Nonparametric Identification and Estimation of Finite Mixture Models of Dynamic Discrete Choices

by

Hiroyuki Kasahara and Katsumi Shimotsu

Research Report # 2006-5 **October 2006**

Department of Economics
Research Report Series

Department of Economics
Social Science Centre
The University of Western Ontario
London, Ontario, N6A 5C2
Canada

This research report is available as a downloadable pdf file on our website
http://economics.uwo.ca/econref/WorkingPapers/departmentresearchreports.html.

Nonparametric Identification and Estimation of Finite Mixture Models of Dynamic Discrete Choices

Hiroyuki Kasahara
Department of Economics
University of Western Ontario
hkasahar@uwo.ca

Katsumi Shimotsu
Department of Economics
Queen's University
shimotsu@econ.queensu.ca

October 15, 2006

Abstract

In dynamic discrete choice analysis, controlling for unobserved heterogeneity is an important issue, and finite mixture models provide flexible ways to account for unobserved heterogeneity. This paper studies nonparametric identifiability of type probabilities and type-specific component distributions in finite mixture models of dynamic discrete choices. We derive sufficient conditions for nonparametric identification for various finite mixture models of dynamic discrete choices used in applied work. Three elements emerge as the important determinants of identification; the time-dimension of panel data, the number of values the covariates can take, and the heterogeneity of the response of different types to changes in the covariates. For example, in a simple case, a time-dimension of $T = 3$ is sufficient for identification, provided that the number of values the covariates can take is no smaller than the number of types, and that the changes in the covariates induce sufficiently heterogeneous variations in the choice probabilities across types. Type-specific components are identifiable even when state dependence is present as long as the panel has a moderate time-dimension ($T \geq 6$). We also develop a series logit estimator for finite mixture models of dynamic discrete choices and derive its convergence rate.

Keywords: Dynamic discrete choice models, finite mixture, nonparametric identification, panel data, sieve estimator, unobserved heterogeneity.

JEL Classification Numbers: C13, C14, C23, C25.

1 Introduction

In dynamic discrete choice analysis, controlling for unobserved heterogeneity is an important issue. Finite mixture models provide flexible ways to account for unobserved heterogeneity, and they are commonly used in empirical analysis. To date, however, the conditions under which finite mixture dynamic discrete choice models are nonparametrically identified are not well understood. This paper studies nonparametric identifiability of finite mixture models of dynamic discrete choices when a researcher has an access to panel data. We also develop a series logit estimator for finite mixture models of dynamic discrete choices.

Finite mixtures have been used in numerous applications, especially in estimating dynamic models. In empirical industrial organization, Crawford and Shum (2005) use finite mixtures to control for patient-level unobserved heterogeneity in estimating a dynamic matching model of pharmaceutical demand. Gowrisankaran, Mitchell, and Moro (2005) estimate a dynamic model of voter behavior with finite mixtures. In labor economics, finite mixtures are a popular choice for controlling for unobserved person-specific effects when dynamic discrete choice models are estimated (cf., Keane and Wolpin (1997), Cameron and Heckman (1998)). Heckman and Singer (1984) use finite mixtures to approximate more general mixture models in the context of duration models with unobserved heterogeneity.

In most applications of finite mixture models, the components of the mixture distribution are assumed to belong to a parametric family. The nonparametric maximum likelihood estimator (NPMLE) of Heckman and Singer (1984) treats the distribution of unobservables nonparametrically but assumes parametric component distributions. Most existing theoretical work on identification of finite mixture models either treats component distributions parametrically or uses training data that are from known component distributions (cf., Titterington, Smith, and Makov (1985), Rao (1992)); as Hall and Zhou (2003) state, "very little is known of the potential for consistent nonparametric inference in mixtures without training data."

This paper studies nonparametric identifiability of type probabilities and type-specific component distributions in finite mixture dynamic discrete choice models. Specifically, we assess the identifiability of type probabilities and type-specific component distributions when no parametric assumption is imposed on them. Our point of departure is Hall and Zhou (2003), who prove nonparametric identifiability of two-type mixture models with independent marginals:

$$F(y) = \pi \prod_{t=1}^{T} F_t^1(y_t) + (1-\pi) \prod_{t=1}^{T} F_t^2(y_t), \tag{1}$$

where $F(y)$ is the distribution function of a T-dimensional variable Y and $F_t^j(y_t)$ is the distribution function of the t-th element of Y conditional on type j. Hall and Zhou show that type probability π and type-specific components F_t^j's are nonparametrically identifiable from $F(y)$ and its marginals when $T \geq 3$, while they are not when $T = 2$. The intuition behind their result

is as follows. Integrating out different elements of y from (1) gives lower-dimensional submodels:

$$F(y_{i_1}, y_{i_2}, \ldots, y_{i_l}) = \pi \prod_{s=1}^{l} F_{i_s}^1(y_{i_s}) + (1-\pi) \prod_{s=1}^{l} F_{i_s}^2(y_{i_s}), \tag{2}$$

where $1 \leq l \leq T$, $1 \leq i_1 < \ldots < i_l \leq T$, and $F(y_{i_1}, y_{i_2}, \ldots, y_{i_l})$ is the l-variate marginal distribution of $F(y)$. Each lower-dimensional submodel implies a different restriction on the unknown elements, i.e., π and F_t^j's. F and its marginals imply $2^T - 1$ restrictions while there are $2T + 1$ unknown elements. When $T = 3$, the number of restrictions is the same as the number of unknowns, and one can solve these restrictions to uniquely determine π and the F_t^j's.

While their analysis provides the insight that lower-dimensional submodels (2) provide important restrictions for identification, it has limited applicability to the finite mixture models of dynamic discrete choices in economic applications. First, it is difficult to generalize their analysis to three or more types.[1] Second, their model (1) does not have any covariates while most empirical models in economics involve covariates. Third, the assumption that elements of y are independent in (1) is not realistic in dynamic discrete choice models.

This paper provides sufficient conditions for nonparametric identification for various finite mixture models of dynamic discrete choices used in applied work. Three elements emerge as the important determinants of identification: the time-dimension of panel data, the number of the values the covariates can take, and the heterogeneity of the response of different types to changes in the covariates. For example, in a simple case, a time-dimension of $T = 3$ is sufficient for identification, provided that the number of values the covariates can take is no smaller than the number of types and that the changes in the covariates induce sufficiently heterogeneous variations in the choice probabilities across types.

The key insight is that, in models with covariates, different *sequences* of covariates imply different identifying restrictions in the lower-dimensional submodels; in fact, if d is the number of support points of the covariates and T is the time-dimension, then the number of restrictions becomes in the order of d^T. As a result, the presence of covariates provides a powerful source of identification in panel data even with a moderate time-dimension T.

We study a variety of finite mixture dynamic discrete choice models. We analyze the case where conditional choice probabilities change over time because time-specific aggregate shocks are present, or agents are finitely-lived. We consider a possibility that the transition function of state variables is different across types. We also examine the case where state dependence is present (for instance, when the lagged choice affects the current choice), and show that type-specific components are identifiable as long as the panel has a moderate time-dimension

[1] When the number of types, M, is more than three, Hall et al. (2005) show that for any number of types, M, there exists T_M such that type probabilities and type-specific component distributions are nonparametrically identifiable when $T \geq T_M$ and that T_M is no larger than $(1 + o(1))6M \ln(M)$ as M increases. But, as they state, such an upper bound is "undoubtedly larger than the minimal value" of T_M.

of $T \geq 6$. This result is important since distinguishing unobserved heterogeneity and state dependence often motivates the use of finite mixture models in empirical analysis.

This paper also develops a series logit estimator for finite mixture models of dynamic discrete choices and derives its convergence rate. Hirano et al. (2003) derives the convergence rate of the (non-mixture) series logit estimator. Therefore, our work may be viewed as a generalization of Hirano et al. (2003) to a finite mixture setting where the objective function is not globally concave. This case is not covered by Chen (2006), who provides a comprehensive survey of series (sieve) estimation methods of semi-nonparametric econometric models. In our Monte Carlo experiment, we find that the performance of our series estimator is almost comparable to that of the parametric maximum likelihood estimator in terms of the accuracy of the estimated conditional choice probabilities.[2]

Nonparametric identification and estimation of finite mixture dynamic discrete choice models are relevant and useful in practical applications for, at least, the following reasons. First, choosing a parametric family for the component distributions is often difficult because of a lack of guidance from economic theory; nonparametric estimation provides a flexible way to reveal the structure hidden in the data. Furthermore, even when theory offers guidance, comparing parametric and nonparametric estimates allows us to examine the validity of the restrictions imposed by economic theory.

Second, analyzing nonparametric identification helps us understand the identification of parametric or semiparametric finite mixture models of dynamic discrete choices. Understanding identification is not a simple task for finite mixture models even with *parametric* component distributions, and formal identification analysis is rarely provided in empirical applications. Once type probabilities and component distributions are nonparametrically identified, the identification analysis of parametric finite mixture models often becomes transparent as it is reduced to the analysis of models without unobserved heterogeneity. As we demonstrate through examples, our nonparametric identification results can be applied to check the identifiability of parametric finite mixture models that are popular in empirical analysis.

Third, the identification results and series estimator of this paper will open the door to applying semiparametric estimators for structural dynamic models to models with unobserved heterogeneity. Recently, by building on the seminal work by Hotz and Miller (1993), computationally attractive semiparametric estimators for structural dynamic models have been developed (Aguirregabiria and Mira (2002), Kasahara and Shimotsu (2006)), and a number of papers in empirical industrial organization have proposed two/multi-step estimators for dynamic games (cf., Bajari, Benkard, and Levin (2005), Pakes, Ostrovsky, and Berry (2005), Pesendorfer and Schmidt-Dengler (2006), Bajari and Hong (2006), and Aguirregabiria and Mira (2006)). To

[2]Houde and Imai (2006) independently study nonparametric identification and estimation of dynamic discrete choice model with unobserved heterogeneity and also find that series logit estimators perform well in their simulations using various models that differ from ours.

date, however, few of these semiparametric estimators have been extended to accommodate unobserved heterogeneity. This is because these estimators often require an initial nonparametric consistent estimate of type-specific component distributions, but it has not been known whether one can obtain a consistent nonparametric estimate in finite mixture models.[3] The identification results and series estimator of this paper provides an apparatus that enables researchers to apply these semiparametric estimators to the models with unobserved heterogeneity. This is important since it is often crucial to control for unobserved heterogeneity in dynamic models (see Aguirregabiria and Mira (2006)).

In a closely related paper, Kitamura (2004) examines nonparametric identifiability of finite mixture models with covariates. Our paper shares his insight that the variation in covariates may provide a source of identification, but the setting as well as the issues we consider is different from Kitamura's. We study discrete choice models in a dynamic setting with panel data, while Kitamura considers regression models with continuous dependent variables with cross-sectional data. We address various issues specific to dynamic discrete choice models including identification in the presence of state dependence and type-dependent transition probabilities for endogenous explanatory variables.

Our work provides yet another angle for analysis that relates current and previous work on dynamic discrete choice models. Honoré and Tamer (2006) study identification of dynamic discrete choice models, including the initial conditions problem, and suggest methods to calculate the identified sets.[4] Rust (1994), Magnac and Thesmar (2002), and Aguirregabiria (2006) study the identification of structural dynamic discrete choice models.[5] Our analysis is also related to an extensive literature on identification of duration models (cf., Elbers and Ridder (1982), Heckman and Singer (1984), Ridder (1990), and Van den Berg (2001)).

The rest of the paper is organized as follows. Section 2 discusses our approach to identification and provide the identification results using a simple "baseline" model. Section 3 extends the identification analysis of Section 2, and studies a variety of finite mixture dynamic discrete choice models. In Section 4, we develop a series logit estimator for finite mixture models. Section 5 reports Monte Carlo simulation results. The proofs are collected in the Appendix.

[3]It is believed that it is not possible to obtain an consistent estimate of choice probabilities. For instance, Aguirregabiria and Mira (2006) propose a pseudo maximum likelihood estimation algorithm for models with unobserved heterogeneity but state that (p.15) "for [models with unobservable market characteristics] it is not possible to obtain consistent nonparametric estimates of [choice probabilities]". Furthermore, Geweke and Keane (2001, p.3490) write that "the [Hotz and Miller's] methods cannot accommodate unobserved state variables."

[4]Honoré and Tamer consider general mixing distributions, but treat the conditional distribution of dependent variable parametrically, and assume strict exogeneity of explanatory variables.

[5]Structural dynamic discrete choice models are not identified generically and, as Magnac and Thesmar (2002) state, "the degree of underidentification is even larger with unobserved heterogeneity." Our identification results imply that the degree of underidentification in models with unobserved heterogeneity can be reduced to that of models *without* unobserved heterogeneity.

2 Nonparametric identification of finite mixture models of dynamic discrete choices

Every period, each individual makes a choice a_t from the discrete and finite set A, given the value of the state variables $(x_t, a_{t-1}) \in X \times A$, where the lagged choice is included as one of the state variables. Each individual belongs to one of M types, and his/her type attribute is unknown. The conditional choice probability and the initial distribution are different across types, and type m's conditional choice probability is denoted by $P^m(a_t|x_t, a_{t-1})$, while its initial distribution (strictly speaking, density or probability mass function) of (x_1, a_1) is denoted by $p^{*m}(x_1, a_1)$.[6] Type m's transition probability function of the state variable x_t is denoted by $f^m(x_t|x_{t-1}, a_{t-1})$. With a slight abuse of notation, we let $p^{*m}(x_1, a_1)$ and $f^m(x_t|x_{t-1}, a_{t-1})$ denote the density of the continuously distributed elements of x_t and the probability mass function of the discretely distributed elements of x_t, respectively.

Suppose we have a panel data with time-dimension equal to T. Each individual observation, $w_{it} = \{a_{it}, x_{it}\}_{t=1}^T$, is drawn randomly from a M-term mixture distribution:

$$P(\{a_t, x_t\}_{t=1}^T) = \sum_{m=1}^M \pi^m p^{*m}(x_1, a_1) \prod_{t=2}^T f^m(x_t|x_{t-1}, a_{t-1}) P^m(a_t|x_t, a_{t-1}), \tag{3}$$

where π^m are positive and sum to one. The left-hand-side of (3) is the distribution function of the observable data while the right-hand-side contains the objects we would like to learn from the observable data.

Remark 1 *In practice, it is sometimes assumed that the distribution of the initial observation, $p^{*m}(x_1, a_1)$, is the stationary distribution satisfying the fixed point constraint*

$$p^{*m}(x_1, a_1) = \sum_{x' \in X} \sum_{a' \in A} P^m(a_1|x_1, a') f^m(x_1|x', a') p^{*m}(x', a'), \tag{4}$$

when all the components of x have finite support. When x is continuously distributed, we replace the summation over x' with integration. Our identification result, however, does not rely on the stationarity assumption of the initial conditions.

The model (3) includes the following examples as special cases.

Example 1 (Dynamic discrete choice model with heterogeneous coefficients) *Let $\theta_i' = (\beta_i', \gamma_i)$ be an individual-specific vector of unobserved variables which are multinomially distributed. Consider a dynamic binary choice model*

$$P^m(a_{it} = 1|x_{it}, a_{i,t-1}) = 1 - \Phi(x_{it}'\beta_i + a_{i,t-1}\gamma_i), \tag{5}$$

[6] Alternatively, we may consider the initial distribution of (x_1, a_0) as primitive and denote it by $p_0^{*m}(x_1, a_0)$. Then, we may define $p^{*m}(x_1, a_1) = \sum_{a' \in A} p_0^{*m}(x_1, a') P^m(a_1|x_1, a')$.

where $\Phi(\cdot)$ denotes standard normal cdf. The distribution of x_{it} conditional on $(x_{i,t-1}, a_{i,t-1})$ is specific to the value of θ_i. Since the evolution of (x_{it}, a_{it})'s in the presample period is not independent of θ_i, the initial distribution of (x_{i1}, a_{i1}) depends on the value of θ_i (cf., Heckman (1981)). Browning and Carro (2006) estimate a version of (5) for the purchase of milk using a Danish consumer "long" panel and provide evidence for heterogeneity in coefficients. Their study illustrates that allowing for such heterogeneity can make a significant difference for outcomes of interest such as the marginal dynamic effect.

Allowing for heterogeneity in coefficients is also important in analyzing the effect of policies or treatments (cf., Heckman, Urzua, and Vytlacil (2006)). For instance, if x_{it} is a treatment variable and a_{it} is a discrete outcome, then the impact of treatment on an outcome is different across individuals. Furthermore, the process of x_{it} may also depend on β_i if the agents who choose x_{it} make their choices based on the gain from the treatment. The model (3) captures such dependence by allowing for the transition function of x_{it}—interpreted as treatment rules—to depend on the value of β_i.

Example 2 (Structural dynamic discrete choice models) *The type m's agent maximizes the expected discounted sum of utilities, $E[\sum_{j=0}^{\infty} \beta^j \{u(x_{t+j}, a_{t+j}; \theta^m) + \epsilon_{t+j}(a_{t+j})\} | a_t, x_t; \theta^m]$, where x_t is observable state variable and $\epsilon_t(a_t)$ is state variable that are known to the agent but not to the researcher. The Bellman equation for this dynamic optimization problem is*

$$V(x) = \int \max_{a \in A} \left\{ u(x, a; \theta^m) + \epsilon(a) + \beta \sum_{x' \in X} V(x') f(x'|x, a; \theta^m) \right\} g(d\epsilon|x), \tag{6}$$

where $g(\epsilon|x)$ is the joint distribution of $\epsilon = \{\epsilon(j) : j \in A\}$ and $f(x'|x, a; \theta^m)$ is type-specific transition function. The conditional choice probability is

$$P_{\theta^m}(a|x) = \int 1 \left\{ a = \arg\max_{j \in A} \left[u(x, j; \theta^m) + \epsilon(j) + \beta \sum_{x' \in X} V_{\theta^m}(x') f(x'|x, j; \theta^m) \right] \right\} g(d\epsilon|x), \tag{7}$$

where V_{θ^m} is the fixed point of (6). Let $P^m(a_t|x_t, a_{t-1}) = P_{\theta^m}(a_t|x_t)$ and $f^m(x_t|x_{t-1}, a_{t-1}) = f(x_t|x_{t-1}, a_{t-1}; \theta^m)$ in (3) and (4). The initial distribution of (x_1, a_1) is given by the stationary distribution (4). Then, the likelihood function for $\{a_t, x_t\}_{t=1}^T$ is given by (3) with (4).

We study the nonparametric identifiability of the type probabilities, the initial distribution, type-specific conditional choice probabilities, and type-specific transition function in equation (3), which we denote by $\theta = \{\pi^m, p^{*m}(x, a), P^m(a'|x, a), f^m(x'|x, a) : (a, a', x, x') \in A \times A \times X \times X\}_{m=1}^M$. Following the standard definition of nonparametric identifiability, θ is said to be nonparametrically identified (or identifiable) if it is uniquely determined by the distribution $P(\{a_t, x_t\}_{t=1}^T)$, without making any parametric assumption about $p^{*m}(x, a)$, $P^m(a'|x, a)$, and $f^m(x'|x, a)$. Because the order of the component distributions can be changed, θ is identified

only up to a permutation of the components. If no two of the π's are identical, we may uniquely determine the components by assuming $\pi^1 < \pi^2 < \cdots < \pi^M$.

2.1 Our approach and identification of the baseline model

The finite mixture models studied by Hall and Zhou (2003) have no covariates as discussed in the introduction. In this subsection, we show that the presence of covariates in our model creates a powerful source of identification.

First, we impose the following simplifying assumptions to the general model (3) and analyze the nonparametric identifiability of the resulting "baseline model." Analyzing the baseline model helps make clear the basic idea of our approach and clarifies the logic behind our main results. In the subsequent sections, we relax Assumption 1 in various ways, and study how it affects the identifiability of the resulting models.

Assumption 1 *(a) The choice probability of a_t is independent of the lagged choice a_{t-1} conditional on x_t so that $P^m(a_t|x_t) = P^m(a_t|x_t, a_{t-1})$, where a_{t-1} is not one of the elements of x_t. (b) $f^m(x_{t+1}|x_t, a_t) > 0$ for all $(x_{t+1}, x_t, a_t) \in X \times X \times A$ and for all m. (c) The transition function is common across types; $f^m(x_{t+1}|x_t, a_t) = f(x_{t+1}|x_t, a_t)$ for all m.*

Under Assumption 1(a), the lagged choice a_{t-1} affects the current choice a_t only through its effect on x_t via $f^m(x_t|x_{t-1}, a_{t-1})$. Assumption 1(b) implies that, starting from any pair of the state and action $(x, a) \in X \times A$, any state $x' \in X$ is reached in the next period with positive probability. With Assumption 1 imposed, the baseline model is

$$P(\{a_t, x_t\}_{t=1}^T) = \sum_{m=1}^M \pi^m p^{*m}(x_1, a_1) \prod_{t=2}^T f(x_t|x_{t-1}, a_{t-1}) P^m(a_t|x_t). \tag{8}$$

Since $f(x'|x, a)$ is nonparametrically identified directly from the data on (x', x, a)'s (cf., Rust (1987)), we may assume $f(x'|x, a)$ is known without affecting the other parts of the argument. Divide $P(\{a_t, x_t\}_{t=1}^T)$ by the transition functions and define

$$\tilde{P}(\{a_t, x_t\}_{t=1}^T) = \frac{P(\{a_t, x_t\}_{t=1}^T)}{\prod_{t=2}^T f(x_t|x_{t-1}, a_{t-1})} = \sum_{m=1}^M \pi^m p^{*m}(x_1, a_1) \prod_{t=2}^T P^m(a_t|x_t), \tag{9}$$

which can be computed from the observed data. Assumption 1 guarantees that $\tilde{P}(\{a_t, x_t\}_{t=1}^T)$ is well-defined for any possible sequence of $\{a_t, x_t\}_{t=1}^T \in (A \times X)^T$.

Let $\mathcal{I} = \{i_1, \ldots, i_l\}$ be a subset of the time indices, so that $\mathcal{I} \subseteq \{1, \ldots, T\}$, where $1 \leq l \leq T$ and $1 \leq i_1 < \ldots < i_l \leq T$. Integrating out different elements from (9) gives l-variate marginal

version of $\tilde{P}(\{a_t, x_t\}_{t=1}^T)$, which we call *lower-dimensional submodels*

$$\tilde{P}(\{a_{i_s}, x_{i_s}\}_{i_s \in \mathcal{I}}) = \sum_{m=1}^{M} \pi^m p^{*m}(a_1, x_1) \prod_{s=2}^{l} P^m(a_{i_s}|x_{i_s}), \quad \text{when } \{1\} \in \mathcal{I}, \tag{10}$$

and

$$\tilde{P}(\{a_{i_s}, x_{i_s}\}_{i_s \in \mathcal{I}}) = \sum_{m=1}^{M} \pi^m \prod_{s=1}^{l} P^m(a_{i_s}|x_{i_s}), \quad \text{when } \{1\} \notin \mathcal{I}. \tag{11}$$

In model (9), a powerful source of identification is provided by the difference in each type's response patterns to the variation of the covariate $(x_1, \ldots, x_T)$. The key insight is that, for each different value of $(x_1, \ldots, x_T)$, (10) and (11) imply different restrictions on the type probabilities and conditional choice probabilities. Let $|X|$ denote the number of elements in X. The variation of $(x_1, \ldots, x_T)$ generates different versions of (10) and (11), providing restrictions whose number is in the order of $|X|^T$, while the number of the parameters $\{\pi^m, p^{*m}(a, x), P^m(a|x) : (a, x) \in A \times X\}_{m=1}^M$ is in the order of $|X|$. This identification approach is much more effective than one without covariates, in particular, when T is small.[7]

To keep the notation simple, we mainly focus on the case where $A = \{0, 1\}$. It is straightforward to extend our analysis to the case with a multinomial choice of a, but with heavier notations. Note also that Chandra (1977) shows that a multivariate finite mixture model is identified if its all marginal models are identified.

Our first proposition provides a sufficient condition for identification under Assumption 1. The proposition extends the idea of the proof of nonparametric identifiability of finite mixture models by Anderson (1954) and Gibson (1955) to models with covariates.[8] The proof is constructive. Define, for $\xi \in X$,

$$\lambda_\xi^{*m} = p^{*m}((a_1, x_1) = (1, \xi)) \quad \text{and} \quad \lambda_\xi^m = P^m(a = 1|x = \xi). \tag{12}$$

Defining $\lambda_\xi^{*m} = p^{*m}((a_1, x_1) = (0, \xi))$ and $\lambda_\xi^m = P^m(a = 0|x = \xi)$ does not change our argument.

Proposition 1 *Suppose that Assumption 1 holds. Assume $T \geq 3$. Let ξ_j, $j = 1, \ldots, M-1$, be elements of X, and define*

$$\underset{(M \times M)}{L} = \begin{bmatrix} 1 & \lambda_{\xi_1}^1 & \cdots & \lambda_{\xi_{M-1}}^1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_{\xi_1}^M & \cdots & \lambda_{\xi_{M-1}}^M \end{bmatrix}.$$

[7] For example, when $T = 3$ and $A = \{0, 1\}$, (10) and (11) imply at least $\binom{|X|+2}{3}$ different restrictions while there are $3M|X| - 1$ parameters.

[8] Anderson (1954) and Gibson (1955) analyze nonparametric identification of finite mixture models similar to (9) but *without covariates* and derive a sufficient condition for nonparametric identifiability under the assumption $T \geq 2M - 1$. Madansky (1960) extends their analysis to obtain a sufficient condition under the assumption $2^{(T-1)/2} \geq M$. When T is small, the number of identifiable types by their method is quite limited.

Suppose that there exists some $\{\xi_1, \ldots, \xi_{M-1}\}$ *such that* L *is nonsingular and that there exists* $k \in X$ *such that* $\lambda_k^{*m} \neq \lambda_k^{*n}$ *for any* $m \neq n$. *Then,* $\{\pi^m, \{\lambda_\xi^{*m}, \lambda_\xi^m\}_{\xi \in X}\}_{m=1}^M$ *is uniquely determined from* $\{\tilde{P}(\{a_t, x_t\}_{t=1}^3) : \{a_t, x_t\}_{t=1}^3 \in (A \times X)^3\}$.

Remark 2

1. *The condition of Proposition 1 implies that all columns in* L *must be linearly independent. Since each column of* L *represents the conditional choice probability of different types for a given value of* x, *the changes in* x *must induce sufficiently heterogeneous variations in the conditional choice probabilities across types. In other words, the observed state variable must be relevant, and different types must respond to its changes differently.*

2. *The condition that* $\lambda_k^{*m} \neq \lambda_k^{*n}$ *for some* $k \in X$ *is satisfied if the initial distributions are different across different types. If this condition is violated, then the initial distribution cannot be used as a source of identification and, as a result, the requirement on* T *becomes* $T \geq 4$ *instead of* $T \geq 3$.

3. *One needs to find only one set of* $M-1$ *points to construct a nonsingular* L. *The identification of choice probabilities at all other points in* X *follows without any further requirement.*

4. *When* X *has* $|X| < \infty$ *support points, the number of identifiable types is at most* $|X| + 1$. *When* x *is continuously distributed, we may potentially identify as many types as we wish.*

5. *By partitioning* X *into M-1 disjoint subsets* $(\Xi_1, \Xi_2, \ldots, \Xi_{M-1})$, *we may characterize a sufficient condition in terms of the conditional choice probabilities given a subset* Ξ_j *of* X *rather than an element* ξ_j *of* X.[9]

Proposition 1 gives a simple and intuitive sufficient condition for identification in terms of the rank of the matrix L and the type-specific choice probabilities evaluated at k. In practice, however, it may be difficult to check the rank condition of L because the elements of L are functions of the component distributions. We develop a corollary that gives sufficient conditions in terms of what we can easily estimate from the observed data.

For convenience, first collect notation. Fix $a_t = 1$ for all t in $\tilde{P}(\{a_t, x_t\}_{t=1}^3)$, and define the resulting function as

$$F_{x_1,x_2,x_3}^* = \tilde{P}(\{1, x_t\}_{t=1}^3) = \sum_{m=1}^M \pi^m \lambda_{x_1}^{*m} \lambda_{x_2}^m \lambda_{x_3}^m, \tag{13}$$

[9]For instance, consider a disjoint partition $(\Xi_1, \ldots, \Xi_{M-1})$ such that $X = \cup_{j=1}^{M-1} \Xi_j$. Define the probability of events $\{a_t, \{x_t \in \Xi_{j_t}\}\}_{t=1}^T$, where $j_t = 1, \ldots, M-1$, by equation (8) but in terms of the initial probability of $(a_1, \{x_1 \in \Xi_{j_1}\})$, the transition function of $\{x_t \in \Xi_{j_t}\}$ given $(a_{t-1}, \{x_{t-1} \in \Xi_{j_{t-1}}\})$, and the conditional choice probabilities given $\{x_t \in \Xi_{j_t}\}$. Then, a similar analysis gives a sufficient condition in terms of the matrix L in which λ_ξ^{*m} and λ_ξ^m are replaced with $\lambda_\Xi^{*m} = p^{*m}(a_1 = 1, \{x_1 \in \Xi\})$ and $\lambda_\Xi^m = P^m(a = 1 | \{x \in \Xi\})$, respectively.

where λ_x^{*m} and λ_x^m are defined in (12). Next, integrate out (a_1, x_1) from $\tilde{P}(\{a_t, x_t\}_{t=1}^3)$ and fix $a_2 = a_3 = 1$, and define the resulting function as

$$F_{x_2,x_3} = \tilde{P}(\{1, x_t\}_{t=2}^3) = \sum_{m=1}^M \pi^m \lambda_{x_2}^m \lambda_{x_3}^m. \tag{14}$$

Similarly, define the following "marginals" by integrating out other elements from $\tilde{P}(\{a_t, x_t\}_{t=1}^3)$ and setting $a_t = 1$:

$$\begin{array}{ll} F^*_{x_1,x_2} = \tilde{P}(\{1, x_t\}_{t=1}^2) = \sum_{m=1}^M \pi^m \lambda_{x_1}^{*m} \lambda_{x_2}^m, & F^*_{x_1,x_3} = \tilde{P}(\{1, x_1, 1, x_3\}) = \sum_{m=1}^M \pi^m \lambda_{x_1}^{*m} \lambda_{x_3}^m, \\ F^*_{x_1} = \tilde{P}(\{1, x_1\}) = \sum_{m=1}^M \pi^m \lambda_{x_1}^{*m}, & \\ F_{x_2} = \tilde{P}(\{1, x_2\}) = \sum_{m=1}^M \pi^m \lambda_{x_2}^m, & F_{x_3} = \tilde{P}(\{1, x_3\}) = \sum_{m=1}^M \pi^m \lambda_{x_3}^m. \end{array} \tag{15}$$

Note that $F^*_{\cdot}$ involves (a_1, x_1) while $F_{\cdot}$ does not contain (a_1, x_1). In fact, $F^*_{x_1,x_2} = F^*_{x_1,x_3}$ if $x_2 = x_3$ because $P^m(a|x)$ does not depend on t, but we keep separate notations for the two because later we analyze the case where the choice probability depends on t.

Corollary 1 *Suppose that Assumption 1 holds. Assume $T \geq 3$. Let $k \in X$ and let ξ_j, $j = 1, \ldots, M-1$, be elements of X. Evaluate $F^*_{x_1,x_2,x_3}, F_{x_2,x_3}$ and their marginals at $x_1 = k$, $x_2 = \xi_1, \ldots, \xi_{M-1}$, and $x_3 = \xi_1, \ldots, \xi_{M-1}$, and arrange them into two $M \times M$ matrices*

$$P = \begin{bmatrix} 1 & F_{\xi_1} & \cdots & F_{\xi_{M-1}} \\ F_{\xi_1} & F_{\xi_1,\xi_1} & \cdots & F_{\xi_1,\xi_{M-1}} \\ \vdots & \vdots & \ddots & \vdots \\ F_{\xi_{M-1}} & F_{\xi_{M-1},\xi_1} & \cdots & F_{\xi_{M-1},\xi_{M-1}} \end{bmatrix}, \quad P_k = \begin{bmatrix} F^*_k & F^*_{k,\xi_1} & \cdots & F^*_{k,\xi_{M-1}} \\ F^*_{k,\xi_1} & F^*_{k,\xi_1,\xi_1} & \cdots & F^*_{k,\xi_1,\xi_{M-1}} \\ \vdots & \vdots & \ddots & \vdots \\ F^*_{k,\xi_{M-1}} & F^*_{k,\xi_{M-1},\xi_1} & \cdots & F^*_{k,\xi_{M-1},\xi_{M-1}} \end{bmatrix}. \tag{16}$$

Suppose that there exists some $\{\xi_1, \ldots, \xi_{M-1}\}$ such that P is of full rank and that all the eigenvalues of $P^{-1}P_k$ take distinct values. Then, $\{\pi^m, \{\lambda_\xi^{*m}, \lambda_\xi^m\}_{\xi \in X}\}_{m=1}^M$ is uniquely determined from $\{\tilde{P}(\{a_t, x_t\}_{t=1}^3) : \{a_t, x_t\}_{t=1}^3 \in (A \times X)^3\}$.

Corollary 1 gives a sufficient condition in terms of P and P_k, both of which can be constructed from the distribution function of the observed data. We may check these conditions by computing the sample counterpart of P and P_k for various $\{\xi_1, \ldots, \xi_{M-1}\}$'s. As discussed in Remark 2.5, we may also check the conditions by computing the sample counterpart of P and P_k for various partitions Ξ_j's instead of elements ξ_j. The latter procedure is especially useful when x is continuously distributed.

For the sake of brevity, in the subsequent analysis we provide sufficient conditions only in terms of the rank of the matrix of the type-specific component distributions (e.g., L). In each of the following propositions, sufficient conditions in terms of the distribution function of the

observed data can easily be deduced from the conditions in terms of the type-specific component distributions.

The identification method of Proposition 1 uses a set of restrictions implied by the joint distribution of only $(a_1, x_1, a_2, x_2, a_3, x_3)$. When the variation of $(x_1, x_2, \ldots, x_T)$ for $T \geq 4$ is available, we may adopt the approach of Madansky (1960) to use the information contained in all x_t's, and extend the maximum number of identifiable types from in the order of $|X|$ to in the order of $|X|^{(T-1)/2}$. Despite being more complex than Proposition 1, this proposition is useful when T is large, making it possible to identify a large number of types even if $|X|$ is small. For notational simplicity, we assume $|X|$ is finite and $X = \{1, 2, \ldots, |X|\}$.

Proposition 2 *Suppose that Assumption 1 holds. Assume $T \geq 3$ is odd and define $u = (T-1)/2$. Suppose $X = \{1, 2, \ldots, |X|\}$. Define*

$$\Lambda_0 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \quad \Lambda_1 = \begin{bmatrix} \lambda_1^1 & \cdots & \lambda_{|X|}^1 \\ \vdots & & \vdots \\ \lambda_1^M & \cdots & \lambda_{|X|}^M \end{bmatrix}.$$

For $l = 2, \ldots, u$, define Λ_l to be a matrix, each column of which is formed by choosing l columns (unordered, with replacement) from the columns of Λ_1 and taking their Hadamard product. There are $\binom{|X|+l-1}{l}$ ways of choosing such columns, thus the dimension of Λ_2 is $M \times \binom{|X|+l-1}{l}$. For example, Λ_2 and Λ_3 take the form

$$\Lambda_2 = \begin{bmatrix} \lambda_1^1\lambda_1^1 & \cdots & \lambda_1^1\lambda_{|X|}^1 & \lambda_2^1\lambda_2^1 & \cdots & \lambda_2^1\lambda_{|X|}^1 & \cdots & \lambda_{|X|}^1\lambda_{|X|}^1 \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots \\ \lambda_1^M\lambda_1^M & \cdots & \lambda_1^M\lambda_{|X|}^M & \lambda_2^M\lambda_2^M & \cdots & \lambda_2^M\lambda_{|X|}^M & \cdots & \lambda_{|X|}^M\lambda_{|X|}^M \end{bmatrix},$$

$$\Lambda_3 = \begin{bmatrix} \lambda_1^1\lambda_1^1\lambda_1^1 & \cdots & \lambda_1^1\lambda_1^1\lambda_{|X|}^1 & \lambda_2^1\lambda_1^1\lambda_2^1 & \cdots & \lambda_2^1\lambda_1^1\lambda_{|X|}^1 & \cdots & \lambda_{|X|}^1\lambda_{|X|}^1\lambda_{|X|}^1 \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots \\ \lambda_1^M\lambda_1^M\lambda_1^M & \cdots & \lambda_1^M\lambda_1^M\lambda_{|X|}^M & \lambda_2^M\lambda_1^M\lambda_2^M & \cdots & \lambda_2^M\lambda_1^M\lambda_{|X|}^M & \cdots & \lambda_{|X|}^M\lambda_{|X|}^M\lambda_{|X|}^M \end{bmatrix}.$$

Define an $M \times (\sum_{l=0}^{u} \binom{|X|+l-1}{l})$ matrix Λ as

$$\Lambda = [\Lambda_0, \Lambda_1, \Lambda_2, \ldots, \Lambda_u].$$

Suppose (a) $\sum_{l=0}^{u} \binom{|X|+l-1}{l} \geq M$, (b) we can construct a nonsingular $M \times M$ matrix $L^\diamond$ by setting its first column as Λ_0 and choosing other $M-1$ columns from the columns of Λ other than Λ_0, and (c) there exists $k \in X$ such that $\lambda_k^{*m} \neq \lambda_k^{*n}$ for any $m \neq n$. Then $\{\pi^m, \{\lambda_j^{*m}, \lambda_j^m\}_{j=1}^{|X|}\}_{m=1}^{M}$ is uniquely determined from $\{\tilde{P}(\{a_t, x_t\}_{t=1}^{T}) : \{a_t, x_t\}_{t=1}^{T} \in (A \times X)^T\}$.

Remark 3 *In a special case where there is no covariates and $|X| = 1$, the matrix Λ becomes*

$$\Lambda = \begin{bmatrix} 1 & \lambda_1^1 & (\lambda_1^1)^2 & \cdots & (\lambda_1^1)^u \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \lambda_1^M & (\lambda_1^M)^2 & \cdots & (\lambda_1^M)^u \end{bmatrix},$$

and the sufficient condition of Proposition 2 reduces to (a) $T \geq 2M-1$, (b) $\lambda_1^m \neq \lambda_1^n$ for any $m \neq n$, and (c) $\lambda_1^{*m} \neq \lambda_1^{*n}$ for any $m \neq n$. Not surprisingly, the condition $T \geq 2M-1$ coincides with the sufficient condition of nonparametric identification of finite mixtures of binomial distributions (Blischke (1964)). This set of sufficient condition also applies to the case where the covariates have no time variation ($x_1 = \cdots = x_T$), such as race and/or sex.

Houde and Imai (2006) study nonparametric identification of finite mixture dynamic discrete choice models by fixing the value of the covariate x (to $\bar{x}$, for instance) and derive a sufficient condition for T. They also consider a model with terminating state.

If the conditional choice probabilities of different types are heterogeneous and the column vectors $(\lambda_x^1, \ldots, \lambda_x^M)'$ for $x = 1, \ldots, |X|$ are linearly independent, the rank condition of this proposition is likely to be satisfied, since the Hadamard products of these column vectors are unlikely to be linearly dependent, unless by a chance. The condition $\sum_{l=0}^{u} \binom{|X|+l-1}{l} \geq M$ with $u = (T-1)/2$ of Proposition 2 is weaker than the condition $|X| + 1 \geq M$ of Proposition 1 when $T \geq 5$.

3 Extensions of the baseline model

In this section, we relax Assumption 1 of the baseline model in various ways to accommodate real-world applications and analyze nonparametric identifiability of resulting models.

3.1 Time-dependent conditional choice probabilities

The baseline model (8) assumes that conditional choice probabilities do not change over periods. However, the agent's decision rules may change over periods in some models, such as a model with time-specific aggregate shocks or a model of finitely-lived individuals. In this subsection, we keep the assumption of the common transition function, but extend our analysis to mixture models with time-dependent choice probabilities and transition functions.

Assumption 2 *For $t = 2, \ldots, T$, (a) $P_t^m(a_t|x_t) = P_t^m(a_t|x_t, a_{t-1})$, where a_{t-1} is not in the elements of x_t. (b) $f_t^m(x_{t+1}|x_t, a_t) = f_t(x_{t+1}|x_t, a_t)$ for all m. (c) $f_t(x_{t+1}|x_t, a_t) > 0$ for all $(x_{t+1}, x_t, a_t) \in X \times X \times A$.*

With time-dependent conditional choice probabilities, the mixture model we consider is

$$P(\{a_t, x_t\}_{t=1}^T) = \sum_{m=1}^M \pi^m p^{*m}(x_1, a_1) \prod_{t=2}^T f_t(x_t|x_{t-1}, a_{t-1}) P_t^m(a_t|x_t),$$

where both conditional choice probabilities and transition functions are indexed by time subscript t. As in the previous section, we assume that the $f_t(x_t|x_{t-1}, a_{t-1})$'s are known and rewrite the above equation as

$$\tilde{P}(\{a_t, x_t\}_{t=1}^T) = \frac{P(\{a_t, x_t\}_{t=1}^T)}{\prod_{t=2}^T f_t(x_t|x_{t-1}, a_{t-1})} = \sum_{m=1}^M \pi^m p^{*m}(a_1, x_1) \prod_{t=2}^T P_t^m(a_t|x_t). \tag{17}$$

The next proposition states a sufficient condition for nonparametric identification of the mixture model (17). In the baseline model (8), the sufficient condition is summarized to the invertibility of a matrix consisting of the conditional choice probabilities. In the time-dependent case, this matrix of conditional choice probability becomes time-dependent, and hence its invertibility needs to hold for each period. We consider the case of $A = \{0, 1\}$. Define, for $\xi \in X$,

$$\lambda_\xi^{*m} = p^{*m}((a_1, x_1) = (1, \xi)) \quad \text{and} \quad \lambda_{t,\xi}^m = P_t^m(a_t = 1|x_t = \xi), \quad t = 2, \ldots, T.$$

Proposition 3 *Suppose that Assumption 2 holds. Assume $T \geq 3$. For $t = 2, \ldots, T-1$, let $\xi_j^t, j = 1, \ldots, M-1$, be elements of X and define*

$$\underset{(M\times M)}{L_t} = \begin{bmatrix} 1 & \lambda_{t,\xi_1^t}^1 & \cdots & \lambda_{t,\xi_{M-1}^t}^1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_{t,\xi_1^t}^M & \cdots & \lambda_{t,\xi_{M-1}^t}^M \end{bmatrix}.$$

Suppose there exists $\{\xi_1^t, \ldots, \xi_{M-1}^t\}$ such that L_t is nonsingular for $t = 2, \ldots, T$ and there exists $k \in X$ such that $\lambda_k^{*m} \neq \lambda_k^{*n}$ for any $m \neq n$. Then, $\{\pi^m, \{\lambda_\xi^{*m}, \{\lambda_{t,\xi}^m\}_{t=2}^T\}_{\xi\in X}\}_{m=1}^M$ is uniquely determined from $\{\tilde{P}(\{a_t, x_t\}_{t=1}^T) : \{a_t, x_t\}_{t=1}^T \in (A \times X)^T\}$.

The following proposition corresponds to Proposition 2 and relaxes the identification condition of Proposition 3 when $T \geq 5$ by utilizing all the marginals of $\tilde{P}(\{a_t, x_t\}_{t=1}^T)$.

Proposition 4 *Suppose Assumption 2 holds. Assume $T \geq 3$ is odd and define $u = (T-1)/2$. Suppose $X = \{1, \ldots, |X|\}$. Define*

$$\bar{\Lambda}_0 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \quad \bar{\Lambda}_1 = \begin{bmatrix} \lambda_{2,1}^1 & \cdots & \lambda_{2,|X|}^1 \\ \vdots & & \vdots \\ \lambda_{2,1}^M & \cdots & \lambda_{2,|X|}^M \end{bmatrix}.$$

For $l = 2, \ldots, u$, define $\bar{\Lambda}_l$ to be a matrix whose elements consists of the l-variate product of the form $\lambda^m_{2,j_2}\lambda^m_{3,j_3}\ldots\lambda^m_{l,j_{l+1}}$, covering all possible l ordered combinations (with replacement) of $(j_2, j_3, \ldots, j_{l+1})$ from $(1, \ldots, |X|)$. For example,

$$\bar{\Lambda}_2 = \begin{bmatrix} \lambda^1_{2,1}\lambda^1_{3,1} & \cdots & \lambda^1_{2,1}\lambda^1_{3,|X|} & \lambda^1_{2,2}\lambda^1_{3,1} & \cdots & \lambda^1_{2,2}\lambda^1_{3,|X|} & \lambda^1_{2,|X|}\lambda^1_{3,1} & \cdots & \lambda^1_{2,|X|}\lambda^1_{3,|X|} \\ \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ \lambda^M_{2,1}\lambda^M_{3,1} & \cdots & \lambda^M_{2,1}\lambda^M_{3,|X|} & \lambda^M_{2,2}\lambda^M_{3,1} & \cdots & \lambda^M_{2,2}\lambda^M_{3,|X|} & \lambda^M_{2,|X|}\lambda^M_{3,1} & \cdots & \lambda^M_{2,|X|}\lambda^M_{3,|X|} \end{bmatrix}.$$

Define an $M \times \sum_{l=0}^{u} |X|^l$ matrix $\bar{\Lambda}$ as $\bar{\Lambda} = [\bar{\Lambda}_0, \bar{\Lambda}_1, \bar{\Lambda}_2, \ldots, \bar{\Lambda}_u]$. Define $\bar{L}^\diamond_1$ to be a $M \times M$ matrix whose first column is $\bar{\Lambda}_0$ and whose other $M-1$ columns are from the columns of $\bar{\Lambda}$ but $\bar{\Lambda}_0$. Define $\bar{L}^\diamond_2$ to be a $M \times M$ matrix whose first column is $\bar{\Lambda}_0$ and whose other columns are from $\bar{\Lambda}_l, 1 \leq l \leq u$ with $\lambda^m_{2,j_2}\lambda^m_{3,j_3}\ldots\lambda^m_{l+1,j_{l+1}}$ replaced with $\lambda^m_{u+2,j_{u+2}}\lambda^m_{u+3,j_{u+3}}\ldots\lambda^m_{u+l+1,j_{u+l+1}}$.

Suppose (a) $\sum_{l=0}^{u} |X|^l \geq M$, (b) $\bar{L}^\diamond_1$ and $\bar{L}^\diamond_2$ are nonsingular, and (c) there exists $k \in X$ such that $\lambda^{*m}_k \neq \lambda^{*n}_k$ for any $m \neq n$. Then, $\{\pi^m, \{\lambda^{*m}_j, \{\lambda^m_{t,j}\}_{t=2}^{T}\}_{j=1}^{|X|}\}_{m=1}^{M}$ is uniquely determined from $\{\tilde{P}(\{a_t, x_t\}_{t=1}^{T}) : \{a_t, x_t\}_{t=1}^{T} \in (A \times X)^T\}$.

The proof is omitted because it is similar to that of Proposition 2. Note that $\sum_{l=0}^{u} |X|^l > \sum_{l=0}^{u} \binom{|X|+l-1}{l}$ and the condition on $|X|$ of Proposition 4 is weaker than that of Proposition 2. This is because the choice probabilities depend on time and the order of the choices becomes relevant for distinguishing different types. As a result, the number of restrictions implied by the submodels, analogously defined to (10)-(11) but with time-subscript, is even larger in the time-dependent case.

3.2 Type-specific transition functions

In empirical applications, we may encounter a case where the transition pattern of state variables is heterogeneous across individuals, even after controlling for other observables.

In this subsection, we extend the baseline model (8) to accommodate type-specific transition functions:

$$P(\{a_t, x_t\}_{t=1}^{T}) = \sum_{m=1}^{M} \pi^m p^{*m}(x_1, a_1) \prod_{t=2}^{T} f^m(x_t|x_{t-1}, a_{t-1}) P^m(a_t|x_t). \tag{18}$$

Namely, we relax Assumption 1(c), but still impose Assumption 1(a)-(b). To facilitate the discussion, define $s_t = (a_t, x_t)$, $q^{*m}(s_1) = p^{*m}(x_1, a_1)$, and $Q^m(s_t|s_{t-1}) = f^m(x_t|x_{t-1}, a_{t-1})P^m(a_t|x_t)$, and rewrite the model (18) as

$$P(\{s_t\}_{t=1}^{T}) = \sum_{m=1}^{M} \pi^m q^{*m}(s_1) \prod_{t=2}^{T} Q^m(s_t|s_{t-1}). \tag{19}$$

Unlike the transformed baseline model (9), s_t appears both in $Q^m(s_t|s_{t-1})$ and $Q^m(s_{t+1}|s_t)$, and creates the dependence between these terms. This dependence causes two potential problems. First, we can no longer integrate out an arbitrary s_t, say s_2, and create the marginals. Second, the variation of s_t affects $P(\{s_t\}_{t=1}^T)$ via both $Q^m(s_t|s_{t-1})$ and $Q^m(s_{t+1}|s_t)$.

The first problem is solved relatively easily; we can still deduce the marginals by sequentially integrating out "backwards," by integrating out s_T first, then s_{T-1}, and so on. We just cannot derive the marginals with respect to pairs of s_t such as (s_1, s_3). We solve the second problem by considering the sequence (s_{t-1}, s_t, s_{t+1}) for various values of s_t while fixing the values of s_{t-1} and s_{t+1}. Once s_{t-1} and s_{t+1} are fixed, the variation of s_t does not affect the state variables in other periods because of the Markovian structure of $Q^m(s_t|s_{t-1})$. As a result, we can use this variation to distinguish different types. Let $\bar{s} \in S = A \times X$ be a fixed value of s, and define

$$\tilde{\pi}_{\bar{s}}^m = \pi^m q^{*m}(\bar{s}), \quad \gamma_{\bar{s}}^m(s) = Q^m(\bar{s}|s)Q^m(s|\bar{s}).$$

Assume T is even, and consider $P(\{s_t\}_{t=1}^T)$ with $s_t = \bar{s}$ for odd t:

$$P(\{s_t\}_{t=1}^T|\ s_t = \bar{s} \text{ for } t \text{ odd}) = \sum_{m=1}^M \tilde{\pi}_{\bar{s}}^m \left(\prod_{t=2,4,\ldots}^{T-2} \gamma_{\bar{s}}^m(s_t) \right) Q^m(s_T|\bar{s}). \tag{20}$$

This conditional mixture model shares the property of independent marginals with (9), and hence we can identify its components for each $\bar{s} \in S$.

The following proposition establishes a sufficient condition for nonparametric identification of model (20). Because of the temporal dependence in s_t, the requirement on T becomes $T \geq 6$.

Proposition 5 *Suppose Assumption 1(a)-(b) holds. Assume $T \geq 6$. Let $\xi_j, j = 1, \ldots, M-1$, be elements of S and define*

$$\underset{(M\times M)}{G_{\bar{s}}} = \begin{bmatrix} 1 & \gamma_{\bar{s}}^1(\xi_1) & \cdots & \gamma_{\bar{s}}^1(\xi_{M-1}) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \gamma_{\bar{s}}^M(\xi_1) & \cdots & \gamma_{\bar{s}}^M(\xi_{M-1}) \end{bmatrix}.$$

Suppose there exists some $\{\xi_1, \ldots, \xi_{M-1}\}$ such that $G_{\bar{s}}$ is nonsingular and there exists $r \in S$ such that $Q^m(r|\bar{s}) \neq Q^n(r|\bar{s})$ for any $m \neq n$. Then, $\{\tilde{\pi}_{\bar{s}}^m, \{\gamma_{\bar{s}}^m(s), Q^m(s|\bar{s})\}_{s\in S}\}_{m=1}^M$ is uniquely determined from $\{P(\{s_t\}_{t=1}^T) : \{s_t\}_{t=1}^T \in S^T\}$.

Having identified $\{\tilde{\pi}_{\bar{s}}^m, \{\gamma_{\bar{s}}^m(s), Q^m(s|\bar{s})\}_{s\in S}\}_{m=1}^M$ for $\bar{s}$, now we turn to the identification of the primitive parameters π^m, $p^{*m}(a,x)$, $f^m(x'|x,a)$, and $P^m(a|x)$. Repeating Proposition 5 for all $\bar{s} \in S$, we obtain $\pi^m q^{*m}(s) = \pi^m p^{*m}(a,x)$ for all $(a,x) \in A \times X$. Then, π^m is determined by $\pi^m = \sum_{(a,x)\in A\times X} \pi^m p^{*m}(a,x)$, and we identify $p^{*m}(a,x) = (\pi^m p^{*m}(a,x))/\pi^m$. For the identification of the transition functions and the conditional choice probabilities, recall

$Q^m(s|\bar{s}) = f^m(x|\bar{x},\bar{a})P^m(a|x)$ with $(\bar{a},\bar{x}) = \bar{s}$. Summing $Q^m(s|\bar{s})$ over $a \in A$ gives $f^m(x|\bar{x},\bar{a})$, and we then have $P^m(a|x) = Q^m(s|\bar{s})/f^m(x|\bar{x},\bar{a})$.

Therefore, a sufficient condition for identifying the primitive parameters π^m, $p^{*m}(a,x)$, $f^m(x'|x,a)$, and $P^m(a|x)$ for all m,a,x,x' is summarized as follows: $T \geq 6$, $|S| \geq M-1$, and, for all $\bar{s} \in S$, the matrix $G_{\bar{s}}$ has rank M for some $\{\xi_1,\ldots,\xi_{M-1}\}$, and there exists $r \in S$ such that $Q^m(r|\bar{s}) \neq Q^n(r|\bar{s})$ for all $m \neq n$. These conditions are likely to hold if $|S| \gg M$ and the transition pattern of s is sufficiently heterogeneous across different types.

When $T > 6$, we can relax the condition $|S| \geq M-1$ of Proposition 5 by applying the argument of Proposition 2. Define $\gamma^m_{\bar{s},1}(s_1) = \gamma^m_{\bar{s}}(s_1)$ and $\gamma^m_{\bar{s},2}(s_1,s_2) = Q^m(\bar{s}|s_1)Q^m(s_1|s_2)Q^m(s_2|\bar{s})$, and similarly define $\gamma^m_{\bar{s},l}(s_1,\ldots,s_l)$ for $l \geq 3$ as a $(l+1)$-variate product of $Q^m(s'|s)$'s of the form $Q^m(\bar{s}|s_1)Q^m(s_1|s_2)\cdots Q^m(s_l|\bar{s})$ for $\{s_t\}_{t=1}^l \in S^l$.

Proposition 6 *Suppose Assumption 1(a)-(b) holds. Assume $T > 6$ and is even and define $u = (T-4)/2$. Suppose that $S = \{1,2,\ldots,|S|\}$. Define*

$$\tilde{\Lambda}_0 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \quad \tilde{\Lambda}_1 = \begin{bmatrix} \gamma^1_{\bar{s},1}(1) & \cdots & \gamma^1_{\bar{s},1}(|S|) \\ \vdots & \ddots & \vdots \\ \gamma^M_{\bar{s},1}(1) & \cdots & \gamma^M_{\bar{s},1}(|S|) \end{bmatrix}.$$

For $l = 2,\ldots,u$, define $\tilde{\Lambda}_l$ to be a matrix whose elements consists of $\gamma^m_{\bar{s},l}(s_1,\ldots,s_l)$, covering all possible unordered combinations (with replacement) of $(s_1,\ldots,s_l)$ from S^l. For example,

$$\underset{(M\times\binom{|S|+1}{2})}{\tilde{\Lambda}_2} = \begin{bmatrix} \gamma^1_{\bar{s},2}(1,1) & \cdots & \gamma^1_{\bar{s},2}(1,|S|) & \gamma^1_{\bar{s},2}(2,2) & \cdots & \gamma^1_{\bar{s},2}(2,|S|) & \cdots & \gamma^1_{\bar{s},2}(|S|,|S|) \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots \\ \gamma^1_{\bar{s},2}(1,1) & \cdots & \gamma^1_{\bar{s},2}(1,|S|) & \gamma^1_{\bar{s},2}(2,2) & \cdots & \gamma^1_{\bar{s},2}(2,|S|) & \cdots & \gamma^1_{\bar{s},2}(|S|,|S|) \end{bmatrix}.$$

Define an $M \times \sum_{l=0}^u \binom{|S|+l-1}{l}$ matrix $\tilde{\Lambda}$ as $\tilde{\Lambda} = [\tilde{\Lambda}_0,\tilde{\Lambda}_1,\tilde{\Lambda}_2,\ldots,\tilde{\Lambda}_u]$, and define $G^\diamond_{\bar{s}}$ to be a $M \times M$ matrix consisting of M columns from $\tilde{\Lambda}$ but with the first column unchanged.

Suppose (a) $\sum_{l=0}^u \binom{|S|+l-1}{l} \geq M$, (b) $G^\diamond_{\bar{s}}$ is nonsingular, and (c) there exists $r \in S$ such that $Q^m(r|\bar{s}) \neq Q^n(r|\bar{s})$ for any $m \neq n$. Then, $\{\tilde{\pi}^m_{\bar{s}}, \{\gamma^m_{\bar{s},l}(s_1,\ldots,s_l) : (s_1,\ldots,s_l) \in S^l\}_{l=1}^u, \{Q^m(s|\bar{s})\}_{s=1}^{|S|}\}_{m=1}^M$ is uniquely determined from $\{P(\{s_t\}_{t=1}^T) : \{s_t\}_{t=1}^T \in (A\times X)^T\}$.

The identification of the primitive parameters $\pi^m, p^{*m}(a,x), f^m(x'|x,a), P^m(a|x)$ follows from using the argument in the paragraph that follows Proposition 5.

In some applications, the model has two types of state variables, z_t and x_t, where the transition function of x_t depends on types, while the transition function of z_t is common across types. In such a case, we may relax the requirement on T in Proposition 5 using the variation of z_t as a main source of identification.

We assume that the transition function of (x',z') conditional on (x,z,a) takes the form $g(z'|x,z,a)f^m(x'|x,a)$, and impose an assumption analogous to Assumption 1(a)-(b):

Assumption 3 *(a)* $P^m(a_t|x_t,z_t) = P^m(a_t|x_t,z_t,a_{t-1})$*, where* (x_t,z_t) *does not include* a_{t-1}*. (b)* $f^m(x'|x,a) > 0$ *for all* $(x',x,a) \in X \times X \times A$ *and* $g(z'|x,z,a) > 0$ *for all* $(z',x,z,a) \in Z \times X \times Z \times A$ *and for* $m = 1,\ldots,M$*.*

Under Assumption 3, consider a model

$$\begin{aligned} P(\{a_t,x_t,z_t\}_{t=1}^T) &= \sum_{m=1}^M \pi^m p^{*m}(x_1,z_1,a_1) \\ &\quad \times \prod_{t=2}^T g(z_t|x_{t-1},z_{t-1},a_{t-1}) f^m(x_t|x_{t-1},a_{t-1}) P^m(a_t|x_t,z_t). \end{aligned} \tag{21}$$

Then, assuming $g(z_t|x_{t-1},z_{t-1},a_{t-1})$ is known and defining $s_t = (a_t,x_t)$, $\tilde{q}^{*m}(s_1,z_1) = p^{*m}(x_1,z_1,a_1)$, and $\tilde{Q}^m(s_t|s_{t-1},z_t) = f^m(x_t|x_{t-1},a_{t-1})P^m(a_t|x_t,z_t)$, we write this equation as

$$\tilde{P}(\{s_t,z_t\}_{t=1}^T) = \frac{P(\{a_t,x_t,z_t\}_{t=1}^T)}{\prod_{t=2}^T g(z_t|x_{t-1},z_{t-1},a_{t-1})} = \sum_{m=1}^M \pi^m \tilde{q}^{*m}(s_1,z_1) \prod_{t=2}^T \tilde{Q}^m(s_t|s_{t-1},z_t). \tag{22}$$

Because s_t appears in both $\tilde{Q}^m(s_t|s_{t-1},z_t)$ and $\tilde{Q}^m(s_{t+1}|s_t,z_{t+1})$, we need to sequentially integrate out s_t's backwards (i.e., s_T, then s_{T-1}, and so on), to obtain the lower dimensional submodels of (22). This is similar to the previously analyzed case without z_t. On the other hand, the presence of an additional state variable z_t provides another source of identification, and we can identify the types by using the variation of z_t, while fixing the value of $\{x_t\}_{t=1}^T$.

The next proposition provides a sufficient condition for nonparametric identification of the model (22). Define, for $\bar{s} \in S$ and $h,\xi \in Z$,

$$\tilde{\pi}_{\bar{s},h}^m = \pi^m \tilde{q}^{*m}(\bar{s},h), \quad \tilde{\gamma}_{\bar{s}}^m(\xi) = \tilde{Q}^m(\bar{s}|\bar{s},\xi).$$

Proposition 7 *Suppose that Assumption 3 holds. Assume* $T \geq 4$*. Define*

$$\underset{(M\times M)}{\bar{G}_{\bar{s}}} = \begin{bmatrix} 1 & \tilde{\gamma}_{\bar{s}}^1(\xi_1) & \cdots & \tilde{\gamma}_{\bar{s}}^1(\xi_{M-1}) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \tilde{\gamma}_{\bar{s}}^M(\xi_1) & \cdots & \tilde{\gamma}_{\bar{s}}^M(\xi_{M-1}) \end{bmatrix}.$$

Suppose that there exists some $\{\xi_1,\ldots,\xi_{M-1}\}$ *such that* $\bar{G}_{\bar{s}}$ *is nonsingular and that there exists* $(r,k) \in S \times Z$ *such that* $\tilde{Q}^m(r|\bar{s},k) \neq \tilde{Q}^n(r|\bar{s},k)$ *for any* $m \neq n$*. Then* $\{\tilde{\pi}_{\bar{s},h}^m, \{\tilde{\gamma}_{\bar{s}}^m(\xi)\}_{\xi\in Z}, \{\tilde{Q}^m(s|\bar{s},\xi)\}_{(s,\xi)\in S\times Z}\}_{m=1}^M$ *is uniquely determined from* $\{\tilde{P}(\{s_t,z_t\}_{t=1}^T) : \{s_t,z_t\}_{t=1}^T \in (S\times Z)^T\}$*.*

Repeating Proposition 7 for all $(\bar{s},h) \in S \times Z$ identifies $\tilde{\pi}_{\bar{s},h}^m$ for all $(\bar{s},h) \in S \times Z$ and $m = 1,\ldots,M$, and we may obtain $\pi^m = \sum_{(\bar{s},h)\in S\times Z} \tilde{\pi}_{\bar{s},h}^m$ and $\tilde{q}^{*m}(s,z) = \tilde{\pi}_{s,z}^m/\pi^m$. The transition functions are obtained as $f^m(x'|x,a) = \sum_{a'\in A} \tilde{Q}^m((a',x')|s,z)$, allowing us to identify the conditional choice probabilities as $P^m(a'|x,z) = \tilde{Q}^m((a',x')|s,z)/f^m(x'|x,a)$.

The requirement of $T = 4$ is weaker than that of $T = 6$ in Proposition 5. This is because Proposition 7 utilizes the variation of z_t rather than that of x_t as a main source of identification. Consequently, its argument is not affected by the temporal dependence of x_t. When $T > 4$, we may apply the argument of Proposition 2 to relax the sufficient condition for identification in Proposition 7, but we do not pursue it here; Proposition 6 provides a similar result.

3.3 Lagged dependent variable

In applications, including the lagged choice in explanatory variables for the current choice is a popular way of specifying dynamic discrete choice models. We show that the models in the previous subsection can be modified to include the lagged choice as an element of current state variable and a similar set of conditions for identification can be derived.

First, consider the finite mixture model (3), which we restate here:

$$P(\{a_t, x_t\}_{t=1}^T) = \sum_{m=1}^M \pi^m p^{*m}(a_1, x_1) \prod_{t=2}^T f^m(x_t|x_{t-1}, a_{t-1}) P^m(a_t|x_t, a_{t-1}), \tag{23}$$

where the conditional choice probability, $P^m(a_t|x_t, a_{t-1})$, contains lagged choice a_{t-1} as one of the conditioning variables. We assume the elements of x_t do not include a_{t-1}, and Assumption 1(b) holds.

Define $q^{*m}(s_1) = p^{*m}(a_1, x_1)$ and $Q^m(s_t|s_{t-1}) = f^m(x_t|x_{t-1}, a_{t-1})P^m(a_t|x_t, a_{t-1})$ with $s_t = (a_t, x_t)$. Then (23) can be written as (19), where the only difference from the original formulation is the definition of $Q^m(s_t|s_{t-1})$. Consequently, we may apply Proposition 5 and Proposition 6 to (23) and find a sufficient condition for nonparametric identification, including $T \geq 6$ and the rank of the matrix $G_{\bar{s}}$ or $G_{\bar{s}}^{\diamond}$.

We may also extend the model (23) to include an additional state variable z_t, similarly to (21). We simply need to replace $P^m(a_t|x_t, z_t)$ in (21) with $P^m(a_t|x_t, z_t, a_{t-1})$ and apply Proposition 7.

Example 3 (Identification of models with heterogeneous coefficients) *Consider the model of Example 1. Denote the i^{th} observation's type by $m_i \in \{1, \ldots, M\}$ so that $(\beta_i, \gamma_i) = (\beta^{m_i}, \gamma^{m_i})$ and $P^{m_i}(a_{it} = 1|x_{it}, a_{i,t-1}) = 1 - \Phi(x_{it}'\beta^{m_i} + a_{i,t-1}\gamma^{m_i})$. The initial observation, (a_{i1}, x_{i1}), is randomly drawn from $p^{*m_i}(a_1, x_1)$ while the transition function of x_{it} is given by $f^{m_i}(x_{it}|x_{i,t-1}, a_{i,t-1})$.*

If the conditions in Proposition 5 including $T \geq 6$, $|S| \geq M - 1$, and the rank of $G_{\bar{s}}$ are satisfied, then $p^{*m}(a_1, x_1)$, $f^m(x_t|x_{t-1}, a_{t-1})$, and $P^m(a_t = 1|x_t, a_{t-1}) = 1 - \Phi(x_t'\beta^m + a_{t-1}\gamma^m)$ are identified for all m. Since $x_t'\beta^m + a_{t-1}\gamma^m = \Phi^{-1}(1 - P^m(a_t = 1|x_t, a_{t-1}))$, the identification of type-specific coefficient (β^m, γ^m) follows if the rank of the $(\dim(\beta^m) + 1) \times |A||X|$ Jacobian matrix consisting of the derivatives of $x_t'\beta^m + a_{t-1}\gamma^m$ with respect to the parameter (β^m, γ^m) is $\dim(\beta^m) + 1$.

Example 4 (Dynamic discrete games) *Consider the model of dynamic discrete games with unobserved market characteristics studied by Aguirregabiria and Mira (2006), section 3.5. There are N_i ex-ante identical "global" firms competing in N_h local markets. There are M market types and each market's type is common knowledge to all firms but unknown to a researcher. In market h, of which type is m_h, a firm i maximizes the expected discounted sum of profits $E[\sum_{s=t}^{\infty} \beta^{s-t}\{\Pi(x_{hs}, a_{hs}; \theta^{m_h}) + \epsilon_{his}(a_{his})\}|a_{ht}, x_{ht}; \theta^{m_h}]$, where x_{ht} is state variable that is common knowledge for all firms, while $\epsilon_{hit}(a_{hit})$ is state variable that is private information to firm i. The state variable x_{ht} may contain the past choice $a_{h,t-1}$. The researcher observes x_{ht} but not ϵ_{hit}. There is no interaction across different markets.*

Denote the strategy of firm i in market h by σ_i^h. Given a set of strategy functions $\sigma^h = \{\sigma_i^h(x, \epsilon_i) : i = 1, \ldots, N_i\}$, the expected behavior of firm i from the viewpoint of the rest of the firms is summarized by the conditional choice probabilities $P_i^{\sigma^h}(a_i|x) = \int 1\{\sigma_i^h(x, \epsilon_i) = a_i\} g(\epsilon_i|x) d\epsilon_i$, where $g(\epsilon_i|x)$ is a density function for $\epsilon = \{\epsilon(a) : a \in A\}$. By assuming that ϵ_i's are iid across firms, the expected profit and the transition probability of x for firm i under σ^h is given by $\pi_i^{\sigma^h}(x, a_i; \theta^{m_h}) = \sum_{a_{-i} \in A^{N-1}} \left(\prod_{j \neq i} P_j^{\sigma^h}(a_j|x) \right) \Pi(x, a_i, a_{-i}; \theta^{m_h})$ and $f_i^{\sigma^h}(x'|x, a_i; \theta^{m_h}) = \sum_{a_{-i} \in A^{N-1}} \left(\prod_{j \neq i} P_j^{\sigma^h}(a_j|x) \right) f(x'|x, a_i, a_{-i}; \theta^{m_h})$, respectively. Then the Bellman equation is

$$V_i^{\sigma^h}(x; \theta^{m_h}) = \int \max_{a_i \in A} \left\{ \pi_i^{\sigma^h}(x, a_i; \theta^{m_h}) + \epsilon_i(a_i) + \beta \sum_{x' \in X} V(x') f^{\sigma^h}(x'|x, a_i; \theta^{m_h}) \right\} g(d\epsilon|x).$$

A set of strategy functions σ^{h*} in a stationary Markov perfect equilibrium satisfies $\sigma^{h*}(x, \epsilon_i) = \arg\max_{a_i \in A} \{\pi_i^{\sigma^{h*}}(x, a_i; \theta^{m_h}) + \epsilon_i(a_i) + \beta \sum_{x' \in X} V(x') f^{\sigma^{h*}}(x'|x, a_i; \theta^{m_h})\}$, and the equilibrium conditional choice probabilities are given by $P^{\sigma^*}(a_i|x; \theta^{m_h}) = \int 1\{a_i = \sigma_i^{h*}(x, \epsilon_i)\} g(d\epsilon_i)$.

Suppose that a panel data $\{\{\{a_{hit}, x_{hit}\}_{t=1}^T\}_{i=1}^{N_i}\}_{h=1}^{N_h}$ is available. Consider the asymptotics where $N_h \to \infty$ with N_i and T fixed. The initial distribution of (a, x) differs across market types and is given by $p^{*m}(a, x)$. Let $P^m(a_{it}|x_{it}) = P^{\sigma^*}(a_{it}|x_{it}; \theta^m)$ and $f^m(x_{it}|x_{i,t-1}, a_{i,t-1}) = f(x_{it}|x_{i,t-1}, a_{i,t-1}; \theta^m)$. Then, for each market, the likelihood function becomes a mixture across different unobserved market types:

$$P(\{\{a_{it}, x_{it}\}_{t=1}^T\}_{i=1}^{N_i}) = \sum_{m=1}^{M} \pi^m \prod_{i=1}^{N_i} p^{*m}(a_{i1}, x_{i1}) \prod_{t=2}^{T} P^m(a_{it}|x_{it}) f^m(x_{it}|x_{i,t-1}, a_{i,t-1}).$$

In this case, N_i plays a similar role to T for identification. If $T = 2$, we may apply the argument of Proposition 5 to show that the choice probabilities are identified when $T = 2$, $N_i \geq 3$, and $|X| \geq M - 1$ (and the corresponding rank conditions hold). When T and N_i are large, we may apply the argument of Proposition 6 and the sufficient condition for identification can be weakened to $T \geq 2$, $N_i \geq 3$, and $\sum_{l=0}^{u} \binom{|X|+l-1}{l} \geq M$, where $u = (N_i - 1)(T - 1)/2$ with N_i odd. Once identification is established, we may consistently estimate the type-specific choice

probabilities P^m's using the series logit estimator discussed in Section 4 below.

3.4 Limited transition pattern

This section analyzes the identification condition of the model (3) when Assumption 1(b) is relaxed. In some applications, the transition pattern of x is limited and not all $x' \in X$ is reachable with a positive probability if one starts from (x, a). In such a case, a set of sequences $\{a_t, x_t\}_{t=1}^T$ that can be realized with a positive probability also becomes limited, and the number of restrictions from a set of the submodels becomes smaller and identification becomes harder.

Example 5 (Bus engine replacement model (Rust, 1987)) *Suppose $a \in \{0, 1\}$ is the replacement decision for a bus engine, where $a = 1$ corresponds to replacing a bus engine. Let x denote the mileage of a bus engine with $X = \{1, 2, \ldots\}$. The transition function of x_t is*

$$f(x_{t+1}|x_t, a_t; \theta) = \begin{cases} \theta_{f,1} & \text{for } x_{t+1} = (1 - a_t)x_t + a_t, \\ \theta_{f,2} & \text{for } x_{t+1} = (1 - a_t)x_t + a_t + 1, \\ 1 - \theta_{f,1} - \theta_{f,2} & \text{for } x_{t+1} = (1 - a_t)x_t + a_t + 2, \\ 0 & \text{otherwise,} \end{cases}$$

and not all $x' \in X$ can be realized from (x, a).

If $f(x'|x, a) = 0$ for some (x', x, a) and not all $x' \in X$ can be reached from (a, x), then $\tilde{P}(\{a_t, x_t\}_{t=1}^T)$ in (9) is not well-defined for some values of $\{a_t, x_t\}$. Consequently, we cannot integrate out arbitrary (x_t, a_t) to construct a lower-dimensional submodel like (10). We handle this problem by applying the approach developed in Proposition 7. Note that integrating out a_T from (8) causes no problem even if $f(x'|x, a) = 0$ for some (x', x, a). Once a_T is integrated out, we can sequentially integrate out backwards x_T, a_{T-1}, x_{T-1}, and so on, and deduce lower-dimensional submodels $P(\{a_t, x_t\}_{t=1}^\tau)$ with $1 \leq \tau \leq T$. As in the previous sections, these submodels constitute restrictions that can be used to pin down π^m and $P^m(a|x)$. We fix the values of (a_1, x_1) and (a_τ, x_τ) and focus on the values of (a_t, x_t) that is realizable between (a_1, x_1) and (a_τ, x_τ). The difference in response patterns between (a_1, x_1) and (a_τ, x_τ) provides a source of identification.

To fix the idea, assume $T = 4$, and fix $a_t = 0$ for all t, $x_1 = h$, and $x_\tau = k$. We set $a_t = 0$ for all t to simplify the presentation, but it is possible to choose different sequences of $\{a_t\}_{t=1}^T$. Let B^h and C^h be subsets of X. We use the variations of x within B^h and C^h as a source of

identification. Specifically, consider the following submodels

$$
\begin{aligned}
&P(x_1 = h, (x_2, x_3) \in B^h \times C^h, x_4 = k; a_t = 0 \text{ for all } t),\\
&P(x_1 = h, x_2 \in B^h, x_3 = k; a_t = 0 \text{ for all } t),\\
&P(x_1 = h, x_2 \in C^h, x_3 = k; a_t = 0 \text{ for all } t),\\
&P(x_1 = h, x_2 = k; a_t = 0 \text{ for all } t).
\end{aligned}
$$

For these submodels to provide restrictions for identification, the value of the transition function $f(x'|x, 0)$ in these submodels must be positive. In other words, all the points in B^h must be reachable from h. If $f(x|h, 0) = 0$ for some $x \in B^h$, then the first and second equation become zero for those values of x, and they provide no information for identifying types. In such a case, we need to exclude a set of x's with $f(x|h, 0) = 0$ from B^h. For the same reason, the first and third equations imply that all the points in C^h must be reachable from all the points in B^h and h. Finally, the first, third, and fourth equations imply that k must be reachable from all the points in B^h and C^h and h.

We develop notations to state the restrictions on B^h and C^h formally. For a singleton $\{x\} \subset X$, let $\Gamma(a, \{x\}) = \{x' \in X : f(x'|x, a) > 0\}$ denote a set of $x' \in X$ that can be reached from (a, x) in the next period with a positive probability. For a subset $W \subseteq X$, define $\Gamma(a, W)$ as the intersection of $\Gamma(a, \{x\})$'s across all x's in W: $\Gamma(a, W) = \cap_{x \in W} \Gamma(a, \{x\})$.

The following assumption summarizes the restrictions on B^h and C^h. Note that the choice of C^h is affected by how B^h is chosen. If Assumption 1(b) holds, it is possible to set $B^h = C^h = X$. The assumption $P^m(a|x) > 0$ is necessary to guarantee that the submodels are well-defined.

Assumption 4 *(a) $P^m(a|x) > 0$ for all $(a, x) \in A \times X$ and $m = 1, \ldots, M$. (b) $h, k \in X$, B^h, and C^h satisfy*

$$
B^h \subseteq \Gamma(0, \{h\}), \quad C^h \subseteq \Gamma(0, B^h) \cap \Gamma(0, \{h\}), \quad \{k\} \subseteq \Gamma(0, C^h) \cap \Gamma(0, B^h) \cap \Gamma(0, \{h\}).
$$

The next proposition provides a sufficient condition for identification when Assumption 1(b) is replaced with Assumption 4. Define, for $h, \xi \in X$,

$$
\pi_h^m = \pi^m p^{*m}(a_1 = 0, x_1 = h) \quad \text{and} \quad \lambda_\xi^m = P^m(a = 0 | x = \xi).
$$

Proposition 8 *Suppose that Assumptions 1(a),(c) and 4 hold. Suppose $T = 4$ and $|B^h|, |C^h| \geq M - 1$. Let $\{\xi_1^b, \ldots, \xi_{M-1}^b\}$ and $\{\xi_1^c, \ldots, \xi_{M-1}^c\}$ be elements of B^h and C^h, respectively. Define*

$$
\underset{(M \times M)}{\tilde{G}_1} = \begin{bmatrix} 1 & \lambda_{\xi_1^b}^1 & \lambda_{\xi_2^b}^1 & \cdots & \lambda_{\xi_{M-1}^b}^1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_{\xi_1^b}^M & \lambda_{\xi_2^b}^M & \cdots & \lambda_{\xi_{M-1}^b}^M \end{bmatrix}, \quad \underset{(M \times M)}{\tilde{G}_2} = \begin{bmatrix} 1 & \lambda_{\xi_1^c}^1 & \lambda_{\xi_2^c}^1 & \cdots & \lambda_{\xi_{M-1}^c}^1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_{\xi_1^c}^M & \lambda_{\xi_2^c}^M & \cdots & \lambda_{\xi_{M-1}^c}^M \end{bmatrix}.
$$

Suppose that $\tilde{G}_1$ and $\tilde{G}_2$ are nonsingular for some $\{\xi_1^b, \ldots, \xi_{M-1}^b\}$ and $\{\xi_1^c, \ldots, \xi_{M-1}^c\}$ and that $\lambda_k^m \neq \lambda_k^n$ for any $m \neq n$. Then $\{\pi_h^m, \lambda_\xi^m : \xi \in B^h \cup C^h\}_{m=1}^M$ is uniquely determined from $\{\tilde{P}(\{a_t, x_t\}_{t=1}^T) : \{a_t, x_t\}_{t=1}^T \in (A \times X)^T\}$.

Assuming that all the values of x's can be realized in the initial period, we may repeat the above argument for all possible values of x_1's to identify λ_ξ^m for any $\xi \in \cup_{h \in X} B^h$. Furthermore, we can repeat the argument for different sequences of $\{a_t\}_{t=1}^4$ to increase the identifiable elements of $P^m(a|x)$'s. For instance, by choosing $B^h = \Gamma(a, \{h\})$, λ_l^m is identified for all $l \in X$ if the union of $\Gamma(a, \{h\})$ across different $(a, h) \in A \times X$ include all the elements of X so that $X = \cup_{(a,h) \in A \times X} \Gamma(a, \{h\})$. This is a weak condition and is satisfied if X is an ergodic set. However, setting $B^h = \Gamma(a, \{h\})$ may lead to a small number of identifiable types.

Example 5 (continued) *In Example 5, assume the initial distribution $p^{*m}(x, a)$ is defined as the fixed point of the type-specific stationary distribution. Setting $a_t = 0$ for $t = 1, \ldots, 4$, we have $\Gamma(0, \{h\}) = \{h, h+1, h+2\}$ for any $h \in X$. Choose $B^h = \{h, h+1\}$, $C^h = \{h+1, h+2\}$, and $k = h+2$, then (h, B^h, C^h, k) satisfy Assumption 4(b). If the other assumptions of Proposition 8 are satisfied, we can identify $M = 3$ types and $\{\pi^m p^{*m}(0, x), P^m(0|x) : x = h, h+1, h+2\}_{m=1,2,3}$. This also identifies $P^m(1|x)$, since $P^m(1|x) = 1 - P^m(0|x)$. Repeating for all $h \in X$, we identify $P^m(a|x)$ for all $(a, x) \in A \times X$. We then identify $p^{*m}(x, a)$ using $P^m(a|x)$, $f(x'|x, a)$ and the fixed point constraint, while π^m is determined as $\pi^m p^{*m}(0, x)/p^{*m}(0, x)$.*

The sufficient condition of Proposition 8 does not allow one to identify many types when the size of B^h or C^h is small. It is possible to identify more types when we can find a subset D of X that is reachable from itself, namely $D \subseteq \Gamma(0, D)$. For example, if the transition pattern is such that $\Gamma(0, \{x\}) = \{x-1, x, x+1\}$ for some $x \in X$, then the set $\{x-1, x, x+1\}$ serves as D. In such cases, we can apply the logic of Proposition 2 to identify many types if $T > 4$.

Assumption 5 *(a) $P^m(a|x) > 0$ for all $(a, x) \in A \times X$ and $m = 1, \ldots, M$. (b) A subset D of X satisfies $D \subseteq \Gamma(0, D)$.*

Set $D = \{d_1, \ldots, d_{|D|}\}$. Define $\lambda_d^{*m} = p^{*m}((a, x) = (1, d))$ and $\lambda_d^m = P^m(a = 1 | x = d)$ for $d \in D$. Under Assumption 5, replacing X with D and simply repeating the proof of Proposition 2 gives the following proposition:

Proposition 9 *Suppose Assumptions 1(a),(c) and 5 hold. Assume $T \geq 4$ is odd and define $u = (T-1)/2$. Define Λ_r, $r = 0, \ldots, u$, analogously to Proposition 2 other than $(X, \lambda_{\xi_j}^{*m}, \lambda_{\xi_j}^m)$ is replaced with $(D, \lambda_{d_j}^{*m}, \lambda_{d_j}^m)$. Define an $M \times (\sum_{l=0}^u \binom{|D|+l-1}{l})$ matrix Λ as $\Lambda = [\Lambda_0, \Lambda_1, \Lambda_2, \ldots, \Lambda_u]$.*

Suppose (a) $\sum_{l=0}^u \binom{|D|+l-1}{l} \geq M$, (b) we can construct a nonsingular $M \times M$ matrix $L^\diamond$ by setting its first column as Λ_0 and choosing the other $M-1$ columns from the columns of Λ but Λ_0, and (c) there exists $d_k \in D$ such that $\lambda_{d_k}^{*m} \neq \lambda_{d_k}^{*n}$ for any $m \neq n$. Then $\{\pi^m, \{\lambda_{d_j}^{*m}, \lambda_{d_j}^m\}_{j=1}^{|D|}\}_{m=1}^M$ is uniquely determined from $\{\tilde{P}(\{a_t, x_t\}_{t=1}^T) : \{a_t, x_t\}_{t=1}^T \in (A \times X)^T\}$.

For example, if $|D| = 3$ and $T = 5$, the number of identifiable types becomes $\binom{2}{0} + \binom{3}{1} + \binom{4}{2} = 10$. Identifying more types is also possible when the model has an additional state variable z_t whose transition pattern is not limited and there is a state $\bar{x}$ such that $P(x_1 = \cdots = x_T = \bar{x}) > 0$ for some sequence of a_t. Then, for $x = \bar{x}$, we can use the variation of z_t and apply Proposition 7. This increases the number of identifiable types to $|Z| + 1$.

3.5 Local identifiability of finite mixture models

Consider a general mixture model of dynamic discrete choices:

$$P(\{s_t\}_{t=1}^T) = \sum_{m=1}^M \pi^m q_1^{*m}(s_1) \prod_{t=2}^T Q_t^m(s_t|s_{t-1}), \tag{24}$$

where the state space of s is given by a finite set S. In the following, extending the argument of Goodman (1974), we derive the sufficient condition for local identifiability for (24).

Consider the case where $T = 3$ and $q_1^{*m}(s) > 0$ and $Q_t^m(s'|s) > 0$ for all possible pairs of (s', s). The model (24) contains $M(2|S|^2 - |S|) - 1$ unknown parameters while it provides a total of $|S|^3 - 1$ restrictions on $q_1^{*m}(s)$ and $Q_t^m(s'|s)$.[10] When $M(2|S|^2 - |S|) > |S|^3$, the number of unknown parameters is necessarily larger than the number of restrictions, and thus the model is not identifiable. While this condition puts a bound on the maximum identifiable number of types, such a bound may not be so informative due to possible redundancy in the restrictions.

We consider next whether the parameter is uniquely determined from the restrictions imposed by the model locally. We say the parameter $\theta = \{\pi^m, \{q_1^m(s), Q_t^m(s'|s) : (s', s) \in S \times S\}_{t=1}^T\}_{m=1}^M$ is locally identifiable if it is uniquely determined from $\{P(\{s_t\}_{t=1}^T) : \{s_t\}_{t=1}^T \in S^T\}$ within some neighborhood of θ. See Rothenberg (1971) for the definition of local identifiability.

There are $|S|^3 - 1$ nonlinear equations for $M(2|S|^2 - |S|) - 1$ unknowns. The local identifiability can be established from the rank condition of a linearized version of these nonlinear equations. Specifically, consider a $M(2|S|^2 - |S|) - 1$ by $|S|^3 - 1$ matrix, that consists of the derivatives of the right-hand side of (24) with respect to the $M(2|S|^2 - |S|) - 1$ parameters evaluated at all the possible $|S|^3 - 1$ points. The parameter θ is locally identifiable if the rank of this matrix evaluated at θ is at least as large as the number of parameters, $|S|^3 - 1$.

The next proposition generalizes the above argument.

Assumption 6 *$q_1^{*m}(s) > 0$ and $Q_t^m(s'|s) > 0$ for all $(s', s) \in S \times S$ and $t = 1, 2, \ldots, T$.*

[10] Since $\sum_{s \in S} q_1^{*m}(s) = 1$ and $\sum_{s' \in S} Q_t^m(s'|s) = 1$, $q_1^{*m}(s)$ and $Q_t^m(s'|s)$ contain $|S| - 1$ and $(|S| - 1)|S|$ unknowns, respectively, for $m = 1, \ldots, M$ and $t = 2, 3$. There are also $M - 1$ unknowns for π^m's. Thus, the model contains $M(|S| - 1 + 2(|S| - 1)|S|) + M - 1 = M(2|S|^2 - |S|) - 1$ parameters. On the other hand, the model (24) provides $|S|^3 - 1$ restrictions because there are $|S|^3$ possible sequences of $\{a_t, x_t\}_{t=1}^T$ under the restriction $\sum_{\{a_t, x_t\}_{t=1}^T} P(\{a_t, x_t\}_{t=1}^T) = 1$.

Proposition 10 *Suppose that Assumptions 6 holds. Let* $J \equiv M(|S|-1+(T-1)(|S|-1)|S|)+M-1$ *and* $K \equiv |S|^T-1$. *Let* θ^0 *be the true parameter.*

(a) If $J > K$, *then* θ *is not uniquely determined from* $\{P(\{s_t\}_{t=1}^T) : \{s_t\}_{t=1}^T \in S^T\}$.

(b) Consider a J *by* K *matrix denoted by* $\Gamma(\theta)$, *which consists of the derivatives of the right-hand side of (24) with respect to the* J *parameters. Assume that, in an open neighborhood of* θ^0, $\Gamma(\theta)$ *has constant rank and the elements of* $\Gamma(\theta)$ *are continuous functions of* θ. *Then,* θ^0 *is locally identifiable from* $\{P(\{s_t\}_{t=1}^T) : \{s_t\}_{t=1}^T \in S^T\}$ *if the rank of* $\Gamma(\theta^0)$ *is* J.

$T \geq 3$ is necessary for identification without imposing further restrictions because the number of parameters is necessarily larger than the number of restrictions when $T = 2$. The proof of (a) follows from counting the number of unknowns and restrictions, while the proof of (b) follows from a standard result on Jacobians. When local identification condition of Proposition 10 fails, we may apply the methods suggested by Honoré and Tamer (2006) to calculate the parameter region identified from the set of restrictions implied by the model.

The local identification results similar to Proposition 10 may be derived for other models we have considered so far. Propositions 1-8 provide sufficient conditions, but they do not always utilize all the restrictions implied by the model. As a result, even in the case where the sufficient conditions of these propositions are not satisfied, these models may be locally identified.

4 Series logit estimation of finite mixture models of dynamic discrete choices

Assuming they are nonparametrically identified, we turn to the estimation of the models in the previous sections. We use a series logit estimator to estimate type-specific conditional choice probabilities $P^m(a|x,a')$ nonparametrically. The number of types, M, is assumed to be known. We assume that the state variable x is continuously distributed and consider a binary choice model with $A = \{0,1\}$; a slight modification accommodates multiple choices but with a more complicated notation. In practice, even when x has a discrete distribution, smoothing estimators are preferred over the frequency estimator when $|X|$ is large and data are sparse (see, for example, a Monte Carlo simulation result in Aguirregabiria and Mira (2006)). There is also a large literature in statistics on applying smoothing methods to discrete data (cf., Simonoff (1995) (1996) and the references therein).

Let $\{r_j(x) : j = 1,2,\ldots\}$ denote a sequence of known basis functions and let $R^K(x) = (r_1(x),\ldots,r_K(x))'$ be a K-vector of functions, where K denotes the number of basis functions to be used. Using an orthogonal polynomial basis such as Chebyshev polynomials avoids multicollinearity problems. When x is one-dimensional, $R^K(x)$ includes the polynomials of x up to power $K-1$. When x is r-dimensional, we need $K = (n+1)^r$ basis functions in order to include powers in all elements of x, at least up to n, into $R^K(x)$ (see Hirano et al. (2003) p. 1177 for

details). We assume $K \to \infty$ and $K/N \to 0$.

Let $\{\{a_{it}, x_{it}\}_{t=1}^T\}_{i=1}^N$ be a panel data such that $w_i = \{a_{it}, x_{it}\}_{t=1}^T$ is randomly drawn from model (3) across i's from the population. The transition function $f^m(x'|x,a)$ is assumed to be known, common across types. Assuming $f^m(x'|x,a)$ is unknown does not affect the results as long as its estimate converges to $f(x'|x,a)$ at an appropriate rate. The initial observation (x_{i1}, a_{i1}) is assumed to be drawn from the type-specific stationary distribution implied by the conditional choice probability and the transition probability. Let $h^* = \{h^*(x,a) : (x,a) \in X \times A, \inf_{x,a} h^*(x,a) > 0\}$ be a possible initial distribution of (x,a), and let $\mathcal{H}$ be the space of h^*. Let $h = \{h(a|x,a') : (a,x,a') \in A \times X \times A, 0 < h(a|x,a') < 1\}$ be a possible conditional choice probability, and let $\mathcal{P}$ be the space of h. Define the operator $\Phi : \mathcal{H} \times \mathcal{P} \to \mathcal{H}$ as (cf., equation (4))

$$\Phi(h^*; h)(x,a) = \int_{x' \in X} \sum_{a' \in A} h(a|x,a') f(x|x',a') h^*(x',a') dz.$$

Letting $\phi(h)$ denote the fixed point of $\Phi(\cdot; h)$, it follows that $p^{*m}(x_1, a_1) = \phi(P^m)(x_1, a_1)$, where $P^m = \{P^m(a|x,a') : (a,x,a') \in A \times X \times A\}$.

We approximate type-specific conditional choice probabilities by a series logit model. Specifically, we estimate $P^m(a|x,a')$ by

$$h_K(a = 1|x, a'; \gamma^{a'm}) = L(R^K(x)'\gamma^{a'm}) = L(R^K(x)'[a'\gamma^{1m} + (1-a')\gamma^{0m}]),$$

and $h_K(a = 0|x, a'; \gamma^{a'm}) = 1 - L(R^K(x)'\gamma^{a'm})$, where $L(z) = \exp(z)/(1+\exp(z))$ is the logistic cdf and $\gamma^{a'm}$ is a series coefficient vector for type m and $a' \in \{0,1\}$. Alternately, we may redefine x to include a' and approximate both $P^m(a|x,0)$ and $P^m(a|x,1)$ by a single series logit model. If the lagged choice is not a part of the state variable, we set $h_K(a = 1|x; \gamma^m) = L(R^K(x)'\gamma^m)$. Define $h_K(\gamma^{0m}, \gamma^{1m}) = \{h_K(a|x,a'; \gamma^{a'm}) : (a,x,a') \in A \times X \times A\}$, then the type-specific likelihood function of the i-th observation is approximated by

$$\ell(w_i; \gamma^{0m}, \gamma^{1m}) = \phi\left(h_K(\gamma^{0m}, \gamma^{1m})\right)(x_{i1}, a_{i1}) \prod_{t=2}^T f(x_{it}|x_{i,t-1}, a_{i,t-1}) h_K(a_{it}|x_{it}, a_{i,t-1}; \gamma^{a_{i,t-1}m}).$$

Define $\zeta = \{\pi^m, \gamma^{0m}, \gamma^{1m}\}_{m=1}^M$. We obtain $\hat{\zeta}_K = \{\hat{\pi}_K^m, \hat{\gamma}_K^{0m}, \hat{\gamma}_K^{1m}\}_{m=1}^M$ as

$$\hat{\zeta}_K = \arg\max_{\zeta \in \Theta_K} \mathcal{L}_N(\zeta),$$

where Θ_K is the space of admissible values of ζ specified later in Assumption 8, and $\mathcal{L}_N(\zeta)$ is the log-likelihood function of the finite mixture series logit model

$$\mathcal{L}_N(\zeta) = \frac{1}{N} \sum_{i=1}^N \ln \left(\sum_{m=1}^M \pi^m \ell(w_i; \gamma^{0m}, \gamma^{1m}) \right).$$

Then the series logit estimator of $P^m(a=1|x,a')$ is $\hat{P}^m(a=1|x,a') = L(R^K(x)'\hat{\gamma}_K^{a'm})$. Define the expectation of $\mathcal{L}_N(\zeta)$ as

$$Q(\zeta) = E\left[\ln\left(\sum_{m=1}^{M}\pi^m \ell(w_i;\gamma^{0m},\gamma^{1m})\right)\right].$$

We assume the following regularity conditions. Let $||\cdot||_\infty$ denote the sup norm, and let $p^*(x,a) = \sum_{m=1}^M \pi^m p^{*m}(x,a)$ denote the stationary distribution of (x,a).

Assumption 7 *(a) The support X of x is a compact subset of R^r. (b) $f(x'|x,a)$ is bounded. (c) $P^m(a|x,a')$ is s times continuously differentiable with respect to x with $s/r \geq 2$ for all m. (d) $\eta = \min_{1\leq m\leq M}\inf_{(x,a')\in X\times A} P^m(1|x,a')(1-P^m(1|x,a')) > 0$. (e) $p^*(x,a)$ is continuous, bounded and bounded away from zero. (f) There exists an integer $N \geq 1$ such that the operator $\Phi^N(\cdot;h)$ is a contraction with modulus $\rho < 1$ with respect to $||\cdot||_\infty$ for any $h \in \mathcal{P}$.*

Assumption 8 *(a) The parameter space is defined as $\Theta_K = \Pi \times \Gamma_K \times \Gamma_K$, where Π is a M-dimensional simplex and Γ_K satisfies*

$$\Gamma_K = \left\{\gamma \in R^K : \left|\inf_{x\in X} L(R^K(x)'\gamma)(1-L(R^K(x)'\gamma))\right| \geq \eta/2\right\}.$$

(b) $Q(\zeta)$ is continuous in $\zeta \in \Theta_K$ and uniquely maximized at ζ_K^*. (c) $Q(\zeta)$ is twice continuously differentiable in a neighborhood of ζ_K^* and $\frac{\partial^2}{\partial\zeta\partial\zeta'}Q(\zeta_K^*)$ is negative definite.
(d) $E|\ln(\sum_{m=1}^M \pi^m \ell(w_i;\gamma^{0m},\gamma^{1m}))| < \infty$ for all $\zeta \in \Theta_K$. (e) $\sup_{x\in X} E\left|\frac{\partial\phi(h^m(\zeta_K^*))(x_1,a_1)}{\partial h^m(a=1|x,a')}\right|^2 < \infty$ for $a' = 0,1$, where $h^m(\zeta_K^*)(a=1|x,a') = L(R^K(x)\gamma^{a'm*})$.

Assumption 7(c) on the smoothness of $P^m(a|x,a')$ can be relaxed by using the splines instead of polynomials (see Newey (1997)). Assumption 7(f) is often implicitly assumed in practice because the stationary distribution is often computed by iterating the operator $\Phi(\cdot;h)$ starting from an arbitrary initial guess until it converges. Assumption 8(a) implies that Θ_K is compact while Assumption 8(b)-(d) are the standard assumptions for consistency. In practice, if $\zeta \notin \Theta_K$, numerical evaluation of $\ln L(R^K(x_{it})'\gamma^{a'm})$ or $\ln(1-L(R^K(x_{it})'\gamma^{a'm}))$ becomes unstable because of the evaluation of $\ln(x)$ for $x \sim 0$. Optimization algorithms for $\mathcal{L}_N(\zeta)$ return error messages in these circumstances. Therefore, the value of ζ outside Θ_K is unlikely to be in the domain of optimization in practice. We can set $\Theta_K = R^K$ if the objective function is globally concave.

Hirano et al. (2003) show the convergence rate of the (non-mixture) series logit estimator. The following lemma extends Lemmas 1 and 2 of Hirano et al. (2003) to a finite mixture series logit estimator. Assumption 8 is stronger than the assumptions in Hirano et al. (2003) mainly because of the lack of the global concavity of our finite-mixture objective function. Let π_0^m denote the true value of the type probabilities.

Lemma 1 *Suppose Assumptions 7 and 8 hold and* $1/N + 1/K + K/N \to 0$. *Then, for* $m = 1, \ldots, M$,

$$\max_{a \in A} \sup_{x \in X} |\phi(\hat{P}^m)(x,a) - \phi(P^m)(x,a)| = O_p(K(K^{-s/(2r)} + \sqrt{K/N})),$$
$$\max_{a' \in A} \sup_{x \in X} |\hat{P}^m(a=1|x,a') - P^m(a=1|x,a')| = O_p(K(K^{-s/(2r)} + \sqrt{K/N})),$$
$$\hat{\pi}^m - \pi_0^m = O_p(K^{-s/(2r)} + \sqrt{K/N}).$$

Therefore, if $s/r \geq 3$ and we choose $K \sim N^\nu$ with $\nu \in (0, 1/3)$, the convergence rate of $\hat{P}^m(a|x,a') - P^m(a|x,a')$ becomes $N^{-\alpha}$ for some $\alpha > 0$. For example, if $s/r = 4$, then setting $K = N^{1/5}$ achieves the optimal convergence rate $N^{-1/5}$. We conjecture it is possible to extend our result to the case where the initial distribution is nonparametrically estimated instead of determined as the fixed point $\phi(\hat{P}^m)$. However, the convergence rate may become slower than the one provided in the lemma.

This result allows us to apply various computationally attractive semiparametric estimators for structural dynamic models listed in the introduction to the models with unobserved heterogeneity. Kasahara and Shimotsu (2006) provides an example of such an application.[11]

Assumption 8(b)(c) imply that the parameter of the series models are uniquely identified. In the following proposition, we provide sufficient conditions for Assumption 8(b)(c) in terms of the primitive condition on nonparametric identifiability of the conditional choice probability. Let $h^m = \{h^m(a|x,a') : (a,x,a') \in A \times X \times A\}$ denote a (generic) conditional choice probability for type m, and consider $\vartheta = \{\pi^m, h^m\}_{m=1}^M$ as an infinite-dimensional parameter. Let $\vartheta^0 = \{\pi_0^m, P^m\}_{m=1}^M$ denote the true parameter value of ϑ. Define the space of ϑ as $\bar{\Theta} = \Delta^M \times \mathcal{P}^M$, where Δ^M is a M-dimensional simplex. Define the expectation of the likelihood function as a function of ϑ as

$$\tilde{Q}(\vartheta) = E\left[\ln\left(\sum_{m=1}^M \pi^m \tilde{\ell}(w_i; h^m)\right)\right], \tag{25}$$

where $\tilde{\ell}(w_i; h^m)$ is the type-specific likelihood function of the i-th observation

$$\tilde{\ell}(w_i; h^m) = \phi(h^m)(x_{i1}, a_{i1}) \prod_{t=2}^T f(x_{it}|x_{i,t-1}, a_{i,t-1}) h^m(a_{it}|x_{it}, a_{i,t-1}).$$

Proposition 11 *Suppose (a)* $\tilde{Q}(\vartheta)$ *is continuous, (b) for any* $\varepsilon > 0$, $\tilde{Q}(\vartheta^0) > \sup_{\vartheta \in \bar{\Theta} \setminus \mathcal{N}_\varepsilon} \tilde{Q}(\vartheta)$, *where* $\mathcal{N}_\varepsilon = \{\vartheta : ||\vartheta - \vartheta^0||_\infty \leq \varepsilon\}$, *and (c)* $\tilde{Q}(\vartheta)$ *is twice continuously Fréchet differentiable and its second-order Fréchet derivative* $D^2\tilde{Q}(\vartheta)$ *satisfies* $D^2\tilde{Q}(\vartheta^0)[\varrho, \varrho] \leq -\lambda < 0$ *for any* $\varrho \neq 0$.[12]

[11] Kasahara and Shimotsu (2006) show that, in structural discrete Markov decision models with unobserved heterogeneity, it is possible to obtain an estimator that is higher-order equivalent to the MLE by iterating the nested pseudo-likelihood (NPL) algorithm of Aguirregabiria and Mira (2002) sufficiently many, but finite times.

[12] Fréchet derivatives are the derivatives defined for mappings from one Banach space X to another Banach

Then Assumption 8(b) and 8(c) hold for sufficiently large K.

5 Monte Carlo Experiments

In order to assess the performance of our series estimators, we use a version of Rust's celebrated bus engine replacement model. The reader is also referred to Rust (1987). Example 2 in this paper provides a brief description of the dynamic discrete choice. We consider the specification with a linear cost function

$$u(x,a;\theta) = \begin{cases} -\alpha_0 & \text{for } a=1, \\ -0.01 \cdot \alpha_1 x & \text{for } a=0. \end{cases}$$

We use the transition function of x in Example 5. We assume that the transition function is known and common across types, where $(\theta_{f,1}, \theta_{f,2})$ is set to $(0.3, 0.3)$ unless stated otherwise. There are M types of buses, where type m is characterized by a type specific parameter $\theta^m = (\alpha_0^m, \alpha_1^m)'$, and the probability of being type m in the population is π^m for $m = 1, 2, \ldots, M$. Define $\pi = (\pi^1, \pi^2, \ldots, \pi^{M-1})'$. Let $\zeta = (\pi', \theta^{1'}, \ldots, \theta^{M'})'$ be the parameter to be estimated.

Let $\{\{a_{it}, x_{it}\}_{t=1}^T\}_{i=1}^N$ be a panel data set such that $w_i = \{a_{it}, x_{it}\}_{t=1}^T$ is randomly drawn across i's from the population. Conditional on being type m, the likelihood of observing w_i is

$$\begin{aligned} \ell(w_i;\theta^m) &= p^*(x_{i1};\theta^m) P_{\theta^m}(a_{i1}|x_{i1}) \prod_{t=2}^T P_{\theta^m}(a_{it}|x_{it}), && (26) \\ \text{s.t.} \quad & P_{\theta^m} = \Psi(P_{\theta^m}, \theta^m), && (27) \\ & p^*(x;\theta^m) = \sum_{x'=1}^{|X|} \sum_{a'=0}^{1} P_{\theta^m}(a'|x') f(x|x',a') p^*(x';\theta^m), && (28) \end{aligned}$$

where P_{θ^m} is the fixed point of $\Psi(\cdot, \theta^m)$ as defined by (27). $p^*(x;\theta^m)$ is the distribution (density) function of the initial observation x_{i1} for type m, which is specified as the stationary distribution of x for type m. The ML estimator of ζ is then defined as

$$\hat{\zeta} = \arg\max_{\zeta \in \Theta_K} \frac{1}{N} \sum_{i=1}^N \ln \left(\sum_{m=1}^M \pi^m \ell(w_i;\theta^m) \right).$$

An $N \times T$ panel dataset is generated as follows. For each individual observation, its type is drawn from a multinomial distribution. Given the realized type, say type m, its initial observation is drawn from the stationary distribution of type m. Then observations of replacement

space Y. When X is a Euclidean space, the Fréchet derivatives coincide with the standard derivatives. Concepts such as the chain rule, product rule, higher-order and partial derivatives, Taylor expansion, and implicit function theorem are defined analogously to the corresponding concepts defined for the functions in Euclidean spaces; see Griffel (1985) and Zeidler (1986) for further details.

decisions are drawn using the choice probabilities obtained by numerically solving the Bellman's equation of type m. Subsequent observations of x's are drawn using the transition function.

The model is estimated using the series logit estimator discussed in section 4. The reported results are based on 100 simulated samples. We first consider the case of two types with the model parameters set to $(\pi^1, \pi^2) = (0.5, 0.5)$, $\alpha^1 = (10, 10)$, and $\alpha^2 = (2, 2)$. Table 1 presents the squared bias, variance, the root (integrated) mean squared error (RMSE) of the estimated conditional choice probabilities, and type probabilities for various degrees of the polynomials in basis functions. It shows a clear pattern for $\hat{P}^m$ that, as the degrees of the polynomials increases, the bias of $\hat{P}^m$ decreases, while its variance increases. Thus, as we expect from Lemma 1, there is a trade-off between bias and variance as we change the degrees of the polynomials. With the sample size $(N, T) = (500, 10)$, the lowest values of the RMSE for $\hat{P}^m$ is achieved with quartic polynomials; when the sample size increases to $(N, T) = (2000, 10)$, the RMSE for $\hat{P}^m$ is lowest at the higher 5 degrees of polynomials.

The last two rows of Table 1 report the frequency at which different degrees of polynomials are chosen by Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC).[13] AIC performs better than BIC in terms of the RMSE; AIC tends to choose lower degrees of polynomials than that which achieves the lowest RMSE of $\hat{P}^m$, but AIC has smaller "bias" in selecting the degrees of polynomials than BIC.

Table 2 compares the performance (bias, variance and RMSE) of the parametric maximum likelihood estimator (MLE) and that of the series logit estimator with a cubic polynomial when the model with two types is estimated. As a benchmark, we also report the results of the MLE under complete data (i.e., in case that the types are completely observed). The table indicates that the series approximation is quite good when the sample size is large, although the bias persists even at $(N, T) = (2000, 10)$. When the sample size is as small as $(N, T) = (500, 3)$, the RMSE is large for the series estimator, but even larger for the parametric MLE. Overall, the series estimator performs comparably to the MLE in terms of the accuracy of the conditional choice probability estimates.

Tables 3 and 4 present the results when three-type mixture models are estimated with the model parameters set to $(\pi^1, \pi^2, \pi^3) = (1/3, 1/3, 1/3)$, $\alpha^1 = (15, 15)$, $\alpha^2 = (1, 1)$, and $\alpha^3 = (4, 4)$. Again, the performance of the series estimators is comparable to that of MLE. Increasing the number of types from two to three leads to higher RMSE, and especially higher variance, as noted from comparison of Tables 3 and 4 with Table 2. As the number of types increases while fixing the degree of the polynomials, the total number of parameters in series logit functions to be estimated also increases, leading to higher variance.

Table 5 reports the performance of the series logit estimator and MLE when transition functions are type-specific and, therefore, are also estimated. The parameters for transition

[13] Stone (1977) shows the asymptotic equivalence of model selection by cross-validation and AIC in maximum likelihood estimation.

functions are set to $(\theta^1_{f,1}, \theta^1_{f,2}) = (0.4, 0.4)$ and $(\theta^2_{f,1}, \theta^2_{f,2}) = (0.2, 0.2)$. Other parameter values are the same as those of Table 2. The performance of the series estimators is as good as that of MLE even when type-specific transition functions are estimated simultaneously. In fact, comparing Table 5 with Table 2, we notice that the RMSEs reported in Table 5 are smaller than those reported in Table 2 for the same sample size. This suggests a possibility that the presence of type-specific functions provides an extra source of identification for different types.

We use Chebyshev polynomials as basis functions. All the starting values for estimating the mixture model with the incomplete data are from the estimates under the complete data where the parameters are estimated for observations with known types. This appears to increase numerical stability although there is a possibility of obtaining local maximum near the starting values. The mixture model is estimated by maximizing the mixture of likelihood by first using Nelder-Mead simplex method to obtain an estimate in the neighborhood of the optimum and then using BFGS to further refine the estimate. We have also experimented with the EM algorithm in place of Nelder-Mead simplex method and found that the simulation results are very similar as long as we use good starting values obtained from estimating the model under the complete data. However, computation time is substantially larger if we use the EM algorithm.

6 Appendix

6.1 Proof of Proposition 1 and Corollary 1

Define $V = \text{diag}(\pi^1, \ldots, \pi^M)$ and $D_k = \text{diag}(\lambda^{*1}_k, \ldots, \lambda^{*M}_k)$. Define P and P_k as in (16). Then P and P_k are expressed as (see (13)-(15))

$$P = L'VL, \quad P_k = L'VD_kL.$$

We now uniquely determine L, V, and D_k from P and P_k constructively. Since L is non-singular, we can construct a matrix $A_k = P^{-1}P_k = L^{-1}D_kL$. Because $A_kL^{-1} = L^{-1}D_k$, the eigenvalues of A_k determine the diagonal elements of D_k while the right-eigenvectors of A_k determine the columns of L^{-1} up to multiplicative constants; denote the right-eigenvectors of A_k by $L^{-1}K$ where K is some diagonal matrix. Now we can determine VK from the first row of $PL^{-1}K$ because $PL^{-1}K = L'VK$ and the first row of L' is a vector of ones. Then L' is determined uniquely by $L' = (PL^{-1}K)(VK)^{-1} = (L'VK)(VK)^{-1}$. Having obtained L', we may determine V from the first column of $(L')^{-1}P$ because $(L')^{-1}P = VL$ and the first column of L is a vector of ones. Therefore, we identify $\{\pi^m, \{\lambda^m_{\xi_j}\}_{j=1}^{M-1}\}_{m=1}^{M}$ as the elements of V and L.

Once V and L are determined, we can uniquely determine $D_\zeta = diag(\lambda^{*1}_\zeta, \ldots, \lambda^{*M}_\zeta)$ for any $\zeta \in X$ by constructing P_ζ in the same way as P_k and using the relationship $D_\zeta = (L'V)^{-1}P_\zeta L^{-1}$. Furthermore, for arbitrary $\zeta, \xi_j \in X$, evaluate F_{x_2,x_3}, F_{x_2}, and F_{x_3} defined in (14) and (15) at

$(x_2, x_3) = (\zeta, \xi_j)$, and define

$$\underset{(M\times 2)}{L^\zeta} = \begin{bmatrix} 1 & \lambda_\zeta^1 \\ \vdots & \vdots \\ 1 & \lambda_\zeta^M \end{bmatrix}, \quad \underset{(2\times M)}{P^\zeta} = \begin{bmatrix} 1 & F_{\xi_1} & \dots & F_{\xi_{M-1}} \\ F_\zeta & F_{\zeta,\xi_1} & \dots & F_{\zeta,\xi_{M-1}} \end{bmatrix}. \tag{29}$$

Since $P^\zeta = (L^\zeta)' V L$, we can uniquely determine $(L^\zeta)' = P^\zeta (VL)^{-1}$. Therefore, $\{\lambda_\zeta^{*m}\}_{m=1}^M$ and $\{\lambda_\zeta^m\}_{m=1}^M$ are identified for any $\zeta \in X$. This completes the proof of Proposition 1, and Corollary 1 follows immediately. □

6.2 Proof of Proposition 2

The proof is similar to the proof of Proposition 1. Let $\mathcal{T} = (\tau_2, \dots, \tau_p)$, $2 \le p \le T$, be a subset of $\{2, \dots, T\}$. Let $\mathcal{X}(\mathcal{T})$ be a subset of $\{x_t\}_{t=2}^T$ with $t \in \mathcal{T}$. For example, if $\mathcal{T} = \{2, 4, 6\}$, then $\mathcal{X}(\mathcal{T}) = \{x_2, x_4, x_6\}$. Starting from $\tilde{P}(\{a_t, x_t\}_{t=1}^T)$, integrating out (a_t, x_t) if $t \notin \mathcal{T}$ and evaluating it at $(a_1, x_1) = (1, k)$ and $a_t = 1$ for $t \in \mathcal{T}$ gives a "marginal" $F^*_{k,\mathcal{X}(\mathcal{T})} = \tilde{P}\left(\{a_1, x_1\} = \{1, k\}, \{1, x_t\}_{\tau\in\mathcal{T}}\right) = \sum_{m=1}^M \pi^m \lambda_k^{*m} \prod_{t\in\mathcal{T}} \lambda_{x_t}^m$. For example, if $\mathcal{T} = \{2, 4, 6\}$, then $F^*_{k,\mathcal{X}(\mathcal{T})} = \sum_{m=1}^M \pi^m \lambda_k^{*m} \lambda_{x_2}^m \lambda_{x_4}^m \lambda_{x_6}^m$. Integrating out (a_1, x_1) additionally and proceeding in a similar way gives $F_{\mathcal{X}(\mathcal{T})} = \tilde{P}(\{1, x_t\}_{\tau\in\mathcal{T}}) = \sum_{m=1}^M \pi^m \prod_{t\in\mathcal{T}} \lambda_{x_t}^m$.

Define $V = \mathrm{diag}(\pi^1, \dots, \pi^M)$ and $D_k = \mathrm{diag}(\lambda_k^{*1}, \dots, \lambda_k^{*M})$. Define $P^\diamond = (L^\diamond)' V L^\diamond$ and $P_k^\diamond = (L^\diamond)' V D_k L^\diamond$, then the elements of $P^\diamond$ take the form $\sum_{m=1}^M \pi^m \prod_{t\in\mathcal{T}} \lambda_{x_t}^m$ and can be expressed as $F_{\mathcal{X}(\mathcal{T})}$ for some $\mathcal{T}$ and $\{x_t\}_{t\in\mathcal{T}} \in X^{|\mathcal{T}|}$. Similarly, the elements of $P_k^\diamond$ can be expressed as $F^*_{k,\mathcal{X}(\mathcal{T})}$. For instance, if $r = 3$, $T = 7$, and both Λ and $L^\diamond$ are $M \times M$, then $P^\diamond$ is given by

$$\begin{bmatrix} 1 & F_1 & \cdots & F_{|X|} & F_{11} & \cdots & F_{|X||X|} & F_{111} & \cdots & F_{|X||X||X|} \\ F_1 & & & & & & & & & \\ \vdots & & & & & & & & & \\ F_{|X|} & & & & F_{|X|11} & & & & & F_{|X||X||X||X|} \\ F_{11} & & & & & & & & & \\ \vdots & & & & & \ddots & & & & \vdots \\ F_{|X||X|} & & & & & & & & & \\ F_{111} & & & & & & & & & F_{111|X||X||X|} \\ \vdots & & & & & & & & & \\ F_{|X||X||X|} & & & & F_{|X||X||X|11} & \cdots & & & & F_{|X||X||X||X||X||X|} \end{bmatrix},$$

where the (i, j)th element of $P^\diamond$ is F_σ, where σ consists of the combined subscripts of the $(i, 1)$th and $(1, j)$th element of $P^\diamond$. For example, the $(|X| + 1, 2)$th element of $P^\diamond$ is $F_{|X|1} (= F_{1|X|})$. $P_k^\diamond$ is given by replacing F_σ in $P^\diamond$ with $F^*_{k,\sigma}$ and setting the $(1, 1)$th element to F_k^*.

Consequently, $P^{\diamond}$ and $P_k^{\diamond}$ can be computed from the distribution function of the observed data. By repeating the argument of the proof of Proposition 1, we determine $L^{\diamond}$, V, and D_k uniquely from $P^{\diamond}$ and $P_k^{\diamond}$ first, and then $D_\zeta = \mathrm{diag}(\lambda_\zeta^{*1}, \ldots, \lambda_\zeta^{*M})$ and L^ζ for any $\zeta \in X$ from $P^{\diamond}$, $P_\zeta^{\diamond}$, $L^{\diamond}$, and P^ζ, where L^ζ and P^ζ are defined in (29). □

6.3 Proof of Proposition 3

The proof is similar to the proof of Proposition 1. Define P_t and $P_{t,k}$ analogously to P and P_k but with λ_{x_2} and λ_{x_3} replaced with λ_{t,x_t} and $\lambda_{t+1,x_{t+1}}$ in the definition of $F_{\cdot}$'s and $F_{\cdot}^{*}$'s. Define V and D_k as before. Then P_t and $P_{t,k}$ are expressed as $P_t = L_t' V L_{t+1}$ and $P_{t,k} = L_t' V D_k L_{t+1}$. Since L_t and L_{t+1} are nonsingular, we have $A_k = P_t^{-1} P_{t,k} = L_{t+1}^{-1} D_k L_{t+1}$. Because $A_k L_{t+1}^{-1} = L_{t+1}^{-1} D_k$, the eigenvalues of A_k determine the diagonal elements of D_k while the right-eigenvectors of A_k determine the columns of L_{t+1}^{-1} up to multiplicative constants; denote the right-eigenvectors of A_k by $L_{t+1}^{-1}K$ where K is some diagonal matrix. Now we can determine VK from the first row of $P_t L_{t+1}^{-1} K$ because $P_t L_{t+1}^{-1} K = L_t' V K$ and the first row of L_t' is a vector of ones. Then L_t' is determined uniquely by $L_t' = (L_t' V K)(VK)^{-1}$. Having obtained L_t', we may determine V and L_{t+1} from $V L_{t+1} = (L_t')^{-1} P$ because the first column of $V L_{t+1}$ equals the diagonal of V and $L_{t+1} = V^{-1}(V L_{t+1})$. Therefore, we determine $\{\pi^m, \{\lambda_{t,\xi_j^t}^m, \lambda_{t+1,\xi_j^{t+1}}^m\}_{j=1}^{M-1}\}_{m=1}^{M}$ as elements of V, L_t, and L_{t+1}. Once V, L_t and L_{t+1} are determined, we can uniquely determine $D_\zeta = diag(\lambda_\zeta^{*1}, \ldots, \lambda_\zeta^{*M})$ for any $\zeta \in X$ by constructing $P_{t,\zeta}$ in the same way as $P_{t,k}$ and using the relationship $D_\zeta = (L_t' V)^{-1} P_{t,\zeta} (L_{t+1})^{-1}$. Furthermore, for arbitrary $\zeta \in X$, define

$$\underset{(M\times 2)}{L_t^\zeta} = \begin{bmatrix} 1 & \lambda_{t,\zeta}^1 \\ \vdots & \vdots \\ 1 & \lambda_{t,\zeta}^M \end{bmatrix}.$$

Then $P_t^\zeta = (L_t^\zeta)' V L_{t+1}$ is a function of the distribution function of the observable data, and we can uniquely determine $(L_t^\zeta)'$ for $2 \le t \le T-1$ as $P_t^\zeta (V L_{t+1})^{-1}$. For $t = T$, we can use the fact that $(L_{T-1})' V L_T^\zeta$ is also a function of the distribution function of the observable data and proceed in the same manner. Therefore, we can determine $\{\lambda_\zeta^{*m}, \lambda_{t,\zeta}^m\}_{j=1}^{M-1}$ for any $\zeta \in X$ and $2 \le t \le T$. □

6.4 Proof of Proposition 5

Integrating out s_t's backwards from $P(\{s_t\}_{t=1}^6)$ and fixing $s_1 = s_3 = s_5 = \bar{s}$ gives the following "marginals" :

$$\begin{aligned}
\tilde{F}^*_{s_2,s_4,s_6} &= \sum_{m=1}^M \tilde{\pi}^m_{\bar{s}} \gamma^m_{\bar{s}}(s_2)\gamma^m_{\bar{s}}(s_4)Q^m(s_6|\bar{s}), \quad \tilde{F}^*_{s_2,s_6} = \sum_{m=1}^M \tilde{\pi}^m_{\bar{s}} \gamma^m_{\bar{s}}(s_2)Q^m(s_6|\bar{s}), \quad \tilde{F}^*_{s_6} = \sum_{m=1}^M \tilde{\pi}^m_{\bar{s}} Q^m(s_6|\bar{s}), \\
\tilde{F}_{s_2,s_4} &= \sum_{m=1}^M \tilde{\pi}^m_{\bar{s}} \gamma^m_{\bar{s}}(s_2)\gamma^m_{\bar{s}}(s_4), \quad \tilde{F}_{s_2} = \sum_{m=1}^M \tilde{\pi}^m_{\bar{s}} \gamma^m_{\bar{s}}(s_2), \quad \tilde{F} = \sum_{m=1}^M \tilde{\pi}^m_{\bar{s}}.
\end{aligned}$$

As in the proof of Proposition 1, evaluate these $\tilde{F}_\cdot$'s at $s_2 = \xi_1, \ldots, \xi_{M-1}$, $s_4 = \xi_1, \ldots, \xi_{M-1}$, and $s_6 = r$, and arrange them into two $M \times M$ matrices

$$\tilde{P} = \begin{bmatrix} \tilde{F} & \tilde{F}_{\xi_1} & \cdots & \tilde{F}_{\xi_{M-1}} \\ \tilde{F}_{\xi_1} & \tilde{F}_{\xi_1,\xi_1} & \cdots & \tilde{F}_{\xi_1,\xi_{M-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{F}_{\xi_{M-1}} & \tilde{F}_{\xi_{M-1},\xi_1} & \cdots & \tilde{F}_{\xi_{M-1},\xi_{M-1}} \end{bmatrix}, \quad \tilde{P}_r = \begin{bmatrix} \tilde{F}^*_r & \tilde{F}^*_{\xi_1,r} & \cdots & \tilde{F}^*_{\xi_{M-1},r} \\ \tilde{F}^*_{\xi_1,r} & \tilde{F}^*_{\xi_1,\xi_1,r} & \cdots & \tilde{F}^*_{\xi_1,\xi_{M-1},r} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{F}^*_{\xi_{M-1},r} & \tilde{F}^*_{\xi_{M-1},\xi_1,r} & \cdots & \tilde{F}^*_{\xi_{M-1},\xi_{M-1},r} \end{bmatrix}.$$

Define $\tilde{V}_{\bar{s}} = \mathrm{diag}(\tilde{\pi}^1_{\bar{s}}, \ldots, \tilde{\pi}^M_{\bar{s}})$ and $\tilde{D}_{r|\bar{s}} = \mathrm{diag}(Q^1(r|\bar{s}), \ldots, Q^M(r|\bar{s}))$. Then $\tilde{P}$ and $\tilde{P}_r$ are expressed as $\tilde{P} = G'_{\bar{s}}\tilde{V}_{\bar{s}}G_{\bar{s}}$ and $\tilde{P}_r = G'_{\bar{s}}\tilde{V}_{\bar{s}}\tilde{D}_{r|\bar{s}}G_{\bar{s}}$. Repeating the argument of the proof of Proposition 1 shows that $G_{\bar{s}}$, $G_{\bar{s}}$, $\tilde{V}_{\bar{s}}$, and $\tilde{D}_{r|\bar{s}}$ are uniquely determined from $\tilde{P}$ and $\tilde{P}_r$, and that $\tilde{D}_{s|\bar{s}}$ and $\gamma^m_{\bar{s}}(s)$ can be uniquely determined for any $s \in S$ and $m = 1, \ldots, M$. □

6.5 Proof of Proposition 6

Define $\tilde{V}_{\bar{s}} = \mathrm{diag}(\tilde{\pi}^1_{\bar{s}}, \ldots, \tilde{\pi}^M_{\bar{s}})$ and $\tilde{D}_{r|\bar{s}} = \mathrm{diag}(Q^1(r|\bar{s}), \ldots, Q^M(r|\bar{s}))$. Applying the argument of the proof of Proposition 5 with $G_{\bar{s}}$ replaced by $G^\diamond_{\bar{s}}$, we can identify $G^\diamond_{\bar{s}}$, $\tilde{V}_{\bar{s}}$, and $\tilde{D}_{r|\bar{s}}$, and then $\tilde{D}_{s|\bar{s}}$ and $\gamma^m_{\bar{s}}(s)$ for any $s \in S$ and $m = 1, \ldots, M$. The stated result immediately follows. □

6.6 Proof of Proposition 7

The proof uses the logic of the proof of Proposition 5. Consider a sequence $\{s_t, z_t\}_{t=1}^4$ with $(s_1, s_2, s_3, s_4) = (\bar{s}, \bar{s}, \bar{s}, r)$ and $(z_1, z_4) = (h, k)$. Summarize the value of s_4 and z_4 into $\zeta = (r, k)$. For $(z_2, z_3) \in Z^2$, define $\tilde{F}^{h*}_{z_2,z_3,\zeta} = \sum_{m=1}^M \tilde{\pi}^m_{\bar{s},h}\tilde{\gamma}^m_{\bar{s}}(z_2)\tilde{\gamma}^m_{\bar{s}}(z_3)\tilde{Q}^m(r|\bar{s},k)$ and $\tilde{F}^h_{z_2,z_3} = \sum_{m=1}^M \tilde{\pi}^m_{\bar{s},h}\tilde{\gamma}^m_{\bar{s}}(z_2)\tilde{\gamma}^m_{\bar{s}}(z_3)$. Define $\tilde{F}^{h*}_{z_2,\zeta} = \sum_{m=1}^M \tilde{\pi}^m_{\bar{s},h}\tilde{\gamma}^m_{\bar{s}}(z_2)\tilde{Q}^m(r|\bar{s},k)$, and define $\tilde{F}^{h*}_\zeta$, $\tilde{F}^h_{z_2}$, and $\tilde{F}^h$ analogously to the proof of Proposition 5.

As in the proof of Proposition 5, arrange these marginals into two matrices $\bar{P}^h$ and $\bar{P}^h_\zeta$. $\bar{P}^h$ and $\bar{P}^h_\zeta$ are the same as $\tilde{P}$ and $\tilde{P}_r$, but $\tilde{F}_\cdot$ and $\tilde{F}^*_{\cdot,r}$ replaced with $\tilde{F}^h_\cdot$ and $\tilde{F}^{h*}_{\cdot,\zeta}$ and subscripts are elements of Z instead of those of S. Define $\tilde{V}^h_{\bar{s}} = \mathrm{diag}(\tilde{\pi}^1_{\bar{s},h}, \ldots, \tilde{\pi}^M_{\bar{s},h})$ and $\tilde{D}_{\zeta|\bar{s}} = \mathrm{diag}(\tilde{Q}^1(r|\bar{s},k), \ldots, \tilde{Q}^M(r|\bar{s},k))$. It then follows that $\bar{P}^h = \bar{G}'_{\bar{s}}\tilde{V}^h_{\bar{s}}\bar{G}_{\bar{s}}$ and $\bar{P}^h_\zeta = \bar{G}'_{\bar{s}}\tilde{V}^h_{\bar{s}}\tilde{D}_{\zeta|\bar{s}}\bar{G}_{\bar{s}}$. By

repeating the argument of the proof of Proposition 1, we can uniquely determine $\bar{G}_{\bar{s}}$, $\tilde{V}^h_{\bar{s}}$, and $\tilde{D}_{\zeta|\bar{s}}$ from $\bar{P}^h$ and $\bar{P}^h_\zeta$, and, having determined $\bar{G}_{\bar{s}}$, determine $\tilde{D}_{(s,z)|\bar{s}}$ for any $(s,z) \in S \times Z$. □

6.7 Proof of Proposition 8

For $(x_2, x_3) \in B^h \times C^h$ and $x_c \in B^h \cup C^h$, define $F^{h*}_{x_2,x_3,k} = \sum_{m=1}^M \tilde{\pi}^m_h \lambda^m_{x_2} \lambda^m_{x_3} \lambda^m_k$, $F^{h*}_{x_c,k} = \sum_{m=1}^M \tilde{\pi}^m_h \lambda^m_{x_c} \lambda^m_k$, $F^{h*}_k = \sum_{m=1}^M \tilde{\pi}^m_h \lambda^m_k$, $F^h_{x_2,x_3} = \sum_{m=1}^M \tilde{\pi}^m_h \lambda^m_{x_2} \lambda^m_{x_3}$, $F^h_{x_c} = \sum_{m=1}^M \tilde{\pi}^m_h \lambda^m_{x_c}$, and $F^h = \sum_{m=1}^M \tilde{\pi}^m_h$. They can be constructed from sequentially integrating out $P(\{a_t, x_t\}_{t=1}^4)$ backwards and then dividing them by a product of $f(x_t|x_{t-1}, 0)$. Note that Assumption 4(b) guarantees $f(x_t|x_{t-1}, 0) > 0$ for all x_t and x_{t-1} in the subsets of X considered.

As in the proof of Proposition 1, arrange these "marginals" into two matrices P^h and P^h_k. P^h and P^h_k are the same as P and P_k but $F_{\cdot}$ and $F^*_{k,\cdot}$ replaced with $F^h_{\cdot}$ and $F^{h*}_{\cdot,k}$. Define $V_h = \text{diag}(\pi^1_h, \ldots, \pi^M_h)$ and $D_k = \text{diag}(\lambda^1_k, \ldots, \lambda^M_k)$. By applying the argument in the proof of Proposition 3, we may show that $\tilde{G}_1$, $\tilde{G}_2$, V_h, and D_k are uniquely determined from $\tilde{P}(\{a_t, x_t\}_{t=1}^4)$ and its marginals and then show that $\{\lambda^m_\xi\}_{m=1}^M$ is determined for $\xi \in B^h \cup C^h$. □

6.8 Proof of Lemma 1

In the following, C denotes a generic positive and finite constant which may take different values in different places. As in Hirano et al. (2003), H03 henceforth, we use the matrix norm $||A|| = (\text{tr}(A'A))^{1/2}$. This norm satisfies the Cauchy-Schwartz inequality $||A'B|| \leq ||A||||B||$. To simplify the notation, we assume that the lagged choice is not a part of the state variable. Thus the conditional choice probability is $P^m(a|x)$ and its estimate is $h_K(a = 1|x; \gamma^m) = L(R^K(x)'\gamma^m)$. Including the lagged choice does not affect the argument of the proof, apart from additional notational complexity. We further assume that Assumption 7(f) holds with $N = 1$ but an analogous argument applies when $N > 1$.

First, we derive a bound of the estimation error of the initial distribution; for any $h^m, \tilde{h}^m \in \mathcal{P}$,

$$||\phi(\tilde{h}^m) - \phi(h^m)||_\infty \leq C(1-\rho)^{-1} \sup_{(a,x)} |\tilde{h}^m(a|x) - h^m(a|x)|. \tag{30}$$

To show (30), we first note that the triangular inequality gives

$$||\phi(\tilde{h}^m) - \phi(h^m)||_\infty \leq ||\Phi(\phi(\tilde{h}^m); \tilde{h}^m) - \Phi(\phi(h^m); \tilde{h}^m)||_\infty + ||\Phi(\phi(h^m); \tilde{h}^m) - \Phi(\phi(h^m); h^m)||_\infty.$$

The first term on the right is no larger than $\rho||\phi(\tilde{h}^m) - \phi(h^m)||_\infty$ by the contraction property of Φ. The second term on the right is no larger than $||\int_{x'} \sum_{a'} (\tilde{h}^m(a|x) - h^m(a|x)) f(x|x', a') \phi(h^m)(x', a') dx'||_\infty$, and (30) follows. Thus, the bound of $||\phi(\hat{P}^m) - \phi(P^m)||_\infty$ in the Lemma follows from that of $|\hat{P}^m(a|x) - P^m(a|x)|$.

The rest of the proof is devoted to deriving the bound of $|\hat{P}^m(a|x) - P^m(a|x)|$. It uses some of the arguments of H03. $p^*(x)$ in H03 corresponds to our $P^m(a|x)$. Following the argument of

H03 pp. 1177-78, leading to their equation (29), we find that there is a γ_K^m such that

$$\sup_{x\in X}\left|P^m(a=1|x) - L(R^K(x)'\gamma_K^m)\right| \leq CK^{-s/r}, \quad m=1,\ldots,M. \tag{31}$$

Hence, it follows from the triangular inequality that, uniformly in x,

$$|\hat{P}^m(a=1|x) - P^m(a=1|x)| \leq |L(R^K(x)'\hat{\gamma}_K^m) - L(R^K(x)'\gamma_K^m)| + CK^{-s/r}. \tag{32}$$

The first term on the right is bounded by, with $\bar{\gamma} \in [\hat{\gamma}_K^m, \gamma_K^m]$,

$$\left|L'(R^K(x)'\bar{\gamma})R^K(x)'(\hat{\gamma}_K^m - \gamma_K^m)\right| \leq C\sup_{x\in X}||R^K(x)|| \cdot ||\hat{\gamma}_K^m - \gamma_K^m|| \leq CK||\hat{\gamma}_K^m - \gamma_K^m||, \tag{33}$$

where the first inequality follow from $\sup_{\gamma\in\Gamma_k}|L'(R^K(x)'\gamma)| < \infty$ by Assumption 8(a), and the second inequality follows from equation (21) of H03. Define $\zeta_K = \{\pi_0^m, \gamma_K^m\}_{m=1}^M$. Then, from (32)-(33) and $||\hat{\gamma}_K^m - \gamma_K^m|| \leq ||\hat{\gamma}_K^m - \gamma_K^{*m}|| + ||\gamma_K^{*m} - \gamma_K^m||$, the results in the Lemma hold if we show there exists a finite constant C_1 such that

$$(a)\ ||\zeta_K - \zeta_K^*|| \leq C_1K^{-s/(2r)} \quad \text{and} \quad (b)\ ||\zeta_K^* - \hat{\zeta}_K|| = O_p(\sqrt{K/N}). \tag{34}$$

We show (34)(a) first. Prior to showing (34)(a), we first need to show $||\zeta_K - \zeta_K^*|| < \eta$ for any $\eta > 0$ and for sufficiently large K. Define ϑ and $\tilde{Q}(\vartheta)$ as in (25), then the information inequality implies $\tilde{Q}(\vartheta^0) \geq \tilde{Q}(\vartheta)$ for any $\vartheta \in \bar{\Theta}$. In particular, since $L(R^K(x)'\gamma_K^m), L(R^K(x)'\gamma_K^{*m}) \in \mathcal{P}$, we have $\tilde{Q}(\vartheta^0) \geq Q(\zeta_K^*) \geq Q(\zeta_K)$. On the other hand, (31) implies $Q(\zeta_K) \geq \tilde{Q}(\vartheta^0) - C_2K^{-s/r}$ for a finite and positive constant C_2. Consequently,

$$Q(\zeta_K) \geq Q(\zeta_K^*) - C_2K^{-s/r}. \tag{35}$$

Let $\mathcal{N}_\eta = \{\zeta : ||\zeta - \zeta_K^*|| < \eta\}$ and suppose $\zeta_K \notin \mathcal{N}_\eta$. Since $Q(\zeta)$ is continuous and Θ_K is compact, we have $\sup_{\zeta\in\Theta_k\setminus\mathcal{N}_\eta} Q(\zeta) < Q(\zeta_K^*)$ (cf., Newey and McFadden (1994), Theorem 2.1) and $Q(\zeta_K^*) - Q(\zeta_K) > \varepsilon$ for some $\varepsilon > 0$. However, (35) implies $Q(\zeta_K^*) - Q(\zeta_K) \leq \varepsilon/2$ for sufficiently large K, contradicting $\zeta_K \notin \mathcal{N}_\eta$. It follows that $||\zeta_K - \zeta_K^*|| < \eta$.

Having established $||\zeta_K - \zeta_K^*|| < \eta$, we proceed to show (34)(a). Let $\lambda_{\min}(A)$ denote the smallest eigenvalue of a matrix A. Assumption 8(c) implies that, if we take η sufficiently small, $\lambda_\eta = \inf_{\zeta\in\mathcal{N}_\eta}\lambda_{\min}[-(\partial^2/\partial\zeta\partial\zeta')Q(\zeta)] > 0$. Choose C_1 so that $\lambda_\eta C_1^2 \geq 4C_2$. Suppose $||\zeta_K - \zeta_K^*|| > C_1K^{-s/(2r)}$. Then, with $\bar{\zeta}$ between ζ_K and ζ_K^*,

$$\begin{aligned} Q(\zeta_K) - Q(\zeta_K^*) &= \frac{\partial Q(\zeta_K^*)}{\partial\zeta'}(\zeta_K - \zeta_K^*) + \frac{1}{2}(\zeta_K - \zeta_K^*)'\frac{\partial^2 Q(\bar{\zeta})}{\partial\zeta\partial\zeta'}(\zeta_K - \zeta_K^*). \\ &\leq -(1/2)\lambda_\eta C_1^2K^{-s/r} \leq -2C_2K^{-s/r}, \end{aligned}$$

which contradicts to (35). This completes the proof of (34)(a).

We proceed to prove (34)(b). First, we show $\hat{\zeta}_K \to_p \zeta_K^*$ by showing that $\mathcal{L}_N(\zeta)$ is equicontinuous. Write $\ell(w_i;\gamma^m)$ as $\phi(h_K(\gamma^m))(x_{i1},a_{i1}) \times g_i(\{R^K(x_{it})'\gamma^m\}_{t=2}^T)$ for a function $g_i(\cdot)$, so that

$$\mathcal{L}_N(\zeta) = \frac{1}{N}\sum_{i=1}^N \ln\left(\sum_{m=1}^M \pi^m \phi(h_K(\gamma^m))(x_{i1},a_{i1}) \times g_i\left(\{R^K(x_{it})'\gamma^m\}_{t=2}^T\right)\right).$$

For $\zeta, \tilde{\zeta} \in \Theta_K$, let h^m and $\tilde{h}^m$ denote the conditional choice probability estimates implied by $R^K(x)'\gamma^m$ and $R^K(x)'\tilde{\gamma}^m$, respectively. Applying the mean value theorem to $\mathcal{L}_N(\zeta) - \mathcal{L}_N(\tilde{\zeta})$, with $\bar{\zeta} \in [\zeta, \tilde{\zeta}]$, gives:

$$\begin{aligned}\mathcal{L}_N(\zeta) - \mathcal{L}_N(\tilde{\zeta}) &= \frac{1}{N}\sum_{i=1}^N\sum_{m=1}^M A_{1i}^m(\bar{\zeta})(\pi^m - \tilde{\pi}^m) + \frac{1}{N}\sum_{i=1}^N\sum_{m=1}^M A_{2i}^m(\bar{\zeta})(\phi(h^m) - \phi(\tilde{h}^m))(x_{i1},a_{i1}) \\ &\quad + \frac{1}{N}\sum_{i=1}^N\sum_{m=1}^M\sum_{t=2}^T A_{3it}^m(\bar{\zeta})R^K(x_{it})'(\gamma^m - \tilde{\gamma}^m), \qquad (36)\end{aligned}$$

where $\sup_{\zeta\in\Theta_K} ||A_{ji(t)}^m(\zeta)|| < \infty$ from Assumption 8(a). The first term on the right of (36) is bounded by $C||\pi^m - \tilde{\pi}^m||$. For the third term on the right of (36), we have

$$E\left|\sum_{m=1}^M\sum_{t=2}^T A_{3it}^m(\bar{\zeta})R^K(x_{it})'(\gamma^m - \tilde{\gamma}^m)\right| \le C\sum_{m=1}^M\sum_{t=2}^T E\left|R^K(x_{it})'(\gamma^m - \tilde{\gamma}^m)\right|.$$

Because $R^K(x)$ may be chosen so that $E[R^K(x_{it})R^K(x_{it})'] = I_K$ (H03, p. 1177), where the expectation is taken with respect to the stationary distribution of x, it follows that

$$\begin{aligned}E\left|R^K(x_{it})'(\gamma^m - \tilde{\gamma}^m)\right| &\le \left(E\left|R^K(x_{it})'(\gamma^m - \tilde{\gamma}^m)\right|^2\right)^{1/2} \\ &= [(\gamma^m - \tilde{\gamma}^m)'E[R^K(x_{it})R^K(x_{it})'](\gamma^m - \tilde{\gamma}^m)]^{1/2} = ||\gamma^m - \tilde{\gamma}^m||. \qquad (37)\end{aligned}$$

Hence the third term on the right of (36) is bounded by $\sum_{m=1}^M ||\gamma^m - \tilde{\gamma}^m||$ in L^1. Finally, the second term on the right of (36) is bounded in L^1 by

$$CE|(\phi(h^m) - \phi(\tilde{h}^m))(x,a)| \le C||\phi(h^m) - \phi(\tilde{h}^m)||_\infty \le C\sup_{(a,x)}|h^m(a|x) - \tilde{h}^m(a|x)|, \qquad (38)$$

where the last inequality follows from (30). Let $f^*(x) = \sum_a p^*(x,a)$ denote the stationary distribution of x. Observe that $\sup_x |h^m(a|x) - \tilde{h}^m(a|x)| \le (\inf_x f^*(x))^{-1}\sup_x\{|h^m(a|x) - \tilde{h}^m(a|x)|f^*(x)\} \le C\int_x |h^m(a|x) - \tilde{h}^m(a|x)|f^*(x)dx$, where the last inequality follows because $|h^m(a|x) - \tilde{h}^m(a|x)|f^*(x)$ is continuous in x and bounded. Because $h^m(a|x) = B_a(R^K(x)'\gamma^m)$ for a continuously differentiable function B_a, the right hand side of (38) is bounded by (see also

(37))

$$C\int_x |R^K(x)'(\gamma^m-\tilde{\gamma}^m)|f^*(x)dx = CE|R^K(x_{it})'(\gamma^m-\tilde{\gamma}^m)| \leq C||\gamma^m-\tilde{\gamma}^m||.$$

Consequently, for any $\zeta,\tilde{\zeta}\in\Theta_K$, $|\mathcal{L}_N(\zeta)-\mathcal{L}_N(\tilde{\zeta})|\leq\Delta_N$ with $E|\Delta_N|<C||\zeta-\tilde{\zeta}||$. Hence, $\mathcal{L}_N(\zeta)$ is stochastically equicontinuous (cf., proof of Lemma 2.9 of Newey and McFadden (1994)). Assumption 8(d) implies $\mathcal{L}_N(\zeta)\to_p Q(\zeta)$ for all $\zeta\in\Theta_K$, and in conjunction with Assumption 8(b), we obtain $\sup_{\zeta\in\Theta_K}|\mathcal{L}_N(\zeta)-Q(\zeta)|\to_p 0$ by Lemma 2.8 of Newey and McFadden (1994). $\hat{\zeta}_K\to_p\zeta_K^*$ follows from their Theorem 2.1.

Having established $\hat{\zeta}_K\to_p\zeta_K^*$, we can expand $\partial\mathcal{L}_N(\hat{\zeta}_K)/\partial\zeta=0$ around ζ_K^* with probability approaching one to obtain

$$0=\frac{\partial}{\partial\zeta}\mathcal{L}_N(\zeta_K^*)+\frac{\partial^2}{\partial\zeta\partial\zeta'}\mathcal{L}_N(\bar{\zeta})(\hat{\zeta}_K-\zeta_K^*).\quad \bar{\zeta}\in[\hat{\zeta}_K,\zeta_K^*]$$

By Assumption 8(c), for N large enough, we have $\lambda_{\min}[-(\partial^2/\partial\zeta\partial\zeta')\mathcal{L}_N(\bar{\zeta})]>\varepsilon>0$. Then (34)(b) follows if we show $\partial\mathcal{L}_N(\zeta_K^*)/\partial\zeta=O_p(\sqrt{K/N})$.

Note that $\partial\mathcal{L}_N(\zeta_K^*)/\partial\gamma^m=\mathcal{L}_N^1+\mathcal{L}_N^2$, where

$$\mathcal{L}_N^1=\frac{1}{N}\sum_{i=1}^N\sum_{m=1}^M A_{2i}^m(\zeta_K^*)\int\frac{\partial\phi(h^m(\zeta_K^*))(x_{i1},a_{i1})}{\partial h^m(a=1|x)}R^K(x)dx,\quad \mathcal{L}_N^2=\frac{1}{N}\sum_{i=1}^N\sum_{m=1}^M\sum_{t=2}^T A_{3it}^m(\zeta_K^*)R^K(x_{it}),$$

and $A_{2i}^m(\cdot)$ and $A_{3it}^m(\cdot)$ are given in (36) and $h^m(\zeta_K^*)(a=1|x)=L(R^K(x)\gamma^{m*})$. For $\mathcal{L}_N^1$, note that $E\left\|\int\frac{\partial\phi(h^m(\zeta_K^*))(x_{i1},a_{i1})}{\partial h^m(a=1|x)}R^K(x)dx\right\|^2\leq C_3\cdot\mathrm{tr}(E[R^K(x_{it})R^K(x_{it})'])=C_3K$ with $C_3=\sup_{x\in X}\left(E\left|\frac{\partial\phi(h^m(\zeta_K^*))(x_1,a_1)}{\partial h^m(a=1|x)}\right|^2\Big/f^*(x)^2\right)<\infty$ by Assumptions 7(e) and 8(e).[14] For $\mathcal{L}_N^2$, note that $E||R^K(x_{it})||^2=K$. In conjunction with $\sup_{\zeta\in\Theta_K}||A_{ji(t)}^m(\zeta)||<\infty$, we obtain $\partial\mathcal{L}_N(\zeta_K^*)/\partial\gamma^m=O_p(\sqrt{K/N})$. Furthermore, $\partial\mathcal{L}_N(\zeta_K^*)/\partial\pi^m=N^{-1}\sum_{i=1}^N\sum_{m=1}^M A_{1i}^m(\zeta_K^*)=O_p(\sqrt{1/N})$ because $E[\sum_{m=1}^M A_{1i}^m(\zeta_K^*)]=0$ by the information inequality. Therefore, $\partial\mathcal{L}_N(\zeta_K^*)/\partial\zeta=O_p(\sqrt{K/N})$, and the required result follows. □

[14] This follows from

$$\begin{aligned}
&E\left\|\int\frac{\partial\phi(h^m(\zeta_K^*))(x_{i1},a_{i1})}{\partial h^m(a=1|x)}R^K(x)dx\right\|^2\\
&\quad=\mathrm{tr}\left(\int\int E\left[\frac{\partial\phi(h^m(\zeta_K^*))(x_1,a_1)}{\partial h^m(a=1|x)}\frac{\partial\phi(h^m(\zeta_K^*))(x_1,a_1)}{\partial h^m(a=1|z)}\right]R^K(x)R^K(z)'dxdz\right)\\
&\quad\leq C_3\mathrm{tr}\left(\int\int R^K(x)R^K(z)'f^*(x)f^*(z)dxdz\right)\leq C_3\mathrm{tr}\left(E[R^K(x)R^K(x)']\right),
\end{aligned}$$

where the expectation is taken with respect to the stationary distribution of (x_1,a_1) and the last two inequalities use the Cauchy-Schwartz inequality.

6.9 Proof of Proposition 11

To simplify the notation, we assume the lagged choice is not a part of the state variable. We show the required result in two steps. Let $\zeta_K^\diamond = \{\pi_K^{\diamond m}, \gamma_K^{\diamond m}\}_{m=1}^M$ be a maximizer of $Q(\zeta)$, which may not necessarily be unique. First, we show $||\zeta_K^\diamond - \zeta_K|| < \eta$ for any $\eta > 0$. Second, we show that $\inf_{\zeta \in \{\zeta : ||\zeta - \zeta_K|| < \delta\}} \lambda_{\min}[-(\partial^2/\partial\zeta\partial\zeta')Q(\zeta)] > 0$ for some $\delta > 0$. Then, it follows that $\zeta_K^\diamond$ is unique and $\lambda_{\min}[-(\partial^2/\partial\zeta\partial\zeta')Q(\zeta_K^\diamond)] > 0$.

Define the conditional choice probability implied by γ_K^m as $h_K^m(1|x) = L(R^K(x)'\gamma_K^m)$, and define $h_K^m = \{h_K^m(a|x) : (a,x) \in A \times X\}$ and $\vartheta_K = \{\pi_0^m, h_K^m\}_{m=1}^M$. Similarly, define $h_K^{\diamond m}(1|x) = L(R^K(x)'\gamma_K^{\diamond m})$, $h_K^{\diamond m} = \{h_K^{\diamond m}(a|x) : (a,x) \in A \times X\}$, and $\vartheta_K^\diamond = \{\pi_K^{\diamond m}, h_K^{\diamond m}\}_{m=1}^M$. Then, repeating the argument leading to (35) gives $\tilde{Q}(\vartheta^0) \geq \tilde{Q}(\vartheta_K^\diamond) \geq \tilde{Q}(\vartheta_K) \geq \tilde{Q}(\vartheta^0) - C_4 K^{-s/r}$ for a finite and positive constant C_4, which implies $||\vartheta_K^\diamond - \vartheta^0||_\infty < \eta/2$ and $||\vartheta_K - \vartheta^0||_\infty < \eta/2$ for any $\eta > 0$ and for sufficiently large K. Hence $||\vartheta_K^\diamond - \vartheta_K||_\infty < \eta$ follows.

We proceed to show $||\zeta_K^\diamond - \zeta_K||^2 \leq C||\vartheta_K^\diamond - \vartheta_K||_\infty^2$. Observe that $||h_K^{\diamond m} - h_K^m||_\infty^2 \geq \sup_x |L(R^K(x)'\gamma_K^{\diamond m}) - L(R^K(x)'\gamma_K^m)|^2$. Since $|L'(R^K(x)'\gamma)|$ is bounded and bounded away from zero for any $\gamma \in \Gamma_K$ by Assumption 8(a), we have $\sup_x |L(R^K(x)'\gamma_K^{\diamond m}) - L(R^K(x)'\gamma_K^m)|^2 \geq C \sup_x |R^K(x)'(\gamma_K^{\diamond m} - \gamma_K^m)|^2$ with $C = \inf_{x,\gamma} |L'(R^K(x)'\gamma)|^2$. Now, since $f^*(x) \in (0, \infty)$, we have

$$\begin{aligned} \sup_x |R^K(x)'(\gamma_K^{\diamond m} - \gamma_K^m)|^2 &\geq C \int |R^K(x)'(\gamma_K^{\diamond m} - \gamma_K^m)|^2 f^*(x) dx \\ &= CE|R^K(x_{it})'(\gamma_K^{\diamond m} - \gamma_K^m)|^2 = ||\gamma_K^{\diamond m} - \gamma_K^m||^2, \end{aligned}$$

and it follows that $||\zeta_K^\diamond - \zeta_K||^2 \leq C||\vartheta_K^\diamond - \vartheta_K||_\infty^2$. Since η is arbitrary, $||\zeta_K^\diamond - \zeta_K|| < \eta$ follows.

For the second result, because $h_K^m(a|x)$ is a function of ζ_K, we may write $\vartheta_K = \vartheta(\zeta_K)$ for a function $\vartheta(\cdot)$. It follows that $Q(\zeta_K) = \tilde{Q}(\vartheta(\zeta_K))$, and taking the second-order derivative of both sides with respect to ζ and multiplying by ζ from the left and the right give, for any $\zeta \in \Theta_K$,

$$\zeta' \frac{\partial^2 Q(\zeta_K)}{\partial\zeta'\partial\zeta}\zeta = D^2\tilde{Q}(\vartheta(\zeta_K)) \left[\frac{\partial\vartheta(\zeta_K)}{\partial\zeta}\zeta, \frac{\partial\vartheta(\zeta_K)}{\partial\zeta}\zeta\right] + 2D\tilde{Q}(\vartheta(\zeta_K)) \left[\zeta' \frac{\partial^2\vartheta(\zeta_K)}{\partial\zeta'\partial\zeta}\zeta\right].$$

Then $\inf_{\zeta \in \{\zeta : ||\zeta - \zeta_K|| < \delta\}} \lambda_{\min}[-(\partial^2/\partial\zeta\partial\zeta')Q(\zeta)] > 0$ follows from $||\vartheta_K - \vartheta^0||_\infty < \eta$ for any $\eta > 0$, Assumption (c), $D\tilde{Q}(\vartheta^0) = 0$ (zero operator) and the continuity of $D\tilde{Q}(\vartheta)$ and $D^2\tilde{Q}(\vartheta)$. □

References

Aguirregabiria, V. (2006). "Another look at the identification of dynamic discrete decision processes." Mimeographed, University of Toronto.

Aguirregabiria, V. and P. Mira (2002). "Swapping the nested fixed point algorithm: a class of estimators for discrete Markov decision models." Econometrica 70(4): 1519-1543.

Aguirregabiria, V. and P. Mira (2006). "Sequential Estimation of Dynamic Discrete Games." Econometrica, forthcoming.

Anderson, T. W. (1954). "On estimation of parameters in latent structure analysis." Psychometrika 19(1): 1-10.

Bajari, P., Benkard, C.L., and Levin, J. (2005). "Estimating dynamic models of imperfect competition." Mimeographed, Stanford University.

Bajari, P. and H. Hong (2006). "Semiparametric estimation of a dynamic game of incomplete information." NBER Technical Working Paper 320.

Blischke, W. R. (1964). "Estimating the parameters of mixtures of binomial distributions." Journal of the American Statistical Association 59: 510-528.

Browning, M. and J. Carro (2006). "Heterogeneity and microeconometrics modelling." Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society, volume 3.

Cameron, S. V. and J. J. Heckman (1998). "Life cycle schooling and dynamic selection bias: models and evidence for five cohorts of American males." Journal of Political Economy 106: 262-333.

Chandra, S. (1977). "On the mixtures of probability distributions." Scandinavian Journal of Statistics 4: 105-112.

Chen, X. (2006). Large sample sieve estimation of semi-nonparametric models. Forthcoming in J. Heckman and E. Leamer (eds.) Handbook of Econometrics, Vol. 6, Elsevier.

Crawford, G. S. and M. Shum (2005). "Uncertainty and learning in pharmaceutical demand." Econometrica 73(4): 1137-1173.

Elbers, C. and G. Ridder (1982). "True and spurious duration dependence: the identifiability of the proportional hazard model." Review of Economic Studies 49(3): 403-09.

Geweke, J. and M. Keane (2001). "Computationally intensive methods for integration in econometrics," in J. Heckman and E. Leamer (eds.) Handbook of Econometrics, Vol. 5, Elsevier.

Gibson, W. A. (1955). "An extension of Anderson's solution for the latent structure equations." Psychometrika 20(1): 69-73.

Goodman, L. A. (1974). "Exploratory latent structure analysis using both identifiable and unidentifiable models." Biometrika 61(2): 215-231.

Gowrisankaran, G., M. F. Mitchell, and A. Moro (2005) "Why do incumbent senators win? evidence from a dynamic selection model." Mimeographed, Washington University in St. Louis.

Griffel, D. H. (1985). *Applied Functional Analysis*. Dover.

Hall, P. and X.-H. Zhou (2003). "Nonparametric estimation of component distributions in a multivariate mixture." Annals of Statistics 31(1): 201-224.

Hall, P., A. Neeman, R. Pakyari and R. Elmore (2005). "Nonparametric inference in multivariate mixtures." Biometrika 92(3): 667-678.

Heckman, J.J. (1981). "The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process," Structural Analysis of Discrete Panel Data with Econometric Applications, C.F. Manski and D. McFadden (eds): pp. 179-195.

Heckman, J.J. and B. Singer (1984). "A method of minimizing the impact of distributional assumptions in econometric models for duration data." Econometrica 52(2): 271-320.

Heckman, J.J., S. Urzua, and E. Vytlacil (2006). "Understanding instrumental variables in models with essential heterogeneity." Review of Economics and Statistics, forthcoming.

Hirano, K., G. W. Imbens, and G. Ridder (2003). "Efficient estimation of average treatment effects using the estimated propensity score." Econometrica 71(4): 1161-1189.

Honoré, B.E. and E. Tamer (2006). "Bounds on parameters in panel dynamic discrete choice models." Econometrica 73(3): 611-629.

Hotz, J. and R. A. Miller (1993). "Conditional choice probabilities and the estimation of dynamic models." Review of Economic Studies 60: 497-529.

Houde, J.-F. and S. Imai (2006). Identification and 2-Step Estimation of DDC Models with Unobserved Heterogeneity. mimeographed, Queen's University.

Kasahara, H. and K. Shimotsu (2006) "Nested pseudo-likelihood estimation and bootstrap-based inference for structural discrete Markov decision models." Queen's University Working Paper.

Keane, M. P., and K. I. Wolpin (1997). "The career decisions of young men." Journal of Political Economy 105: 473-522.

Kitamura, Y. (2004) "Nonparametric identifiability of finite mixtures." Mimeographed, Yale University.

Madansky, A. (1960). "Determinantal methods in latent class analysis." Psychometrika 25(2): 183-198.

Magnac T., and D. Thesmar (2002). "Identifying dynamic discrete decision processes." Econometrica 70: 801-816.

Newey, W. K. and D. McFadden (1994). "Large Sample Estimation and Hypothesis Testing," in R. F. Engle and D. L. McFadden (eds.) Handbook of Econometrics, Vol. 4, Elsevier.

Pakes, A., M. Ostrovsky, and S. Berry (2005). "Simple estimators for the parameters of discrete dynamic games (with entry/exit examples)." Mimeographed, Harvard University.

Pesendorfer, M. and P. Schmidt-Dengler (2006). "Asymptotic Least Squares Estimators for Dynamic Games." Mimeographed, LSE.

Rao, P. (1992). Identifiability in stochastic models. Academic Press.

Ridder, G. (1990) "The non-parametric identification of generalized accelerated failure-time models." Review of Economic Studies 57(2): 167-181.

Rothenberg, T. J. (1971). "Identification in parametric models." Econometrica 39(3): 577-591.

Rust, J. (1987). "Optimal replacement of GMC bus engines: an empirical model of Harold Zurcher." Econometrica 55(5): 999-1033.

Rust, J. (1994). "Estimation of dynamic structural models, problems and prospects: discrete decision processes," in C. Sims (ed.) Advances in Econometrics. Sixth World Congress, Cambridge University Press.

Simonoff, J. S. (1995). "Smoothing categorical data." Journal of Statistical Planning and Inference 47: 41-69.

Simonoff, J. S. (1996). Smoothing Methods in Statistics, Springer.

Stone, M. (1977). "An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion." Journal of the Royal Statistical Society B 39: 44-47.

Titterington, D. M., Smith, A.F.M. and Makov, U.E. (1985). Statistical Analysis of Finite Mixture Distributions, Wiley.

Van den Berg, G. J. (2001). "Duration models: specification, identification and multiple durations," in J. Heckman and E. Leamer (eds.) Handbook of Econometrics, Vol. 5, Elsevier.

Zeidler, E. (1986). *Nonlinear Functional Analysis and its Applications I: Fixed-Point Theorems.* New York, Springer-Verlag.

Table 1: Performance of Series Logit Estimator under different degrees of the polynomials

	$(N,T)=(500,10)$					$(N,T)=(2000,10)$				
Degree of Polynomials	2	3	4	5	6	2	3	4	5	6
$1000\times\text{Bias}^2$ of $\hat{P}^m$	1.0780	0.3210	0.0657	0.0449	0.0254	1.0236	0.2891	0.0648	0.0476	0.0317
$1000\times$Variance of $\hat{P}^m$	0.2707	0.3510	0.5263	0.6079	0.9139	0.0618	0.0846	0.1299	0.1364	0.1927
RMSE of $\hat{P}^m$	0.0367	0.0259	0.0243	0.0256	0.0306	0.0329	0.0193	0.0140	0.0136	0.0150
$1000\times\text{Bias}^2$ of $\hat{\pi}$	0.5392	0.0857	0.0249	0.2519	0.4111	0.3159	0.0112	0.0010	0.0031	0.0385
$1000\times$Variance of $\hat{\pi}$	1.2825	1.6346	1.9402	3.1487	3.3385	0.3105	0.3929	0.4815	0.5822	0.8096
RMSE of $\hat{\pi}$	0.0427	0.0415	0.0443	0.0583	0.0612	0.0250	0.0201	0.0220	0.0242	0.0291
Frequency Selected by										
AIC	24	53	17	4	2	1	24	65	4	6
BIC	87	13	0	0	0	11	85	4	0	0

Notes: Based on 100 simulated samples. The number of types is set to two. The sample size is $(N,T)=(500,10)$. The model parameters are set to: $(\pi^1,\pi^2)=(0.5,0.5)$, $\alpha^1=(10,10)$, and $\alpha^2=(2,2)$. Reported numbers for $\hat{P}^m$ are average across states and types.

Table 2: Performance of Series Logit Estimator and Parametric Maximum Likelihood Estimator for P^m and π (two types)

		P^m						π		
		Incomplete Data				Complete Data		Incomplete		Complete
		Series Logit		MLE		MLE		Series	MLE	MLE
		$m=1$	$m=2$	$m=1$	$m=2$	$m=1$	$m=2$			
$N=500$	$1000\times\text{Bias}^2$	0.1542	0.3678	0.1316	0.7211	0.0206	0.0029	3.3513	1.8457	0.0096
$T=3$	$1000\times$Variance	2.2603	2.7179	2.6998	5.3706	0.5136	0.4563	9.1492	15.1517	0.6004
	RMSE	0.0491	0.0555	0.0532	0.0780	0.0231	0.0214	0.1118	0.1304	0.0247
$N=500$	$1000\times\text{Bias}^2$	0.1260	0.4121	0.0846	0.1582	0.0036	0.0017	0.3549	0.0937	0.0060
$T=5$	$1000\times$Variance	1.1455	0.8314	1.9855	2.5664	0.4168	0.3786	5.3843	8.8271	0.5317
	RMSE	0.0357	0.0353	0.0455	0.0522	0.0205	0.0195	0.0758	0.0944	0.0232
$N=500$	$1000\times\text{Bias}^2$	0.0578	0.5890	0.0070	0.0003	0.0032	0.0045	0.0911	0.0431	0.0114
$T=10$	$1000\times$Variance	0.4561	0.2584	0.4478	0.4857	0.2269	0.2172	1.6148	1.8202	0.5839
	RMSE	0.0227	0.0291	0.0213	0.0220	0.0152	0.0149	0.0413	0.0432	0.0244
$N=2000$	$1000\times\text{Bias}^2$	0.1138	0.4408	0.0291	0.0393	0.0026	0.0002	0.6471	0.0068	0.0000
$T=3$	$1000\times$Variance	0.3714	0.3164	0.5185	1.0567	0.0887	0.1086	2.2959	4.0910	0.1075
	RMSE	0.0220	0.0275	0.0234	0.0331	0.0096	0.0104	0.0542	0.0640	0.0104
$N=2000$	$1000\times\text{Bias}^2$	0.0778	0.4745	0.0025	0.0056	0.0000	0.0004	0.3570	0.0514	0.0010
$T=5$	$1000\times$Variance	0.2590	0.1956	0.3012	0.4281	0.0779	0.0875	1.2365	1.9300	0.1126
	RMSE	0.0184	0.0259	0.0174	0.0208	0.0088	0.0094	0.0399	0.0445	0.0107
$N=2000$	$1000\times\text{Bias}^2$	0.0530	0.5252	0.0020	0.0040	0.0014	0.0005	0.0112	0.0000	0.0000
$T=10$	$1000\times$Variance	0.0926	0.0766	0.0975	0.1410	0.0453	0.0603	0.3929	0.4917	0.0927
	RMSE	0.0121	0.0245	0.0100	0.0120	0.0068	0.0078	0.0201	0.0222	0.0096

Notes: Based on 100 simulated samples. Cubic polynomials are used in the series logit estimator. The model parameters are set to: $(\pi^1,\pi^2)=(0.5,0.5)$, $\alpha^1=(10,10)$, and $\alpha^2=(2,2)$.

Table 3: Performance of Series Logit Estimator and Parametric Maximum Likelihood Estimator for P^m (three types)

		Incomplete Data						Complete Data		
		Series Logit			MLE			MLE		
		$m=1$	$m=2$	$m=3$	$m=1$	$m=2$	$m=3$	$m=1$	$m=2$	$m=3$
$N=500$	$1000\times\text{Bias}^2$	0.4580	0.3132	4.0997	1.0423	1.8160	2.8257	0.0116	0.0376	0.0028
$T=5$	$1000\times$Variance	8.1276	11.5557	25.4937	12.3535	20.6446	31.1327	0.8243	1.1771	0.3056
	RMSE	0.0927	0.1089	0.1720	0.1157	0.1499	0.1843	0.0289	0.0349	0.0176
$N=500$	$1000\times\text{Bias}^2$	0.4645	0.2978	0.3810	0.7467	0.4774	0.2678	0.0236	0.0039	0.0011
$T=10$	$1000\times$Variance	3.8790	5.3353	2.0695	5.4977	10.8451	3.9569	0.3413	0.7509	0.2166
	RMSE	0.0659	0.0751	0.0495	0.0790	0.1064	0.0650	0.0191	0.0275	0.0148
$N=2000$	$1000\times\text{Bias}^2$	0.0460	1.3345	1.1407	0.2542	0.0910	0.9240	0.0032	0.0023	0.0008
$T=5$	$1000\times$Variance	1.3847	1.4514	6.9258	3.1676	7.0027	6.9142	0.1577	0.2445	0.0721
	RMSE	0.0378	0.0528	0.0898	0.0585	0.0842	0.0885	0.0127	0.0157	0.0085
$N=2000$	$1000\times\text{Bias}^2$	0.0127	0.6411	0.4888	0.0209	0.0217	0.0138	0.0015	0.0015	0.0000
$T=10$	$1000\times$Variance	0.3429	0.3446	0.3165	0.6456	0.6921	0.3464	0.0863	0.1869	0.0456
	RMSE	0.0189	0.0314	0.0284	0.0258	0.0267	0.0190	0.0094	0.0137	0.0068

Notes: Based on 100 simulated samples. Cubic polynomials are used in the series logit estimator. The model parameters are set to: $(\pi^1, \pi^2, \pi^3) = (1/3, 1/3, 1/3)$, $\alpha^1 = (15, 15)'$, $\alpha^2 = (1, 1)'$, and $\alpha^3 = (4, 4)'$.

Table 4: Performance of Series Logit Estimator and Parametric Maximum Likelihood Estimator for π^m (three types)

		Incomplete Data				Complete Data	
		Series Logit		MLE		MLE	
		$m=1$	$m=2$	$m=1$	$m=2$	$m=1$	$m=2$
$N=500$	$1000\times\text{Bias}^2$	1.5946	2.7148	0.3482	2.1969	0.0006	0.0001
$T=5$	$1000\times$Variance	28.7290	17.7394	29.5289	17.7247	0.3978	0.4551
	RMSE	0.1741	0.1430	0.1729	0.1411	0.0200	0.0213
$N=500$	$1000\times\text{Bias}^2$	0.6430	0.0220	0.1005	0.1745	0.0001	0.0151
$T=10$	$1000\times$Variance	12.1714	5.0756	16.4884	6.0962	0.5204	0.5029
	RMSE	0.1132	0.0714	0.1288	0.0792	0.0228	0.0228
$N=2000$	$1000\times\text{Bias}^2$	2.3321	0.0556	0.0065	0.7193	0.0003	0.0004
$T=5$	$1000\times$Variance	9.9087	6.4760	15.1737	7.8472	0.1012	0.1350
	RMSE	0.1106	0.0808	0.1232	0.0926	0.0101	0.0116
$N=2000$	$1000\times\text{Bias}^2$	2.8377	0.0448	0.0081	0.0508	0.0002	0.0000
$T=10$	$1000\times$Variance	2.0779	0.8650	4.1446	0.8099	0.1137	0.1314
	RMSE	0.0701	0.0302	0.0644	0.0293	0.0107	0.0115

Notes: Based on 100 simulated samples. Cubic polynomials are used in the series logit estimator. The model parameters are set to: $(\pi^1, \pi^2, \pi^3) = (1/3, 1/3, 1/3)$, $\alpha^1 = (15, 15)'$, $\alpha^2 = (1, 1)'$, and $\alpha^3 = (4, 4)'$.

Table 5: Performance of Series Logit Estimator and Parametric Maximum Likelihood Estimator under the model with type-specific transition functions (two types)

		P^m						π		
		Incomplete Data				Complete Data		Incomplete		Complete
		Series Logit		MLE		MLE		Series	MLE	MLE
		$m=1$	$m=2$	$m=1$	$m=2$	$m=1$	$m=2$			
$N=500$	$1000\times\text{Bias}^2$	0.1143	0.2277	0.2364	0.0051	0.0666	0.0009	0.6424	0.2164	0.0071
$T=3$	$1000\times$Variance	3.0484	0.8486	3.0553	0.7756	0.8445	0.3586	6.5723	8.3861	0.5959
	RMSE	0.0562	0.0328	0.0574	0.0279	0.0302	0.0190	0.0849	0.0927	0.0246
$N=500$	$1000\times\text{Bias}^2$	0.0222	0.1990	0.0065	0.0102	0.0020	0.0027	0.1130	0.0160	0.0060
$T=5$	$1000\times$Variance	1.3496	0.4278	1.1417	0.4057	0.5980	0.2965	2.2066	2.2994	0.5317
	RMSE	0.0370	0.0250	0.0339	0.0204	0.0245	0.0173	0.0482	0.0481	0.0232
$N=500$	$1000\times\text{Bias}^2$	0.0359	0.2677	0.0144	0.0073	0.0077	0.0070	0.0136	0.0046	0.0114
$T=10$	$1000\times$Variance	0.6176	0.1675	0.5379	0.1678	0.3206	0.1506	0.7165	0.7462	0.5839
	RMSE	0.0256	0.0209	0.0235	0.0132	0.0181	0.0126	0.0270	0.0274	0.0244
$N=2000$	$1000\times\text{Bias}^2$	0.0067	0.2310	0.0067	0.0004	0.0101	0.0005	0.3722	0.0017	0.0000
$T=3$	$1000\times$Variance	0.6572	0.1927	0.6567	0.1588	0.1706	0.0913	1.4178	1.7934	0.1075
	RMSE	0.0258	0.0206	0.0258	0.0126	0.0134	0.0096	0.0423	0.0424	0.0104
$N=2000$	$1000\times\text{Bias}^2$	0.0591	0.2134	0.0039	0.0002	0.0004	0.0005	0.0741	0.0042	0.0010
$T=5$	$1000\times$Variance	0.2741	0.0926	0.2397	0.0875	0.1197	0.0678	0.4172	0.4522	0.1126
	RMSE	0.0183	0.0175	0.0156	0.0094	0.0110	0.0083	0.0222	0.0214	0.0107
$N=2000$	$1000\times\text{Bias}^2$	0.0788	0.2296	0.0011	0.0004	0.0013	0.0002	0.0012	0.0003	0.0000
$T=10$	$1000\times$Variance	0.1011	0.0470	0.0894	0.0491	0.0528	0.0426	0.1526	0.1571	0.0927
	RMSE	0.0134	0.0166	0.0095	0.0070	0.0074	0.0065	0.0124	0.0125	0.0096

Notes: Based on 100 simulated samples. Cubic polynomials are used in the series logit estimator. The type probabilities are set to $(\pi^1, \pi^2) = (0.5, 0.5)$. The model parameters are set to: $(\pi^1, \pi^2) = (0.5, 0.5)$, $\alpha^1 = (10, 10)$, and $\alpha^2 = (2, 2)$. The parameters for transition functions are $(\theta^1_{f,1}, \theta^1_{f,2}) = (0.4, 0.4)$ and $(\theta^2_{f,1}, \theta^2_{f,2}) = (0.2, 0.2)$.