University of Toronto
Department of Economics

Working Paper 603

Sufficient Statistics for Unobserved Heterogeneity in Structural Dynamic Logit Models

By Victor Aguirregabiria, Jiaying Gu and Yao Luo

May 10, 2018

Sufficient Statistics for Unobserved Heterogeneity in Structural Dynamic Logit Models

Victor Aguirregabiria* — University of Toronto and CEPR
Jiaying Gu* — University of Toronto
Yao Luo* — University of Toronto

May 8th, 2018

Abstract

We study the identification and estimation of structural parameters in dynamic panel data logit models where decisions are forward-looking and the joint distribution of unobserved heterogeneity and observable state variables is nonparametric, i.e., fixed-effects model. We consider models with two endogenous state variables: the lagged decision variable, and the time duration in the last choice. This class of models includes as particular cases important economic applications such as models of market entry-exit, occupational choice, machine replacement, inventory and investment decisions, or dynamic demand of differentiated products. The identification of structural parameters requires a sufficient statistic that controls for unobserved heterogeneity not only in current utility but also in the continuation value of the forward-looking decision problem. We obtain the minimal sufficient statistic and prove identification of some structural parameters using a conditional likelihood approach. We apply this estimator to a machine replacement model.

Keywords: Panel data discrete choice models; Dynamic structural models; Fixed effects; Unobserved heterogeneity; Structural state dependence; Identification; Sufficient statistic.

JEL: C23; C25; C41; C51; C61.

Victor Aguirregabiria. Address: 150 St. George Street. Toronto, ON, M5S 3G7, Canada.
E-mail: victor.aguirregabiria@utoronto.ca

Jiaying Gu. Address: 150 St. George Street. Toronto, ON, M5S 3G7, Canada.
E-mail: jiaying.gu@utoronto.ca

Yao Luo. Address: 150 St. George Street. Toronto, ON, M5S 3G7, Canada.
E-mail: yao.luo@utoronto.ca

*We have benefited from comments and conversations with Bo Honoré, Kyoo il Kim, Roger Koenker, Bob Miller, Pedro Mira, Ismael Mourifié, Whitney Newey, Elie Tamer, Yuanyuan Wan, Jeff Wooldridge, and from participants in seminars at Harvard-MIT, Michigan State, Princeton, Toronto, the conference at UCL on *Implementation of Structural Dynamic Models*, and the conference at Vanderbilt on *Identification in Econometrics*.

1 Introduction

Persistent unobserved heterogeneity is pervasive in empirical applications using panel data of individuals, households, or firms. An important challenge in these applications consists of distinguishing between *true dynamics* due to state dependence and *spurious dynamics* due to unobserved heterogeneity (Heckman, 1981). The identification of *true dynamics*, when persistent unobserved heterogeneity is present, should deal with two key econometric issues: the *incidental parameters problem*, and the *initial conditions problem*. The first one establishes that a simple dummy-variables estimator, that treats each individual unobservable as a parameter to be estimated jointly with the parameters of interest, is inconsistent in most nonlinear panel data models when T is fixed (Neyman and Scott, 1948, Lancaster, 2000). Given this issue, it would seem reasonable to consider a nonparametric (or a flexible) joint distribution of the unobserved heterogeneity and the observables variables, and construct a likelihood function that is integrated over unobservables. In this context, the *initial conditions problem* establishes that the joint distribution of the unobserved heterogeneity and the initial values of the observable variables is not nonparametrically identified, but the misspecification of this joint distribution can generate important biases in the estimation of the parameters of interest (Heckman, 1981, Chamberlain, 1985, among others).

There are two general approaches to deal with this issue: random effects and fixed effects models/methods. Random-effects models impose restrictions on the distribution of unobserved heterogeneity (e.g., parametric, finite mixture), and on the joint distribution of these unobservables and the initial conditions of the observable explanatory variables. Under these restrictions, the parameters of interest and the distribution of the unobserved heterogeneity are jointly estimated. In contrast, fixed-effects methods focus on the estimation of the parameters of interest and they do not try to identify the distribution of the unobserved heterogeneity. These methods are more robust because they are fully nonparametric in the specification of the joint distribution of unobserved heterogeneity and exogenous or predetermined explanatory variables.[1]

A fixed effect method, pioneered by Andersen (1970) and extended by Chamberlain (1980), is based on the derivation of sufficient statistics for the incidental parameters (fixed effects) and the maximization of a likelihood function conditional on these sufficient statistics. This paper deals

[1] See Arellano and Honoré (2001), and Arellano and Bonhomme (2012, 2017) for recent surveys on the econometrics of nonlinear panel data models.

with this fixed effects - sufficient statistics - conditional maximum likelihood approach (FE-CML hereinafter). We study the applicability of this approach to structural dynamic discrete choice models where agents are forward-looking.[2]

There is a wide class of nonlinear panel data models where the FE-CML approach cannot identify the structural parameters.[3] In general, a sufficient statistic of the incidental parameters always exists.[4] The identification problem appears when the minimal sufficient statistic is such that the likelihood conditional on this statistic does not depend on the structural parameters. For instance, in the context of binary choice models, Chamberlain (1993, 2010) shows that a necessary and sufficient condition for (point) identification under the FE-CML approach is that the distribution of the time-varying unobservable is logistic.[5] Similarly, identification is not possible in discrete choice models where unobserved heterogeneity appears in the slope parameters, interacting with predetermined explanatory variables [6] This has important implications for structural dynamic discrete choice models. In these models, an agent's optimal decision depends not only on her current utility but also on the continuation value function, which is an endogenous object. In general, unobserved heterogeneity enters non-additively in the continuation value function and interacts with the observable state variables, even when this unobserved heterogeneity is additively separable in the one-period utility function. This interaction between the unobserved heterogeneity and the

[2] Among the class of fixed-effects estimators in short panels, the dummy-variables estimator is the simplest of these methods. However, as mentioned above, this estimator is inconsistent in most nonlinear panel data models when T is fixed. Two-step bias reduction methods, both analytical and simulation-based, have been proposed to correct for the asymptotic bias of these dummy-variables fixed-effect estimators (e.g., Hahn and Newey, 2004, Browning and Carro, 2010, and Hahn and Kuersteiner, 2011, among others). Other fixed-effects estimators are based on a transformation of the model that eliminates the fixed effects, e.g., Manski's maximum score estimator (Manski, 1987). However, in nonlinear models, these estimators require strict exogeneity of the explanatory variables, ruling out nonlinear dynamic models. Bonhomme (2012) presents a *functional differencing approach* that includes as particular cases different fixed effects estimators in the literature.

[3] In this paper, the concepts of identification and consistent estimation, as N goes not infinity and T is fixed, are used as synonymous.

[4] For instance, we could define as sufficient statistic the complete choice history of an individual. Obviously, the conditional likelihood function based on this sufficient statistic does not depend neither on incidental nor on structural parameters. Though this is an extreme example, it illustrates that the key identification problem is not finding a sufficient statistic for the incidental parameters but showing that there are sufficient statistics for which the conditional likelihood still depends on the structural parameters.

[5] Chamberlain (1993, 2010) considers the model where the time-varying unobservables are independently and identically distributed. Magnac (2004) studies a two-period model where the two time-varying unobservables have a general joint distribution. Honorè and Tamer (2006) study partial identification of the dynamic Probit model and derive sharp bounds on parameters.

[6] Browining and Carro (2014) study the identification of this type of dynamic binary choice model with *maximal heterogeneity* in short panels. The fixed-effects model (nonparametric specification of the unobserved heterogeneity) is not identified. They consider a finite mixture specification of the heterogeneous parameters. This is in the same spirit as Kasahara and Shimotsu (2009), though these other authors consider a nonparametric Markov chain with finite mixture unobserved heterogeneity.

endogenous state variables implies that structural parameters are not identified in the fixed-effects model.

For non-structural (i.e., myopic) dynamic logit models with unobserved heterogeneity only in the intercept, Chamberlain (1985) and Honoré and Kyriazidou (2000) have shown that the FE-CML approach can identify the parameters of interest.[7] In contrast, all the methods and applications for structural dynamic discrete choice models have considered random-effects models with a finite mixture distribution, e.g., Keane and Wolpin (1997), Aguirregabiria and Mira (2007), Kasahara and Shimotsu (2009), Arcidiacono and Miller (2011), among many others. This random-effects approach imposes important restrictions: the number of points in the support of the unobserved heterogeneity is finite and is typically reduced to a small number of points; furthermore, the joint distribution of the unobserved heterogeneity and the initial conditions of the observable state variables is restricted.

In this paper, we revisit the applicability of FE-CML methods to the identification and estimation of structural dynamic discrete choice models. We follow the sufficient statistics approach to study the identification of payoff function parameters in structural dynamic logit models with a fixed-effects specification of the time-invariant unobserved heterogeneity. We consider multinomial models with two types of endogenous state variables: the lagged value of the decision variable, and the time duration in the last choice. The main challenge for the identification of this model comes from the fact that unobserved heterogeneity enters not only in current utility but also in the continuation value of the forward-looking decision problem. In general, this continuation value is a nonlinear function of unobserved heterogeneity and state variables.[8] Therefore, identification requires a sufficient statistic that controls for this continuation value but implies a conditional likelihood that still depends on the structural parameters that capture true state dependence. We derive the minimal sufficient statistic and show that some structural parameters are identified. The forward-looking model where the only state variable is the lagged decision is identified under the same conditions as the myopic version of the model. Instead, with duration dependence, there are some parameters identified in the myopic model but not in the forward-looking model.

[7] In the models of these papers, the only endogenous (predetermined) explanatory variable is the lagged decision. For instance, time duration in the last choice is not an explanatory variable. In our model, we include both lagged decision and duration as state variables.

[8] In fact, before solving the model, we do not know how unobserved heterogeneity and state variables enter this continuation value function. Therefore, for fixed-effects estimation, it is as if we had a nonparametric specification of this function.

Based on our identification results, we consider a conditional maximum likelihood estimator, and a test for the validity of a correlated random effects specification. We apply this estimator and the test to the bus engine model Rust (1987) using both simulated and actual data.

In most empirical applications of structural models, the researcher is not only interested in the value of the structural parameters but also on the estimation of marginal effects and counterfactual experiments. The identification of marginal effects and counterfactuals requires the identification of the distribution of the observed heterogeneity. Point identification requires imposing restrictions on the joint distribution of unobserved heterogeneity and the initial conditions of the state variables. Alternatively, the researcher may prefer not to impose these restrictions and then set-identify the distribution of the unobservables and the marginal effects (Chernozhukov, Fernandez-Val, Hahn, and Newey, 2013). We discuss this problem is section 3.4.

This paper contributes to the literature on structural dynamic discrete choice models. The structure of the payoff function and of the endogenous state variables that we consider in this paper includes as particular cases important economic applications in the literature of dynamic discrete choice structural models, such as models of market entry and exit either binary (Roberts and Tybout, 1997, Aguirregabiria and Mira, 2007) or multinomial (Sweeting, 2013; Caliendo et al, 2015); occupational choice models (Miller, 1984; Keane and Wolpin, 1997); machine replacement models (Rust, 1987; Das, 1992; Kennet, 1993; and Kasahara, 2009); inventory and investment decision models (Aguirregabiria 1999; Ryan, 2013; Kalouptsidi, 2014); demand of differentiated products with consumer brand switching costs (Erdem, Keane, and Sun, 2008) or storable products (Erdem, Imai, and Keane, 2003; Hendel and Nevo, 2006); and dynamic pricing models with menu costs (Willis, 2006), or with duration dependence due to inflation or other forms of depreciation (Slade, 1998; Aguirregabiria, 1999; Kano, 2013); among others.[9] Our paper also contributes to the literature on nonlinear dynamic panel data models by providing new identification results of fixed effects dynamic logit models with duration dependence (Frederiksen, Honoré, and Hu, 2007).

The rest of the paper is organized as follows. Section 2 describes the class of models that we study in this paper. Section 3 presents our identification results. Section 4 deals with estimation and

[9] Note that most of the empirical applications cited above in this paragraph do not allow for time-invariant unobserved heterogeneity. This is still a common restriction in empirical applications of dynamic structural models, and it is mostly justified by computational convenience. The exceptions, within the cited papers, are Keane and Wolpin (1997), Erdem, Imai, and Keane (2003), Willis (2006), Aguirregabiria and Mira (2007), and Erdem, Keane, and Sun (2008).

inference. In section 5, we illustrate our identification results in the context of a bus replacement model. Section 6 summarizes and concludes. Proofs of Lemmas and Propositions are in the Appendix. Also in the Appendix, we show that our identification results extend to an extended version of our model where the endogenous state variables have a stochastic transition rule.

2 Model

Time is discrete and indexed by t that belongs to $\{1, 2, ..., \infty\}$.[10] Agents are indexed by i. Every period t, agent i chooses a value of the discrete variable $y_{it} \in \mathcal{Y} = \{0, 1, ..., J\}$ to maximize her expected and discounted intertemporal utility $\mathbb{E}_t \left[\sum_{j=0}^{\infty} \delta_i^j \; \Pi_{i,t+j}(y_{i,t+j})\right]$, where $\delta_i \in (0, 1)$ is agent i's time discount factor, and $\Pi_{it}(y)$ is her one-period utility if she chooses action y. This utility is a function of four types of state variables which are known to the agent at period t:

$$\Pi_{it}(y) = \alpha\left(y, \boldsymbol{\eta}_i, \mathbf{z}_{it}\right) + \beta\left(y, \mathbf{x}_{it}, \mathbf{z}_{it}\right) + \varepsilon_{it}(y). \tag{1}$$

$\mathbf{z}_{it}$ and $\mathbf{x}_{it}$ are observable to the researcher, and ε_{it} and $\boldsymbol{\eta}_i$ are unobservable. The vector $\mathbf{z}_{it}$ contains exogenous state variables and it follows a Markov process with transition probability function $f_{\mathbf{z}}(\mathbf{z}_{i,t+1}|\mathbf{z}_{it})$. The vector $\mathbf{x}_{it}$ contains endogenous state variables. We describe below the nature of these endogenous state variables and their transition rules. Both $\mathbf{z}_{it}$ and $\mathbf{x}_{it}$ have discrete supports $\mathcal{Z}$ and $\mathcal{X}$, respectively. The unobservable variables $\{\varepsilon_{it}(y) : y \in \mathcal{Y}\}$ are *i.i.d.* over (i, t, y) with an extreme value type I distribution. The vector $\boldsymbol{\eta}_i$ represents time-invariant unobserved heterogeneity from the point of view of the researcher. Let $\theta_i \equiv (\boldsymbol{\eta}_i, \delta_i)$ represent the unobserved heterogeneity from individual i. The probability distribution of θ_i conditional on the history of observable state variables $\{\mathbf{z}_{it}, \mathbf{x}_{it} : t = 1, 2, ...\}$ is unrestricted and nonparametrically specified, i.e., fixed effects model. Functions $\alpha\left(y, \boldsymbol{\eta}, \mathbf{z}\right)$ and $\beta\left(y, \mathbf{x}, \mathbf{z}\right)$ are nonparametrically specified but they are bounded.

Our specification of the utility function represents a general semiparametric fixed-effect logit model. It builds on *Rust model* (Rust, 1987, 1994) and generalizes it in two directions. First, Rust assumes that all the unobservables satisfy the conditions of *additive separability* and *conditional independence*, and they have an extreme value distribution. While our time-varying unobservables $\varepsilon_{it}(y)$ satisfy these conditions, our time-invariant unobserved heterogeneity interacts, in an

[10] The time horizon of the decision problem is infinite.

unrestricted way, with the exogenous state variables and the choice, and they do not satisfy the conditional independence assumption. Second, we allow for unobserved heterogeneity in the discount factor.

The assumption of additive separability between $\boldsymbol{\eta}_i$ and the endogenous state variables in $\mathbf{x}_{it}$ is key for the identification results in this paper. This condition does not imply that the conditional-choice value functions, that describe the solution of the dynamic model, are additive separability between $\boldsymbol{\eta}_i$ and $\mathbf{x}_{it}$. In general, the solution of the dynamic programming problem implies a value function that is not additively separable in $\boldsymbol{\eta}_i$ and $\mathbf{x}_{it}$ even when the utility function is additive in these variables.

The model includes two types of endogenous state variables that correspond to two different types of state dependence, $\mathbf{x}_{it} = (y_{i,t-1}, d_{it})$: (a) dependence on the the lagged decision variable, $y_{i,t-1}$; and (b) *duration dependence*, where $d_{it} \in \{1, 2, ..., \infty\}$ is the number of periods since the last change in choice. The lagged decision has the obvious transition rule. The transition rule for the duration variable is $d_{i,t+1} = 1\{y_{it} = y_{i,t-1}\}\ d_{it} + 1$, where $1\{.\}$ is the indicator function.[11]

The term $\beta(y, \mathbf{x}_{it}, \mathbf{z}_{it})$ in the payoff function captures the dynamics, or structural state dependence, in the model. We distinguish in this function two additive components that correspond to the two forms of state dependence in the model:

$$\beta(y, \mathbf{x}_{it}, \mathbf{z}_{it}) = 1\{y = y_{i,t-1}\}\ \beta_d(y, d_{it}, \mathbf{z}_{it}) + 1\{y \neq y_{i,t-1}\}\ \beta_y(y, y_{i,t-1}, \mathbf{z}_{it}) \tag{2}$$

Function $\beta_d(y, d_{it}, \mathbf{z}_{it})$ captures duration dependence. For instance, in an occupational choice model, this term captures the return on earnings of job experience in the current occupation. Function $\beta_y(y, y_{i,t-1}, \mathbf{z}_{it})$ represents switching costs. In an occupational choice model, this term represents the cost of switching from occupation $y_{i,t-1}$ to occupation y. The additive separability between switching costs and "returns to experience" is not without loss of generality. For instance, the cost of switching occupation could depend on experience in the current job not only through the loss of the returns of experience, i.e., $\beta_y(.)$ could depend on d_{it}. However, this additive separability facilitates our analysis of identification and the model is still more general than previous fixed-effects discrete choice models.

[11] Note that these endogenous state variables follow deterministic transition rules. In the Appendix, we present a version of the model that allows for stochastic transition rules for the endogenous state variables.

We impose a restriction on the structural function $\beta_d(y, d, \mathbf{z}_{it})$ that plays a role in our identification results for this function. We assume that there is not duration dependence in choice alternative $y = 0$, i.e., $\beta_d(0, d, \mathbf{z}_{it}) = 0$ for any value of d. Also, but without loss of generality, we set $\beta_y(y, y, \mathbf{z}_{it}) = 0$, i.e., the switching cost of no-switching is zero.[12] Assumption 1 summarizes our basic conditions on the model. For the rest of the paper, we assume that this assumption holds.

ASSUMPTION 1. (A) The time horizon is infinite and $\delta_i \in (0, 1)$. (B) The utility function has the form given by equations (1) and (2), and functions $\alpha(y, \eta, \mathbf{z})$, $\beta_d(y, d, \mathbf{z})$, and $\beta_y(y, y_{-1}, \mathbf{z})$ are bounded. (C) $\beta_y(y, y, \mathbf{z}) = 0$, $\beta_d(0, d, \mathbf{z}) = 0$. (D) $\{\varepsilon_{it}(y) : y \in \mathcal{Y}\}$ are i.i.d. over (i, t, y) with a extreme value type I distribution. (E) $\mathbf{z}_{it}$ has discrete and finite support $\mathcal{Z}$ and follows a time-homogeneous Markov process. (F) The probability distribution of $\theta_i \equiv (\boldsymbol{\eta}_i, \delta_i)$ conditional on $\{\mathbf{z}_{it}, \mathbf{x}_{it} : t = 1, 2, ...\}$ is nonparametrically specified and completely unrestricted. ■

Since the model does not have duration dependence when at choice alternative 0, it is convenient for notation to make duration equal to zero at state $y_{t-1} = 0$. In other words, we consider the following modification in the transition rule for duration:

$$d_{i,t+1} = \begin{cases} 1\{y_{it} = y_{i,t-1}\}\ d_{it} + 1 & \text{if} \quad y_{it} > 0 \\ 0 & \text{if} \quad y_{it} = 0 \end{cases} \tag{3}$$

For our identification results in forward-looking models with duration dependence, we also impose the following assumption.

ASSUMPTION 2. For any $y \in \mathcal{Y}$ there is a finite value of duration, $d_y^* < \infty$, such that the marginal return of duration is zero for values greater that d_y^*:[13]

$$\beta_d(y, d, \mathbf{z}) = \beta_d(y, d_y^*, \mathbf{z}) \quad \textit{for any } d \geq d_y^* \qquad ■ \tag{4}$$

For the moment, we assume that the researcher knows the values of d_y^*. In section 4, we show that these values $\{d_y^*\}$ are identified from the data.

The following are some examples of models within the class defined by Assumption 1.

(a) Market entry-exit models. In its simpler version, there is only one market, and the choice variable is binary and represents a firm's decision of being active in the market ($y_{it} = 1$) or not

[12] Given the payoff function in equation (2), the parameter $\beta_y(y, y)$ is completely irrelevant for an individual's optimal decision. When $y_{it} = y_{i,t-1} = y$, we have that $\beta(y, \mathbf{x}_{it}) = \beta_d(y, d_{it}) + 0$ such that the term $\beta_y(y, y)$ never enters in the relevant payoff function. Therefore, $\beta_y(y, y)$ can be normalized to zero without loss of generality.

[13] The assumption of no duration dependence in choice alternative $y = 0$ is equivalent to assuming $d_0^* = 1$.

($y_{it} = 0$), e.g., Dunne et al. (2013). The only endogenous state variable is the lagged decision, $y_{i,t-1}$. The parameter $-\beta_y(1, 0, \mathbf{z})$ represents the cost of entry in the market. Similarly, the parameter $-\beta_y(0, 1, \mathbf{z})$ represents the cost of exit from the market. An extension of the basic entry model includes as an endogenous state variable the number of periods of experience since last entry in the market, d_{it}, which follows the transition rule $d_{i,t+1} = d_{it} + 1$ if $y_{it} = 1$ and $d_{i,t+1} = 0$ if $y_{it} = 0$. The parameter $\beta_d(1, d, \mathbf{z})$ represents the effect of market experience on the firm's profit (Roberts and Tybout, 1997). The model can be extended to J markets (Sweeting, 2013; Caliendo et al, 2015). The two endogenous state variables are the index of the market where the firm was active at the previous period ($y_{i,t-1}$) and the number of periods of experience in the current market (d_{it}). The parameter $\beta_y(y, y_{-1}, \mathbf{z})$ represents the cost of switching from market y_{-1} to market y. There is not duration dependence if a firm is not active in any market (if $y = 0$), and the marginal return to experience in market y is zero after d_y^* periods in the market.

(b) Occupational choice models (Miller, 1984; Keane and Wolpin, 1997). A worker chooses between J occupations and the choice alternative of not working ($y = 0$). There are costs of switching occupations such that a worker's occupation at previous period, y_{it-1}, is a state variable of the model. There is (passive) learning that increases productivity in the current occupation. There is not duration dependence if the worker is unemployed.

(c) Machine replacement models (Rust, 1987; Das, 1992; Kennet, 1993; and Kasahara, 2009). The choice variable is binary and it represents the decision of keeping a machine ($y_{it} = 1$) or replacing it ($y_{it} = 0$). The only endogenous state variable is the number of periods since the last replacement, d_{it}, i.e., the machine age. The evolution of the machine age is $d_{i,t+1} = d_{it} + 1$ if $y_{it} = 1$ and $d_{i,t+1} = 0$ if $y_{it} = 0$. The parameter $\beta_d(1, d, \mathbf{z})$ represents the effect of age on the firm's profit, e.g., productivity declines and maintenance costs increase with age.[14] More generally, the class of models in this paper includes binary choice models of investment in capital, inventory, or capacity (Aguirregabiria 1999; Ryan, 2013; Kalouptsidi, 2014), as long as the depreciation of the stock is deterministic.

(d) Dynamic demand of differentiated products (Erdem, Imai, and Keane, 2003; Hendel and Nevo,

[14] In some versions of this model, such as Rust (1987), the endogenous state variable represents cumulative usage of the machine and it can follow a stochastic transition rule. We consider this stochastic version of the model in the Appendix.

2006). A differentiated product has J varieties and a consumer chooses which one, if any, to purchase (no purchase is represented by $y = 0$). Brand switching costs imply that the brand in the last purchase is a state variable (Erdem, Keane, and Sun, 2008). For storable products, the duration since last purchase, d_{it}, represents (or proxies) the consumer's level of inventory that is an endogenous state variable. Function $\beta_d(y, d, \mathbf{z})$ captures the effect of inventory on the consumer's utility, and function $\beta_y(y, y_{-d}, , \mathbf{z})$ represents brand switching costs.

(e) Menu costs models of pricing (Slade, 1998; Aguirregabiria, 1999; Willis, 2006; Kano, 2013). A firm sells a product and chooses its price to maximize intertemporal profits. The firm's profit has two components: a variable profit that depends on the real price (in logarithms), r_{it}; and a fixed menu cost that is paid only if the firm changes its nominal price. There is a constant inflation rate, π, that erodes the real price. Every period, the firm decides whether to keep its nominal price ($y_{it} = 1$) or to adjust it ($y_{it} = 0$) such that current real price becomes r^*. The evolution of log-real-price is: $r_{it+1} = r_{it} - \pi$ if $y_{it} = 1$, and $r_{it+1} = r^* - \pi$ if $y_{it} = 0$. If d_{it} represents the time duration since the last nominal price change, we can represent the transition rule of the real price as follows: $(r_{it+1} - r^*)/\pi = d_{it} + 1$ if $y_{it} = 1$, and $(r_{it+1} - r^*)/\pi = 0$ if $y_{it} = 0$. This model has a similar structure as the machine replacement models described above. ■

We now derive the optimal decision rule and the conditional choice probabilities in this model. Agent i chooses y_{it} to maximize its expected and discounted intertemporal utility. Given the infinite horizon and the time-homogeneous utility and transition probability functions, Blackwell's Theorem establishes that the value function and the optimal decision rule are time-invariant (Blackwell, 1965). Let $V_{\theta_i}(y_t, d_t, \mathbf{z}_t)$ be the integrated (or smoothed) value function for agent type θ_i, as defined by Rust (1994).[15] The optimal choice at period t can be represented as:

$$y_{it} \quad = \quad \arg\max_{y \in \mathcal{Y}} \{ \; \alpha(y, \boldsymbol{\eta}_i, \mathbf{z}_{it}) + \beta(y, \mathbf{x}_{it}, \mathbf{z}_{it}) + \varepsilon_{it}(y) + \delta_i \; \mathbb{E}\left[V_{\theta_i}(y, d_{i,t+1}, \mathbf{z}_{i,t+1}) \;\mid\; y, \; \mathbf{x}_{it}, \mathbf{z}_{it}\right] \; \} \tag{5}$$

Note that $d_{i,t+1}$ is a deterministic function of $(y, \; \mathbf{x}_{it})$, i.e., conditional on $y_{it} = y$. Therefore, we can represent the continuation value $\mathbb{E}[V_{\theta_i}(y, d_{i,t+1}, \mathbf{z}_{i,t+1}) \mid y, \; \mathbf{x}_{it}, \mathbf{z}_{it}]$ using a function $v_{\theta_i}(y, d_{t+1}[y, \mathbf{x}_{it}]), \mathbf{z}_{it})$ with $d_{t+1}[y, \mathbf{x}_{it}] = 0$ if $y = 0$ and $d_{t+1}[y, \mathbf{x}_{it}] = 1\{y = y_{it-1}\}d_{it} + 1$ if $y > 0$.

[15] The integrated value function is defined as the integral of the value function over the distribution of the i.i.d. unobservable state variables ε.

The extreme value type 1 distribution of the unobservables ε implies that the *conditional choice probability* (CCP) function has the following form:

$$P_{\theta_i}(y \mid \mathbf{x}_{it}, \mathbf{z}_{it}) = \frac{\exp\{\ \alpha(y, \boldsymbol{\eta}_i, \mathbf{z}_{it}) + \beta(y, \mathbf{x}_{it}, \mathbf{z}_{it}) + v_{\theta_i}(y, d_{t+1}[y, \mathbf{x}_{it}], \mathbf{z}_{it})\ \}}{\sum\limits_{j \in \mathcal{Y}} \exp\{\ \alpha(j, \boldsymbol{\eta}_i, \mathbf{z}_{it}) + \beta(j, \mathbf{x}_{it}, \mathbf{z}_{it}) + v_{\theta_i}(j, d_{t+1}[j, \mathbf{x}_{it}], \mathbf{z}_{it})\ \}} \tag{6}$$

The continuation value function v_{θ_i} has two properties which play an important role in our identification results. These properties establish conditions under which the continuation values do not depend on current endogenous state variables, $(y_{i,t-1}.d_{it})$.

Property 1. In a model without duration dependence (i.e., $\beta_d = 0$), the continuation value function becomes $v_{\theta_i}(y, \mathbf{z}_{it})$ that does not depend on the state variable, y_{it-1}.

Property 2. In a model with duration dependence, the continuation $v_{\theta_i}(y, d_{t+1}[y, \mathbf{x}_{it}], \mathbf{z}_{it})$ is equal to $v_{\theta_i}(y, d_y^*, \mathbf{z}_{it})$ for any duration $d_{t+1}[y, \mathbf{x}_{it}] \geq d_y^*$.

3 Identification

3.1 Preliminaries

The researcher has a panel dataset of N individuals over T periods of time, $\{y_{it},\ \mathbf{x}_{it}\ ,\ \mathbf{z}_{it} : i = 1, 2, ..., N\ ;\ t = 1, 2, ..., T\}$. We consider microeconometric applications where N is large and T is small. More precisely, our identification results and the asymptotic properties of the proposed estimator assume that N goes to infinity and T is small and fixed.[16] We are interested in the identification of the functions β_y and β_d that represent the dependence of utility with respect to the endogenous state variables.

For the rest of this section, we omit the individual subindex i in most of the expressions, and instead we include θ as an argument (or subindex) in those functions that depend on the time-invariant unobserved heterogeneity, i.e., $\alpha_\theta(y, \mathbf{z})$ and $v_\theta(\mathbf{x}, \mathbf{z})$. We use β to represent the vector of structural parameters that define the functions β_y and β_d.[17]

As in Honoré and Kyriazidou (2000), our sufficient statistics include the condition that the exogenous state variables, $\mathbf{z}$, remains constant over several periods. For notational simplicity,

[16] Note that T represents the number of periods with data on the decision variable and the state variables for all the individuals. The set of observable state variables includes the endogenous state variables $y_{i,t-1}$ and d_{it}. Knowing the values of these state variables at the initial period $t = 1$ (i.e., y_{i0} and d_{i1}) may require data on the individual's choices for periods before $t = 1$. Therefore, the time dimension T may not correspond to the actual time dimension of the required panel dataset.

[17] Since $(y_t, \mathbf{x}_t, \mathbf{z}_t)$ has finite support, we can represent the structural functions $\beta_y(y_t, y_{t-1}, \mathbf{z}_t)$ and $\beta_d(y_t, d_t, \mathbf{z}_t)$ using a finite vector of parameters.

we omit $\mathbf{z}$ as an argument in most of the expressions for the rest of this section. In section 4, we explain how to deal with this condition in the implementation of the conditional maximum likelihood estimator.

Let $\widetilde{\mathbf{y}} = \{y_1, y_2, ..., y_T\}$ be an individual's observed history of choices and exogenous state variables, respectively. The model implies that:

$$\mathbb{P}(\widetilde{\mathbf{y}} \mid \mathbf{x}_1, \theta, \beta) = \prod_{t=1}^{T} \frac{\exp\{\ \alpha_\theta(y_t) + \beta(y_t, \mathbf{x}_t) + v_\theta(y_t, d_{t+1}[y_t, \mathbf{x}_t])\ \}}{\sum_{j \in \mathcal{Y}} \exp\{\ \alpha_\theta(j) + \beta(j, \mathbf{x}_t) + v_\theta(j, d_{t+1}[j, \mathbf{x}_t])\ \}} \tag{7}$$

Our identification results, for different versions of the model, have the following common features. First, we show that the log-probability function $\ln \mathbb{P}(\widetilde{\mathbf{y}} \mid \mathbf{x}_1, \theta, \beta)$ has the following structure:

$$\ln \mathbb{P}(\widetilde{\mathbf{y}} \mid \mathbf{x}_1, \theta, \beta) = U(\widetilde{\mathbf{y}}, \mathbf{x}_1)' g_\theta + S(\widetilde{\mathbf{y}}, \mathbf{x}_1)' \beta^* \tag{8}$$

where $U(\widetilde{\mathbf{y}}, \mathbf{x}_1)$ and $S(\widetilde{\mathbf{y}}, \mathbf{x}_1)$ are vectors of statistics (i.e., deterministic functions of the history $(\widetilde{\mathbf{y}}, \mathbf{x}_1)$), g_θ is a vector of functions of θ, and β^* is a vector of linear combinations of the original vector of structural parameters β. This representation is such that each of the vectors, U, g_θ, S, and β^*, has elements which are linearly independent.[18] Based on this representation of the log-probability of a choice history, we establish the following results. For notational simplicity, we use U and S to represent $U(\widetilde{\mathbf{y}}, \mathbf{x}_1)$ and $S(\widetilde{\mathbf{y}}, \mathbf{x}_1)$, respectively.

(i) Sufficiency. U is a sufficient statistic for θ, i.e., for any $(\widetilde{\mathbf{y}}, \mathbf{x}_1)$ and θ, $\mathbb{P}(\widetilde{\mathbf{y}} | \mathbf{x}_1, \theta, U)$ does not depend on θ. By definition, $\ln \mathbb{P}(\widetilde{\mathbf{y}} | \mathbf{x}_1, \theta, U)$ is equal to $\ln \mathbb{P}(\widetilde{\mathbf{y}} | \mathbf{x}_1, \theta) - \ln \mathbb{P}(U | \mathbf{x}_1, \theta)$, and taking into account the form of the log-probability in equation (8), we have:

$$\begin{aligned} \ln \mathbb{P}(\widetilde{\mathbf{y}} | \mathbf{x}_1, U, \beta^*) &= U' g_\theta + S' \beta^* - \ln \left(\sum_{j: U(j) = U} \exp\{U(j)' g(\theta) + S(j)' \beta^*\} \right) \\ &= S' \beta^* - \ln \left(\sum_{j: U(j) = U} \exp\{S(j)' \beta^*\} \right) \end{aligned} \tag{9}$$

where $\sum_{j: U(j) = U}$ represents the sum over all the histories that imply the same value U of the vector of sufficient statistics. Equation (9) shows that the structure of the log-probability in (8) implies that U is a sufficient statistic for θ.

[18] Suppose that S and β are $K \times 1$ vectors, and only $K^* < K$ elements in S are linearly independent. Then, $S = [S_a, S_b]$ where S_a contains K^* linearly independent elements, and $S_b = \mathbf{A}\ S_a$ where $\mathbf{A}$ is a $(K - K^*) \times K^*$ matrix. This implies that $S'\beta = S_a' \beta^*$ with $\beta^* = [\mathbf{I} \vdots \mathbf{A}]' \beta$, such that S_a and β^* are vectors with linearly independent elements.

(ii) Minimal sufficiency. U is a *minimal sufficient statistic*, i.e., it does not contain redundant information. More formally, let $\mathbf{U}$ be a matrix where each row corresponds to a value of the choice history $(\widetilde{\mathbf{y}}, \mathbf{x}_1)$. Then, U is *minimal* if and only if matrix $\mathbf{U}$ is full-column rank.

(iii) Identification. Define the conditional log-likelihood function, in the population, $\ell(\beta^*) \equiv \mathbb{E}_{(\widetilde{\mathbf{y}},\mathbf{x}_1)}[\ln \mathbb{P}(\widetilde{\mathbf{y}}|\mathbf{x}_1, U, \beta^*)]$. The vector of parameters β^* is point identified if the population likelihood is uniquely maximized at the true value of β^*. Lemma 1 establishes a necessary and sufficient condition for identification that is simple to verify. Let K be the dimension of the vector of statistics S and of the vector of parameters β^*.

LEMMA 1. Given $K+1$ histories $(\widetilde{\mathbf{y}}, \mathbf{x}_1)$, say $\{A_j : j = 0, 1, 2..., K\}$, define a $K \times K$ matrix $\mathbf{S}$ such that every row j is associated to a history and contains the vector of statistics $S(A_j)' - S(A_0)'$. The vector of parameters β^* is identified if and only if there exist $K+1$ histories with the same value of the statistic U and a non-singular matrix $\mathbf{S}$. ■

Corollary: If $K = 1$, parameter β^* is identified iff there are two histories, A and B, such that $U(A) = U(B)$ and $S(A) \neq S(B)$.

The derivation of these sufficient statistics should deal with two issues that do not appear in the previous literature on FE-CMLE of non-structural (or myopic) nonlinear panel data models. First, we consider models with duration dependence. Second, we should take into account that unobserved heterogeneity enters in the continuation value function, v_θ. This implies that the sufficient statistic U should control not only for $\alpha_\theta(y_t)$ but also for the continuation values $v_\theta(y_t, d_{t+1})$. This is challenging because, in general, these continuation values depend on the endogenous state variables. We cannot fully control for (or condition on) the value of the state variables because the identification condition (iii) would not hold. We show that there are states where the continuation value does not depend on current state variables once we condition on current choices.

The presentation of our identification results tries to emphasize both the links and extensions with previous results in the literature. For this reason, we start presenting identification results for the binary choice model, that is the model more extensively studied in the literature of nonlinear dynamic panel data. For this binary choice model, we present new identification results for the myopic model with duration dependence and for the forward-looking model with and without duration dependence. Then, we present our identification results for multinomial models.

Some useful statistics. We show below that, in our model, the log-probability of a choice history, $\mathbb{P}(\widetilde{\mathbf{y}} \mid y_0, d_1, \theta, \beta)$, can be written in terms of several sets of statistics or functions of $(y_0, d_1, \widetilde{\mathbf{y}})$: the initial and final choices, $\{y_0, y_T\}$; the initial and final durations, $\{d_1, d_{T+1}\}$; and the statistics that we denote as *hits*, *dyads*, *histogram of states,* and *histogram of choice-states*. We now define these statistics. Note that each of these statistics for a single history $(y_0, d_1, \widetilde{\mathbf{y}})$.

Hit statistics. For any choice alternative $y \in \mathcal{Y}$, the *hit* statistic $T^{(y)}$ represents the number of times that alternative y is visited (or *hit*) during the choice history $\widetilde{\mathbf{y}}$, i.e., $T^{(y)} \equiv \sum_{t=1}^{T} 1\{y_t = y\}$.

Dyad statistics. For y_{-1} and y in $\mathcal{Y}$, the *dyad* statistic $D^{(y_{-1},y)}$ is the number of times that the sequence (y_{-1}, y) is observed at two consecutive periods in the choice history $(y_0, \widetilde{\mathbf{y}})$, i.e., $D^{(y_{-1},y)} \equiv \sum_{t=1}^{T} 1\{y_{t-1} = y_{-1},\, y_t = y\}$.

Histogram of states. Given a history $(y_0, d_1, \widetilde{\mathbf{y}})$, the statistic $H^{(y)}(d)$ (for $y \in \mathcal{Y}$ and $d \geq 0$) is the number of times that we observe state $(y_{t-1}, d_t) = (y, d)$, i.e., $H^{(y)}(d) = \sum_{t=1}^{T} 1\{y_{t-1} = y,\, d_t = d\}$.

Extended histogram of states. For any $y \in \mathcal{Y}$ and $d \geq 0$, the statistic $X^{(y)}(d)$ represents the number of times that we observe state $(y_{t-1}, d_t) = (y, d)$ and the individual decides to continue one more period in choice y. By definition, $X^{(y)}(d) = \sum_{t=1}^{T} 1\{y_{t-1} = y_t = y,\, d_t = d\}$.

Difference between final and initial states. For any $y \in \mathcal{Y}$ and $d \geq 0$, the statistic $\Delta^{(y)}(d)$ is defined as $1\{y_T = y,\, d_{T+1} = d\} - 1\{y_0 = y,\, d_1 = d\}$. When the difference applies only to the choice variable, we represent it as $\Delta^{(y)} \equiv 1\{y_T = y\} - 1\{y_0 = y\}$.

Table 1 summarizes our definition of statistics. The following Lemma 2 establishes several properties of these statistics that we apply in our derivations.

LEMMA 2. For any history $(y_0, d_1, \widetilde{\mathbf{y}})$ and value $y > 0$ the following properties apply: (i) $H^{(y)}(0) = 0$; (ii) $X^{(y)}(0) = 0$; (iii) $\sum_{d \geq 1} H^{(y)}(d) = T^{(y)} - \Delta^{(y)}$; (iv) $\sum_{d \geq 1} X^{(y)}(d) = D^{(y,y)}$; (v) for $d \geq 1$, $X^{(y)}(d) = H^{(y)}(d+1) + \Delta^{(y)}(d+1)$; (vi) $\sum_{d \geq 1} \Delta^{(y)}(d) = \Delta^{(y)}$; and (vii) for $y \geq 1$, $\sum_{y_{-1} \neq y} D^{(y_{-1},y)} = H^{(y)}(1) + \Delta^{(y)}$. ■

Table 1
Definition of statistics for a choice history $\{y_0, d_1 \mid \widetilde{\mathbf{y}}\}$

Name: Symbol	*Definition*
Hits: $T^{(y)}$	$\sum_{t=1}^{T} 1\{y_t = y\}$
Dyad: $D^{(y_{-1},y)}$	$\sum_{t=1}^{T} 1\{y_{t-1} = y_{-1},\, y_t = y\}$
Histogram of states: $H^{(y)}(d)$	$\sum_{t=1}^{T} 1\{y_{t-1} = y,\, d_t = d\}$
Extended histogram of states: $X^{(y)}(d)$	$\sum_{t=1}^{T} 1\{y_{t-1} = y_t = y,\, d_t = d\}$
Diff. final-initial states: $\Delta^{(y)}(d)$	$1\{y_T = y, d_{T+1} = d\} - 1\{y_0 = y, d_1 = d\}$
$\Delta^{(y)}$	$1\{y_T = y\} - 1\{y_0 = y\}$

3.2 Binary choice models

Consider the binary choice version of the model characterized by Assumption 1. The optimal decision rule in this model is:

$$y_t = 1\left\{\begin{array}{l} \alpha_\theta(1) - \alpha_\theta(0) + \beta(1, y_{t-1}, d_t) - \beta(0, y_{t-1}, d_t) \\ \\ +v_\theta\,(1, d_t + 1) - v_\theta\,(0) + \varepsilon_t(1) - \varepsilon_t(0) \geq 0 \end{array}\right\} \tag{10}$$

where for choice $y = 0$ we use $v_\theta\,(0)$ instead $v_\theta\,(0, 0)$ to emphasize that there is not duration dependence when the state is $y = 0$. We now present identification results for different versions of this model, starting with the myopic model without duration dependence that has been studied by Chamberlain (1985) and Honoré and Kyriazidou (2000).

3.2.1 Myopic dynamic model without duration dependence

Consider the model in equation (10) under the restrictions of myopic behavior (i.e., $\delta = 0$) and no duration dependence (i.e., $\beta_d(y, d) = 0$). These restrictions imply that the continuation values, $v_\theta\,(1, d_t + 1)$ and $v_\theta\,(0)$, become zero, and the term $\beta(1, y_{t-1}, d_t) - \beta(0, y_{t-1}, d_t)$ becomes equal to $\beta_y(1, 0) - y_{t-1}\,[\beta_y(1, 0) + \beta_y(0, 1)]$. We can present this model using the more standard representation,

$$y_t = 1\left\{\widetilde{\alpha}_\theta + \widetilde{\beta}_y\; y_{t-1} + \widetilde{\varepsilon}_t \geq 0\right\} \tag{11}$$

with $\widetilde{\alpha}_\theta \equiv \alpha_\theta(1) - \alpha_\theta(0) + \beta_y(1, 0)$, $\widetilde{\beta}_y \equiv -\beta_y(1, 0) - \beta_y(0, 1)$, and $\widetilde{\varepsilon}_t \equiv \varepsilon_t(1) - \varepsilon_t(0)$. In a model of market entry-exit, the parameter $\widetilde{\beta}_y$ represents the sum of the costs of entry and exit, or equivalently the sunk cost of entry. This is an important structural parameter.

Define function $\sigma_\theta(y_{t-1}) \equiv -\ln\left(1+\exp\left\{\widetilde{\alpha}_\theta + \widetilde{\beta}_y y_{t-1}\right\}\right)$. The log-probability of the choice history $\widetilde{\mathbf{y}}$ conditional on (y_0, θ) is:

$$\ln \mathbb{P}\left(\widetilde{\mathbf{y}} \mid y_0, \theta\right) \quad = \quad \sum_{t=1}^{T} y_t \left[\widetilde{\alpha}_\theta + \widetilde{\beta}_y y_{t-1}\right] + (1-y_{t-1})\ \sigma_\theta(0) + y_{t-1}\ \sigma_\theta(1) \tag{12}$$

Proposition 1 establishes (i) the sufficient statistic, (ii) minimal sufficiency, and (iii) identification for this model.

PROPOSITION 1. In the myopic binary choice model without duration dependence the log-probability of a choice history has the form

$$\ln \mathbb{P}\left(\widetilde{\mathbf{y}} \mid y_0, \theta\right) \quad = \quad T^{(1)}\, g_{\theta,1} + \Delta^{(1)}\, g_{\theta,2} + \widetilde{\beta}_y\ D^{(1,1)} \tag{13}$$

with $g_{\theta,1} \equiv \widetilde{\alpha}_\theta + \sigma_\theta(1) - \sigma_\theta(0)$, *and* $g_{\theta,2} \equiv \sigma_\theta(0) - \sigma_\theta(1)$, *such that* $U = \{T^{(1)}, \Delta^{(1)}\}$, $S = D^{(1,1)}$, *and* $\beta^* = \widetilde{\beta}_y$. *We have that: (i)* $U = \{T^{(1)}, \Delta^{(1)}\}$ *is a sufficient statistic; (ii)* $T^{(1)}$ *and* $\Delta^{(1)}$ *are linearly independent such that* U *is a minimal sufficient statistic; and (iii) for* $T \geq 3$ *there is a pair of histories* $\{y_0|\widetilde{\mathbf{y}}\}$, *say* A *and* B, *with* $U(A) = U(B)$ *and* $S(A) \neq S(B)$ *such that the parameter* $\widetilde{\beta}_y$ *is identified as* $\left[\ln \mathbb{P}\left(A|U\right) - \ln \mathbb{P}\left(B|U\right)\right] / \left[D_A^{(1,1)} - D_B^{(1,1)}\right]$. *For instance, with* $T = 3$, $A = \{0|0, 1, 1\}$ *and* $B = \{0|1, 0, 1\}$. ■

This Proposition 1 is almost identical to the identification result in Chamberlain (1985). Chamberlain shows that the vector of statistics $\{T^{(1)}, y_0,\ y_T\}$ is sufficient for θ, and conditional on this vector the parameter $\widetilde{\beta}_y$ is identified. Our Proposition 1 shows that Chamberlain's sufficient statistic is not minimal and the minimal statistic is $\{T^{(1)}, y_T - y_0\}$. However, it turns out that, in this binary choice model, the extra variation left by the minimal sufficient statistic does not help in the identification of $\widetilde{\beta}_y$, so the two CMLEs are equivalent.

3.2.2 Forward-looking dynamic model without duration dependence

Consider a forward-looking version of the model in equation (10) but still without duration dependence. Since the model is of forward-looking behavior, now we have the continuation values $v_\theta\left(1, d_t + 1\right) - v_\theta\left(0\right)$. However, there is not duration dependence, and the only state variable is y_{t-1}. Therefore, for this version of the model we have that $v_\theta\left(1, d_t + 1\right) - v_\theta\left(0\right)$ becomes $v_\theta(1) - v_\theta(0) \equiv$

$\widetilde{v}_\theta$, i.e., continuation values depend on current choices but not on the current state variable y_{t-1}. We can represent this model as,

$$y_t = 1\{\widetilde{\alpha}_\theta + \widetilde{v}_\theta + \widetilde{\beta}_y y_{t-1} + \widetilde{\varepsilon}_t \geq 0\} \tag{14}$$

The only difference between this model and the myopic model is that now the fixed effect has two components: $\widetilde{\alpha}_\theta$ that comes from current profit, and $\widetilde{v}_\theta$ that comes from the continuation values. However, from the point of view of identification and estimation, the two models are observationally equivalent.

A key feature of this model, that determines the observational equivalence with the myopic model, is the property that the state variable at period $t+1$ depends on the choice at period t but not on the state variable at period t, i.e., $x_{t+1} = y_t$.

Proposition 2 establishes this equivalence.

PROPOSITION 2. In the forward-looking binary choice model without duration dependence the log-probability of a choice history has the form

$$\ln \mathbb{P}(\widetilde{\mathbf{y}} \mid y_0, \theta) \;=\; T^{(1)}\, g_{\theta,1} + \Delta^{(1)}\, g_{\theta,2} + \widetilde{\beta}_y \; D^{(1,1)} \tag{15}$$

with $g_{\theta,1} \equiv \widetilde{\alpha}_\theta + \widetilde{v}_\theta + \sigma_\theta(1) - \sigma_\theta(0)$, *and* $g_{\theta,2} \equiv \sigma_\theta(0) - \sigma_\theta(1)$, *such that* $U = \{T^{(1)}, \Delta^{(1)}\}$, $S = D^{(1,1)}$, *and* $\beta^* = \widetilde{\beta}_y$. *We have that: (i)* $U = \{T^{(1)}, \Delta^{(1)}\}$ *is a sufficient statistic; (ii)* $T^{(1)}$ *and* $\Delta^{(1)}$ *are linearly independent such that* U *is a minimal sufficient statistic; and (iii) for* $T \geq 3$ *there is a pair of histories* $\{y_0|\widetilde{\mathbf{y}}\}$, *say* A *and* B, *with* $U(A) = U(B)$ *and* $S(A) \neq S(B)$ *such that the parameter* $\widetilde{\beta}_y$ *is identified as* $[\ln \mathbb{P}(A|U) - \ln \mathbb{P}(B|U)] \,/\, \left[D_A^{(1,1)} - D_B^{(1,1)}\right]$. ■

3.2.3 Myopic dynamic model with duration dependence

The continuation values are zero, and the term $\beta(1, y_{t-1}, d_t) - \beta(0, y_{t-1}, d_t)$ is equal to $(1 - y_{t-1})$ $\beta_y(1,0) +$ y_{t-1} $\beta_d(1, d_t) -$ $y_{t-1}\beta_y(0,1)$, and it can be represented as $\beta_y(1,0) + \widetilde{\beta}_y$ $y_{t-1} + \beta_d(1, d_t)$ y_{t-1}. Therefore, we can present this model as

$$y_t = 1\{\widetilde{\alpha}_\theta + \widetilde{\beta}_y y_{t-1} + \beta_d(1, d_t)\; y_{t-1} + \widetilde{\varepsilon}_t \geq 0\} \tag{16}$$

For this model, the log-probability of the choice history $\widetilde{\mathbf{y}}$ conditional on (y_0, d_1, θ) is:

$$\ln \mathbb{P}(\widetilde{\mathbf{y}} \mid y_0, d_1, \theta) \;=\; \sum_{t=1}^{T} y_t \left[\widetilde{\alpha}_\theta + \widetilde{\beta}_y\; y_{t-1} + \beta_d(1, d_t)\; y_{t-1}\right] + \sigma_\theta(y_{t-1}, d_t) \tag{17}$$

where $\sigma_\theta(y_{t-1}, d_t) \equiv -\ln\left(1 + \exp\left\{\widetilde{\alpha}_\theta + \widetilde{\beta}_y\ y_{t-1} + \beta_d(1, d_t)\ y_{t-1}\right\}\right)$. In order to emphasize that $\sigma_\theta(y_{t-1}, d_t)$ does not depend on d_t when $y_{t-1} = 0$, we use the notation $\sigma_\theta(0)$ to represent $\sigma_\theta(0, 0)$.

Proposition 3 establishes the minimal sufficient statistic and identification of structural parameters in this model.

PROPOSITION 3. In the myopic binary choice model with duration dependence under Assumption 1, the log-probability of a choice history has the form

$$\ln \mathbb{P}\left(\widetilde{\mathbf{y}}|y_0, d_1\right) \quad = \quad \sum_{d\geq 1} H^{(1)}(d)\ g_{\theta,1}(d) + \Delta^{(1)}\ g_{\theta,2} + \sum_{d\geq 1} \Delta^{(1)}(d)\ \gamma(d-1) \tag{18}$$

with $g_{\theta,1}(d) \equiv \widetilde{\alpha}_\theta + \sigma_\theta(1, d) - \sigma_\theta(0) + \gamma(d-1)$, $g_{\theta,2} \equiv \widetilde{\alpha}_\theta$, $\gamma(d) \equiv \widetilde{\beta}_y + \beta_d(1, d)$, *and* $\gamma(0) = 0$, *such that* $U = \{H^{(1)}(d) : d \geq 1,\ \Delta^{(1)}\}$, $S = \{\Delta^{(1)}(d) : d \geq 1\}$, *and* $\beta^* = \{\gamma(d) : d \geq 1\}$. *Then, we have that: (i)* U *is a sufficient statistic. (ii) The elements in the vector* U *are linearly independent such that* U *is a minimal sufficient statistic. (iii) Conditional on* U, *the statistics* $\{\Delta^{(1)}(d) : d \geq 1\}$ *have variation and the structural parameters* $\{\gamma(d) : 1 \leq d \leq T-2\}$ *are identified, i.e., for any* $1 \leq d \leq T-2$, *there is a pair of histories,* A *and* B, *such that* $U(A) = U(B)$ *and* $\gamma(d) = \ln \mathbb{P}(A|U) - \ln \mathbb{P}(B|U)$. ■

Proof. The derivation of equation (18) is in the Appendix. Proof of (iii). For any duration n, with $1 \leq n \leq T-2$, define a sub-history $\{y_0, d_1 \mid y^{n+2}\}$, and consider the sub-histories $A = \{0, 0 \mid 0, \mathbf{1}_{n+1}\}$ and $B = \{0, 0 \mid \mathbf{1}_n, 0, 1\}$, where $\mathbf{1}_n$ represents a sequence of n consecutive $1'$s. The corresponding histories of durations $\{d_t : t = 1, ..., n+2\}$ are: for A, $\{0, 0, 1, ..., n\}$; and for B, $\{0, 1, ..., n, 0\}$. It is clear that the histogram of durations is the same under the two histories: $H^{(1)}_A(d) = H^{(1)}_B(d) = 1$ for any $1 \leq d \leq n$, and $H^{(1)}_A(d) = H^{(1)}_B(d) = 0$ for $d \geq n+1$. Also, $\Delta^{(1)}_A = y_{n+2,A} - y_{0,A} = 1$ and $\Delta^{(1)}_B = y_{n+2,B} - y_{0,B} = 1$. Therefore, we conclude that $U(A) = U(B)$. For the statistics associated to the structural parameters: $d_{1,A} = 0$ and $d_{n+3,A} = n+1$, such that $\Delta^{(1)}_A(n+1) = 1$ and $\Delta^{(1)}_A(d) = 0$ for any $d \neq n+1$; $d_{1,B} = 0$ and $d_{n+3,B} = 1$, such that $\Delta^{(1)}_B(d) = 0$ for any $d \geq 2$. Therefore, $\ln \mathbb{P}(A|U) - \ln \mathbb{P}(B|U) = [\Delta^{(1)}_A(n+1) - \Delta^{(1)}_B(n+1)]\ \gamma(n) = \gamma(n)$, and this structural parameter is identified. ■

For this model, the vector of sufficient statistics include the histogram of durations, $\{H^{(1)}(d) : d \geq 1\}$. Conditional on these statistics, the identification of the structural parameter $\gamma(d)$ comes from the difference between the final and the initial value of duration, $\Delta^{(1)}(d+1) = 1\{d_{T+1} =$

$d+1\}- 1\{d_1 = d+1\}$. The identification result in Proposition 3 for the myopic model with duration dependence does not depend on Assumption 2.

In this binary choice model, the parameters $\widetilde{\beta}_y$ and $\beta_d(1,n)$ cannot be separately identified. However, given the parameters $\{\gamma(d) : 1 \leq d \leq T-2\}$, we can identify the marginal returns to experience $\beta_d(1,d) - \beta_d(1,d-1)$ as $\gamma(d) - \gamma(d-1)$ for any value d between 2 and $T-2$.[19]

3.2.4 Forward-looking dynamic model with duration dependence

Now, the optimal decision rule includes the difference of continuation values $v_\theta(1, d_t+1) - v_\theta(0)$. Therefore, the model is:

$$y_t = 1\left\{\widetilde{\alpha}_\theta + \widetilde{\beta}_y \; y_{t-1} + \beta_d(1,d_t) \; y_{t-1} + v_\theta(1, d_t+1) + \widetilde{\varepsilon}_t \geq 0\right\} \tag{19}$$

where now $\widetilde{\alpha}_\theta \equiv \alpha_\theta(1) - \alpha_\theta(0) + \beta_y(1,0) - v_\theta(0)$. For this model, the log-probability of the choice history $\widetilde{\mathbf{y}}$ conditional on (y_0, d_1, θ) is:

$$\ln \mathbb{P}(\widetilde{\mathbf{y}} \mid y_0, d_1, \theta) \;\; = \;\; \sum_{t=1}^{T} y_t \left[\widetilde{\alpha}_\theta + \widetilde{\beta}_y y_{t-1} + \beta_d(1,d_t) y_{t-1} + v_\theta(1, d_t+1)\right] + \sigma_\theta(y_{t-1}, d_t) \tag{20}$$

with $\sigma_\theta(y_{t-1}, d_t) \equiv -\ln(1+ \exp\{\widetilde{\alpha}_\theta + \widetilde{\beta}_y y_{t-1} + \beta_d(1,d_t) y_{t-1} + v_\theta(1, d_t+1)\})$. Comparing equation (20) with (17) we can see the forward looking model has the additional term $\sum_{t=1}^{T} y_t \; v_\theta(1, d_t+1)$.

Proposition 4 establishes that under Assumption 1 (and without Assumption 2) there is not identification of any structural parameter.

PROPOSITION 4. In the forward-looking binary choice model with duration dependence under Assumption 1, the log-probability of a choice history has the following form

$$\ln \mathbb{P}(\widetilde{\mathbf{y}} | y_0, d_1, \theta) \;\; = \;\; \sum_{d \geq 1} H^{(1)}(d) \; g_{\theta,1}(d) + \sum_{d \geq 1} \Delta^{(1)}(d) \; g_{\theta,2}(d) \tag{21}$$

with $g_{\theta,1}(d) \equiv \widetilde{\alpha}_\theta + \sigma_\theta(1,d) - \sigma_\theta(0) + \gamma(d-1) + v_\theta(1,d)$, $g_{\theta,2}(d) \equiv \widetilde{\alpha}_\theta + v_\theta(1,d) + \gamma(d-1)$, $\gamma(d) \equiv \widetilde{\beta}_y + \beta_d(1,d)$, *and* $\gamma(0) = 0$, *such that* $S = \{\Delta^{(1)}(d) : d \geq 1\}$ *and* $U = \{H^{(1)}(d) : d \geq 1, \Delta^{(1)}(d) : d \geq 1\}$. *The minimal sufficient statistic* U *includes the whole vector* S, *and therefore, the structural parameters* $\gamma(d)$ *are not identified.* ■

[19] In this binary choice model with both switching costs and duration dependence, it is not possible to separately identify the switching cost parameter $\widetilde{\beta}_y$ and the level of the return to experience $\beta_d(1,d)$. This result resembles the under-identification of the autoregressive of the order two model studied by Chamberlain (1985). In that model, we have $y_{it} = 1\{\widetilde{\alpha}_i + \beta_1 \; y_{i,t-1} + \beta_2 \; y_{i,t-2} + \widetilde{\varepsilon}_{it} \geq 0\}$. Chamberlain shows that the parameter β_2 is identified but the parameter β_1 is not.

In terms of the minimal sufficient statistic, the difference between this forward-looking model and its myopic counterpart is that now we need to control for the difference between final and initial duration, $\Delta^{(1)}(d+1)$. These additional statistics are also the only statistics associated with the structural parameter $\gamma(d)$. Therefore, after controlling for the vector of sufficient statistics U, there is not variation left that can identify structural parameters in this model.

The under-identification result in Proposition 4 applies to the model under Assumption 1 but without Assumption 2. Under Assumption 2, continuation values are such that $v_\theta(1,d) = v_\theta(1,d^*)$ for any $d \geq d^*$. This property provides identification of some structural parameters. Proposition 5 establishes this result.

PROPOSITION 5. In the forward-looking binary choice model with duration dependence under Assumptions 1 and 2, the log-probability of a choice history has the following form

$$\begin{aligned}
\ln \mathbb{P}(\widetilde{\mathbf{y}}|y_0,\theta) \quad = \quad & \sum_{d \leq d^*-1} H^{(1)}(d)\ g_{\theta,1}(d) + \left[\sum_{d \geq d^*} H^{(1)}(d)\right] g_{\theta,1}(d^*) \\
+ \quad & \sum_{d \leq d^*-1} \Delta^{(1)}(d)\ g_{\theta,2}(d) + \left[\sum_{d \geq d^*} \Delta^{(1)}(d)\right] g_{\theta,2}(d^*) \\
+ \quad & \Delta^{(1)}(d^*)\ \left[\beta_d(1,d^*-1) - \beta_d(1,d^*)\right]
\end{aligned} \tag{22}$$

with $g_{\theta,1}(d) \equiv \widetilde{\alpha}_\theta + \sigma_\theta(1,d) - \sigma_\theta(0) + \gamma(d-1) + v_\theta(1,d)$, and $g_{\theta,2}(d) \equiv \widetilde{\alpha}_\theta + v_\theta(1,d) + \gamma(d-1)$. We have that: (i) $U = \{H^{(1)}(d) : d \leq d^*-1,\ \sum_{d \geq d^*} H^{(1)}(d),\ \Delta^{(1)}(d) : d \leq d^*-1,\ \sum_{d \geq d^*} \Delta^{(1)}(d)\}$ is a sufficient statistic for θ. (ii) The elements in the vector U are linearly independent such that U is a minimal sufficient statistic. (iii) Conditional on U, the statistic $\Delta^{(1)}(d^*)$ has variation and the structural parameter $\Delta\beta_d(d^*) \equiv \beta_d(1,d^*) - \beta_d(1,d^*-1)$ is identified, i.e., there is a pair of histories, A and B, such that $U(A) = U(B)$ and$\Delta\beta_d(d^*) = [\ln \mathbb{P}(A|U) - \ln \mathbb{P}(B|U)]/[\Delta_A^{(1)}(d^*) - \Delta_B^{(1)}(d^*)]$. ■

Proof. The derivation of equation (22) is in the Appendix. Proof of (iii). Given a choice history $\{y_0, d_1 \mid \widetilde{\mathbf{y}}\}$ consider the sub-history $\{y_0, d_1 \mid y_1, y_2, ..., y_{2d^*+1}\}$. Consider the choice histories $A = \{0, 0 \mid \mathbf{1}_{d^*-1}, 0, \mathbf{1}_{d^*+1}\}$ and $B = \{0, 0 \mid \mathbf{1}_{d^*}, 0, \mathbf{1}_{d^*}\}$. The corresponding histories of durations $\{d_t : t = 1, ..., 2d^*+1\}$ are: for A, $\{0,\ 1,\ 2, ...,\ d^*-1,\ 0,\ 1,\ 2,\ ..., d^*\}$; and for B, $\{0,\ 1,\ 2,\ ...,\ d^*,\ 0,\ 1,\ 2,\ ...,\ d^*-1\}$. We verify that $U(A) = U(B)$: (a) for any $d \leq d^*-1$, $H_A(d) = H_B(d) = 2$; (b) $\sum_{d \geq d^*} H_A(d) = \sum_{d \geq d^*} H_B(d) = 1$; (c) for any $d \leq d^*-1$, $\Delta_A^{(1)}(d) = \Delta_B^{(1)}(d) = 0$; and

(d) $\sum_{d \geq d^*} \Delta_A^{(1)}(d) = \sum_{d \geq d^*} \Delta_B^{(1)}(d) = 1$. The two histories have different values for the statistic $\Delta^{(1)}(d^*)$, i.e., $\Delta_A^{(1)}(d^*) = 0$ and $\Delta_B^{(1)}(d^*) = 1$. Therefore, $\ln \mathbb{P}(A|U) - \ln \mathbb{P}(B|U) = [\Delta_A^{(1)}(d^*) - \Delta_B^{(1)}(d^*)] \, [-\Delta\beta_d(d^*)] = \Delta\beta_d(d^*)$. ■

In the forward-looking binary choice model with duration dependence, only $\Delta\beta_d(d^*)$ is identified. This result contrasts with the myopic model where we can identify $\Delta\beta_d(d)$ for any duration $2 \leq d \leq T-1$ (Proposition 3).

Table 2 summarizes the identification results for the binary choice model.

Table 2
Identification of Dynamic Binary Logit Models

Panel 1: Models without duration dependence

Myopic Model			*Forward-Looking Model*		
Minimal sufficient stat.	Identified parameters	Identifying statistics	Minimal sufficient stat.	Identified parameters	Identifying statistics
$T^{(1)}$, $\Delta^{(1)}$	$\widetilde{\beta}_y$	$D^{(1,1)}$	$T^{(1)}$, $\Delta^{(1)}$	$\widetilde{\beta}_y$	$D^{(1,1)}$

Panel 2: Models with duration dependence

Myopic Model			*Forward-Looking Model*		
Minimal sufficient stat.	Identified parameters	Identifying statistics	Minimal sufficient stat.	Identified parameters	Identifying statistics
$\Delta^{(1)}$, $H^{(1)}(d) : d \geq 1$	$\widetilde{\beta}_y + \beta_d(1,d)$ for $d \leq T-2$	$\Delta^{(1)}(d)$	$H^{(1)}(d) : d \leq d^*-1$, $\sum_{d \geq d^*} H^{(1)}(d)$, $\Delta^{(1)}(d) : d \leq d^*-1$, $\sum_{d \geq d^*} \Delta^{(1)}(d)$	$\Delta\beta_d(d^*) \equiv \beta_d(1,d^*) - \beta_d(1,d^*-1)$	$\Delta^{(1)}(d^*)$

Identification of d^* in the forward-looking model. We have assumed so far that the value of d^* is known to the researcher. We now establish the identification of d^*. Let n be any duration such that $2n+1 \leq T$. Consider the pair of histories $A_n = \{0, 0 \mid \mathbf{1}_{n-1}, 0, \mathbf{1}_{n+1}\}$ and $B_n = \{0, 0 \mid \mathbf{1}_n, 0, \mathbf{1}_n\}$. We have that:

$$\begin{cases} \text{For } n > d^*, \quad U(A_n) = U(B_n), \text{ and } \ln \mathbb{P}(A_n|U) - \ln \mathbb{P}(B_n|U) = \Delta\beta_d(n) = 0 \\ \\ \text{For } n = d^*, \quad U(A_n) = U(B_n), \text{ and } \ln \mathbb{P}(A_n|U) - \ln \mathbb{P}(B_n|U) = \Delta\beta_d(d^*) \neq 0 \\ \\ \text{For } n < d^*, \quad U(A_n) \neq U(B_n) \end{cases} \tag{23}$$

Note that $\ln \mathbb{P}(A_n|U_n) - \ln \mathbb{P}(B_n|U_n)$ identifies the parameter $\Delta\beta_d(n)$ only if $n \geq d^*$. Given a dataset with T time periods, we can construct histories A_n and B_n only if $2n+1 \leq T$. Putting these two conditions together, the identification of the value of d^* requires that $T \geq 2d^*+1$ or equivalently, $d^* \leq (T-1)/2$. Under this condition, we can describe the parameter d^* as the maximum value of n such that $\ln \mathbb{P}(A_n|U_n) - \ln \mathbb{P}(B_n|U_n) \neq 0$. This condition uniquely identifies d^*.

PROPOSITION 6. Consider the forward-looking binary choice model with duration dependence under Assumptions 1 and 2. For any duration n with $2n+1 \leq T$, define the pair of histories $A_n = \{0, 0 \mid \mathbf{1}_{n-1}, 0, \mathbf{1}_{n+1}\}$ and $B_n = \{0, 0 \mid \mathbf{1}_n, 0, \mathbf{1}_n\}$. Then, if $d^* \leq (T-1)/2$, we have that the value of d^* is point identified as:

$$d^* = \max\{n : \ \ln \mathbb{P}(A_n|U_n) - \ln \mathbb{P}(B_n|U_n) \neq 0\} \qquad \blacksquare \tag{24}$$

3.3 Multinomial choice models

3.3.1 Multinomial myopic model without duration dependence

We can represent this model as $y_t = \arg\max_{y \in \mathcal{Y}}\{\alpha_\theta(y) + \beta_y(y, y_{t-1}) + \varepsilon_t(y)\}$. The log-probability of the choice history $\widetilde{\mathbf{y}}$ conditional on (y_0, θ) is:

$$\ln \mathbb{P}(\widetilde{\mathbf{y}}|y_0, \theta) \quad = \quad \sum_{t=1}^{T} [\alpha_\theta(y_t) + \beta_y(y_t, y_{t-1})] + \sigma_\theta(y_{t-1}) \tag{25}$$

where $\sigma_\theta(y_{t-1}) \equiv -\ln\left[\sum_{y=0}^{J} \exp\{\alpha_\theta(y) + \beta_y(y, y_{t-1})\}\right]$. Proposition 7 presents our identification result for this model.

PROPOSITION 7. In the myopic multinomial model without duration dependence under Assumption 1, the log-probability has the following form

$$\ln \mathbb{P}(\widetilde{\mathbf{y}}|y_0, \theta) \quad = \quad \sum_{y=1}^{J} T^{(y)} \ g_{\theta,1}(y) + \sum_{y=1}^{J} \Delta^{(y)} \ g_{\theta,2}(y) + \sum_{y_{-1}=1}^{J}\sum_{y=1}^{J} D^{(y_{-1},y)} \ \widetilde{\beta}_y(y, y_{-1}) \tag{26}$$

where $g_{\theta,1}(y) \equiv \alpha_\theta(y) - \alpha_\theta(0) + \sigma_\theta(y) - \sigma_\theta(0) + \beta_y(0, y) + \beta_y(y, 0)$, $g_{\theta,2}(y) \equiv -\sigma_\theta(y) + \sigma_\theta(0) - \beta_y(0, y)$, and $\widetilde{\beta}_y(y, y_{-1}) \equiv \beta_y(y, y_{-1}) - \beta_y(0, y_{-1}) - \beta_y(y, 0)$ for any $y, y_{-1} \in \mathcal{Y}$. Then: (i) $U = \{T^{(y)} : y \geq 1, \ \Delta^{(y)} : y \geq 1\}$ is a sufficient statistic for θ. (ii) The elements in the vector U are linearly independent such that U is a minimal sufficient statistic. (iii) Conditional on U, the vector of statistics $\{D^{(y_{-1},y)} : y_{-1}, y \in \mathcal{Y} - \{0\}\}$ are linearly independent such that they can identify the vector *of parameters* $\{\widetilde{\beta}_y(y, y_{-1}) : y_{-1}, y \in \mathcal{Y} - \{0\}\}$*, i.e., for every pair of choices* $y_{-1}, y \in \mathcal{Y} - \{0\}$*, there is a pair of histories,* A *and* B*, such that* $U(A) = U(B)$ *and* $\widetilde{\beta}_y(y, y_{-1}) = [\ln \mathbb{P}(A|U) - \ln \mathbb{P}(B|U)]/[D_A^{(y_{-1},y)} - D_B^{(y_{-1},y)}]$. ■

The following example illustrates a pair of histories that identifies $\widetilde{\beta}_y(y, y_{-1})$.

EXAMPLE 1. Suppose that $T = 3$ and consider the following two realizations of the history $(y_0|\widetilde{\mathbf{y}})$: $A = \{0 \mid 0, j, k\}$ and $B = \{0 \mid j, 0, k\}$ with $j, k \neq 0$. We first confirm that $U(A) = U(B)$: $T_A^{(j)} = T_B^{(j)} = 1$, $T_A^{(k)} = T_B^{(k)} = 1$, and $T_A^{(y)} = T_B^{(y)} = 0$ for any $y \neq 0, j, k$. The identifying statistics $D^{(y_{-1},y)}$ take the following values: $D_A^{(j,k)} - D_B^{(j,k)} = 1$, $D_A^{(j,0)} - D_B^{(j,0)} = -1$, $D_A^{(0,k)} - D_B^{(0,k)} = -1$, and $D_A^{(y_{-1},y)} - D_B^{(y_{-1},y)} = 0$ for any other pair (y_{-1}, y). Therefore, we have that $\ln \mathbb{P}(A|U) - \ln \mathbb{P}(B|U) = \widetilde{\beta}_y(k, j) - \widetilde{\beta}_y(0, j) - \widetilde{\beta}_y(k, 0) = \widetilde{\beta}_y(k, j)$. A particular case of this example is when $j = k$, such that $A = \{0 \mid 0, j, j\}$ and $B = \{0 \mid j, 0, j\}$. In this case, $\ln \mathbb{P}(A|U) - \ln \mathbb{P}(B|U)$ identifies $\widetilde{\beta}_y(j, j)$ that is equal to the sunk cost $-\beta_y(0, j) - \beta_y(j, 0)$. ■

As in the binary choice model, we cannot identify the whole switching cost function β_y. With $J + 1$ choice alternatives, we can identify J^2 switching cost parameters. However, the structural parameter $\widetilde{\beta}_y(y, y_{-1}) \equiv \beta_y(y, y_{-1}) - \beta_y(0, y_{-1}) - \beta_y(y, 0)$ has a clear interpretation: it is the difference in switching cost between a *direct switch* from y_{-1} to y and an *indirect switch* via alternative 0. For this identification result, there is nothing special with alternative 0 and we could choose any other alternative as the baseline. Note also that the set of identified structural parameters $\widetilde{\beta}_y(y, y_{-1})$ includes the *sunk cost of entry* in market y, i.e., for any $y > 0$, $\widetilde{\beta}_y(y, y) = -\beta_y(0, y) - \beta_y(y, 0)$, because $\beta_y(y, y) = 0$.

3.3.2 Multinomial forward-looking model without duration dependence

The optimal decision rule for this model is $y_t = \arg\max_{y \in \mathcal{Y}} \{\alpha_\theta(y) + v_\theta(y) + \beta_y(y, y_{t-1}) + \varepsilon_t(y)\}$, where $v_\theta(y)$ is the continuation value of choosing alternative y. The log-probability of the choice history $\widetilde{\mathbf{y}}$ conditional on (y_0, θ) has a similar form as in the myopic model, but now the incidental parameter θ enters through the function $\alpha_\theta(y) + v_\theta(y)$.

$$\ln \mathbb{P}(\widetilde{\mathbf{y}}|y_0, \theta) \quad = \quad \sum_{t=1}^{T} [\alpha_\theta(y_t) + v_\theta(y_t) + \beta_y(y_t, y_{t-1})] + \sigma_\theta(y_{t-1}) \tag{27}$$

Therefore, the identification of the structural parameters is the same as in the myopic model without duration dependence.

PROPOSITION 8. In the multinomial forward-looking model without duration dependence under Assumption 1, the log-probability of a choice history has the following form

$$\ln \mathbb{P}(\widetilde{\mathbf{y}}|y_0, \theta) = \sum_{y=1}^{J} T^{(y)}\ g_{\theta,1}(y) + \sum_{y=1}^{J} \Delta^{(y)}\ g_{\theta,2}(y) + \sum_{y_{-1}=1}^{J}\sum_{y=1}^{J} D^{(y_{-1},y)}\ \widetilde{\beta}_y(y, y_{-1}) \tag{28}$$

where $g_{\theta,1}(y) \equiv \alpha_\theta(y) - \alpha_\theta(0) + v_\theta(y) - v_\theta(0) + \sigma_\theta(y) - \sigma_\theta(0) + \beta_y(0,y) + \beta_y(y,0)$*, and* $g_{\theta,2}(y) \equiv \sigma_\theta(y) - \sigma_\theta(0) - \beta_y(0,y)$*. Then: (i)* $U = \{T^{(y)} : y \geq 1,\ \Delta^{(y)} : y \geq 1\}$ *is a sufficient statistic for* θ*. (ii) The elements in the vector* U *are linearly independent such that* U *is a minimal sufficient statistic. (iii) Conditional on* U*, the vector of statistics* $\{D^{(y_{-1},y)} : y_{-1}, y \in \mathcal{Y} - \{0\}\}$ *are linearly independent such that they can identify the vector of parameters* $\{\widetilde{\beta}_y(y, y_{-1}) : y_{-1}, y \in \mathcal{Y} - \{0\}\}$*.*
■

3.3.3 Multinomial myopic model with duration dependence

The model is $y_t = \arg\max_{y \in \mathcal{Y}}\{\alpha_\theta(y) + 1\{y \neq y_{t-1}\}\ \beta_y(y, y_{t-1}) + 1\{y = y_{t-1}\}\ \beta_d(y, d_t) + \varepsilon_t(y)\}$, and the log-probability of a choice history $\widetilde{\mathbf{y}}$ conditional on (y_0, d_1, θ) is:

$$\ln \mathbb{P}(\widetilde{\mathbf{y}}|y_0, d_1, \theta) = \sum_{t=1}^{T} [\alpha_\theta(y_t) + 1\{y_t \neq y_{t-1}\}\beta_y(y_t, y_{t-1}) + 1\{y_t = y_{t-1}\}\beta_d(y_t, d_t)] + \sigma_\theta(y_{t-1}) \tag{29}$$

Proposition 9 presents identification results for the structural parameters β_y and β_d.

PROPOSITION 9. In the multinomial myopic model with duration dependence under Assumption 1, the log-probability of a choice history has the form

$$\begin{aligned} \ln \mathbb{P}(\widetilde{\mathbf{y}}|y_0, d_1) = & \sum_{y=1}^{J}\sum_{d\geq 1} H^{(y)}(d)\ g_{\theta,1}(y,d) + \sum_{y=1}^{J} \Delta^{(y)}\ g_{\theta,2}(y) \\ & + \sum_{y_{-1}=1}^{J}\ \sum_{y=1, y\neq y_{-1}}^{J} D^{(y_{-1},y)}\ \widetilde{\beta}_y(y, y_{-1}) + \sum_{y=1}^{J}\sum_{d\geq 1} \Delta^{(y)}(d)\ \gamma(y, d-1) \end{aligned} \tag{30}$$

with $g_{\theta,1}(y,d) \equiv \alpha_\theta(y) - \alpha_\theta(0) + \sigma_\theta(y,d) - \sigma_\theta(0) + \beta_y(0,y) + \beta_y(y,0) + \gamma(y, d-1)$*,* $g_{\theta,2}(y) \equiv \alpha_\theta(y) - \alpha_\theta(0) + \beta_y(y,0)$*,* $\widetilde{\beta}_y(y, y_{-1}) \equiv \beta_y(y, y_{-1}) - \beta_y(0, y_{-1}) - \beta_y(y, 0)$*, and* $\gamma(y,d) \equiv \beta_d(y,d) - \beta_y(y,0) - \beta_y(0,y)$*. Then: (i)* $U = \{H^{(y)}(d) : y \geq 1,\ d \geq 1,\ \Delta^{(y)} : y \geq 1\}$ *is a sufficient statistic of* θ*. (ii) The elements in the vector* U *are linearly independent such that* U *is a minimal sufficient statistic. (iii) Conditional on* U*, the vector of statistics* $\{D^{(y_{-1},y)} : y_{-1},\ y \geq 1;\ \Delta^{(y)}(d) : y \geq 1,\ d \geq 1\}$ *are linearly independent and they identify the vectors of structural parameters* $\{\widetilde{\beta}_y(y, y_{-1}) : y_{-1},\ y \geq 1,\ y \neq y_{-1};\ \gamma(y,d) : y \geq 1,\ d \geq 1\}$*.* ■

The following examples present choice histories that identify structural parameters $\widetilde{\beta}_y(y, y_{-1})$ and $\gamma(y, d)$ according to Proposition 9.

EXAMPLE 2. Suppose that $T = 3$ and consider two realizations of the history $(y_0, d_1|\widetilde{\mathbf{y}})$: for $j \neq k$, $A = \{0, 0 \mid 0, j, k\}$ and $B = \{0, 0 \mid j, 0, k\}$. It is straightforward to verify that $U(A) = U(B)$ and that $\ln \mathbb{P}(A|U) - \ln \mathbb{P}(B|U) = \beta_y(k, j) - \beta_y(k, 0) - \beta_y(0, j) = \widetilde{\beta}_y(k, j)$. ■

EXAMPLE 3. Given an arbitrary positive integer n, consider the pair of choice histories $(y_0, d_1|\widetilde{\mathbf{y}})$ with $T = n + 2$: $A = \{0, 0 \mid 0, \mathbf{y}_{n+1}\}$ and $B = \{0, 0 \mid \mathbf{y}_n, 0, y\}$, where $\mathbf{y}_n$ represents a vector of dimension n with all its elements equal to y. It is simple to verify that $U(A) = U(B)$ (i.e., same values for $H^{(y)}(d)$ and $\Delta^{(y)}$). Furthermore, $\Delta_A^{(y)}(n+1) = 1$ and $\Delta_B^{(y)}(n+1) = 0$, such that we have $\ln \mathbb{P}(A|U) - \ln \mathbb{P}(B|U) = \gamma(y, n)$. ■

3.3.4 Multinomial forward-looking model with duration dependence

The model is $y_t = \arg\max_{y \in \mathcal{Y}}\{\alpha_\theta(y) + \beta_y(y, y_{t-1}) + 1\{y = y_{t-1}\}\beta_d(y, d_t) + v_\theta(y, d_{t+1}[y, y_{t-1}, d_t]) + \varepsilon_t(y)\}$, where $d_{t+1}[y, y_{t-1}, d_t] = 0$ if $y = 0$, and $d_{t+1}[y, y_{t-1}, d_t] = 1\{y = y_{t-1}\}d_t + 1$ if $y \neq 0$. In contrast to the binary choice model, in the multinomial choice model it is possible to identify switching cost parameters without imposing Assumption 2. Proposition 10 establishes the identification of switching costs parameters under Assumption 1.

PROPOSITION 10. In the multinomial forward-looking model with duration dependence under Assumption 1, the log-probability a choice history has the form

$$\begin{aligned} \ln \mathbb{P}(\widetilde{\mathbf{y}}|y_0, d_1) \quad = \quad & \sum_{y=1}^{J} \sum_{d \geq 1} H^{(y)}(d) \; g_{\theta,1}(y, d) + \sum_{y=1}^{J} \sum_{d \geq 1} \Delta^{(y)}(d) \; g_{\theta,2}(y, d) \\ + \quad & \sum_{y_{-1}=1}^{J} \sum_{y \neq y_{-1}} D^{(y_{-1}, y)} \; \widetilde{\beta}_y(y, y_{-1}) \end{aligned} \tag{31}$$

with $g_{\theta,1}(y, d) \equiv \alpha_\theta(y) - \alpha_\theta(0) + \sigma_\theta(y, d) - \sigma_\theta(0) + \beta_y(0, y) + \beta_y(y, 0) + v_\theta(y, d) - v_\theta(0) + \gamma(y, d-1)$, and $g_{\theta,2}(y, d) \equiv \alpha_\theta(y) - \alpha_\theta(0) + \beta_y(0, y) + v_\theta(y, d) - v_\theta(0) + \gamma(y, d-1)$. Then: (i) $U = \{H^{(y)}(d) : y \geq 1$, $d \geq 1$, $\Delta^{(y)}(d) : y \geq 1$, $d \geq 1\}$ is a sufficient statistic of θ. (ii) The elements in the vector U are linearly independent such that U is a minimal sufficient statistic. (iii) Conditional on U, the vector of statistics $\{D^{(y_{-1}, y)} : y_{-1}, y \geq 1\}$ are linearly independent and they identify the vectors of structural parameters $\{\widetilde{\beta}_y(y, y_{-1}) : y_{-1}, y \geq 1, y \neq y_{-1}\}$. The duration dependence parameters $\gamma(y, d)$ are not identified. ■

For instance, the pair of choice histories in Example 2, $A = \{0,0 \mid 0,j,k\}$ and $B = \{0,0 \mid j,0,k\}$, have the same continuation values. In this forward-looking model, it is simple to very that these histories satisfy the conditions in Proposition 10 such that $U(A) = U(B)$ and $\ln \mathbb{P}(A|U) - \ln \mathbb{P}(B|U) = \widetilde{\beta}_y(k,j)$.

For the identification of duration dependence parameters, we impose the restriction in Assumption 2. Proposition 11 presents this identification result.

PROPOSITION 11. In the multinomial forward-looking model with duration dependence under Assumptions 1 and 2, the log-probability a choice history has the form

$$\begin{aligned} \ln \mathbb{P}(\widetilde{\mathbf{y}}|y_0,d_1) &= \sum_{y=1}^{J} \sum_{d \leq d_y^*-1} H^{(y)}(d)\; g_{\theta,1}(y,d) + \left[\sum_{y=1}^{J} \sum_{d \geq d_y^*} H^{(y)}(d)\right] g_{\theta,1}(y,d_y^*) \\ & \sum_{y=1}^{J} \sum_{d \leq d_y^*-1} \Delta^{(y)}(d)\; g_{\theta,2}(y,d) + \left[\sum_{y=1}^{J} \sum_{d \geq d_y^*} \Delta^{(y)}(d)\right] g_{\theta,2}(y,d_y^*) \\ &+ \sum_{y_{-1}=1}^{J} \sum_{y=1, y \neq y_{-1}}^{J} D^{(y_{-1},y)}\; \widetilde{\beta}_y(y,y_{-1}) - \sum_{y=1}^{J} \Delta^{(y)}(d_y^*)\; \Delta\beta_d(y,d_y^*) \end{aligned} \tag{32}$$

with $g_{\theta,1}(y,d) \equiv \alpha_\theta(y) - \alpha_\theta(0) + \sigma_\theta(y,d) - \sigma_\theta(0) + \beta_y(0,y) + \beta_y(y,0) + v_\theta(y,d) - v_\theta(0) + \gamma(y,d-1)$*, and* $g_{\theta,2}(y,d) \equiv \alpha_\theta(y) - \alpha_\theta(0) + \beta_y(y,0) + v_\theta(y,d) - v_\theta(0) + \gamma(y,d-1)$*, and* $\Delta\beta_d(y,d_y^*) \equiv \beta_d(y,d_y^*) - \beta_d(y,d_y^*-1)$*. (i)* $U = \{H^{(y)}(d) : y \geq 1,\ d \leq d_y^*-1,\ \sum_{d \geq d_y^*} H^{(y)}(d),\ \Delta^{(y)}(d) : y \geq 1,\ d \leq d_y^*-1,\ \sum_{d \geq d_y^*} \Delta^{(y)}(d)\}$ *is a sufficient statistic of* θ*. (ii) The elements in the vector* U *are linearly independent such that* U *is a minimal sufficient statistic. (iii) Conditional on* U*, the vector of statistics* $\{D^{(y_{-1},y)} : y_{-1},\ y \geq 1\}$ *are linearly independent and they identify the vector of structural parameters* $\{\widetilde{\beta}_y(y,y_{-1}) : y_{-1},\ y \geq 1, y \neq y_{-1}\}$*. Furthermore, the vector of statistics* $\{\Delta^{(y)}(d^*) : y \geq 1\}$ *are also linearly independent and they identify the vector of structural parameters* $\{\Delta\beta_d(y,d_y^*) : y \geq 1\}$*.* ■

EXAMPLE 4. Given $y \geq 1$ with $d_y^* \geq 2$, consider the pair of choice histories $A = \left\{0,0 \mid \mathbf{y}_{d_y^*-1}, 0, \mathbf{y}_{d_y^*+1}\right\}$ and $B = \left\{0,0 \mid \mathbf{y}_{d_y^*}, 0, \mathbf{y}_{d_y^*}\right\}$. The two choice histories have the same statistics $H^{(y)}(d)$ for all $1 \leq d \leq d_y^*-1$ and $\sum_{d \geq d_y^*} H^{(y)}(d)$, and $\min\{d_1, d_y^*\}$ and $\min\{d_{T+1}, d_y^*\}$ agrees between A and B. Therefore, we have that $U(A) = U(B)$. It is straightforward to show that $\Delta_A^{(y)}(d_y^*) = 0$ and $\Delta_B^{(y)}(d_y^*) = 1$, and this implies that $\ln \mathbb{P}(A|U) - \ln \mathbb{P}(B|U) = \Delta\beta_d(y,d_y^*)$. ■

Table 3 summarizes the identification results for the multinomial model.

Table 3
Identification of Dynamic Multinomial Logit Models

Panel 1: Models without duration dependence

Myopic Model			*Forward-Looking Model*		
Minimal sufficient stat.	Identified parameters	Identifying statistics	Minimal sufficient stat.	Identified parameters	Identifying statistics
$T^{(y)}$, $\Delta^{(y)}$: $y \geq 1$	$\widetilde{\beta}_y(y, y_{-1})$ $y_{-1}, y \geq 1$	$D^{(y_{-1},y)}$: $y_{-1}, y \geq 1$	$T^{(y)}$, $\Delta^{(y)}$: $y \geq 1$	$\widetilde{\beta}_y(y, y_{-1})$ $y_{-1}, y \geq 1$	$D^{(y_{-1},y)}$ $y_{-1}, y \geq 1$

Panel 2: Models with duration dependence

Myopic Model			*Forward-Looking Model*		
Minimal sufficient stat.	Identified parameters	Identifying statistics	Minimal sufficient stat.	Identified parameters	Identifying statistics
$\Delta^{(y)}$: $y \geq 1$, $H^{(y)}(d)$: $y \geq 1, d \geq 1$	$\widetilde{\beta}_y(y, y_{-1})$: $y_{-1}, y \geq 1$ and $\gamma(y, d)$: $y \geq 1, d \geq 1$	$D^{(y_{-1},y)}$: $y_{-1}, y \geq 1$ and $\Delta^{(y)}(d)$: $y \geq 1, d \geq 1$	$H^{(y)}(d)$: $y \geq 1, d \leq d_y^* - 1$; $\sum_{d \geq d^*} H^{(y)}(d) : y \geq 1$; $\Delta^{(y)}(d)$: $y \geq 1, d \leq d_y^* - 1$; $\sum_{d \geq d^*} \Delta^{(y)}(d) : y \geq 1$	$\widetilde{\beta}_y(y, y_{-1})$: $y_{-1} \neq y \geq 1$ and $\Delta\beta_d(y, d_y^*)$: $y \geq 1$	$D^{(y_{-1},y)}$ $y_{-1} \neq y \geq 1$ and $\Delta^{(y)}(d_y^*) : y \geq 1$

3.4 Identification of the distribution of unobserved heterogeneity

In empirical applications of dynamic structural models, the answer to some important empirical questions requires the identification of the distribution of the unobserved heterogeneity. For instance, the researcher can be interested in the *average marginal effects* $\int [\partial P_\theta (y|\ \mathbf{x}, \beta^*) / \partial \mathbf{x}]\ f(\theta)\ d\theta$ or $\int [\partial P_\theta (y|\ \mathbf{x}, \beta^*) / \partial \beta^*]\ f(\theta)\ d\theta$, where $f(\theta)$ is the density function of the unobserved heterogeneity. Without further restrictions, the density function $f(\theta)$ is not (nonparametrically) point identified, i.e., initial conditions problem. In this section, we briefly describe this identification problem, and two possible approaches that the researcher can take to deal with this problem: (a) nonparametric finite mixture; and (b) set identification.

Let $f(\theta \mid \mathbf{x}_1)$ be the density function of θ conditional on the initial value of the state variables $\mathbf{x}_1 \equiv (y_0, d_1)$. After the identification/estimation of the structural parameters, β^*, the model implies the following restrictions for the identification of $f(\theta \mid \mathbf{x}_1)$. For any choice history $\widetilde{\mathbf{y}}$, we have that:

$$\mathbb{P}(\widetilde{\mathbf{y}}|\mathbf{x}_1) = \int \left[\prod_{t=1}^{T} P(y_t \mid \mathbf{x}_t, \beta^*, \theta) \right] f(\theta|\mathbf{x}_1)\ d\theta \tag{33}$$

The probabilities of choice histories $\mathbb{P}(\widetilde{\mathbf{y}}|\mathbf{x}_1)$ are identified from the data. Also, for a fixed value of θ, the probabilities $P(y_t \mid \mathbf{x}_t, \beta^*, \theta)$ are also known to the researcher after the identification of the structural parameters β^*. Therefore, the identification of the density function $f(\theta|\mathbf{x}_1)$ can be seen as the solution to a system of linear equations.

Let $|\Theta|$ be the dimension of the support of θ. This dimension can be infinite. Equation (33) can be written in vector form as,

$$\mathbb{P}_{\mathbf{x}_1} = \mathbf{L}_{\mathbf{x}_1}\ \mathbf{f}_{\mathbf{x}_1} \tag{34}$$

$\mathbb{P}_{\mathbf{x}_1}$ is a vector of dimension $(J+1)^T \times 1$ with the probabilities of all the possible choice histories with initial conditions $\mathbf{x}_1$. $\mathbf{L}_{\mathbf{x}_1}$ is a matrix with dimension $(J+1)^T \times |\Theta|$ such that each row contains the probabilities $\prod_{t=1}^{T} P(y_t \mid \mathbf{x}_t, \beta^*, \theta)$ for a given choice history and for every value of θ. Finally, $\mathbf{f}_{\mathbf{x}_1}$ is a $|\Theta| \times 1$ vector with the probabilities $f(\theta|\mathbf{x}_1)$. Given this representation, it is clear that $\mathbf{f}_{\mathbf{x}_1}$ is point identified if and only if matrix $\mathbf{L}_{\mathbf{x}_1}$ is full column rank.

If the distribution of θ is continuous, then $|\Theta| = \infty$ and $\mathbf{L}_{\mathbf{x}_1}$ cannot be full-column rank. In fact, the number of rows in matrix $\mathbf{L}_{\mathbf{x}_1}$ (i.e., the number of possible choice histories, $(J+1)^T$) provides an upper bound to the dimension of the support $|\Theta|$ for which the density is nonparametrically (point) identified. The researcher may be willing to impose the restriction that the support of θ is discrete such that matrix $\mathbf{L}_{\mathbf{x}_1}$ is full column rank. Under this condition, $\mathbf{f}_{\mathbf{x}_1}$ can be identified as the linear projection:

$$\mathbf{f}_{\mathbf{x}_1} = \left[\mathbf{L}'_{\mathbf{x}_1}\mathbf{L}_{\mathbf{x}_1}\right]^{-1}\mathbf{L}'_{\mathbf{x}_1}\mathbb{P}_{\mathbf{x}_1} \tag{35}$$

Note that the estimator β^* is still a fixed-effect estimator that is robust to this finite-mixture restriction on the distribution of the unobservables. However, under this approach, the estimation of marginal effects depends on this assumption. Alternatively, the researcher may prefer not to impose this finite support restriction and set-identify the distribution of the unobservables. This is the approach in Chernozhukov, Fernandez-Val, Hahn, and Newey (2013).

Finally, we would like to comment on a practical issue in the implementation of the finite-mixture estimation described above. For the evaluation of the choice probabilities $P\left(y_t \mid \mathbf{x}_t, \beta^*, \theta\right)$ in matrix $\mathbf{L}_{\mathbf{x}_1}$, the vector of unobserved heterogeneity θ is multidimensional. That is, we need to choose a grid of points for the parameters $\alpha_\theta(y)$ but also for the continuation values $v_\theta(y,d)$. In the forward-looking model without duration dependence, unobserved heterogeneity enters through the term $\tau_\theta(y) \equiv \alpha_\theta(y) + v_\theta(y)$. Therefore, for this model we need to fix a grid of points for the J incidental parameters $\{\tau_\theta(y) : y > 1\}$. Using a grid of κ points for each parameter $\tau_\theta(y)$ we have that the dimension of the density vector $\mathbf{f}_{\mathbf{x}_1}$ is $|\Theta| = \kappa^J$ that should be smaller that $(J+1)^T$ in order to have identification. In the forward-looking model with duration dependence, unobserved heterogeneity enters through the term $\tau_\theta(y,d) \equiv \alpha_\theta(y)+v_\theta(y,d)$. Therefore, we need to fix a grid of points for the JT incidental parameters $\{\tau_\theta(y,d) : y > 1;\, 1 \leq d \leq T\}$. Using a grid of κ points for each parameter $\tau_\theta(y,d)$ we have that the dimension of $\mathbf{f}_{\mathbf{x}_1}$ is $|\Theta| = \kappa^{JT}$ that should be smaller that $(J+1)^T$. This is a strong restriction on the dimension of unobserved heterogeneity, κ. However, this approach is not taking into account that the continuation values $v_\theta(y,d)$ are endogenous objects that can be obtained given $\alpha'_\theta s$ and β^* by solving the Bellman equation of the model. Taking into account this structure of the model, we can reduce substantially the dimensionality of θ. Given a value of the J incidental parameters $\{\alpha_\theta(y) : y > 1\}$, we can solve the Bellman equation to obtain all the continuation values $v_\theta(y,d)$. Therefore, the dimension of θ in the structural model with duration dependence is also equal to the dimension of $\{\alpha_\theta(y) : y > 1\}$, as in the model without duration dependence.

4 Estimation and Inference

Since the identification is based on the conditional MLE approach, the estimator for the structural parameters of interest (β_y, β_d) will be an Andersen (1970) type of estimator. We illustrate the estimator for the forward-looking multinomial choice model with duration dependence under Assumption 1 and 2, since estimators for the structural parameters in the other models can be defined in a similar fashion.

4.1 Estimation of β^* (given d^*)

Let $\beta^* = \{\widetilde{\beta}_y', \gamma'\}'$ be the vector of identified structural parameters. Let U_i be the vector of sufficient statistics (associated to θ), and let S_i be the vector of identifying statistics associated to β^*. Then, the conditional MLE for $\beta^*$is defined as the maximizer of the conditional log-likelihood function:

$$\mathcal{L}_N(\beta^*) = \sum_{i=1}^{N} \mathcal{L}_i(\beta^*) = \sum_{i=1}^{N} S_i'\beta^* - \left(\sum_{j:U(j)=U_i} \exp\left\{S(j)'\beta^*\right\} \right) \tag{36}$$

where the condition $\{j : U(j) = U_i\}$ represents all the choice histories $(y_0, d_1, \widetilde{\mathbf{y}})$ with the same value of U as observation i. This log-likelihood function is globally concave in β^*, and therefore the computation of the CMLE is straightforward using Newton-Raphson or BHHH algorithm. Using standard arguments (Newey and McFadden, 1994), we have

$$\sqrt{N}(\widehat{\beta}^* - \beta^*) \Rightarrow \mathcal{N}(0, J(\beta^*)^{-1}) \tag{37}$$

The consistent estimator for the Fisher information is $J_N(\widehat{\beta}^*) = -N^{-1}\sum_{i=1}^{N} \nabla_{\beta\beta}\mathcal{L}_i(\widehat{\beta}^*)$.

4.2 Estimation of d^*

We describe here a CML estimator for the joint estimation of (d^*, β^*). Let d_0^* represent the true value of the parameter d^*. And let $\beta_0(n)$ be the true value of the parameter $\beta(n) \equiv \beta_d(y, n) - \beta_d(y, n-1)$. By definition, we have that $\beta_0(d_0^*) \neq 0$ and $\beta_0(n) = 0$ for any $n > d_0^*$. For notational simplicity, we use β^* and β_0^* to represent $\beta(d^*)$ and and $\beta_0(d_0^*)$, respectively. We are interested in the joint identification of (d_0^*, β_0^*) from the maximization of the conditional likelihood function.

Based on Proposition 6, we consider the following representation of the sufficient statistic U. For any $2 \leq n \leq L$ with $L \leq (T-1)/2$, define the pair of histories $A_n = \{0, 0|\mathbf{y}_{n-1}, 0, \mathbf{y}_{T-n}\}$ and $B_n = \{0, 0|\mathbf{y}_n, 0, \mathbf{y}_{T-n-1}\}$. Then, $U_i = \{\mathbf{y}_i \in A_n \cup B_n \text{ for some } 2 \leq n \leq L\}$. Given this statistic, the conditional likelihood function is:

$$\mathcal{L}_N(\nu) \quad = \quad \sum_{n=2}^{L}\sum_{i=1}^{N} 1\{\mathbf{y}_i = A_n\} \ln\left[\frac{\exp\{\nu(n)\}}{1+\exp\{\nu(n)\}}\right] + 1\{\mathbf{y}_i = B_n\} \ln\left[\frac{1}{1+\exp\{\nu(n)\}}\right] \tag{38}$$

where $\nu(n)$ is a parameter that represents the value $\beta_d(y,n)-\beta_d(y,n-1)+\int[v_\theta(y,n+1)-v_\theta(y,n)]$ $dF(\theta|\mathbf{x}_1)$, and ν is the vector of parameters $\{\nu(n) : n = 2, 3, ..., L\}$. The model implies the following relationship between the parameters $\nu(n)$ and the structural parameters (d^*, β^*).

$$\nu(n) = \begin{cases} \text{unrestricted} & \text{if} \quad n < d^* \\ \\ \beta^* & \text{if} \quad n = d^* \\ \\ 0 & \text{if} \quad n > d^* \end{cases} \tag{39}$$

The unconstrained likelihood function $\mathcal{L}_N(\nu)$ is globally concave in each of the parameters $\nu(n)$. It is straightforward to show that the unconstrained CML estimator of $\nu(n)$ is $\widehat{v}(n) = \ln\widehat{\mathbb{P}}(A_n) - \ln\widehat{\mathbb{P}}(B_n)$, where $\widehat{\mathbb{P}}(A_n)$ and $\widehat{\mathbb{P}}(B_n)$ are the sample frequencies $N^{-1}\sum_{i=1}^{N}1\{\mathbf{y}_i = A_n\}$ and $N^{-1}\sum_{i=1}^{N}1\{\mathbf{y}_i = B_n\}$, respectively. For a given value of d^*, let $\widehat{\nu}^c_{d^*}$ be the constrained estimator of ν that imposes the restriction in equation (39) such that: $\widehat{\nu}^c_{d^*}(n) = \widehat{\nu}(n)$ (unconstrained) for $n \leq d^*$; and $\widehat{\nu}^c_{d^*}(n) = 0$ (constrained) for $n > d^*$. Furthermore, the estimator of the structural parameter β^* is $\widehat{\beta^*} = \widehat{\nu}(d^*)$.

Let $\ell_N(d^*)$ be the concentrated likelihood function $\ell_N(d^*) \equiv \mathcal{L}_N(\widehat{\nu}^c_{d^*})$, i.e., the value of the likelihood given a value of d^* and where the parameters ν have been estimated under this restriction. By definition, we have that:

$$\begin{aligned} \ell_N(d^*) \quad &= \quad N\sum_{n=2}^{d^*}\widehat{\mathbb{P}}(A_n)\ \ln\left[\frac{\widehat{\mathbb{P}}(A_n)}{\widehat{\mathbb{P}}(A_n)+\widehat{\mathbb{P}}(B_n)}\right] + \widehat{\mathbb{P}}(B_n)\ \ln\left[\frac{\widehat{\mathbb{P}}(B_n)}{\widehat{\mathbb{P}}(A_n)+\widehat{\mathbb{P}}(B_n)}\right] \\ &+ \quad N\sum_{n=d^*+1}^{L}\widehat{\mathbb{P}}(A_n)\ \ln\left[\frac{1}{2}\right] + \widehat{\mathbb{P}}(B_n)\ \ln\left[\frac{1}{2}\right] \end{aligned} \tag{40}$$

The following Proposition 12 establishes some properties of this concentrated likelihood function.

PROPOSITION 12. (A) As $N \rightarrow \infty$, the concentrated likelihood function $N^{-1}\ell_N(d^*)$ converges uniformly in d^* to its population counterpart $\ell_0(d^*)$. (B) $\ell_0(d^*_0) > \ell_0(d^*)$ for any $d^* < d^*_0$, and $\ell_0(d^*_0) = \ell_0(d^*)$ for any $d^* > d^*_0$. Therefore, d^*_0 is point identified as the minimum value of d^* that maximizes the concentrated likelihood function: $d^*_0 = \min\{n : n \in \arg\max_{2\leq d^*\leq L}\ \ell_0(d^*)\}$. ■

Given this result, a possible estimator for d^*_0 would be the sample analog $\widehat{d^*} = \min\{n : n \in \arg\max_{2\leq d^*\leq L}\ \ell_N(d^*)\}$. However, this estimator has a an important limitation in finite samples.

Though the population likelihood function $\ell_0(d^*)$ is flat for values of d^* greater than the true d_0^*, in a finite sample this likelihood increases with d^* and reaches its maximum at the largest possible value of d^*, i.e., $d^* = L$. This is because any value of d^* smaller than L implies restrictions on the parameters $\nu(n)$ of the model, i.e., $\nu(n) = 0$ for $n > d^*$. The larger the value of d^*, the smaller the number of these restrictions and the largest the value of the likelihood function in a finite sample.

To deal with this problem we consider an estimator of d_0^* that maximizes the *Bayesian Information Criterion* (BIC). This criterion function introduces a penalty that increases with the number of free parameters $\{v(n)\}$ in the model. In this model, the number of free parameters is equal to d^*. The BIC function is defined as:

$$BIC_N(d^*) = \ell_N(d^*) - \frac{d^*}{2}\ln(N) \tag{41}$$

Our estimator of d_0^* is defined as the value of d^* that maximizes $BIC_N(d^*)$.

PROPOSITION 13. Consider the estimator $\widehat{d_N^*} = \arg\max_{2\leq d^*\leq L} BIC_N(d^*)$. *As* $N \to \infty$, $\mathbb{P}(\widehat{d_N^*} = d_0^*) \to 1$. ■

The joint estimation of (d^*, β^*) has the analogy of model selection where d^* determines the model dimension and β^* is the parameter of interest. We can use standard inference for the CML estimator for β^* in this joint estimation method since Proposition 13 shows that $\widehat{d_N^*}$ is a consistent estimator for d_0^*. This is in the same spirit that under consistent model selection: the asymptotic property of the estimator for parameters in the selected model is unaffected (see Pötscher, 1991). However, Pötscher (1991) also pointed out that inference for parameters post model selection can be problematic in finite samples if the parameter is too close to zero and the true model is not selected with probability close to one. In our Monte Carlo experiments, we found that the probability of selecting the true d_0^* is very close to 1 throughout different data generating processes.[20]

[20]For example, for DGP 1 with Sample B, described in Table 5, out of 1000 repetition, 99% of the times $\widehat{d_N^*}$ agrees with the true d_0^*.

5 Empirical Application

Here we revisit the model and data in the seminal article by Rust (1987). The model belongs to the class of *machine replacement models* that we have briefly described in section 2. The superintendent of maintenance at the Madison (Wisconsin) Metropolitan Bus Company has a fleet of N buses indexed by i. For every bus i and at every period t, the superintendent decides whether to keep the bus engine ($y_{it} = 1$) or to replace it ($y_{it} = 0$). In Rust's model, if the engine is replaced, the payoff is equal to $-RC + \varepsilon_{it}(0)$, where RC is a parameter that represents the replacement cost. If the manager decides to keep the engine, the payoff is equal to $-c_0 - c_1(m_{it}) + \varepsilon_{it}(1)$, where m_{it} is a state variable that represents the engine cumulative mileage, and $c_0 + c_1(m_{it})$ is the maintenance cost. We incorporate two modifications in this model. First, we replace cumulative mileage m_{it} with duration since last replacement, d_{it}. The transition rule for this state variable is $d_{it+1} = y_{it}[d_{it} + 1]$, such that $d_{it} \in \{0, 1, 2, ...\}$. Using Rust's actual data, the correlation between the variables m_{it} and d_{it} is 0.9552. Second, we allow for time-invariant unobserved heterogeneity in the replacement cost, RC_i, and in the constant term in the maintenance cost function, c_{0i}. Using our notation, the payoff function is $\alpha_i(0) + \varepsilon_{it}(0)$ if $y_{it} = 0$ (replacing the engine), and $\alpha_i(1) + \beta_d(d_{it}) + \varepsilon_{it}(1)$ if $y_{it} = 1$ (keeping the engine), where $\alpha_i(0) = -RC_i$, $\alpha_i(1) = -c_{0i}$, and $\beta_d(d_{it}) = -c_1(d_{it})$.

In section 5.1, we present evidence from several Monte Carlo experiments using this model. The purpose of these experiments is threefold. First, showing that the FE-CMLE provides precise and robust estimates of structural parameters, even when the sample size is not large. Second, showing that the bias of misspecifying the distribution of the unobserved heterogeneity. And third, showing that a Hausman test, based on the comparison of the FE-CMLE and a CRE-MLE, has enough power to reject specifications that wrongly ignore unobserved heterogeneity, or that misspecified its probability distribution or its joint distribution with the initial conditions of the state variables. In section 5.2, we apply the FE-CMLE method, our procedure to estimate d^*, and the Hausman test to the actual dataset in Rust (1987).

5.1 Monte Carlo experiments

We present experiments using simulated data from four different Data Generating Processes (DGPs). Table 4 describes these DGPs. The difference between the four DGPs is in the specification of the distribution of the unobserved heterogeneity for the replacement cost RC_i. In DGP 1, the distribution of the replacement cost is normal with mean 8 and standard deviation 2. In DGPs 2 and 3, this distribution has only two types. Finally, DGP 4 is a model without unobserved heterogeneity.

For each of these DGPs, we do not estimate the model using the whole sample of $T = 25$ periods. Instead, we construct three samples: sample A, from period 1 to 7; sample B, from period 1 to 14; and Sample C, from period 8 to 21. Therefore, we present results from 12 Monte Carlo experiments, i.e., four DGPs times 3 samples. We analyze the effect of increasing the number of time periods T, by comparing the experiments with sample A (with $T = 7$) and sample B (with $T = 14$). We study the effect of the initial conditions problem by comparing the experiments for sample B (where at $t = 1$ all the buses have the same initial condition, $(y_{i0}, d_{i1}) = (0, 0)$) and sample C, that is subject to the initial conditions problem.

Table 4
Description of DGPs in the Monte Carlo experiments

Parameter / Constant	*DGP 1*	*DGP 2*	*DGP 3*	*DGP 4*
$\alpha_i(0) = -RC_i$	$N(\mu, \sigma^2)$	Two types	Two types	1 type
Random draws from:	$\mu = 8, \sigma = 2$	$RC_1 = 4.5, RC_2 = 9$	$RC_1 = 8, RC_2 = 9$	$RC = 8$
		$\lambda_1 = \lambda_2 = 0.5$	$\lambda_1 = \lambda_2 = 0.5$	
$\alpha_i(1) = -c_{0i}$	0	0	0	0
$\beta_d(d) = \beta\ d$ if $d \leq d^*$	$\beta = 1$	$\beta = 1$	$\beta = 1$	$\beta = 1$
d^*	3	3	3	3
Discount factor (δ)	0.95	0.95	0.95	0.95
Initial y_0, d_1	$0, 0$	$0, 0$	$0, 0$	$0, 0$
Maximum T	25	25	25	25
N *(number of buses)*	1000	1000	1000	1000
# simulated samples	1000	1000	1000	1000

The structural parameter of interest is parameter β in the maintenance cost function, $\beta_d(d) = \beta\ d$. We apply four estimators to each of the samples: the *FE-CMLE* using the true value of d^* (that we denote as *CMLE-true-d**); *FE-CMLE* using the BIC estimator of d^* (that we denote as *CMLE-BIC-d**); a MLE that imposes the restriction of no unobserved heterogeneity (that we denote as *MLE-noUH*), and a MLE that assumes that there are two types of replacement costs and ignores the potential initial conditions problem (that we denote as *MLE-2types*). We compare the bias and variance of these estimators.[21] We also implement two Hausman tests: a test of the null hypothesis

[21] The code for this experiment is in Matlab. For the two ML estimators, we use the Nested Fixed Point Algorithm.

of no unobserved heterogeneity, that compares estimators *CMLE-BIC-d** and *MLE-noUH*; and a test of the null hypothesis of two-types, that compares estimators *CMLE-BIC-d** and *MLE-2types*. We present the results of these experiments in tables 5 to 8, one table for each DGP.

Table 5
Monte Carlo Experiments with DGP 1 (Normal RCs)

Estimator of β	**Sample A** ($t = 1$to 7) Estimate[(1)] Mean	Median	St. dev.	**Sample B** ($t = 1$to 14) Estimate[(1)] Mean	Median	St. dev.	**Sample C** ($t = 8$to 21) Estimate[(1)] Mean	Median	St. dev.
CMLE-true-d*	1.0073	1.0086	0.1436	0.9990	1.0003	0.0801	0.9954	0.9978	0.0731
CMLE-BIC-d*	1.0073	1.0086	0.1436	0.9935	1.0001	0.1054	0.9873	0.9971	0.1146
MLE-2types	0.9778	0.9765	0.0528	0.8956	0.8962	0.0325	0.8565	0.8554	0.0308
MLE-noUH	0.6204	0.6191	0.0295	0.5842	0.5835	0.0232	0.5444	0.5439	0.0229
Testing null hypothesis	**Frequency of Ho rejection with significance level** 1%	5%	10%	**Frequency of Ho rejection with significance level** 1%	5%	10%	**Frequency of Ho rejection with significance level** 1%	5%	10%
No Unob. Het.	0.541	0.777	0.874	0.999	1.000	1.000	1.000	1.000	1.000
Two types	0.008	0.042	0.096	0.125	0.308	0.429	0.281	0.515	0.658

Note (1): Mean, median, and standard deviation of estimated parameter over the 1,000 replications.

Table 5 deals with DGP 1, with normally distributed replacement costs. The MLEs are substantially biased, especially in sample C (with the initial conditions problem) and sample B (with large T). When T increases there are multiple spells per bus and this implies stronger correlation between observed durations and unobserved heterogeneity. This generates a larger bias of the MLE of a misspecified model. In contrast, the biases of the *CMLEs* (either with true or estimated d^*) are negligible. The BIC method provides precise estimates of d^*: in all our DGPs, the estimated value of d^* is equal to its true value for more than 95% of the Monte Carlo replications. As a result, the bias of the CMLE estimator of β with estimated d^* is very similar to the bias of the CMLE

The maximization of the log-likelihood function applies a quasi-newton method (procedure `fminunc`) using the true value of the vector of parameters as the starting value. For the MLE with 2-types, during the search algorithm we often get a singular Hessian matrix. When this happens, we switch to the BHHH method.

with *true* d^*. As expected, the CMLEs have larger variance than the MLEs, and the CMLE with estimated d^* has larger variance than the CMLE with true d^*. However, the *CMLE-BIC-d^** has a Mean Square Error (MSE, variance plus square bias) that is substantially smaller than the one of the *MLE-noUH* in the three samples, and of the *MLE-2types* in samples B and C. In sample A, the *MLE-2types* has a MSE comparable to the one of the *CMLE.* That is, in a DGP without initial conditions problem and with one duration spell for most of the buses, a misspecified random effects model with only two types has good properties. However, this is not longer the case in samples B and C. Hausman test has very strong power to reject the model without unobserved heterogeneity.[22] It has also substantial power to reject the model with two types in samples B and C. However, the rejection rates for the model with two types in sample A are practically equal to the nominal size or significance level of the test.

Table 6
Monte Carlo Experiments with DGP 2 (Two types: RC = 4.5, 9)

Estimator	**Sample A** (t = 1to 7) Estimate$^{(1)}$			**Sample B** (t = 1to 14) Estimate$^{(1)}$			**Sample C** (t = 8to 21) Estimate$^{(1)}$		
of β	Mean	Median	St. dev.	Mean	Median	St. dev.	Mean	Median	St. dev.
CMLE-true-d*	1.0094	1.0060	0.1598	1.0027	1.0033	0.0813	0.9992	0.9948	0.0813
CMLE-BIC-d*	1.0094	1.0060	0.1598	0.9952	1.0025	0.1216	0.9886	0.9941	0.1384
MLE-2types	1.0018	0.9990	0.0513	1.0007	1.0001	0.0289	0.9954	0.9941	0.0288
MLE-noUH	0.5556	0.5557	0.0229	0.5283	0.5284	0.0156	0.5009	0.5004	0.0146
Testing	Frequency of Ho rejection with significance level			Frequency of Ho rejection with significance level			Frequency of Ho rejection with significance level		
null hypothesis	1%	5%	10%	1%	5%	10%	1%	5%	10%
No Unob. Het.	0.590	0.820	0.902	1.000	1.000	1.000	1.000	1.000	1.000
Two types	0.005	0.044	0.094	0.005	0.054	0.096	0.005	0.047	0.107

Note (1): Mean, median, and standard deviation of estimated parameter over the 1,000 replications.

[22] Though the distribution of types in DGP 1 is continuous, the level of unobserved heterogeneity is modest. In the distribution of RC_i, the ratio between the standard deviation and the mean is only 25%. Continuous distributions with higher variance imply higher rejection rates of the model with only two types, even in sample A.

Table 6 presents results under DGP 2, with two types of replacement costs, $RC_1 = 4.5$ and $RC_2 = 9$, with equal probabilities. In this case, the *MLE-2types* and our *CMLEs* are consistent estimators. Both estimators have negligible finite-sample biases in the three samples. As expected, the *MLE-2types* has smaller variance, especially in sample A. In the three samples, the *MLE-noUH* is still extremely biased and the Hausman test that compares this estimator with *CMLE-BIC-d** has strong power to reject the model without unobserved heterogeneity. For the rejection of the true model with two types, Hausman test exhibits a rejection rate that is practically identical to the nominal size or significance level.

Table 7
Monte Carlo Experiments with DGP 3 (Two types: RC = 8, 9)

Estimator of β	**Sample A** ($t = 1$to 7) Estimate$^{(1)}$ Mean	Median	St. dev.	**Sample B** ($t = 1$to 14) Estimate$^{(1)}$ Mean	Median	St. dev.	**Sample C** ($t = 8$to 21) Estimate$^{(1)}$ Mean	Median	St. dev.
CMLE-true-d*	1.0088	1.0058	0.1371	1.0014	1.0035	0.0744	0.9978	0.9957	0.0726
CMLE-BIC-d*	1.0088	1.0058	0.1371	0.9905	1.0026	0.1313	0.9923	0.9941	0.1040
MLE-2types	1.0111	1.0064	0.0626	1.0026	1.0012	0.0374	0.9990	0.9982	0.0389
MLE-noUH	0.9628	0.9609	0.0451	0.9576	0.9564	0.0317	0.9501	0.9492	0.0334
Testing null hypothesis	**Frequency of Ho rejection with significance level** 1%	5%	10%	**Frequency of Ho rejection with significance level** 1%	5%	10%	**Frequency of Ho rejection with significance level** 1%	5%	10%
No Unob. Het.	0.014	0.057	0.117	0.031	0.088	0.163	0.032	0.121	0.187
Two types	0.014	0.051	0.104	0.008	0.053	0.100	0.009	0.065	0.115

Note (1): Mean, median, and standard deviation of estimated parameter over the 1,000 replications.

Table 7 deals with DGP 3, that has also two types of replacement costs, but now these types are very similar: $RC_1 = 8$ and $RC_2 = 9$, with equal probabilities. The main purpose of the experiments with this DGP is to investigate the bias of the *MLE-noUH* and the power of this Hausman test in an scenario with a very modest amount of unobserved heterogeneity. Even in this scenario, for samples B and C, the bias of the *MLE-noUH* is approximately 5% of the true value of the parameter, and the Hausman test rejects the null hypothesis of no unobserved heterogeneity with probability that is more than twice the nominal size of the test.

Finally, Table 8 presents results of experiments under DGP 4 where there is not unobserved heterogeneity and $RC = 8$. The purpose of these experiments is to study possible biases in the size of Hausman test for the null hypothesis of no unobserved heterogeneity. We can see that, for the three samples, the size of this test is very close to the nominal size.

Table 8
Monte Carlo Experiments with DGP 4 (No UH, RC = 8)

Estimator of β	**Sample A** ($t = 1$to 7) Estimate(1) Mean	Median	St. dev.	**Sample B** ($t = 1$to 14) Estimate(1) Mean	Median	St. dev.	**Sample C** ($t = 8$to 21) Estimate(1) Mean	Median	St. dev.
CMLE-true-d*	1.0030	1.0029	0.1237	0.9979	0.9942	0.0660	0.9994	0.9994	0.0660
CMLE-BIC-d*	1.0030	1.0029	0.1237	0.9900	0.9937	0.1140	0.9889	0.9986	0.1201
MLE-2types	1.0203	1.0156	0.0513	1.0070	1.0063	0.0312	1.0079	1.0061	0.0318
MLE-noUH	1.0011	1.0004	0.0414	1.0001	0.9990	0.0293	1.0017	1.0005	0.0302
Testing null hypothesis	**Frequency of Ho rejection with significance level 1%**	**5%**	**10%**	**Frequency of Ho rejection with significance level 1%**	**5%**	**10%**	**Frequency of Ho rejection with significance level 1%**	**5%**	**10%**
No Unob. Het.	0.007	0.045	0.094	0.009	0.05	0.097	0.014	0.052	0.108
Two types	0.008	0.056	0.104	0.012	0.063	0.109	0.019	0.053	0.107

Note (1): Mean, median, and standard deviation of estimated parameter over the 1,000 replications.

5.2 Estimation using Rust's dataset

Rust's full sample contains a total of 124 buses that are classified in eight groups according to the bus size and the engine manufacturer. For the estimation of the structural model, Rust focuses on groups 1 to 4 that account for 104 buses. For every bus, the choice history in the data starts with the actual initial condition of the engine, i.e., the first month where the engine was installed. For these 104 buses, the distribution of the number of engine replacements is the following: 0 engine replacements for 45 buses; 1 replacement for 58 buses; and 2 replacements for 1 bus. For the implementation of our FE-CMLE, choice histories with zero replacements do not contain any useful information. Therefore, for the CMLE we use only 59 buses. For our analysis, we consider that the frequency of the superintendent's decisions is at the annual level. Table 9 presents the empirical distribution of choice histories for the 59 buses with at least one engine replacement, of which 27 are observed during 6 years, and 32 over 10 years.

Table 9
Bus Engine Replacement (Rust, 1987)
Empirical distribution of choice histories with replacement

	Frequency		
Choice history	*Absolute*	*%*	*% cumulative*
110111	2	3.39	3.39
111011	7	11.86	15.25
111101	7	11.86	27.12
111110	11	1864	45.76
1101111111	1	1.69	47.46
1110111111	4	6.78	54.24
1111011111	2	3.39	57.63
1111101111	7	11.86	69.49
1111110111	7	11.86	81.35
1111111011	5	8.47	89.83
1111111101	3	5.08	94.91
1111111110	2	3.39	98.30
1101110111	1	1.69	100.00
Total	59	100.00	

Table 10 presents ML estimates of the model with three different specifications of the maintenance cost function $\beta_d(d)$ according to: the value of the parameter d^* (at which function $\beta_d(d)$ becomes flat); and the functional for durations smaller than d^*, i.e., linear, quadratic, and square-root. We report estimates of the replacement cost parameter and of the parameter $\beta_d^* \equiv \beta_d(d^*) - \beta_d(d^* - 1)$. We consider a model with two unobserved types. However, for all the specifications, we always converge to a model with a single type. We have tried thousands of initial values for the vector of parameters (i.e., RC_1, RC_2, λ, and β_d), and we have also estimated the model using grid search. Regardless the computational strategy, we always converge to the same estimate with only one type. The specification of the function $\beta_d(d)$ that provides the maximum value of the likelihood function is the the square-root function with a value d^* equal to six. For this specification, the estimate of the replacement cost parameter is $\widehat{RC} = 10.8566$ ($s.e. = 1.5247$), and the estimate of the parameter of β_d^* is $\widehat{\beta_d^*} = 0.3054$ ($s.e. = 0.0496$).

Table 10
Bus Engine Replacement (Rust, 1987)

Maximum Likelihood Estimates

Model		*RC*		$\beta_d^* \equiv -\Delta\beta_d(d^*)$		
$\beta_d(d)$	d^*	$\widehat{RC}$	$se\left(\widehat{RC}\right)$	$\widehat{\beta_d^*}$	$se\left(\widehat{\beta_d^*}\right)$	log-*likelihood*
Square root	3	28.2218	6.9110	2.0110	0.5149	-162.7081
$\beta_d(d) = \beta\sqrt{d}$	4	16.5364	3.0438	0.7777	0.1546	-160.7515
	5	12.8403	1.9959	0.4486	0.0774	-158.5760
	6	**10.8566**	**1.5247**	**0.3054**	**0.0496**	**-158.2108****
	7	9.6817	1.2821	0.2317	0.0372	-158.7021
	8	8.9953	1.1623	0.1909	0.0313	-159.4693
	9	8.6517	1.1183	0.1682	0.0285	-160.0868
Linear	3	18.2995	4.1695	2.0388	0.4977	-162.7529
$\beta_d(d) = \beta\ d$	4	11.4552	1.9053	0.8418	0.1566	-160.9650
	5	9.2473	1.2769	0.5103	0.0817	-158.8536
	6	**7.9817**	**0.9809**	**0.3623**	**0.0548**	**-158.8132**
	7	7.1859	0.8219	0.2856	0.0434	-159.7641
	8	6.7030	0.7411	0.2448	0.0388	-160.9912
	9	6.4612	0.7114	0.2259	0.0379	-161.9368
Square	3	13.1481	2.7300	2.1006	0.4804	-162.8699
$\beta_d(d) = \beta\ d^2$	4	8.7707	1.2806	0.9603	0.1628	-161.4943
	5	**7.3081**	**0.8850**	**0.6257**	**0.0921**	**-159.4992**
	6	6.3777	0.6844	0.4709	0.0673	-160.0882
	7	5.7404	0.5689	0.3905	0.0583	-161.9366
	8	5.3323	0.5072	0.3535	0.0578	-164.0680
	9	5.1227	0.4837	0.3515	0.0636	-165.6751

Table 11 presents estimates of the parameter $\beta_d^* \equiv \beta_d(d^*) - \beta_d(d^* - 1)$ using the CMLE and under different values of d^*. Given the observed histories in this dataset (as shown in Table 9), the parameter β_d^* is identified only under two possible values of d^*: $d^* = 3$ and $d^* = 4$.[23] We report

[23] To identify β_d^* for $d^* = 2$, we need histories with a replacement when duration is equal to 1 ($d^* - 1$). To identify β_d^* when $d^* \geq 5$, we need histories with at least 5 years without replacement both before and after an observed

the value of the concentrated log-likelihood function and of the BIC function. According to the BIC function, the estimate of d^* is $\widehat{d^*} = 3$, and the corresponding estimator of β_d^* is $\widehat{\beta_d^*} = 1.7009$ (s.e. $= 1.0244$). Note also that for $d^* = 3$, the estimate of β_d^* is significantly different to zero for a significance level of 10% parameter (p-value = 0.0968). In contrast, for $d^* = 4$, this parameter is not significantly different to zero for any standard significance level (p-value = 0.8446). Therefore, the estimate $\widehat{d^*} = 3$ and $\widehat{\beta_d^*} = 1.7009$ is consistent with the definition of d^* as the maximum duration with $\beta_d(d) - \beta_d(d-1)$ different to zero.

Table 11
Bus Engine Replacement (Rust, 1987)

Fixed-Effects-Conditional Maximum Likelihood

	β_d^*		*p-value*	*concentrated*	
d^*	$\widehat{\beta_d^*}$	$se\left(\widehat{\beta_d^*}\right)$	$H_0 : \beta_d^* = 0$	*log-likelihood*	*BIC(d^*)*
3	**1.7009**	**1.0244**	**0.0968**	**-102.1215**	**-108.2378**
4	0.1178	0.6009	0.8446	-102.1020	-110.2571

Table 12 compares the CMLE estimate of the parameter β_d^* with the corresponding MLE using the estimates in Table 10. Given the very small sample size and the corresponding large standard error of the CMLE estimates, we cannot reject the null hypothesis of no unobserved heterogeneity, despite the magnitude of the difference between MLE and CMLE estimates is important and it generates important differences in distribution of durations.

Table 12
Bus Engine Replacement (Rust, 1987)

Hausman Test of Unobserved Heterogeneity

Model	$\widehat{\beta_d^*}$ *(se)* *MLE*	$\widehat{\beta_d^*}$ *(se)* *CMLE*	*Hausman*	*p-value*
Square root	0.4548 (0.0739)	1.7009 (1.0244)	1.4873	0.2226
Linear	0.3623 (0.0549)	1.7009 (1.0244)	1.7123	0.1907
Square	0.3476 (0.0512)	1.7009 (1.0244)	1.7494	0.186

replacement. In this small sample, we do not observe these types of histories.

6 Conclusions

This paper presents the first identification results of structural parameters in forward-looking dynamic discrete choice models where the joint distribution of time-invariant unobserved heterogeneity and endogenous state variables is nonparametrically specified. This unobserved heterogeneity can have multiple components and can have continuous support. The dependence between the unobserved heterogeneity and the initial values of the state variables is also unrestricted. We consider models with two endogenous state variables: the lagged decision variable, and the time duration in the last choice. We show that structural parameters that capture switching costs are identified under mild conditions. The identification of structural parameters that capture duration dependence require additional restrictions. In particular, to obtain identification of these parameters we assume that the marginal return of an additional period of experience (duration) becomes equal to zero after a finite number of periods.

Based on our identification results, we propose tests for the validity of restricted models without unobserved heterogeneity or with a parametric specification of the correlated random effects. Our Monte Carlo experiments show that the Conditional MLE provides precise estimates of structural parameters and the specification test has strong power to reject misspecified correlated random effects models.

Appendix 1. Proofs

Proof of Lemma 2.

(i) For any $y > 0$, we have that $1\{y_{t-1} = y,\ d_t = 0\} = 0$ because $y_{t-1} > 0$ implies $d_t > 0$. Therefore, $H^{(y)}(0) = \sum_{t=1}^{T} 1\{y_{t-1} = y,\ d_t = 0\} = 0$.

(ii) For any $y > 0$, we have that $1\{y_{t-1} = y_t = y,\ d_t = 0\} = 0$ because $y_{t-1} > 0$ implies $d_t > 0$. Therefore, $X^{(y)}(0) = \sum_{t=1}^{T} 1\{y_{t-1} = y_t = y, d_t = 0\} = 0$.

(iii) For any $y > 0$, $\sum_{d\geq 1} H^{(y)}(d) = \sum_{d\geq 1}\sum_{t=1}^{T} 1\{y_{t-1} = y,\ d_t = d\} = \sum_{t=1}^{T} 1\{y_{t-1} = y\} = T^{(y)} + 1\{y_0 = y\} - 1\{y_T = y\}$.

(iv) For any $y > 0$, $\sum_{d\geq 1} X^{(y)}(d) = \sum_{d\geq 1}\sum_{t=1}^{T} 1\{y_{t-1} = y_t = y, d_t = d\} = \sum_{t=1}^{T} 1\{y_{t-1} = y_t = y\} = T^{(y)} + 1\{y_0 = y\} - 1\{y_T = y\}$.

(v) First, note that $y_{t-1} = y > 0$ implies $d_t \geq 1$. Therefore, for any $y > 0$ and $d \geq 1$, the event $\{y_{t-1} = y_t = y, d_t = d\}$ is equivalent to the event $\{y_t = y,\ d_{t+1} = d + 1\}$ for any $1 \leq t \leq T$. Therefore, $X^{(y)}(d) = \sum_{t=1}^{T}\{y_t = y, d_{t+1} = d + 1\} = \sum_{t=2}^{T+1} 1\{y_{t-1} = y, d_t = d + 1\} = H^{(y)}(d + 1) - 1\{y_0 = y, d_1 = d + 1\} + 1\{y_T = y, d_{T+1} = d + 1\}$. ■

Proof of Propositions 1 and 2. Remember that $T^{(y)}$ represents the number of times that choice alternative y is visited in the choice history $\widetilde{\mathbf{y}}$, and $D^{(y)}$ is the number of times that choice alternative y is observed at two consecutive periods over the history $(y_0, \widetilde{\mathbf{y}})$. For the binary choice model, we have that $\sum_{t=1}^{T} y_t = T^{(1)}$, $\sum_{t=1}^{T} y_{t-1} y_t = D^{(1,1)}$, and $y_T - y_0 = \Delta^{(1)}$.

$$\ln \mathbb{P}(\widetilde{\mathbf{y}} \mid y_0, \theta) \quad = \quad T^{(1)}\,[\widetilde{\alpha}_\theta + \sigma_\theta(1) - \sigma_\theta(0)] + \Delta^{(1)}\ \ [\sigma_\theta(0) - \sigma_\theta(1)] + \widetilde{\beta}_y\ D^{(1,1)} \tag{A.1}$$

where we have omitted the term $T\ \sigma_\theta(0)$ because T is constant over all the histories. Consider choice histories $A = \{0|0, 1, 1\}$ and $B = \{0|1, 0, 1\}$. It is clear that $T_A^{(1)} = T_B^{(1)} = 2$, and $\Delta_A^{(1)} = \Delta_B^{(1)} = 1$, such that $U_A = U_B$. Also, $D_A^{(1,1)} = 1$ and $D_B^{(1,1)} = 0$. Therefore, $\ln \mathbb{P}(A|U) - \ln \mathbb{P}(B|U) = \widetilde{\beta}_y$. ■

Proof of Proposition 3. The log-probability of this model is:

$$\ln \mathbb{P}(\widetilde{\mathbf{y}} \mid y_0, d_1, \theta) \quad = \quad \sum_{t=1}^{T} y_t \left[\widetilde{\alpha}_\theta + \widetilde{\beta}_y\ y_{t-1} + \beta_d(1, d_t)\ y_{t-1}\right] + \sigma_\theta(y_{t-1}, d_t) \tag{A.2}$$

We have that $\ln \mathbb{P}(\widetilde{\mathbf{y}} \mid y_0, d_1, \theta) = T^{(1)}\widetilde{\alpha}_\theta + [T^{(0)} - \Delta^{(0)}]\sigma_\theta(0) + D^{(1,1)}\widetilde{\beta}_y + \sum_{d\geq 1} X^{(1)}(d)\ \beta_d(1, d) + \sum_{d\geq 1} H^{(1)}(d)\ \sigma_\theta(1, d)$. Taking into account that $\sum_{d\geq 1} H^{(1)}(d) = T^{(1)} - \Delta^{(1)}$ and $D^{(1,1)} = \sum_{d\geq 1} X^{(1)}(d)$, we obtain:

$$\begin{aligned} \ln \mathbb{P}\left(\widetilde{\mathbf{y}}|y_0,\theta\right) &= \sum_{d\geq 1} H^{(1)}(d)\ \left[\widetilde{\alpha}_\theta + \sigma_\theta(1,d) - \sigma_\theta(0)\right] + \Delta^{(1)}\widetilde{\alpha}_\theta \\ &+ \sum_{d\geq 1} X^{(1)}(d)\ \gamma(d) \end{aligned} \tag{A.3}$$

where we have omitted the term $T\ \sigma_\theta(0)$ because T is constant over all the histories, and we define $\gamma(d) \equiv \widetilde{\beta}_y + \beta_d(1,d)$. Now, Lemma 2(v) establishes that $X^{(1)}(d) = H^{(1)}(d+1) + \Delta^{(1)}(d+1)$. Note that $\sum_{d\geq 1}\left[H^{(1)}(d+1) + \Delta^{(1)}(d+1)\right]\ \gamma(d)$ is equal to $\sum_{d\geq 1}\left[H^{(1)}(d) + \Delta^{(1)}(d)\right]\ \gamma(d-1)$, if we define $\gamma(0) = 0$. Then, we have that,

$$\begin{aligned} \ln \mathbb{P}\left(\widetilde{\mathbf{y}}|y_0,\theta\right) &= \sum_{d\geq 1} H^{(1)}(d)\ \left[\widetilde{\alpha}_\theta + \sigma_\theta(1,d) - \sigma_\theta(0)\right] + \Delta^{(1)}\widetilde{\alpha}_\theta \\ &+ \sum_{d\geq 1}\left[H^{(1)}(d) + \Delta^{(1)}(d)\right]\ \gamma(d-1) \\ &= \sum_{d\geq 1} H^{(1)}(d)\ \left[\widetilde{\alpha}_\theta + \sigma_\theta(1,d) - \sigma_\theta(0) + \gamma(d-1)\right] + \Delta^{(1)}\widetilde{\alpha}_\theta \\ &+ \sum_{d\geq 1} \Delta^{(1)}(d)\ \gamma(d-1) \qquad \blacksquare \end{aligned} \tag{A.4}$$

Proof of Proposition 4. The log-probability of this model is:

$$\ln \mathbb{P}\left(\widetilde{\mathbf{y}} \mid y_0, d_1, \theta\right) = \sum_{t=1}^{T} y_t \left[\widetilde{\alpha}_\theta + \widetilde{\beta}_y y_{t-1} + \beta_d(1,d_t) y_{t-1} + v_\theta\left(1, d_t+1\right)\right] + \sigma_\theta(y_{t-1}, d_t) \tag{A.5}$$

Comparing this log-probability with the one for the myopic model with duration, we can see that the only difference is in the term $\sum_{t=1}^{T} y_t\ v_\theta\left(1, d_t+1\right)$, that can be written as $\sum_{d\geq 0}\sum_{t=1}^{T} y_t\ 1\{d_t = d\}$ $v_\theta\left(1, d+1\right)$. For the statistic $\sum_{t=1}^{T} y_t\ 1\{d_t = d\}$ we can distinguish two cases: (a) if $d = 0$, then $\sum_{t=1}^{T} y_t\ 1\{d_t = 0\} = \sum_{t=1}^{T} y_t\ (1 - y_{t-1}) = T^{(1)} - D^{(1,1)}$; and (b) if $d \geq 1$, then $\sum_{t=1}^{T} y_t\ 1\{d_t = d\} = \sum_{t=1}^{T} y_t\ y_{t-1}\ 1\{d_t = d\} = X^{(1)}(d)$. Therefore,

$$\begin{aligned} \sum_{d\geq 0}\sum_{t=1}^{T} y_t 1\{d_t = d\} v_\theta\left(1, d+1\right) &= \left[T^{(1)} - D^{(1,1)}\right]\ v_\theta\left(1,1\right) + \sum_{d\geq 1} X^{(1)}(d)\ v_\theta\left(1, d+1\right) \\ &= T^{(1)}\ v_\theta\left(1,1\right) + \sum_{d\geq 1} X^{(1)}(d)\ \left[v_\theta\left(1, d+1\right) - v_\theta\left(1,1\right)\right] \end{aligned} \tag{A.6}$$

where for the second equality we have applied Lemma 2(iv), $D^{(1,1)} = \sum_{d\geq 1} X^{(1)}(d)$. Then, the log-probability is equal to

$$
\begin{aligned}
\ln \mathbb{P}\left(\widetilde{\mathbf{y}} | y_0, \theta\right) &= \sum_{d\geq 1} H^{(1)}(d) \;\; [\widetilde{\alpha}_\theta + \sigma_\theta(1,d) - \sigma_\theta(0) + \gamma(d-1)] + \Delta^{(1)}\widetilde{\alpha}_\theta \\
&+ \sum_{d\geq 1} \Delta^{(1)}(d) \; \gamma(d-1) \\
&+ T^{(1)} \; v_\theta\left(1,1\right) + \sum_{d\geq 1} X^{(1)}(d) \;\; [v_\theta\left(1,d+1\right) - v_\theta\left(1,1\right)]
\end{aligned} \tag{A.7}
$$

From Lemma 2, we have that: (iii) $T^{(1)} = \sum_{d\geq 1} H^{(1)}(d) + \Delta^{(1)}$; and (v) $X^{(1)}(d) = H^{(1)}(d+1)+$ $\Delta^{(1)}(d+1)$, and solving these expressions in (A.13), we have that:

$$
\begin{aligned}
\ln \mathbb{P}\left(\widetilde{\mathbf{y}} | y_0, \theta\right) &= \sum_{d\geq 1} H^{(1)}(d) \;\; [\widetilde{\alpha}_\theta + \sigma_\theta(1,d) - \sigma_\theta(0) + \gamma(d-1) + v_\theta\left(1,d\right)] \\
&+ \Delta^{(1)} \;\; [\widetilde{\alpha}_\theta + v_\theta\left(1,1\right)] \\
&+ \sum_{d\geq 1} \Delta^{(1)}(d) \;\; [v_\theta\left(1,d\right) - v_\theta\left(1,1\right) + \gamma(d-1)]
\end{aligned} \tag{A.8}
$$

Taking into account that $\Delta^{(1)} = \sum_{d\geq 1}\Delta^{(1)}(d)$, we have:

$$
\ln \mathbb{P}\left(\widetilde{\mathbf{y}} | y_0, \theta\right) = \sum_{d\geq 1} H^{(1)}(d) \; g_{\theta,1}(d) + \sum_{d\geq 1} \Delta^{(1)}(d) \;\; [\widetilde{\alpha}_\theta + v_\theta\left(1,d\right) + \gamma(d-1)] \tag{A.9}
$$

with $g_{\theta,1}(d) \equiv \widetilde{\alpha}_\theta + \sigma_\theta(1,d) - \sigma_\theta(0) + \gamma(d-1) + v_\theta\left(1,d\right)$. ■

Proof of Proposition 5. Define $Z \equiv \sum_{d\geq 1}\Delta^{(1)}(d) \; [v_\theta\left(1,d\right) + \gamma(d-1)]$. Under Assumption 2, we have that $v_\theta\left(1,d\right) = v_\theta\left(1,d^*\right)$ for any $d \geq d^*$, and $\gamma\left(d-1\right) = \gamma\left(d^*\right)$ for any $d \geq d^*+1$. Therefore, we have:

$$
\begin{aligned}
Z &= \sum_{d\leq d^*-1} \Delta^{(1)}(d) \; v_\theta\left(1,d\right) + \left[\sum_{d\geq d^*} \Delta^{(1)}(d)\right] v_\theta\left(1,d^*\right) \\
&+ \sum_{d\leq d^*} \Delta^{(1)}(d) \; \gamma(d-1) + \left[\sum_{d\geq d^*+1} \Delta^{(1)}(d)\right] \gamma(d^*) \\
&= \sum_{d\leq d^*-1} \Delta^{(1)}(d) \;\; [v_\theta\left(1,d\right) + \gamma(d-1)] + \left[\sum_{d\geq d^*} \Delta^{(1)}(d)\right] [v_\theta\left(1,d^*\right) + \gamma(d^*)] \\
&+ \Delta^{(1)}(d^*) \;\; [\gamma(d^*-1) - \gamma(d^*)]
\end{aligned} \tag{A.10}
$$

Then, the log-probability becomes:

$$\begin{aligned}\ln \mathbb{P}\left(\widetilde{\mathbf{y}}|y_0,\theta\right) &= \sum_{d\geq 1} H^{(1)}(d)\ g_{\theta,1}(d) \\ &+ \sum_{d\leq d^*-1} \Delta^{(1)}(d)\ g_{\theta,2}(d) + \left[\sum_{d\geq d^*} \Delta^{(1)}(d)\right] g_{\theta,2}(d^*) \\ &+ \Delta^{(1)}(d^*)\ \left[\gamma(d^*-1)-\gamma(d^*)\right]\end{aligned} \tag{A.11}$$

with $g_{\theta,1}(d) \equiv \widetilde{\alpha}_\theta + \sigma_\theta(1,d) - \sigma_\theta(0) + \gamma(d-1) + v_\theta\,(1,d)$, and $g_{\theta,2}(d) \equiv \widetilde{\alpha}_\theta + v_\theta\,(1,d) + \gamma(d-1)$. Note that $g_{\theta,1}(d) = g_{\theta,1}(d^*)$ for any $d \geq d^*$. Therefore, we have $\sum_{d\geq 1} H^{(1)}(d)\ g_{\theta,1}(d) = \sum_{d\leq d^*-1} H^{(1)}(d)$ $g_{\theta,1}(d)+ \left[\sum_{d\geq d^*} H^{(1)}(d)\right] g_{\theta,1}(d^*)$, such that

$$\begin{aligned}\ln \mathbb{P}\left(\widetilde{\mathbf{y}}|y_0,\theta\right) &= \sum_{d\leq d^*-1} H^{(1)}(d)\ g_{\theta,1}(d) + \left[\sum_{d\geq d^*} H^{(1)}(d)\right] g_{\theta,1}(d^*) \\ &+ \sum_{d\leq d^*-1} \Delta^{(1)}(d)\ g_{\theta,2}(d) + \left[\sum_{d\geq d^*} \Delta^{(1)}(d)\right] g_{\theta,2}(d^*) \\ &+ \Delta^{(1)}(d^*)\ \left[\gamma(d^*-1)-\gamma(d^*)\right] \qquad \blacksquare\end{aligned} \tag{A.12}$$

Proof of Propositions 7 and 8. For this model, the log probability is $\sum_{j=0}^{J}\sum_{t=1}^{T} 1\{y_t = j\}$ $\alpha_\theta(j)+ \sum_{j=0}^{J}\sum_{k=0}^{J}\sum_{t=1}^{T} 1\{y_{t-1} = j,\ y_t = k\}\ \beta_y(k,j)+ \sum_{j=0}^{J}\sum_{t=1}^{T} 1\{y_{t-1} = j\}\ \sigma_\theta(j)$. Using the definitions of our statistics, we have that:

$$\ln \mathbb{P}\left(\widetilde{\mathbf{y}}|y_0,\theta\right) = \sum_{j=0}^{J} T^{(j)} \alpha_\theta(j) + \sum_{j=0}^{J}\sum_{k=0}^{J} D^{(j,k)} \beta_y(k,j) + \sum_{j=0}^{J} \left[T^{(j)} - \Delta^{(j)}\right] \sigma_\theta(j) \tag{A.13}$$

where $\Delta^{(j)} \equiv 1\{y_T = j\} - 1\{y_0 = j\}$. Note that $T^{(0)} = T - \sum_{j=1}^{J} T^{(j)}$, and $\Delta^{(0)} = 1 - \sum_{j=1}^{J} \Delta^{(j)}$, such that:

$$\begin{aligned}\ln \mathbb{P}\left(\widetilde{\mathbf{y}}|y_0,\theta\right) &= \sum_{j=1}^{J} T^{(j)} \left[\alpha_\theta(j) - \alpha_\theta(0) + \sigma_\theta(j) - \sigma_\theta(0)\right] + \sum_{j=1}^{J} \Delta^{(j)}\ \left[-\sigma_\theta(j) + \sigma_\theta(0)\right] \\ &+ \sum_{j=0}^{J}\sum_{k=0}^{J} D^{(j,k)} \beta_y(k,j)\end{aligned} \tag{A.14}$$

where we have omitted the term $T\alpha_\theta(0) + \sigma_\theta(0)$ because it is constant over all the choice histories. For the term, $\sum_{j=0}^{J}\sum_{k=0}^{J} D^{(j,k)} \beta_y(k,j)$, note that: $\sum_{j=0}^{J} D^{(j,k)} = T^{(k)}$ such that $D^{(0,k)} = T^{(k)} -$

$\sum_{j=1}^{J} D^{(j,k)}$; and $\sum_{k=0}^{J} D^{(j,k)} = T^{(j)} - \Delta^{(j)}$ such that $D^{(j,0)} = T^{(j)} - \Delta^{(j)} - \sum_{k=1}^{J} D^{(j,k)}$. Therefore,

$$\begin{aligned}\sum_{j=0}^{J}\sum_{k=0}^{J} D^{(j,k)}\beta_y(k,j) &= \sum_{j=0}^{J}\left[\sum_{k=1}^{J} D^{(j,k)}\beta_y(k,j) + \left[T^{(j)} - \Delta^{(j)} - \sum_{k=1}^{J} D^{(j,k)}\right]\beta_y(0,j)\right] \\ &= \sum_{j=0}^{J}\sum_{k=1}^{J} D^{(j,k)}\left[\beta_y(k,j) - \beta_y(0,j)\right] + \sum_{j=1}^{J}\left[T^{(j)} - \Delta^{(j)}\right]\beta_y(0,j)\end{aligned} \tag{A.15}$$

where we have omitted the term $(T-1)\beta_y(0,0)$ because it is constant over every choice history (also we have normalized $\beta_y(y,y) = 0$ for every y). Now, applying a similar property to the term $\sum_{j=0}^{J}\sum_{k=1}^{J} D^{(j,k)}\ [\beta_y(k,j) - \beta_y(0,j)]$, we have:

$$\begin{aligned}\sum_{j=0}^{J}\sum_{k=1}^{J} D^{(j,k)}\left[\beta_y(k,j) - \beta_y(0,j)\right] &= \sum_{k=1}^{J}\left[\sum_{j=1}^{J} D^{(j,k)}\left[\beta_y(k,j) - \beta_y(0,j)\right] + \left[T^{(k)} - \sum_{j=1}^{J} D^{(j,k)}\right]\left[\beta_y(k,0) - \beta_y(0,0)\right]\right] \\ &= \sum_{k=1}^{J}\sum_{j=1}^{J} D^{(j,k)}\left[\beta_y(k,j) - \beta_y(0,j) - \beta_y(k,0)\right] + \sum_{k=1}^{J} T^{(k)}\beta_y(k,0)\end{aligned} \tag{A.16}$$

Putting together (A.15) and (A.16), we have that:

$$\sum_{j=0}^{J}\sum_{k=0}^{J} D^{(j,k)}\beta_y(k,j) = \sum_{k=1}^{J}\sum_{j=1}^{J} D^{(j,k)}\widetilde{\beta}_y(k,j) + \sum_{j=1}^{J} T^{(j)}\left[\beta_y(0,j) + \beta_y(j,0)\right] - \sum_{j=1}^{J}\Delta^{(j)}\beta_y(0,j) \tag{A.17}$$

where $\widetilde{\beta}_y(k,j) \equiv \beta_y(k,j) - \beta_y(0,j) - \beta_y(k,0)$. And plugging this expression into equation (A.14) for the log-probability, we obtain:

$$\ln \mathbb{P}\left(\widetilde{\mathbf{y}}|y_0,\theta\right) = \sum_{j=1}^{J} T^{(j)}\ g_{\theta,1}(j) + \sum_{j=1}^{J}\Delta^{(j)}\ g_{\theta,2}(j) + \sum_{j=1}^{J}\sum_{k=1}^{J} D^{(j,k)}\widetilde{\beta}_y(k,j) \tag{A.18}$$

where $g_{\theta,1}(j) \equiv \alpha_\theta(j) - \alpha_\theta(0) + \sigma_\theta(j) - \sigma_\theta(0) + \beta_y(0,j) + \beta_y(j,0)$, and $g_{\theta,2}(j) \equiv -\sigma_\theta(j) + \sigma_\theta(0) - \beta_y(0,j)$. ■

Proof of Proposition 9. For this model, the log probability is $\sum_{j=0}^{J}\sum_{t=1}^{T} 1\{y_t = j\}\ \alpha_\theta(j) + \sum_{j=0}^{J}\sum_{k=0}^{J}\sum_{t=1}^{T} 1\{y_{t-1} = j,\ y_t = k\}\ \beta_y(k,j) + \sum_{j=1}^{J}\sum_{d=1}^{J}\sum_{t=1}^{T} 1\{y_{t-1} = y_t = j,\ d_t = d\}\ \beta_d(j,d) + \sum_{j=0}^{J}\sum_{d=0}^{J}\sum_{t=1}^{T} 1\{y_{t-1} = j,\ d_t = d\}\ \sigma_\theta(j,d)$. Using the definition of the statistics in Table 1, we can write this log-probability as follows:

$$\begin{aligned}\ln \mathbb{P}\left(\widetilde{\mathbf{y}}|y_0,d_1,\theta\right) &= \sum_{j=0}^{J} T^{(j)}\alpha_\theta(j) + \left[T^{(0)} - \Delta^{(0)}\right]\sigma_\theta(0) + \sum_{j=1}^{J}\sum_{d\geq 1} H^{(j)}(d)\sigma_\theta(j,d) \\ &+ \sum_{j=0}^{J}\sum_{k=0}^{J} D^{(j,k)}\beta_y(k,j) + \sum_{j=1}^{J}\sum_{d\geq 1} X^{(j)}(d)\beta_d(j,d)\end{aligned} \tag{A.19}$$

Taking into account that: $T^{(0)} = T - \sum_{j=1}^{J} T^{(j)}$, we have that $\sum_{j=0}^{J} T^{(j)} \alpha_\theta(j) +$ $T^{(0)} \sigma_\theta(0) =$ $T[\alpha_\theta(0) + \sigma_\theta(0)] +$ $\sum_{j=1}^{J} T^{(j)} [\alpha_\theta(j) - \alpha_\theta(0) - \sigma_\theta(0)]$. And using equation (A.17) from the proof of Propositions 7-8, we have:

$$\begin{aligned}
\ln \mathbb{P}(\tilde{\mathbf{y}}|y_0, d_1, \theta) &= \sum_{j=1}^{J} T^{(j)} \left[\alpha_\theta(j) - \alpha_\theta(0) - \sigma_\theta(0) + \beta_y(0,j) + \beta_y(j,0)\right] + \sum_{j=1}^{J} \Delta^{(j)} \left[\sigma_\theta(0) - \beta_y(0,j)\right] \\
&+ \sum_{j=1}^{J} \sum_{d \geq 1} H^{(j)}(d) \sigma_\theta(j,d) \\
&+ \sum_{k=1}^{J} \sum_{j=1}^{J} D^{(j,k)} \widetilde{\beta}_y(k,j) + \sum_{j=1}^{J} \sum_{d \geq 1} X^{(j)}(d) \beta_d(j,d)
\end{aligned} \tag{A.20}$$

where we have omitted the term $T\alpha_\theta(0) + (T-1)\sigma_\theta(0)$ because the are constant across all the histories. Given that $T^{(j)} = \Delta^{(j)} + \sum_{d \geq 1} H^{(j)}(d)$ and $D^{(j,j)} = \sum_{d \geq 1} X^{(j)}(d)$ and $\tilde{\beta}_y(j,j) = -\beta_y(j,0) - \beta_y(0,j)$ by construction, we get:

$$\begin{aligned}
\ln \mathbb{P}(\tilde{\mathbf{y}}|y_0, d_1, \theta) &= \sum_{j=1}^{J} \sum_{d \geq 1} H^{(j)}(d) \left[\alpha_\theta(j) - \alpha_\theta(0) + \sigma_\theta(j,d) - \sigma_\theta(0) + \beta_y(0,j) + \beta_y(j,0)\right] \\
&+ \sum_{j=1}^{J} \Delta^{(j)} \left[\alpha_\theta(j) - \alpha_\theta(0) + \beta_y(j,0)\right] \\
&+ \sum_{k=1}^{J} \sum_{j \neq k} D^{(j,k)} \widetilde{\beta}_y(k,j) + \sum_{j=1}^{J} \sum_{d \geq 1} X^{(j)}(d) \ \gamma(j,d)
\end{aligned} \tag{A.21}$$

Now, consider the term $\sum_{j=1}^{J} \sum_{d \geq 1} X^{(j)}(d) \ \beta_d(j,d)$. By Lemma 2, for $d \geq 1$, $X^{(j)}(d) = H^{(j)}(d+1) - \Delta^{(j)}(d+1)$. Therefore,

$$\begin{aligned}
\sum_{j=1}^{J} \sum_{d \geq 1} X^{(j)}(d) \ \gamma(j,d) &= \sum_{j=1}^{J} \sum_{d \geq 1} \left[H^{(j)}(d+1) + \Delta^{(j)}(d+1)\right] \ \gamma(j,d) \\
&= \sum_{j=1}^{J} \sum_{d \geq 1} \left[H^{(j)}(d) + \Delta^{(j)}(d)\right] \ \gamma(j,d-1)
\end{aligned} \tag{A.22}$$

where, for the second equality, we take into account the normalization $\beta_d(j,0) = 0$ for any $j \geq 1$. Solving equation (A.22) into (A.21), and taking into account that $\sum_{d \geq 1} \Delta^{(j)}(d) = \Delta^{(j)}$, we obtain:

$$\begin{aligned}
\ln \mathbb{P}(\tilde{\mathbf{y}}|y_0, d_1, \theta) &= \sum_{j=1}^{J} \sum_{d \geq 1} H^{(j)}(d) \ g_{\theta,1}(j,d) + \sum_{j=1}^{J} \Delta^{(j)} g_{\theta,2}(j) \\
&+ \sum_{k=1}^{J} \sum_{j=1, j \neq k}^{J} D^{(j,k)} \widetilde{\beta}_y(k,j) + \sum_{j=1}^{J} \sum_{d \geq 1} \Delta^{(j)}(d) \ \gamma(j,d-1)
\end{aligned} \tag{A.23}$$

with $g_{\theta,1}(j,d) \equiv \alpha_\theta(j) - \alpha_\theta(0) + \sigma_\theta(j,d) - \sigma_\theta(0) + \beta_y(0,j) + \beta_y(j,0) + \gamma(j,d-1)$, $g_{\theta,2}(j) \equiv \alpha_\theta(j) - \alpha_\theta(0) + \beta_y(j,0)$, $\widetilde{\beta}_y(y,y_{-1}) \equiv \beta_y(y,y_{-1}) - \beta_y(0,y_{-1}) - \beta_y(y,0)$, and $\gamma(j,d) \equiv \beta_d(j,d) - \beta_y(j,0) - \beta_y(0,j)$. ■

Proof of Proposition 10. The expression of the log-probability is similar as in Proposition 9 but now we have the additional term $\sum_{t=1}^{T} v_\theta(y_t, d_{t+1}[y, y_{t-1}, d_t])$. This term is equal to $T^{(0)} v_\theta(0) + \sum_{j=1}^{J}\sum_{d\geq 1}\sum_{t=1}^{T} 1\{y_t = j,\ d_{t+1} = d\}\ v_\theta(j,d) = T^{(0)} v_\theta(0) + \sum_{j=1}^{J}\sum_{d\geq 1} v_\theta(j,d)\Big[H^{(j)}(d) + \Delta^{(j)}(d)\Big]$. Given $T^{(0)} = T - \sum_{j=1}^{J} T^{(j)} = T - \sum_{j=1}^{J}\sum_{d\geq 1} H^{(j)}(d) - \sum_{j=1}^{J}\sum_{d\geq 1} \Delta^{(j)}(d)$ and using equation (A.23) from the proof of Proposition 9, we have

$$\begin{aligned}
\ln \mathbb{P}(\widetilde{\mathbf{y}}|y_0, d_1, \theta) &= \sum_{j=1}^{J}\sum_{d\geq 1} H^{(j)}(d)\ g_{\theta,1}(j,d) + \sum_{j=1}^{J} \Delta^{(j)} g_{\theta,2}(j) \\
&+ \sum_{k=1}^{J}\sum_{j\neq k} D^{(j,k)} \widetilde{\beta}_y(k,j) + \sum_{j=1}^{J}\sum_{d\geq 1} \Delta^{(j)}(d)\ (\gamma(j,d-1) + v_\theta(j,d))
\end{aligned} \tag{A.24}$$

with $g_{\theta,1}(j,d) \equiv \alpha_\theta(j) - \alpha_\theta(0) + \sigma_\theta(j,d) - \sigma_\theta(0) + \beta_y(0,y) + \beta_y(y,0) + \gamma(y,d-1) + v_\theta(j,d) - v_\theta(0)$, $g_{\theta,2}(j) \equiv \alpha_\theta(j) - \alpha_\theta(0) + \beta_y(y,0) - v_\theta(0)$

Taking into account that $\sum_{d\geq 1} \Delta^{(j)}(d) = \Delta^{(j)}$ for any $j \geq 1$, we have

$$\ln \mathbb{P}(\widetilde{\mathbf{y}}|y_0, d_1, \theta) = \sum_{j=1}^{J}\sum_{d\geq 1} H^{(j)}(d)\ g_{\theta,1}(j,d) + \sum_{j=1}^{J}\sum_{d\geq 1} \Delta^{(j)}(d)\ g_{\theta,2}(j,d) + \sum_{k=1}^{J}\sum_{j\neq k} D^{(j,k)} \widetilde{\beta}_y(k,j) \tag{A.25}$$

where $g_{\theta,2}(j,d) \equiv \gamma(j,d-1) + v_\theta(j,d) + \alpha_\theta(j) - \alpha_\theta(0) + \beta_y(y,0) - v_\theta(0)$.

Proof of Proposition 11. Define $Z^{(j)} \equiv \sum_{d\geq 1} \Delta^{(j)}(d)[v_\theta(j,d) + \gamma(j,d-1)]$. Under Assumption 2, we have $v_\theta(j,d) = v_\theta(j,d^*)$ for any $d \geq d_j^*$, and $\gamma(j,d-1) = \gamma(j,d^*)$ for any $d \geq d_j^* + 1$. Therefore, we have for all $j \geq 1$,

$$\begin{aligned}
Z^{(j)} &= \textstyle\sum_{d\leq d_j^*-1} \Delta^{(j)}(d) v_\theta(j,d) + \Big[\sum_{d\geq d_j^*} \Delta^{(j)}(d)\Big] v_\theta(j,d_j^*) \\
&+ \textstyle\sum_{d\leq d_j^*} \Delta^{(j)}(d) \gamma(j,d-1) + \Big[\sum_{d\geq d_j^*+1} \Delta^{(j)}(d)\Big] \gamma(j,d_j^*) \\
&= \textstyle\sum_{d\leq d_j^*-1} \Delta^{(j)}(d)[v_\theta(j,d) + \gamma(j,d-1)] + \Big[\sum_{d\geq d_j^*} \Delta^{(j)}(d)\Big][v_\theta(j,d_j^*) + \gamma(j,d_j^*)] \\
&+ \Delta^{(j)}(d_j^*)[\gamma(j,d_j^*-1) - \gamma(j,d_j^*)]
\end{aligned} \tag{A.26}$$

Then the log-probability becomes:

$$\begin{aligned}\ln \mathbb{P}\left(\widetilde{\mathbf{y}}|y_0, d_1, \theta\right) \;&=\; \sum_{j=1}^{J}\sum_{d\geq 1} H^{(j)}(d)\;\, g_{\theta,1}(j,d) + \sum_{k=1}^{J}\sum_{j\neq k} D^{(j,k)}\widetilde{\beta}_y(k,j) \\ &+\; \sum_{j=1}^{J}\sum_{d\geq d_j^*-1} \Delta^{(j)}(d) g_{\theta,2}(j,d) + \sum_{j=1}^{J}\left[\textstyle\sum_{d\geq d_j^*}\Delta^{(j)}(d)\right] g_{\theta,2}(j,d_j^*) \\ &+\; \sum_{j=1}^{J}\Delta^{(j)}(d_j^*)\;\left(\gamma(j,d_j^*-1)-\gamma(j,d_j^*)\right)\end{aligned} \tag{A.27}$$

with $g_{\theta,1}(j,d) \equiv \alpha_\theta(j)-\alpha_\theta(0)+\sigma_\theta(j,d)-\sigma_\theta(0)+\beta_y(0,y)+\beta_y(y,0)+\gamma(y,d-1)+v_\theta(j,d)-v_\theta(0)$ and $g_{\theta,2}(j,d) \equiv \alpha_\theta(j)-\alpha_\theta(0)+\beta_y(y,0)-v_\theta(0)+v_\theta(j,d)+\gamma(j,d-1)$. Note that $g_{\theta,1}(j,d) = g_{\theta,1}(j,d_j^*)$ for $d \geq d_j^*$. Therefore, we have $\sum_{d\geq 1} H^{(j)}(d) g_{\theta,1}(d) = \sum_{d\leq d_j^*-1} H^{(j)}(d) g_{\theta,1}(d) + \left[\sum_{d\geq d_j^*} H^{(j)}(d)\right] g_{\theta,1}(d_j^*)$, such that

$$\begin{aligned}\ln \mathbb{P}\left(\widetilde{\mathbf{y}}|y_0, d_1, \theta\right) \;&=\; \sum_{j=1}^{J}\sum_{d\leq d_j^*-1} H^{(j)}(d)\;\, g_{\theta,1}(j,d) + \sum_{j=1}^{J}\Big[\sum_{d\geq d_j^*} H^{(j)}(d)\Big] g_{\theta,1}(j,d_j^*) \\ &+\; \sum_{k=1}^{J}\sum_{j=1}^{J} D^{(j,k)}\widetilde{\beta}_y(k,j) \\ &+\; \sum_{j=1}^{J}\sum_{d\geq d_j^*-1} \Delta^{(j)}(d) g_{\theta,2}(j,d) + \sum_{j=1}^{J}\Big[\sum_{d\geq d_j^*}\Delta^{(j)}(d)\Big] g_{\theta,2}(j,d_j^*) \\ &+\; \sum_{j=1}^{J}\Delta^{(j)}(d_j^*)\;\left(\gamma(j,d_j^*-1)-\gamma(j,d_j^*)\right) \qquad \blacksquare\end{aligned} \tag{A.28}$$

Proof of Proposition 12. It is clear that $\widehat{\mathbb{P}}\left(A_n\right) \rightarrow_p \mathbb{P}_0\left(A_n\right)$ and $\widehat{\mathbb{P}}\left(B_n\right) \rightarrow_p \mathbb{P}_0\left(B_n\right)$ such that the concentrated likelihood function $N^{-1}\ell_N(d^*)$ converges uniformly to the function:

$$\begin{aligned}\ell_0(d^*) \;&=\; \sum_{n=2}^{d^*} \mathbb{P}_0\left(A_n\right)\;\ln\left[\frac{\mathbb{P}_0\left(A_n\right)}{\mathbb{P}_0\left(A_n\right)+\mathbb{P}_0\left(B_n\right)}\right] + \mathbb{P}_0\left(B_n\right)\;\ln\left[\frac{\mathbb{P}_0\left(B_n\right)}{\mathbb{P}_0\left(A_n\right)+\mathbb{P}_0\left(B_n\right)}\right] \\ &+\; \sum_{n=d^*+1}^{L} \mathbb{P}_0\left(A_n\right)\;\ln\left[\frac{1}{2}\right] + \mathbb{P}_0\left(B_n\right)\;\ln\left[\frac{1}{2}\right]\end{aligned} \tag{A.29}$$

Lemma. Consider the function $f(q) = p_1\,\ln(q) + p_2\,\ln(1-q)$ where $p_1, p_2, q \in (0,1)$. This function is uniquely maximized at $q = p_1/[p_1+p_2]$.

Taking into account this Lemma, we have that for any value of n:

$$\begin{aligned}&\mathbb{P}_0\left(A_n\right)\;\ln\left[\frac{\mathbb{P}_0\left(A_n\right)}{\mathbb{P}_0\left(A_n\right)+\mathbb{P}_0\left(B_n\right)}\right] + \mathbb{P}_0\left(B_n\right)\;\ln\left[\frac{\mathbb{P}_0\left(B_n\right)}{\mathbb{P}_0\left(A_n\right)+\mathbb{P}_0\left(B_n\right)}\right] \\ &\geq \mathbb{P}_0\left(A_n\right)\;\ln\left[\frac{1}{2}\right] + \mathbb{P}_0\left(B_n\right)\;\ln\left[\frac{1}{2}\right]\end{aligned} \tag{A.30}$$

and the inequality is strict if and only if $\mathbb{P}_0(A_n) = \mathbb{P}_0(B_n)$. Given this result, it is straightforward to show that: $\ell_0(d_0^*) > \ell_0(d^*)$ for any $d^* < d_0^*$; and $\ell_0(d_0^*) = \ell_0(d^*)$ for any $d^* > d_0^*$. ■

Proof of Proposition 13. Let n be a value of the parameter d^* different to the true value d_0^*. Given our BIC function, we favor $\widehat{d_N^*} = n$ over $\widehat{d_N^*} = d_0^*$ if and only if $BIC_N(n) > BIC_N(d_0^*)$ and this is equivalent to:

$$2\left[\ell_N(n) - \ell_N(d_0^*)\right] > \left[n - d_0^*\right] \;\ln(N) \tag{A.31}$$

We show below that, as $N \to \infty$, $\mathbb{P}(2\left[\ell_N(n) - \ell_N(d_0^*)\right] > \left[n - d_0^*\right] \ln(N)) \to 0$, and therefore, $\mathbb{P}(\widehat{d_N^*} = d_0^*) \to 1$.

First, we show that $\mathbb{P}(\widehat{d_N^*} > d_0^*) \to 0$ as $N \to \infty$. By definition,

$$\mathbb{P}\left(\widehat{d_N^*} > d_0^*\right) = \mathbb{P}\left(\exists n > d_0^* : 2\left[\ell_N(n) - \ell_N(d_0^*)\right] > \left[n - d_0^*\right] \;\ln(N)\right) \tag{A.32}$$

Proposition 12 implies that, for any $n \geq d_0^*$, $N^{-1}\ell_N(n) \to_p \ell_0(d_0^*)$ and $2[\ell_N(n) - \ell_N(d_0^*)] \to_d \chi^2_{n-d_0^*} = O_p(1)$. Therefore, $\mathbb{P}\left(\widehat{d_N^*} > d_0^*\right) = \mathbb{P}\left(O_p(1) > \left[n - d_0^*\right] \;\ln(N)\right)$ that goes to zero as $N \to \infty$.

Now, we show that $\mathbb{P}(\widehat{d_N^*} < d_0^*) \to 0$ as $N \to \infty$. We need to prove that, for any $n < d_0^*$, the probability that $2\left[\ell_N(d_0^*) - \ell_N(n)\right] < \left[d_0^* - n\right] \ln(N)$ goes to zero as $N \to \infty$. We can write

$$2\left[\ell_N(d_0^*) - \ell_N(n)\right] = 2\left[\ell_N(d_0^*) - \ell_N(d_0^* - 1)\right] + \sum_{j=n+1}^{d_0^*-1} 2\left[\ell_N(j) - \ell_N(j-1)\right] \tag{A.33}$$

Since $\beta_0(d_0^*) \neq 0$, classical results imply that: (a) there exist constants c and $\mathcal{C}$ such that $cN \leq 2\left[\ell_N(d_0^*) - \ell_N(d_0^* - 1)\right] \leq \mathcal{C}N$; and (b) $\sum_{j=n+1}^{d_0^*-1} 2\left[\ell_N(j) - \ell_N(j-1)\right] = O_p(N)$ for all $n < d_0^*$, therefore $\mathbb{P}(2\left[\ell_N(d_0^*) - \ell_N(n)\right] < \left[d_0^* - n\right] \ln(N)) \to 0$ as $N \to \infty$. ■

Appendix 2. Model with stochastic transition of the endogenous state variables

Consider a model with the same structure as the model in Section 2 and Assumption 1 but now the vector of endogenous state variables is $\mathbf{x}_t = (x_t^y, x_t^d)$ and variables x_t^y and x_t^d *stochastic versions* of the variables y_{t-1} and d_t, respectively. We now describe precisely the stochastic process of these variables.

The support of state variable x_t^y is the choice set $\mathcal{Y}$, and its transition rule is $x_{t+1}^y = f_y(y_t, \xi_{t+1}^y)$ where ξ_{t+1}^y is i.i.d. over time and independent of $\mathbf{x}_t$. The support of state variable x_t^d is the set of possible durations, $\{1, 2, ..., \infty\}$, and its transition rule is $x_{t+1}^d = 1\{y_t > 0\}[\ 1\{y_t = x_t^y\}\ x_t^d + 1 + \xi_{t+1}^d]$, where ξ_{t+1}^d has support $\{0, 1, ..., \infty\}$, and it is i.i.d. over time and independent of $\mathbf{x}_t$. Importantly, the stochastic shocks ξ_{t+1}^y and ξ_{t+1}^d are not known to the agent when she makes her decision at period t. Note that this model becomes our model in the main text when these shocks have a degenerate probability distribution with $p(\xi_{t+1}^y = 0) = p(\xi_{t+1}^d = 0) = 1$.

Assumption 1' below is simply an extension of our Assumption 1 to this stochastic version of the model. We omit the exogenous state variables $\mathbf{z}_t$ for notational simplicity.

ASSUMPTION 1'. (A) The time horizon is infinite and $\delta \in (0, 1)$. (B) The utility function is $\Pi_t(y) = \alpha_\theta(y) + 1\{y = x_t^y\}\ \beta_d(y, x_t^d) + 1\{y \neq x_t^y\}\ \beta_y(y, x_t^y) + \varepsilon_t(y)$, and functions $\alpha_\theta(y)$, $\beta_d(y, x_t^d)$, and $\beta_y(y, x_t^y)$ are bounded. (C) $\beta_y(y, y) = 0$, $\beta_d(0, x^d) = 0$. (D) $\{\varepsilon_t(y) : y \in \mathcal{Y}\}$ are i.i.d. over (i, t, y) with a extreme value type I distribution. (E) $\mathbf{z}_t$ has discrete and finite support $\mathcal{Z}$ and follows a time-homogeneous Markov process. (F) The probability distribution of θ conditional on $\{\mathbf{z}_t, \mathbf{x}_t : t = 1, 2, ...\}$ is nonparametrically specified and completely unrestricted. (G) $x_t^y \in \mathcal{Y}$, and $x_{t+1}^y = f_y(y_t, \xi_{t+1}^y)$ where ξ_{t+1}^y is i.i.d. over time and independent of $\mathbf{x}_t$; $x_t^d \in \{0, 1, ..., \infty\}$, and $x_{t+1}^d = 1\{y_t > 0\}[\ 1\{y_t = x_t^y\}\ x_t^d + 1 + \xi_{t+1}^d]$, where ξ_{t+1}^d has support $\{0, 1, ..., \infty\}$, and it is i.i.d. over time and independent of $\mathbf{x}_t$. ■

The model has the following integrated Bellman equation:

$$V_\theta(\mathbf{x}_t) = \ln\left(\sum_{y\in\mathcal{Y}} \exp\left\{\ \alpha_\theta(y) + \beta(y, \mathbf{x}_t) + \delta\ \mathbb{E}_{\xi_{t+1}}\left[V_\theta\left(f_y(y_t, \xi_{t+1}^y),\ 1\{y_t = x_t^y\}\, x_t^d + 1 + \xi_{t+1}^d\right)\right]\ \right\}\right)$$

where $\mathbb{E}_{\xi_{t+1}}(.)$ the expectation over the distribution of $(\xi_{t+1}^y, \xi_{t+1}^d)$. Let $v_{\theta,t}$ be the continuation value function $\delta\ \mathbb{E}_{\xi_{t+1}}[V_\theta\left(f_y(y_t, \xi_{t+1}^y),\ 1\{y_t = x_t^y\}\, x_t^d +\ 1 + \xi_{t+1}^d\right)]$. Under our assumptions on the distribution of $(\xi_{t+1}^y, \xi_{t+1}^d)$, the continuation value function has very similar properties as in the model with a deterministic transition of the endogenous state variables. More specifically, (a) it depends only y_t and $1\{y_t = x_t^y\} x_t^d + 1$, i.e., $v_{\theta,t} = v_\theta(y_t, 1\{y_t = x_t^y\} x_t^d + 1)$; (b) If $y_t \neq x_t^y$, then $v_{\theta,t} = v_\theta(y_t, 1)$; (c) If $y_t = x_t^y$, then $v_{\theta,t} = v_\theta(y_t, x_t^d + 1)$; and (D) if $x_t^d \geq d_y^* - 1$ and $y_t = x_t^y$, then $v_{\theta,t} = v_\theta(y_t, d_y^*)$.

References

[1] Aguirregabiria, V. (1999): "The dynamics of markups and inventories in retailing firms," *The Review of Economic Studies*, 66, 275-308.

[2] Aguirregabiria, V. and P. Mira (2007): "Sequential Estimation of Dynamic Discrete Games," *Econometrica*, 75, 1-53.

[3] Andersen, E (1970): "Asymptotic Properties of Conditional Maximum Likelihood Estimators," *Journal of the Royal Statistical Society, Series B*, 32, 283 301.

[4] Arcidiacono, P., and R. Miller (2011): "Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity," *Econometrica*, 79(6), 1823-1867.

[5] Arellano, M., and S. Bonhomme (2012): "Nonlinear Panel Data Analysis", *Annual Review of Economics*, 3, 395-424.

[6] Arellano, M., and S. Bonhomme (2017): "Nonlinear Panel Data Methods for Dynamic Heterogeneous Agent Models, *Annual Review of Economics*, forthcoming.

[7] Arellano, M., and B. Honoré (2001): "Panel Data Models: Some Recent Developments," in J. J. Heckman and E. Leamer (eds.) *Handbook of Econometrics*, Volume 5, Chapter 53, North-Holland, 3229-3296.

[8] Blackwell, D. (1965): "Discounted dynamic programming," *The Annals of Mathematical Statistics*, 36(1), 226-235.

[9] Bonhomme, S. (2012): "Functional differencing," *Econometrica*, 80(4), 1337-1385.

[10] Browning, M., and J. Carro (2010): "Heterogeneity in dynamic discrete choice models," *Econometrics Journal*, 13(1), 1-39.

[11] Browning, M., and J. Carro (2014): "Dynamic binary outcome models with maximal heterogeneity," *Journal of Econometrics*, 178(2), 805-823.

[12] Caliendo, L., M. Dvorkin, and F. Parro (2015): "Trade and Labor Market Dynamics," NBER WP No. 21149.

[13] Chamberlain, G. (1980): "Analysis of Covariance with Qualitative Data," *Review of Economic Studies*, 47(1), 225-238.

[14] Chamberlain, G. (1985): "Heterogeneity, Omitted Variable Bias, and Duration Dependence," in Longitudinal Analysis of Labor Market Data, edited by J. J. Heckman and B. Singer. Cambridge: Cambridge University Press.

[15] Chamberlain, G. (1993): “Feedback in Panel Data Models,” unpublished manuscript, No. 1656. Harvard-Institute of Economic Research.

[16] Chamberlain, G. (2010): “Binary response models for panel data: Identification and information,” *Econometrica*, 78(1), 159-168.

[17] Chernozhukov, V., I. Fernandez-Val, J. Hahn, and W. Newey (2013): "Average and Quantile Effects in Nonseparable Panel Models," *Econometrica*, 81(2), 535-580.

[18] Chintagunta, P., E. Kyriazidou, and P. Perktold (2001): “Panel Data Analysis of Household Brand Choices,” *Journal of Econometrics*, 103(1), 111-153.

[19] Das, M. (1992): "A Micro-econometric Model of Capital Utilization and Retirement: The Case of the Cement Industry," *Review of Economic Studies*, 59, 277-297.

[20] Dunne, T., S. Klimek, M. Roberts, and D. Xu (2013): "Entry, Exit and the Determinants of Market Structure," *RAND Journal of Economics*, 44(3), 462-487.

[21] Erdem, T., S. Imai and M. P. Keane (2003): "Brand and Quantity Choice Dynamics under Price Uncertainty," *Quantitative Marketing and Economics*, 1, 5-64.

[22] Erdem, T., M. Keane, and B. Sun (2008): "A dynamic model of brand choice when price and advertising signal product quality," *Marketing Science*, 27(6), 1111-1125.

[23] Ericson, R. and A. Pakes (1995): "Markov Perfect Industry Dynamics: A Framework for Empirical Work," *Review of Economic Studies*, 62, 53-82.

[24] Frederiksen, A., B. Honoré, B., and L. Hu, L. (2007): "Discrete time duration models with group-level heterogeneity," *Journal of Econometrics*, 141(2), 1014-1043.

[25] Hahn, J.,and W. Newey (2004): "Jacknife and analytical bias reduction for nonlinear panel data models," *Econometrica*, 72(4), 1295-1319.

[26] Hahn, J. anf G. Kuersteiner (2011): "Bias reduction for dynamic nonlinear panel models with fixed effects," *Econometric Theory*, 27(6), 1152-1191.

[27] Heckman, J. (1981): “The incidental parameters problem and the problem of initial conditions in estimating a discrete time - discrete data stochastic process,” in C. Manski and D. McFadden (eds.), *Structural Analysis of Discrete Data with Econometric Applications*. MIT Press.

[28] Hendel, I., and A. Nevo (2006): "Measuring the Implications of Sales and Consumer Inventory Behavior. Econometrica 74, 1637-1674.

[29] Honoré, B. , and E. Kyriazidou (2000): "Panel data discrete choice models with lagged dependent variables," *Econometrica*, 68(4), 839-874.

[30] Honoré, B. , and E. Kyriazidou (2017): "Panel Vector Autoregressions with Binary Data," manuscript. Princeton University.

[31] Honorè, B., and E. Tamer (2006): "Bounds on parameters in panel dynamic discrete choice models," *Econometrica*, 74(3), 611-629.

[32] Kalouptsidi, M. (2014): "Time to build and fluctuations in bulk shipping," *American Economic Review*, 104(2), 564-608.

[33] Kalouptsidi, M., P. Scott, and E. Souza-Rodrigues (2015): "Identification of counterfactuals in dynamic discrete choice models," NBER working paper, No. 21527.

[34] Kano, K. (2013): "Menu Costs and Dynamic Duopoly," *International Journal of Industrial Organization*, 31(1), 102-118.

[35] Kasahara, H. (2009): "Temporary Increases in Tariffs and Investment: The Chilean Case," Journal of Business and Economic Statistics, 27(1), 113-127.

[36] Kasahara, H. and K. Shimotsu (2009): "Nonparametric identification of finite mixture models of dynamic discrete choices," *Econometrica*, 77(1), 135-175.

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

[38] Kennet, M. (1993) "Did Deregulation Affect Aircraft Engine Maintenance? An Empirical Policy Analysis," *The RAND Journal of Economics,* 24, 542-558.

[39] Lancaster, T. (2000): "The incidental parameter problem since 1948," *Journal of Econometrics*, 95(2), 391-413.

[40] Magnac, T. (2004): "Panel Binary Variables and Sufficiency: Generalizing Conditional Logit," *Econometrica*, 72(6), 1859-1876.

[41] Magnac, T., and D. Thesmar (2002): "Identifying dynamic discrete decision processes," *Econometrica*, 70(2), 801-816.

[42] Manski, C. (1987): "Semiparametric analysis of random effects linear models from binary panel data," *Econometrica*, 55(2), 357-362.

[43] Miller, R. (1984): "Job Matching and Occupational Choice", *Journal of Political Economy*, 92, 1086-1120.

[44] Newey, W. and D. McFadden (1994): "Large sample estimation and hypothesis testing," in R. E. Engle and McFadden (eds.) *Handbook of Econometrics Volume 4*, 2111-2245. North-Holland. Amsterdam.

[45] Neyman, J. and E. Scott (1948): "Consistent Estimates Based on Partially Consistent Observations," *Econometrica*, 16(1), 1-32.

[46] Pötscher, B.M. (1991): "Effects of model selection on inference," *Econometric Theory*, 7, 163 - 185.

[47] Roberts, M. and J. Tybout (1997): "The Decision to Export in Colombia: An Empirical Model of Entry with Sunk Costs," *American Economic Review*, 87(4), 545-564.

[48] Rust, J. (1987): "Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher," *Econometrica,* 55, 999-1033.

[49] Rust, J. (1994): "Structural estimation of Markov decision processes," in R. E. Engle and McFadden (eds.) *Handbook of Econometrics Volume 4*, North-Holland. Amsterdam.

[50] Ryan, S. (2013) "The costs of environmental regulation in a concentrated industry," *Econometrica,* 80(3), 1019-1061.

[51] Slade, M. (1998): "Optimal Pricing with Costly Adjustment: Evidence from Retail Grocery Stores", *Review of Economic Studies*, 65, 87-108.

[52] Sweeting, A. (2013): "Dynamic product positioning in differentiated product markets: The effect of fees for musical performance rights on the commercial radio industry," *Econometrica*, 81(5), 1763-1803.

[53] Willis, J. (2006): "Magazine prices revisited," *Journal of Applied Econometrics*, 21(3), 337-344.

[54] Wooldridge, J. (2005): "Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity," *Journal of Applied Econometrics*, 20(1), 39-54.