

Essays on the Econometrics of Dynamic Discrete Choice

by

Jackson Bunting

Department of Economics
Duke University

Date: _____

Approved:

Federico A Bugni, Advisor

Adam Rosen

Arnaud P V Maurel

Matthew Masten

Dissertation submitted in partial fulfillment of the requirements for the
degree of Doctor of Philosophy in the Department of Economics
in the Graduate School of Duke University

2021

ABSTRACT

Essays on the Econometrics of Dynamic Discrete Choice

by

Jackson Bunting

Department of Economics
Duke University

Date: _____

Approved:

Federico A Bugni, Advisor

Adam Rosen

Arnaud P V Maurel

Matthew Masten

An abstract of a dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in the Department of Economics
in the Graduate School of Duke University

2021

Copyright © 2021 by Jackson Bunting
All rights reserved

Abstract

Disentangling the causal effect of policy from that of behavior—i.e. controlling for selection—is a foundational empirical challenge in economics. Dynamic discrete choice models are a structural approach that posits that selection is driven by forward-looking, optimizing decision makers. The resultant econometric problem is to recover the structural parameters that characterize the model.

This dissertation contributes to the econometrics of dynamic discrete choice models in several directions. Chapter two shows identification of dynamic discrete choice models with continuous permanent unobserved heterogeneity. That is, the model allows for infinitely many persistent unobserved differences between decision making agents. The previous literature allowed for only finitely many types of persistent unobserved differences.

The third chapter provides a hypothesis test for an important modeling assumption, that of ‘homogeneity’. Commonly, it is assumed that behavior is sufficiently similar across time or markets that data can be pooled across these dimensions. However, this assumption may fail in the presence of a structural break or multiple equilibria. The chapter proposes a hypothesis test to evaluate whether the homogeneity assumption holds in the data. As an approximate randomization test, the hypothesis test is valid even without a large sample.

The fourth chapter provides a computationally advantageous estimator for dynamic discrete choice models. The estimator is based on the observation that dynamic discrete choice models possess a multiple index structure. The chapter shows that index sufficiency can be used to construct a set of equality constraints, which restrict the parameter of interest to belong to a subspace of the parameter space. The chapter proposes an estimator that imposes these restrictions, thus attaining

computational gains by reducing the effective dimension of the parameter space.

Contents

Abstract	iv
List of Figures	ix
List of Tables	x
1 Introduction	1
2 Continuous permanent unobserved heterogeneity in dynamic discrete choice models	3
2.1 Introduction	3
2.1.1 Related Literature	8
2.2 Model and identification	10
2.2.1 Infinite horizon model	10
2.2.2 Finite horizon model	17
2.3 Estimation	22
2.3.1 A general two-step seminonparametric estimator	23
2.3.2 Fixed grid estimation	24
2.3.3 Estimating the support of unobserved heterogeneity	26
2.4 Simulations	27
2.5 Application to a labor force participation model	30
2.6 Conclusion	34
3 Testing homogeneity in dynamic discrete games in finite samples	35
3.1 Introduction	35
3.2 The econometric model and the testing problem	38
3.2.1 The econometric model	39
3.2.2 The hypothesis testing problem	41

3.3	Our hypothesis test	42
3.3.1	The MCMC algorithm	44
3.4	Our test as an approximate randomization test	46
3.4.1	An alternative representation of the likelihood	47
3.4.2	A transformation group related to the proposed MCMC algorithm	48
3.4.3	A randomization test	50
3.4.4	An MCMC approximation to the randomization test	51
3.5	Monte Carlo simulations	52
3.6	Empirical application	58
3.7	Conclusions	61
4	Dimension reduction in dynamic discrete choice models via index sufficiency	62
4.1	Introduction	62
4.2	Model and estimator	64
4.3	Nullity of Σ_0	70
4.4	Estimation of Σ_0	71
4.5	Conclusion	73
5	Conclusion	74
A	Appendix for Chapter 1	75
A.1	Supplementary identification results	75
A.1.1	Random intercepts	75
A.1.2	Identification without the terminal period	76
A.1.3	Finite dependence	78
A.2	Identification proofs	79
A.2.1	Infinite horizon model	79

A.2.2	Finite horizon model	91
A.2.3	Proof of supplementary identification results	95
A.3	Estimation appendix	102
A.3.1	General two-step seminonparametric estimation	102
A.3.2	Fixed grid estimation	104
A.3.3	Estimating the support of unobserved heterogeneity	107
B	Appendix for Chapter 2	108
B.1	Appendix to Section 3.3	108
B.1.1	Implementation of Step 2 in Algorithm 3.3.1	108
B.1.2	Implementation of Step 3 in Algorithm 3.3.1	116
B.1.3	Proof of Theorem 5	117
B.2	Appendix to Section 3.4	118
B.2.1	Proof of lemmas	118
B.2.2	Auxiliary results	123
C	Appendix for Chapter 3	136
C.1	Proofs	136
C.1.1	Proof of Theorem 7	136
C.1.2	Proof of Theorem 8	141
C.1.3	Proof of Theorem 9	142
C.1.4	Proof of Theorem 10	142
	References	148

List of Figures

2.1	Simulation results for estimation of f_β for each sample size.	30
2.2	Estimated distribution of ν_i	33
3.1	Histogram of market capacity per year, measured in thousand of tons.	59

List of Tables

2.1	Some applications of DDC models with discrete permanent unobserved heterogeneity	7
2.2	Simulation results for estimation of γ and f_β	29
2.3	Estimates of γ	33
3.1	Simulation results under H_0	55
3.2	Simulation results under H_1	56
3.3	Summary statistics for market capacity per year, measured in thousand of tons.	58
3.4	Results of testing (3.1) separately before and after the passing of the 1990 Amendments.	60

Chapter 1

Introduction

Disentangling the causal effect of policy from that of behavior—i.e. controlling for selection—is a foundational empirical challenge in economics. Dynamic discrete choice models are a structural approach that posits that selection is driven by forward-looking, optimizing decision makers. The resultant econometric problem is to recover the structural parameters that characterize the model. This dissertation makes contributions to the econometrics of dynamic discrete choice models in several directions.

In Chapter 2, 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.

The literature on dynamic discrete games often assumes that the conditional choice probabilities and the state transition probabilities are homogeneous across markets and over time. I refer to this as the “homogeneity assumption” in dynamic discrete games. This homogeneity assumption enables empirical studies to estimate the game’s structural parameters by pooling data from multiple markets and from many time periods. Chapter 3 proposes a hypothesis test to evaluate whether the homogeneity assumption holds in the data. The hypothesis test is the result of an approximate randomization test, implemented via a Markov chain Monte Carlo (MCMC) algorithm. The chapter shows that the hypothesis test becomes valid as the (user-defined) number of MCMC draws diverges, for any fixed number of markets, time periods, and players. The chapter includes an application of the test to the empirical study of the U.S. Portland cement industry in Ryan 2012.

Chapter 4 provides a computationally advantageous estimator for dynamic discrete choice models. The estimator is based on the observation that dynamic discrete choice models possess a multiple index structure. The chapter shows that index sufficiency can be used to construct a set of equality constraints, which restrict the parameter of interest to belong to a strict subspace of the parameter space. The chapter proposes an estimator that imposes these restrictions, thus attaining computational gains by reducing the effective dimension of the parameter space.

Chapter 3 reflects collaborative work with Federico Bugni and Takuya Ura, and Chapter 4 is collaborative work with Takuya Ura. In both cases, I contributed to all parts of the development of the chapters, including theoretical work, drafting and editing.

Chapter 2

Continuous permanent unobserved heterogeneity in dynamic discrete choice models

2.1 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

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

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 (2.2)) — 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 case¹ (Aguirre-

¹In 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.

gabiria and Mira 2002b; Arcidiacono and Miller 2011b) 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 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 het-

erogeneity 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 2.1, 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.

Table 2.1: Some applications of DDC models 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.

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

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 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 4.2). Section 2.2.1 treats the infinite horizon model and Section 2.2.2 the finite horizon model. Variants on these baseline models are found in the appendix (Section A.1). Section 2.3 proposes the two-step sieve M-estimator, and shows its consistency. Section 2.5 considers an application, section 2.4 presents simulation results and finally section 2.6 concludes.

2.1.1 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 probabilities² and the discrete distribution of unobserved heterogeneity³ via the measurement

²The 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 .

³The 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.

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.

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).

2.2 Model and identification

2.2.1 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:

$$V(x_t, \epsilon_t) = \max_{(a_t, a_{t+1}, \dots)} E \left[\sum_{s=1}^{\infty} \rho^{s-t} u(x_s, \epsilon_s, a_s) \mid x_t, \epsilon_t, a_t \right] \quad (2.1)$$

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 1994, Theorem 2.3), they admit a recursive formulation:

$$V(x_t, \epsilon_t) = \max_{a_t \in A} \{u(x_t, \epsilon_t, a_t) + \rho E[V(x_{t+1}, \epsilon_{t+1}) \mid x_t, \epsilon_t, a_t]\}. \quad (2.2)$$

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

$$dF_S(s_{t+1}; s_t, a_t) = dF_\epsilon(\epsilon_{t+1})dF_x(x_{t+1}; x_t, a_t). \quad (2.3)$$

(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 $\tilde{A} = \{1, 2, \dots, |A|\}$, 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:*

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

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 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 not 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 (Kawahara 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 reached⁴. 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 $\mathbb{R}^{1+|A|}$*

⁴More 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).

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: $F_x(x'; x, a) = \pi \mathbf{1}(x' = 0) + (1 - \pi)F_+(x'; x, a)$, 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 (2.2) is a contraction between spaces of bounded functions (Kristensen et al. 2020). This could be relaxed, but proving equation (2.2) 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 $\gamma = (\gamma_a : a \in \tilde{A}) \in \mathbb{R}^{|\tilde{A}|(k-1)}$, 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 (2.1) 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.1.1 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 A.2.1, 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:

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

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

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

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 $(\gamma, f_{\beta|x_1})$.

2.2.2 Finite horizon model

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

$$V_{iT_0}(x_{T_0}, \epsilon_{T_0}) = \max_{(a_{T_0}, a_{T_0+1}, \dots)} E \left[\sum_{t=T_0}^{T_1} \rho^t u_{it}(x_t, \epsilon_t, a_t) \mid x_{T_0}, \epsilon_{T_0}, a_{T_0} \right] \quad (2.4)$$

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 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 (2.1), 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 (2.3)), 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 $\beta_i = ((\beta_{1ia}, \beta_{2ia}) : a \in \tilde{A}) \in \mathbb{R}^b$ for $b = (1 + p)|A|$ enters the model through the period utility function as follows:

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

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 $\tilde{A} = \{1, 2, \dots, A\}$, 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

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

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*

$$L_{T_1, \beta}^{E, \gamma, \tilde{E}, \tilde{\gamma}} : \mathcal{L}_{S_\beta} \rightarrow \mathcal{L}_{S_\beta} \quad [L_{T_1, \beta}^{E, \gamma, \tilde{E}, \tilde{\gamma}} m](b) = \left[\left((L_{T_1, \beta}^{E, \gamma})^{-1} L_{T_1, \beta}^{E, \tilde{\gamma}} - (L_{T_1, \beta}^{\tilde{E}, \gamma})^{-1} L_{T_1, \beta}^{\tilde{E}, \tilde{\gamma}} \right) m \right] (b) \quad (2.5)$$

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 , $L_{T_1, \beta}^{E, \gamma}$ is injective so that $L_{T_1, \beta}^{E, \gamma, \tilde{E}, \tilde{\gamma}}$ 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 $\forall x_2 \in S_{X_2}, \gamma \neq \tilde{\gamma}$

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

where the distribution of unobserved heterogeneity $f_{X_1 | \beta}^{X_2}$ 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

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

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

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

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 (2.5) 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 A.2.4 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 (2.4) 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 A.2.2 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.1.2 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.1.3 considers a limited form of finite dependence, and shows a binary choice model with random coefficients is identified.

2.3 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

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

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 (1993b) 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.

2.3.1 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 i th 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

$$\frac{1}{n} \sum_{i=1}^n \psi(y_i, \hat{F}_x, \hat{\gamma}, \hat{f}_{\beta|x_1}) \geq \sup_{(\gamma, f) \in \Gamma \times \mathcal{M}_n} \frac{1}{n} \sum_{i=1}^n \psi(y_i, \hat{F}_x, \gamma, f) - o_p(1/n) \quad (2.6)$$

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 2.6 is consistent for $(\gamma, f_{\beta|x_1})$.*

The full statement of Theorem 3 and its proof are contained in Appendix A.3.1.

2.3.2 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 a subset of $\{f : S_\beta \times S_1 \rightarrow [0, 1]\}$ and defined as:

$$\left\{ f : f(b, x_1) = \sum_{j=1}^{B(n)} \sum_{k=1}^{X(n)} P_{j,k} 1(b_j \leq b) 1(x_1 \in \mathcal{X}_{k,n}), P_{j,k} \geq 0, \sum_{j=1}^{B(n)} P_{j,k} = 1, b_j \in \mathcal{B}_n, \mathcal{X}_{k,n} \in \mathcal{X}_n \right\}, \quad (2.7)$$

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 (2.6), 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 A.3.2 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 2.6 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 par-

ticular, 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 (2.7) this is a convex optimization problem, with a unique global optimum and can be solved very efficiently. The profile estimator is formed as

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

2.3.3 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 6 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 A.3.3. 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.

2.4 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 2.3. 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:

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

Following the model of Altuğ and Miller (1998), x_1 can be interpreted as an average consumption value (see Section 2.5 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 $(a_{it}, x_{it} : t = 1, \dots, 8)_{i=1}^n$ 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.

Table 2.2 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

Table 2.2: 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.

		$n = 100$	$n = 500$	$n = 1,000$	$n = 10,000$
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

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 2.1 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

⁵To 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 $\frac{1}{M} \sum_{m=1}^M \int (\hat{f}_{\beta,m}(b) - f_\beta(b))^2 db$ where $M = 1,000$ is the number of replications.

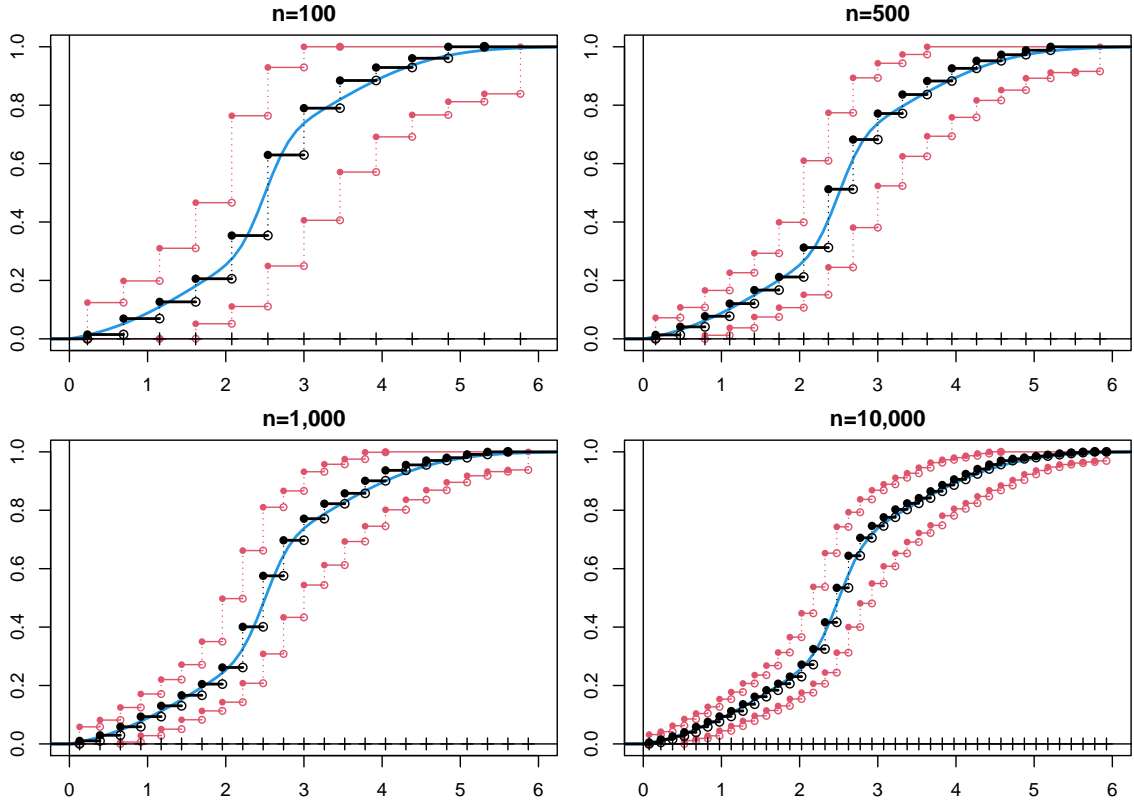


Figure 2.1: Simulation results for estimation of f_β for each sample size. The black curve represents the median estimate, the red curves pointwise 97.5%, 2.5% quantiles, and the blue curve the true distribution. The ticks on the x-axis represent the grid points.

the larger sample sizes.

2.5 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 4.2 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:

$$\eta_i \lambda_t \nu_i \omega_t \exp(\gamma'_3 x_{Wit}) l_{it} \quad (2.8)$$

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 4.2 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 par-

ticular, suppose the period payoff from entering the labor market for individual of type ν_i is:

$$u_i(x, \epsilon, 1) = x'(\nu_i, \gamma) + \epsilon_1 \quad (2.9)$$

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.

Table 2.3 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

Table 2.3: Estimates of γ .

Variable	Point estimate	Standard errors
Intercept	-0.4307	0.0774
Head-of-household income	2.530×10^{-3}	1.5572×10^{-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

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

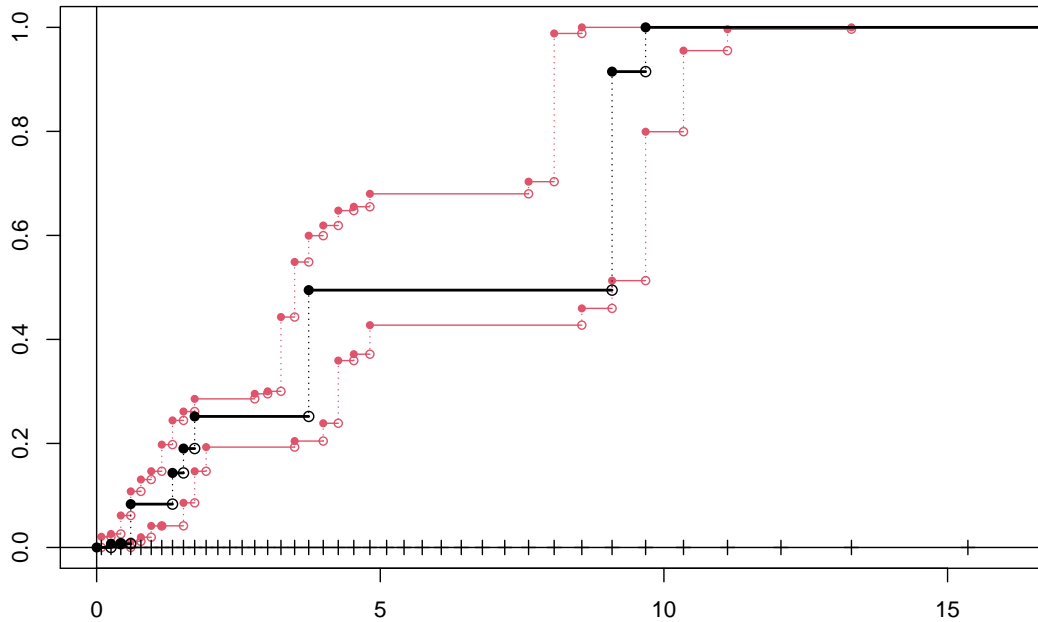


Figure 2.2: Estimated distribution of ν_i . The black curve represents the point estimate, the red curves represent bootstrapped 95% pointwise confidence intervals. The ticks on the x-axis represent the grid points.

2.6 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.

Chapter 3

Testing homogeneity in dynamic discrete games in finite samples

3.1 Introduction

In applications of dynamic discrete games, practitioners often assume that the conditional choice probabilities and the state transition probabilities are invariant across time and markets.¹ We refer to this as the “homogeneity assumption” in dynamic discrete games. This is a convenient assumption, as it allows the estimation of the model’s structural parameters by pooling data from multiple markets and from many time periods.

Despite the widespread use of the homogeneity assumption in dynamic discrete games, it is plausible for this condition to fail in applications. We now provide a few examples. First, a game could suffer from a structural change, which would invalidate the homogeneity assumption across time. Second, markets could be affected by permanent time-invariant heterogeneity that is observed to the players but not to the econometrician (e.g., Arcidiacono and Miller (2011a)). This would invalidate the homogeneity assumption across markets. Third and relatedly, there may be multiplicity of equilibria, and different markets could be playing different equilibria. The literature has considered hypothesis testing for the multiplicity of equilibria in games. In particular, de Paula and Tang (2012) propose a test for the multiplicity of equilibria across markets in static games, while Otsu, Pesendorfer, and Takahashi (2016) consider a test for the multiplicity of equilibria across markets in dynamic games.

¹In this paper, we use “market” to denote a cross-sectional unit.

In this paper, we propose a hypothesis test for the homogeneity assumption. Our test is implemented via Markov chain Monte Carlo (MCMC) methods, and it is justified by the theory of randomization tests (cf. Lehmann and Romano (2005, Section 15.2)). While our test is not exactly a randomization test, we establish its validity by coupling it with an underlying randomization test. The latter is exactly valid yet computationally infeasible. In this sense, we can interpret our proposed MCMC algorithm as a computationally feasible way to implement the infeasible randomization test. We formally show that the approximation error vanishes as the (user-defined) number of MCMC draws diverges. It is worth mentioning that our results hold for any fixed and finite number of players, markets, and time periods. This is an important aspect of our contribution, as the datasets used in empirical applications often have a small number of time periods and markets. For example, our empirical application is based on Ryan (2012), and has only $n = 23$ markets and either $T = 9$ or $T = 10$ time periods.

The econometric framework considered in this paper is arguably very general. It includes the single-agent dynamic discrete choice model (e.g., Rust (1987), Hotz and Miller (1993a), Hotz et al. (1994), and Aguirregabiria and Mira (2002a)) and the Markov equilibrium dynamic game model (e.g., Pakes, Ostrovsky, and Berry (2007), Aguirregabiria and Mira (2007a), Bajari, Benkard, and Levin (2007), Pesendorfer and Schmidt-Dengler (2008a), and Pesendorfer and Schmidt-Dengler (2010)). Furthermore, it includes the Markov dynamic game model of Aguirregabiria and Magesan (2020), which allows some players to have biased beliefs.

In a recent paper, Otsu, Pesendorfer, and Takahashi (2016) propose several hypothesis tests for dynamic discrete games. Some of their proposals are related to the problem considered in this paper. Specifically, they consider a method to test the homogeneity across markets of the conditional choice probabilities and the state

transition probabilities, under the maintained assumption that these functions are time-homogeneous. Their inference method is based on the bootstrap, and its validity is shown in an asymptotic framework in which the number of time periods T diverges to infinity. Unfortunately, the number of time periods in applications is often small. Besides the aforementioned application of Ryan (2012) with $T = 9$ or $T = 10$, we can mention Sweeting (2013) with $T = 4$, Collard-Wexler (2013) with $T = 24$, and Dunne et al. (2013) with $T = 5$.

The most critical step of our MCMC algorithm is based on the so-called Euler algorithm (Kandel et al. 1996). In related work, Besag and Mondal (2013) use this algorithm to test whether a time series of data has a time-homogeneous Markov structure. In terms of our setup, this corresponds to testing whether the data from a single market has a time-homogeneous state transition probability. Relative to this work, our paper incorporates several essential features of dynamic Markov discrete games. First, we recognize that the dataset in a typical dynamic game has information about actions and states. Second, our construction exploits the typical economic structure imposed in dynamic games, such as the conditional independence assumption (i.e., conditional on the current state variable, the current action variable is independent from the past information). This can be clearly evidenced in our MCMC algorithm, where we first transform the state variable data and then we transform the action variable data. Finally, while Besag and Mondal (2013) focus on data from a single market, our MCMC algorithm exploits the possibility that the data includes observations from multiple markets.² This is an important aspect of our contribution, as the datasets used in empirical applications usually include data multiple markets, e.g.,

²Besag and Mondal (2013, Section 5) briefly mentions how their methods could be extended to the multiple market case. Unfortunately, they do not explain how this can be implemented nor they justify its validity. In contrast, the hypothesis test that we propose is the result of a different procedure than theirs, and we prove its validity by connecting it with the theory of randomization tests.

Ryan (2012) with $n = 23$, Sweeting (2013) with $n = 102$, Collard-Wexler (2013) with $n = 1,600$, and Dunne et al. (2013) with $n = 639$.

We explore the performance of our hypothesis test in Monte Carlo simulations. Our results show that our method provides excellent size control even in small samples, and can successfully detect relatively small deviations from the homogeneity hypothesis. In these two accounts, our test appears to work favorably in comparison with the bootstrap-based test in Otsu, Pesendorfer, and Takahashi (2016). As an empirical application, we investigate the homogeneity of the decisions in the U.S. Portland cement industry data used in Ryan (2012). This is a key assumption in Ryan (2012), as it allows him to pool data from multiple markets to estimate the model's parameters. Unlike Otsu, Pesendorfer, and Takahashi (2016)'s test, our test finds no evidence against the homogeneity hypothesis in the data.

The rest of the paper is organized as follows. Section 4.2 describes the dynamic discrete choice model and the hypothesis test. Section 3.3 specifies our hypothesis test and its implementation via the MCMC algorithm. In Section 3.4, we show that our test is an approximate implementation of a computationally infeasible randomization test. In Section 3.5, we evaluate the performance of our test in finite samples via Monte Carlo simulations. Section 3.6 considers an empirical application based on Ryan (2012). Section 4.5 concludes. The paper's appendix collects all of the proofs, several auxiliary results, and computational details related to the proposed MCMC algorithm.

3.2 The econometric model and the testing problem

3.2.1 The econometric model

We begin by describing the dynamic discrete game under consideration. We observe the outcome of n markets in which J players choose actions over T time periods. Our setup allows for $J = 1$, i.e., single-agent problems, or $J > 1$, i.e., multiple-agent games. This paper's inference results are valid for all finite n , T , and J .

We consider a setup in which the observed actions and state variables are discretely distributed, which is common in the dynamic discrete choice literature. For every market $i = 1, \dots, n$ and period $t = 1, \dots, T$, let $A_{i,t}$ be the random variable that specifies the actions chosen by the players in market i and period t , and let $S_{i,t}$ be the random variable that specifies the state variable of market i and period t . We define the following $n \times T$ matrices:

$$S \equiv (S_{i,t} : i = 1, \dots, n, t = 1, \dots, T),$$

$$A \equiv (A_{i,t} : i = 1, \dots, n, t = 1, \dots, T).$$

In this notation, the data are then given by

$$X \equiv (S, A).$$

We assume the common support of $S_{i,t}$ is a finite set \mathcal{S} , and the common support of $A_{i,t}$ is a finite set \mathcal{A} . Then, the support of X is represented by $\mathcal{X} \equiv \mathcal{S}^{nT} \times \mathcal{A}^{nT}$.

Remark 4. We have thus far assumed that we observe a balanced panel, i.e., all n markets are fully observed over T time periods. This is only for the simplicity of notation and exposition. All of the arguments in our paper extend immediately to the case in which each market $i = 1, \dots, n$ is observed over a market-specific time period T_i .

Throughout this paper, we maintain the following assumption.

Assumption 3.2.1. *The following conditions hold:*

- (a) $((S_{i,t}, A_{i,t}) : t = 1, \dots, T)$ are independent across $i = 1, \dots, n$.
- (b) $(S_{i,t}, A_{i,t})$ and $(S_{i,1}, A_{i,1}, \dots, S_{i,t-2}, A_{i,t-2})$ are conditionally independent given $(S_{i,t-1}, A_{i,t-1})$ for every $i = 1, \dots, n$ and $t = 3, \dots, T$.
- (c) $A_{i,t}$ and $(S_{i,t-1}, A_{i,t-1})$ are conditionally independent given $S_{i,t}$ for every $i = 1, \dots, n$ and $t = 2, \dots, T$.

Assumption 3.2.1 is standard in much of the literature on dynamic discrete games. Assumption 3.2.1(a) imposes that markets are independently distributed. Assumption 3.2.1(b) indicates that the observations of state and actions are a Markov process. Assumption 3.2.1(c) imposes that the current actions are independent of past information once we condition on the current state. Assumptions 3.2.1(b)-(c) are high-level restrictions that are typically imposed on the equilibrium strategies used by the players. In particular, they follow from the assumption that players use Markov strategies Maskin and Tirole (2001), as assumed in Pakes, Ostrovsky, and Berry (2007), Aguirregabiria and Mira (2007a), Bajari, Benkard, and Levin (2007), and Pesendorfer and Schmidt-Dengler (2008a). These conditions are implied even in models in which the players' beliefs are allowed to be out of equilibrium, i.e., do not coincide with the true equilibrium probabilities (e.g., Aguirregabiria and Magesan (2020)).

We now introduce necessary notation to express our hypothesis of interest. We use $\sigma_{i,t}$ to denote the conditional choice probability for market i and period t , i.e., for every $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$\sigma_{i,t}(a|s) \equiv P(A_{i,t} = a | S_{i,t} = s).$$

We use $g_{i,t+1}$ to denote the state transition probability from period t to $t + 1$ for market i , i.e., for every $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$,

$$g_{i,t+1}(s'|a, s) \equiv P(S_{i,t+1} = s' | (S_{i,t}, A_{i,t}) = (s, a)).$$

We use $m_i(s)$ to denote the marginal state distribution for market i in period 1, i.e., for every $s \in \mathcal{S}$,

$$m_i(s) \equiv P(S_{i,1} = s).$$

With this notation in place, we specify our hypothesis testing problem in the next section.

3.2.2 The hypothesis testing problem

Our goal is to test whether the “homogeneity assumption” holds in the data, i.e., whether the conditional choice probabilities and state transition probabilities are homogeneous across time and markets. That is,

$$H_0 : \sigma_{i,t}(a|s) = \sigma(a|s) \text{ and } g_{i,t+1}(s'|a, s) = g(s'|a, s) \quad \text{vs.} \quad H_1 : H_0 \text{ is false.} \quad (3.1)$$

Note that H_0 in (3.1) represents two types of homogeneity: time and market homogeneity, and involves two functions: conditional choice probabilities and state transition probabilities. In this sense, our hypothesis test evaluates four homogeneity conditions: time homogeneity of the conditional choice probabilities, market homogeneity of the conditional choice probabilities, time homogeneity of the state transition probabilities, and market homogeneity of the state transition probabilities. A rejection of H_0 in (3.1) would be indicative that one or more of these homogeneity conditions is violated. In certain applications, however, one may feel comfortable that some of the conditions are satisfied and should be part of our maintained assumptions. For example, in a given application, one may be confident that the conditional choice probability and state transition probability are time-homogeneous. Then, one could reinterpret H_0 in (3.1) as testing the market homogeneity of the conditional choice probabilities and state transition probabilities.

Under Assumption 3.2.1 and H_0 in (3.1), Lemma B.2.1 in the appendix shows that the likelihood of the data $X = (S, A)$ evaluated at any realization $\tilde{X} = (\tilde{S}, \tilde{A}) \in \mathcal{X}$

is as follows:

$$P(X = \tilde{X}) = \prod_{i=1}^n \left(m_i(\tilde{S}_{i,1}) \left(\prod_{t=1}^T \sigma(\tilde{A}_{i,t} | \tilde{S}_{i,t}) \right) \left(\prod_{t=1}^{T-1} g(\tilde{S}_{i,t+1} | \tilde{S}_{i,t}, \tilde{A}_{i,t}) \right) \right). \quad (3.2)$$

This expression reveals that the markets are independently distributed (Assumption 3.2.1(a)), but they are not necessarily identically distributed because $m_i(\cdot)$ depends on i . Even though the conditional choice probabilities and state transition probabilities are homogeneous under H_0 , markets can still be heterogeneous due to differences in their initial state values. This is a desired feature in our testing problem, as the dynamic discrete choice literature usually allows the initial state distribution to be market-specific.

3.3 Our hypothesis test

In this paper, we propose to reject H_0 in (3.1) whenever the significance level α is larger than or equal to our p -value, which we denote by \hat{p}_K . That is,

$$\phi_K(X) \equiv 1\{\hat{p}_K \leq \alpha\}. \quad (3.3)$$

In turn, our p -value \hat{p}_K is the result of constructing K transformations of the data via our MCMC algorithm, which we describe in Section 3.3.1 (see Algorithm 3.3.1). Our MCMC algorithm produces K sequential transformations of the data X , denoted by $(X^{(1)}, \dots, X^{(K)})$. Our p -value is then computed as follows

$$\hat{p}_K \equiv \frac{1}{K} \sum_{k=1}^K 1\{\tau(X^{(k)}) \geq \tau(X)\}, \quad (3.4)$$

where $\tau : \mathcal{X} \rightarrow \mathbb{R}$ denotes the test statistic designed to detect departures from H_0 in the data.³ One notable feature of our hypothesis test is that its validity will not

³For example, (3.20) in Section 3.5 considers two possible test statistics designed to detect departures from market homogeneity in the conditional choice probability function.

depend on the choice of the test statistic. The following is the main result of this paper.

Theorem 5. *Under H_0 in (3.1), the test in (3.3) satisfies*

$$\limsup_{K \rightarrow \infty} E[\phi_K(X)] \leq \alpha, \quad (3.5)$$

where the expectation is taken with respect to the randomness in $(X, X^{(1)}, \dots, X^{(K)})$, i.e., both in the data X and in the random draws $(X^{(1)}, \dots, X^{(K)})$ generated by our MCMC algorithm.

Theorem 5 establishes that the proposed test in (3.3) controls size as the length of the MCMC draws diverges. While this is an approximate result for a finite K , we note that the researcher controls the number of MCMC draws and that the approximation error becomes negligible by choosing a large value of K . Remarkably, Theorem 5 holds regardless of the number of markets n , time periods T , and players J , which can remain constant in our analysis. In addition and as promised, this result also holds irrespective of the specific choice of test statistic τ used in the construction of the p -value in (3.4).

Our proposed test can be related to randomization tests, e.g., Lehmann and Romano (2005, Chapter 15.2). In particular, Theorem 5 follows from showing that the p -value in (3.4) approximates the p -value of a (computationally infeasible) randomization test for H_0 in (3.1). This observation provides intuition as to why Theorem 5 does not require the number of markets n , time periods T , or players J to grow. We provide a detailed explanation and additional results in Section 3.4. In the rest of this section, we introduce the MCMC algorithm used to construct $(X^{(1)}, \dots, X^{(K)})$.

3.3.1 The MCMC algorithm

Our MCMC algorithm requires some notation. Let $I = \{I_1, I_2\}$ denote an arbitrary pair of markets I_1 and I_2 in the data, i.e., $I_1, I_2 = 1, 2, \dots, n$. We allow for $I_1 = I_2$. We use \mathcal{I} to denote the collection of all such pairs of markets, i.e., $|\mathcal{I}| = n^2$. We also define several sets.

Definition 3.3.1. For any $I \in \mathcal{I}$ and $\check{S} \in \mathcal{S}^{nT}$, $R_S(I, \check{S})$ is the set of all $\tilde{S} \in \mathcal{S}^{nT}$ satisfying the following conditions:

- (a) $\tilde{S}_{i,1} = \check{S}_{i,1}$ for all $i = 1, \dots, n$,
- (b) $\sum_{t=1}^{T-1} 1\{\tilde{S}_{i,t} = s, \tilde{S}_{i,t+1} = s'\} = \sum_{t=1}^{T-1} 1\{\check{S}_{i,t} = s, \check{S}_{i,t+1} = s'\}$ for all $s, s' \in \mathcal{S}$ and $i \in I^c$,
- (c) $\sum_{i \in I} \sum_{t=1}^{T-1} 1\{\tilde{S}_{i,t} = s, \tilde{S}_{i,t+1} = s'\} = \sum_{i \in I} \sum_{t=1}^{T-1} 1\{\check{S}_{i,t} = s, \check{S}_{i,t+1} = s'\}$ for all $s, s' \in \mathcal{S}$.

In words, $R_S(I, \check{S})$ is the set of all state configurations that result from “mixing” the hypothetical state data \check{S} , subject to certain restrictions (given by conditions (a)-(c)). Under H_0 , these restrictions imply that each state configuration in $R_S(I, \check{S})$ has the same value of the likelihood function, provided that it is paired with a suitable action configuration. The corresponding suitable action configurations are precisely those in next definition.

Definition 3.3.2. For any $\tilde{S}, \check{S} \in \mathcal{S}^{nT}$ and $\check{A} \in \mathcal{A}^{nT}$, $R_A(\tilde{S}, (\check{S}, \check{A}))$ is the set of all $\tilde{A} \in \mathcal{A}^{nT}$ satisfying the following conditions:

- (a) $\sum_{i=1}^n \sum_{t=1}^{T-1} 1\{\tilde{S}_{i,t} = s, \tilde{A}_{i,t} = a, \tilde{S}_{i,t+1} = s'\} = \sum_{i=1}^n \sum_{t=1}^{T-1} 1\{\check{S}_{i,t} = s, \check{A}_{i,t} = a, \check{S}_{i,t+1} = s'\}$ for all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$,

(b) $\sum_{i=1}^n 1\{\tilde{S}_{i,T} = s, \tilde{A}_{i,T} = a\} = \sum_{i=1}^n 1\{\check{S}_{i,T} = s, \check{A}_{i,T} = a\}$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

By definition, $R_A(\tilde{S}, (\check{S}, \check{A}))$ is the set of action configurations that result from “mixing” the hypothetical action data \check{A} , subject to certain restrictions (given by conditions (a)-(b)). Under H_0 , these restrictions imply that the hypothetical data (\check{S}, \check{A}) has the same likelihood as the state configuration \tilde{S} paired with any action configuration in $R_A(\tilde{S}, (\check{S}, \check{A}))$. With these definitions in place, we can now specify our MCMC algorithm.

Algorithm 3.3.1 (MCMC algorithm). Let $(X^{(1)}, \dots, X^{(K)})$ denote the following Markov chain.

Initiation. Initiate the chain with $X^{(1)} = X$.

Iteration. The rest of the chain is iteratively generated as follows. For any $k = 2, \dots, K$ and given $(X^{(1)}, \dots, X^{(k-1)})$, $X^{(k)} = (S^{(k)}, A^{(k)})$ is randomly generated as follows:

Step 1: Draw $I^{(k)}$ uniformly from \mathcal{I} , independently of $(X^{(1)}, \dots, X^{(k-1)})$.

Step 2: Given $(X^{(1)}, \dots, X^{(k-1)}, I^{(k)})$, draw $S^{(k)}$ uniformly from $R_S(I^{(k)}, S^{(k-1)})$.

Step 3: Given $(X^{(1)}, \dots, X^{(k-1)}, I^{(k)}, S^{(k)})$, draw $A^{(k)}$ uniformly from $R_A(S^{(k)}, X^{(k-1)})$. ■

Several comments are in order. Steps 2 and 3 require randomly drawing state and action configurations uniformly over the sets $R_S(I^{(k)}, S^{(k-1)})$ and $R_A(S^{(k)}, X^{(k-1)})$, respectively. On the one hand, Step 3 is relatively easy to implement by permuting the action data in $A^{(k-1)}$ subject to the restrictions in $R_A(S^{(k)}, X^{(k-1)})$. Algorithm B.1.3 in Section B.1.2 explains how to implement this in practice and provides a

justification (Lemma B.1.5). On the other hand, Step 2 is more involved. We propose to implement it using a modified version of the Euler algorithm (Kandel et al. 1996; Besag and Mondal 2013). Section B.1.1 describes the original Euler algorithm (Algorithm B.1.1), our modification (Algorithm B.1.2), and formally shows that the latter satisfy Step 2 in Lemma B.1.2.

For any $k = 2, \dots, K$, $X^{(1)}, \dots, X^{(k-1)} \in \mathcal{X}$, $I \in \mathcal{I}$, and $\tilde{X} = (\tilde{S}, \tilde{A}) \in \mathcal{X}$, our MCMC algorithm implies the following transition probabilities:

$$P(I^{(k)} = I \mid X^{(1)}, \dots, X^{(k-1)}) = \frac{1}{|\mathcal{I}|}, \quad (3.6)$$

$$P(S^{(k)} = \tilde{S} \mid I^{(k)}, X^{(1)}, \dots, X^{(k-1)}) = \frac{1\{\tilde{S} \in R_S(I^{(k)}, S^{(k-1)})\}}{|R_S(I^{(k)}, S^{(k-1)})|}, \quad (3.7)$$

$$P(A^{(k)} = \tilde{A} \mid S^{(k)}, I^{(k)}, X^{(1)}, \dots, X^{(k-1)}) = \frac{1\{\tilde{A} \in R_A(S^{(k)}, X^{(k-1)})\}}{|R_A(S^{(k)}, X^{(k-1)})|}. \quad (3.8)$$

Note that (3.7) and (3.8) are well defined, as both denominators can be shown to be positive. In turn, these transition probabilities imply that our MCMC algorithm has the following transition probability:

$$P(X^{(k)} = \tilde{X} \mid X^{(1)}, \dots, X^{(k-1)}) = \begin{cases} \sum_{I \in \mathcal{I}} \frac{1\{\tilde{S} \in R_S(I, S^{(k-1)}), \tilde{A} \in R_A(\tilde{S}, X^{(k-1)})\}}{|\mathcal{I}| \times |R_S(I, S^{(k-1)})| \times |R_A(\tilde{S}, X^{(k-1)})|} & \text{if } |R_S(I, S^{(k-1)})| \times |R_A(\tilde{S}, X^{(k-1)})| > 0, \\ 0 & \text{otherwise.} \end{cases} \quad (3.9)$$

3.4 Our test as an approximate randomization test

This section provides the formal arguments that are necessary to prove Theorem 5 and, thus, justify our hypothesis test in (3.3). In particular, we show that our MCMC-based p -value in (3.4) is an approximation of the p -value of a specific ran-

domization test. We argue that this randomization test is valid in finite samples but computationally infeasible, which explains why we propose the MCMC algorithm to approximate its p -value.

This section is organized as follows. Section 3.4.1 provides an alternative representation of the likelihood of the data under H_0 in (3.1). This result allows us to define a sufficient statistic of the data under H_0 , denoted by $U(X)$. Section 3.4.2 introduces a transformation group of the data, which does not change the value of the sufficient statistic $U(X)$.⁴ Section 3.4.3 defines a specific randomization test for (3.1), and argues that is both finite-sample valid and computationally infeasible. Section 3.4.4 shows that our MCMC-based test in (3.3) can successfully approximate the infeasible randomization test as the number of MCMC draws diverges.

3.4.1 An alternative representation of the likelihood

The following result provides an alternative representation of the likelihood under H_0 in (3.1).

Lemma 3.4.1. *Under Assumption 3.2.1 and H_0 in (3.1), the likelihood of the data $X = (S, A)$ evaluated at $\tilde{X} = (\tilde{S}, \tilde{A}) \in \mathcal{X}$ with $\tilde{S} = (\tilde{S}_{i,t} : i = 1, \dots, n, t = 1, \dots, T) \in \mathcal{S}^{nT}$ and $\tilde{A} = (\tilde{A}_{i,t} : i = 1, \dots, n, t = 1, \dots, T) \in \mathcal{A}^{nT}$ is*

$$P(X = \tilde{X}) = P(A = \tilde{A} | S = \tilde{S}) \times P(S = \tilde{S}), \quad (3.10)$$

where

$$\begin{aligned} P(A = \tilde{A} | S = \tilde{S}) = & \prod_{(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}} \left(\frac{\sigma(a|s)g(s'|s,a)}{\sum_{\tilde{a} \in \mathcal{A}} g(s'|\tilde{a},s)\sigma(\tilde{a}|s)} \right)^{\sum_{i=1}^n \sum_{t=1}^{T-1} 1_{\{\tilde{S}_{i,t}=s, \tilde{A}_{i,t}=a, \tilde{S}_{i,t+1}=s'\}}} \\ & \times \prod_{(s,a) \in \mathcal{S} \times \mathcal{A}} \sigma(a|s)^{\sum_{i=1}^n 1_{\{\tilde{S}_{i,T}=s, \tilde{A}_{i,T}=a\}}} \end{aligned} \quad (3.11)$$

⁴Hereafter, we use “transformation group” to denote the set defined in Lehmann and Romano (2005, pages 693-4).

and

$$P(S = \tilde{S}) = \left(\prod_{i=1}^n m_i(\tilde{S}_{i,1}) \right) \times \prod_{(s,s') \in \mathcal{S} \times \mathcal{S}} \left(\sum_{a \in \mathcal{A}} g(s'|a, s) \sigma(\tilde{a}|s) \right)^{\sum_{i=1}^n \sum_{t=1}^{T-1} 1\{\tilde{S}_{i,t}=s, \tilde{S}_{i,t+1}=s'\}}. \quad (3.12)$$

From this result, we can deduce the following corollary.

Corollary 2. *Under Assumption 3.2.1 and H_0 in (3.1), the sufficient statistic for $X = (S, A)$ is*

$$U(X) = \left(\begin{array}{l} (S_{i,1} : i = 1 \dots, n), \\ \left(\sum_{i=1}^n \sum_{t=1}^{T-1} 1\{S_{i,t} = s, A_{i,t} = a, S_{i,t+1} = s'\} : (s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \right), \\ \left(\sum_{i=1}^n 1\{S_{i,T} = s, A_{i,T} = a\} : (s, a) \in \mathcal{S} \times \mathcal{A} \right) \end{array} \right). \quad (3.13)$$

Corollary 2 implies that, under H_0 , any transformation of the data that does not change the value of $U(X)$ does not affect the value of the likelihood. This observation provides the basis of the randomization test that we consider in the remaining sections.

3.4.2 A transformation group related to the proposed MCMC algorithm

Our proposed MCMC algorithm can be understood as an iteration of transformations to the data X . In particular, $X^{(1)} = X$ is the identity transformation, $X^{(2)}$ follows from applying Steps 1-3 to X , $X^{(3)}$ follows from applying Steps 1-3 twice to X , and so forth. In this section, we define a transformation group that is related to the transformations in our MCMC algorithm. To define this properly, we first require the following definition.

Definition 3.4.1. For any pair of markets $I = \{I_1, I_2\} \in \mathcal{I}$, let $\mathbf{G}(I)$ denote the set of all transformations of \mathcal{X} onto itself such that, for any $g \in \mathbf{G}(I)$ and $(\check{S}, \check{A}) \in \mathcal{X}$, $(\tilde{S}, \tilde{A}) = g(\check{S}, \check{A})$ satisfies $\tilde{S} \in R_S(I, \check{S})$ and $\tilde{A} \in R_A(\tilde{S}, (\check{S}, \check{A}))$.

Lemma B.2.3 in the appendix shows that $\mathbf{G}(I)$ is a transformation group. By Definition 3.4.1, $\mathbf{G}(I)$ is the group representation of Steps 2-3 of our MCMC algorithm. Given a randomly chosen pair of markets $I^{(k)}$ in Step 1, Steps 2-3 obtain the next element of the Markov chain $X^{(k)} = (S^{(k)}, A^{(k)})$ by applying a randomly chosen transformation in $\mathbf{G}(I^{(k)})$ to the preceding element of the Markov chain, $X^{(k-1)}$. In this sense, Steps 2-3 of our MCMC algorithm are a specific way of choosing a particular transformation in $\mathbf{G}(I^{(k)})$.

By the description in the previous paragraph, our MCMC algorithm randomly chooses transformations in $\mathbf{G}(I)$ for random pairs of markets I , and iteratively applies them to the data. These iterative transformations are related to the set that we define next.

Definition 3.4.2. Let \mathbf{G} be the set of all finitely many compositions of the elements in $\bigcup_{I \in \mathcal{I}} \mathbf{G}(I)$.

The next result states that \mathbf{G} is a transformation group with desirable properties.

Lemma 3.4.2. $\mathbf{G} : \mathcal{X} \rightarrow \mathcal{X}$ is a transformation group of \mathcal{X} such that, for any $g \in \mathbf{G}$ and $\tilde{X} \in \mathcal{X}$, \tilde{X} and $g\tilde{X}$ have the same sufficient statistic in (3.13), i.e., $U(\tilde{X}) = U(g\tilde{X})$.

The properties shown in Lemma 3.4.2 imply that we can use \mathbf{G} to define a valid randomization test. We do this in Section 3.4.3.

3.4.3 A randomization test

Following Lehmann and Romano (2005, Chapter 15.2), we can use the transformation group \mathbf{G} to define a randomization test. This test rejects H_0 in (3.1) whenever the significance level α is larger than or equal to the randomization p -value, which we denote by \hat{p} . That is,

$$\phi(X) \equiv 1\{\hat{p} \leq \alpha\}, \quad (3.14)$$

where

$$\hat{p} \equiv \frac{1}{|\mathbf{G}|} \sum_{g \in \mathbf{G}} 1\{\tau(gX) \geq \tau(X)\}. \quad (3.15)$$

By the arguments in Lehmann and Romano (2005, Page 636), the randomization test in (3.14) is finite-sample valid. We record this in the next result.

Lemma 3.4.3. *Under H_0 in (3.1) and for any $\alpha \in (0, 1)$, the test in (3.14) satisfies*

$$E[\phi(X)] \leq \alpha. \quad (3.16)$$

Unlike our proposed test in (3.3), the hypothesis test in (3.14) is computationally infeasible in typical applications of dynamic discrete choice games. The basic reason is that the transformation group \mathbf{G} is usually impossible to enumerate. To see why, note that \mathbf{G} is a set of transformations restricted by the condition on the sufficient statistics in (3.13). This condition is hard to impose without explicitly verifying that it holds. In turn, an explicit verification of this condition is not practically feasible, as it would require exploring all possible transformations that map \mathcal{X} to \mathcal{X} . Even in applications in which n , T , and $|\mathcal{A}|$ and $|\mathcal{S}|$ are relatively small, the resulting state space of the data $\mathcal{X} = \mathcal{S}^{nT} \times \mathcal{A}^{nT}$ can be overwhelming.

In the randomization testing literature, it is not uncommon for the transformation set \mathbf{G} to be huge. As Lehmann and Romano (2005, page 636) explains, one can still implement a random version of the test in (3.14) by drawing randomly from \mathbf{G} in a

uniform fashion. This point is routinely exploited in standard settings to construct tests based on permutations or sign changes. To the best of our knowledge, however, there is no feasible way of obtaining such random draws in the current setup, as the condition on the sufficient statistics in (3.13) is hard to impose without explicitly checking whether it holds. This explains why we cannot directly exploit the finite-sample result in Lemma 3.4.3. In any case, Section 3.4.4 reveals that our MCMC-based p -value in (3.4) approximates the infeasible p -value in (3.15) as the length of the MCMC diverges.

3.4.4 An MCMC approximation to the randomization test

Our main theoretical result is Theorem 5, which shows that the test in (3.3) controls size as the number of MCMC draws K diverges to infinity. The following lemma provides a fundamental stepping stone to prove this result.

Lemma 3.4.4. *Conditional on X ,*

$$\sup_{t \in \mathbb{R}} \left| \frac{1}{K} \sum_{k=1}^K 1\{\tau(X^{(k)}) \geq t\} - \frac{1}{|\mathbf{G}|} \sum_{g \in \mathbf{G}} 1\{\tau(gX) \geq t\} \right| \xrightarrow{a.s.} 0 \quad \text{as } K \rightarrow \infty.$$

Lemma 3.4.4 shows that, as the number of MCMC draws diverges, the conditional distribution based on the MCMC algorithm converge to the conditional distribution of the computationally infeasible randomization test described in Section 3.4.3. By applying Lemma 3.4.4 with $t = \tau(X)$, we can deduce that the p -value in (3.4) approximates the p -value in (3.15) as the number of MCMC draws K diverges. That is, conditional on X ,

$$\hat{p}_K \xrightarrow{a.s.} \hat{p} \quad \text{as } K \rightarrow \infty.$$

By combining this observation with the finite-sample validity of the infeasible randomization test in (3.14) (Lemma 3.4.3), it follows that our proposed MCMC-test

becomes valid as the number of MCMC draws K diverges. This argument provides the intuition behind Theorem 5 (see Section B.1.3 of the appendix for the proof), and why it holds regardless of the number of markets n , time periods T , and players J .

3.5 Monte Carlo simulations

In this section, we explore the performance of our proposed test in Monte Carlo simulations. We consider the Monte Carlo design used by Otsu, Pesendorfer, and Takahashi (2016, Section 4), which follows from the dynamic oligopoly discrete game in Pesendorfer and Schmidt-Dengler (2008a, Section 7.1). We refer to these papers for the details on the setup. The simulated data are generated by two oligopolic firms deciding whether to enter or not into n markets, and over T time periods. This dynamic game has multiple equilibria, which we exploit to generate departures from the homogeneity assumption.

In each period $t = 1, \dots, T$ and market $i = 1, \dots, n$, there are four possible actions in this game: $A_{i,t} = 1$ denotes that neither firm entered the market, $A_{i,t} = 2$ denotes that only firm 2 enters, $A_{i,t} = 3$ denotes that only firm 1 enters, and $A_{i,t} = 4$ denotes that both firms enter. This implies that $\mathcal{A} = \{1, 2, 3, 4\}$. In addition, the state is equal to the last period's action, i.e.,

$$S_{i,t} = A_{i,t-1}, \tag{3.17}$$

and so $\mathcal{S} = \{1, 2, 3, 4\}$. Note that (3.17) implies that the state transition probabilities are homogeneous, and given by

$$g_{i,t+1}(s'|a, s) = 1\{s' = a\}. \tag{3.18}$$

We presume that (3.18) is known to the researcher, who replaces H_0 in (3.1) with the homogeneity of the conditional choice probabilities. In other words, the relevant

hypothesis testing problem is

$$H_0 : \sigma_{i,t}(a|s) = \sigma(a|s) \quad \text{vs.} \quad H_1 : H_0 \text{ is false.} \quad (3.19)$$

Following Otsu, Pesendorfer, and Takahashi (2016, Eq. (4), (7)), we consider the following test statistics

$$\begin{aligned} \tau_1(X) &\equiv \sum_{i=1}^n \sum_{(a,s) \in \mathcal{A} \times \mathcal{S}} (\hat{\sigma}_i(a|s) - \hat{\sigma}(a|s))^2 \left(\frac{\sum_{t=1}^T \mathbf{1}\{S_{i,t} = s\}}{\hat{\sigma}(a|s)} \right) \\ \tau_2(X) &\equiv 2 \sum_{i=1}^n \sum_{(a,s) \in \mathcal{A} \times \mathcal{S}} \hat{\sigma}_i(a|s) \log \left(\frac{\hat{\sigma}_i(a|s)}{\hat{\sigma}(a|s)} \right) \sum_{t=1}^T \mathbf{1}\{S_{i,t} = s\}, \end{aligned} \quad (3.20)$$

where we interpret $0/0 = 0$ and $0 \times \log(0) = 0$, and we define

$$\begin{aligned} \hat{\sigma}_i(a|s) &\equiv \frac{\sum_{t=1}^T \mathbf{1}\{A_{i,t} = a, S_{i,t} = s\}}{\sum_{t=1}^T \mathbf{1}\{S_{i,t} = s\}} \\ \hat{\sigma}(a|s) &\equiv \frac{\sum_{i=1}^n \sum_{t=1}^T \mathbf{1}\{A_{i,t} = a, S_{i,t} = s\}}{\sum_{i=1}^n \sum_{t=1}^T \mathbf{1}\{S_{i,t} = s\}}. \end{aligned}$$

The statistics in (3.20) compute weighted differences between market-specific empirical conditional choice probabilities and the overall counterpart.

The data produced by this game is a matrix $X = (S, A) \in \mathcal{X}$ constructed exactly as in Otsu, Pesendorfer, and Takahashi (2016, Section 4). We simulate data from a mixture of two data generating processes: DGP 1 and DGP 2. They represent Markov perfect equilibria of the dynamic game, which differ in the conditional choice probabilities $\sigma(a|s)$. In DGP 1, the matrix of conditional choice probabilities is

$$\begin{pmatrix} 0.19 & 0.30 & 0.12 & 0.18 \\ 0.08 & 0.09 & 0.08 & 0.07 \\ 0.53 & 0.48 & 0.46 & 0.53 \\ 0.20 & 0.13 & 0.34 & 0.22 \end{pmatrix},$$

where the columns index the value of the state $s \in \mathcal{S} = \{1, 2, 3, 4\}$, and the rows index the value of the action $a \in \mathcal{A} = \{1, 2, 3, 4\}$. In DGP 2, the corresponding

matrix of conditional choice probabilities is

$$\begin{pmatrix} 0.18 & 0.48 & 0.03 & 0.16 \\ 0.20 & 0.21 & 0.14 & 0.23 \\ 0.29 & 0.22 & 0.13 & 0.26 \\ 0.33 & 0.09 & 0.70 & 0.35 \end{pmatrix}.$$

Each market is sampled independently. Market $i = 1, \dots, n$ behaves according to DGP 1 if $i/n \leq \lambda$ and to DGP 2 if $i/n > \lambda$. Therefore, $\lambda \in [0, 1]$ represents the proportion of markets that are in DGP 1. Each market is initialized with state equal to 1, and we simulate the corresponding action according to the corresponding conditional choice probabilities. This, in turn, determines the next period's state according to (3.17). We then proceed iteratively until we have simulated $T + 100$ periods for each market. Then, the first 100 periods are discarded, producing a sample of T periods for n markets, which is the data observed by the researcher.

For each simulated data, we implement our proposed test in (3.3) with $K = 10,000$. We consider simulations with $n \in \{20, 40, 80, 160\}$, $T \in \{5, 10, 20, 40, 80\}$, and $\lambda \in \{1, 0, 0.5, 0.9\}$. As explained earlier, λ represents the proportion of markets that are in DGP 1. If $\lambda = 1$ or $\lambda = 0$, all markets are sampled from the same distribution, and so the conditional choice probabilities are homogeneous across markets. This means that H_0 in (3.19) holds. In turn, if $\lambda = 0.5$ or $\lambda = 0.9$, each data is composed of markets from both distributions, and so the conditional choice probabilities are not homogeneous across markets. This means that H_0 in (3.19) fails to hold. Note that $\lambda = 0.5$ generates data in which both distributions are equally represented, and so the heterogeneity in the conditional choice probabilities should be more salient. On the other hand, the case with $\lambda = 0.9$ produces data with a vast majority of markets in DGP 1, and so the heterogeneity in the conditional choice probabilities should be harder to detect. For each simulation design, we compute rejection rates based on 1,000 independently simulated datasets.

The results from the Monte Carlo simulation are shown in Table 3.1 for $\lambda \in \{0, 1\}$ and Table 3.2 for $\lambda \in \{0.5, 0.9\}$, respectively. For the sake of comparison, we also include the results from the test proposed by Otsu, Pesendorfer, and Takahashi (2016). Their test compares the same test statistics in (3.20) with critical values based on the bootstrap. As mentioned earlier, they show the validity of their test in an asymptotic framework with $T \rightarrow \infty$ and n fixed. In contrast, our main result in Theorem 5 is valid for any finite n and T .

Table 3.1: Simulation results under H_0 for $\alpha = 5\%$ based on 1,000 i.i.d. simulation draws. The results for $\lambda = 1$ corresponds to data sampled from DGP 1 and $\lambda = 0$ corresponds to data sampled from DGP 2. The test statistics $\tau_1(X)$ and $\tau_2(X)$ are defined in (3.20). Our test is computed according to (3.3) with $K = 10,000$. OPT refers to Otsu, Pesendorfer, and Takahashi (2016), whose results were copied from Tables 1 and 2 in their paper.

n	T	DGP 1 ($\lambda = 1$)				DGP 2 ($\lambda = 0$)			
		Our test		OPT's test		Our test		OPT's test	
		$\tau_1(X)$	$\tau_2(X)$	$\tau_1(X)$	$\tau_2(X)$	$\tau_1(X)$	$\tau_2(X)$	$\tau_1(X)$	$\tau_2(X)$
20	5	5.6	5.2	13.2	5.9	4.8	4.6	13.5	12.7
20	10	5.5	4.9	7.0	4.5	6.4	5.4	7.9	8.4
20	20	4.1	4.9	4.4	5.0	4.9	5.2	5.8	7.1
20	40	6.0	5.9	5.1	6.2	5.4	4.9	4.8	5.4
20	80	5.6	5.0	5.7	6.6	5.7	5.5	5.1	5.2
40	5	5.5	4.8	6.5	2.3	3.9	4.3	8.0	6.4
40	10	5.5	4.9	3.8	2.7	5.8	6.2	5.2	4.9
40	20	4.6	4.3	4.3	3.4	4.8	5.0	6.2	6.9
40	40	5.7	5.8	4.5	5.3	6.2	4.9	3.9	5.3
40	80	5.4	5.2	5.3	5.3	4.7	4.9	5.6	4.5
80	5	6.1	6.0	5.3	1.5	4.6	4.5	4.6	3.7
80	10	5.4	4.6	3.2	1.2	5.5	5.4	5.6	5.1
80	20	4.9	4.3	5.2	3.5	5.8	4.8	4.9	5.7
80	40	6.5	6.1	3.9	3.9	4.7	4.6	4.7	5.1
80	80	5.2	5.6	4.7	4.6	6.0	6.3	5.3	5.0
160	5	7.0	8.0	4.9	0.6	6.1	6.3	4.0	1.4
160	10	5.5	5.2	3.4	0.9	4.3	5.2	4.7	3.9
160	20	5.2	5.0	3.3	2.4	5.9	5.8	4.5	4.5
160	40	4.7	5.4	4.8	4.8	6.0	5.6	6.3	5.5
160	80	4.7	4.8	4.5	4.6	5.6	6.0	5.5	5.2

Table 3.2: Simulation results under H_1 for $\alpha = 5\%$ based on 1,000 i.i.d. simulation draws. The results for $\lambda = 0.5$ corresponds to data sampled from DGP 1 and DGP 2 in equal proportions, and the results for $\lambda = 0.9$ corresponds to data sampled from DGP 1 and DGP 2 with proportions 0.9 and 0.1, respectively. The test statistics $\tau_1(X)$ and $\tau_2(X)$ are defined in (3.20). Our test is computed according to (3.3) with $K = 10,000$. OPT refers to Otsu, Pesendorfer, and Takahashi (2016), whose results were copied from Tables 3 and 4 in their paper.

n	T	Mixture with $\lambda = 0.5$				Mixture with $\lambda = 0.9$			
		Our test		OPT's test		Our test		OPT's test	
		$\tau_1(X)$	$\tau_2(X)$	$\tau_1(X)$	$\tau_2(X)$	$\tau_1(X)$	$\tau_2(X)$	$\tau_1(X)$	$\tau_2(X)$
20	5	4.5	5.9	10.3	8.3	4.1	5.0	10.7	6.0
20	10	8.6	13.5	6.5	7.4	5.4	5.8	6.5	4.8
20	20	38.2	51.3	27.8	27.4	12.0	15.0	11.7	12.8
20	40	96.3	98.2	79.7	76.1	38.5	40.1	32.7	35.3
20	80	100	100	99.9	99.8	85.9	86.9	75.8	76.5
40	5	4.8	8.1	4.7	4.1	5.1	5.5	4.5	2.5
40	10	9.9	17.8	7.4	5.5	6.3	6.4	5.4	4.2
40	20	63.3	76.3	44.6	36.2	20.2	24.0	16.0	14.8
40	40	100	100	97.4	94.3	59.6	63.7	49.0	50.1
40	80	100	100	100	100	98.5	99.1	93.5	92.5
80	5	4.3	9.3	3.3	2.3	4.8	6.8	3.4	1.7
80	10	13.3	25.2	10.8	5.8	6.3	9.0	5.9	3.2
80	20	87.3	95.4	68.5	55.5	28.5	34.0	23.3	19.7
80	40	100	100	100	99.9	85.1	88.2	72.8	73.2
80	80	100	100	100	100	100	100	99.7	99.6
160	5	4.1	11.5	2.9	0.9	4.9	7.7	4.0	0.9
160	10	21.4	44.5	12.4	5.8	9.1	12.6	6.0	2.1
160	20	99.2	100	92.3	78.6	44.4	53.0	38.2	30.6
160	40	100	100	100	100	97.7	98.3	93.4	92.4
160	80	100	100	100	100	100	100	100	100

Table 3.1 reveals that our test achieves relatively good size control for all values of time periods and market sizes under consideration. The table shows the result of running 80 hypothesis tests for different data configurations that satisfy H_0 in (3.19) (four market sizes, five time periods, two test statistics, and two distributions). Across these 80 numbers, our proposed test has an average rejection rate of 5.3, with a standard deviation of 0.7, and a range of 3.9 to 8. We note that Theorem 5 implies that our test should not produce over-rejection as K becomes large, but it is silent about the possibility of under-rejection. Table 3.1 reveals that our test does not seem to suffer from under-rejection in these simulations. For Otsu, Pesendorfer, and Takahashi (2016)'s test, the average rejection rate is 5.1, with a standard deviation is 2.2, and a range of 0.6 to 13.5. We note that these extreme rejection rates occur in simulations with $T = 5$, which is reasonable for a test whose validity is proven in an asymptotic framework in which T diverges.

Table 3.2 explores the performance of these tests for data configurations that do not satisfy H_0 in (3.19) due to the multiplicity of equilibria. We begin by explaining the results of the table that are common to both hypothesis tests. First, the value of λ denotes the proportion of the n markets in the data that are in DGP 1. As λ becomes closer to either zero or one, the data are increasingly coming from a single distribution, making the departure from the H_0 harder to detect. Second, as the number of markets n grows, the inference methods gain more evidence of the presence of multiplicity, resulting in higher rejection rates. The same phenomenon occurs as the number of time periods T increases. Third, $\tau_2(X)$ is designed to be a more efficient test statistic than $\tau_1(X)$, which explains why it produces higher rejection rates across the various simulation designs. We now turn to compare rejection rates between the two tests. In most simulation designs, our test appears to have a higher or equal rejection rate than Otsu, Pesendorfer, and Takahashi (2016)'s test. The few

Table 3.3: Summary statistics for market capacity per year, measured in thousand of tons.

Sample	Average	Std. dev.	Minimum	Maximum
1980-1990	4,226.8	2,284.4	1,321.3	12,578.0
1991-1998	3,857.2	2,107.9	1,084.0	9,564.8

exceptions occur in designs with $n = 20$ and $T \in \{5, 10\}$, which include cases where their test over-rejects under the null hypothesis.

3.6 Empirical application

In this section, we revisit the application in Ryan (2012), as studied in Otsu, Pesendorfer, and Takahashi (2016, Section 5). Ryan (2012) considers a dynamic discrete game to study the welfare costs of the 1990 Amendments to the Clean Air Act on the U.S. Portland cement industry. He develops a dynamic oligopoly game based on Ericson and Pakes (1995), and estimates it using the two-stage method developed by Bajari, Benkard, and Levin (2007). This method’s first stage is to estimate optimal entry, exit, and investment decisions as a function of production capacity, and it relies on the assumptions that markets are homogeneous. Our hypothesis test can be used to investigate the validity of this assumption.

We use the same data as in Otsu, Pesendorfer, and Takahashi (2016, Section 5). For each year in 1980-1998 and 23 geographically separated U.S. markets, we observe the sum of the production capacities for all the firms in that market. Table 3.3 provides summary statistics of this aggregate production capacity before and after the 1990 Amendments, and Figure 3.1 provides the corresponding histogram.

These data represent the result of the firms’ optimal entry, exit, and investment decisions in the dynamic game estimated by Ryan (2012). We follow Otsu, Pesendorfer, and Takahashi (2016) and discretize the production capacity into 50 bins with equal intervals of 250 thousand tons each (0-250 thousand tons, 250-500 thou-

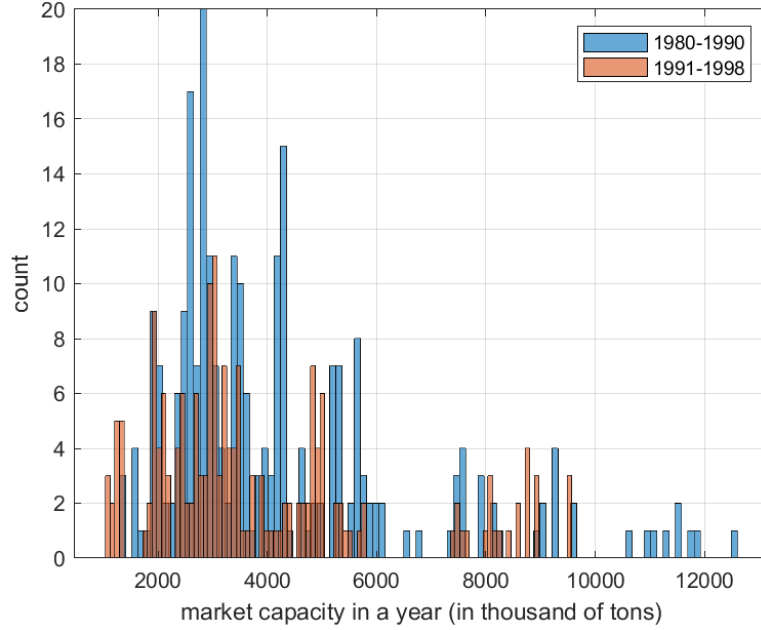


Figure 3.1: Histogram of market capacity per year, measured in thousand of tons.

sand tons, and so on). For each $i = 1, \dots, n = 23$ and year $t = 1, \dots, 19$, we use $A_{i,t} \in \mathcal{A} = \{1, \dots, 50\}$ to denote the production capacity bin. The state variable in any market is the previous period's action, i.e.,

$$S_{i,t} = A_{i,t-1}, \quad (3.21)$$

and so $S_{i,t} \in \mathcal{S} = \{1, \dots, 50\}$. We note that (3.21) implies that the state transition probabilities are homogeneous, and so H_0 in (3.1) is equivalent to the homogeneity of the conditional choice probabilities.

Following Ryan (2012) and Otsu, Pesendorfer, and Takahashi (2016), we allow the 1990 Amendments to affect the decision of the firms. We then test the homogeneity of the conditional choice probabilities for two subsets of data: before and after 1990.

Table 3.4: Results of testing (3.1) separately before and after the passing of the 1990 Amendments. The test statistics $\tau_1(X)$ and $\tau_2(X)$ are defined in (3.20). Our test is computed according to (3.3) with $K = 10,000$. OPT refers to Otsu, Pesendorfer, and Takahashi (2016), whose results were copied from Table 6 in their paper.

	Before 1990		After 1990	
	$\tau_1(X)$	$\tau_2(X)$	$\tau_1(X)$	$\tau_2(X)$
Test statistic	199.48	159.43	89.44	90.58
Our p -value	0.17	0.07	0.62	0.56
OPT's p -value	0.009	0.01	0.09	0.055

That is, we implement the following hypothesis tests:

$$H_0^{\text{before}} : \sigma_{i,t}(a|s) = \sigma(a|s) \text{ for } i = 1, \dots, 23, t = 1, \dots, 10 \quad \text{vs.} \quad H_1^{\text{before}} : H_0^{\text{before}} \text{ is false} \quad (3.22)$$

$$H_0^{\text{after}} : \sigma_{i,t}(a|s) = \sigma(a|s) \text{ for } i = 1, \dots, 23, t = 11, \dots, 19 \quad \text{vs.} \quad H_1^{\text{after}} : H_0^{\text{after}} \text{ is false} \quad (3.23)$$

We note that the two samples used to test the hypotheses in (3.22) and (3.23) have a relatively small number of time periods ($T = 10$ and $T = 9$ for (3.22) and (3.23), respectively) and markets (in both cases, $n = 23$). In this sense, this represents an ideal scenario for our proposed test, as its validity does not rely on either one of these dimensions diverging.

Table 3.4 shows the results of applying our procedure to test the hypotheses in (3.22) and (3.23). We consider both test statistics in (3.20), and we use $K = 10,000$. At a significance level of $\alpha = 5\%$, we do not reject the homogeneity of the conditional choice probabilities. Table 3.4 also shows the results of the bootstrap-based tests proposed by Otsu, Pesendorfer, and Takahashi (2016), using the same test statistics. As opposed to our results, their methods reject the hypothesis of homogeneity of the conditional choice probabilities in the sample prior to 1990.

3.7 Conclusions

This paper proposes a hypothesis test for the “homogeneity assumption” in dynamic discrete games. Our test is implemented by an MCMC algorithm. We show that our test is valid as the (user-defined) number of MCMC draws diverges, regardless of the number of markets and time periods in the data. This result contrasts with that of available methods in the literature, which require the number of time periods to diverge. We establish our validity result by showing that our proposed test is an MCMC approximation to a computationally infeasible randomization test, which happens to be finite-sample valid. Our Monte Carlo simulations confirm that our test has an excellent performance in finite samples, both in terms of size control and power.

Chapter 4

Dimension reduction in dynamic discrete choice models via index sufficiency

4.1 Introduction

A common objective of dynamic discrete choice modeling is to estimate the underlying structural equations that govern economic decisions. Most available estimators for the structural parameter of interest $\theta_0 \in \Theta \subseteq \mathbb{R}^K$ are M-estimators:

$$\hat{\theta}^* = \arg \max_{\theta} \hat{Q}(\theta). \quad (4.1)$$

For instance, the criterion function \hat{Q} may be the likelihood function (Rust 1988), a pseudo-likelihood function (Hotz and Miller 1993b; Arcidiacono and Miller 2011b) or minimum distance function (Pesendorfer and Schmidt-Dengler 2008b). Although these estimators have desirable theoretical properties, they often prove to be computationally demanding. First, computation can be intensive since forming the criterion function requires solving the model, often by fixed-point iteration. Second, the lack of global concavity in the criterion function means that local optima are a concern, which may necessitate trying many different starting values. This is especially true of models with permanent unobserved heterogeneity which are commonly estimated using the EM-algorithm.

In this paper, we propose a computationally advantageous estimator for θ_0 which is first-order asymptotically equivalent to $\hat{\theta}^*$. The computational gains of our method come from effectively reducing the dimension of the parameter space. The method is based on the observation that DDC models fall in the class of invertible index models (Ahn et al. 2018). This implies that if conditional choice probabilities (CCPs) at

different state values are nearly equal, then the structural index must be nearly equal also. This implication can be formalized as a set of equality constraints which restrict θ to belong in a strict subspace of Θ , thus simplifying the computational challenge of estimation.

In particular, we consider a class of DDC models in which the CCP—defined as the probability of taking action 1 in state z —has a multiple index structure:

$$\Pi_0(z) = p_0(\gamma_0^\top z, \delta_0^\top z)$$

where $\Pi_0(z) = \Pr(A_{it} = 1 | Z_{it} = z)$ is the CCP, p_0 is an unknown function which depends on the structural parameter of interest $\theta_0 = (\gamma_0^\top, \rho_0^\top)^\top \in \Theta \subseteq \mathbb{R}^K$, and $\delta_0 \in \mathbb{R}^{K \times J}$ (with $J \leq K$) governs the transition of the state variable. Since the state transition can be estimated without reference to the structural model, we assume a consistent estimator for δ_0 and focus on estimation of θ_0 . The structural parameter θ contains two components— γ that enters the model as a linear index and ρ which does not. For example, γ may be the payoff parameter in a linear flow payoff function and ρ may be the parameter governing permanent unobserved heterogeneity.

The dimension reduction is attained by exploiting index sufficiency. In the context of this model, index sufficiency means that if the observed CCPs at different values of z are approximately equal, then the structural index must be approximately equal. That is, for values of the state variable such that $\delta_0^\top(z_1 - z_2) = 0$,

$$\Pi_0(z_1) = \Pi_0(z_2) \iff (z_1 - z_2)^\top \gamma_0 = 0.$$

This identity forms the basis of our dimension reduction technique. First define the matrix

$$\Sigma_0 \equiv E[(Z_1 - Z_2)(Z_1 - Z_2)^\top | \Pi_0(Z_1) = \Pi_0(Z_2), \delta_0^\top(Z_1 - Z_2) = 0].$$

From the identity, it follows that the true structural parameter satisfies $\Sigma_0 \theta_0 = 0$. That is, θ_0 belongs in the nullspace of Σ_0 . As such, the dimension of the nullspace—

the nullity of Σ_0 —characterizes the extent of dimension reduction our method permits. We show that the nullity of Σ_0 is related to the number of continuous state variables and the dimension of δ_0 .

Our estimator is constructed as follows. First, given $\hat{\Sigma}$, a consistent estimator for Σ_0 , set

$$\tilde{\theta} = \arg \max_{\theta: \hat{\Sigma}\theta=0} \hat{Q}(\theta).$$

Relative to the baseline estimator $\hat{\theta}^*$, $\tilde{\theta}$ will be computationally advantageous, as it is necessary to search only within values that satisfy $\hat{\Sigma}\theta = 0$. Second, our estimator $\hat{\theta}$ is constructed by taking a number of Newton-Raphson updates from $\tilde{\theta}$. We show that $\hat{\theta}$ is first-order asymptotically equivalent to the more computationally intensive $\hat{\theta}^*$.

Section 4.2 outlines the estimator and derives its equivalency to the baseline estimator. A key issue is the nullity of Σ_0 —that is, the dimension of the subspace to which θ_0 is restricted. Subsection 4.3 derives some results on the nullity of Σ_0 . Finally, section 4.4 proposes a consistent estimator for Σ_0 and derives its rate of convergence.

Notation: We consider the Euclidean norm for all the vectors. For the matrices, we consider the induced operator norm.

4.2 Model and estimator

We assume we have a sample $\{Z_{i1}, A_{i1}, Z_{i2}, \dots, A_{iT}\}_{i=1}^N$ which are independent and identically distributed across i with $A_{it} \in \{0, 1\}$ representing an action and $Z_{it} \in \mathbb{R}^K$ the vector of state variables. The sample size N tends to infinity as T remains fixed. In what follows we omit the index (i, t) for notational simplicity. Indeed, the panel length T is unimportant in our analysis.

Define $\Pi_{0t}(z) = \Pr(A_{it} = 1 \mid Z_{it} = z)$ to be the observed conditional choice probabilities. For instance in the presence of permanent unobserved heterogeneity λ , these might be $\Pi_{0t}(z) = \int \Pr(A_{it} = 1 \mid Z_{it} = z, \lambda = v) \rho(\lambda = v \mid Z_{it} = z) dv$ where ρ is the conditional distribution of $\lambda \mid Z_{it}$. We suppose there is a structural model for the CCPs parameterized by the parameter $\theta_0 = (\gamma_0^\top, \rho_0^\top)^\top \in \Theta \subseteq \mathbb{R}^P$ and consistent estimator

$$\hat{\theta}^* = \arg \max_{\theta} \hat{Q}(\theta).$$

We also allow for a nuisance parameter δ_0 , which we assume is estimable outside the structural model. For example, in a DDC model δ_0 may govern the transition kernel—that is, the distribution of Z_{t+1} conditional upon (Z_t, A_t) . Since the transition kernel is directly observed, δ_0 can be consistently estimated without reference to the choice model.

Our first modeling assumption supposes that the binary outcome model possesses a multiple index structure:

Assumption 4.2.1. *For every pair of points, z_1 and z_2 , in the support of Z with $\delta_0^T(z_1 - z_2) = 0$,*

$$\Pi_0(z_1) = \Pi_0(z_2) \iff (z_1 - z_2)^T \gamma_0 = 0.$$

This assumption holds in a broad class of dynamic binary choice models:

Example 1 (Dynamic binary choice model with linear flow payoffs). An agent selects a sequence of actions $\{A_{i1}, A_{i2}, \dots\}$ to maximize the expected sum of discounted utilities, $E[\sum_{t=0}^{\infty} \beta^t \{u(z_t, a_t) + \epsilon_t(a_t)\} \mid a_0, z_0]$, where $\epsilon_t(a_t)$ is a state variable observed by the decision maker, but not by the econometrician. The structural parameter $\theta = (\gamma, \rho)$ with $u(z, 1) = \gamma^\top z$, and $\rho = \beta$ the discount factor. The structural model

for the CCPs is

$$\Pi(z) = \Pr \left(\gamma^\top z + \epsilon(1) + \beta \int v(z') F_Z(dz'; z, 1) \geq \epsilon(0) + \beta \int v(z') F_Z(dz'; z, 0) \mid Z = z \right) \quad (4.2)$$

where $v(z)$ is the integrated value function defined by the Bellman equation

$$v(z) = E \left[\max_{a \in \{0,1\}} \left\{ u(z, a) + \epsilon(a) + \beta \int v(z') F_Z(dz'; z, a) \right\} \right],$$

and F_Z is the distribution of the future state Z' conditional upon the current state and action (Z, A) . There is $\delta \in \mathbb{R}^{K \times J}$ with $J \leq K$ and some function G_Z such that $G_Z(z'; \delta^\top z) = F_Z(z'; z, 1) - F_Z(z'; z, 0)$, so that $\Pi(z)$ can be written

$$\Pr \left(\gamma^\top z + \beta \int v(z') G_Z(dz'; \delta^\top z) \geq \epsilon(0) - \epsilon(1) \mid Z = z \right).$$

Under the standard assumption that $\epsilon(0) - \epsilon(1)$ is independent of Z , then this model satisfies Assumption 4.2.1. Full- or pseudo-likelihood methods can be used to consistently estimate θ_0 (Aguirregabiria and Mira 2010).

Example 2 (Random effects dynamic discrete choice model). Consider the model of Example 1, but now let $u(z, 1) = \lambda + \gamma^\top z$ where λ is a random effect. If λ is assumed to have V points of support, then this is a finite-mixture binary choice model with

$$\Pr \left(\lambda + \gamma^\top z + \beta \int v(z') G_Z(dz'; \delta^\top z) \geq \epsilon(0) - \epsilon(1) \mid Z = z \right)$$

and Assumption 1 is satisfied. The structural parameter $\theta = (\gamma, \rho)$ with $\rho = (\beta, \{\Pr(\lambda = v) : v = 1, \dots, V\})$ can be consistently estimated using the expectation-maximization algorithm (Arcidiacono and Miller 2011b).

In order to exploit index sufficiency, we first define the matrix

$$\Sigma_0 \equiv E[(Z_1 - Z_2)(Z_1 - Z_2)^\top \mid \Pi_0(Z_1) = \Pi_0(Z_2), \delta_0^\top(Z_1 - Z_2) = 0].$$

Notice that under our sampling framework Z_1 and Z_2 are independent random variables whose marginal distribution is equal to that of Z . Our first result shows that Σ_0 characterizes the equality constraints that are implied by the index sufficiency:

Theorem 6. *Under Assumption 4.2.1,*

$$\Sigma_0 = E[(Z_1 - Z_2)(Z_1 - Z_2)^T \mid [\gamma_0, \delta_0]^T(Z_1 - Z_2) = 0], v$$

and

$$\Sigma_0 \gamma_0 = 0. \tag{4.3}$$

Proof. The first equation follows from Assumption 4.2.1. The second equation follows from

$$\begin{aligned} \Sigma_0 \gamma_0 &= E[(Z_1 - Z_2)(Z_1 - Z_2)^T \gamma_0 \mid [\gamma_0, \delta_0]^T(Z_1 - Z_2) = 0] \\ &= E[(Z_1 - Z_2)0 \mid [\gamma_0, \delta_0]^T(Z_1 - Z_2) = 0] = 0. \end{aligned}$$

□

Our proposed estimator involves imposing the linear equality constraints of equation 4.3. In order to do so, an estimator of Σ_0 is required. For now, we assume there is a $n^{-2-(L+1)}$ -consistent estimator $\tilde{\Sigma}$ as in the following assumption. We will provide a construction of $\tilde{\Sigma}$ in Section 4.4.

Assumption 4.2.2. *There is an integer L such that*

$$\tilde{\Sigma} - \Sigma_0 = o_p(n^{-2-(L+1)}).$$

and $\tilde{\Sigma}$ is symmetric and positive

The condition $\Sigma_0 \gamma_0 = 0$ states that the structural parameter of interest belongs in the nullspace of Σ_0 . Therefore, the effective dimension of the parameter space

is equal to the dimension of the nullspace of Σ_0 . Equivalently, by the rank-nullity theorem, the effective dimension reduction is equal to the rank of Σ_0 . For example, a nonparametric estimator of Σ_0 often yields a full rank $\tilde{\Sigma}$, so that the only γ satisfying $\tilde{\Sigma}\gamma = 0$ is the zero vector. It is therefore important that the rank of the estimate for Σ_0 is equal to the rank of Σ_0 itself.

To ensure this, we will consider a low rank approximation of $\tilde{\Sigma}$ by truncating the eigenvalues.¹ Let $\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_K$ be the eigenvalues for $\tilde{\Sigma}$, and $\hat{\nu}_1, \dots, \hat{\nu}_K$ be the corresponding eigenvectors. Define the low rank approximation

$$\hat{\Sigma} = [\hat{\nu}_1, \dots, \hat{\nu}_K]^T \text{diag} \left(\hat{\lambda}_1 \cdot 1\{\hat{\lambda}_1 > \kappa\}, \dots, \hat{\lambda}_K \cdot 1\{\hat{\lambda}_K > \kappa\} \right) [\hat{\nu}_1, \dots, \hat{\nu}_K],$$

where κ is a threshold value satisfying the following condition.

Assumption 4.2.3. $Pr(\|\tilde{\Sigma} - \Sigma_0\| \leq \kappa \leq \min\{\lambda_k : \lambda_k > 0\} - \|\tilde{\Sigma} - \Sigma_0\|) = 1 + o(1)$, where $\lambda_1 \geq \dots \geq \lambda_K$ are the eigenvalues for Σ_0 .

With a rank-consistent estimator for Σ_0 , we are now in a position to introduce the estimator for θ_0 . First consider the optimization problem

$$\text{maximize } \hat{Q} \left([\gamma^T, \rho^T]^T \right) \text{ subject to } \hat{\Sigma}\gamma = 0,$$

and denote its solution by

$$\tilde{\theta} \equiv [\tilde{\gamma}^T, \tilde{\rho}^T]^T. \tag{4.4}$$

¹Instead of this construction of $\hat{\Sigma}$, we may be able to apply a rank estimator, e.g., in Chen and Fang 2019. Since our results rely only on the convergence rate of $\tilde{\theta}$ in (4.4) and the rank is correctly estimated with probability approaching one, we conjecture that estimating the rank does not change our main result.

Second, consider L iterations of Newton-Raphson update from $\tilde{\theta}$:

$$\begin{aligned}\tilde{\theta}_1 &\equiv \tilde{\theta} - \hat{Q}^{(2)}(\tilde{\theta})^{-1} \hat{Q}^{(1)}(\tilde{\theta}) \\ \tilde{\theta}_2 &\equiv \tilde{\theta}_1 - \hat{Q}^{(2)}(\tilde{\theta}_1)^{-1} \hat{Q}^{(1)}(\tilde{\theta}_1) \\ &\vdots \\ \hat{\theta} &\equiv \tilde{\theta}_{L-1} - \hat{Q}^{(2)}(\tilde{\theta}_{L-1})^{-1} \hat{Q}^{(1)}(\tilde{\theta}_{L-1}),\end{aligned}$$

where $\hat{Q}^{(1)}(\theta)$ is the first derivative of $\hat{Q}(\theta)$ and $\hat{Q}^{(2)}(\theta)$ is the second derivative.

The final assumption imposes regularity conditions on the parameter space Θ and criterion function Q :

Assumption 4.2.4. (i) Θ is compact. (ii) θ_0 is the unique maximizer of $Q_0(\theta)$ over $\theta \in \Theta$. (iii) Q_0 is twice differentiable with the first derivative $Q_0^{(1)}$ and the second derivative $Q_0^{(2)}$. $Q_0^{(1)}$ is bounded. $Q_0^{(2)}(\theta_0)$ is nonsingular. (iv) $\hat{Q}(\theta)$ is three-times differentiable with bounded third derivatives. $\sup_{\theta \in \Theta} \left\| \hat{Q}^{(1)}(\theta) - Q_0^{(1)}(\theta) \right\| = o_p(n^{-2-(L+1)})$. $\hat{Q}^{(2)}(\theta_0) = Q_0^{(2)}(\theta_0) + o_p(1)$.

Theorem 7. Under Assumptions 4.2.1-4.2.4,

$$\hat{\theta} = \hat{\theta}^* + o_p(n^{-1/2}).$$

Theorem 7 is the main result of this paper. It says that the estimator $\hat{\theta}$ is first-order asymptotically equivalent to the computationally more intensive $\hat{\theta}^*$. The computational savings come from imposing linear constraints on θ , which restrict the structural parameter of interest to belong to a strict subset of Θ . This means that the desirable asymptotic properties of the commonly used estimator $\hat{\theta}^*$ can be attained at a lower computational cost. In the next section we consider the extent of computational savings that our method provides.

4.3 Nullity of Σ_0

In this section we consider the nullity of Σ_0 —that is, the dimension of the nullspace of Σ_0 . Under Assumption 4.2.1, the structural parameter of interest θ_0 belongs in the nullspace of Σ_0 . This means that the smaller the nullity of Σ_0 , the greater the computational advantage of imposing the equality constraints.

Theorem 8. *Suppose $Z = [Z_A^T, Z_B^T]^T$ and there is a support point $z = [z_A^T, z_B^T]^T$ of Z such that z_A is an interior point of the conditional support of Z_A given $Z_B = z_B$. Then $\text{nullity}(\Sigma_0) \leq \dim(Z_B) + \text{rank}(\text{Var}([\gamma_0, \delta_0]^T [Z_A^T, 0^T]^T))$.*

The term $\text{rank}(\text{Var}([\gamma_0, \delta_0]^T [Z_A^T, 0^T]^T))$ represents how many components in $[\Pi_0(Z), Z^T \delta_0]^T$ are continuously distributed. If the transition probability does not involve continuous state variables, we can simplify the statement as follows.

Corollary 3. *Suppose the assumptions in Theorem 8. If $\delta_0^T [Z_A^T, 0^T]^T = 0$ is discrete, then*

$$\text{nullity}(\Sigma_0) \leq \dim(Z_B) + 1.$$

In addition to the continuity, the support condition can also help to guarantee the lower bound on the rank of Σ_0 . We modify the arguments of Horowitz and Härdle 1996 to the current framework.

Theorem 9. *Suppose the same assumptions in Theorem 8. If, in addition, the conditional support of $[\gamma_0, \delta_0]^T [Z_A^T, z_B^T]^T$ given $Z_B = z_B$ is the same as the support of $[\gamma_0, \delta_0]^T Z$, then*

$$\text{nullity}(\Sigma_0) \leq \dim(Z_B) - \text{rank}(\text{Var}(Z_B)) + \text{rank}(\text{Var}([\gamma_0, \delta_0]^T [Z_A^T, 0^T]^T)).$$

Again, if the transition probability does not involve continuous state variables, we can simplify the statement as follows.

Corollary 4. *Suppose the assumptions in Theorem 9. If $\delta_0^T [Z_A^T, 0^T]^T = 0$ is discrete, then*

$$\text{nullity}(\Sigma_0) \leq \dim(Z_B) - \text{rank}(\text{Var}(Z_B)) + 1.$$

If, in addition, $\text{Var}(Z_B)$ has a full rank, then

$$\text{nullity}(\Sigma_0) \leq 1.$$

4.4 Estimation of Σ_0

Assumption 4.2.2 supposed a consistent estimator for Σ_0 , which is required for our main result (Theorem 7). In this section we propose a nonparametric estimator for Σ_0 and provide conditions for consistency. To allow for discrete state variables, in this section we write $[\Pi_0(Z_{it}), \delta_0^T Z_{it}]^T = [U_{it}^T, V_{it}^T]^T$, where U_{it} is continuously distributed and V_{it} is not. The proposed estimator will be use kernel smoothing, and therefore we require conditions on both the kernel function \mathbf{K} and the bandwidth h :

Assumption 4.4.1. *(i) \mathbf{K} is a differentiable function of \mathbb{R}^{J+1} to \mathbb{R} with the first derivative $\mathbf{K}^{(1)}$. (ii) $\mathbf{K} \left([u^T, v^T]^T \right) = 0$ and $\mathbf{K}^{(1)} \left([u^T, v^T]^T \right) = 0$ when $\|v\|$ is sufficiently large. $\mathbf{K} \left([u^T, v^T]^T \right) = 0$ and $\mathbf{K}^{(1)} \left([u^T, v^T]^T \right) = 0$ when $[u^T, v^T]^T$ is sufficiently large. (iii) $\int \mathbf{K} \left([u^T, 0^T]^T \right) du = 1$, $\int \mathbf{K} \left([u^T, 0^T]^T \right) u du = 0$, $\int \mathbf{K} \left([u^T, 0^T]^T \right) \|u\|^2 du < \infty$, and $\int \|\mathbf{K}^{(1)} \left([u^T, 0^T]^T \right)\| du < \infty$. (iv) $nh^{\dim(U_{it})/(2-2^L)} \rightarrow \infty$ and $nh^{2L+2} \rightarrow 0$.*

To construct an estimator for Σ_0 , we assume that there is a consistent estimator $(\hat{\delta}, \hat{\Pi})$ for (δ_0, Π_0) . As discussed earlier, in DDC models δ_0 may govern the state transition kernel, and is thus consistently estimable from data on the state transition. Similarly, the CCPs Π_0 are identified directly from the data.

Assumption 4.4.2. *There is a positive constant τ such that*

$$\frac{n^{-\tau}}{h} + \frac{n^{-2\tau}(n^{-\tau} + \log(n)h)^{\dim(U_{it})}}{h^{\dim(U_{it})+2}} = o(n^{-2-(L+1)})$$

and that, with probability approaching one,

$$\sup_{(i_1, t_1, i_2, t_2): i_1 \neq i_2} \left\| \hat{\zeta}_{i_1 t_1 i_2 t_2} - \zeta_{i_1 t_1 i_2 t_2} \right\| \leq n^{-\tau}, \quad (4.5)$$

where

$$\begin{aligned} \zeta_{i_1 t_1 i_2 t_2} &\equiv \left[\Pi_0(Z_{i_1 t_1}) - \Pi_0(Z_{i_2 t_2}), \delta_0^T(Z_{i_1 t_1} - Z_{i_2 t_2}) \right]^T. \\ \hat{\zeta}_{i_1 t_1 i_2 t_2} &\equiv \left[\hat{\Pi}(Z_{i_1 t_1}) - \hat{\Pi}(Z_{i_2 t_2}), \hat{\delta}^T(Z_{i_1 t_1} - Z_{i_2 t_2}) \right]^T. \end{aligned}$$

Assumption 4.4.3. (i) *Each component of $\Xi(u)$ and $f_{U_1 - U_2 | V_1 = V_2}(u)$ is twice continuously differentiable with bounded second derivatives, where $\Xi(u) = E[(Z_1 - Z_2)(Z_1 - Z_2)^T \mid U_1 - U_2 = u, V_1 = V_2] f_{U_1 - U_2 | V_1 = V_2}(u)$. (ii) $f_{U_1 - U_2}$, $f_{U_1 - U_2 | Z_1}$, $E[\|Z_2\| \mid U_1 - U_2, Z_1]$, $E[\|Z_2\|^2 \mid U_1 - U_2, Z_1]$, and $E[\|Z_1 - Z_2\|^4 \mid U_1 - U_2, V_1 - V_2]$ are bounded. (iii) $Pr(V_1 = V_2) > 0$.*

With these assumptions in hand, we define our estimator as

$$\tilde{\Sigma} \equiv \frac{\sum_{(i_1, i_2): i_1 \neq i_2} \sum_{t_1, t_2} \mathbf{K} \left(\hat{\zeta}_{i_1 t_1 i_2 t_2} / h \right) (Z_{i_1 t_1} - Z_{i_2 t_2})(Z_{i_1 t_1} - Z_{i_2 t_2})^T}{\sum_{(i_1, i_2): i_1 \neq i_2} \sum_{t_1, t_2} \mathbf{K} \left(\hat{\zeta}_{i_1 t_1 i_2 t_2} / h \right)}.$$

The main result of this section is Theorem 10. It states that $\hat{\Sigma}$ is consistent for Σ_0 .

Theorem 10. *Assumption 4.4.1-4.4.3 imply Assumption 4.2.2.*

4.5 Conclusion

In this paper we provide a method to simplify estimation of dynamic discrete choice models by exploiting index sufficiency. Index sufficiency implies a set of equality constraints which restrict the structural parameter of interest to belong in a subspace of the parameter space. We propose an estimator that imposes the equality constraints, and show it is first-order asymptotically equivalent to the unconstrained estimator. The proposed constrained estimator may be computationally advantageous due to the effective reduction in the dimension of the parameter space. Furthermore, we provide a number of results on the extent of effective dimension reduction, and show that if there is sufficient variation in the observed state variables, the parameter $\gamma \in \mathbb{R}^K$ can be restricted to a line.

Chapter 5

Conclusion

This dissertation has considered three distinct aspects of the econometrics of dynamic discrete choice models. These models are widely used in applied microeconomics as a structural approach to understanding selection. The second chapter considered a dynamic discrete choice model, and showed that a continuum of agent types can be allowed for. The third chapter provides a finite-sample valid test for an important and common modeling assumption. The fourth and final chapter suggests a computationally attractive estimator for dynamic discrete choice models, exploiting semiparametric estimation techniques.

Appendix A

Appendix for Chapter 1

A.1 Supplementary identification results

A.1.1 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 $\beta_i = (\beta_{ia} : a \in \tilde{A}) \in \mathbb{R}^b$ for $b = |A|$ enters the model through the period utility function as follows.*

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

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 5. *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 (2.1) satisfying assumptions I1, I2.1, I3.1 and I4.1. Then $(\gamma, f_{\beta|x_1})$ is point identified.*

The proof to Corollary 5 is contained in section A.2.3. It follows from the proofs of Theorems 1 and 2.

A.1.2 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:*

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

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 $(T + 1, \dots, T_1)$, 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 6. *Assume the distribution of $(x_t, a_t)_{t=1}^T$ is observed for $T = 4$, generated from agents solving the model of equation (2.1) satisfying assumptions F1, F2.1-F4.1, F6 and F7. Then $(\gamma, f_{\beta|X_1})$ is point identified.*

The argument for Corollary 6 is found in section A.2.3. 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.1.3 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 7. *Assume the distribution of $(x_t, a_t)_{t=1}^4$ is observed, generated from agents solving the model of equation (2.1) 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 A.2.3 contains the proof to corollary 7. The most substantial steps follow the proof of Theorem 1. The key difference is in showing identification of the finite parameter γ .

A.2 Identification proofs

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

A.2.1 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

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

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 $\mathcal{L}_{\mathcal{A}} = \{f : \mathcal{A} \rightarrow \mathbb{R} : \sup_{a \in \mathcal{A}} |f(a)| < \infty\}$, S_3 the support of x_3 satisfying Assumption I4(ii), and S_2 the support of x_2 satisfying Assumption I4(iii). Let $L_{3,4,2} : \mathcal{L}_{S_2} \rightarrow A \times \mathcal{L}_{S_3}$ and $L_{3,2} : \mathcal{L}_{S_2} \rightarrow A \times \mathcal{L}_{S_3}$ be defined as follows:

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

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

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

$$L_{3,\beta} : \mathcal{L}_{S_\beta} \rightarrow A \times \mathcal{L}_{S_3} \quad [L_{3,\beta}m](a_3, x_3) = \int P(a_3; x_3, b) m(b) db$$

$$D_\beta^4 : \mathcal{L}_{S_\beta} \rightarrow \mathcal{L}_{S_\beta} \quad [D_\beta^4 m](b) = P(a_4; x_4, b) m(b)$$

$$D_\beta : \mathcal{L}_{S_\beta} \rightarrow \mathcal{L}_{S_\beta} \quad [D_\beta m](b) = P(a_1; x_1, b) f_{\beta|x_1}(b; x_1) m(b)$$

$$L_{\beta,2} : \mathcal{L}_{S_2} \rightarrow \mathcal{L}_{S_\beta} \quad [L_{\beta,2}m](b) = \int P(0; x_2, b) m(x_2) dx_2$$

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 adjoint¹ 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 $(\gamma, f_{\beta|x_1})$.

¹The adjoint of a linear operator between Hilbert Spaces $L : U \rightarrow V$ is the operator $L^* : V \rightarrow U$ 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.

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

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

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

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 $S_\beta \rightarrow [0, 1]$, as

$$\mathcal{H} = \{h : S_\beta \rightarrow [0, 1] : h(b) = P(0; x, b), x \in S_t\}. \quad (\text{A.1})$$

Lemma A.2.1 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

$$\tilde{\mathcal{H}} = \{h : S_\beta \rightarrow [0, 1] : h(b) = P(0; x, b), x \in \mathbb{R}^{b+1} \times S_t^r\}, \quad (\text{A.2})$$

where S_t^r is the restriction of S_t to the final $k - (1 + |A|)$ elements of x_t . Lemma A.2.3 implies that if the functions $P(a; b, x) : \mathbb{R}^{b+1} \rightarrow [0, 1]$ are real analytic functions in the first $1 + |A|$ elements 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 A.2.1 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 A.2.1 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 A.2.2) is to show that the span of $\tilde{\mathcal{H}}$ is uniformly dense in the set

$$\mathcal{H}_1 = \{h: S_\beta \rightarrow [0, 1] : h(b) = \widetilde{\cos}(\alpha'_1 b + \alpha_0), (\alpha_1, \alpha_0) \in \mathbb{R}^{b+1}\}, \quad (\text{A.3})$$

where $\widetilde{\cos}(x) = (1 + \cos[x + 3\pi/2])(1/2)\mathbf{1}_{|x| \leq \pi/2} + \mathbf{1}_{x > \pi/2}$ 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

$$\mathcal{H}_2 = \{h: S_\beta \rightarrow [-\pi, \pi] : h(b) = \cos(\alpha'_1 b + \alpha_0), (\alpha_1, \alpha_0) \in \mathbb{R}^{b+1}\}.$$

We thus have the following containment: $\overline{\text{sp}}\tilde{\mathcal{H}} \supset \text{sp}\mathcal{H}_1 \supset \mathcal{H}_2$ where $\text{sp}\mathcal{A}$ is the linear span of \mathcal{A} , and $\overline{\text{sp}}\mathcal{A}$ 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 $\overline{\text{sp}}\mathcal{H}_2 \supset C(S_\beta)$ and it follows from the previous containment that $\overline{\text{sp}}\tilde{\mathcal{H}} \supset C(S_\beta)$ — 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

$$L_{4,3,2}L_{3,2}^{-1} = L_{3,\beta}D_\beta^4L_{3,\beta}^{-1} \quad (\text{A.4})$$

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 (A.4). 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 (A.4). To be more explicit, for any injective function R that generates another index $\tilde{\beta}$ as $\beta = R(\tilde{\beta})$, it holds that $L_{3,\beta}D_{\beta}^4L_{3,\beta}^{-1} = L_{3,\tilde{\beta}}D_{\tilde{\beta}}^4L_{3,\tilde{\beta}}^{-1}$ where

$$L_{3,\tilde{\beta}} : \mathcal{L}_{S_{\tilde{\beta}}} \rightarrow A \times \mathcal{L}_{S_3} \quad [L_{3,\tilde{\beta}}m](a, x) = \int \Pr(a_{i3} = a \mid x_{i3} = x, \tilde{\beta}_i = b)m(b)db$$

$$D_{\tilde{\beta}}^4 : \mathcal{L}_{S_{\tilde{\beta}}} \rightarrow \mathcal{L}_{S_{\tilde{\beta}}} \quad [D_{\tilde{\beta}}^4m](b) = \Pr(a_{i4} = a_4 \mid x_{i4} = x_4, \tilde{\beta}_i = b)m(b)$$

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 $(a_3, x_3) \in A \times S_3$,

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

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 $a \in A$

$$(R(b_a) \quad \tilde{\gamma}'_a) x_3 = (b_a \quad \gamma'_a) x_3$$

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

$$\frac{f_{a_2a_1x_2|x_1}(0, a_1, x_2; x_1)}{F_x(x_2; x_1, a_1)} = [L_{\beta,2}^*(P(a_1; x_1, \cdot)f_{\beta|x_1}(\cdot; x_1))] (x_2).$$

$L_{\beta,2}^*$ 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.

²This equality is shown explicitly in Hu and Schennach (2008, Supplement S.3)

□

Lemma A.2.1 (Properties of the CCP function). *Under assumptions I1, I2, I4 and I5, the sets \mathcal{H} and $\tilde{\mathcal{H}}$ defined in equations (A.1) and (A.2) 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*

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

are real analytic functions.

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

$$h(b) = P(a; x, b) = \frac{\exp(x'(b_a, \gamma_a) + \rho \int v(x'; b, \gamma) dF_x(x'|x, a))}{\sum_{\tilde{a} \in A} \exp(x'(b_{\tilde{a}}, \gamma_{\tilde{a}}) + \rho \int v(x'; b, \gamma) dF_x(x'|x, \tilde{a}))}. \quad (\text{A.5})$$

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 $|h(b)| \leq 1$ for all $b \in S_\beta$,

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

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 by 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.

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

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 $\check{\alpha}_n = \int \tilde{\alpha}_n(a, w, x') dx'$, which exists since the defining integral is supported on a bounded set.

□

Lemma A.2.2 (Approximation). *Under Assumptions I1, I2, I3 and I5 the span of $\tilde{\mathcal{H}}$ (equation (A.2)) is uniformly dense in \mathcal{H}_1 (equation (A.3)), that is for any $(\alpha_0, \alpha_1) \in \mathbb{R}^{b+1}$:*

$$\begin{aligned} \forall \epsilon > 0 \exists f \in \overline{\text{sp}} \{h: S_\beta \rightarrow [0, 1] : h(b) = P(a; (z, w), b), (a, z) \in A \times \mathbb{R}^{1+|A|}\} \\ \text{s.t. } \sup_{b \in S_\beta} |\widetilde{\text{cos}}(\alpha'_1 b + \alpha_0) - f(b)| < \epsilon, \end{aligned} \quad (\text{A.6})$$

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

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

$$P(a; x, b) = \frac{\exp(x'(b_a, \gamma_a) + \rho \int v(x'; b, \gamma)(dF_x(x'|x, a) - dF_x(x'|x, 0)))}{1 + \sum_{\tilde{a} \in \tilde{A}} \exp(x'(b_{\tilde{a}}, \gamma_{\tilde{a}}) + \rho \int v(x'; b, \gamma)(dF_x(x'|x, \tilde{a}) - dF_x(x'|x, 0)))}. \quad (\text{A.7})$$

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 (A.6).

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

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

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 A.2.1, 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

$$h(b; c) \in \begin{pmatrix} (1 - \eta, 1] & \text{if } \prod_j \mathbf{1}[\text{sgn}(\alpha_{1j})(b_j - c_j) > \epsilon] > 1 \\ [0, \eta) & \text{if } \prod_j \mathbf{1}[\text{sgn}(\alpha_{1j})(b_j - c_j) < -\epsilon] = 0 \\ [0, 1] & \text{otherwise} \end{pmatrix}. \quad (\text{A.8})$$

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

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

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 $d \rightarrow \infty$, it can be seen that there exists a d such that this function satisfies (A.8).

Second, let $c_1 < c_2 < \dots < c_n$ 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 5), so I do not repeat them here.

□

Remark 5 (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 A.2.1 applies largely directly, except for the second step of the proof of Lemma A.2.2, 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

$i = 1, 2, \dots, N$, set

$$z_i = \frac{2P}{\widetilde{\cos}^{-1}(i/N) - \widetilde{\cos}^{-1}((i-1)/N)} \alpha_1$$

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

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 $\sup_{b \in \mathcal{S}_\beta} |\widetilde{\cos}(\alpha_1' b + \alpha_0) - f(b)| < \epsilon$, first consider the i th 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 $j = 1, 2, \dots, N$, $\widetilde{\cos}(\alpha_1' b + \alpha_0) \in [(j-1)/N, j/N]$, otherwise $\widetilde{\cos}(\alpha_1' b + \alpha_0) = 1$. In the former case, the $i = 1, \dots, j-1$ components of f take values between $(1 - \bar{\epsilon}/2N)/N$ and $1/N$; the j th component takes a value between 0 and $1/N$; and the $i = j+1, \dots, N$ 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

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

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 $\bar{\epsilon}/2N < \epsilon$. \square

Lemma A.2.3 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 A.2.3. *Let F be a signed measure with compact support \mathcal{Y} and \mathcal{D} be a finite set. If*

$$\forall x \in \mathbb{R}^k, \int f(x, y) dF(y) = 0 \Rightarrow \forall y \in \mathcal{Y}, F(y) = 0 \quad (\text{A.9})$$

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 (A.9) 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 (A.9), F vanishes on \mathcal{Y} . \square

A.2.2 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

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

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

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

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:

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

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 A.2.3 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

$$\forall (a, z) \in A \times \mathbb{R}^p, \quad \int P_T(a; x, b) m(b) db = 0 \quad (\text{A.10})$$

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:

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

Fix a and consider the first-order partial derivative ($n = 1$) with respect to the i th 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 $\{P_T(a; x, b) : a \in A\}$ 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

$$\forall z \in \mathbb{R}^p, \int \tilde{P}_T(x; b)^\alpha m(b) db = 0 \tag{A.11}$$

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

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

In other words, m_z is the density of the random variable $\tilde{P}_T(x; \beta)$. Thus from equation (A.11),

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

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

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

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 t th period. With the CCPs known, the model can be solved for the finite parameters γ_t by backwards recursion. □

A.2.3 Proof of supplementary identification results

Finite horizon without terminal period

Proof of Corollary 6. For ease of notation, relabel the time index so that $t_1 = 4$. Denote $\mathcal{L}_{\mathcal{A}} = \{f : \mathcal{A} \rightarrow \mathbb{R} : \sup_{a \in \mathcal{A}} |f(a)| < \infty\}$, 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 $L_{4,2,3} : \mathcal{L}_{S_{X_3}} \rightarrow A \times \mathcal{L}_{S_{X_4}}$ and $L_{4,3} : \mathcal{L}_{S_{X_3}} \rightarrow A \times \mathcal{L}_{S_{X_4}}$ defined as

$$[L_{4,2,3}m](a_4, x_4) = \int \frac{f_{A_4 A_3 A_2 A_1 X_4 X_3 X_2 | X_1}(a_4, 0, a_2, a_1, x_4, x_3, x_2; x_1)}{F_{x_4}(x_4; x_3, 0) F_{x_3}(x_3; x_2, a_2) F_{x_2}(x_2; x_1, a_1)} m(x_3) dx_3$$

$$[L_{4,3}m](a_4, x_4) = \int \sum_{a_2 \in A} \frac{f_{A_4 A_3 A_2 A_1 X_4 X_3 X_2 | X_1}(a_4, 0, a_2, a_1, x_4, x_3, x_2; x_1)}{F_{x_4}(x_4; x_3, 0) F_{x_3}(x_3; x_2, a_2) F_{x_2}(x_2; x_1, a_1)} m(x_3) dx_3$$

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:

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

and conclude $L_{4,2,3} = L_{4,\beta}D_\beta^2D_\beta L_{\beta,3}$, 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 $L_{4,3,2}L_{3,2}^{-1} = L_{3,\beta}D_\beta^4L_{3,\beta}^{-1}$ 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 A.2.1 and A.2.2 apply for the new kernel function P_4 . It is now shown that this is the case.

The arguments of Lemma A.2.1 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 A.2.1 are redundant.

Let $x = (z, w)$ with z the first $1 + |A|$ elements of x . In the context of the finite horizon model, A.2.2 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 $\gamma_{t+} = (\gamma_s)_{s=t}^T$, 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 A.2.2 will apply and part(ii) will hold.

To show the uniform bound, given the arguments in A.2.2, 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 2007b). For $t = T - 1$,

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

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

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

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 A.2.2 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 A.2.4 (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 $x = (z, w) \in S_{x_T}$,

$$\int \Lambda(\beta_1 + \beta'_2 z + \gamma' w) df_{\beta|x_1}(b; x_1) = \int \Lambda(\beta_1 + \beta'_2 z + \tilde{\gamma}' w) d\tilde{f}_{\beta|x_1}(b; x_1).$$

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

$$\int \Lambda(\beta_1 + \beta'_2 z + (\gamma^c)' w^c) df_{\beta|x_1}(b; x_1) = \int \Lambda(\beta_1 + \beta'_2 z + (\tilde{\gamma}^c)' w^c) d\tilde{f}_{\beta|x_1}(b; x_1).$$

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 $\forall(z, w) \in S_{x_T}$,

$$\int \frac{\partial}{\partial w_{k'}} \Lambda(b_1 + b'_2 z + \gamma' w) f_{\beta|x_1}(b; x_1) db = \int \frac{\partial}{\partial w_{k'}} \Lambda(b_1 + b'_2 z + \tilde{\gamma}' w) \tilde{f}_{\beta|x_1}(b; x_1) db.$$

It is well known that the derivative of $\Lambda(x)$ is $\Lambda(x)(1 - \Lambda(x))$, so the above display is equivalent to $\forall(z, w) \in S_{x_T}$,

$$\gamma_{k'} \int [\Lambda(1-\Lambda)](b_1 + b'_2 z + \gamma' w) f_{\beta|x_1}(b; x_1) db = \tilde{\gamma}_{k'} \int [\Lambda(1-\Lambda)](b_1 + b'_2 z + \tilde{\gamma}' w) \tilde{f}_{\beta|x_1}(b; x_1) db.$$

By assumption $\gamma_k = \tilde{\gamma}_k = 1$, so we have that $\forall(z, w) \in S_{x_T}$,

$$\int [\Lambda(1-\Lambda)](b_1 + b'_2 z + \gamma' w) f_{\beta|x_1}(b; x_1) db = \int [\Lambda(1-\Lambda)](b_1 + b'_2 z + \tilde{\gamma}' w) \tilde{f}_{\beta|x_1}(b; x_1) db,$$

which is non-zero. Thus

$$\forall(z, w) \in S_{x_T}, (\gamma_k - \tilde{\gamma}_k) \int [\Lambda(1-\Lambda)](b_1 + b'_2 z + \gamma' w) f_{\beta|x_1}(b; x_1) db = 0,$$

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 $w_{k'} = 1$,

$$\begin{aligned} \Lambda(\beta_1 + \beta'_2 z + (\gamma^c)' w^c + \gamma_{k'} w_{k'}) \\ < \Lambda(\beta_1 + \beta'_2 z + (\gamma^c)' w^c + \tilde{\gamma}_{k'} w_{k'}). \end{aligned}$$

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

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

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

Infinite horizon model with random intercepts

Proof of Corollary 5. The proof follows closely the structure of the proof to Theorem 1. As in that proof, Assumptions I1 and I4.1 enable the decompositions $L_{3,4,2} = L_{3,\beta}D_{\beta}^4D_{\beta}L_{\beta,2}$ 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

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

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 $\beta \rightarrow \tilde{P}(x; \beta)$ 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.

□

Finite dependence

Proof of Corollary 7. 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 $\mathcal{L}_{\mathcal{A}} = \{f : \mathcal{A} \rightarrow \mathbb{R} : \sup_{a \in \mathcal{A}} |f(a)| < \infty\}$ and define the following operators. First define $L_{3,4,2} : \mathcal{L}_{S_2} \rightarrow \mathcal{L}_{S_3}$ and $L_{3,2} : \mathcal{L}_{S_2} \rightarrow \mathcal{L}_{S_3}$ as:

$$[L_{3,4,2}m](x_3) = \int \frac{f_{A_4 A_3 A_2 A_1 X_4 X_3 X_2 | X_1}(1, 1, 1, a_1, x_4, x_3, x_2; x_1)}{F_{x_4}(x_4; x_3, 1) F_{x_3}(x_3; x_2, 1) F_{x_1}(x_2; x_1, a_1)} m(x_2) dx_2$$

$$[L_{3,2}m](x_3) = \int \sum_{a_2 \in A} \frac{f_{A_4 A_3 A_2 A_1 X_4 X_3 X_2 | X_1}(1, 1, a_2, a_1, x_4, x_3, x_2; x_1)}{F_{x_4}(x_4; x_3, 1) F_{x_3}(x_3; x_2, 1) F_{x_1}(x_2; x_1, a_1)} m(x_2) dx_2$$

In addition, define

$$L_{3,\beta} : \mathcal{L}_{S_\beta} \rightarrow \mathcal{L}_{S_3} \quad [L_{3,\beta}m](x_3) = \int P_3(1; x_3, b) m(b) db$$

$$D_\beta^4 : \mathcal{L}_{S_\beta} \rightarrow \mathcal{L}_{S_\beta} \quad [D_\beta^4 m](b) = P_4(1; x_4, b) m(b)$$

$$D_\beta : \mathcal{L}_{S_\beta} \rightarrow \mathcal{L}_{S_\beta} \quad [D_\beta m](b) = P_1(a_1; x_1, b) f_{\beta | X_1}(b; x_1) m(b)$$

$$L_{\beta,2} : \mathcal{L}_{S_2} \rightarrow \mathcal{L}_{S_\beta} \quad [L_{\beta,2}m](b) = \int P_2(1; x_2, b) m(x_2) dx_2$$

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

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

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

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

$$g \left(x'_3(R(b), \gamma_3) + \int v(x'; R(b), \gamma) (F_{x_4}(dx'|x_3, 1) - F_{x_4}(dx'|x_3, 0)) \right)$$

where g is the logistic 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 $t_0 \leq s < t_1 - 1$ proceeds inductively as in the proof to Corollary 6. \square

A.3 Estimation appendix

A.3.1 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:

$$\hat{F}_{X_t,n}(x'; x, a) = \frac{\sum_{i=1}^N K_{X',h_{X'}}(x' - x_{t,i}) K_{X,h_X}(x - x_{t,i}) 1(a_{it} = a)}{\sum_{i=1}^N K_{X,h_X}(x - x_{t,i}) 1(a_{it} = a)} \quad (\text{A.12})$$

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

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

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 A.12 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 $\|\hat{\gamma} - \gamma\| + d_{\mathcal{M}}(\hat{f}_{\beta|x_1}, f_{\beta|x_1}) = o_p(1)$.

A.3.2 Fixed grid estimation

The choice of tuning parameters must satisfy the following condition:

Assumption E3.1. *The sieve space defined in (2.7) is such that (i) $\mathcal{M}_n \subseteq \mathcal{M}_{n+1}$ and as $n \rightarrow \infty$, (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. Assumption 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 $E[\psi(y_i, F_x, \gamma, f_{\beta|x_1})]$ 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 infimum

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

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

3.3), the condition $\log N(w, \{\psi(\cdot, F_x, \gamma, f_{\beta|x_1}) : (F_x, \gamma, f_{\beta|x_1}) \in \mathcal{N}_{f,n} \times \Gamma \times \mathcal{M}_n\}, \|\cdot\|_1) = o_p(n)$ 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

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

□

A.3.3 Estimating the support of unobserved heterogeneity

Proof of Corollary 1. From the definitions in the proof to Theorem 1 and Corollary 6, 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 . □

Appendix B

Appendix for Chapter 2

B.1 Appendix to Section 3.3

B.1.1 Implementation of Step 2 in Algorithm 3.3.1

For any $k = 2, \dots, K$, $S^{(k-1)} \in \mathcal{S}^{nT}$, and $I^{(k)}$ selected in Step 1 of Algorithm 3.3.1, Step 2 of Algorithm 3.3.1 draws $S^{(k)}$ uniformly within $R_S(I^{(k)}, S^{(k-1)})$. To implement this step, we propose a modification of the Euler Algorithm. For a description of the Euler Algorithm, see Kandel et al. (1996) and Besag and Mondal (2013). We first describe the original Euler Algorithm in Algorithm B.1.1 and then introduce our modification in Algorithm B.1.2. Throughout this section, we use 0 to represent an auxiliary value for the state variable that does not belong to the observed values of the state variable, as $0 \notin \mathcal{S} = \{1, 2, \dots, |\mathcal{S}|\}$.

Algorithm B.1.1 (Euler Algorithm). Given any integer $V \geq 2$ and any $\check{\xi} \in (\mathcal{S} \cup \{0\})^V$, $\tilde{\xi} = (\tilde{\xi}_1, \dots, \tilde{\xi}_V)$ is randomly generated as follows:

Step 1: For every $s, s' \in \mathcal{S} \cup \{0\}$, define

$$N^{(0)}(s, s') = 1\{(\check{\xi}_V, \check{\xi}_1) = (s, s')\} + \sum_{v=1}^V 1\{(\check{\xi}_v, \check{\xi}_{v+1}) = (s, s')\}.$$

Step 2: Define $\zeta_1 = \check{\xi}_V$. Set $v = 1$ and do the following.

(a) Generate ζ_{v+1} according to the following distribution.

$$P(\zeta_{v+1} = s \mid \zeta_v = s') = \frac{N^{(0)}(s, s')}{\sum_{s'' \in \mathcal{S} \cup \{0\}} N^{(0)}(s'', s')},$$

(b) If $(\mathcal{S} \cup \{0\}) \not\subset \{\zeta_1, \dots, \zeta_{v+1}\}$, then increase v by one and go back to (a).

If $(\mathcal{S} \cup \{0\}) \subset \{\zeta_1, \dots, \zeta_{v+1}\}$, then set $\bar{v} = v + 1$ and go to Step 3.

Step 3: Define $\check{\xi}_1 = \check{\xi}_1$. Also, for every $s, s' \in \mathcal{S} \cup \{0\}$, set

$$N^{(1)}(s, s') = \sum_{v=1}^V 1\{(\check{\xi}_v, \check{\xi}_{v+1}) = (s, s')\} - 1\{s' = \zeta_{(\min\{v=1, \dots, \bar{v}: \zeta_v=s\}-1)}\}.$$

Step 4: For every $v = 2, \dots, V$, generate $\check{\xi}_v$ iteratively according to $P(\check{\xi}_v = s' \mid \check{\xi}_{v-1} = s)$ which equals

$$= \begin{cases} \frac{N^{(v-1)}(s, s')}{\sum_{s'' \in \mathcal{S} \cup \{0\}} N^{(v-1)}(s, s'')} & \text{if } \sum_{s'' \in \mathcal{S} \cup \{0\}} N^{(v-1)}(s, s'') \geq 1, \\ 1\{s' = \zeta_{(\min\{v=1, \dots, \bar{v}: \zeta_v=s\}-1)}\} & \text{otherwise,} \end{cases}$$

where, for every $s, s' \in \mathcal{S} \cup \{0\}$, $N^{(v)}(s, s') = N^{(v-1)}(s, s') - 1\{(\check{\xi}_{v-1}, \check{\xi}_v) = (s, s')\}$. ■

Before we describe the central property of the Euler algorithm, we first introduce the following definition.

Definition B.1.1. For any $\check{\xi} \in (\mathcal{S} \cup \{0\})^V$, let $R_{S_0}(\check{\xi})$ denote the set of all $\check{\xi} \in (\mathcal{S} \cup \{0\})^V$ that satisfy the following conditions:

(a) $\check{\xi}_1 = \check{\xi}_1$,

(b) $\sum_{v=1}^V 1\{\check{\xi}_v = s, \check{\xi}_{v+1} = s'\} = \sum_{v=1}^V 1\{\check{\xi}_v = s, \check{\xi}_{v+1} = s'\}$ for all $s, s' \in \mathcal{S} \cup \{0\}$.

Note that $\check{\xi} \in R_{S_0}(\check{\xi})$, and so $R_{S_0}(\check{\xi}) \neq \emptyset$. Next, we give the main property of the Euler algorithm.

Lemma B.1.1. For any $\check{\xi} \in (\mathcal{S} \cup \{0\})^V$, the outcome of the Euler algorithm given $\check{\xi}$ (i.e., Algorithm B.1.1) is uniformly distributed over $R_{S_0}(\check{\xi})$ conditional on $\check{\xi}$.

Proof. See Kandel et al. (1996, Theorem 2). □

We now introduce our modification of the Euler algorithm to construct $S^{(k)}$ for any $k = 2, \dots, K$.

Algorithm B.1.2 (Generation of $S^{(k)}$). For any $k = 2, \dots, K$ and given $(X^{(1)}, \dots, X^{(k-1)}, I^{(k)})$, $S^{(k)}$ is randomly generated as follows:

Case 1: $I_1^{(k)} \neq I_2^{(k)}$.

Step 1: Set $\xi^{(k-1)} = (S_{I_1^{(k)},1}^{(k-1)}, \dots, S_{I_1^{(k)},T}^{(k-1)}, 0, S_{I_2^{(k)},1}^{(k-1)}, \dots, S_{I_2^{(k)},T}^{(k-1)}, 0)$.

Step 2: Generate $\xi^{(k)}$ as follows:

- (a) Generate a random draw of ξ using the Euler algorithm given $\xi^{(k-1)}$.
- (b) If $\xi_{T+1} = 0$, set $\xi^{(k)} = \xi$ and go to Step 3. Otherwise, return to (a).

Step 3: Given $\xi^{(k)}$, generate $S^{(k)}$ as follows:

- (a) For every $i \notin I^{(k)}$, generate $(S_{i,1}^{(k)}, \dots, S_{i,T}^{(k)})$ using the Euler algorithm given $(S_{i,1}^{(k-1)}, \dots, S_{i,T}^{(k-1)})$.
- (b) $(S_{I_1^{(k)},1}^{(k)}, \dots, S_{I_1^{(k)},T}^{(k)}) = (\xi_1^{(k)}, \dots, \xi_T^{(k)})$.
- (c) $(S_{I_2^{(k)},1}^{(k)}, \dots, S_{I_2^{(k)},T}^{(k)}) = (\xi_{T+1}^{(k)}, \dots, \xi_{2T+1}^{(k)})$.

Case 2: $I_1^{(k)} = I_2^{(k)}$. For every $i = 1, \dots, n$, generate $(S_{i,1}^{(k)}, \dots, S_{i,T}^{(k)})$ using the Euler algorithm given $(S_{i,1}^{(k-1)}, \dots, S_{i,T}^{(k-1)})$. ■

Lemma B.1.2 shows that $S^{(k)}$ generated by Algorithm B.1.2 has the desired properties.

Lemma B.1.2. *For any $k = 2, \dots, K$, $S^{(k)}$ generated by Algorithm B.1.2 satisfies the requirements of Step 2 of Algorithm 3.3.1, i.e., (3.7) holds.*

Proof. We fix $k = 2, \dots, K$, $(X^{(1)}, \dots, X^{(k-1)})$, and a generic $\check{S} \in \mathcal{S}^{nT}$ arbitrarily throughout this proof. We divide the proof in two cases.

Case 1: $I_1^{(k)} \neq I_2^{(k)}$. For $S^{(k-1)}$ and $S^{(k)}$ determined by $X^{(k-1)} = (S^{(k-1)}, A^{(k-1)})$ and $X^{(k)} = (S^{(k)}, A^{(k)})$, and for a generic $\check{S} \in \mathcal{S}^{nT}$, we set

$$\begin{aligned}\xi^{(k-1)} &= (S_{I_1^{(k-1)},1}, \dots, S_{I_1^{(k-1)},T}, 0, S_{I_2^{(k-1)},1}, \dots, S_{I_2^{(k-1)},T}, 0), \\ \xi^{(k)} &= (S_{I_1^{(k)},1}, \dots, S_{I_1^{(k)},T}, 0, S_{I_2^{(k)},1}, \dots, S_{I_2^{(k)},T}, 0), \\ \check{\xi} &= (\check{S}_{I_1^{(k)},1}, \dots, \check{S}_{I_1^{(k)},T}, 0, \check{S}_{I_2^{(k)},1}, \dots, \check{S}_{I_2^{(k)},T}, 0).\end{aligned}$$

Step 3 of Algorithm B.1.2 implies

$$P(S^{(k)} = \check{S} \mid I^{(k)}, X^{(1)}, \dots, X^{(k-1)}) = \left\{ \begin{array}{l} P(\xi^{(k)} = \check{\xi} \mid \xi^{(k-1)}) \times \\ \prod_{i \in (I^{(k)})^c} P((S_{i,1}^{(k)}, \dots, S_{i,T}^{(k)}) = (\check{S}_{i,1}, \dots, \check{S}_{i,T}) \mid S_{i,1}^{(k-1)}, \dots, S_{i,T}^{(k-1)}) \end{array} \right\}, \quad (\text{A-1})$$

Lemma B.1.1 implies that $P((S_{i,1}^{(k)}, \dots, S_{i,T}^{(k)}) = (\check{S}_{i,1}, \dots, \check{S}_{i,T}) \mid S_{i,1}^{(k-1)}, \dots, S_{i,T}^{(k-1)})$ is equal to

$$\frac{1\{(\check{S}_{i,1}, \dots, \check{S}_{i,T}) \in R_{S0}(S_{i,1}^{(k-1)}, \dots, S_{i,T}^{(k-1)})\}}{|R_{S0}(S_{i,1}^{(k-1)}, \dots, S_{i,T}^{(k-1)})|} \quad (\text{A-2})$$

for every $i \in (I^{(k)})^c$. In turn, Lemma B.1.3 implies that

$$P(\xi^{(k)} = \check{\xi} \mid \xi^{(k-1)}) = \frac{1\{\check{\xi} \in R_{S0}(\xi^{(k-1)}) : \check{\xi}_{T+1} = 0\}}{| \{\check{\xi} \in R_{S0}(\xi^{(k-1)}) : \check{\xi}_{T+1} = 0\} |}. \quad (\text{A-3})$$

By combining (A-1), (A-2), and (A-3),

$$P(S^{(k)} = \check{S} \mid I^{(k)}, X^{(1)}, \dots, X^{(k-1)}) = \frac{1\{\check{\xi} \in R_{S0}(\xi^{(k-1)}) : \check{\xi}_{T+1} = 0\} \times \prod_{i \in (I^{(k)})^c} 1\{(\check{S}_{i,1}, \dots, \check{S}_{i,T}) \in R_{S0}(S_{i,1}^{(k-1)}, \dots, S_{i,T}^{(k-1)})\}}{| \{\check{\xi} \in R_{S0}(\xi^{(k-1)}) : \check{\xi}_{T+1} = 0\} | \times \prod_{i \in (I^{(k)})^c} |R_{S0}(S_{i,1}^{(k-1)}, \dots, S_{i,T}^{(k-1)})|}. \quad (\text{A-4})$$

To complete the proof, it suffices to show that the right-hand side of (A-4) is equal to the right-hand side of (3.7). To this end, it suffices to show that

$$\begin{aligned} & 1\{\check{\xi} \in R_{S_0}(\xi^{(k-1)}) : \check{\xi}_{T+1} = 0\} \times \prod_{i \in (I^{(k)})^c} 1\{(\check{S}_{i,1}, \dots, \check{S}_{i,T}) \in R_{S_0}(S_{i,1}^{(k-1)}, \dots, S_{i,T}^{(k-1)})\} \\ &= 1\{\check{S} \in R_S(I^{(k)}, S^{(k-1)})\} \end{aligned} \quad (\text{A-5})$$

and

$$| \{ \check{\xi} \in R_{S_0}(\xi^{(k-1)}) : \check{\xi}_{T+1} = 0 \} | \times \prod_{i \in (I^{(k)})^c} |R_{S_0}(S_{i,1}^{(k-1)}, \dots, S_{i,T}^{(k-1)})| = |R_S(I^{(k)}, S^{(k-1)})|. \quad (\text{A-6})$$

To show (A-5), consider the following derivation.

$$\begin{aligned} & 1\{\check{\xi} \in R_{S_0}(\xi^{(k-1)}) : \check{\xi}_{T+1} = 0\} \times \prod_{i \in (I^{(k)})^c} 1\{(\check{S}_{i,1}, \dots, \check{S}_{i,T}) \in R_{S_0}(S_{i,1}^{(k-1)}, \dots, S_{i,T}^{(k-1)})\} \\ &= \left\{ \begin{array}{l} 1\{(\check{S}_{I_1^{(k)},1}, \dots, \check{S}_{I_1^{(k)},T}, 0, \check{S}_{I_2^{(k)},1}, \dots, \check{S}_{I_2^{(k)},T}, 0) \in R_{S_0}(\xi^{(k-1)})\} \\ \times \prod_{i \in (I^{(k)})^c} 1\{(\check{S}_{i,1}, \dots, \check{S}_{i,T}) \in R_{S_0}(S_{i,1}^{(k-1)}, \dots, S_{i,T}^{(k-1)})\} \end{array} \right\} \\ &\stackrel{(1)}{=} 1 \left\{ \begin{array}{l} \check{S}_{i,1} = S_{i,1}^{(k-1)} \text{ for all } i = 1, \dots, n, \\ \sum_{i \in I^{(k)}} \sum_{t=1}^{T-1} 1\{\check{S}_{i,t} = s, \check{S}_{i,t+1} = s'\} = \sum_{i \in I^{(k)}} \sum_{t=1}^{T-1} 1\{S_{i,t}^{(k-1)} = s, S_{i,t+1}^{(k-1)} = s'\} \forall s, s' \in \mathcal{S}, \\ \sum_{i \in I^{(k)}} 1\{\check{S}_{i,T} = s\} = \sum_{i \in I^{(k)}} 1\{S_{i,T}^{(k-1)} = s\} \text{ for all } s \in \mathcal{S}, \\ \sum_{t=1}^{T-1} 1\{\check{S}_{i,t} = s, \check{S}_{i,t+1} = s'\} = \sum_{t=1}^{T-1} 1\{S_{i,t}^{(k-1)} = s, S_{i,t+1}^{(k-1)} = s'\} \text{ for all } s, s' \in \mathcal{S}, i \in (I^{(k)})^c \end{array} \right\} \\ &\stackrel{(2)}{=} 1 \left\{ \begin{array}{l} \check{S}_{i,1} = S_{i,1}^{(k-1)} \text{ for all } i = 1, \dots, n, \\ \sum_{i \in I^{(k)}} \sum_{t=1}^{T-1} 1\{\check{S}_{i,t} = s, \check{S}_{i,t+1} = s'\} = \sum_{i \in I^{(k)}} \sum_{t=1}^{T-1} 1\{S_{i,t}^{(k-1)} = s, S_{i,t+1}^{(k-1)} = s'\} \forall s, s' \in \mathcal{S}, \\ \sum_{t=1}^{T-1} 1\{\check{S}_{i,t} = s, \check{S}_{i,t+1} = s'\} = \sum_{t=1}^{T-1} 1\{S_{i,t}^{(k-1)} = s, S_{i,t+1}^{(k-1)} = s'\} \text{ for all } s, s' \in \mathcal{S}, i \in (I^{(k)})^c \end{array} \right\} \\ &\stackrel{(3)}{=} 1\{\check{S} \in R_S(I^{(k)}, S^{(k-1)})\}, \end{aligned}$$

as desired, where (1) follows from $I_1^{(k)} \neq I_2^{(k)}$ and applying Definition B.1.1, (2) follows from Lemma B.1.4, and (3) follows from Definition 3.3.1. To show (A-6), consider

the following argument.

$$\begin{aligned}
1 &= \sum_{\check{S} \in \mathcal{S}} P(S^{(k)} = \check{S} \mid I^{(k)}, X^{(1)}, \dots, X^{(k-1)}) \\
&\stackrel{(1)}{=} \frac{\sum_{\check{S} \in \mathcal{S}} 1\{\check{S} \in R_S(I^{(k)}, S^{(k-1)})\}}{|\{\tilde{\xi} \in R_{S_0}(\xi^{(k-1)}) : \tilde{\xi}_{T+1} = 0\}| \times \prod_{i \in (I^{(k)})^c} |R_{S_0}(S_{i,1}^{(k-1)}, \dots, S_{i,T}^{(k-1)})|} \\
&= \frac{|R_S(I^{(k)}, S^{(k-1)})|}{|\{\tilde{\xi} \in R_{S_0}(\xi^{(k-1)}) : \tilde{\xi}_{T+1} = 0\}| \times \prod_{i \in (I^{(k)})^c} |R_{S_0}(S_{i,1}^{(k-1)}, \dots, S_{i,T}^{(k-1)})|},
\end{aligned}$$

where (1) follows from combining (A-4) and (A-5). From here, (A-6) follows.

Case 2: $I_1^{(k)} = I_2^{(k)}$. Algorithm B.1.2 implies $P(S^{(k)} = \check{S} \mid I^{(k)}, X^{(1)}, \dots, X^{(k-1)}) =$ is equal to

$$\prod_{i=1}^n P((S_{i,1}^{(k)}, \dots, S_{i,T}^{(k)}) = (\check{S}_{i,1}, \dots, \check{S}_{i,T}) \mid S_{i,1}^{(k-1)}, \dots, S_{i,T}^{(k-1)}). \quad (\text{A-7})$$

Lemma B.1.1 implies that for every $i = 1, \dots, n$, $P((S_{i,1}^{(k)}, \dots, S_{i,T}^{(k)}) = (\check{S}_{i,1}, \dots, \check{S}_{i,T}) \mid S_{i,1}^{(k-1)}, \dots, S_{i,T}^{(k-1)})$ is equal to

$$= \frac{1\{(\check{S}_{i,1}, \dots, \check{S}_{i,T}) \in R_{S_0}(S_{i,1}^{(k-1)}, \dots, S_{i,T}^{(k-1)})\}}{|R_{S_0}(S_{i,1}^{(k-1)}, \dots, S_{i,T}^{(k-1)})|}. \quad (\text{A-8})$$

By combining (A-7) and (A-8)

$$\begin{aligned}
P(S^{(k)} = \check{S} \mid I^{(k)}, X^{(1)}, \dots, X^{(k-1)}) &= \prod_{i=1}^n \frac{1\{(\check{S}_{i,1}, \dots, \check{S}_{i,T}) \in R_{S_0}(S_{i,1}^{(k-1)}, \dots, S_{i,T}^{(k-1)})\}}{|R_{S_0}(S_{i,1}^{(k-1)}, \dots, S_{i,T}^{(k-1)})|} \\
&= \frac{\prod_{i=1}^n 1\{(\check{S}_{i,1}, \dots, \check{S}_{i,T}) \in R_{S_0}(S_{i,1}^{(k-1)}, \dots, S_{i,T}^{(k-1)})\}}{\prod_{i=1}^n |R_{S_0}(S_{i,1}^{(k-1)}, \dots, S_{i,T}^{(k-1)})|}.
\end{aligned} \quad (\text{A-9})$$

To complete the proof, it suffices to show that the right-hand side of (A-9) is equal

to the right-hand side of (3.7). To this end, it suffices to show that

$$\prod_{i=1}^n \mathbb{1}\{\check{S}_{i,1}, \dots, \check{S}_{i,T} \in R_{S_0}(S_{i,1}^{(k-1)}, \dots, S_{i,T}^{(k-1)})\} = \mathbb{1}\{\check{S} \in R_S(I^{(k)}, S^{(k-1)})\} \quad (\text{A-10})$$

$$\prod_{i=1}^n |R_{S_0}(S_{i,1}^{(k-1)}, \dots, S_{i,T}^{(k-1)})| = |R_S(I^{(k)}, S^{(k-1)})|. \quad (\text{A-11})$$

To show (A-10), consider the following derivation.

$$\begin{aligned} & \prod_{i=1}^n \mathbb{1}\{\check{S}_{i,1}, \dots, \check{S}_{i,T} \in R_{S_0}(S_{i,1}^{(k-1)}, \dots, S_{i,T}^{(k-1)})\} \\ & \stackrel{(1)}{=} \mathbb{1}\left\{ \begin{array}{l} \check{S}_{i,1} = S_{i,1}^{(k-1)} \text{ for all } i = 1, \dots, n, \\ \sum_{t=1}^T \mathbb{1}\{\check{S}_{i,t} = s, \check{S}_{i,t+1} = s'\} = \sum_{t=1}^T \mathbb{1}\{S_{i,t}^{(k-1)} = s, S_{i,t+1}^{(k-1)} = s'\} \text{ for all } s, s' \in \mathcal{S} \text{ and } i = 1, \dots, n \end{array} \right\} \\ & \stackrel{(2)}{=} \mathbb{1}\{\check{S} \in R_S(I^{(k)}, S^{(k-1)})\}, \end{aligned}$$

as desired, where (1) follows from $I_1^{(k)} = I_2^{(k)}$ and applying Definition B.1.1 for each $i = 1, \dots, n$, and (2) follows from Definition 3.3.1. Finally, (A-11) can be shown by using an argument that is analogous to the one used to prove (A-6). We omit this for the sake of brevity. \square

Lemma B.1.3. *For any $k = 2, \dots, K$, if $\xi^{(k)}$ is generated by Algorithm B.1.2, then $\xi^{(k)}$ is uniformly distributed over the set $\{\xi \in R_{S_0}(\xi^{(k-1)}) : \xi_{T+1} = 0\}$ conditional on $(I^{(k)}, X^{(1)}, \dots, X^{(k-1)})$.*

Proof. By Lemma B.1.1, $\tilde{\xi}$ in Step 2(a) of Algorithm B.1.2 follows the uniform distribution on $R_{S_0}(\xi^{(k-1)})$, conditional on $(I^{(k)}, X^{(1)}, \dots, X^{(k-1)})$. Steps 2(b) of Algorithm B.1.2 truncates the variable to the set $\{\xi \in R_{S_0}(\xi^{(k-1)}) : \xi_{T+1} = 0\}$. The desired result then follows from the fact that a truncated version of a discrete uniform distribution is uniformly distributed on the truncated set. \square

Lemma B.1.4. *For any $I \in \mathcal{I}$, if $\check{S}, \tilde{S} \in \mathcal{S}^{nT}$ satisfy the following conditions:*

$$(a) \quad \check{S}_{i,1} = \tilde{S}_{i,1} \text{ for all } i \in I,$$

$$(b) \sum_{i \in I} \sum_{t=1}^{T-1} 1\{\tilde{S}_{i,t} = s, \tilde{S}_{i,t+1} = s'\} = \sum_{i \in I} \sum_{t=1}^{T-1} 1\{\check{S}_{i,t} = s, \check{S}_{i,t+1} = s'\} \text{ for all } s, s' \in \mathcal{S},$$

then, $\sum_{i \in I} 1\{\tilde{S}_{i,T} = s\} = \sum_{i \in I} 1\{\check{S}_{i,T} = s\}$ for all $s \in \mathcal{S}$.

Proof. For every $i \in I$ and $s \in \mathcal{S}$, note that

$$\begin{aligned} 1\{\check{S}_{i,T} = s\} &= \sum_{t=1}^{T-1} 1\{\check{S}_{i,t+1} = s\} - \sum_{t=1}^{T-1} 1\{\check{S}_{i,t} = s\} + 1\{\check{S}_{i,1} = s\} \\ &= \sum_{\bar{s} \in \mathcal{S}} \sum_{t=1}^{T-1} 1\{\check{S}_{i,t} = \bar{s}, \check{S}_{i,t+1} = s\} - \sum_{\bar{s} \in \mathcal{S}} \sum_{t=1}^{T-1} 1\{\check{S}_{i,t} = s, \check{S}_{i,t+1} = \bar{s}\} + 1\{\check{S}_{i,1} = s\}. \end{aligned} \tag{A-12}$$

By the same argument applied to $\tilde{S} \in R_S(I, \check{S})$, we have that for every $i \in I$ and $s \in \mathcal{S}$,

$$1\{\tilde{S}_{i,T} = s\} = \sum_{\bar{s} \in \mathcal{S}} \sum_{t=1}^{T-1} 1\{\tilde{S}_{i,t} = \bar{s}, \tilde{S}_{i,t+1} = s\} - \sum_{\bar{s} \in \mathcal{S}} \sum_{t=1}^{T-1} 1\{\tilde{S}_{i,t} = s, \tilde{S}_{i,t+1} = \bar{s}'\} + 1\{\tilde{S}_{i,1} = s\}. \tag{A-13}$$

To show this lemma, fix $s \in \mathcal{S}$ arbitrarily and consider the following argument.

$$\begin{aligned} \sum_{i \in I} 1\{\check{S}_{i,T} = s\} &\stackrel{(1)}{=} \sum_{\bar{s} \in \mathcal{S}} \sum_{i \in I} \sum_{t=1}^{T-1} 1\{\check{S}_{i,t} = \bar{s}, \check{S}_{i,t+1} = s\} - \\ &\quad \sum_{\bar{s} \in \mathcal{S}} \sum_{i \in I} \sum_{t=1}^{T-1} 1\{\check{S}_{i,t} = s, \check{S}_{i,t+1} = \bar{s}\} + \sum_{i \in I} 1\{\check{S}_{i,1} = s\} \\ &\stackrel{(2)}{=} \sum_{\bar{s} \in \mathcal{S}} \sum_{i \in I} \sum_{t=1}^{T-1} 1\{\tilde{S}_{i,t} = \bar{s}, \tilde{S}_{i,t+1} = s\} - \\ &\quad \sum_{\bar{s} \in \mathcal{S}} \sum_{i \in I} \sum_{t=1}^{T-1} 1\{\tilde{S}_{i,t} = s, \tilde{S}_{i,t+1} = \bar{s}\} + \sum_{i \in I} 1\{\tilde{S}_{i,1} = s\} \\ &\stackrel{(3)}{=} \sum_{i \in I} 1\{\tilde{S}_{i,T} = s\}, \end{aligned}$$

where (1) holds by (A-12), (2) holds by conditions (a)-(b), and (3) holds by (A-13). \square

B.1.2 Implementation of Step 3 in Algorithm 3.3.1

For any $k = 2, \dots, K$, $X^{(k-1)} \in \mathcal{X}$, and $S^{(k)} \in \mathcal{S}^{nT}$, Step 3 of Algorithm 3.3.1 draws $A^{(k)}$ uniformly within $R_A(S^{(k)}, X^{(k-1)})$. This can be implemented by the following algorithm.

Algorithm B.1.3 (Generation of $A^{(k)}$). For any $k = 2, \dots, K$ and given $(X^{(1)}, \dots, X^{(k-1)}, I^k, S^{(k)})$, $A^{(k)}$ is randomly generated as follows

Step 1: For every $(s, s') \in \mathcal{S} \times \mathcal{S}$, define

$$\text{Index}^{(k-1)}(s, s') = \{(i, t) \in \{1, \dots, n\} \times \{1, \dots, T-1\} : (S_{i,t}^{(k-1)}, S_{i,t+1}^{(k-1)}) = (s, s')\}$$

$$\text{Index}^{(k)}(s, s') = \{(i, t) \in \{1, \dots, n\} \times \{1, \dots, T-1\} : (S_{i,t}^{(k)}, S_{i,t+1}^{(k)}) = (s, s')\}$$

$$\text{Index}^{(k-1)}(s) = \{(i, T) : i \in \{1, \dots, n\}, S_{i,T}^{(k-1)} = s\}$$

$$\text{Index}^{(k)}(s) = \{(i, T) : i \in \{1, \dots, n\}, S_{i,T}^{(k)} = s\}.$$

Step 2: For every $(s, s') \in \mathcal{S} \times \mathcal{S}$, we generate $(A_{i,t}^{(k)} : (i, t) \in \text{Index}^{(k)}(s, s'))$ by uniformly sampling from $(A_{i,t}^{(k-1)} : (i, t) \in \text{Index}^{(k-1)}(s, s'))$ without replacement, i.e., a uniformly chosen permutation.

Step 3: For every $s \in \mathcal{S}$, we construct $(A_{i,T}^{(k)} : (i, T) \in \text{Index}^{(k)}(s))$ by uniformly sampling from the discrete set $(A_{i,T}^{(k-1)} : (i, T) \in \text{Index}^{(k-1)}(s))$ without replacement, i.e., a uniformly chosen permutation. \blacksquare

Lemma B.1.5 shows that $A^{(k)}$ generated by Algorithm B.1.3 has the desired properties.

Lemma B.1.5. *For any $k = 2, \dots, K$, the outcome, $A^{(k)}$, of Algorithm B.1.3 satisfies the requirements of Step 3 of Algorithm 3.3.1, i.e., (3.8) holds.*

Proof. This follows from noting that any element of $R_A(S^{(k)}, X^{(k-1)})$ corresponds to a restricted set of permutations of the action data, and Algorithm B.1.3 chooses an element uniformly within this set. \square

B.1.3 Proof of Theorem 5

By (3.3), (3.5) is equivalent to $\liminf_{K \rightarrow \infty} (\alpha - P(\hat{p}_K \leq \alpha)) \geq 0$. In this proof, we are going to show a stronger statement (cf. Lehmann and Romano 2005, Eq. (15.6)):

$$\liminf_{K \rightarrow \infty} \inf_{u \in [0,1]} (u - P(\hat{p}_K \leq u)) \geq 0.$$

Fix $\varepsilon > 0$ and $u \in [0, 1]$ arbitrarily. The rest of the proof is going to show

$$u - P(\hat{p}_K \leq u) \geq -2\varepsilon$$

for sufficiently large K . For any positive integer K , let

$$\mathcal{E}_K \equiv \left\{ \sup_{t \in \mathbb{R}} \left| \frac{1}{K} \sum_{k=1}^K 1\{\tau(X^{(k)}) \geq t\} - \frac{1}{|\mathbf{G}|} \sum_{g \in \mathbf{G}} 1\{\tau(g(X)) \geq t\} \right| > \varepsilon \right\}.$$

By Lemma 3.4.4, for sufficiently large K ,

$$P(\mathcal{E}_K) \leq \varepsilon. \tag{A-14}$$

For any positive integer K , consider the following derivation:

$$\begin{aligned}
P(\hat{p}_K \leq u) &= P\left(\frac{1}{K} \sum_{k=1}^K 1\{\tau(X^{(k)}) \geq \tau(X)\} \leq u\right) \\
&= P\left(\left\{\frac{1}{K} \sum_{k=1}^K 1\{\tau(X^{(k)}) \geq \tau(X)\} \leq u\right\} \cup \mathcal{E}_K^c\right) \tag{A-15}
\end{aligned}$$

$$\begin{aligned}
&+ P\left(\left\{\frac{1}{K} \sum_{k=1}^K 1\{\tau(X^{(k)}) \geq \tau(X)\} \leq u\right\} \cup \mathcal{E}_K\right) \\
&\leq P\left(\frac{1}{|\mathbf{G}|} \sum_{g \in \mathbf{G}} 1\{\tau(g(X)) \geq \tau(X)\} \leq u + \varepsilon\right) + P(\mathcal{E}_K) \\
&\stackrel{(1)}{\leq} u + \varepsilon + P(\mathcal{E}_K), \tag{A-16}
\end{aligned}$$

where (1) holds by Lemma 3.4.3. By (A-16) and (A-14), we conclude that, for sufficiently large K , $P(\hat{p}_K \leq u) \leq u + 2\varepsilon$ or, equivalently, $u - P(\hat{p}_K \leq u) \geq -2\varepsilon$, as desired. \blacksquare

B.2 Appendix to Section 3.4

B.2.1 Proof of lemmas

Proof of Lemma 3.4.1. Note that

$$P(S = \tilde{S}) = \prod_{i=1}^n \left(m_i(\tilde{S}_{i,1}) \prod_{t=1}^{T-1} \left(\sum_{a \in \mathcal{A}} g(\tilde{S}_{i,t+1}|a, \tilde{S}_{i,t}) \sigma(a|\tilde{S}_{i,t}) \right) \right). \tag{B-17}$$

This equation follows from the following derivation

$$\begin{aligned}
& P(S = \tilde{S}) \\
& \stackrel{(1)}{=} \prod_{i=1}^n \left(P(S_{i,1} = \tilde{S}_{i,1}) \prod_{t=1}^{T-1} P(S_{i,t} = \tilde{S}_{i,t} | (S_{i,1}, \dots, S_{i,t-1}) = (\tilde{S}_{i,1}, \dots, \tilde{S}_{i,t-1})) \right) \\
& \stackrel{(2)}{=} \prod_{i=1}^n \left(P(S_{i,1} = \tilde{S}_{i,1}) \prod_{t=1}^{T-1} P(S_{i,t} = \tilde{S}_{i,t} | S_{i,t-1} = \tilde{S}_{i,t-1}) \right) \\
& = \prod_{i=1}^n \left(P(S_{i,1} = \tilde{S}_{i,1}) \prod_{t=1}^{T-1} \left(\sum_{a \in \mathcal{A}} \left(P(S_{i,t} = \tilde{S}_{i,t} | A_{i,t-1} = a, S_{i,t-1} = \tilde{S}_{i,t-1}) \right) \times P(A_{i,t-1} = a | S_{i,t-1} = \tilde{S}_{i,t-1}) \right) \right) \\
& \stackrel{(3)}{=} \prod_{i=1}^n \left(m_i(\tilde{S}_{i,1}) \prod_{t=1}^{T-1} \left(\sum_{a \in \mathcal{A}} g(\tilde{S}_{i,t+1} | a, \tilde{S}_{i,t}) \sigma(a | \tilde{S}_{i,t}) \right) \right),
\end{aligned}$$

where (1) holds by Assumption 3.2.1(a), (2) holds by Lemma B.2.2, and (3) holds under H_0 in (3.1).

To conclude the proof, it suffices to show (3.11) and (3.12). To this end, consider the following derivation.

$$\begin{aligned}
P(X = \tilde{X}) & \stackrel{(1)}{=} \prod_{i=1}^n \left(m_i(\tilde{S}_{i,1}) \sigma(\tilde{A}_{i,T} | \tilde{S}_{i,T}) \prod_{t=1}^{T-1} \left(\sigma(\tilde{A}_{i,t} | \tilde{S}_{i,t}) g(\tilde{S}_{i,t+1} | \tilde{S}_{i,t}, \tilde{A}_{i,t}) \right) \right) \\
& \stackrel{(2)}{=} P(S = \tilde{S}) \left(\sigma(\tilde{A}_{i,T} | \tilde{S}_{i,T}) \left(\prod_{t=1}^{T-1} \frac{\sigma(\tilde{A}_{i,t} | \tilde{S}_{i,t}) g(\tilde{S}_{i,t+1} | \tilde{S}_{i,t}, \tilde{A}_{i,t})}{\sum_{a \in \mathcal{A}} g(\tilde{S}_{i,t+1} | a, \tilde{S}_{i,t}) \sigma(a | \tilde{S}_{i,t})} \right) \right),
\end{aligned} \tag{B-18}$$

where (1) holds by (3.2), which is shown in Lemma B.2.1, and (2) holds by (B-17).

By combining (3.10) and (B-18), we conclude that

$$P(A = \tilde{A} | S = \tilde{S}) = \prod_{i=1}^n \left(\sigma(\tilde{A}_{i,T} | \tilde{S}_{i,T}) \left(\prod_{t=1}^{T-1} \frac{\sigma(\tilde{A}_{i,t} | \tilde{S}_{i,t}) g(\tilde{S}_{i,t+1} | \tilde{S}_{i,t}, \tilde{A}_{i,t})}{\sum_{a \in \mathcal{A}} g(\tilde{S}_{i,t+1} | a, \tilde{S}_{i,t}) \sigma(a | \tilde{S}_{i,t})} \right) \right).$$

By re-expressing this equation in terms of counts of $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, (3.11) follows. Moreover, (3.12) follows from re-expressing (B-17) in terms of individual counts of each $(s, s') \in \mathcal{S} \times \mathcal{S}$. \square

Proof of Lemma 3.4.2. We first show that \mathbf{G} is a collection of transformations from \mathcal{X} onto itself. Consider any $g \in \mathbf{G}$. By definition, g is the composition of a finite number of transformations in $\bigcup_{I \in \mathcal{I}} \mathbf{G}(I)$, i.e., $g = g^{(K)} \circ \dots \circ g^{(1)}$ with $(g^{(1)}, \dots, g^{(K)}) \in \mathbf{G}(I^{(1)}) \times \dots \times \mathbf{G}(I^{(K)})$ with $I^{(j)} \in \mathcal{I}$ for $j = 1, \dots, K$. By Lemma B.2.3, $g^{(j)} \in \mathbf{G}(I^{(j)})$ are onto transformations from \mathcal{X} to itself. From this, we can conclude that $g = g^{(K)} \circ \dots \circ g^{(1)}$ is an onto transformation from \mathcal{X} to itself, as desired.

Second, we show that \mathbf{G} is a group. To this end, it suffices to verify conditions (i)-(iv) in Lehmann and Romano (2005, Section A.1). To verify condition (i), consider arbitrary $g_1, g_2 \in \mathbf{G}$. By definition, this implies g_1 and g_2 are compositions of a finite number of transformations in $\bigcup_{I \in \mathcal{I}} \mathbf{G}(I)$. Then, $g_2 \circ g_1$ is a composition of a finite number of elements in $\bigcup_{I \in \mathcal{I}} \mathbf{G}(I)$, and so $g_2 \circ g_1 \in \mathbf{G}$. Condition (ii) follows from the argument in Lehmann and Romano (2005, page 693). Condition (iii) follows from the fact that $\mathbf{G}(I)$ is a group for any for any $I \in \mathcal{I}$ (shown in Lemma B.2.3), and so it includes the identity transformation. To verify condition (iv), consider the following argument for any arbitrary $g \in \mathbf{G}$. By definition, g is the composition of a finite number of transformations in $\bigcup_{I \in \mathcal{I}} \mathbf{G}(I)$, i.e., $g = g^{(K)} \circ \dots \circ g^{(1)}$ with $(g^{(1)}, \dots, g^{(K)}) \in \mathbf{G}(I^{(1)}) \times \dots \times \mathbf{G}(I^{(K)})$ with $I^{(j)} \in \mathcal{I}$ for $j = 1, \dots, K$. By Lemma B.2.3, $\mathbf{G}(I^{(j)})$ is a group for each $j = 1, \dots, K$. From this, we can conclude that $\exists (g^{(j)})^{-1} \in \mathbf{G}(I^{(j)})$ for each $j = 1, \dots, K$. Since $g \circ \tilde{g}$ and $\tilde{g} \circ g$ are equal to the identity transformation, $\tilde{g} = g^{-1}$. Finally, note that $g^{-1} = (g^{(1)})^{-1} \circ \dots \circ (g^{(K)})^{-1}$ is the compositions of a finite number of transformations in $\bigcup_{I \in \mathcal{I}} \mathbf{G}(I)$ and so $g^{-1} \in \mathbf{G}$, as desired.

To complete the proof, it suffices to show that, for any $\tilde{X} \in \mathcal{X}$ and $g \in \mathbf{G}$, \tilde{X} and $g\tilde{X}$ have the same sufficient statistics in (3.13). g is the composition of a finite number of transformations in $\bigcup_{I \in \mathcal{I}} \mathbf{G}(I)$, i.e., $g = g^{(K)} \circ \dots \circ g^{(1)}$ with $(g^{(1)}, \dots, g^{(K)}) \in \mathbf{G}(I^{(1)}) \times \dots \times \mathbf{G}(I^{(K)})$ with $I^{(j)} \in \mathcal{I}$ for $j = 1, \dots, K$. Therefore, $g\tilde{X} = g^{(K)} \circ \dots \circ$

$g^{(1)}\tilde{X}$. For each $j = 1, \dots, K$, Lemma B.2.4 implies that, for any $\check{X} \in \mathcal{X}$ and $g^{(j)} \in \mathbf{G}(I^{(j)})$, $g^{(j)}\check{X}$ and \check{X} have the same sufficient statistic in (3.13). From these observations and by finite induction, it follows that \tilde{X} and $g\tilde{X}$ have the same sufficient statistics in (3.13), as desired. \square

Proof of Lemma 3.4.3. By Lemma 3.4.2, we know (i) \mathbf{G} is a finite group of transformations of \mathcal{X} onto itself, and (ii) if X satisfies H_0 in (3.19), then X and gX have the same sufficient statistics in (3.13) for any $g \in \mathbf{G}$. The second statement, together with Lemma 3.4.1, implies that the randomization hypothesis holds (Lehmann and Romano (2005, Definition 15.2.1)), i.e., if X satisfies H_0 in (3.1), its distribution is invariant under the transformations in \mathbf{G} . Under these conditions, the result follows from Lehmann and Romano (2005, Eq. (15.6) and Problem 15.2). \square

Proof of Lemma 3.4.4. We condition on $X \in \mathcal{X}$ throughout this proof. Let $(G^{(1)}, \dots, G^{(K)})$ be as in Definition B.2.1. By Lemma B.2.5, it suffices to show that

$$\sup_{t \in \mathbb{R}} \left| \frac{1}{K} \sum_{k=1}^K 1\{\tau(G^{(k)}X) \geq t\} - \frac{1}{|\mathbf{G}|} \sum_{g \in \mathbf{G}} 1\{\tau(gX) \geq t\} \right| \xrightarrow{a.s.} 0 \quad \text{as } K \rightarrow \infty. \quad (\text{B-19})$$

For any $k = 1, \dots, K$, Definition B.2.1 implies that $G^{(k)}X \in \mathcal{X}$. Thus, $\tau(G^{(k)}X)$ takes values in the finite set $\{\tau(\tilde{X}) : \tilde{X} \in \mathcal{X}\}$. It then suffices to show the pointwise version of (B-19), i.e.,

$$\frac{1}{K} \sum_{k=1}^K 1\{\tau(G^{(k)}X) \geq t\} \xrightarrow{a.s.} \frac{1}{|\mathbf{G}|} \sum_{g \in \mathbf{G}} 1\{\tau(gX) \geq t\} \quad \text{as } K \rightarrow \infty.$$

By Definition B.2.1, $(G^{(1)}, \dots, G^{(K)})$ is the result of a Markov chain with transition probability given in (B-24). By Robert and Casella (2004, Algorithm A-24 and pages 270-1), we can equivalently interpret $(G^{(1)}, \dots, G^{(K)})$ as the outcome of a Metropolis-Hastings algorithm. For any $g, \check{g} \in \mathbf{G}$, this Metropolis-Hastings algorithm has a

conditional density $q(\check{g} | g) \equiv P(G^{(k+1)} = \check{g} | G^{(k)} = g)$, a target probability f defined by

$$f(g) \equiv \frac{1\{g \in \mathbf{G}\}}{|\mathbf{G}|}, \quad (\text{B-20})$$

and Metropolis-Hastings acceptance probability equal to one. To show the latter, note that, for every $g, \check{g} \in \mathbf{G}$,

$$\rho(g, \check{g}) = \min \left\{ \frac{f(\check{g})}{f(g)} \times \frac{q(g | \check{g})}{q(\check{g} | g)}, 1 \right\} \stackrel{(1)}{=} 1,$$

where (1) uses that $f(\check{g}) = f(g) = 1/|\mathbf{G}|$ and $q(g | \check{g}) = q(\check{g} | g)$ by (B-20) and Lemma B.2.9, respectively. By this and Robert and Casella (2004, Theorem 7.4), it suffices to show that the conditional density $q(\check{g} | g)$ is f -irreducible. By Robert and Casella (2004, Theorem 6.15, part (i)), this follows from showing that, for any $g, \check{g} \in \mathbf{G}$ (and so $f(g) > 0$ and $f(\check{g}) > 0$), the Markov chain has a positive probability of transitioning from g to \check{g} after a sufficient number of steps. We devote the rest of the proof to show this.

Consider any arbitrary choice of $g, \check{g} \in \mathbf{G}$. Since \mathbf{G} is the group generated by finitely many compositions of elements in $\bigcup_{I \in \mathcal{I}} \mathbf{G}(I)$, there are $(g^{(1)}, \dots, g^{(K_1+K_2)}) \in \mathbf{G}(I^{(1)}) \times \dots \times \mathbf{G}(I^{(K_1+K_2)})$ with $I^{(j)} \in \mathcal{I}$ for $j = 1, \dots, K$ such that $g = g^{(K_1)} \circ \dots \circ g^{(1)}$ and $\check{g} = g^{(K_1+K_2)} \circ \dots \circ g^{(K_1+1)}$. By Lemma B.2.3, $\mathbf{G}(I^{(j)})$ is a group for all $j = 1, \dots, K_1 + K_2$, and so $(g^{(j)})^{-1} \in \mathbf{G}(I^{(j)})$ for every $j = 1, \dots, K_1 + K_2$. Then, note that

$$\begin{aligned} \check{g} &= \check{g} \circ g^{-1} \circ g \\ &\stackrel{(1)}{=} g^{(K_1+K_2)} \circ \dots \circ g^{(K_1+1)} \circ (g^{(1)})^{-1} \circ \dots \circ (g^{(K_1)})^{-1} \circ g \\ &\stackrel{(2)}{=} \check{g}^{(K_1+K_2)} \circ \dots \circ \check{g}^{(K_1+1)} \circ \check{g}^{(K_1)} \circ \dots \circ \check{g}^{(1)} \circ g, \end{aligned} \quad (\text{B-21})$$

where (1) holds by setting $\check{g} = g^{(K_1+K_2)} \circ \dots \circ g^{(K_1+1)}$ and $g^{-1} = (g^{(1)})^{-1} \circ \dots \circ (g^{(K_1)})^{-1}$, and (2) holds by defining $(\check{g}^{(1)}, \dots, \check{g}^{(K_1+K_2)}) =$

$((g^{(K_1)})^{-1}, \dots, (g^{(1)})^{-1}, g^{(K_1+1)}, \dots, g^{(K_1+K_2)})$. Note that (B-21) provides a specific path for transitioning from g to \check{g} after $K_1 + K_2$ steps. We complete the proof by showing that $P(G^{(K_1+K_2+k)} = \check{g} | G^{(k)} = g) > 0$ for any positive integer k . To this end, we define $(\check{I}^{(1)}, \dots, \check{I}^{(K_1+K_2)}) = (I^{(K_1)}, \dots, I^{(1)}, I^{(K_1+1)}, \dots, I^{(K_1+K_2)})$ and we consider the following argument:

$$\begin{aligned} P(G^{(K_1+K_2+k)} = \check{g} | G^{(k)} = g) &\stackrel{(1)}{\geq} q(\check{g}^{(1)} \circ g | g) \prod_{k=2}^K q(\check{g}^{(k)} \circ \dots \circ \check{g}^{(1)} \circ g | \check{g}^{(k-1)} \circ \dots \circ \check{g}^{(1)} \circ g) \\ &\stackrel{(2)}{\geq} \prod_{j=1}^{K_1+K_2} \frac{1}{|\mathcal{I}||\mathbf{G}(\check{I}^j)|} \stackrel{(3)}{>} 0, \end{aligned}$$

where (1) uses the fact that the conditional distribution of $G^{(j+1)}$ given $G^{(j)}$ is q for all $j = 1, \dots, K_1 + K_2$, (2) holds by (B-24) and $\check{g}^{(j)} \in \check{I}^{(j)}$ for all $j = 1, \dots, K_1 + K_2$, and (3) holds because $\check{I}^{(j)} \in \mathcal{I}$ for all $j = 1, \dots, K_1 + K_2$. \square

B.2.2 Auxiliary results

Lemma B.2.1. *Under Assumptions 3.2.1 and H_0 in (3.1), (3.2) holds.*

Proof. Consider the following derivation.

$$\begin{aligned} P(X = \tilde{X}) &\stackrel{(1)}{=} \prod_{i=1}^n P((S_{i,t}, A_{i,t}) = (\tilde{S}_{i,t}, \tilde{A}_{i,t}) : t = 1, \dots, T) \\ &\stackrel{(2)}{=} \prod_{i=1}^n \left[\prod_{t=2}^T P((S_{i,t}, A_{i,t}) = (\tilde{S}_{i,t}, \tilde{A}_{i,t}) | (S_{i,t-1}, A_{i,t-1}) = (\tilde{S}_{i,t-1}, \tilde{A}_{i,t-1})) \times \right. \\ &\quad \left. P(S_{i,1} = \tilde{S}_{i,1}, A_{i,1} = \tilde{A}_{i,1}) \right] \\ &\stackrel{(3)}{=} \prod_{i=1}^n \left[P(S_{i,1} = \tilde{S}_{i,1}) \left(\prod_{t=1}^T P(A_{i,t} = \tilde{A}_{i,t} | S_{i,t} = \tilde{S}_{i,t}) \right) \right. \\ &\quad \left. \times \left(\prod_{t=1}^{T-1} P(S_{i,t+1} = \tilde{S}_{i,t+1} | S_{i,t} = \tilde{S}_{i,t}, A_{i,t} = \tilde{A}_{i,t}) \right) \right] \\ &\stackrel{(4)}{=} \prod_{i=1}^n \left[m_i(\tilde{S}_{i,1}) \left(\prod_{t=1}^T \sigma(\tilde{A}_{i,t} | \tilde{S}_{i,t}) \right) \left(\prod_{t=1}^{T-1} g(\tilde{S}_{i,t+1} | \tilde{S}_{i,t}, \tilde{A}_{i,t}) \right) \right], \end{aligned}$$

where (1) holds by Assumption 3.2.1(a), (2) holds by Assumption 3.2.1(b), (3) holds by Assumption 3.2.1(c), and (4) holds under H_0 in (3.1). \square

Lemma B.2.2. *Under Assumptions 3.2.1(b)-(c), the state variable is Markovian, i.e., for every $i = 1, \dots, n$ and $t = 2, \dots, T$ and every $\tilde{S} \in \mathcal{S}^{nT}$,*

$$P(S_{i,t} = \tilde{S}_{i,t} | S_{i,t-1} = \tilde{S}_{i,t-1}) = P(S_{i,t} = \tilde{S}_{i,t} | (S_{i,1}, \dots, S_{i,t-1}) = (\tilde{S}_{i,1}, \dots, \tilde{S}_{i,t-1})). \quad (\text{B-22})$$

Proof. Fix $i = 1, \dots, n$, $t = 2, \dots, T$, and $\tilde{S} \in \mathcal{S}^{nT}$ arbitrarily. Consider the following argument.

$$\begin{aligned} & P((S_{i,1}, \dots, S_{i,t}) = (\tilde{S}_{i,1}, \dots, \tilde{S}_{i,t})) \\ &= \sum_{(a_1, \dots, a_{t-1}) \in \mathcal{A}^{t-1}} \left(\begin{array}{c} P((S_{i,1}, A_{i,1}, \dots, S_{i,t-1}, A_{i,t-1}) = (\tilde{S}_{i,1}, a_1, \dots, \tilde{S}_{i,t-1}, a_{t-1})) \times \\ P(S_{i,t} = \tilde{S}_{i,t} | (S_{i,1}, A_{i,1}, \dots, S_{i,t-1}, A_{i,t-1}) = (\tilde{S}_{i,1}, a_1, \dots, \tilde{S}_{i,t-1}, a_{t-1})) \end{array} \right) \\ &\stackrel{(1)}{=} \sum_{(a_1, \dots, a_{t-1}) \in \mathcal{A}^{t-1}} \left(\begin{array}{c} P((S_{i,1}, A_{i,1}, \dots, S_{i,t-1}, A_{i,t-1}) = (\tilde{S}_{i,1}, a_1, \dots, \tilde{S}_{i,t-1}, a_{t-1})) \\ \times P(S_{i,t} = \tilde{S}_{i,t} | (S_{i,t-1}, A_{i,t-1}) = (\tilde{S}_{i,t-1}, a_{t-1})) \end{array} \right) \\ &= \sum_{a_{t-1} \in \mathcal{A}} \left(\begin{array}{c} P((S_{i,1}, \dots, S_{i,t-1}) = (\tilde{S}_{i,1}, \dots, \tilde{S}_{i,t-1}), A_{i,t-1} = a_{t-1}) \\ \times P(S_{i,t} = \tilde{S}_{i,t} | (S_{i,t-1}, A_{i,t-1}) = (\tilde{S}_{i,t-1}, a_{t-1})) \end{array} \right) \\ &= \sum_{a_{t-1} \in \mathcal{A}} \left(\begin{array}{c} P(A_{i,t-1} = a_{t-1} | (S_{i,1}, \dots, S_{i,t-1}) = (\tilde{S}_{i,1}, \dots, \tilde{S}_{i,t-1})) \\ \times P((S_{i,1}, \dots, S_{i,t-1}) = (\tilde{S}_{i,1}, \dots, \tilde{S}_{i,t-1})) \\ \times P(S_{i,t} = \tilde{S}_{i,t} | (S_{i,t-1}, A_{i,t-1}) = (\tilde{S}_{i,t-1}, a_{t-1})) \end{array} \right) \\ &\stackrel{(2)}{=} \sum_{a_{t-1} \in \mathcal{A}} \left(\begin{array}{c} P(A_{i,t-1} = a_{t-1} | S_{i,t-1} = \tilde{S}_{i,t-1}) P((S_{i,1}, \dots, S_{i,t-1}) = (\tilde{S}_{i,1}, \dots, \tilde{S}_{i,t-1})) \\ \times P(S_{i,t} = \tilde{S}_{i,t} | (S_{i,t-1}, A_{i,t-1}) = (\tilde{S}_{i,t-1}, a_{t-1})) \end{array} \right) \\ &= P(S_{i,t} = \tilde{S}_{i,t} | S_{i,t-1} = \tilde{S}_{i,t-1}) P((S_{i,1}, \dots, S_{i,t-1}) = (\tilde{S}_{i,1}, \dots, \tilde{S}_{i,t-1})) \end{aligned}$$

where (1) holds by Assumption 3.2.1(b) and (2) holds by Assumption 3.2.1(c). Therefore,

$$P(S_{i,t} = \tilde{S}_{i,t} | S_{i,t-1} = \tilde{S}_{i,t-1}) = P(S_{i,t} = \tilde{S}_{i,t} | (S_{i,1}, \dots, S_{i,t-1}) = (\tilde{S}_{i,1}, \dots, \tilde{S}_{i,t-1})),$$

as desired. \square

Lemma B.2.3. *For any $I \in \mathcal{I}$, $\mathbf{G}(I)$ is a group.*

Proof. We fix $I \in \mathcal{I}$ arbitrarily. It suffices to verify conditions (i)-(iv) in Lehmann and Romano (2005, Section A.1). Note that we can verify condition (ii) using the same argument as in Lehmann and Romano (2005, page 693).

We begin with condition (i). First, for any arbitrary $g_1, g_2 \in \mathbf{G}(I)$, we now verify that $g_2 \circ g_1 \in \mathbf{G}(I)$. Since $g_1, g_2 \in \mathbf{G}(I)$, g_1 and g_2 are both onto transformations of \mathcal{X} onto itself, then $g_2 \circ g_1$ is an onto transformation of \mathcal{X} onto itself. Now we will show that, for any $(\check{S}, \check{A}) \in \mathcal{X}$, the data configuration $(\tilde{S}, \tilde{A}) = (g_2 \circ g_1)(\check{S}, \check{A})$ satisfies $\tilde{S} \in R_S(I, \check{S})$ and $\tilde{A} \in R_A(\tilde{S}, (\check{S}, \check{A}))$. Define $(\dot{S}, \dot{A}) = g_1(\check{S}, \check{A})$. Now $(\dot{S}, \dot{A}) = g_1(\check{S}, \check{A})$ and $(\tilde{S}, \tilde{A}) = g_2(\dot{S}, \dot{A})$. Since $g_1, g_2 \in \mathbf{G}(I)$, all the conditions in Definitions 3.3.1 and 3.3.2 satisfy the transitive property as the equality condition, so that $\tilde{S} \in R_S(I, \check{S})$ and $\tilde{A} \in R_A(\tilde{S}, (\check{S}, \check{A}))$, as desired. By combining these results, we conclude that $g_2 \circ g_1 \in \mathbf{G}(I)$, as desired.

To verify condition (iii), we now show that the identity transformation belongs to $\mathbf{G}(I)$. To this end, we note that the identity transformation is an onto transformation of \mathcal{X} onto itself, and $\check{S} \in R_S(I, \check{S})$ and $\check{A} \in R_A(\check{S}, (\check{S}, \check{A}))$.

To verify condition (iv), we now show that for any $g \in \mathbf{G}(I)$, $g^{-1} \in \mathbf{G}(I)$ holds. By definition $\mathbf{G}(I)$ is a collection of onto transformations that map a finite set \mathcal{X} onto itself. By the pigeonhole principle, the transformations in $\mathbf{G}(I)$ are one to one, i.e., bijective, implying that g^{-1} is well defined. First, note that g^{-1} is a bijective transformation (hence, an onto transformation) of \mathcal{X} onto itself. For the rest of the verification of Condition (iv), pick $\tilde{X} \in \mathcal{X}$ arbitrarily. Second, we would like to show that, for any $(\check{S}, \check{A}) \in \mathcal{X}$, the data configuration $(\tilde{S}, \tilde{A}) = g^{-1}(\check{S}, \check{A})$ satisfies $\tilde{S} \in R_S(I, \check{S})$ and $\tilde{A} \in R_A(\tilde{S}, (\check{S}, \check{A}))$. Since $g \in \mathbf{G}(I)$ and $g(\tilde{S}, \tilde{A}) = g(g^{-1}(\check{S}, \check{A})) = (\check{S}, \check{A})$, we have $\check{S} \in R_S(I, \tilde{S})$ and $\check{A} \in R_A(\check{S}, (\tilde{S}, \tilde{A}))$. Note that all the conditions in Definitions 3.3.1 and 3.3.2 treat (\tilde{S}, \tilde{A}) and (\check{S}, \check{A}) symmetrically. Therefore, we have $\tilde{S} \in R_S(I, \check{S})$ and $\tilde{A} \in R_A(\tilde{S}, (\check{S}, \check{A}))$, as desired. \square

Lemma B.2.4. For any $I \in \mathcal{I}$ and any $g \in \mathbf{G}(I)$, \check{X} and $g\check{X}$ have the same sufficient statistic in (3.13), i.e., $U(\check{X}) = U(g\check{X})$.

Proof. Let $\check{X} = (\check{S}, \check{A})$ and $\tilde{X} = (\tilde{S}, \tilde{A}) = g(\check{S}, \check{A})$. By definition 3.4.1, this implies that $\tilde{S} \in R_S(\check{S})$ and $\tilde{A} \in R_A(\tilde{S}, (\check{S}, \check{A}))$. By (3.13), it then suffices to show the following statements:

1. $\check{S}_{i,1} = \tilde{S}_{i,1}$ for all $i = 1, \dots, n$,
2. $\sum_{i=1}^n \sum_{t=1}^{T-1} 1\{\tilde{S}_{i,t} = s, \tilde{A}_{i,t} = a, \tilde{S}_{i,t+1} = s'\} = \sum_{i=1}^n \sum_{t=1}^{T-1} 1\{\check{S}_{i,t} = s, \check{A}_{i,t} = a, \check{S}_{i,t+1} = s'\}$ for all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$,
3. $\sum_{i=1}^n 1\{\tilde{S}_{i,T} = s, \tilde{A}_{i,T} = a\} = \sum_{i=1}^n 1\{\check{S}_{i,T} = s, \check{A}_{i,T} = a\}$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

The first statement follows from $\tilde{S} \in R_S(\check{S})$ and condition (a) in Definition 3.3.1. The second and third statements follow from $\tilde{A} \in R_A(\tilde{S}, (\check{S}, \check{A}))$ and conditions (a) and (b) in Definition 3.3.2, respectively. \square

Several upcoming results involve a Markov chain of transformations in \mathbf{G} , specified in Definition B.2.1.

Definition B.2.1. Let $(G^{(1)}, \dots, G^{(K)})$ denote a Markov chain of transformations of \mathcal{X} onto itself that is defined as follows:

- $G^{(1)} : \mathcal{X} \rightarrow \mathcal{X}$ be equal to the identity transformation, i.e., $x = G^{(1)}x$ for any $x \in \mathcal{X}$
- For any $k = 2, \dots, K$ and given $(G^{(1)}, \dots, G^{(k-1)}, X)$, $G^{(k)} : \mathcal{X} \rightarrow \mathcal{X}$ is a random transformation distributed according to the following transition probability:

$$P(G^{(k)} = \tilde{g} \mid G^{(1)}, \dots, G^{(k-1)}, X) = P(G^{(k)} = \tilde{g} \mid G^{(k-1)}) \quad (\text{B-23})$$

$$= \sum_{I \in \mathcal{I}} \sum_{g \in \mathbf{G}(I)} \frac{1\{\tilde{g} = g \circ (G^{(k-1)})\}}{|\mathcal{I}| \times |\mathbf{G}(I)|}. \quad (\text{B-24})$$

Lemma B.2.5. *Conditional on X , $(X^{(1)}, \dots, X^{(K)})$ generated by Algorithm 3.3.1 and $(G^{(1)}X, \dots, G^{(K)}X)$ with $(G^{(1)}, \dots, G^{(K)})$ as in Definition B.2.1 have the same distribution.*

Proof. We condition on X throughout this proof. First, note that Algorithm 3.3.1 and Definition B.2.1 imply that $X = X^{(1)} = G^{(1)}X$. Second, note that $(X^{(1)}, \dots, X^{(K)})$ and $(G^{(1)}X, \dots, G^{(K)}X)$ are both Markov chains in \mathcal{X} . To complete the proof, it suffices to show that they have the same transition probabilities. The transition probability of $(X^{(1)}, \dots, X^{(K)})$ is specified in (3.9). It then suffices to show that, for any $k = 2, \dots, K$, $\tilde{X} = (\tilde{S}, \tilde{A}) \in \mathcal{X}$, and $G^{(k-1)}X = \check{X} = (\check{S}, \check{A}) \in \mathcal{X}$,

$$\begin{aligned} & P(G^{(k)}X = \tilde{X} \mid G^{(1)}X, \dots, G^{(k-2)}X, G^{(k-1)}X = \check{X}, X) \\ &= P(G^{(k)}X = \tilde{X} \mid G^{(k-1)}X = \check{X}, X) \\ &= \begin{cases} \sum_{I \in \mathcal{I}} \frac{\mathbf{1}\{\tilde{S} \in R_S(I, \check{S}), \tilde{A} \in R_A(\tilde{S}, \check{X})\}}{|\mathcal{I}| \times |R_S(I, \check{S})| \times |R_A(\tilde{S}, \check{X})|} & \text{if } |R_S(I, \check{S})| \times |R_A(\tilde{S}, \check{X})| > 0, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{B-25}$$

For the rest of the proof, we fix $k = 2, \dots, K$, and $\tilde{X} = (\tilde{S}, \tilde{A}), \check{X} = (\check{S}, \check{A}) \in \mathcal{X}$ arbitrarily. To show (B-25), consider the following derivation.

$$\begin{aligned} & P(G^{(k)}X = \tilde{X} \mid G^{(1)}X, \dots, G^{(k-2)}X, G^{(k-1)}X = \check{X}, X) \\ &\stackrel{(1)}{=} E[P(G^{(k)}X = \tilde{X} \mid G^{(1)}, \dots, G^{(k-1)}, X) \mid G^{(1)}X, \dots, G^{(k-2)}X, G^{(k-1)}X = \check{X}, X] \\ &\stackrel{(2)}{=} E[P(G^{(k)}X = \tilde{X} \mid G^{(k-1)}, X) \mid G^{(1)}X, \dots, G^{(k-2)}X, G^{(k-1)}X = \check{X}, X], \end{aligned} \tag{B-26}$$

where (1) holds by the law of total probability, and (2) holds by (B-24). From (B-26), (B-25) follows if we show that, for $G^{(k-1)}X = \check{X}$, $P(G^{(k)}X = \tilde{X} \mid G^{(k-1)}, X)$ is equal to

$$\begin{cases} \sum_{I \in \mathcal{I}} \frac{\mathbf{1}\{\tilde{S} \in R_S(I, \check{S}), \tilde{A} \in R_A(\tilde{S}, \check{X})\}}{|\mathcal{I}| \times |R_S(I, \check{S})| \times |R_A(\tilde{S}, \check{X})|} & \text{if } |R_S(I, \check{S})| \times |R_A(\tilde{S}, \check{X})| > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{B-27}$$

To show (B-27), consider the following derivation.

$$\begin{aligned}
& P(G^{(k)}X = \tilde{X} \mid G^{(k-1)}, X) && \text{(B-28)} \\
& \stackrel{(1)}{=} P(G^{(k)}(G^{(k-1)})^{-1}\check{X} = \tilde{X} \mid G^{(k-1)}, X) \\
& = \sum_{g \in \mathbf{G}} P(G^{(k)} = g \mid G^{(k-1)}, X) \mathbf{1}\{g(G^{(k-1)})^{-1}\check{X} = \tilde{X}\} \\
& \stackrel{(2)}{=} \sum_{g \in \mathbf{G}} \frac{\sum_{I \in \mathcal{I}} \frac{1}{|\mathcal{I}|} \sum_{\tilde{g} \in \mathbf{G}(I)} \mathbf{1}\{g = \tilde{g} \circ (G^{(k-1)})^{-1}\}}{|\mathbf{G}(I)|} \mathbf{1}\{g(G^{(k-1)})^{-1}\check{X} = \tilde{X}\} \\
& = \frac{1}{|\mathcal{I}|} \sum_{I \in \mathcal{I}} \frac{\sum_{\tilde{g} \in \mathbf{G}(I)} \sum_{g \in \mathbf{G}} \mathbf{1}\{g = \tilde{g} \circ (G^{(k-1)})\} \mathbf{1}\{g(G^{(k-1)})^{-1}\check{X} = \tilde{X}\}}{|\mathbf{G}(I)|} \\
& \stackrel{(3)}{=} \frac{1}{|\mathcal{I}|} \sum_{I \in \mathcal{I}} \frac{\sum_{\tilde{g} \in \mathbf{G}(I)} \sum_{g \in \mathbf{G}} \mathbf{1}\{g = \tilde{g} \circ (G^{(k-1)})\} \mathbf{1}\{\tilde{g}\check{X} = \tilde{X}\}}{|\mathbf{G}(I)|} \\
& = \frac{1}{|\mathcal{I}|} \sum_{I \in \mathcal{I}} \frac{\sum_{\tilde{g} \in \mathbf{G}(I)} \mathbf{1}\{\tilde{g}\check{X} = \tilde{X}\} \sum_{g \in \mathbf{G}} \mathbf{1}\{g = \tilde{g} \circ (G^{(k-1)})\}}{|\mathbf{G}(I)|} \\
& \stackrel{(4)}{=} \frac{1}{|\mathcal{I}|} \sum_{I \in \mathcal{I}} \frac{\sum_{\tilde{g} \in \mathbf{G}(I)} \mathbf{1}\{\tilde{g}\check{X} = \tilde{X}\}}{|\mathbf{G}(I)|}, && \text{(B-29)}
\end{aligned}$$

where (1) holds by $G^{(k-1)}X = \check{X}$ and the fact that $(G^{(k-1)})^{-1} \in \mathbf{G}$ since \mathbf{G} is a group (by Lemma 3.4.2), (2) holds by (B-24), (3) holds because $\{g = \tilde{g} \circ (G^{(k-1)})\}$ occurs if and only if $\{g(G^{(k-1)})^{-1} = \tilde{g}\}$, and (4) holds because $\sum_{g \in \mathbf{G}} \mathbf{1}\{g = \tilde{g} \circ (G^{(k-1)})\} = 1$, as we show in the next paragraph.

To show $\sum_{g \in \mathbf{G}} \mathbf{1}\{g = \tilde{g} \circ (G^{(k-1)})\} = 1$, consider the following argument. Since $g, \tilde{g}, G^{(k-1)} \in \mathbf{G}$, and \mathbf{G} is a group, $\tilde{g} \circ (G^{(k-1)}) \in \mathbf{G}$, and so $\exists g \in \mathbf{G}$ s.t. $g = \tilde{g} \circ (G^{(k-1)})$, i.e., $\sum_{g \in \mathbf{G}} \mathbf{1}\{g = \tilde{g} \circ (G^{(k-1)})\} \geq 1$. Now, suppose that $\sum_{g \in \mathbf{G}} \mathbf{1}\{g = \tilde{g} \circ (G^{(k-1)})\} > 1$. This implies that $\exists g_1, g_2 \in \mathbf{G}$ with $g_1 \neq g_2$ s.t. $g_1 = \tilde{g} \circ (G^{(k-1)}) = g_2$. But using again that \mathbf{G} is a group, $\exists g_1^{-1} \in \mathbf{G}$ and so $g_1^{-1}g_2 = g_1^{-1}g_1$ and $g_2g_1^{-1} = g_1g_1^{-1}$ and both equal to the identity transformation. This would imply that $g_1^{-1} = g_2^{-1}$, and

since the inverse transformation is unique, reach a contradiction.

Fix $I \in \mathcal{I}$ arbitrarily. By (B-29), (B-27) then follows from showing that $\sum_{\tilde{g} \in \mathbf{G}(I)} \frac{1_{\{\tilde{g}\check{X} = \check{X}\}}}{|\mathbf{G}(I)|}$ equals

$$\begin{cases} \frac{1_{\{\tilde{S} \in R_S(I, \check{S}), \tilde{A} \in R_A(\tilde{S}, \check{X})\}}}{|R_S(I, \check{S})| \times |R_A(\tilde{S}, \check{X})|} & \text{if } |R_S(I, \check{S})| \times |R_A(\tilde{S}, \check{X})| > 0, \\ 0 & \text{otherwise.} \end{cases} \quad (\text{B-30})$$

We divide our argument into two cases. First, consider $|R_S(I, \check{S})| \times |R_A(\tilde{S}, \check{X})| = 0$. In this case, we have $\nexists g \in \mathbf{G}(I)$ s.t. $g\check{X} = \check{X}$, and therefore

$$\sum_{g \in \mathbf{G}(I)} \frac{1_{\{g\check{X} = \check{X}\}}}{|\mathbf{G}(I)|} = 0,$$

which verifies (B-30).

Second, consider $|R_S(I, \check{S})| \times |R_A(\tilde{S}, \check{X})| > 0$. Then, consider the following derivation.

$$\begin{aligned} \frac{\sum_{g \in \mathbf{G}(I)} 1_{\{g\check{X} = \check{X}\}}}{|\mathbf{G}(I)|} &\stackrel{(1)}{=} \frac{\sum_{g \in \mathbf{G}(I)} 1_{\{g\check{X} = \check{X}\}}}{|\mathbf{G}(I)|} 1_{\{\tilde{S} \in R_S(I, \check{S}), \tilde{A} \in R_A(\tilde{S}, \check{X})\}} \\ &\stackrel{(2)}{=} \frac{\sum_{g \in \mathbf{G}(I)} 1_{\{g\check{X} = \check{X}\}}}{|\mathbf{G}(I)|} 1_{\{\tilde{S} \in R_S(I, \check{S}), \tilde{A} \in R_A(\tilde{S}, \check{X})\}} \\ &\stackrel{(3)}{=} \frac{1_{\{\tilde{S} \in R_S(I, \check{S}), \tilde{A} \in R_A(\tilde{S}, \check{X})\}}}{|R_S(I, \check{S})| \times |R_A(\tilde{S}, \check{X})|} \\ &\stackrel{(4)}{=} \frac{1_{\{\tilde{S} \in R_S(I, \check{S}), \tilde{A} \in R_A(\tilde{S}, \check{X})\}}}{|R_S(I, \check{S})| \times |R_A(\tilde{S}, \check{X})|}, \end{aligned} \quad (\text{B-31})$$

which verifies (B-30), where (1) follows from Definition 3.4.1, as it implies that $\{\tilde{g}\check{X} = \check{X}\}$ with $\tilde{g} \in \mathbf{G}(I)$, $\check{X} = (\check{S}, \check{A})$, and $\check{X} = (\tilde{S}, \tilde{A})$ implies that $\{\tilde{S} \in R_S(I, \check{S})\}$ and $\{\tilde{A} \in R_A(\tilde{S}, \check{X})\}$, (2) follows from Lemma B.2.6, and (3) is shown in (B-32), and (4) follows from Lemma B.2.7 (which applies because the expression is multiplied by $1_{\{\tilde{S} \in R_S(I, \check{S})\}}$).

To show (3) in (B-31), consider the following argument.

$$\begin{aligned}
|\mathbf{G}(I)| &\stackrel{(1)}{=} \sum_{\check{X} \in \mathcal{X}} \left(\sum_{g \in \mathbf{G}(I)} 1\{g\check{X} = \check{X}\} \right) \\
&\stackrel{(2)}{=} \sum_{\check{S} \in R_S(I, \check{S})} \sum_{\check{A} \in R_A(\check{S}, \check{X})} \left(\sum_{g \in \mathbf{G}(I)} 1\{g\check{X} = \check{X}\} \right) \\
&\stackrel{(3)}{=} \left(\sum_{g \in \mathbf{G}(I)} 1\{g\check{X} = \check{X}\} \right) \sum_{\check{S} \in R_S(I, \check{S})} \sum_{\check{A} \in R_A(\check{S}, \check{X})} \\
&= \left(\sum_{g \in \mathbf{G}(I)} 1\{g\check{X} = \check{X}\} \right) \sum_{\check{S} \in R_S(I, \check{S})} |R_A(\check{S}, \check{X})| \\
&\stackrel{(4)}{=} \left(\sum_{g \in \mathbf{G}(I)} 1\{g\check{X} = \check{X}\} \right) |R_A(\check{S}, \check{X})| \sum_{\check{S} \in R_S(I, \check{S})} 1 \\
&= \left(\sum_{g \in \mathbf{G}(I)} 1\{g\check{X} = \check{X}\} \right) |R_A(\check{S}, \check{X})| \times |R_S(I, \check{S})|, \tag{B-32}
\end{aligned}$$

where (1) follows from partitioning $\mathbf{G}(I)$ into its possible range of outcomes when applied to $\check{X} \in \mathcal{X} = \mathcal{S}^{nT} \times \mathcal{A}^{nT}$, (2) follows from Definition 3.4.1, as it implies that $\{g\check{X} = \check{X}\}$ with $\check{g} \in \mathbf{G}(I)$, $\check{X} = (\check{S}, \check{A})$, and $\check{X} = (\check{S}, \check{A})$ if and only if $\{\check{S} \in R_S(I, \check{S})\}$ and $\{\check{A} \in R_A(\check{S}, \check{X})\}$, (3) follows from Lemma B.2.6, and (4) follows from Lemma B.2.7. \square

Lemma B.2.6. *Fix $\check{X} = (\check{S}, \check{A}) \in \mathcal{X}$, $\check{X} = (\check{S}, \check{A}) \in \mathcal{X}$, and $I \in \mathcal{I}$ arbitrarily. Then, $\check{S} \in R_S(I, \check{S})$ and $\check{A} \in R_A(\check{S}, \check{X})$ implies that $\sum_{g \in \mathbf{G}(I)} 1\{g\check{X} = \check{X}\} = \sum_{g \in \mathbf{G}(I)} 1\{g\check{X} = \check{X}\}$.*

Proof. Fix $\check{X} = (\check{S}, \check{A}) \in \mathcal{X}$, $\check{X} \in \mathcal{X}$, and $I \in \mathcal{I}$ arbitrarily, and assume that $\check{S} \in R_S(I, \check{S})$ and $\check{A} \in R_A(\check{S}, \check{X})$. By definition of $\mathbf{G}(I)$, $R_S(I, \check{S})$, and $R_A(\check{S}, \check{X})$,

$\tilde{S} \in R_S(I, \check{S})$ and $\tilde{A} \in R_A(\tilde{S}, \check{X})$ implies that $\exists \check{g} \in \mathbf{G}(I)$ s.t. $\check{g}\check{X} = \tilde{X}$. Therefore,

$$\sum_{g \in \mathbf{G}(I)} 1\{g\check{X} = \tilde{X}\} \stackrel{(1)}{=} \sum_{g \in \mathbf{G}(I)} 1\{g\check{X} = \check{g}\check{X}\} \stackrel{(2)}{=} \sum_{g \in \mathbf{G}(I)} 1\{\check{g}^{-1}g\check{X} = \check{X}\} \stackrel{(3)}{=} \sum_{g \in \mathbf{G}(I)} 1\{g\check{X} = \check{X}\},$$

where (1) holds by $\check{g}\check{X} = \tilde{X}$, (2) holds because $\check{g} \in \mathbf{G}(I)$ and that $\mathbf{G}(I)$ is a group (by Lemma B.2.3), (3) holds by $\{\check{g}^{-1}g : g \in \mathbf{G}(I)\} = \mathbf{G}(I)$, as $\mathbf{G}(I)$ is a group (again, by Lemma B.2.3). \square

Lemma B.2.7. *Fix $\check{X} = (\check{S}, \check{A}) \in \mathcal{X}$ and $I \in \mathcal{I}$ arbitrarily. Then, $\tilde{S} \in R_S(I, \check{S})$ implies that $|R_A(\tilde{S}, \check{X})| = |R_A(\check{S}, \check{X})|$.*

Proof. Fix $\check{X} = (\check{S}, \check{A}) \in \mathcal{X}$ and $I \in \mathcal{I}$ arbitrarily, and assume that $\tilde{S} \in R_S(I, \check{S})$.

We first show that $|R_A(\check{S}, (\check{S}, \check{A}))| \leq |R_A(\tilde{S}, (\check{S}, \check{A}))|$. Let (A^1, \dots, A^C) enumerate the (distinct) elements in $R_A(\check{S}, (\check{S}, \check{A}))$. By $\tilde{S} \in R_S(I, \check{S})$ and Lemma B.2.8, there is a permutation π s.t. $\tilde{S} = \check{S}_\pi$ and $A_\pi^c \in R_A(\tilde{S}, (\check{S}, A^c))$ for each $c = 1, \dots, C$. We now show that $(A_\pi^1, \dots, A_\pi^C)$ are all distinct elements. To this end, suppose that $\exists c_1, c_2 \in \{1, \dots, C\}$ s.t. $A_\pi^{c_1} = A_\pi^{c_2}$. If that were the case, and by the fact that a permutation is a bijective relationship, we conclude that $A^{c_1} = A^{c_2}$. Since (A^1, \dots, A^C) are distinct, we conclude that $c_1 = c_2$, as desired. To conclude the argument, it suffices to show that $A_\pi^c \in R_A(\tilde{S}, (\check{S}, \check{A}))$ for all $c = 1, \dots, C$. To this end, choose $c = 1, \dots, C$ arbitrarily. Since $\check{S} \in R_S(I, \check{S})$ (trivially) and $A^c \in R_A(\check{S}, (\check{S}, \check{A}))$, Definition 3.4.1 implies that $\exists g_1 \in \mathbf{G}(I)$ s.t. $g_1(\check{S}, \check{A}) = (\check{S}, A^c)$. Since $\tilde{S} \in R_S(I, \check{S})$ and $A_\pi^c \in R_A(\tilde{S}, (\check{S}, A^c))$, Definition 3.4.1 implies that $\exists g_2 \in \mathbf{G}(I)$ s.t. $g_2(\tilde{S}, A^c) = (\tilde{S}, A_\pi^c)$. Since $\mathbf{G}(I)$ is a group (by Lemma B.2.3), we conclude that $g_3 = g_2 \circ g_1 \in \mathbf{G}(I)$. Since $g_3(\check{S}, \check{A}) = (\tilde{S}, A_\pi^c)$ and $g_3 \in \mathbf{G}(I)$, Definition 3.4.1 implies that $A_\pi^c \in R_A(\tilde{S}, (\check{S}, \check{A}))$, as desired.

We next show that $|R_A(\check{S}, (\check{S}, \check{A}))| \geq |R_A(\tilde{S}, (\check{S}, \check{A}))|$. Let (A^1, \dots, A^C) enumerate the (distinct) elements in $R_A(\tilde{S}, (\check{S}, \check{A}))$. Since $\tilde{S} \in R_S(I, \check{S})$ and by the fact that

the Definition 3.3.1 treats \tilde{S} and \check{S} symmetrically, we conclude that $\check{S} \in R_S(I, \tilde{S})$. In turn, by $\check{S} \in R_S(I, \tilde{S})$ and Lemma B.2.8, there is a permutation π s.t. $\check{S} = \tilde{S}_\pi$ and $A_\pi^c \in R_A(\check{S}, (\tilde{S}, A^c))$ for each $c = 1, \dots, C$. By repeating the previous argument, we can show that $(A_\pi^1, \dots, A_\pi^C)$ are all distinct elements. To conclude the proof, it suffices to show that $A_\pi^c \in R_A(\check{S}, (\check{S}, \check{A}))$ for all $c = 1, \dots, C$. To this end, choose $c = 1, \dots, C$ arbitrarily. Since $\tilde{S} \in R_S(I, \check{S})$ and $A^c \in R_A(\tilde{S}, (\check{S}, \check{A}))$, Definition 3.4.1 implies that $\exists g_1 \in \mathbf{G}(I)$ s.t. $g_1(\check{S}, \check{A}) = (\tilde{S}, A^c)$. Since $\check{S} \in R_S(I, \tilde{S})$ and $A_\pi^c \in R_A(\check{S}, (\tilde{S}, A^c))$, Definition 3.4.1 implies that $\exists g_2 \in \mathbf{G}(I)$ s.t. $g_2(\tilde{S}, A^c) = (\check{S}, A_\pi^c)$. Since $\mathbf{G}(I)$ is a group (by Lemma B.2.3), we conclude that $g_3 = g_2g_1 \in \mathbf{G}(I)$. Since $g_3(\check{S}, \check{A}) = (\check{S}, A_\pi^c)$ and $g_3 \in \mathbf{G}(I)$, Definition 3.4.1 implies that $A_\pi^c \in R_A(\check{S}, (\check{S}, \check{A}))$, as desired. \square

Lemma B.2.8. *For any $\check{S} \in \mathcal{S}^{nT}$, $I \in \mathcal{I}$ and $\tilde{S} \in R_S(I, \check{S})$, there exists a permutation $\pi : \{1, \dots, n\} \times \{1, \dots, T\} \rightarrow \{1, \dots, n\} \times \{1, \dots, T\}$ such that $\check{S} = \tilde{S}_\pi$ and $\check{A}_\pi \in R_A(\tilde{S}, (\check{S}, \check{A}))$ for every $\check{A} \in \mathcal{A}^{nT}$.*

Proof. Fix $\check{S} \in \mathcal{S}^{nT}$ and $I \in \mathcal{I}$ arbitrarily and assume that $\tilde{S} \in R_S(I, \check{S})$. For every $s, s' \in \mathcal{S}$, let

$$\text{Index}_1(s, s') = \{(i, t) \in \{1, \dots, n\} \times \{1, \dots, T-1\} : (\check{S}_{i,t}, \check{S}_{i,t+1}) = (s, s')\},$$

$$\text{Index}_2(s, s') = \{(i, t) \in \{1, \dots, n\} \times \{1, \dots, T-1\} : (\tilde{S}_{i,t}, \tilde{S}_{i,t+1}) = (s, s')\},$$

$$\text{Index}_1(s) = \{(i, T) : i \in \{1, \dots, n\}, \check{S}_{i,T} = s\},$$

$$\text{Index}_2(s) = \{(i, T) : i \in \{1, \dots, n\}, \tilde{S}_{i,T} = s\}.$$

We use

$$C(s, s') \equiv |\text{Index}_1(s, s')| \stackrel{(1)}{=} |\text{Index}_2(s, s')|,$$

$$C(s) \equiv |\text{Index}_1(s)| \stackrel{(2)}{=} |\text{Index}_2(s)|,$$

where (1) and (2) hold by $\tilde{S} \in R_S(I, \check{S})$.

For every $s, s' \in \mathcal{S}$, we can enumerate $\text{Index}_1(s, s')$ by $(\nu_1(1, s, s'), \dots, \nu_1(C(s, s'), s, s'))$, $\text{Index}_2(s, s')$ by $(\nu_2(1, s, s'), \dots, \nu_2(C(s, s'), s, s'))$, $\text{Index}_1(s)$ by $(\nu_1(1, s), \dots, \nu_1(C(s), s))$, and $\text{Index}_2(s)$ by $(\nu_2(1, s), \dots, \nu_2(C(s), s))$. By definition, $(\nu_1(1, s, s'), \dots, \nu_1(C(s, s'), s, s'))$ represent the (i, t) indices that satisfy $(\check{S}_{i,t}, \check{S}_{i,t+1}) = (s, s')$ and $(\nu_1(1, s), \dots, \nu_1(C(s), s))$ represent the (i, T) indices that satisfy $\check{S}_{i,T} = s$, $(\nu_2(1, s, s'), \dots, \nu_2(C(s, s'), s, s'))$ represent the (i, t) indices that satisfy $(\tilde{S}_{i,t}, \tilde{S}_{i,t+1}) = (s, s')$, and $(\nu_2(1, s), \dots, \nu_2(C(s), s))$ represent the (i, T) indices that satisfy $\tilde{S}_{i,T} = s$.

These enumerations allows us to interpret \check{S} as a permutation of the values of \tilde{S} . We denote this permutation by $\pi : \{1, \dots, n\} \times \{1, \dots, T\} \rightarrow \{1, \dots, n\} \times \{1, \dots, T\}$, and we characterize it next. For any $(i, t) \in \{1, \dots, n\} \times \{1, \dots, T-1\}$, there exists $(s, s') \in \mathcal{S}$ and $c = 1, \dots, C(s, s')$ s.t. $(i, t) = \nu_1(c, s, s') \in \text{Index}_1(s, s')$. In this case, set $\pi(i, t) = \nu_2(c, s, s')$. By this construction,

$$\check{S}_{i,t} = \check{S}_{\nu_1(c,s,s')} = \tilde{S}_{\nu_2(c,s,s')} = \tilde{S}_{\pi(i,t)},$$

Similarly, for any $i \in \{1, \dots, n\}$, there exists $s \in \mathcal{S}$ and $c = 1, \dots, C(s)$ s.t. $(i, T) = \nu_1(c, s) \in \text{Index}_1(s)$. In this case, set $\pi(i, T) = \nu_2(c, s)$. By this construction,

$$\check{S}_{i,T} = \check{S}_{\nu_1(c,s)} = \tilde{S}_{\nu_2(c,s)} = \tilde{S}_{\pi(i,T)}.$$

To show the second part, for any $\check{A} \in \mathcal{A}^{nT}$, consider $\tilde{A} = \check{A}_\pi$. For each $s, s' \in \mathcal{S}$, note that

$$\begin{aligned} \tilde{A}_{\nu_2(c,s,s')} &= \check{A}_{\nu_1(c,s,s')} \quad \text{for } c = 1, \dots, C(s, s') \\ \tilde{A}_{\nu_2(c,s)} &= \check{A}_{\nu_1(c,s)} \quad \text{for } c = 1, \dots, C(s). \end{aligned} \tag{B-33}$$

To complete the proof, it suffices to show that $\tilde{A} \in R_A(\tilde{S}, \check{X})$. To this end, it suffices to verify conditions (a)-(b) in Definition 3.3.2. We only show condition (a),

as condition (b) can be shown using an analogous argument. For any $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$, consider the following derivation.

$$\begin{aligned}
\sum_{i=1}^n \sum_{t=1}^{T-1} 1\{\check{S}_{i,t} = s, \check{A}_{i,t} = a, \check{S}_{i,t+1} = s'\} &\stackrel{(1)}{=} \sum_{(i,t) \in \text{Index}_1(s,s')} 1\{\check{A}_{i,t} = a\} \\
&= \sum_{c=1}^{C(s,s')} 1\{\check{A}_{\nu_1(c,s,s')} = a\} \\
&\stackrel{(2)}{=} \sum_{c=1}^{C(s,s')} 1\{\tilde{A}_{\nu_2(c,s,s')} = a\} \\
&= \sum_{(i,t) \in \text{Index}_2(s,s')} 1\{\tilde{A}_{i,t} = a\} \\
&\stackrel{(3)}{=} \sum_{i=1}^n \sum_{t=1}^{T-1} 1\{\tilde{S}_{i,t} = s, \tilde{A}_{i,t} = a, \tilde{S}_{i,t+1} = s'\},
\end{aligned} \tag{B-34}$$

where (1) follows from the fact that $\text{Index}_1(s, s')$ is the collection of all indices $(i, t) \in \{1, \dots, n\} \times \{1, \dots, T-1\}$ s.t. $(\check{S}_{i,t}, \check{S}_{i,t+1}) = (s, s')$, (2) holds by (B-33), and (3) follows from the fact that $\text{Index}_2(s, s')$ is the collection of all indices $(i, t) \in \{1, \dots, n\} \times \{1, \dots, T-1\}$ s.t. $(\tilde{S}_{i,t}, \tilde{S}_{i,t+1}) = (s, s')$. \square

Lemma B.2.9. *The transition probability in (B-24) is symmetric, i.e., for any $g, \check{g} \in \mathbf{G}$, $P(G^{(k+1)} = \check{g} | G^{(k)} = g) = P(G^{(k+1)} = g | G^{(k)} = \check{g})$.*

Proof. Fix $g, \check{g} \in \mathbf{G}$ arbitrarily and consider the following argument.

$$\begin{aligned}
P(G^{(k+1)} = \check{g} | G^{(k)} = g) &= \sum_{I \in \mathcal{I}} \frac{1}{|\mathcal{I}|} \sum_{\tilde{g} \in \mathbf{G}(I)} \frac{1\{\check{g} = \tilde{g} \circ g\}}{|\mathbf{G}(I)|} \\
&\stackrel{(1)}{=} \sum_{I \in \mathcal{I}} \frac{1}{|\mathcal{I}|} \sum_{\tilde{g} \in \mathbf{G}(I)} \frac{1\{g = \tilde{g}^{-1} \circ \check{g}\}}{|\mathbf{G}(I)|} \\
&\stackrel{(2)}{=} \sum_{I \in \mathcal{I}} \frac{1}{|\mathcal{I}|} \sum_{\tilde{g} \in \mathbf{G}(I)} \frac{1\{g = \tilde{g} \circ \check{g}\}}{|\mathbf{G}(I)|} \\
&= P(G^{(k+1)} = g | G^{(k)} = \check{g}),
\end{aligned}$$

where (1) follows from the fact that $\mathbf{G}(I)$ is a group (by Lemma B.2.3), and so $\exists \tilde{g}^{-1} \in \mathbf{G}(I)$ for any $\tilde{g} \in \mathbf{G}(I)$, and that $1\{\check{g} = \tilde{g} \circ g\} = 1\{\tilde{g}^{-1} \circ \check{g} = g\}$, and (2) follows from defining $\mathbf{G}(I) = \{\tilde{g}^{-1} : \tilde{g} \in \mathbf{G}(I)\}$, which holds because $\mathbf{G}(I)$ is a group (again, by Lemma B.2.3). \square

Appendix C

Appendix for Chapter 3

C.1 Proofs

C.1.1 Proof of Theorem 7

The proof of Theorem 7 uses the following lemmas.

Lemma C.1.1. *Under Assumptions 4.2.2 and 4.2.3, $\hat{\Sigma}^\dagger \hat{\Sigma} = \Sigma_0^\dagger \Sigma_0 + o_p(n^{-2^{-(L+1)}})$.*

Proof. By Wang et al. (2018, Corollary 8.1.4),

$$\frac{\|\hat{\Sigma}^\dagger - \Sigma_0^\dagger\|}{\|\Sigma_0^\dagger\|} \leq C \frac{\|\Sigma_0^\dagger\| \|\hat{\Sigma} - \Sigma_0\|}{1 - \|\Sigma_0^\dagger\| \|\hat{\Sigma} - \Sigma_0\|} = O_p(\|\hat{\Sigma} - \Sigma_0\|)$$

for some constant C . Therefore, it suffices to show $\hat{\Sigma} - \Sigma_0 = o_p(n^{-2^{-(L+1)}})$. Since $Pr(\|\tilde{\Sigma} - \Sigma_0\| < \kappa < \min\{\lambda_k : \lambda_k > 0\}) = 1 + o(1)$, in this proof, we assume $\|\tilde{\Sigma} - \Sigma_0\| < \kappa < \min\{\lambda_k : \lambda_k > 0\} - \|\tilde{\Sigma} - \Sigma_0\|$ without loss of generality. As long as $\|\tilde{\Sigma} - \Sigma_0\| \leq \kappa \leq \min\{\lambda_k : \lambda_k > 0\} - \|\tilde{\Sigma} - \Sigma_0\|$, by Weyl's inequality on the eigenvalue perturbations,

$$1\{\hat{\lambda}_k > \kappa\} = 1\{\lambda_k > 0\}$$

for every $k = 1, \dots, K$. It implies $\text{rank}(\hat{\Sigma}) = \text{rank}(\Sigma_0)$, and, by the Eckart-Young-Mirsky theorem,

$$\|\hat{\Sigma} - \tilde{\Sigma}\| \leq \|\Sigma_0 - \tilde{\Sigma}\|.$$

Therefore,

$$\|\hat{\Sigma} - \Sigma_0\| \leq \|\hat{\Sigma} - \tilde{\Sigma}\| + \|\tilde{\Sigma} - \Sigma_0\| \leq 2\|\tilde{\Sigma} - \Sigma_0\| = o_p(n^{-2^{-(L+1)}}).$$

□

Lemma C.1.2. *Suppose the assumptions in Theorem 7. There are an open convex neighborhood \mathcal{N} of θ_0 and a constant C such that*

$$\|\theta_1 - \theta_0\| \leq C \left\| Q_0^{(1)}(\theta_1) \right\|$$

for every $\theta_1 \in \mathcal{N}$.

Proof. By the second-order Taylor expansion, there is a constant C_1 such that such that, for every $\theta_1 \in \Theta$,

$$Q_0^{(1)}(\theta_1) = Q_0^{(1)}(\theta_0) + Q_0^{(2)}(\theta_0)(\theta_1 - \theta_0) + R(\theta_1)$$

with $\|R(\theta_1)\| \leq C_1 \|\theta_1 - \theta_0\|^2$. By the first-order condition $Q_0^{(1)}(\theta_0) = 0$, we have

$$Q_0^{(1)}(\theta_1) = Q_0^{(2)}(\theta_0)(\theta_1 - \theta_0) + R(\theta_1). \quad (\text{A-1})$$

Since $Q_0^{(2)}(\theta_0)$ is invertible,

$$\theta_1 - \theta_0 = Q_0^{(2)}(\theta_0)^{-1} Q_0^{(1)}(\theta_1) - Q_0^{(2)}(\theta_0)^{-1} R(\theta_1),$$

and therefore

$$\|\theta_1 - \theta_0\| \leq C_2 \left\| Q_0^{(1)}(\theta_1) \right\| + C_1 C_2 \|\theta_1 - \theta_0\|^2$$

for some constant C_2 . Let $\mathcal{N} = \{\theta_1 \in \Theta : C_1 C_2 \|\theta_1 - \theta_0\| < 1/2\}$. Then

$$\|\theta_1 - \theta_0\| \leq 2C_2 \left\| Q_0^{(1)}(\theta_1) \right\|$$

for every $\theta \in \mathcal{N}$. □

Lemma C.1.3. *Under the assumptions in Theorem 7,*

$$\hat{\theta}_{\text{inf}} = \theta_0 + o_p(n^{-2^{-(L+1)}}).$$

Proof. By Newey and McFadden (1994, Theorem 2.1), we have $\hat{\theta}_{\text{inf}} = \theta_0 + O_p(1)$. Since $\hat{Q}^{(1)}(\hat{\theta}_{\text{inf}}) = 0$, we have $Q_0^{(1)}(\hat{\theta}_{\text{inf}}) = Q_0^{(1)}(\hat{\theta}_{\text{inf}}) - \hat{Q}^{(1)}(\hat{\theta}_{\text{inf}})$. By Assumption 4.2.4, $Q_0^{(1)}(\hat{\theta}_{\text{inf}}) = o_p(n^{-2^{-(L+1)}})$. By Lemma C.1.2, the statement of this lemma holds. \square

Lemma C.1.4. *Suppose the assumptions in Theorem 7. There are an open convex neighborhood \mathcal{N} of θ_0 and a constant C such that*

$$\| [((I - \Sigma_0^\dagger \Sigma_0)\gamma)^T, \rho^T]^T - \theta_0 \| \leq C \left\| \frac{\partial}{\partial \theta} Q_0 \left([((I - \Sigma_0^\dagger \Sigma_0)\gamma)^T, \rho^T]^T \right) \right\|$$

for every $\theta \in \mathcal{N}$.

Proof. Let $\theta_2 = [((I - \Sigma_0^\dagger \Sigma_0)\gamma_2)^T, \rho_2^T]^T$. Since $\theta_0 = \begin{pmatrix} I - \Sigma_0^\dagger \Sigma_0 & O \\ O & I \end{pmatrix} \theta_0$, we can rewrite (A-1) as

$$Q_0^{(1)}(\theta_2) = Q_0^{(2)}(\theta_0) \begin{pmatrix} I - \Sigma_0^\dagger \Sigma_0 & O \\ O & I \end{pmatrix} (\theta_2 - \theta_0) + R(\theta_2).$$

Since

$$\left. \frac{\partial}{\partial \theta} Q_0 \left([((I - \Sigma_0^\dagger \Sigma_0)\gamma)^T, \rho^T]^T \right) \right|_{\theta=\theta_2} = \begin{pmatrix} I - \Sigma_0^\dagger \Sigma_0 & O \\ O & I \end{pmatrix} Q_0^{(1)}(\theta_2)$$

and $\begin{pmatrix} I - \Sigma_0^\dagger \Sigma_0 & O \\ O & I \end{pmatrix}$ is idempotent, we have

$$\begin{aligned} & (\theta_2 - \theta_0)^T \begin{pmatrix} I - \Sigma_0^\dagger \Sigma_0 & O \\ O & I \end{pmatrix} \left. \frac{\partial}{\partial \theta} Q_0 \left([((I - \Sigma_0^\dagger \Sigma_0)\gamma)^T, \rho^T]^T \right) \right|_{\theta=\theta_2} \\ &= (\theta_2 - \theta_0)^T \begin{pmatrix} I - \Sigma_0^\dagger \Sigma_0 & O \\ O & I \end{pmatrix} Q_0^{(2)}(\theta_0) \begin{pmatrix} I - \Sigma_0^\dagger \Sigma_0 & O \\ O & I \end{pmatrix} (\theta_2 - \theta_0) \\ &+ (\theta_2 - \theta_0)^T \begin{pmatrix} I - \Sigma_0^\dagger \Sigma_0 & O \\ O & I \end{pmatrix} R(\theta_2). \end{aligned}$$

Since the minimum eigenvalue of $Q_0^{(2)}(\theta_0)$ is positive, there is some positive constant C_1 such that

$$\begin{aligned} & \left\| \begin{pmatrix} I - \Sigma_0^\dagger \Sigma_0 & O \\ O & I \end{pmatrix} (\theta_2 - \theta_0) \right\| \left\| \frac{\partial}{\partial \theta} Q_0 \left([((I - \Sigma_0^\dagger \Sigma_0)\gamma)^T, \rho^T]^T \right) \Big|_{\theta=\theta_2} \right\| \\ & \geq C_1 \left\| \begin{pmatrix} I - \Sigma_0^\dagger \Sigma_0 & O \\ O & I \end{pmatrix} (\theta_2 - \theta_0) \right\|^2 - \left\| \begin{pmatrix} I - \Sigma_0^\dagger \Sigma_0 & O \\ O & I \end{pmatrix} (\theta_2 - \theta_0) \right\|^3. \end{aligned}$$

Let $\mathcal{N} = \left\{ \theta \in \Theta : \left\| \begin{pmatrix} I - \Sigma_0^\dagger \Sigma_0 & O \\ O & I \end{pmatrix} \theta - \theta_0 \right\| < 1/(2C_1) \right\}$. Then

$$\left\| \begin{pmatrix} I - \Sigma_0^\dagger \Sigma_0 & O \\ O & I \end{pmatrix} (\theta - \theta_0) \right\| \leq 2C_1 \left\| \frac{\partial}{\partial \theta} Q_0 \left([((I - \Sigma_0^\dagger \Sigma_0)\gamma)^T, \rho^T]^T \right) \right\|$$

for every $\theta \in \mathcal{N}$. □

Lemma C.1.5. *Under the assumptions in Theorem 7,*

$$\tilde{\theta} - \theta_0 = o_p(n^{-2-(L+1)}).$$

Proof. First, we are going to show $\tilde{\theta} - \theta_0 = o_p(1)$. By the compactness of Θ and the uniqueness of θ_0 , it suffices to show that $Q_0(\tilde{\theta}) - Q_0(\theta_0) = o_p(1)$. Note that $Q_0(\tilde{\theta}) - Q_0(\theta_0) \leq 0$ and that

$$\begin{aligned} Q_0(\tilde{\theta}) - Q_0(\theta_0) &= Q_0(\tilde{\theta}) - \hat{Q}(\tilde{\theta}) \\ &\quad + \hat{Q}(\tilde{\theta}) - \hat{Q}([((I - \hat{\Sigma}^\dagger \hat{\Sigma})\gamma_0)^T, \rho_0^T]^T) \\ &\quad + \hat{Q}([((I - \hat{\Sigma}^\dagger \hat{\Sigma})\gamma_0)^T, \rho_0^T]^T) - \hat{Q}([((I - \Sigma_0^\dagger \Sigma_0)\gamma_0)^T, \rho_0^T]^T) \\ &\quad + \hat{Q}(\theta_0) - Q_0(\theta_0) \\ &\geq -2 \sup_{\theta \in \Theta} |\hat{Q}(\theta) - Q_0(\theta)| \\ &\quad + \hat{Q}([((I - \hat{\Sigma}^\dagger \hat{\Sigma})\gamma_0)^T, \rho_0^T]^T) - \hat{Q}([((I - \Sigma_0^\dagger \Sigma_0)\gamma_0)^T, \rho_0^T]^T), \end{aligned}$$

where the equality uses $\gamma_0 = (I - \Sigma_0^\dagger \Sigma_0)\gamma_0$ and the inequality uses

$$\hat{Q}(\tilde{\theta}) = \max_{\theta \in \Theta} \hat{Q}(\left[((I - \hat{\Sigma}^\dagger \hat{\Sigma})\gamma)^T, \rho^T \right]^T) \geq \hat{Q}(\left[((I - \hat{\Sigma}^\dagger \hat{\Sigma})\gamma_0)^T, \rho_0^T \right]^T).$$

Since the mean-value expansion yields

$$\begin{aligned} & \left| \hat{Q}(\left[((I - \hat{\Sigma}^\dagger \hat{\Sigma})\gamma_0)^T, \rho_0^T \right]^T) - \hat{Q}(\left[((I - \Sigma_0^\dagger \Sigma_0)\gamma_0)^T, \rho_0^T \right]^T) \right| \\ & \leq \sup_{\theta \in \Theta} \left\| \hat{Q}^{(1)}(\theta) \right\| \left\| \hat{\Sigma}^\dagger \hat{\Sigma} - \Sigma_0^\dagger \Sigma_0 \right\| \|\gamma_0\|, \end{aligned}$$

Lemma C.1.1 implies $Q_0(\tilde{\theta}) - Q_0(\theta_0) = o_p(1)$.

Second, we are going to show $\tilde{\theta} - \theta_0 = o_p(n^{-2^{-(L+1)}})$. By Lemmas C.1.1 and C.1.4, it suffices to show

$$\left. \frac{\partial}{\partial \theta} Q_0 \left(\left[((I - \Sigma_0^\dagger \Sigma_0)\gamma)^T, \rho^T \right]^T \right) \right|_{\theta=\tilde{\theta}} = o_p(n^{-2^{-(L+1)}}).$$

Some calculus operations yield

$$\begin{aligned} & \left\| \frac{\partial}{\partial \theta} Q_0 \left(\left[((I - \Sigma_0^\dagger \Sigma_0)\gamma)^T, \rho^T \right]^T \right) \right\|_{\theta=\tilde{\theta}} \\ & \leq \left\| \frac{\partial}{\partial \theta} \hat{Q} \left(\left[((I - \hat{\Sigma}^\dagger \hat{\Sigma})\gamma)^T, \rho^T \right]^T \right) \right\|_{\theta=\tilde{\theta}} \\ & \quad + \left\| \begin{pmatrix} I - \hat{\Sigma}^\dagger \hat{\Sigma} & O \\ O & I \end{pmatrix} \left(Q_0^{(1)}(\left[((I - \hat{\Sigma}^\dagger \hat{\Sigma})\tilde{\gamma})^T, \tilde{\rho}^T \right]^T) - \hat{Q}^{(1)}(\left[((I - \hat{\Sigma}^\dagger \hat{\Sigma})\tilde{\gamma})^T, \tilde{\rho}^T \right]^T) \right) \right\| \\ & \quad + \left\| \begin{pmatrix} I - \hat{\Sigma}^\dagger \hat{\Sigma} & O \\ O & I \end{pmatrix} \left(Q_0^{(1)}(\left[((I - \Sigma_0^\dagger \Sigma_0)\tilde{\gamma})^T, \tilde{\rho}^T \right]^T) - Q_0^{(1)}(\left[((I - \hat{\Sigma}^\dagger \hat{\Sigma})\tilde{\gamma})^T, \tilde{\rho}^T \right]^T) \right) \right\| \\ & \quad + \left\| \begin{pmatrix} \Sigma_0^\dagger \Sigma_0 - \hat{\Sigma}^\dagger \hat{\Sigma} & O \\ O & O \end{pmatrix} Q_0^{(1)}(\left[((I - \Sigma_0^\dagger \Sigma_0)\tilde{\gamma})^T, \tilde{\rho}^T \right]^T) \right\|. \end{aligned}$$

By the definition of $\tilde{\theta}$, we have

$$\left. \frac{\partial}{\partial \theta} \hat{Q} \left(\left[((I - \hat{\Sigma}^\dagger \hat{\Sigma})\gamma)^T, \rho^T \right]^T \right) \right|_{\theta=\tilde{\theta}} = 0.$$

By Assumption 4.2.4 and Lemma C.1.1, we have

$$\frac{\partial}{\partial \theta} Q_0 \left(\left[((I - \Sigma_0^\dagger \Sigma_0) \gamma)^T, \rho^T \right]^T \right) \Big|_{\theta = \tilde{\theta}} = o_p(n^{-2^{-(L+1)}}).$$

□

Proof of Theorem 7. By Lemmas C.1.3 and C.1.5, $\tilde{\theta} - \theta_0 = o_p(1)$ and $\|\tilde{\theta} - \hat{\theta}_{\text{inf}}\|^{2L} = o_p(n^{-1/2})$. Thus the statement of this theorem follows from Robinson (1988, Theorem 2). Assumption A1 in Robinson 1988 follows from Assumption 4.2.4 and Lemma C.1.3. Assumption A3 in Robinson 1988 follows from Assumption 4.2.4. □

C.1.2 Proof of Theorem 8

Proof. Let $R_A = \dim(Z_A) - \text{rank}(\text{Var}([\gamma_0, \delta_0]^T [Z_A^T, 0^T]^T))$. There are a $(J + 1) \times \text{rank}(\text{Var}([\gamma_0, \delta_0]^T [Z_A^T, 0^T]^T))$ matrix \mathbf{M}_1 and a $\text{rank}(\text{Var}([\gamma_0, \delta_0]^T [Z_A^T, 0^T]^T)) \times \dim(Z_A)$ matrix \mathbf{M}_2 such that

$$[\gamma_0, \delta_0]^T [Z_A^T, 0^T]^T = \mathbf{M}_1 \mathbf{M}_2 Z_A \text{ almost surely.}$$

Let $\bar{\nu}_1, \dots, \bar{\nu}_{R_A}$ be R_A linearly independent vectors in the image of

$$\begin{pmatrix} I - \mathbf{M}_2^\dagger \mathbf{M}_2 & O \\ O & O \end{pmatrix}.$$

Note that $[\gamma_0, \delta_0]^T \bar{\nu}_r = 0$ for every $r = 1, \dots, R_A$. By Theorem 6, it suffices to show that, even if $[\gamma_0, \delta_0]^T (Z_1 - Z_2) = 0$, there is a non-zero variation in $\bar{\nu}_r^T (Z_1 - Z_2)$ for every $r = 1, \dots, R_A$. Consider the point z in the assumption of Theorem 8. Since z_A is an interior point, there is a positive constant c such that $[z_A^T, z_B^T]^T + c\bar{\nu}_r$ belongs to the support of Z . Define $z_1 = z$ and $z_2 = z + c\bar{\nu}_r$. This z_2 and z_1 are support points of Z such that $[\gamma_0, \delta_0]^T (z_2 - z_1) = 0$ and $\bar{\nu}_r^T (z_2 - z_1) = c\bar{\nu}_r^T \bar{\nu}_r \neq 0$. □

C.1.3 Proof of Theorem 9

Proof. We use R_A and $(\bar{\nu}_1, \dots, \bar{\nu}_{R_A})$ in the proof of Theorem 8. Let $R_B = \text{rank}(\text{Var}(Z_B))$. There are linearly independent vectors $\bar{\nu}_{R_A+1}, \dots, \bar{\nu}_{R_A+R_B}$ in the support of $[0^T, (Z_{2,B} - Z_{1,B})^T]^T$. Note that $\bar{\nu}_1, \dots, \bar{\nu}_{R_A+R_B}$ are linearly independent. By Theorem 6, it suffices to show that, even if $[\gamma_0, \delta_0]^T(Z_1 - Z_2) = 0$, there is a non-zero variation in $\bar{\nu}_r^T(Z_1 - Z_2)$ for every $r = 1, \dots, R_A + R_B$. The proof for $r = 1, \dots, R_A$ is the same as in the proof of Theorem 8. Consider $r = R_A + 1, \dots, R_A + R_B$. There are $z_{1,B}$ and $z_{2,B}$ in the support of Z_B such that

$$[0^T, (z_{1,B} - z_{2,B})^T]^T = \bar{\nu}_r.$$

Let $z_{1,A}$ be any point such that $[z_{1,A}^T, z_{1,B}^T]^T$ is in the support of Z . By the assumption of this theorem, we can find a point $z_{2,A}$ such that

$$[\gamma_0, \delta_0]^T [z_{2,A}^T, z_{2,B}^T]^T = [\gamma_0, \delta_0]^T [z_{1,A}^T, z_{1,B}^T]^T$$

and $[z_{2,A}^T, z_{2,B}^T]^T$ is in the support of Z . Define $z_1 = [z_{1,A}^T, z_{1,B}^T]^T$ and $z_2 = [z_{2,A}^T, z_{2,B}^T]^T$. This z_2 and z_1 are support points of Z such that $[\gamma_0, \delta_0]^T(z_2 - z_1) = 0$ and $\bar{\nu}_r^T(z_2 - z_1) = \bar{\nu}_r^T \bar{\nu}_r \neq 0$. \square

C.1.4 Proof of Theorem 10

We use the following lemmas to prove Theorem 10. Define

$$\hat{W}_{i_1 i_2}^N \equiv T^{-2} \sum_{t_1, t_2} \frac{1}{h^{\dim(U_{it})}} \mathbf{K} \left(\hat{\zeta}_{i_1 t_1 i_2 t_2} / h \right) (Z_{i_1 t_1} - Z_{i_2 t_2})(Z_{i_1 t_1} - Z_{i_2 t_2})^T,$$

$$\hat{W}_{i_1 i_2}^D \equiv T^{-2} \sum_{t_1, t_2} \frac{1}{h^{\dim(U_{it})}} \mathbf{K} \left(\hat{\zeta}_{i_1 t_1 i_2 t_2} / h \right),$$

and

$$W_{i_1 i_2}^N \equiv T^{-2} \sum_{t_1, t_2} \frac{1}{h^{\dim(U_{it})}} \mathbf{K} (\zeta_{i_1 t_1 i_2 t_2} / h) (Z_{i_1 t_1} - Z_{i_2 t_2})(Z_{i_1 t_1} - Z_{i_2 t_2})^T.$$

Lemma C.1.6. Under the assumptions in Theorem 10,

$$\frac{1}{n(n-1)} \sum_{(i_1, i_2): i_1 \neq i_2} \left(\hat{W}_{i_1 i_2}^N - W_{i_1 i_2}^N \right) = o_p(n^{-2-(L+1)}).$$

Proof. Since the probability of (4.5) is approaching one, we can assume (4.5) without loss of generality. Thus

$$\|\hat{\zeta}_{i_1 t_1 i_2 t_2}\| \geq n^{-\tau} + \log(n)h \implies \|\hat{\zeta}_{i_1 t_1 i_2 t_2}\| \geq \log(n)h$$

and therefore

$$\begin{aligned} \left| \mathbf{K} \left(\hat{\zeta}_{i_1 t_1 i_2 t_2} / h \right) - \mathbf{K} \left(\zeta_{i_1 t_1 i_2 t_2} / h \right) \right| &= 1 \{ \|\zeta_{i_1 t_1 i_2 t_2}\| \\ &\leq n^{-\tau} + \log(n)h \} \left| \mathbf{K} \left(\hat{\zeta}_{i_1 t_1 i_2 t_2} / h \right) - \mathbf{K} \left(\zeta_{i_1 t_1 i_2 t_2} / h \right) \right| \end{aligned}$$

for sufficiently large n . By the second-order Taylor expansion, there is some constant C such that

$$\mathbf{K} \left(\hat{\zeta}_{i_1 t_1 i_2 t_2} / h \right) - \mathbf{K} \left(\zeta_{i_1 t_1 i_2 t_2} / h \right) = \frac{1}{h} \mathbf{K}^{(1)} \left(\zeta_{i_1 t_1 i_2 t_2} / h \right) \left(\hat{\zeta}_{i_1 t_1 i_2 t_2} - \zeta_{i_1 t_1 i_2 t_2} \right) + \frac{1}{h^2} R_{i_1 t_1 i_2 t_2}$$

with $\|R_{i_1 t_1 i_2 t_2}\| \leq C \left\| \hat{\zeta}_{i_1 t_1 i_2 t_2} - \zeta_{i_1 t_1 i_2 t_2} \right\|^2$. Therefore,

$$\begin{aligned} \left| \mathbf{K} \left(\hat{\zeta}_{i_1 t_1 i_2 t_2} / h \right) - \mathbf{K} \left(\zeta_{i_1 t_1 i_2 t_2} / h \right) \right| &\leq \frac{1}{h} \left\| \mathbf{K}^{(1)} \left(\zeta_{i_1 t_1 i_2 t_2} / h \right) \right\| \left\| \hat{\zeta}_{i_1 t_1 i_2 t_2} - \zeta_{i_1 t_1 i_2 t_2} \right\| \\ &\quad + \frac{1}{h^2} 1 \{ \|\zeta_{i_1 t_1 i_2 t_2}\| \leq n^{-\tau} + \log(n)h \} \|R_{i_1 t_1 i_2 t_2}\|. \end{aligned}$$

Since

$$\left\| \hat{W}_{i_1 i_2}^N - W_{i_1 i_2}^N \right\| \leq T^{-2} \sum_{t_1, t_2} \frac{1}{h^{\dim(U_{it})}} \left| \mathbf{K} \left(\hat{\zeta}_{i_1 t_1 i_2 t_2} / h \right) - \mathbf{K} \left(\zeta_{i_1 t_1 i_2 t_2} / h \right) \right| \|Z_{i_1 t_1} - Z_{i_2 t_2}\|^2,$$

we have

$$\begin{aligned}
& \left\| \frac{1}{n(n-1)} \sum_{(i_1, i_2): i_1 \neq i_2} \left(\hat{W}_{i_1 i_2}^N - W_{i_1 i_2}^N \right) \right\| \\
& \leq \frac{1}{n(n-1)} \sum_{(i_1, i_2): i_1 \neq i_2} T^{-2} \sum_{t_1, t_2} \frac{1}{h^{\dim(U_{it})+1}} \left\| \mathbf{K}^{(1)}(\zeta_{i_1 t_1 i_2 t_2} / h) \right\| \\
& \quad \times \|Z_{i_1 t_1} - Z_{i_2 t_2}\|^2 \left\| \hat{\zeta}_{i_1 t_1 i_2 t_2} - \zeta_{i_1 t_1 i_2 t_2} \right\| \\
& \quad + \frac{1}{n(n-1)} \sum_{(i_1, i_2): i_1 \neq i_2} T^{-2} \sum_{t_1, t_2} \frac{1}{h^{\dim(U_{it})+2}} \\
& \quad \times 1\{\|\zeta_{i_1 t_1 i_2 t_2}\| \leq n^{-\tau} + \log(n)h\} \|Z_{i_1 t_1} - Z_{i_2 t_2}\|^2 \|R_{i_1 t_1 i_2 t_2}\| \\
& \leq \mathcal{U}_1 \frac{1}{h} \sup_{(i_1, t_1, i_2, t_2): i_1 \neq i_2} \left\| \hat{\zeta}_{i_1 t_1 i_2 t_2} - \zeta_{i_1 t_1 i_2 t_2} \right\| \\
& \quad + C\mathcal{U}_2 \frac{(n^{-\tau} + \log(n)h)^{\dim(U_{it})}}{h^{\dim(U_{it})+2}} \sup_{(i_1, t_1, i_2, t_2): i_1 \neq i_2} \left\| \hat{\zeta}_{i_1 t_1 i_2 t_2} - \zeta_{i_1 t_1 i_2 t_2} \right\|^2 \\
& \leq \mathcal{U}_1 \frac{n^{-\tau}}{h} + C\mathcal{U}_2 \frac{n^{-2\tau}(n^{-\tau} + \log(n)h)^{\dim(U_{it})}}{h^{\dim(U_{it})+2}},
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{U}_1 & \equiv \frac{1}{n(n-1)} \sum_{(i_1, i_2): i_1 \neq i_2} T^{-2} \sum_{t_1, t_2} \frac{1}{h^{\dim(U_{it})}} \left\| \mathbf{K}^{(1)}(\zeta_{i_1 t_1 i_2 t_2} / h) \right\| \|Z_{i_1 t_1} - Z_{i_2 t_2}\|^2 \\
\mathcal{U}_2 & \equiv \frac{1}{n(n-1)} \sum_{(i_1, i_2): i_1 \neq i_2} T^{-2} \sum_{t_1, t_2} \frac{1}{(n^{-\tau} + \log(n)h)^{\dim(U_{it})}} \\
& \quad \times 1\{\|\zeta_{i_1 t_1 i_2 t_2}\| \leq n^{-\tau} + \log(n)h\} \|Z_{i_1 t_1} - Z_{i_2 t_2}\|^2.
\end{aligned}$$

By Assumption 4.4.2, it suffices to show $\mathcal{U}_1 = O_p(1)$ and $\mathcal{U}_2 = O_p(1)$. Note that

$$\begin{aligned}
E[\|\mathcal{U}_1\|] & \leq \frac{1}{n(n-1)} \sum_{(i_1, i_2): i_1 \neq i_2} T^{-2} \sum_{t_1, t_2} E\left[\frac{1}{h^{\dim(U_{it})}} \left\| \mathbf{K}^{(1)}(\zeta_{i_1 t_1 i_2 t_2} / h) \right\| \right. \\
& \quad \left. \times E[\|Z_{i_1 t_1} - Z_{i_2 t_2}\|^2 \mid \zeta_{i_1 t_1 i_2 t_2}] \right] \\
& \leq CE \left[\frac{1}{h^{\dim(U_{it})}} \left\| \mathbf{K}^{(1)}([(U_1 - U_2)^T, (V_1 - V_2)^T]^T / h) \right\| \right]
\end{aligned}$$

for some constant C . For sufficiently small h ,

$$E[|\mathcal{U}_1|] \leq CE \left[\frac{1}{h^{\dim(U_{it})}} \|\mathbf{K}^{(1)}([(U_1 - U_2)^T/h, 0^T]^T)\| \right].$$

Using the change of variable,

$$\begin{aligned} E[|\mathcal{U}_1|] &\leq C \int \frac{1}{h^{\dim(U_{it})}} \|\mathbf{K}^{(1)}([u^T/h, 0^T]^T)\| f_{U_1-U_2}(u) du \\ &= C \int \|\mathbf{K}^{(1)}([u^T, 0^T]^T)\| f_{U_1-U_2}(uh) du \\ &= O(1). \end{aligned}$$

Similarly, we can show $\mathcal{U}_2 = O_p(1)$. □

Lemma C.1.7. *Under the assumptions in Theorem 10, $\frac{1}{n(n-1)} \sum_{(i_1, i_2): i_1 \neq i_2} (W_{i_1 i_2}^N - E[W_{i_1 i_2}^N]) = o_p(n^{-2-(L+1)})$.*

Proof. Based on the variance formula for U-statistics, it suffices to show

$$\text{Var} \left(\frac{1}{h^{\dim(U_{it})}} \mathbf{K}(\zeta_{i_1 t_1 i_2 t_2}/h) \|Z_{i_1 t_1} - Z_{i_2 t_2}\|^2 \right) = o(n^{2-2^{-L}})$$

and

$$\text{Var} \left(E \left[\frac{1}{h^{\dim(U_{it})}} \mathbf{K}(\zeta_{i_1 t_1 i_2 t_2}/h) \|Z_{i_1 t_1} - Z_{i_2 t_2}\|^2 \mid Z_{i_1 t_1} \right] \right) = o(n^{1-2^{-L}}).$$

First, we are going to show $\text{Var} \left(\frac{1}{h^{\dim(U_{it})}} \mathbf{K}(\zeta_{i_1 t_1 i_2 t_2}/h) \|Z_{i_1 t_1} - Z_{i_2 t_2}\|^2 \right) = o(n^{2-2^{-L}})$. Note that

$$\begin{aligned} &\text{Var} \left(\frac{1}{h^{\dim(U_{it})}} \mathbf{K}(\zeta_{i_1 t_1 i_2 t_2}/h) \|Z_{i_1 t_1} - Z_{i_2 t_2}\|^2 \right) \\ &\leq E \left[\frac{1}{h^{2 \dim(U_{it})}} \mathbf{K}(\zeta_{i_1 t_1 i_2 t_2}/h)^2 \|Z_{i_1 t_1} - Z_{i_2 t_2}\|^4 \right] \\ &= O(1) E \left[\frac{1}{h^{2 \dim(U_{it})}} \mathbf{K}(\zeta_{i_1 t_1 i_2 t_2}/h)^2 \right]. \end{aligned}$$

For sufficiently small h ,

$$\begin{aligned} \text{Var} \left(\frac{1}{h^{\dim(U_{it})}} \mathbf{K}(\zeta_{i_1 t_1 i_2 t_2} / h) \|Z_{i_1 t_1} - Z_{i_2 t_2}\|^2 \right) \\ \leq O(1) E \left[\frac{1}{h^{2 \dim(U_{it})}} \mathbf{K} \left([(U_{i_1 t_1} - U_{i_2 t_2})^T / h, 0^T]^T \right)^2 \right] \end{aligned}$$

Using a change of variables,

$$\begin{aligned} & \text{Var} \left(\frac{1}{h^{\dim(U_{it})}} \mathbf{K}(\zeta_{i_1 t_1 i_2 t_2} / h) \|Z_{i_1 t_1} - Z_{i_2 t_2}\|^2 \right) \\ &= O(1) \int \frac{1}{h^{2 \dim(U_{it})}} \mathbf{K} \left([u^T / h, 0^T]^T \right)^2 f_{U_1 - U_2}(u) du \\ &= O(1) \int \frac{1}{h^{\dim(U_{it})}} \mathbf{K} \left([u^T, 0^T]^T \right)^2 f_{U_1 - U_2}(uh) du \\ &= O(h^{-(\dim(U_{it}))}) \\ &= o(n^{2-2^{-L}}). \end{aligned}$$

Second, we are going to show $\text{Var} \left(E \left[\frac{1}{h^{\dim(U_{it})}} \mathbf{K}(\zeta_{i_1 t_1 i_2 t_2} / h) \|Z_{i_1 t_1} - Z_{i_2 t_2}\|^2 \mid Z_{i_1 t_1} \right] \right) = o(n^{1-2^{-L}})$. It suffices to show that

$$\begin{aligned} & E \left[\frac{1}{h^{\dim(U_{it})}} \mathbf{K}(\zeta_{i_1 t_1 i_2 t_2} / h) ((Z_{i_1 t_1} - Z_{i_2 t_2})^T \omega)^2 \mid Z_{i_1 t_1} \right] \\ & \leq \int \mathbf{K} \left([u^T, 0^T]^T \right) f_{U_{i_1 t_1} - U_{i_2 t_2} \mid Z_{i_1 t_1}}(uh) du (C_0 + C_1 \|Z_{i_1 t_1}\| + C_2 \|Z_{i_1 t_1}\|^2) \end{aligned}$$

for some constants C_0, C_1, C_2 . For sufficiently small h ,

$$\begin{aligned} & E \left[\frac{1}{h^{\dim(U_{it})}} \mathbf{K}(\zeta_{i_1 t_1 i_2 t_2} / h) ((Z_{i_1 t_1} - Z_{i_2 t_2})^T \omega)^2 \mid Z_{i_1 t_1} \right] \\ & \leq E \left[\frac{1}{h^{\dim(U_{it})}} \mathbf{K} \left([(U_{i_1 t_1} - U_{i_2 t_2})^T / h, 0^T]^T \right) ((Z_{i_1 t_1} - Z_{i_2 t_2})^T \omega)^2 \mid Z_{i_1 t_1} \right]. \end{aligned}$$

Since $E[\|Z_{i_2 t_2}\|^2 \mid U_{i_1 t_1} - U_{i_2 t_2}, Z_{i_1 t_1}]$ and $E[\|Z_{i_2 t_2}\|^2 \mid U_{i_1 t_1} - U_{i_2 t_2}, Z_{i_1 t_1}]$ are

bounded, there are some constants C_0, C_1, C_2 such that

$$\begin{aligned} & E \left[\frac{1}{h^{\dim(U_{it})}} \mathbf{K}(\zeta_{i_1 t_1 i_2 t_2} / h) ((Z_{i_1 t_1} - Z_{i_2 t_2})^T \omega)^2 \mid Z_{i_1 t_1} \right] \\ & \leq E \left[\frac{1}{h^{\dim(U_{it})}} \mathbf{K} \left([(U_{i_1 t_1} - U_{i_2 t_2})^T / h, 0^T]^T \right) (C_0 + C_1 \|Z_{i_1 t_1}\| + C_2 \|Z_{i_1 t_1}\|^2) \mid Z_{i_1 t_1} \right] \end{aligned}$$

Using a change of variables,

$$\begin{aligned} & E \left[\frac{1}{h^{\dim(U_{it})}} \mathbf{K}(\zeta_{i_1 t_1 i_2 t_2} / h) ((Z_{i_1 t_1} - Z_{i_2 t_2})^T \omega)^2 \mid Z_{i_1 t_1} \right] \\ & \leq \int \frac{1}{h^{\dim(U_{it})}} \mathbf{K} \left([u^T / h, 0^T]^T \right) f_{U_{i_1 t_1} - U_{i_2 t_2} \mid Z_{i_1 t_1}}(u) du (C_0 + C_1 \|Z_{i_1 t_1}\| + C_2 \|Z_{i_1 t_1}\|^2) \\ & = \int \mathbf{K} \left([u^T, 0^T]^T \right) f_{U_{i_1 t_1} - U_{i_2 t_2} \mid Z_{i_1 t_1}}(uh) du (C_0 + C_1 \|Z_{i_1 t_1}\| + C_2 \|Z_{i_1 t_1}\|^2). \end{aligned}$$

□

Lemma C.1.8. *Under the assumptions in Theorem 10, $\frac{1}{n(n-1)} \sum_{(i_1, i_2): i_1 \neq i_2} E[W_{i_1 i_2}^N] = Pr(V_1 = V_2) \Xi(0) + o(n^{-2-(L+1)})$.*

Proof. Note that

$$\begin{aligned} E[W_{i_1 i_2}^N] & = E \left[\frac{1}{h^{\dim(U_{it})}} \mathbf{K} \left([(U_1 - U_2)^T, (V_1 - V_2)^T]^T / h \right) (Z_1 - Z_2)(Z_1 - Z_2)^T \right] \\ & = E \left[\frac{1}{h^{\dim(U_{it})}} \mathbf{K} \left([(U_1 - U_2)^T, (V_1 - V_2)^T]^T / h \right) \Xi(U_1 - U_2) \right]. \end{aligned}$$

By Assumption 4.4.1, for sufficiently small h , we have

$$E[W_{i_1 i_2}^N] = Pr(V_1 = V_2) E \left[\frac{1}{h^{\dim(U_{it})}} \mathbf{K} \left([(U_1 - U_2)^T / h, 0^T]^T \right) \Xi(U_1 - U_2) \mid V_1 = V_2 \right].$$

Using a change of variables, we have

$$\begin{aligned} E[W_{i_1 i_2}^N] & = Pr(V_1 = V_2) \int \frac{1}{h^{\dim(U_{it})}} \mathbf{K} \left([u^T / h, 0^T]^T \right) \Xi(u) du \\ & = Pr(V_1 = V_2) \int \mathbf{K} \left([u^T, 0^T]^T \right) \Xi(uh) du. \end{aligned}$$

By Assumptions 4.4.1 and 4.4.3,

$$E[W_{i_1 i_2}^N] = Pr(V_1 = V_2)\Xi(0) + O(h^2).$$

□

Proof of Theorem 10. By Lemmas C.1.6, C.1.7, and C.1.8,

$$\frac{1}{n(n-1)} \sum_{(i_1, i_2): i_1 \neq i_2} \hat{W}_{i_1 i_2}^N = Pr(V_1 = V_2)\Xi(0) + o_p(n^{-2-(L+1)}).$$

For the denominator, in a similar fashion, we can show that

$$\frac{1}{n(n-1)} \sum_{(i_1, i_2): i_1 \neq i_2} \hat{W}_{i_1 i_2}^D = Pr(V_1 = V_2) + o_p(n^{-2-(L+1)}).$$

Combining these two statements, we can establish the statement of this theorem. □

Bibliography

- Aguirregabiria, V. and Mira, P. (July 2002a). “Swapping the Nested Fixed Point Algorithm: A Class of Estimators for Discrete Markov Decision Models”. *Econometrica* 70.2, pp. 1519–1543.
- (Jan. 2007a). “Sequential Estimation of Dynamic Discrete Games”. *Econometrica* 75.1, pp. 1–53.
- Aguirregabiria, V., Gu, J., and Luo, Y. (2020). “Sufficient statistics for unobserved heterogeneity in structural dynamic logit models”. *Journal of Econometrics*.
- Aguirregabiria, V. and Magesan, A. (2020). “Identification and estimation of dynamic games when players’ beliefs are not in equilibrium”. *The Review of Economic Studies* 87.2, pp. 582–625.
- Aguirregabiria, V. and Mira, P. (2002b). “Swapping the nested fixed point algorithm: A class of estimators for discrete Markov decision models”. *Econometrica* 70.4, pp. 1519–1543.
- (2007b). “Sequential estimation of dynamic discrete games”. *Econometrica* 75.1, pp. 1–53.
- (2010). “Dynamic discrete choice structural models: A survey”. *Journal of Econometrics* 156.1, pp. 38–67.
- Ahn, H., Ichimura, H., Powell, J. L., and Ruud, P. A. (2018). “Simple estimators for invertible index models”. *Journal of Business & Economic Statistics* 36.1, pp. 1–10.
- 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 Miller, R. A. (Nov. 2011a). “Conditional Choice Probability Estimation of Dynamic Discrete Choice Models with Unobserved Heterogeneity”. *Econometrica* 79.6, pp. 1823–1867.
- 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. (2011b). “Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity”. *Econometrica* 79.6, pp. 1823–1867.
- (2020). “Identifying dynamic discrete choice models off short panels”. *Journal of Econometrics* 215.2, pp. 473–485.

- Bajari, P., Benkard, D., and Levin, J. (Sept. 2007). “Estimating Dynamic Models of Imperfect Competition”. *Econometrica* 75.5, pp. 1331–1370.
- 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.
- Besag, J. and Mondal, D. (2013). “Exact Goodness-of-Fit Tests for Markov Chains”. *Biometrics* 69.2, pp. 488–496.
- 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.
- 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, Q. and Fang, Z. (2019). “Improved inference on the rank of a matrix”. *Quantitative Economics* 10.4, pp. 1787–1824.
- 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.
- Collard-Wexler, A. (2013). “Demand fluctuations in the ready-mix concrete industry”. *Econometrica* 81.3, pp. 1003–1037.
- de Paula, A. and Tang, X. (Jan. 2012). “Inference of Signs of Interaction Effects in Simultaneous Games With Incomplete Information”. *Econometrica* 80.1, pp. 143–172.
- Dunne, T., Klimek, S. D., Roberts, M. J., and Xu, D. Y. (2013). “Entry, exit, and the determinants of market structure”. *The RAND Journal of Economics* 44.3, pp. 462–487.
- 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.

- Ericson, R. and Pakes, A. (Jan. 1995). “Markov-Perfect Industry Dynamics: A Framework for Empirical Work”. *The Review of Economic Studies* 62.1, pp. 53–82.
- 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.
- Hahn, J., Liao, Z., and Ridder, G. (2018). “Nonparametric two-step sieve M estimation and inference”. *Econometric Theory* 34.6, pp. 1281–1324.
- 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.
- 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.
- Horowitz, J. L. and Härdle, W. (1996). “Direct semiparametric estimation of single-index models with discrete covariates”. *Journal of the American Statistical Association* 91.436, pp. 1632–1640.
- Hotz, J. V. and Miller, R. T. A. (July 1993a). “Conditional Choice Probabilities and the Estimation of Dynamic Models”. *Review of Economics Studies* 60.3, pp. 497–529.
- Hotz, J. V., Miller, R. T. A., Sanders, S., and Smith, J. (Apr. 1994). “A Simulation Estimator for Dynamic Models of Discrete Choice”. *Review of Economics Studies* 61.2, pp. 265–289.
- Hotz, V. J. and Miller, R. A. (1993b). “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.
- Kandel, D., Yossi, M., Unger, R., and Winkler, P. (1996). “Shuffling biological sequences”. *Discrete Applied Mathematics* 71.1-3, pp. 171–185.
- 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.
- (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.
- Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypothesis: Third edition*. Springer.

- 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.
- Maskin, E. and Tirole, J. (Oct. 2001). “Markov Perfect Equilibrium I: Observable Actions”. *Journal of Economic Theory* 100.2, pp. 191–219.
- 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.
- Nevo, A., Turner, J. L., and Williams, J. W. (2016). “Usage-based pricing and demand for residential broadband”. *Econometrica* 84.2, pp. 411–443.
- Newey, W. K. and McFadden, D. (1994). “Large sample estimation and hypothesis testing”. *Handbook of econometrics* 4, pp. 2111–2245.
- 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.
- Otsu, T., Pesendorfer, M., and Takahashi, Y. (2016). “Pooling data across markets in dynamic Markov Games”. *Quantitative Economics* 7.2, pp. 523–559.
- Pakes, A., Ostrovsky, M., and Berry, S. (Summer 2007). “Simple estimators for the parameters of discrete dynamic games (with entry/exit examples)”. *RAND Journal of Economics* 38.2, pp. 373–399.
- Pesendorfer, M. and Schmidt-Dengler, P. (Sept. 2008a). “Asymptotic Least Squares Estimators for Dynamic Games”. *Review of Economic Studies* 75.3, pp. 901–928.
- (Mar. 2010). “Sequential Estimation of Dynamic Discrete Games: A Comment”. *Econometrica* 78.2, pp. 833–842.
- Pesendorfer, M. and Schmidt-Dengler, P. (2008b). “Asymptotic least squares estimators for dynamic games”. *The Review of Economic Studies* 75.3, pp. 901–928.
- Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer Texts in Statistics.
- Robinson, P. M. (1988). “The stochastic difference between econometric statistics”. *Econometrica: Journal of the Econometric Society*, pp. 531–548.

- Rudin, W. (1964). *Principles of mathematical analysis*. Vol. 3. McGraw-hill New York.
- Rust, J. (Sept. 1987). “Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher”. *Econometrica* 55 (1), pp. 999–1033.
- Rust, J. (1988). “Maximum likelihood estimation of discrete control processes”. *SIAM journal on control and optimization* 26.5, pp. 1006–1024.
- (1994). “Structural estimation of Markov decision processes”. *Handbook of econometrics* 4.4, pp. 3081–3143.
- (2008). “Dynamic programming”. *The New Palgrave Dictionary of Economics* 1, p. 8.
- Ryan, S. (2012). “The costs of environmental regulation in a concentrated industry”. *Econometrica* 80.3, pp. 1019–1061.
- 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.
- Sweeting, A. (2013). “Dynamic product positioning in differentiated product markets: The effect of fees for musical performance rights on the commercial radio industry”. *Econometrica* 81.5, pp. 1763–1803.
- 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.
- Wang, G., Wei, Y., Qiao, S., Lin, P., and Chen, Y. (2018). *Generalized inverses: theory and computations*. Vol. 53. Springer.
- Williams, B. (2019). “Nonparametric identification of discrete choice models with lagged dependent variables”. *Journal of Econometrics*.