{i0} is drawn from its stationary distribution where $\mathbb{E}(y_{i0})=\alpha_i/(1-\rho)$ and $Var(y_{i0})=\sigma^2/\left(1-\rho^2\right)$, where $|\rho|<1$. In this case, $A_n \xrightarrow{p} 2\sigma^2/(1+\rho)\neq 0$. As a result, $\sigma^2/A-1<0$. Thus, the consistent root is the smaller root of the quadratic equation. Now, consider the case where $\rho=1$. Note that the large-n limit of A_n is such that $\sigma^2/A-1<0$ since $y_{i1}-y_{i0}=\alpha_i+\varepsilon_{i1}$ implies that $\mathbb{E}(y_{i1}-y_{i0})^2=\mathbb{E}(\alpha_i+\varepsilon_{i1})^2=\mathbb{E}\left(\alpha_i^2\right)+\sigma^2>\sigma^2$. As a result, the consistent root is still the smaller root of the quadratic equation.

Dhaene and Jochmans (2015a) extensively document the behavior of the resulting likelihood obtained after integrating the adjusted profile score. They have shown that the profile score has a bias that depends only on the common parameters and not on the incidental parameters. The adjusted profile score is then the difference between the profile score and its bias. They also propose a procedure to choose among the multiple critical points of the adjusted likelihood. Extensions of the model that allow for incidental trends can be found in Moon and Phillips (2004), where they also link the second-order projected score to their proposed moment condition.

Allowing for further lags should be straightforward for the projected score because a scalar pth order difference equation can be written as a vector first-order difference equation. Therefore, the quadratic equation derived for the AR(1) case is still going to be a quadratic equation with coefficients that are matrices. Allowing for regressors, whether strictly exogenous or predetermined, will not remove the multiple root problem and will have to be examined on a case-by-case basis. ■

To explore the effect of including a predetermined regressor, consider an extension of the previous example that automatically allows for two individual-specific fixed effects.

Example 4.2.5. (Linear panel VAR(1) model) Consider the following structural model for the dynamics of two variables (y_{it}, x_{it}):

$$\begin{aligned} y_{it} &= \phi_{11} y_{i,t-1} + \phi_{12} x_{i,t-1} + \eta_{xi} + \varepsilon_{1it} \\ x_{it} &= \phi_{21} y_{i,t-1} + \phi_{22} x_{i,t-1} + \eta_{yi} + \varepsilon_{2it} \end{aligned}$$

where the idiosyncratic errors have the following distribution:

$$\begin{pmatrix} \varepsilon_{1it} \\ \varepsilon_{2it} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} \right).$$

for $i = 1, \ldots, n$ and $t = 1, 2$. Assume that (i) Σ is positive definite, i.e., $\det(\Sigma) = \sigma_{11}\sigma_{22} - \sigma_{12}^2 > 0$, (ii) the initial observations (y_{i0}, x_{i0}) are available, and (iii) the distribution of the fixed effects and initial observations are left unspecified. The structural parameters are $\boldsymbol{\theta} = (\phi_{11}, \phi_{12}, \phi_{21}, \phi_{22}, \sigma_{11}, \sigma_{22}, \sigma_{12})$. The MLEs for η_{xi} and η_{yi} given the other parameters are

$$\begin{aligned} \widehat{\eta_{xi}} &= \frac{1}{2}(y_{i2} - y_{i1}(\phi_{11} - 1) - \phi_{11} y_{i0} - \phi_{12} x_{i0} - \phi_{12} x_{i1}) \\ \widehat{\eta_{yi}} &= \frac{1}{2}(x_{i2} - \phi_{21} y_{i0} - \phi_{21} y_{i1} - x_{i1}(\phi_{22} - 1) - \phi_{22} x_{i0}) \end{aligned}$$

The explicit calculations for the projected score can be carried out in *Mathematica*. The expectation of the plug-in first-order projected score for the ith cross-sectional

unit has nonzero bias, i.e.

$$\mathbb{E}\left(\widehat{\mathbf{U}_{i,1}}\right) = \left(-\frac{1}{2}, 0, 0, -\frac{1}{2}, -\frac{\sigma_{22}}{2\det(\Sigma)}, -\frac{\sigma_{11}}{2\det(\Sigma)}, -\frac{\sigma_{12}}{\det(\Sigma)}\right).$$

Notice that this nonzero bias does not depend on η_{xi} and η_{yi}. As a result, this fits into Case 2 of Dhaene and Jochmans (2015b), where the profile score has expectation free of the incidental parameters. Similarly, calculations in *Mathematica* show that the expectation of the plug-in second-order projected score for the ith cross-sectional unit has zero bias. ■

Next, I consider a nonlinear model where the score of some conditional likelihood for the model is an unbiased estimating equation.

Example 4.2.6. (Static logit model with strictly exogenous regressors) Suppose $y_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2} \sim \text{Bernoulli}(p_{it})$ with probability of success $p_{it} = \mathbb{E}(y_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}) = F(\alpha_i + \mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta})$ for $i = 1, \ldots, n$ and $t = 1, 2$. Assume that F is the logistic cdf. For $j = 0, 1, 2$, define $N_j = \{i : y_{i1} + y_{i2} = j\}$. Following the computations in the appendix, the second-order projected score using all i can be computed as

$$\begin{aligned}\sum_{i=1}^{n} U_2^i &= \sum_{i\in N_0} y_{i2} + \sum_{i\in N_2}(y_{i2}-1) + \sum_{i\in N_1}(\mathbf{x}_{i2}-\mathbf{x}_{i1})^{\mathrm{T}}\left[y_{i2} - \frac{1}{1+e^{(\mathbf{x}_{i2}-\mathbf{x}_{i1})^{\mathrm{T}}\boldsymbol{\beta}}}\right] \\ &= \sum_{i\in N_1}(\mathbf{x}_{i2}-\mathbf{x}_{i1})^{\mathrm{T}}\left[y_{i2} - \frac{1}{1+e^{(\mathbf{x}_{i2}-\mathbf{x}_{i1})^{\mathrm{T}}\boldsymbol{\beta}}}\right]. \qquad (4.2.15)\end{aligned}$$

Note that the individuals in N_0 and N_2 have zero contribution to the plug-in second-order projected score. Although the above expression is monotonically decreasing in β, there is no closed-form solution to the above estimating equation. Despite this, the plug-in second-order projected score can be shown that this coincides with score of the conditional likelihood formed from the units for which $y_{i1} + y_{i2} = 1$. Since Chamberlain (1980) shows that the conditional MLE is $\sqrt{n}$-consistent, the same goes for the root of the plug-in second-order projected score.

Arellano and Bonhomme (2009) derive a bias-reducing prior for this model for general T that removes the $O\left(T^{-1}\right)$ bias. Their Monte Carlo simulations include an estimator where the adjustment was iterated. The simulations indicate that the iterated adjustment will mimic the properties of the conditional score when n is fixed and T increased to around 20. In contrast, Dhaene and Jochmans (2015b), who also consider the case of $T = 2$, show that the conditional score can be obtained either by an infinite-order profile score adjustment or by rescaling the profile score by the total number of movers. It is unclear whether rescaling will extend to the case where $T > 2$. ■

4.3 Simulations

In this section, I show that the finite sample performance of the plug-in second-order projected score is as good as or sometimes better than some existing competitors. I focus on panels with a very small value of T for the following reasons. Panels obtained from developing countries or panels formed from small-scale experiments usually have single-digit T. In practice, applied researchers will also use a subset of the data, especially when there are structural breaks in the time series or when the data are unbalanced. Therefore, it seems appropriate to choose small values of T to gauge finite sample performance.

I implement the projected score method and other alternatives using *Mathematica*.[7] *Mathematica* allows us to calculate the symbolic representation of the projected score and to compute the roots using the `FindRoot` command. Thus, the user only needs to specify the likelihood function and modify the code for the situation he considers without recoding the actual expressions of the corrections.[8] Furthermore, the calculations become much more compact and organized. I use two starting points, namely, the MLEs for the pooled and fixed effects model, for the root-finding algorithm. I use the software `R` to generate the data for the Monte Carlo experiments and to compute the MLEs (using the routine `glm`) under the pooled and fixed effects model.[9] The draws for the individual-specific fixed effects α_i are fixed across 5000 replications.

The implementation exploits the comparative advantages of both `R` and *Mathematica*. `R` can be used to generate samples from a user-specified data generating process and to perform routine estimation procedures, while *Mathematica* can be used to symbolically calculate the adjusted score and find its roots. The coding style in the *Mathematica* notebook allows any end user to do the following:

1. Specify either an objective function or an estimating function based on some parametric model.

2. Use the built-in commands for differentiation and calculation of expectations to produce symbolic representations of the adjustment found in (4.2.12) .

3. Import data and estimation results. The data and estimation results can come from any statistical software capable of exporting its outputs to a text file.

4. Use the programmed functions to generate empirical counterparts of the symbolic representations, to calculate roots and produce output for diagnostics, and to generate routine estimation results such as standard errors.

[7] All *Mathematica* notebooks and `R` code are available upon request.

[8] Coding the actual expressions would take an inordinate number of lines of code and would only be valid for a specific model.

[9] Whenever the MLE does not exist, I take notice of this and I increase the number of replications so that I could attain the target of 5000 replications.

The coding style almost creates the feeling of a built-in package which may attract more users. But the user only has to change the parametric model in the *Mathematica* notebook whenever the user contemplates changes in the model.

To construct the plug-in second-order projected score, I compute the projected score as discussed in (4.2.12) and use an estimator for α_i. Rather than recompute $\widehat{\alpha_i}(\theta)$ at every iteration of the root-finding algorithm, I use a linear approximation of $\widehat{\alpha_i}(\theta)$ around $\widehat{\alpha_i}$ suggested by Bellio and Sartori (2003), i.e.,

$$\widehat{\alpha_i}(\theta) = \widehat{\alpha_i} + j^{-1}_{\alpha_i\alpha_i}\left(\widehat{\theta},\widehat{\alpha_i}\right) j_{\alpha_i\theta}\left(\widehat{\theta},\widehat{\alpha_i}\right)\left(\widehat{\theta}-\theta\right),$$

where $j_{\alpha_i\alpha_i}$ and $j_{\alpha_i\theta}$ are the corresponding (α_i,α_i) and (α_i,θ) blocks of the observed information matrix

$$j(\theta,\alpha_i) = \begin{bmatrix} j_{\theta\theta} & j_{\theta\alpha_i} \\ j_{\alpha_i\theta} & j_{\alpha_i\alpha_i} \end{bmatrix},$$

respectively. Other alternatives may be possible, for instance, using penalized likelihood estimator proposed by Firth (1993) and Kosmidis and Firth (2010) or the EM-based estimator proposed by Chen (2014). The idea behind these estimators is to improve the quality and stability of the plug-in values for α_i. These alternatives may be helpful in models where the plug-in values for α_i are either extreme or even undefined.

The first data generating process I consider is the static probit model. I use the following design adapted from Fernandez-Val (2009) with some modifications. The original design included a stationary AR(1) model with a linear time trend for the exogenous regressor x_{it}. Omitting this feature leads to the following modified specification:

$$\begin{aligned} &y_{it}|x_{i1},\ldots,x_{iT},\alpha_i \sim Ber\left(p_{it}\right), \;\; p_{it} = \Phi\left(\alpha_i + \beta_0 x_{it}\right), \\ &x_{it} \sim \text{iid } N\left(0,1\right), \alpha_i \sim \text{iid } N\left(0,1\right), x_{it} \perp \alpha_i, \\ &n = 125, T = 4, \beta_0 = 0.5, \end{aligned}$$

where $\Phi(\cdot)$ is the standard normal CDF. An important thing to note is that the regressor is already independent of the fixed effects.[10] I choose this design because I stripped it down to the simplest elements. I already explored the static logit case in an example found in the previous section.

I also compare the performance of the projected score to the uncorrected MLE, the corrected estimator by Fernandez-Val (2009), and the score corrections by Carro (2007) and Woutersen (2003). Table 4.3.1 contains simulation results for the static probit model based on 5000 replications. The results indicate good finite sample performance of the projected score relative to all the other corrections. The Monte Carlo estimate of the bias is almost reduced by 90% relative to the uncorrected MLE.

[10]The exogenous regressor x is redrawn for every replication for all the experiments in this section.

As a result, taking higher-order projections may not be needed as the gains will be marginal relative to computational cost. Furthermore, the standard deviation of the estimator obtained from the projected score is comparable to the standard deviation of the other estimators. The results clearly indicate that score-based corrections may be preferable in terms of RMSE. Although the number of nonconvergent cases is very small relative to the number of replications, I recommend obtaining a log of the iterations produced by the root-finding algorithm when implementing score-based corrections.

Table 4.3.1: Finite sample performance of estimators of β_0

Estimator	Mean bias	Median bias	Standard deviation	Median AD	RMSE
Uncorrected MLE	0.210	0.203	0.135	0.089	0.723
Fernandez-Val (2009)	0.162	0.156	0.123	0.081	0.674
Woutersen (2003)	0.069	0.064	0.099	0.066	0.577
(12 cases nonconvergent)					
Carro (2007)	0.071	0.066	0.100	0.066	0.580
(13 cases nonconvergent)					
Projected score	0.030	0.025	0.095	0.063	0.538

Note: True value of β_0 is equal to 0.5. Results are based on 5000 replications.

The second data generating process is the first-order dynamic logit model. Once more, I adapt the design from Fernandez-Val (2009) with some modifications.

$$
\begin{aligned}
&y_{it}|y_{i,t-1},\ldots,y_{i0},x_{i0},x_{i1},\ldots,x_{iT},\alpha_i \sim Ber\left(p_{it}\right),\ \ p_{it}=F\left(\alpha_i+\rho_0 y_{i,t-1}+\beta_0 x_{it}\right),\\
&y_{i0}|x_{i0},x_{i1},\ldots,x_{iT},\alpha_i \sim Ber\left(p_{i0}\right),\ \ p_{i0}=F\left(\alpha_i+\beta_0 x_{i0}\right),\\
&x_{it}\sim \text{iid } L(0,1),\alpha_i \sim \text{iid } L(0,1), x_{it}\perp\alpha_i,\\
&n=125, T=3, \beta_0=1, \rho_0=0.5.
\end{aligned}
$$

In this design, $F(\cdot)$ is the logistic CDF and $L(0,1)$ is the logistic distribution with mean 0 and scale 1. The original design assumes that $x_{it} \sim N\left(0,\pi^2/3\right)$ and the individual-specific fixed effects were generated as an average of the four oldest values of x_{it}. I choose to use $L(0,1)$ because it is quite similar to $N\left(0,\pi^2/3\right)$ but with heavier tails. I condition on y_{i0} instead of using the information from the distribution $y_{i0}|x_{i0},x_{i1},\ldots,x_{iT}$ in the likelihood function. For this model, the alternatives are the fixed -T consistent estimator proposed by Honoré and Kyriazidou (2000), the corrected estimators by Fernandez-Val (2009) and Hahn and Kuersteiner (2011), and the score-based corrections by Carro (2007) and Woutersen (2003).

Recall that Hahn and Kuersteiner (2011) obtain a characterization of the nonzero center of the asymptotic distribution of the MLE as discussed in Example 1.2.3. Estimator-based corrections will have to rely on an estimator of this nonzero center. This nonzero center depends on the cross-spectrum of the α_i-score and the deriva-

tive of the θ-score with respect to α_i at the zero frequency and the spectrum of the α_i-score at the zero frequency. Since the cross-spectrum and spectrum are infinite sums of cross-covariances and covariances, respectively, a feasible procedure would require some lag truncation. As a result, we would require an integer bandwidth of lower order than $T^{1/2}$ for trimming purposes and for the asymptotic theory to hold. Since $T = 3$, I set the bandwidth at values 0, 1, and 2.

In contrast, Honoré and Kyriazidou (2000) propose an estimator based on the maximizer of a likelihood conditioned on the subset of observations for which $\{x_{i2} = x_{i3}\}$. Since this set is a zero probability event given the DGP, a kernel with a corresponding bandwidth is used to give higher weight to observations where x_{i2} is close to x_{i3} and give lower weight to observations otherwise. I use a standard normal kernel for this purpose. Furthermore, I use the optimal bandwidth derived by Honoré and Kyriazidou (2000) which is a constant multiple of $T^{-1/5}$. I set this constant to values 1, 8, and 64 just as Honoré and Kyriazidou (2000) do so in their own simulations.

The Monte Carlo results in Table 4.3.2 indicate that score-based corrections are performing quite well relative to estimator-based corrections for the design I consider. The bias of the root of the projected score is almost eliminated for both coefficients of the linear predictor. In contrast, the other score-based estimators are having problems eliminating the bias in the autoregressive coefficient. There seems to be a point at which a higher bandwidth will not improve finite sample performance of estimator-based corrections. In fact, the estimator-based correction by Hahn and Kuersteiner (2011) almost has the same performance as uncorrected MLE when the bandwidth is equal to 2. Furthermore, the dispersion of the corrected estimators is less than half that of the uncorrected MLE with the exception of the correction by Hahn and Kuersteiner (2011). The dispersion of the root of the projected score is more in line with that of the uncorrected MLE.

I also present two power curves in Table 4.3.1 for the projected score in the dynamic logit model. I do not present the results for the competing procedures because the estimated biases are large relative to the estimated standard deviation. The rejection probability of the test $\rho = 0.5$ is almost 5% while that of the test $\beta = 1$ is about 2%. Unfortunately, power is relatively low but this is expected as the asymptotics require a large value for T.

It is clear from the Monte Carlo results that the projected score is a competitive alternative to some of the competing bias-reduction procedures (especially with respect to finite sample bias but not in RMSE terms). The biggest downside is the computational time. For the designs considered, setting up of the projected score, the calculation of the root, and the standard error calculations took about 2 to 5 minutes for every replication on a laptop with 8 GB memory and an i7-processor. Even if we exploit parallel processing, the memory requirement is almost too great for all cores to be used all at once, especially when conducting Monte Carlo simulations. The reason for the high memory requirement is in the nature of the correction – a

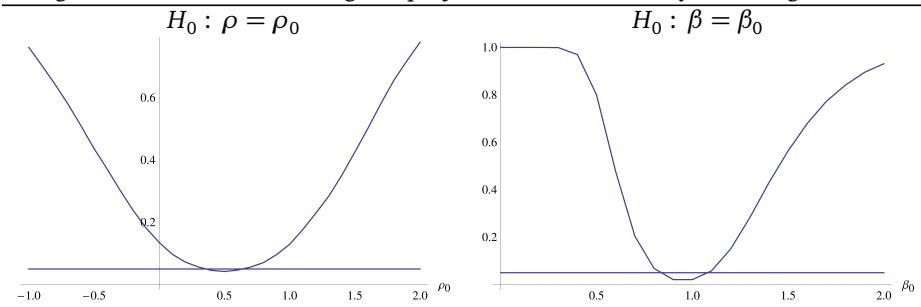
Table 4.3.2: Finite sample performance of estimators of $\overline{\beta}$ and $\overline{\rho}$

	Mean bias		Median bias		Standard deviation		Median AD		RMSE	
Estimator	ρ	β	ρ	β	ρ	β	ρ	β	ρ	β
Uncorrected MLE	-2.660	0.867	-2.588	0.792	0.902	0.457	0.558	0.272	2.341	1.921
Hahn and Kuersteiner (2011)										
Bandwidth 0	-0.552	-0.759	-1.066	-0.301	3.049	2.477	0.606	0.371	3.049	2.488
Bandwidth 1	-0.341	-0.028	-0.723	0.242	2.870	2.019	0.689	0.390	2.874	2.240
Bandwidth 2	-1.957	0.865	-1.870	0.783	0.897	0.485	0.538	0.281	1.711	1.927
Fernandez-Val (2009)										
Bandwidth 0	-1.994	0.225	-1.984	0.232	0.555	0.213	0.363	0.099	1.593	1.243
Bandwidth 1	-1.807	0.217	-1.795	0.226	0.554	0.212	0.363	0.096	1.419	1.235
Bandwidth 2	-1.948	0.211	-1.940	0.220	0.552	0.212	0.359	0.095	1.550	1.230
Honoré and Kyriazidou (2000)										
Bandwidth 1			0.268	0.550			1.771	0.885		
Bandwidth 8			-0.049	0.126			0.561	0.265		
Bandwidth 64			-0.059	0.131			0.541	0.250		
Woutersen (2003)	-0.183	-0.047	-0.181	-0.052	0.348	0.124	0.236	0.080	0.471	0.961
Carro (2007)	-0.505	-0.047	-0.506	-0.056	0.329	0.121	0.219	0.080	0.329	0.961
(1 case nonconvergent)										
Projected score	0.011	0.048	0.004	0.023	1.170	0.402	0.366	0.136	1.277	1.122
(24 cases nonconvergent)										

Note: The true values are given by $\rho_0 = 0.5$ and $\beta_0 = 1$. Results are based on 5000 replications.

symbolic representation is created and the data are substituted into this representation. Despite these issues, the implementation is very straightforward and would not require us to program new procedures every time we make changes to the model.

Figure 4.3.1: Inference using the projected score for the dynamic logit model



Note: Significance level set at 5% and represented as horizontal line

4.4 Concluding remarks

This paper develops a bias reduction method for the estimators of common parameters of a linear or nonlinear panel data model with individual-specific fixed effects. The past decades saw a spur of research on bias reduction methods. It is easier to see what these methods have in common by considering what is called the projected score. This projected score is calculated by projecting the score vector for the common parameters onto the orthogonal complement of a space characterized by incidental parameter fluctuations.

I show that projected scores reduce the asymptotic bias of the estimators of common parameters in panel data models. Although the projected score has been introduced two decades ago, its widespread use has been hindered by computational issues. Relative to other bias reduction procedures, computation (in terms of processor time and memory) may be prohibitive but programming is less error-prone and more intuitive. I hope that this will encourage applied researchers to use the projected score. Monte Carlo simulations indicate that the bias-reducing properties of the projected score already take effect even for very small sample sizes usually encountered when panel data models are estimated for subsamples. Finally, the applied researcher need not choose a bandwidth anymore.

Future work on practical aspects include extensions to nonsmooth functions arising, say, in quantile regression. In addition, the projection idea has to be modified when one wants to extend to non-likelihood settings and when one wants to include time effects. I intend to pursue these extensions in the future.

4.5 Appendix

Proof of Lemma 4.2.1

To show that $U_{i,2}$ is an unbiased estimating equation, we have to show that $\mathbb{E}\left[U_{i,0}\right] = 0$, $\mathbb{E}[V_i] = 0$, and $\mathbb{E}\left[vec\left[V_i^{(2)}\right]\right] = 0$. The first two statements follow from the zero-mean property of the scores. Since $\mathbb{E}\left[vec\left[V_i^{(2)}\right]\right] = vec\left(\mathbb{E}\left[V_i^{(2)}\right]\right)$, we have to show that $\mathbb{E}\left[V_i^{(2)}\right] = 0$. Differentiating $\mathbb{E}[V_i] = 0$ with respect to α_i gives the desired result. Thus, we have shown that $U_{i,2}$ is an unbiased estimating equation. To show second-order $\mathbb{E}$-ancillarity, we can show that (4.2.12) satisfies the moment conditions in (4.2.3) for $k = 1, 2$. This follows by construction. ■

Proof of Proposition 4.2.2

To simplify the exposition, I return to the case where incidental parameter is scalar. To show (4.2.13), consider a second-order Taylor series expansion of the plug-in second-order projected score for the ith individual about the true value α_{i0}, i.e.

$$\begin{aligned}\widehat{U_{i,2}}(\theta_0) &= U_{i,2} + \partial_{\alpha_i} U_{i,2}(\widehat{\alpha_i}(\theta_0) - \alpha_{i0}) + \frac{1}{2}\partial^2_{\alpha_i} U_{i,2}(\widehat{\alpha_i}(\theta_0) - \alpha_{i0})^2 \\ &\quad + O_p\left(T^{-1/2}\right). \end{aligned} \tag{4.5.1}$$

Under regularity conditions for maximum likelihood estimation, the three terms in (4.5.1) are $O_p\left(T^{1/2}\right)$, $O_p\left(T^{1/2}\right)$, and $O_p(1)$. The final term is a zero mean $O_p\left(T^{-1/2}\right)$ term. Note that the first-order conditions used to obtain a plug-in estimator for α_i can be expanded in the following way:

$$\widehat{V_i}(\theta_0) = V_i + \partial_{\alpha_i} V_i(\widehat{\alpha_i}(\theta_0) - \alpha_{i0}) + O_p(1).$$

Since the right hand size is equal to zero, we can write

$$\widehat{\alpha_i}(\theta_0) - \alpha_{i0} = -\frac{V_i}{\mathbb{E}\left(\partial_{\alpha_i} V_i\right)} + O_p\left(T^{-1}\right). \tag{4.5.2}$$

Furthermore, the square of $\widehat{\alpha_i}(\theta_0) - \alpha_{i0}$ can be written as

$$\begin{aligned} &(\widehat{\alpha_i}(\theta_0) - \alpha_{i0})^2 \\ = \;& \frac{V_i^2}{\left[\mathbb{E}\left(\partial_{\alpha_i} V_i\right)\right]^2} - 2\frac{V_i}{\mathbb{E}\left(\partial_{\alpha_i} V_i\right)} O_p\left(T^{-1}\right) + O_p\left(T^{-2}\right) \\ = \;& \frac{V_i^2}{\left[\mathbb{E}\left(\partial_{\alpha_i} V_i\right)\right]^2} - 2\frac{\mathbb{E}(V_i)}{\mathbb{E}\left(\partial_{\alpha_i} V_i\right)} O_p\left(T^{-1}\right) - 2\frac{O_p\left(T^{1/2}\right)}{\mathbb{E}\left(\partial_{\alpha_i} V_i\right)} O_p\left(T^{-1}\right) + O_p\left(T^{-2}\right) \end{aligned}$$

$$
\begin{aligned}
&= \frac{\mathbb{E}\left(V_i^2\right)}{\left[\mathbb{E}\left(\partial_{\alpha_i} V_i\right)\right]^2} + \frac{O_p\left(T^{1/2}\right)}{\left[\mathbb{E}\left(\partial_{\alpha_i} V_i\right)\right]^2} + O_p\left(T^{-3/2}\right) \\
&= \frac{\mathbb{E}\left(V_i^2\right)}{\left[\mathbb{E}\left(\partial_{\alpha_i} V_i\right)\right]^2} + O_p\left(T^{-3/2}\right). \qquad (4.5.3)
\end{aligned}
$$

Note that $\mathbb{E}(V_i) = 0$ because the α_i-score is an unbiased estimating equation. Central limit theorems for V_i and V_i^2 allow us to obtain (4.5.3). After substituting (4.5.2) into $\partial_{\alpha_i} U_{i,2}(\widehat{\alpha_i}(\theta_0) - \alpha_{i0})$, we have

$$
\begin{aligned}
\partial_{\alpha_i} U_{i,2}(\widehat{\alpha_i}(\theta_0) - \alpha_{i0}) &= -\frac{V_i \partial_{\alpha_i} U_{i,2}}{\mathbb{E}\left(\partial_{\alpha_i} V_i\right)} + \partial_{\alpha_i} U_{i,2} O_p\left(T^{-1}\right) \\
&= -\frac{V_i \partial_{\alpha_i} U_{i,2}}{\mathbb{E}\left(\partial_{\alpha_i} V_i\right)} + \mathbb{E}\left(\partial_{\alpha_i} U_{i,2}\right) O_p\left(T^{-1}\right) + O_p\left(T^{-1/2}\right) \\
&= -\frac{V_i \partial_{\alpha_i} U_{i,2}}{\mathbb{E}\left(\partial_{\alpha_i} V_i\right)} + O_p\left(T^{-1/2}\right) \qquad (4.5.4)
\end{aligned}
$$

A central limit theorem for $\partial_{\alpha_i} U_{i,2}$ and second-order local $\mathbb{E}$-ancillarity allow us to produce the previous derivation. The expression in (4.5.4) involves the product of $\partial_{\alpha_i} U_{i,2}$ and V_i and a zero mean $O_p\left(T^{-1/2}\right)$ term. As a result, the expectation of the term $\partial_{\alpha_i} U_{i,2}(\widehat{\alpha_i}(\theta_0) - \alpha_{i0})$ is $O\left(T^{-1}\right)$.

Next, we substitute (4.5.3) into $\partial^2_{\alpha_i} U_{i,2}(\widehat{\alpha_i}(\theta_0) - \alpha_{i0})^2$. As a result, we obtain

$$
\begin{aligned}
&\partial^2_{\alpha_i} U_{i,2}(\widehat{\alpha_i}(\theta_0) - \alpha_{i0})^2 \\
&= \frac{\partial^2_{\alpha_i} U_{i,2} \mathbb{E}\left(V_i^2\right)}{\left[\mathbb{E}\left(\partial_{\alpha_i} V_i\right)\right]^2} + \partial^2_{\alpha_i} U_{i,2} O_p\left(T^{-3/2}\right) \\
&= \frac{\partial^2_{\alpha_i} U_{i,2} \mathbb{E}\left(V_i^2\right)}{\left[\mathbb{E}\left(\partial_{\alpha_i} V_i\right)\right]^2} + \mathbb{E}\left(\partial^2_{\alpha_i} U_{i,2}\right) O_p\left(T^{-3/2}\right) + O_p\left(T^{1/2}\right) O_p\left(T^{-3/2}\right) \\
&= \frac{\partial^2_{\alpha_i} U_{i,2} \mathbb{E}\left(V_i^2\right)}{\left[\mathbb{E}\left(\partial_{\alpha_i} V_i\right)\right]^2} + O_p\left(T^{-1}\right) \qquad (4.5.5)
\end{aligned}
$$

The expression in (4.5.5) involves $\partial^2_{\alpha_i} U_{i,2}$, which has zero expectation because of second-order local $\mathbb{E}$-ancillarity, and an $O_p\left(T^{-1}\right)$ term. As a result, the expectation of the term $\partial^2_{\alpha_i} U_{i,2}(\widehat{\alpha_i}(\theta_0) - \alpha_{i0})$ is $O\left(T^{-1}\right)$.

Proof of Proposition 4.2.3

Assume that the system of equations implied by the plug-in second-order projected score has a solution in some neighborhood of the true value θ_0. We denote this

solution by $\widehat{\theta^c}$ and it satisfies $\sum_{i=1}^{N} \widehat{U_{i,2}}\left(\widehat{\theta^c}\right) = 0$. Consider the following first-order Taylor series expansion of the plug-in second-order projected score around θ_0, i.e.

$$\sum_{i=1}^{n} \widehat{U_{i,2}}\left(\widehat{\theta^c}\right) = \sum_{i=1}^{n} \widehat{U_{i,2}}(\theta_0) + \sum_{i=1}^{n} \frac{d}{d\theta} \widehat{U_{i,2}}(\bar{\theta})\left(\widehat{\theta^c} - \theta_0\right). \tag{4.5.6}$$

Note that the left hand side of (4.5.6) is equal to zero because $\widehat{\theta^c}$ is the root of the plug-in second-order projected score. Rewrite (4.5.6) as

$$\sqrt{nT}\left(\widehat{\theta^c} - \theta_0\right) = \left(\frac{1}{nT}\sum_{i=1}^{n} \frac{d}{d\theta} \widehat{U_{i,2}}(\bar{\theta})\right)^{-1} \frac{1}{\sqrt{nT}} \sum_{i=1}^{n} \widehat{U_{i,2}}(\theta_0). \tag{4.5.7}$$

Let $n, T \to \infty$ and $n/T \to c \in (0, \infty)$. Note that

$$\frac{1}{\sqrt{nT}}\sum_{i=1}^{n}\left[\widehat{U_{i,2}}(\theta_0) - U_{i,2}\right] = \frac{1}{\sqrt{nT}}\sum_{i=1}^{n}\mathbb{E}\left[\widehat{U_{i,2}}(\theta_0) - U_{i,2}\right] + O_p(1) = O_p\left(\sqrt{\frac{n}{T^3}}\right) + O_p(1)$$

The first equality comes from replacing the empirical mean with an expectation and leaving behind a zero-mean $O_p(1)$ term. The second equality comes from the order calculation in Proposition 4.2.2. Provided that $n/T^3 \to 0$, $\frac{1}{\sqrt{nT}}\sum_i \widehat{U_{i,2}}(\theta_0)$ can be approximated by $\frac{1}{\sqrt{nT}}\sum_i U_{i,2}$ and the latter quantity is asymptotically normal. A central limit theorem applies to $\frac{1}{\sqrt{nT}}\sum_i U_{i,2}$ (similar to the score in likelihood settings), i.e.

$$\frac{1}{\sqrt{nT}}\sum_{i=1}^{n} U_{i,2} \xrightarrow{d} N\left(0, \lim_{n,T\to\infty} \frac{1}{nT}\sum_{i=1}^{n}\mathbb{E}\left[U_{i,2}U_{i,2}^{\mathrm{T}}\right]\right).$$

Next, note that

$$\left.\frac{d}{d\theta}\widehat{U_{i,2}}(\theta)\right|_{\theta=\bar{\theta}} = \left.\left[\partial_\theta \widehat{U_{i,2}}(\theta) + \left(\partial_{\alpha_i}\widehat{U_{i,2}}(\theta)\right)\left(\partial_\theta \widehat{\alpha_i}(\theta)\right)\right]\right|_{\theta=\bar{\theta}} \tag{4.5.8}$$

by the chain rule. Replacing $\widehat{U_{i,2}}(\theta)$ with its Taylor series expansion

$$\widehat{U_{i,2}}(\theta) = U_{i,2}(\theta, \alpha_{i0}) + \partial_{\alpha_i} U_{i,2}(\theta, \alpha_{i0})(\widehat{\alpha_i}(\theta) - \alpha_{i0}) + O_p(1) \tag{4.5.9}$$

and calculating the derivatives in (4.5.8) yields

$$\begin{aligned}
\partial_\theta \widehat{U_{i,2}}(\theta) &= \partial_\theta U_{i,2}(\theta, \alpha_{i0}) + \partial^2_{\theta\alpha_i} U_{i,2}(\theta, \alpha_{i0})(\widehat{\alpha_i}(\theta) - \alpha_{i0}) \\
&\quad + \partial_{\alpha_i} U_{i,2}(\theta, \alpha_{i0})(\partial_\theta \widehat{\alpha_i}(\theta)) + O_p(1), \qquad (4.5.10) \\
\partial_{\alpha_i} \widehat{U_{i,2}}(\theta) &= \partial_{\alpha_i} U_{i,2}(\theta, \alpha_{i0}) + \partial^2_{\alpha_i} U_{i,2}(\theta, \alpha_{i0})(\widehat{\alpha_i}(\theta) - \alpha_{i0}) \\
&\quad + \partial_{\alpha_i} U_{i,2}(\theta, \alpha_{i0})\left(\partial_{\alpha_i} \widehat{\alpha_i}(\theta)\right) + O_p(1). \qquad (4.5.11)
\end{aligned}$$

Taking probability limits, we have the following components:

$$\begin{aligned}
\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\operatorname*{plim}_{T\to\infty}\frac{1}{T}\partial_\theta U_{i,2}(\theta,\alpha_{i0})\Big|_{\theta=\bar\theta} &= \lim_{n,T\to\infty}\frac{1}{nT}\sum_{i=1}^{n}\mathbb{E}\big[\partial_\theta U_{i,2}\big],\\
\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\operatorname*{plim}_{T\to\infty}\frac{1}{T}\partial^2_{\theta\alpha_i}U_{i,2}(\theta,\alpha_{i0})(\widehat{\alpha_i}(\theta)-\alpha_{i0})\Big|_{\theta=\bar\theta} &= 0,\\
\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\operatorname*{plim}_{T\to\infty}\frac{1}{T}\partial_{\alpha_i}U_{i,2}(\theta,\alpha_{i0})\big(\partial_{\alpha_i}\widehat{\alpha_i}(\theta)\big)\Big|_{\theta=\bar\theta} &= 0,\\
\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\operatorname*{plim}_{T\to\infty}\frac{1}{T}\partial_{\alpha_i}U_{i,2}(\theta,\alpha_{i0})\Big|_{\theta=\bar\theta} &= 0,\\
\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\operatorname*{plim}_{T\to\infty}\frac{1}{T}\partial^2_{\alpha_i}U_{i,2}(\theta,\alpha_{i0})(\widehat{\alpha_i}(\theta)-\alpha_{i0})\Big|_{\theta=\bar\theta} &= 0,\\
\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\operatorname*{plim}_{T\to\infty}\frac{1}{T}\partial_{\alpha_i}U_{i,2}(\theta,\alpha_{i0})\big(\partial_{\alpha_i}\widehat{\alpha_i}(\theta)\big)\Big|_{\theta=\bar\theta} &= 0.
\end{aligned}$$

Note that as $T\to\infty$, we have both $\bar\theta\xrightarrow{p}\theta_0$ and $\widehat{\alpha_i}\big(\bar\theta\big)\xrightarrow{p}\alpha_{i0}$. The second and fifth equalities follow $\widehat{\alpha_i}\big(\bar\theta\big)\xrightarrow{p}\alpha_{i0}$ as $T\to\infty$. The third, fourth, and sixth equalities would follow from the law of large numbers and second-order ancillarity. The $O_p(1)$ terms in (4.5.9), (4.5.10), and (4.5.11) all converge to zero because $\widehat{\alpha_i}\big(\bar\theta\big)\xrightarrow{p}\alpha_{i0}$ as $T\to\infty$. We can then conclude that

$$\operatorname*{plim}_{n,T\to\infty}\frac{1}{nT}\sum_{i=1}^{n}\frac{d}{d\theta}\widehat{U_{i,2}}(\theta)\Big|_{\theta=\bar\theta}=\lim_{n,T\to\infty}\frac{1}{nT}\sum_{i=1}^{n}\mathbb{E}\big[\partial_\theta U_{i,2}\big].$$

Notice that $U_{i,2}$ behaves like $U_{i,0}$ asymptotically because the correction in (4.2.12) has expectation zero at the true value. As long as the information identity holds, we have $\mathbb{E}\Big[U_{i,2}U_{i,2}^{\mathrm{T}}\Big]=\mathbb{E}\Big[U_{i,0}U_{i,0}^{\mathrm{T}}\Big]=\mathbb{E}\big[\partial_\theta U_{i,0}\big]$. Otherwise, we have the usual sandwich-type asymptotic covariance matrix.

Alternative proof of (4.2.14)

In this appendix, we prove the main results in the spirit of the papers by Hahn and Newey (2004) and Hahn and Kuersteiner (2011). We also note some departures from their proof. Let F_i and $\widehat{F_i}$ denote the CDF and its empirical counterpart for the ith individual. Define $F_i(\epsilon)=F_i+\epsilon\sqrt{T}(\widehat{F_i}-F_i)$ and $\Delta_{iT}=\sqrt{T}(\widehat{F_i}-F_i)$, where $\epsilon\in[0,T^{-1/2}]$. We have $\mathbf{F}(\epsilon)=\mathbf{F}+\epsilon\sqrt{T}(\widehat{\mathbf{F}}-\mathbf{F})$ in vector form.

Let $\alpha_i(\boldsymbol{\theta},F_i(\epsilon))$ and $\boldsymbol{\theta}(\mathbf{F}(\epsilon))$ be the solutions to the estimating equations below:

$$\int V_i(\boldsymbol{\theta},\alpha_i(\boldsymbol{\theta},F_i(\epsilon));\mathbf{z}_i)\,dF_i(\epsilon) = 0 \qquad (4.5.12)$$

$$\sum_{i=1}^{n} \int U_{i,2}(\boldsymbol{\theta}(\mathbf{F}(\epsilon)), \alpha_i(\boldsymbol{\theta}(\mathbf{F}(\epsilon)), F_i(\epsilon)); \mathbf{z})\, dF_i(\epsilon) \quad = \quad 0 \tag{4.5.13}$$

The plug-in used for the α_i's in the second-order projected score can be written as $\widehat{\alpha_i}(\theta) = \alpha_i(\boldsymbol{\theta}, F_i(T^{-1/2}))$. The root for the plug-in version of the second-order projected score can be written as $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}(\mathbf{F}(T^{-1/2}))$. On the other hand, the true values can be written as $\boldsymbol{\theta}_0 = \boldsymbol{\theta}(\mathbf{F}(0)) = \boldsymbol{\theta}(\mathbf{F})$ and $\alpha_{i0} = \alpha_i(\boldsymbol{\theta}, F_i)$.

Expand the functional $\boldsymbol{\theta}(\widehat{\mathbf{F}})$ about the true value $\boldsymbol{\theta}(\mathbf{F})$ up to the third order, i.e.

$$\boldsymbol{\theta}(\widehat{\mathbf{F}}) - \boldsymbol{\theta}(\mathbf{F}) = \frac{1}{\sqrt{T}}\boldsymbol{\theta}^{\epsilon}(0) + \frac{1}{2}\left(\frac{1}{\sqrt{T}}\right)^2 \boldsymbol{\theta}^{\epsilon\epsilon}(0) + \frac{1}{6}\left(\frac{1}{\sqrt{T}}\right)^3 \boldsymbol{\theta}^{\epsilon\epsilon\epsilon}(\tilde{\epsilon}) \tag{4.5.14}$$

where

$$\boldsymbol{\theta}^{\epsilon}(0) = \partial_{\epsilon}\boldsymbol{\theta}(\mathbf{F}(\epsilon))|_{\epsilon=0},\ \boldsymbol{\theta}^{\epsilon\epsilon}(0) = \partial_{\epsilon}^2\boldsymbol{\theta}(\mathbf{F}(\epsilon))|_{\epsilon=0},\ \boldsymbol{\theta}^{\epsilon\epsilon\epsilon}(0) = \partial_{\tilde{\epsilon}}^3\boldsymbol{\theta}(\mathbf{F}(\epsilon))|_{\epsilon=\tilde{\epsilon}\in[0,T^{-1/2}]} \tag{4.5.15}$$

Define the object

$$h_i(\epsilon) = U_{i,2}(\boldsymbol{\theta}(\mathbf{F}(\epsilon)), \alpha_i(\boldsymbol{\theta}(\mathbf{F}(\epsilon)), F_i(\epsilon))) \tag{4.5.16}$$

where the dependence on the data is suppressed. Hahn and Newey (2004) and Hahn and Kuersteiner (2011) use $U_{i,1}$ instead of $U_{i,2}$. It follows that (4.5.13) can be rewritten as

$$\frac{1}{n}\sum_{i=1}^{n} \int h_i(\epsilon)\, dF_i(\epsilon) = 0 \tag{4.5.17}$$

We show that when $n, T \to \infty$ such that $n/T \to c \in (0, \infty)$,

$$\sqrt{nT}(\boldsymbol{\theta}(\widehat{\mathbf{F}}) - \boldsymbol{\theta}(\mathbf{F})) \xrightarrow{d} N\left(\mathbf{0}, \left(\lim_{n,T\to\infty} \frac{1}{nT}\sum_{i=1}^{n} \mathbf{I}_i\right)^{-1}\right) \tag{4.5.18}$$

in the following manner:

1. Differentiate (4.5.17) with respect to ϵ twice. The resulting expressions can be decomposed into two terms: a term that requires integration with respect to $F_i(\epsilon)$ and a term that characterizes the "tail" or the remainder. We have

$$\frac{1}{n}\sum_{i=1}^{n} \int \frac{dh_i(\epsilon)}{d\epsilon}\, dF_i(\epsilon) + \frac{1}{n}\sum_{i=1}^{n} \int h_i(\epsilon)\, d\Delta_{iT} \quad = \quad 0 \tag{4.5.19}$$

$$\frac{1}{n}\sum_{i=1}^{n} \int \frac{d^2h_i(\epsilon)}{d\epsilon^2}\, dF_i(\epsilon) + \frac{2}{n}\sum_{i=1}^{n} \int \frac{dh_i(\epsilon)}{d\epsilon}\, d\Delta_{iT} \quad = \quad 0 \tag{4.5.20}$$

2. Compute the total derivatives in the previous equations noting the dependence of $\boldsymbol{\theta}(\mathbf{F}(\epsilon))$ and $\alpha_i(\boldsymbol{\theta}(\mathbf{F}(\epsilon)), F_i(\epsilon))$ on ϵ.

$$
\begin{aligned}
&\frac{dh_i(\epsilon)}{d\epsilon} \\
=\;& \partial_{\boldsymbol{\theta}} h_i(\epsilon)\partial_\epsilon \boldsymbol{\theta} + \partial_{\alpha_i} h_i(\epsilon)(\partial_{\boldsymbol{\theta}}\alpha_i)^{\mathrm{T}}\partial_\epsilon\boldsymbol{\theta} + \partial_{\alpha_i} h_i(\epsilon)\partial_\epsilon\alpha_i \\
&\frac{d^2h_i(\epsilon)}{d\epsilon^2} \\
=\;& \underbrace{\partial_\epsilon\boldsymbol{\theta}\left[\partial_\theta^2 h_i(\epsilon)\partial_\epsilon\boldsymbol{\theta} + \partial^2_{\theta,\alpha_i} h_i(\epsilon)\partial_{\boldsymbol{\theta}}\alpha_i\partial_\epsilon\boldsymbol{\theta} + \partial^2_{\theta,\alpha_i} h_i(\epsilon)\partial_\epsilon\alpha_i\right] + \partial_{\boldsymbol{\theta}} h_i(\epsilon)\partial_\epsilon^2\boldsymbol{\theta}}_{\partial_\epsilon(\partial_{\boldsymbol{\theta}} h_i(\epsilon)\partial_\epsilon\boldsymbol{\theta})} \\
&+\partial_{\boldsymbol{\theta}}\alpha_i\partial_\epsilon\boldsymbol{\theta}\underbrace{\left[\partial^2_{\theta,\alpha_i} h_i(\epsilon)\partial_\epsilon\boldsymbol{\theta} + \partial^2_{\alpha_i} h_i(\epsilon)\partial_{\boldsymbol{\theta}}\alpha_i\partial_\epsilon\boldsymbol{\theta} + \partial^2_{\alpha_i} h_i(\epsilon)\partial_\epsilon\alpha_i\right]}_{\partial_\epsilon(\partial_{\alpha_i} h_i(\epsilon))} \\
&+\partial_{\alpha_i} h_i(\epsilon)\underbrace{\left[(\partial_\epsilon\boldsymbol{\theta})^2\partial^2_{\boldsymbol{\theta}}\alpha_i + \partial^2_{\boldsymbol{\theta},\epsilon}\alpha_i\partial_\epsilon\boldsymbol{\theta} + \partial_{\boldsymbol{\theta}}\alpha_i\partial^2_\epsilon\boldsymbol{\theta}\right]}_{\partial_\epsilon(\partial_{\boldsymbol{\theta}}\alpha_i\partial_\epsilon\boldsymbol{\theta})} \\
&+\partial_\epsilon\alpha_i\left[\partial^2_{\theta,\alpha_i} h_i(\epsilon)\partial_\epsilon\boldsymbol{\theta} + \partial^2_{\alpha_i} h_i(\epsilon)\partial_{\boldsymbol{\theta}}\alpha_i\partial_\epsilon\boldsymbol{\theta} + \partial^2_{\alpha_i} h_i(\epsilon)\partial_\epsilon\alpha_i\right] \\
&+\partial_{\alpha_i} h_i(\epsilon)\underbrace{\left[(\partial^2_{\boldsymbol{\theta},\epsilon}\alpha_i)^{\mathrm{T}}\partial_\epsilon\boldsymbol{\theta} + \partial^2_\epsilon\alpha_i\right]}_{\partial_\epsilon(\partial_\epsilon\alpha_i)}
\end{aligned}
$$

3. Next, we have to derive $\boldsymbol{\theta}^\epsilon(0)$ and $\boldsymbol{\theta}^{\epsilon\epsilon}(0)$. This means that we have to evaluate the expressions in (b) at $\epsilon = 0$. Use the definitions of $\boldsymbol{\theta}_0$, α_{i0} and (4.5.16) to rewrite the resulting expressions. As a consequence, we have

$$
\begin{aligned}
&\left.\frac{dh_i(\epsilon)}{d\epsilon}\right|_{\epsilon=0} \\
=\;& \left[\partial_{\boldsymbol{\theta}} U_{i,2}(\boldsymbol{\theta}_0, \alpha_{i0}) + \partial_{\alpha_i} U_{i,2}(\boldsymbol{\theta}_0, \alpha_{i0})(\partial_{\boldsymbol{\theta}}\alpha_i(\boldsymbol{\theta}_0, F_i))^{\mathrm{T}}\right]\boldsymbol{\theta}^\epsilon(0) \\
&+\partial_{\alpha_i} U_{i,2}(\boldsymbol{\theta}_0, \alpha_{i0})\partial_{\boldsymbol{\theta}}\alpha_i(\boldsymbol{\theta}_0, F_i)\boldsymbol{\theta}^\epsilon(0) \\
&+\partial_{\alpha_i} U_{i,2}(\boldsymbol{\theta}_0, \alpha_{i0})\partial_\epsilon\alpha_i(\boldsymbol{\theta}_0, F_i)
\end{aligned}
\tag{4.5.21}
$$

$$
\begin{aligned}
&\left.\frac{d^2h_i(\epsilon)}{d\epsilon^2}\right|_{\epsilon=0} \\
=\;& \boldsymbol{\theta}^\epsilon(0)\left[\partial^2_{\boldsymbol{\theta}} U_{i,2}(\boldsymbol{\theta}_0, \alpha_{i0})\boldsymbol{\theta}^\epsilon(0) + \partial^2_{\boldsymbol{\theta},\alpha_i} U_{i,2}(\boldsymbol{\theta}_0, \alpha_{i0})\partial_{\boldsymbol{\theta}}\alpha_i(\boldsymbol{\theta}_0, F_i)\boldsymbol{\theta}^\epsilon(0)\right] \\
&+\boxed{\boldsymbol{\theta}^\epsilon(0)\partial^2_{\boldsymbol{\theta},\alpha_i} U_{i,2}(\boldsymbol{\theta}_0, \alpha_{i0})\partial_\epsilon\alpha_i(\boldsymbol{\theta}_0, F_i)} + \partial_{\boldsymbol{\theta}} U_{i,2}(\boldsymbol{\theta}_0, \alpha_{i0})\boldsymbol{\theta}^{\epsilon\epsilon}(0) \\
&+\partial_{\boldsymbol{\theta}}\alpha_i(\boldsymbol{\theta}_0, F_i)\boldsymbol{\theta}^\epsilon(0)\left[\partial^2_{\boldsymbol{\theta},\alpha_i} U_{i,2}(\boldsymbol{\theta}_0, \alpha_{i0})\boldsymbol{\theta}^\epsilon(0)\right] \\
&+\boxed{\partial_{\boldsymbol{\theta}}\alpha_i(\boldsymbol{\theta}_0, F_i)\boldsymbol{\theta}^\epsilon(0)\left[\partial^2_{\alpha_i} U_{i,2}(\boldsymbol{\theta}_0, \alpha_{i0})\partial_{\boldsymbol{\theta}}\alpha_i(\boldsymbol{\theta}_0, F_i)\boldsymbol{\theta}^\epsilon(0)\right]}
\end{aligned}
$$

$$
\begin{aligned}
&+ \boxed{\partial_{\boldsymbol{\theta}}\alpha_i(\boldsymbol{\theta}_0,F_i)\boldsymbol{\theta}^{\epsilon}(0)\left[\partial^2_{\alpha_i}U_{i,2}(\boldsymbol{\theta}_0,\alpha_{i0})\partial_{\epsilon}\alpha_i(\boldsymbol{\theta}_0,F_i)\right]}\\
&+ \boxed{\partial_{\alpha_i}U_{i,2}(\boldsymbol{\theta}_0,\alpha_{i0})(\boldsymbol{\theta}^{\epsilon}(0))^{\mathrm{T}}\partial^2_{\boldsymbol{\theta}}\alpha_i(\boldsymbol{\theta}_0,F_i)\boldsymbol{\theta}^{\epsilon}(0)}\\
&+ \boxed{\partial_{\alpha_i}U_{i,2}(\boldsymbol{\theta}_0,\alpha_{i0})\partial^2_{\boldsymbol{\theta},\epsilon}\alpha_i(\boldsymbol{\theta}_0,F_i)\boldsymbol{\theta}^{\epsilon}(0)}\\
&+ \boxed{\partial_{\alpha_i}U_{i,2}(\boldsymbol{\theta}_0,\alpha_{i0})[\partial_{\boldsymbol{\theta}}\alpha_i(\boldsymbol{\theta}_0,F_i)\boldsymbol{\theta}^{\epsilon\epsilon}(0)]}\\
&+ \boxed{\partial_{\epsilon}\alpha_i(\boldsymbol{\theta}_0,F_i)\left[\partial^2_{\boldsymbol{\theta},\alpha_i}U_{i,2}(\boldsymbol{\theta}_0,\alpha_{i0})\boldsymbol{\theta}^{\epsilon}(0)\right]}\\
&+ \boxed{\partial_{\epsilon}\alpha_i(\boldsymbol{\theta}_0,F_i)\left[\partial^2_{\alpha_i}U_{i,2}(\boldsymbol{\theta}_0,\alpha_{i0})\partial_{\boldsymbol{\theta}}\alpha_i(\boldsymbol{\theta}_0,F_i)\boldsymbol{\theta}^{\epsilon}(0)\right]}\\
&+ \boxed{\boxed{\partial_{\epsilon}\alpha_i(\boldsymbol{\theta}_0,F_i)\left[\partial^2_{\alpha_i}U_{i,2}(\boldsymbol{\theta}_0,\alpha_{i0})\partial_{\epsilon}\alpha_i(\boldsymbol{\theta}_0,F_i)\right]}}\\
&+ \boxed{\partial_{\alpha_i}U_{i,2}(\boldsymbol{\theta}_0,\alpha_{i0})\partial^2_{\boldsymbol{\theta},\epsilon}\alpha_i(\boldsymbol{\theta}_0,F_i)\boldsymbol{\theta}^{\epsilon}(0)}\\
&+ \boxed{\boxed{\partial_{\alpha_i}U_{i,2}(\boldsymbol{\theta}_0,\alpha_{i0})\partial^2_{\epsilon}\alpha_i(\boldsymbol{\theta}_0,F_i)}} \qquad (4.5.22)
\end{aligned}
$$

4. Substitute the above expressions into (4.5.19) and (4.5.20). The first sum in (4.5.19) and (4.5.20) when evaluated at $\epsilon = 0$ becomes the expectation with respect to the true values while the second sum becomes a "tail" term characterizing the difference between the realized distribution $\widehat{F_i}$ and the true one F_i. Since $\boldsymbol{\theta}^{\epsilon}(0)$ and $\boldsymbol{\theta}^{\epsilon\epsilon}(0)$ do not depend on the data, they can be treated as constants with respect to the expectation.

5. We need to derive the expressions for the first and second derivatives of $\alpha_i(\boldsymbol{\theta}, F_i)$ with respect to $\boldsymbol{\theta}$ and ϵ. Differentiate (4.5.12) with respect to $\boldsymbol{\theta}$ and ϵ. Solve the resulting system of two equations in two unknowns for $\partial_{\boldsymbol{\theta}}\alpha_i(\boldsymbol{\theta}, F_i(\epsilon))$ and $\partial_{\epsilon}\alpha_i(\boldsymbol{\theta}, F_i(\epsilon))$. Next, get the second derivatives of (4.5.12) with respect to $\boldsymbol{\theta}$ and ϵ. Solve the resulting system of three equations in three unknowns for $\partial^2_{\boldsymbol{\theta}}\alpha_i(\boldsymbol{\theta}, F_i(\epsilon))$, $\partial^2_{\boldsymbol{\theta},\epsilon}\alpha_i(\boldsymbol{\theta}, F_i(\epsilon))$, and $\partial^2_{\epsilon}\alpha_i(\boldsymbol{\theta}, F_i(\epsilon))$.[11] In effect, we are applying the Implicit Function Theorem and evaluating at $\epsilon = 0$ and $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. The resulting first derivatives would be

$$
\partial_{\boldsymbol{\theta}}\alpha_i(\boldsymbol{\theta}_0,F_i) \quad = \quad -\frac{\mathbb{E}(\partial_{\boldsymbol{\theta}}V_i^{(1)})}{\mathbb{E}(\partial_{\alpha_i}V_i^{(1)})} = O_p(1) \qquad (4.5.23)
$$

[11] The systems of equations can be found in the appendix of Hahn and Kuersteiner (2011). Refer to pages 1178 and 1181. Solving the system of equations is not as hard as it sounds because the coefficient matrix is diagonal.

$$\partial_\epsilon \alpha_i(\boldsymbol{\theta}_0, F_i) = -\frac{T^{1/2}\frac{1}{T}\left(V_i^{(1)} - \mathbb{E}(V_i^{(1)})\right)}{\mathbb{E}(\partial_{\alpha_i} V_i^{(1)})} = O_p(T^{-1}) \quad (4.5.24)$$

The resulting second derivatives would be

$$\begin{aligned}
&\partial^2_{\boldsymbol{\theta}} \alpha_i(\boldsymbol{\theta}_0, F_i) \\
= &-\frac{1}{\mathbb{E}(\partial_{\alpha_i} V_i^{(1)})}\left[\mathbb{E}(\partial^2_{\boldsymbol{\theta}} V_i^{(1)}) + \partial_{\boldsymbol{\theta}} \alpha_i(\boldsymbol{\theta}_0, F_i)\mathbb{E}(\partial^2_{\boldsymbol{\theta},\alpha_i} V_i^{(1)})^{\mathrm{T}}\right] \\
&-\frac{1}{\mathbb{E}(\partial_{\alpha_i} V_i^{(1)})}\left[\mathbb{E}(\partial^2_{\boldsymbol{\theta},\alpha_i} V_i^{(1)})(\partial_{\boldsymbol{\theta}} \alpha_i(\boldsymbol{\theta}_0, F_i))^{\mathrm{T}}\right] \\
&-\frac{1}{\mathbb{E}(\partial_{\alpha_i} V_i^{(1)})}\left[\mathbb{E}(\partial^2_{\alpha_i} V_i^{(1)})\partial_{\boldsymbol{\theta}} \alpha_i(\boldsymbol{\theta}_0, F_i)\partial_{\boldsymbol{\theta}} \alpha_i(\boldsymbol{\theta}_0, F_i)^{\mathrm{T}}\right] \\
= &\ O_p(1) \qquad (4.5.25) \\
&\partial^2_{\boldsymbol{\theta},\epsilon} \alpha_i(\boldsymbol{\theta}_0, F_i) \\
= &-\frac{1}{\mathbb{E}(\partial_{\alpha_i} V_i^{(1)})}\left[\mathbb{E}(\partial^2_{\boldsymbol{\theta},\alpha_i} V_i^{(1)})\partial_\epsilon \alpha_i(\boldsymbol{\theta}_0, F_i) + T^{1/2}\frac{1}{T}\left(\partial_{\boldsymbol{\theta}} V_i^{(1)} - \mathbb{E}(\partial_{\boldsymbol{\theta}} V_i^{(1)})\right)\right] \\
&-\frac{1}{\mathbb{E}(\partial_{\alpha_i} V_i^{(1)})}\left[\mathbb{E}(\partial^2_{\alpha_i} V_i^{(1)})\partial_{\boldsymbol{\theta}} \alpha_i(\boldsymbol{\theta}_0, F_i)\partial_\epsilon \alpha_i(\boldsymbol{\theta}_0, F_i)\right] \\
&-\frac{1}{\mathbb{E}(\partial_{\alpha_i} V_i^{(1)})}\left[T^{1/2}\frac{1}{T}\left(\partial_{\alpha_i} V_i^{(1)} - \mathbb{E}(\partial_{\alpha_i} V_i^{(1)})\right)\partial_{\boldsymbol{\theta}} \alpha_i(\boldsymbol{\theta}_0, F_i)\right] \\
= &\ O_p(T^{-1}) \qquad (4.5.26) \\
&\partial^2_{\epsilon} \alpha_i(\boldsymbol{\theta}_0, F_i) \\
= &-\frac{1}{\mathbb{E}(\partial_{\alpha_i} V_i^{(1)})}\left[\mathbb{E}(\partial^2_{\alpha_i} V_i^{(1)})(\partial_\epsilon \alpha_i(\boldsymbol{\theta}_0, F_i))^2\right] \\
&-\frac{1}{\mathbb{E}(\partial_{\alpha_i} V_i^{(1)})}\left[2T^{1/2}\frac{1}{T}\left(\partial_{\alpha_i} V_i^{(1)} - \mathbb{E}(\partial_{\alpha_i} V_i^{(1)})\right)\partial_\epsilon \alpha_i(\boldsymbol{\theta}_0, F_i)\right] \\
= &\ O_p(T^{-2}) \qquad (4.5.27)
\end{aligned}$$

Central limit theorems are applied to $\partial_{\alpha_i} V_i^{(1)}$ and $\partial_{\boldsymbol{\theta}} V_i^{(1)}$, so that the resulting order of magnitude calculations can be obtained.

6. We are now in a position to simplify $\boldsymbol{\theta}^\epsilon(0)$ and $\boldsymbol{\theta}^{\epsilon\epsilon}(0)$.

   (a) First, we find an expression for $\boldsymbol{\theta}^\epsilon(0)$. Calculate the expectation of every term in (4.5.21) at the true values. Note that $\boldsymbol{\theta}^\epsilon(0)$ do not depend on the data. Further note that (4.5.23) is already a constant while (4.5.24) depends on the data through $V_i^{(1)}$.[12] Second-order E-ancillarity implies

[12] A curious aspect of the proof in Hahn and Newey (2004) and Hahn and Kuersteiner (2011) is that they treat (4.5.24), (4.5.26), and (4.5.27) as constants yet they still depend on the data. We solve the

that $\mathbb{E}(\partial_{\alpha_i} U_{i,2}(\boldsymbol{\theta}_0, \alpha_{i0})) = 0$ and $\mathbb{E}(V_i^{(1)} \partial_{\alpha_i} U_{i,2}(\boldsymbol{\theta}_0, \alpha_{i0})) = 0$, as in (4.2.7). As a result, the first sum in (4.5.19) is given by

$$\frac{1}{n}\sum_{i=1}^{n} \int \frac{dh_i(0)}{d\epsilon}\, dF_i = \left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(\partial_{\boldsymbol{\theta}} U_{i,2}(\boldsymbol{\theta}_0, \alpha_{i0}))\right) \boldsymbol{\theta}^{\epsilon}(0) \tag{4.5.28}$$

The remaining term in (4.5.19) is given by

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n} \int h_i(\epsilon)\, d\Delta_{iT} \\ = \; & \frac{1}{n}\sum_{i=1}^{n} \int U_{i,2}(\boldsymbol{\theta}_0, \alpha_{i0})\, d\Delta_{iT} \\ = \; & \frac{\sqrt{T}}{n}\sum_{i=1}^{n} \frac{1}{T}\left(U_{i,2}(\boldsymbol{\theta}_0, \alpha_{i0}) - \mathbb{E}(U_{i,2}(\boldsymbol{\theta}_0, \alpha_{i0}))\right) \end{aligned} \tag{4.5.29}$$

Define $\mathbf{I}_i$ as follows, provided integration and differentiation can be interchanged:

$$\mathbf{I}_i = \mathbb{E}[(U_{i,2}(\boldsymbol{\theta}_0, \alpha_{i0}))(U_{i,2}(\boldsymbol{\theta}_0, \alpha_{i0}))^{\mathrm{T}}] = \mathbb{E}(\partial_{\boldsymbol{\theta}} U_{i,2}(\boldsymbol{\theta}_0, \alpha_{i0})) \tag{4.5.30}$$

Thus, we have the following expression for $\boldsymbol{\theta}^{\epsilon}(0)$, whose asymptotic distribution we seek:

$$\boldsymbol{\theta}^{\epsilon}(0) = \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{I}_i\right)^{-1} \frac{\sqrt{T}}{n}\sum_{i=1}^{n} \frac{1}{T}\left(U_{i,2}(\boldsymbol{\theta}_0, \alpha_{i0}) - \mathbb{E}(U_{i,2}(\boldsymbol{\theta}_0, \alpha_{i0}))\right) \tag{4.5.31}$$

Assume that a central limit theorem holds for $U_{i,2}(\boldsymbol{\theta}_0, \alpha_{i0})$, i.e.

$$\sqrt{nT}\left(\frac{1}{nT}\sum_{i=1}^{n} U_{i,2}(\boldsymbol{\theta}_0, \alpha_{i0})\right) \xrightarrow{d} N\left(0, \lim_{n,T\to\infty} \frac{1}{nT}\sum_{i=1}^{n} \mathbf{I}_i\right) \tag{4.5.32}$$

As a consequence, we have

$$\begin{aligned} \sqrt{nT}\frac{1}{\sqrt{T}}\boldsymbol{\theta}^{\epsilon}(0) &= \left(\frac{1}{nT}\sum_{i=1}^{n} \mathbf{I}_i\right)^{-1} \left(\frac{1}{\sqrt{nT}}\sum_{i=1}^{n} U_{i,2}(\boldsymbol{\theta}_0, \alpha_{i0})\right) \\ &\xrightarrow{d} N\left(\mathbf{0}, \left(\lim_{n,T\to\infty} \frac{1}{nT}\sum_{i=1}^{n} \mathbf{I}_i\right)^{-1}\right) \end{aligned} \tag{4.5.33}$$

Therefore,

$$\boldsymbol{\theta}^{\epsilon}(0) = O_p(n^{-1/2}) \tag{4.5.34}$$

system of equations mentioned in Step 5 and make the order of magnitude calculations explicit to take into account the latter fact.

(b) Next we find an expression for $\boldsymbol{\theta}^{\epsilon\epsilon}(0)$. Calculate the expectation of every term in (4.5.22) at the true values while noting the orders of magnitude in (4.5.23), (4.5.24), (4.5.25), (4.5.26), (4.5.27), and (4.5.34). The boxed, double-boxed, oval-boxed and unboxed terms in (4.5.22) are $O_p(n^{-1/2})$, $O_p(T^{-1})$, $\mathbf{0}$, and $O_p(Tn^{-1})$ respectively. The first sum in (4.5.20) can now be written as

$$\frac{1}{n}\sum_{i=1}^{n}\int \frac{d^2h_i(\epsilon)}{d\epsilon^2}\, dF_i(\epsilon) = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{I}_i\right)\boldsymbol{\theta}^{\epsilon\epsilon}(0) + O_p\left(\frac{T}{n}\right) + O_p\left(\frac{1}{T}\right) + O_p\left(\frac{1}{\sqrt{n}}\right) \tag{4.5.35}$$

After applying central limit theorems for $\partial_{\boldsymbol{\theta}} U_{i,2}(\boldsymbol{\theta}_0, \alpha_{i0})$ and $\partial_{\alpha_i} U_{i,2}(\boldsymbol{\theta}_0, \alpha_{i0})$ and noting the order of magnitude calculations in (4.5.23), (4.5.24), and (4.5.34), the "tail" term in (4.5.20) can now be written as

$$\frac{2}{n}\sum_{i=1}^{n}\int \frac{dh_i(\epsilon)}{d\epsilon}\, d\Delta_{iT} = O_p\left(\frac{1}{\sqrt{n}}\right) + O_p\left(\frac{1}{T}\right) \tag{4.5.36}$$

As a consequence, we have

$$\begin{aligned} & \sqrt{nT}\left(\frac{1}{\sqrt{T}}\right)^2 \boldsymbol{\theta}^{\epsilon\epsilon}(0) \\ = \; & \left(\frac{1}{nT}\sum_{i=1}^{n}\mathbf{I}_i\right)^{-1}\left[O_p\left(\frac{1}{\sqrt{nT}}\right) + O_p\left(\frac{1}{\sqrt{T^3}}\right) + O_p\left(\sqrt{\frac{n}{T^5}}\right)\right] \\ & + \left(\frac{1}{nT}\sum_{i=1}^{n}\mathbf{I}_i\right)^{-1}\left[O_p\left(\frac{1}{\sqrt{nT^2}}\right) + O_p\left(\frac{1}{T^2}\right)\right] \end{aligned}$$

Under the conditions that $n, T \to \infty$ and $n/T \to c \in (0, \infty)$, the distribution of $\boldsymbol{\theta}^{\epsilon\epsilon}(0)$ becomes degenerate at $\mathbf{0}$.

(c) The last term in the Taylor series expansion (4.5.14) can be shown to be $o_p(1)$. This step mimics the derivation in Hahn and Kuersteiner (2011).

Projected score for the AR(1) linear dynamic panel data model

The model specification is as follows:

$$\begin{aligned} Y_{i,t-1} &= \{y_{i0}, y_{i1}, \ldots, y_{i,t-1}\}, \\ y_{it}|Y_{i,t-1} &\sim \text{iid}\, N(\alpha_i + \rho y_{i,t-1}, \sigma^2),\ i = 1, \ldots, n; t = 1 \ldots, T \end{aligned} \tag{4.5.37}$$

Assume y_{i0} is available and we calculate expectations conditional on y_{i0} (so that $\mathbb{E}[\cdot]$ is the expectation of some expression conditional on y_{i0}). Let $u_{it} = y_{it} - \alpha_i - \rho y_{i,t-1}$. The scores for the common parameters ρ and σ^2 and the incidental parameter α_i

are given by:

$$\begin{aligned} U_{i,0}^{\rho} &= \frac{1}{\sigma^2}\sum_{t=1}^{T} u_{it} y_{i,t-1}, \\ U_{i,0}^{\sigma^2} &= -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{t=1}^{T} u_{it}^2, \\ V_i^{(1)} &= \frac{1}{\sigma^2}\sum_{t=1}^{T} u_{it}. \end{aligned}$$

To calculate the second-order projected score, we need the following elements:

$$\begin{aligned} \mathbb{E}\left[(V_i^{(1)})^2\right] &= \frac{1}{\sigma^4} E\left[\sum_{t=1}^{T} u_{it}^2 + 2\sum_{t=2}\sum_{s<t} u_{is}u_{it}\right] \\ &= \frac{1}{\sigma^4}\left[\sum_{t=1}^{T}\sigma^2 + 2\sum_{t=2}\sum_{s<t}\mathbb{E}(u_{it}u_{is})\right] \\ &= \frac{T}{\sigma^2}. \end{aligned}$$

Note that only the second moments of (4.5.37) are used for the above calculation

$$\begin{aligned} \mathbb{E}\left(V_i^{(1)}V_i^{(2)}\right) &= -\frac{T}{\sigma^4}\sum_{t=1}^{T}\mathbb{E}(u_{it}) + \frac{1}{\sigma^6}\mathbb{E}\left(\sum_{t=1}^{T} u_{it}\right)^3 \\ &= \frac{1}{\sigma^6}\mathbb{E}\left[\left(\sum_{t=1}^{T} u_{it}\right)\left(\sum_{t=1}^{T} u_{it}^2 + 2\sum_{t=2}^{T}\sum_{s<t} u_{is}u_{it}\right)\right] \\ &= \frac{1}{\sigma^6}\left[\sum_{t=1}^{T}\mathbb{E}\left(u_{it}^3\right) + \sum_{t=1}^{T}\sum_{s\neq t}\mathbb{E}\left(u_{it}^2 u_{is}\right) + 2\sum_{r=1}^{T}\sum_{t=2}^{T}\sum_{s<t} u_{is}u_{it}u_{ir}\right] \\ &= 0. \end{aligned}$$

Thus, $V_i^{(1)}$ and $V_i^{(2)}$ are orthogonal. Note that we used the third moments of (4.5.37) for the preceding calculation

$$\begin{aligned} \mathbb{E}\left((V_i^{(2)})^2\right) &= \frac{T^2}{\sigma^4} - \frac{2T}{\sigma^6}\mathbb{E}\left(\sum_{t=1}^{T} u_{it}\right)^2 + \frac{1}{\sigma^8}\mathbb{E}\left(\sum_{t=1}^{T} u_{it}^2 + 2\sum_{t=2}^{T}\sum_{s<t} u_{is}u_{it}\right)^2 \\ &= \frac{T^2}{\sigma^4} - \frac{2T^2}{\sigma^4} + \frac{1}{\sigma^8}\mathbb{E}\left[\sum_{t=1}^{T} u_{it}^4 + 2\sum_{t=2}^{T}\sum_{s<t} u_{it}^2 u_{is}^2 + 4\sum_{t=2}^{T}\sum_{s<t}\left(\sum_{r=2}^{T}\sum_{q<r} u_{iq}u_{ir}\right)u_{is}u_{it}\right] \\ &= -\frac{T^2}{\sigma^4} + \frac{1}{\sigma^4}[3T + T(T-1) + 2(T-1)(T)] \\ &= \frac{2T^2}{\sigma^4}. \end{aligned}$$

We have used fourth moments of (4.5.37) for the preceding calculation

$$\begin{aligned}
\mathbb{E}\left(U_{i,0}^{\rho}V_i\right) &= \frac{1}{\sigma^4}\mathbb{E}\left[\left(\sum_{t=1}^{T}u_{it}y_{i,t-1}\right)\left(\sum_{t=1}^{T}u_{it}\right)\right] \\
&= \frac{1}{\sigma^4}\left[\sum_{t=1}^{T}\mathbb{E}\left(u_{it}^2y_{i,t-1}\right)+\sum_{t=2}^{T}\sum_{s<t}\mathbb{E}\left(u_{it}u_{is}y_{i,s-1}\right)+\sum_{t=2}^{T}\sum_{s<t}\mathbb{E}\left(u_{it}u_{is}y_{i,t-1}\right)\right] \\
&= \frac{1}{\sigma^2}\sum_{t=1}^{T}\mathbb{E}\left(y_{i,t-1}\right) \\
&= \frac{1}{\sigma^2}\left[\left(1+\rho+\cdots+\rho^{T-1}\right)y_{i0}+(T-t)\rho^{t-1}\alpha_i\right].
\end{aligned}$$

The last line follows from recursive substitution. Alternatively, we can impose mean stationarity. Note that

$$\begin{aligned}
\mathbb{E}\left(U_{i,0}^{\sigma^2}V_i\right) &= -\frac{T}{2\sigma^4}\sum_{t=1}^{T}\mathbb{E}(u_{it})+\frac{1}{2\sigma^6}\mathbb{E}\left[\left(\sum_{t=1}^{T}u_{it}^2\right)\left(\sum_{t=1}^{T}u_{it}\right)\right] \\
&= \frac{1}{2\sigma^6}\mathbb{E}\left[\sum_{s=1}^{T}\sum_{t=1}^{T}u_{it}^2u_{is}\right] \\
&= \frac{1}{2\sigma^6}\left[\sum_{t=1}^{T}\mathbb{E}\left(u_{it}^3\right)+\sum_{t=2}^{T}\sum_{s<t}\mathbb{E}\left(u_{it}^2u_{is}\right)+\sum_{t=2}^{T}\sum_{s<t}\mathbb{E}\left(u_{is}^2u_{it}\right)\right] \\
&= 0,
\end{aligned}$$

$$\begin{aligned}
E(U_{i,0}^{\sigma^2}V_i^{(2)}) &= \frac{T^2}{2\sigma^4}-\frac{T}{2\sigma^6}E\left[\sum_{t=1}^{T}u_{it}\right]^2+\frac{1}{2\sigma^8}E\left[\left(\sum_{t=1}^{T}u_{it}^2\right)\left(\sum_{t=1}^{T}u_{it}^2+2\sum_{t=2}^{T}\sum_{s<t}u_{it}u_{is}\right)\right] \\
&\quad -\frac{T}{2\sigma^6}E\left[\sum_{t=1}^{T}u_{it}\right]^2 \\
&= \frac{T^2}{2\sigma^4}-\frac{T^2}{2\sigma^4}+\frac{1}{2\sigma^8}T(3\sigma^4+(T-1)\sigma^4))-\frac{T^2}{2\sigma^4} \\
&= \frac{T}{\sigma^4},
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left(U_{i,0}^{\rho}V_i^{(2)}\right) &= -\frac{T}{\sigma^4}\mathbb{E}\left[\sum_{t=1}^{T}u_{it}y_{i,t-1}\right]+\frac{1}{\sigma^6}\mathbb{E}\left[\left(\sum_{t=1}^{T}u_{it}y_{i,t-1}\right)\left(\sum_{t=1}^{T}u_{it}^2+2\sum_{t=2}^{T}\sum_{s<t}u_{it}u_{is}\right)\right] \\
&= \frac{1}{\sigma^6}\mathbb{E}\left[\sum_{t=1}^{T}u_{it}^3y_{i,t-1}+\sum_{t=1}^{T}\sum_{s\neq t}u_{is}^2u_{it}y_{i,t-1}+2\sum_{t=2}^{T}\sum_{s=t-1}u_{is}u_{it}^2y_{is}+2\sum_{t=2}^{T}\sum_{s\neq t-1}u_{is}u_{it}^2y_{is}\right] \\
&= \frac{2}{\sigma^6}\mathbb{E}\left[\sum_{t=2}^{T}\sum_{s=t-1}u_{is}u_{it}^2y_{is}\right]
\end{aligned}$$

$$= \frac{2}{\sigma^2}(T-t)\rho^{t-1}.$$

Thus, the second-order projected score for an arbitrary value of T can be computed in a straightforward manner using all the components cited above.

Projected score for the static binary choice model with an exogenous regressor

Suppose $y_t|x_1, x_2 \sim Ber(p_t)$ with

$$p_t = \mathbb{E}(y_t|x_1, x_2) = F(\alpha + x_t^{\mathrm{T}}\beta) \equiv F_t$$

for $i = 1, \ldots, n$ and $t = 1, 2$. The uncentered moments of this conditional distribution are all equal to F_t. For this discussion, I suppress the dependence of the expression on i. Calculations in a separate *Mathematica* file give the following analytical results specific to the static logit model with one exogenous regressor for $T = 2$. Let

$$\begin{aligned} D_1 &= \left(e^{\alpha+\beta x_1}+1\right)\left(e^{\alpha+\beta x_2}+1\right), \\ D_2 &= 4e^{\alpha+\beta x_1+\beta x_2} + e^{2\alpha+2\beta x_1+\beta x_2} + e^{2\alpha+\beta x_1+2\beta x_2} + e^{\beta x_1} + e^{\beta x_2}. \end{aligned}$$

The scores for β and α are given by

$$\begin{aligned} U_0 &= \frac{x_1\left(e^{\alpha+\beta x_2}+1\right)\left(y_1\left(e^{\alpha+\beta x_1}+1\right)-e^{\alpha+\beta x_1}\right) + x_2\left(e^{\alpha+\beta x_1}+1\right)\left(y_2\left(e^{\alpha+\beta x_2}+1\right)-e^{\alpha+\beta x_2}\right)}{D_1} \\ V^{(1)} &= \frac{-e^{\alpha}\left(2e^{\alpha+\beta x_1+\beta x_2}+e^{\beta x_1}+e^{\beta x_2}\right)+(y_1+y_2)D_1}{D_1} \end{aligned}$$

The components of the second-order projected score are calculated below. First,

$$\mathbb{E}\left(V^{(1)}\right)^2 = \frac{e^{\alpha}D_2}{D_1^2}$$

$$\begin{aligned} \mathbb{E}\left(V^{(1)}V^{(2)}\right) &= \frac{e^{\alpha}\left(e^{3\alpha+3\beta x_1+\beta x_2} - e^{4\alpha+3\beta x_1+2\beta x_2} + e^{3\alpha+\beta x_1+3\beta x_2} - e^{4\alpha+2\beta x_1+3\beta x_2}\right)}{D_1^3} \\ &+ \frac{e^{\alpha}\left(-e^{\alpha+2\beta x_1} - e^{\alpha+2\beta x_2} + e^{\beta x_1} + e^{\beta x_2}\right)}{D_1^3} \\ &+ \frac{e^{\alpha}\left(6e^{\alpha+\beta x_1+\beta x_2} - 6e^{3\alpha+2\beta x_1+2\beta x_2}\right)}{D_1^3} \end{aligned}$$

$$\mathbb{E}\left(U_0 V^{(1)}\right) = \frac{e^{\alpha}\left(x_2 e^{\beta x_2}\left(e^{\alpha+\beta x_1}+1\right)^2 + x_1 e^{\beta x_1}\left(e^{\alpha+\beta x_2}+1\right)^2\right)}{D_1^2}$$

Next, $V^{(2)}$ is given by

$$\begin{aligned}
V^{(2)}\Big|_{y_1=0,y_2=0} &= \frac{e^{\alpha}\left(e^{\alpha+2\beta x_1} - 2e^{\alpha+\beta x_1+\beta x_2} + 3e^{2\alpha+2\beta x_1+\beta x_2} + e^{\alpha+2\beta x_2}\right)}{D_1^2} \\
&\quad + \frac{e^{\alpha}\left(3e^{2\alpha+\beta x_1+2\beta x_2} + 4e^{3\alpha+2\beta x_1+2\beta x_2} - e^{\beta x_1} - e^{\beta x_2}\right)}{D_1^2} \\
V^{(2)}\Big|_{y_1=1,y_2=1} &= \frac{3e^{\alpha+\beta x_1} + e^{2(\alpha+\beta x_1)} + 3e^{\alpha+\beta x_2} + e^{2(\alpha+\beta x_2)}}{D_1^2} \\
&\quad + \frac{4 - 2e^{2\alpha+\beta x_1+\beta x_2} - e^{3\alpha+2\beta x_1+\beta x_2} - e^{3\alpha+\beta x_1+2\beta x_2}}{D_1^2} \\
V^{(2)}\Big|_{y_1=0,y_2=1} &= \frac{e^{\alpha}\left(e^{\alpha+2\beta x_1} - 2e^{\alpha+\beta x_1+\beta x_2} + 3e^{2\alpha+2\beta x_1+\beta x_2} + e^{\alpha+2\beta x_2}\right)}{D_1^2} \\
&\quad + \frac{e^{\alpha}\left(3e^{2\alpha+\beta x_1+2\beta x_2} + 4e^{3\alpha+2\beta x_1+2\beta x_2} - e^{\beta x_1} - e^{\beta x_2}\right)}{D_1^2} \\
V^{(2)}\Big|_{y_1=1,y_2=0} &= \frac{e^{\alpha}\left(e^{\alpha+2\beta x_1} - 2e^{\alpha+\beta x_1+\beta x_2} + 3e^{2\alpha+2\beta x_1+\beta x_2} + e^{\alpha+2\beta x_2}\right)}{D_1^2} \\
&\quad + \frac{e^{\alpha}\left(3e^{2\alpha+\beta x_1+2\beta x_2} + 4e^{3\alpha+2\beta x_1+2\beta x_2} - e^{\beta x_1} - e^{\beta x_2}\right)}{D_1^2}
\end{aligned}$$

To orthogonalize $V^{(2)}$, we compute

$$V^{(2),*} = V^{(2)} - \frac{E(V^{(1)}V^{(2)})}{E[(V^{(1)})^2]} V^{(1)}.$$

Depending on the binary patterns of the sequence (y_1, y_2), we have

$$\begin{aligned}
V^{(2),*}\Big|_{y_1=0,y_2=0} &= \frac{2\left(e^{\beta x_1} + e^{\beta x_2}\right)e^{2\alpha+\beta x_1+\beta x_2}}{D_2}, \\
V^{(2),*}\Big|_{y_1=1,y_2=1} &= \frac{2\left(e^{\beta x_1} + e^{\beta x_2}\right)}{D_2}, \\
V^{(2),*}\Big|_{y_1=0,y_2=1} &= -\frac{4e^{\alpha+\beta x_1+\beta x_2}}{D_2}, \\
V^{(2),*}\Big|_{y_1=1,y_2=0} &= -\frac{4e^{\alpha+\beta x_1+\beta x_2}}{D_2}.
\end{aligned}$$

Hence, we have

$$\mathbb{E}\left(U_0 V^{(2),*}\right) = \frac{2(x_1 - x_2)\left(e^{\beta x_2} - e^{\beta x_1}\right) e^{2\alpha+\beta x_1+\beta x_2}}{D_2},$$

$$\mathbb{E}\left(V^{(2),*}\right)^2 = \frac{4\left(e^{\beta x_1} + e^{\beta x_2}\right) e^{2\alpha+\beta x_1+\beta x_2}}{D_1 D_2}.$$

The second-order projected score can now be written as

$$\begin{aligned} U_2 = & \frac{(x_1 - x_2)\left(y_1^2\left(e^{\beta x_1} - e^{\beta x_2}\right) + y_1\left(-e^{\beta x_1} + 3e^{\beta x_2} + 2y_2\left(e^{\beta x_1} - e^{\beta x_2}\right)\right)\right)}{2\left(e^{\beta x_1} + e^{\beta x_2}\right)} \\ & \frac{+(x_1 - x_2)\, y_2\left(-3e^{\beta x_1} + e^{\beta x_2} + y_2\left(e^{\beta x_1} - e^{\beta x_2}\right)\right)}{2\left(e^{\beta x_1} + e^{\beta x_2}\right)} \end{aligned}$$

Note that we have $U_2 = 0$ over cross-sectional units for which $y_1 + y_2 = 0$ or $y_1 + y_2 = 2$. For cross-sectional units for which $y_1 + y_2 = 1$, i.e., substituting in $y_1 = 1 - y_2$ in the expression for U_2, gives the expression one sees in (4.2.15).

Chapter 5

The role of sparsity in panel data models

5.1 Introduction

We have seen increased collection of longitudinal or panel data through active or passive means in recent years. We can study these repeated measurements in three ways – (a) analyze the repeated measurements for each cross-sectional unit separately, (b) analyze the cross-sectional information, and (c) pool the both cross-sectional and time-series information together.

Methods in time series analysis can be used in situation (a) but will only be feasible when the number of repeated measurements is sufficiently large. The latter case precludes studying panels with a short time series dimension, typically collected for purposes of crafting policy. Methods in cross-sectional analysis can be used in situation (b) but precludes the study of the dynamics of change unless the time series dimension is also large. A compromise would then be to use methods that accomplish (c).

Unfortunately, there is much leeway as to how we should pool information available in panel data. Traditionally, econometricians have introduced cross-sectional heterogeneity in the parameters of a panel data model. Research during the 1960s up to the 1980s, cross-sectional heterogeneity is usually accomplished via the variance components model and the random coefficients model. These models typically impose parametric assumptions on the distribution of heterogeneity so that the dimension of the parameter space can be reduced substantially. Recent research has been aimed at completely removing these parametric assumptions. Success in this area has been mixed but a lot of progress has been made.

In particular, recent results have been negative with respect to fixed-T identification and fixed-T consistent estimation (see the most recent survey by Arellano and Bonhomme (2011)). However, a major insight behind recent results is the need for reducing the support of the fixed effects relative to the support of the dependent variable. Bonhomme (2012) show how this reduction in the support aids in constructing moment conditions for the structural parameters. Despite these negative results, Browning and Carro (2010) argue that we have actually not allowed for full heterogeneity at all. In particular, they argue for a fully heterogeneous setup where slope coefficients are allowed to vary across observations but still be time-invariant. Another way to interpret heterogeneity is to allow for time-invariant heterogeneity in the inverse link functions (and not the coefficients of the linear predictor) for single-index panel data models as proposed by Chen, Gao, and Li (2013).

Notice that the previous descriptions of heterogeneity assume that cross-sectional units are totally different from one another. At the other extreme, all cross-sectional units are assumed to be the same (with respect to model parameters). There is a large middle ground that needs to be explored. Grouping and clustering methods come to mind because they allow the data to determine which units can be pooled and which cannot. Furthermore, partial pooling allows for a possibility to implement the reduction in the support of the fixed effects. Recent research on grouped heterogeneity by Bonhomme and Manresa (2015) point toward this possibility. They even allow the grouping to vary over time. Yet another way to implement partial pooling is proposed by Sarafidis and Weber (2011) where they allow for an unknown number of clusters in the data and full homogeneity is assumed within each cluster.

In this chapter, I argue that sparsity may be a useful device to accomplish a reduction in the support of the fixed effects and to allow the data to determine the groups that may be present in the data. In particular, there are economic and empirical situations for which some cross-sectional units can have the same value for the individual-specific effect. For instance, an econometric method should be able to accommodate the situation where only a subset of units obey conditional moment restrictions implied by an economic model. This is where we must account for partial pooling and where a sparsity assumption on the individual-specific effects can be a useful technical device. Furthermore, it is of interest to try to identify these deviations in the same manner in which we want to be able to detect outliers to obtain some form of robustness.

Recent work by Fan, Tang, and Shi (2012) indicate that it is possible to estimate the structural parameter of a linear model with exogenous covariates with just $T = 1$ despite allowing for the intercept to vary across observations. Their idea was to divide the incidental parameters into three types – those that are very large that they can be treated as outliers, those that are zero, and those that are non-zero but small enough that they can be treated as zero asymptotically. I show how to extend their arguments to the linear panel data case but allowing for contemporaneously

exogenous variables. I also modify their procedure for selecting the data-driven regularization parameter. Unfortunately, not all the results in Fan, Tang, and Shi (2012) survive the extension as we will see in the next section.

Although sparsity has been used in machine learning and big data situations, the focus has always been settings where the number of covariates is extremely large relative to the sample size. I restrict myself to the setting where the regressor vector is still finite-dimensional. In contrast, Kock (2013) differences out the incidental parameters first before proposing a penalty method for the differenced model and allows the regressor vector to be high-dimensional. Kock (2014) extends the previous paper to allow the possibility that the incidental parameters are weakly sparse. In his context, weak sparsity means that the L_1 norm of all the incidental parameters is small. As a result, the values of the incidental parameters need not be zero at all. In contrast, I explicitly have zero-valued incidental parameters but allow for some of these parameters to be small enough that they can be taken as zero asymptotically. Furthermore, these two papers by Kock are confined to regressors that are strictly exogenous. Kock and Tang (2014) extends these papers further to dynamic panels and allow for predetermined regressors. All these developments are under a framework where n and T are allowed to vary. Furthermore, their results are in the form of oracle inequalities. These inequalities provide upper bounds for the estimation error (in some suitable norm) as a function of the design matrix and the dimensions of the problem.

In contrast, my modifications to Fan, Tang, and Shi (2012) for the panel data case allow me to consider contemporaneously exogenous regressors and a fixed number of time periods T. I introduce these modifications and the resulting consequences in Section 5.2. I use Monte Carlo simulations to study the finite sample performance of the two-step panel lasso estimator in Section 5.3. I revisit the relationship between inequality and growth using the microdata collected by van der Weide and Milanovic (2014). I end with some concluding remarks, suggestions for future research, and a technical appendix containing some proofs of the main results.

5.2 Panel lasso for the linear model

5.2.1 Setup and notation

Consider the data generating process where

$$y_{it} = \alpha_{i0} + \mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}_0 + \epsilon_{it}, \;\; i = 1, \ldots, n; \;\; t = 1, \ldots, T \tag{5.2.1}$$

where $\left(\alpha_{10}, \alpha_{20}, \ldots, \alpha_{N0}, \boldsymbol{\beta}_0\right)$ are the true values of the parameters and $\boldsymbol{\beta}_0 \in \mathbb{R}^d$. In contrast to the machine learning literature, I assume that d is fixed and does not grow with sample size. Define the averages $\bar{\mathbf{x}}_i = T^{-1}\sum_{t=1}^{T}\mathbf{x}_{it}$ and $\bar{\epsilon}_i = T^{-1}\sum_{t=1}^{T}\epsilon_{it}$.

Let $a_+ = \max\{a, 0\}$ be the positive part of a, $\text{sgn}(a)$ be the sign function, and $\|\cdot\|_2$ be the L_2-norm. Let $B_C(\boldsymbol{\beta}_0) = \{\boldsymbol{\beta} \in \mathbb{R}^d : |\beta_j - \beta_{j0}| \leq C,\ 1 \leq j \leq d\}$ for some constant $C > 0$. Finally, $a_n \ll b_n$ is shorthand for a_n is of smaller order than b_n and $a_n \gg b_n$ is shorthand for a_n is of larger order than b_n.

I impose the following assumptions:

A1 (Independence) The errors ϵ_{it} are independent across i.

A2-1 (Predeterminedness) The errors ϵ_{it} and the covariates $\mathbf{x}_i^t = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{it})$ satisfy $\mathbb{E}(\epsilon_{it} \mid \mathbf{x}_i^t) = 0$ for all i and t.

A2-2 (Contemporaneous exogeneity) The errors ϵ_{it} and the covariates $\mathbf{x}_{it}$ satisfy $\mathbb{E}(\epsilon_{it} \mid \mathbf{x}_{it}) = 0$ for all i and t.

A3 (Behavior of averages) Assume that $\mathbb{E}\left(\|\bar{\mathbf{x}}_i\|_2\right) < \infty$ and $\mathbb{E}(\bar{\epsilon}_i) < \infty$. There exists $\kappa_n, \gamma_n \ll \sqrt{n}$ such that, as $n \to \infty$, we have

$$\Pr\left(\max_{1 \leq i \leq n} \|\bar{\mathbf{x}}_i\|_2 > \kappa_n\right) \to 0, \tag{5.2.2}$$

$$\Pr\left(\max_{1 \leq i \leq n} |\bar{\epsilon}_i| > \gamma_n\right) \to 0. \tag{5.2.3}$$

A4 (Sparsity) Each cross-sectional unit i belongs to one and only one of the three possible index sets $\{1, \ldots, s_1\}$, $\{s_1 + 1, \ldots, s\}$, and $\{s + 1, \ldots, n\}$. If $i \in \{1, \ldots, s_1\}$, then $\min_{1 \leq i \leq s_1} |\alpha_{i0}| \gg \max\{\kappa_n, \gamma_n\}$. If $i \in \{s_1 + 1, \ldots, s\}$, then $\max_{1 \leq i \leq n} |\alpha_{i0}| < \gamma_n$. If $i \in \{s + 1, \ldots, n\}$, then $\alpha_{i0} = 0$.

Assumption A1 is a standard assumption imposed in panel data models without cross-sectional dependence. Assumptions A2-1 or A2-2 allow us to consider dynamics or feedback effects.[1] Implementations of GMM estimators for panel data models usually maintain Assumption A2-1 (see Bun and Sarafidis (2015) for a survey). Assumption A2-2 is usually imposed in pooled OLS (see Wooldridge (2010)). Fan, Tang, and Shi (2012) impose assumptions on the behavior of the covariates and the errors similar to A3. The difference is that we impose tail behavior assumptions on the time series averages for every i rather than on the individual values. The existence of $\kappa_n, \gamma_n \ll \sqrt{n}$ is guaranteed by A1 and Markov's inequality.

Finally, there are three types of incidental parameters by A4 – s_1 of them are "large" incidental parameters, $s - s_1$ of them are bounded, and $n - s$ of them are zero. Note that A4 imposes an assumption on the number and the size of the incidental parameters. Furthermore, the number of each type of incidental parameter is unknown. With respect to the size of the incidental parameters,

[1] It is a priori unclear how changing the lasso penalty to other convex or concave penalty functions will affect the main results.

1. Cross-sectional units that belong to the index set $\{1,\ldots,s_1\}$ have a "large" value for α_{i0} in the sense that the tail behavior of both the time series averages of the regressors and the errors are dominated.

2. Cross-sectional units that belong to the index set $i \in \{s_1+1,\ldots,s\}$ have a value for α_{i0} that is bounded by the tail behavior of the time series average of the errors.

3. Cross-sectional units that belong to the index set $i \in \{s+1,\ldots,n\}$ have a zero value for the incidental parameter.

If we can detect which of the cross-sectional units are zero, then these units can now be pooled together to recover a consistent estimator for $\boldsymbol{\beta}_0$. Unfortunately, the panel lasso will have problems distinguishing whether a cross-sectional unit that is not a member of $\{1,\ldots,s_1\}$ will be classified as bounded or zero. As a result, the panel lasso under the assumptions laid out will not have the oracle property, i.e. the panel lasso cannot perform as good as an oracle who knows which of the cross-sectional units have $\alpha_{i0}=0$. Nevertheless, the panel lasso produces a fixed-T consistent estimator because it shrinks bounded incidental parameters to zero and this shrinkage has an asymptotically negligible effect.

Note that under large-n asymptotics, the number of each type of incidental parameter may grow with n. We will see later how the growth in the number of each type of incidental parameter has to be restricted so that consistency and asymptotic normality would be obtained. In addition, the size of each type of incidental parameter may depend on n (at least for the bounded and "large" incidental parameters) as seen in assumption A3 and A4.

Under what circumstances would it be plausible for assumption A4 to hold? Consider the following linear model where

$$y_{it} = \mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}_0 + \omega_{it}, \quad i=1,\ldots,n;\ t=1,\ldots,T. \tag{5.2.4}$$

Decompose ω_{it} into $\mathbb{E}(\omega_{it}|\mathbf{x}_{it})$ and its residual $\omega_{it}-\mathbb{E}(\omega_{it}|\mathbf{x}_{it})$. Let $\alpha_{i0}=\mathbb{E}(\omega_{it}|\mathbf{x}_{it})$ be the portion of the error ω_{it} representing some model deficiency specific to the ith unit that is correlated with the included regressors. Let the residual $\omega_{it}-\mathbb{E}(\omega_{it}|\mathbf{x}_{it})$ be equal to ϵ_{it}. We have now produced (5.2.1) that can potentially satisfy the assumptions laid out above from (5.2.4). Therefore, the units for which $\alpha_{i0}=0$ can represent the units for which the conditional moment restriction $\mathbb{E}(y_{it}|\mathbf{x}_{it})=\mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}_0$ is appropriate. The units for which α_{i0} are close enough to zero may be treated as zero asymptotically using the proposed panel lasso estimator. It then becomes important to detect the units for which there is some serious model deficiency.

The described setting may also apply to situations where we have endogenous regressors but are unable to find valid instruments. Think of α_{i0} as the unit-specific correlation between the error ω_{it} and $\mathbf{x}_{it}$. It is possible that only a subset of the units

have a regressor vector that is endogenous. It is therefore of interest to detect these units so that we are still able to consistently estimate $\boldsymbol{\beta}_0$ after removing these units from the sample.

We can interpret Assumption A4 as allowing for 3 groups in the cross-sectional dimension. This is neither more general or less general than grouping cross-sectional units in advance. Allowing for more than 3 groups in Assumption A4 is possible but the usefulness is unclear. The main results later on suggest that we are only able to detect "large" incidental parameters and not the ones that are bounded. Further refining the partitioning of incidental parameters will necessitate more tuning parameters and would only obscure the main results.

5.2.2 Estimation and inference

To develop an estimator for $\boldsymbol{\beta}_0$, consider minimizing the least squares objective function subject to an L_1-penalty[2] on the incidental parameters, i.e.

$$\min_{(\alpha_1,\alpha_2,\ldots,\alpha_n,\boldsymbol{\beta})} \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(y_{it}-\alpha_i-\mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}\right)^2+\sum_{i=1}^{n}2\lambda|\alpha_i|, \tag{5.2.5}$$

where $\lambda \geq 0$ is some user-specified regularization parameter.[3] This parameter takes on nonnegative values and governs the rate at which shrinkage toward zero is being applied to each of the α_i. Large values of λ will tend to shrink the α_i's toward zero. Therefore, the minimizer of (5.2.5) is the pooled OLS estimator when $\lambda \to \infty$. On the other hand, we obtain the within estimator when $\lambda \to 0$.

A minimizer $(\alpha_1, \alpha_2, \ldots, \alpha_n, \boldsymbol{\beta})$ of (5.2.5) satisfies the following first-order conditions:[4]

$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(y_{it}-\alpha_i-\mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}\right)\mathbf{x}_{it} = 0, \tag{5.2.6}$$

$$\frac{1}{T}\sum_{t=1}^{T}\left(y_{it}-\alpha_i-\mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}\right)-\lambda\frac{\alpha_i}{|\alpha_i|} = 0. \tag{5.2.7}$$

[2] Imposing an L_2-penalty leads to ridge regression. I do not use this penalty because I am working with Assumption A4. The L_2-penalty only shrinks estimators toward zero.

[3] An intercept should be included in the model. The data generating process considered in the Monte Carlo simulations sets the intercept to zero.

[4] To get the derivative of the absolute value function $|\alpha|$, note that $|\alpha| = \sqrt{\alpha^2}$. So, $\partial_\alpha |\alpha| = \partial_\alpha \sqrt{\alpha^2} = \left(\alpha^2\right)^{-1/2} \times 2\alpha/2 = \alpha/|\alpha|$ provided that $\alpha \neq 0$.

For an arbitrary $\boldsymbol{\beta}$, we can solve for α_i from (5.2.7) as[5]

$$\begin{cases} \frac{1}{T}\sum_{t=1}^{T}\left(y_{it}-\widehat{\alpha_i}(\boldsymbol{\beta})-\mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}\right)=\lambda\operatorname{sgn}(\widehat{\alpha_i}(\boldsymbol{\beta})) & \text{if } \widehat{\alpha_i}(\boldsymbol{\beta})\neq 0, \\ \left|\frac{1}{T}\sum_{t=1}^{T}\left(y_{it}-\mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}\right)\right|-\lambda\leq 0 & \text{if } \widehat{\alpha_i}(\boldsymbol{\beta})=0. \end{cases} \tag{5.2.8}$$

(5.2.8) can be rewritten as a soft-threshold estimator, i.e.

$$\widehat{\alpha_i}(\boldsymbol{\beta})=\left(\left|\frac{1}{T}\sum_{t=1}^{T}\left(y_{it}-\mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}\right)\right|-\lambda\right)_{+}\operatorname{sgn}\left(\sum_{t=1}^{T}\left(y_{it}-\mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}\right)\right). \tag{5.2.9}$$

Substituting this into (5.2.6) gives a profiled estimating function for $\boldsymbol{\beta}$:

$$g(\boldsymbol{\beta})=\left(\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{x}_{it}\mathbf{x}_{it}^{\mathrm{T}}\right)\boldsymbol{\beta}-\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{x}_{it}\left(y_{it}-\widehat{\alpha_i}(\boldsymbol{\beta})\right) \tag{5.2.10}$$

The panel lasso estimator for $\boldsymbol{\beta}_0$, denoted by $\widehat{\boldsymbol{\beta}}$, solves $g\left(\widehat{\boldsymbol{\beta}}\right)=0$. Let $\widehat{\alpha_i}=\widehat{\alpha_i}\left(\widehat{\boldsymbol{\beta}}\right)$.

Since the objective is to derive the asymptotic properties of $\widehat{\boldsymbol{\beta}}$, we need to determine how well (5.2.9) classifies the ith unit into one of the sets $\{1,\ldots,s_1\}$, $\{s_1+1,\ldots,s\}$, and $\{s+1,\ldots,N\}$. Note that (5.2.9) depends on the signs of $\left|\frac{1}{T}\sum_{t=1}^{T}\left(y_{it}-\mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}\right)\right|-\lambda$ and $\sum_{t=1}^{T}\left(y_{it}-\mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}\right)$. Substitute the model into the preceding expressions and define the following index sets:

$$\begin{aligned} S_{10} &= \left\{s+1\leq i\leq n:\left|\bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0-\boldsymbol{\beta}\right)+\bar{\epsilon}_i\right|\leq\lambda\right\}, \\ S_{11} &= \left\{1\leq i\leq s_1:\left|\alpha_{i0}+\bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0-\boldsymbol{\beta}\right)+\bar{\epsilon}_i\right|\leq\lambda\right\}, \\ S_{12} &= \left\{s_1+1\leq i\leq s:\left|\alpha_{i0}+\bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0-\boldsymbol{\beta}\right)+\bar{\epsilon}_i\right|\leq\lambda\right\}. \end{aligned}$$

Call S_{20}, S_{21}, and S_{22} the sets where we drop the absolute values and replace $\leq\lambda$ with $>\lambda$ in the definitions of S_{10}, S_{11}, and S_{12} respectively. Finally, call S_{30}, S_{31}, and S_{32} the sets where we drop the absolute values and replace $\leq\lambda$ with $<-\lambda$ in the definitions of S_{10}, S_{11}, and S_{12} respectively. By the definitions above, S_{10}, S_{20}, and S_{30} are mutually disjoint. The same applies to S_{11}, S_{21}, and S_{31} and S_{12}, S_{22}, and S_{32}. By assumption A4, $\alpha_{i0}=0$ for all $i\in S_{10},S_{20},S_{30}$. Note that these index sets will

[5]There are two cases to consider given the nondifferentiability of the absolute value function at 0. The Karush-Kuhn-Tucker conditions state the necessary and sufficient conditions for a minimizer of the optimization problem, i.e., the subdifferential at $\widehat{\alpha_i}(\boldsymbol{\beta})$ is zero. The first case where $\widehat{\alpha_i}(\boldsymbol{\beta})\neq 0$ follows from the requirement that the ordinary first derivative is equal to zero. The second case where $\widehat{\alpha_i}(\boldsymbol{\beta})=0$ requires that the subdifferential at $\widehat{\alpha_i}(\boldsymbol{\beta})$ has to include the zero element, i.e., $\frac{1}{T}\sum_{t=1}^{T}\left(y_{it}-\mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}\right)-\lambda e=0$ for some e that satisfies $-1\leq e\leq 1$. Recall that the inequality $|x|\leq\lambda$ is equivalent to $-\lambda\leq x\leq\lambda$. As a result, we get the expression in (5.2.8).

depend on $\boldsymbol{\beta}$. For an arbitrary index set S, I use $\widehat{S}$ to denote the result if we plug in the panel lasso estimator $\widehat{\boldsymbol{\beta}}$ into S.

To analyze whether the panel lasso estimator is consistent, I have to analyze the components of the estimating equation $g\left(\widehat{\boldsymbol{\beta}}\right) = 0$ after substituting (5.2.1) into (5.2.10):

$$\underbrace{\left(\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{x}_{it}\mathbf{x}_{it}^{\mathrm{T}}\right)}_{\mathbf{W}_n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right) = \frac{1}{n}\sum_{i=1}^{n}\bar{\mathbf{x}}_i\alpha_{i0}+\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{x}_{it}\epsilon_{it}-\frac{1}{n}\sum_{i=1}^{n}\bar{\mathbf{x}}_i\widehat{\alpha_i}. \quad (5.2.11)$$

The last term of (5.2.11) can be rewritten depending on which set i belongs, i.e.

$$\sum_{i\in S}\bar{\mathbf{x}}_i\widehat{\alpha_i} = \begin{cases} \sum_{i\in S}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}\right)+\sum_{i\in S}\bar{\mathbf{x}}_i\bar{\epsilon}_i-\lambda\sum_{i\in S}\bar{\mathbf{x}}_i & \text{if } S=\widehat{S_{20}} \\ \sum_{i\in S}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}\right)+\sum_{i\in S}\bar{\mathbf{x}}_i\alpha_{i0}+\sum_{i\in S}\bar{\mathbf{x}}_i\bar{\epsilon}_i-\lambda\sum_{i\in S}\bar{\mathbf{x}}_i & \text{if } S=\widehat{S_{21}},\ \widehat{S_{22}} \\ \sum_{i\in S}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}\right)+\sum_{i\in S}\bar{\mathbf{x}}_i\bar{\epsilon}_i+\lambda\sum_{i\in S}\bar{\mathbf{x}}_i & \text{if } S=\widehat{S_{30}} \\ \sum_{i\in S}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}\right)+\sum_{i\in S}\bar{\mathbf{x}}_i\alpha_{i0}+\sum_{i\in S}\bar{\mathbf{x}}_i\bar{\epsilon}_i+\lambda\sum_{i\in S}\bar{\mathbf{x}}_i & \text{if } S=\widehat{S_{31}},\ \widehat{S_{32}} \\ \mathbf{0} & \text{if } S=\widehat{S_{10}},\ \widehat{S_{11}},\ \widehat{S_{12}} \end{cases} \quad (5.2.12)$$

To further simplify (5.2.11), we need to say something about the contents of the index sets defined earlier as $n\to\infty$. But then, we would have to specify how to tune the regularization parameter, i.e. I assume that

$$\kappa_n \ll \lambda,\ \alpha\gamma_n \leq \lambda,\ \lambda \ll \sqrt{n}, \quad (5.2.13)$$

where $\alpha > 2$. This means that λ should be large enough to overrule the tail behavior of the time series averages of the regressors and the idiosyncratic error but small enough that it does not overrule the smallest of the "large" incidental parameters. Recall that assumption A4 imposes a particular behavior for the smallest of the "large" incidental parameters. As a result, I am able to extend Lemma 3.1 of Fan, Tang, and Shi (2012) to the panel data case and I present the details of the proof in the appendix.

Lemma 5.2.1 (Contents of index sets)**.** *Assume that A1, A2-1 (or A2-2), A3, and A4 hold. Let $n\to\infty$. For every $C>0$ and for every $\boldsymbol{\beta}\in B_C\left(\boldsymbol{\beta}_0\right)$, with probability going to 1,*

$$\begin{array}{lll} S_{10}=S_{10}^*, & S_{11}=\varnothing, & S_{12}=S_{12}^* \\ S_{20}=\varnothing, & S_{21}=S_{21}^*, & S_{22}=\varnothing \\ S_{30}=\varnothing, & S_{31}=S_{31}^*, & S_{32}=\varnothing \end{array}$$

where $S_{10}^*=\{s+1,s+2,\ldots,n\}$, $S_{12}^*=\{s_1+1,s_1+2,\ldots,s\}$, $S_{21}^*=\{1\leq i\leq s_1:\alpha_{i0}>0\}$ and $S_{31}^*=\{1\leq i\leq s_1:\alpha_{i0}<0\}$.

Notice that the preceding lemma enables us to allocate the indices $\{1,\ldots,n\}$ into four sets asymptotically – (a) S_{10}^{*} contain the indices for the units whose incidental parameter values are equal to zero, (b) S_{12}^{*} contain the indices for the units whose incidental parameter values are "bounded", (c) S_{21}^{*} contain the indices for the units whose incidental parameter values are "large" and positive, and (d) S_{31}^{*} contain the indices for the units whose incidental parameter values are "large" and negative. However, this result is not enough to guarantee consistency of the panel lasso estimator for $\boldsymbol{\beta}_0$.

Notice that the left hand side of (5.2.11) still contains terms that do not disappear in the limit unless we impose additional restrictions on the rate of growth of the number of the " bounded" and "large" incidental parameters. The proof of the next theorem can be found in the appendix.

Theorem 5.2.2 (Consistency of the panel lasso estimator)**.** *Assume that A1, A2-1 (or A2-2), A3, and A4 hold. Further assume that (i)* $\mathbf{W}_n \xrightarrow{p} \mathbf{W}$*, where* $\mathbf{W}$ *is nonsingular, (ii)* $s - s_1 = o\left(n/\left(\kappa_n\gamma_n\right)\right)$*, (iii)* λ *obeys (5.2.13), and (iv)* $s_1 = O(1)$*. Then, for some* $\overline{C} > 0$*, wpg 1, there exists a unique estimator* $\widehat{\boldsymbol{\beta}} \in B_{\overline{C}}\left(\boldsymbol{\beta}_0\right)$ *such that* $g\left(\widehat{\boldsymbol{\beta}}\right) = 0$ *hold and* $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$*.*

The theorem provides us with an alternative consistent estimator for the linear dynamic panel data model (possibly with feedback effects) but it would require a sparsity assumption. This theorem also differs from Theorem 3.2 of Fan, Tang, and Shi (2012) because (iv) is present. This condition bounds the number of "large" incidental parameters by a constant even as $n \to \infty$. However, it buys us the possibility to include predetermined and even contemporaneously exogenous variables and still be able to obtain a consistent estimator. Furthermore, Fan, Tang, and Shi (2012) assume a zero mean for the covariates.[6] I do not impose that assumption at all. Had we imposed this assumption, we can use a less restrictive condition on the number of "large" incidental parameters, i.e., $s_1 = o\left(n/\left(\kappa_n\gamma_n\right)\right)$.

I now construct a two-step estimator. First, define the following events:

$$\begin{aligned}\mathscr{E}_1 &= \left\{\widehat{\alpha_i} \neq 0 \text{ for } i = 1,\ldots,s_1\right\},\\ \mathscr{E}_2 &= \left\{\widehat{\alpha_i} = 0 \text{ for } i = s_1+1,\ldots,s,s+1,\ldots,n\right\}.\end{aligned}$$

The next lemma allows us to construct a two-step estimator by choosing the subset of the n units whose α_{i0} was estimated to be $\widehat{\alpha_i} = 0$. As long as we have a consistent estimator for $\boldsymbol{\beta}_0$, we would be able to detect the indices of the units who have "large" incidental parameters with high probability. Unfortunately, the lemma states that we are unable to estimate their values consistently. More importantly, we are able to

[6] Centering and standardization is typical in the high-dimensional statistics literature, especially for fixed design matrices. Centering and standardization are not so clear-cut in panel data situations given the presence of two dimensions of the data and the rather loose exogeneity assumptions I impose.

detect the zero-valued incidental parameters correctly but we shrink all the bounded incidental parameters to zero. In other words, s_1 is identified but not s. The proof is available in the appendix.

Lemma 5.2.3 (Partial consistency). *Let* $\mathscr{E} = \mathscr{E}_1 \cap \mathscr{E}_2$. *If* $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$, *then* $\Pr(\mathscr{E}) \to 1$ *under A1 to A4.*

In principle, the lemma applies to any initial consistent estimator of $\boldsymbol{\beta}_0$, even those that do not explicitly encourage sparsity. An example would be the usual GMM estimator. Unfortunately, the GMM estimator is not available under Assumption A2-2.

We now reestimate (5.2.1) using the data from the subset for which $\widehat{\alpha_i} = 0$ using this lemma. Define

$$\widehat{I_0} = \{1 \leq i \leq n : \widehat{\alpha_i} = 0\}$$

to be the subset under consideration. Similarly, define the corresponding true index set, i.e.,

$$I_0 = \{1 \leq i \leq n : \alpha_{i0} = 0\}.$$

The two-step panel lasso estimator $\widetilde{\boldsymbol{\beta}}$ can now be defined as the minimizer of

$$\min_{\boldsymbol{\beta}} \frac{1}{nT} \sum_{i \in \widehat{I_0}} \sum_{t=1}^{T} \left(y_{it} - \mathbf{x}_{it}^{\mathrm{T}} \boldsymbol{\beta}\right)^2, \tag{5.2.14}$$

Notice that (5.2.14) is exactly the least squares objective function restricted to observations belonging to $\widehat{I_0}$. Define the following matrices for $i \in \widehat{I_0}$:

$$\mathbf{X}_i = \begin{pmatrix} \mathbf{x}_{i1}^{\mathrm{T}} \\ \mathbf{x}_{i2}^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_{iT}^{\mathrm{T}} \end{pmatrix}, \mathbf{y}_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{pmatrix}, \boldsymbol{\epsilon}_i = \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{iT} \end{pmatrix}$$

The solution to (5.2.14) is given by the usual pooled least squares estimator, i.e.,

$$\widetilde{\boldsymbol{\beta}} = \left(\sum_{i \in \widehat{I_0}} \mathbf{X}_i^{\mathrm{T}} \mathbf{X}_i\right)^{-1} \left(\sum_{i \in \widehat{I_0}} \mathbf{X}_i^{\mathrm{T}} \mathbf{y}_i\right). \tag{5.2.15}$$

The next theorem shows that two-step panel lasso estimator $\widetilde{\boldsymbol{\beta}}$ is consistent as $n \to \infty$. The underlying idea is to use the previous lemma and apply the system OLS consistency theorem (Theorem 7.1) of Wooldridge (2010), along with an assumption on the rate of growth of the number of bounded incidental parameters.

To be specific, we can rewrite (5.2.15) as

$$\begin{aligned}
\widetilde{\boldsymbol{\beta}} &= \left(\sum_{i\in\widehat{I_0}}\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\right)^{-1}\left(\sum_{i\in\widehat{I_0}}\mathbf{X}_i^{\mathrm{T}}\mathbf{y}_i\right) \\
&= \left(\sum_{i\in\widehat{I_0}}\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\right)^{-1}\left(\sum_{i\in\widehat{I_0}}\mathbf{X}_i^{\mathrm{T}}\mathbf{y}_i\right)\left(\mathbf{1}\left\{\widehat{I_0}\neq I_0\right\}+\mathbf{1}\left\{\widehat{I_0}=I_0\right\}\right) \\
&= \underbrace{\left(\sum_{i\in\widehat{I_0}}\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\right)^{-1}\left(\sum_{i\in\widehat{I_0}}\mathbf{X}_i^{\mathrm{T}}\mathbf{y}_i\right)\mathbf{1}\left\{\widehat{I_0}\neq I_0\right\}}_{R}+\left(\sum_{i\in I_0}\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\right)^{-1}\left(\sum_{i\in I_0}\mathbf{X}_i^{\mathrm{T}}\mathbf{y}_i\right) \\
&\xrightarrow{p} \boldsymbol{\beta}_0+\left(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i\in I_0}\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\right)^{-1}\left(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i\in I_0}\mathbf{X}_i^{\mathrm{T}}\iota_T\alpha_{i0}+\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i\in I_0}\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\epsilon}_i\right),
\end{aligned}$$

where ι_T is a $T\times 1$ vector of ones. The term R involve the event $\left\{\widehat{I_0}\neq I_0\right\}$, whose probability converges to 0 as $n\to\infty$ by Lemma 5.2.3. Using assumptions A3 and A4, the term involving the incidental parameters can be evaluated as follows:

$$\left\|\frac{1}{n}\sum_{i\in I_0}\mathbf{X}_i^{\mathrm{T}}\iota_T\alpha_{i0}\right\|_2=\left\|\frac{1}{n}\sum_{i\in I_0}T\bar{\mathbf{x}}_i\alpha_{i0}\right\|_2\leq\frac{T}{n}\sum_{i=s_1+1}^{s}\|\bar{\mathbf{x}}_i\|_2\,|\alpha_{i0}|\leq T\left(\frac{s-s_1}{n}\right)\kappa_n\gamma_n.$$

Here we used assumption A4 to show that $\alpha_{i0}=0$ for $i=s+1,\ldots,n$. As a result, the term becomes $o_p(1)$ when $s-s_1=o\left(n/\left(\kappa_n\gamma_n\right)\right)$. Take note that T is taken as fixed here. The term involving the errors has probability limit equal to zero because of A2-2. As a result, we have the following theorem:

Theorem 5.2.4 (Consistency of the two-step panel lasso estimator)**.** *Suppose that the conditions in Theorem 5.2.2 hold. Further assume that (i)* $\mathbf{A}=E\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{X}_i\right)$ *is nonsingular and (ii)* $s-s_1=o\left(n/\left(\kappa_n\gamma_n\right)\right)$*. Then* $\widetilde{\boldsymbol{\beta}}\xrightarrow{p}\boldsymbol{\beta}_0$*.*

I now show the asymptotic normality of the two-step panel lasso estimator. Note that the event $\left\{\sqrt{n}R=0\right\}$ is a subset of the event $\left\{\widehat{I_0}\neq I_0\right\}$. As a result, $\Pr\left(\sqrt{n}R=0\right)\leq\Pr\left\{\widehat{I_0}\neq I_0\right\}=0$. Consider again the argument earlier for consistency. Recall that

$$\left\|\frac{1}{\sqrt{n}}\sum_{i\in I_0}\mathbf{X}_i^{\mathrm{T}}\iota_T\alpha_{i0}\right\|_2\leq T\left(\frac{s-s_1}{\sqrt{n}}\right)\kappa_n\gamma_n.$$

The term involving the incidental parameters will only be $o_p(1)$ when $s-s_1=o\left(\sqrt{n}/\left(\kappa_n\gamma_n\right)\right)$ to asymptotically remove the influence of the bounded incidental parameters. A central limit theorem can then be applied to the term involving the

errors. The result then follows from OLS asymptotic normality theorem for systems of equations (Theorem 7.2) in Wooldridge (2010).

Theorem 5.2.5 (Asymptotic normality of the two-step panel lasso estimator)**.** *Suppose that the conditions in Theorem 5.2.2 hold. Further assume that (i)* $\mathbf{A} = E\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{X}_i\right)$ *is nonsingular and (ii)* $s - s_1 = o\left(\sqrt{n}/\left(\kappa_n\gamma_n\right)\right)$*. Then* $\sqrt{n}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \xrightarrow{d} N\left(0, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\right)$*, where* $\mathbf{B} = Var\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\epsilon}_i\right)$*.*

Following standard arguments like those in Wooldridge (2010), a consistent estimator of the asymptotic variance of $\widetilde{\boldsymbol{\beta}}$ is given by

$$\widehat{\mathbf{V}} = \left(\sum_{i\in\widehat{I_0}} \mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\right)^{-1} \left(\sum_{i\in\widehat{I_0}} \mathbf{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\epsilon}}_i\widehat{\boldsymbol{\epsilon}}_i^{\mathrm{T}}\mathbf{X}_i\right) \left(\sum_{i\in\widehat{I_0}} \mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i\right)^{-1},$$

where $\widehat{\boldsymbol{\epsilon}}_i = \mathbf{y}_i - \mathbf{X}_i\widetilde{\boldsymbol{\beta}} = \boldsymbol{\epsilon}_i - \mathbf{X}_i\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)$ is a consistent estimator of $\boldsymbol{\epsilon}_i$. Note that this estimator is a robust variance estimator which allows for arbitrary serial correlation and time-series heteroscedasticity. Finally, note that, in principle, the two-step panel lasso reduces to the usual pooled OLS estimator when there is no heterogeneity in α_{i0}.

5.2.3 Choice of regularization parameter

So far we only have a theoretical specification for the regularization parameter λ as seen in (5.2.13). In practice, the choice of λ would have to be data-driven. A feasible procedure for the proposed method is as follows:[7]

1. Apply OLS to the model where $y_{i1} = \mathbf{x}_{i1}^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_{i1}$ for $i = 1, \ldots, n$. Obtain the residuals from the resulting regression, i.e. $\widehat{\epsilon_{i1}} = y_{i1} - \mathbf{x}_{i1}^{\mathrm{T}}\widehat{\boldsymbol{\beta}^{OLS}}$ for every i.

2. Select n_{pure} observations that correspond to the smallest values of $\{|\widehat{\epsilon_{11}}|, \ldots, |\widehat{\epsilon_{n1}}|\}$.

3. Apply OLS once again to the model where $y_{i1} = \mathbf{x}_{i1}^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_{i1}$ but only for the selected n_{pure} observations. Obtain the new set of residuals $\widetilde{\epsilon_{i1}} = y_{i1} - \mathbf{x}_{i1}^{\mathrm{T}}\widehat{\boldsymbol{\beta}^{OLS}}$ for every $i = 1, \ldots, n$.

4. Repeat Step 2 for the new set of residuals $\{|\widetilde{\epsilon_{11}}|, \ldots, |\widetilde{\epsilon_{n1}}|\}$.

5. Apply the lasso to $y_{i1} = \alpha_{i0} + \mathbf{x}_{i1}^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_{i1}$ but only for the n_{pure} observations. The regularization parameter for this step is the value that minimizes the extended BIC (EBIC) criterion.

[7] This feasible procedure has not been studied analytically. Hence, there are no guarantees that the procedure will mimic the theoretical properties of the panel lasso established earlier.

6. The units for which α_{i0} was estimated to be nonzero are removed from the dataset completely.

7. Apply the panel lasso to $y_{it} = \alpha_{i0} + \mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_{it}$ for all the remaining units i and the remaining time periods $t = 2, \ldots, T$. The regularization parameter for this step is the value that minimizes the EBIC criterion.

8. Determine the set $\widehat{I_0}$ and apply the two-step panel lasso estimator.

The first five steps can be thought of as applying the lasso to a testing set. The reasoning behind Steps 1 to 5 is to try to select the subset of the data that would most likely have $\alpha_{i0} = 0$. If the ith unit is of this type, then it is quite likely that the absolute value of the residuals in Steps 1 and 3 would be quite small relative to the other two types of incidental parameters. Steps 3 and 4 essentially repeat the first two steps to reduce the possibility of selecting units with a large absolute value for the idiosyncratic error. Notice that the first five steps use only the earliest information available in the panel. In contrast, the final three steps use all the remaining information in the panel. Finally, note that Step 5 is really a special case of the panel lasso for $T = 1$.

The extended BIC criterion, proposed by Chen and Chen (2008), modifies the usual BIC criterion so that the latter can still be applied in the context where the number of regressors P grow with sample size at some polynomial rate, i.e., $P = O\left(n^k\right)$ for some $k > 0$. The EBIC is indexed by some $\phi \in [0, 1]$ and is given by

$$BIC_{\phi}(s) = BIC(s) + 2\phi \log \begin{pmatrix} P \\ s \end{pmatrix},$$

where s is the size of the model. Notice the extra term in the criterion. This extra term penalizes models that are too large because the model space is growing large as the number of regressors also grow with sample size. Chen and Chen (2008) have shown the selection consistency of the extended BIC and when $P = O\left(n^k\right)$ and $\phi > 1 - 1/(2k)$. In the panel data case, we have $k = 1$ and we have to set $\phi > 1/2$.

The only remaining issue is that the user has to specify n_{pure}. Choosing the size of n_{pure} will ultimately depend on the user's faith in the sparsity assumption. If n is very large, n_{pure} can be set at a relatively small value. The value of n_{pure} should not be too small or too large for two reasons. First, there should still be enough degrees of freedom so that we are still able to implement Step 1, i.e., there should be enough observations so that $\boldsymbol{\beta}$ can still be estimated.[8] Second, the data-based procedure might produce a regularization parameter that overshrinks the large incidental parameters.

[8] One should take care that there are enough degrees of freedom so that the coefficients of a large but finite number of regressors can still be estimated.

5.3 Monte Carlo

In this section, I study the finite-sample performance of the two-step panel-lasso estimator. I start with the dynamic linear panel data model because this is often used in empirical applications. Furthermore, the empirical application discussed in Section 5.4 involves the estimation of a dynamic linear panel data model. The experiments follow the Monte Carlo design of Bun and Sarafidis (2015). Their design attempts to encompass existing Monte Carlo designs while ensuring comparability across different simulations. They consider the following model for $i = 1, \ldots, n$ and $t = 1, \ldots, T$:

$$\begin{aligned} y_{it} &= 0.8y_{i,t-1} + 0.2x_{it} + \alpha_i + \varepsilon_{it}, \\ x_{it} &= 0.95x_{i,t-1} + 0.25\alpha_i + \upsilon_{it}, \\ \upsilon_{it} &= \nu_{it} - 0.1\varepsilon_{it}. \end{aligned}$$

I set n to be either 50 or 1000. Assume that $\varepsilon_{it} \overset{iid}{\sim} N(0,1)$ and $\nu_{it} \overset{iid}{\sim} N\left(0, \sigma_\nu^2\right)$. To better control the experimental conditions, they suggest fixing the values of four parameters: the signal-to-noise ratio (SNR), the variance ratio (VR) and the correlation between the deviation of the initial condition of the x (and y) process(es) from its long run steady state path and the level of the steady state path itself (r_x and r_y respectively). The signal-to-noise ratio SNR represents the additional variance provided by the explained portion of the model conditional on α_{i0} after netting out the variance of ε_{it}. The variance ratio VR measures the relative magnitudes of the cumulative impact of the two error components on the average variance of y_{it} over time.[9]

I set $SNR = 3$, $VR = 100$, and $r_x = r_y = 0.5$. As a consequence, there is a lot of noise coming from α_{i0} relative to other components and the initial conditions are above the long-run steady state path. I set 40 periods for burn-in.[10] The incidental parameters α_{i0} are iid draws from the following mixed discrete-continuous distribution:

$$\alpha_{i0} = \begin{cases} 0 & \text{with probability } p_0 \\ W_1\left(0.5 + W_2\right) & \text{with probability } p_1 \\ U\left[-0.5, 0.5\right] & \text{with probability } 1 - p_0 - p_1 \end{cases},$$

where $W_1 = -1$ with probability 0.75 and $W_1 = 1$ with probability 0.25 and W_2 are iid draws from the exponential distribution with mean κ. I choose the value of κ so that I would be able to match the standard deviation of the distribution for the incidental

[9] For more details on the derivation of these quantities, I refer the reader to Bun and Sarafidis (2015).

[10] Although the number of burn-in periods may be high, mean stationarity is not exactly a requirement for the panel lasso to work.

parameters and the distribution of ν_{it} (denoted by σ_η and σ_ν respectively). The chosen values for κ can be found in the tables containing the simulation results.

I carried out all computations with R and the `glmnet` package (see Friedman, Hastie, and Tibshirani (2010).[11] In the Monte Carlo experiments, I set $T = 2$ where initial condition y_{i0} is available. Since there are no moment conditions that can be used to construct a GMM estimator, I compare the two-step panel lasso estimator to the pooled OLS estimator which treats all $\alpha_{i0} = 0$ and the pooled OLS estimator where I only use the units for which $\alpha_{i0} = 0$. I expect that the pooled OLS estimator which treats all $\alpha_{i0} = 0$ to be inconsistent. The pooled OLS estimator where I only use the units for which $\alpha_{i0} = 0$ is also called the oracle estimator. Clearly, the oracle should be consistent since it uses true knowledge not available to the econometrician.

I now describe the ability of the first step of the panel lasso to predict the type of incidental parameter in Table 5.3.1. The table provides us with an understanding of the theoretical results when applied to finite samples. The reported statistics include the mean biases and standard deviation for the estimators. I also report the rejection rates for individual Wald t-tests of the true null hypotheses $\gamma = 0.8$ and $\beta = 0.2$.

I consider two designs, referred to as Designs A and B. Designs A and B consider the situation where $(p_0, p_1) = (0.966, 0.02)$ and $(p_0, p_1) = (0.83, 0.1)$, respectively. The latter design allow us to consider what happens when the conditions on the number of "bounded" and "large" incidental parameters are violated. The former design is well within the conditions specified in the theorems discussed in the previous section.

The results in the table are in line with what was developed in the Section 5.2. It would seem that varying n_{pure}/n did not matter so much.[12] Furthermore, the pooled OLS estimator seems to be tracking the behavior of the panel lasso estimator when $n = 50$. Once we increase the value of n twenty-fold, we see that the inference properties of the pooled OLS estimator has suffered relative to the panel lasso. Clearly, the oracle is performing the best in every aspect.

The results in Table 5.3.1 may give the impression that the proposed estimator is not performing well when there is a larger sample size. The consistency of the panel lasso estimator requires that the number of large incidental parameters is bounded as sample size grows large. There are relatively more draws for the large incidental parameters when $n = 1000$ compared to $n = 50$, especially when one looks at Design B.

[11] R scripts used for the computations are available upon request. The EBIC introduced in the previous section can be used in conjunction with the LARS algorithm. However, the implementation in the Monte Carlo simulations uses coordinate descent rather than LARS.

[12] Varying ϕ in the EBIC did not change the results much either.

Table 5.3.1: Finite sample performance of estimators from 1000 replications

	Mean bias γ		Mean bias β		Sd γ		Sd β		Rej. Rate $\gamma = 0.8$		Rej. rate $\beta = 0.2$	
Design A ($\kappa = 7.74$)	$n = 50$	$n = 1000$	$n = 50$	$n = 1000$	$n = 50$	$n = 1000$	$n = 50$	$n = 1000$	$n = 50$	$n = 1000$	$n = 50$	$n = 1000$
Lasso $n_{pure}/n = 0.5, \phi = 1$	0.0170	0.0124	0.0071	0.0080	0.0795	0.0167	0.1232	0.0246	0.0590	0.1130	0.0710	0.0580
Lasso $n_{pure}/n = 0.9, \phi = 1$	0.0169	0.0124	0.0082	0.0080	0.0775	0.0167	0.1225	0.0245	0.0700	0.1140	0.0750	0.0580
Pooled OLS	0.0109	0.0124	0.0078	0.0080	0.0552	0.0122	0.0805	0.0170	0.0530	0.2020	0.0470	0.0730
Oracle	0.0003	-0.0008	0.0009	-0.0004	0.0575	0.0127	0.0838	0.0170	0.0540	0.0560	0.0490	0.0420
Design B ($\kappa = 3.367$)												
Lasso $n_{pure}/n = 0.5, \phi = 1$	0.0444	0.0463	0.0308	0.0302	0.0705	0.0142	0.1190	0.0244	0.1190	0.8990	0.0680	0.2470
Lasso $n_{pure}/n = 0.9, \phi = 1$	0.0463	0.0463	0.0291	0.0302	0.0693	0.0142	0.1189	0.0244	0.1210	0.8930	0.0700	0.2420
Pooled OLS	0.0441	0.0463	0.0300	0.0300	0.0475	0.0102	0.0790	0.0172	0.1750	0.9930	0.0630	0.4300
Oracle	-0.0014	-0.0008	0.0034	-0.0007	0.0631	0.0135	0.0902	0.0194	0.0590	0.0580	0.0540	0.0520

Note: The implied $\sigma_\eta = 1.477$ and $\sigma_\nu = 0.425$.

5.4 Inequality and income growth

The relationship between inequality and income growth has long been a subject of intense economic debate. Kuznets (1955), one of the top 20 papers chosen to celebrate the centennial of the American Economic Review, precisely deals with this relationship. The first page alone already lists down the key issues with studying this relationship. I interpret the approach by van der Weide and Milanovic (2014) as one way to address some of the issues with studying this relationship, in particular, the need to start with the family as the unit of analysis. A casual Google search of these keywords will already point out the immense number of studies (usually at the country level) devoted to studying this relationship. As more and more data are collected and disaggregated, we see more complicated methods being applied like those that involve dynamic linear panel data methods.

In this section, I revisit the evidence found by van der Weide and Milanovic (2014) where high levels of inequality reduce the income growth of the poorest percentiles of the distribution. Instead of using readily available aggregate measures of inequality and income, they construct these aggregate measures using state-level data from the United States. Individual-level data from the Integrated Public Use Microdata Survey for 1960, 1970, 1980, 1990, 2000, and 2010 were used to construct state-level measures of income per capita (`lnyxx`), inequality (`gini`), educational shortfalls (`edushort1518`), educational attainment beyond the college level (`edu_ms_age2139`), share of women outside the labor force (`olf_female`), share of household members that are too young (`age015`) and too old to work (`age65`). Details regarding the computation and definition of these variables can be found in their paper.

van der Weide and Milanovic (2014) estimate a Solow-type growth regression at the state level that includes a measure of inequality as one of the regressors. The parameter of interest is the effect of income inequality on income growth. They estimate (5.2.1) with variables defined as follows:

1. y_{it} represents income growth at specific percentiles. This is coded as `dlnyxx` where $xx \in (05, 10, 25, 50, 75, 90, 95, 99)$.[13]

2. $\mathbf{x}_{it}^{\mathrm{T}}$ contain the first-order lags of `gini`, `edu_ms_age2139`, `edushort1518`, `age015`, `age65`, `olf_female`, `lnyxx`, and time dummies. They also use two alternative measures of `gini`, namely, the state-level Gini of the bottom 40% (`gini_b40`) and top 40% (`gini_t40`) of the population.[14]

[13] It is unclear whether estimating separate regressions for each percentile is preferable over quantile regressions. I leave this to future work.

[14] They wanted to "unpack" the effect of inequality at the bottom and at the top on income growth at different percentiles.

3. They consider data from $n = 51$ states and $T = 5$ time periods (representing every decade since 1960). Alaska and the District of Columbia were considered outliers and were excluded from the sample.

I apply the panel lasso estimator to the model just described for $T = 2$ and for all states.[15] Why would the panel lasso be appropriate in this empirical application? The model chosen in the empirical application can be thought of as a growth regression possibly based on an augmented Solow growth model. Just as I discussed in Section 5.2 where I introduced the panel lasso, we can think of the incidental parameters as representing the suitability or fit of the Solow growth model to the data. In particular, large non-zero incidental parameters represent states for which the growth regression may not be a good approximation. The bounded incidental parameters represent moderate state-specific deviations from the growth model that can be shrunk toward zero (as this would have no effect on the estimator properties, at least asymptotically). An alternative justification is that sparsity can be useful when n and T are both small. The Monte Carlo experiments already provide some evidence in this regard. Furthermore, the two-step panel lasso estimator can be applied even for $T = 2$ and even accommodates contemporaneously exogenous regressors. For the moment, there is no estimator that could match these advantages.

The results of the one-step panel lasso estimator indicate that it is possible to pool all the states, regardless of whether I use `gini` as the inequality measure or `gini_b40` and `gini_t40` as the inequality measures. As a result, the exclusion of Alaska and DC from the sample by van der Weide and Milanovic (2014) may be unwarranted given the results of the one-step panel lasso estimator. The overall conclusion seems to be that heterogeneity across states may not be as large as one might think.

I then estimate using the proposed two-step panel lasso estimator for $T = 2$. Robust standard errors are used to construct the confidence intervals. Since the main interest of van der Weide and Milanovic (2014) is the effect of inequality on income growth, I only present 95% confidence intervals for the slopes of `gini`, `gini_b40`, and `gini_t40`. I set $n_{pure}/n = 0.8$ and $\phi = 1$ for the computation of the regularization parameter. I also report results from the pooled OLS estimator where there is only an overall intercept and no state-specific fixed effects. All other results are available upon request.

Figure 5.4.1 already gives an impression that the effect of inequality on income growth for the top 50% of the population has not changed so much over time. Although the effect of inequality is mostly positive for the top 50% of the population, the estimated effects are not as large as suggested by the system GMM results of van der Weide and Milanovic (2014). In contrast, the effect of inequality on income growth for the bottom 50% of the population has had substantial changes over time.

[15] Other configurations were implemented but the patterns obtained in Figures 5.4.1 and 5.4.2 remain.

If we compare data from 1970-1980 and the decades after, we see that even if the estimated effects are negative (and sometimes close to zero when looking at the median), the absolute values of these estimated effects are getting smaller over time.

Results of pooled OLS estimation can be found in Figure 5.4.2.[16] The results are substantially different from Figure 5.4.1 in two respects. First, the confidence intervals obtained by pooled OLS for 1970-1980 are strikingly different from those obtained from the two-step panel lasso. Second, the standard errors are much larger for the panel lasso. I interpret the results of Figure 5.4.2 as evidence that we may have to conduct a separate analysis of the 1970-1980 decade. Furthermore, the figure casts doubt on whether there is parameter constancy over time. The panel lasso has somehow stabilized this parameter nonconstancy.

Whichever figure one uses, there seems to be a sharp change in the relationship of inequality (whether bottom or top or as a whole) and income growth across all percentiles after 1970-1980. After this sharp change, this relationship has not changed so much after 1990, especially at the top percentiles. Most of the estimated effects of bottom inequality on income growth are statistically different from zero, especially for the percentiles above the median in recent years. I find that higher bottom inequality has a positive relationship with income growth at the top percentiles just like van der Weide and Milanovic (2014) but the magnitudes are slightly smaller. The estimated effects of top inequality on income growth are statistically not different from zero, especially for the percentiles above the median. If we look at Figures 5.4.1 and 5.4.2, there is reason to be optimistic because of the gradual reduction in the absolute effect of inequality (whether bottom or top or as a whole) on income growth.

To summarize, the results are strikingly different from the reported impression of massive inequality during 1990-2010. The absolute effect of bottom or top inequality on income growth has been getting smaller across time and across percentiles, especially when one looks at the bottom 50% of the population. The sharp change after the 1970-1980 decade might be driving the rather negative results (in the sense that they find that inequality is good for the rich but not for the poor) of van der Weide and Milanovic (2014). The notion that inequality (whether bottom or top) benefits only the rich may be a lot more nuanced than we think.

5.5 Concluding remarks

We show how the penalized least squares approach in the presence of incidental parameters of Fan, Tang, and Shi (2012) can be extended to panel data models. Not

[16] Pooled OLS is not the same as the panel lasso whenever there are some units for which the incidental parameter value is not equal to zero. Consider the case where there are cross-sectional units with "large" incidental parameters. The panel lasso removes these units while pooled OLS treats them as if these units were no different from units with zero or bounded values for the incidental parameters.

all of their results survive the extension. The most serious change in terms of consistency and valid inference is the need to bound the number of "large" incidental parameters by a constant. Despite this, I was able to allow for contemporaneously exogenous regressors. The sparsity of incidental parameters has been useful in deriving consistent estimators for the structural parameters. They come at the cost of specifying a particular structure for the asymptotic growth in the different types of incidental parameters in order to obtain consistency and asymptotic normality. The latter has been problematic in the context of estimators that encourage sparsity, as discussed extensively by Leeb and Pötscher (2005; 2008).

I also propose a data-based procedure for choosing the regularization parameter which uses the extended BIC criterion since the usual BIC criterion is inconsistent when the number of parameters grow at a polynomial rate with sample size. This data-based procedure is still at its infancy and would require further study to provide guarantees that coincide with the theory established in the past sections. It is likely that there would be other algorithms that would perform better than the one I have proposed.

Results of Monte Carlo experiments indicate good finite sample performance of the two-step panel lasso estimator for very small T. Unfortunately, departures from the assumed sparsity of the incidental parameters create substantial problems for consistent estimation and valid inference. As a result, the two-step panel lasso estimator is unable to match the performance of the oracle estimator but is preferable to simply using pooled OLS.

I also use the two-step panel lasso estimator to shed light on the relationship between inequality and income growth by revisiting the evidence of van der Weide and Milanovic (2014). The small sample size deters us from making stronger conclusions about what makes the excluded states different from the others. Perhaps the most optimistic aspect of the results is the gradual move toward the reduced impact of inequality on income growth across all percentiles.

Although the focus has been on fixed-T consistent estimation, an analysis of the performance of the penalized least squares estimator under alternative asymptotic embeddings such as letting $n, T \to \infty$ jointly or at a particular rate, say $n/T \to c \in (0, \infty)$ would be of practical value. The effect of a larger value of T might help us reduce the restrictions on the growth in the different types of incidental parameters. This analysis will also give insight as to the statistical benefits of a repeated observation and determine whether the cost of collecting panel data is justifiable. Furthermore, the derivation of the asymptotic properties of the two-step panel lasso estimator uses results from seemingly unrelated regressions, as introduced by Zellner (1962). Seemingly unrelated regressions give a natural framework for allowing varying coefficients not just for the intercept term but for the slope coefficients as well. By allowing for this extension, we may be able to develop alternative estimators for the varying coefficients model. Extensions to the case of nonlinear panel data

models will also be needed. Finally, linking the properties of the pretest estimator after a test of poolability to that of the two-step panel lasso estimator may also be of practical value. I leave all these to future research.

5.6 Appendix

Proof of Lemma 5.2.1

The argument follows Fan, Tang, and Shi (2012) but with some modifications and corrections. Let C be an arbitrary (small) positive number and $\boldsymbol{\beta} \in B_C\left(\boldsymbol{\beta}_0\right)$.

We first prove that $S_{10} = S_{10}^*, S_{20} = \varnothing, S_{30} = \varnothing$ wpg 1. It is always true that $S_{10} \subseteq S_{10}^*$. Thus we have to show that $\Pr\left(S_{10} \supseteq S_{10}^*\right) \to 1$. Define the events $\mathscr{B} = \left\{\max_{1\leq i\leq n} \|\bar{\mathbf{x}}_i\|_2 \leq \kappa_n\right\}$ and $\mathscr{D} = \{s+1 \leq i \leq n : |\bar{\epsilon}_i| < \gamma_n\}$. Note that $\Pr\left(S_{10} \supseteq S_{10}^*\right) \geq \Pr\left(S_{10} \supseteq S_{10}^* \mid \mathscr{B}\right)\Pr(\mathscr{B})$. Since $\Pr(\mathscr{B}) \to 1$ by assumption A3, it suffices to show that $\Pr\left(S_{10} \supseteq S_{10}^* \mid \mathscr{B}\right) \to 1$. For large n, $\Pr\left(S_{10}^* \subseteq \mathscr{D}\right) \to 1$. Conditional on $\mathscr{B}$, $|\bar{\epsilon}_i| \leq \gamma_n$ implies that

$$\left|\bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0 - \boldsymbol{\beta}\right) + \bar{\epsilon}_i\right| \leq \left|\bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0 - \boldsymbol{\beta}\right)\right| + |\bar{\epsilon}_i| \leq \left\|\bar{\mathbf{x}}_i^{\mathrm{T}}\right\|_2 \left\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}\right\|_2 + |\bar{\epsilon}_i| \leq \kappa_n\sqrt{d}C + \gamma_n,$$

and we have $\kappa_n\sqrt{d}C + \gamma_n \leq \lambda$ for large n by (5.2.13). As a result, $\mathscr{D} \subseteq S_{10}$. Thus, $\Pr\left(S_{10}^* \subseteq \mathscr{D} \subseteq S_{10} \mid \mathscr{B}\right) \to 1$. Since $S_{10} \cup S_{20} \cup S_{30} = S_{10}^*$ is always true, we must have $S_{20} = \varnothing$ and $S_{30} = \varnothing$ wpg 1.

Next we prove that $S_{11} = \varnothing, S_{21} = S_{21}^*, S_{31} = S_{31}^*$ wpg 1. We show that $S_{21} = S_{21}^*$ wpg 1 as the case for $S_{31} = S_{31}^*$ wpg 1 is analogous. . Let $S_{211} = S_{21} \cap S_{21}^*$ and $S_{212} = S_{21} \cap \left(S_{21}^*\right)^c$. It suffices to show that $\Pr\left(S_{211} = S_{21}^*\right) \to 1$ and $\Pr(S_{212} = \varnothing) \to 1$. It is always true that $S_{21} \subseteq S_{21}^*$. Thus we have to show that $\Pr\left(S_{21} \supseteq S_{21}^*\right) \to 1$. Define the events $\mathscr{B} = \left\{\max_{1\leq i\leq n} \|\bar{\mathbf{x}}_i\|_2 \leq \kappa_n\right\}$ and $\mathscr{D} = \{1 \leq i \leq s_1 : \bar{\epsilon}_i \geq -\gamma_n\}$. Note that $\Pr\left(S_{21} \supseteq S_{21}^*\right) \geq \Pr\left(S_{21} \supseteq S_{21}^* \mid \mathscr{B}\right)\Pr(\mathscr{B})$. Since $\Pr(\mathscr{B}) \to 1$ by assumption A3, it suffices to show that $\Pr\left(S_{211} \supseteq S_{21}^* \mid \mathscr{B}\right) \to 1$. For large n, $\Pr\left(S_{21}^* \subseteq \mathscr{D}\right) \to 1$. Conditional on $\mathscr{B}$ and noting that $\alpha_{i0} > 0$, $\bar{\epsilon}_i > -\gamma_n$ implies that

$$\alpha_{i0} + \bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0 - \boldsymbol{\beta}\right) + \bar{\epsilon}_i > \alpha^* - \kappa_n\sqrt{d}C - \gamma_n,$$

and we have $\alpha^* - \kappa_n\sqrt{d}C - \gamma_n \geq \lambda$ for large n by (5.2.13). As a result, $\mathscr{D} \subseteq S_{211}$. Thus, $\Pr\left(S_{21}^* \subseteq \mathscr{D} \subseteq S_{211} \mid \mathscr{B}\right) \to 1$.

Now, we show that $\Pr(S_{212} = \varnothing) \to 1$. It is always true that $\varnothing \subseteq S_{212}$. Thus we have to show that $\Pr(S_{212} \subseteq \varnothing) \to 1$. Define the events $\mathscr{B} = \left\{\max_{1\leq i\leq n} \|\bar{\mathbf{x}}_i\|_2 \leq \kappa_n\right\}$ and $\mathscr{D} = \{1 \leq i \leq s_1 : \bar{\epsilon}_i > \gamma_n\}$. Note that $\Pr(S_{212} \subseteq \varnothing) \geq \Pr(S_{212} \subseteq \varnothing \mid \mathscr{B})\Pr(\mathscr{B})$. Since $\Pr(\mathscr{B}) \to 1$ by assumption A3, it suffices to show that $\Pr(S_{212} \subseteq \varnothing \mid \mathscr{B}) \to 1$. For large n, $\Pr(\mathscr{D} \subseteq \varnothing) \to 1$. Conditional on $\mathscr{B}$, noting that $\alpha_{i0} < 0$ and $\gamma_n - \alpha^* +$

$\kappa_n\sqrt{d}C < \lambda$ for large n by (5.2.13), $\alpha_{i0} + \bar{\mathbf{x}}_i^{\mathrm{T}}(\boldsymbol{\beta}^* - \boldsymbol{\beta}) + \bar{\epsilon}_i > \lambda$ implies that

$$\bar{\epsilon}_i > \lambda - \alpha_{i0} - \bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0 - \boldsymbol{\beta}\right) > \lambda + \alpha^* - \kappa_n\sqrt{d}C > \gamma_n.$$

As a result, $S_{212} \subseteq \mathscr{D}$. Thus, $\Pr(S_{212} \subseteq \mathscr{D} \subseteq \varnothing \mid \mathscr{B}) \to 1$. Along with $\Pr\left(S_{21}^* \subseteq S_{211} \mid \mathscr{B}\right) \to 1$, we have $S_{21} = S_{21}^*$ wpg 1.

Finally, we prove that $S_{12} = S_{12}^*, S_{22} = \varnothing, S_{32} = \varnothing$ wpg 1. It is always true that $S_{12} \subseteq S_{12}^*$. Thus we have to show that $\Pr\left(S_{12} \supseteq S_{12}^*\right) \to 1$. Define the events $\mathscr{B} = \left\{\max_{1\le i\le n}\|\bar{\mathbf{x}}_i\|_2 \le \kappa_n\right\}$ and $\mathscr{D} = \{s_1 + 1 \le i \le s : |\bar{\epsilon}_i| < \gamma_n\}$. Note that $\Pr\left(S_{12} \supseteq S_{12}^*\right) \ge \Pr\left(S_{12} \supseteq S_{12}^* \mid \mathscr{B}\right)\Pr(\mathscr{B})$. Since $\Pr(\mathscr{B}) \to 1$ by assumption A3, it suffices to show that $\Pr\left(S_{12} \supseteq S_{12}^* \mid \mathscr{B}\right) \to 1$. For large n, $\Pr\left(S_{12}^* \subseteq \mathscr{D}\right) \to 1$. Conditional on $\mathscr{B}$, $|\bar{\epsilon}_i| \le \gamma_n$ implies that

$$\alpha_{i0} - \kappa_n\sqrt{d}C - \gamma_n \le \alpha_{i0} + \bar{\mathbf{x}}_i^{\mathrm{T}}\left(\boldsymbol{\beta}_0 - \boldsymbol{\beta}\right) + \bar{\epsilon}_i \le \alpha_{i0} + \kappa_n\sqrt{d}C + \gamma_n,$$

and we have $-\lambda - \alpha_{i0} + \kappa_n\sqrt{d}C < -\gamma_n$ and $\lambda - \alpha_{i0} - \kappa_n\sqrt{d}C > \gamma_n$ for large n by (5.2.13). As a result, $\mathscr{D} \subseteq S_{12}$. Thus, $\Pr\left(S_{12}^* \subseteq \mathscr{D} \subseteq S_{12} \mid \mathscr{B}\right) \to 1$. Since $S_{12} \cup S_{22} \cup S_{32} = S_{12}^*$ is always true, we must have $S_{22} = \varnothing$ and $S_{32} = \varnothing$ wpg 1.

Proof of Theorem 5.2.2

We now analyze every term in (5.2.11) after substituting in (5.2.12) and applying Lemma 5.2.1:

1. Collect all the terms that involve α_{i0}. Under A3 and A4, we have

$$\left\|\frac{1}{n}\sum_{i\in S_{12}^*}\bar{\mathbf{x}}_i\alpha_{i0}\right\|_2 \le \frac{1}{n}\sum_{i\in S_{12}^*}\|\bar{\mathbf{x}}_i\|_2\,|\alpha_{i0}| \le \frac{s-s_1}{n}\kappa_n\gamma_n.$$

   Provided that $s - s_1 = o\left(n/\left(\kappa_n\gamma_n\right)\right)$, we have $\dfrac{1}{n}\sum_{i=1}^{n}\bar{\mathbf{x}}_i\alpha_{i0} = o_p(1)$.

2. By the law of large numbers along with A1,

$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{x}_{it}\epsilon_{it} \xrightarrow{p} \lim_{n\to\infty}\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbb{E}(\mathbf{x}_{it}\epsilon_{it}) = \mathbf{0}.$$

   The latter equality follows from A2-1 or A2-2 or even strict exogeneity.

3. Strict exogeneity of $\mathbf{x}_{it}$ allows us to conclude that $\dfrac{1}{n}\sum_{i=1}^{n}\bar{\mathbf{x}}_i\bar{\epsilon}_i \xrightarrow{p} \mathbf{0}$. If any of the variables in $\mathbf{x}_{it}$ is predetermined or weakly exogenous, the argument has to

change slightly, i.e.

$$\left\| \frac{1}{n} \sum_{i \in S_{21}^* \cup S_{31}^*} \bar{\mathbf{x}}_i \bar{\epsilon}_i \right\|_2 \leq \frac{1}{n} \sum_{i \in S_{21}^* \cup S_{31}^*} \|\bar{\mathbf{x}}_i \bar{\epsilon}_i\|_2 \leq \frac{s_1}{n} \kappa_n \gamma_n.$$

The latter is $o(1)$ when $s_1 = o(n/(\kappa_n \gamma_n))$.

4. Let $S = S_{21}^*, S_{31}^*$. Note that

$$\left\| \frac{\lambda}{n} \sum_{i \in S} \bar{\mathbf{x}}_i \right\|_2 \leq \frac{\lambda}{n} \sum_{i \in S} \|\bar{\mathbf{x}}_i\|_2 \leq \frac{\lambda}{n} s_1 \kappa_n = \frac{\lambda}{\sqrt{n}} s_1 \frac{\kappa_n}{\sqrt{n}}.$$

The latter will only be $o_p(1)$ when A3 holds, λ obeys (5.2.13), and $s_1 = O(1)$.

Proof of Lemma 5.2.3

Let $\mathscr{E} = \mathscr{E}_1 \cap \mathscr{E}_2$. Assume that $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$. Define the following probabilities:

$$\begin{aligned}
T_0 &= \Pr\left(\bigcap_{i=s+1}^{n} \left\{ \left| \bar{\mathbf{x}}_i^{\mathrm{T}} \left(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}} \right) + \bar{\epsilon}_i \right| \leq \lambda \right\} \right) \\
T_1 &= \Pr\left(\bigcap_{i=1}^{s_1} \left\{ \alpha_{i0} + \bar{\mathbf{x}}_i^{\mathrm{T}} \left(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}} \right) + \bar{\epsilon}_i > \lambda \right\} \right) \\
T_2 &= \Pr\left(\bigcap_{i=s_1+1}^{s} \left\{ \left| \alpha_{i0} + \bar{\mathbf{x}}_i^{\mathrm{T}} \left(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}} \right) + \bar{\epsilon}_i \right| \leq \lambda \right\} \right)
\end{aligned}$$

Notice that $\Pr(\mathscr{E}) = T_0 T_1 T_2$. Therefore, to show that $\Pr(\mathscr{E}) \to 1$, it suffices to show that $T_0 \to 1$, $T_1 \to 1$, and $T_2 \to 1$. Note that

$$\begin{aligned}
1 - T_1 &= \Pr\left(\bigcup_{i=1}^{s_1} \left\{ \alpha_{i0} + \bar{\mathbf{x}}_i^{\mathrm{T}} \left(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}} \right) + \bar{\epsilon}_i \leq \lambda \right\} \right) \\
&\leq \underbrace{\Pr\left(\bigcup_{i \in S_{21}^*} \left\{ \alpha_{i0} + \bar{\mathbf{x}}_i^{\mathrm{T}} \left(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}} \right) + \bar{\epsilon}_i \leq \lambda \right\} \right)}_{T_{11}} \\
&\quad + \underbrace{\Pr\left(\bigcup_{i \in S_{31}^*} \left\{ \alpha_{i0} + \bar{\mathbf{x}}_i^{\mathrm{T}} \left(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}} \right) + \bar{\epsilon}_i \leq \lambda \right\} \right)}_{T_{12}}.
\end{aligned}$$

We just have to show that $T_{11} \to 0$ and $T_{12} \to 0$. Define $\mathscr{C} = \left\{ \left\| \boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}} \right\|_2 < \overline{C} \right\}$ where $\overline{C} = (\alpha - 1) / \left(M \alpha \sqrt{d} \right) > 0$, for some choice of M. Note that

$$\begin{aligned} T_{11} &\leq \Pr\left(\bigcup_{i \in S_{21}^*} \left[\left\{ \alpha_{i0} + \bar{\mathbf{x}}_i^{\mathrm{T}} \left(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}} \right) + \bar{\epsilon}_i \leq \lambda \right\} \cap \mathscr{C} \right] \right) + \Pr(\mathscr{C}^c) \\ &\leq \Pr\left(\bigcup_{i \in S_{21}^*} \left[\left\{ \bar{\epsilon}_i \leq \lambda - \alpha^* + \kappa_n \overline{C} \sqrt{d} \right\} \right] \right) + \Pr(\mathscr{C}^c) \\ &\leq \Pr\left(\bigcup_{i \in S_{21}^*} \left[\{ \bar{\epsilon}_i \leq -\gamma_n \} \right] \right) + \Pr(\mathscr{C}^c) \\ &\leq s_1 \Pr(\{ \bar{\epsilon}_i \leq -\gamma_n \}) + \Pr(\mathscr{C}^c) \to 0. \end{aligned}$$

The first inequality follows from the law of total and probability and the monotonicity of the probability function. The second inequality follows from the definition of $\mathscr{C}$ and the characteristics of the incidental parameters belong to the set S_{21}^*. The third and fourth inequalities follows from the specification of the regularization parameter found in (5.2.13) and subadditivity. The convergence to zero follows from assumption A3 and the consistency of the panel lasso. An analogous derivation will show that $T_{12} \to 0$.

To show that $T_0 \to 1$, note that

$$\begin{aligned} T_0 &\geq \Pr\left(\bigcap_{i=s+1}^{n} \left\{ -\lambda - \bar{\mathbf{x}}_i^{\mathrm{T}} \left(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}} \right) \leq \bar{\epsilon}_i \leq \lambda - \bar{\mathbf{x}}_i^{\mathrm{T}} \left(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}} \right) \right\} \cap \mathscr{C} \right) \\ &\geq \Pr\left(\bigcap_{i=s+1}^{n} \{ -\gamma_n \leq \bar{\epsilon}_i \leq \gamma_n \} \right) \to 1. \end{aligned}$$

The first inequality follows from the monotonicity of the probability function and some algebra. The second inequality arises because λ obeys (5.2.13) and

$$\begin{aligned} -\lambda - \bar{\mathbf{x}}_i^{\mathrm{T}} \left(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}} \right) &\leq -\lambda + \left\| \bar{\mathbf{x}}_i^{\mathrm{T}} \right\|_2 \left\| \boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}} \right\|_2 \leq -\lambda + \kappa_n \overline{C} \sqrt{d} \\ &\leq -\lambda + \lambda M \overline{C} \sqrt{d} = \lambda \left(M \overline{C} \sqrt{d} - 1 \right) < \lambda \left(-\frac{1}{\alpha} \right) \leq -\gamma_n \end{aligned}$$

for the choice of $\overline{C}$ indicated earlier.

Figure 5.4.1: 95% confidence intervals obtained from the panel lasso

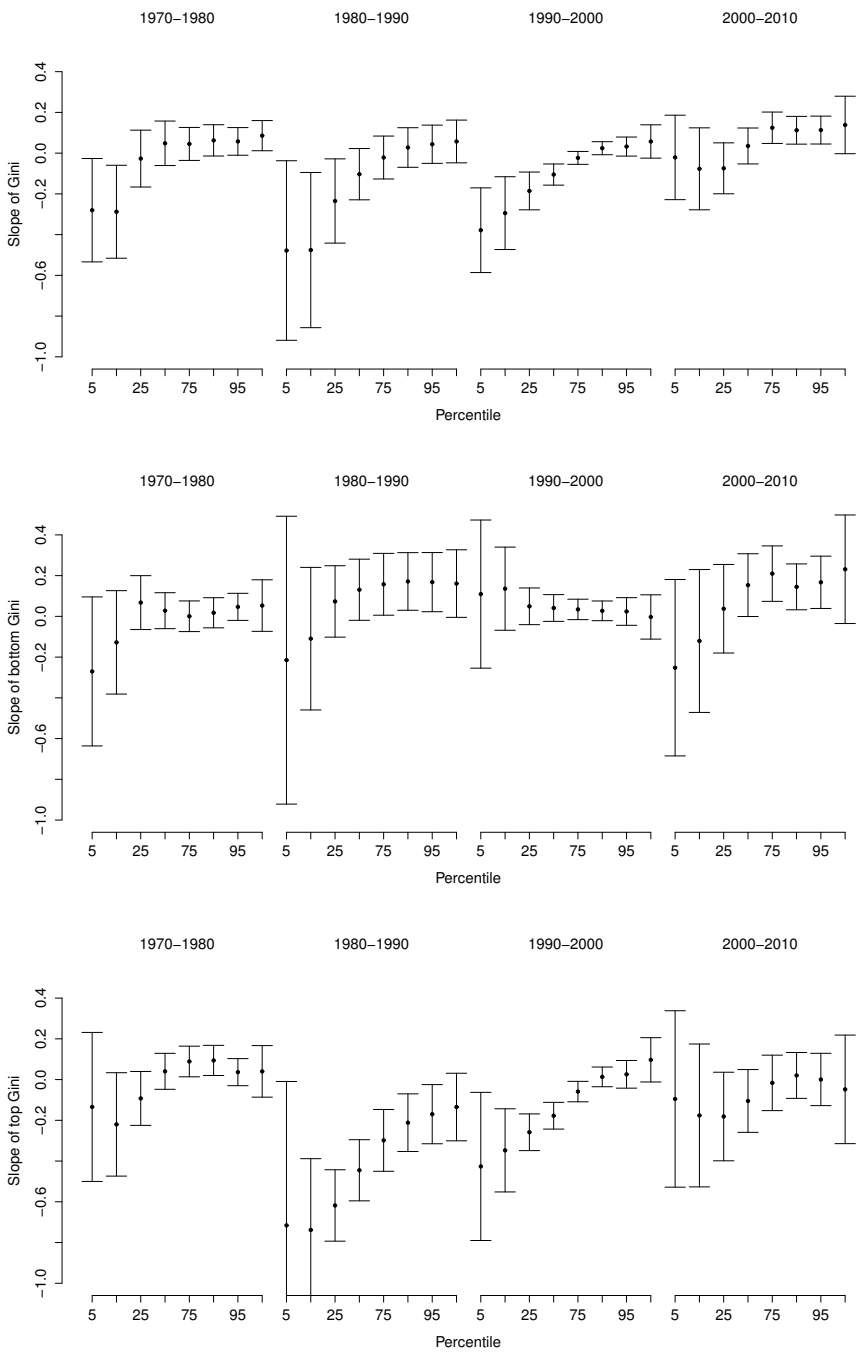
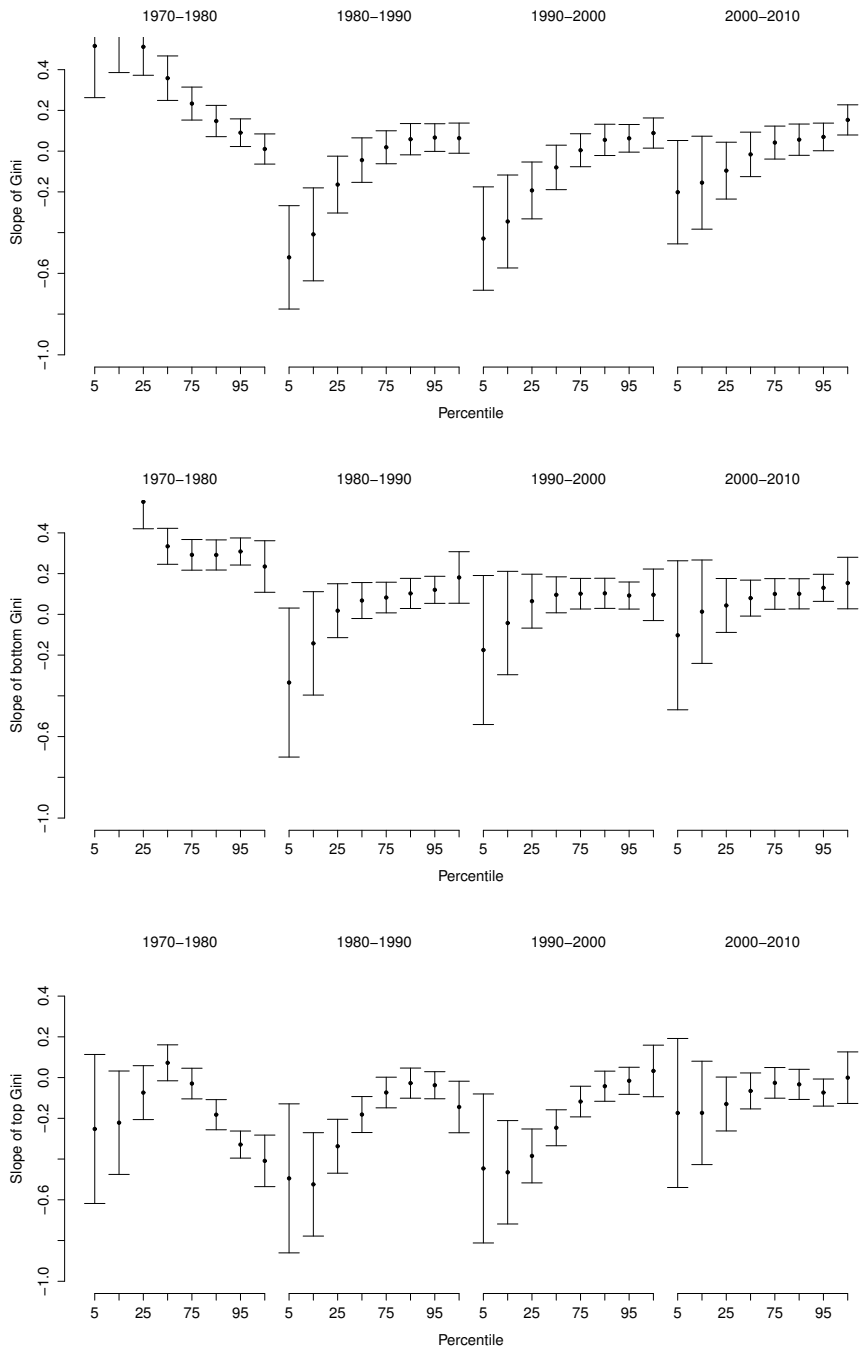


Figure 5.4.2: 95% confidence intervals obtained from pooled OLS



Chapter 6

Summary

My thesis is a collection of essays with a common theme: what practices and methods can be considered appropriate responses to the incidental parameter problem in panel data models. In recent years, we have seen an explosion of data collected from individuals, firms, or countries across short or long periods of time. This type of data gives us an opportunity to study the dynamics of change while controlling for time-invariant unobserved heterogeneity. Unfortunately, time-invariant unobserved heterogeneity, which is usually in the form of individual-specific fixed effects, creates problems for identification, estimation, and inference, especially if we continue to use default procedures without modification.

In Chapter 1, I introduce the reader to what I consider to be the main developments in the panel data literature over the past decades that would be relevant for understanding the motivation behind the remaining chapters in my thesis. Chapters 2 to 5 document my contributions to the panel data literature.

In Chapter 2, I show the folly of the usual empirical practice in top journals of using a simple linear probability model (LPM) to approximate average marginal effects from a nonlinear binary choice model in panel data settings. Setting aside the possibility that the average marginal effect may not be point-identified, directly applying IV estimators to a dynamic LPM delivers inconsistent estimators for the true average marginal effect regardless of whether cross-sectional or time series dimension diverge.

In Chapter 3, I develop a method to use panel data so that we are able to estimate a simultaneous equations model with discrete outcomes that allow for individual-specific unobserved heterogeneity and dynamics. This type of model has been considered quite frequently (but avoided) in empirical applications and no encompassing theory has yet been developed. I use the method to revisit empirical results from a model documenting the interaction of liquidity constraints and quantity constraints on labor supply for male household heads in the Panel Study of Income Dynamics.

In Chapter 4, I use orthogonal projections to construct a bias correction method for common parameters in panel data models. The proposed method involves a corrected score which is calculated by projecting the score vector for the structural parameters onto the orthogonal complement of a space characterized by incidental parameter fluctuations. Assuming that the individual-specific effect could take on almost any finite value and that the densities for the data are correctly specified, I show that the asymptotic distribution of the structural parameters is normal and centered at zero mimicking the results of bias correction procedures considered in this literature. Furthermore, the construction of the projected score lends itself to situations where there are multiple fixed effects. Numerical experiments show that the finite sample performance of projected scores is at least as good or better than existing competitors, especially when there are three or four time periods.

In the penultimate and speculative chapter, I exploit the strong parallels between extracting usable low-dimensional information from panel data even after controlling for individual-specific unobserved heterogeneity and extracting usable low-dimenstional information from the high volume but low informational content of big data. It seemed natural to ask exactly how a machine learning method like the lasso can offer a way to obtain consistency of the structural parameters (rather than predictive power) in linear dynamic panel data models with a fixed number of time periods (typically short) if we are willing to make an assumption that the individual-specific fixed effects are sparse. Results in this chapter indicate that the asymptotic theory requires stringent conditions on the growth rate of the number and size of the individual-specific fixed effects so that consistent estimation and valid inference are possible.

I wrote the essays with a research agenda in mind. Future work that I consider a priority should explore the following ideas. Just as in Chapter 2, I need to further document situations for which the linear probability model works or does not work. Developing a nonparametric identification argument and procedures for estimation and inference for the approach considered in Chapter 3 will definitely be of value to future empirical work that seeks to avoid imposing parametric restrictions. When the second-order orthogonal projection developed in Chapter 4 is carried out to the infinite order, it would be of interest to show that either we have a score from a conditional likelihood (if it exists), a score from a marginal likelihood (if it exists), or some other object that is a function of the structural parameters alone. Finally, the stage is set for extending the ideas in Chapter 5 to nonlinear panel data models.

Bibliography

Abrevaya, J (1999). Leapfrog estimation of a fixed-effects model with unknown transformation of the dependent variable. *Journal of Econometrics* **93**.(2), 203–228.

Acemoglu, D, S Johnson, JA Robinson, and P Yared (2009). Reevaluating the modernization hypothesis. *Journal of Monetary Economics* **56**.(8), 1043–1058.

Altonji, JG and RL Matzkin (2005). Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors. *Econometrica* **73**.(4), 1053–1102.

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

Anderson, T and C Hsiao (1981). Estimation of Dynamic Models with Error Components. *Journal of the American Statistical Association* **76**.(375), 598–606.

Anderson, T and C Hsiao (1982). Formulation and estimation of dynamic models using panel data. *Journal of Econometrics* **18**.(1), 47–82.

Angrist, JD and JS Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Angrist, JD (2001). Estimations of Limited Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice (with discussion). *Journal of Business & Economic Statistics* **19**.(1), 2–16.

Arellano, M and S Bond (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *Review of Economic Studies* **58**.(2), 277–97.

Arellano, M and S Bonhomme (2009). Robust Priors in Nonlinear Panel Data Models. *Econometrica* **77**.(2), 489–536.

Arellano, M and S Bonhomme (2011). Nonlinear panel data analysis. *Annual Review of Economics* **3**, 395–424.

Arellano, M and J Hahn (2006). *A Likelihood-Based Approximate Solution To The Incidental Parameter Problem In Dynamic Nonlinear Models With Multiple Effects*. Working Papers. CEMFI. `http://ideas.repec.org/p/cmf/wpaper/wp2006_0613.html`.

Arellano, M and J Hahn (2007). “Understanding Bias in Nonlinear Panel Models: Some Recent Developments”. In: *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*. Ed. by R Blundell, W Newey, and T Persson. Vol. 3. Cambridge University Press. Chap. 12, pp.381–409.

Arellano, M and B Honoré (2001). “Panel data models: some recent developments”. In: *Handbook of Econometrics*. Ed. by J Heckman and E Leamer. Vol. 5. Handbook of Econometrics. Elsevier. Chap. 53, pp.3229–3296.

Bajari, P, J Hahn, H Hong, and G Ridder (2011). A Note on Semiparametric Estimation of Finite Mixtures of Discrete Choice Models with Application to Game Theoretic Models. *International Economic Review* **52**.(3), 807–824.

Bartolucci, F, R Bellio, A Salvan, and N Sartori (2014). *Modified Profile Likelihood for Fixed-Effects Panel Data Models*. Tech. rep. Forthcoming in Econometric Reviews.

Basu, D (1977). On the Elimination of Nuisance Parameters. *Journal of the American Statistical Association* **72**.(358), 355–366.

Bellio, R and N Sartori (2003). Extending conditional likelihood in models for stratified binary data. *Statistical Methods and Applications* **12**.(2), 121–132.

Berger, JO, B Liseo, and RL Wolpert (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statist. Sci.* **14**.(1), 1–28.

Bernard, AB and JB Jensen (2004). Why Some Firms Export. *The Review of Economics and Statistics* **86**.(2), 561–569.

Bester, CA and C Hansen (2009a). A penalty function approach to bias reduction in nonlinear panel models with fixed effects. *Journal of Business and Economic Statistics* **27**.(2), 131–148.

Bester, CA and C Hansen (2009b). Identification of Marginal Effects in a Nonparametric Correlated Random Effects Model. *Journal of Business and Economic Statistics* **27**.(2), 235–250.

Bhanja, J and JK Ghosh (1992a). Efficient Estimation with Many Nuisance Parameters (Part I). *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* **54**.(1), 1–39.

Bhanja, J and JK Ghosh (1992b). Efficient Estimation with Many Nuisance Parameters (Part II). *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* **54**.(2), 135–156.

Bhanja, J and JK Ghosh (1992c). Efficient Estimation with Many Nuisance Parameters (Part III). *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* **54**.(3), 297–308.

Bhattacharyya, A (1946). On Some Analogues of the Amount of Information and Their Use in Statistical Estimation. *Sankhyā: The Indian Journal of Statistics (1933-1960)* **8**.(1), 1–14.

Bhattacharyya, A (1947). On Some Analogues of the Amount of Information and Their Use in Statistical Estimation (Contd.) *Sankhyā: The Indian Journal of Statistics (1933-1960)* **8**.(3), 201–218.

Bhattacharyya, A (1948). On Some Analogues of the Amount of Information and Their Use in Statistical Estimation (Concluded). *Sankhyā: The Indian Journal of Statistics (1933-1960)* **8**.(4), 315–328.

Bickel, PJ, CAJ Klaassen, Y Ritov, and J Wellner (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag New York, Inc.

Bickel, P and C Klaassen (1986). Empirical Bayes estimation in functional and structural models, and uniformly adaptive estimation of location. *Advances in Applied Mathematics* **7**.(1), 55–69.

Bjorn, PA and QH Vuong (1984). *Simultaneous Equations Models for Dummy Endogenous Variables: A Game Theoretic Formulation with an Application to Labor Force Participation*. Working Papers 537. California Institute of Technology, Division of the Humanities and Social Sciences. `http://ideas.repec.org/p/clt/sswopa/537.html`.

Blundell, RW and JL Powell (2004). Endogeneity in Semiparametric Binary Response Models. *Review of Economic Studies* **71**, 655–679.

Blundell, RW and RJ Smith (1993). "Simultaneous Microeconometric Models witrh Censored or Qualitative Dependent Variables". In: *Handbook of Statistics*. Ed. by GS Maddala, CR Rao, and HD Vinod. Vol. 11. Elsevier Science Publishers. Chap. 5, pp.117–143.

Blundell, R and RJ Smith (1994). Coherency and Estimation in Simultaneous Models with Censored or Qualitative Dependent Variables. *Journal of Econometrics* **64**.(1–2), 445–471.

Blundell, R and I Walker (1986). A Life-Cycle Consistent Empirical Model of Family Labour Supply Using Cross-Section Data. *The Review of Economic Studies* **53**.(4),

Bonhomme, S (2012). Functional Differencing. *Econometrica* **80**.(4), 1337–1385.

Bonhomme, S and E Manresa (2015). Grouped Patterns of Heterogeneity in Panel Data. *Econometrica* **83**.(3), 1147–1184.

Bresnahan, T and PC Reiss (1991). Empirical models of discrete games. *Journal of Econometrics* **48**.(1-2), 57–81.

Browning, M and JM Carro (2010). Heterogeneity in dynamic discrete choice models. *Econometrics Journal* **13**.(1), 1–39.

Bun, MJG and MA Carree (2005). Bias-Corrected Estimation in Dynamic Panel Data Models. *Journal of Business & Economic Statistics* **23**.(2), 200–210.

Bun, MJG and V Sarafidis (2015). "Chapter 3 – Dynamic Panel Data Models". In: *The Oxford Handbook of Panel Data*. Ed. by BH Baltagi. Oxford University Press, pp.76–110.

Cameron, SV and C Taber (2004). Estimation of Educational Borrowing Constraints Using Returns to Schooling. *Journal of Political Economy* **112**.(1), 132–182.

Carro, JM (2007). Estimating dynamic panel data discrete choice models with fixed effects. *Journal of Econometrics* **140**.(2), 503–528.

Carro, JM and A Traferri (2012). State Dependence and Heterogeneity in Health Using a Bias-corrected Fixed Effects Estimator. *Journal of Applied Econometrics*.

Cerra, V and SC Saxena (2008). Growth Dynamics: The Myth of Economic Recovery. *American Economic Review* **98**.(1), 439–57.

Chamberlain, G (1980). Analysis of Covariance with Qualitative Data. *Review of Economic Studies* **47**.(1), 225–38.

Chamberlain, G (1984). "Panel data". In: *Handbook of Econometrics*. Ed. by Z Griliches and MD Intriligator. Vol. 2. Handbook of Econometrics. Elsevier. Chap. 22, pp.1247–1318.

Chamberlain, G (1985). "Heterogeneity, Omitted Variable Bias, and Duration Dependence". In: *Longitudinal Analysis of Labor Market Data*. Ed. by JJ Heckman and B Singer. Cambridge University Press. Chap. 1, pp.3–38.

Chamberlain, G (2010). Binary Response Models for Panel Data: Identification and Information. *Econometrica* **78**.(1), 159–168.

Chen, J, J Gao, and D Li (2013). Estimation in Single-Index Panel Data Models with Heterogeneous Link Functions. *Econometric Reviews* **32**.(8), 928–955.

Chen, J and Z Chen (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95**.(3), 759–771.

Chen, M (2014). *Estimation of Nonlinear Panel Models with Multiple Unobserved Effects*. Tech. rep. `http://blogs.bu.edu/mlchen/files/2014/11/JMP-Nov15th-non-url.pdf`.

Chernozhukov, V, I Fernández-Val, J Hahn, and W Newey (2013). Average and Quantile Effects in Nonseparable Panel Models. *Econometrica* **81**.(2), 535–580.

Chesher, A and A Rosen (2012). *Simultaneous equations for discrete outcomes:coherence, completeness, and identification*. CeMMAP working papers CWP21/12. Centre for Microdata Methods and Practice, Institute for Fiscal Studies. `http://ideas.repec.org/p/ifs/cemmap/21-12.html`.

Choirat, C and R Seri (2012). Estimation in Discrete Parameter Models. *Statistical Science* **27**.(2), 278–293.

Christensen, BJ and NM Kiefer (2000). Panel data, local cuts and orthogeodesic models. *Bernoulli* **6**.(4), 667–678.

Clerides, SK, S Lach, and JR Tybout (1998). Is Learning By Exporting Important? Micro-Dynamic Evidence From Colombia, Mexico, And Morocco. *The Quarterly Journal of Economics* **113**.(3), 903–947.

Cornwell, C, P Schmidt, and D Wyhowski (1992). Simultaneous equations and panel data. *Journal of Econometrics* **51**.(1), 151–181.

Cox, DR and N Reid (1987). Parameter orthogonality and approximate conditional inference (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)* **49**.(1), 1–39.

Dagenais, MG (1999). A Simultaneous Probit Model. *Cahiers Economiques de Bruxelles* **163**, 325–346.

De Bin, R, N Sartori, and TA Severini (2015). Integrated likelihoods in models with stratum nuisance parameters. *Electronic Journal of Statistics* **9**, 1474–1491.

Dhaene, G and K Jochmans (2015a). *Likelihood inference in an autoregression with fixed effects*. Tech. rep. Forthcoming in Econometric Theory.

Dhaene, G and K Jochmans (2015b). *Profile-score adjustments for nonlinear fixed-effect models*. Tech. rep. Katholieke Universiteit Leuven.

Dubin, JA and DL McFadden (1984). An Econometric Analysis of Residential Electric Appliance Holdings and Consumption. *Econometrica* **52**.(2), 345–362.

Fan, J, R Tang, and X Shi (2012). Partial Consistency with Sparse Incidental Parameters. *ArXiv e-prints*.

Fernandez-Val, I and M Weidner (2013). Individual and Time Effects in Nonlinear Panel Models with Large N, T. *ArXiv e-prints*. arXiv: `1311.7065 [stat.ME]`.

Fernandez-Val, I (2009). Fixed effects estimation of structural parameters and marginal effects in panel probit models. *Journal of Econometrics* **150**, 71–85.

Firth, D and IR Harris (1991). Quasi-likelihood for Multiplicative Random Effects. *Biometrika* **78**.(3), 545–555.

Firth, D (1993). Bias Reduction of Maximum Likelihood Estimates. *Biometrika* **80**.(1), 27–38.

Freedman, DA and JS Sekhon (2010). Endogeneity in Probit Response Models. *Political Analysis* **18**.(2), 138–150.

Friedman, J, T Hastie, and R Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33**.(1), 1–22.

Galvao, AF and K Kato (2014). Estimation and Inference for Linear Panel Data Models Under Misspecification When Both n and T are Large. *Journal of Business & Economic Statistics* **32**.(2), 285–309.

Ghanem, D (2015). *Testing Identifying Assumptions in Nonseparable Panel Data Models*. Tech. rep.

Gourieroux, C, JJ Laffont, and A Monfort (1980). Coherency Conditions in Simultaneous Linear Equation Models with Endogenous Switching Regimes. *Econometrica* **48**.(3), 675–695.

Hahn, J (2001). Comment: Binary Regressors in Nonlinear Panel-Data Models with Fixed Effects. **19**.(1), 16–17.

Hahn, J and G Kuersteiner (2011). Bias Reduction for Dynamic Nonlinear Panel Models with Fixed Effects. *Econometric Theory* **27**.(6), 1152–1191.

Hahn, J and HR Moon (2010). Panel Data Models with Finite Number of Multiple Equilibria. *Econometric Theory* **26**.(3), 863–881.

Hahn, J and W Newey (2004). Jackknife and Analytical Bias Reduction for Nonlinear Panel Models. *Econometrica* **72**.(4), 1295–1319.

Hajivassiliou, V and F Savignac (2011). *Novel Approaches to Coherency Conditions in LDV Models with an Application to Interactions between Financing Constraints and a Firm's Decision and Ability to Innovate*. Tech. rep.

Hajivassiliou, VA and YM Ioannides (1995). *Unemployment and Liquidity Constraints*. Cowles Foundation Discussion Papers 1090. Cowles Foundation for Research in Economics, Yale University. `http://ideas.repec.org/p/cwl/cwldpp/1090.html`.

Hajivassiliou, VA and YM Ioannides (2007). Unemployment and liquidity constraints. *Journal of Applied Econometrics* **22**.(3), 479–510.

Han, S and EJ Vytlacil (2015). *Identification in a Generalization of Bivariate Probit Models with Endogenous Regressors*. Tech. rep.

Hausman, JA and ML Pinkovskiy (2013). "A Nonlinear Least Squares Approach to Estimating Fixed Effects Panel Data Models with Lagged Dependent Variables, with Applications to the Incidental Parameters Problem".

Heckman, JJ (1978). Dummy Endogenous Variables in a Simultaneous Equation System. *Econometrica* **46**.(4), 931–59.

Hoderlein, S and H White (2012). Nonparametric identification in nonseparable panel data models with generalized fixed effects. *Journal of Econometrics* **168**.(2), 300–314.

Honoré, BE (1992). Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects. *Econometrica* **60**.(3), 533–65.

Honoré, BE (1993). Orthogonality conditions for Tobit models with fixed effects and lagged dependent variables. *Journal of Econometrics* **59**.(1-2), 35–61.

Honoré, BE and E Kyriazidou (2000). Panel Data Discrete Choice Models with Lagged Dependent Variables. *Econometrica* **68**.(4), 839–874.

Honoré, BE and A Lewbel (2002). Semiparametric Binary Choice Panel Data Models Without Strictly Exogeneous Regressors. *Econometrica* **70**.(5), 2053–2063.

Honoré, BE and E Tamer (2006). Bounds on Parameters in Panel Dynamic Discrete Choice Models. *Econometrica* **74**.(3), 611–629.

Horrace, WC and RL Oaxaca (2006). Results on the bias and inconsistency of ordinary least squares for the linear probability model. *Economics Letters* **90**.(3), 321–327.

Hyslop, DR (1999). State Dependence, Serial Correlation and Heterogeneity in Intertemporal Labor Force Participation of Married Women. *Econometrica* **67**.(6), 1255–1294.

Jiménez, G, S Ongena, JL Peydró, and J Saurina (2014). Hazardous Times for Monetary Policy: What Do Twenty-Three Million Bank Loans Say About the Effects of Monetary Policy on Credit Risk-Taking? *Econometrica* **82**.(2), 463–505.

Kalbfleisch, JD and DA Sprott (1970). Application of Likelihood Methods to Models Involving Large Numbers of Parameters. *Journal of the Royal Statistical Society. Series B (Methodological)* **32**.(2), 175–208.

Khan, S, A Maurel, and Y Zhang (2015). *Informational Content of Factor Structures in Simultaneous Discrete Response Models*. Tech. rep. http://sites.duke.edu/yichongzhang/files/2015/10/msu2015.pdf.

Kim, MS and Y Sun (2009). "k-step bootstrap bias correction for fixed effects estimators in nonlinear panel models".

Kock, AB (2013). Oracle Efficient Variable Selection in Random and Fixed Effect Panel Data Models. *Econometric Theory* **29** (01), 115–152.

Kock, AB (2014). *Oracle inequalities and Variable Selection in High-Dimensional Panel Data Models*. Tech. rep. https://sites.google.com/site/andersbkock/LassoPanel.pdf.

Kock, AB and H Tang (2014). *Inference in high-dimensional dynamic panel data models*. Tech. rep. https://sites.google.com/site/andersbkock/KockTang_v11_20141224_3.pdf.

Kooreman, P (1994). Estimation of Econometric Models of Some Discrete Games. *Journal of Applied Econometrics* **9**.(3), 255–68.

Kosmidis, I and D Firth (2010). A generic algorithm for reducing bias in parametric estimation. *Electronic Journal of Statistics* **4**, 1097–1112.

Kuznets, S (1955). Economic growth and income inequality. *The American Economic Review* **45**.(1), 1–28.

Lancaster, T (2000). The incidental parameter problem since 1948. *Journal of Econometrics* **95**.(2), 391–413.

Lancaster, T (2002). Orthogonal Parameters and Panel Data. *Review of Economic Studies* **69**.(3), 647–66.

Leeb, H and BM Pötscher (2005). Model selection and inference: Facts and fiction. *Econometric Theory* **21**, 21–59.

Leeb, H and BM Pötscher (2008). Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics* **142**, 201–221.

Leon-Gonzalez, R (2003). A Panel Data Simultaneous Equation Model with a Dependent Categorical Variable and Selectivity. *Journal of Computational and Graphical Statistics* **12**.(1), 230–242.

Lewbel, A (2007). Coherency and Completeness of Structural Models Containing a Dummy Endogenous Variable. *International Economic Review* **48**.(4), 1379–1392.

Lewbel, A, Y Dong, and TT Yang (2012). Viewpoint: Comparing features of convenient estimators for binary choice models with endogenous regressors. *Canadian Journal of Economics* **45**.(3), 809–829.

Li, H, BG Lindsay, and RP Waterman (2003). Efficiency of projected score methods in rectangular array asymptotics. *Journal of the Royal Statistical Society B* **65**.(1), 191–208.

Maddala, GS (1987). Limited Dependent Variable Models Using Panel Data. *Journal of Human Resources* **22**.(3), 307–338.

Maddala, GS and LF Lee (1976). Recursive Models with Qualitative Endogenous Variables. *Annals of Economic and Social Measurement* **5**.(4), 525–545.

Manski, C (1987a). Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data. *Econometrica* **55**.(2), 357–62.

Manski, CF (1985). Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of Econometrics* **27**.(3), 313–333.

Manski, CF (1987b). Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data. *Econometrica* **55**.(2), 357–62.

Manski, CF (1988). Identification of Binary Response Models. English. *Journal of the American Statistical Association* **83**.(403), 729–738.

Massacci, D (2010). *Identification and Estimation of Bivariate Simultaneous Discrete Response Model without Sign Restrictions*. Tech. rep.

Masten, M (2015). *Random coefficients on endogenous variables in simultaneous equations models*. Tech. rep. `http://ideas.repec.org/p/ifs/cemmap/25-15.html`.

Matzkin, RL (2008). Identification in Nonparametric Simultaneous Equations Models. *Econometrica* **76**.(5), 945–978.

Matzkin, RL (2012). Identification in nonparametric limited dependent variable models with simultaneity and unobserved heterogeneity. *Journal of Econometrics* **166**.(1), 106–115.

McLeish, DL and CG Small (1994). *Hilbert Space Methods in Probability and Statistical Inference*. John Wiley & Sons, Inc.

Meango, R and I Mourifie (2013). *A note on the identification in two equations probit model with dummy endogenous regressor*. Working Papers tecipa-503. University of Toronto, Department of Economics. `http://ideas.repec.org/p/tor/tecipa/tecipa-503.html`.

Moon, HR and PCB Phillips (2004). GMM Estimation of Autoregressive Roots Near Unity with Panel Data. *Econometrica* **72**.(2), 467–522.

Moral-Benito, E (2013). Likelihood-Based Estimation of Dynamic Panels With Predetermined Regressors. *Journal of Business & Economic Statistics* **31**.(4), 451–472.

Moral-Benito, E (2014). Growth Empirics in Panel Data Under Model Uncertainty and Weak Exogeneity. *Forthcoming in the Journal of Applied Econometrics*.

Moran, PAP (1971). Estimating structural and functional relationships. *Journal of Multivariate Analysis* **1**.(2), 232–255.

Mundlak, Y (1978). On the Pooling of Time Series and Cross Section Data. *Econometrica* **46**.(1), 69–85.

Murtazashvili, I and JM Wooldridge (2008). Fixed effects instrumental variables estimation in correlated random coefficient panel data models. *Journal of Econometrics* **142**.(1), 539–552.

Nash, JC (2014). On Best Practice Optimization Methods in R. *Journal of Statistical Software* **60**.(2), 1–14.

Nash, JC and R Varadhan (2011). Unifying Optimization Algorithms to Aid Software System Users: optimx for R. *Journal of Statistical Software* **43**.(9), 1–14.

Nelsen, RB (2006). *An Introduction to Copulas (Springer Series in Statistics)*. Springer-Verlag New York, Inc.

Neyman, J and EL Scott (1948). Consistent Estimates Based on Partially Consistent Observations. *Econometrica* **16**.(1), 1–32.

Nickell, SJ (1981). Biases in Dynamic Models with Fixed Effects. *Econometrica* **49**.(6), 1417–26.

Pfanzagl, J (1993). Incidental Versus Random Nuisance Parameters. *Ann. Statist.* **21**.(4), 1663–1691.

Phillips, PCB and HR Moon (1999). Linear Regression Limit Theory for Nonstationary Panel Data. *Econometrica* **67**.(5), 1057–1111.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. `http://www.R-project.org`.

Reid, N (2013). Aspects of likelihood inference. *Bernoulli* **19**.(4), 1404–1418.

Sarafidis, V and N Weber (2011). *A Partially Heterogeneous Framework for Analyzing Panel Data*. Tech. rep.

Schmidt, P (1981). "Constraints on the Parameters in Simultaneous Tobit and Probit Models". In: *Structural Analysis of Discrete Data and Econometric Applications*. Ed. by CF Manski and DL McFadden. MIT Press. Chap. 12, pp.422–434.

Sims, C (2000). Using a likelihood perspective to sharpen econometric discourse: Three examples. *Journal of Econometrics* **95**.(2), 443–462.

Small, CG, J Wang, and Z Yang (2000). Eliminating multiple root problems in estimation (with discussion). *Statistical Science* **15**.(4), 313–341.

Sobel, ME and G Arminger (1992). Modeling Household Fertility Decisions: A Nonlinear Simultaneous Probit Model. *Journal of the American Statistical Association* **87**.(417), 38–47.

Stewart, MB (2004). Semi-nonparametric estimation of extended ordered probit models. *Stata Journal* **4**.(1), 27–39(13).

Stratmann, T (1992). The Effects of Logrolling on Congressional Voting. *The American Economic Review* **82**.(5), 1162–1176.

Tamer, E (2003). Incomplete Simultaneous Discrete Response Model with Multiple Equilibria. *The Review of Economic Studies* **70**.(1), 147–165. eprint: `http://restud.oxfordjournals.org/content/70/1/147.full.pdf+html`.

Tibshirani, R (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**.(1), 267–288.

Tibshirani, R (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**.(3), 273–282.

Tibshirani, R and L Wasserman (1994). Some aspects of the reparametrization of statistical models. *Canadian Journal of Statistics* **22**.(1), 163–173.

Torgovitsky, A (2015). *Partial Identification of State Dependence*. Tech. rep. `http://ssrn.com/abstract=2564305`.

Trivedi, PK and DM Zimmer (2007). Copula Modeling: An Introduction for Practitioners. *Foundations and Trends(R) in Econometrics* **1**.(1), 1–111.

van der Weide, R and B Milanovic (2014). *Inequality is bad for growth of the poor (but not for that of the rich)*. Policy Research Working Paper Series. The World Bank. `http://ideas.repec.org/p/wbk/wbrwps/6963.html`.

Waterman, RP (1993). "Projective Score Methods". PhD thesis. Pennsylvannia State University.

Winkelmann, R (2012). Copula Bivariate Probit Models: With an Application to Medical Expenditures. *Health Economics* **21**.(12), 1444–1455.

Wooldridge, JM (2000). A framework for estimating dynamic, unobserved effects panel data models with possible feedback to future explanatory variables. *Economics Letters* **68**.(3), 245–250.

Wooldridge, JM (2005a). Fixed-Effects and Related Estimators for Correlated Random-Coefficient and Treatment-Effect Panel Data Models. *The Review of Economics and Statistics* **87**.(2), 385–390.

Wooldridge, JM (2005b). Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics* **20**.(1), 39–54.

Wooldridge, JM (2010). *Econometric Analysis of Cross-Section and Panel Data*. MIT Press.

Woutersen, T (2003). *Robustness against incidental parameters*. Tech. rep. `http://www.yaroslavvb.com/papers/woutersen-robustness.pdf`.

Woutersen, T (2011). "Consistent estimation and orthogonality". In: *Missing Data Methods: Cross-sectional Methods and Applications*. Vol. 27A. Advances in Econometrics. Emerald Group Publishing Limited, pp.155–178.

Zellner, A (1962). An efficient method of estimating seemingly unrelated regression equations and tests for aggregation bias. *Journal of the American Statistical Association* **57**, 348–368.

Nederlandse Samenvatting (Summary in Dutch)

Dit proefschrift bestaat uit een aantal artikelen rond een gezamenlijk thema: het incidentele-parameterprobleem in paneldata modellen. De laatste decennia is er een grote toename in de beschikbaarheid van datasets waarbij een groep individuen, bedrijven of landen over een korte of lange periode worden waargenomen. Dit type data stelt ons in de gelegenheid om dynamisch gedrag te analyseren, waarbij rekening wordt gehouden met tijdsinvariante niet-waargenomen heterogeniteit. Doorgaans heeft deze heterogeniteit de vorm van individu-specifieke effecten, welke een probleem opleveren voor identificatie, schatting en toetsing van parameters, in het bijzonder wanneer standaardmethoden zonder aanpassing worden toegepast op panel data.

In Hoofdstuk 1 wordt een inleiding gegeven op de belangrijkste ontwikkelingen in de paneldata literatuur over de afgelopen decennia, voor zover die relevant zijn voor de motivatie voor de overige hoofdstukken van het proefschrift. Hoofdstukken 2 tot en met 5 bevatten mijn bijdragen aan de paneldata literatuur.

Hoofdstuk 2 laat zien dat het benaderen van gemiddelde marginale effecten in discrete-keuzemodellen op basis van het lineaire kansmodel – een gebruikelijke empirische aanpak in toptijdschriften – zeer onverstandig is in paneldata situaties. Afgezien van het feit dat het gemiddelde marginale effect soms niet uniek geïdentificeerd is, blijkt de instrumentele-variabelenmethode toegepast op een dynamisch lineair kansmodel te leiden tot een inconsistente schatter van het gemiddelde marginale effect, ongeacht welk type asymptotische benadering wordt gebruikt (grote cross-sectie of tijdreeksdimensie).

In Hoofdstuk 3 wordt een paneldata schattingsmethode ontwikkeld voor een simultane-vergelijkingenmodel met discrete endogene variabelen, niet-waargenomen heterogeniteit en dynamiek. Dit type model komt geregeld voor in empirische toepassingen, maar tot nu toe is er nog geen alomvattende econometrische aanpak voor ontwikkeld. De nieuw ontwikkelde methode wordt toegepast op een model voor de interactie van liquiditeits- en hoeveelheidseffecten op het arbeidsaanbod van man-

nelijke hoofden van de huishouding in de *Panel Study of Income Dynamics*. Daarbij worden de resultaten op basis van de nieuwe methode vergeleken met bestaande empirische resultaten.

In Hoofdstuk 4 wordt een methode ontwikkeld om de vertekening in de schatting van parameters in paneldata modellen te corrigeren op basis van orthogonale projecties. De voorgestelde methode is gebaseerd op een gecorrigeerde scorevector, verkregen door de scorevector voor de structurele parameters te projecteren op het orthogonale complement van een ruimte opgespannen door functies die de fluctuatie in incidentele parameters karakteriseren. Onder de aanname dat de individuspecifieke effecten vrijwel elke eindige waarde kunnen aannemen, en dat de kansdichtheden van de data correct gespecificeerd zijn, wordt afgeleid dat de asymptotische verdeling van de parameterschatters normaal is, gecentreerd rond de werkelijke waarde. Dit resultaat correspondeert met dat van bestaande biascorrectiemethoden in de literatuur. De constructie van de gecorrigeerde scorevector kan worden uitgebreid naar situaties waarin de niet-waargenomen heterogeniteit meerdimensionaal is. Monte Carlo simulaties laten zien dat de eindige-steekproefeigenschappen van de nieuwe methode minstens even goed, en in sommige gevallen beter zijn dan die van bestaande methoden, in het bijzonder als er slechts 3 of 4 tijdswaarnemingen zijn.

In Hoofdstuk 5 verken ik de parallellen tussen paneldata methoden die individuspecifieke niet-waargenomen heterogeniteit toelaten, en *big data* methoden gebaseerd op grote hoeveelheden gegevens met een relatief lage informatiewaarde, voor het verkrijgen van informatie over laag-dimensionale parameters. De vraag is hoe *machine learning* methoden zoals Lasso consistente schatters van de structurele parameters kunnen opleveren in lineaire dynamische paneldata modellen met een vast (en klein) aantal tijdswaarnemingen, als we bereid zijn om aan te nemen dat de individuele effecten *sparse* zijn. De resultaten van dit hoofdstuk geven aan dat sterke aannames nodig zijn wat betreft de omvang en het aantal individuele effecten ongelijk aan nul, om consistente schatting en betrouwbare inferentie over de structurele parameters mogelijk te maken.

Elk van de hoofdstukken passen in de onderzoeksagenda zoals hiervoor omschreven. Mijn vervolgonderzoek zal zich richten op de volgende ideeën. In aansluiting op Hoofdstuk 2 is verder onderzoek noodzakelijk in welke situaties het lineaire kansmodel wel of niet werkt. Het ontwikkelen van een identificatieaanpak en van schattings- en toetstingsprocedures aansluitend op de methode zoals ontwikkeld in Hoofdstuk 3 zal van pas komen bij toekomstig empirisch werk dat niet gebaseerd is op parametrische restricties. De tweede-orde orthogonale projectie van Hoofdstuk 4 kan worden uitgebreid naar hogere ordes; het is van belang te onderzoeken of dit uiteindelijk convergeeert naar de scorevector van een conditionele of marginale likelihood (als deze bestaan), of naar een andere functie van slechts de structurele parameters. De aanpak in Hoofdstuk 5, tenslotte, kan worden uitgebreid naar niet-lineaire panel data modellen.

Notes

Notes

Notes

Notes