A Contraction Fixed Point Method for Infinite Mixture Models and Direct Counterfactual Analysis

Yuya Sasaki*

Johns Hopkins University

November 16, 2015

Abstract

For infinitely mixed moment equality models, this paper proposes a θ-dependent self map, whose fixed point exists if and only if θ belongs to the identified set. Its contraction property in large sample is discussed. This method provides a computationally attractive way to implement the ELVIS, especially when the number of moment equalities is not small as is often the case with panel data. Applying this method to the PSID, I directly obtain set estimates for the average counterfactual effects of an exogenous birth and an exogenous child on female labor supply, accounting for multi-dimensional continuous heterogeneity, state-dependence, and endogenous fertility decisions in observed data.

Keywords: contraction, counterfactual, female labor supply, fixed point, infinite mixture

*Email: sasaki@jhu.edu. Address: Department of Economics, 440 Mergenthaler Hall, 3400 N. Charles St., Baltimore, MD 21218. I wish to thank Robert Moffitt for helpful guidance on the empirical application. I benefited from useful comments and suggestions by John Rust and Susanne Schennach and seminar participants at Georgetown University. All remaining errors are mine.

1 Introduction

Consider the infinitely mixed model represented by the d_g equality restrictions

$$\mathrm{E}_{\tilde{\pi}\times\tilde{\mu}}\left[g(U,Z;\theta)\right]=0, \tag{1.1}$$

where $\tilde{\mu}$ is the probability measure of the continuous unobservables U supported on $\mathcal{U}\subset\mathbb{R}^{d_u}$, and $\tilde{\pi}$ is the conditional probability measure of observables Z given U, where Z is supported on $\mathcal{Z}\subset\mathbb{R}^{d_z}$. A researcher is interested in θ, which may be a structural primitive or a counterfactual outcome. Reversely let π denote the probability measure of Z and let μ denote the conditional probability measure of U given Z. Then, $\tilde{\pi}\times\tilde{\mu}=\mu\times\pi$ is true, and (1.1) can be rewritten as

$$\mathrm{E}_{\mu\times\pi}\left[g(U,Z;\theta)\right]=0. \tag{1.2}$$

With $\mathcal{P}_{\mathcal{U}|\mathcal{Z}}$ denoting a collection of absolutely continuous regular conditional probability measures supported on $\mathcal{U}\subset\mathbb{R}^{d_u}$ given measurable subsets of $\mathcal{Z}$, the identified set for θ is

$$\Theta_0=\left\{\theta\in\Theta\,\middle|\,\inf_{\mu\in\mathcal{P}_{\mathcal{U}|\mathcal{Z}}}\left\|\mathrm{E}_{\mu\times\pi}\left[g(U,Z;\theta)\right]\right\|=0\right\} \tag{1.3}$$

for a compact set $\Theta\subset\mathbb{R}^{d_\theta}$.

The optimization problem in (1.3) is practically challenging as μ is nonparametric. Schennach (2014) proposes a finite-dimensional optimization problem (ELVIS) to characterize the identified set Θ_0, and therefore substantially alleviates the dimensionality problem. While a variety of computational methods are available to implement the ELVIS, this paper proposes a practically appealing contraction fixed point approach. The proposed method has an advantage especially when the number d_g of moment equalities is not small, as is often the case with panel data. While the Newton-Raphson method, for example, could be also used to implement the ELVIS, it requires to execute d_u-fold numerical integration as many as $2d_g^2+d_g$ times per each iteration step. On the other hand, the contraction fixed point method proposed in this paper requires to execute d_u-fold numerical integration only d_g times per each iteration step. For not so small d_g and/or d_u, this difference may well matter for even the most patient empirical researcher sitting to wait for dozens of empirical estimates. This practical computational procedure was feasible to derive due to a specific form of the first-order condition for the ELVIS.

In Section 2, I propose a θ-dependent self map, whose fixed point exists if and only if θ belongs to the identified set Θ_0. In order to ensure convergence of the iterative procedure to fixed points in large sample, I adapt the innovations of Kasahara and Shimotsu (2012a)[1] to my methodological framework in Section 3. I apply this method to partial identification of marginal effects for the binary choice dynamic panel model (Honoré and Tamer, 2006) extended with correlated random coefficients. After presenting simulation studies in Section 4, I take this model to the Panel Survey of Income Dynamics (PSID) with the empirical framework of Hyslop (1999) and Keane and Sauer (2009) together with correlated random coefficients in Section 5. The counterfactuals of my interest are the marginal effects of an exogenous birth or an exogenous child on female labor supply decisions, accounting for continuous heterogeneity and state dependence as well as the endogeneity of the fertility decisions in observed data. I directly obtain partial identification of counterfactual marginal effects without having to identify the heterogeneous primitives and without having to impose a full structural model in a similar spirit to Bhattacharya (2015) and Hausman and Newey (2014). Obtained set estimates are consistent with plausible economic stories of female labor supply. The empirical framework of this paper is situated in the middle ground between the full structural approach and the IV approach – it flexibly allows for policy-relevant counterfactual analysis that is not instrument-dependent, it

[1] Kasahara and Shimotsu (2012a) analyze contraction properties and propose the relaxation method to reinforce the effectiveness of the iterative procedure of Aguirregabiria and Mira (2007).

accommodates infinite heterogeneous types in a structural framework, and it is robust against model mis-specifications.

2 The Fixed-Point Characterization

In this section, I construct a θ-dependent self map whose fixed point exists if and only if $\theta \in \Theta_0$. Section 2.1 provides the fixed point characterization in population. Section 2.2 provides the fixed point characterization in a random sample.

2.1 In Population

Provided the moment functions g, the observed probability measure π, auxiliary functions $h_j : \mathcal{U} \times \mathcal{Z} \times \Theta \to \mathbb{R}$ for $j \in \{1, \cdots, d_g\}$, and an auxiliary conditional probability measure $\rho \in \mathcal{P}_{\mathcal{U}|\mathcal{Z}}$, we define the self map $\Psi(\,\cdot\,, \theta; g, h, \rho, \pi) : \mathbb{R}^{d_g} \to \mathbb{R}^{d_g}$ for each $\theta \in \Theta$ by

$$\Psi(\gamma, \theta; g, h, \rho, \pi) \equiv \begin{bmatrix} \int\int \frac{h_1(u,z;\theta)}{\int h_1(v,z;\theta) d\rho(v|z;\theta)} g(u,z;\theta)' d\rho(u \mid z) d\pi(z) \\ \vdots \\ \int\int \frac{h_{d_g}(u,z;\theta)}{\int h_{d_g}(v,z;\theta) d\rho(v|z;\theta)} g(u,z;\theta)' d\rho(u \mid z) d\pi(z) \end{bmatrix}^{-1} \times \tag{2.1}$$

$$\begin{bmatrix} \int\int \left(1 + \ln\left(\frac{e^{\gamma' g(u,z;\theta)}}{\int e^{\gamma' g(v,z;\theta)} d\rho(v|z;\theta)}\right)\right)\left(\frac{h_1(u,z;\theta)}{\int h_1(v,z;\theta) d\rho(v|z;\theta)} - \frac{e^{\gamma' g(u,z;\theta)}}{\int e^{\gamma' g(v,z;\theta)} d\rho(v|z;\theta)}\right) d\rho(u \mid z) d\pi(z) \\ \vdots \\ \int\int \left(1 + \ln\left(\frac{e^{\gamma' g(u,z;\theta)}}{\int e^{\gamma' g(v,z;\theta)} d\rho(v|z;\theta)}\right)\right)\left(\frac{h_{d_g}(u,z;\theta)}{\int h_{d_g}(v,z;\theta) d\rho(v|z;\theta)} - \frac{e^{\gamma' g(u,z;\theta)}}{\int e^{\gamma' g(v,z;\theta)} d\rho(v|z;\theta)}\right) d\rho(u \mid z) d\pi(z) \end{bmatrix},$$

where $h = (h_1, \cdots, h_{d_g})'$. This function Ψ has a convenient property to indicate the identified set. In short, it will be claimed that θ belongs to the identified set Θ_0 if and only if this self map $\Psi(\,\cdot\,, \theta; g, h, \rho, \pi)$ has a fixed point. A researcher may select any auxiliary functions $h_j : \mathcal{U} \times \mathcal{Z} \times \Theta \to \mathbb{R}$ and any auxiliary conditional probability measure $\rho \in \mathcal{P}_{\mathcal{U}|\mathcal{Z}}$, subject to the following conditions.

Condition 1. (i) $\mathrm{E}_{\rho\times\pi}\left[g(U,Z;\theta)\left(\frac{h_1(U,Z;\theta)}{\int h_1(v,Z;\theta)d\rho(v|Z;\theta)},\cdots,\frac{h_{d_g}(U,Z;\theta)}{\int h_{d_g}(v,Z;\theta)d\rho(v|Z;\theta)}\right)\right]$ is non-singular. (ii) $\mathrm{supp}\rho(\ \cdot \ \mid z)=\mathcal{U}$ for each $z\in\mathcal{Z}$. (iii) $\mathrm{E}_\pi\left[\ln \mathrm{E}_\rho\left[\exp(\gamma' g(U,Z;\theta)\mid Z\right]\right]$ exists and is twice differentiable in γ.

Condition (i) makes sense of the definition of the self map Ψ containing the inverse of $\mathrm{E}_{\rho\times\pi}\left[g(U,Z;\theta)\left(\frac{h_1(U,Z;\theta)}{\int h_1(v,Z;\theta)d\rho(v|Z;\theta)},\cdots,\frac{h_{d_g}(U,Z;\theta)}{\int h_{d_g}(v,Z;\theta)d\rho(v|Z;\theta)}\right)\right]'$. Since both h and ρ are known to the researcher as her choices, it is simply a matrix that consists of moments with respect to the probability measure π of observables Z. Therefore, this condition (i) is empirically testable with the standard matrix rank tests (e.g., Kleibergen and Paap, 2006). The last two conditions, (ii) and (iii), are the same as the requirements for ρ by Schennach (2014), and are discussed therein. Define the function $\tilde{g}:\mathcal{Z}\times\Theta\times\mathbb{R}^{d_g}\to\mathbb{R}^{d_g}$ by

$$\tilde{g}(z;\theta,\gamma)\equiv\frac{\int g(u,z;\theta)e^{\gamma' g(u,z;\theta)}d\rho(u\mid z;\theta)}{\int e^{\gamma' g(u,z;\theta)}d\rho(u\mid z;\theta)}$$

and we obtain the following lemma characterizing the solutions γ to $\mathrm{E}_\pi\left[\tilde{g}(Z;\theta,\gamma)\right]=0$ by a fixed point of the self map $\Psi(\ \cdot\ ,\theta;g,h,\rho,\pi):\mathbb{R}^{d_g}\to\mathbb{R}^{d_g}$.

Lemma 1. *Suppose that the auxiliary functions h and the auxiliary measure ρ are chosen subject to Condition 1. The moment equality $E_\pi\left[\tilde{g}(Z;\theta,\gamma)\right]=0$ holds if and only if γ is a fixed point of the self map $\Psi(\ \cdot\ ,\theta;g,h,\rho,\pi):\mathbb{R}^{d_g}\to\mathbb{R}^{d_g}$.*

Combining the 'iff' statement in this lemma with the 'iff' statement of Schennach (2014), we can characterize the identified set Θ_0 via the existence of a fixed point of the self map $\Psi(\ \cdot\ ,\theta;g,h,\rho,\pi)$. In Schennach, it is shown that $\inf_{\mu\in\mathcal{P}_{\mathcal{U}|\mathcal{Z}}}\left\|\mathrm{E}_{\mu\times\pi}\left[g(U,Z;\theta)\right]\right\|=0$ holds if and only if $\inf_{\gamma\in\mathbb{R}^{d_g}}\left\|\mathrm{E}_\pi\left[\tilde{g}(Z;\theta,\gamma)\right]\right\|=0$ under Condition 1 (ii) and (iii). Furthermore, the minimizing γ exists in $\mathbb{R}^{d_g}$ as we focus on the collection $\mathcal{P}_{\mathcal{U}|\mathcal{Z}}$ of absolutely continuous conditional probability measures μ. Therefore, the following theorem follows from Lemma 1.

Theorem 1 (Fixed-Point Characterization of Identified Set). *Suppose that the auxiliary functions h and the auxiliary measure ρ are chosen subject to Condition 1. Then, $\theta \in \Theta_0$ is true if and only if the self map $\Psi(\,\cdot\,,\theta; g, h, \rho, \pi) : \mathbb{R}^{d_g} \to \mathbb{R}^{d_g}$ has a fixed point.*

We emphasize that the auxiliary functions h and the auxiliary probability measure ρ can be freely chosen by researchers as far as Condition 1 is satisfied. The existence of a fixed point does not rely on the choice of them. By the theorem, a researcher can make decisions about $\theta \in \Theta_0$ by checking if $\Psi(\,\cdot\,,\theta; g, h, \rho, \pi)$ has a fixed point.

2.2 In A Random Sample

To make decisions on the existence of a fixed point in a random sample, consider the sample-counterpart self map $\Psi(\,\cdot\,,\theta; g, h, \rho, \widehat{\pi}_N)$, where $\widehat{\pi}_N$ denotes the empirical probability measure of Z with sample size N drawn from the population probability measure π. For a small $\varepsilon > 0$, consider the set of ε-fixed points defined by

$$\gamma^*_\varepsilon(\theta; g, h, \rho, \widehat{\pi}_N) \equiv \{\gamma \in \Gamma \mid \|\Psi(\gamma, \theta; g, h, \rho, \widehat{\pi}_N) - \gamma\| < \varepsilon\}$$

for a compact subset $\Gamma \subset \mathbb{R}^{d_g}$. Practically, we reach one of the points in $\gamma^*_\varepsilon(\theta; g, h, \rho, \widehat{\pi}_N)$ by stopping iterations after finite steps based on the small tolerance level ε under a suitable contraction property to be discussed in the next section. The current section argues that this approximation set $\gamma^*_\varepsilon(\theta; g, h, \rho, \widehat{\pi}_N)$ of ε-fixed points may be used in a large sample to make decisions on the existence of a fixed point of the population self map $\Psi(\,\cdot\,,\theta; g, h, \rho, \pi)$ under a given $\theta \in \Theta$. The following sampling process is assumed.

Assumption 1. *(U, Z) is i.i.d. following $\mu \times \pi$.*

In addition to Condition 1, we also require that the choice of the auxiliary functions h and the auxiliary conditional probability measure ρ satisfies the following conditions.

Condition 2. (i) $\mathrm{E}_{\rho\times\pi}\left[g_j(U,Z;\theta)h_k(U,Z;\theta)\right]$ exists for each coordinate pair $j,k \in \{1,\cdots,d_g\}$.
(ii) $(\gamma, z) \mapsto \int \left(1+\ln\left(\frac{e^{\gamma' g(u,z;\theta)}}{\int e^{\gamma' g(v,z;\theta)} d\rho(v|z;\theta)}\right)\right)\left(\frac{h_j(u,z;\theta)}{\int h_j(v,z;\theta) d\rho(v|z;\theta)} - \frac{e^{\gamma' g(u,z;\theta)}}{\int e^{\gamma' g(v,z;\theta)} d\rho(v|z;\theta)}\right) d\rho(u \mid z)$ is continuous for each coordinate $j \in \{1,\cdots,d_g\}$.
(iii) $\mathrm{E}_{\pi}\left[\sup_{\gamma\in\Gamma}\left|\mathrm{E}_{\rho}\left[\left(1+\ln\left(\frac{e^{\gamma' g(U,Z;\theta)}}{\int e^{\gamma' g(v,Z;\theta)} d\rho(v|Z;\theta)}\right)\right)\left(\frac{h_j(U,Z;\theta)}{\int h_j(v,Z;\theta) d\rho(v|Z;\theta)} - \frac{e^{\gamma' g(U,Z;\theta)}}{\int e^{\gamma' g(v,Z;\theta)} d\rho(v|Z;\theta)}\right)\middle| Z\right]\right|\right] < \infty$ for each coordinate $j \in \{1,\cdots,d_g\}$.

These conditions, together with Condition 1 and Assumption 1, guarantee that the sample-counterpart self map $\Psi(\,\cdot\,,\theta; g, h, \rho, \widehat{\pi}_N)$ converges almost surely to the population self map $\Psi(\,\cdot\,,\theta; g, h, \rho, \pi)$ uniformly over $\Gamma \subset \mathbb{R}^{d_g}$ (see Proposition 3 in Section B.1 in the appendix). Using this uniform consistency result in turn yields the following proposition, showing that the set $\gamma^*_\varepsilon(\theta; g, h, \rho, \widehat{\pi}_N)$ of ε-fixed points is informative for the existence of a fixed point of $\Psi(\,\cdot\,,\theta; g, h, \rho, \pi)$ in a large sample.

Proposition 1. *Suppose that Assumption 1 holds and that the auxiliary functions h and the auxiliary measure ρ are chosen subject to Conditions 1 and 2. If $\Psi(\,\cdot\,,\theta; g, h, \rho, \pi)$ has a fixed point $\gamma^* \in \Gamma$, then*

$$\lim_{N\to\infty} P\left(\gamma^*_\varepsilon(\theta; g, h, \rho, \widehat{\pi}_N) \neq \emptyset\right) = 1 \textit{ for all } \varepsilon \in (0,\infty).$$

Conversely, if $\Psi(\,\cdot\,,\theta; g, h, \rho, \pi)$ has no fixed point in Γ, then for some $\bar{\varepsilon} > 0$

$$\lim_{N\to\infty} P\left(\gamma^*_\varepsilon(\theta; g, h, \rho, \widehat{\pi}_N) \neq \emptyset\right) = 0 \textit{ for all } \varepsilon \in (0,\bar{\varepsilon}).$$

The first part of this proposition shows that, if there indeed exists a fixed point $\gamma^* \in \Gamma$, then $\gamma^*_\varepsilon(\theta; g, h, \rho, \widehat{\pi}_N)$ is non-empty with probability approaching one in a large sample for any tolerance level $\varepsilon > 0$. The second part shows that, if there exists no fixed point in Γ, then $\gamma^*_\varepsilon(\theta; g, h, \rho, \widehat{\pi}_N)$ is empty with probability approaching one in a large sample for small tolerance levels.[2] Combining this proposition with the fixed point characterization by Theorem 1 in population, we can make decisions about $\theta \in \Theta_0$ using the set $\gamma^*_\varepsilon(\theta; g, h, \rho, \widehat{\pi}_N)$ of ε-fixed points in a large sample with small values of $\varepsilon > 0$. The following section proposes contraction mapping methods for finding an ε-fixed point.

3 The Iterative Procedure

3.1 Contraction Properties

In this section, we assume the null hypothesis that $\Psi(\,\cdot\,, \theta; g, h, \rho, \pi)$ has a fixed point $\gamma^* \in \mathrm{int}\Gamma$ under $\theta \in \Theta$, and study the convergence properties of the iteration algorithm with the sample counterpart self map $\Psi(\,\cdot\,, \theta; g, h, \rho, \widehat{\pi}_N)$. We let $\widehat{\gamma}_N$ denote the fixed point of $\Psi(\,\cdot\,, \theta; g, h, \rho, \widehat{\pi}_N)$, if any exists. As a first auxiliary step, we show that this sample-counterpart fixed point $\widehat{\gamma}_N$ converges almost surely to the population fixed point γ^* under Assumption 1 and the following assumption, as well as Conditions 1 and 2 for a choice of the auxiliary measure ρ – see Proposition 4 in Section B.2 in the appendix.

Assumption 2. *(i)* $\inf_{\gamma:\|\gamma-\gamma^*\|\geqslant r} \|\Psi(\gamma, \theta; g, h, \rho, \pi) - \gamma\| > 0$ *is true for each* $r > 0$. *(ii)* $\frac{\partial}{\partial \gamma_j}\Psi(\gamma^*, \theta; g, h, \rho, \pi) \neq e_j$ *for each coordinate* $j \in \{1, \cdots, d_g\}$*, where* $e_j \in \mathbb{R}^{d_g}$ *denotes the* j*-th unit vector.*

Part (i) states that γ^* is the unique fixed point of the population self map $\Psi(\,\cdot\,, \theta; g, h, \rho, \pi)$ in Γ. Part (ii) in addition requires that the population self map $\Psi(\,\cdot\,, \theta; g, h, \rho, \pi)$ in Γ is "regular" at this fixed point γ^*, in the sense that it is not tangent to the identity map $\gamma \mapsto \gamma$

[2]Apparently, the second result of Proposition 1 may sound vacuous if Γ is convex, in light of Brouwer's Fixed Point Theorem. However, I note that the self map $\Psi(\,\cdot\,, \theta; g, h, \rho, \pi)$ with the domain restricted to Γ does not necessarily map into Γ. As such, it is possible that $\Psi(\,\cdot\,, \theta; g, h, \rho, \pi)$ has no fixed point in Γ even if Γ is convex.

at $\gamma = \gamma^*$. One can imagine that, if the population self map were tangent at the fixed point, then a sample counterpart fixed point $\widehat{\gamma}_N$ can even fail to exist with non-vanishing probability.

To study contraction properties of the sample counterpart self map $\Psi(\,\cdot\,, \theta; g, h, \rho, \widehat{\pi}_N)$, it is convenient to have the uniform convergence of its first and second derivatives. In the following condition, we strengthen Condition 2 (ii) and (iii) of the requirements for a choice of the auxiliary functions h and the auxiliary conditional probability measure ρ, in order to have the first and second derivatives of $\Psi(\,\cdot\,, \theta; g, h, \rho, \widehat{\pi}_N)$ converge almost surely to those of $\Psi(\,\cdot\,, \theta; g, h, \rho, \pi)$ uniformly over $\Gamma \subset \mathbb{R}^{d_g}$ (see Proposition 5 in Section B.3 in the appendix).

Condition 3. For any partial derivation operator $D_\gamma^\alpha \equiv \frac{\partial^{|\alpha|}}{\partial \gamma^\alpha}$ with $\alpha \in \mathbb{Z}_+^{d_g}$ and $|\alpha| \leqslant 2$:
(i) $(\gamma, z) \mapsto D_\gamma^\alpha \int \left(1 + \ln\left(\frac{e^{\gamma' g(U,Z;\theta)}}{\int e^{\gamma' g(\upsilon,Z;\theta)} d\rho(\upsilon|Z;\theta)}\right)\right)\left(\frac{h_j(U,Z;\theta)}{\int h_j(\upsilon,Z;\theta) d\rho(\upsilon|Z;\theta)} - \frac{e^{\gamma' g(U,Z;\theta)}}{\int e^{\gamma' g(\upsilon,Z;\theta)} d\rho(\upsilon|Z;\theta)}\right) d\rho(u \mid z)$ is continuous for each coordinate $j \in \{1, \cdots, d_g\}$.
(ii) $\mathrm{E}_\pi\left[\sup_{\gamma\in\Gamma}\left|D_\gamma^\alpha \mathrm{E}_\rho\left[\left(1 + \ln\left(\frac{e^{\gamma' g(U,Z;\theta)}}{\int e^{\gamma' g(\upsilon,Z;\theta)} d\rho(\upsilon|Z;\theta)}\right)\right)\left(\frac{h_j(U,Z;\theta)}{\int h_j(\upsilon,Z;\theta) d\rho(\upsilon|Z;\theta)} - \frac{e^{\gamma' g(U,Z;\theta)}}{\int e^{\gamma' g(\upsilon,Z;\theta)} d\rho(\upsilon|Z;\theta)}\right)\right| Z\right]\right|\right] < \infty$ for each coordinate $j \in \{1, \cdots, d_g\}$.

We let $\widehat{\gamma}^0 \in \Gamma$ denote the initial point of iterations, and let $\widehat{\gamma}^\iota = \Psi(\widehat{\gamma}^{\iota-1}, \theta; g, h, \rho, \widehat{\pi}_N)$ denote the point of γ obtained after the ι-th iteration of the sample counterpart self map $\Psi(\,\cdot\,, \theta; g, h, \rho, \widehat{\pi}_N)$. We also introduce the notation $\sigma(M)$ to denote the largest eigenvalue of the $d_g \times d_g$ matrix $M \in \mathcal{M}(d_g, d_g)$. The following proposition shows that, if $\sigma(D_\gamma \Psi(\gamma^*, \theta; g, h, \rho, \pi)) < 1$ is true where D_γ denotes the gradient operator, then the sample counterpart self map $\Psi(\,\cdot\,, \theta; g, h, \rho, \widehat{\pi}_N)$ has a local contraction property in large sample.

Proposition 2. *Suppose that Assumptions 1 and 2 hold and that the auxiliary functions h and the auxiliary measure ρ are chosen subject to Conditions 1, 2 and 3. If $\sigma(D_\gamma \Psi(\gamma^*, \theta; g, h, \rho, \pi)) < 1$ is true, then there exists a neighborhood $\mathcal{N} \subset \Gamma$ of γ^* such that for any initial value $\widehat{\gamma}^0 \in \mathcal{N}$ we have $\lim_{\iota\to\infty} \widehat{\gamma}^\iota = \widehat{\gamma}_N$ almost surely.*

When $\Psi(\cdot ,\theta; g,h,\rho,\widehat{\pi}_N)$ does not have a contraction property, we can resort to the relaxation method (Başar, 1988; Ljungqvist and Sargent, 2004; Kasahara and Shimotsu, 2012a). Consider the self map $\tilde{\Psi}^{\lambda}(\cdot ,\theta; g,h,\rho,\pi) : \mathbb{R}^{d_g} \to \mathbb{R}^{d_g}$ defined by the log linear combination of $\Psi(\cdot ,\theta; g,h,\rho,\pi)$ and the identity map $\gamma \mapsto \gamma$, i.e.,

$$\tilde{\Psi}^{\lambda}(\gamma,\theta; g,h,\rho,\pi) \equiv \Psi(\gamma,\theta; g,h,\rho,\pi)^{\lambda}\gamma^{1-\lambda}$$

Kasahara and Shimotsu (2012a; Proposition 5) show that there exist values λ such that the property $\sigma(D_\gamma \tilde{\Psi}^{\lambda}(\gamma^*,\theta; g,h,\rho,\pi)) < 1$ holds under alternative cases depending on the eigenvalues of $D_\gamma \Psi(\gamma^*,\theta; g,h,\rho,\pi)$.

In practice, there are a few caveats regarding the numerical implementation of the relaxation method. First, since γ is not positive in general, the computer program should be designed appropriately to handle this feature.[3] Second, since we do not know the true eigenvalues $\sigma(D_\gamma \tilde{\Psi}^{\lambda}(\gamma^*,\theta; g,h,\rho,\pi))$ in empirical applications, one needs to implement the relaxation method over a list of values for the relaxation parameter λ in order not to miss a possible fixed point – see Section 5 for a demonstration of this approach in our empirical application. This feature of the proposed method may sound inconvenient, but it is not unique to this particular method. For example, the Newton-Raphson method also lets a researcher choose a step size parameter (or take its default value) in standard software packages. Third, note that the point $\gamma = \vec{0}$ is an absorbing state under the relaxation method, and hence starting iterations near the origin had better be avoided.

In light of the first and third points listed in the previous paragraph, we ideally want the contraction mapping to start from a point that is close to the fixed point $\widehat{\gamma}_N$. I use

[3]There are a couple of alternative ways to handle this problem. The first approach is to let the iteration procedure operate on complex numbers. The second approach is to fix the sign, and to let the iteration procedure operate on absolute values. The latter approach works well only when the initial value $\widehat{\gamma}^0$ has the correct signs.

$\widehat{\gamma}^0 = -\left[\frac{1}{N}\sum_{i=1}^N Var_\rho(g(U,Z_i;\theta) \mid Z_i)\right]^{-1}\left[\frac{1}{N}\sum_{i=1}^N \mathrm{E}_\rho\left[g(U,Z_i;\theta) \mid Z_i\right]\right]$ as the initial point. This expression is derived from a linear approximation of the sample counterpart of the first-order conditions used in the proof of Lemma 1, and thus may be reasonably close to the fixed point $\widehat{\gamma}_N$. With all these devices, the contraction and fixed point approach are shown to perform well with Monte Carlo simulations and the empirical application in Sections 4 and 5.

3.2 Computational Advantage of the Proposed Method

What is a practical advantage of using our contraction fixed point approach over alternative methods? An answer to this question can be found in the components

$$\text{(I)} \quad \begin{bmatrix} \int\int \frac{h_1(u,z;\theta)}{\int h_1(v,z;\theta)d\rho(v|z;\theta)} g(u,z;\theta)' d\rho(u \mid z) d\widehat{\pi}_N(z) \\ \vdots \\ \int\int \frac{h_{d_g}(u,z;\theta)}{\int h_{d_g}(v,z;\theta)d\rho(v|z;\theta)} g(u,z;\theta)' d\rho(u \mid z) d\widehat{\pi}_N(z) \end{bmatrix} \quad \text{and}$$

$$\text{(II)} \quad \begin{bmatrix} \int\int \left(1+\ln\left(\frac{e^{\gamma' g(u,z;\theta)}}{\int e^{\gamma' g(v,z;\theta)}d\rho(v|z;\theta)}\right)\right)\left(\frac{h_1(u,z;\theta)}{\int h_1(v,z;\theta)d\rho(v|z;\theta)} - \frac{e^{\gamma' g(u,z;\theta)}}{\int e^{\gamma' g(v,z;\theta)}d\rho(v|z;\theta)}\right) d\rho(u \mid z) d\widehat{\pi}_N(z) \\ \vdots \\ \int\int \left(1+\ln\left(\frac{e^{\gamma' g(u,z;\theta)}}{\int e^{\gamma' g(v,z;\theta)}d\rho(v|z;\theta)}\right)\right)\left(\frac{h_{d_g}(u,z;\theta)}{\int h_{d_g}(v,z;\theta)d\rho(v|z;\theta)} - \frac{e^{\gamma' g(u,z;\theta)}}{\int e^{\gamma' g(v,z;\theta)}d\rho(v|z;\theta)}\right) d\rho(u \mid z) d\widehat{\pi}_N(z) \end{bmatrix}$$

in the definition of our sample-counterpart self map $\Psi(\,\cdot\,,\theta;g,h,\rho,\widehat{\pi}_N)$. The first component (I) is a $d_g \times d_g$ matrix, and the second component (II) is a $d_g \times 1$ matrix. Every entry in each of these matrices involves a d_u-fold integral. A useful feature of our self map is that its larger part (I) does not involve γ at all. By this property, when we implement iterations of our self map or its relaxed version, we do not need to compute this $d_g \times d_g$ matrix in each iteration step as γ evolves through contraction mapping. Instead, it is sufficient to compute this $d_g \times d_g$ matrix only once before the first iteration of contraction mapping starts. Therefore, our computational procedure needs to implement the d_u-fold numerical integration only for each of the d_g elements of part (II) in each iteration step.

In contrast, the Newton-Raphson method, a popular alternative numerical method which could also be used to implement the ELVIS, requires more computation. The Newton-Raphson method uses (I$'$) the Jacobian matrix of the ELVIS criterion, as well as (II$'$) the criterion itself, which are displayed below.

$$
(\mathrm{I}') \quad \begin{bmatrix} \frac{\partial}{\partial \gamma_1} \int \frac{\int g_1(u,z;\theta) e^{\gamma' g(u,z;\theta)} d\rho(u|z)}{\int e^{\gamma' g(u,z;\theta)} d\rho(u|z;\theta)} d\widehat{\pi}_N(z) & \cdots & \frac{\partial}{\partial \gamma_{d_g}} \int \frac{\int g_1(u,z;\theta) e^{\gamma' g(u,z;\theta)} d\rho(u|z)}{\int e^{\gamma' g(u,z;\theta)} d\rho(u|z;\theta)} d\widehat{\pi}_N(z) \\ \vdots & & \vdots \\ \frac{\partial}{\partial \gamma_1} \int \frac{\int g_{d_g}(u,z;\theta) e^{\gamma' g(u,z;\theta)} d\rho(u|z)}{\int e^{\gamma' g(u,z;\theta)} d\rho(u|z;\theta)} d\widehat{\pi}_N(z) & \cdots & \frac{\partial}{\partial \gamma_{d_g}} \int \frac{\int g_{d_g}(u,z;\theta) e^{\gamma' g(u,z;\theta)} d\rho(u|z)}{\int e^{\gamma' g(u,z;\theta)} d\rho(u|z;\theta)} d\widehat{\pi}_N(z) \end{bmatrix}
$$

$$
(\mathrm{II}') \quad \begin{bmatrix} \int \frac{\int g_1(u,z;\theta) e^{\gamma' g(u,z;\theta)} d\rho(u|z)}{\int e^{\gamma' g(u,z;\theta)} d\rho(u|z;\theta)} d\widehat{\pi}_N(z) \\ \vdots \\ \int \frac{\int g_{d_g}(u,z;\theta) e^{\gamma' g(u,z;\theta)} d\rho(u|z)}{\int e^{\gamma' g(u,z;\theta)} d\rho(u|z;\theta)} d\widehat{\pi}_N(z) \end{bmatrix}
$$

Similarly to our self map components, part (I$'$) is a $d_g \times d_g$ matrix, and part (II$'$) is a $d_g \times 1$ matrix. Notice that every element in each of these matrices contains γ in a non-trivial manner. Therefore, both (I$'$) and (II$'$) need to be computed in each iteration step as γ evolves. Computation of part (I$'$) requires d_g^2 numerical derivatives of the criterion function,[4] each of which in turn contains a d_u-fold numerical integral. Computation of part (II$'$) requires d_u-fold numerical integration d_g times. Thus, in total, the Newton-Raphson method needs to execute d_u-fold numerical integration as many as $2d_g^2 + d_g$ times in each iteration step.

In summary, our approach has much computational advantage over popular alternative methods. Our method requires d_u-fold numerical integrals to be executed only d_g times in each iteration step. On the other hand, the Newton-Raphson method, for example, requires to compute d_u-fold numerical integrals as many as $2d_g^2 + d_g$ times in each iteration step. When the number d_g of moment equalities is not small as is often the case with panel data models,

[4]Numerical computation of the first order derivative (e.g., Judd, 1998; Section 2.5) requires double computation of the criterion.

these two numbers, d_g and $2d_g^2 + d_g$, of times to implement d_u-fold numerical integration make a huge difference.[5] In such cases, the convenient feature of our self map will therefore have much practical appeal to even the most patient empirical researchers sitting to wait for a large number of empirical results. I remark that the Newton-Raphson method I picked for comparison in this subsection is perhaps one of the most favorable alternative, and other methods that I did not pick for comparison may require even more computation – see Judd (1998; Section 4.2) for a list of alternative methods. Finally, since it uses a specific form of the first-order condition of the ELVIS, our computationally attractive method may not be generalizable to non-ELVIS objective functions.

4 Monte Carlo Simulations

Consider the binary choice dynamic panel model with correlated random coefficients:

$$Y_{i,t} = \mathbb{1}\{U_{i,1} + U_{i,2}Y_{i,t-1} + U_{i,3}X_{i,t} \geqslant \epsilon_{i,t}\} \qquad t = 2, \cdots, T.$$

Econometricians observe the choice outcomes $Y_i = (Y_{i,1}, \cdots, Y_{i,T})'$ and covariates $X_i = (X_{i,1}, \cdots, X_{i,T})'$ for T time periods for each individual $i = 1, \cdots, N$. Write $Z_i = (Y_i', X_i')'$ for a short-hand notation. Econometricians do not observe correlated random coefficients $U_i = (U_{i,1}, U_{i,2}, U_{i,3})'$, and do not know their distribution.[6] The remaining unobserved variable $\epsilon_{i,t}$ is a random shock following a known distribution with its cdf denoted by Φ. An unknown

[5] See Judd (1998; Section 7.5) for details of computational complexity for multivariate quadrature.

[6] Whereas a rich set of results are available for random-coefficient linear panel models (e.g., Hsiao and Pesaran, 2008), relatively less is known for random-coefficient binary choice models. Ichimura and Thompson (1998) and Gautier and Kitamura (2013) develop conditions (including independence and large support restrictions) to obtain point identification of the distribution of random coefficients for binary choice models. In this paper, I take the position to be agnostic on these restrictions, but instead only obtain set estimates for certain counterfactuals.

infinite-dimensional parameter governs the joint distribution of $(Z_i', U_i')'$, and it in particular features the initial conditions problem addressed by Wooldridge (2005) and Honoré and Tamer (2006) in a closely related context. Provided the standard independence (predeterminedness) condition $\epsilon_{i,t} \perp\!\!\!\perp (U_i', Y_{i,t-1}, X_{i,t})'$ for the random shocks, we can form the moment equalities

$$\mathrm{E}_{\mu\times\pi}\left[(1, Y_{i,t-1}, X_{i,t})' \left(Y_{i,t} - \Phi\left(U_{i,1} + U_{i,2}Y_{i,t-1} + U_{i,3}X_{i,t}\right)\right)\right] = 0 \tag{4.1}$$

for all $t \in \{2, \cdots, T\}$, where π is the probability measure of Z_i and μ is the conditional probability measure of U_i given Z_i.

Suppose that a researcher is interested in the marginal effect (cf. Honoré and Tamer, 2006; Section 3) of $X_{i,t}$ on $Y_{i,t}$, i.e., $\theta \equiv \mathrm{E}_{\mu\times\pi}\left[U_{i,3} \cdot \Phi'\left(U_{i,1} + U_{i,2}Y_{i,t-1} + U_{i,3}X_{i,t}\right)\right]$ where Φ' denotes the first derivative of Φ. In this case, she may concatenate the moment equality $\mathrm{E}_{\mu\times\pi}\left[\theta - U_{i,3} \cdot \Phi'\left(U_{i,1} + U_{i,2}Y_{i,t-1} + U_{i,3}X_{i,t}\right)\right]$ to (4.1) and form the moment equality of the generic form (1.2), or $\mathrm{E}_{\mu\times\pi}\left[g(U_i, Z_i; \theta)\right] = \vec{0}$, where

$$g(U_i, Z_i; \theta) \equiv \begin{bmatrix} (1, Y_{i,1}, X_{i,2})' \left(Y_{i,2} - \Phi\left(U_{i,1} + U_{i,2}Y_{i,1} + U_{i,3}X_{i,2}\right)\right) \\ \vdots \\ (1, Y_{i,T-1}, X_{i,T})' \left(Y_{i,T} - \Phi\left(U_{i,1} + U_{i,2}Y_{i,T-1} + U_{i,3}X_{i,T}\right)\right) \\ \theta - U_{i,3} \cdot \Phi'\left(U_{i,1} + U_{i,2}Y_{i,t-1} + U_{i,3}X_{i,t}\right) \end{bmatrix}.$$

This formulation conveniently allows one to directly partially identify the counterfactual effects $\theta = \mathrm{E}_{\mu\times\pi}\left[U_{i,3} \cdot \Phi'\left(U_{i,1} + U_{i,2}Y_{i,t-1} + U_{i,3}X_{i,t}\right)\right]$ of marginal increase in $X_{i,t}$ without having to identify the heterogeneous primitives. This approach shares a similar spirit to Bhattacharya (2015) and Hausman and Newey (2014), where they point- and partially identify the equivalent variation and the compensating variation without having to identify the nonparametric primitive structure.

We generate data by $U_i \sim N(\mu, \Sigma)$, where $\mu = (-1.0, 0.0, 1.0)'$ and Σ is a 3×3 matrix with 1.0^2 at the diagonal positions and $0.4 \cdot 1.0^2$ at the off-diagonal positions.[7] The covariates are generated endogenously by $X_{i,t} = \frac{1}{3}(U_{i,1} + U_{i,2} + U_{i,3}) + \eta_{i,t}$ with an independent random shock $\eta_{i,t} \sim N(0, 1)$. The random shock ϵ is generated independently from the standard logistic distribution with its cdf denoted by Φ. Based on this specification, we run 100 time periods of simulated binary choices before the initial time period $t = 1$ in order to have an approximately stationary distribution of $(Y_{i,t-1}, X_{i,t}) \mid U_i$ by $t = 1$. By construction, therefore, there is a non-trivial statistical dependence of $(Y_{i,1}, X_{i,2})$ on U_i, which is the source of the aforementioned initial conditions problem. The panel data sets of size $N = 1,000$ and $T \in \{3, 4, 5, 6, 7, 8, 9, 10\}$ are generated. For each size, we run 1,000 Monte Carlo iterations. Like Honoré and Tamer (2006), we aim to observe how the identified set becomes more informative as the data length T increases.

To construct our self map Ψ, we use the auxiliary functions $h \equiv exp(g)$ and use the probability measure of $N((0, 0, 0)', I_3)$ for the auxiliary conditional probability measure ρ, where I_3 is the 3×3 identity matrix. For this artificial model, negative values of the parameter λ allow for contraction of the relaxation method. It would be necessary to run iterations with various values of $\lambda \in \mathbb{R}$ for actual empirical data for which a researcher does not know the true eigenvalues. In the current section, on the other hand, we set $\lambda = -1.0$ throughout simulations in the interest of time in finishing thousands of simulations.[8]

Table 1 displays the Monte Carlo probabilities of convergence to an ε-approximate fixed point for the tolerance level $\varepsilon = 0.0001$. The probabilities are computed based on 1,000 Monte Carlo iterations. In each Monte Carlo iteration, we run up to 100 self map iterations. The

[7] These numbers are selected in order to have the population mean marginal effects to have the approximate value $\mathrm{E}_{\mu\times\pi}[U_{i,3} \cdot \Phi'(U_{i,1} + U_{i,2}Y_{i,t-1} + U_{i,3}X_{i,t})] \approx 0.10$ in the stationary distribution of $(Z_i', U_i')'$.

[8] For the empirical application presented in Section 5, I do not know the true eigenvalues. Therefore, I will use a list of various relaxation parameter values λ in order to avoid missing any potential fixed point.

	$T=3$	$T=4$	$T=5$	$T=6$	$T=7$	$T=8$	$T=9$	$T=10$
$\theta=0.00$	0.018	0.018	0.039	0.024	0.023	0.038	0.040	0.029
$\theta=0.02$	0.398	0.262	0.202	0.163	0.151	0.117	0.087	0.088
$\theta=0.04$	0.884	0.770	0.722	0.688	0.621	0.589	0.563	0.536
$\theta=0.06$	0.985	0.961	0.927	0.905	0.908	0.853	0.864	0.840
$\theta=0.08$	0.996	0.984	0.966	0.953	0.922	0.944	0.922	0.920
$\theta=0.10$	0.996	0.987	0.958	0.959	0.942	0.952	0.941	0.929
$\theta=0.12$	0.996	0.977	0.962	0.947	0.945	0.938	0.932	0.943
$\theta=0.14$	0.996	0.967	0.973	0.924	0.944	0.941	0.917	0.926
$\theta=0.16$	0.995	0.980	0.948	0.930	0.905	0.895	0.871	0.856
$\theta=0.18$	0.998	0.968	0.914	0.859	0.840	0.833	0.797	0.781
$\theta=0.20$	0.993	0.945	0.864	0.799	0.747	0.739	0.712	0.672
$\theta=0.22$	0.984	0.882	0.764	0.679	0.624	0.600	0.518	0.497
$\theta=0.24$	0.980	0.813	0.661	0.554	0.446	0.408	0.376	0.341
$\theta=0.26$	0.958	0.708	0.504	0.363	0.291	0.244	0.220	0.203
$\theta=0.28$	0.922	0.577	0.296	0.219	0.171	0.138	0.108	0.103
$\theta=0.30$	0.859	0.381	0.178	0.101	0.068	0.073	0.049	0.036
$\theta=0.32$	0.748	0.211	0.077	0.046	0.038	0.019	0.023	0.013
$\theta=0.34$	0.524	0.089	0.022	0.021	0.013	0.008	0.008	0.002

Table 1: Monte Carlo probabilities of achieving an ε-fixed point. The results are based on the sample size of $N = 1,000$, the tolerance level of $\varepsilon = 0.0001$, the relaxation parameter of $\lambda = -1.0$, and 1,000 Monte Carlo iterations. The approximate true value, $\theta = \mathrm{E}[U_{i,3} \cdot \Phi'(U_{i,1} + U_{i,2}Y_{i,t-1} + U_{i,3}X_{i,t})] \approx 0.10$, is indicated by a pair of border lines.

	$T=3$	$T=4$	$T=5$	$T=6$	$T=7$	$T=8$	$T=9$	$T=10$
$\theta=0.00$	0.024	0.016	0.022	0.033	0.024	0.020	0.028	0.027
$\theta=0.02$	0.442	0.274	0.200	0.158	0.132	0.103	0.086	0.086
$\theta=0.04$	0.949	0.870	0.810	0.770	0.742	0.689	0.692	0.668
$\theta=0.06$	0.995	0.987	0.959	0.937	0.933	0.911	0.905	0.911
$\theta=0.08$	0.999	0.991	0.974	0.962	0.952	0.940	0.940	0.925
$\theta=0.10$	0.999	0.990	0.976	0.956	0.958	0.948	0.946	0.929
$\theta=0.12$	1.000	0.987	0.965	0.964	0.953	0.947	0.953	0.943
$\theta=0.14$	0.999	0.982	0.966	0.959	0.946	0.927	0.926	0.919
$\theta=0.16$	0.996	0.985	0.958	0.916	0.886	0.862	0.856	0.851
$\theta=0.18$	1.000	0.973	0.930	0.845	0.836	0.782	0.749	0.736
$\theta=0.20$	0.996	0.965	0.881	0.796	0.679	0.665	0.601	0.602
$\theta=0.22$	0.997	0.924	0.782	0.621	0.555	0.505	0.434	0.347
$\theta=0.24$	0.990	0.872	0.637	0.467	0.349	0.292	0.266	0.238
$\theta=0.26$	0.987	0.773	0.469	0.320	0.187	0.165	0.118	0.112
$\theta=0.28$	0.965	0.613	0.286	0.146	0.089	0.065	0.051	0.037
$\theta=0.30$	0.940	0.389	0.106	0.057	0.023	0.026	0.021	0.014
$\theta=0.32$	0.813	0.158	0.039	0.011	0.010	0.013	0.006	0.008
$\theta=0.34$	0.616	0.058	0.011	0.007	0.006	0.002	0.003	0.003

Table 2: Monte Carlo probabilities of achieving an ε-fixed point. The results are based on the sample size of $N = 2,000$, the tolerance level of $\varepsilon = 0.0001$, the relaxation parameter of $\lambda = -1.0$, and 1,000 Monte Carlo iterations. The approximate true value, $\theta = \mathrm{E}[U_{i,3} \cdot \Phi'(U_{i,1} + U_{i,2}Y_{i,t-1} + U_{i,3}X_{i,t})] \approx 0.10$, is indicated by a pair of border lines.

event of $\widehat{\gamma}^{\iota} \notin \gamma_{\varepsilon}^{*}(\theta; g, h, \rho, \widehat{\pi}_N)$ for all $\iota \in \{1, \cdots, 100\}$ is considered as non-convergence. The approximate true value, $\theta \approx 0.10$, is indicated by a pair of border lines in the table. First, looking at any column in the table, observe that the convergence probability is high at and around the approximate true value $\theta \approx 0.10$. Second, looking at the row for the approximate true value $\theta \approx 0.10$, observe that the convergence probability does not substantially drop as T increases in this row. Third, looking at any row far away from the approximate true value $\theta \approx 0.10$, observe that the convergence probability substantially drops as T increases. From these observations, we can see that the set estimates tend to shrink toward the approximate true value $\theta \approx 0.10$ as T increases. This pattern of the simulation results is analogous to the analysis by Honoré and Tamer (2006) on a related model. Table 2 displays the results based on a larger sample size ($N = 2,000$). Not surprisingly, the obtain bounds are approximately the same as those in Table 1, because the identified sets will not change with the sample size. On the other hand, we see that the sampling variation of the boundaries in Table 2 is sharper than that of Table 1.

5 Inter-Temporal Female Labor Supply Decisions

Hyslop (1999) and Keane and Sauer (2009) study Markov models of inter-temporal female labor supply decisions, where the current decision is modeled to depend on the previous labor force participation state and individual unobserved heterogeneity as well as observed covariates. In this framework, the spurious serial correlation of the labor supply decisions is explicitly distinguished into two causal factors - the state dependence and heterogeneity. It is found that the decisions are substantially affected by both of the two factors.

In this paper, I consider the Markov model of inter-temporal female labor supply decisions extended with correlated random coefficients to allow for nonseparable interactions between unobserved heterogeneity and observed explanatory factors. With this model, I aim to obtain set estimates for the counterfactual marginal effects of an exogenous birth or an exogenous child on labor supply decisions, controlling for the incidental parameters (heterogeneity) and lagged labor supply (state dependence) as well as other observed characteristics.

There are two broad approaches used to answer similar empirical questions in the existing literature. The first approach uses full structural models to allow for identification of all the economic components of the model and counterfactual implications. While it is useful to answer various economic and policy questions, this approach may be subject to mis-specifications biases, and is generally incapable of handling infinite heterogeneous types. The second approach uses instrumental variables. While it imposes less structural restrictions and allows for infinite heterogeneous types, this approach obtains causal effects only among certain subpopulations characterized by the instrumental variables, and may not always be able to answer relevant economic and policy questions. The method proposed in this paper is situated in the middle ground, allows for continuous heterogeneity, is robust against model mis-specifications, and is capable of partially identifying various policy-relevant counterfactual implications of the model.

In studying female labor supply, it is crucial to account for the simultaneity or the endogeneity of fertility decisions and labor supply decisions – see Rosenzweig and Wolpin (1980), Moffitt (1984), Mroz (1987), Hotz and Miller (1988), Jakubson (1988), Browning (1992), Angrist and Evans (1998), Hyslop (1999), Keane and Sauer (2009) and Keane and Wolpin (2010). We model the endgeneity through an infinite dimensional nuisance parameter, i.e., $\mu \times \pi$, which reflects an implicit process where unobserved heterogeneity U_i simultaneously affects the current labor supply and fertility decisions $Z_{i,t}$, given $Z_{i,t-1}$. One could of course impose an explicit structural model for this simultaneous decision process. Constructing such a full model may certainly help the identified set to shrink, but it may incur a bias in case the shrinkage is due to model mis-specifications. Because we can allow for fully nonparametric family of nuisance parameters $\mu \times \pi$, it effectively identifies the counterfactual marginal effects robustly against concrete structural specifications of the simultaneous decision process.

5.1 Data

Following earlier work (e.g., Hyslop, 1999), I use the Panel Survey of the Income Dynamics (PSID) for the calendar years 1979-1985. This period contains waves 12–19 of the PSID. I use the balanced portion of the panel through the period. The households with male heads with continuously married couples are selected. The data consist of female individuals aged between 25 and 54 all the way from 1979 to 1985. Since the PSID truncates very high numbers for earnings and income, I drop observations with $\geqslant$99,999 dollars of earnings and/or income at any year in the period. To compute the "other family income" that is supposed to affect the female labor supply decision, I take the difference of the total annual family income and the annual wage income of the female. Since the other family income will be transformed into the logarithms for the econometric analysis, I drop observations with zero or negative values for this field. After dropping observations with these attributes and missing values, the sample size reduces to $N = 1,413$.

Table 3 displays summary statistics of the sample. The values indicate the sample means. The values in parentheses indicate the standard errors. Since the survey asks for labor supply, income, and earnings for the previous calendar year, we shift the first year from 1979 to 1978, and the final year from 1985 to 1984, in order to reflect the actual time of their realizations. The other time-varying variables, such as fertility and family composition, are associated with the current survey years. The first row shows that the female labor supply rates exhibit an

	Calendar Year						
$N = 1,413$	1978	1979	1980	1981	1982	1983	1984
Labor Supply	0.692	0.707	0.707	0.702	0.697	0.723	0.752
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.011)
Log Other Income	9.738	9.859	9.942	9.995	10.051	10.104	10.179
(Nominal Value)	(0.018)	(0.015)	(0.017)	(0.020)	(0.018)	(0.019)	(0.020)
Log Other Income	9.738	9.752	9.708	9.663	9.659	9.681	9.713
(Deflated to 1978 Value)	(0.018)	(0.015)	(0.017)	(0.020)	(0.018)	(0.019)	(0.020)
Fertility		0.085	0.061	0.052	0.042	0.042	0.038
		(0.007)	(0.006)	(0.006)	(0.005)	(0.005)	(0.005)
# Children Aged 1–2		0.280	0.271	0.260	0.235	0.187	0.156
		(0.013)	(0.013)	(0.013)	(0.012)	(0.011)	(0.011)
# Children Aged 3–5		0.312	0.316	0.304	0.287	0.281	0.270
		(0.014)	(0.014)	(0.014)	(0.014)	(0.013)	(0.013)
# Children Aged 6–13		0.800	0.816	0.810	0.831	0.841	0.808
		(0.025)	(0.025)	(0.024)	(0.025)	(0.025)	(0.025)
Black							0.230
							(0.011)
Education							12.427
							(0.059)

Table 3: Summary statistics of the Panel Survey of Income Dynamics (PSID). The displayed values are the sample means. The values in parentheses indicate the standard errors.

overall increasing trend over time. The second row shows the logarithm of other family income, which we assume explains a part of the female labor supply decision. The third row deflates the values in the second row to the 1978 value using the average annual consumer price index (CPI) provided by the U.S. Department of Labor, Bureau of Labor Statistics. These real values, unlike the nominal ones, stay roughly constant over time during the period of interest. The fourth row shows that fertility, which may well be a partially alternative choice outcome to the female labor supply, steadily decreases over time during the period of interest. Accordingly, the number of younger children, shown in the fifth and sixth rows, decrease over time during the period. On the other hand, the number of children aged 6–13 shown in the seventh row does not exhibit any monotonic trend over time. The time-constant variables, such as years of education and race, are displayed in the table for the purpose of providing a better idea about the underlying population of the sample. However, these constant variables will not be used in our econometric analysis in the following subsections because any constant variable will be absorbed by the time-constant heterogeneous parameters.

5.2 Empirical Specifications

With the standard logistic cdf Φ, we consider the following model of female labor supply:

$$\Pr(Y_{i,t} = 1 \mid Y_{i,t-1}, X_{i,t}, \alpha_i, \beta_i, \gamma_i) \;=\; \Phi\left(\alpha_i + \beta_i Y_{i,t-1} + X'_{i,t}\gamma_i\right) \qquad \text{for } t = 2, \cdots, T.$$

The outcome variable, $Y_{i,t}$, is the binary indicator of labor supply by female individual i in year t. Note that the inclusion of the lagged labor supply indicator allows for state dependence, while the inclusion of the correlated random coefficients $(\alpha_i, \beta_i, \gamma'_i)'$ allows for heterogeneity. Distinguishing these two causal factors has long been discussed in a broad literature at least since Feller (1943), and particularly in economics (e.g., Heckman, 1981ab).

The covariate vector $X_{i,t}$ includes the fertility dummy, the logarithm of other family income, the number of children between 1 and 2, the number of children between 3 and 5, and the number of children between 6 and 13. For this list of explanatory variables, I follow the empirical specification of Hyslop (1999; Section 4). As noted in the previous subsection, we will not include the time-constant variables, such as years of education and race, because they are anyway absorbed by the time-constant heterogeneous parameter α_i.

To have the notations of this model consistent with those in the general formulation of this paper, we let $U_i = (\alpha_i, \beta_i, \gamma_i')'$ denote the vector of all the unobserved variables and let $Z_i = (Y_{i,1}, \cdots, Y_{i,T}, X_{i,2}', \cdots, X_{i,T}')'$ denote the vector of all the observed variables. The counterfactual marginal effects of exogenous birth at time t are

$$\Phi\left(\alpha_i + \beta_i Y_{i,t-1} + X(1)_{i,t}'\gamma_i\right) - \Phi\left(\alpha_i + \beta_i Y_{i,t-1} + X(0)_{i,t}'\gamma_i\right),$$

where $X(b)_{i,t}$ is the same as $X_{i,t}$ except that the coordinate for the fertility dummy is set exogenously to b for each $b \in \{0, 1\}$. Note that, with the fertility variable being binary, this effect is the discrete analog of the counterfactual marginal effect considered in Section 4. We partially identify the population average of these counterfactual marginal effects using the same approach as in Section 4. Specifically, we use the moment equality (1.2) where g is defined by

$$g(U_i, Z_i; \theta) \equiv \begin{bmatrix} (1, Z_{i,1}')'\left(Y_{i,2} - \Phi\left(\alpha_i + \beta_i Y_{i,1} + X_{i,2}'\gamma_i\right)\right) \\ \vdots \\ (1, Z_{i,T-1}')'\left(Y_{i,T} - \Phi\left(\alpha_i + \beta_i Y_{i,T-1} + X_{i,T}'\gamma_i\right)\right) \\ \theta - \frac{1}{T-1}\sum_{t=2}^{T}\left(\Phi\left(\alpha_i + \beta_i Y_{i,t-1} + X(1)_{i,t}'\gamma_i\right) - \Phi\left(\alpha_i + \beta_i Y_{i,t-1} + X(0)_{i,t}'\gamma_i\right)\right) \end{bmatrix}$$

where $Z_{i,t-1}$ is (a subvector of) $(Y_{i,1}, \cdots, Y_{i,t-1}, X_{i,2}', \cdots, X_{i,t}')'$. With this device, we directly identify the counterfactual marginal effects, without having to identify the heterogeneous primitives in a similar spirit to Bhattacharya (2015) and Hausman and Newey (2014). It is reported in the literature that, even if bounds for structural parameters may be wide, bounds for counterfactuals tend to be narrow enough to be informative – see Norets and Tang (2014) for example.

5.3 Empirical Results

Table 4 displays set estimates for the average counterfactual marginal effects of exogenous births on female labor supply. The displayed numbers indicate the maximum values of the relaxation parameter $\lambda \in \{-1, -\frac{1}{2}, \cdots, -\frac{1}{2^7}, -\frac{1}{2^8}\} \cup \{\frac{1}{2^8}, \frac{1}{2^7}, \cdots, \frac{1}{2}, 1\}$ at which the contraction mapping converged to an $\varepsilon = 0.0001$-fixed point.[9] The blank cells indicate that a convergence did not occur for any $\lambda \in \{-1, -\frac{1}{2}, \cdots, -\frac{1}{2^7}, -\frac{1}{2^8}\} \cup \{\frac{1}{2^8}, \frac{1}{2^7}, \cdots, \frac{1}{2}, 1\}$. In other words, the set estimate in each column of the table includes all the values of θ for which there is a non-empty row entry. These cells are shaded for visual convenience.

Columns (1) through (5) show results for the population average counterfactual marginal effects $\theta = \mathrm{E}\left[\Phi\left(\alpha_i + \beta_i Y_{i,t-1} + X(1)'_{i,t}\gamma_i\right) - \Phi\left(\alpha_i + \beta_i Y_{i,t-1} + X(0)'_{i,t}\gamma_i\right)\right]$ using various combinations of included regressors $(Y_{i,t-1}, X'_{i,t})'$. When we include nothing but the fertility decision, the set estimate is $[-1.0, -0.4]$ as shown in column (1). When we in addition include the state dependence or the first lag of the labor supply indicator $Y_{i,t-1}$, the set estimate expands to $[-1.0, -0.1]$ as shown in column (2). It is wide, but the zero effect is still excluded from the set estimate. This expansion of the set features the general difficulties in identifying the models where heterogeneity and state dependence co-exist (e.g., Feller,1943, Heckman, 1981ab, Honoré and Tamer, 2006). Including other income sources does not change the set estimate much – see

[9]In Monte Carlo simulations in Section 4, we used a fixed relaxation parameter λ due to our knowledge of the true data generating process. In the current empirical application, we do not know the data generating process. Therefore, I use various values of λ in order to avoid missing possible fixed points.

Average Marginal Effects of Exogenous Births on Female Labor Supply: Set Estimates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Subpopulation	All Individuals					Births	No Birth
Lag Labor Supply	N	Y	Y	Y	Y	Y	Y
Other Income	N	N	Y	Y	Y	Y	Y
Children	N	N	N	Y	Y	Y	Y
Time	N	N	N	N	Y	Y	Y
$\theta = -1.00$	2^{-0}	2^{-0}	2^{-0}	2^{-0}	2^{-0}	2^{-0}	2^{-1}
$\theta = -0.90$	2^{-0}	2^{-0}	2^{-0}	2^{-0}	2^{-0}	2^{-0}	2^{-1}
$\theta = -0.80$	2^{-0}	2^{-0}	2^{-0}	2^{-0}	2^{-0}	2^{-0}	2^{-1}
$\theta = -0.70$	2^{-0}	2^{-0}	2^{-0}	2^{-0}	2^{-0}	2^{-0}	2^{-3}
$\theta = -0.60$	2^{-0}	2^{-0}	2^{-0}	2^{-0}	2^{-1}	2^{-0}	
$\theta = -0.50$	2^{-0}	2^{-0}	2^{-0}			2^{-0}	
$\theta = -0.40$	2^{-0}	2^{-0}	2^{-0}			2^{-0}	
$\theta = -0.30$		2^{-0}	2^{-0}			2^{-1}	
$\theta = -0.20$		2^{-0}	2^{-1}			2^{-1}	
$\theta = -0.10$		2^{-2}	2^{-2}				
$\theta = \ \ 0.00$							

Table 4: Set estimates for the counterfactual marginal effects of exogenous births on female labor supply indicated by the maximum contractionary relaxation parameter λ. Columns (1)–(5) show results for the population average marginal effects. Column (6) shows results for the average marginal effects in the subpopulation of those who actually gave births. Column (7) shows results for the average marginal effects in the subpopulation of those who did not.

column (3). When we include the three age categories of children, however, the set estimate now shrinks down to $[-1.0, -0.6]$ as shown in column (4). Finally, adding time trends[10] does not change the set estimate much – see column (5). With this full specification, we obtain that the set estimate for the population average counterfactual marginal effects of exogenous birth on female labor supply is $[-1.0, -0.6]$.

While the first five columns show results for the population average, we can also obtain set estimates for subpopulation averages in a similar spirit to the "treatment on the treated" used in the literature of causal inference. Because female individuals endogenously do or do not self-select into pregnancy, the causal effects of fertility on labor supply may well be different between the subpopulation of individuals with births and the subpopulation of individuals without births. Column (6) shows the set estimate for the average counterfactual marginal effects of exogenous birth among the subpopulation of females who actually gave births, i.e., $\theta = \mathrm{E}\left[\Phi\left(\alpha_i + \beta_i Y_{i,t-1} + X(1)'_{i,t}\gamma_i\right) - \Phi\left(\alpha_i + \beta_i Y_{i,t-1} + X(0)'_{i,t}\gamma_i\right) \mid \text{Birth}_{i,t} = 1\right]$.[11] The set estimate, $[-1.0, -0.2]$, for this subpopulation is wider than that for the population, $[-1.0, -0.6]$. This result is consistent with plausible economic stories. For example, those females who actually made the endogenous decisions to give births may tend to have types of skills and occupations for which work is more feasible during maternity than the others. As such, this subpopulation may well have smaller (in absolute value) effects of births on labor supply. In contrast, the set estimate, $[-1.0, -0.7]$, shown in column (7) for the subpopulation of females who actually did not give births is narrower toward -1.0. Those females who endogenously decided not to give births may tend to have types of skills and occupations for which work is

[10]The time trends are linear in time. But the trend slopes are heterogeneous across individuals i, as are γ_i.

[11]Insert $\mathbb{1}\{\text{Birth}_{i,t} = 1\}\left[\theta - \frac{1}{T-1}\sum_{t=2}^{T}\left(\Phi\left(\alpha_i + \beta_i Y_{i,t-1} + X(1)'_{i,t}\gamma_i\right) - \Phi\left(\alpha_i + \beta_i Y_{i,t-1} + X(0)'_{i,t}\gamma_i\right)\right)\right]$ instead of just $\theta - \frac{1}{T-1}\sum_{t=2}^{T}\left(\Phi\left(\alpha_i + \beta_i Y_{i,t-1} + X(1)'_{i,t}\gamma_i\right) - \Phi\left(\alpha_i + \beta_i Y_{i,t-1} + X(0)'_{i,t}\gamma_i\right)\right)$ in the last row in the definition of $g(U_i, Z_i; \theta)$ in order to obtain set estimates for the subpopulation average effects.

more difficult during maternity, and hence the average effects of birth on labor supply may well be large in absolute value for this subpopulation.

Table 5 shows set estimates for the average counterfactual marginal effects among various subpopulations defined by observed characteristics. As a reference, the first column in Table 5 displays the set estimate for the population average effects copied from column (5) of Table 4. Columns (8) and (9) compare set estimates between the subpopulations defined by the presence of a child aged 1 through 5. Likewise, columns (10) and (11) compare set estimates between the subpopulations defined by the presence of a child aged 1 through 13. In both of these pairs of columns, the bound for the subpopulation of females who have a child is closer to -1.0. In other words, the average marginal effects of exogenous birth on labor supply may be larger in absolute value if it is the first birth, or if it is the first birth after many years of no birth.

Columns (12) and (13) compare set estimates between the subpopulations defined by the amount of household income other than the earnings by the female individual. Two income categories are constructed using the borderline of 20,000 US dollars in the 1978 value. The bound for the subpopulation with larger amounts of other income is closer to -1.0. This is consistent with economic stories. If one has enough income from other sources, then the marginal benefit from labor supply during maternity may well be smaller. Therefore, it is natural for this subpopulation to have larger (in absolute value) average marginal effects of exogenous birth on female labor supply.

While all the results presented so far concern about the marginal effects of exogenous fertility, it is also of interest to study the causal effects of an exogenous child on female labor supply. As a related matter, one may be interested in how the magnitude of such effects change with the age of a child. Table 6 shows set estimates for the average counterfactual effects of exogenously having a child of various age categories. Specifically, column (14) shows the set

Average Marginal Effects of Exogenous Births on Female Labor Supply: Set Estimates

	(5)	(8)	(9)	(10)	(11)	(12)	(13)
		Infant		Child		Other Income	
Subpopulation	All	Y	N	Y	N	<20K	⩾20K
Lag Labor Supply	Y	Y	Y	Y	Y	Y	Y
Other Income	Y	Y	Y	Y	Y	Y	Y
Children	Y	Y	Y	Y	Y	Y	Y
Time	Y	Y	Y	Y	Y	Y	Y
$\theta = -1.00$	2^{-0}	2^{-2}	2^{-4}	2^{-1}	2^{-2}	2^{-1}	2^{-3}
$\theta = -0.90$	2^{-0}	2^{-2}	2^{-5}	2^{-2}	2^{-2}	2^{-2}	2^{-3}
$\theta = -0.80$	2^{-0}	2^{-3}	2^{-6}	2^{-2}	2^{-2}	2^{-2}	2^{-4}
$\theta = -0.70$	2^{-0}	2^{-4}		2^{-3}	2^{-2}	2^{-2}	
$\theta = -0.60$	2^{-1}	2^{-4}		2^{-4}		2^{-2}	
$\theta = -0.50$		2^{-4}					
$\theta = -0.40$							
$\theta = -0.30$							
$\theta = -0.20$							
$\theta = -0.10$							
$\theta = \ \ 0.00$							

Table 5: Set estimates for the counterfactual marginal effects of exogenous births on female labor supply indicated by the maximum contractionary relaxation parameter λ. Column (5), copied from Table 4, shows the result for the population average marginal effects as a reference. Columns (8)–(13) show results for the average marginal effects among various subpopulations.

Average Marginal Effects of Exogenous Child on Female Labor Supply: Set Estimates

	(5)	(14)	(15)	(16)	(17)
Child Aged 0 (Birth)	1 – 0	1 – 0	0 – 0	0 – 0	0 – 0
Child Aged 1–2	Sample	0 – 0	1 – 0	0 – 0	0 – 0
Child Aged 3–5	Sample	0 – 0	0 – 0	1 – 0	0 – 0
Child Aged 6–13	Sample	0 – 0	0 – 0	0 – 0	1 – 0
$\theta = -1.00$	2^{-0}	2^{-0}	2^{-0}	2^{-1}	2^{-0}
$\theta = -0.90$	2^{-0}	2^{-0}	2^{-0}	2^{-1}	2^{-0}
$\theta = -0.80$	2^{-0}	2^{-1}	2^{-1}	2^{-1}	2^{-0}
$\theta = -0.70$	2^{-0}	2^{-1}	2^{-3}	2^{-2}	2^{-0}
$\theta = -0.60$	2^{-1}				2^{-0}
$\theta = -0.50$					2^{-0}
$\theta = -0.40$					2^{-0}
$\theta = -0.30$					2^{-2}
$\theta = -0.20$					2^{-5}
$\theta = -0.10$					
$\theta = \ \ 0.00$					

Table 6: Set estimates for the counterfactual marginal effects of exogenous child on female labor supply indicated by the maximum contractionary relaxation parameter λ. Column (5) is copied from Table 4. Column (14) shows the population average marginal effects of having a birth in the presence of no other child. Column (15) shows the population average marginal effects of having a child aged 1–2 against no child. Column (16) shows the population average marginal effects of having a child aged 3–5 against no child. Column (17) shows the population average marginal effects of having a child aged 6–13 against no child. $X_{i,t}$ includes everything.

estimate for the population average of

$$\Phi\left(\alpha_i + \beta_i Y_{i,t-1} + X(1,0,0,0)'_{i,t}\gamma_i\right) \; - \; \Phi\left(\alpha_i + \beta_i Y_{i,t-1} + X(0,0,0,0)'_{i,t}\gamma_i\right),$$

where $X(b_0, b_1, b_2, b_3)_{i,t}$ is the same as $X_{i,t}$ except that the coordinate for the fertility dummy is set exogenously to b_0, the coordinate for a child of age 1–2 is set exogenously to b_1, the coordinate for a child of age 3–5 is set exogenously to b_2, and the coordinate for a child of age 6–13 is set exogenously to b_3 for each $(b_0, b_1, b_2, b_3) \in \{0,1\}^4$. Likewise, columns (15), (16) and (17) show the results for the population average of

$$\begin{aligned}
&\Phi\left(\alpha_i + \beta_i Y_{i,t-1} + X(0,1,0,0)'_{i,t}\gamma_i\right) \quad - \quad \Phi\left(\alpha_i + \beta_i Y_{i,t-1} + X(0,0,0,0)'_{i,t}\gamma_i\right), \\
&\Phi\left(\alpha_i + \beta_i Y_{i,t-1} + X(0,0,1,0)'_{i,t}\gamma_i\right) \quad - \quad \Phi\left(\alpha_i + \beta_i Y_{i,t-1} + X(0,0,0,0)'_{i,t}\gamma_i\right), \text{ and} \\
&\Phi\left(\alpha_i + \beta_i Y_{i,t-1} + X(0,0,0,1)'_{i,t}\gamma_i\right) \quad - \quad \Phi\left(\alpha_i + \beta_i Y_{i,t-1} + X(0,0,0,0)'_{i,t}\gamma_i\right),
\end{aligned}$$

respectively. Note that columns (14), (15), and (16) show the same set estimates, $[-1.0, -0.7]$, i.e., the set estimate for the average causal effects of an exogenous child of age 0 is the same as the set estimate for the average causal effects of an exogenous child of age 1–2 or 3–5. On the other hand, column (17) shows that the set estimate, $[-1.0, -0.2]$, for the average causal effects of an exogenous child of age 6–13 is much wider. This pattern of the results is natural, as female individuals may find it more feasible to work after her child reaches the school age.

6 Concluding Remarks

This paper proposes a contraction fixed point approach to partial identification problems for a class of incomplete models with continuously distributed endogenous latent variables. This approach provides a practically appealing method to implement the ELVIS (Schennach, 2014). In particular, when the number d_g of moment equalities is not small as is often the case with panel data models, it outperforms alternative methods. While the Newton-Raphson method, for example, requires to execute d_u-fold numerical integration as many as $2d_g^2+d_g$ times per each iteration step, the contraction fixed point method proposed in this paper requires to execute d_u-fold numerical integration only d_g times per each iteration step. For not so small d_g and/or d_u, this difference can hugely affect the efficiency of empirical research. This convenient property was obtained by using a specific form of the first-order condition for the ELVIS. With the contraction property of the relaxation method, the proposed approach was shown to perform well in Monte Carlo simulations and in the empirical application.

I remark some methodological analogies between finite and infinite mixture models. For finite mixture models, there is a growing literature on identification, e.g., Hall and Zhou (2003), Hu (2008), Kasahara and Shimotsu (2009) and Henry, Kitamura and Salanié (2014). For finite mixture models, Heckman and Singer (1984) and Arcidiacono and Jones (2003) propose optimization-based methods of estimation, and Aguirregabiria and Mira (2007) and Kasahara and Shimotsu (2012a) propose a contraction-based method of estimation. In contrast, I focus on infinite unobserved types U. For infinite mixture models, Hu and Schennach (2008) present an identification result. For infinite mixture models, Schennach (2014) proposes optimization-based methods of estimation, and no paper to my knowledge has ever proposed a contraction-based method of estimation.[12] This paper fills the last part in the methodological parallel between the finite and infinite mixture frameworks.

Turning to the empirical application, we allow a large extent of model flexibility accounting

[12]If an infinite mixture model could be represented by a certain class of contraction operator equation (e.g., Rust, Traub, and Wozniakowski, 2002), then a contraction approach may be practically feasible. Furthermore, the extremal characterization of Schennach (2014) can be also technically translated into a fixed point problem through the Newton-Raphson method.

for heterogeneity, state dependence, and endogenous fertility decisions in observed data. I obtain set estimates without having to specify a distribution assumption and without having to impose a specific simultaneous decision model in a similar manner to Honoré and Tamer (2006). Furthermore, I directly obtain counterfactual effects without having to identify the structural primitives in a similar manner to Bhattacharya (2015) and Hausman and Newey (2014). The cost of imposing less assumptions certainly is the width of the identified set. However, as reported by Norets and Tang (2014) in a different context, identified sets for counterfactuals tend to be informative even if those for structural primitives may be wide. The identified sets for the counterfactuals obtained in this paper was indeed informative enough to fit plausible economic stories of female labor supply.

A Proofs of the Main Results

A.1 Proof of Lemma 1

First, note that as the first-order conditions for the constrained entropy minimization problem, we have

$$1 + \ln f(u \mid z) = \gamma' g(u, z) + \phi(z), \tag{A.1}$$

where $f(u \mid z) \equiv d\mu(u \mid z)/d\rho(u \mid z)$ for all $u \in \mathcal{U}$ and $z \in \mathcal{Z}$, γ is the Lagrange multiplier vector for the equality constraints $\mathrm{E}_{\mu \times \pi}[g(U, Z; \theta)] = 0$, and ϕ is the Lagrange multiplier function for the equality constraints $\int f(u \mid z) d\rho(u \mid z) = 1$ for all $u \in \mathcal{U}$ and $z \in \mathcal{Z}$. Solving the system of these equations yields

$$f(u \mid z) = \tilde{f}(u \mid z; \theta, \gamma) \tag{A.2}$$

for all $u \in \mathcal{U}$ and $z \in \mathcal{Z}$, where $\tilde{f}(u \mid z;\theta,\gamma) \equiv e^{\gamma' g(u,z;\theta)} \big/ \int e^{\gamma' g(v,z;\theta)} d\rho(v \mid z;\theta)$. For brevity, we also use the short-hand notation $\tilde{h}_j(u,z;\theta) \equiv h_j(u,z;\theta) / \int h_j(v,z;\theta) d\rho(v \mid z;\theta)$ for each $j \in \{1,\cdots,d_g\}$, and $\tilde{h}(u,z;\theta) \equiv \left(\tilde{h}_1(u,z;\theta),\cdots,\tilde{h}_{d_g}(u,z;\theta)\right)'$.

Take the product of (A.1) and $\tilde{h}_j(u,z;\theta)$, and integrate this product with respect to $\rho \times \pi$ to get

$$\mathrm{E}_{\rho\times\pi}\left[(1+\ln f(U \mid Z))\,\tilde{h}_j(U,Z;\theta)\right] \;=\; \gamma' \,\mathrm{E}_{\rho\times\pi}\left[g(U,Z;\theta)\tilde{h}_j(U,Z;\theta)\right] + \mathrm{E}_\pi\left[\phi(Z)\right]$$

where we use $\mathrm{E}_{\rho\times\pi}\left[\phi(Z)\tilde{h}_j(U,Z;\theta)\right] = \mathrm{E}_\pi\left[\phi(Z)\,\mathrm{E}_\rho\left[\tilde{h}_j(U,Z;\theta) \mid Z\right]\right] = \mathrm{E}_\pi\left[\phi(Z)\right]$. Likewise, take the product of (A.1) and $\tilde{f}(u \mid z;\theta,\gamma)$, and integrate this product with respect to $\rho \times \pi$ to get

$$\mathrm{E}_{\rho\times\pi}\left[(1+\ln f(U \mid Z))\,\tilde{f}(U \mid Z;\theta,\gamma)\right] \;=\; \gamma' \,\mathrm{E}_\pi\left[\tilde{g}(Z;\theta,\gamma)\right] + \mathrm{E}_\pi\left[\phi(Z)\right]$$

where we use $\mathrm{E}_{\rho\times\pi}\left[g(U,Z;\theta)\tilde{f}(U \mid Z;\theta,\gamma)\right] = \mathrm{E}_\mu\left[\tilde{g}(Z;\theta,\gamma)\right]$ and $\mathrm{E}_{\rho\times\pi}\left[\phi(Z)\tilde{f}(U \mid Z;\theta,\gamma)\right] = \mathrm{E}_\pi\left[\phi(Z)\,\mathrm{E}_\rho\left[\tilde{f}(U \mid Z;\theta,\gamma) \mid Z\right]\right] = \mathrm{E}_\pi\left[\phi(Z)\right]$. Taking the difference of the above two equations yields

$$\begin{aligned}
&\mathrm{E}_{\rho\times\pi}\left[(1+\ln f(U \mid Z))\left(\tilde{h}_j(U,Z;\theta) - \tilde{f}(U \mid Z;\theta,\gamma)\right)\right] \\
&\quad = \gamma' \left\{\mathrm{E}_{\rho\times\pi}\left[g(U,Z;\theta)\tilde{h}_j(U,Z;\theta)\right] - \mathrm{E}_\pi\left[\tilde{g}(Z;\theta,\gamma)\right]\right\}
\end{aligned}$$

Substituting (A.2) thus yields

$$\begin{aligned}
\mathrm{E}_{\rho\times\pi}&\left[\left(1+\ln\left(\frac{e^{\gamma' g(U,Z;\theta)}}{\int e^{\gamma' g(v,Z;\theta)} d\rho(v \mid Z;\theta)}\right)\right)\left(\tilde{h}_j(U,Z;\theta) - \frac{e^{\gamma' g(U,Z;\theta)}}{\int e^{\gamma' g(v,Z;\theta)} d\rho(v \mid Z;\theta)}\right)\right] \\
&= \gamma' \left\{\mathrm{E}_{\rho\times\pi}\left[g(U,Z;\theta)\tilde{h}_j(U,Z;\theta)\right] - \mathrm{E}_\pi\left[\tilde{g}(Z;\theta,\gamma)\right]\right\}
\end{aligned}$$

Now, in order to prove the lemma, observe from the last equation that $\mathrm{E}_\pi\left[\tilde{g}(Z;\theta,\gamma)\right] = 0$ holds if and only if

$$\begin{aligned}
\mathrm{E}_{\rho\times\pi}&\left[\left(1+\ln\left(\frac{e^{\gamma' g(U,Z;\theta)}}{\int e^{\gamma' g(v,Z;\theta)} d\rho(v \mid Z;\theta)}\right)\right)\left(\tilde{h}_j(U,Z;\theta) - \frac{e^{\gamma' g(U,Z;\theta)}}{\int e^{\gamma' g(v,Z;\theta)} d\rho(v \mid Z;\theta)}\right)\right] \\
&= \gamma' \,\mathrm{E}_{\rho\times\pi}\left[g(U,Z;\theta)\tilde{h}_j(U,Z;\theta)\right]
\end{aligned}$$

for each $j \in \{1, \cdots, d_g\}$, which is in turn equivalent to

$$\gamma = \mathrm{E}_{\rho\times\pi}\left[\tilde{h}(U,Z;\theta)g(U,Z;\theta)'\right]^{-1} \times$$
$$\mathrm{E}_{\rho\times\pi}\left[\begin{array}{c} \left(1+\ln\left(\frac{e^{\gamma' g(U,Z;\theta)}}{\int e^{\gamma' g(\upsilon,Z;\theta)}d\rho(\upsilon|Z;\theta)}\right)\right)\left(\tilde{h}_1(U,Z;\theta)-\frac{e^{\gamma' g(U,Z;\theta)}}{\int e^{\gamma' g(\upsilon,Z;\theta)}d\rho(\upsilon|Z;\theta)}\right) \\ \vdots \\ \left(1+\ln\left(\frac{e^{\gamma' g(U,Z;\theta)}}{\int e^{\gamma' g(\upsilon,Z;\theta)}d\rho(\upsilon|Z;\theta)}\right)\right)\left(\tilde{h}_{d_g}(U,Z;\theta)-\frac{e^{\gamma' g(U,Z;\theta)}}{\int e^{\gamma' g(\upsilon,Z;\theta)}d\rho(\upsilon|Z;\theta)}\right) \end{array}\right],$$

under Condition 1 (i). This completes a proof of the lemma. $\square$

A.2 Proof of Proposition 1

Suppose that $\gamma^* \in \Gamma$ is a fixed point of $\Psi(\,\cdot\,, \theta; g, h, \rho, \pi)$. Then,

$$\begin{aligned}
\lim_{N\to\infty} P(\gamma^*_\varepsilon(\theta; g, h, \rho, \widehat{\pi}_N) \neq \emptyset) &\geqslant \lim_{N\to\infty} P(\gamma^* \in \gamma^*_\varepsilon(\theta; g, h, \rho, \widehat{\pi}_N)) \\
&= \lim_{N\to\infty} P(\|\Psi(\gamma^*, \theta; g, h, \rho, \widehat{\pi}_N) - \gamma^*\| < \varepsilon) \\
&= \lim_{N\to\infty} P(\|\Psi(\gamma^*, \theta; g, h, \rho, \widehat{\pi}_N) - \Psi(\gamma^*, \theta; g, h, \rho, \pi)\| < \varepsilon) \\
&\geqslant \lim_{N\to\infty} P\left(\sup_{\gamma\in\Gamma} \|\Psi(\gamma, \theta; g, h, \rho, \widehat{\pi}_N) - \Psi(\gamma, \theta; g, h, \rho, \pi)\| < \varepsilon\right) = 1,
\end{aligned}$$

where the first equality is due to the definition of $\gamma^*_\varepsilon(\theta; g, h, \rho, \widehat{\pi}_N)$, the second equality is due to the definition of γ^* as the fixed point of $\Psi(\,\cdot\,, \theta; g, h, \rho, \pi)$, and the last equality follows from Proposition 3 under Assumption 1 and Conditions 1 and 2.

Conversely, suppose that $\Psi(\,\cdot\,, \theta; g, h, \rho, \pi)$ has no fixed point in Γ. In this case, we have $\|\Psi(\gamma, \theta; g, h, \rho, \pi) - \gamma\| \neq 0$ for all $\gamma \in \Gamma$. Since $\Psi(\,\cdot\,, \theta; g, h, \rho, \pi)$ is continuous on Γ and Γ is compact, there exists $\min_{\gamma\in\Gamma} \|\Psi(\gamma, \theta; g, h, \rho, \pi) - \gamma\| \in (0, \infty)$. In this light, we let $\bar{\varepsilon} \equiv$

$\frac{1}{2}\min_{\gamma\in\Gamma}\|\Psi(\gamma,\theta;g,h,\rho,\pi)-\gamma\|$ and $\varepsilon\in(0,\bar{\varepsilon})$. Note that $\gamma\in\gamma_{\varepsilon}^{*}(\theta;g,h,\rho,\widehat{\pi}_N)$ implies

$$\begin{aligned}
2\varepsilon &\leqslant \|\Psi(\gamma,\theta;g,h,\rho,\pi)-\gamma\| \\
&\leqslant \|\Psi(\gamma,\theta;g,h,\rho,\widehat{\pi}_N)-\Psi(\gamma,\theta;g,h,\rho,\pi)\|+\|\Psi(\gamma,\theta;g,h,\rho,\widehat{\pi}_N)-\gamma\| \\
&< \|\Psi(\gamma,\theta;g,h,\rho,\widehat{\pi}_N)-\Psi(\gamma,\theta;g,h,\rho,\pi)\|+\varepsilon.
\end{aligned}$$

In other words, $\gamma_{\varepsilon}^{*}(\theta;g,h,\rho,\widehat{\pi}_N)\neq\emptyset$ implies $\varepsilon<\sup_{\gamma\in\Gamma}\|\Psi(\gamma,\theta;g,h,\rho,\widehat{\pi}_N)-\Psi(\gamma,\theta;g,h,\rho,\pi)\|$. Therefore,

$$\begin{aligned}
\lim_{N\to\infty}P(\gamma_{\varepsilon}^{*}(\theta;g,h,\rho,\widehat{\pi}_N)\neq\emptyset) &\leqslant \lim_{N\to\infty}P\left(\sup_{\gamma\in\Gamma}\|\Psi(\gamma,\theta;g,h,\rho,\widehat{\pi}_N)-\Psi(\gamma,\theta;g,h,\rho,\pi)\|>\varepsilon\right) \\
&= 0
\end{aligned}$$

follows from Proposition 3 under Assumption 1 and Conditions 1 and 2. □

A.3 Proof of Proposition 2

Given the auxiliary result presented in Proposition 6, I prove this proposition for large parts by following the arguments in the proof (Kasahara and Shimotsu, 2012b; pp. 1–2) of Proposition 1 of Kasahara and Shimotsu (2012a).

Proof. Let $\delta>0$ such that $\sigma(D_{\gamma}\Psi(\gamma^{*},\theta;g,h,\rho,\pi))+2\delta<1$. There exists a matrix norm $\|\cdot\|_M$ such that $\|D_{\gamma}\Psi(\gamma^{*},\theta;g,h,\rho,\pi)'\|_M\leqslant\sigma(D_{\gamma}\Psi(\gamma^{*},\theta;g,h,\rho,\pi))+\delta$. Define the vector norm $\|\cdot\|_v$ on $\mathbb{R}^{d_g}$ by $\|\gamma\|_v\equiv\|[\gamma\;0\ldots0]\|_M$. Since $\gamma^{*}\in\text{int}\Gamma$, we have $\mathcal{N}\equiv\left\{\gamma\in\mathbb{R}^{d_g}:\|\gamma-\gamma^{*}\|_v<\delta\right\}$ $\subset\Gamma$ (by choosing δ small enough at the beginning). Note that this set $\mathcal{N}$ is a neighborhood of γ^{*} because any norm on $\mathbb{R}^{d_g}$ induces the equivalent topology. If $\widehat{\gamma}^{\iota-1}\in\mathcal{N}$, then we have $\|\widehat{\gamma}^{\iota-1}-\widehat{\gamma}_N\|_v\leqslant\|\widehat{\gamma}^{\iota-1}-\gamma^{*}\|_v+\|\gamma^{*}-\widehat{\gamma}_N\|_v<\delta$ almost surely by Proposition 4 under Assumptions 1 and 2 and Conditions 1 and 2. Furthermore, by Proposition 6 under Assumptions 1 and 2 and Conditions 1, 2 and 3, we can write $\|\widehat{\gamma}^{\iota}-\widehat{\gamma}_N\|_v\leqslant\|D_{\gamma}\Psi(\gamma^{*},\theta;g,h,\rho,\pi)\|_M\|(\widehat{\gamma}^{\iota-1}-\widehat{\gamma}_N)\|_v+$

$\left\|\widehat{Q}_N(\widehat{\gamma}^{\iota-1}-\widehat{\gamma}_N)\right\|_v$. By our choice of the matrix norm $\|\cdot\|_M$, the first term can be bounded as $\|D_\gamma\Psi(\gamma^*,\theta;g,h,\rho,\pi)\|_M \|(\widehat{\gamma}^{\iota-1}-\widehat{\gamma}_N)\|_v \leqslant (\sigma(D_\gamma\Psi(\gamma^*,\theta;g,h,\rho,\pi))+\delta)\|(\widehat{\gamma}^{\iota-1}-\widehat{\gamma}_N)\|_v$. From an earlier argument, together with the equivalence of all the norms on $\mathbb{R}^{d_g}$ and Proposition 6, the second term can be bounded as $\left\|\widehat{Q}_N(\widehat{\gamma}^{\iota-1}-\widehat{\gamma}_N)\right\|_v \leqslant o(1)\|\widehat{\gamma}^{\iota-1}-\widehat{\gamma}_N\|_v + O(1)\|\widehat{\gamma}^{\iota-1}-\widehat{\gamma}_N\|_v^2 \leqslant \delta\|\widehat{\gamma}^{\iota-1}-\widehat{\gamma}_N\|_v$ almost surely. Combining the above inequalities together, we get $\|\widehat{\gamma}^{\iota}-\widehat{\gamma}_N\|_v \leqslant (\sigma(D_\gamma\Psi(\gamma^*,\theta;g,h,\rho,\pi))+2\delta)\ \|(\widehat{\gamma}^{\iota-1}-\widehat{\gamma}_N)\|_v$ almost surely, where $\sigma(D_\gamma\Psi(\gamma^*,\theta;g,h,\rho,\pi))+2\delta<1$. Therefore, it follows that $\lim_{\iota\to\infty}\widehat{\gamma}^{\iota}=\widehat{\gamma}_N$ almost surely if $\widehat{\gamma}^0\in\mathcal{N}$. $\square$

B Auxiliary Results and Their Proofs

B.1 Uniform Consistency of $\Psi(\,\cdot\,,\theta;g,h,\rho,\widehat{\pi}_N)$ for $\Psi(\,\cdot\,,\theta;g,h,\rho,\pi)$

The following auxiliary proposition shows the uniform consistency of $\Psi(\,\cdot\,,\theta;g,h,\rho,\widehat{\pi}_N)$ for $\Psi(\,\cdot\,,\theta;g,h,\rho,\pi)$ on compact $\Gamma\subset\mathbb{R}^{d_g}$. This result is used in turn to prove Propositions 1 and 4.

Proposition 3. *Suppose that Assumption 1 holds and that the auxiliary functions h and the auxiliary measure ρ are chosen subject to Conditions 1 and 2. We have the uniform convergence $\sup_{\gamma\in\Gamma}\|\Psi(\gamma,\theta;g,h,\rho,\widehat{\pi}_N)-\Psi(\gamma,\theta;g,h,\rho,\pi)\| \xrightarrow{a.s.} 0$ as $N\to\infty$ if Γ is a compact subset of $\mathbb{R}^{d_g}$.*

Proof. We can rewrite the self map succinctly as

$$\Psi(\gamma,\theta;g,h,\rho,\pi)=\left[\int\varphi(z;\theta)d\pi(z)\right]^{-1}\int\Phi(z;\theta,\gamma)d\pi(z),$$

where $\varphi(z;\theta) \equiv \left[\int g(u,z;\theta)\left(\frac{h_1(u,z;\theta)}{\int h_1(u,z;\theta)d\rho(u|z;\theta)},\cdots,\frac{h_{d_g}(u,z;\theta)}{\int h_{d_g}(u,z;\theta)d\rho(u|z;\theta)}\right)d\rho(u\mid z)\right]'$ and

$$\Phi(z;\theta,\gamma) \equiv \begin{bmatrix} \int\left(1+\ln\left(\frac{e^{\gamma' g(u,z;\theta)}}{\int e^{\gamma' g(v,z;\theta)}d\rho(v|z;\theta)}\right)\right)\left(\frac{h_1(u,z;\theta)}{\int h_1(u,z;\theta)d\rho(u|z;\theta)} - \frac{e^{\gamma' g(u,z;\theta)}}{\int e^{\gamma' g(v,z;\theta)}d\rho(v|z;\theta)}\right)d\rho(u\mid z) \\ \vdots \\ \int\left(1+\ln\left(\frac{e^{\gamma' g(u,z;\theta)}}{\int e^{\gamma' g(v,z;\theta)}d\rho(v|z;\theta)}\right)\right)\left(\frac{h_{d_g}(u,z;\theta)}{\int h_{d_g}(u,z;\theta)d\rho(u|z;\theta)} - \frac{e^{\gamma' g(u,z;\theta)}}{\int e^{\gamma' g(v,z;\theta)}d\rho(v|z;\theta)}\right)d\rho(u\mid z) \end{bmatrix}.$$

First, note that $\int \varphi_{r,c}(z;\theta)d\widehat{\pi}_N(z) \xrightarrow{a.s.} \int \varphi_{r,c}(z;\theta)d\pi(z)$ holds for each row and column $r,c \in \{1,\cdots,d_g\}$ under Assumption 1 and Condition 2 (i) by Kolmogorov's strong law of large numbers. Second, we show the uniform consistency of $\int \Phi_j(z;\theta,\,\cdot\,)d\widehat{\pi}_N(z)$ to $\int \Phi_j(z;\theta,\,\cdot\,)d\pi(z)$ for each coordinate $j \in \{1,\cdots,d_g\}$. To this end, we note that Γ is compact, that $\Phi_j(\,\cdot\,;\theta,\,\cdot\,)$ is continuous under Condition 2 (ii), and that $\Phi_j(\,\cdot\,;\theta,\gamma)$ is dominated by an $L^1(\pi)$ function uniformly for all $\gamma \in \Gamma$ under Condition 2 (iii). Therefore, by the uniform law of large numbers, we have $\sup_{\gamma\in\Gamma}\left|\int \Phi_j(z;\theta,\gamma)d\widehat{\pi}_N(z) - \int \Phi_j(z;\theta,\gamma)d\pi(z)\right| \xrightarrow{a.s.} 0$ under Assumption 1. An application of the continuous mapping theorem under Condition 1 (i) completes a proof of the proposition. □

B.2 Almost Sure Convergence of $\widehat{\gamma}_N$ for γ^*

The following auxiliary proposition shows the almost sure convergence of the fixed point estimate $\widehat{\gamma}_N$ for the population fixed point γ^*. This auxiliary result is in turn used to prove the iteration expansion in Proposition 6. Furthermore, this result, together with Proposition 6, is also used to prove the contraction property in Proposition 2.

Proposition 4. *Suppose that Assumptions 1 and 2 hold and that the auxiliary functions h and the auxiliary measure ρ are chosen subject to Conditions 1 and 2. We have $\widehat{\gamma}_N \xrightarrow{a.s.} \gamma^*$ as $N \to \infty$.*

Proof. Let $\varepsilon > 0$. Because $\frac{\partial}{\partial\gamma_j}\Psi(\gamma^*,\theta;g,h,\rho,\pi) \neq e_j$ for each $j \in \{1,\cdots,d_g\}$ under Assumption 2 (ii) and $\Psi(\,\cdot\,,\theta; g,h,\rho,\pi)$ is continuously differentiable on Γ, there exist $r,c \in (0,\infty)$ such that $\|\Psi(\gamma,\theta;g,h,\rho,\pi)-\gamma\| \geqslant c\,\|\gamma-\gamma^*\|$ holds for all $\gamma \in B_r(\gamma^*) \equiv \{\gamma \in \Gamma : \|\gamma-\gamma^*\| < r\}$. Furthermore, since Γ is compact, $\Gamma\backslash B_r(\gamma^*)$ is also compact, and there exists $\delta(r) \equiv \min_{\gamma\in\Gamma\backslash B_r(\gamma^*)} \|\Psi(\gamma,\theta;g,h,\rho,\pi)-\gamma\|$. By Assumption 2 (i), $\delta(r)>0$ is true. Therefore, it follows that $\{\gamma \in \Gamma : \|\Psi(\gamma,\theta;g,h,\rho,\pi)-\gamma\| < \min\{c\varepsilon,\delta(r)\}\} \subset \{\gamma\in\Gamma : \|\gamma-\gamma^*\| < \varepsilon\}$ holds. This implies

$$
\begin{aligned}
&\left\{\omega\in\Omega : \sup_{\gamma\in\Gamma}\|\Psi(\gamma,\theta;g,h,\rho,\widehat{\pi}_N(\omega)) - \Psi(\gamma,\theta;g,h,\rho,\pi)\| < \min\{c\varepsilon,\delta(r)\}\right\}\\
\subset\;\; & \{\omega\in\Omega : \|\Psi(\widehat{\gamma}_N(\omega),\theta;g,h,\rho,\widehat{\pi}_N(\omega)) - \Psi(\widehat{\gamma}_N(\omega),\theta;g,h,\rho,\pi)\| < \min\{c\varepsilon,\delta(r)\}\}\\
=\;\; & \{\omega\in\Omega : \|\widehat{\gamma}_N(\omega) - \Psi(\widehat{\gamma}_N(\omega),\theta;g,h,\rho,\pi)\| < \min\{c\varepsilon,\delta(r)\}\} \subset \{\omega\in\Omega : \|\widehat{\gamma}_N(\omega)-\gamma^*\| < \varepsilon\},
\end{aligned}
$$

where the equality $\Psi(\widehat{\gamma}_N,\theta;g,h,\rho,\widehat{\pi}_N) = \widehat{\gamma}_N$, that follows from the definition of $\widehat{\gamma}_N$ as the fixed point of $\Psi(\,\cdot\,,\theta;g,h,\rho,\widehat{\pi}_N)$, is used. Therefore, we obtain

$$
\begin{aligned}
&P\left(\liminf_{N\to\infty}\{\omega\in\Omega : \|\widehat{\gamma}_N(\omega)-\gamma^*\| < \varepsilon\}\right) \geqslant\\
&P\left(\liminf_{N\to\infty}\left\{\omega\in\Omega : \sup_{\gamma\in\Gamma}\|\Psi(\gamma,\theta;g,h,\rho,\widehat{\pi}_N(\omega)) - \Psi(\gamma,\theta;g,h,\rho,\pi)\| < \min\{c\varepsilon,\delta(r)\}\right\}\right) = 1,
\end{aligned}
$$

where the last equality follows from Proposition 3 under Assumption 1 and Conditions 1 and 2. This completes a proof that $\widehat{\gamma}_N \xrightarrow{a.s.} \gamma^*$ as $N\to\infty$. $\square$

B.3 Uniform Consistency of $D_\gamma^\alpha\Psi(\,\cdot\,,\theta;g,h,\rho,\widehat{\pi}_N)$ for $D_\gamma^\alpha\Psi(\,\cdot\,,\theta;g,h,\rho,\pi)$

The following auxiliary proposition shows the uniform consistency of $D_\gamma^\alpha\Psi(\,\cdot\,,\theta;g,h,\rho,\widehat{\pi}_N)$ for $D_\gamma^\alpha\Psi(\,\cdot\,,\theta;g,h,\rho,\pi)$ for any partial derivation operator $D_\gamma^\alpha = \frac{\partial^{|\alpha|}}{\partial\gamma^\alpha}$ with $\alpha\in\mathbb{Z}_+^{d_g}$ and $|\alpha|\leqslant 2$ on compact $\Gamma\subset\mathbb{R}^{d_g}$. This auxiliary result is in turn used to prove the iteration expansion in Proposition 6.

Proposition 5. *Suppose that Assumption 1 holds and that the auxiliary functions h and the auxiliary measure ρ are chosen subject to Conditions 1, 2 (i) and 3. We have the uniform convergence* $\sup_{\gamma\in\Gamma}\left\|D_\gamma^\alpha\Psi(\gamma,\theta;g,h,\rho,\widehat{\pi}_N) - D_\gamma^\alpha\Psi(\gamma,\theta;g,h,\rho,\pi)\right\| \xrightarrow{a.s.} 0$ *as* $N\to\infty$ *for any partial derivation operator* $D_\gamma^\alpha = \frac{\partial^\alpha}{\partial\gamma^\alpha}$ *with* $\alpha\in\mathbb{Z}_+^{d_g}$ *and* $|\alpha|\leqslant 2$ *if* Γ *is a compact subset of* $\mathbb{R}^{d_g}$.

Proof. Let $D_\gamma^\alpha = \frac{\partial^\alpha}{\partial\gamma^\alpha}$ denote a partial derivation operator with $\alpha\in\mathbb{Z}_+^{d_g}$ and $|\alpha|\leqslant 2$. We can rewrite the derivative of the self map with respect to γ succinctly as

$$D_\gamma^\alpha\Psi(\gamma,\theta;g,h,\rho,\pi) = \left[\int\varphi(z;\theta)d\pi(z)\right]^{-1}\int D_\gamma^\alpha\Phi(z;\theta,\gamma)d\pi(z)$$

under Condition 3 (ii), where φ and Φ are defined in the same way as in the proof of Proposition 3.

Note first that we have the consistency $\int\varphi_{r,c}(z;\theta)d\widehat{\pi}_N(z) \xrightarrow{a.s.} \int\varphi_{r,c}(z;\theta)d\pi(z)$ holds for each row and column $r,c\in\{1,\cdots,d_g\}$ under Assumption 1 and Condition 2 (i) by the same lines of the argument as in the proof of Proposition 3. Note second that Γ is compact, that $D_\gamma^\alpha\Phi_j(\ \cdot\ ;\theta,\ \cdot\ )$ is continuous under Condition 3 (i), and that $D_\gamma^\alpha\Phi_j(\ \cdot\ ;\theta,\gamma)$ is dominated by an $L^1(\pi)$ function uniformly for all $\gamma\in\Gamma$ under Condition 3 (ii). Therefore, we have $\sup_{\gamma\in\Gamma}\left|\int D_\gamma^\alpha\Phi_j(z;\theta,\gamma)d\widehat{\pi}_N(z) - \int D_\gamma^\alpha\Phi_j(z;\theta,\gamma)d\pi(z)\right| \xrightarrow{a.s.} 0$ by the uniform law of large numbers under Assumption 1 for each coordinate $j\in\{1,\cdots,d_g\}$. An application of the continuous mapping theorem under Condition 1 (i) completes a proof of the proposition. □

B.4 Iteration Expansion

The following auxiliary proposition shows an iteration expansion of the sample counterpart self map. This result, together with Proposition 4, is used in turn to prove the contraction property in Proposition 2.

Proposition 6. *Suppose that Assumptions 1 and 2 hold and that the auxiliary functions h and the auxiliary measure ρ are chosen subject to Conditions 1, 2 and 3. Let $\widehat{\gamma}^{\iota}$ denote the vector obtained in the ι-th iteration of $\Psi(\,\cdot\,, \theta; g, h, \rho, \widehat{\pi}_N)$ starting from $\widehat{\gamma}^0 \in \Gamma$. If $\widehat{\gamma}^{\iota-1} \in \Gamma$, then we have $\widehat{\gamma}^{\iota} - \widehat{\gamma}_N = [D_{\gamma}\Psi(\gamma^*, \theta; g, h, \rho, \pi)]' (\widehat{\gamma}^{\iota-1} - \widehat{\gamma}_N) + \widehat{Q}_N(\widehat{\gamma}^{\iota-1} - \widehat{\gamma}_N)$, where D_{γ} denotes the gradient operator and $\left\|\widehat{Q}_N(\widehat{\gamma}^{\iota-1} - \widehat{\gamma}_N)\right\| = o(1) \|\widehat{\gamma}^{\iota-1} - \widehat{\gamma}_N\| + O(1) \|\widehat{\gamma}^{\iota-1} - \widehat{\gamma}_N\|^2$ almost surely.*

Proof. Since $\widehat{\gamma}_N \in \Gamma$ is the fixed point of $\Psi(\,\cdot\,, \theta; g, h, \rho, \widehat{\pi}_N)$ and $\Psi(\,\cdot\,, \theta; g, h, \rho, \widehat{\pi}_N)$ is twice continuously differentiable, we can write the j-th coordinate of the difference $\widehat{\gamma}^{\iota} - \widehat{\gamma}_N$ by Taylor expansion as

$$
\begin{aligned}
&\widehat{\gamma}^{\iota}_j - \widehat{\gamma}_{N,j} = \Psi_j(\widehat{\gamma}^{\iota-1}, \theta; g, h, \rho, \widehat{\pi}_N) - \widehat{\gamma}_{N,j} = \Psi_j(\widehat{\gamma}^{\iota-1}, \theta; g, h, \rho, \widehat{\pi}_N) - \Psi_j(\widehat{\gamma}_N, \theta; g, h, \rho, \widehat{\pi}_N) \\
&= [D_{\gamma}\Psi_j(\widehat{\gamma}_N, \theta; g, h, \rho, \widehat{\pi}_N)]' (\widehat{\gamma}^{\iota-1} - \widehat{\gamma}_N) + (\widehat{\gamma}^{\iota-1} - \widehat{\gamma}_N)' \widehat{R}_j(\widehat{\gamma}^{\iota-1}, \theta; g, h, \rho, \widehat{\pi}_N)(\widehat{\gamma}^{\iota-1} - \widehat{\gamma}_N)
\end{aligned}
$$

where the function $\widehat{R}_j(\,\cdot\,, \theta; g, h, \rho, \widehat{\pi}_N) : \Gamma \to \mathcal{M}(d_g, d_g)$ is bounded as $\left|\widehat{R}_j(\gamma, \theta; g, h, \rho, \widehat{\pi}_N)_{(r,c)}\right| \leqslant \frac{1}{2} \max_{|\alpha|=2} \sup_{\tilde{\gamma}\in\Gamma} \left|D^{\alpha}_{\gamma}\Psi_j(\tilde{\gamma}, \theta; g, h, \rho, \widehat{\pi}_N)\right|$ for all $\gamma \in \Gamma$ for each row and column $r, c \in \{1, \cdots, d_g\}$. Furthermore, since $D_{\gamma}\Psi_j(\,\cdot\,, \theta; g, h, \rho, \pi)$ is continuously differentiable on Γ, we can use Taylor expansion to write

$$
D_{\gamma}\Psi_j(\widehat{\gamma}_N, \theta; g, h, \rho, \pi) = D_{\gamma}\Psi_j(\gamma^*, \theta; g, h, \rho, \pi) + R_j(\widehat{\gamma}_N, \theta; g, h, \rho, \pi)(\widehat{\gamma}_N - \gamma^*)
$$

where the function $R_j(\,\cdot\,, \theta; g, h, \rho, \pi) : \Gamma \to \mathcal{M}(d_g, d_g)$ is bounded as $\left|R_j(\gamma)_{(r,c)}\right| \leqslant \frac{1}{2} \max_{|\alpha|=2} \sup_{\tilde{\gamma}\in\Gamma} \left|D^{\alpha}_{\gamma}\Psi_j(\tilde{\gamma}, \theta; g, h, \rho, \pi)\right|$ for all $\gamma \in \Gamma$ for each row and column $r, c \in \{1, \cdots, d_g\}$, and $\widehat{\gamma}_N - \gamma^* = o(1)$ almost surely by Proposition 4 under Assumptions 1 and 2 as well as Conditions 1 and 2.

Combining all above, we can now express the j-th coordinate of the difference $\widehat{\gamma}^{\iota} - \widehat{\gamma}_N$ by

$$\begin{aligned}
\widehat{\gamma}_j^{\iota} - \widehat{\gamma}_{N,j} &= [D_\gamma \Psi_j(\gamma^*, \theta; g, h, \rho, \pi)]' (\widehat{\gamma}^{\iota-1} - \widehat{\gamma}_N) \\
&\quad + [D_\gamma \Psi_j(\widehat{\gamma}_N, \theta; g, h, \rho, \widehat{\pi}_N) - D_\gamma \Psi_j(\widehat{\gamma}_N, \theta; g, h, \rho, \pi)]' (\widehat{\gamma}^{\iota-1} - \widehat{\gamma}_N) \\
&\quad + [R_j(\widehat{\gamma}_N, \theta; g, h, \rho, \pi)(\widehat{\gamma}_N - \gamma^*)]' (\widehat{\gamma}^{\iota-1} - \widehat{\gamma}_N) \\
&\quad + (\widehat{\gamma}^{\iota-1} - \widehat{\gamma}_N)' \widehat{R}_j(\widehat{\gamma}^{\iota-1}, \theta; g, h, \rho, \widehat{\pi}_N)(\widehat{\gamma}^{\iota-1} - \widehat{\gamma}_N).
\end{aligned}$$

Note first that we have

$$\left| [D_\gamma \Psi_j(\widehat{\gamma}_N, \theta; g, h, \rho, \widehat{\pi}_N) - D_\gamma \Psi_j(\widehat{\gamma}_N, \theta; g, h, \rho, \pi)]_r \right| = o(1)$$

almost surely for each row $r \in \{1, \cdots, d_g\}$ by Proposition 5 under Assumption 1 and Conditions 1, 2 (i) and 3. Second, note that

$$\left| [R_j(\widehat{\gamma}_N, \theta; g, h, \rho, \pi)(\widehat{\gamma}_N - \gamma^*)]_r \right| = o(1)$$

almost surely for each row $r \in \{1, \cdots, d_g\}$ by Proposition 4 under Assumptions 1 and 2 as well as Conditions 1 and 2. Finally, note that

$$\begin{aligned}
\left| \widehat{R}_j(\widehat{\gamma}^{\iota-1}, \theta; g, h, \rho, \widehat{\pi}_N)_{(r,c)} \right| &\leqslant \frac{1}{2} \max_{|\alpha|=2} \sup_{\gamma \in \Gamma} \left| D_\gamma^\alpha \Psi_j(\gamma, \theta; g, h, \rho, \pi) \right| \\
&\quad + \frac{1}{2} \max_{|\alpha|=2} \sup_{\gamma \in \Gamma} \left| D_\gamma^\alpha \Psi_j(\gamma, \theta; g, h, \rho, \widehat{\pi}_N) - D_\gamma^\alpha \Psi_j(\gamma, \theta; g, h, \rho, \pi) \right| = O(1)
\end{aligned}$$

almost surely for each row and column $r, c \in \{1, \cdots, d_g\}$ by Proposition 5 under Assumption 1 and Conditions 1, 2 (i) and 3. Therefore, letting $\|\cdot\|_\infty$ denote the max norm on $\mathbb{R}^{d_g}$, we obtain

$$\widehat{\gamma}_j^{\iota} - \widehat{\gamma}_{N,j} = [D_\gamma \Psi_j(\gamma^*, \theta; g, h, \rho, \pi)]' (\widehat{\gamma}^{\iota-1} - \widehat{\gamma}_N) + \widehat{Q}_{N,j}(\widehat{\gamma}^{\iota-1} - \widehat{\gamma}_N)$$

where

$$\left| \widehat{Q}_{N,j}(\widehat{\gamma}^{\iota-1} - \widehat{\gamma}_N) \right| = o(1) \left\| \widehat{\gamma}^{\iota-1} - \widehat{\gamma}_N \right\|_\infty + O(1) \left\| \widehat{\gamma}^{\iota-1} - \widehat{\gamma}_N \right\|_\infty^2$$

almost surely. Using the equivalence of norms on $\mathbb{R}^{d_g}$ completes a proof of the proposition. $\square$

C Inference

The literature provide a method of inference for finite dimensional parameters robustly against non-identification and infinite-dimensional nuisance parameters (e.g., Chen, Tamer and Torgovitsky, 2011). The fixed-point approach presented in this paper, on the other hand, reduces the infinite-dimensional nuisance parameters to finite-dimensional ones γ. Therefore, it is readily applicable to existing methods of set inference (e.g., Chernozhukov, Hong and Tamer, 2007; Romano and Shaikh, 2010) and the existing methods of identification-robust inference (e.g., Stock and Wright, 2000; Kleibergen, 2005; Andrews and Mikusheva, 2014) for subvectors, that are based on zeros of criterion functions. We focus on testing the null hypothesis $H_0 : \theta \in \Theta_0$. By Theorem 1, the null hypothesis $H_0 : \theta_0 \in \Theta_0$ leads to the fixed point restriction $\Psi(\gamma^*(\theta_0; g, h, \rho), \theta_0; g, h, \rho, \pi) = \gamma^*(\theta_0; g, h, \rho)$, where $\gamma^*(\theta; g, h, \rho)$ is the fixed point of the contraction mapping $\Psi(\,\cdot\,, \theta; g, h, \rho, \pi)$. In other words, we have $\mathrm{E}_\pi\left[\tilde{M}(Z, \gamma^*(\theta_0; g, h, \rho), \theta_0; g, h, \rho)\right]$ $= 0$ under the null hypothesis H_0, where

$$\tilde{M}(z,\gamma,\theta;g,h,\rho) \equiv \begin{bmatrix} \int \frac{h_1(u,z;\theta)}{\int h_1(v,z;\theta)d\rho(v|z;\theta)}\gamma' g(u,z;\theta) - \left(1+\ln\left(\frac{e^{\gamma' g(u,z;\theta)}}{\int e^{\gamma' g(v,z;\theta)}d\rho(v|z;\theta)}\right)\right)\left(\frac{h_1(u,z;\theta)}{\int h_1(v,z;\theta)d\rho(v|z;\theta)} - \frac{e^{\gamma' g(u,z;\theta)}}{\int e^{\gamma' g(v,z;\theta)}d\rho(v|z;\theta)}\right) d\rho(u \mid z) \\ \vdots \\ \int \frac{h_{d_g}(u,z;\theta)}{\int h_{d_g}(v,z;\theta)d\rho(v|z;\theta)}\gamma' g(u,z;\theta) - \left(1+\ln\left(\frac{e^{\gamma' g(u,z;\theta)}}{\int e^{\gamma' g(v,z;\theta)}d\rho(v|z;\theta)}\right)\right)\left(\frac{h_{d_g}(u,z;\theta)}{\int h_{d_g}(v,z;\theta)d\rho(v|z;\theta)} - \frac{e^{\gamma' g(u,z;\theta)}}{\int e^{\gamma' g(v,z;\theta)}d\rho(v|z;\theta)}\right) d\rho(u \mid z) \end{bmatrix}.$$

Define $G_N(\gamma,\theta;g,h,\rho) = \frac{1}{\sqrt{N}}\sum_{i=1}^N \tilde{M}(Z_i,\gamma,\theta;g,h,\rho)$. Note that $G_N(\widehat{\gamma}_N(\theta_0;g,h,\rho),\theta_0;g,h,\rho)$ $= 0$ is true for the sample-counterpart fixed point $\widehat{\gamma}_N(\theta_0;g,h,\rho)$. In particular, this implies that $\widehat{\gamma}_N(\theta_0;g,h,\rho) = \arg\min_{\gamma\in\Gamma} G_N(\gamma,\theta_0;g,h,\rho)' W G_N(\gamma,\theta_0;g,h,\rho)$ for any positive definite $d_g \times d_g$ matrix W.

We can thus use Andrews and Mikusheva (2014; Section 5) for inference of the subvector θ of $(\gamma', \theta')'$ – also see Kleibergen (2005; Section 3.2). The covariance function for the Gaussian process to which $\left((G_N(\gamma, \theta; g, h, \rho) - \mathrm{E}\, G_N(\gamma, \theta; g, h, \rho))', \sqrt{N}(\widehat{\gamma}_N(\theta; g, h, \rho) - \gamma^*(\theta; g, h, \rho))'\right)'$ asymptotically converges to can be estimated by

$$\hat{\hat{\Sigma}}(\gamma_1, \theta_1, \gamma_2, \theta_2; g, h, \rho) \equiv \left(I_{d_g}\ \hat{B}(\gamma_1, \theta_1; g, h, \rho)'\right)'\ \tilde{\Omega}(\gamma_1, \theta_1, \gamma_2, \theta_2; g, h, \rho) \left(I_{d_g}\ \hat{B}(\gamma_2, \theta_2; g, h, \rho)'\right),$$

where

$$\hat{B}(\gamma, \theta; g, h, \rho) \equiv \frac{1}{\sqrt{N}} \left([D_\gamma G_N(\gamma, \theta; g, h, \rho)]' W \left[D_\gamma G_N(\gamma, \theta; g, h, \rho)\right]\right)^{-1} [D_\gamma G_N(\gamma, \theta; g, h, \rho)]' W$$

and

$$\begin{aligned} \tilde{\Omega}(\gamma_1, \theta_1, \gamma_2, \theta_2; g, h, \rho) \equiv\ & \frac{1}{N} \sum_{i=1}^{N} \tilde{M}(Z_i, \gamma_1, \theta_1; g, h, \rho) \tilde{M}(Z_i, \gamma_2, \theta_2; g, h, \rho)' - \\ & \left[\frac{1}{N} \sum_{i=1}^{N} \tilde{M}(Z_i, \gamma_1, \theta_1; g, h, \rho)\right] \left[\frac{1}{N} \sum_{i=1}^{N} \tilde{M}(Z_i, \gamma_2, \theta_2; g, h, \rho)\right]'. \end{aligned}$$

Thus, the covariance function for the moment function $G_N(\widehat{\gamma}_N(\theta; g, h, \rho), \theta; g, h, \rho)$ can be estimated by

$$\begin{aligned} \hat{\Sigma}(\theta_1, \theta_2; g, h, \rho) =\ & \left(I_{d_g}\ [D_\gamma G_N(\widehat{\gamma}_N(\theta_1; g, h, \rho), \theta_1; g, h, \rho)]\right) \times \\ & \hat{\hat{\Sigma}}(\widehat{\gamma}_N(\theta_1; g, h, \rho), \theta_1, \widehat{\gamma}_N(\theta_2; g, h, \rho), \theta_2; g, h, \rho) \times \\ & \left(I_{d_g}\ [D_\gamma G_N(\widehat{\gamma}_N(\theta_2; g, h, \rho), \theta_2; g, h, \rho)]\right)', \end{aligned}$$

and we can then construct the quasi-likelihood ratio statistic

$$\begin{aligned} QLR(\theta_0; g, h, \rho) =\ & G_N(\widehat{\gamma}_N(\theta_0; g, h, \rho), \theta_0; g, h, \rho)' \hat{\Sigma}(\theta_0, \theta_0; g, h, \rho)^{-1} G_N(\widehat{\gamma}_N(\theta_0; g, h, \rho), \theta_0; g, h, \rho) \\ & - \inf_\theta G_N(\widehat{\gamma}_N(\theta; g, h, \rho), \theta; g, h, \rho)' \hat{\Sigma}(\theta, \theta; g, h, \rho)^{-1} G_N(\widehat{\gamma}_N(\theta; g, h, \rho), \theta; g, h, \rho) \end{aligned}$$

to test the hypothesis $H_0 : \theta = \theta_0$. In order to obtain the quantiles of the QLR statistic, generate a random sample of QLR statistics using

$$G^*_N(\theta; g, h, \rho) = H_N(\theta; g, h, \rho) + \hat{\Sigma}(\theta, \theta_0; g, h, \rho) \hat{\Sigma}(\theta_0, \theta_0; g, h, \rho)^{-1} \xi^*$$

where ξ^* is drawn from $N(0, \hat{\Sigma}(\theta_0, \theta_0; g, h, \rho))$, and $H_N(\theta; g, h, \rho) \equiv G_N(\widehat{\gamma}_N(\theta; g, h, \rho), \theta; g, h, \rho) - \hat{\Sigma}(\theta, \theta_0; g, h, \rho)\hat{\Sigma}(\theta_0, \theta_0; g, h, \rho)^{-1} G_N(\widehat{\gamma}_N(\theta_0; g, h, \rho), \theta_0; g, h, \rho)$. See Andrews and Mikusheva (2014) for a list of required assumptions for this method of inference.

References

Aguirregabiria, Victor and Pedro Mira (2007) "Sequential Estimation of Dynamic Discrete Games," *Econometrica,* 75, (1), 1–53.

Andrews, Isaiah and Anna Mikusheva (2014) "Conditional Inference with a Functional Nuisance Parameter," MIT Working Paper

Angrist, Joshua D. and William N. Evans (1998) "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," *American Economic Review,* 88 (3), 450–477.

Aridiacono, Peter and John Bailey Jones (2003) "Finite Mixture Distributions, Sequential Likelihood and the EM Algorithm," *Econometrica,* 71 (3), 933–946.

Başar, Tamer (1988) "Asynchronous Algorithms in Non-Cooperative Games," *Journal of Economic Dynamics and Control,* 12 (1), 167–172.

Bhattacharya, Debopam (2015) "Nonparametric Welfare Analysis for Discrete Choice," *Econometrica,* 83 (2), 617–649.

Browning, Martin (1992) "Children and Household Economic Behavior," *Journal of Economic Literature,* 30 (3), 1434–1475.

Chen, Xiaohong, Elie Tamer, and Alexander Torgovitsky (2011) "Sensitivity Analysis in Semiparametric Likelihood Models," Cowles Foundation Discussion Paper No. 1836.

Chernozhukov, Victor, Han Hong, and Elie Tamer (2007) "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica,* 75 (5), 1243–1284.

Feller, William (1943) "On a General Class of Contagious Distributions," *Annals of Mathematical Statistics,* 14 (4), 389–400.

Gautier, Eric and Yuichi Kitamura (2013) "Nonparametric Estimation in Random Coefficients Binary Choice Models," *Econometrica,* 81 (2), 581–607.

Hall, Peter and Xiao-Hua Zhou (2003) "Nonparametric Estimation of Component Distributions in a Multivariate Mixture," *Annals of Statistics*, 31 (1), 201–224.

Hausman, Jerry and Whitney Newey (2014) "Individual Heterogeneity and Average Welfare," Working Paper, MIT.

Heckman, James J. (1981a) "Heterogeneity and State Dependence," In: Rosen, S. (Ed.), Studies in Labor Markets. University of Chicago Press, 91–139.

Heckman, James J. (1981b) "The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time-Discrete Data Stochastic Process," In: Manski, C.F., McFadden, D. (Eds.), Structural Analysis of Discrete Data with Econometric Applications. MIT Press, 179–195.

Heckman, James J. and Burton Singer (1984) "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica,* 52 (2), 271–320.

Henry, Marc, Yuichi Kitamura and Benard Salanié (2014) "Partial Identification of Finite Mixtures in Econometric Models," *Quantitative Economics,* 5 (1), 123–144.

Honoré, Bo E. and Elie Tamer (2006) "Bounds on Parameters in Panel Dynamic Discrete Choice Models," *Econometrica,* 74 (3), 611–629.

Hotz, V. Joseph and Robert A. Miller (1988) "An Empirical Analysis of Life Cycle Fertility and Female Labor Supply," *Econometrica,* 56 (1), 91–118.

Hsiao, Cheng, and M. Hashem Pesaran (2008) "Random Coefficient Models," *The Econometrics of Panel Data, Advanced Studies in Theoretical and Applied Econometrics,* Vol. 46, pp. 185–213.

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

Hu, Yingyao and Susanne Schennach (2008) "Instrumental Variable Treatment of Nonclassical Measurement Error Models," *Econometrica,* 76 (1), 195–216.

Hyslop, Dean R. (1999) "State Dependence, Serial Correlation and Heterogeneity in Intertemporal Labor Force Participation of Married Women," *Econometrica,* 67 (6), 1255–1294.

Ichimura, Hidehiko and T. Scott Thompson (1998) "Maximum Likelihood Estimation of a Binary Choice Model with Random Coefficients of Unknown Distribution," *Journal of Econometrics,* 86 (2), 269–295.

Jakubson, George (1988) "The Sensitivity of Labor-Supply Parameter Estimates to Unobserved

Individual Effects: Fixed- and Random-Effects Estimates in a Nonlinear Model Using Panel Data," *Journal of Labor Economics,* 6 (3), 302–329.

Judd, Kenneth L. (1998) *Numerical Methods in Economics,* MIT Press.

Keane, Michael P. and Robert M. Sauer (2009) "Classification Error in Dynamic Discrete Choice Models: Implications for Female Labor Supply Behavior," *Econometrica,* 77 (3), 975–991.

Keane, Michael P. and Kenneth I. Wolpin (2010) "The Role of Labor and Marriage Markets, Preference Heterogeneity and the Welfare System in the Life Cycle Decisions of Black, Hispanic and White Women," *International Economic Review,* 51 (3), 851–892.

Kasahara, Hiroyuki and Katsumi Shimotsu (2009) "Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices," *Econometrica,* 77 (1), 135–175.

Kasahara, Hiroyuki and Katsumi Shimotsu (2012a) "Sequential Estimation of Structural Models With a Fixed Point Constraint," *Econometrica,* 80 (5), 2303-2319.

Kasahara, Hiroyuki and Katsumi Shimotsu (2012b) "Supplement to 'Sequential Estimation of Structural Models With a Fixed Point Constraint'," *Econometrica Supplemental Material,* 80 (5), https://www.econometricsociety.org/sites/default/files/8291_proofs_0.pdf

Kleibergen, Frank (2005) "Testing Parameters in GMM Without Assuming that They Are Identified," *Econometrica,* 73 (4), 1103–1123.

Kleibergen, Frank and Richard Paap (2006) "Generalized Reduced Rank Tests Using the Singular Value Decomposition," *Journal of Econometrics,* 133 (1), 97–126.

Ljungqvist, Lars and Thomas J. Sargent (2004) *Recursive Macroeconomic Theory,* 2nd Edition, MIT Press.

Moffitt, Robert (1984) “Profiles of Fertility, Labour Supply and Wages of Married Women: A Complete Life-Cycle Model ,” *Review of Economic Studies,* 51 (2), 263–278.

Mroz, Thomas A. (1987) “The Sensitivity of an Empirical Model of Married Women’s Hours of Work to Economic and Statistical Assumptions,” *Econometrica,* 55 (4), 765–799.

Norets, Andriy and Xun Tang (2014) “Semiparametric Inference in Dynamic Binary Choice Models,” *Review of Economic Studies,* 81 (3), 2014, 1229–1262.

Romano, Joseph P. and Azeem M. Shaikh (2010) “Inference for the Identified Set in Partially Identified Econometric Models,” *Econometrica,* 78 (1), 169–211.

Rosenzweig, Mark R. and Kenneth I. Wolpin (1980) “Life-Cycle Labor Supply and Fertility: Causal Inferences from Household Models,” *Journal of Political Economy,* 88 (2), 328–348.

Rust, John P., Joseph F. Traub, and Henry Wozniakowski (2002) “Is There a Curse of Dimensionality for Contraction Fixed Points in the Worst Case?” *Econometrica,* 70 (1), 285–329.

Schennach, Susanne (2014) “Entropic Latent Variable Integration via Simulation,” *Econometrica,* 82 (1), 345–385.

Stock, James H. and Jonathan H. Wright (2000) “GMM with Weak Identification,” *Econometrica,* 68 (5), 1055–1096.

Wooldridge, Jeffrey M. (2005) “Simple Solutions to the Initial Conditions Problem in Dynamic, Nonlinear Panel Data Models with Unobserved Heterogeneity,” *Journal of Applied Econometrics,* 20 (1), 39–54.