CCP Estimation of Dynamic Discrete/Continuous Choice Models with Generalized Finite Dependence and Correlated Unobserved Heterogeneity

Wayne-Roy Gayle*

August 7, 2017

Abstract

This paper investigates conditional choice probability estimation of dynamic structural discrete and continuous choice models. I extend the concept of finite dependence in a way that accommodates non-stationary, irreducible transition probabilities. I show that under this new definition of finite dependence, one-period dependence is obtainable in any dynamic structural model with non-degenerate transition functions. This finite dependence property also provides a convenient and computationally cheap representation of the optimality conditions for the continuous choice variables. I allow for discrete-valued unobserved heterogeneity in utilities, transition probabilities, and production functions. The unobserved heterogeneity may be correlated with the observable state variables. I show the estimator is root-n–asymptotically normal. I develop a new and computationally cheap algorithm to compute the estimator, and analyse the finite sample properties of this estimator via Monte Carlo techniques. I apply the proposed method to estimate a model of education and labor supply choices to investigate properties of the distribution of returns to education, using data from the National Longitudinal Survey of Youth 1979.

KEYWORDS: Conditional Choice Probabilities Estimator, Discrete/Continuous Choice, Finite Dependence, Correlated Random Effects.
JEL: C14, C31, C33, C35.

*Department of Resource Economics, University of Massachusetts, 80 Campus Way, Amherst, MA 01003, E-mail: wayneroy@gmail.com. The author is grateful to Victor Aguirregabiria, George-Levi Gayle, Lala Zun Ma, Holger Sieg, and Steven Stern for insightful comments. I also thank the seminar participants at the University of Virginia and the University of North Carolina, Chapel Hill. All errors are my own.

1 Introduction

In this paper, I investigate conditional choice probability (CCP) estimation of dynamic structural discrete/continuous choice models with unobserved individual heterogeneity. I show that an extension to the definition of finite dependence proposed in Altug and Miller [1998] and Arcidiacono and Miller [2011] accommodates general non-stationary and irreducible transition functions, as well as a general form of correlated unobserved heterogeneity in the utility functions, production functions, and the transition functions. I propose a generalized method of moments (GMM) estimator for the structural parameters of the model and derive their asymptotic distributions. I also propose a simple algorithm to implement the estimator. I investigate the finite sample properties of the estimator by way of Monte Carlo analysis and implement this method to estimate a model of education and labor supply choices to investigate the distribution of returns to education, using data from the NLSY79.

Since its introduction by Hotz and Miller [1993], CCP estimation of dynamic structural models has flourished in empirical labor economics and industrial organization, largely because of its potential for an immense reduction in computational costs compared to the more traditional backward recursive- and contraction mapping-based full maximum likelihood estimation pioneered by Rust [1987], referred to as the nested fixed-point algorithm (NFXP). The CCP estimator circumvents having to solve the dynamic programming problem for each trial value of the structural parameters, by making use of a one-to-one mapping between the normalized value functions and the CCPs established in Hotz and Miller [1993]. Therefore, nonparametric estimates of the CCPs can be inverted to obtain estimates of the normalized value functions, which can then be used to estimate the structural parameters.

Empirical application of the early formulation of CCP estimation had important limitations relative to the NFXP method. The emerging literature has focused on separate but related drawbacks. The first is that nonparametric estimation of the CCPs results in less efficient estimates of the structural parameters, as well as relatively poor finite sample performance. The second is the difficulty of accounting for unobserved individual heterogeneity, mainly due to having to estimate the CCPs by nonparametric methods. A limitation of both the CCP and NFXP approaches to estimation is that they are largely restricted to discrete choice, discrete states models.

Aguirregabiria and Mira [2002] proposed a solution to the issue of efficiency and finite sample performance of the CCP estimator relative to the NFXP estimator. They show that for a given value of the preference parameters, the fixed point problem in the value function space can be transformed into a fixed-point problem in the probability space. Aguirregabiria and Mira (2002) propose swapping the nesting of the NFXP, and show the resulting estimator is asymptotically equivalent to the NFXP estimator. Furthermore, Aguirregabiria and Mira [2002] show in simulation studies that their method produces estimates 5 to 15 times faster than NFXP. The method that Aguirregabiria and Mira [2002] propose is restricted to discrete choice models in stationary environments, and is not designed to account for unobserved individual heterogeneity.

Recent developments in accounting for unobserved heterogeneity in CCP estimators include Aguirregabiria and Mira [2007], and Arcidiacono and Miller [2011]. Aguirregabiria and Mira (2007) allow for permanent unobserved heterogeneity in stationary, dynamic discrete games. Their method requires multiple inversion of potentially large dimensional matrices. Arcidiacono and Miller [2011] propose a more general method for incorporating time-specific or time-invariant unobserved heterogeneity into CCP estimators. Their method modifies the expectations-maximization algorithm in a way similar to Acidiacono (2002). However, Arcidiacono and Miller's method is only applicable to discrete dynamic models.

Altug and Miller [1998] proposed an approach that allows for continuous choices in the CCP framework. By assuming complete markets, estimates of individual effects and aggregate shocks are obtained, which are then used in the second stage to form (now) observationally equivalent individuals. These observationally equivalent individuals are used to compute counterfactual continuous choices. Bajari et al. [2007] modify the methods of Hotz and Miller [1993] and Hotz et al. [1994], to estimate dynamic games. Their method of modeling unobserved heterogeneity in continuous choices is inconsistent with the dynamic selection.

The finite dependence property – when two different policies associated with different initial choices lead to the same distribution of states after a few periods – is critical for the computational feasibility and finite sample performance of CCP estimators. Finite dependence combined with the invertibility result of Hotz and Miller [1993] results in a significant reduction in the computational cost of estimating dynamic structural models. Essentially, the smaller the order of dependence, the faster and more precise the estimator, because fewer

future choice probabilities have to be estimated or updated, depending of the method of estimation. The concept of finite dependence was first introduced by Hotz and Miller [1993], extended by Altug and Miller [1998], and further by Arcidiacono and Miller [2011]. Despite these generalizations, the concept of finite dependence is largely restricted to discrete choice models with either stationary transitions or the renewal property.

This paper makes three separate, but closely related contributions to the literature on CCP estimation of dynamic structural models. I extend the concept of finite dependence to allow for general non-stationary and irreducible transition probabilities. Although its definition is precise and well understood, the strategy to construct finite dependence in dynamic structural models have been largely ad hoc, and often achieved by relying on assumptions that are either theoretically unjustified or by significantly restricting the data. Altug and Miller [1998], Gayle and Miller [2003], and Gayle [2006] rely on complete markets and degenerate transition probability assumptions to form counterfactual strategies that obtain finite dependence. A key insight of Arcidiacono and Miller [2011] is that: "the expected value of future utilities from optimal decision making can always be expressed as functions of the flow payoffs and conditional choice probabilities for *any* sequence of future choices, optimal or not." This insight is the basis of our extension of the finite dependence property. I show the expected value of future utilities from optimal decision making can be expressed as *any linear combination* of flow payoffs and conditional CCPs, as long as the weights sum to one. This insight converts the difficult problem of finding one pair of sequences of choices that obtains finite dependence to a potential continuum of finite dependencies from which to choose.

Given I am now able to choose from a class of finite dependence representations, the question becomes whether a choice of weights exists that obtains one-period finite dependence. Indeed, one-period finite dependence is achievable regardless of the form of the transition functions, so long as they are non-degenerate. The resulting form of the conditional value function provides a simple method to accommodate continuous choices.

The approach taken to model continuous choices may be considered a dynamic version of the Roy [1951] model and parallels the method for estimating discrete/continuous static structural models of Dubin and McFadden [1984], and Hanemann [1984]. Particularly, in each period and for each discrete alternative, the agent observes the period-specific shocks and solves for the associated conditional continuous choices (henceforth CCCs) that maximizes the corresponding alternative-specific value of the discrete choice. The agent then chooses

the alternative with the highest (maximized) value. This dynamic selection on unobservables implies the distribution of observed CCCs differ from the distribution of optimal CCCs, rendering first-stage estimation of optimal CCCs biased without additional restrictions, such as the Pareto Optimality condition imposed by Altug and Miller [1998] and subsequent authors.

Along with individual-time–specific shocks, the model developed in this paper allows for discrete-valued, permanent unobserved heterogeneity in the utility functions, production functions, and the transition probabilities. The distribution of these unobserved random variables may be correlated with observable covariates of the model. I provide sufficient conditions for identification of all the parameters of the model: those governing preferences, transitions, production functions, and the distribution of the unobservable heterogeneity.

I propose a GMM estimator for these parameters and an iterative algorithm to compute them. Relative to maximum likelihood-based estimators, the GMM estimator has the advantage that it does not require specification of the distribution of measurement errors, which is of particular concern in this framework because continuous outcome variables are often measured with errors. Also, the GMM estimator is robust to the initial conditions problem in that consistent estimation of the parameters does not require observing the initial state variables nor estimating the initial conditions. A consequence of opting for the GMM approach to estimation is the method developed in Arcidiacono and Miller [2011] to account for unobserved heterogeneity is no longer available. I address this deficiency by developing an iterative Bayes method which uses information from the CCPs alone.

I investigate the finite sample properties of the proposed estimator by way of Monte Carlo methods. Three environments are considered. The first two are models for which the weights that achieve one-period finite dependence are closed-form while the third requires computation of these weights. The results show the proposed estimator performs well in all three environments.

I apply the proposed model and estimator to estimate a model of educational attainment and labor supply to investigate properties of the distribution of the returns to education, using data from the National Longitudinal Survey of Youth 1979 (NLSY79). Key features of the model are that: (1) I allow for individuals to choose to simultaneously participate in the labor market and enroll in school, (2) I treat hours worked as a continuous choice variable and allow for it to affect the probability of completing the grade level enrolled in, (3) I allow for

psychic costs of school attendance and labor market activities, and (4) Returns to education is modeled a random coefficient with a finite mixture distribution with type probabilities depending on racial categories.

The rest of the paper proceeds as follows. Section 2 outlines the class of dynamic structural models investigated in this paper and presents the new alternative representation of the value functions that I use to obtain finite dependence. Section 3 then defines generalized finite dependence, shows that one period finite dependence can be obtained in my class of models, defines first-order optimality conditions for optimal choices, and outlines my approach to incorporating correlated unobserved heterogeneity in the model. I provide sufficient conditions for identification of the parameters of the model in section 4. Section 5 proposes a GMM estimator for the parameters. Section 6 outlines the algorithm I propose to compute the estimator, and section 7 presents the asymptotic properties of my estimator. The Monte Carlo analysis of the finite sample properties of the proposed estimator is presented in Section 8 and in section 9, I implement my method to estimate a model of educational attainment and labor supply. Section 10 concludes. The Appendix contains the proofs and the tables reporting the estimation results from my empirical application in Section 9.

2 Model

2.1 General framework

This section outlines the class of dynamic structural discrete/continuous choice models that I consider and corresponding alternative representation. This framework only modifies that of Arcidiacono and Miller [2011] to include the CCCs, and I maintain the notation of notation of Arcidiacono and Miller [2011] where feasible for consistency. However, exposition of the framework is necessary to make clear the approach to obtaining finite dependence.

In each period, t, an individual chooses among J discrete, mutually exclusive, and exhaustive alternatives. Let d_{jt} be 1 if the discrete action $j \in \{1,\cdots,J\}$ is taken in period t, and zero otherwise, and define $d_t = (d_{1t},\cdots,d_{Jt})$. Associated with each discrete alternative, j, the individual chooses L_j continuous alternatives. Let $c_{l_jt} \in \Re_+$, $l_j \in 1,\cdots,L_j$, be the continuous actions associated with alternative j, with $c_{l_jt} > 0$ if $d_{jt} = 1$. Define $c_{jt} = (c_{1t},\cdots,c_{L_jt}) \in \Re_+^{L_j}$,

and $c_t = (c_{1t}, \cdots, c_{Jt}) \in \Re_+^L$, where $L = \sum_{j=1}^J L_j$. Also, let (j, c_{jt}) be the vector of discrete and continuous actions associated with alternative j. The current-period payoff associated with action (j, c_{jt}) depends on the observed state $x_t \in \mathcal{X} \subseteq \Re^{D_x}$, where D_x is the dimension of x_t, the unobserved state $s_t \in \Re^{D_s}$, where D_s is the dimension of s_t, the unidimensional discrete-choice–specific shock $\varepsilon_{jt} \in \Re$, and the L_j-dimensional vector of continuous-choice–specific shocks $r_{jt} = (r_{1t}, \cdots, r_{L_j t}) \in \Re^{L_j}$. Let $z_t = (x_t, s_t) \in \mathcal{Z} \subseteq \Re^{D_x + D_s}$, $e_{jt} = (\varepsilon_{jt}, r_{jt})$, and $e_t = (e_{1t}, \cdots, e_{Jt})$. The probability function of (z_{t+1}, e_{t+1}) given (z_t, e_t) and that action (j, c_{jt}) is taken in period t is denoted by $f_{jt}(z_{t+1}, e_{t+1} | z_t, c_{jt}, e_t)$. The vector of shocks, e_t, are observed to the individual at the beginning of period t and the individual's conditional direct current-period payoff from choosing alternative (j, c_{jt}) in period t is denoted by $u_{jt}(z_t, c_{jt}, r_{jt}) + \varepsilon_{jt}$.

Define $y_{jt} = (d_{jt}, c_{jt})$. The individual chooses the vector $y_t = (y_{1t}, \cdots, y_{Jt})$ to sequentially maximize the expected discounted sum of payoffs:

$$E\left\{\sum_{t=1}^{T}\sum_{j=1}^{J}\beta^{t-1} d_{jt}[u_{jt}(z_t, c_{jt}, r_{jt}) + \varepsilon_{jt}]\right\}, \tag{2.1}$$

where $\beta \in (0,1)$ is the discount factor. In each period, t, the expectation is taken with respect to the joint distribution of $z_{t+1}, \cdots, z_T$ and $e_{t+1}, \cdots, e_T$. Let the policy rule at period t be given by $y_t^0 = \{(d_{jt}^0(z_t, e_t), c_{jt}^0(z_t, e_t)), j = 1, \cdots, J\}$. Let the ex-ante value function in period t, $V_t(z_t, r_t)$, be the discounted sum of expected future payoffs, prior to observing ε_t, given the policy rule:

$$V_t(z_t, r_t) = E\left\{\sum_{\tau=t}^{T}\sum_{j=1}^{J}\beta^{\tau-t} d_{j\tau}^0(z_\tau, e_\tau)[u_{j\tau}(z_\tau, c_{j\tau}^0(z_\tau, e_\tau), r_{j\tau}) + \varepsilon_{j\tau}]\right\}.$$

As is standard in dicrete/continuous models, the additive separability of the utility function implies the CCCs are functions of their associated shocks and not of ε_t. Assume $f_{jt}(z_{t+1}, e_{t+1} | z_t, c_{jt}, e_t) = f_{jt}(z_{t+1} | z_t, c_{jt}, r_{jt}) g_r(r_{t+1}) g_\varepsilon(\varepsilon_{t+1})$, where g_r is the density function of r_t and g_ε is the density function of ε. The expected value function in period $t+1$, given z_t, r_t, the discrete choice, j, and CCCs is

$$\bar{V}_{jt+1}(z_t, c_{jt}, r_{jt}) = \beta \int V_{t+1}(z_{t+1}, r_{t+1}) f_{jt}(z_{t+1} | z_t, c_{jt}, r_{jt}) g_r(r_{t+1}) dr_{t+1} dz_{t+1}.$$

The ex-ante value function can then written recursively:

$$\begin{aligned} V_t(z_t, r_t) &= E\left\{ \sum_{j=1}^{J} d_{jt}^0(z_t, e_t) \left[u_{jt}(z_t, c_{jt}^0(z_t, r_{jt}), r_{jt}) + \varepsilon_{jt} + \beta \bar{V}_{t+1,j}(z_t, r_{jt}) \right] \right\} \\ &= \int \sum_{j=1}^{J} d_{jt}^0(z_t, e_t) \left[u_{jt}(z_t, c_{jt}^0(z_t, r_{jt}), r_{jt}) + \varepsilon_{jt} + \beta \bar{V}_{t+1,j}(z_t, r_{jt}) \right] g_\varepsilon(\varepsilon_t) d\varepsilon_t, \\ &= \int \sum_{j=1}^{J} d_{jt}^0(z_t, e_t) \left[v_{jt}(z_t, c_{jt}^0(z_t, r_{jt}), r_{jt}) + \varepsilon_{jt} \right] g_\varepsilon(\varepsilon_t) d\varepsilon_t \end{aligned}$$

where

$$v_{jt}(z_t, c_{jt}, r_{jt}) = u_{jt}(z_t, c_{jt}, r_{jt}) + \bar{V}_{jt+1}(z_t, c_{jt}, r_{jt}), \tag{2.2}$$

is the choice-specific conditional value function excluding ε_{jt} and

$$\bar{V}_{jt+1}(z_t, r_{jt}) = \bar{V}_{jt+1}(z_t, c_{jt}^0(z_t, r_{jt}), r_{jt}).$$

The optimal CCCs given the discrete alternative j being chosen in period t, satisfy

$$\frac{\partial}{\partial c_{l_j t}} v_{jt}(z_t, c_{jt}^0(z_t, r_{jt}), r_{jt}) = 0, \tag{2.3}$$

for $l_j = 1, \cdots, L_j$. Given the optimal CCCs, $c_t^0(z_t, r_t) = (c_{kt}^0(z_t, r_{kt}), k = 1, \cdots, J)$, the individual's discrete choice of alternative j is optimal if

$$d_{jt}^0(z_t, c_t^0(z_t, r_t), e_t) = \begin{cases} 1 & \text{if} \quad v_{jt}(z_t, c_{jt}^0(z_t, r_{jt}), r_{jt}) + \varepsilon_{jt} > v_{kt}(z_t, c_{kt}^0(z_t, r_{kt}), r_{tk}) + \varepsilon_{kt} \quad \forall k \neq j \\ 0 & \text{otherwise.} \end{cases} \tag{2.4}$$

Finally, the optimal unconditional continuous choice, $c_{jt}(z_t, r_{jt})$, is given by

$$c_{jt}^*(z_t, e_{jt}) = d_{jt}^0(z_t, c_t^0(z_t, r_t), e_t) c_{jt}^0(z_t, r_{jt}). \tag{2.5}$$

2.2 Alternative representation

The probability of choosing alternative j at time t, conditional on z_t, r_t, and the vector of choice-specific optimal conditional continuous choices, $c_t^0(z_t, r_t)$, is given by

$$p_{jt}^0(z_t, r_t) = E[d_{jt}^0(z_t, c_t^0(z_t, r_t), e_t)|z_t, r_t] = \int d_{jt}^0(z_t, c_t^0(z_t, r_t), r_t, \varepsilon_t) g_\varepsilon(\varepsilon_t) d\varepsilon_t, \tag{2.6}$$

so that, for all (z_t, r_t), $\sum_{j=1}^J p_{jt}^0(z_t, r_t) = 1$, and $p_{jt}^0(z_t, r_t) > 0$ for all j. Let $p_t^0(z_t, r_t) = (p_{1t}^0(z_t, r_t), \cdots, p_{Jt}^0(z_t, r_t))'$ be the vector of CCPs. Lemma 1 of Arcidiacono and Miller [2011] shows a function $\psi : [0,1]^J \mapsto \Re$ exists such that, for $k = 1, \cdots, J$,

$$\psi_k(p_t^0(z_t, r_t)) \equiv V_t(z_t, r_t) - v_{tk}(z_t, c_{kt}^0(z_t, r_{kt}), r_{kt}). \tag{2.7}$$

Equation (2.7) is simply equation (3.5) of Arcidiacono and Miller [2011], modified so the choice probabilities and value functions are also conditional on the i.i.d. shocks associated with the CCCs.

The key insight is: if (2.7) holds for $k = 1, \cdots, J$, then for any J-dimensional vector of real numbers $a = (a_1, \cdots, a_J)$ with $\sum_{k=1}^J a_k = 1$,

$$V_t(z_t, r_t) = \sum_{k=1}^J a_k [v_{kt}(z_t, c_{kt}^0(z_t, r_{kt}), r_{kt}) + \psi_k(p_t^0(z_t, r_t))]. \tag{2.8}$$

Let $a_{jt+1} = (a_{1jt+1}, \cdots, a_{Jjt+1})$, possibly depending on $(z_1 \cdots, z_{t+1})$, be the weights associated with the initial discrete choice, j, in period t. Substituting equation (2.8) into equation (2.2) gives:

$$\begin{aligned} v_{jt}(z_t, c_{jt}, r_{jt}) &= u_{jt}(z_t, c_{jt}, r_{jt}) \\ &+ \beta \sum_{k=1}^J \int [v_{k,t+1}(z_{t+1}, c_{kt+1}^0(z_{t+1}, r_{kt+1}), r_{kt+1}) \\ &+ \psi_k(p_{t+1}^0(z_{t+1}, r_{t+1}))] a_{kjt+1} g_r(r_{t+1}) dr_{t+1} f_{jt}(z_{t+1}|z_t, c_{jt}, r_{jt}) dz_{t+1}, \end{aligned} \tag{2.9}$$

Equation (2.9) shows the value function conditional on (z_t, r_t) can be written as the flow payoff of the choice plus any weighted sum of a function of the one-period-ahead CCPs plus the one-period-ahead conditional value functions, where the weights sum to 1. This extension

of the results of Arcidiacono and Miller [2011] provides a powerful tool for obtaining finite dependence in any models that can be formulated as the one developed in the previous section.

Clarifying example

To clarify the alternative representation, I provide a "stripped down" example of the model formation for which $J = 2$. In this example, I assume the individual-time–specific discrete-choice shock, ε_{ijt}, is distributed i.i.d., type 1 extreme value. I also suppress the dependency of the optimal conditional continuous choices, $c_t^0(z_t, r_t)$ on (z_t, r_t) and abstract away from the shocks associated to the conditional continuous choices. The choice-specific conditional value function in equation (2.2) becomes

$$v_{jt}(z_t, c_{jt}) = u_{jt}(z_t, c_{jt}) + \bar{V}_{t+1,j}(z_t, c_{jt}), \tag{2.10}$$

where

$$\bar{V}_{jt+1}(z_t, c_{jt}) = \beta \int \ln \sum_{k=1}^{2} e^{v_{kt+1}(z_{t+1}, c_{kt+1}^0)} f_{jt}(z_{t+1}|z_t, c_{jt}) dz_{t+1} + \beta\gamma,$$

and γ is the Euler constant. Equation (2.2) becomes

$$\begin{aligned} v_{jt}(z_t, c_{jt}) &= u_{jt}(z_t, c_{jt}) \\ &+ \beta \int \ln \sum_{k=1}^{2} e^{v_{kt+1}(z_{t+1}, c_{kt+1}^0)} f_{jt}(z_{t+1}|z_t, c_{jt}) dz_{t+1} + \beta\gamma. \end{aligned} \tag{2.11}$$

Also, the period $t+1$ conditional choice probability of alternative $j = 1, 2$ is given by

$$p_{jt+1}^0(z_{t+1}, c_{t+1}^0) = \frac{e^{v_{jt+1}(z_{t+1}, c_{jt+1}^0)}}{\sum_{k=1}^{2} e^{v_{kt+1}(z_{t+1}, c_{kt+1}^0)}}. \tag{2.12}$$

From equation (2.12), the following equality holds for $j = 1, 2$:

$$\ln \sum_{k=1}^{2} e^{v_{kt+1}(z_{t+1}, c_{kt+1}^0)} = v_{jt+1}(z_{t+1}, c_{jt+1}^0) - \ln p_{jt+1}^0(z_{t+1}, c_{t+1}^0). \tag{2.13}$$

Notice equation (2.13) is simply equation (2.7) under the assumptions of this example and evaluated at period $t+1$. Also, note the LHS of equation (2.13) is a term inside the integral

on the RHS of equation (2.11). For alternative $j = 1,2$, let a_{kjt+1} be weights associated with alternative j in period t and alternative k in period $t+1$, with $a_{1jt+1} + a_{2jt+1} = 1, j = 1,2$. Then from equation (2.12),

$$\begin{aligned}
&\ln \sum_{k=1}^{2} e^{v_{kt+1}(z_{t+1}, c^0_{kt+1})} \\
&= \sum_{k=1}^{2} a_{kjt+1}[v_{kt+1}(z_{t+1}, c^0_{kt+1}) - \ln p^0_{kt+1}(z_{t+1}, c^0_{t+1})]. \qquad (2.14)
\end{aligned}$$

Substituting equation (2.14) into equation (2.11) obtains

$$\begin{aligned}
&\bar{V}_{t+1,j}(z_t, c_{jt}) \\
&= \beta \int \sum_{k=1}^{2} [v_{kt+1}(z_{t+1}, c^0_{kt+1}) - \ln p^0_{kt+1}(z_{t+1}, c^0_{t+1})] \\
&\times a_{kjt+1} f_{jt}(z_{t+1}|z_t, c_{jt}) dz_{t+1} + \beta\gamma. \qquad (2.15)
\end{aligned}$$

Now, substituting $\bar{V}_{jt+1}$ from equation (2.15) into equation (2.11), obtains

$$\begin{aligned}
&v_{jt}(z_t, c_{jt}) = u_{jt}(z_t, c_{jt}) \\
&+ \beta \int \sum_{k=1}^{2} [v_{kt+1}(z_{t+1}, c^0_{kt+1}) - \ln p^0_{kt+1}(z_{t+1}, c^0_{t+1})] \\
&\times a_{kjt+1} f_{jt}(z_{t+1}|z_t, c_{jt}) dz_{t+1} + \beta\gamma. \qquad (2.16)
\end{aligned}$$

3 Generalized finite dependence

The purpose of this section is to show how the weights, $\{a_{kj\tau}, \tau \geq t+1,\ k, j = 1, \cdots, J\}$, may be used to obtain finite dependence. I begin by showing this result holds for the clarifying example.

Clarifying example contd.

Evaluating equation (2.16) at period $t+1$, and substituting into equation (2.10) obtains

$$\begin{aligned}
&v_{jt}(z_t, c_{jt}) = u_{jt}(z_t, c_{jt}) \\
&+\beta \int \sum_{k=1}^{2} [u_{kt+1}(z_{t+1}, c^0_{kt+1}) - \ln p^0_{kt+1}(z_{t+1}, c^0_{kt+1})] a_{kjt+1} f_{jt}(z_{t+1}|z_t, c_{jt}) dz_{t+1} \\
&+\beta^2 \int V_{t+2}(z_{t+2}) \left[\int \sum_{k=1}^{2} f_{kt+1}(z_{t+2}|z_{t+1}, c^0_{kt+1}) a_{kjt+1} f_{jt}(z_{t+1}|z_t, c_{jt}) dz_{t+1} \right] dz_{t+2} \\
&+\beta\gamma. \qquad (3.1)
\end{aligned}$$

Equation (3.1) can be used to write the difference in the choice-specific conditional value function as follows:

$$\begin{aligned}
&v_{2t}(z_t, c_{2t}) - v_{1t}(z_t, c_{1t}) = u_{2t}(z_t, c_{2t}) - u_{1t}(z_t, c_{1t}) \\
&+\beta \int \sum_{k=1}^{2} [u_{kt+1}(z_{t+1}, c^0_{kt+1}) - \ln p^0_{kt+1}(z_{t+1}, c^0_{t+1})] \\
&\times [a_{k2t+1} f_{2t}(z_{t+1}|z_t, c_{2t}) - a_{k1t+1} f_{1t}(z_{t+1}|z_t, c_{1t})] dz_{t+1} \\
&+\beta^2 \int V_{t+2}(z_{t+2}) \\
&\times \left[\int \sum_{k=1}^{2} f_{kt+1}(z_{t+2}|z_{t+1}, c^0_{kt+1}) (a_{k2t+1} f_{2t}(z_{t+1}|z_t, c_{2t}) - a_{k1t+1} f_{1t}(z_{t+1}|z_t, c_{1t})) dz_{t+1} \right] dz_{t+2}. \qquad (3.2)
\end{aligned}$$

Finite dependence is obtained if $\{a_{kjt+1},\ \ k, j = 1, 2\}$ satisfies

$$\begin{aligned}
&\int \sum_{k=1}^{2} f_{kt+1}(z_{t+2}|z_{t+1}, c^0_{kt+1}) a_{k2t+1} f_{2t}(z_{t+1}|z_t, c_{2t}) dz_{t+1} \\
&- \int \sum_{k=1}^{2} f_{kt+1}(z_{t+2}|z_{t+1}, c^0_{kt+1}) a_{k1t+1} f_{1t}(z_{t+1}|z_t, c_{1t}) dz_{t+1} = 0 \quad \text{a.e. } z_{t+2}, \qquad (3.3) \\
&\sum_{k=1}^{2} a_{kjt+1} = 1, \text{ and}, \qquad (3.4) \\
&a_{k^*2t+1} f_{2t}(z_{t+1}|z_t, c_{2t}) \neq a_{k^*1t+1} f_{1t}(z_{t+1}|z_t, c_{1t}) \quad \text{for at least one } k^* \in \{1, 2\}. \qquad (3.5)
\end{aligned}$$

Note the last expression in bracket of equation (3.2) is the equivalent to the expression on the

RHS of equation (3.3). Equation (3.3) is written to make clear that $(a_{k2t+1} f_{2t}(z_{t+1}|z_t, c_{2t}) - a_{k1t+1} f_{1t}(z_{t+1}|z_t, c_{1t})) = 0$ is not necessary for equation (3.3) to hold and therefore weights typically exist for which both equations (3.3) and (3.5) hold.

In general, one would have to solve equation (3.10) to obtain weights that achieve finite dependence. However, special cases exist in the literature for which the weights that solve this system of equations are closed-form. Define

$$f_{kj}(z_{t+2}, z_{t+1}) = f_{kt+1}(z_{t+2}|z_{t+1}, c^0_{kt+1}) f_{jt}(z_{t+1}|z_t, c_{jt}),$$

where we suppress dependence of (z_t, c_{jt}), and let $z_t = (z_{1t}, z_{2t})$, where z_{1t} is a vector of strictly exogenous variables. I discuss three such cases in the following.

Simple Transition A simple transition function is defined by the restriction that the period $t+1$ conditional distribution of the endogenous state variables is independent of the period t endogenous state variables given the joint distribution of the period t and $t+1$ strictly exogenous state variables. A necessary condition for this property to hold is the optimal CCCs (if they exist) are not functions of the endogenous state variables, given the exogenous state variables, which is satisfied if the derivative of the period-specific utility functions with respect to the CCCs are not functions of the endogenous state variables. Particularly, the transition function takes the form

$$f_{jt}(z_{t+1}|z_t, c^0_{jt}(z_t)) = f_{jt}(z_{2t+1}|z_{1t+1}, z_{1t}, c^0_{jt}(z_{1t})) f_t(z_{1t+1}|z_{1t}), \tag{3.6}$$

in which case for $k = 1, 2$,

$$\int f_{kj}(z_{t+2}, z_{t+1}) dz_{2t+1} = f_{kt+1}(z_{2t+2}|z_{1t+2}, z_{1t+1}, c^0_{kt+1}(z_{1t+1})) f_{t+1}(z_{1t+2}, z_{1t+1}|z_{1t}),$$

resulting in the two period ahead distribution of the state variables being independent of the action taken in the current period. Therefore, setting $a_{11t+1} = a_{12t+1} = \gamma$, for any $\gamma \in \Re$ satisfies equation (3.3).

Renewal. A model with the renewal property is one for which an action, say alternative one, can be taken in period $t+1$ so the conditional distribution of the period $t+2$ endogenous state variables does not depend on the action taken in period t, given the joint distribution of the periods $(t, \cdots, t+2)$ exogenous state variables. In other words, equation (3.6) holds for

only $j = 1$. Then one-period finite dependence is obtained by setting $a_{11t+1} = a_{12t+1} = 1$ in equation (3.2). The bus engine replacement model of Rust [1987] is the central example of a model with the renewal property, where the state variable of interest is mileage of the bus and the renewal action of replacing the bus engine (alternative 1 in our example) in period $t+1$ resets mileage to zero, thus making the distribution of mileage in period $t+2$ independent of the decision of whether to replace the bus engine in period t.

Exchangeability. A model with the exchangeability property is one for which taking "opposing" discrete actions in periods t and $t+1$ result in the same distribution of the two-period ahead state variables. Consider the following restriction

$$f_{kj}(z_{t+2}, z_{t+1}) = f_{kt+1}(z_{t+2}|z_{t+1})\, f_{jt}(z_{t+1}|z_t).$$

Then

$$\int f_{kj}(z_{t+2}, z_{t+1}) dz_{2t+1} = f_{t+1}(z_{t+2}|d_{kt+1}, d_{jt}, z_{1t+1}, z_t).$$

The exchangeability property is satisfied if

$$f_{t+1}(z_{t+2}|d_{kt+1} = 0, d_{jt} = 1, z_{1t+1}, z_t) = f_{t+1}(z_{t+2}|d_{kt+1} = 1, d_{jt} = 0, z_{1t+1}, z_t), \quad j \neq k,$$

in which case, setting $a_{kjt+1} = a_{jkt+1} = 1$ for $j \neq k$ obtains one-period finite dependence. The exchangeability restriction holds in the typical labor supply model where, say alternative 2 is the decision to work and the endogenous state variable is years of experience, $\sum_{\tau=1}^{t-1} d_{2\tau}$, which enters the classical Mincerian wage offer function (with formal education completed). This exchangeability condition is typically violated when the transition function depends on CCCs, such as hours worked. Exchangeability is, however, maintained CCCs enter only the period-specific utility functions, which is the case in the typical labor supply model where hours worked does not enter the wage offer equation.

It is important to note that none of these examples require

$$a_{k2t+1} f_{2t}(z_{t+1}|z_t, c_{2t}) = a_{k1t+1} f_{1t}(z_{t+1}|z_t, c_{1t})$$

for $k = 1, 2$ so that equation (3.5) is generically satisfied. There may be additional model

frameworks for which there are closed-form weights that satisfy the system of equations (3.3) - (3.5). However, this system of equations has to be solved numerically to obtain the weights that obtain finite dependence in more general frameworks. To further illuminate how this system may be solved, suppose z_t is discrete with cardinality $|z|$ and define

$$\begin{aligned} f_{kj}(z) &= (f_{kj}(z,z_1),\cdots,f_{kj}(z,z_{|z|})), \\ f_{kj} &= (f_{kj}(z_1)',\cdots,f_{kj}(z_{|z|})')', \\ a_{kj} &= (a_{kjt+1}(z_1),\cdots,a_{kjt+1}(z_{|z|}))'. \end{aligned} \tag{3.7}$$

Then, imposing condition (3.4), equation (3.3) may be written as follows.

$$f_{12}a_{12} + f_{22}(\iota - a_{12}) - f_{11}a_{11} - f_{21}(\iota - a_{11}) = 0, \tag{3.8}$$

$$\Leftrightarrow (f_{12} - f_{22})a_{12} - (f_{11} - f_{21})a_{11} + (f_{22} - f_{21})\iota_{|z|} = 0, \tag{3.9}$$

where $\iota_{|z|}$ is the $|z|$-dimensional vector of ones. The three examples given above can be seen clearly by observing their implications for equation (3.8). The simple transition restriction implies that $f_{k2} = f_{k1}, k = 1,2$, so that equation (3.8) is satisfied by setting $a_{12} = a_{11} = \gamma\iota_{|z|}$. The renewal property implies $f_{k2} = f_{k1}$ for only $k = 1$, so that equation (3.8) is satisfied by setting $a_{12} = a_{11} = \iota_{|z|}$. The exchangeability restriction implies $f_{21} = f_{12}$ so that equation (3.8) is satisfied by setting $a_{21} = a_{12} = \iota_{|z|}$. In general, equation (3.9) can be solved by fixing a_{11} and solving for a_{12} as follows.

$$a_{12} = [(f_{12} - f_{22})'(f_{12} - f_{22})]^{+}(f_{12} - f_{22})'[(f_{11} - f_{21})a_{11} + (f_{21} - f_{22})\iota_{|z|}] \tag{3.10}$$

where $^{+}$ denote a generalized inverse.

I now present the conditions for the generalized $\rho \geq 1$ dependence in the model outlined in Section 2.2. For any initial choice (j, c_{jt}), for periods $\tau = \{t+1,\cdots,t+\rho\}$, and any corresponding sequence $a_\tau = \{a_{kj\tau}, k, j = 1,\cdots,J\}$ with $\sum_{k=1}^{J} a_{kj\tau} = 1$, define

$$\kappa_{j\tau}(z_{\tau+1},|z_t,c_{jt},r_{jt}) = \begin{cases} f_{jt}(z_{t+1}|z_t,c_{jt},r_{jt}) \quad \text{for } \tau = t \\ \int \sum_{k=1}^{J} f_{k\tau}(z_{\tau+1}|z_\tau,c^0_{k\tau},r_{k\tau})a_{kj\tau}g_r(r_{k\tau})\kappa_{\tau-1,j}(z_\tau|z_t,c_{jt},r_{jt})dr_{k\tau}dz_\tau \\ \quad \text{for } \tau = t+1,\cdots,t+\rho, \end{cases} \tag{3.11}$$

where $\int \kappa_{j\tau}(z_{\tau+1},|z_t,c^0_{jt},r_{jt})dz_{\tau+1}=1$, because $\sum_{k=1}^{J} a_{kj\tau+1}=1$. This restriction does not require $a_j \geq 0$. By forward substitution, equations (2.9) and (3.11) obtain

$$\begin{aligned}
&v_{jt}(z_t,c_{jt},r_{jt}) = u_{jt}(z_t,c_{jt},r_{jt}) \\
&+\sum_{\tau=t+1}^{t+\rho}\sum_{k=1}^{J}\int \beta^{\tau-t}[u_{\tau k}(z_\tau,c^0_{k\tau},r_{k\tau})+\psi_k[p^0_\tau(z_\tau,c^0_\tau,r_\tau)]] \\
&\times a_{kj\tau}g_r(r_\tau)\kappa_{\tau-1,j}(z_\tau|z_t,c_{jt},r_{jt})dr_\tau dz_\tau \\
&+\beta^{t+\rho+1}\int V_{t+\rho+1}(z_{t+\rho+1},r_{t+\rho+1}) \\
&\times g_r(r_{t+\rho+1})\kappa_{jt+\rho+1}(z_{t+\rho+1}|z_t,c_{jt},r_{jt})dr_{t+\rho+1}dz_{t+\rho+1} \\
&+\gamma\sum_{\tau=1}^{\rho}\beta^\tau. \qquad (3.12)
\end{aligned}$$

Using equation (3.12), the difference in the conditional value functions associated with two alternative initial choices, j and j' becomes

$$\begin{aligned}
&v_{jt}(z_t,c_{jt},r_{jt})-v_{j't}(z_t,c_{j't},r_{j't}) = u_{jt}(z_t,c_{jt},r_{jt})-u_{j't}(z_t,c_{j't},r_{j't}) \\
&+\sum_{\tau=t+1}^{t+\rho}\sum_{k=1}^{J}\int \beta^{\tau-t}[u_{\tau k}(z_\tau,c^0_{k\tau},r_{k\tau})+\psi_k[p^0_\tau(z_\tau,r_\tau)]] \\
&\times[a_{kj\tau}\kappa_{j\tau-1}(z_\tau|z_t,c_{jt},r_{jt})-a_{\tau kj'}\kappa_{j'\tau-1}(z_\tau|z_t,c_{j't},r_{j't})]g_r(r_\tau)dr_\tau dz_\tau \\
&+\beta^{t+\rho+1}\int V_{t+\rho+1}(z_{t+\rho+1},r_{t+\rho+1})g_r(r_{t+\rho+1})dr_{t+\rho+1} \\
&\times[\kappa_{jt+\rho}(z_{t+\rho+1}|z_t,c_{jt},r_{jt})-\kappa_{j't+\rho}(z_{t+\rho+1}|z_t,c_{j't},r_{j't})]dz_{t+\rho+1}. \qquad (3.13)
\end{aligned}$$

Therefore, a pair of initial choices, (j,c_{jt}) and $(j',c_{j't})$, exhibits **generalized ρ-period dependence** if corresponding sequences, $(a_{jt+1},\cdots,a_{jt+\rho})$ and $(a_{j't+1},\cdots,a_{j't+\rho})$, (recall $a_{jt}:=(a_{1jt},\cdots,a_{Jjt})$) exist such that

$$\kappa_{jt+\rho}(z_{t+\rho+1}|z_t,c_{jt},r_{jt})=\kappa_{j't+\rho}(z_{t+\rho+1}|z_t,c_{j't},r_{j't})$$

almost everywhere,

$$\sum_{k=1}^{J} a_{\tau kj}=1, \quad j=1,\cdots J, \quad \tau=t+1,\cdots,t+\rho,$$

and for at least one $k^* \in \{1,\cdots,J\}$ and $\tau \in \{t+1,\cdots,t+\rho\}$,

$$a_{\tau k^* j}\kappa_{j\tau-1}(z_\tau|z_t,c_{jt},r_{jt}) \neq a_{k^* j'\tau}\kappa_{j'\tau-1}(z_\tau|z_t,c_{j't},r_{j't}).$$

Now, for initial choice (j,c_{jt}) and $\rho = 1$,

$$\kappa_{jt+1}(z_{t+\rho+1}|z_t,c_{jt},r_{jt}) = \int \sum_{k=1}^{J} a_{kjt+1} f_{kt+1}(z_{t+2}|z_{t+1}) f_{jt}(z_{t+1}|z_t,c_{jt},r_{jt})dz_{t+1},$$

so for any pair of initial choices, (j,c_{jt}) and $(j',c_{j't})$,

$$\begin{aligned}
&\kappa_{jt+1}(z_{t+\rho+1}|z_t,c_{jt},r_{jt}) - \kappa_{j't+1}(z_{t+\rho+1}|z_t,c_{j't},r_{j't}) \\
&= \int \sum_{k=1}^{J} f_{kt+1}(z_{t+2}|z_{t+1})[a_{kjt+1} f_{jt}(z_{t+1}|z_t,c_{jt},r_{jt}) - a_{t+1,kj'} f_{j't}(z_{t+1}|z_t,c_{j't},r_{j't})]dz_{t+1}.
\end{aligned} \tag{3.14}$$

Then sufficient condition for one-period dependence in the model set out in Section 2.2 is that $\{(a_{kjt+1}, a_{kj't+1}),\ k = 1,\cdots,J\}$ satisfies

$$\begin{aligned}
&\int \sum_{k=1}^{J} f_{kt+1}(z_{t+2}|z_{t+1}) \\
&\times [a_{kjt+1} f_{jt}(z_{t+1}|z_t,c_{jt},r_{jt}) - a_{kj't+1} f_{j't}(z_{t+1}|z_t,c_{j't},r_{j't})]dz_{t+1} = 0, \quad j,k = 1,2, \\
&\sum_{k=1}^{J} a_{kjt+1} = 1, \quad j = 1\cdots,J, \text{ and,} \\
&a_{k^* jt+1} f_{jt}(z_{t+1}|z_t,c_{jt},r_{jt}) \neq a_{k^* j't+1} f_{j't}(z_{t+1}|z_t,c_{j't},r_{j't}) \quad \text{for at least one } k^* \in \{1,\cdots,J\}.
\end{aligned}$$

The rest of the paper does not rely on the order of depends, so I proceed by assuming $\rho = 1$, in which case, equation (3.12) reduces to

$$\begin{aligned}
&v_{jt}(z_t,c_{jt},r_{jt}) - v_{j't}(z_t,c_{j't},r_{j't}) = u_{jt}(z_t,c_{jt},r_{jt}) - u_{j't}(z_t,c_{j't},r_{j't}) \\
&+ \beta \int \left(\int \sum_{k=1}^{J} [u_{kt+1}(z_{t+1},c^0_{kt+1},r_{kt+1}) + \psi_k[p^0_{t+1}(z_{t+1},r_{t+1})]]g_r(r_{t+1})dr_{t+1} \right) \\
&\times [a_{kjt+1} f_{jt}(z_{t+1}|z_t,c_{jt},r_{jt}) - a_{kj't+1} f_{j't}(z_{t+1}|z_t,c_{j't},r_{j't})]dz_{t+1}.
\end{aligned} \tag{3.15}$$

3.1 Optimal continuous choice

The alternative representation of the difference in conditional value functions provides a simple and convenient representation of the condition for optimal conditional continuous choice, c^0_{jt}, given that alternative j is chosen. The key is to note $\partial v_{j't}(z_t, c^0_{j't}, r_{j't})/\partial c_{l_{jt}} = 0$ for $j' \neq j$ and $l_j = 1, \cdots, L_j$. This equality and equation (3.15) imply $c^0_{jt}(z_t, r_{jt})$ solves

$$
\begin{aligned}
0 = \frac{\partial}{\partial c^0_{jt}} \Big(v_{jt}(z_t, c^0_{jt}, r_{jt}) - v_{j't}(z_t, c^0_{j't}, r_{j't}) \Big) = \frac{\partial}{\partial c_{tl_j}} u_{jt}(z_t, c^0_{jt}, r_{jt}) \\
+ \beta \int \left(\int \sum_{k=1}^{J} \frac{\partial}{\partial c_{l_{jt}}} [u_{kt+1}(z_{t+1}, c^0_{kt+1}, r_{kt+1}) + \psi_k[p^0_{t+1}(z_{t+1}, r_{t+1})]] g_r(r_{t+1}) dr_{t+1} \right) \\
\times [a_{kjt+1} f_{tj}(z_{t+1}|z_t, c^0_{jt}, r_{jt}) - a_{kj't+1} f_{tl}(z_{t+1}|z_t, c^0_{j't}, r_{j't})] dz_{t+1} \\
+ \beta \int \left(\int \sum_{k=1}^{J} [u_{kt+1}(z_{t+1}, c^0_{kt+1}, r_{kt+1}) + \psi_k[p^0_{t+1}(z_{t+1}, r_{t+1})]] g_r(r_{t+1}) dr_{t+1} \right) \\
\times \frac{\partial}{\partial c_{l_{jt}}} [a_{kjt+1} f_{jt}(z_{t+1}|z_t, c^0_{jt}, r_{jt}) - a_{kj't+1} f_{lt}(z_{t+1}|z_t, c^0_{j't}, r_{j't})] dz_{t+1}
\end{aligned} \tag{3.16}
$$

The key to this solution is to note $a_{kj't+1}$ is not a function of $c_{j't}$. Sufficient conditions for uniqueness of the solution to equation (3.16) are presented in Assumption 4.1, particularly 4.1.4.

3.2 Correlated unobserved heterogeneity

Recall that $z_t = (x_t, s_t)$, where x_t is a D_x-dimensional vector of observable state variables and s_t is a D_s-dimensional vector of unobserved state variables. I impose the restriction that the permanent unobserved heterogeneity is independent of x_t, given $w \in \mathcal{W} \subseteq \Re^{D_w}$, a D_w-dimensional time-invariant subset of x_t. While neither of these restrictions are necessary, they improve the exposition of the method by which we account for these unobserved heterogeneity in estimation. Assume that for $j = 1, \cdots, J$,

$$
f_{jt}(z_{t+1}|z_t, c_{jt}, r_{jt}) = f_{jt}(x_{t+1}|x_t, s, c_{jt}) \pi(s|w).
$$

Then equation (3.15) obtains

$$
\begin{aligned}
&v_{jt}(x_t,s,c_{jt},r_{jt}) - v_{jt'}(x_t,s,c_{jt'},r_{jt'}) = u_{jt}(x_t,s,c_{jt},r_{jt}) - u_{jt'}(x_t,s,c_{jt'},r_{jt'}) \\
&+\beta\int\left(\int\sum_{k=1}^{J}[u_{kt+1}(x_{t+1},c^0_{kt+1},s,r_{kt+1}) + \psi_k[p^0_{t+1}(x_{t+1},s,r_{t+1})]]g_r(r_{t+1})dr_{t+1}\right) \\
&\times[a_{kjt+1}f_{jt}(x_{t+1}|x_t,s,c_{jt}) - a_{kj't+1}f_{j't}(x_{t+1}|x_t,s,c_{j't})]dx_{t+1}. \qquad (3.17)
\end{aligned}
$$

The probability of choosing alternative j at time t, conditional on x_t, s, r_t, and the vector of choice-specific optimal conditional continuous choices, $c^0_t = (c^0_{t1},\cdots,c^0_{jt})$, is given by

$$
p^0_{jt}(x_t,s,r_t) = E[d^0_{jt}(z_t,e_t)|x_t,s,c^0_{jt},r_t] = \int d^0_{jt}(z_t,r_t,\varepsilon_t)g_\varepsilon(\varepsilon_t)d\varepsilon_t. \qquad (3.18)
$$

The probability of choosing alternative j at time t and the corresponding optimal continuous choice, conditional on s and x_t, are

$$
p^0_{jt}(x_t,s) = \int p^0_{jt}(x_t,s,r_t)g_r(r_t)dr_t. \qquad (3.19)
$$

Assume that s is discretely distributed with Q support points, $s \in \{s_1,\cdots,s_Q\}$ and denote $\pi(s_q|w) = \pi_q(w)$. The probability of choosing alternative j at time t and the corresponding optimal continuous choice,, conditional on x_t, are

$$
p^0_{jt}(x_t) = \sum_{q=1}^{Q} p^0_{jt}(x_t,s_q)\pi_q(w). \qquad (3.20)
$$

4 Identification

This section presents sufficient conditions for identification of the parameters of the model. We will assume that w is a categorical variable with K mutually exclusive and exhaustive categories. Define $\pi(w) = (\pi_1(w),\cdots,\pi_Q(w))'$, $\pi = \{\pi(w), w \in \mathcal{W}\}$, $\Pi(w) = (\{s_1,\cdots,s_Q\}$, $\pi(w), Q)$, and $\Pi = (\{s_1,\cdots,s_Q\},\pi,Q)$. The following parametric restrictions are imposed: $u_{jt}(z_t,s,c_{jt},r_{jt}) = u_{jt}(z_t,s,c_{jt},r_{jt};B_1,)$ is known up to $B_1 \in \Re^{D_{B_1}}$; $g_r(r_t) = g_r(r_t;B_2)$ is known up to $B_2 \in \Re^{D_{B_2}}$; and $f_{jt}(x_{t+1}|x_t,s,c_{jt}) = f_{jt}(x_{t+1}|x_t,s,c_{jt};B_3)$ is known up to $B_3 \in \Re^{D_{B_3}}$. Define $B = (\beta,B_1,B_2,B_3) \in \mathcal{B} \subseteq \Re^{D_{B_1}+D_{B_2}+D_{B_3}}$.

For each individual unit, the random variables (d_t, c_t, x_t), $t = 1, \cdots, T$ are observable. Hence, in the population, the joint distribution F_{d_t, c_t, x_t} is observed. For $t = 1, \cdots, T$, $j = 1, \cdots, J, k \neq j$, define

$$\begin{aligned} u_{jkt}(x_t, s, c_t, r_t; B_1) &= u_{jt}(x_t, s, c_{jt}, r_{jt}; B_1) - u_{kt}(x_t, s, c_{kt}, r_{kt}; B_1), \quad \text{and} \\ v_{jkt}(x_t, s, c_t, r_t; B) &= v_{jt}(x_t, s, c_{jt}, r_{jt}; B) - v_{kt}(x_t, s, c_{kt}, r_{kt}; B). \end{aligned} \tag{4.1}$$

Assumption 4.1. *1. Rank* $E[x_t' x_t | w] = D_x$, *Rank* $E[w'w] = D_w$, *and the conditional density function of* x_t *given* w, $f_{x_t|w} > 0$.
2. ε_j *and* ε_k *are independent and* g_ε *is twice continuously differentiable and log-concave with support* $\Re$.
3. For each $j \in \{1, \cdots, J\}$, *at least one* $k \in \{1, \cdots, J\} \setminus \{j\}$ *and at least one* $t \in \{2, \cdots, T-1\}$ *exist such that*

$$\begin{aligned} &Rank\left\{ \left(E\left[\frac{\partial}{\partial (B,s)} v_{jkt}(x_t, s, c_t, r_t; B) \middle| w \right] \right) \left(E\left[\frac{\partial}{\partial (B,s)} v_{jkt}(x_t, s, c_t, r_t; B) \middle| w \right] \right)' \right\} \\ &= D_B + D_s. \end{aligned}$$

4. $u_{jt}(x_t, s_q, c_{jt}, r_{jt}; B_1)$ *is strictly increasing, strictly concave, and twice continuously differentiable in* c_{tl_j}, $l_j = 1, \cdots, L_j$, *and* $f_{jt}(x_{t+1} | x_t, s_q, c_{jt}; B_3)$ *is twice continuously differentiable in* $c_{tl_j}, l_j = 1, \cdots, L_j$.
5. For at least one $t \in \{1, \cdots, T\}$, *for some* $j^* \in \{1, \cdots, J\}$, *for any* w, *and for all* $k \in \{1, \cdots, J\} \setminus \{j^*\}$, *a non-empty set of* x_t, $\tilde{\mathcal{X}} \subseteq \mathcal{X}$ *exists for which the following hold:*

i. For any $B \in \mathcal{B}$, *for any* r_t, *and* $s_{q'} \neq s_q$, *either*
$v_{j^*kt}(x_t, s_{q'}, r_t; B) > v_{j^*kt}(x_t, s_q, r_t; B)$ *or* $v_{j^*kt}(x_t, s_{q'}, r_t; B) < v_{j^*kt}(x_t, s_q, r_t; B)$, *for all* $x_t \in \tilde{\mathcal{X}}$

ii. For $\tilde{B} \neq B$, *for any pair* $(\tilde{s}_q, s_q)$ *and any* r_t, $(\tilde{x}_t, x_t) \in \tilde{\mathcal{X}}^2$ *exists for which*
$v_{j^*kt}(\tilde{x}_t, \tilde{s}_q, r_t; \tilde{B}) < v_{j^*kt}(\tilde{x}_t, s_q, r_t; B)$, *and* $v_{j^*kt}(x_t, \tilde{s}_q, r_t; \tilde{B}) > v_{j^*kt}(x_t, s_q, r_t; B)$.

Define $\mathbf{P}(x; B, \Pi) = (p^0_{jt}(x_t; B, \Pi(w)), j = 1, \cdots, J, t = 1, \cdots, T,)$. Let (B_0, Π_0) be the true parameter vector; that is, the probabilities generated from the model at (B_0, Π_0) coincides with the population probabilities: $\mathbf{P}(x; B_0, \Pi_0) = \mathbf{P}^0(x)$.

Theorem 4.2. *Suppose assumption 4.1 holds. Then* (B_0, Π_0) *is identified in the sense that any* $(\tilde{B}, \tilde{\Pi})$ *satisfying* $\mathbf{P}(x; \tilde{B}, \tilde{\Pi}) = \mathbf{P}^0(x)$ *implies* $(\tilde{B}, \tilde{\Pi}) = (B_0, \Pi_0)$.

The proof of Theorem 4.2 is provided in Appendix A.1. It follows the proof of identification in Gayle [2013], which proves identification in a more general framework than the current one.

5 Estimator

In this section, I propose a GMM estimator for the parameters of the model, B and Π. I choose to propose a GMM estimator instead of the ML estimator for two reasons. First, the definition of the GMM estimator does not require specifying the distribution of measurement errors, which is of particular concern in discrete and continuous choice models because observed continuous choice variables are often measured with errors. Second, the GMM estimator is robust to the initial conditions problem: consistent estimation of the parameters does not require observing the initialization of x_t given s or for it to be specified.

The estimator proposed in this section assumes that Q is known by the investigator and w is discrete valued.

Assumption 5.1. *i. The investigator has prior knowledge of number of types, Q, and ii.* w *is discrete valued with K points.*

Suppose n observations of the random vectors $y_i = \{(d_{it}', c_{it}')', t = 1, \cdots, T\}$ and $x_i = \{x_{it}', t = 1, \cdots, T\}$ are accessible to the investigator. Define $\theta_q = (s_q, B)$ and $\theta = (\{s_1, \cdots, s_Q\}, B)$. For each i, and for $t = 1, \cdots, T$, define

$$\begin{aligned}
\tilde{p}_t^0(x_{it}; \theta_q) &= (p_{1t}^0(x_{it}; \theta_q), \cdots, p_{J-1,t}^0(x_{it}; \theta_q))', \\
c_t^0(x_{it}; \theta_q) &= (d_{1it} c_{1t}^0(x_{it}; \theta_q)', \cdots, d_{Jit} c_{Jt}^0(x_{it}; \theta_q)')', \\
h_t(x_{it}; \theta_q) &= (\tilde{p}_t^0(x_{it}; \theta_q)', c_t^0(x_{it}; \theta_q)')', \\
h(x_i; \theta_q) &= (h_1(x_{i1}; \theta_q)', h_T(x_{iT}; \theta_q)')', \\
h(x_i; \theta_q) &= (h(x_i; \theta_1), \cdots, h(x_i; \theta_Q)), \\
\rho(y_i, x_i; \theta, \pi) &= y_i - h(x_i; \theta)\pi(w_i),
\end{aligned} \tag{5.1}$$

where the vector y_i is defined analogously. The vector $\rho(y_i, x_i; \theta, \pi)$ is of $(L+J-1)T$ dimension. Let X_{it} be a vector of instruments with dimension $N_{X_t} \geq J+L-1$, and define $X_i = diag\{X_{it}, t = 1, \cdots, T\}$. Pre-multiplying equation (5.1) by X_i obtains the following $N_X := \sum_{t=1}^{T} N_{X_t}$-dimensional vector.

$$m_i(\theta; \pi) = X_i \rho(y_i, x_i; \theta, \pi). \tag{5.2}$$

Let Ω be a $N_X \times N_X$-dimensional symmetric, positive definite weighting matrix. The population moment condition and objective function are defined as follows.

$$m_0(\theta; \pi) = E[m_i(\theta; \pi)], \tag{5.3}$$

$$S_0(\theta; \pi) = m_0(\theta; \pi)' \Omega m_0(\theta; \pi). \tag{5.4}$$

To define the estimator, let

$$\hat{m}(\theta; \pi) = \frac{1}{n} \sum_{i=1}^{n} m_i(\theta; \pi). \tag{5.5}$$

Then the sample objective function is defined as

$$\hat{S}(\theta; \pi) = \hat{m}(\theta; \pi)' \hat{\Omega} \hat{m}(\theta; \pi), \tag{5.6}$$

where $\hat{\Omega}$ is positive definite and is a consistent estimator for Ω.

6 Computing The Estimator

In this section, I present a method for computing the estimator proposed in the previous section. I describe the algorithm at the $o+1$ iteration with $\left(\pi^{[o]}, \mathbf{c}^{[o]}, \mathbf{p}^{[o]}, \theta^{[o]}\right)$ in hand. In the development of the algorithm, I maintain the assumption of one-period finite dependence, so that period t conditional values functions at iteration $o+1$ depend only on $\left(\pi_{t+1}^{[o]}, c_{t+1}^{[o]}, p_{t+1}^{[o]}, \theta^{[o]}\right)$. Note this assumption is not necessary, but it simplifies the notation. To further conserve on notation, I also suppress the dependence of all functions on the shocks associated with the CCCs and the individual subscripts noting that application of the following algorithm should incorporate these additional dimensions.

6.1 Updating the CCCs

These CCCs, $c_{tj}^{[o+1]}(x_t, \theta_q^{[o]})$, are updated by solving equation (3.16) as follows:

$$\left.\frac{\partial}{\partial c}\right|_{c=c_{tj}^{[o+1]}(x_t,\theta_q^{[o]})} \left(v_{tj}(x_t, c; c_{t'}^{[o]}, p_{t'}^{[o]}, \theta_q^{[o]}) - v_{tj'}(x_t, c; c_{t'}^{[o]}, p_{t'}^{[o]}, \theta_q^{[o]})\right) = 0, \tag{6.1}$$

where $t' = t+1$. Conditions for uniqueness of these solutions are given in Assumption 4.5.

6.2 Updating the CCPs

Let

$$\begin{aligned} &v_t(x_t, c_{tq}^{[o+1]}; c_{t'}^{[o]}, p_{t'}^{[o]}, \theta_q^{[o]}) = \\ &\left(v_{t1}(x_t, c_{t1q}^{[o+1]}; c_{t'}^{[o]}, p_{t'}^{[o]}, \theta_q^{[o]}), \cdots, v_{tJ}(x_t, c_{tJq}^{[o+1]}; c_{t'}^{[o]}, p_{t'}^{[o]}, \theta_q^{[o]})\right)'. \end{aligned}$$

For each $j \in \{1, \cdots, J\}$, Hotz and Miller [1993] show the existence of a function $\Psi_j : \Re^J \mapsto [0,1]$ satisfying

$$p_{tj}^{[o+1]}(x_t, \theta_q^{[o]}) = \Psi_j\left(v_t(x_t, c_{tq}^{[o+1]}; c_{t'}^{[o]}, p_{t'}^{[o]}, \theta_q^{[o]})\right), \tag{6.2}$$

where the functional form of Ψ_j is determined by the distribution of the alternative-specific shocks, ε. For example, if ε is distributed i.i.d. type one extreme value, then

$$p_{tj}^{[o+1]}(x_t, \theta_q^{[o]}) = \frac{e^{v_{tj}(x_t, c_{tjq}^{[o+1]}; c_{t'}^{[o+1]}, p_{t'}^{[o]}, \theta_q^{[o]})}}{\sum_{j'=1}^{J} e^{v_{tj'}(x_t, c_{tj'q}^{[o+1]}; c_{t'}^{[o+1]}, p_{t'}^{[o]}, \theta_q^{[o]})}}.$$

Computing the difference in the continuation values requires integration, which in turn requires values of the (counterfactual) CCCs and CCPs for the relevant states attainable by the individual given his current state. Similar to Arcidiacono and Miller [2011], these values may be computed by interpolation (or cell averaging in the case of discrete states). To this end, $K_\sigma(\cdot)$ with bandwidth σ. Then these counterfactual CCCs and CCPs may be computed

as follows:

$$c_{tj}^{[o+1]}(x_t, \theta_q^{[o]}) = \frac{\sum_{i=1}^{n} c_{tj}^{[o+1]}(x_{it}, \theta_q^{[o]}) K_\sigma(x_t - x_{it})}{\sum_{i=1}^{n} K_\sigma(x_t - x_{it})}, \tag{6.3}$$

$$p_{tj}^{[o+1]}(x_t, \theta_q^{[o]}) = \frac{\sum_{i=1}^{n} p_{tj}^{[o+1]}(x_{it}, \theta_q^{[o]}) K_\sigma(x_t - x_{it})}{\sum_{i=1}^{n} K_\sigma(x_t - x_{it})}. \tag{6.4}$$

In the case where the state vector is discrete valued, the kernel function is taken to be the indicator function $1\{\cdot\}$.

6.3 Updating π

Arcidiacono and Miller [2011] propose and Expectations-Maximization (EM) algorithm to compute estimates of their proposed model. The approach proposed by them provides an internally consistent way to account for unobserved heterogeneity in the form of finite mixtures. The parameters of the model proposed in this paper may also be estimated using their method if one assumes the observed conditional choices are not measured with errors, or if the distributions measurement errors are specified up to finite dimensional parameters. While the EM algorithm approach presented in Arcidiacono and Miller [2011] is not immediately accessible if the observed continuous choices are measured with errors and the investigator is unwilling to specify the distribution of these errors, their intuition of exploiting Bayes' rule to update the type probabilities using the CCPs is still valid. I formalize this process of updating the type probabilities in this section.

In what follows, I suppress the dependence of the type probabilities on w and reintroduce them at the end. The likelihood of d_{it} given (x_{it}, θ) implied by the model is given by

$$f(d_t | x_t, \theta_q) = \left(p_{tj}(x_t, \theta_q)\right)^{d_{tj}}. \tag{6.5}$$

For given type probabilities, π, Bayes's rule implies the following vector of posterior type probabilities

$$\pi_q(d_{it}, x_{it}) = \frac{f(d_{it} | x_{it}, \theta_q)\pi_q}{\sum_{q'=1}^{Q} f(d_{it} | x_{it}, \theta_{q'})\pi_{q'}}, \quad q = 1, \cdots, Q. \tag{6.6}$$

At $\theta = \theta_0$, equation (6.6) obtains

$$\begin{aligned} E[\pi_q(d_{it}, x_{it})|x] &= E\left[\frac{f(d_{it}|x_{it}, \theta_{0q})\pi_q}{\sum_{q'=1}^{Q} f(d_{it}|x_{it}, \theta_{0q'})\pi_{q'}}\middle| x\right] \\ &= \int \frac{f(d_{it}|x, \theta_{0q})\pi_q}{\sum_{q'=1}^{Q} f(d_{it}|x, \theta_{0q'})\pi_{q'}} f_0(d_{it}|x)\mathrm{d}d_{it} \\ &= \pi_q \int \frac{f(d_{it}|x, \theta_{0q})}{\sum_{q'=1}^{Q} f(d_{it}|x, \theta_{0q'})\pi_{q'}} \sum_{q'=1}^{Q} f(d_{it}|x, \theta_{0q'})\pi_{0q'}\mathrm{d}d_{it} \end{aligned} \tag{6.7}$$

for almost every $x \in \mathcal{X}$ where $f_0(d_{it}|x_{it})$ and is the population conditional P.M.F. of d_{it} given x. The last equality is a result of identification of the parameters of the model so that $\sum_{q=1}^{Q} f(d_{it}|x, \theta_{0q})\pi_{0q} = f_0(d_{it}|x_{it})$. Equation (6.7) implies that for each $q \in \{1, \cdots, Q\}$, $E[\pi_q(d_{it}, x_{it})|x] = \pi_q$ for almost every $x \in \mathcal{X}$ if, and only if $\pi_q = \pi_{0q}$. Define

$$\boldsymbol{f}_{it}(\theta, \pi) = diag\left\{\frac{f(d_{it}|x_{it}, \theta_q)}{\sum_{q'=1}^{Q} f(d_{itt}|x_{it}, \theta_{q'})\bar{\pi}_{q'}(\theta)}, q = 1, \cdots, Q\right\}.$$

Stacking equation (6.6) in q obtains

$$\pi(d_{it}, x_{it}) = \boldsymbol{f}_{it}(\theta, \pi)\pi.$$

The above shows that π_0 uniquely minimizes

$$E\left[(\pi - \boldsymbol{f}_{it}(\theta_0, \pi)\pi)^2\right]$$

over $\pi \in \Delta^{Q-1}$ with the solution satisfying

$$\pi_0 = E\left[\boldsymbol{f}_{it}(\theta_0, \pi_0)\pi_0\right]. \tag{6.8}$$

To operationalize this result, define the smooth mapping $\theta \to (\theta', \pi_0(\theta)')'$ (see Gayle and Namoro [2013] for details of this definition) where for $\theta \in \Theta$, $\pi_0(\theta)$ satisfies

$$\pi_0(\theta) = E[\boldsymbol{f}_{it}(\theta, \pi_0(\theta))\pi_0(\theta)]. \tag{6.9}$$

Then θ_0 minimizes $S_0(\theta; \pi_0(\theta))$ over Θ so that, consistent with equation (6.8), $\pi_0 = \pi_0(\theta_0)$.

Analogously, for $\theta \in \Theta$, let $\hat{\pi}(\theta)$ satisfy

$$\hat{\pi}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{f}_{it}(\theta, \hat{\pi}(\theta))\hat{\pi}(\theta). \tag{6.10}$$

Then $\hat{\theta}$ minimizes $\hat{S}(\theta, \hat{\pi}(\theta))$ over Θ so that $\hat{\pi} = \hat{\pi}(\hat{\theta})$. The reintroduction of dependence on w_k only requires that all of the above theoretical and empirical expectations be taken conditioned on $\{w_i = w_k\}$. Define $I_i(w) = 1\{w_i = w\}$ and $n_k = \sum_{i=1}^{n} I_i(w_k)$ for $k \in \{1, \cdots, K\}$, then with these in hand, the type probabilities may be updated at iteration $o+1$ by solving

$$\pi^{[o+1]}(w_k, \theta^{[o]}) = \frac{1}{n_k}\sum_{i=1}^{n}\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{f}_{it}^{[o+1]}(\theta^{[o]}, \pi^{[o+1]}(w_k, \theta^{[o]}))\pi^{[o+1]}(w_k, \theta^{[o]})I_i(w_k). \tag{6.11}$$

Equation (6.11) may be solved recursively by iterating

$$\pi^{[o'+1]}(w_k; \theta^{[o]}) = \frac{1}{n_k}\sum_{i=1}^{n}\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{f}_{it}^{[o+1]}(\theta^{[o]}, \pi^{[o']}(w_k; \theta^{[o]}))\pi^{[o']}(w_k; \theta^{[o]})I_i(w_k). \tag{6.12}$$

in o' until convergence, where $\pi^{[o]}(w_k; \theta^{[o-1]})$ is taken as the initial prior (see Richardson [1972] for an analysis of Bayesian-based iterative algorithms). The updated type probabilities $\pi^{[o+1]}(w_k; \theta^{[o]})$ are these convergent values.

6.4 Updating θ

With the values $\left(\mathbf{c}^{[o+1]}, \mathbf{p}^{[o+1]}, \pi^{[o+1]}\right)$, the updated values of θ is obtained as follows. Define

$$\begin{aligned}
&\hat{m}^{[o]}(\theta, \pi^{[o]}(\theta)) = \hat{m}\left(\mathbf{c}^{[o]}(\theta), \mathbf{p}^{[o]}(\theta), \pi^{[o]}(\theta)\right), \\
&\hat{M}^{[o]}(\theta, \pi^{[o]}(\theta)) = \frac{\partial}{\partial\theta}\hat{m}^{[o]}(\theta, \pi^{[o]}(\theta)), \\
&\varphi^{[o]}(\theta, \pi^{[o]}(\theta)) = -\left[\hat{M}^{[o]}(\theta, \pi^{[o]}(\theta))'\hat{\Omega}\hat{M}^{[o]}(\theta, \pi^{[o]}(\theta))\right]^{-1}\hat{M}^{[o]}(\theta, \pi^{[o]}(\theta))'\hat{\Omega}\hat{m}^{[o]}(\theta, \pi^{[o]}(\theta)).
\end{aligned}$$

The updated values of θ, $\theta^{[o+1]}$ is given by

$$\theta^{[o+1]} = \theta^{[o]} + \varphi^{[o+1]}(\theta^{[o]}, \pi^{[o+1]}(\theta^{[o]})). \tag{6.13}$$

The full algorithm for computing the estimates of the model is as follows.

Algorithm
1 - Initialize $\left(\pi^{[0]}, \mathbf{c}^{[0]}, \mathbf{p}^{[0]}, \theta^{[0]}\right)$
2 - For $o \geq 0$:
    2.1 - Compute $\mathbf{c}^{[o+1]}$ by solving equation (6.1).
    2.2 - Compute $\mathbf{p}^{[o+1]}$ using equation (6.2).
    2.3 - Compute $\pi^{[o+1]}$ using equations (6.11).
    2.4 - Compute $\theta^{[o+1]}$ using equations (6.13)
until convergence in θ.

Convergence of the Gauss-Newton algorithm is not guaranteed for a variety of reasons, and if it does, convergence may be slow (see Dennis Jr. and Shanbel 1996 for discussion). Methods to improve the success and rate of convergence of the Gauss-Newton algorithm have been proposed in recent years (see Fan and Yuan 2005, Zhou and Chen 2010, and Ferreira et al. 2011). While I adopt components of these proposed algorithms in the simulation exercise and empirical application, a detailed discussion of these modifications is beyond the scope of the current paper.

What we do know, however, is that good initialization of the parameters of the model does improve the likelihood that the algorithm convergences, as well as reduce the number of iterations required to achieve convergence. Initial estimates of the CCCs and CCPs may be used as good initial values for the above algorithm, which iterates on the quantities in a way similar to Aguirregabiria and Mira [2002] and Arcidiacono and Miller [2011]. Therefore, there is not loss of efficiency due to multistage estimation of the parameters of the model such as in Hotz and Miller [1993], Altug and Miller [1998], Bajari et al. [2007] and other papers that apply such methods.

7 Asymptotic properties of the estimator

To derive the asymptotic properties of the estimator $(\hat{\theta}, \hat{\pi})$, some regularity conditions are needed. I use the following notations in all assumptions, theorems, and proofs: $\sup_\theta = \sup_{\theta \in \Theta}$, $\sup_\pi = \sup_{\pi \in \Delta^{Q-1}}$, $\sup_{\theta,\pi} = \sup_{\theta \in \Theta} \sup_{\pi \in \Delta^{Q-1}}$, and $\sum_k = \sum_{k=1}^{K}$. The first assumption imposes the typical random-sampling restriction of the sampling process.

Assumption 7.1. *As sample of n independent realizations is drawn from $F(d,c,x)$. For each $i=1,\cdots,n$, $(d_{it},c_{it},x_{it},t=1,\cdots,T)$ is observed.*

The next assumption imposes restrictions on the parameter space and the admissible functional forms of the period-specific utility functions.

Assumption 7.2. *1. The sets $\mathcal{X}$ and Θ are compact; 2. $\theta_0 \in int(\Theta)$ and for j=0,1,2, and $k=1,\cdots,K$, $\|\partial^j \pi_0(w_k;\theta)/\partial\theta^j\|$ exists at each $\theta \in int(\Theta)$; 3. $E[\|c\|^2]<\infty$ and $E[\|h(x_i,\theta_0)\|^2]<\infty$; 4. $h(x;\theta)$ is twice continuously differentiable at each $\theta \in int(\Theta)$, with $\|\partial^j h(x;\theta)/\partial\theta^j\| \leq \tilde{h}_j(x)$, $j=0,1,2$ for all $\theta\in\Theta$ and some $\tilde{h}_j(x)$ satisfying $E[\tilde{h}_j(x)]<\infty$ for $j=0,1,2$; and 5. For $k=1,\cdots,K$, $n_k/n \to c_k > 0$.*

Assumption 7.3. *$\hat{\Omega}$ is symmetric and positive definite with $\|\hat{\Omega}-\Omega\| = o_p(1)$.*

Some additional definitions and notations are required to proceed. Define

$$m_{it}^{\pi}(\theta,\pi) = \pi - \boldsymbol{f}_{it}(\theta,\pi)\pi, \quad m_i^{\pi}(\theta,\pi) = \frac{1}{T}\sum_{t=1}^{T} m_{it}^{\pi}(\theta,\pi),$$
$$m_0^{\pi}(w;\theta,\pi) = E\left[m_i^{\pi}(\theta,\pi)|w\right], \quad \hat{m}^{\pi}(w_k;\theta,\pi) = \frac{1}{n_k}\sum_{i=1}^{n} m_i^{\pi}(\theta,\pi)I_i(w_k).$$

Let I and ι be the $q-$ dimensional identity matrix and iota vector respectively. Define also $\boldsymbol{\pi} = diag\{\pi_1,\cdots,\pi_Q\}$, and let

$$M_{it}^{\pi}(\theta,\pi) = I - \boldsymbol{f}_{it}(\theta,\pi) + \boldsymbol{\pi}\boldsymbol{f}_{it}(\theta,\pi)\iota'\boldsymbol{f}_{it}(\theta,\pi), \quad M_i^{\pi}(\theta,\pi) = \frac{1}{T}\sum_{t=1}^{T} M_{it}^{\pi}(\theta,\pi),$$
$$M_0^{\pi}(w;\theta,\pi) = E\left[M_i^{\pi}(\theta,\pi)|w\right], \quad \hat{M}^{\pi}(w_k;\theta,\pi) = \frac{1}{n_k}\sum_{i=1}^{n} M_i^{\pi}(\theta,\pi)I_i(w_k).$$

The proof of the following consistency theorem is in Appendix A.2.

Theorem 7.4. *Suppose (i) Assumption 4.1 holds, (ii) Assumption 5.1 holds, and (iii) Assumptions 7.1, 7.2, and 7.3 hold. Then $\hat{\theta} \xrightarrow{p} \theta_0$, and for $k \in \{1,\cdots,D_w\}$,*

$$\left\| \frac{\partial^j}{\partial\theta^j}\hat{\pi}(w_k;\hat{\theta}) - \frac{\partial^j}{\partial\theta^j}\pi_0(w_k;\theta_0) \right\| \xrightarrow{p} 0, \quad j=0,1.$$

Let $m_i = m_i(\theta_0, \pi_0)$, $M_i(\theta, \pi) = \partial m_i(\theta, \pi)/\partial\theta$, and $M_0 = E[M_i(\theta_0, \pi_0)]$, $M_{\pi i}(\theta) = X_i h(x_i, \theta)$, $M_{\pi 0}(w) = E[M_{\pi i}(\theta_0)|w]$, $m_0^\pi(w) = m_0^\pi(w; \theta_0, \pi_0)$ and $M_0^\pi(w) = M_0^\pi(w; \theta_0, \pi_0)$.

Theorem 7.5. *Suppose the conditions of theorem 7.4 hold, and* θ_0 *is in the interior of* Θ. *Then,*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{p} N(0, V),$$

where $V = (M_0'\Omega M_0)^{-1}(M_0'\Omega\Sigma\Omega M_0)(M_0'\Omega M_0)^{-1}$, *and*

$$\Sigma = E\left[\left(m_i + M_{\pi 0}(w_i)M_0^\pi(w_i)^{-1}m_i^\pi\right)'\left(m_i + M_{\pi 0}(w_i)M_0^\pi(w_i)^{-1}m_i^\pi\right)\right].$$

The proof of Theorem 7.5 is in Appendix A.3.

In practice, a consistent estimator for the asymptotic variance V is required. One can be obtained via the plug-in approach, where the parameters in V are replaced with their respective estimators, and the expectations are replaced with sample averages. The proof for consistency of this plug-in estimator is standard and can also be found in Newey and McFadden [1994].

As is standard, the choice of the weighting matrix which minimizes the asymptotic variance of the estimator is $\Omega = \Sigma^{-1}$. A brief discussion of the efficiency of the proposed estimator is in order. Efficiency of an estimator is defined with respect to the class of models considered. Therefore, efficiency of the estimator in this paper cannot be compared to maximum likelihood estimation methods where the distribution of all the shocks is specified (up to a finite dimensional set of parameters), or where the econometrician observes the initial conditions. As discussed in Section 6, the estimator is not a multistage estimator, but rather a profile-likelihood–type estimator where the type probabilities and "concentrated out". The issue of efficiency then becomes whether the estimator incorporates for all information contained in the distribution of the observables with respect to the model specification, which boils down to whether the profile-likelihood type method of estimating θ uses all information to "concentrate out" the type probabilities. Indeed, the type probabilities are updated using all information contained in the fully specified CCPs. I conjecture, therefore, that the estimator proposed is indeed efficient in the class of models set out in Section 2 although a formal proof is beyond the scope of the paper.

8 Monte-Carlo Evidence

In this section, I present three Monte-Carlo exercises to illustrate the finite sample performance of the proposed estimator. The first two of these exercises are based on the clarifying examples presented in Section 3 for which the weights that produce one-period finite dependence are known. The third example is a combination of the previous two for which the weights that produce one-period finite dependence have to be computed numerically to satisfy the conditions specified in equations (3.3) - (3.5).

In all three cases, the number of discrete alternatives is $J = 2$ and the period-specific utility functions is specified as follows:

$$u_t(z_t, c_t, r_t, \varepsilon_t) = d_{2t}\left(\theta_1 + \theta_2 r_{2t} c_{2t} + \theta_3 c_{2t}^2 + s x_{1t} + \theta_4 x_{2t} + \varepsilon_{2t}\right) + d_{1t}\varepsilon_{1t}. \tag{8.1}$$

Define $x_{3t} = \sum_{\tau=1}^{t-1} d_{2\tau}$ and $x_t = (x_{1t}, x_{2t}, x_{3t})$. In all three exercises, $f_{jt}(x_{1t+1}|x_t, c_{jt})$ is distributed discretized normal $N(\mu_{jt}(x_t, c_{jt}), 1)$ with $|x_1| = 5$ support points, where

$$\mu_{jt}(x_t, c_{jt}) = \theta_5 + \theta_6\phi(t) + \theta_7 x_{3t} + \theta_8 d_{2t} + \theta_9 d_{2t} c_{2t}, \text{ and} \tag{8.2}$$

$$\phi(t) = 0.067t - 0.001t^2. \tag{8.3}$$

The random variable x_{2t} is distributed discretized normal $N(\theta_{10} + \theta_{11}\phi(t), 1)$ with $|x_2| = 5$ support points, where $\phi(t)$ is defined in equation (8.3), and r_{it} is distributed discretized normal $N(0, 0.5)$ with $|r| = 3$ support points. In all three designs, I set $\theta_1 = -1$, $\theta_2 = 0.15$, $\theta_3 = -1$, $\theta_4 = 0.2$, $\theta_5 = 2.5$, $\theta_6 = 0.05$, $\theta_{10} = 2.5$ and $\theta_{11} = 0.2$. I set the number of types in the population to two with support $(0.4, 0.6)$ and probabilities $(0.6, 0.4)$, and the shocks, ε_{jt}, $j = 1, 2$ to be distributed i.i.d., type 1 extreme value. I estimate θ_2, θ_4 and the distribution of s in each design.

For the first design, I set $\theta_7 = 0$, $\theta_8 = 0.1$, and $\theta_9 = -0.2$ in which case, Design 1 corresponds to the first clarifying example in Section 2.2. To obtain one-period finite dependence, I set $a_{1,2,t+1} = a_{1,1,t+1} = 0.5$. For Design 2, I set $\theta_7 = -0.002$, $\theta_8 = 0$, and $\theta_9 = 0$ which corresponds to the third clarifying example in Section 2.2. I set $a_{1,2,t+1} = a_{2,1,t+1} = 1$ to obtain finite dependence in Design 2.

For Design 3, I set $\theta_7 = -0.002$, $\theta_8 = 0.1$, and $\theta_9 = -0.2$. It is not know if closed-form

solutions to the weights $a_{kj,t+1}, k, j = 1,2$ that achieves one-period finite-dependence exist, in which case, they must to be computed numerically by solving equations (3.3) - (3.5). Let $\{r_l,\ l = 1 \cdots, |r|\}$ be the support points of r_t and define

$$\begin{aligned} f_{kt+1}(x_{1t+2}|x_{t+1}, s) &= \sum_{l=1}^{|r|} f_{kt+1}(x_{1t+2}|x_{t+1}, c^0_{kt+1}(x_{t+1}, s, r_l)) g_r(r_l), \\ g_{kjt+1}(x_{1t+2}, x_{1t+1}, x_{1t}, s, r) &= f_{kt+1}(x_{1t+2}|x_{t+1}, s) \\ &\quad \times f_{jt}(x_{1t+1}|x_t, c^0_{jt}(x_t, s, r)). \end{aligned} \tag{8.4}$$

By setting $a_{11} = 0.5$ (defined in equation (3.7), I solve for a_{12} using equation (3.10).

I verify equivalence between the alternative representation and the solution by backward recursion in designs 1 and 2 by comparing the implied CCCs and CCPs obtained from the solution of the model to updated CCCs and CCPs obtained from alternative representation when the initial their initial values and are those obtained from the solution. I also perform this comparison for design 3, where the weights that generate one-period finite dependence are calculated by the above method. The difference between the true and updated CCCs are as large as 2E-6 and are approximately 4E-7 on average. The difference between the true and updated CCPs are as large as 3E-5 and are approximately 1E-6 on average.

For each design, the simulated data is generated by solving the dynamic programming for 60 periods and simulating 100 replications of 500 individuals, with the initialization $d_{11} = 1$. Estimation is based on periods 30 - 50, and I contaminate the log CCCs with additive measurement errors, which are distributed i.i.d. normal with zero mean and variance equal to 10% of the variance in the simulated CCCs. Estimating the parameters require the last period (50) CCCs and CCPs, which are obtained by way of extrapolation using equations (6.3) and (6.4).

I compute the estimates of the models using the modifications to the algorithm discussed at the end of Section 6 and I stop the iteration if either

$$\hat{S}\left(\theta^{[o]}\right) - \hat{S}\left(\theta^{[o+1]}\right) < 1E-5, \text{ and } \max\left|\theta^{[o+1]} - \theta^{[0]}\right| < 1E-5.$$

The initial values of θ, CCCs and CCPs are obtained by random perturbation of the true values and convergence is obtained in all simulations.

Table 1 presents the results of the simulation exercises based on the three model designs. The estimator performs well in recovering the true parameters in all three designs.

Table 1: Finite sample properties of the estimator.

$$u_t(z_t, c_t, r_t, \varepsilon_t) = d_{2t}\left(\theta_1 + \theta_2 r_{2t} c_{2t} + \theta_3 c_{2t}^2 + s x_{1t} + \theta_4 x_{2t} + \varepsilon_{2t}\right) + d_{1t}\varepsilon_{1t}$$

	θ_2	θ_4	s_1	s_2	π_1	π_2
True Value	0.15	0.20	0.40	0.60	0.60	0.40
Design 1						
MB	0.0028	-0.0061	-0.0036	-0.0125	-0.0013	0.0013
MAB	0.0030	0.0135	0.0157	0.0153	0.0026	0.0026
RMSE	0.0032	0.0195	0.0252	0.0382	0.0035	0.0035
Design 2						
MB	0.0003	-0.0335	-0.0210	-0.0083	-0.0009	0.0009
MAB	0.0005	0.0390	0.0277	0.0146	0.0020	0.0020
RMSE	0.0007	0.0564	0.0451	0.0226	0.0025	0.0025
Design 3						
MB	-0.0031	-0.0038	-0.0057	0.0061	-0.0009	0.0009
MAB	0.0031	0.0095	0.0115	0.0088	0.0024	0.0024
RMSE	0.0032	0.0133	0.0178	0.0163	0.0028	0.0028

9 Education and Labor Market Choices and the Heterogenous Returns to Education

In this section, I implement the method developed in the previous sections to investigate the life-cycle educational and labor market choices of a sample of young men from the 1979 cohort of the National Longitudinal Survey of Youth (NLSY79), and the resulting long-run "ex-ante" returns to education. Key ingredients of the theoretical model are as follows: I allow for the decision to work while enrolled in school, for hours worked to affect the

likelihood that an individual will advance a grade level, and for heterogeneity in the returns to education, which may be dependent on observed characteristics of the individual.

The inadequacy of the classical Mincer equation to obtain policy-relevant estimates of the returns to education has been well documented over the last two decades (see Heckman et al. [2006] for a review of the relevant literature). Heckman et al. [2008] cite notable extensions to the classical Mincer wage equation that are likely to reduce the biases in estimates of returns to education. These extensions include direct and psychic costs of schooling, non-separability between experience and schooling, heterogeneity in returns to education, and disentangling marginal and average returns to schooling. Other important factors that may affect estimates of the returns to education include the endogeneity of schooling and work experience choices and uncertainty about the completed level of education.

9.1 The theoretical model

The structural model specified below incorporates psychic costs of schooling and working. It accounts for direct benefits from working; the income earned from working and the additional years of experience gained. Nonseparability between working and schooling is accounted for by the simultaneity of these choices, intra-temporal nonseparability in schooling and employment decisions, and dynamic selection. The specification of the log wage equation allows for heterogeneity in the returns to education with the distribution varying across racial groups. Furthermore, ex-ante and ex-post returns to education is disentangled by allowing the returns to education to be uncertain due to uncertainty in the level of completed education, and by allowing for individual-time specific shocks to wages.

The environment a generic individual faces is modeled as follows. In each period, t, the individual is endowed with a fixed amount of time which is normalized to 1. He faces four mutually exclusive and exhaustive alternatives, j: to stay home ($j = 1$); to not attend school and work ($j = 2$); to not work and attend school ($j = 3$); and to both work and attend school ($j = 4$). Let d_{jt} be equal to 1 if the individual chooses alternative j in period t, and zero otherwise. If the individual chooses to work in period t, he must decide how to allocate his time endowment between leisure, l_t, and labor supply, h_t, so that $l_t + h_t = 1$. If the individual chooses not to work in period t, then $l_t = 1$ ($h_t = 0$). Define d_t^h to be equal to 1 if the individual chooses to work in period t, and zero otherwise.

If the individual decides to work in period t, he gains an additional year of experience. If he decides to enroll in school, he advances the grade level with probability $F_t(h_t, x_t^a; \theta^a)$, where $x_t^a = (1, h_t, EDU_t, EDU_t^2, AGE_t, BLACK, AFQT)$, EDU_t is his years of education at period t, AGE_t is his age at period t, $BLACK$ is equal to 1 if the individual is black, and zero otherwise, and $AFQT$ is his Armed Forces qualification test score. This specification is a key (though not only) source of uncertainty affecting completed level of education. The individual considers the benefits from working while in school, which include income generated and the level of labor market experience earned, to the costs, which include loss of leisure time and the potentially negative impact of working while in school on the probability of advancing the grade level.

In each period, the individual receives a wage offer $wage_t^o$, which is parameterized as follows:

$$\ln(wage_t^0) = \theta_1^w + \theta_2^w EXPER_t + \theta_3^w EXPER_t^2 + \theta_4^w AFQT + \theta_6^w BLACK + sEDU_t + r_t$$

where $EXPER_t$ is his years of experience as at period t, s is returns to his level of education, and r_t is the period-specific shock to his wage offer which is assumed to be distributed i.i.d. $N(0, \sigma_r)$. We assume returns to education, s, is discretely distributed with Q support points, $s \in \{s_1, \cdots, s_Q\}$ and corresponding PMFs $\pi(w) = (\pi_1(w), \cdots, \pi_Q(w))$, where w is a set of observed characteristics. I allow for the log of wage offer to be measured with additive error, which is assumed to have zero mean and uncorrelated with the all variables in the model. I select w to represent the two groups of race in the data, black and white.

The contemporaneous utility function is given by

$$\begin{aligned} u_t(z_t) = {} & wage_t^o \ln(wage_t^o h_t + 1 - d_t^h) + \exp(\theta_1^u x_t^l) \ln(l_t) \\ & + \theta_2^u d_t^E x_t^E + \theta_3^u d_t^h x_t^h + \theta_4^u d_t^h d_t^E + d_t' \varepsilon_t, \end{aligned} \tag{9.1}$$

where d_t^E is equal to 1 if the individual enrolls in school at period t, and zero otherwise, x_t^l, x_t^E, and x_t^h are observed demographic characteristics that capture individual variation leisure, the psychic costs of schooling, and the psychic costs of working, respectively, $d_t = (d_{1t}, \cdots, d_{Jt})'$, and $\varepsilon_t = (\varepsilon_{1t}, \cdots, \varepsilon_{Jt})'$, where ε_{jt} is the alternative j specific shock to utility, which is distributed i.i.d., type 1 extreme value.

9.2 Data

The data are taken from the 1979 youth cohort of the National Longitudinal Survey of Labor Market Experience (NLSY79), a comprehensive panel data set that follows individuals over the period 1979 to 2000, who were 14 to 21 years of age as of January 1, 1979. The data set initially consisted of 12,686 individuals: a representative sample of 6,111 individuals, a supplemental sample of 5,295 Hispanics, non-Hispanic blacks, and economically disadvantaged, non-black, non-Hispanics, and a supplemental sample of 1,280 military youth. Interviews were conducted on an annual basis though 1994, after which they adopted a biennial interview schedule. This study makes use of the individuals observed for each of the first 27 years of interviews, from 1979 to 2006. The data are restricted to include non-Hispanic males and respondents with missing observations on the highest grade level completed that cannot be recovered with high confidence from other data information. (Further details on the sample restrictions are provided in see Gayle [2006]). With these restriction, the data used in this application consists of 1637 individuals observed over 27 year.

9.3 Estimation

There are three key endogenous state variables that need to be considered to obtain finite dependence; wage offer, years of labor market experience, and level of formal education. The only distribution that needs to be estimated within the model is the wage offer distribution, because the conditional distribution of years of experience is degenerate and the probability of advancing the a grade level can be estimated outside the model, which is done in this application. In the following, I describe the weights used to achieve one-period finite dependence in equation (3.15) for this application.

To compute $v_{2t} - v_{1t}$, I set $a_{12,t+1} = a_{21,t+1} = 1$. To compute $v_{3t} - v_{1t}$, I estimate θ^a (denoted $\hat{\theta}^a$) outside the model and construct

$$\begin{aligned} f_{kj}(z_{t+2}, z_{t+1}) = f_{kt+1}(EDU_{t+2}|x^a_{t+1}, h^0_{t+1} = 0; \hat{\theta}^a) \\ \times f_{jt}(EDU_{t+1}|x^a_t, h_t = 0; \hat{\theta}^a), \end{aligned} \tag{9.2}$$

where $f_{jt}(EDU_{t+1}|x^a_t, h_t; \theta^a)$ is the conditional distribution of EDU_{t+1}, which is a function

of the probability of advancing. The weights are then computed using equation (3.10), with weight associated with the conditionally degenerate transition function, a_{11} to zero.

Let $x_t^w = (1, EXPER_T, EXPER_t^2, AFQT, BLACK, EDU_t)$ and define

$$f_{kt+1}(EDU_{t+2}|x_{t+1}^a, x_{t+1}^l, x_{t+1}^w; \gamma) = \int f_{kt+1}(EDU_{t+2}|x_{t+1}^a, h_{t+1}^0(x_{t+1}^a, x_{t+1}^l, wage_{t+1}^o); \theta^a) \times f_{t+1}^w(wage_{t+1}^0|x_{t+1}^w) dwage_{t+1}^0, \tag{9.3}$$

which is transition function for EDU_{t+2} unconditioned on $wage_{t+1}^0$. To compute $v_{4t} - v_{1t}$, I estimate γ (denoted $\hat{\gamma}$) outside the model, construct

$$f_{kj}(z_{t+2}, z_{t+1}) = f_{kt+1}(EDU_{t+2}|x_{t+1}^a, x_{t+1}^l, x_{t+1}^w; \hat{\gamma}) f_{jt}(EDU_{t+1}|x_t^a, h_t; \hat{\theta}^a),$$

and compute the weights using equation (3.10) with $a_{11} = 0$. To understand the validity of these weights, notice the wage offer distribution is affected by the individual's actions only through education and experience. Therefore, the weights can be computed by first integrating the conditional distribution of EDU_{t+2} with respect to the period $t+1$ wage offer distribution and then use these integrated conditional distributions to compute the weights that achieve one-period finite dependence.

9.4 Results

9.4.1 Grade-promotion probability

Table 1 presents the results from estimation of the grade-transition probability under the assumption of a logit probability of advancing to the next grade level given enrollment. The results indicate that given employment, additional hours worked reduces the probability of advancing to the next grade level. However, a positive and statistically significant extensive margin of employment also exists, which is captured by the coefficient on d_t^h. These two results capture the crowding-out and congruence hypothesis. The former hypothesis states that working while enrolled in school crowds out time spent on school activities, thus reducing the chances of completing the grade level. The latter hypothesis states that working moderate hours while enrolled in school improves organization skills, which results in more

efficient school activities and higher chances of completing the grade level.

The results imply that, ceteris paribus, working more than 14 hours per week during the school year has a negative total impact on the probability of completing the grade level. In the data, approximately 47% of black males work while enrolled in school, and 24% of them work more than 14 hours, whereas 64% of white males work while enrolled in school, and 35% of them work for more 14 hours. White makes also receive (80 cents) higher hourly wage rates while enrolled in school. For any given level of hours worked, the probability of completing the grade level is higher for whites than for blacks because of the higher level of the AFQT score, which has a positive and significant coefficient in the grade promotion probability estimate. These results suggest that white males take advantage of their higher schooling ability/preparedness (as measured by the AFQT score) by accumulating higher income and labor market experience during their school years at a lower risk of lower completed level of education, which is due to the fact that their labor market activities on the chances of them completing a grade level is mitigated by their higher average abilities/preparedness.

Table 2 presents the result from estimation of the approximation of the unconditional (on hours worked) grade transition probability. I employ a flexible logit model with regressors (x_t^a, x_t^l, x_t^w), as implied by equation (9.3). Both models of grade transition probabilities are estimated with a sample of 3601 individuals over a average period of 4 years, and all coefficients are estimated precisely.

9.4.2 Period-specific utility

Table 3 presents the estimates of the parameters governing the period-specific utility. The results indicate the utility of leisure is increasing and concave in age, and higher for blacks compared to whites. A key quantity of interest is the psychic value of school attendance and labor market participation. As Heckman et al. [2008] discuss, the existence of psychic costs of schooling drives a wedge between the Mincer rate of return and the internal rate of return to education. Indeed, Heckman et al. [2008] show that if psychic costs of schooling are significant and ignored, the Mincer coefficient is expected to be larger than the internal rate of return. The results in panel two of Table 3 imply the existence of psychic costs of school attendance which increases with age. I also find evidence of the preference for continuous schooling. However, these quantities seem too small to be of economic relevance. Complete

analysis of the economic relevance of the psychic cost of schooling requires the model to simulated to see the size of the impact of these variables on individual choices and outcomes.

Performing the same analysis as done in Heckman et al. [2008] where the psychic cost of working is included shows this cost also drives a wedge between the Mincer coefficient and the internal rate of return to education. However, the size of the Mincer coefficient relative to the internal rate of return is ambiguous. The results in panel 3 of Table 3 suggest the existence of significant psychic *values* of labor market participation which is lower for blacks and decreases with age. The results also suggest significant preference for continuous employment and distaste for working and attending school at the same time. These quantities are relatively large and are almost certainly economically relevant.

9.5 Wage offer equation

Table 4 presents the results from estimation of the wage offer function. I include a third-order polynomial in time to account for aggregate movements in wage offer. This inclusion implicity assumes the individuals have perfect foresight of aggregate shocks to wage offers. A alternative specification, which is outside the scope of this application, is to model aggregate movements in wage offers as a dynamic random process. I opt for the assumption of perfect foresight in aggregate wages over ignoring these aggregate movements because the NLSY79 follows a single cohort over time, in which case the effect of labor market experience and education achievement may be confounded with omitted aggregate movements in wage offers.

I assume the support of returns to education takes on three values ($Q = 3$). I estimated the model with four support points and found two of these points to be statistically indistinguishable. The results in Table 4 indicate the support of returns to education ranges from 0.07 to 0.08 and the Wald test supports the claim of support points at the 5% level of significance. This range of returns to education lie within the range of the estimates from other studies that implement instrumental variables methods and data from similar time periods (see Card [1999] for a review of these studies), but at the lower end of estimates produced in Heckman et al. [2008]. The rest of the results from estimation of the wage offer equation are standard: the coefficient on AFQT scores is positive and significant while the coefficient of black is small and statistically insignificant (see Neal and Johnson [1996]), wage offer

is increasing and concave in labor market experience (although the negative coefficient on squared experience is statistically insignificant). The results indicate positive and significant individual-time specific shocks to wages. The standard deviation of wage offer shocks are lower than those estimated in standard static sample selection models, implying the existence of significant measurement errors in observed wages.

9.5.1 Returns to education

Table 5 reports the estimates of the distribution of the returns to education by racial groups, along with their respective means and standard deviations. The results show significant heterogeneity in the returns to education for both black and white males. In particular, the results suggests that white males receive rates of returns to education of 0.071 with probability 0.14, 0.078 with probability 0.11, and 0.081 with probability 0.75. On the other hand black males receive rates of returns to education of 0.071 with probability 0.39, 0.078 with probability 0.20, and 0.081 with probability 0.41. Therefore, white males are 34 percentage points more likely than black males to receive high returns to education, while black males are 25 percentage points more likely than white males to receive low returns to education.

10 Conclusion

CCP estimation of dynamic structural models has flourished over the two decades, largely because of the potential to dramatically reduce computational costs. The current state of art shows the expected value of future utilities from optimal decision making can always be expressed as functions of the flow payoffs and conditional choice probabilities for any sequence of future choices. Any future choice sequence chosen for a given initial choice generates a corresponding sequence of distributions of states. The term ρ-period finite dependence is obtained if two distinct initial choices with two corresponding future choice sequences can be constructed so that their respective distributions of states are the same after ρ-periods in the future. The smaller ρ is, the better, in that the computational cost reduces and, in many cases, the accuracy of the estimator improves. Though innovative, this property is restrictive and application of the CCP method often requires strong assumptions about the choice sequence, or the state transition probabilities, or both.

I generalize the concept of ρ-period dependence in a way that overcomes these two issues. I show that my generalization obtains one-period dependence ($\rho = 1$) in a large class of dynamic structural models, which includes most models that have been estimated using the CCP method. Moreover, the method proposed in this paper imposes minimal restrictions on the state transition probabilities, greatly increasing the scope of models the CCP method can estimate.

The class of models I consider is not restricted to models of discrete choices, but also includes models of both discrete and continuous choices. The representation developed in this model allows for a simple and elegant optimality condition for the continuous choices.

I show how to include discrete-valued unobserved heterogeneity into my model. I allow for the distribution of the unobserved heterogeneity to depend on observed covariates of the model. I provide sufficient conditions for identification of all the parameters of the model, including the conditional distribution of unobserved heterogeneity, and I propose a GMM estimator for these parameters. I propose an algorithm to compute the estimator. A key feature of the algorithm is how the distribution of unobserved heterogeneity is updated. I propose an iterated Bayes method, which jointly estimates the number of types. I am not aware of other methods with this property in the GMM estimation framework.

I derive the asymptotic distribution of the estimator and investigate its finite sample properties by way of Monte Carlo methods. Three environments are considered. The first two are models for which the weights that achieve one-period finite dependence are closed form while the third requires computation of these weights. The results show that the proposed estimator performs well in all three environments.

Finally, I implement my method to estimate a dynamic structural model of educational attainment of labor supply using data from the NLSY79 with the ultimate goal of estimating the distribution of the returns to education. The results suggest the existence of significant heterogeneity in the returns to education, ranging between 0.07 to 0.08, which aligns with studies of IV estimation of the returns to education, using the same data set as well as other data sets with comparable cohorts. The distribution depends on categories of race in that white males are significantly more likely to receive high levels of returns to education compared to black males.

A LEMMA AND THEOREMS

A.1 Proof of Theorem 4.2

Proof. It is sufficient to prove identification for $J = 2$ with scalar random effects, using information from the CCPs and suppressing the CCCs and the time dimension. Define

$$\mathcal{P}_1(B) = \left\{ p^0(x, B, s_q) : x \in \mathcal{X}, B \in \mathcal{B}, s_q \in \Re, q \geq 1, 2, \cdots \right\}, \tag{A.1}$$

$$\mathcal{P}_2(w, B, \Pi(w)) = \left\{ p^0(x, B, \Pi(w)) : p^0(x, B, \Pi(w)) = \sum_{q=1}^{Q} p^0(x, B, s_q)\pi_q(w), \pi_q(w) > 0, \right.$$
$$\left. \sum_{q=1}^{Q} \pi_q(w) = 1, p^0(x, B, s_q) \in \mathcal{P}_1(B), w \in \Re^{D_w}, Q \geq 1, 2, \cdots \right\}. \tag{A.2}$$

Under Assumptions 4.1.1 - 4.1.4, $\mathcal{P}_1(B)$ is a linearly independent set for any fixed $B \in \mathcal{B}$ with probability one. Therefore, $p^0(x, B, \Pi(w)) \in \mathcal{P}_2(w, B)$ has a unique representation as a linear combination of finitely may elements of $\mathcal{P}_1(B)$ (see Kreyszig [1989]), that is, any $\beta \in \mathcal{B}$ induces a unique corresponding finite mixing distribution $\Pi(w; B) = (\{s_1(B), \cdots, s_Q(B)\}, \pi(w; B), Q(B))$ (s_q and Q are not functions of w by assumption). By hypothesis,

$$E[d|x] = p^0(x, B_0, \Pi(w; B_0)) = \sum_{q=1}^{Q_0} p^0(x, B_0, s_q(B_0))\pi_q(w; B_0).$$

Suppose $\tilde{B} \neq B_0$ exists for which $E[d|x] = p^0(x, \tilde{B}, \Pi(w; \tilde{B}))$ so that

$$\sum_{q=1}^{Q(\tilde{B})} p^0(x, \tilde{B}, s_q(\tilde{B}))\pi_q(w; \tilde{B}) = \sum_{q=1}^{Q(B_0)} p^0(x, B_0, s_q(B_0))\pi_q(w; B_0) \tag{A.3}$$

for almost every x in the support of $f_{x|w}$. Because the weights sum to one, equation (A.3) can be written as follows:

$$p^0(x,\tilde{B},s_1(\tilde{B}))-p^0(x,B_0,s_1(B_0))=\sum_{q=1}^{Q(B_0)}(p^0(x,B_0,s_q(B_0))-p^0(x,B_0,s_1(B_0)))\pi_q(w;B_0)$$
$$-\sum_{q=1}^{Q(\tilde{B})}(p^0(x,\tilde{B},s_q(\tilde{B}))-p^0(x,\tilde{B},s_1(\tilde{B})))\pi_q(w;\tilde{B}) \quad \text{(A.4)}$$

for almost every x in the support of $f_{x|w}$. Suppose $v_{21}(x_t,s_{q'};B)>v_{21}(x_t,s_q;B)$, $s_{q'}\neq s_q$ on $\tilde{\mathcal{X}}$. Then, under Assumption 4.1.2, $p^0(x,B,s_{q'})>p^0(x,B,s_q)$ on $\tilde{\mathcal{X}}$, in which case, Assumption 4.1.5.i allows for rearrangement of the elements of s_0 and $\tilde{s}$ to have $p^0(x,B_0,s_1(B_0))<p^0(x,B_0,s_q(B_0)),\ q=2,\cdots,Q(B_0)$ and $p^0(x,\tilde{B},s_1(\tilde{B}))>p^0(x,\tilde{B},s_q(\tilde{B})),\ q=2,\cdots,Q(\tilde{B})$. Then, the RHS of equation (A.4) is non-negative for all $x\in\tilde{\mathcal{X}}$, because the weight are strictly positive, but Assumption 4.1.5.ii implies that an element $\tilde{x}\in\tilde{\mathcal{X}}$ exists for which $p^0(\tilde{x},\tilde{B},s_1(\tilde{B}))<p^0(\tilde{x},B_0,s_1(B_0))$ making the LHS of equation (A.4) negative, contradicting equation (A.4). So $\tilde{B}=B_0$, which in turn implies $\Pi(w;\tilde{B})=\Pi(w;B_0)=\Pi_0(w)$. The rest of the proof $(v_{21}(x_t,s_{q'};B)<v_{21}(x_t,s_q;B))$ follows the same argument.

□

A.2 Proof of Theorem 7.4

Proof. For $k=1,\cdots,D_w$, define $\hat{h}(w_k,\theta)=\sum_{i=1}^n h(x_i,\theta)I_i(w_k)/n_k$ and $h_0(w_k\theta)=E[h(x_i,\theta)|w_i=w_k]$. Under Assumptions 7.1, 7.2.1, 7.2.3, and 7.2.5, Lemma 2.4 of Newey and McFadden [1994] implies

$$\sup_\theta\left\|\frac{\partial^j}{\partial\theta^j}\hat{h}(w_k,\theta)-\frac{\partial^j}{\partial\theta^j}h_0(w_k,\theta)\right\|\xrightarrow{p}0,\quad j=0,1,2. \quad \text{(A.5)}$$

I suppress the dependence of the type probabilities and empirical and theoretical expectations and on w. Note the denominator term in $\boldsymbol{f}_{it}(\theta,\pi)$ satisfies $0<\sum_q f(d_{it}|x_{it},\theta_q)\pi_q<1$ uniformly over $\theta\in\Theta$ and $\pi\in\triangle^{Q-1}$. Then, by defining $\boldsymbol{f}_i(\theta,\pi)=\sum_{i=1}^T\boldsymbol{f}_{it}(\theta,\pi)/T$, $\hat{\boldsymbol{f}}(\theta,\pi)=$

$\sum_{i=1}^{n} \boldsymbol{f}_i(\theta,\pi)/n$ and $\boldsymbol{f}_0(\theta,\pi) = E[\boldsymbol{f}_i(\theta,\pi)]$, equation (A.5) implies

$$\sup_{\theta,\pi} \left\| \frac{\partial^{j+l}}{\partial\theta^j \pi^l} \hat{\boldsymbol{f}}(\theta,\pi) - \frac{\partial^{j+l}}{\partial\theta^j \pi^l} \boldsymbol{f}_0(\theta,\pi) \right\| \xrightarrow{p} 0, \quad j,l = 0,1,2. \tag{A.6}$$

Consistent with equation (6.7), $\hat{m}_q^{\pi}(\theta,\hat{\pi}(\theta)) = 0$ for any $\theta \in \Theta$. The mean value expansion around $\pi_0(\theta)$ obtains

$$\begin{aligned} 0 = {} & \hat{m}_q^{\pi}(\theta,\pi_0(\theta)) \\ & + \frac{1}{nT} \sum_{i,t} \left[\left(1 - \frac{f(d_{it}|x_{it},\theta_q)}{\sum_{q'=1}^{Q} f(d_{it}|x_{it},\theta_{q'})\overline{\pi}_{q'}(\theta)} \right) \left(\hat{\pi}_q(\theta) - \pi_{0q}(\theta) \right) \right. \\ & + \frac{f(d_{it}|x_{it},\theta_q)\overline{\pi}_q}{\sum_{q'=1}^{Q} f(d_{it}|x_{it},\theta_{q'}(\theta))\overline{\pi}_{q'}(\theta)} \\ & \left. \times \sum_{q' \neq q} \frac{f(d_{it}|x_{it},\theta_{q'})}{\sum_{q''=1}^{Q} f(d_{it}|x_{it},\theta_{q''})\overline{\pi}_{q''}(\theta)} \left(\hat{\pi}_{q'}(\theta) - \pi_{0q'}(\theta) \right) \right] \end{aligned} \tag{A.7}$$

identically over θ over Θ, where $\overline{\pi}(\theta)$ are mean values. Stacking equation (A.7) in q obtains

$$0 = \hat{m}^{\pi}(\theta,\pi_0(\theta)) + \hat{M}^{\pi}(\theta,\overline{\pi}(\theta)) \left(\hat{\pi}(\theta) - \pi_0(\theta) \right). \tag{A.8}$$

Note that each $\overline{\pi}_q(\theta)$ is strictly positive so $M_{it}^{\pi}(\theta,\overline{\pi}(\theta))$ can be written as follows

$$M_{it}^{\pi}(\theta,\overline{\pi}(\theta)) = \overline{\boldsymbol{\pi}}(\theta) \left[\overline{\boldsymbol{\pi}}(\theta)^{-1} \left(I - \boldsymbol{f}_{it}(\theta,\overline{\pi}) \right) + \boldsymbol{f}_{it}(\theta.\overline{\pi}) \iota\iota' \boldsymbol{f}_{it}(\theta,\overline{\pi}) \right].$$

Note also that $\overline{\boldsymbol{\pi}}(\theta)^{-1} \left(I - \boldsymbol{f}_{it}(\theta,\overline{\pi}) \right) + \boldsymbol{f}_{it}(\theta.\overline{\pi}) \iota\iota' \boldsymbol{f}_{it}(\theta,\overline{\pi})$ is symmetric with strictly positive elements, and

$$\begin{aligned} & det \left[\overline{\boldsymbol{\pi}}(\theta)^{-1} \left(I - \boldsymbol{f}_{it}(\theta,\overline{\pi}) \right) + \boldsymbol{f}_{it}(\theta.\overline{\pi}) \iota\iota' \boldsymbol{f}_{it}(\theta,\overline{\pi}) \right] \\ & = \left[1 + \iota' \boldsymbol{f}_{it}(\theta,\overline{\pi}) \left(I - \boldsymbol{f}_{it}(\theta,\overline{\pi}) \right)^{-1} \overline{\boldsymbol{\pi}}(\theta) \boldsymbol{f}_{it}(\theta.\overline{\pi}) \iota \right] det \left[\overline{\boldsymbol{\pi}}(\theta)^{-1} \left(I - \boldsymbol{f}_{it}(\theta,\overline{\pi}) \right) \right] \\ & > 0 \end{aligned} \tag{A.9}$$

so that $\overline{\boldsymbol{\pi}}(\theta)^{-1} \left(I - \boldsymbol{f}_{it}(\theta,\overline{\pi}) \right) + \boldsymbol{f}_{it}(\theta.\overline{\pi}) \iota\iota' \boldsymbol{f}_{it}(\theta,\overline{\pi})$ is symmetric and positive definite, which in turn implies that $\frac{1}{nT} \sum_{i,t} \overline{\boldsymbol{\pi}}(\theta)^{-1} \left(I - \boldsymbol{f}_{it}(\theta,\overline{\pi}) \right) + \boldsymbol{f}_{it}(\theta.\overline{\pi}) \iota\iota' \boldsymbol{f}_{it}(\theta,\overline{\pi})$ is symmetric and pos-

itive definite. The diagonal matrix $\boldsymbol{\pi}(\theta)$ is also symmetric and positive definite. From these results, conclude that $\hat{M}^{\pi}(\theta,\bar{\pi}(\theta))$ is invertible so that equation (A.8) obtains

$$\hat{\pi}(\theta)-\pi_0(\theta)=-\hat{M}^{\pi}(\theta,\bar{\pi}(\theta))^{-1}\hat{m}^{\pi}(\theta,\pi_0(\theta)) \tag{A.10}$$

identically in θ over Θ. Dy definition,

$$\hat{\pi}(\theta)=\hat{\boldsymbol{f}}(\theta,\hat{\pi}(\theta))\hat{\pi}(\theta) \text{ and} \tag{A.11}$$

$$\pi_0(\theta)=\boldsymbol{f}_0(\theta,\pi_0(\theta))\pi_0(\theta) \tag{A.12}$$

hold identically in θ over Θ, and equation (A.12) implies that $\hat{m}^{\pi}_{it}(\theta,\pi_0(\theta))$ can be written as follows.

$$m^{\pi}_i(\theta,\pi_0(\theta))=\boldsymbol{f}_0(\theta,\pi_0(\theta))\pi_0(\theta)-\boldsymbol{f}_i(\theta,\pi_0(\theta))\pi_0(\theta). \tag{A.13}$$

Equations (A.6) and (A.13), along with Assumption 7.2.3 imply

$$\sup_{\theta}\left\|\frac{\partial^j}{\partial\theta^j}\hat{m}^{\pi}(\theta,\pi_0(\theta))\right\|\xrightarrow{p}0 \quad j=0,1,2. \tag{A.14}$$

Because $\boldsymbol{f}_{it}(\theta,\pi)$ is a diagonal matrix with probabilities that sum to one, $\sup_{\theta,\pi}\|\boldsymbol{f}_{it}(\theta,\pi)\|=1$. Therefore,

$$\begin{aligned}\|M^{\pi}_{it}(\theta,\pi)\|&\leq\|Q+1+\|\boldsymbol{\pi}\|\|\boldsymbol{f}_{it}(\theta,\pi)\iota\iota'\boldsymbol{f}_{it}(\theta,\pi)\|\leq\mathcal{T}_1<\infty\\ \left\|\frac{\partial}{\partial\pi}M^{\pi}_{it}(\theta,\pi)\right\|&=\left\|(2\boldsymbol{\pi}\boldsymbol{f}_{it}(\theta,\pi)\iota\iota'-I)\frac{\partial}{\partial\pi}\boldsymbol{f}_{it}(\theta,\pi)+\boldsymbol{f}_{it}(\theta,\pi)\iota\iota'\boldsymbol{f}_{it}(\theta,\pi)\right\|\\ &\leq\mathcal{T}_2+\mathcal{T}_3\left\|\frac{\partial}{\partial\pi}\boldsymbol{f}_{it}(\theta,\pi)\right\|,\end{aligned}$$

where $\mathcal{T}_2$ and $\mathcal{T}_3$ are positive and finite constants. These results and equation (A.6) implies that

$$\sup_{\theta,\pi}\left\|\frac{\partial^j}{\partial\pi^j}\hat{M}^{\pi}(\theta,\pi)-\frac{\partial^j}{\partial\pi^j}M^{\pi}_0(\theta,\pi)\right\|\xrightarrow{p}0, \quad j=0,1. \tag{A.15}$$

Equations (A.10), (A.14) and (A.15) obtain

$$\sup_{\theta} \|\hat{\pi}(\theta) - \pi_0(\theta)\| \xrightarrow{p} 0. \tag{A.16}$$

The mean-value theorem, equations (A.6), (A.15), (A.16), and Assumption 7.2.2 imply

$$\begin{aligned}
&\sup_{\theta} \left\| \hat{M}^{\pi}(\theta, \hat{\pi}(\theta)) - M_0^{\pi}(\theta, \pi_0(\theta)) \right\| \\
&= \sup_{\theta} \left\| \hat{M}^{\pi}(\theta, \hat{\pi}(\theta)) - M_0^{\pi}(\theta, \hat{\pi}(\theta)) + \frac{\partial}{\partial \pi} M_0^{\pi}(\theta, \bar{\pi}(\theta))(\hat{\pi}(\theta) - \pi_0(\theta)) \right\| \\
&\leq \sup_{\theta, \pi} \left\| \hat{M}^{\pi}(\theta, \pi) - M_0^{\pi}(\theta, \pi) \right\| + \sup_{\theta, \pi} \left\| \frac{\partial}{\partial \pi} M_0^{\pi}(\theta, \pi) \right\| \sup_{\theta} \|\hat{\pi}(\theta) - \pi_0(\theta)\| \\
&= o_p(1).
\end{aligned} \tag{A.17}$$

$$\begin{aligned}
&\sup_{\theta} \left\| \frac{\partial}{\partial \theta} \hat{\boldsymbol{f}}(\theta, \hat{\pi}(\theta)) - \frac{\partial}{\partial \theta} \boldsymbol{f}_0(\theta, \pi_0(\theta)) \right\| \\
&= \sup_{\theta} \left\| \frac{\partial}{\partial \theta} \hat{\boldsymbol{f}}(\theta, \hat{\pi}(\theta)) - \frac{\partial}{\partial \theta} \boldsymbol{f}_0(\theta, \hat{\pi}(\theta)) + \frac{\partial^2}{\partial \theta \partial \pi} \boldsymbol{f}_0(\theta, \bar{\pi}(\theta))(\hat{\pi}(\theta) - \pi_0(\theta)) \right\| \\
&\leq \sup_{\theta, \pi} \left\| \frac{\partial}{\partial \theta} \hat{\boldsymbol{f}}(\theta, \pi) - \frac{\partial}{\partial \theta} \boldsymbol{f}_0(\theta, \pi) \right\| + \sup_{\theta, \pi} \left\| \frac{\partial^2}{\partial \theta \partial \pi} \boldsymbol{f}_0(\theta, \pi) \right\| \sup_{\theta, \pi} \|\hat{\pi}(\theta) - \pi_0(\theta)\| \\
&= o_p(1).
\end{aligned} \tag{A.18}$$

Because equations (A.11) and (A.12) holding identically in θ over Θ, the envelope condition implies, that

$$\hat{M}^{\pi}(\theta, \hat{\pi}(\theta)) \frac{\partial}{\partial \theta} \hat{\pi}(\theta) + \hat{\pi}(\theta)' \otimes I \frac{\partial}{\partial \theta} vec\left(\hat{\boldsymbol{f}}(\theta, \hat{\pi}(\theta)) \right) = 0, \tag{A.19}$$

$$M_0^{\pi}(\theta, \pi_0(\theta)) \frac{\partial}{\partial \theta} \pi_0(\theta) + \pi_0(\theta)' \otimes I \frac{\partial}{\partial \theta} vec\left(\boldsymbol{f}_0(\theta, \pi_0(\theta)) \right) = 0, \tag{A.20}$$

hold identically in θ over $int(\Theta)$, where $\otimes$ is the Kronecker product operator and *vec* is the vectorization operator. By noting the equality $\hat{a}\hat{b} - ab = (\hat{a} - a)(\hat{b} - b) + a(\hat{b} - b) + (\hat{a} - a)b$,

the difference between equations (A.20) and (A.19) gives

$$\hat{M}^{\pi}(\theta,\hat{\pi}(\theta))\left(\frac{\partial}{\partial\theta}\hat{\pi}(\theta)-\frac{\partial}{\partial\theta}\pi_0(\theta)\right)=-\left(\hat{M}^{\pi}(\theta,\hat{\pi}(\theta))-M_0^{\pi}(\theta,\pi_0(\theta))\right)\frac{\partial}{\partial\theta}\pi_0(\theta)$$
$$+(\hat{\pi}(\theta)-\pi_0(\theta))'\otimes I\left(\frac{\partial}{\partial\theta}vec\left(\hat{\boldsymbol{f}}(\theta,\hat{\pi}(\theta))\right)-\frac{\partial}{\partial\theta}vec\left(\boldsymbol{f}_0(\theta,\pi_0(\theta))\right)\right)$$
$$+\pi_0(\theta)'\otimes I\left(\frac{\partial}{\partial\theta}vec\left(\hat{\boldsymbol{f}}(\theta,\hat{\pi}(\theta))\right)-\frac{\partial}{\partial\theta}vec\left(\boldsymbol{f}_0(\theta,\pi_0(\theta))\right)\right)$$
$$+(\hat{\pi}(\theta)-\pi_0(\theta))'\otimes I\frac{\partial}{\partial\theta}vec\left(\boldsymbol{f}_0(\theta,\pi_0(\theta))\right). \tag{A.21}$$

Equations (A.16) - (A.21) and Assumption 7.2.2 imply that

$$\sup_{\theta}\left\|\frac{\partial}{\partial\theta}\hat{\pi}(\theta)-\frac{\partial}{\partial\theta}\pi_0(\theta)\right\|\xrightarrow{p}0. \tag{A.22}$$

Now, under Assumptions 7.1, 7.2.1, and 7.2.4,

$$\|m_i(\theta,\pi_0(\theta))\|\leq\|X_i\|(\|y_i\|+\|h(x_i,\theta)\|\|\pi_0(\theta)\|)\leq\mathcal{T}_1\|y_i\|+\mathcal{T}_2\tilde{h}_0(x_i)$$

for positive and finite constants $\mathcal{T}_1$ and $\mathcal{T}_2$ so that Assumption 7.2.3 and Lemma 2.4 of Newey and McFadden [1994] imply

$$\sup_{\theta}\|\hat{m}(\theta,\pi_0(\theta))-m_0(\theta,\pi_0(\theta))\|\xrightarrow{p}0. \tag{A.23}$$

Therefore, equations (A.29), (A.16), and (A.23) imply

$$\|\hat{m}(\theta,\hat{\pi}(\theta))-m_0(\theta,\pi_0(\theta))\|\leq\|\hat{m}(\theta,\hat{\pi}(\theta))-m_0(\theta,\hat{\pi}(\theta))\|$$
$$+\|E[X_ih(x_i,\theta)]\|\,\|\hat{\pi}(\theta)-\pi_0(\theta)\|$$
$$+\left\|\hat{M}_{\pi}(\theta)-M_{\pi0}(\theta)\right\|\|\hat{\pi}(\theta)-\pi_0(\theta)\|\xrightarrow{p}0, \tag{A.24}$$

which, along with Assumption 7.3 obtains $\hat{\theta}\xrightarrow{p}\theta_0$. Finally, this result, Assumption 7.2.2,

equations (A.16), (A.22), and (A.23) imply

$$\begin{aligned}\|\hat{\pi}(\hat{\theta})-\pi_0(\theta_0)\| &= \left\|\hat{\pi}(\hat{\theta})-\pi_0(\hat{\theta})+\frac{\partial\pi_0(\bar{\theta})}{\partial\theta}(\hat{\theta}-\theta_0)\right\| \\ &\leq \sup_\theta\|\hat{\pi}(\theta)-\pi_0(\theta)\|+\sup_\theta\left\|\frac{\partial\pi_0(\theta)}{\partial\theta}\right\|\|\hat{\theta}-\theta_0\| \xrightarrow{p} 0, \qquad \text{(A.25)}\end{aligned}$$

and

$$\begin{aligned}\left\|\frac{\partial}{\partial\theta}\hat{\pi}(\hat{\theta})-\frac{\partial}{\partial\theta}\pi_0(\theta_0)\right\| &= \left\|\frac{\partial}{\partial\theta}\hat{\pi}(\hat{\theta})-\frac{\partial}{\partial\theta}\pi_0(\hat{\theta})+\frac{\partial^2\pi_0(\bar{\theta})}{\partial\theta^2}(\hat{\theta}-\theta_0)\right\| \\ &\leq \sup_\theta\|\hat{\pi}(\theta)-\pi_0(\theta)\|+\sup_\theta\left\|\frac{\partial^2\pi_0(\theta)}{\partial\theta^2}\right\|\|\hat{\theta}-\theta_0\| \xrightarrow{p} 0. \qquad \text{(A.26)}\end{aligned}$$

The reintroduction of the type probabilities depending on w only involves perform the above analysis with all theoretical and empirical expectations replace with their corresponding conditional expectations, conditioned on the event $\{w_i = w_k\}, k = 1,\cdots,D_w$ and noting that $n/n_k \to c_k < \infty$ for $k = 1,\cdots,D_w$. With these modifications, conclude that

$$\max_k\left\|\frac{\partial^j}{\partial\theta^j}\hat{\pi}(w_k;\hat{\theta})-\frac{\partial^j}{\partial\theta^j}\pi_0(w_k;\theta_0)\right\| \xrightarrow{p} 0, \quad j = 0,1.$$

□

A.3 Proof of Theorem 7.5

Proof. By recalling that $\boldsymbol{f}_i(\theta,\pi)$ is a diagonal matrix with probabilities that sum to one,

$$E[\|\pi_0(w_k;\theta)'\boldsymbol{f}_i(\theta,\pi_0(w_k;\theta))'\boldsymbol{f}_i(\theta,\pi_0(w_k;\theta))\pi_0(w_k;\theta)\||w_k] < \infty$$

fpr $k = 1,\cdots,D_w$. This result implies that for any $\theta\in\Theta$, $m_i^\pi(\theta,\pi_0(w_k;\theta))$ defined in equation (A.13) satisfies $E[\left|m_i^\pi(\theta,\pi_0(w_k;\theta))'m_i^\pi(\theta,\pi_0(w_k;\theta))\right||w_k] < \infty$. Therefore, under Assumptions 7.1 and 7.2, and by noting that $n_k/n \to c_k > 0, k = 1,\cdots,D_w$ and equation (A.17), application of the Lindeberg-Levy CLT to equation (A.10) gives $\|\hat{m}^\pi(w_k;\theta,\pi_0(w_k;\theta))\| =$

$O_p(1/\sqrt{n}),\ k=1,\cdots,D_w$. This, equations (A.10) and (A.15) imply that

$$\hat{\pi}(w_k;\theta)-\pi_0(w_k;\theta)=-M_0^{\pi}(w_k;\theta,\pi_0(w_k;\theta))^{-1}\frac{1}{n_k}\sum_{i=1}^{n}m_i^{\pi}(w_k;\theta,\pi_0(w_k;\theta))I_i(w_k)+o_p(1/\sqrt{n})$$

for any $\theta\in\Theta$ and for $k=1,\cdots,D_w$. Also, under Assumptions 7.1 and 7.2, $E[\|m_i'm_i\|]<\infty$ so that, by the Lindegerg-Levi CLT,

$$\|\hat{m}\|=O_p(1/\sqrt{n}), \tag{A.27}$$

Define $\hat{M}_{\pi}(w_k;\theta)=\sum_{i=1}^{n}M_{\pi i}(\theta)I_i(w_k)/n_k$, for $k=1,\cdots,D_w$ and let $\hat{M}(\theta,\pi)=\sum_{i=1}^{n}M_i(\theta,\pi)/n$. The mean value expansion obtains

$$\begin{aligned}
&\hat{m}(\hat{\theta},\hat{\pi}(\hat{\theta}))=\\
&\hat{m}+\left[\hat{M}(\bar{\theta},\hat{\pi}(\bar{\theta}))\right](\hat{\theta}-\theta_0)-\frac{1}{n}\sum_{i=1}^{n}X_ih(x_i,\theta_0)(\hat{\pi}(w_i;\theta_0)-\pi_0(w_i;\theta_0))\\
&=\hat{m}+\left[\hat{M}(\bar{\theta},\hat{\pi}(\bar{\theta}))\right](\hat{\theta}-\theta_0)-\frac{1}{n}\sum_{i=1}^{n}X_ih(x_i,\theta_0)\sum_{k}I_i(w_k)(\hat{\pi}(w_k;\theta_0)-\pi_0(w_k;\theta_0))\\
&=\hat{m}+\left[\hat{M}(\bar{\theta},\hat{\pi}(\bar{\theta}))\right](\hat{\theta}-\theta_0)-\sum_{k}\left[\frac{1}{n_k}\sum_{i=1}^{n}X_ih(x_i,\theta_0)I_i(w_k)\right]\frac{n_k}{n}(\hat{\pi}(w_k;\theta_0)-\pi_0(w_k;\theta_0))\\
&=\hat{m}+\left[\hat{M}(\bar{\theta},\hat{\pi}(\bar{\theta}))\right](\hat{\theta}-\theta_0)-\sum_{k}\hat{M}_{\pi}(w_k;\theta_0)\frac{n_k}{n}(\hat{\pi}(w_k;\theta_0)-\pi_0(w_k;\theta_0))\\
&=\hat{m}+\left[\hat{M}(\bar{\theta},\hat{\pi}(\bar{\theta}))\right](\hat{\theta}-\theta_0)+\sum_{k}M_{\pi 0}(w_k)M_0^{\pi}(w_k)^{-1}\frac{1}{n}\sum_{i=1}^{n}m_i^{\pi}(w_k;\theta_0,\pi_0(\theta_0))I_i(w_k)+o_p(1/\sqrt{n})\\
&=\hat{m}+\left[\hat{M}(\bar{\theta},\hat{\pi}(\bar{\theta}))\right](\hat{\theta}-\theta_0)+\frac{1}{n}\sum_{i=1}^{n}\sum_{k}M_{\pi 0}(w_k)M_0^{\pi}(w_k)^{-1}m_i^{\pi}(w_k)I_i(w_k)+o_p(1/\sqrt{n})\\
&=\hat{m}+\left[\hat{M}(\bar{\theta},\hat{\pi}(\bar{\theta}))\right](\hat{\theta}-\theta_0)+\frac{1}{n}\sum_{i=1}^{n}M_{\pi 0}(w_i)M_0^{\pi}(w_i)^{-1}m_i^{\pi}+o_p(1/\sqrt{n})\\
&=\left[\hat{M}(\bar{\theta},\hat{\pi}(\bar{\theta}))\right](\hat{\theta}-\theta_0)+\frac{1}{n}\sum_{i=1}^{n}\left[m_i+M_{\pi 0}(w_i)M_0^{\pi}(w_i)^{-1}m_i^{\pi}\right]+o_p(1/\sqrt{n})
\end{aligned}\tag{A.28}$$

where $\bar{\theta}$ and $\bar{\pi}$ are mean values. Next, under the same conditions that obtains equation (A.5),

$$\sup_{\theta}\left\|\frac{\partial^j}{\partial\theta^j}\hat{M}_{\pi}(w_k,\theta)-\frac{\partial^j}{\partial\theta^j}M_{\pi 0}(w_k,\theta)\right\|\xrightarrow{p}0,\quad j=0,1,2. \tag{A.29}$$

$$\hat{M}(\theta,\pi(\theta)) = \sum_k \hat{M}_\pi(w_k;\theta)\frac{\partial}{\partial\theta}\pi(w_k;\theta) + \sum_k \pi(w_k;\theta)' \otimes I_{N_X}\frac{\partial}{\partial\theta} vec\left(\hat{M}_\pi(w_k;\theta)\right),$$

so that

$$\begin{aligned}
&\left\|\hat{M}(\hat{\theta},\hat{\pi}(\hat{\theta})) - M_0(\theta_0,\pi_0(\theta_0))\right\| \\
&\leq \sum_k \left\| \hat{M}_\pi(w_k;\hat{\theta})\frac{\partial}{\partial\theta}\hat{\pi}(w_k;\hat{\theta}) - M_{\pi 0}(w_k;\theta_0)\frac{\partial}{\partial}\pi_0(\theta_0) \right\| \\
&+ \sum_k \left\| \hat{\pi}(w_k;\hat{\theta})' \otimes I_{N_X}\frac{\partial}{\partial\theta} vec\left(\hat{M}_\pi(w_k;\hat{\theta})\right) - \pi_0(w_k;\theta_0)' \otimes I_{N_X}\frac{\partial}{\partial\theta} vec\left(M_{\pi 0}(w_k;\theta_0)\right) \right\| \\
&= o_p(1). \qquad \text{(A.30)}
\end{aligned}$$

where the last equality is obtained using Assumptions 7.1 and 7.2, the consistency results of Theorem 7.4, equation (A.29), and calculations similar to equation (A.21). This result also holds when $\hat{\theta}$ is replaced with the mean values $\bar{\theta}$.

The first-order condition $\hat{M}(\hat{\theta},\hat{\pi}(\hat{\theta}))'\hat{\Omega}\hat{m}(\hat{\theta},\hat{\pi}(\hat{\theta})) = 0$, equations (A.28) and (A.30), and Assumption 7.3 obtain

$$\sqrt{n}(\hat{\theta}-\theta_0) = -(M_0'\Omega M_0)^{-1}M_0'\Omega\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[m_i + M_{\pi 0}(w_i)M_0^{\pi}(w_i)^{-1}m_i^{\pi}\right] + o_p(1).$$

Application of the Lindeberg-Levi CLT completes the proof.

□

B Tables

Table 1: Conditional Probability of Grade Promotion, Conditioned on Hours Worked

Variable	Estimate	Std. Err.
Constant	6.5172	0.1846
Hours worked	-3.1819	0.4151
Employment	0.1941	0.0617
Education	-0.4184	0.0201
Labor market experience	0.0784	0.0112
Age	-0.2473	0.0109
Black	1.4697	0.0790
AFQT	0.7230	0.0233

Table 2: Unconditional Probability of Grade Promotion

Variable	Estimate	Std. Err.
Constant	9.6010	0.5520
Employment	-0.0975	0.0491
Education	-0.4090	0.0222
Labor market experience	0.1719	0.0209
Squared Labor market experience	-0.0098	0.0015
Age	-0.5424	0.0537
Squared age	0.0062	0.0011
Black	1.4863	0.0791
AFQT	0.0730	0.0023

Table 3: Period-specific Utility

Variable	Estimate	Std. Err.
Utility of leisure $\exp(\theta_1^u x_t^l)(1-h_t)$		
Age/10	0.2791	0.0598
$(\text{Age}/10)^2$	-0.0427	0.0091
Black	0.2917	0.1690
Psychic value of school attendance $\theta_3^u d_t^E x_t^E$		
Constant	0.0021	0.0017
Lagged Enrollment	0.0235	0.0018
Black	0.0005	0.0017
Age/10	-0.0626	0.0041
Psychic value of labor supply $\theta_4^u d_t^h x_t^h$		
Constant	1.8315	0.2192
Lagged labor supply	7.1726	0.4384
Black	-0.9806	0.3769
Age/10	-0.0626	0.0041
Employment and Enrollment	-0.0622	0.0012

References

V. Aguirregabiria and P. Mira. Swapping the nested fixed point algorithm: A class of estimators for distrete markov decision models. *Econometrica*, 70(4):1519–1543, Jul 2002.

V. Aguirregabiria and P. Mira. Sequential estimation of dynamic discrete games. *Econometrica*, 75(1):1–53, Jan 2007.

Sumru Altug and Robert A. Miller. Effect of work experience of female wages and labour supply. *The Review of Economic Studies*, 65(1):45–85, 1998.

Peter Arcidiacono and Robert A. Miller. Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity. *Econometrica*, 79(6):1823–1867, Nov 2011.

P. Bajari, C. Benkard, and J. Levin. Estimating dynamic models of imperfect competition. *Econometrica*, 75(5), Sep 2007.

David Card. *The Causal Effect of Education on Earnings*. Elsevier Science Publishers, 1999.

J.E. Dennis Jr. and Robert B. Shanbel. *Numerical Methods for Unconstrained Optimization and Nonliear Equations,in: Classics in Applied Mathematics,* volume 16. Society of Industrial and Applied Mathematics (SIAM), Philadelphia, PA., 1996.

Table 4: Wage Offer Equation

Variable	Estimate	Std. Err.
Constant	0.6525	0.0513
t/T	-1.5804	0.2842
$(t/T)^2$	0.1727	0.2555
$(t/T)^3$	-1.4939	0.0821
Experience	0.0773	0.0103
Squared experience	-0.0012	0.0033
AFQT	0.0207	0.0006
Black	-0.0233	0.0372
Wage shock St.Dev	0.0847	0.0184
s_1	0.0708	0.0136
s_2	0.0781	0.0068
s_3	0.0814	0.0093

Table 5: Distribution of Returns to Education

	0.0708	0.0781	0.0814	Mean	Std. Dev.
White					
	0.1409	0.1096	0.7495	0.0795	0.0037
Black					
	0.3872	0.1967	0.4161	0.0766	0.0048

Jeffrey A. Dubin and Daniel L. McFadden. An econometric analysis of residential electric applicance holdings and consumption. *Econometrica*, 52(2):345–362, March 1984.

Jin-Yan Fan and Ya-xing Yuan. On the quadratic convergence of the levenberg-marquardt method without nonsingularity assumption. *Computing*, 74:23–39, 2005.

O.P. Ferreira, M.L.N. Goncalves, and P.R. Oliviera. Local convergence analysis of the gauss-newton method under a matorant condition. *Journal of Complexity*, 27:111–125, 2011.

George-Levi Gayle and Robert A. Miller. Life-cycle fertility and human capital accumulation. *Unpublished Manuscript, Carnegie Mellon University*, 2003.

Wayne-Roy Gayle. A dynamic structural model of labor market supply and educational attainment. *MIMEO: University of Virginia*, 2006.

Wayne-Roy Gayle. Identification and estimation of semiparametric, correlated finite mixture models. *MIMEO: University of Virginia*, 2013.

Wayne-Roy Gayle and Soiliou D. Namoro. Estimation of a nonlinear panel data model with semiparametric individual effects. *Journal of Econometrics*, 175:46–59, July 2013.

Michael W. Hanemann. Discrete/continuous models of consumer demand. *Econometrica*, 52(3):541–561, May 1984.

James J. Heckman, Lance J. Lochner, and Petra E. Todd. *Earnings Functions, Rates of Return and Treatment Effects: The Mincer Equation and Beyond*, volume 1. Elsevier Science B.V., 2006.

James J. Heckman, Lance J. Lochner, and Petra E. Todd. Earnings functions and rates of return. *Journal of Human Capital*, 2(1):1–31, 2008.

Joseph V. Hotz and Robert A. Miller. Conditional choice probabilities and the estimation of dynamic models. *Review of Economic Studies*, 60:497–529, 1993.

Joseph V. Hotz, Robert A. Miller, Seth Sanders, and Jeffrey Smith. A simulation estimator for dynamic models of discrete choice. *Review of Economic Studies*, 61(2):265–289, 1994.

Erwin Kreyszig. *Introductory Functional Analysis with Applications*. John Wiley and Sons, 1989.

Derek A. Neal and William R. Johnson. The role of premarket factors i black-white wage differences. *Journal of Political Economics*, 105(5):869–895, 1996.

Whitney K. Newey and Daniel McFadden. *Large Sample Estimation and Hypothesis Testing*. Elsevier Science Publishers, 1994.

W.H. Richardson. Bayesian-based iterative method of image restoration. *Journal of the Optical Society of America*, 62(1):55–59, Jan 1972.

A. D. Roy. Some thoughts on the distribution of earnings. *Oxford Economic Papers*, 3(2): 135–146, Jun 1951.

John Rust. Optimal replacement of gmc bus engines: An empirical model of harold zurcher. *Econometrica*, 55(5):999–1033, 1987.

Weijun Zhou and Xiaojun Chen. Global convergence of a new hybrid gauss-newton structured bfgs method for nonlinear least squares problems. *SIAM Journal of Optimization*, 20(5):2422–2441, 2010.