Continuous permanent unobserved heterogeneity in dynamic discrete choice models

Jackson Bunting[*]

December 16, 2020

Latest version here

Abstract

In this paper, I show that dynamic discrete choice (DDC) models with continuous permanent unobserved heterogeneity are identified. The existing DDC literature controls for permanent unobserved heterogeneity through finite mixtures — that is, by assuming there is a finite number of agent 'types'. In contrast, I show that DDC models with infinitely many agent types are identified. Relative to the existing literature, I exploit commonly imposed assumptions to show identification under low-level conditions. My results apply to both finite- and infinite-horizon DDC models, do not require a full support assumption, nor a large panel, and place no parametric restriction on the distribution of unobserved heterogeneity.

The results provide a number of advantages for applied work. First, commonly used structural models can be estimated with more flexible heterogeneity. Second, my results do not require that the number of types be known *a priori*. Although there is rarely a theoretical reason for the number of types to be known, it is a common assumption in applied and theoretical work. Finally, the proposed seminonparametric estimator can be implemented using familiar parametric methods. I illustrate these advantages by applying my results to the labor force participation model of Altuğ and Miller (1998). In this model, permanent unobserved heterogeneity may be interpreted as individual-specific labor productivity, and my results imply that the distribution of labor productivity can be estimated from the participation model.

Keywords: dynamic discrete choice problems, nonparametric identification, unobserved heterogeneity.

JEL Classification Codes: C14, C61

[*] Duke University (jjb56@duke.edu). I am grateful to Federico Bugni, Adam Rosen, Matt Masten, Arnaud Maurel, Takuya Ura, Craig Fratrik, and Paul Diegert.

1 Introduction

Dynamic discrete choice (DDC) models provide a tractable way to learn about selection, while capturing the dynamic nature of economic decisions. A worker's decision to enter the labor market, a student's decision to attend a charter school, a hospital's decision to discharge a patient, a firm's decision to enter a market, a family's decision to migrate — these are applications of DDC found in the economics literature (Keane and Wolpin 2009; Einav, Finkelstein, and Mahoney 2018; Walters 2018). By accounting for selection, DDC models can be used to understand the effect of policy (such as the effect of expanding charter school access (Walters 2018)), or simply to explain important economic phenomena (such as US wage inequality (Heckman, Lochner, and Taber 1998)).

While DDC models allow for flexible dynamics, agent heterogeneity is somewhat limited. The models reflect the dynamic nature of many economic decisions: individuals are forward looking, and consider how current choices impact future outcomes. However, these sophisticated dynamics mean identification is delicate and generally requires many strong assumptions (Magnac and Thesmar 2002; Norets and Tang 2013). Of these, a key restriction is on the homogeneity of the agents. To be precise, in models without permanent unobserved heterogeneity, it is common to assume agents differ only by observed covariates and by random preference shocks drawn from a common distribution. This rules out, for example, permanent unobserved differences between individuals. Given this restrictive heterogeneity in baseline models, allowing for other forms of agent heterogeneity is of major interest.

The existing literature does provide for identification of DDC models with discrete permanent unobserved heterogeneity (Kasahara and Shimotsu 2009; Hu and Shum 2012). That is, agents are assumed to be one of a finite number of 'types', which is unobserved by the econometrician. This is an important relaxation of the baseline model: for example, in understanding the effect of price on drug purchases, different types of individuals may have different levels of unobserved health (Einav, Finkelstein, and Schrimpf 2015). In a model of migration, different types of individuals may have different costs of moving (Kennan and Walker 2011).

The main contribution of this paper is to show identification of DDC models with continuous permanent unobserved heterogeneity — that is, DDC models that allow, but do not require, there to be infinitely many types of individuals. This generalization is especially compelling because there is seldom a theoretical reason for the number of types to be finite. For instance, there may be no reason to think that unobserved health or unmeasured moving costs vary discretely among individuals. I show identification of the distribution of types and the type-specific component distributions in a short panel. The main results pertain to identification of DDC models with random coefficients, though I can also allow for random intercepts (fixed effects) under additional conditions. The results do not require the covariates to have full support, nor place parametric restrictions on the distribution of unobserved heterogeneity.

The second major difference from the existing literature is that I provide low-level conditions for identification. As is the case for finite mixtures, the key high-level condition for identification is that types display 'adequate variation' in behavior — that is, that each type responds adequately differently to changes in covariates. I show that commonly made assumptions, such as linear period payoffs, help ensure 'adequate variation.' To elaborate on this point, consider a binary choice DDC model. This model simplifies to a non-linear threshold crossing model, in which the choice of individual i at time t, given covariates x_{it} and the agent's type β_i, is equal to

$$1\left(v(x_{it}, \beta_i) + \epsilon_{it} \geq 0\right),$$

where v represents the value of choosing 1 over 0, and ϵ_{it} is a random preference shock. The key issue for identification is that, in general, v does not have a known analytical form and depends on its arguments non-linearly. An important insight of this paper is that common assumptions, such as random preference shocks and linear period payoffs, impose structure on v that is useful for proving 'adequate variation.' First, with parametric random preference shocks, the smoothness properties of v are well understood (Norets 2010; Kristensen et al. 2020). In particular, the smoothness of the state transition determines the smoothness of v . Second, v is entirely determined by the period payoff function — for example, in the infinite horizon case v is characterized by a fixed point that depends on the functional form of period payoffs (see equation (2)) — and it is common to assume the period payoff function is linear in x_{it}. In this paper, I exploit linearity and smoothness to show that the function v 'adequately varies' across different values of β_i in such a way as to achieve identification. This is in contrast to the canonical papers in the identification literature (Kasahara and Shimotsu 2009; Hu and Shum 2012), which impose 'adequate variation' at a high-level.

To implement the identification results, I propose a novel estimation method. Existing DDC estimation methods which focus on the parametric case[1] (Aguirregabiria and Mira 2002; Arcidiacono and Miller 2011) do not apply to the model of this paper, as the distribution of unobserved heterogeneity may be an infinite dimensional parameter of interest. Similarly, the computational complexity of DDC models means that immediately available nonparametric methods (such as

[1] In principle, standard DDC models may be semiparametric in the presence of continuous covariates, but, in practice, continuous covariates are often discretized and treated as such for estimation.

sieve likelihood estimation) may be impractical. To address these issues I propose a two-step sieve M-estimator, and show it is consistent for the distribution of permanent unobserved heterogeneity. I also propose a computationally convenient sieve space based on Heckman and Singer (1984). Intuitively, the estimator approximates the possibly continuous distribution of permanent unobserved heterogeneity by a discrete distribution. In this set up, the 'fixed grid' of support points of the approximating distribution is a tuning parameter of the sieve estimator. Computationally, this estimator is identical to an estimator for a model with finite types, but instead of the support points being a key identifying assumption, they are simply a tuning parameter.

As an alternative use for the identification results, I consider the case that the applied econometrician wishes to maintain the standard assumption that permanent unobserved heterogeneity is discrete. In this case, a key modeling decision is how to choose the number of support points of unobserved heterogeneity. There are rigorous methods to estimate the number of support points (Kasahara and Shimotsu 2014; Kwon and Mbakop 2019), which have been used in practice (Igami and Yang 2016). However, without additional assumptions, the estimators are consistent for only a *lower bound* on the number of support points in general. I show that my identification arguments imply that the estimator of Kwon and Mbakop (2019) is consistent for the number of support points of unobserved heterogeneity, if it is assumed to be finite.

To summarize, the above theoretical results may be of interest to applied economists for a number of reasons. First, the results imply that more flexible heterogeneity can be allowed for in commonly used structural models. In practice, permanent unobserved heterogeneity is often estimated with a small number of support points. See, for example, Table 1, which collects some important applications of DDC models with discrete unobserved heterogeneity. While there may be valid computational reasons for imposing this restriction on unobserved heterogeneity, the results of this paper imply that much richer patterns of heterogeneity can be identified. Second, the results mean that the economist need not know *a priori* the number of support points of permanent unobserved heterogeneity, as is commonly assumed in practice. As mentioned, without a long panel, point identification generally requires an upper bound on the number of types to be known. Despite this, economic theory seldom provides guidance for knowing the true number of agent types. Finally, the identification results can be implemented using familiar parametric methods which are computationally attractive.

To illustrate these advantages I apply my results to the labor supply model of Altuğ and Miller (1998). In this model, agents value consumption and leisure, and decide in each period whether

Table 1: Some applications of DDC models with discrete permanent unobserved heterogeneity

Authors (Year)	Journal	Support points
Keane and Wolpin (1997)	JPE	4
Lee and Wolpin (2006)	ECMA	5
Todd and Wolpin (2006)	AER	3
Kennan and Walker (2011)	ECMA	2
Scott (2014)	AER (R&R)	2
Einav, Finkelstein, and Schrimpf (2015)	QJE	5
Traiberman (2019)	AER	2

Note: These papers estimate a DDC model with discrete permanent unobserved heterogeneity. 'Support points' are the number of support points of the discrete unobserved heterogeneity. Journal names are: AER: American Economic Review; ECMA: Econometrica; JPE: Journal of Political Economy; QJE: Quarterly Journal of Economics.

or not to enter the workforce. The authors incorporate permanent unobserved heterogeneity by assuming that individual-specific labor productivity can be identified from a panel of wages. This assumption may be invalid if the length of the panel does not diverge or if the productivity term cannot be expressed as a deterministic function of observed variables. My identification results provide a means to avoid this assumption: in the context of their model, individual-specific labor productivity is identified as permanent unobserved heterogeneity in the labor force participation model.

To investigate the finite-sample properties of the estimator, I consider a suite of Monte Carlo simulations based on a simplified version of the model of Altuğ and Miller (1998). This section also demonstrates the computational attractiveness of the estimator.

After discussing related literature, I introduce the model and provide the main identification results (Section 2). Section 2.1 treats the infinite horizon model and Section 2.2 the finite horizon model. Variants on these baseline models are found in the appendix (Section A). Section 3 proposes the two-step sieve M-estimator, and shows its consistency. Section 5 considers an application, section 4 presents simulation results and finally section 6 concludes.

1.1 Related Literature

The canonical papers on point identification of DDC models with permanent unobserved heterogeneity are Kasahara and Shimotsu (2009) and Hu and Shum (2012). These papers use a short panel to identify type-specific conditional choice probabilities[2] and the discrete distribution of unobserved heterogeneity[3] via the measurement error approach of Hu and Schennach (2008). As discussed above, an important assumption in these papers is that choice behavior 'adequately varies' across types. In particular, they assume that a matrix of conditional choice probabilities is full rank, which is the precise meaning of 'adequate variation.' In the continuous case, the matrix rank condition generalizes to the injectivity of an integral operator. In contrast to their approach, I show that injectivity is implied by linear payoffs, parametric random preference shocks, continuous state variables and a non-zero homogeneous coefficient. On the other hand, their approach allows unobserved heterogeneity to enter the model in any way, restricted only by their high-level assumptions. For example, my assumptions rules out type-specific transition functions, considered in Kasahara and Shimotsu (2009, Section 3.2).

There is a large literature on identifying the distribution of continuous unobserved heterogeneity in binary response models. One stream exploits a linear index and full support covariates, while retaining nonparametric random preference shocks (Ichimura and Thompson 1998; Lewbel 2000; Gautier and Kitamura 2013). Relative to these papers, a DDC model yields a non-linear index with additive parametric preference shocks. To be precise, although period payoffs may be linear in the state variable, because agents are forward looking, the future value enters the choice equation also, and the net effect is a non-linear index. The second, more closely related stream considers the identifying power of parametric random preference shocks. Fox et al. (2012) show identification of random coefficients in a static model with a linear index. Williams (2019) identifies the distribution of univariate continuous unobserved heterogeneity in a dynamic multinomial choice model. However their results do not apply to the DDC models considered in this paper. In particular their Assumption 3.1 imposes that the third period state variable is independent of the second period choice, conditional upon the second period state and permanent unobserved heterogeneity. A key feature of the DDC models considered in this paper is that the transition of the state variable depends on the agent's choice.

[2] The conditional choice probability (CCP) is $\Pr(a_{it} = a \mid x_{it} = x, \beta_i = b)$: the probability that agent i chooses action a in period t given their observed state variable is x and their level of unobserved heterogeneity is b.

[3] The main result in Hu and Shum (2012) identifies continuous unobserved heterogeneity, but that theorem does not directly apply to the DDC model of their section 4.1.

Another stream of literature considers the identification of finite dimensional parameters, viewing permanent unobserved heterogeneity as a nuisance parameter. Aguirregabiria, Gu, and Luo (2020) consider the identification of homogeneous parameters in the presence of fixed effects in a particular DDC model. They adopt the conditional likelihood approach of Chamberlain (1980) and use particular sequences of choice variables to difference away the fixed effect. Chernozhukov et al. (2013) provide a semiparametric estimator that is robust to set identification of the finite dimensional parameter.

The seminonparametric estimator I propose is based on Heckman and Singer (1984). Similar 'fixed grid' estimators have been analyzed for both the parametric and static cases (Fox et al. 2011; Fox, Kim, and Yang 2016), and are increasingly used in applied work (e.g. Nevo, Turner, and Williams 2016).

2 Model and identification

2.1 Infinite horizon model

In an infinite horizon dynamic discrete choice model, the agent chooses a sequence of actions $(a_t, a_{t+1}, \dots)$ to maximize lifetime utility:

$$V(x_t, \epsilon_t) = \max_{(a_t, a_{t+1}, \dots)} E\left[\sum_{s=1}^{\infty} \rho^{s-t} u(x_s, \epsilon_s, a_s,) \mid x_t, \epsilon_t, a_t\right] \tag{1}$$

where $a_t \in \{0, \dots, |A|\} \equiv A$ is the control variable, ρ is the discount factor, $s_t = (x_t, \epsilon_t)$ is the state variable of which x_t is observed by the econometrician and $\epsilon_t = (\epsilon_{at} : a \in A)$ is not, and $u(x_t, \epsilon_t, a_t)$ is the period payoff which may be agent specific. This formulation makes clear the dynamic aspect of DDC problems: first, agents value current and future payoffs and, second, through the conditional expectation, they take into account how today's choice impacts future outcomes. Assuming the state variables follow a Markov process, the agent's problem becomes a Markovian Dynamic Discrete Choice problem. The distinguishing feature of infinite horizon DDC problems is that, under mild conditions (e.g. Rust et al. 1994, Theorem 2.3), they admit a recursive formulation:

$$V(x_t, \epsilon_t) = \max_{a_t \in A} \{u(x_t, \epsilon_t, a_t) + \rho E[V(x_{t+1}, \epsilon_{t+1}) \mid x_t, \epsilon_t, a_t]\}. \tag{2}$$

The formulation also makes transparent the non-linearity of DDC models — even if the period payoffs u are linear in x_t, that property is unlikely to be inherited by the integrated value function V. The non-linearity is a key challenge for identification.

In this section I present conditions for identification of the distribution of continuous unobserved heterogeneity within the above model. The first assumption imposes restrictions that are standard for DDC models without permanent unobserved heterogeneity.

Assumption I1. *(i)* $u(x_t, \epsilon_t, a_t) = u(x_t, a_t) + \epsilon_{a_t t}$. *(ii)* $\rho \in [0, 1)$ *is known. (iii)* (x_t, ϵ_t) *satisfy*

$$dF_S(s_{t+1}; s_t, a_t) = dF_\epsilon(\epsilon_{t+1}) dF_x(x_{t+1}; x_t, a_t). \tag{3}$$

(iv) $u_i(x_t, 0) = 0$. *(v) The distribution of* ϵ_{at} *is extreme value type I.*

Assumption I1 contains standard identifying assumptions for DDC models (Magnac and Thesmar 2002; Aguirregabiria and Mira 2010), including additive separability of the state variables, that the discount factor is known, conditional independence, and the outside good assumption. These assumptions are not innocuous — for example, Norets and Tang (2013) show that the choice of outside good may affect predicted counterfactual outcomes, and is therefore not a true normalization. Nevertheless, it is standard to assume the unobserved state variables have a known distribution, of which normal and extreme value type I are common choices. Denote $\tilde{A} = \{1, 2, \ldots, |A|\}$, the choice set excluding choice 0.

Assumption I2. *Permanent unobserved heterogeneity* $\beta_i = (\beta_{ia} : a \in \tilde{A}) \in \mathbb{R}^b$ *for* $b = |A|$ *enters the model through the period utility function as follows:*

$$u_i(x, a) = x' \left(\beta_{ia},\ \gamma_a\right),$$

where $x \in \mathbb{R}^k$ *is the vector of observed state variables, and the agent index* i *is shown for explicitness.* S_β*, the support of* β_i*, is a bounded subset of* $\mathbb{R}^b$. β_i *conditional upon* $x_1 = x$ *is either discrete or absolutely continuous, in which case its density function* $f_{\beta|x_1}$ *is bounded.*

Assumption I2 states that permanent unobserved heterogeneity enters the model as random coefficients. The restrictions placed on its distribution are mild. First, it allows, but does require, the distribution to have uncountable support. This is a point of departure from the existing literature, where the support is assumed to have a known finite number of support points (Kasahara and Shimotsu 2009; Hu and Shum 2012). Assumption I2 allows there to be infinitely many types of agents, but nests the standard finite-types assumption as a special case. Second, no restrictions are placed on the dependence between the observed state variable and permanent unobserved heterogeneity, which is standard in this literature.

Assumption I3. *Let $\gamma_{a|A|}$ be the first $|A|$ components of the vector γ_a and let Γ_A be the $|A| \times |A|$ matrix with columns $\gamma_{a|A|}$. Then the matrix Γ_A has full rank, as do all of its principal submatrices.*

Assumption I3 imposes that the state variable cannot affect payoffs for each choice in a similar fashion. For example, in the binary choice case ($|A| = 1$), the assumption states that $\gamma_0 \neq 0$. The final two assumptions place restrictions on directly observed objects. First, broadly speaking, the support of the state variable is required to contain an open set:

Assumption I4. *(i) The restriction of the support of x_2 conditional upon $(x_1, a_1) = (x, a)$ to the first $1 + |A|$ elements of x_2 is bounded and contains a non-empty open set. (ii) The support of x_3 conditional upon $(x_2, a_2) = (x, 0)$ for some x in the support of part (i) contains k linearly independent elements and its restriction to the first $1 + |A|$ elements of x_3 is bounded and contains a non-empty open set. (iii) The intersection over a_3 of the support of x_4 conditional upon x_3 in the support of part (ii) and a_3 contains k linearly independent elements.*

Assumption I4 places restrictions on the support of the observed state variable. It allows the support to be bounded, but requires that it be uncountable. The conditions do rule out some transition patterns found in applied work, but may be less onerous than other support conditions in the literature. First, parts (ii) and (iii) rule out the case that the state variable x_t contains the lagged choice a_{t-1}. However, it does not rule out 'machine replacement models' such as the Rust model of Kasahara and Shimotsu (2009, Section 3.3). Finally, unlike some results in the literature, it does not require that the support be 'rectangular' — which requires that, starting from any sequence of choices and past state variables, any state can be reached[4]. state variable.

Assumption I5. *The state transition kernel $F_x(x_{t+1}; x_t, a_t)$ has bounded support and may be decomposed into absolutely continuous and discrete components, and the associated density and probabilities are real analytic functions of the first $1 + |A|$ elements of x_t. Furthermore, these functions have analytic continuations to $\mathbb{R}^{1+|A|}$ which are bounded.*

Assumption I5 allows the state transition to be a mixture of an absolutely continuous and discrete random variable, but restricts the component functions to be smooth functions of the conditioning state variable. In particular, they must be real analytic functions — that is, functions that have a convergent power series representation. An example of a state transition satisfying Assumption I5 is a mixture of a mass point at $x_{t+1} = 0$ and a truncated normal: $F_x(x'; x, a) =$

[4] More precisely, that is for each (x, a), $F_x(x'; x, a) > 0$ for all x' in its support. The assumption is made in Kasahara and Shimotsu (2009, Propositions 1-9).

$\pi 1(x' = 0) + (1 - \pi)F_+(x'; x, a)$, where $F_+(x'; x, a)$ is a truncated normal whose mean and variance are real analytic functions of (x, a).

Other examples of real analytic functions include polynomials, the logistic function, trigonometric functions, the Gaussian function, and linear combinations of these functions. These functions are known to be good approximators to square-integrable functions (e.g. Chen 2007, Section 2.3), and can therefore approximate any density function arbitrarily well. The bounded support assumption is a technical requirement to ensure that the mapping (2) is a contraction between spaces of bounded functions (Kristensen et al. 2020). This could be relaxed, but proving equation (2) has the contraction property is more delicate (Norets 2010, see).

With these assumptions in hand, the model parameters are $(F_x, \gamma, f_{\beta|X_1})$: the state transition, the homogeneous payoff parameter $\gamma = (\gamma_a : a \in \tilde{A}) \in \mathbb{R}^{|A|(k-1)}$, and the conditional distribution of permanent unobserved heterogeneity. As the state transition is identified by direct observation, the following result deals with the other parameters:

Theorem 1. *Assume the distribution of $(x_t, a_t)_{t=1}^T$ is observed for $T \geq 4$, generated from agents solving the model of equation (1) satisfying assumptions I1-I5. Then $(\gamma, f_{\beta|X_1})$ is point identified.*

Remark 1 (Random intercepts)*.* In applied work, it is common to impose that permanent unobserved heterogeneity enters the model as a random intercept — a fixed effect. That is, the period utility function of Assumption I2 is replaced by

$$u_i(x, a) = \beta_{ia} + x'\gamma_a.$$

This parsimonious model often gives a natural interpretation. For example, if the choice set includes home production, schooling and various occupations, β_{ia} can be interpreted as choice-specific skill endowments (Keane and Wolpin 1997).

Section A.1 considers identification of an infinite-horizon DDC model with random intercepts. It shows point identification can be attained under an additional restriction on the state transition. Specifically, there must be some point in the support of x_{it} for which the state transition is not choice dependent. For instance, the machine replacement model of Kasahara and Shimotsu (2009, Example 9) displays this property.

Remark 2 (Panel length)*.* Theorem 1 requires at least four observations per individual. In contrast Kasahara and Shimotsu (2009) require only $T = 3$. With three periods, identification of the model in Theorem 1 is possible under a high-level assumption on the joint distribution of permanent unobserved heterogeneity and the first period state variable. For example, independence would be sufficient for identification. However, the advantage of $T = 4$ is to allow unrestricted dependence between the state variable and permanent unobserved heterogeneity, while achieving identification under above the low-level conditions.

The proof to Theorem 1 is found in section B.1, though I now sketch the key ideas. For simplicity, consider the binary choice case ($|A| = 1$). By the law of total probability the distribution of observed choices and states can be expressed as follows:

$$\begin{aligned} f_{a_2 a_1 x_2|x_1}(1, a_1, x_2; x_1) &= \int f_{a_2 a_1 x_2 \beta|x_1}(1, a_1, x_2, b; x_1) db \\ &= \int f_{a_2|a_1 x_2 x_1 \beta}(1; a_1, x_2, x_1, b) f_{x_2|a_1 x_1 \beta}(x_2; a_1, x_1, b) f_{a_1|x_1 \beta}(a_1; x_1, b) f_{\beta|x_1}(b; x_1) db \end{aligned}$$

Then under the independence conditions of Assumption I1, this simplifies to

$$f_{a_2 a_1 x_2|x_1}(1, 1, x_2; x_1) = \int P(1; x_2, b) F_{x_2}(x_2; x_1, a_1) P(a_1; x_1, b) f_{\beta|x_1}(b; x_1) db,$$

Where the notation $P(a; x, b)$ represents the conditional choice probabilities $\Pr(a_{it} = a | x_{it} = x, \beta_i = b)$. Since the state transition is assumed to be common across agents, it can be treated as observed, and whenever it has positive measure:

$$\frac{f_{a_2 a_1 x_2|x_1}(1, 1, x_2; x_1)}{F_{x_2}(x_2; x_1, a_1)} = \int P(1; x_2, b) P(a_1; x_1, b) f_{\beta|x_1}(b; x_1) db,$$

Although not necessary for the proof, if the state transition has some common support, it is possible to sum over a_1:

$$\sum_{a_1 \in \{0,1\}} \frac{f_{a_2 a_1 x_2|x_1}(1, a_1, x_2; x_1)}{F_{x_2}(x_2; x_1, a_1)} = \int P(1; x_2, b) f_{\beta|x_1}(b; x_1) db,$$

The left-hand side is observed but the right-hand is not. The ‘adequate variation’ condition for identification is precisely whether this integral operator is injective: that is, denoting the observed left-hand side function $\rho(x_2)$ with x_1 fixed, does the system $\rho = Pf$ have a unique solution f.

In the proof, it is shown that injectivity is equivalent to the functions

$$\{P \colon S_\beta \to [0, 1] : P(b) = P(1; x_2, b), x_2 \in S_2\}$$

being good approximators for square-integrable functions on S_β. The proof proceeds by showing the functions P have this universal approximation property.

To show the conditional choice probability functions are good approximators, it proves useful to exploit smoothness and the linear period payoff function. First, I show that the real analytic property of F_x is inherited by P. This is useful for the following reason. Intuitively, since real analytic functions, like polynomials, are determined by their values on an open set, their approximation qualities can be understood by considering any open set. The most convenient open set, of course, is the Euclidean space. Second, with the artificial full support, linearity is useful to constructively prove the universal approximation property.

Proving injectivity, however, is only one part of the proof. With injectivity in hand, measurement error arguments based on Hu and Schennach (2008) are used to identify $(\gamma, f_{\beta|x_1})$.

2.2 Finite horizon model

In a finite horizon dynamic discrete choice model, the agent chooses a sequence of actions $(a_{T_0}, a_{T_0+1}, \ldots, a_{T_1})$ to maximize lifetime utility:

$$V_{iT_0}(x_{T_0}, \epsilon_{T_0}) = \max_{(a_{T_0}, a_{T_0+1}, \ldots)} E\left[\sum_{t=T_0}^{T_1} \rho^t u_{it}(x_t, \epsilon_t, a_t) \mid x_{T_0}\epsilon_{T_0}, a_{T_0}\right] \tag{4}$$

where $a_t \in \{0, \ldots, |A|\} \equiv A$ is the control variable, ρ is the discount factor, $s_t = (x_t, \epsilon_t)$ is the state variable of which x_t is observed by the econometrician and ϵ_t is not, and $u_i(x_t, \epsilon_t, a_t)$ is period utility which may be agent *and* period specific (that is, non-stationary). Relative to the infinite horizon choice problem (1), there is some finite period T_1 after which the agent does not consider payoffs. Because of this, even with the conditional independence assumption (equation (3)), the problem does not admit a contraction mapping structure.

In this section I consider a finite horizon dynamic discrete choice model in which the terminal period is observed. By definition, the decision-maker has no future utility flows to consider in the terminal period and thus a different proof strategy is adopted. This argument allows for identification of random intercepts ('fixed effects'), which was not the case in the infinite horizon model. However, there are many settings where it is not realistic to expect the terminal period is observed. In this case, identification is still possible (see remark 3).

Assumption F1. *(i)* $u_t(x_t, \epsilon_t, a_t) = u_t(x_t, a_t) + \epsilon_t(a_t)$. *(ii)* ρ *is known. (iii)* $s_t = (x_t, \epsilon_t)$ *satisfies*

$$dF_{S_{t+1}}(s_{t+1}; s_t, a_t) = dF_\epsilon(\epsilon_{t+1}) dF_{x_{t+1}}(x_{t+1}; x_t, a_t).$$

(iv) $u_i(x_t, 0) = 0$. *(v) The distribution of* $\epsilon_t(a)$ *extreme value type I.*

Assumption F2. *Permanent unobserved heterogeneity* $\beta_i = \big((\beta_{1ia}, \beta_{2ia}) : a \in \tilde{A}\big) \in \mathbb{R}^b$ *for* $b = (1 + p)|A|$ *enters the model through the period utility function as follows:*

$$u_{it}(x, a) = \beta_{1ia} + x'\left(\beta_{2ia},\ \gamma_{at}\right),$$

where $x \in \mathbb{R}^k$ is the vector of observed state variables, and the agent index i is shown for explicitness. S_β, the support of β_i, is a bounded subset of $\mathbb{R}^b$. If $f_{\beta|x_1}(\cdot; x)$, the distribution of β_i conditional upon $x_1 = x$, admits a density function, it is bounded.

Assumption F2 states that permanent unobserved heterogeneity enters the model as a random coefficient. The restrictions are weaker than those in the infinite horizon model (Assumption I2). First, the permanent unobserved heterogeneity can include a random intercept. Second, the probability distribution of β_i need not be bounded. As was the case for the infinite horizon model, the support of permanent unobserved heterogeneity may be finite, but it need not be.

Assumption F3. *Let $\gamma_{aT_1,|A|}$ be the first $|A|$ components of the vector γ_{aT_1} and let Γ_{AT_1} be the $|A| \times |A|$ matrix with columns $\gamma_{aT_1,|A|}$. Then the matrix Γ_{AT_1} is full rank.*

Like Assumption I3, Assumption F3 imposes that the state variable cannot affect payoffs for each choice in a similar fashion. It is mildly weaker than its infinite-horizon counterpart. The final two assumptions place restrictions on directly observed objects.

Assumption F4. *For each x_1 and $(a_t)_{t=1}^{T_1-1}$, there is a sequence of state variables $(x_t)_{t=1+1}^{T_1}$ such that the restriction of the support of x_T to its first $p+1$ elements contains a non-empty open set.*

Assumption F4 places restrictions on the support of the observed state variable. It is substantially weaker than the assumption required for the infinite horizon model (Assumption I4).

To introduce the final assumption, denote $\tilde{A} = \{1, 2, \ldots, A\}$, S_{T_1} the support of x_{T_1} of Assumption F4. Let E be a subset of $S_{T_1} \times \tilde{A}$ whose projection on the first $p+1$ elements contains an open set. Define the operator

$$L_{T_1,\beta}^{E,\gamma} : \mathcal{L}_{S_\beta} \to \mathcal{L}_E \qquad [L_{T_1,\beta}^{E,\gamma} m](x_{T_1}) = \int f_{A_{T_1}|X_{T_1}\beta}(1; x_{T_1}, b; \gamma) m(b) db.$$

Denote $(L_{T_1,\beta}^{E,\gamma})^{-1}$ as the left inverse of $L_{T_1,\beta}^{E,\gamma}$ which exists if it is injective.

Assumption F5. *For every $\gamma \neq \tilde{\gamma}$, there exists $(E, \tilde{E}) \subseteq S_{X_{T_1}} \times \tilde{A}$ whose projections on the first $p+1$ elements of x_{T_1} are non-empty open sets such that the operator*

$$L_{T_1,\beta}^{E,\gamma,\tilde{E},\tilde{\gamma}} : \mathcal{L}_{S_\beta} \to \mathcal{L}_{S_\beta} \qquad [L_{T_1,\beta}^{E,\gamma,\tilde{E},\tilde{\gamma}} m](b) = \left[\left((L_{T_1,\beta}^{E,\gamma})^{-1} L_{T_1,\beta}^{E,\tilde{\gamma}} - (L_{T_1,\beta}^{\tilde{E},\gamma})^{-1} L_{T_1,\beta}^{\tilde{E},\tilde{\gamma}}\right) m\right](b) \qquad (5)$$

is injective.

This high-level condition ensures that the parameter γ_{T_1} can be identified without knowledge of the distribution of unobserved heterogeneity. A few comments on Assumption F5 are in order.

First, it is shown in the proof to Theorem 2 that assumptions F1-F4 imply that, for any E, $L^{E,\gamma}_{T_1,\beta}$ is injective so that $L^{E,\gamma,\tilde{E},\tilde{\gamma}}_{T_1,\beta}$ exists. Second, this assumption is not required in a model without random intercepts (Remark 3). Third, the condition is stated in terms of observed objects, and thus can be verified prior to estimation.

Fourth, the condition can be related to the high-level necessary conditions for identification of a common parameter in discrete choice panel data given in Johnson (2004) and Chamberlain (2010). To describe their result, fix $x \equiv (x_1, x_2, \dots, x_{T_1})$ and γ which is time-invariant for convenience, and let $p(\beta; x, \gamma)$ be the length 2^{T_1} vector of choice probabilities $\left\{\prod_{t=1}^{T_1} f_{A_t|X_t\beta}(a_t; x_t, b; \gamma) : (a_t)_{t=1}^{T_1} \in \{0,1\}^{T_1} \setminus \{0_{T_1}\}\right\}$ in the $(2^{T_1} - 1)$-dimensional hypercube. Johnson (2004, Theorem 2.2) states that the common parameter γ will not be identified if the set $\{p(\beta; x, \gamma) : \beta \in S_\beta\}$ does not lie in a hyperplane for some x. For the static binary choice model with $T = 2$, Chamberlain (2010) shows that the hyperplane restriction is satisfied if and only if the unobserved state variables are iid extreme-value type I. This is suggestive that the $T = 2$ dynamic binary choice model does not satisfy Johnson (2004)'s condition and therefore γ is not identified. If that is the case, then $\forall x_2 \in S_{X_2}$, $\gamma \neq \tilde{\gamma}$

$$\exists (f^{X_2}, \tilde{f}^{X_2}) : \left[L^{S_{X_2},\gamma}_{2,\beta} f^{X_2}_{\beta|X_1}(\cdot; x_1, x_2)\right](x_2) = \left[L^{S_{X_2},\tilde{\gamma}}_{2,\beta} \tilde{f}^{X_2}_{\beta|X_1}(\cdot; x_1, x_2)\right](x_2),$$

where the distribution of unobserved heterogeneity $f^{X_2}_{X_1|\beta}$ is allowed to depend on x_2, as in Johnson (2004) and Chamberlain (2010). If the distribution is restricted to be the same for all $x_2 \in S_{X_2}$, the above condition implies that for $\gamma \neq \tilde{\gamma}$, $\exists x_2 \in S_{X_2}$, $(f, \tilde{f})$ such that

$$\left[L^{S_{X_2},\gamma}_{2,\beta} f_{\beta|X_1}(\cdot; x_1)\right](x_2) = \left[L^{S_{X_2},\tilde{\gamma}}_{2,\beta} \tilde{f}_{\beta|X_1}(\cdot; x_1)\right](x_2).$$

However, since the distribution of unobserved heterogeneity is required to be the same for all x_2, there may be some other $\tilde{x}_2 \in S_{X_2}$ such that

$$\left[L^{S_{X_2},\gamma}_{2,\beta} f_{\beta|X_1}(\cdot; x_1)\right](\tilde{x}_2) \neq \left[L^{S_{X_2},\tilde{\gamma}}_{2,\beta} \tilde{f}_{\beta|X_1}(\cdot; x_1)\right](\tilde{x}_2).$$

Let $E, \tilde{E}$ be neighborhoods of $(x_2, \tilde{x_2})$, respectively. In the proof to Theorem 2 it is shown that, without knowing f or $\tilde{f}$, we know there does exist such an $\tilde{x}_2$ if the operator defined in equation (5) is injective. This can be viewed as a partial converse to Johnson (2004)'s high-level condition: in that case, without knowing f or $\tilde{f}$, we know there does *not* exist such an $\tilde{x}_2$ if their 'rank' condition does not apply. In principle, the logic of Assumption F5 can be extended to the general discrete choice panel model of Johnson (2004), if the distribution of unobserved heterogeneity is required to be independent of covariates.

Finally, should Assumption F5 not hold, I show in lemma B.4 that under Assumptions F1-F4 and a scale restriction on γ_{T_1}, that γ_{T_1} and distribution of unobserved heterogeneity is identified.

Theorem 2. *Assume the distribution of $Y \equiv (X_t, A_t)_{t=1}^{T_1}$ is observed for $1 < T_1$, generated from agents solving the model of equation (4) satisfying assumptions F1-F5. Then the homogeneous payoff parameter $(\gamma_t)_{t=1}^{T_1}$ and the conditional distribution of permanent unobserved heterogeneity $f_{\beta|x_1}$ are identified.*

Section B.2 contains the proof of Theorem 2. The outline is broadly similar to that of the proof to Theorem 1, though the details are substantially different. In particular, injectivity is shown directly exploiting the properties of the link function.

Remark 3 (Identification without the terminal period). In many empirical settings, the time horizon of the decision maker extends beyond the period of observation. For example, a worker's labor force participation decisions may not be observed for their entire working life. This poses an issue for identification since in-sample decisions reflect payoff parameters for both in- and out-of-sample time periods.

One approach to this issue is to impose restrictions on out-of-sample payoffs. Section A.2 adopts this approach and shows that the model without random intercepts is identified. That is, where the payoff function equals

$$u_i(x_{it}, a_{it}) = \beta_{ia_{it}} z_{it} + \gamma'_{a_{it}g} w_{it}.$$

A different approach is to impose that the state transition exhibits finite dependence: when multiple sequences of actions leads to the same distribution of the state variable (Arcidiacono and Ellickson 2011). Finite dependence limits the number of out-of-sample time periods that affect in-sample decisions. Section A.3 considers a limited form of finite dependence, and shows a binary choice model with random coefficients is identified.

3 Estimation

This section considers consistent estimation of the model parameters $(F_x, \gamma, f_{\beta|x_1})$ in a short panel. The distribution of $y = ((a_t, x_t)_{t=2}^{T}, a_1)$ conditional upon x_1 can be written as

$$f_{y|x_1}(y; x_1) = \int \prod_{t=2}^{T} \left(P_t(a_t; x_t, b; \gamma, F_x) F_{x_t}(x_t; x_{t-1}, a_{t-1})\right) P_1(a_1; x_1, b; \gamma, F_x) df_{\beta|x_1}(b; x_1).$$

I propose two-step sieve M-estimation based on the above expression. The first step consists of estimating the state transition F_x. The second step consists of forming the pseudo-likelihood function using the fact that the CCPs P_t are known up to the state transition and payoff parameter (F_x, γ), and using sieve M-estimation methods to estimate $(\gamma, f_{\beta|x_1})$.

It is of course possible to estimate the model in a single step as a sieve maximum likelihood problem. The advantage of the proposed two-step approach is computational: for example, in the infinite horizon case, the integrated value function does not have to be recomputed within the second step optimization. This is similar to the idea of using the Hotz and Miller (1993) inversion to avoid full solution estimation of parametric DDC models, although there may be efficiency loss for the standard pseudo likelihood estimator.

Although I show consistency for a general sieve space, this may be computationally infeasible to implement, since estimation requires computing the CCPs for every point in the support of the sieve. To circumvent this issue, I suggest a 'fixed grid' estimator, based on Heckman and Singer (1984)'s first-order monotone spline sieve, which reduces the computational burden by having a finite number of support points.

In this section, I focus on estimating the cumulative distribution function of β. While it would be possible to present conditions for consistent estimation of the density function, smoothness restrictions would rule out the possibility that β_i has discrete support, which is the standard assumption in the literature.

As a final comment, in practice there will be an approximation error in the evaluation of the CCPs. This problem is inherent to DDPs with large or infinite state spaces, and has received significant attention in the recent literature (Rust 2008; Kristensen et al. 2020). I assume away the effect of these errors on estimation — that is, that the approximation error is negligible relative to sampling error. In principle, the results of Kristensen et al. (2020) could be used to explicitly consider the effect of value function approximation error on estimation, though I do not pursue this here. Of course, the approximation error can be made arbitrarily small at increased computational cost.

3.1 A general two-step seminonparametric estimator

In this section, I briefly outline the two-step sieve M-estimator and present the general consistency result. Denote the true parameters as $\theta_0 = (F_x, \gamma, f_{\beta|x_1}) \in \Theta = \mathcal{F} \times \Gamma \times \mathcal{M}$, where $\mathcal{F}$ is the space of state transitions, $\Gamma \subseteq \mathbb{R}^p$, and $\mathcal{M}$ is the space of distribution functions on $S_\beta \times S_1$ with S_1 the support of x_1. The first step consists of forming a consistent estimator $\hat{F}_x$ for the state transition F_x. Since the state transition is directly observed, standard non-parametric methods are available. For the second step, the log-likelihood for the ith observation is

$$\psi(y_i, \hat{F}_x, \gamma, f_{\beta|x_1}) = \log \int \prod_{t=1}^{T} P_t(a_{it}, x_{it}, b; \hat{F}_x, \gamma) df_{\beta|x_1}(b; x_{i1})$$

Given a sieve space $\mathcal{M}_n$, which approximates $\mathcal{M}$ arbitrarily well for large n, the second step estimator is defined as

$$\frac{1}{n}\sum_{i=1}^{n}\psi(y_i,\hat{F}_x,\hat{\gamma},\hat{f}_{\beta|x_1}) \geq \sup_{(\gamma,f)\in\Gamma\times\mathcal{M}_n}\frac{1}{n}\sum_{i=1}^{n}\psi(y_i,\hat{F}_x,\gamma,f) - o_p(1/n) \tag{6}$$

The following result states that under standard regularity conditions, the estimator is consistent.

Theorem 3. *Let $(a_{it},x_{it}:t=1,\dots,T)_{i=1}^{n}$ be iid data generated from the DDC model satisfying either Assumptions I1-I5 or Assumptions F1-F5. If Assumptions E1-E4 hold, then the estimator $(\hat{\gamma},\hat{f}_{\beta|x_1})$ defined in equation 6 is consistent for $(\gamma,f_{\beta|x_1})$.*

The full statement of Theorem 3 and its proof are contained in Appendix C.1.

3.2 Fixed grid estimation

In this section I propose a particular choice of sieve which has the advantage of being simple to implement: the first-order monotone spline sieve. This is a popular choice of sieve for seminonparametric models, see for example Heckman and Singer (1984), Chen (2007), and Fox, Kim, and Yang (2016). To define the sieve, let $\mathcal{B}_n$ be a set of knots that partition S_β and $\mathcal{X}_n$ be a partition of S_1, the support of x_{1i}. The sieve space $\mathcal{M}_n$ is defined as follows:

$$\left\{f\colon S_\beta\times S_1\to[0,1]:f(b,x_1)=\sum_{j=1}^{B(n)}\sum_{k=1}^{X(n)}P_{j,k}1(b_j\leq b)1(x_1\in\mathcal{X}_{k,n}),P_{j,k}\geq 0,\sum_{j=1}^{B(n)}P_{j,k}=1,b_j\in\mathcal{B}_n,\mathcal{X}_{k,n}\in\mathcal{X}_n\right\}, \tag{7}$$

where the number of knots $B(n)$, $X(n)$ and their location b_j and $\mathcal{X}_{k,n}$ are tuning parameters. For given $(\mathcal{B}_n,\mathcal{X}_n)$, an element of $\mathcal{M}_n$ is a piecewise constant function with jumps of size $P_{j,k}$ at point b_j. The computational advantages of this sieve are clear: to find the supremum in (6), the CCP functions need only be evaluated for the values $b_j\in\mathcal{B}_n$. This would not be the case if the sieve space consisted of continuous functions.

A theoretical advantage of this sieve space is that many of the high-level conditions for consistency are attained as long as the number of knots does not grow too fast. See Appendix C.2 for details.

Theorem 4. *Let $(a_{it},x_{it}:t=1,\dots,T)_{i=1}^{n}$ be iid data generated from the DDC model satisfying either Assumptions I1-I5 or Assumptions F1-F5. If Assumptions E1, E3.1 and E4.1 hold, then the estimator $(\hat{\gamma},\hat{f}_{\beta|x_1})$ defined in equation 6 is consistent for $(\gamma,f_{\beta|x_1})$.*

To implement the estimator, the number and location of grid points must be chosen. For consistency, it is enough that $B(n)X(n)\log(B(n)X(n)) = o(n)$ and that the grid points become dense in the support of $\beta \times x_1$. In principle, convergence rates for this estimator could be derived to determine optimal growth rates $B(n), X(n)$, but I do not pursue this here.

For computation, it may be computationally attractive to use profiling. In particular, to form $(\hat{\gamma}, \hat{f}_{\beta|x_1})$, fix γ and let

$$\hat{f}_{\beta|x_1}(\gamma) = \arg \sup_{f \in \mathcal{M}_n} \frac{1}{n} \sum_{i=1}^{n} \psi(y_i, \hat{F}_x, \gamma, f).$$

For the sieve space (7) this is a convex optimization problem, with a unique global optimum and can be solved very efficiently. The profile estimator is formed as

$$\frac{1}{n} \sum_{i=1}^{n} \psi(y_i, \hat{F}_x, \hat{\gamma}, \hat{f}_{\beta|x_1}(\gamma)) \geq \sup_{\gamma \in \Gamma} \frac{1}{n} \sum_{i=1}^{n} \psi(y_i, \hat{F}_x, \gamma, \hat{f}_{\beta|x_1}(\gamma)) - o_p(1/n).$$

3.3 Estimating the support of unobserved heterogeneity

In the existing DDC literature, it is common to assume permanent unobserved heterogeneity is discrete. When this assumption is made, a key parameter is the number of support points of permanent unobserved heterogeneity. In practice, it is common to assume the number of support points is known, although there are methods to identify a lower bound on the number of support points (Kasahara and Shimotsu 2009; Kasahara and Shimotsu 2014; Kwon and Mbakop 2019) which have been applied in economics (Igami and Yang 2016). However, in general, these methods can only identify the number of support points if an upper bound is known. This is because there is no guarantee *a priori* that the data is rich enough to identify any arbitrarily large number of types. Intuitively, the population likelihood may be flat as a mixture component is added, but this may be because the initial likelihood had the true number of mixture components *or* because the data is not rich enough to distinguish the model from one with an additional mixture component. Technically, this issue can be resolved by a rank assumption on an unobserved matrix (Kasahara and Shimotsu 2009, Proposition 3; Kwon and Mbakop 2019, Assumption 2.1).

The purpose of this section is to show the models of Theorem 1 and Corollary 3 satisfy a condition equivalent to Kwon and Mbakop (2019, Assumption 2.1) when the distribution of unobserved heterogeneity is discrete. This means the number of types is identified, without knowledge of an upper bound on the number of types.

Corollary 1. *Assume the distribution of $Y = (x_t, a_t)_{t=1}^{T}$ is observed for $T \geq 3$, generated from the DDC model satisfying either Assumptions I1-I5 or Assumptions F1, F2.1-F4.1, F6 and F7. In* *addition, suppose that the support of β conditional upon x_1 has $R < \infty$ points of support. Then R is identified as the rank of the operator*

$$[Lu](x_3) = \int u(x_2) \frac{f_{A_3 A_2 A_1 X_3 X_2 | X_1}(a_3, a_2, a_1, x_3, x_2; x_1)}{F_{x_3}(x_3; x_2, a_2) F_{x_2}(x_2; x_1, a_1)} dx_2.$$

That is, R equals the dimension of the range of L.

The proof to Corollary 1 is found in Section C.3. The result means that the techniques of Kasahara and Shimotsu (2014) and Kwon and Mbakop (2019) can be used to consistently estimate the number of types should the applied econometrician wish to maintain the standard assumption that permanent unobserved heterogeneity is discrete. Broadly speaking, these estimators consist of forming a matrix of observed choice probabilities with values of x_3 varying over the rows, and x_2 over the columns. The identification result means that, at the population level, the rank of the matrix equals the true number of types.

4 Simulations

This section investigates the fixed grid estimator in a Monte Carlo simulation. The main goals of this section are threefold: first, to explore the finite sample performance of the estimator; second, to demonstrate the computational requirements; and, third, to verify the asymptotic results of Section 3. I simulate data using a simple labor force participation model based on Altuğ and Miller (1998, Section 6).

In each period, each individual decides whether or not to enter the labor force, upon observation of the state variable. Thus $A = \{0, 1\}$, with $a_{it} = 1$ representing that individual i enters the labor force at time t. The state variables are $s_{it} = (x_{it}, \epsilon_{it})$ where $\epsilon_{it} = (\epsilon_{0it}, \epsilon_{1it})$ is unobserved and $x_{it} \in X \subseteq \mathbb{R}^2$ is observed. The period period payoff from entering the labor market depends on individual-specific labor productivity β_i as follows:

$$u_i\left(x = (x_1, x_2), \epsilon, 1\right) = \beta_i x_1 + \gamma x_2 + \epsilon_1$$

Following the model of Altuğ and Miller (1998), x_1 can be interpreted as an average consumption value (see Section 5 for details) and x_2 is equal to the income of the primary earner in the household. The period payoff from not entering is $u_i(x, \epsilon, 0) = \epsilon_0$. The random preference shock ϵ_{ait} is assumed to be i.i.d. extreme value type I, and the agents' time horizon is assumed to be infinite. In addition, I assume that β_i is independent of x_{it} and follows a mixture of three truncated normal distributions.

In particular

$$\beta_i \sim \begin{cases} \mathcal{N}_{tr}(1.5, 1) & \text{with prob. } 1/3 \\ \mathcal{N}_{tr}(2.5, 0.25) & \text{with prob. } 1/3 \\ \mathcal{N}_{tr}(3.5, 1) & \text{with prob. } 1/3 \end{cases}$$

Where $\mathcal{N}_{tr}(\mu, \sigma)$ is the truncated normal distribution with parameters (μ, σ), minimum value 0 and maximum value 50. The simulation results are the average of 1,000 i.i.d. datasets $(a_{it}, x_{it} : t = 1, \ldots, 8)_{i=1}^{n}$ drawn from this model. Results are presented for four sample sizes: $n = 100$, $n = 500$, $n = 1,000$ and $n = 10,000$. For estimation I choose the number of grid points equal to $4n^{1/4}$, which satisfies the rate conditions required for Theorem 4.

		$n = 100$	$n = 500$	$n = 1,000$	$n = 10,000$
γ	Bias	-0.328	-0.211	-0.093	0.074
	Var	2.750	2.890	2.840	2.720
	MSE	2.860	2.930	2.850	2.730
Time		27.3	31.9	41.4	131.9
MISE		0.0754	0.040	0.032	0.020
IAE	Mean	0.458	0.334	0.302	0.240
	Min	0.255	0.199	0.199	0.18
	Max	0.925	0.520	0.482	0.337
$4n^{1/4}$		13	19	23	40
No. types	Mean	5.2	6.8	7.6	10.1
	Min	2	4	5	6
	Max	9	9	11	24

Table 2: Simulation results for estimation of γ and f_β. "γ" denotes results for estimation of γ, which includes scaled average empirical bias ("Bias"), variance ("Var") and mean-squared error ("MSE"). "Time" denotes median computation time in seconds. "MISE" denotes empirical mean integrated squared error, "IAE" denotes empirical integrated absolute error, and "No. types" denotes the number of support points.

Table 2 presents results for the estimator of (γ, f_{β_i}), in addition to computation times. First consider results for γ. Here, empirical variance is significantly larger than empirical bias, which diminishes with sample size. Scaled empirical mean squared error is largely flat across sample sizes. In terms of computational burden, the fixed grid estimator takes around 30 seconds to run for the smaller sample sizes, though it takes around 2 minutes for $n = 10,000$.

Turning to results for the estimation of f_β, both measures of integrated error diminish with sample size.[5] The number of grid points increases slowly with sample size — indeed slower than the growth of the number of support points selected by the estimator. For $n = 100$, on average 3.8 points are selected. This increases to 10.1 for the large sample size. This pattern is broadly similar to previous simulation results for a parametric variant of this estimator (Fox et al. 2011).

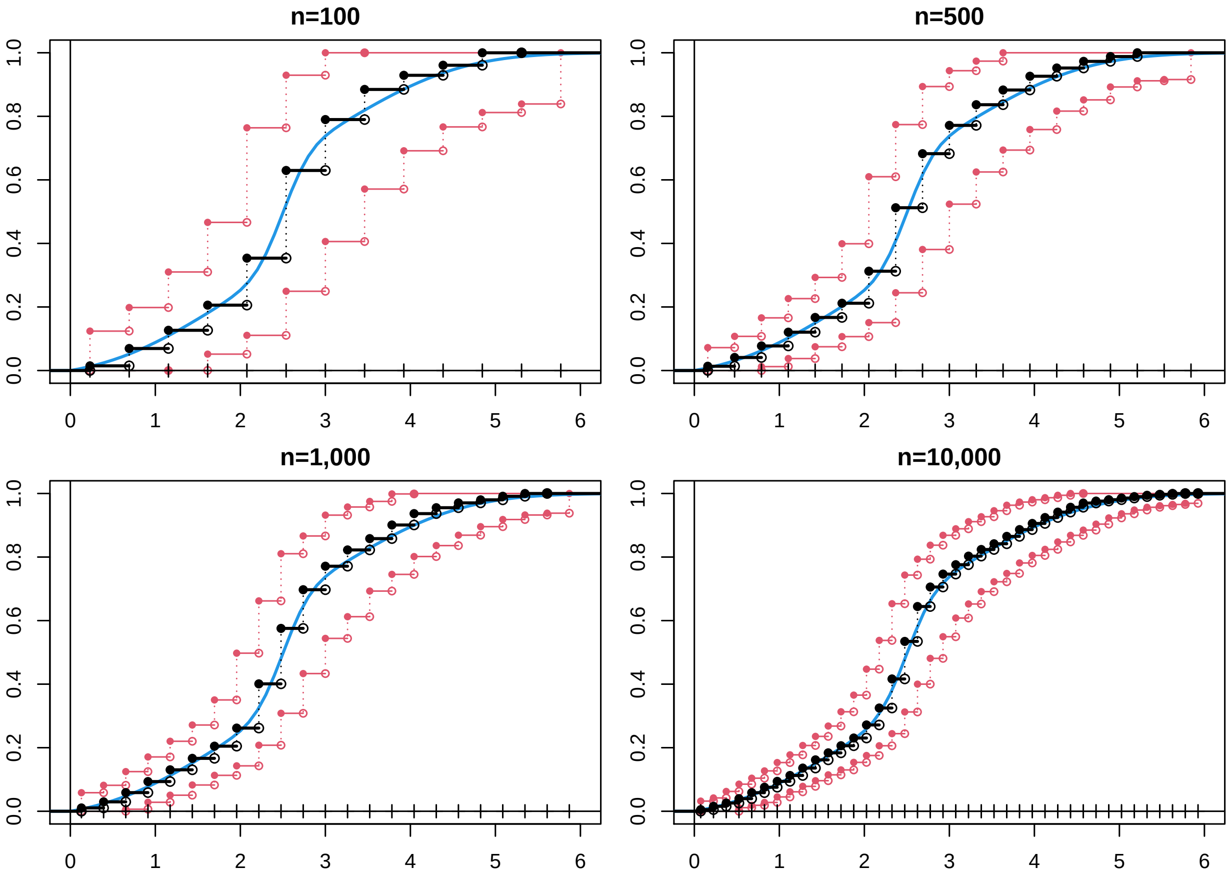

Figure 1: Simulation results for estimation of f_β for each sample size. The black curve represents the median estimate, the red curves pointwise 97.5%, 2.5% quantiles, and the blue curve the true distribution. The ticks on the x-axis represent the grid points.

Figure 1 presents empirical quantiles for the estimator of f_β. For each sample size the median estimate (the black curve) falls close to the true distribution (the blue curve). The empirical pointwise confidence bands are substantially narrower for the larger sample sizes.

[5] To be precise, integrated absolute error for simulation run m with estimate $\hat{f}_{\beta,m}$ is $\int |\hat{f}_{\beta,m}(b) - f_\beta(b)| db$ and mean integrated squared error is equal to $\frac{1}{M} \sum_{m=1}^{M} \int \left(\hat{f}_{\beta,m}(b) - f_\beta(b)\right)^2 db$ where $M = 1,000$ is the number of replications.

5 Application to a labor force participation model

This section revisits the female labor supply model of Altuğ and Miller (1998). I combine the life-cycle model of Altuğ and Miller (1998) with the identification results of Section 2 to estimate the distribution of labor productivity from data on labor force participation. Before introducing the econometric model used in this section, I discuss the approach of Altuğ and Miller (1998).

Altuğ and Miller (1998) introduces a framework to understand female labor supply that takes into account aggregate shocks and time non-separable preferences. In their model, agents gain utility from consumption and leisure. Under their specification of consumption and Pareto optimality, the consumption component of flow utility is:

$$\eta_i \lambda_t \nu_i \omega_t \exp(\gamma_3' x_{Wit}) l_{it} \tag{8}$$

The term $(\eta_i \lambda_t)$ is the shadow value of consumption, which is estimated from data on consumption. The term $(\nu_i \omega_t \exp(\gamma_3' x_{Wit}) l_{it})$ represents an individual's earnings, which is equal to the amount of time they spend working conditional on participating, l_{it}, multiplied by their marginal product. The individual-specific marginal product of labor consists of unobserved aggregate and individual productivity effects (ω_t, ν_i) in addition to a component that depends on covariates x_{it}. These terms are estimated from the wage equation, which is as follows:

$$\tilde{w}_{it} = \omega_t \nu_i \exp(\gamma_3' x_{Wit}) \exp(\tilde{\epsilon}_{it}).$$

Altuğ and Miller (1998) consider two estimators for the individual-specific productivity ν_i. First, they use the fixed effects estimator from the wage equation above. Of course, in the asymptotic framework considered in this paper where n is large but T is fixed, this estimator is subject to the incidental parameters problem is not consistent in general. Second, the authors assume that the fixed effect can be written as a deterministic function of observables, and then estimate that function non-parametrically. The observed variables consists of demographic data such as race, marital status and education levels. The second estimator is inconsistent if individual productivity cannot be written as a function of observed data. Furthermore, their estimators requires that $\tilde{\epsilon}_{it} - \tilde{\epsilon}_{i1}$ is mean independent of ν_i, which restricts an individual's wage schedule.

The identification results of Section 2 obviate the need to estimate individual-specific productivity from the wage equation. Instead, ν_i can be interpreted as a random coefficient in the discrete choice model of labor force participation. In particular, suppose the period payoff from entering the labor market for individual of type ν_i is:

$$u_i(x, \epsilon, 1) = x' (\nu_i, \gamma) + \epsilon_1 \tag{9}$$

with $x_{it} = (\hat{z}_{it}, 1, \text{hinc}_{it}, \text{age}_{it}, \text{nkids}_{it}, \text{educ}_{it}, \text{lagged.work}_{it})$. Here $\hat{z}_{it}$ is constructed following the schema of Altuğ and Miller (1998). Precisely, $\hat{z}_{it} = \hat{\eta}_i \hat{\lambda}_t \hat{\omega}_t \exp(\hat{\gamma}_3' x_{it}) \hat{l}_{it}$. The remaining observed state variables are, respectively, annual head-of-household income, age, a dummy variable for the presence of children in the household and, a dummy variable that equals 1 if the individual worked in either of the previous two periods.

Relative to the participation model in Altuğ and Miller (1998, Equation (6.7)), ν_i is treated as an unobserved random variable. In their model ν_i is replaced by first-stage estimator $\hat{\nu}_i$, so that $\nu_i \hat{z}_{it}$ is treated as a known constant. Like Altuğ and Miller (1998), I make the outside good assumption and assume that ϵ_{ait} is i.i.d. extreme value type I. For simplicity, the agents' time horizon is assumed to be infinite.

As in Altuğ and Miller (1998), the labor force participation model is estimated using a subset of data from the PSID. The construction of the subset largely followed the details in Altuğ and Miller (1998, Appendix B). The final data set contains 3084 individuals, each of whom have between four and ten panel observations, with an average close to eight. For estimation, the sieve space is chosen to have 40 grid points.

Variable	Point estimate	Standard errors
Intercept	-0.4307	0.0774
Head-of-household income	2.530×10^{-3}	$1.5572{\times}10^{-6}$
Age	0.1833	0.0419
Number of kids	-0.7024	0.0277
Education	0.2358	0.0499
Lagged work status	1.4462	0.0440

Table 3: Estimates of γ.

Table 3 presents point estimates of the finite parameter γ alongside bootstrapped standard errors. The results can be compared to Altuğ and Miller (1998, Section 6). Some results are broadly similar: for example, both sets of result indicate that disutility from work increases with age. Others are different: for example, the positive coefficient on lagged work status can be interpreted as indicating that current and previous work are complements. That is, work in the past increases utility from work in the present. This is in contrast to results in Altuğ and Miller (1998), which suggest that current and past leisure choices are substitutes.

Figure 2 presents the estimated distribution of labor productivity. The estimator puts mass on nine out of the 40 points of support.

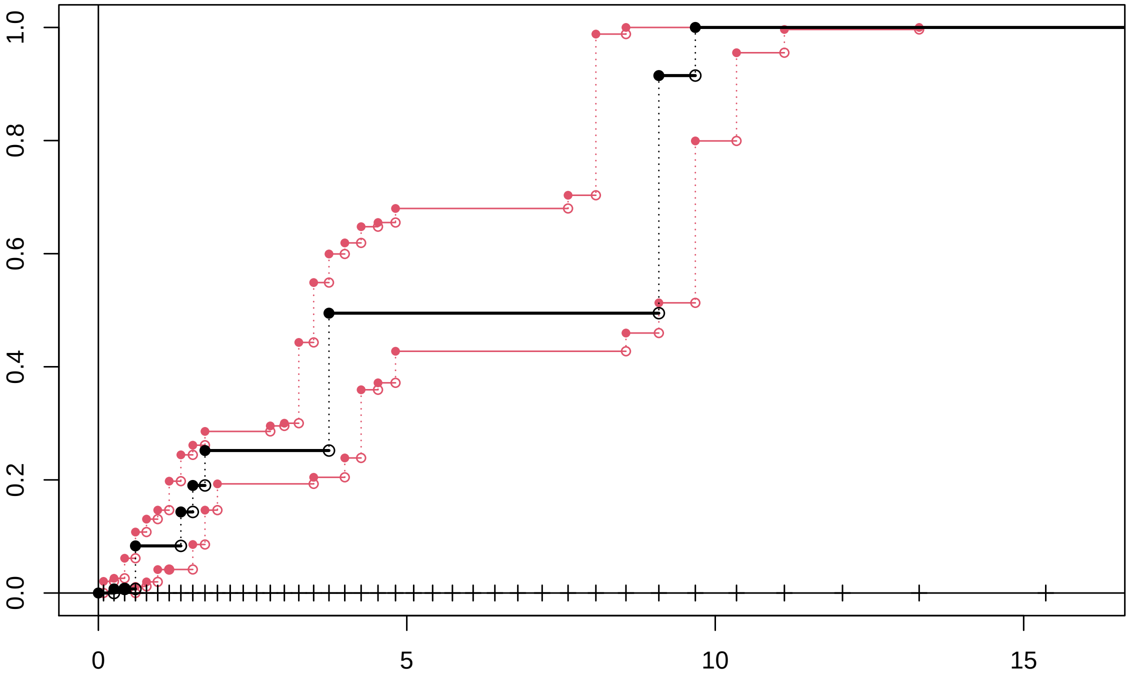

Figure 2: Estimated distribution of ν_i. The black curve represents the point estimate, the red curves represent bootstrapped 95% pointwise confidence intervals. The ticks on the x-axis represent the grid points.

6 Conclusion

In this paper I show point identification of the distribution of continuous unobserved heterogeneity in a dynamic discrete choice model in a short panel. This improves upon existing methods (Kasahara and Shimotsu 2009; Hu and Shum 2012), in not restricting permanent unobserved heterogeneity in to have finite support. Unlike these canonical papers which exploit only the Markov structure of DDC models, I impose common functional form assumptions made in the DDC literature and show that they have identifying power. This result may be surprising due to the non-linearity of DDC models. My results encompass both finite and infinite horizon models, and do not rely on a full support condition, nor parametric assumptions on the distribution on permanent unobserved heterogeneity.

I propose a seminonparametric estimator for the distribution of continuous permanent unobserved heterogeneity in the style of Heckman and Singer (1984). The estimator is computationally simple, and coincides with the estimator for a semiparametric model. As a result, the applied econometrician can proceed as they would for discrete permanent unobserved heterogeneity, providing they commit to increasing the number of support points as the sample size grows. In this way, my paper provides a solution to the problem of choosing the number of support points for discrete permanent unobserved heterogeneity.

References

Aguirregabiria, V., Gu, J., and Luo, Y. (2020). "Sufficient statistics for unobserved heterogeneity in structural dynamic logit models". *Journal of Econometrics*.

Aguirregabiria, V. and Mira, P. (2002). "Swapping the nested fixed point algorithm: A class of estimators for discrete Markov decision models". *Econometrica* 70.4, pp. 1519–1543.

— (2010). "Dynamic discrete choice structural models: A survey". *Journal of Econometrics* 156.1, pp. 38–67.

Altuğ, S. and Miller, R. A. (1998). "The effect of work experience on female wages and labour supply". *The Review of Economic Studies* 65.1, pp. 45–85.

Arcidiacono, P. and Ellickson, P. B. (2011). "Practical methods for estimation of dynamic discrete choice models". *Annu. Rev. Econ.* 3.1, pp. 363–394.

Arcidiacono, P. and Miller, R. A. (2011). "Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity". *Econometrica* 79.6, pp. 1823–1867.

Chamberlain, G. (1980). "Analysis of Covariance with Qualitative Data". *The Review of Economic Studies* 47.1, pp. 225–238.

— (2010). "Binary response models for panel data: Identification and information". *Econometrica* 78.1, pp. 159–168.

Chen, X. (2007). "Large sample sieve estimation of semi-nonparametric models". *Handbook of econometrics* 6, pp. 5549–5632.

Chernozhukov, V., Fernández-Val, I., Hahn, J., and Newey, W. (2013). "Average and quantile effects in nonseparable panel models". *Econometrica* 81.2, pp. 535–580.

Einav, L., Finkelstein, A., and Mahoney, N. (2018). "Provider Incentives and Healthcare Costs: Evidence From Long-Term Care Hospitals". *Econometrica* 86.6, pp. 2161–2219.

Einav, L., Finkelstein, A., and Schrimpf, P. (2015). "The response of drug expenditure to nonlinear contract design: Evidence from Medicare Part D". *The Quarterly Journal of Economics* 130.2, pp. 841–899.

Fox, J., Kim, K., Ryan, S., and Bajari, P. (2011). "A simple estimator for the distribution of random coefficients". *Quantitative Economics* 2.3, pp. 381–418.

Fox, J., Kim, K. I., Ryan, S., and Bajari, P. (2012). "The random coefficients logit model is identified". *Journal of Econometrics* 166.2, pp. 204–212.

Fox, J. T., Kim, K. I., and Yang, C. (2016). "A simple nonparametric approach to estimating the distribution of random coefficients in structural models". *Journal of Econometrics* 195.2, pp. 236–254.

Gautier, E. and Kitamura, Y. (2013). "Nonparametric estimation in random coefficients binary choice models". *Econometrica* 81.2, pp. 581–607.

Heckman, J. and Singer, B. (1984). "A method for minimizing the impact of distributional assumptions in econometric models for duration data". *Econometrica*, pp. 271–320.

Heckman, J. J., Lochner, L., and Taber, C. (1998). "Explaining rising wage inequality: Explorations with a dynamic general equilibrium model of labor earnings with heterogeneous agents". *Review of economic dynamics* 1.1, pp. 1–58.

Hotz, V. J. and Miller, R. A. (1993). "Conditional choice probabilities and the estimation of dynamic models". *The Review of Economic Studies* 60.3, pp. 497–529.

Hu, Y. and Schennach, S. M. (2008). "Instrumental variable treatment of nonclassical measurement error models". *Econometrica* 76.1, pp. 195–216.

Hu, Y. and Shum, M. (2012). "Nonparametric identification of dynamic models with unobserved state variables". *Journal of Econometrics* 171.1, pp. 32–44.

Ichimura, H. and Thompson, T. S. (1998). "Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution". *Journal of Econometrics* 86.2, pp. 269–295.

Igami, M. and Yang, N. (2016). "Unobserved heterogeneity in dynamic games: Cannibalization and preemptive entry of hamburger chains in Canada". *Quantitative Economics* 7.2, pp. 483–521.

Johnson, E. G. (2004). "Identification in discrete choice models with fixed effects". *Working paper, Bureau of Labor Statistics.* Citeseer.

Kasahara, H. and Shimotsu, K. (2009). "Nonparametric identification of finite mixture models of dynamic discrete choices". *Econometrica* 77.1, pp. 135–175.

— (2014). "Non-parametric identification and estimation of the number of components in multivariate mixtures". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.1, pp. 97–111.

Keane, M. P. and Wolpin, K. I. (1997). "The career decisions of young men". *Journal of Political Economy* 105.3, pp. 473–522.

Keane, M. P. and Wolpin, K. I. (2009). “Empirical applications of discrete choice dynamic programming models”. *Review of Economic Dynamics* 12.1, pp. 1–22.

Kennan, J. and Walker, J. R. (2011). “The effect of expected income on individual migration decisions”. *Econometrica* 79.1, pp. 211–251.

Kristensen, D., Mogensen, P. K., Moon, J. M., and Schjerning, B. (2020). “Solving dynamic discrete choice models using smoothing and sieve methods”. *Journal of Econometrics*.

Kwon, C. and Mbakop, E. (2019). “Estimation of the Number of Components of Non-Parametric Multivariate Finite Mixture Models”. *arXiv:1908.03656*.

Lee, D. and Wolpin, K. I. (2006). “Intersectoral labor mobility and the growth of the service sector”. *Econometrica* 74.1, pp. 1–46.

Lewbel, A. (2000). “Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables”. *Journal of Econometrics* 97.1, pp. 145–177.

Magnac, T. and Thesmar, D. (2002). “Identifying dynamic discrete decision processes”. *Econometrica* 70.2, pp. 801–816.

Nevo, A., Turner, J. L., and Williams, J. W. (2016). “Usage-based pricing and demand for residential broadband”. *Econometrica* 84.2, pp. 411–443.

Norets, A. (2010). “Continuity and differentiability of expected value functions in dynamic discrete choice models”. *Quantitative economics* 1.2, pp. 305–322.

Norets, A. and Tang, X. (2013). “Semiparametric inference in dynamic binary choice models”. *Review of Economic Studies* 81.3, pp. 1229–1262.

Rust, J. et al. (1994). “Structural estimation of Markov decision processes”. *Handbook of econometrics* 4.4, pp. 3081–3143.

Rust, J. (2008). “Dynamic programming”. *The New Palgrave Dictionary of Economics* 1, p. 8.

Scott, P. (2014). “Dynamic discrete choice estimation of agricultural land use”.

Stinchcombe, M. and White, H. (1998). “Consistent specification testing with nuisance parameters present only under the alternative”. *Econometric Theory* 14.3, pp. 295–325.

Todd, P. E. and Wolpin, K. I. (2006). “Assessing the impact of a school subsidy program in Mexico: Using a social experiment to validate a dynamic behavioral model of child schooling and fertility”. *American Economic Review* 96.5, pp. 1384–1417.

Traiberman, S. (2019). “Occupations and import competition: Evidence from Denmark”. *American Economic Review* 109.12, pp. 4260–4301.

Walters, C. R. (2018). “The demand for effective charter schools”. *Journal of Political Economy* 126.6, pp. 2179–2223.

Williams, B. (2019). "Nonparametric identification of discrete choice models with lagged dependent variables". *Journal of Econometrics.*

A Supplementary identification results

A.1 Random intercepts

This subsection provides conditions for identification of an infinite-horizon DDC model with random intercepts (fixed effects) (see remark 1).

Assumption I2.1. *Permanent unobserved heterogeneity* $\beta_i = \left(\beta_{ia} : a \in \tilde{A}\right) \in \mathbb{R}^b$ *for* $b = |A|$ *enters the model through the period utility function as follows.*

$$u_i(x, a) = \beta_{ia} + x'\gamma_a$$

where $x \in \mathbb{R}^k$ *is the vector of observed state variables, and the agent index* i *is shown for explicitness.* S_β*, the support of* β_i*, is a bounded subset of* $\mathbb{R}^b$*.* β_i *conditional upon* $x_1 = x$ *is either discrete or absolutely continuous, in which case its density function* $f_{\beta|x_1}$ *is bounded.*

Assumption I3.1. *Let* $\gamma_{a|A|}$ *be the first* $|A|$ *components of the vector* γ_a *and let* Γ_A *be the* $|A| \times |A|$ *matrix with columns* $\gamma_{a|A|}$*. Then the matrix* Γ_A *is full rank.*

Assumption I4.1. *(i) The restriction of the support of* x_2 *conditional upon* $(x_1, a_1) = (x, a)$ *to the first* $1 + |A|$ *elements of* x_2 *is bounded and contains a non-empty open set for which* $F_x(x'; x, a)$ *does not depend on* a*. (ii) The support of* x_3 *conditional upon* $(x_2, a_2) = (x, 0)$ *for some* x *in the support of part (i) contains* k *linearly independent elements and its restriction to the first* $1 + |A|$ *elements of* x_3 *is bounded and contains a non-empty open set for which* $F_x(x'; x, a)$ *does not depend on* a*. (iii) The intersection over* a_3 *of the support of* x_4 *conditional upon* x_3 *in the support of part (ii) and* a_3 *contains* k *linearly independent elements.*

This strengthens Assumption I4 by requiring the state transition to be constant across choices.

Corollary 2. *Assume the distribution of* $(x_t, a_t)_{t=1}^T$ *is observed for* $T \geq 4$*, generated from agents solving the model of equation (1) satisfying assumptions I1, I2.1, I3.1 and I4.1. Then* $(\gamma, f_{\beta|x_1})$ *is point identified.*

The proof to Corollary 2 is contained in section B.3.2. It follows from the proofs of Theorems 1 and 2.

A.2 Identification without the terminal period

In many realistic contexts the terminal period is not observed. This is the model I consider in this section. The result follows from arguments similar to the proof of Theorem 1, and the assumptions reflect this. First the condition on permanent unobserved heterogeneity is strengthened relative to Assumption F2. In particular, random intercepts are ruled out.

Assumption F2.1. *Permanent unobserved heterogeneity $\beta_i = (\beta_{ia} : a \in \tilde{A}) \in \mathbb{R}^b$ for $b = |A|$ enters the model through the period utility function as follows:*

$$u_{it}(x,a) = x' \left(\beta_{ia}, \gamma_{at}\right),$$

where $x \in \mathbb{R}^k$ is the vector of observed state variables, and the agent index i is shown for explicitness. S_β, the support of β_i, is a bounded subset of $\mathbb{R}^b$. β_i conditional upon $x_1 = x$ is either discrete or absolutely continuous, in which case its density function $f_{\beta|x_1}$ is bounded.

The next three assumptions are similar to Assumptions I3-I5

Assumption F3.1. *Let $\gamma_{at|A|}$ be the first $|A|$ components of the vector γ_{at} and let Γ_{At} be the $|A|\times|A|$ matrix with columns $\gamma_{at|A|}$. Then the matrix Γ_{At} has full rank, as do all of its principal submatrices.*

Assumption F4.1. *(i) The restriction of the support of x_2 conditional upon $(x_1, a_1) = (x,a)$ to the first $1+|A|$ elements of x_2 contains k linearly independent elements. (ii) The intersection over $a_2 \in A$ of the support of x_3 conditional upon $(x_2, a_2) = (x,a)$ for some x in the support of part (i) restricted to the first $1+|A|$ elements of x_3 is bounded and contains a non-empty open set. (iii) The support of x_4 conditional upon x_3 in the support of part (ii) and $a_3 = 0$ contains k linearly independent elements and its restriction to the first $1+|A|$ elements of x_4 is bounded and contains a non-empty open set.*

Assumption F6. *For each t, the state transition kernel $F_{x_{t+1}}(x_{t+1}; x_t, a_t)$ has bounded support and may be decomposed into absolutely continuous and discrete components, and the associated density and probabilities are real analytic functions of the first $1+|A|$ elements of x_t. Furthermore, these functions have analytic continuations to $\mathbb{R}^{1+|A|}$ which are bounded.*

These assumptions are very similar to Assumptions I2-I5, the difference being that the homogenous parameter γ_t and the transition kernel F_{x_t} are non-stationary. Since we do not observe behavior in periods $(T+1, \ldots, T_1)$, the following restriction is placed on out-of-sample behavior:

Assumption F7. *Let $\gamma_t = (\gamma_{at} : a \in \tilde{A})$. For all $t \in (T+1, \dots, T_1)$, $\gamma_t = \gamma_T$. In addition, $F_{T+1}(x'; x, a)$ is identified.*

No restriction is placed on the homogeneous parameter in periods $t < 1$, since it has no bearing on in-sample behavior. This type of restriction is avoided in other 'censored' finite horizon models by exploiting features of the transition function, such as finite dependence (Arcidiacono and Miller 2020). Identification of the state kernel is typically attained by assuming the data is of the form $(x_{it}, a_{it}, x_{i,t+1} : i = 1, \dots, N; t = 1, \dots, T)$. Since I do not adopt this structure, I directly assume identification of the final period transition. Let $\gamma = (\gamma_t)_{t=1}^T$

Corollary 3. *Assume the distribution of $(x_t, a_t)_{t=1}^T$ is observed for $T = 4$, generated from agents solving the model of equation (1) satisfying assumptions F1, F2.1-F4.1, F6 and F7. Then $(\gamma, f_{\beta|X_1})$ is point identified.*

The argument for Corollary 3 is found in section B.3.1. It is broadly similar to the argument for Theorem 1. For notational simplicity I assume exactly 4 periods are observed, the same arguments apply if additional periods are observed.

A.3 Finite dependence

A DDC model exhibits finite dependence if there are multiple sequences of actions that yield the same distribution over the state variable. Finite dependence is useful for estimation as it allows the continuation value term to be expressed in terms of CCPs (Arcidiacono and Ellickson 2011). This fact also makes finite dependence useful for identification in models without permanent unobserved heterogeneity, as it reduces the number of periods of out-of-sample behavior that must be assumed known (Arcidiacono and Miller 2020, Section 3.3).

In this section I show a similar feature is present for models with continuous permanent unobserved heterogeneity. In particular, I assume the transition function exhibits a special case of finite-dependence: the renewal action. The canonical example of renewal is machine replacement, but turnover and job matching also display this pattern (Arcidiacono and Miller 2020).

Assumption F7.1. *There is a choice $a \in A$ such that the transition does not depend on the initial state: $F_{x_t}(x'; x, a) = F_{x_t}(x'; \tilde{x}, a)$ for all $(x, \tilde{x})$ in the support of x_t.*

Let $\gamma = (\gamma_t)_{t=t_0}^{t_1 - 1}$.

Corollary 4. *Assume the distribution of $(x_t, a_t)_{t=1}^4$ is observed, generated from agents solving the model of equation (1) satisfying assumptions F1, F2.1, F3.1, I4, F6 and F7.1. Then $(\gamma, f_{\beta|X_1})$ is point identified.*

As before, a panel of length 4 is assumed for notational ease, but the arguments also apply for longer panels.

Section B.3.3 contains the proof to corollary 4. The most substantial steps follow the proof of Theorem 1. The key difference is in showing identification of the finite parameter γ.

B Identification proofs

Notation $P_t(a; x, b)$ denotes the conditional choice probabilities at period t, that is $\Pr(a_{it} = a \mid x_{it} = x, \beta_i = b)$.

B.1 Infinite horizon model

Proof of Theorem 1. By assumption I1,

$$\begin{aligned} f_{a_4 a_3 a_2 a_1 x_4 x_3 x_2 | x_1}(a_4, a_3, 0, a_1, x_4, x_3, x_2; x_1) = \int & P(a_4; x_4, b) F_x(x_4; x_3, a_3) P(a_3; x_3, b) \\ & \times F_x(x_3; x_2, 0) P(0; x_2, b) F_x(x_2; x_1, a_1) P(a_1; x_2, b) f_{\beta|x_1}(db; x_1) \end{aligned}$$

Where the transition kernel has positive measure, we can write

$$\frac{f_{a_4 a_3 a_2 a_1 x_4 x_3 x_2 | x_1}(a_4, a_3, 0, a_1, x_4, x_3, x_2; x_1)}{F_x(x_4; x_3, a_3) F_x(x_3; x_2, 0) F_x(x_2; x_1, a_1)} = \int P(a_4; x_4, b) P(a_3; x_3, b) P(0; x_2, b) P(a_1; x_2, b) f_{\beta|x_1}(db; x_1)$$

In this structure, the choices and states (a_t, x_t) can be framed as repeated measurements of β_i, and measurement error methods can be adapted to prove identification. To this end, denote $\mathcal{L}_{\mathcal{A}} = \{f : \mathcal{A} \to \mathbb{R} : \sup_{a \in \mathcal{A}} |f(a)| < \infty\}$, S_3 the support of x_3 satisfying Assumption I4(ii), and S_2 the support of x_2 satisfying Assumption I4(iii). Let $L_{3,4,2} : \mathcal{L}_{S_2} \to A \times \mathcal{L}_{S_3}$ and $L_{3,2} : \mathcal{L}_{S_2} \to A \times \mathcal{L}_{S_3}$ be defined as follows:

$$\begin{aligned} [L_{3,4,2} m](a_3, x_3) &= \int \frac{f_{a_4 a_3 a_2 a_1 x_4 x_3 x_2 | x_1}(a_4, a_3, 0, a_1, x_4, x_3, x_2; x_1)}{F_x(x_4; x_3, a_3) F_x(x_3; x_2, 0) F_x(x_2; x_1, a_1)} m(x_2) dx_2 \\ [L_{3,2} m](a_3, x_3) &= \int \frac{f_{a_3 a_2 a_1 x_4 x_3 x_2 | x_1}(a_3, 0, a_1, x_4, x_3, x_2; x_1)}{F_x(x_3; x_2, 0) F_x(x_2; x_1, a_1)} m(x_2) dx_2 \end{aligned}$$

Under Assumption I4 the above operators are observed and well-defined for $x_4 \in S_4$ where S_4 is the support of X_4 satisfying Assumption I4(iii).

These operators can be decomposed into constituent parts. For this purpose define

$$\begin{aligned}
L_{3,\beta} : \mathcal{L}_{S_\beta} \to A \times \mathcal{L}_{S_3} &\qquad [L_{3,\beta}m](a_3,x_3) = \int P(a_3;x_3,b)m(b)db \\
D_\beta^4 : \mathcal{L}_{S_\beta} \to \mathcal{L}_{S_\beta} &\qquad [D_\beta^4 m](b) = P(a_4;x_4,b)m(b) \\
D_\beta : \mathcal{L}_{S_\beta} \to \mathcal{L}_{S_\beta} &\qquad [D_\beta m](b) = P(a_1;x_1,b)f_{\beta|x_1}(b;x_1)m(b) \\
L_{\beta,2} : \mathcal{L}_{S_2} \to \mathcal{L}_{S_\beta} &\qquad [L_{\beta,2}m](b) = \int P(0;x_2,b)m(x_2)dx_2
\end{aligned}$$

It is straightforward to derive that $L_{3,4,2} = L_{3,\beta}D_\beta^4 D_\beta L_{\beta,2}$ and $L_{3,2} = L_{3,\beta}D_\beta L_{\beta,2}$.

The proof proceeds in two steps. First it is shown that the operators $L_{3,\beta}$ and $L_{\beta,2}^*$ are injective, where $L_{\beta,2}^*$ is the adjoint[6] of $L_{\beta,2}$. As argued below, injectivity of these operators implies that $L_{3,2}$ has a right inverse: In particular, that the equivalency $L_{4,3,2}L_{3,2}^{-1} = L_{3,\beta}D_\beta^4 L_{3,\beta}^{-1}$ holds. The second step of the proof is to use this eigendecomposition to identify the CCP functions P, and subsequently $(\gamma, f_{\beta|x_1})$.

Part I: Injectivity of $L_{3,\beta}$ and $L_{\beta,2}^*$

$L_{\beta,2}^*$ is defined as

$$L_{\beta,2}^* : \mathcal{L}_{S_\beta} \to \mathcal{L}_{S_2} \qquad [L_{\beta,2}^* m](x_2) = \int P(0;x_2,b)m(b)db.$$

Given the common structure of $L_{3,\beta}$ and $L_{\beta,2}^*$ when $a_3 = 0$, set $a_3 = 0$ and the following argument applies for $t = 2, 3$.

Define $\mathcal{H}$, a subset of functions $S_\beta \to [0,1]$, as

$$\mathcal{H} = \left\{h \colon S_\beta \to [0,1] : h(b) = P(0;x,b), x \in S_t\right\}. \tag{10}$$

Lemma B.1 shows $\mathcal{H}$ is a subset of $L^2(S_\beta)$, the square integrable functions on measure space $(S_\beta, \mathcal{B}, \lambda)$ where $\mathcal{B}$ is the Borel sigma field on S_β and λ is the Lebesgue measure. In the language of Stinchcombe and White (1998, Definition 2.1), $\mathcal{H}$ is *totally revealing* if and only if the operator is injective.

Now consider a superset of $\mathcal{H}$, $\tilde{\mathcal{H}}$ defined as

$$\tilde{\mathcal{H}} = \left\{h \colon S_\beta \to [0,1] : h(b) = P(0;x,b), x \in \mathbb{R}^{b+1} \times S_t^r\right\}, \tag{11}$$

where S_t^r is the restriction of S_t to the final $k - (1 + |A|)$ elements of x_t. Lemma B.3 implies that if the functions $P(a;b,x) : \mathbb{R}^{b+1} \to [0,1]$ are real analytic functions in the first $1 + |A|$ elements

[6] The adjoint of a linear operator between Hilbert Spaces $L : U \to V$ is the operator $L^* : V \to U$ that satisfies $\langle Lu, v\rangle_V = \langle u, L^* v\rangle_U$ where $\langle \cdot,\cdot\rangle_W$ is the inner product on W. See Carrasco, Florens, and Renault (2007) for further discussion.

of x and the restriction of S_t to those elements contains a non-empty open set, then $\tilde{\mathcal{H}}$ is totally revealing if and only if $\mathcal{H}$ is totally revealing. Lemma B.1 verifies that these functions are indeed real analytic and Assumption I4(i),(ii) ensures the open set condition is satisfied, so it remains to show $\tilde{\mathcal{H}}$ is totally revealing.

Stinchcombe and White (1998, Theorem 3.1) states that a norm bounded subset of L^2 is totally revealing if and only if its span is weakly dense in L^2. Lemma B.1 verifies $\tilde{\mathcal{H}}$ is a norm bounded subset of L^2, thus it is sufficient to show the weak density of $\tilde{\mathcal{H}}$ in $L^2(S_\beta)$.

The first step (lemma B.2) is to show that the span of $\tilde{\mathcal{H}}$ is uniformly dense in the set

$$\mathcal{H}_1 = \left\{h\colon S_\beta \to [0,1] : h(b) = \widetilde{\cos}(\alpha_1' b + \alpha_0), (\alpha_1, \alpha_0) \in \mathbb{R}^{b+1}\right\}, \tag{12}$$

where $\widetilde{\cos}(x) = (1+\cos[x+3\pi/2])(1/2)1_{|x|\leq\pi/2} + 1_{x>\pi/2}$ is the cosine squasher of Hornik, Stinchcombe, and White (1989). Next, observe that on any compact domain, a finite linear combination of elements of $\mathcal{H}_1$ can be made equal to any element of

$$\mathcal{H}_2 = \left\{h\colon S_\beta \to [-\pi,\pi] : h(b) = \cos(\alpha_1' b + \alpha_0), (\alpha_1, \alpha_0) \in \mathbb{R}^{b+1}\right\}.$$

We thus have the following containment: $\overline{\mathrm{sp}}\tilde{\mathcal{H}} \supset \mathrm{sp}\mathcal{H}_1 \supset \mathcal{H}_2$ where $\mathrm{sp}\mathcal{A}$ is the linear span of $\mathcal{A}$, and $\overline{\mathrm{sp}}\mathcal{A}$ is its uniform closure. It is simple to verify that the linear span of $\mathcal{H}_2$ satisfies the conditions of the Stone-Weierstrass theorem, and thus is uniformly dense in continuous functions on S_β, $C(S_\beta)$ (Rudin 1964, Theorem 7.32). That is $\overline{\mathrm{sp}}\mathcal{H}_2 \supset C(S_\beta)$ and it follows from the previous containment that $\overline{\mathrm{sp}}\tilde{\mathcal{H}} \supset C(S_\beta)$ — that the span of $\tilde{\mathcal{H}}$ is uniformly dense in continuous functions on S_β. Uniform density in $L^2(S_\beta)$ follows from Hornik, Stinchcombe, and White (1989, Corollary 2.2). Finally, since the uniform closure of a set is contained within its weak closure, uniform denseness of $\tilde{\mathcal{H}}$ in $L^2(S_\beta)$ implies weak denseness and we conclude $L_{3,\beta}$ and $L^*_{\beta,2}$ are injective.

Now suppose that the measure $f_{\beta|X_1}(b;x_1)$ has $S < \infty$ points of support. In this case, the operators $L_{3,\beta}$ and $L^*_{\beta,2}$ are a matrix of probabilities with rows $(P(0;x_t,b_s))_{S=1,\dots,S}$. Let $x_t = (z_t, w_t)$ with z_t the first $b+1$ elements of x_t. From the above approximation result, for each s, a sequence of $z_{n,s,t} \in \mathbb{R}^{b+1}$ can be found such that $\lim P(0;(z_{n,s,t},w_t),b_{s_+}) = 1$ for $s_+ \geq s$ and $\lim P(0;(z_{n,s,t},w_t),b_{s_-}) = 0$ for $s_- < s$. For each t, these S sequences define a sequence of square matrices whose limit is full rank. Therefore for n large enough, the matrix $(P(0;(z,w_t),b_s))_{z\in z_{n,s,t};s=1,\dots,S}$ is full rank.

Part II: Eigendecomposition

Since D_β is invertible (as $P(a_1;x_1,b)f_{\beta|X_1}(b;x_1) > 0$ almost surely-S_β), and $L_{3,\beta}$ and $L^*_{\beta,2}$ are injective, $L_{3,2}$ has a right inverse, the equivalence

$$L_{4,3,2}L_{3,2}^{-1} = L_{3,\beta}D_\beta^4 L_{3,\beta}^{-1} \tag{13}$$

holds, and $L_{4,3,2}L_{3,2}^{-1}$ admits a unique spectral decomposition (Williams 2019, Lemma A.1). In particular, the right-hand side is the eigenvalue-eigenfunction decomposition of the operator $L_{4,3,2}L_{3,2}^{-1}$. The eigenfunctions are $(a_3, x_3) \mapsto P(a_3; x_3, b)$ corresponding to the eigenvalue $P(a_4; x_4, b)$. Each b indexes an eigenvalue and the corresponding eigenfunction $(a_3, x_3) \mapsto P(a_3; x_3, b)$. As in Hu and Schennach (2008), the decomposition is unique up to (1) uniqueness of the eigenvalues, (2) scaling of the eigenfunctions and (3) a reindexing of the eigenvalues ("ordering").

The uniqueness problem is that if two eigenfunctions share the same eigenvalue, then any linear combination of the eigenfunctions is also an eigenfunction. For eigenvalue uniqueness it is sufficient that for each $b \neq \tilde{b} \in S_\beta \subseteq \mathbb{R}^b$, there exist some $(a_4, x_4) \in A \times \mathbb{R}^k$ such that $P(a_4; x_4, b) \neq P(a_4; x_4, \tilde{b})$ (Hu and Schennach 2008). This condition is exactly the condition that homogeneous parameters can be identified from the conditional choice probabilities. It applies in the model under consideration here due to assumption I4(iii), due to the argument provided below for identification of the ordering function.

The scale problem is that each eigenfunction may be multiplied by a constant (that may depend on the eigenvalue), yielding a different eigenvalue-eigenfunction decomposition that is nevertheless consistent with equation (13). If $s(b)$ is the unknown constant, then we conclude 'identification up to scale' means $s(b)P(a_3; x_3, b)$ is identified. The scale of the eigenfunctions is set by the requirement that $\sum_{a_3 \in A} P(a_3; x_3, b) = 1$.

The problem of ordering is that the index for the eigenvalues β can be reordered by some function R yielding a decomposition consistent with equation (13). To be more explicit, for any injective function R that generates another index $\tilde{\beta}$ as $\beta = R(\tilde{\beta})$, it holds that $L_{3,\beta} D_\beta^4 L_{3,\beta}^{-1} = L_{3,\tilde{\beta}} D_{\tilde{\beta}}^4 L_{3,\tilde{\beta}}^{-1}$[7] where

$$\begin{aligned} L_{3,\tilde{\beta}} : \mathcal{L}_{S_{\tilde{\beta}}} \to A \times \mathcal{L}_{S_3} \qquad & [L_{3,\tilde{\beta}} m](a, x) = \int \Pr(a_{i3} = a \mid x_{i3} = x, \tilde{\beta}_i = b) m(b) db \\ D_{\tilde{\beta}}^4 : \mathcal{L}_{S_{\tilde{\beta}}} \to \mathcal{L}_{S_{\tilde{\beta}}} \qquad & [D_{\tilde{\beta}}^4 m](b) = \Pr(a_{i4} = a_4 \mid x_{i4} = x_4, \tilde{\beta}_i = b) m(b) \end{aligned}$$

Notice that $\Pr(a_{i3} = a \mid x_{i3} = x, \tilde{\beta}_i = b) = \Pr(a_{i3} = a \mid x_{i3} = x, \beta_i = R(b)) = P(a; x, R(b))$, so 'identification up to ordering' means the function $P(a_3; x_3, R(b))$ is identified with the injective function R unknown.

To show R is identified, suppose that for all $(a_3, x_3) \in A \times S_3$,

$$P(a_3; x_3, R(b)) = P(a_3; x_3, b).$$

[7] This equality is shown explicitly in Hu and Schennach (2008, Supplement S.3)

By standard arguments for identification of homogenous parameters in DDC models (e.g. Bajari et al. 2015, Section 3.5), it follows that for each b and $a \in A$

$$\begin{pmatrix} R(b_a) & \tilde{\gamma}_a' \end{pmatrix} x_3 = \begin{pmatrix} b_a & \gamma_a' \end{pmatrix} x_3$$

Under Assumption I4(ii) S_3 contains k linearly independent vectors, so it follows that $(R(b_a), \tilde{\gamma}_a) = (b_a, \gamma_a)$ and thus γ and $P(a_3; x_3, \beta)$ are identified.

To identify $f_{\beta|x_1}$, notice that

$$\frac{f_{a_2 a_1 x_2 | x_1}(0, a_1, x_2; x_1)}{F_x(x_2; x_1, a_1)} = \left[L^*_{\beta,2}(P(a_1; x_1, \cdot) f_{\beta|x_1}(\cdot\,; x_1)) \right](x_2).$$

$L^*_{\beta,2}$ is injective and identified, since its kernel (the CCP function) is identified. Applying the left inverse of $L^*_{\beta,2}$, $P(a_1; x_1, b) f_{\beta|x_1}(b; x_1)$ and thus $f_{\beta|x_1}(b; x_1)$ is identified.

In the case that β_i conditional upon x_{i1} has $S < \infty$ points of support, the above arguments apply directly with matrices replacing integral operators where appropriate.

□

Lemma B.1 (Properties of the CCP function)**.** *Under assumptions I1,I2,I4 and I5, the sets $\mathcal{H}$ and $\tilde{\mathcal{H}}$ defined in equations (10) and (11) are norm bounded subsets of $L^2(S_\beta, \mathcal{B}, \lambda)$ where $\mathcal{B}$ is the Borel sigma field on S_β, λ is the Lebesgue measure. Let $x_t = (z_t, w_t)$ with $z_t \in \mathbb{R}^{1+|A|}$, then*

$$\left\{ h : \mathbb{R}^b \to [0,1] : h(z) = P(a; (z, w), b), z \in \mathbb{R}^{1+|A|} \right\}.$$

are real analytic functions.

Proof. Under Assumptions I1 and I2 an element $h : S_\beta \to [0,1]$ of the set $\tilde{\mathcal{H}}$ is defined as:

$$h(b) = P(a; x, b) = \frac{\exp\left(x'(b_a,\ \gamma_a) + \rho \int v(x'; b, \gamma) dF_x(x'|x, a)\right)}{\sum_{\tilde{a} \in A} \exp\left(x'(b_{\tilde{a}},\ \gamma_{\tilde{a}}) + \rho \int v(x'; b, \gamma) dF_x(x'|x, \tilde{a})\right)}. \tag{14}$$

Let $x = (z, w)$ where $z \in \mathbb{R}^{1+|A|}$ is the first $1 + |A|$ elements of x. P is well-defined for all $z \in \mathbb{R}^{1+|A|}$ since the state transition $dF_x(x'|x, a)$ is well-defined for all $z \in \mathbb{R}^{1+|A|}$ as the analytic continuation of $dF_x(x'|x, \tilde{a})$ for z in its support, which contains an open set under Assumption I4.

Since the set S_β is a compact subset of $\mathbb{R}^b$ and $|h(b)| \le 1$ for all $b \in S_\beta$,

$$\|h\|_2^2 = \int_{S_\beta} P(a_t; x_t, b)^2 d\lambda(b) \le \int_{S_\beta} d\lambda(b) < \infty,$$

and thus $h \in L^2(S_\beta, \mathcal{B}, \lambda)$.

It remains to show that the functions $P(a;x,b)$ are real analytic functions of z. Since the sum, composition and ratio of strictly positive real analytic functions are real analytic it is sufficient to show the following function is real analytic:

$$z \mapsto \int v(x';b,\gamma)dF(x'|x,a).$$

By Assumption I5, the transition kernel can be partitioned into a component represented by a density f_c, and a part represented my a mass function f_d:

$$\int v(x';b,\gamma)dF(x'|x,a) = \int v(x';b,\gamma)f_c(x'|x,a)dx' + \sum_{i=1}^{N} v(i;b,\gamma)f_d(i;x,a)$$

Since f_d is a real analytic function of z, it is enough to show $\int v(x';b,\gamma)f_c(x'|x,a)dx'$ is real analytic. By assumption I5, $f_c(x'|x,a)$ is real analytic on $z \in \mathbb{R}^{1+|A|}$. That is, for each $a \in A$, $x' \in \mathbb{R}^k$ and w in its support, there is a unique power series representation, such that for all $z \in \mathbb{R}^{1+|A|}$,

$$f_c(x'|x,a) = \sum_{n\in\mathbb{N}^{1+|A|}} \alpha_n(a,w,x')z^n, .$$

Furthermore, for any x' outside its bounded support and any (w,a), since $f_c(x'|x,a) = 0$ for z in its support, it follows that $f_c(x'|x,a) = 0$ for $z \in \mathbb{R}^{1+|A|}$ since the support of z contains an open set (a real analytic function that is zero on an open set is zero everywhere it is defined). We are now in a position to show the result.

$$\begin{aligned}\int v(x';b,\gamma)f_c(x'|x,a)dx' &= \int v(x';b,\gamma)\sum_{n\in\mathbb{N}^{b+1}} \alpha_n(a,w,x')z^n dx' \\ &= \int \sum_{n\in\mathbb{N}^{b+1}} \tilde{\alpha}_n(a,w,x')z^n dx' \\ &= \sum_{n\in\mathbb{N}^{b+1}} \left(\int \tilde{\alpha}_n(a,w,x')dx'\right) z^n = \sum_{n\in\mathbb{N}^{b+1}} \breve{\alpha}_n z^n\end{aligned}$$

The first equality holds by definition. The second holds from defining $\tilde{\alpha}_n(a,w,x') = v(x';b,\gamma)\alpha_n(a,w,x')$. The third equality holds from the bounded convergence theorem because, the integral being supported on a bounded set, $\tilde{\alpha}_n(a,w,x')$ is dominated by its supremum taken over its bounded support. The final equality is by definition of $\breve{\alpha}_n = \int \tilde{\alpha}_n(a,w,x')dx'$, which exists since the defining integral is supported on a bounded set.

□

Lemma B.2 (Approximation)**.** *Under Assumptions I1, I2, I3 and I5 the span of $\tilde{\mathcal{H}}$ (equation (11)) is uniformly dense in $\mathcal{H}_1$ (equation (12)), that is for any $(\alpha_0, \alpha_1) \in \mathbb{R}^{b+1}$:*

$$\begin{aligned} \forall \epsilon > 0 \; \exists f \in \overline{sp}\left\{h\colon S_\beta \to [0,1] : h(b) = P(a;(z,w),b), (a,z) \in A \times \mathbb{R}^{1+|A|}\right\} \\ s.t. \; \sup_{b \in S_\beta} |\widetilde{\cos}(\alpha_1' b + \alpha_0) - f(b)| < \epsilon, \end{aligned} \tag{15}$$

where $\widetilde{\cos}(x) = (1 + \cos[x + 3\pi/2])(1/2)1_{|x| \le \pi/2} + 1_{x > \pi/2}$ and $\overline{sp}\mathcal{A}$ is the uniform closure of the linear span of $\mathcal{A}$.

Proof. An element of $\mathcal{H}$ has the form

$$P(a;x,b) = \frac{\exp\left(x'(b_a,\ \gamma_a) + \rho \int v(x';b,\gamma)(dF_x(x'|x,a) - dF_x(x'|x,0))\right)}{1 + \sum_{\tilde{a} \in \tilde{A}} \exp\left(x'(b_{\tilde{a}},\ \gamma_{\tilde{a}}) + \rho \int v(x';b,\gamma)(dF_x(x'|x,\tilde{a}) - dF_x(x'|x,0))\right)}. \tag{16}$$

The proof will proceed in two steps. Again, let $x = (z,w)$ where z are the first $1 + |A|$ elements of x. First I show that the function

$$(a,z,b) \mapsto \int v(x';b,\gamma)(dF_x(x'|x,a) - dF_x(x'|x,0))$$

is uniformly bounded in $(a,z,b) \in \tilde{A} \times \mathbb{R}^{b+1} \times S_\beta$. Using this fact, I then construct a function satisfying (15).

For the first step, denote $S_{x'}$ as the support of the state transition kernel, consider that

$$\begin{aligned} \left|\int v(x';b,\gamma)(dF_x(x'|x,a) - dF_x(x'|x,0))\right| &\le \int \left|v(x';b,\gamma)\right| \left|dF_x(x'|x,a) - dF_x(x'|x,0)\right| dx' \\ &= \int_{x' \in S_{x'}} \left|v(x';b,\gamma)\right| \left|dF_x(x'|x,a) - dF_x(x'|x,0)\right| dx' \\ &\quad + \int_{x' \notin S_{x'}} \left|v(x';b,\gamma)\right| \left|dF_x(x'|x,a) - dF_x(x'|x,0)\right| dx' \\ &\le M_1(b) \int_{x' \in S_{x'}} M_2(a,w,x')dx' + 0 \le M(a,w,b) < \infty \end{aligned}$$

The second inequality follows because (a) the value function $v(x;b,\gamma)$ is bounded when the state space is contained in a compact set (Kristensen et al. 2020), (b) the transition kernels are bounded functions of z (Assumption I5), and (c) as argued in Lemma B.1, the transition kernels are identically zero for x' outside its bounded support. The final inequality follows from the existence of the integral over $S_{x'}$, a bounded set. The uniform bound is attained as $M(w) = \sup_{(a,b) \in \tilde{A} \times S_\beta} M(a,w,b)$. Since w is fixed throughout, I suppress the dependence of the uniform bound on its value.

The second step consists of showing that there exists a function in the linear span of $\tilde{\mathcal{H}}$ that is uniformly dense in the cosine squashers. I proceed in several parts. Let $\text{sgn}(\alpha_1)$ be the length $|A|$

vector of the sign of the components of α_1. First, I show that for any for any $\epsilon, \eta > 0$ and $(c_j)_{j=1}^{1+|A|}$, there exists a function in $\tilde{\mathcal{H}}$ that satisfies

$$h(b;c) \in \begin{pmatrix} (1-\eta, 1] & \text{if } \prod_j 1[\operatorname{sgn}(\alpha_{1j})(b_j - c_j) > \epsilon] > 1 \\ [0,\eta) & \text{if } \prod_j 1[\operatorname{sgn}(\alpha_{1j})(b_j - c_j) < -\epsilon] = 0 \\ [0,1] & \text{otherwise} \end{pmatrix}. \tag{17}$$

Let A^-, A^+ be the negative and positive components of α_1 respectively. Denote 2^{A^-} be the power set of A^-, and $|A^-|$ the cardinality of set A^-. Now define the function $f(x;c)$ as

$$\sum_{\mathcal{A}\in 2^{A^-}} (-1)^{|\mathcal{A}|} \frac{1}{1 + \sum_{a\in A^+\cup\mathcal{A}} \exp(-d(b_a - c_a)) + \sum_{a\in A^-\setminus\mathcal{A}} \exp(-d(b_a + d)) + \int v(x';b,\gamma)(f(x'|x_{\mathcal{A}},a) - f(x'|x_{\mathcal{A}},0))dx'}$$

where $x_{\mathcal{A}} = (z, w_{\mathcal{A}})$ for $z = -d$ and $w_{\mathcal{A}}$ a solution to the system of linear equations $dc_a = \gamma_a' w$ for $a \in \mathcal{A}\cup A^+$ and $-d^2 = \gamma_a' w$ for $a \in A^- \setminus \mathcal{A}$, which exists due to Assumption I3. For fixed ϵ, η, by taking $d \to \infty$, it can be seen that there exists a d such that this function satisfies (17).

Second, let $c_1 < c_2 < \cdots < c_n$ be equally spaced vectors on the curve $\{c \in S_\beta : \alpha_1' c + \alpha_0 = 0\}$ with c_1, c_n on the boundaries of the convex hull of S_β. Then set

$$h(b) = \sum_{i=1}^{n} h(b;c_i)/n$$

For n large enough, if $|\alpha_1' b + \alpha_0| > \epsilon$, then $|1(\alpha_1' b + \alpha_0 > 0) - g(b)| < \eta$.

Third, these approximate indicator functions can be made uniformly close to any cosine squasher on the compact support of β_i following the arguments in Hornik, Stinchcombe, and White (1989, Lemma A.2). The steps are fully elaborated for the binary choice case (Remark 4), so I do not repeat them here.

□

Remark 4 (Binary choice). In the binary choice case, it is possible to replace the period utility function of Assumption I2 with

$$u_i(x,1) = x'(\beta_{i1}, \gamma_1),$$

where now $\beta_{i1} \in \mathbb{R}^b$ is a vector. The proof to Theorem 1 provided in Section B.1 applies largely directly, except for the second step of the proof of Lemma B.2, which I now show.

Proof. The second step is the construction of a function

$$h(b) = \sum_{i=1}^{N} c_i P(1;(x_i, b))$$

for $N \in \mathbb{N}$, $c_i \in \mathbb{R}$, $x_i = (z_i, w_{1i}, w_{-1})$ with $(z_i, w_{1i}) \in \mathbb{R}^{b+1}$ and w_{-1} fixed at some value in the support, that is uniformly close to $\widetilde{\cos}(\alpha_1' b + \alpha_0)$ on $b \in S_\beta \subseteq \mathbb{R}^b$.

Let $\epsilon > 0$ and $(\alpha_0, \alpha_1) \in \mathbb{R}^{1+b}$ and $M > 0$, the uniform bound from the first step, be given. Set $\bar{\epsilon} = \epsilon \vee 1$, $N > 2/\bar{\epsilon}$ and $P = M - g^{-1}(\bar{\epsilon}/2N)$. Let $c_i = 1/N$. For $i = 1, 2, \ldots, N$, set

$$z_i = \frac{2P}{\widetilde{\cos}^{-1}(i/N) - \widetilde{\cos}^{-1}((i-1)/N)}\alpha_1, \quad w_{1i} = -\frac{P(\widetilde{\cos}^{-1}(i/N) + \widetilde{\cos}^{-1}((i-1)/N) - 2\alpha_0)}{\gamma_1(\widetilde{\cos}^{-1}(i/N) - \widetilde{\cos}^{-1}((i-1)/N))} - \frac{\gamma_{-1}{}' w_{-1}}{\gamma_1},$$

where $\widetilde{\cos}^{-1}(x) = \arccos(1-2x) - \pi/2$, the inverse of $x \mapsto \widetilde{\cos}(x)$ defined on $|x| \leq \pi/2$. With $x_i = (z_i, w_{1i}, w_{-1})$, the function h is defined.

To verify that h satisfies $\sup_{b \in S_\beta} |\widetilde{\cos}(\alpha_1' b + \alpha_0) - f(b)| < \epsilon$, first consider the ith component in the sum defining f. For $\alpha_1' b + \alpha_0 < \widetilde{\cos}^{-1}((i-1)/N)$,

$$b' z_i + \gamma_1 w_{1i} + \gamma_{-1}' w_{-1} + \int v(x'; b, \gamma)(dF_x(x'|x_i, 1) - dF_x(x'|x_i, 0)) < -P + M = g^{-1}(\bar{\epsilon}/2N).$$

The inequality follows from the choice of (z_i, w_i) and the uniform bound shown in the first step. Similarly for $\alpha_1' b + \alpha_0 > \widetilde{\cos}^{-1}(i/N)$,

$$b' z_i + \gamma_1 w_{1i} + \gamma_{-1}' w_{-1} + \int v(x'; b, \gamma)(dF_x(x'|x_i, 1) - dF_x(x'|x_i, 0)) > P - M = -g^{-1}(\bar{\epsilon}/2N) = g^{-1}(1 - \bar{\epsilon}/2N).$$

For any $j = 1, 2, \ldots, N$, $\widetilde{\cos}(\alpha_1' b + \alpha_0) \in [(j-1)/N, j/N)$, otherwise $\widetilde{\cos}(\alpha_1' b + \alpha_0) = 1$. In the former case, the $i = 1, \ldots, j-1$ components of f take values between $(1 - \bar{\epsilon}/2N)/N$ and $1/N$; the jth component takes a value between 0 and $1/N$; and the $i = j+1, \ldots, N$ components take values between 0 and $\bar{\epsilon}/2N$. This means a lower bound for f is $(j-1)(1 - \bar{\epsilon}/2N)/N$ and an upper bound is $j/N + (N-j)\bar{\epsilon}/2N$. The difference between f and $\widetilde{\cos}(\alpha_1' b + \alpha_0)$ is therefore bounded above by

$$\max\left\{|j/N + (N-j)\bar{\epsilon}/2N - (j-1)/N|, |j/N - (j-1)(1 - \bar{\epsilon}/2N)/N|\right\},$$

which is strictly less than ϵ. In the case that $\widetilde{\cos}(\alpha_1' b + \alpha_0) = 1$, all N components of the sum defining f take values between $(1 - \bar{\epsilon}/2N)/N$ and $1/N$. So the difference between the functions is at most $\bar{\epsilon}/2N < \epsilon$. □

Lemma B.3 is a straightforward generalization of Stinchcombe and White (1998, Theorem 3.8) that allows for non-linear kernel functions. The results states that an integral operator is injective if the relevant covariates have support containing an open set, if the operators are injective when the covariates have full support.

Lemma B.3. *Let F be a signed measure with compact support $\mathcal{Y}$ and $\mathcal{D}$ be a finite set. If*

$$\forall x \in \mathbb{R}^k, \int f(x, y) dF(y) = 0 \Rightarrow \forall y \in \mathcal{Y}, F(y) = 0 \tag{18}$$

and f is a real analytic function on $x \in \mathbb{R}^k$, then for any $T \subseteq \mathbb{R}^k$ open and non-empty,

$$\forall x \in T, \int f(x,y)dF(y) = 0 \Rightarrow \forall y \in \mathcal{Y}, F(y) = 0$$

Proof. Suppose that equation (18) holds and that $\forall x \in T, \int f(x,y)dF(y) = 0$, for some $T \subseteq \mathbb{R}^k$ open and non-empty. Since f is real analytic for each y and $\mathcal{Y}$ is bounded, $\int f(x,y)dF(y)$ is a real analytic function of x (Mattner 1999). Since $\int f(x,y)dF(y)$ is zero on an open set, it is zero on the Euclidean space and by equation (18), F vanishes on $\mathcal{Y}$. □

B.2 Finite horizon model

Proof of Theorem 2. Let $y = ((a_t, x_t)_{t=2}^T, a_1)$, then by Assumption F1, the distribution of y conditional upon $x_1 = x$ is

$$f_{y|x_1}(y; x_1) = \int \prod_{t=2}^{T} \left(P_t(a_t; x_t, b) F_{x_t}(x_t; x_{t-1}, a_{t-1})\right) P_1(a_1; x_1, b) f_{\beta|x_1}(db; x_1)$$

Where the transition kernel has positive measure, we can write

$$\frac{f_{y|x_1}(y; x_1)}{\prod_{t=2}^{T} F_{x_t}(x_t; x_{t-1}, a_{t-1})} = \int \prod_{t=1}^{T} P_t(a_t; x_t, b) f_{\beta|x_1}(db; x_1)$$

Define $g(b; (a_t)_{t=1}^{T-1}) = \prod_{t=1}^{T-1} P_t(a_t; x_t, b) f_{\beta|x_1}(b; x_1)$, then the right-hand side of the above equation can be written as $\int P_T(a_t; x_t, b) g(b; (a_t)_{t=1}^{T-1}) db$. With this integral equation representation, the proof follows a similar structure as the proof to Theorem 1. First, the operator

$$L_{T,\beta} : L_{S_\beta} \to A \times \mathcal{L}_{S_T} \qquad [L_{T,\beta} m](a_T, x_T) = \int P_T(a_T; x_T, b) m(b) db$$

is shown to be injective. Second, injectivity is used to identify $(\gamma_t)_{t=t_0}^T$ and $f_{\beta|x_1}$.

Part I: Injectivity of $L_{T,\beta}$

First notice that the CCP function has the form:

$$P_T(a; x, b) = \frac{\exp\left(\beta_{1a} + x'(\beta_{2a},\ \gamma_{aT})\right)}{1 + \sum_{\tilde{a} \in A} \left(\beta_{1\tilde{a}} + x'(\beta_{2\tilde{a}},\ \gamma_{\tilde{a}T})\right)}$$

Let $x = (z, w)$ where z is the first p elements of x, and denote w_A as the first $|A|$ elements of w. The CCP function is real analytic in (z, w_A) whose support contains a non-empty open set by Assumption F4. Since the support of β is compact, Lemma B.3 applies and $L_{T,\beta}$ is injective if and only if it is injective when the support of x_T is $\mathbb{R}^{p+|A|} \times S_T^r$, where S_T^r is the restriction of the support of x_T to the final $k - p - |A|$ elements of x_T. I show injectivity directly. Begin by assuming $m(b)$ is a finite signed measure satisfying

$$\forall (a, z) \in A \times \mathbb{R}^p, \int P_T(a; x, b) m(b) db = 0 \tag{19}$$

for any fixed w. Viewed as a function of a $w_A \in \mathbb{R}^{|A|}$ this object is infinitely differentiable and since it is identically zero, all of its derivatives are zero. Furthermore, since both P_T and m are bounded, we can exchange the order of differentiation and integration, so that:

$$\forall n \in \mathbb{N}_+ , \forall (a, z) \in A \times \mathbb{R}^p, \int \frac{\partial^n}{\partial {w_{At}}^n} P_T(a; x, b) m(b) db = 0.$$

Fix a and consider the first-order partial derivative ($n = 1$) with respect to the ith element of w_A:

$$\forall z \in \times \mathbb{R}^p, \; \gamma_{aT,a} \int P_T(a; x, b) dm(b) - \sum_{i \in \tilde{A}} \gamma_{aT,i} \int P_T(a; x, b)(i; x, b) dm(b) = 0.$$

From the preceding two equations, it follows that for all i,

$$\forall (a, z) \in A \times \mathbb{R}^p, \; \sum_{i \in \tilde{A}} \gamma_{aT,i} \int P_T(a; x, b) P_T(i; x, b) dm(b) = 0.$$

Repeating the argument for all $i \in \tilde{A}$ yields the system of linear equations

$$\Gamma_A \int P_T(a; x, b) \otimes \tilde{P}_T(x, b) dm(b) = 0_{|A|}$$

where $\tilde{P}_T(x; b)$ is the vector $\{P_T(a; x, b) \colon a \in A\}$ and $\otimes$ is the Kronecker product. Thus $\int P_T(a; x, b) \otimes P_T(x, b) dm(b) = 0_{|A|}$ and, repeating the argument for each a,

$$\forall z \in \mathbb{R}^p, \; \int \tilde{P}_T(x; b)^{\alpha} m(b) db = 0$$

for $\alpha = 2$ the multi-index of length A. Repeating the argument for higher order derivatives, we conclude that

$$\forall z \in \mathbb{R}^p, \; \int \tilde{P}_T(x; b)^{\alpha} m(b) db = 0 \tag{20}$$

for all multi-indices $\alpha \geq 1$. Let m_z be the signed measure induced by the transformation $\beta \to \tilde{P}_T(x; \beta)$, or more precisely:

$$m_{\tilde{z}}(B) = \int m(b) 1[\tilde{P}_T(x; b) \in B] db.$$

In other words, m_z is the density of the random variable $\tilde{P}_T(x; \beta)$. Thus from equation (20),

$$\forall z \in \mathbb{R}^p, \int x^{\alpha} m_z(x) dx = 0$$

for all multi-indices α. It follows that the Fourier transform of $\tilde{P}_T(x; \beta)$ is identically zero, and thus the measure m_z is zero for each $z \in \mathbb{R}^p$ (Hornik 1993, Theorem 1 Proof). Since the random variable $\tilde{P}_T(x; \beta)$ can be injectively mapped to $\{\beta_{1a} + x'(\beta_{2a}, \; \gamma_{aT}) : a \in \tilde{A}\}$, $m_z(B) = 0$ implies

$$\tilde{m}_z(B) = \int m(b) 1[b : \{b_{1a} + x'(b_{2a}, \; \gamma_{aT}) : a \in \tilde{A}\} \in B] = 0.$$

From here standard arguments (Masten 2018, Lemma 1) give that the characteristic function of β is zero, given fixed (γ_T, w_T), and thus the signed measure $m(b) = 0$. We conclude that $L_{T,\beta}$ is injective.

Part II: Identification of $(\gamma_t)_{t=t_0}^T$

Since $L_{T,\beta}$ is injective for any arbitrary γ and support satisfying Assumptions F3-F4, $L_{T,\beta}^{E,\gamma}$ is also. This implies that the operator defined in Assumption F5 exists. Under that assumption, γ_T is identified as follows: Given $\gamma_T \neq \tilde{\gamma}_T$, let $E, \tilde{E}$ be as in Assumption F5 and suppose that for all $x_T \in E$, there exists distributions $f_{\beta|X_{t_0}}, \tilde{f}_{\beta|X_{t_0}}$ such that

$$\int f_{A_T|X_T,\beta}(1; x_T, b; \gamma_T) f_{\beta|X_{t_0}}(b; x_{t_0}) db = \int f_{A_T|X_T,\beta}(1; x_T, b; \tilde{\gamma}_T) \tilde{f}_{\beta|X_{t_0}}(b; x_{t_0}) db$$

In different notation, this equation is: $[L_{T,\beta}^{E,\gamma_T} f_{\beta|X_{t_0}}](x_T) = [L_{T,\beta}^{E,\tilde{\gamma}_T} \tilde{f}_{\beta|X_{t_0}}](x_T)$ for all $x_T \in E$. By injectivity, it follows that $f_{\beta|X_{t_0}}(b; x_{t_0}) = [(L_{T,\beta}^{E,\gamma_T})^{-1} L_{T,\beta}^{E,\tilde{\gamma}_T} \tilde{f}_{\beta|X_{t_0}}](b)$. Suppose the same equality holds for all $x_T \in \tilde{E}$, that is $f_{\beta|X_{t_0}}(b; x_{t_0}) = [(L_{T,\beta}^{\tilde{E},\gamma_T})^{-1} L_{T,\beta}^{\tilde{E},\tilde{\gamma}_T} \tilde{f}_{\beta|X_{t_0}}](b)$. It follows that

$$0 = \left[\left((L_{T,\beta}^{E,\gamma_T})^{-1} L_{T,\beta}^{E,\tilde{\gamma}_T} - (L_{T,\beta}^{\tilde{E},\gamma_T})^{-1} L_{T,\beta}^{\tilde{E},\tilde{\gamma}_T}\right) \tilde{f}_{\beta|X_{t_0}}\right](b),$$

which contradicts the assumption that $L_{T,\beta}^{E,\gamma_T,\tilde{E},\tilde{\gamma}_T} \equiv (L_{T,\beta}^{E,\gamma_T})^{-1} L_{T,\beta}^{E,\tilde{\gamma}_T} - (L_{T,\beta}^{\tilde{E},\gamma_T})^{-1} L_{T,\beta}^{\tilde{E},\tilde{\gamma}_T}$ is injective, so γ_T is point identified.

To identify $f_{\beta|x_1}$, notice that

$$\frac{f_{y|x_1}(y; x_1)}{\prod_{t=2}^T F_{x_t}(x_t; x_{t-1}, a_{t-1})} = [L_{T,\beta} g(\cdot; (a_t)_{t=1}^{T-1}](a_T, x_T)$$

with $L_{T,\beta}$ injective and identified, since its kernel (the CCP function) is identified. Applying the left inverse of $L_{T,\beta}$, $g(b; (a_t)_{t=1}^{T-1}) = \prod_{t=1}^{T-1} (P_t(a_t; x_t, b;)) f_{\beta|x_1}(b; x_1)$ is identified. Repeating this argument for each choice sequence $(a_t)_{t=1}^T$, $f_{\beta|x_1}(b; x_1)$ is identified as $\sum_{\vec{a} \in A^{(T-2)}} g(b; \vec{a})$.

To identify γ_t for $t_0 \leq t < T$, first P_t is identified by summing $g(b, (a_t)_{t=1}^{T-1})$ over the support of $(a_t)_{t=1}^{T-1}$ for all periods except the tth period. With the CCPs known, the model can be solved for the finite parameters γ_t by backwards recursion.

□

B.3 Proof of supplementary identification results

B.3.1 Finite horizon without terminal period

Proof of Corollary 3. For ease of notation, relabel the time index so that $t_1 = 4$. Denote $\mathcal{L}_{\mathcal{A}} = \{f : \mathcal{A} \to \mathbb{R} : \sup_{a \in \mathcal{A}} |f(a)| < \infty\}$, and S_3 be the support of x_3 satisfying Assumption F4.1(ii), and S_4 the support of x_4 satisfying Assumption F4.1(ii). As in the proof to Theorem 1, under Assumptions F1, F4.1, the operators $L_{4,2,3} : \mathcal{L}_{S_{X_3}} \to A \times \mathcal{L}_{S_{X_4}}$ and $L_{4,3} : \mathcal{L}_{S_{X_3}} \to A \times \mathcal{L}_{S_{X_4}}$ defined as

$$[L_{4,2,3}m](a_4, x_4) = \int \frac{f_{A_4A_3A_2A_1X_4X_3X_2|X_1}(a_4, 0, a_2, a_1, x_4, x_3, x_2; x_1)}{F_{x_4}(x_4; x_3, 0)F_{x_3}(x_3; x_2, a_2)F_{x_2}(x_2; x_1, a_1)} m(x_3)dx_3$$
$$[L_{4,3}m](a_4, x_4) = \int \sum_{a_2 \in A} \frac{f_{A_4A_3A_2A_1X_4X_3X_2|X_1}(a_4, 0, a_2, a_1, x_4, x_3, x_2; x_1)}{F_{x_4}(x_4; x_3, 0)F_{x_3}(x_3; x_2, a_2)F_{x_2}(x_2; x_1, a_1)} m(x_3)dx_3$$

are well-defined and observed for $x_2 \in S_2$ where S_2 is the support of x_2 satisfying Assumption F4.1(i). As before, define the following operators:

$$\begin{aligned} L_{4,\beta} : \mathcal{L}_{S_\beta} \to A \times \mathcal{L}_{S_{X_4}} \qquad & [L_{4,\beta}m](a_4, x_4) = \int P_4(a_4; x_4, b)m(b)db \\ D^2_\beta : \mathcal{L}_{S_\beta} \to \mathcal{L}_{S_\beta} \qquad & [D^2_\beta m](b) = P_2(a_2; x_2, b)m(b) \\ D_\beta : \mathcal{L}_{S_\beta} \to \mathcal{L}_{S_\beta} \qquad & [D_\beta m](b) = P_1(a_1; x_1, b)f_{\beta|X_1}(b; x_1)m(b) \\ L_{\beta,3} : \mathcal{L}_{S_{X_3}} \to \mathcal{L}_{S_\beta} \qquad & [L_{\beta,3}m](b) = \int P_3(0; x_3, b)m(x_3)dx_3 \end{aligned}$$

and conclude $L_{4,2,3} = L_{4,\beta}D^2_\beta D_\beta L_{\beta,3,}$ and $L_{4,3} = L_{4,\beta}D_\beta L_{\beta,3,}$.

The proof follows structure of the proof to Theorem 1. First it is shown that the operators $L_{4,\beta}$ and $L^*_{\beta,3}$ are injective. This which implies that $L_{3,2}$ has a right inverse and, therefore, that the equivalency $L_{4,3,2}L^{-1}_{3,2} = L_{3,\beta}D^4_\beta L^{-1}_{3,\beta}$ holds. The second step of the proof is to use this eigendecomposition to identify the CCP functions P_4, and subsequently $(\gamma, f_{\beta|x_1})$.

Part I: Injectivity of $L_{4,\beta}$ and $L^*_{\beta,3}$ I focus on injectivity of $L_{4,\beta}$, since injectivity of $L^*_{\beta,3}$ follows by the same argument. Defining $\mathcal{H}, \tilde{\mathcal{H}}, \mathcal{H}_1, \mathcal{H}_2$ as in the proof to theorem 1, it follows that $\tilde{L}_{4,\beta}$ is injective if analogies to Lemmas B.1 and B.2 apply for the new kernel function P_4. It is now shown that this is the case.

The arguments of Lemma B.1 apply directly to the CCP function and we conclude that (i) $P_4(0; x_t, b)$ is real analytic function of the first $1 + |A|$ elements of x_t, and (ii) that $\tilde{\mathcal{H}}$ is a norm bounded subset of L^2. Indeed if $t = T$, then for part (i), many parts of the argument in lemma B.1 are redundant.

Let $x = (z, w)$ with z the first $1 + |A|$ elements of x. In the context of the finite horizon model, B.2 consists of two steps: (i) showing

$$(z, b) \mapsto \int v_5(x'; b, \gamma_{5+})(dF_{x_5}(x'; x, a) - dF_5(x'; x, 0))$$

is uniformly bounded in $z, b \in \mathbb{R}^{b+1} \times S_\beta$, where $\gamma_{t+} = (\gamma_s)^T_{s=t}$, and (ii) using the uniform bound to construct an approximation to the cosine squasher. If a uniform bound can be shown, then the construction in of B.2 will apply and part(ii) will hold.

To show the uniform bound, given the arguments in B.2, it is sufficient to show that $v_6(x';b,\gamma_{t+})$ is uniformly bounded on the support of $F_{x_5}(x';x,a) - F_5(x';x,0)$. Given this support and the support of b is bounded, it is enough to show that $v_6(x;b,\gamma_{5+})$ is finite for each (x,b,γ_{5+}). The argument is by induction. First define $e(a,x) = E[\epsilon_t(a)|x,a \text{ is optimal strategy}]$. Under Assumption F1, the function $e(a,x)$ is known and bounded (Aguirregabiria and Mira 2007). For $t = T-1$,

$$v_{t+1}(x;b,\gamma_{t+}) = \sum_{a\in A} f_{A_{t+1}|X_{t+1}\beta}(a;x,b)\left(b_a z + \gamma'_{a,t+1}w + e(a,x)\right),$$

which is bounded because the CCP functions are. For $t < T-1$, suppose that $v_{t+2}(x;b,\gamma_{t+1,+})$ is finite. $v_{t+1}(x;b,\gamma_{t+})$ is equal to

$$v_{t+1}(x;b,\gamma_{t+}) = \sum_{a\in A} P_{t+1}(a;x,b)\left(x'(b_a,\gamma_{a,t+1}) + e(a,x) + \rho\int v_{t+2}(x';b,\gamma_{t+1,t+})dF_{x_{t+1}}dx'\right)$$

and is finite also. Thus for all t, $v_{t+1}(x;b,\gamma_{t+})$ is finite for any (x,b) and a uniform bound is given by the supremum over the support. Therefore the construction in B.2 goes through directly to show part (ii). We conclude that $\tilde{L}_{4,\beta}$ is injective.

Part II: Identification of γ

Here the argument is the same as in Part II of the proof of Theorem 1, except that the operators are defined slightly differently.

The same arguments as in the proof to Theorem 1 imply that $L_{4,3,2} = L_{4,\beta}D_\beta^2 D_\beta L_{\beta,3}$ and $L_{4,3} = L_{4,\beta}D_\beta L_{\beta,3}$, and also that the spectral decomposition

$$L_{4,3,2}L_{4,3}^{-1} = L_{4,\beta}D_\beta^2 L_{4,\beta}^{-1}$$

identifies γ_4. Again, identification of γ_4 and injectivity of $L_{4,\beta}$ imply that $f_{\beta|X_1}(b;x_1)$ is point identified.

To identify γ_t for $t < 4$ we proceed inductively. Since $L_{4,3} = L_{4,\beta}D_\beta L_{\beta,3}$ and $L_{4,\beta}D_\beta$ is injective, the operator $L_{\beta,3}$ is identified which is equivalent to knowing its kernel function $P_3(1;x_3,b)$. Since the CCPs are known and the value function v_4 is known since γ_4 is identified, inversion of the CCP function identifies the linear index $x'_3(b,\gamma_3)$ and thus γ_3. Identification of (γ_2,γ_1) follows the same argument.

With identification of P_2, $f_{\beta|x_1}$ is identified by the same argument as in the proof to Theorem 1

□

Lemma B.4 (Result without rank condition)**.** *Suppose the Assumptions of Theorem 2 hold, excluding Assumption F5, and that the first component of γ_T is known. Further, assume the model* *is saturated in the discrete components of x, and that $|A| = 1$. Then γ and the distribution of unobserved heterogeneity are identified.*

Proof. The difference from the proof of Theorem 2 is that γ is identified, up to normalization, without using injectivity of the operator $L_{T,\beta}$. Since the proof of injectivity is the same, I show only identification of γ_T. Assume that for all $x = (z, w) \in S_{x_T}$,

$$\int \Lambda\left(\beta_1 + \beta_2' z + \gamma' w\right) df_{\beta|x_1}(b; x_1) = \int \Lambda\left(\beta_1 + \beta_2' z + \tilde{\gamma}' w\right) d\tilde{f}_{\beta|x_1}(b; x_1).$$

In particular, this must be true for all the indicator functions switched off. Allowing w^c to be the elements of w with support containing an open ball, it follows that

$$\int \Lambda\left(\beta_1 + \beta_2' z + (\gamma^c)' w^c\right) df_{\beta|x_1}(b; x_1) = \int \Lambda\left(\beta_1 + \beta_2' z + (\tilde{\gamma}^c)' w^c\right) d\tilde{f}_{\beta|x_1}(b; x_1).$$

For notational simplicity, let $w = (w^c, 0)$ — that is, setting the components of w with discrete support to zero. Viewed as a function of the continuous elements of w, this object is infinitely differentiable. Since both Λ and $f_{\beta|x_1}$ are bounded, the limits defining differentiation and integration may be exchanged, so that:

$$\forall (z, w) \in S_{x_T},\ \int \frac{\partial}{\partial w_{k'}} \Lambda\left(b_1 + b_2' z + \gamma' w\right) f_{\beta|x_1}(b; x_1) db = \int \frac{\partial}{\partial w_{k'}} \Lambda\left(b_1 + b_2' z + \tilde{\gamma}' w\right) \tilde{f}_{\beta|x_1}(b; x_1) db.$$

It is well known that the derivative of $\Lambda(x)$ is $\Lambda(x)(1 - \Lambda(x))$, so the above display is equivalent to

$$\forall (z, w) \in S_{x_T},\ \gamma_{k'} \int [\Lambda(1-\Lambda)]\left(b_1 + b_2' z + \gamma' w\right) f_{\beta|x_1}(b; x_1) db = \tilde{\gamma}_{k'} \int [\Lambda(1-\Lambda)]\left(b_1 + b_2' z + \tilde{\gamma}' w\right) \tilde{f}_{\beta|x_1}(b; x_1) db.$$

By assumption $\gamma_k = \tilde{\gamma}_k = 1$, so we have that

$$\forall (z, w) \in S_{x_T},\ \int [\Lambda(1-\Lambda)]\left(b_1 + b_2' z + \gamma' w\right) f_{\beta|x_1}(b; x_1) db = \int [\Lambda(1-\Lambda)]\left(b_1 + b_2' z + \tilde{\gamma}' w\right) \tilde{f}_{\beta|x_1}(b; x_1) db,$$

which is non-zero. Thus

$$\forall (z, w) \in S_{x_T}, (\gamma_k - \tilde{\gamma}_k) \int [\Lambda(1 - \Lambda)]\left(b_1 + b_2' z + \gamma' w\right) f_{\beta|x_1}(b; x_1) db = 0,$$

so $\gamma_k = \tilde{\gamma}_k$. This procedure can be repeated for all elements of γ whose corresponding covariates have support containing an open set.

With identification of the components of γ whose corresponding state variables have continuous support, the arguments of Theorem 2 can be used to identify $f_{\beta|x_1}$.

Now consider the discrete components of γ. For discrete component $w_{k'}$, assume $\gamma_{k'} < \tilde{\gamma}_{k'}$. Since the logistic function is strictly increasing, for $w_{k'} = 1$,

$$\Lambda\left(\beta_1 + \beta_2' z + (\gamma^c)' w^c + \gamma_{k'} w_{k'}\right) < \Lambda\left(\beta_1 + \beta_2' z + (\gamma^c)' w^c + \tilde{\gamma}_{k'} w_{k'}\right).$$

Since $f_{\beta|x_1}$ is positive,

$$\int \Lambda\left(\beta_1 + \beta_2' z + (\gamma^c)' w^c + \gamma_{k'} w_{k'}\right) df_{\beta|x_1}(\beta; x_1) < \int \Lambda\left(\beta_1 + \beta_2' z + (\gamma^c)' w^c + \tilde{\gamma}_{k'} w_{k'}\right) df_{\beta|x_1}(\beta; x_1).$$

Since the model is saturated, there is some (z, w^c) for which $x = (z, w^c, w^d_{k'} = 1, (w^d_{-k'}) = 0)$ is in the support of x_2. Thus $\gamma_{k'}$ is identified. □

B.3.2 Infinite horizon model with random intercepts

Proof of Corollary 2. The proof follows closely the structure of the proof to Theorem 1. As in that proof, Assumptions I1 and I4.1 enable the decompositions $L_{3,4,2} = L_{3,\beta} D^4_\beta D_\beta L_{\beta,2}$ and $L_{3,2} = L_{3,\beta} D_\beta L_{\beta,2}$ where the operators are defined in proof to Theorem 1. As before, Part I is to show injectivity of $L_{3,\beta}$ and $L^*_{\beta,2}$.

Let x_A be the first $|A|$ elements of x_2. By Assumption I4.1, the support of x_A contains a non-empty open set for which

$$P(a; x, b) = \frac{\exp\left(\beta_a + x'\gamma_a\right)}{1 + \sum_{\tilde{a} \in \tilde{A}} \exp\left(\beta_{\tilde{a}} + x'\gamma_{\tilde{a}}\right)}.$$

Given this functional form, the arguments from Part I of the proof to Theorem 2 give that

$$\int \tilde{P}(x; b)^\alpha dm(b) = 0$$

for all multi-indices $\alpha \geq 1$ where $\tilde{P}(x; b) = \{P(a; x, b) : a \in \tilde{A}\}$. It follows that the measure induced by the mapping $\beta \to \tilde{P}(x; \beta)$ is identically zero. Because this mapping is injective, the measure $m(b)$ is identically zero and thus $L_{3,\beta}$ and $L^*_{2,\beta}$ are injective.

With injectivity in hand, identification follows from Part II of the proof to Theorem 1, which applies under Assumptions I4.1.

□

B.3.3 Finite dependence

Proof of Corollary 4. For ease of notation, assume $t_1 - t_0 = 3$ and relabel T such that let $t_0 = 1$ and $t_1 = 4$. If the panel is longer than 4 (i.e. if $t_1 - t_0 > 4$), then γ_t for $t \leq t_1 - 2$ can be the same arguments as below. For notational ease, set $t_1 = 4$. Let $\mathcal{L}_{\mathcal{A}} = \{f : \mathcal{A} \to \mathbb{R} : \sup_{a \in A} |f(a)| < \infty\}$ and define the following operators:

$$\begin{aligned}
L_{3,4,2} : \mathcal{L}_{S_2} \to \mathcal{L}_{S_3} \qquad & [L_{4,3,2}m](x_3) = \int \frac{f_{A_4A_3A_2A_1X_4X_3X_2|X_1}(1,1,1,a_1,x_4,x_3,x_2;x_1)}{F_{x_4}(x_4;x_3,1)F_{x_3}(x_3;x_2,1)F_{x_1}(x_2;x_1,a_1)} m(x_2)dx_2 \\
L_{3,2} : \mathcal{L}_{S_2} \to \mathcal{L}_{S_3} \qquad & [L_{4,3}m](x_3) = \int \sum_{a_2 \in A} \frac{f_{A_4A_3A_2A_1X_4X_3X_2|X_1}(1,1,a_2,a_1,x_4,x_3,x_2;x_1)}{F_{x_4}(x_4;x_3,1)F_{x_3}(x_3;x_2,1)F_{x_1}(x_2;x_1,a_1)} m(x_2)dx_2 \\
L_{3,\beta} : \mathcal{L}_{S_\beta} \to \mathcal{L}_{S_3} \qquad & [L_{3,\beta}m](x_3) = \int P_3(1;x_3,b)m(b)db \\
D_\beta^4 : \mathcal{L}_{S_\beta} \to \mathcal{L}_{S_\beta} \qquad & [D_\beta^4 m](b) = P_4(1;x_4,b)m(b) \\
D_\beta : \mathcal{L}_{S_\beta} \to \mathcal{L}_{S_\beta} \qquad & [D_\beta m](b) = P_1(a_1;x_1,b)f_{\beta|X_1}(b;x_1)m(b) \\
L_{\beta,2} : \mathcal{L}_{S_2} \to \mathcal{L}_{S_\beta} \qquad & [L_{\beta,2}m](b) = \int P_2(1;x_2,b)m(x_2)dx_2
\end{aligned}$$

Under Assumptions F1 and I4 these operators are well-defined and observed. The same arguments as in the proof to Theorem 1 imply that $L_{4,3,2} = L_{3,\beta}D_\beta^4 D_\beta L_{\beta,2}$ and $L_{3,2} = L_{3,\beta}D_\beta L_{\beta,2}$.

For injectivity, as assumptions F1, I4, and F2.1-F3.1 apply, and thus the spectral decomposition

$$L_{4,3,2}L_{3,2}^{-1} = L_{3,\beta}D_\beta^4 L_{3,\beta}^{-1}$$

is unique. I now show the eigenvalue-eigenfunction representation is unique. Since the model is binary choice with real valued β, the function $P_4(1;x_4,b)$ is injective in b. It follows that the eigenvalues are unique, and, up to the ordering function R, $P_4(1;x_4,R(b))$ is identified. The eigenfunctions of the decomposition identify $P_3(1;x_3,R(b))$, which equal

$$g\left(x_3'(R(b),\gamma_3) + \int v(x';R(b),\gamma)\left(F_{x_4}(dx'|x_3,1) - F_{x_4}(dx'|x_3,0)\right)\right)$$

where g is the logistic function function. Under Assumption F7.1, the continuation value can be expressed in terms of $P_4(1;x_4,R(b))$, and is therefore identified. Therefore identification consists of showing that $(R(b),\gamma_3)$ can be identified from $x_3'(R(b),\gamma_3)$, which follows from the support assumption.

With γ_{t_1-1} identified, identification of γ_s for $t_0 \le s < t_1 - 1$ proceeds inductively as in the proof to Corollary 3. □

C Estimation appendix

C.1 General two-step seminonparametric estimation

This section details the assumptions of Theorem 3 that provide for consistent estimation of $\theta_0 = (F_x, \gamma, f_{\beta|x_1}) \in \Theta = \mathcal{F} \times \Gamma \times \mathcal{M}$. Here $\mathcal{F}$ is the space of state transitions, $\Gamma \subseteq \mathbb{R}^p$, and $\mathcal{M}$ is the space of distribution functions on $S_\beta \times S_1$. The first assumption postulates the existence of a consistent estimator for the state transition F_x:

Assumption E1. *There exists an estimator $\hat{F}_{X,n}$ that satisfies $\left\|\hat{F}_{X,n} - F_x\right\|_{\mathcal{F}} = o_p(1)$, where $\|\cdot\|_{\mathcal{F}}$ is a norm on $\mathcal{F}$.*

One such estimator that satisfies Assumption E1 is the kernel estimator of the conditional density:

$$\hat{F}_{X_t,n}(x';x,a) = \frac{\sum_{i=1}^N K_{X',h_{X'}}(x'-x_{t,i})K_{X,h_X}(x-x_{t,i})1(a_{it}=a)}{\sum_{i=1}^N K_{X,h_X}(x-x_{t,i})1(a_{it}=a)} \tag{21}$$

where K_{Z,h_Z} are multivariate kernel functions with bandwidth h_Z.

Let $\mathcal{M}_n$ be a sieve space that approximates $\mathcal{M}$, and denote $d_{\mathcal{M}}(\cdot,\cdot)$ as the Prokhorov metric. The Prokhorov distance between two measures $f, \tilde{f}$ on S_β is

$$\inf\left\{\delta > 0 \colon f(B) \leq \tilde{f}(B_\delta) + \delta \vee \tilde{f}(B) \leq f(B_\delta) + \delta,\ \forall A \in \mathcal{B}(S_\beta)\right\},$$

where B_δ be the δ neighborhood of $B \subseteq S_\beta$ and $\mathcal{B}(S_\beta)$ is the Borel sigma field. The next assumption requires that the true parameter values be a well-separated maximum.

Assumption E2. *For all $\epsilon > 0$ there exists some decreasing sequence of positive numbers $c_n(\epsilon)$ satisfying $\liminf c_n(\epsilon) > 0$ such that*

$$E[\psi(y_i, F_X, \gamma, f_{\beta|x_1})] - \sup_{\{(\tilde{\gamma},\tilde{f})\in\Gamma\times\mathcal{M}_n : \|\tilde{\gamma}-\gamma\| + d_{\mathcal{M}}(\tilde{f}, f_{\beta|x_1}) \geq \epsilon\}} E[\psi(y_i, F_X, \tilde{\gamma}, \tilde{f})] \geq c_n(\epsilon).$$

Assumption E2 is the condition of Remark 3.1(2) in Chen (2007) that strengthens their Condition 3.1. If the strict inequality restriction on c_n were replaced by a weak inequality, then the assumption would be implied by the identification result.

Assumption E3. *The sieve space (i) satisfies $\mathcal{M}_n \subseteq \mathcal{M}_{n+1} \subseteq \mathcal{M}$ and (ii) is such that there exists a sequence $f_n \in \mathcal{M}_n$ that converges to $f_{\beta|x_1}$ and satisfies*

$$\left|E[\psi(y_i, F_X, \gamma, f_n)] - E[\psi(y_i, F_X, \gamma, f_{\beta|x_1})]\right| = o(1).$$

These are standard restrictions on the sieve space and the population criterion function (Chen 2007, Condition 3.2, 3.3(ii)). The second condition is a local continuity assumption. As per Chen (2007, Remark 2.1), it is implied by compactness of the sieve space and continuity of the population criterion function on $\mathcal{M}_n$.

Define $\mathcal{F}_n$ to be the set of possible values that the estimator $\hat{f}_n$ can take. For example, if the conditional density kernel estimator is chosen, then an element of the set $\mathcal{F}_n$ takes the form in equation 21 and the set $\mathcal{F}_n$ is defined by ranging $(x_{it+1}, x_{it}, a_{it})$ over its support. Define the neighborhood $\mathcal{N}_{F_x,n} = \{\tilde{F}_X \in \mathcal{F}_n : \|\tilde{F}_X - F_X\|_{\mathcal{F}} \leq \epsilon_{1,n}\}$ where $\|\cdot\|_{\mathcal{F}}$ is the norm in Assumption E1.

Assumption E4. *Assume the following two conditions hold*

$$\sup_{(\tilde{F}_X,\tilde{\gamma},\tilde{f}) \in \mathcal{N}_{F_x,n} \times \Gamma \times \mathcal{M}_n} \left| \frac{1}{n} \sum_{i=1}^{n} \psi(y_i, \tilde{F}_x, \tilde{\gamma}, \tilde{f}) - E[\psi(y_i, \tilde{F}_x, \tilde{\gamma}, \tilde{f})] \right| = o_p(1),$$

$$\sup_{(\tilde{F}_X,\tilde{\gamma},\tilde{f}) \in \mathcal{N}_{F_x,n} \times \Gamma \times \mathcal{M}_n} \left| E[\psi(y_i, \tilde{F}_x, \tilde{\gamma}, \tilde{f})] - E[\psi(y_i, F_x, \tilde{\gamma}, \tilde{f})] \right| = o(1).$$

This is similar to Hahn, Liao, and Ridder (2018, Assumption 5.3), which is based on Chen (2007, Condition 3.5) but includes an additional condition to account for the presence of a first-step estimator.

Theorem 3 is a direct consequence of Hahn, Liao, and Ridder (2018, Theorem 5.1), so the proof is omitted. In the proof, by consistency it is meant that $\|\hat{\gamma} - \gamma\| + d_{\mathcal{M}}(\hat{f}_{\beta|x_1}, f_{\beta|x_1}) = o_p(1)$.

C.2 Fixed grid estimation

The choice of tuning parameters must satisfy the following condition:

Assumption E3.1. *The sieve space defined in (7) is such that (i) $\mathcal{M}_n \subseteq \mathcal{M}_{n+1}$ and as $n \to \infty$, (ii) $\mathcal{B}_n \times \mathcal{X}_n$ becomes dense in $S_\beta \times S_1$ and (iii) $I(n) \log I(n) = o(n)$ where $I(n) = B(n)X(n)$.*

We also place some restrictions on the complexity of $\mathcal{N}_{F_X,n}$, the neighborhood to which the estimator $\hat{F}_{X,n}$ belongs with probability approaching one. For this purpose define $N(w, \mathcal{G}, \|\cdot\|_{\mathcal{G}})$ as the covering number of set $\mathcal{G}$ with balls of radius w under the norm $\|\cdot\|_{\mathcal{G}}$.

Assumption E4.1. *(i) $(\mathcal{N}_{F_X,n}, \|\cdot\|_{\mathcal{F}})$ and Γ are compact. (ii) P_t is Lipschitz continuous in $\gamma \in \Gamma$ and continuous in $F_X \in \mathcal{N}_{F_X,n}$. (iii) $\log N(w/\sqrt{I(n)}, \mathcal{N}_{f,n}, \|\cdot\|_{\mathcal{F}}\}) = o(n)$ with $I(n)$ as in Assumption E3.1.*

Proof of Theorem 4. The proof consists of verifying the assumptions of Theorem 4 imply those of Theorem 3. Assumptions E1 is assumed.

To verify assumption E2, suppose that (i) $\mathcal{M}_n$ and $\mathcal{M}$ are compact in the weak topology and (ii) that $E[(y_i, F_x, \gamma, f_{\beta|x_1})]$ is continuous on $f_{\beta|x_1} \in \mathcal{M} \supset \mathcal{M}_n$ in the weak topology and $\gamma \in \Gamma$. Then consider that since θ_0 is identified, this value uniquely maximizes the expected log likelihood, so that for any $(\tilde{\gamma}, \tilde{f}_{\beta|x_1}) \neq (\gamma, f_{\beta|x_1})$,

$$E[\psi(y_i, F_x, \gamma, f_{\beta|x_1})] - E[\psi(y_i, F_x, \tilde{\gamma}, \tilde{f}_{\beta|x_1})] > 0$$

Because $\{(\tilde{\gamma}, \tilde{f}) \in \Gamma \times \mathcal{M}_n : \|\tilde{\gamma} - \gamma\| + d_{\mathcal{M}}(\tilde{f}, f_{\beta|x_1}) \geq \epsilon\}$ is closed in the compact set $\mathcal{M}_n \times \Gamma$, it is compact and the following infinum

$$E[\psi(y_i, F_x, \gamma, f_{\beta|x_1})] - \sup_{\{(\tilde{\gamma}, \tilde{f}) \in \Gamma \times \mathcal{M}_n : \|\tilde{\gamma} - \gamma\| + d_{\mathcal{M}}(\tilde{f}, f_{\beta|x_1}) \geq \epsilon\}} E[\psi(y_i, F_x, \tilde{\gamma}, \tilde{f}_{\beta|x_1})]$$

is attained for each (ϵ, n). Setting this difference to $c_n(\epsilon)$ guarantees it is positive. It remains to show that $\liminf c_n(\epsilon) > 0$. Consider that

$$\begin{aligned} c_n(\epsilon) =& E[\psi(y_i, F_x, \gamma, f_{\beta|x_1})] - \sup_{\{(\tilde{\gamma}, \tilde{f}) \in \Gamma \times \mathcal{M}_n : \|\tilde{\gamma} - \gamma\| + d_{\mathcal{M}}(\tilde{f}, f_{\beta|x_1}) \geq \epsilon\}} E[\psi(y_i, F_x, \tilde{\gamma}, \tilde{f}_{\beta|x_1})] \\ \geq& E[\psi(y_i, F_x, \gamma, f_{\beta|x_1})] - \sup_{\{(\tilde{\gamma}, \tilde{f}) \in \Gamma \times \mathcal{M} : \|\tilde{\gamma} - \gamma\| + d_{\mathcal{M}}(\tilde{f}, f_{\beta|x_1}) \geq \epsilon\}} E[\psi(y_i, F_x, \tilde{\gamma}, \tilde{f}_{\beta|x_1})] > 0 \end{aligned}$$

The weak inequality is because $\mathcal{M}_n \subseteq \mathcal{M}$. The strict inequality is because the set $\{(\tilde{\gamma}, \tilde{f}) \in \Gamma \times \mathcal{M}_n : \|\tilde{\gamma} - \gamma\| + d_{\mathcal{M}}(\tilde{f},$ is compact and $E[(y_i, F_x, \gamma, f_{\beta|x_1})]$ is continuous. Since $c_n(\epsilon)$ is bounded above zero by a universal constant for all n, its limit inferior is strictly positive.

To complete the argument, it must be shown that (i) $\mathcal{M}_n$ and $\mathcal{M}$ are compact in the weak topology and (ii) that $E[\psi(y_i, F_x, \gamma, f_{\beta|x_1})]$ is continuous on $\mathcal{M} \supset \mathcal{M}_n$ in the weak topology and $\gamma \in \Gamma$ Compactness of $\mathcal{M}$ and $\mathcal{M}_n$, both in the weak topology, is shown in Fox, Kim, and Yang (2016, pp. 240, 247). Since the CCP functions P_t are continuous in (b, γ) (Norets 2010), the argument of Fox, Kim, and Yang (2016, Remark 2) implies the function $f_{\beta|x_1} \mapsto \int \prod_{t=1}^{t_1} P_t(a_{it}, x_{it}, b; F_X, \gamma) df_{\beta|x_1}(b, x_{i1})$ is continuous. Since it is bounded away from zero, $f_{\beta|x_1} \mapsto \log \int \prod_{t=1}^{t_1} P_t(a_{it}, x_{it}, b; F_X, \gamma) df_{\beta|x_1}(b, x_{i1})$ is also continuous. And since this function is bounded away from negative infinity, $f_{\beta|x_1} \mapsto E[\log \int \prod_{t=1}^{t_1} P_t(a_{it}, x_{it}, b; F_X, \gamma) df_{\beta|x_1}(b, x_{i1})]$ is continuous by the bounded convergence theorem as required.

Assumption E3(i) is guaranteed by Assumption E3.1(i). For Assumption E3(ii), Fox, Kim, and Yang (2016, p. 247) show the existence of such a sequence $f_n \subseteq \mathcal{M}$ that converges to $f_{\beta|x_1} \in \mathcal{M}$. Since the sequence $(f_n)_{n \in \mathbb{N}}$ takes values in $\mathcal{M}$ and $E[\psi(y_i, F_x, \gamma, f_{\beta|x_1})]$ is continuous on $\mathcal{M}$, we have that

$$\left| E[\psi(Y_i, F_x, \gamma, f_n)] - E[\psi(y_i, F_x, \gamma, f_{\beta|x_1})] \right| = o(1).$$

For Assumption E4(i), note that

$$\begin{aligned}\left|E[\psi(y_i, F_x, \gamma, f_{\beta|x_1})]\right| &\leq E[\left|\psi(y_i, F_x, \gamma, f_{\beta|x_1})\right|] \\ &\leq E\left[\left|\sum_{t=2}^{T} \log F_{x_t}(x_{it}, x_{i,t-1}, a_{i,t-1})\right|\right] + E\left[\left|\log \int \prod_{t=1}^{T} P_t(a_{it}, x_{it}, b; F_x, \gamma) df_{\beta|x_1}(b; x_{i1})\right|\right] < \infty.\end{aligned}$$

The left term in the sum is finite by construction. Since $\mathcal{N}_{f,n} \times \Gamma \times S_\beta$ is compact and P_t is strictly positive for each (b, F_x, γ), P_t is uniformly bounded away from zero, so the right term is finite also. Then based on the discussion around Chen (2007, Remark 3.3), the condition $\log N(w, \{\psi(\cdot, F_x, \gamma, f_{\beta|x_1}) \colon (F_x, \gamma, f_{\beta|x_1}) \in \mathcal{N}_{f,n} \times \Gamma \times \mathcal{M}_n\}, \|\cdot\|_1) = o_p(n)$ is equivalent to Assumption E4(i). This entropy is bounded above by the sum of the entropies associated with $\mathcal{N}_{F_x,n}$, Γ and $\mathcal{M}_n$, so it sufficient that each are $o_p(n)$. Fox, Kim, and Yang (2016, p. 248) show the entropies associated with Γ and $\mathcal{M}_n$ are $o_p(n)$ under Assumption E3.1(iii). By Assumption E4.1(iii), the entropy associated with $\mathcal{N}_{F_x,n}$ is $o_p(n)$.

Assumption E4(ii) follows easily from the continuity of the population criterion function on the compact set $\mathcal{N}_{F_x,n} \times \Gamma \times \mathcal{M}_n$, so that

$$\sup_{\mu \in \mathcal{M}_n, f \in \mathcal{N}_{f,n}} \left|E[\psi(Y_i, \mu, f)] - E[\psi(Y_i, \mu, f_0)]\right| = o(1)$$

□

C.3 Estimating the support of unobserved heterogeneity

Proof of Corollary 1. From the definitions in the proof to Theorem 1 and Corollary 3, it is immediate that $L = L_{3,\beta} D_\beta L_{\beta,2}$. From those proofs, $L_{3,\beta}$, D_β and $L_{\beta,2}$ are matrices with rank R. □

References for Appendix

Aguirregabiria, V. and Mira, P. (2007). "Sequential estimation of dynamic discrete games". *Econometrica* 75.1, pp. 1–53.

Arcidiacono, P. and Ellickson, P. B. (2011). "Practical methods for estimation of dynamic discrete choice models". *Annu. Rev. Econ.* 3.1, pp. 363–394.

Arcidiacono, P. and Miller, R. A. (2020). "Identifying dynamic discrete choice models off short panels". *Journal of Econometrics* 215.2, pp. 473–485.

Bajari, P., Chernozhukov, V., Hong, H., and Nekipelov, D. (2015). *Identification and efficient semiparametric estimation of a dynamic discrete game.* Tech. rep. National Bureau of Economic Research.

Carrasco, M., Florens, J.-P., and Renault, E. (2007). “Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization”. *Handbook of econometrics* 6, pp. 5633–5751.

Chen, X. (2007). “Large sample sieve estimation of semi-nonparametric models”. *Handbook of econometrics* 6, pp. 5549–5632.

Fox, J. T., Kim, K. I., and Yang, C. (2016). “A simple nonparametric approach to estimating the distribution of random coefficients in structural models”. *Journal of Econometrics* 195.2, pp. 236–254.

Hahn, J., Liao, Z., and Ridder, G. (2018). “Nonparametric two-step sieve M estimation and inference”. *Econometric Theory* 34.6, pp. 1281–1324.

Hornik, K. (1993). “Some new results on neural network approximation”. *Neural networks* 6.8, pp. 1069–1072.

Hornik, K., Stinchcombe, M., and White, H. (1989). “Multilayer feedforward networks are universal approximators.” *Neural networks* 2.5, pp. 359–366.

Hu, Y. and Schennach, S. M. (2008). “Instrumental variable treatment of nonclassical measurement error models”. *Econometrica* 76.1, pp. 195–216.

Kristensen, D., Mogensen, P. K., Moon, J. M., and Schjerning, B. (2020). “Solving dynamic discrete choice models using smoothing and sieve methods”. *Journal of Econometrics.*

Masten, M. A. (2018). “Random coefficients on endogenous variables in simultaneous equations models”. *The Review of Economic Studies* 85.2, pp. 1193–1250.

Mattner, L. (1999). *Complex differentiation under the integral.* Universität Hamburg. Institut für Mathematische Stochastik.

Norets, A. (2010). “Continuity and differentiability of expected value functions in dynamic discrete choice models”. *Quantitative economics* 1.2, pp. 305–322.

Rudin, W. (1964). *Principles of mathematical analysis.* Vol. 3. McGraw-hill New York.

Stinchcombe, M. and White, H. (1998). “Consistent specification testing with nuisance parameters present only under the alternative”. *Econometric Theory* 14.3, pp. 295–325.

Williams, B. (2019). “Nonparametric identification of discrete choice models with lagged dependent variables”. *Journal of Econometrics.*