Journal of Econometrics 223 (2021) 280–311

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Sufficient statistics for unobserved heterogeneity in structural dynamic logit models

Victor Aguirregabiria [a,b,*,1], Jiaying Gu [a,1], Yao Luo [a,1]

[a] University of Toronto, 150 St. George Street, Toronto, ON, M5S 3G7, Canada
[b] CEPR, Canada

ARTICLE INFO

Article history:
Received 10 April 2018
Received in revised form 2 April 2019
Accepted 23 July 2019
Available online 18 September 2020

JEL classification:
C23
C25
C41
C51
C61

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

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. We prove the identification of the structural parameters using a conditional likelihood approach. The structure of the model implies that there is a sufficient statistic such that the likelihood function conditional on this statistic no longer depends on the unobserved heterogeneity – neither through the current utility nor through the continuation value of the forward-looking decision problem – but still depends on the structural parameters. We apply this estimator to a machine replacement model.

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 *incidental parameters problem* 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

* Corresponding author at: University of Toronto, 150 St. George Street, Toronto, ON, M5S 3G7, Canada.
E-mail addresses: victor.aguirregabiria@utoronto.ca (V. Aguirregabiria), jiaying.gu@utoronto.ca (J. Gu), yao.luo@utoronto.ca (Y. Luo).

[1] 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 Emory, Georgetown, Harvard-MIT, Michigan State, Pittsburgh, Princeton, PUC-Rio, Toronto, Yale, the conference at UCL on *Implementation of Structural Dynamic Models* (September, 2017), and the conference at Vanderbilt on *Identification in Econometrics* (May 2018). This research was supported by the *Social Sciences and Humanities Research Council of Canada.*

https://doi.org/10.1016/j.jeconom.2019.07.010
0304-4076/

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 main approaches to deal with this identification problem: random effects and fixed effects models/methods. Random-effects models impose restrictions – parametric and finite mixture restrictions – on the joint distribution of the unobserved heterogeneity and the initial conditions of the explanatory variables. Some of these restrictions can provide identification of the parameters of interest and the distribution of the unobserved heterogeneity. In contrast, fixed-effects methods focus on the identification of the parameters of interest and 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 probability distribution of the unobserved heterogeneity and the initial conditions of the explanatory variables.[2]

A fixed effect conditional likelihood method (Cox, 1958; Rasch, 1961; Andersen, 1970; 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 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. We consider that the researcher has data from a short panel — the number of periods T is fixed.[3]

There are nonlinear panel data models where the FE-CML approach cannot identify the structural parameters. 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 a panel data binary choice model where the choice probability is given by a distribution function – F – evaluated at a linear index of the explanatory variables plus an individual fixed effect, Chamberlain (2010) shows that if the explanatory variables have bounded support, then a necessary condition for the (point) identification of the parameters of interest is that the distribution function F is logistic.[5]

Another example of panel data model where parameters may not be identified is a dynamic binary choice model with fixed effect unobserved heterogeneity in the slope parameters (Browning and Carro, 2014).[6] As we explain below, this result has important implications for the identification of structural dynamic discrete choice models with fixed-effects unobserved heterogeneity.

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

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

3 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; Hahn and Kuersteiner, 2011, among others). Another fixed-effect estimator is Manski's maximum score method (Manski, 1987). Honoré and Kyriazidou (2000) have developed a maximum score estimator for dynamic discrete choice models. Bonhomme (2012) presents a *functional differencing approach* that includes as particular cases different fixed effects estimators in the literature.

4 For instance, we could choose the complete choice history of an individual as a sufficient statistic. Obviously, the conditional likelihood function based on this sufficient statistic depends 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 (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. Honoré and Tamer (2006) study partial identification of the dynamic panel data Probit model and derive sharp bounds for the identified set of the parameter of interest.

6 Browning 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. To obtain identification, these authors consider a finite mixture specification of the heterogeneous parameters. Their approach is in the same spirit as Kasahara and Shimotsu (2009).

7 Chamberlain (1985) and Honoré and Kyriazidou (2000) consider discrete choice logit models where the explanatory variables are the dependent variable lagged one and two periods, i.e., AR(1) and AR(2) models. D'Addio and Honoré (2010) study more comprehensively the AR(2) model. They do not incorporate time duration in the last choice as an explicit explanatory variable, though they interpret a non-zero value for the parameter associated with the second lag as evidence consistent with duration dependence. In our model, we include both lagged decision and duration as explicit state variables.

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. 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.[8] Identification can be obtained if there is a sufficient statistic that controls for this continuation value and implies a conditional likelihood that still depends on the structural parameters.

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.

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 of 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 in the effects of marginal changes of the explanatory variables or the structural parameters. The identification of marginal effects 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 et al., 2013). We discuss this problem in Section 3.7.

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 market entry/exit models with either binary choices (Roberts and Tybout, 1997; Aguirregabiria and Mira, 2007) or multinomial choices (Sweeting, 2013; Caliendo et al., 2019); occupational choice models (Miller, 1984; Keane and Wolpin, 1997); machine replacement models (Rust, 1987; Das, 1992; Kennet, 1993; Kasahara, 2009); inventory and investment decision models (Aguirregabiria, 1999; Ryan, 2013; Kalouptsidi, 2014); demand of differentiated products with consumer brand switching costs (Erdem et al., 2008) or storable products (Erdem et al., 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 for fixed effects dynamic logit models with duration dependence (Frederiksen et al., 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 inference. In Section 5, we illustrate our identification results in the context of Rust's bus replacement model and data (Rust, 1987). Section 6 summarizes and concludes. Proofs of Lemmas and Propositions are in 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 $\{1, 2, \ldots, \infty\}$.[10] Agents are indexed by i. In every period t, agent i chooses a value of the discrete variable $y_{it} \in \mathcal{Y} = \{0, 1, \ldots, J\}$ 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 $\Pi_{it}(j)$ is her one-period utility if she chooses action $y_{it} = j$. This utility is a function of four types of state variables which are known to the agent at period t:

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

Variables $\mathbf{z}_{it}$ and $\mathbf{x}_{it}$ are observable to the researcher, and $\boldsymbol{\varepsilon}_{it}$ 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. Vectors $\mathbf{z}_{it}$ and $\mathbf{x}_{it}$ have supports $\mathcal{Z}$ and $\mathcal{X}$, respectively. The unobservable variables $\{\varepsilon_{it}(j) : j \in \mathcal{Y}\}$ are *i.i.d.* over (i, t, j) with

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

[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 et al. (2003), Willis (2006), Aguirregabiria and Mira (2007), and Erdem et al. (2008).

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

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 $\boldsymbol{\theta}_i \equiv (\boldsymbol{\eta}_i, \delta_i)$ represent the unobserved heterogeneity from individual i. The probability distribution of $\boldsymbol{\theta}_i$ conditional on the history of observable state variables $\{\mathbf{z}_{it}, \mathbf{x}_{it} : t = 1, 2, \ldots\}$ is unrestricted and nonparametrically specified, i.e., fixed effects model. Functions $\alpha(j, \boldsymbol{\eta}, \mathbf{z})$ and $\beta(j, \mathbf{x}, \mathbf{z})$ are nonparametrically specified.

Our specification of the utility function represents a general semiparametric fixed-effect logit model. It extends *Rust's model* in two directions (Rust, 1987, 1994). 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 $\varepsilon_{it}(j)$ satisfy these conditions, our time-invariant unobserved heterogeneity interacts, in an 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 separable 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 lagged decision variable, $y_{i,t-1}$; and (b) *duration dependence*, where $d_{it} \in \{1, 2, \ldots, \infty\}$ is the number of periods since the last change in choice. The lagged decision has an 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 $\beta(j, \mathbf{x}_{it}, \mathbf{z}_{it})$ 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:

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

Function $\beta_d(j, d_{it}, \mathbf{z}_{it})$ captures duration dependence. For instance, in an occupational choice model, this term captures the return on earnings of job experience in occupation j. Function $\beta_y\left(j, y_{i,t-1}, \mathbf{z}_{it}\right)$ represents switching value (or switching costs with negative sign). In an occupational choice model, this term represents the (negative) cost of switching from occupation $y_{i,t-1}$ to occupation j. The additive separability between switching costs and *returns to experience* is not without loss of generality. For instance, we are restricting the cost of switching occupations to not depend on experience in the current job. However, this additive separability facilitates our analysis of identification and the model is still more general than previous fixed-effects discrete choice models.

We impose a restriction on the structural function $\beta_d(j, d, \mathbf{z}_{it})$ that plays a role in our identification results for this function. We assume that there is no 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 $\beta_y(j, j, \mathbf{z}_{it}) = 0$, 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 Eqs. (1) and (2). (C) $\beta_y(j, j, \mathbf{z}) = 0$, $\beta_d(0, d, \mathbf{z}) = 0$. (D) $\{\varepsilon_{it}(j) : j \in \mathcal{Y}\}$ are *i.i.d.* over (i, t, j) with an extreme value type I distribution. (E) $\mathbf{z}_{it}$ follows a time-homogeneous Markov process. (F) The probability distribution of $\boldsymbol{\theta}_i \equiv (\boldsymbol{\eta}_i, \delta_i)$ conditional on $\{\mathbf{z}_{it} : t = 1, 2, \ldots\}$ and on the initial condition $\mathbf{x}_{i1}$ is nonparametrically specified and completely unrestricted. ■

Assumption 1 implies that the model is stationary. Therefore, it rules out time trends and time dummies as explanatory variables. This setting can be unrealistic in some empirical applications. However, this stationarity assumption is the *status quo* in applications of dynamic structural models with infinite horizon, which are common in industrial organization.

Since the model does not have duration dependence 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.

[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.

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

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

$$\beta_d(j, d, \mathbf{z}) = \beta_d\left(j, d_j^*, \mathbf{z}\right) \text{ for any } d \geq d_j^*. \quad \blacksquare \tag{4}$$

For the moment, we assume that the researcher knows the values of d_j^*. In Section 4, we show that these values $\{d_j^*\}$ 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 ($y_{it} = 0$), 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., 2019). The two endogenous state variables are the index of the market where the firm was active in the previous period ($y_{i,t-1}$) and the number of periods of experience in the current market (d_{it}). The parameter $\beta_y(j, k, \mathbf{z})$ represents the (negative) cost of switching from market k to market j. There is no duration dependence if a firm is not active in any market (if $j = 0$), and the marginal return to experience in market j is zero after d_j^* 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 in the 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 no duration dependence if the worker is unemployed.

(c) Machine replacement models (Rust, 1987; Das, 1992; Kennet, 1993; 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} – that is, 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 et al., 2003; Hendel and Nevo, 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 previous purchase is a state variable (Erdem et al., 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 $\beta_d(j, d, \mathbf{z})$ captures the effect of inventory on the consumer's utility, and function $\beta_y(j, y_{-d}, \mathbf{z})$ 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 log 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^*. Let d_{it} represent the time duration since the last nominal price change, such that $d_{it} \in \{0, 1, 2, \ldots\}$ with $d_{i,t+1} = d_{it} + 1$ if $y_{it} = 1$ and $d_{i,t+1} = 0$ if $y_{it} = 0$. There is a simple relationship between this duration variable and the logarithm of real price: $r_{it} = r^* - \pi\, d_{it}$. This model has a similar structure as the machine replacement models described above. $\blacksquare$

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-homogeneity of the utility and the 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 $\boldsymbol{\theta}_i$, as defined by Rust (1994).[15] The optimal choice at period t can be represented as:

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

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

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

[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 Appendix.

[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 I distribution of the unobservables ε implies that the *conditional choice probability* (CCP) function has the following form:

$$P_{\theta_i}(j \mid \mathbf{x}_{it}, \mathbf{z}_{it}) = \frac{\exp\left\{\alpha\left(j, \eta_i, \mathbf{z}_{it}\right) + \beta\left(j, \mathbf{x}_{it}, \mathbf{z}_{it}\right) + v_{\theta_i}(j, d_{t+1}[j, \mathbf{x}_{it}], \mathbf{z}_{it})\right\}}{\sum_{k \in \mathcal{Y}} \exp\left\{\alpha\left(k, \eta_i, \mathbf{z}_{it}\right) + \beta\left(k, \mathbf{x}_{it}, \mathbf{z}_{it}\right) + v_{\theta_i}(k, d_{t+1}[k, \mathbf{x}_{it}], \mathbf{z}_{it})\right\}}. \quad (6)$$

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 of choosing alternative j becomes $v_{\theta_i}(j, \mathbf{z}_{it})$, which does not depend on the state variable, y_{it-1}.

Property 2. Under Assumption 2, for $j = y_{it-1}$ and any $d_{it} \geq d_j^* - 1$, the continuation value $v_{\theta_i}(j, d_{t+1}[j, y_{it-1}, d_{it}], \mathbf{z}_{it})$ becomes $v_{\theta_i}(j, d_j^*, \mathbf{z}_{it})$.

3. Identification

3.1. Preliminaries

The researcher has a panel dataset of N individuals who are observed over T periods: $\{y_{it}, \mathbf{x}_{it}, \mathbf{z}_{it} : i = 1, 2, \ldots, N ; t = 1, 2, \ldots, T\}$. We consider microeconometric applications where T is small — short panels.[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 $\boldsymbol{\theta}$ as an argument (or subindex) in those functions that depend on the time-invariant unobserved heterogeneity, i.e., $\alpha_{\boldsymbol{\theta}}(y, \mathbf{z})$ and $v_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})$.

Similarly as in Honoré and Kyriazidou (2000), our sufficient statistics include the condition that the vector of exogenous state variables $\mathbf{z}$ remains constant over several consecutive periods in the sample. For notational simplicity, we omit $\mathbf{z}$ as an argument in most of the expressions for the rest of this section. We use $\boldsymbol{\beta}$ to represent the vector of structural parameters that define the functions β_y and β_d.[17]

In discrete choice models, we can only identify utility differences relative to the utility of a baseline choice alternative. This implies that we cannot identify all the parameters in the functions β_y and β_d — regardless of whether the model has fixed effects unobserved heterogeneity or not, or whether agents are myopic or forward-looking. Therefore, we start by presenting a reparameterization of the model that defines the set of parameters in β_y and β_d that can be identified in a version of the model without unobserved heterogeneity and with myopic agents. Lemma 1 presents this reparameterization. The proof is in Appendix.

Lemma 1. *The model can be represented using the following equation:*

$$y_t = \arg\max_{j \in \mathcal{Y}} \left\{ \widetilde{\alpha}_{\boldsymbol{\theta}}(j) + \sum_{k \neq \{0,j\}} 1\{y_{t-1} = k\}\, \widetilde{\beta}_y(j,k) + 1\{y_{t-1} = j\}\widetilde{\beta}_d(j, d_t) + \widetilde{v}_{\boldsymbol{\theta}}(j, d_{t+1}) + \varepsilon_t(j) \right\} \quad (7)$$

with $\widetilde{\alpha}_{\boldsymbol{\theta}}(j) \equiv \alpha_{\boldsymbol{\theta}}(j) - \alpha_{\boldsymbol{\theta}}(0) + \beta_y(j,0)$; $\widetilde{\beta}_y(j,k) \equiv \beta_y(j,k) - \beta_y(0,k) - \beta_y(j,0)$; $\widetilde{\beta}_d(j,d) \equiv \beta_d(j,d) - \beta_y(0,j) - \beta_y(j,0)$; *and* $\widetilde{v}_{\boldsymbol{\theta}}(j, d_{t+1}) \equiv v_{\boldsymbol{\theta}}(j, d_{t+1}) - v_{\boldsymbol{\theta}}(0,0)$. ■

Lemma 1 establishes that in the best case scenario of a model without time invariant unobserved heterogeneity and with myopic agents, the parameters $\{\widetilde{\beta}_y(j,k) : j, k \geq 1, j \neq k\}$ and $\{\widetilde{\beta}_d(j,d) : j \geq 1, d \geq 1\}$ represent all the information that we can obtain about the functions β_y and β_d. Therefore, for the rest of the paper, we only consider these structural parameters. These parameters have a clear economic interpretation. Parameter $\widetilde{\beta}_y(j,k)$ represents the difference in switching cost between a *direct (one-period) switch* from k to j and an *indirect (two periods) switch* via alternative 0. Parameter $\widetilde{\beta}_d(j,d)$ is the sum of two components: $\beta_d(j,d)$ is the return of d periods of experience in occupation/market j; and the term $-\beta_y(j,0) - \beta_y(0,j)$ is the sum of the cost of entry into occupation/market j ($-\beta_y(j,0)$) and the cost of exit from occupation/market j ($-\beta_y(0,j)$). The sum of these two costs is typically described as the *sunk cost* of entry in occupation/market j. Given the parameters $\widetilde{\beta}_d(j,d)$, we can obtain the marginal return to experience $\beta_d(j,d) - \beta_d(j,d-1)$ for values of experience d greater or equal than two, i.e., $\beta_d(j,d) - \beta_d(j,d-1) = \widetilde{\beta}_d(j,d) - \widetilde{\beta}_d(j,d-1)$.

[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 for the initial period $t = 1$ (i.e., knowing 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 $(y_t, \mathbf{x}_t)$ has finite support, for a given value of $\mathbf{z}$ we can represent the structural functions $\beta_y(y_t, y_{t-1}, \mathbf{z})$ and $\beta_d(y_t, d_t, \mathbf{z})$ using a finite vector of parameters.

Given this description of the model, we can summarize our main identification results as follows. First, all the switching cost parameters $\{\widetilde{\beta}_y(j,k) : j,k \geq 1, j \neq k\}$ are identified regardless fixed effects unobserved heterogeneity or agents' forward-looking behavior (see Propositions 1, 2 and 7–11). Though these parameters are always identified, the set of choice histories in the data that provide information about these parameters depends crucially on whether the model has unobserved heterogeneity and/or agents are forward-looking. Second, all the return to experience parameters $\{\widetilde{\beta}_d(j,d) : j \geq 1, d \geq 1\}$ are identified in a model with unobserved heterogeneity when agents are myopic (see Propositions 3 and 9). However, without further restrictions, we cannot identify any return to experience parameter when agents are forward-looking (see Propositions 4 and 10). Third, in the forward-looking model, under the additional restriction of Assumption 2, we can identify the returns to experience parameters $\{\widetilde{\beta}_d(j,d_j^*) - \widetilde{\beta}_d(j,d_j^*-1) : j \geq 1\}$ (see Propositions 5 and 11). Finally, we show that the value of the parameters $\{d_j^* : j \geq 1\}$ in Assumption 2 is identified (see Proposition 6).

3.2. A general description of the conditional likelihood approach

The data for an individual in the sample consist of the history of choices between periods 1 and T, $\{y_1, y_2, \ldots, y_T\}$, and the initial values of the endogenous state variables, (y_0, d_1). We represent these data using the vector $\widetilde{\mathbf{y}} \equiv (d_1, y_0; y_1, y_2, \ldots, y_T)$ and we refer to this vector as an *individual's history*. The model implies the following probability:

$$\mathbb{P}(\widetilde{\mathbf{y}} \mid \boldsymbol{\theta}, \boldsymbol{\beta}) = \sum_{t=1}^{T} \frac{\exp\left\{\widetilde{\alpha}_{\boldsymbol{\theta}}(y_t) + \widetilde{\beta}(y_t, \mathbf{x}_t) + \widetilde{v}_{\boldsymbol{\theta}}(y_t, d_{t+1})\right\}}{\sum_{j \in \mathcal{Y}} \exp\left\{\widetilde{\alpha}_{\boldsymbol{\theta}}(j) + \widetilde{\beta}(j, \mathbf{x}_t) + \widetilde{v}_{\boldsymbol{\theta}}(j, d_{t+1})\right\}} p(y_0, d_1 \mid \boldsymbol{\theta}). \tag{8}$$

In a fixed effects model, the probability distribution of the initial values of the endogenous state variables conditional on the incidental parameters, $p(y_0, d_1 \mid \boldsymbol{\theta})$, is nonparametrically specified. Our identification results, for different versions of the model, have the following common features. First, we show that the log-probability function $\ln \mathbb{P}(\widetilde{\mathbf{y}} \mid \boldsymbol{\theta}, \boldsymbol{\beta})$ has the following structure (up to a constant term that does not depend on the data $\widetilde{\mathbf{y}}$):

$$\ln \mathbb{P}(\widetilde{\mathbf{y}} \mid \boldsymbol{\theta}, \boldsymbol{\beta}) = U(\widetilde{\mathbf{y}})' g_{\boldsymbol{\theta}} + S(\widetilde{\mathbf{y}})' \boldsymbol{\beta}^*, \tag{9}$$

where $U(\widetilde{\mathbf{y}})$ and $S(\widetilde{\mathbf{y}})$ are vectors of statistics (i.e., deterministic functions of the data $\widetilde{\mathbf{y}}$), $g_{\boldsymbol{\theta}}$ is a vector of functions of $\boldsymbol{\theta}$, and $\boldsymbol{\beta}^*$ is a vector of linear combinations of the original vector of structural parameters $\boldsymbol{\beta}$. This representation is such that each of the vectors, $U(\widetilde{\mathbf{y}})$, $g_{\boldsymbol{\theta}}$, $S(\widetilde{\mathbf{y}})$, and $\boldsymbol{\beta}^*$, has elements which are linearly independent.[18] The exact elements included in these vectors depend on the version of the model. 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 $U(\widetilde{\mathbf{y}})$ and $S(\widetilde{\mathbf{y}})$, respectively.

(i) Sufficiency. Definition: U is a sufficient statistic for $\boldsymbol{\theta}$ if and only if, for any $\widetilde{\mathbf{y}}$ the probability $\mathbb{P}(\widetilde{\mathbf{y}} \mid \boldsymbol{\theta}, \boldsymbol{\beta}, U)$ does not depend on $\boldsymbol{\theta}$. We now show that, given the structure in Eq. (9), U is a sufficient statistic for $\boldsymbol{\theta}$. Since U is a deterministic function $\widetilde{\mathbf{y}}$, we have that: (a) $\ln \mathbb{P}(\widetilde{\mathbf{y}} \mid \boldsymbol{\theta}, \boldsymbol{\beta}, U)$ is equal to $\ln \mathbb{P}(\widetilde{\mathbf{y}} \mid \boldsymbol{\theta}, \boldsymbol{\beta}) - \ln \mathbb{P}(U \mid \boldsymbol{\theta}, \boldsymbol{\beta})$; and (b) $\mathbb{P}(U \mid \boldsymbol{\theta}, \boldsymbol{\beta})$ is the sum of probabilities of all the possible histories $\widetilde{\mathbf{y}}'$ with the same value U. Therefore, we have that $\ln \mathbb{P}(\widetilde{\mathbf{y}} \mid \boldsymbol{\theta}, \boldsymbol{\beta}, U)$ is equal to $\ln \mathbb{P}(\widetilde{\mathbf{y}} \mid \boldsymbol{\theta}, \boldsymbol{\beta}) - \ln[\sum_{\widetilde{\mathbf{y}}':\, U(\widetilde{\mathbf{y}}')=U} \mathbb{P}(\widetilde{\mathbf{y}}' \mid \boldsymbol{\theta}, \boldsymbol{\beta})]$. Combining this expression with the form of the log-probability in Eq. (9), we have that:

$$\begin{aligned} \ln \mathbb{P}(\widetilde{\mathbf{y}} \mid \boldsymbol{\theta}, \boldsymbol{\beta}, U) &= U' g_{\boldsymbol{\theta}} + S' \boldsymbol{\beta}^* - \ln\left(\sum_{\widetilde{\mathbf{y}}':\, U(\widetilde{\mathbf{y}}')=U} \exp\left\{U(\widetilde{\mathbf{y}}')' g_{\boldsymbol{\theta}} + S(\widetilde{\mathbf{y}}')' \boldsymbol{\beta}^*\right\}\right) \\ &= S' \boldsymbol{\beta}^* - \ln\left(\sum_{\widetilde{\mathbf{y}}':\, U(\widetilde{\mathbf{y}}')=U} \exp\left\{S(\widetilde{\mathbf{y}}')' \boldsymbol{\beta}^*\right\}\right). \end{aligned} \tag{10}$$

Since the right hand side of Eq. (10) does not depend $\boldsymbol{\theta}$, we have that the structure of the log-probability in Eq. (9) implies that U is a sufficient statistic for $\boldsymbol{\theta}$.

Eq. (8) implies that the term $\ln p(y_0, d_1 \mid \boldsymbol{\theta})$ enters additively in the logarithm of the probability of an individual's data. Since the probability function $p(y_0, d_1 \mid \boldsymbol{\theta})$ is nonparametrically specified in a fixed effects model, any vector of sufficient statistics for the incidental parameters $\boldsymbol{\theta}$ should include the initial value of the endogenous state variables, (y_0, d_1).

(ii) Minimal sufficiency. U is a *minimal sufficient statistic,* that is, it does not contain redundant information. More formally, let $\mathbf{U}$ be a matrix consisting of $U(\widetilde{\mathbf{y}})'$ for all possible values of $\widetilde{\mathbf{y}}$ as row vectors. 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, $\ell(\boldsymbol{\beta}^*) \equiv \mathbb{E}_{\widetilde{\mathbf{y}}}\left[\ln \mathbb{P}(\widetilde{\mathbf{y}} \mid U, \boldsymbol{\beta}^*)\right]$. The vector of parameters $\boldsymbol{\beta}^*$ is point identified if the population likelihood is uniquely maximized at the true value of $\boldsymbol{\beta}^*$. Lemma 2 establishes a necessary and sufficient condition for identification. Let K be the dimension of the vector of parameters $\boldsymbol{\beta}^*$.

[18] Suppose that S and $\boldsymbol{\beta}$ 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 $S'\boldsymbol{\beta} = S_a' \boldsymbol{\beta}^*$ with $\boldsymbol{\beta}^* = [\mathbf{I} \,\vdots\, \mathbf{A}]'\boldsymbol{\beta}$, such that S_a and $\boldsymbol{\beta}^*$ are vectors with linearly independent elements.

Table 1
Definition of statistics for a choice history $\widetilde{\mathbf{y}}$.

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

Lemma 2. *Given $K+1$ histories, say $\{A_j : j = 0, 1, ..., K\}$, let $\mathbf{S}$ be a $K \times K$ matrix consisting of row vectors $S(A_j)' - S(A_0)'$ for all $j = 1, ..., K$. The vector $\boldsymbol{\beta}^*$ is identified if and only if there exist $K+1$ histories with the same value of the statistic U and a non-singular matrix $\mathbf{S}$.* ■

For example, if $\boldsymbol{\beta}^*$ is a scalar such that $K = 1$, then this parameter is identified if and only if 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. Duration dependence reflects that the payoff and thus the choice probability depends on the number of periods since the last change in choice. In some applications, this may represent an important source of persistence. Second, we should take into account that unobserved heterogeneity enters into the continuation value function, $v_{\boldsymbol{\theta}}$. This implies that the sufficient statistic U should control not only for $\widetilde{\alpha}_{\boldsymbol{\theta}}(y_t)$ but also for the continuation values $\widetilde{v}_{\boldsymbol{\theta}}(y_t, d_{t+1})$. 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. Instead, 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.

3.3. Some useful statistics

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

Table 1 summarizes our definition of statistics.

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

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

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

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

Difference between final and initial states. For any choice alternative $j \in \mathcal{Y}$ and any duration $d \geq 0$, the statistic $\Delta^{(j)}(d)$ is defined as $1\{y_T = j, d_{T+1} = d\} - 1\{y_0 = j, d_1 = d\}$.

3.4. Binary choice models

Given the general representation of the model in Eq. (7), we can particularize it to the binary choice model to have:

$$y_t = 1\left\{\widetilde{\alpha}_{\boldsymbol{\theta}} + y_{t-1}\widetilde{\beta}_d(d_t) + \widetilde{v}_{\boldsymbol{\theta}}(d_t + 1) + \widetilde{\varepsilon}_t \geq 0\right\} \tag{11}$$

where $\widetilde{\alpha}_{\boldsymbol{\theta}} \equiv \alpha_{\boldsymbol{\theta}}(1) - \alpha_{\boldsymbol{\theta}}(0) + \beta_y(1, 0)$, $\widetilde{\beta}_d(d) = \beta_d(1, d) - \beta_y(1, 0) - \beta_y(0, 1)$, $\widetilde{v}_{\boldsymbol{\theta}}(d) \equiv v_{\boldsymbol{\theta}}(1, d) - v_{\boldsymbol{\theta}}(0, 0)$, and $\widetilde{\varepsilon}_t \equiv \varepsilon_t(1) - \varepsilon_t(0)$. 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).

3.4.1. Myopic dynamic model without duration dependence

Consider the model in Eq. (11) under the restrictions of myopic behavior (i.e., $\delta = 0$) and no duration dependence (i.e., $\beta_d(1, d) = 0$). These restrictions imply that the continuation values $\widetilde{v}_\theta(d_t + 1)$ become zero, and the term $\widetilde{\beta}_d(d_t)$ becomes equal to $-\beta_y(1, 0) - \beta_y(0, 1)$. The parameter $-\beta_y(1, 0) - \beta_y(0, 1)$ represents the sum of the costs of market entry and exit, or equivalently the sunk cost of entry. We use β_y to denote this sunk cost parameter. We can present this model using the standard representation,

$$y_t = 1\left\{\widetilde{\alpha}_\theta + \widetilde{\beta}_y\, y_{t-1} + \widetilde{\varepsilon}_t \geq 0\right\}. \tag{12}$$

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 $\widetilde{\mathbf{y}}$ is:

$$\ln \mathbb{P}\left(\widetilde{\mathbf{y}} \mid \boldsymbol{\theta}\right) \quad = \quad \ln p_\theta(y_0) + \sum_{t=1}^{T} y_t \left[\widetilde{\alpha}_\theta + \widetilde{\beta}_y y_{t-1}\right] + (1 - y_{t-1})\, \sigma_\theta(0) + y_{t-1}\, \sigma_\theta(1). \tag{13}$$

Proposition 1 establishes (i) the sufficient statistic, (ii) minimal sufficiency, and (iii) identification for this model. The identification result in this Proposition was established in Chamberlain (1985).

Proposition 1. *In the myopic binary choice model without duration dependence, the log-probability of a choice history has the form* $\ln \mathbb{P}\left(\widetilde{\mathbf{y}} \mid \boldsymbol{\theta}, \boldsymbol{\beta}\right) = U' g_\theta + S' \boldsymbol{\beta}^*$ *with*

$$U = \left(y_0,\ y_T,\ T^{(1)}\right) \quad ; \quad S = D^{(1,1)} \quad ; \quad \boldsymbol{\beta}^* = \widetilde{\beta}_y. \tag{14}$$

U is a minimal sufficient statistic for $\boldsymbol{\theta}$. For $T \geq 3$ – conditional on U – there is variation in S such that the parameter $\widetilde{\beta}_y$ is identified. ■

Example 1. Suppose that $T = 3$ such that the history of an individual is $\{y_0 \mid y_1, y_2, y_3\}$. Consider the pair of histories $A = (0 \mid 0, 1, 1)$ and $B = (0 \mid 1, 0, 1)$. Applying Eq. (13) to these histories, we have that $\ln \mathbb{P}(A) = \ln p_\theta(0) + 2\widetilde{\alpha}_\theta + 2\sigma_\theta(0) + \sigma_\theta(1) + \widetilde{\beta}_y$, and $\ln \mathbb{P}(B) = \ln p_\theta(0) + 2\widetilde{\alpha}_\theta + 2\sigma_\theta(0) + \sigma_\theta(1)$, such that $\ln \mathbb{P}(A) - \ln \mathbb{P}(B) = \widetilde{\beta}_y$. Therefore, the parameter $\widetilde{\beta}_y$ is identified as $\ln \mathbb{P}(0|0, 1, 1) - \ln \mathbb{P}(0|1, 0, 1)$. Intuitively, the sunk cost parameter is identified from the logarithm of the ratio between the frequency of "stayers" – individuals with histories $(0|0, 1, 1)$ – and the frequency of "switchers" – individuals with histories $(0|1, 0, 1)$. We can also obtain this identification result using the representation in Proposition 1. The vector of sufficient statistics U consists of y_0, y_3, and $y_1 + y_2 + y_3$. The identifying statistic S is $y_0 y_1 + y_1 y_2 + y_2 y_3$. Histories A and B have the same value for the sufficient statistic vector, $U(A) = U(B) = (y_0, y_3, y_1 + y_2 + y_3) = (0, 1, 2)$, but they have different values for the identifying statistic, $D^{(1,1)}(A) = 1$ and $D^{(1,1)}(B) = 0$. ■

With $T \geq 3$, the parameter $\widetilde{\beta}_y$ is over-identified. For instance, following up with the case with $T = 3$ in Example 1, we can consider the pair of histories $(1 \mid 1, 0, 0)$ and $(1 \mid 0, 1, 0)$, and it is simple to verify that $\widetilde{\beta}_y$ can be also identified as $\ln \mathbb{P}(1|1, 0, 0) - \ln \mathbb{P}(1|0, 1, 0)$. Therefore, the model implies the testable over-identifying restriction $\ln \mathbb{P}(0|0, 1, 1) - \ln \mathbb{P}(0|1, 0, 1) = \ln \mathbb{P}(1|1, 0, 0) - \ln \mathbb{P}(1|0, 1, 0)$, which is an implication of the assumptions of stationarity and no duration dependence.

3.4.2. Forward-looking dynamic model without duration dependence

Consider a forward-looking version of the model in Eq. (11) but without duration dependence. We can represent this model as,

$$y_t = 1\{\widetilde{\alpha}_\theta + \widetilde{v}_\theta + \widetilde{\beta}_y\, y_{t-1} + \widetilde{\varepsilon}_t \geq 0\} \tag{15}$$

where $\widetilde{v}_\theta = v_\theta(1) - v_\theta(0)$, and we omit argument d in this subsection because there is no duration dependence. 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, the two models are observationally equivalent.

Proposition 2. *In the forward-looking binary choice model without duration dependence, the log-probability of a history has the form* $\ln \mathbb{P}\left(\widetilde{\mathbf{y}} \mid \boldsymbol{\theta}, \boldsymbol{\beta}\right) = U' g_\theta + S' \boldsymbol{\beta}^*$ *with*

$$U = \left(y_0,\ y_T,\ T^{(1)}\right) \quad ; \quad S = D^{(1,1)} \quad ; \quad \boldsymbol{\beta}^* = \widetilde{\beta}_y. \tag{16}$$

U is a minimal sufficient statistic for $\boldsymbol{\theta}$. For $T \geq 3$, conditional on U there is variation in S such that the parameter $\widetilde{\beta}_y$ is identified. ■

Example 2. Example 1 applies to this model as well such that, with $T = 3$, the sunk cost parameter $\widetilde{\beta}_y$ is identified from the logarithm of the ratio between the frequency of "stayers" and the frequency of "switchers". That is, $\widetilde{\beta}_y = \ln \mathbb{P}(0, 0, 1, 1) - \ln \mathbb{P}(0, 1, 0, 1)$ and also, $\widetilde{\beta}_y = \ln \mathbb{P}(1, 1, 0, 0) - \ln \mathbb{P}(1, 0, 1, 0)$. ■

3.4.3. Myopic dynamic model with duration dependence

Consider the model in Eq. (11) with duration dependence but where agents are myopic. We can present this model as

$$y_t = 1\left\{\widetilde{\alpha}_{\boldsymbol{\theta}} + y_{t-1}\widetilde{\beta}_d(d_t) + \widetilde{\varepsilon}_t \geq 0\right\}. \tag{17}$$

For this model, the log-probability of the choice history $\widetilde{\mathbf{y}} = (y_0, d_1; y_1, \ldots, y_T)$ is:

$$\ln \mathbb{P}\left(\widetilde{\mathbf{y}} \mid \boldsymbol{\theta}\right) \quad = \quad \ln p_{\boldsymbol{\theta}}(y_0, d_1) + \sum_{t=1}^{T} y_t \left[\widetilde{\alpha}_{\boldsymbol{\theta}} + y_{t-1}\, \widetilde{\beta}_d(d_t)\right] + \sigma_{\boldsymbol{\theta}}(y_{t-1}, d_t) \tag{18}$$

where $\sigma_{\boldsymbol{\theta}}(y_{t-1}, d_t) \equiv -\ln\left(1 + \exp\left\{\widetilde{\alpha}_{\boldsymbol{\theta}} + y_{t-1}\, \widetilde{\beta}_d(d_t)\right\}\right)$, and we use $\sigma_{\boldsymbol{\theta}}(0)$ to represents $\sigma_{\boldsymbol{\theta}}(0, d)$.

Proposition 3 establishes the minimal sufficient statistic and the 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* $\ln \mathbb{P}\left(\widetilde{\mathbf{y}} \mid \boldsymbol{\theta}, \boldsymbol{\beta}\right) = U' g_{\boldsymbol{\theta}} + S' \boldsymbol{\beta}^*$ *with*

$$\begin{cases} U = \left[d_1,\ y_0,\ y_T,\ \{H^{(1)}(d) : d \geq 1\}\right] \\ S = \left[\Delta^{(1)}(d) : 2 \leq d \leq T-1\right] \\ \boldsymbol{\beta}^* = \left[\widetilde{\beta}_d(d) : 1 \leq d \leq T-2\right]. \end{cases} \tag{19}$$

U is a minimal sufficient statistic for $\boldsymbol{\theta}$. Conditional on U, the elements in the vector of statistics S are linearly independent such that the structural parameters $\boldsymbol{\beta}^*$ are identified. ■

For this model, the vector of sufficient statistics includes the histogram of durations $\{H^{(1)}(d) : d \geq 1\}$. Conditional on these statistics, the identification of the structural parameter $\widetilde{\beta}_d(d)$ comes from the difference between the final and the initial value of duration, $\Delta^{(1)}(d+1) = 1\{d_{T+1} = d+1\} - 1\{d_1 = d+1\}$ for $d \geq 1$. The identification result in Proposition 3 for the myopic model with duration dependence does not depend on Assumption 2.

Note that under the assumption of myopic individual behavior, we can identify the same duration dependence parameters $\widetilde{\beta}_d(d)$ regardless of whether the model has fixed effects unobserved heterogeneity or not. However, the set of choice histories that contain identifying information about these parameters is substantially reduced when we have unobserved heterogeneity.

Example 3(a). Suppose that $T = 3$ such that a choice history is $\{y_0, d_1 \mid y_1, y_2, y_3\}$. Consider the histories $A = \{0, 0 \mid 0, 1, 1\}$ and $B = \{0, 0 \mid 1, 0, 1\}$. Applying Eq. (18) to these histories, we have that $\ln \mathbb{P}(A) = \ln p_{\boldsymbol{\theta}}(0, 0) + 2\widetilde{\alpha}_{\boldsymbol{\theta}} + 2\sigma_{\boldsymbol{\theta}}(0) + \sigma_{\boldsymbol{\theta}}(1, 1) + \widetilde{\beta}_d(1)$, and $\ln \mathbb{P}(B) = \ln p_{\boldsymbol{\theta}}(0, 0) + 2\widetilde{\alpha}_{\boldsymbol{\theta}} + 2\sigma_{\boldsymbol{\theta}}(0) + \sigma_{\boldsymbol{\theta}}(1, 1)$, such that $\ln \mathbb{P}(A) - \ln \mathbb{P}(B) = \widetilde{\beta}_d(1)$. This implies that the parameter $\widetilde{\beta}_d(1)$ is identified from $\ln \mathbb{P}(0, 0|0, 1, 1) - \ln \mathbb{P}(0, 0|1, 0, 1)$. We can also confirm this identification result using the representation in Proposition 3. Histories A and B have the same value of the initial condition, $(y_0, d_1) = (0, 0)$, and of the final choice, $y_3 = 1$. Under history A, the series of durations $\{d_1, d_2, d_3\}$ is $\{0, 0, 1\}$, and under history B the evolution of durations is $\{0, 1, 0\}$. Therefore, the histogram of durations between periods 1 and 3 is the same under the two histories such that they have the same value for the sufficient statistic vector, $U(A) = U(B)$. However, the two histories have different final durations d_4. We have that $d_4 = 2$ under history A, and it is equal to 1 under history B. Therefore, we have that $S(A)'\boldsymbol{\beta}^* = \widetilde{\beta}_d(1)$ and $S(B)'\boldsymbol{\beta}^* = 0$, and this implies that the parameter $\widetilde{\beta}_d(1)$ is identified from $\ln \mathbb{P}(0, 0|0, 1, 1) - \ln \mathbb{P}(0, 0|1, 0, 1)$. ■

Example 3(b). Suppose that $T \geq 3$, let n be any integer such that $1 \leq n \leq T-2$, and define a sub-history $\{y_0, d_1 \mid y_1, \ldots, y_{n+2}\}$. 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 ones. Applying Eq. (18) to these histories, we have that $\ln \mathbb{P}(A) = (n+1)\,\widetilde{\alpha}_{\boldsymbol{\theta}} + \sum_{d=1}^{n} \widetilde{\beta}_d(d) + 2\sigma_{\boldsymbol{\theta}}(0) + \sum_{d=1}^{n} \sigma_{\boldsymbol{\theta}}(1, d)$, and $\ln \mathbb{P}(B) = (n+1)\,\widetilde{\alpha}_{\boldsymbol{\theta}} + \sum_{d=1}^{n-1} \widetilde{\beta}_d(d) + 2\sigma_{\boldsymbol{\theta}}(0) + \sum_{d=1}^{n} \sigma_{\boldsymbol{\theta}}(1, d)$, such that $\ln \mathbb{P}(A) - \ln \mathbb{P}(B) = \widetilde{\beta}_d(n)$. This implies that the parameter $\widetilde{\beta}_d(n)$ is identified from $\ln \mathbb{P}(0, 0|0, \mathbf{1}_{n+1}) - \ln \mathbb{P}(0, 0|\mathbf{1}_n, 0, 1)$. We can also confirm this identification result using the representation in Proposition 3. Histories A and B have the same value of the initial condition, $(y_0, d_1) = (0, 0)$, and of the final choice, $y_{n+2} = 1$. Under history A, the series of durations $\{d_1, d_2, \ldots, d_{n+2}\}$ is $\{0, 0, 1, \ldots, n\}$, and under history B the evolution of durations is $\{0, 1, \ldots, n, 0\}$. The histogram of durations is the same under the two histories such that $U(A) = U(B)$. The two histories have different final durations d_{n+3}. We have that $d_{n+3} = n+1$ under history A, and $d_{n+3} = 1$ under history B. This implies that $S(A)'\boldsymbol{\beta}^* - S(B)'\boldsymbol{\beta}^* = \widetilde{\beta}_d(n)$. ■

3.4.4. Forward-looking dynamic model with duration dependence

Consider the general binary choice model in Eq. (11), with duration dependence and with forward-looking agents. For this model, the log-probability of the choice history $\widetilde{\mathbf{y}}$ conditional on $(y_0, d_1, \boldsymbol{\theta})$ is:

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

with $\sigma_\theta(y_{t-1}, d_t) \equiv -\ln(1+\exp\{\widetilde{\alpha}_\theta + y_{t-1}\,\widetilde{\beta}_d(d_t) + \widetilde{v}_\theta(d_t+1)\})$, and $\sigma_\theta(0) \equiv \sigma_\theta(0,d)$. Compared to Eq. (18), the forward looking model in Eq. (20) includes the additional term $\sum_{t=1}^{T} y_t\,\widetilde{v}_\theta(d_t+1)$.

Proposition 4 establishes that under Assumption 1 – and without Assumption 2 – the *CMLE approach* does not provide 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 form* $\ln \mathbb{P}(\widetilde{\mathbf{y}}|\,\boldsymbol{\theta},\boldsymbol{\beta}) = U'g_\theta + S'\boldsymbol{\beta}^*$, *with*

$$U \;=\; \left[d_1,\; y_0,\; \{H^{(1)}(d),\; \Delta^{(1)}(d) : d \geq 1\}\right]. \tag{21}$$

U is a minimal sufficient statistic for $\boldsymbol{\theta}$. *The structural parameters* β_y *and* β_d*are not identified because U includes all the statistics associated with these structural parameters.* ■

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, $\Delta^{(1)}(d)$. These additional statistics are also the only statistics associated with the structural parameter $\widetilde{\beta}_d(d)$. Therefore, after controlling for the vector of sufficient statistics U, there is no variation left that can identify structural parameters in this model.

Example 4(a). Suppose that $T = 3$ such that a history is $\{y_0, d_1 \mid y_1, y_2, y_3\}$. Consider histories $A = \{0, 0 \mid 0, 1, 1\}$ and $B = \{0, 0 \mid 1, 0, 1\}$. Taking into account the form of the log-probability in Eq. (20), we have that $\ln \mathbb{P}(A) = 2\widetilde{\alpha}_\theta + \widetilde{\beta}_d(1) + \widetilde{v}_\theta(1) + \widetilde{v}_\theta(2) + 2\sigma_\theta(0) + \sigma_\theta(1,1)$, and $\ln \mathbb{P}(B) = 2\widetilde{\alpha}_\theta + 2\,\widetilde{v}_\theta(1) + 2\sigma_\theta(0) + \sigma_\theta(1,1)$, such that

$$\ln \mathbb{P}(A) - \ln \mathbb{P}(B) = \widetilde{\beta}_d(1) + \widetilde{v}_\theta(2) - \widetilde{v}_\theta(1). \tag{22}$$

The right-hand side includes the expected future return of a second year of experience, $\widetilde{v}_\theta(2) - \widetilde{v}_\theta(1)$, which depends on the incidental parameters. Therefore, $\ln \mathbb{P}(A) - \ln \mathbb{P}(B)$ does not identify any structural parameter. In particular, it does not identify $\widetilde{\beta}_d(1)$. ■

Example 4(b). Suppose that $T \geq 5$, let n be any integer such that $2 \leq n \leq (T-1)/2$, and consider the sub-histories: $A = \{0, 0 \mid \mathbf{1}_{n-1}, 0, \mathbf{1}_{n+1}\}$ and $B = \{0, 0 \mid \mathbf{1}_n, 0, \mathbf{1}_n\}$. Given the expression for the log-probability in Eq. (20), we have that $\ln \mathbb{P}(A) = 2n\widetilde{\alpha}_\theta + 2\sum_{d=1}^{n-2}\widetilde{\beta}_d(d) + \widetilde{\beta}_d(n-1) + \widetilde{\beta}_d(n) + 2\sum_{d=1}^{n-1}\widetilde{v}_\theta(d) + \widetilde{v}_\theta(n) + \widetilde{v}_\theta(n+1) + 2\sigma_\theta(0) + 2\sum_{d=1}^{n-1}\sigma_\theta(1,d) + \sigma_\theta(1,n)$, and $\ln \mathbb{P}(B) = 2n\widetilde{\alpha}_\theta + 2\sum_{d=1}^{n-2}\widetilde{\beta}_d(d) + 2\widetilde{\beta}_d(n-1) + 2\sum_{d=1}^{n}\widetilde{v}_\theta(d) + 2\sigma_\theta(0) + 2\sum_{d=1}^{n-1}\sigma_\theta(1,d) + \sigma_\theta(1,n)$, such that

$$\ln \mathbb{P}(A) - \ln \mathbb{P}(B) = \widetilde{\beta}_d(n) - \widetilde{\beta}_d(n-1) + \widetilde{v}_\theta(n+1) - \widetilde{v}_\theta(n). \tag{23}$$

This difference in log-probabilities depends on the incidental parameters through the continuation values. Though this pair of histories identifies the structural parameter $\widetilde{\beta}_d(n) - \widetilde{\beta}_d(n-1)$ in a myopic model with duration dependence, it does not identify any structural parameter in the forward-looking model. ■

Examples 4(a) and 4(b), and more specifically Eqs. (22) and (23), suggest a restriction that provides identification of the structural parameters. A sufficient condition for the identification of $\widetilde{\beta}_d(n) - \widetilde{\beta}_d(n-1)$ from $\ln \mathbb{P}(A) - \ln \mathbb{P}(B)$ is that $\widetilde{v}_\theta(n+1) - \widetilde{v}_\theta(n) = 0$ for any possible value of the incidental parameters.[19] By Property 2, under Assumption 2 there is a value d^* such that $\widetilde{v}_\theta(n+1) - \widetilde{v}_\theta(n) = 0$ for any duration n greater or equal than 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 form* $\ln \mathbb{P}(\widetilde{\mathbf{y}} \mid \boldsymbol{\theta},\boldsymbol{\beta}) = U'g_\theta + S'\boldsymbol{\beta}^*$ *with*

$$\begin{cases} U = \left[d_1,\; y_0,\; \{H^{(1)}(d),\; \Delta^{(1)}(d) : d \leq d^*-1\},\; \sum_{d\geq d^*} H^{(1)}(d),\; \sum_{d\geq d^*}\Delta^{(1)}(d)\right] \\ S = H^{(1)}(d^*) + \Delta^{(1)}(d^*); \quad \boldsymbol{\beta}^* = \beta_d(1, d^*-1) - \beta_d(1, d^*). \end{cases} \tag{24}$$

U is a minimal sufficient statistic for $\boldsymbol{\theta}$. *Conditional on U, the statistic* $\Delta^{(1)}(d^*)$ *has variation and the structural parameter* $\beta_d(1,d^*) - \beta_d(1,d^*-1)$ *is identified.* ■

Example 5(a). Consider the data in Example 4(a) with $T = 3$ and histories $A = \{0, 0 \mid 0, 1, 1\}$ and $B = \{0, 0 \mid 1, 0, 1\}$. We have shown that $\ln \mathbb{P}(A) - \ln \mathbb{P}(B) = \widetilde{\beta}_d(1) + \widetilde{v}_\theta(2) - \widetilde{v}_\theta(1)$. Suppose that $d^* = 1$ such that there is return for one period of experience but not for additional experience, that is, $\widetilde{\beta}_d(d) = \widetilde{\beta}_d(1)$ for $d \geq 1$. Under this assumption, as established in Property 2 of the model, we have that $\widetilde{v}_\theta(d) - \widetilde{v}_\theta(1) = 0$ for any $d \geq 1$. Therefore, the parameter $\widetilde{\beta}_d(1)$ is identified as

[19] In principle, it would be sufficient that $v_\theta(1, n+1) - v_\theta(1, n)$ does not depend on $\boldsymbol{\theta}$, i.e., $v_\theta(1, n+1) - v_\theta(1, n) = f(n)$. If we could obtain this type of condition, then $\ln \mathbb{P}(A) - \ln \mathbb{P}(B)$ would identify the parameter $\widetilde{\beta}_y + \beta_d(1, n) + f(n)$ where $f(n)$ would have an economic interpretation as a continuation value. However, $v_\theta(1, n)$ is a nonlinear function of $\boldsymbol{\theta}$, i.e., $v_\theta(1, n) = \ln(\exp\{\delta\; v_\theta(0)\} + \exp\{\delta[\alpha_\theta + \widetilde{\beta}_y + \beta_d(1, n) + v_\theta(1, n+1)]\})$. Given this structure, it seems that the only restrictions on the primitives of the model that can make $v_\theta(1, n+1) - v_\theta(1, n)$ independent of $\boldsymbol{\theta}$ are those that make it equal to zero.

Table 2
Identification of dynamic binary logit models.

Myopic model			*Forward-Looking model*		
Minimal sufficient stat.	Identified parameters	Identifying statistics	Minimal sufficient stat.	Identified parameters	Identifying statistics
Panel 1: Models without duration dependence					
$T^{(1)}$, y_0, y_T	$\widetilde{\beta}_y$	$D^{(1,1)}$	$T^{(1)}$, y_0, y_T	$\widetilde{\beta}_y$	$D^{(1,1)}$
Panel 2: Models with duration dependence					
y_0, d_1, y_T, $H^{(1)}(d) : d \geq 1$	$\widetilde{\beta}_d(d)$ for $1 \leq d \leq T-2$	$\Delta^{(1)}(d)$ $2 \leq d \leq T-1$	$H^{(1)}(d) : d \leq d^*-1$; $\sum_{d \geq d^*} H^{(1)}(d)$; $\Delta^{(1)}(d) : d \leq d^*-1$; $\sum_{d \geq d^*} \Delta^{(1)}(d)$	$\widetilde{\beta}_d(n)$ $-\widetilde{\beta}_d(n-1)$ for $n \geq d^*$	$H^{(1)}(n) + \Delta^{(1)}(n)$ for $n \geq d^*$

$\ln \mathbb{P}(0,0 \mid 0,1,1) - \ln \mathbb{P}(0,0 \mid 1,0,1)$. With $d^* = 1$, the sufficient statistic is $U = [d_1, y_0, \sum_{d \geq 1} H^{(1)}(d), \sum_{d \geq 1} \Delta^{(1)}(d)]$, or taking into account that $\sum_{d \geq 1} H^{(1)}(d) = T^{(1)} + y_0 - y_T$ and $\sum_{d \geq 1} \Delta^{(1)}(d) = y_T - y_0$, we have that $U = [d_1, y_0, y_T, T^{(1)}]$. The identifying statistic is $S = \Delta^{(1)}(1) = y_T \ 1\{d_T = 1\} - y_0 \ 1\{d_1 = 1\}$. ■

Example 5(b). Consider the data in Example 4(b) with $T \geq 5$, and sub-histories $A = \{0,0 \mid \mathbf{1}_{n-1}, 0, \mathbf{1}_{n+1}\}$ and $B = \{0,0 \mid \mathbf{1}_n, 0, \mathbf{1}_n\}$ for $n \leq (T-1)/2$. We have shown that $\ln \mathbb{P}(A) - \ln \mathbb{P}(B) = \widetilde{\beta}_d(n) - \widetilde{\beta}_d(n-1) + \widetilde{v}_\theta(n+1) - \widetilde{v}_\theta(n)$. Suppose that Assumption 2 holds, and consider values of n such that $n \geq d^*$. Under these conditions, we have that $\widetilde{v}_\theta(n+1) - \widetilde{v}_\theta(n) = 0$ such that:

$$\ln \mathbb{P}(0,0 \mid \mathbf{1}_{n-1}, 0, \mathbf{1}_{n+1}) - \ln \mathbb{P}(0,0 \mid \mathbf{1}_n, 0, \mathbf{1}_n) = \widetilde{\beta}_d(n) - \widetilde{\beta}_d(n-1). \tag{25}$$

For $n = d^*$, we have that $\ln \mathbb{P}(A) - \ln \mathbb{P}(B)$ identifies $\beta_d(1, d^*) - \beta_d(1, d^*-1)$. For values of n strictly greater than d^*, the model implies that $\ln \mathbb{P}(A) - \ln \mathbb{P}(B) = \beta_d(1,n) - \beta_d(1,n-1) = 0$. As we show below, this restriction for $n > d^*$ provides identification of the parameter d^*. ■

In Examples 5(a) and 5(b), the parameter $\boldsymbol{\beta}^*$ is identified because histories A and B have different values for the $\Delta^{(1)}(d^*)$. It is possible to obtain other examples where the identification of $\boldsymbol{\beta}^*$ is based on two histories with different values for the identifying statistic $H^{(1)}(d^*)$.[20]

In the forward-looking binary choice model with duration dependence, only $\widetilde{\beta}_d(d^*) - \widetilde{\beta}_d(d^*-1)$ is identified. This result contrasts with the myopic model where, as shown in Examples 3(a) and 3(b), we can identify $\widetilde{\beta}_d(d)$ for any duration $1 \leq d \leq T-2$.

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

3.5. 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:

$$\begin{cases} \text{For } n > d^*, \quad U(A_n) = U(B_n), \text{ and } \ln \mathbb{P}(A_n) - \ln \mathbb{P}(B_n) = \Delta\beta_d(n) = 0. \\ \text{For } n = d^*, \quad U(A_n) = U(B_n), \text{ and } \ln \mathbb{P}(A_n) - \ln \mathbb{P}(B_n) = \Delta\beta_d(d^*) \neq 0. \\ \text{For } n < d^*, \quad U(A_n) \neq U(B_n). \end{cases} \tag{26}$$

Note that $\ln \mathbb{P}(A_n) - \ln \mathbb{P}(B_n)$ identifies the parameter $\widetilde{\beta}_d(n) - \widetilde{\beta}_d(n-1)$ 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 $\ln \mathbb{P}(A_n) - \ln \mathbb{P}(B_n) \neq 0$. 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:*

$$d^* = \max\{n : \ \ln \mathbb{P}(A_n) - \ln \mathbb{P}(B_n) \neq 0\} \quad ■ \tag{27}$$

[20] For instance, consider the histories $A = \{0,0 \mid 1,0,1,1,1,1\}$ and $B = \{0,0 \mid 1,1,0,1,1,1\}$. The sequences of durations under these histories are $\mathbf{d}(A) = \{0,1,0,1,2,3\}$ and $\mathbf{d}(B) = \{0,1,2,0,1,2\}$, respectively. We have that $\Delta^{(1)}(2)(A) = \Delta^{(1)}(2)(B) = 0$, but $H^{(1)}(2)(A) = 1$ and $H^{(1)}(2)(B) = 2$ such that – when $d^* = 2$ – the parameter $\boldsymbol{\beta}^*$ is identified from the frequencies of these two histories.

Example 6. Suppose that $T = 7$. Consider the following three pairs of histories: $A_1 = \{0, 0 \mid 0, 1, 1\}$ and $B_1 = \{0, 0 \mid 1, 0, 1\}$; $A_2 = \{0, 0 \mid 1, 0, 1, 1, 1\}$ and $B_2 = \{0, 0 \mid 1, 1, 0, 1, 1\}$; and $A_3 = \{0, 0| 1, 1, 0, 1, 1, 1, 1\}$ and $B_3 = \{0, 0 |1, 1, 1, 0, 1, 1, 1\}$.[21] Without knowing the true value of d^*, all we can say is that:

$$\begin{cases} \ln \mathbb{P}(A_1) - \ln \mathbb{P}(B_1) &= \widetilde{\beta}_d(1) + \widetilde{v}_{\boldsymbol{\theta}}(2) - \widetilde{v}_{\boldsymbol{\theta}}(1). \\ \ln \mathbb{P}(A_2) - \ln \mathbb{P}(B_2) &= \widetilde{\beta}_d(2) - \widetilde{\beta}_d(1) + \widetilde{v}_{\boldsymbol{\theta}}(3) - \widetilde{v}_{\boldsymbol{\theta}}(2). \\ \ln \mathbb{P}(A_3) - \ln \mathbb{P}(B_3) &= \widetilde{\beta}_d(3) - \widetilde{\beta}_d(2) + \widetilde{v}_{\boldsymbol{\theta}}(4) - \widetilde{v}_{\boldsymbol{\theta}}(3). \end{cases} \tag{28}$$

Given that $T = 7$, to identify d^* we need to assume that $d^* \in \{1, 2, 3\}$. The following table describes the pattern of the log-probability differences $\ln \mathbb{P}(A_1) - \ln \mathbb{P}(B_1)$, $\ln \mathbb{P}(A_2) - \ln \mathbb{P}(B_2)$, and $\ln \mathbb{P}(A_3) - \ln \mathbb{P}(B_3)$ for each of the three possible values of d^*.

True d^*	$\ln\left[\frac{\mathbb{P}(A_1)}{\mathbb{P}(B_1)}\right]$	$\ln\left[\frac{\mathbb{P}(A_2)}{\mathbb{P}(B_2)}\right]$	$\ln\left[\frac{\mathbb{P}(A_3)}{\mathbb{P}(B_3)}\right]$
$d^* = 1$	$\neq 0$	0	0
$d^* = 2$	any value	$\neq 0$	0
$d^* = 3$	any value	any value	$\neq 0$

We can distinguish between these different patterns and therefore we can identify d^*. ■

3.6. Multinomial choice models

3.6.1. Multinomial myopic model without duration dependence

Consider the general multinomial choice model in Eq. (7) but particularized to the case with myopic agents, $\widetilde{v}_{\boldsymbol{\theta}}(j, d) = 0$, and without duration dependence, $\widetilde{\beta}_d(j, d) = 0$. We have:

$$y_t = \arg\max_{j \in \mathcal{Y}} \left\{ \widetilde{\alpha}_{\boldsymbol{\theta}}(j) + \sum_{k \neq 0} 1\{y_{t-1} = k\} \, \widetilde{\beta}_y(j, k) + \varepsilon_t(j) \right\}. \tag{29}$$

The log-probability of the choice history $\widetilde{\mathbf{y}} = \{y_0, y_1, \ldots, y_T\}$ conditional on $\boldsymbol{\theta}$ is:

$$\ln \mathbb{P}(\widetilde{\mathbf{y}} | \boldsymbol{\theta}) = \ln p_{\boldsymbol{\theta}}(y_0) + \sum_{t=1}^{T} \sum_{j \neq 0} 1\{y_t = j\} \left[\widetilde{\alpha}_{\boldsymbol{\theta}}(j) + \widetilde{\beta}_y(j, y_{t-1}) \right] + \sum_{t=1}^{T} \sigma_{\boldsymbol{\theta}}(y_{t-1}) \tag{30}$$

where $\sigma_{\boldsymbol{\theta}}(y_{t-1}) \equiv -\ln\left[1 + \sum_{j \neq 0} \exp\{\widetilde{\alpha}_{\boldsymbol{\theta}}(j) + \widetilde{\beta}_y(j, y_{t-1})\}\right]$. 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 form* $\ln \mathbb{P}(\widetilde{\mathbf{y}} \mid \boldsymbol{\theta}, \boldsymbol{\beta}) = U' g_{\boldsymbol{\theta}} + S' \boldsymbol{\beta}^*$ *with*

$$\begin{cases} U = \left[y_0, \; y_T, \; \{T^{(j)} : j \geq 1\} \right]. \\ S = \left[D^{(j,k)} : j, k \geq 1 \right]. \\ \boldsymbol{\beta}^* = \left[\widetilde{\beta}_y(j, k) : j, k \geq 1 \right]. \end{cases} \tag{31}$$

U is a minimal sufficient statistic for $\boldsymbol{\theta}$. Conditional on U, the elements in the vector of statistics S are linearly independent such that the vector of parameters $\boldsymbol{\beta}^*$ is identified. ■

The following example presents a pair of histories that identifies $\widetilde{\beta}_y(j, k)$.

Example 7. Suppose that $T = 3$ and consider the following two realizations of the history $(y_0| \, y_1, y_2, y_3)$: $A = \{0 \mid 0, k, j\}$ and $B = \{0 \mid k, 0, j\}$ with $j, k \neq 0$. Using the formula for the log-probability of a choice history in Eq. (30), we have that $\ln \mathbb{P}(A) = \ln p_{\boldsymbol{\theta}}(0) + \widetilde{\alpha}_{\boldsymbol{\theta}}(k) + \widetilde{\alpha}_{\boldsymbol{\theta}}(j) + 2\sigma_{\boldsymbol{\theta}}(0) + \sigma_{\boldsymbol{\theta}}(k) + \widetilde{\beta}_y(k, 0) + \widetilde{\beta}_y(j, k)$, and $\ln \mathbb{P}(B) = \ln p_{\boldsymbol{\theta}}(0) + \widetilde{\alpha}_{\boldsymbol{\theta}}(k) + \widetilde{\alpha}_{\boldsymbol{\theta}}(j) + 2\sigma_{\boldsymbol{\theta}}(0) + \sigma_{\boldsymbol{\theta}}(k) + \widetilde{\beta}_y(k, 0) + \widetilde{\beta}_y(0, k) + \widetilde{\beta}_y(j, 0)$, such that $\ln \mathbb{P}(A) - \ln \mathbb{P}(B)$ identifies the parameter $\widetilde{\beta}_y(k, j) - \widetilde{\beta}_y(0, j) - \widetilde{\beta}_y(k, 0)$ which is equal to $\widetilde{\beta}_y(k, j)$ because, by definition, $\widetilde{\beta}_y(0, j) = 0$ and $\widetilde{\beta}_y(0, k) = 0$. We can also obtain this identification result by using the representation in Proposition 7. Histories A and B have the same value for the initial condition, y_0, the final

[21] Note that the pair of histories A_1 and B_1 may come from periods $t = 0$ to $t = 3$ or from any other sequence of four consecutive periods within the original history of length $T = 7$. The same comment applies to the pair of histories A_2 and B_2.

choice, y_3, and the statistics $T^{(j)}$ and $T^{(k)}$, such that $U(A) = U(B)$. The identifying statistics $D^{(y_{-1},y)}$ take the following values: $D^{(j,k)}(A) = 1$, $D^{(j,k)}(B) = 0$, $D^{(j,0)}(A) = 0$, $D^{(j,0)}(B) = 1$, $D^{(0,k)}(A) = 0$, $D^{(0,k)}(B) = 1$, and $D^{(y,y_{-1})}(A) = D^{(y,y_{-1})}(B) = 0$ for any other pair (y, y_{-1}). Therefore, $S(A)'\boldsymbol{\beta}^* - S(B)'\boldsymbol{\beta}^* = [D^{(j,k)}(A) - D^{(j,k)}(B)]\ \widetilde{\beta}_y(j,k) + [D^{(j,0)}(A) - D^{(j,0)}(B)]\ \widetilde{\beta}_y(j,0) + [D^{(0,k)}(A) - D^{(0,k)}(B)]\ \beta_y(0,k) = \widetilde{\beta}_y(k,j)$. A particular case of this example is when $j = k$, such that $A = \{0|\ 0, j, j\}$ and $B = \{0 \mid j, 0, j\}$. In this case, $\ln \mathbb{P}(A) - \ln \mathbb{P}(B)$ identifies the sunk cost for choice j, $-\beta_y(0,j) - \beta_y(j,0)$. ■

3.6.2. Multinomial forward-looking model without duration dependence

Consider the general multinomial choice model in Eq. (7) with forward-looking agents but without duration dependence, $\widetilde{\beta}_d(j,d) = 0$. We can represent this model as:

$$y_t = \arg\max_{j\in\mathcal{Y}} \left\{ \widetilde{\alpha}_{\boldsymbol{\theta}}(j) + \widetilde{v}_{\boldsymbol{\theta}}(j) + \sum_{k\neq 0} 1\{y_{t-1} = k\}\ \widetilde{\beta}_y(j,k) + \varepsilon_t(j) \right\}. \tag{32}$$

The log-probability of the choice history $\widetilde{\mathbf{y}}$ conditional on $\boldsymbol{\theta}$ has a similar form as in the myopic model, but now the incidental parameter $\boldsymbol{\theta}$ enters through the function $\widetilde{\alpha}_{\boldsymbol{\theta}}(j) + \widetilde{v}_{\boldsymbol{\theta}}(j)$.

$$\ln \mathbb{P}(\widetilde{\mathbf{y}}|\boldsymbol{\theta}) \quad = \quad \ln p_{\boldsymbol{\theta}}(y_0) + \sum_{t=1}^{T}\sum_{j\neq 0} 1\{y_t = j\}\ \left[\widetilde{\alpha}_{\boldsymbol{\theta}}(j) + \widetilde{v}_{\boldsymbol{\theta}}(j) + \widetilde{\beta}_y(j, y_{t-1})\right] + \sum_{t=1}^{T} \sigma_{\boldsymbol{\theta}}(y_{t-1}) \tag{33}$$

where $\sigma_{\boldsymbol{\theta}}(y_{t-1}) \equiv -\ln\left[1 + \sum_{j\neq 0} \exp\{\widetilde{\alpha}_{\boldsymbol{\theta}}(j) + \widetilde{v}_{\boldsymbol{\theta}}(j) + \widetilde{\beta}_y(j, y_{t-1})\}\right]$. 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* $\ln \mathbb{P}(\widetilde{\mathbf{y}} \mid \boldsymbol{\theta}, \boldsymbol{\beta}) = U' g_{\boldsymbol{\theta}} + S'\boldsymbol{\beta}^*$ *with* $U = [y_0, y_T, \{T^{(j)} : j \geq 1\}]$, $S = [D^{(j,k)} : j, k \geq 1]$, *and* $\boldsymbol{\beta}^* = [\widetilde{\beta}_y(j,k) : j, k \geq 1]$. *$U$ is a minimal sufficient statistic for $\boldsymbol{\theta}$.Conditional on U, the elements in the vector of statistics S are linearly independent such that the vector of parameters $\boldsymbol{\beta}^*$ is identified.* ■

Example 8. Example 7 also applies to the forward-looking model. With $T = 3$, we have that the parameter $\widetilde{\beta}_y(j,k)$ is identified from $\ln \mathbb{P}(0|0,k,j) - \ln \mathbb{P}(0 \mid k, 0, j)$. ■

3.6.3. Multinomial myopic model with duration dependence

Consider the multinomial choice model in Eq. (7) with duration dependence but with myopic agents. We can represent this model as follows:

$$y_t = \arg\max_{j\in\mathcal{Y}} \left\{ \widetilde{\alpha}_{\boldsymbol{\theta}}(j) + \sum_{k\neq\{0,j\}} 1\{y_{t-1} = k\}\ \widetilde{\beta}_y(j,k) + 1\{y_{t-1} = j\}\widetilde{\beta}_d(j,d_t) + \varepsilon_t(j) \right\}. \tag{34}$$

The log-probability of a choice history $\widetilde{\mathbf{y}}$ conditional on $\boldsymbol{\theta}$ is:

$$\begin{aligned} \ln \mathbb{P}(\widetilde{\mathbf{y}}|\boldsymbol{\theta}) \quad &= \quad \ln p_{\boldsymbol{\theta}}(y_0, d_1) + \sum_{t=1}^{T}\sum_{j\neq 0} 1\{y_t = j\}\ \widetilde{\alpha}_{\boldsymbol{\theta}}(j) + \sum_{t=1}^{T} \sigma_{\boldsymbol{\theta}}(y_{t-1}, d_t) \\ &+ \quad \sum_{t=1}^{T}\sum_{j\neq 0} \left[\sum_{k\neq\{0,j\}} 1\{y_t = j, y_{t-1} = k\}\ \widetilde{\beta}_y(j,k) + 1\{y_t = y_{t-1} = j\}\ \widetilde{\beta}_d(j,d_t) \right] \end{aligned} \tag{35}$$

where $\sigma_{\boldsymbol{\theta}}(y_{t-1}, d_t) \equiv -\ln[1 + \sum_{j\neq 0} \exp\{\widetilde{\alpha}_{\boldsymbol{\theta}}(j) + \sum_{k\neq\{0,j\}} 1\{y_{t-1} = k\}\widetilde{\beta}_y(j,k) + 1\{y_{t-1} = j\}\widetilde{\beta}_d(j,d_t)\}]$. Proposition 9 presents identification results.

Proposition 9. *In the multinomial myopic model with duration dependence under Assumption 1, the log-probability of a choice history has the form* $\ln \mathbb{P}(\widetilde{\mathbf{y}}|\ \boldsymbol{\theta}, \boldsymbol{\beta}) = U' g_{\boldsymbol{\theta}} + S'\boldsymbol{\beta}^*$ *with*

$$\begin{cases} U = \left[d_1,\ y_0,\ y_T,\ \{H^{(j)}(d) : j \geq 1,\ d \geq 1\}\right]. \\ S = \left[D^{(j,k)} : j, k \geq 1,\ k \neq j;\ \Delta^{(j)}(d) : j \geq 1;\ d \geq 2\right]. \\ \boldsymbol{\beta}^* = \left[\widetilde{\beta}_y(j,j) : j \geq 1,\ k \neq j;\ \widetilde{\beta}_d(j,d) : j \geq 1;\ d \geq 1\right]. \end{cases} \tag{36}$$

U is a minimal sufficient statistic for $\boldsymbol{\theta}$. Conditional on U, the elements in the vector of statistics S are linearly independent such that the vector of parameters $\boldsymbol{\beta}^*$ is identified. ■

The following examples present choice histories that identify $\widetilde{\beta}_y(j,k)$ and $\widetilde{\beta}_d(j,d)$.

Example 9(a). Suppose that $T = 3$ such that a choice history is $(y_0, d_1 \mid y_1, y_2, y_3)$. For $j, k \neq 0$ and $j \neq k$, consider the pair of histories $A = \{0, 0 \mid 0, j, k\}$ and $B = \{0, 0 \mid j, 0, k\}$. Using the expression for the log-probability of a choice history in Eq. (35) we have that $\ln \mathbb{P}(A) = \ln p_\theta(0,0) + \widetilde{\alpha}_\theta(j) + \widetilde{\alpha}_\theta(k) + 2\sigma_\theta(0) + \sigma_\theta(j,1) + \widetilde{\beta}_y(k,j)$, and $\ln \mathbb{P}(B) = \ln p_\theta(0,0) + \widetilde{\alpha}_\theta(j) + \widetilde{\alpha}_\theta(k) + 2\sigma_\theta(0) + \sigma_\theta(j,1)$, such that $\ln \mathbb{P}(A) - \ln \mathbb{P}(B) = \widetilde{\beta}_y(k,j)$. Therefore, the parameter $\widetilde{\beta}_y(k,j)$ is identified from $\ln \mathbb{P}(0,0 \mid 0,j,k) - \ln \mathbb{P}(0,0 \mid j,0,k)$. We can also obtain this identification result by using Proposition 9. The initial condition, $(d_1, y_0) = (0,0)$, and the final choice, $y_3 = k$, are the same in the two histories. The histories also have the same histogram for the states (y_{t-1}, d_t): the state $(0,0)$ occurs twice, state $(j,1)$ occurs once, and the other possible states never happen. Therefore, we have that $U(A) = U(B)$. As for the values of the identifying statistics in the vector S, we have that: $D^{(j,k)} = 1$ under history A and $D^{(j,k)} = 0$ under history B; since $d_1 = 0$ and $d_4 = 1$ in both histories, we have that for any $d \geq 2$ the statistics $\Delta^{(k)}(d) \equiv 1\{y_3 = k, d_4 = d\} - 1\{y_0 = k, d_1 = d\}$ are zero for both A and B. Therefore, $S(A)'\boldsymbol{\beta}^* - S(B)'\boldsymbol{\beta}^* = \widetilde{\beta}_y(k,j)$. ■

Example 9(b). Suppose that $T \geq 5$, let n be any integer such that $2 \leq n \leq (T-1)/2$, and define a sub-history $\{y_0, d_1 \mid y_1, \ldots, y_{2n+1}\}$. Consider the sub-histories $A = \{0, 0 \mid \mathbf{j}_{n-1}, 0, \mathbf{j}_{n+1}\}$ and $B = \{0, 0 \mid \mathbf{j}_n, 0, \mathbf{j}_n\}$, where $\mathbf{j}_n$ represents a sequence of n consecutive values of the choice alternative j. Applying Eq. (35) to these histories, we have that $\ln \mathbb{P}(A) = \ln p_\theta(0,0) + 2n\,\widetilde{\alpha}_\theta(j) + 2\sigma_\theta(0) + 2[\sum_{d=1}^{n-1} \sigma_\theta(j,d)] + \sigma_\theta(j,n) + 2[\sum_{d=1}^{n-2} \widetilde{\beta}_d(j,d)] + \widetilde{\beta}_d(j,n-1) + \widetilde{\beta}_d(j,n)$, and $\ln \mathbb{P}(B) = \ln p_\theta(0,0) + 2n\,\widetilde{\alpha}_\theta(j) + 2\sigma_\theta(0) + 2[\sum_{d=1}^{n-1} \sigma_\theta(j,d)] + \sigma_\theta(j,n) + 2[\sum_{d=1}^{n-2} \widetilde{\beta}_d(j,d)] + 2\widetilde{\beta}_d(j,n-1)$, such that $\ln \mathbb{P}(A) - \ln \mathbb{P}(B) = \widetilde{\beta}_d(j,n) - \widetilde{\beta}_d(j,n-1)$. Therefore, the marginal return of going from $n-1$ to n periods of experience in alternative j, $\beta_d(j,n) - \beta_d(j,n-1)$, is identified from $\ln \mathbb{P}(0,0 \mid \mathbf{j}_{n-1}, 0, \mathbf{j}_{n+1}) - \ln \mathbb{P}(0,0 \mid \mathbf{j}_n, 0, \mathbf{j}_n)$. We can also obtain this identification result using the representation of the log-probability in Proposition 9. Histories A and B have the values for the vector of sufficient statistics U: the initial condition, $(y_0, d_1) = (0,0)$, the final choice, $y_{2n+1} = j$, and the histogram of states (y_{t-1}, d_t). As for the identifying statistics, we have that the dyad statistics $D^{(j,k)}$ are the same in the two histories ($D^{(j,j)} = 2n-2$, $D^{(j,0)} = 1$, $D^{(0,j)} = 2$, and for the rest of the dyads $D^{(j,k)} = 0$), but the statistic $\Delta^{(j)}(n+1)$ is equal to 1 for history A and it is zero for history B, the statistic $\Delta^{(j)}(n)$ is equal to 0 for history A and it is one for history B. Therefore, $S(A)'\boldsymbol{\beta}^* - S(B)'\boldsymbol{\beta}^* = \beta_d(j,n) - \beta_d(j,n-1)$. ■

3.6.4. Multinomial forward-looking model with duration dependence

Consider the general multinomial choice model in Eq. (7) with duration dependence and forward-looking agents. The log-probability of a choice history $\widetilde{\mathbf{y}}$ conditional on $\boldsymbol{\theta}$ is:

$$\begin{aligned} \ln \mathbb{P}(\widetilde{\mathbf{y}} \mid \boldsymbol{\theta}) \;&=\; \ln p_\theta(y_0, d_1) + \sum_{t=1}^{T} \sum_{j \neq 0} 1\{y_t = j\}\; [\widetilde{\alpha}_\theta(j) + \widetilde{v}_\theta(j, d_{t+1})] + \sum_{t=1}^{T} \sigma_\theta(y_{t-1}, d_t) \\ &+\; \sum_{t=1}^{T} \sum_{j \neq 0} \left[\sum_{k \neq \{0,j\}} 1\{y_t = j, y_{t-1} = k\}\; \widetilde{\beta}_y(j,k) + 1\{y_t = y_{t-1} = j\}\; \widetilde{\beta}_d(j, d_t) \right]. \end{aligned} \tag{37}$$

In this multinomial choice model, it is possible to identify switching cost parameters without imposing Assumption 2. Proposition 10 establishes the identification of switching costs parameters.

Proposition 10. *In the multinomial forward-looking model with duration dependence under Assumption 1, the log-probability of a choice history has the form* $\ln \mathbb{P}(\widetilde{\mathbf{y}} \mid \boldsymbol{\theta}, \boldsymbol{\beta}) = U' g_\theta + S' \boldsymbol{\beta}^*$ *with*

$$\begin{cases} U = \left[d_1,\; y_0,\; y_T,\; \{H^{(j)}(d), \Delta^{(j)}(d) : j \geq 1,\, d \geq 1\} \right]. \\ S = \left[D^{(j,k)} : j, k \geq 1,\; j \neq k \right]. \\ \boldsymbol{\beta}^* = \left[\widetilde{\beta}_y(k,j) : j, k \geq 1, j \neq k \right]. \end{cases} \tag{38}$$

U is a minimal sufficient statistic for $\boldsymbol{\theta}$. Conditional on U, the elements in the vector of statistics S are linearly independent such that the vector of parameters $\boldsymbol{\beta}^*$ is identified. The duration dependence parameters $\widetilde{\beta}_d(j,d)$ are not identified. ■

Now, in contrast to the result in Proposition 9, the vector of sufficient statistics U includes also the statistics $\{\Delta^{(j)}(d) : j \geq 1, d \geq 1\}$. This implies that, without additional restrictions, we cannot identify the duration dependence parameters $\widetilde{\beta}_d(j,d)$. However, the dyad statistics $D^{(j,k)}$ are not part of the sufficient statistic U and they still provide identification of the parameters $\widetilde{\beta}_y(k,j)$. Example 10 presents a pair of histories that identifies $\widetilde{\beta}_y(k,j)$.

Example 10. Consider the same data and histories as in Example 9(a) but now in a forward-looking model. That is, $T = 3$ and the pair of histories is $A = \{0, 0 \mid 0, j, k\}$ and $B = \{0, 0 \mid j, 0, k\}$ with $j, k \neq 0$ and $j \neq k$. Using the expression for the log-probability of a choice history in Eq. (37) we have that $\ln \mathbb{P}(A) = \ln p_\theta(0,0) + \widetilde{\alpha}_\theta(j) + \widetilde{\alpha}_\theta(k) + 2\sigma_\theta(0) + \sigma_\theta(j,1) + \widetilde{v}_\theta(j,1)$

Table 3
Identification of dynamic multinomial logit models.

Myopic model			*Forward-Looking model*		
Minimal sufficient stat.	Identified parameters	Identifying statistics	Minimal sufficient stat.	Identified parameters	Identifying statistics
Panel 1: Models without duration dependence					
$T^{(j)},\ \Delta^{(j)}: j \geq 1$	$\widetilde{\beta}_y(j,k)$ $j,k \geq 1$	$D^{(k,j)}$: $j,k \geq 1$	$T^{(j)},\ \Delta^{(j)}: j \geq 1$	$\widetilde{\beta}_y(j,k)$ $j,k \geq 1$	$D^{(k,j)}$: $j,k \geq 1$
Panel 2: Models with duration dependence					
For any $j \geq 1$ $\Delta^{(j)}$, $H^{(j)}(d): d \geq 1$	For any $j \geq 1$ $\widetilde{\beta}_y(j,k): j \neq k$ and $\widetilde{\beta}_d(j,d)$: $d \geq 1$	For any $j \geq 1$ $D^{(j,k)}: j \neq k$ and $\Delta^{(j)}(d)$: $d \geq 1$	For any $j \geq 1$ $H^{(j)}(d): d \leq d_j^* - 1$; $\sum_{d \geq d_j^*} H^{(j)}(d)$; $\Delta^{(j)}(d): d \leq d_j^* - 1$; $\sum_{d \geq d_j^*} \Delta^{(j)}(d)$	For any $j \geq 1$ $\widetilde{\beta}_y(j,k): j \neq k$ and $\widetilde{\beta}_d(j,d) - \widetilde{\beta}_d(j,d-1)$ $d \geq d_j^*$	For any $j \geq 1$ $D^{(j,k)}: j \neq k$ and $H^{(j)}(d) + \Delta^{(j)}(d)$ $d \geq d_j^*$

$+\ \widetilde{v}_\theta(k,1) + \widetilde{\beta}_y(k,j)$, and $\ln \mathbb{P}(B) = \ln p_\theta(0,0) + \widetilde{\alpha}_\theta(j) + \widetilde{\alpha}_\theta(k) + 2\sigma_\theta(0) + \sigma_\theta(j,1) + \widetilde{v}_\theta(j,1) + \widetilde{v}_\theta(k,1)$, such that $\ln \mathbb{P}(A) - \ln \mathbb{P}(B) = \widetilde{\beta}_y(k,j)$. Therefore, in this forward-looking model we can still identify the switching cost parameter $\widetilde{\beta}_y(k,j)$ from $\ln \mathbb{P}(0,0 \mid 0,j,k) - \ln \mathbb{P}(0,0 \mid j,0,k)$. ■

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 of a choice history has the form* $\ln \mathbb{P}(\widetilde{\mathbf{y}} \mid \boldsymbol{\theta}, \boldsymbol{\beta}) = U' g_{\boldsymbol{\theta}} + S' \boldsymbol{\beta}^*$ *with*

$$\begin{cases} U = \begin{bmatrix} d_1,\ y_0,\ y_T, \\ \left\{H^{(j)}(d),\ \Delta^{(j)}(d): j \geq 1,\ 1 \leq d \leq d_j^* - 1\right\}, \\ \left\{\sum_{d \geq d_j^*} H^{(j)}(d),\ \sum_{d \geq d_j^*} \Delta^{(j)}(d): j \geq 1\right\} \end{bmatrix}. \\ S = \left[D^{(j,k)}: j,k \geq 1,\ j \neq k;\ H^{(j)}(d_j^*) + \Delta^{(j)}(d_j^*): j \geq 1\right]. \\ \boldsymbol{\beta}^* = \left[\widetilde{\beta}_y(k,j): j,k \geq 1, j \neq k;\ \beta_d(j,d_j^* - 1) - \beta_d(j,d_j^*): j \geq 1\right]. \end{cases} \tag{39}$$

U is a minimal sufficient statistic for $\boldsymbol{\theta}$. Conditional on U, the elements in the vector S are linearly independent such that the vector of parameters $\boldsymbol{\beta}^*$ is identified. ■

Example 11. Suppose that $T \geq 2d_j^* + 1$, and let n be any integer such that $d_j^* \leq n \leq (T-1)/2$. Consider the pair of choice histories $A = \{0,0 \mid \mathbf{j}_{n-1}, 0, \mathbf{j}_{n+1}\}$ and $B = \{0,0 \mid \mathbf{j}_n, 0, \mathbf{j}_n\}$. Applying Eq. (37) to these histories, we have that $\ln \mathbb{P}(A) = \ln p_\theta(0,0) + \alpha_\theta(0) + 2n\,\alpha_\theta(j) + 2\sigma_\theta(0) + 2[\sum_{d=1}^{n-1} \sigma_\theta(j,d)] + \sigma_\theta(j,n) + 2\beta_y(j,0) + \beta_y(0,j) + 2[\sum_{d=1}^{n-2} \beta_d(j,d)] + \beta_d(j,n-1) + \beta_d(j,n) + v_\theta(0) + 2[\sum_{d=1}^{d_j^*-1} v_\theta(j,d)] + 2(n - d_j^* + 1)\, v_\theta(j,d_j^*)$, and $\ln \mathbb{P}(B) = \ln p_\theta(0,0) + \alpha_\theta(0) + 2n\,\alpha_\theta(j) + 2\sigma_\theta(0) + 2[\sum_{d=1}^{n-1} \sigma_\theta(j,d)] + \sigma_\theta(j,n) + 2\beta_y(j,0) + \beta_y(0,j) + 2[\sum_{d=1}^{n-2} \beta_d(j,d)] + 2\beta_d(j,n-1) + v_\theta(0) + 2[\sum_{d=1}^{d_j^*-1} v_\theta(j,d)] + 2(n - d_j^* + 1)\, v_\theta(j,d_j^*)$, such that $\ln \mathbb{P}(A) - \ln \mathbb{P}(B) = \beta_d(j,n) - \beta_d(j,n-1)$. Therefore, the marginal return of experience $\beta_d(j,n) - \beta_d(j,n-1)$ is identified for any value $n \geq d_j^*$. We can also obtain this result using the conditions in Proposition 11. The two choice histories have the same value for the sufficient statistic U, and for $n \geq d_j^*$ we have that $\Delta^{(j)}(n)$ is equal to 0 for history A and equal to 1 for history B. ■

Table 3 summarizes the identification results for the multinomial model.

3.7. 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* $\int [\partial P_{\boldsymbol{\theta}}(y \mid \mathbf{x}, \boldsymbol{\beta}^*)/\partial \mathbf{x}]\, f(\boldsymbol{\theta})\, d\boldsymbol{\theta}$ or $\int [\partial P_{\boldsymbol{\theta}}(y \mid \mathbf{x}, \boldsymbol{\beta}^*)/\partial \boldsymbol{\beta}^*]\, f(\boldsymbol{\theta})\, d\boldsymbol{\theta}$, where $f(\boldsymbol{\theta})$ is the density function of the unobserved heterogeneity.

Without further restrictions, the density function $f(\boldsymbol{\theta})$ is not (nonparametrically) point identified. This is the 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 after the structural parameters $\boldsymbol{\beta}^*$ have been identified: (a) nonparametric finite mixture; and (b) set identification.

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

$$\mathbb{P}\left(\widetilde{\mathbf{y}}|\mathbf{x}_1\right) = \int \left[\prod_{t=1}^{T} P\left(y_t \mid \mathbf{x}_t, \boldsymbol{\beta}^*, \boldsymbol{\theta}\right)\right] f(\boldsymbol{\theta}|\mathbf{x}_1)\, d\boldsymbol{\theta}. \tag{40}$$

The probabilities of choice histories $\mathbb{P}\left(\widetilde{\mathbf{y}}|\mathbf{x}_1\right)$ are identified from the data. Also, for a fixed value of $\boldsymbol{\theta}$, the probabilities $P\left(y_t \mid \mathbf{x}_t, \boldsymbol{\beta}^*, \boldsymbol{\theta}\right)$ are also known to the researcher after the identification of the structural parameters $\boldsymbol{\beta}^*$. Eq. (40) can be seen as a system of linear equations (with a potentially infinite dimension), and the identification of the density function $f(\boldsymbol{\theta}|\mathbf{x}_1)$ is equivalent to finding a unique solution to this system.

Let $|\Theta|$ be the dimension of the support of $\boldsymbol{\theta}$. This dimension can be infinite. Eq. (40) can be written in vector form as,

$$\mathbb{P}_{\mathbf{x}_1} = \mathbf{L}_{\mathbf{x}_1}\, \mathbf{f}_{\mathbf{x}_1}. \tag{41}$$

The term $\mathbb{P}_{\mathbf{x}_1}$ is a vector of dimension $(J+1)^T \times 1$ with the probabilities of all the possible choice histories with initial conditions $\mathbf{x}_1$. The term $\mathbf{L}_{\mathbf{x}_1}$ is a matrix with dimension $(J+1)^T \times |\Theta|$ such that each row contains the probabilities $\prod_{t=1}^{T} P\left(y_t \mid \mathbf{x}_t, \boldsymbol{\beta}^*, \boldsymbol{\theta}\right)$ for a given choice history and for every value of $\boldsymbol{\theta}$. Finally, the term $\mathbf{f}_{\mathbf{x}_1}$ is a $|\Theta| \times 1$ vector with the probabilities $f(\boldsymbol{\theta}|\mathbf{x}_1)$. 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 $\boldsymbol{\theta}$ is continuous, then $|\Theta| = \infty$ and $\mathbf{L}_{\mathbf{x}_1}$ cannot be full-column rank. In fact, the number of rows in matrix $\mathbf{L}_{\mathbf{x}_1}$ – 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 $\boldsymbol{\theta}$ is discrete, and choose the points in the support of the fixed effects, 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:

$$\mathbf{f}_{\mathbf{x}_1} = \left[\mathbf{L}_{\mathbf{x}_1}'\mathbf{L}_{\mathbf{x}_1}\right]^{-1} \mathbf{L}_{\mathbf{x}_1}'\mathbb{P}_{\mathbf{x}_1}. \tag{42}$$

Note that the identification of $\boldsymbol{\beta}^*$ is still based on a fixed-effects model that is robust to this finite-mixture restriction on the distribution of the unobservables. However, under this approach, the point identification of marginal effects depends on this restriction on the points of support of the unobserved heterogeneity.

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 et al. (2013).

Finally, we would like to comment on a practical issue in the identification of the finite-mixture model described above. For the evaluation of the choice probabilities $P\left(y_t \mid \mathbf{x}_t, \boldsymbol{\beta}^*, \boldsymbol{\theta}\right)$ in matrix $\mathbf{L}_{\mathbf{x}_1}$, the vector of unobserved heterogeneity $\boldsymbol{\theta}$ is multidimensional. That is, we need to choose a grid of points for the parameters $\widetilde{\alpha}_{\boldsymbol{\theta}}(j)$ but also for the continuation values $\widetilde{v}_{\boldsymbol{\theta}}(j,d)$.

The selection of this grid of points is relatively simple in the forward-looking model without duration dependence. In this version of the model, the unobserved heterogeneity enters through the term $\tau_{\boldsymbol{\theta}}(j) \equiv \widetilde{\alpha}_{\boldsymbol{\theta}}(j) + \widetilde{v}_{\boldsymbol{\theta}}(j)$. Therefore, for this model we need to fix a grid of points for the J incidental parameters $\{\tau_{\boldsymbol{\theta}}(j) : j > 1\}$. Using a grid of κ points for each parameter $\tau_{\boldsymbol{\theta}}(j)$ we have that the dimension of the density vector $\mathbf{f}_{\mathbf{x}_1}$ is $|\Theta| = \kappa^J$ that should be smaller that $(J+1)^T$ such that the order condition of identification holds.

This problem becomes more complicated in the forward-looking model with duration dependence. In this model, unobserved heterogeneity enters through the term $\tau_{\boldsymbol{\theta}}(j,d) \equiv \widetilde{\alpha}_{\boldsymbol{\theta}}(j) + \widetilde{v}_{\boldsymbol{\theta}}(j,d)$. Therefore, we need to fix a grid of points for the JT incidental parameters $\{\tau_{\boldsymbol{\theta}}(j,d) : j > 1;\ 1 \le d \le T\}$. Using a grid of κ points for each parameter $\tau_{\boldsymbol{\theta}}(j,d)$, we have that the dimension of $\mathbf{f}_{\mathbf{x}_1}$ is $|\Theta| = \kappa^{JT}$ that should be smaller than $(J+1)^T$. This condition imposes a strong restriction on the dimension of unobserved heterogeneity, κ.

However, this approach is not taking into account that the continuation values $\widetilde{v}_{\boldsymbol{\theta}}(j,d)$ are endogenous objects that can be obtained given $\widetilde{\alpha}_{\boldsymbol{\theta}}'s$ and $\boldsymbol{\beta}^*$ by solving the Bellman equation of the model. Taking into account this structure of the model, we can reduce substantially the dimensionality of $\boldsymbol{\theta}$. Given a value of the J incidental parameters $\{\widetilde{\alpha}_{\boldsymbol{\theta}}(j) : j > 1\}$, we can solve the Bellman equation to obtain all the continuation values $\widetilde{v}_{\boldsymbol{\theta}}(j,d)$. Therefore, the dimension of $\boldsymbol{\theta}$ in the structural model with duration dependence is also equal to the dimension of $\{\widetilde{\alpha}_{\boldsymbol{\theta}}(j) : j > 1\}$, 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 will be an Andersen (1970) type of estimator. We illustrate the estimator for the forward-looking multinomial choice model with duration dependence under Assumptions 1 and 2, since estimators for the structural parameters in the other models can be defined in a similar fashion.

4.1. Estimation of $\boldsymbol{\beta}^*$ (given d^*)

Let $\boldsymbol{\beta}^*$ be the vector of identified structural parameters. Let U_i and S_i be the vectors of sufficient and identifying statistics, respectively, for observation i. The conditional MLE for $\boldsymbol{\beta}^*$ is defined as the maximizer of the conditional log-likelihood function:

$$\mathcal{L}_N(\boldsymbol{\beta}^*) = \sum_{i=1}^{N} \mathcal{L}_i(\boldsymbol{\beta}^*) = \sum_{i=1}^{N} S_i'\boldsymbol{\beta}^* - \ln\left(\sum_{\widetilde{\mathbf{y}}: U(\widetilde{\mathbf{y}})=U_i} \exp\left\{ S(\widetilde{\mathbf{y}})'\boldsymbol{\beta}^* \right\} \right) \tag{43}$$

where the condition $\{\widetilde{\mathbf{y}} : U(\widetilde{\mathbf{y}}) = U_i\}$ represents all the choice histories with the same value of U as observation i. This log-likelihood function is globally concave in $\boldsymbol{\beta}^*$, and therefore the computation of the CMLE is straightforward using Newton–Raphson or BHHH algorithm. Under the condition that the spaces of the structural parameters $\boldsymbol{\beta}^*$ and of the incidental parameters $\boldsymbol{\theta}$ are bounded, and using standard arguments (Newey and McFadden, 1994), we have

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}^*) \Rightarrow \mathcal{N}(0, J(\boldsymbol{\beta}^*)^{-1}). \tag{44}$$

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

The main cost in the computation of this estimator comes from the calculation of the statistics $U(\widetilde{\mathbf{y}})$ and $S(\widetilde{\mathbf{y}})$ for every possible choice history $\widetilde{\mathbf{y}}$ (in the sample or not), and from the calculation of the sums of the terms $\exp\left\{S(\widetilde{\mathbf{y}})'\boldsymbol{\beta}^*\right\}$ over all these possible histories. The number of possible histories increases exponentially with the number of time periods, T. For instance, if the number of choice alternatives is six, the number of possible choice histories is close to 8,000 when $T = 5$ but it becomes larger than 60 million when $T = 10$. An approach to deal with this computational burden consists of splitting the original histories in the data into shorter sub-histories. In the new transformed dataset, we have more individual histories but with a shorter time dimension, and we treat two histories from the same individual as if they were from different individuals. This approach is perfectly feasible for the estimation of our model. The Conditional MLE applied to the transformed data has the same asymptotic properties as described above but it implies a loss of efficiency (a larger asymptotic variance) due to the splitting of the original histories.

4.2. Joint estimation of $\boldsymbol{\beta}^*$ and d^*

We describe here a CML estimator for the joint estimation of $(d^*, \boldsymbol{\beta}^*)$ in the case of the binary choice model with forward-looking agents. Let d_0^* represent the true value of the parameter d^*. And let β_0^* be the true value of the parameter $\widetilde{\beta}_d(d_0^*) - \widetilde{\beta}_d(d_0^* - 1)$. We use d^* and β^* to represent arbitrary values of these parameters. We are interested in the joint estimation of (d_0^*, β_0^*) from the maximization of the conditional likelihood function. We consider a *profiling* method. First, for every guess of d^*, we estimate the structural parameter β^* using a constrained CML estimator. Second, given the CML estimates $\widehat{\beta}^*(d^*)$ for $d^* = 1, 2, ..., T-2$ obtained in the first stage, we construct a concentrated likelihood function where the only unknown parameter is d^*, and use a BIC-based method to estimate d_0^*. Finally, given the BIC estimator $\widehat{d}^*$ in stage two and the profile estimates $\widehat{\beta}^*(1), \ldots, \widehat{\beta}^*(T-2)$ in stage one, we obtain the estimator of β_0^* as $\widehat{\beta}^*(\widehat{d}^*)$.

Let L_T be the equal to $\lfloor (T-1)/2 \rfloor$ where $\lfloor . \rfloor$ represent the floor function. For any integer n such that $2 \leq n \leq L_T$, define the pair of histories $A_n = \{0, 0 \mid \mathbf{1}_{n-1}, 0, \mathbf{1}_{T-n}\}$ and $B_n = \{0, 0 \mid \mathbf{1}_n, 0, \mathbf{1}_{T-n-1}\}$. Then, $U_i = \{\widetilde{\mathbf{y}}_i \in A_n \cup B_n \text{ for some } 2 \leq n \leq L_T\}$. Given this statistic, the conditional likelihood function is:

$$\mathcal{L}_N(\boldsymbol{\nu}) = \sum_{n=2}^{L_T}\sum_{i=1}^{N} 1\{\widetilde{\mathbf{y}}_i = A_n\} \ln\left[\frac{\exp\{\nu(n)\}}{1+\exp\{\nu(n)\}}\right] + 1\{\widetilde{\mathbf{y}}_i = B_n\} \ln\left[\frac{1}{1+\exp\{\nu(n)\}}\right] \tag{45}$$

where $\nu(n)$ is a parameter that represents the value $\widetilde{\beta}_d(n) - \widetilde{\beta}_d(n-1) + \int [\widetilde{v}_{\boldsymbol{\theta}}(n+1) - \widetilde{v}_{\boldsymbol{\theta}}(n)]\, f(\boldsymbol{\theta})\, d\boldsymbol{\theta}$, and $\boldsymbol{\nu}$ is the vector of parameters $\{\nu(n) : n = 2, 3, \ldots, L\}$. The unconstrained likelihood function $\mathcal{L}_N(\boldsymbol{\nu})$ is globally concave in each of the parameters $\nu(n)$. It is straightforward to show that the unconstrained CML estimator of $\nu(n)$ is $\widehat{\nu}(n) = \ln \widehat{\mathbb{P}}(A_n) - \ln \widehat{\mathbb{P}}(B_n)$, where $\widehat{\mathbb{P}}(A_n)$ and $\widehat{\mathbb{P}}(B_n)$ are the sample frequencies $N^{-1}\sum_{i=1}^{N} 1\{\widetilde{\mathbf{y}}_i = A_n\}$ and $N^{-1}\sum_{i=1}^{N} 1\{\widetilde{\mathbf{y}}_i = B_n\}$, respectively. However, the model imposes nontrivial constraints on $\nu(n)$, which leads to a constrained CMLE. In particular, the model implies the following relationship between the parameters $\nu(n)$ and the structural parameters (d^*, β^*).

$$\nu(n) = \begin{cases} \text{unrestricted} & \text{if} \quad n < d^*. \\ \beta^* & \text{if} \quad n = d^*. \\ 0 & \text{if} \quad n > d^*. \end{cases} \tag{46}$$

For a given value of d^*, let $\widehat{\boldsymbol{\nu}}_{d^*}^c$ be the constrained estimator of $\boldsymbol{\nu}$ that imposes the restriction in Eq. (46) such that: $\widehat{\nu}_{d^*}^c(n) = \widehat{\nu}(n)$ (unconstrained) for $n \leq d^*$; and $\widehat{\nu}_{d^*}^c(n) = 0$ (constrained) for $n > d^*$. Furthermore, the estimator of the structural parameter β^* is $\widehat{\beta}^*(d^*) = \widehat{\nu}(d^*)$.

We now consider the estimation of d^*. Let $\ell_N(d^*)$ be the concentrated likelihood function $\ell_N(d^*) \equiv \mathcal{L}_N(\widehat{\boldsymbol{\nu}}^c_{d^*})$, i.e., the value of the likelihood given a value of d^* and where the parameters $\boldsymbol{\nu}$ have been estimated under the model restriction in Eq. (46). By definition, we have that:

$$\begin{aligned} \ell_N(d^*) &= N \sum_{n=2}^{d^*} \widehat{\mathbb{P}}(A_n) \ln\left[\frac{\widehat{\mathbb{P}}(A_n)}{\widehat{\mathbb{P}}(A_n)+\widehat{\mathbb{P}}(B_n)}\right] + \widehat{\mathbb{P}}(B_n) \ln\left[\frac{\widehat{\mathbb{P}}(B_n)}{\widehat{\mathbb{P}}(A_n)+\widehat{\mathbb{P}}(B_n)}\right] \\ &+ N \sum_{n=d^*+1}^{L_T} \widehat{\mathbb{P}}(A_n) \ln\left[\frac{1}{2}\right] + \widehat{\mathbb{P}}(B_n) \ln\left[\frac{1}{2}\right]. \end{aligned} \tag{47}$$

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

Proposition 12. *(A) As $N \to \infty$, $N^{-1}\ell_N(d^*)$ 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: $d_0^* = \min\{n : n \in \arg\max_{2\le d^*\le L_T} \ell_0(d^*)\}$.* ■

Given this result, a possible estimator for d_0^* would be the sample analog $\widehat{d}^* = \min\{n : n \in \arg\max_{2\le d^*\le L} \ell_N(d^*)\}$. However, this estimator has 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^*. This is because any value of d^* smaller than L_T implies restrictions on the parameters $\nu(n)$, i.e., $\nu(n) = 0$ for $n > d^*$. The larger the value of d^*, the smaller the number of these restrictions and the larger the value of the likelihood 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 $\{\nu(n)\}$ in the model. In this model, the number of free parameters is equal to d^*. The BIC function is defined as:

$$BIC_N(d^*) = \ell_N(d^*) - \frac{d^*}{2}\ln(N). \tag{48}$$

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

Proposition 13. *Consider the estimator $\widehat{d}_N^* = \arg\max_{2\le d^*\le L_T} BIC_N(d^*)$. 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 as *consistent model selection*: the asymptotic property of the estimator for parameters in the selected model is unaffected (see Pötscher, 1991). However, Pötscher (1991) also points 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 find that the probability of selecting the true d_0^* is very close to 1.[22]

5. Empirical application

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's cumulative mileage, and $c_0 + c_1(m_{it})$ is the maintenance cost.

We incorporate two modifications to 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 $d_{it} \in \{0, 1, 2, \ldots\}$. Using Rust's 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 $y_{it} = 0$ – replace the engine – and it is $\alpha_i(1) + \beta_d(1, d_{it}) + \varepsilon_{it}(1)$ if $y_{it} = 1$ – keep the engine – where $\alpha_i(0) = -RC_i$, $\alpha_i(1) = -c_{0i}$, and $\beta_d(1, d_{it}) = -c_1(d_{it})$.

In Section 5.1, we present evidence from several Monte Carlo experiments using this model. The purpose of conducting these experiments is threefold. First, we want to show that the FE-CMLE can provide precise and robust estimates of structural parameters, even when the sample size is not large. Second, we want to measure the bias of misspecifying the distribution of the unobserved heterogeneity. And third, we want to study the power of a Hausman test – based

[22] For example, for DGP 1 with Sample B – described in Table 5 – $\widehat{d}_N^*$ agrees with the true d_0^* 99% of the times.

Table 4
Description of DGPs in the Monte Carlo experiments.

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

on the comparison of the FE-CMLE and a *Correlated Random Effects* MLE – to reject specifications that wrongly ignore unobserved heterogeneity, or that misspecify its probability distribution. In Section 5.2, we use Rust's dataset to implement the FE-CMLE method, our procedure to estimate d^*, and Hausman test.

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 only has two types. Finally, in DGP 4 there is no 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 — three samples for each of four DGPs. 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.

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**); an MLE that imposes the restriction of no unobserved heterogeneity (that we denote as *MLE-noUH*), and an 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.[23]

We also implement two Hausman tests: a test of the null hypothesis 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 the experiments with DGP 1 in Table 5. The results with the other DGPs are presented in the Appendix.[24]

Table 5 displays results from 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 a stronger correlation between observed durations and unobserved heterogeneity. This generates a larger bias for 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 with *true* d^*. As expected, the CMLEs have larger variance than

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

[24] The *MLE-noUH* estimates a logit model where the error term is $\alpha_i + \varepsilon_{it}$. Therefore, the *MLE-noUH* estimates the parameter $\beta/(1+\sigma)$ where σ is the standard deviation of α_i. In contrast, the other estimators, *FE-CMLE* and *MLE-2types*, control for α_i such that they estimate the parameter β. This implies that there are two sources of discrepancy between the *MLE-noUH* and the other estimators: the bias due to ignoring unobserved heterogeneity; and the different scaling. Since our model includes only one parameter, we cannot control for the different scaling by reporting estimates of ratios relative to a baseline parameter, say β/β_1. Nevertheless, the Hausman test that compares the *MLE-noUH* and the *FE-CMLE* is still a valid test of the null hypothesis of no unobserved heterogeneity because under the null hypothesis we have that $\sigma = 0$ such that there are no different scales.

Table 5
Monte Carlo experiments with DGP 1 (Normal RCs).

Estimator of β	**Sample A** ($t = 1$ to 7)			**Sample B** ($t = 1$ to 14)			**Sample C** ($t = 8$ to 21)		
	Estimate(1)			Estimate(1)			Estimate(1)		
	Mean	Median	St. dev.	Mean	Median	St. dev.	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			Frequency of Ho rejection with significance level			Frequency of Ho rejection with significance level		
	1%	5%	10%	1%	5%	10%	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 1000 replications.

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 Squared Error 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 that of the *CMLE*. That is, in a DGP without an 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. This is no longer the case in samples B and C.

The Hausman test has very strong power to reject the model without unobserved heterogeneity.[25] 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.

5.2. Estimation using Rust's dataset

In Rust's dataset, the full sample contains a total of 124 buses that are classified in eight groups according to bus size and engine manufacturer. For the estimation of the structural model, Rust focuses on groups 1 to 4 — which account for 104 buses. We use this sample of 104 buses. For every bus in the sample, the choice history starts with the actual initial condition of the engine — the month in which the engine was installed. For our analysis, we assume the frequency of the superintendent's decisions to be at the annual level.

Table 6 presents the empirical distribution of the choice histories for the 104 buses. The panel is unbalanced: buses are observed for 2, 4, 6, or 10 years. The observations from all these buses contribute to the Maximum Likelihood estimation of the model without unobserved heterogeneity — that we report in Table 7.

However, not all these choice histories contribute to the conditional maximum likelihood estimation. For instance, the 45 buses without any engine replacement do not contribute to the conditional likelihood function. More generally, a choice history contributes to the conditional likelihood function if there exists other possible choice history – in the sample or not – with the same value for the sufficient statistic U and a different value for the identifying statistic S. The third column of Table 6 indicates whether a choice history contributes or not to the CML estimation of the model. The number of buses that contribute to the CMLE is 46.

Table 7 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 of 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 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 8 presents estimates of the parameter $\beta_d^* \equiv \beta_d(d^*) - \beta_d(d^* - 1)$ using the CMLE and under different values of d^*. We report the value of the concentrated log-likelihood function and of the BIC function for $d^* = 3$ and $d^* = 4$. 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 from zero for a significance level of 10% parameter (p-value = 0.0968). In contrast, for $d^* = 4$, this parameter is not significantly different from zero for any standard significance level (p-value = 0.8446). Therefore, the estimates $\widehat{d}^* = 3$ and $\widehat{\beta}_d^* = 1.7009$ are consistent with the definition of d^* as the maximum duration with $\beta_d(d) - \beta_d(d - 1)$ different from zero.

25 Though the distribution of types in DGP 1 is continuous, the level of unobserved heterogeneity is modest. In the distribution of RC_i, the coefficient of variation is only 25%. Continuous distributions with higher variance imply higher rejection rates of the model with only two types.

Table 6
Bus engine replacement (Rust, 1987)
Empirical distribution of histories with replacement.

Choice history	*Absolute Frequency*	*Does the history contribute to the CMLE?*
With 0 replacements	(45)	
11	15	NO
1111	4	NO
111111	21	NO
1111111111	5	NO
With 1 replacement & T=6	(27)	
110111	2	YES
111011	7	YES
111101	7	YES
111110	11	NO
With 1 replacement & T=10	(31)	
1101111111	1	YES
1110111111	4	YES
1111011111	2	YES
1111101111	7	YES
1111110111	7	YES
1111111011	5	YES
1111111101	3	YES
1111111110	2	NO
With 2 replacements & T=10	(1)	
1101110111	1	YES
Total	104	Sample size for CMLE = 46

Table 7
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 9 compares the CMLE estimate of the parameter β_d^* with the corresponding MLE using the estimates in Table 7. 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 that the magnitude of the difference between MLE and CMLE estimates is substantial and it generates important differences in the distribution of durations.

Table 8
Bus engine replacement (Rust, 1987).

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

Table 9
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

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 A. Proofs

Proof of Lemma 1. We choose alternative $j = 0$ as the baseline. We can write the optimal decision using utilities in deviations with respect to alternative 0. That is,

$$y_t = \arg\max_{j\in\mathcal{Y}} \left\{ \begin{array}{c} \alpha_{\theta}(j) - \alpha_{\theta}(0) + \beta_y(j, y_{t-1}) - \beta_y(0, y_{t-1}) \\ +1\{y_{t-1} = j\}\ \beta_d(j, d_t) + v_{\theta}(j, d_{t+1}) - v_{\theta}(0, 0) + \varepsilon_t(j) \end{array} \right\} \tag{A.1}$$

where we have imposed the restriction that $\beta_d(0, d_t) = 0$, that comes from Assumption 1. For the term related to the switching cost, we have that $\beta_y(j, y_{t-1}) - \beta_y(0, y_{t-1}) = 1\{y_{t-1} = 0\}\ \beta_y(j, 0) + \sum_{k\neq 0} 1\{y_{t-1} = k\}\ [\beta_y(j, k) - \beta_y(0, k)]$, and given that $1\{y_{t-1} = 0\} = 1 - \sum_{k\neq 0} 1\{y_{t-1} = k\}$ we can write this expression as:

$$\beta_y(j, y_{t-1}) - \beta_y(0, y_{t-1}) = \beta_y(j, 0) + \sum_{k\neq 0} 1\{y_{t-1} = k\}[\beta_y(j, k) - \beta_y(0, k) - \beta_y(j, 0)]. \tag{A.2}$$

As for the term associated to the return of experience, $1\{y_{t-1} = j\}\ \beta_d(j, d_t)$, note that it appears multiplied by the dummy variable $1\{y_{t-1} = j\}$. This dummy variable also appears associated to the parameter $-\beta_y(0, j) - \beta_y(j, 0)$ in Eq. (A.2) (note that $\beta_y(j, j) = 0$). Therefore, we cannot separately identify the parameter $-\beta_y(0, j) - \beta_y(j, 0)$ and the parameters in the duration dependence function $\beta_d(j, d_t)$. To avoid this perfect collinearity problem, we can put together the terms $1\{y_{t-1} = j\}\ [-\beta_y(0, j) - \beta_y(j, 0)]$ and $1\{y_{t-1} = j\}\ \beta_d(j, d_t)$. That is,

$$\begin{aligned} &\beta_y(j, y_{t-1}) - \beta_y(0, y_{t-1}) + 1\{y_{t-1} = j\}\ \beta_d(j, d_t) = \\ &1\{y_{t-1} = j\}\ \left[\beta_d(j, d_t) - \beta_y(0, j) - \beta_y(j, 0)\right] + \sum_{k\neq\{0,j\}} 1\{y_{t-1} = k\}\ \left[\beta_y(j, k) - \beta_y(0, k) - \beta_y(j, 0)\right]. \end{aligned} \tag{A.3}$$

Plugging Eq. (A.3) into Eq. (A.1), we have the following reparameterization of the model:

$$y_t = \arg\max_{j\in\mathcal{Y}} \left\{ \widetilde{\alpha}_\theta(j) + \sum_{k\neq\{0,j\}} 1\{y_{t-1}=k\}\,\widetilde{\beta}_y(j,k) + 1\{y_{t-1}=j\}\widetilde{\beta}_d(j,d_t) + \widetilde{v}_\theta(j,d_{t+1}) + \varepsilon_t(j) \right\} \tag{A.4}$$

where $\widetilde{\alpha}_\theta(j) \equiv \alpha_\theta(j) - \alpha_\theta(0) + \beta_y(j,0)$; $\widetilde{\beta}_y(j,k) \equiv \beta_y(j,k) - \beta_y(0,k) - \beta_y(j,0)$; $\widetilde{\beta}_d(j,d) \equiv \beta_d(j,d_t) - \beta_y(0,j) - \beta_y(j,0)$; and $\widetilde{v}_\theta(j,d_{t+1}) \equiv v_\theta(j,d_{t+1}) - v_\theta(0,0)$. ■

Lemma 3 and Proof. The proofs of the Propositions exploit some properties or relationships between the statistics. We summarize these properties in the following Lemma.

Lemma 3. *For any history $\widetilde{\mathbf{y}}$ and choice alternative $j > 0$, the following properties apply: (i) $H^{(j)}(0) = 0$; (ii) $X^{(j)}(0) = 0$; (iii) $\sum_{d\geq 1} H^{(j)}(d) = T^{(j)} + 1\{y_0 = j\} - 1\{y_T = j\}$; (iv) $\sum_{d\geq 1} X^{(j)}(d) = D^{(j,j)}$; (v) for $d \geq 1$, $X^{(j)}(d) = H^{(j)}(d+1) + \Delta^{(j)}(d+1)$; (vi) $\sum_{d\geq 1} \Delta^{(j)}(d) = 1\{y_T = j\} - 1\{y_0 = j\}$; and (vii) $\sum_{k\neq j} D^{(j,k)} = T^{(j)} - D^{(j,j)}$.* ■

Proof of Lemma 3.

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

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

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

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

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

(vi) For any $j > 0$, $\sum_{d\geq 1} \Delta^{(j)}(d) = \sum_{d\geq 1} 1\{y_T = j,\, d_{T+1} = d\} - 1\{y_0 = j,\, d_1 = d\} = 1\{y_T = j\} - 1\{y_0 = j\}$.

(vii) For any $j \geq 1$, $\sum_{k\neq j} D^{(k,j)} = \sum_{t=1}^{T}\sum_{k\neq j} 1\{y_{t-1} = k,\, y_t = j\} = \sum_{t=1}^{T} 1\{y_t = j\} - 1\{y_{t-1} = y_t = j\} = T^{(j)} - D^{(j,j)}$. ■

Proof of Proposition 1. From Eq. (13) we have that $\ln \mathbb{P}(\widetilde{\mathbf{y}} \mid \boldsymbol{\theta}) = \sum_{t=1}^{T} y_t \left[\widetilde{\alpha}_\theta + \widetilde{\beta}_y y_{t-1}\right] + (1 - y_{t-1})\,\sigma_\theta(0) + y_{t-1}\,\sigma_\theta(1) + y_0 \ln p_\theta(1) + (1 - y_0) \ln p_\theta(0)$, and we can write this expression as $\sigma_\theta(0) + \ln p_\theta(0) + \left[\sum_{t=1}^{T} y_t\right] \widetilde{\alpha}_\theta + \left[\sum_{t=1}^{T} y_t y_{t-1}\right] \widetilde{\beta}_y + \left[\sum_{t=1}^{T} y_{t-1}\right] \widetilde{\sigma}_\theta + y_0 \ln \widetilde{p}_\theta$, where $\widetilde{\sigma}_\theta \equiv \sigma_\theta(1) - \sigma_\theta(0)$ and $\ln \widetilde{p}_\theta \equiv \ln p_\theta(1) - \ln p_\theta(0)$. Remember that by definition the statistic $T^{(1)}$ is equal to $\sum_{t=1}^{T} y_t$, and the statistic $D^{(1,1)}$ is equal to $\sum_{t=1}^{T} y_{t-1} y_t$. Also, not that $\sum_{t=1}^{T} y_{t-1} = T^{(1)} + y_0 - y_T$. Therefore, we can write $\ln \mathbb{P}(\widetilde{\mathbf{y}} \mid \boldsymbol{\theta})$ as $T^{(1)}\,[\widetilde{\alpha}_\theta + \widetilde{\sigma}_\theta] + D^{(1,1)}\,\widetilde{\beta}_y + [y_0 - y_T]\,\widetilde{\sigma}_\theta + y_0 \ln \widetilde{p}_\theta$. Or equivalently,

$$\begin{aligned} \ln \mathbb{P}(\widetilde{\mathbf{y}}|\boldsymbol{\theta}) &= y_0\,[\ln \widetilde{p}_\theta + \widetilde{\sigma}_\theta] + y_T\,[-\widetilde{\sigma}_\theta] + T^{(1)}\,[\widetilde{\alpha}_\theta + \widetilde{\sigma}_\theta] \\ &+ D^{(1,1)}\,\widetilde{\beta}_y \end{aligned} \tag{A.5}$$

where we have omitted the term $T\,\sigma_\theta(0) + \ln p_\theta(0)$ because it is constant over all the histories. We can write Eq. (A.5) as $U' g_\theta + S' \boldsymbol{\beta}^*$ with $U = (y_0, y_T, T^{(1)})$, $g_\theta = (\ln \widetilde{p}_\theta + \widetilde{\sigma}_\theta,\, -\widetilde{\sigma}_\theta,\, \widetilde{\alpha}_\theta + \widetilde{\sigma}_\theta)'$, $S = D^{(1,1)}$, and $\boldsymbol{\beta}^* = \widetilde{\beta}_y$. For $T \geq 3$, it is always possible to find a pair of histories, A and B, with the same values for the initial condition y_0, the final choice y_T, and the number of 1's $T^{(1)}$, but with $D_A^{(1,1)} \neq D_B^{(1,1)}$ such that $\widetilde{\beta}_y$ is identified as $[\ln \mathbb{P}(A) - \ln \mathbb{P}(B)]/[D_A^{(1,1)} - D_B^{(1,1)}]$. See Example 1. ■

Proof of Proposition 2. The only difference between the expression for $\ln \mathbb{P}(\widetilde{\mathbf{y}} \mid \boldsymbol{\theta})$ in this forward-looking model and in the myopic model of Proposition 1 is that now $\widetilde{\alpha}_\theta + \widetilde{v}_\theta$ replaces $\widetilde{\alpha}_\theta$ in the vector g_θ. This does not have any influence in the sufficient statistic U or the identifying statistic S. ■

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

$$\ln \mathbb{P}(\widetilde{\mathbf{y}}|\boldsymbol{\theta}) = \sum_{t=1}^{T} y_t \left[\widetilde{\alpha}_\theta + y_{t-1}\,\widetilde{\beta}_d(d_t)\right] + \sigma_\theta(y_{t-1}, d_t) + \ln p_\theta(y_0, d_1). \tag{A.6}$$

We can write this log-probability as $\widetilde{\alpha}_\theta \sum_{t=1}^{T} y_t + \sum_{d\geq 1}[\sum_{t=1}^{T} y_t y_{t-1}\ 1\{d_t = d\}]\ \widetilde{\beta}_d(d) + \sigma_\theta(0) \sum_{t=1}^{T}(1 - y_{t-1}) + \sum_{d\geq 1}[\sum_{t=1}^{T} y_{t-1}\ 1\{d_t = d\}]\ \sigma_\theta(1,d) + \ln p_\theta(y_0, d_1)$. Using the definition of the statistics in Table 1, this expression becomes: $T^{(1)}\widetilde{\alpha}_\theta + \sum_{d\geq 1} X^{(1)}(d)\ \widetilde{\beta}_d(d) + [T - (T^{(1)} + y_0 - y_T)]\ \sigma_\theta(0) + \sum_{d\geq 1} H^{(1)}(d)\ \sigma_\theta(1,d) + \ln p_\theta(y_0, d_1)$. We have that

$\sum_{d\geq 1} H^{(1)}(d) = T^{(1)} + y_0 - y_T$ by Lemma 3(iii). We obtain:

$$\begin{aligned} \ln \mathbb{P}(\widetilde{\mathbf{y}}|\boldsymbol{\theta}) &= \ln p_{\boldsymbol{\theta}}(y_0, d_1) + (y_T - y_0)\,\widetilde{\alpha}_{\boldsymbol{\theta}} + \sum_{d\geq 1} H^{(1)}(d)\;[\widetilde{\alpha}_{\boldsymbol{\theta}} + \widetilde{\sigma}_{\boldsymbol{\theta}}(d)] \\ &+ \sum_{d\geq 1} X^{(1)}(d)\,\widetilde{\beta}_d(d) \end{aligned} \tag{A.7}$$

where $\widetilde{\sigma}_{\boldsymbol{\theta}}(d) \equiv \sigma_{\boldsymbol{\theta}}(1,d) - \sigma_{\boldsymbol{\theta}}(0)$ and we have omitted the term $T\,\sigma_{\boldsymbol{\theta}}(0)$ because T is constant over all the histories. Now, Lemma 3(v) establishes that $X^{(1)}(d) = H^{(1)}(d+1) + \Delta^{(1)}(d+1)$. Then, we have that,

$$\begin{aligned} \ln \mathbb{P}(\widetilde{\mathbf{y}}|\boldsymbol{\theta}) &= \ln p_{\boldsymbol{\theta}}(y_0, d_1) + (y_T - y_0)\,\widetilde{\alpha}_{\boldsymbol{\theta}} + \sum_{d\geq 1} H^{(1)}(d)\;\left[\widetilde{\alpha}_{\boldsymbol{\theta}} + \widetilde{\sigma}_{\boldsymbol{\theta}}(d) + \widetilde{\beta}_d(d-1)\right] \\ &+ \sum_{d\geq 2} \Delta^{(1)}(d)\,\widetilde{\beta}_d(d-1) \end{aligned} \tag{A.8}$$

with $\widetilde{\beta}_d(0) \equiv 0$, just for notational convenience and without loss of generality. We can write Eq. (A.8) as $U'g_{\boldsymbol{\theta}} + S'\boldsymbol{\beta}^*$ with $U = (d_1, y_0, y_T, H^{(1)}(d) : d \geq 1)$, $S = (\Delta^{(1)}(d) : 2 \leq d \geq T-1)$, and $\boldsymbol{\beta}^* = (\widetilde{\beta}_d(d) : 1 \leq d \leq T-2)$. ■

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

$$\ln \mathbb{P}(\widetilde{\mathbf{y}}|\boldsymbol{\theta}) = \ln p_{\boldsymbol{\theta}}(y_0, d_1) + \sum_{t=1}^{T} y_t \left[\widetilde{\alpha}_{\boldsymbol{\theta}} + y_{t-1}\;\widetilde{\beta}_d(d_t) + \widetilde{v}_{\boldsymbol{\theta}}(d_t+1)\right] + \sigma_{\boldsymbol{\theta}}(y_{t-1}, d_t). \tag{A.9}$$

Comparing this log-probability with the one for the myopic model with duration, we can see that the only difference is in the term $\sum_{t=1}^{T} y_t\;\widetilde{v}_{\boldsymbol{\theta}}(d_t+1)$, that can be written as $\sum_{d\geq 0} \widetilde{v}_{\boldsymbol{\theta}}(d+1)\;(\sum_{t=1}^{T} y_t\;1\{d_t = d\})$. Then, taking into account (A.8), we have:

$$\begin{aligned} \ln \mathbb{P}(\widetilde{\mathbf{y}}|\boldsymbol{\theta}) &= \ln p_{\boldsymbol{\theta}}(y_0, d_1) + (y_T - y_0)\,\widetilde{\alpha}_{\boldsymbol{\theta}} + \sum_{d\geq 1} H^{(1)}(d)\;\left[\widetilde{\alpha}_{\boldsymbol{\theta}} + \widetilde{\sigma}_{\boldsymbol{\theta}}(d) + \widetilde{\beta}_d(d-1)\right] \\ &+ \sum_{d\geq 2} \Delta^{(1)}(d)\,\widetilde{\beta}_d(d-1) + \sum_{d\geq 0}\left[\sum_{t=1}^{T} y_t 1\{d_t = d\}\right]\widetilde{v}_{\boldsymbol{\theta}}(d+1)\,. \end{aligned} \tag{A.10}$$

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,

$$\begin{aligned} \sum_{d\geq 0}\left[\sum_{t=1}^{T} y_t\;1\{d_t = d\}\right]\widetilde{v}_{\boldsymbol{\theta}}(d+1) &= \left[T^{(1)} - D^{(1,1)}\right]\;\widetilde{v}_{\boldsymbol{\theta}}(1) + \sum_{d\geq 1} X^{(1)}(d)\,\widetilde{v}_{\boldsymbol{\theta}}(d+1) \\ &= T^{(1)}\,\widetilde{v}_{\boldsymbol{\theta}}(1) + \sum_{d\geq 1} X^{(1)}(d)\;[\widetilde{v}_{\boldsymbol{\theta}}(d+1) - \widetilde{v}_{\boldsymbol{\theta}}(1)] \end{aligned} \tag{A.11}$$

where, for the second equality, we have applied Lemma 3(iv), $D^{(1,1)} = \sum_{d\geq 1} X^{(1)}(d)$. Then, plugging (A.11) into (A.10), we have:

$$\begin{aligned} \ln \mathbb{P}(\widetilde{\mathbf{y}}|\boldsymbol{\theta}) &= \ln p_{\boldsymbol{\theta}}(y_0, d_1) + (y_T - y_0)\,\widetilde{\alpha}_{\boldsymbol{\theta}} + \sum_{d\geq 1} H^{(1)}(d)\;\left[\widetilde{\alpha}_{\boldsymbol{\theta}} + \widetilde{\sigma}_{\boldsymbol{\theta}}(d) + \widetilde{\beta}_d(d-1)\right] \\ &+ \sum_{d\geq 2} \Delta^{(1)}(d)\,\widetilde{\beta}_d(d-1) + T^{(1)}\,\widetilde{v}_{\boldsymbol{\theta}}(1) + \sum_{d\geq 1} X^{(1)}(d)\;[\widetilde{v}_{\boldsymbol{\theta}}(d+1) - \widetilde{v}_{\boldsymbol{\theta}}(1)]\,. \end{aligned} \tag{A.12}$$

From Lemma 3(iii), we have that $T^{(1)} = \sum_{d\geq 1} H^{(1)}(d) + (y_T - y_0)$; and from Lemma 3(v) $X^{(1)}(d) = H^{(1)}(d+1) + \Delta^{(1)}(d+1)$. Solving these expressions in (A.12), we have that:

$$\begin{aligned} \ln \mathbb{P}(\widetilde{\mathbf{y}}|\boldsymbol{\theta}) &= \ln p_{\boldsymbol{\theta}}(y_0, d_1) + (y_T - y_0)\;[\widetilde{\alpha}_{\boldsymbol{\theta}} + \widetilde{v}_{\boldsymbol{\theta}}(1)] + \sum_{d\geq 1} H^{(1)}(d)\left[\widetilde{\alpha}_{\boldsymbol{\theta}} + \widetilde{\sigma}_{\boldsymbol{\theta}}(d) + \widetilde{v}_{\boldsymbol{\theta}}(1) + \widetilde{\beta}_d(d-1)\right] \\ &+ \sum_{d\geq 1}\left[H^{(1)}(d) + \Delta^{(1)}(d)\right][\widetilde{v}_{\boldsymbol{\theta}}(d) - \widetilde{v}_{\boldsymbol{\theta}}(1)] + \sum_{d\geq 2} \Delta^{(1)}(d)\,\widetilde{\beta}_d(d-1). \end{aligned} \tag{A.13}$$

Taking into account that, by Lemma 3(iii), $y_T - y_0 = \sum_{d\geq 1}\Delta^{(1)}(d)$, we have:

$$\begin{aligned}\ln \mathbb{P}\left(\widetilde{\mathbf{y}}|\boldsymbol{\theta}\right) &= \ln p_{\boldsymbol{\theta}}(y_0, d_1) + \sum_{d\geq 1} H^{(1)}(d)\ \left[\widetilde{\alpha}_{\boldsymbol{\theta}} + \widetilde{\sigma}_{\boldsymbol{\theta}}(d) + \widetilde{\nu}_{\boldsymbol{\theta}}(d) + \widetilde{\beta}_d(d-1)\right] \\ &+ \sum_{d\geq 1}\Delta^{(1)}(d)\ \left[\widetilde{\alpha}_{\boldsymbol{\theta}} + \widetilde{\nu}_{\boldsymbol{\theta}}(d) + \widetilde{\beta}_d(d-1)\right].\end{aligned} \tag{A.14}$$

We can present this equation as follows:

$$\ln \mathbb{P}\left(\widetilde{\mathbf{y}}|\boldsymbol{\theta}\right) = \ln p_{\boldsymbol{\theta}}(y_0, d_1) + \sum_{d\geq 1} H^{(1)}(d)\ \left[g_{\boldsymbol{\theta},1}(d) + \widetilde{\beta}_d(d-1)\right] + \sum_{d\geq 1}\Delta^{(1)}(d)\ \left[g_{\boldsymbol{\theta},2}(d) + \widetilde{\beta}_d(d-1)\right] \tag{A.15}$$

with $g_{\boldsymbol{\theta},1}(d) \equiv \widetilde{\alpha}_{\boldsymbol{\theta}} + \widetilde{\sigma}_{\boldsymbol{\theta}}(d) + \widetilde{\nu}_{\boldsymbol{\theta}}(d)$;and $g_{\boldsymbol{\theta},2}(d) \equiv \widetilde{\alpha}_{\boldsymbol{\theta}} + \widetilde{\nu}_{\boldsymbol{\theta}}(d)$. Therefore, the vector of sufficient statistics U is $[y_0, d_1, \{H^{(1)}(d), \Delta^{(1)}(d) : d \geq 1\}]$. All the statistics associated to the parameters $\widetilde{\beta}_d$ are also part of U. Therefore, the parameters $\widetilde{\beta}_d$ are not identified. ■

Proof of Proposition 5. Let $g_{\boldsymbol{\theta},1}(d)$ and $g_{\boldsymbol{\theta},2}(d)$ be the functions defined in the proof of Proposition 4. Under Assumption 2, we have that $\widetilde{\nu}_{\boldsymbol{\theta}}(d) = \widetilde{\nu}_{\boldsymbol{\theta}}(d^*)$ and $\widetilde{\sigma}_{\boldsymbol{\theta}}(d) = \widetilde{\sigma}_{\boldsymbol{\theta}}(d^*)$ for any $d \geq d^*$. This implies that $g_{\boldsymbol{\theta},1}(d) = g_{\boldsymbol{\theta},1}(d^*)$ and $g_{\boldsymbol{\theta},2}(d) = g_{\boldsymbol{\theta},2}(d^*)$ for any $d \geq d^*$. Under Assumption 2, we can re-write Eq. (A.15) as:

$$\begin{aligned}\ln \mathbb{P}\left(\widetilde{\mathbf{y}}|\boldsymbol{\theta}\right) &= \ln p_{\boldsymbol{\theta}}(y_0, d_1) + \sum_{1\leq d\leq d^*-1} H^{(1)}(d)\, g_{\boldsymbol{\theta},1}(d) + \left[\sum_{d\geq d^*} H^{(1)}(d)\right] g_{\boldsymbol{\theta},1}(d^*) \\ &+ \sum_{1\leq d\leq d^*-1}\Delta^{(1)}(d)\, g_{\boldsymbol{\theta},2}(d) + \left[\sum_{d\geq d^*}\Delta^{(1)}(d)\right] g_{\boldsymbol{\theta},2}(d^*) \\ &+ \sum_{d\geq 2}\left[H^{(1)}(d) + \Delta^{(1)}(d)\right]\widetilde{\beta}_d(d-1).\end{aligned} \tag{A.16}$$

Under Assumption 2, we have that $\widetilde{\beta}_d(d-1) = \widetilde{\beta}_d(d^*)$ for any $d \geq d^* + 1$. This implies that we can represent $\sum_{d\geq 2}[H^{(1)}(d) + \Delta^{(1)}(d)]\ \widetilde{\beta}_d(d-1)$ as the sum of three terms:

$$\begin{aligned}&\sum_{d\geq 2}[H^{(1)}(d) + \Delta^{(1)}(d)]\ \widetilde{\beta}_d(d-1) = \sum_{2\leq d\leq d^*-1}[H^{(1)}(d) + \Delta^{(1)}(d)]\ \widetilde{\beta}_d(d-1) \\ &+ \left[H^{(1)}(d^*) + \Delta^{(1)}(d^*)\right]\left[\widetilde{\beta}_d(d^*-1) - \widetilde{\beta}_d(d^*)\right] \\ &+ \left[\sum_{d\geq d^*} H^{(1)}(d) + \Delta^{(1)}(d)\right]\ \widetilde{\beta}_d(d^*).\end{aligned} \tag{A.17}$$

Plugging Eq. (A.17) into (A.16), we get:

$$\begin{aligned}\ln \mathbb{P}\left(\widetilde{\mathbf{y}}|\boldsymbol{\theta}\right) &= \ln p_{\boldsymbol{\theta}}(y_0, d_1) \\ &+ \sum_{1\leq d\leq d^*-1} H^{(1)}(d)\ \left[g_{\boldsymbol{\theta},1}(d) + \widetilde{\beta}_d(d-1)\right] + \left[\sum_{d\geq d^*} H^{(1)}(d)\right]\left[g_{\boldsymbol{\theta},1}(d^*) + \widetilde{\beta}_d(d^*)\right] \\ &+ \sum_{1\leq d\leq d^*-1}\Delta^{(1)}(d)\ \left[g_{\boldsymbol{\theta},2}(d) + \widetilde{\beta}_d(d-1)\right] + \left[\sum_{d\geq d^*}\Delta^{(1)}(d)\right]\left[g_{\boldsymbol{\theta},2}(d^*) + \widetilde{\beta}_d(d^*)\right] \\ &+ \left[H^{(1)}(d^*) + \Delta^{(1)}(d^*)\right]\ \left[\widetilde{\beta}_d(d^*-1) - \widetilde{\beta}_d(d^*)\right].\end{aligned} \tag{A.18}$$

Eq. (A.18) implies that the vector of sufficient statistics U is $[d_1, y_0, \{H^{(1)}(d), \Delta^{(1)}(d) : d \leq d^*-1\}, \sum_{d\geq d^*} H^{(1)}(d), \sum_{d\geq d^*}\Delta^{(1)}(d)]$, the identifying statistic is $S = H^{(1)}(d^*) + \Delta^{(1)}(d^*)$, and it identifies the parameter $\boldsymbol{\beta}^* = \beta_d(1, d^*-1) - \beta_d(1, d^*)$. ■

Proof of Propositions 7 and 8. For this model, the log probability is $\ln p_{\boldsymbol{\theta}}(y_0) + \sum_{t=1}^{T}\sum_{j\neq 0} 1\{y_t = j\}\ \widetilde{\alpha}_{\boldsymbol{\theta}}(j) + \sum_{t=1}^{T}\sum_{j\neq 0}\sum_{k\neq 0} 1\{y_t = j,\, y_{t-1} = k\}\ \widetilde{\beta}_y(j,k) + \sum_{t=1}^{T}\sum_{j=0}^{J} 1\{y_{t-1} = j\}\ \sigma_{\boldsymbol{\theta}}(j)$. Using the definitions of our statistics, we have

that:

$$\ln \mathbb{P}\left(\widetilde{\mathbf{y}}|\boldsymbol{\theta}\right) = \ln p_{\boldsymbol{\theta}}(y_0) + \sum_{j=1}^{J} T^{(j)}\, \widetilde{\alpha}_{\boldsymbol{\theta}}(j) + \sum_{j\neq 0}\sum_{k\neq 0} D^{(j,k)} \widetilde{\beta}_y(j,k) + \sum_{j=0}^{J} \left[T^{(j)} - \Delta^{(j)}\right] \sigma_{\boldsymbol{\theta}}(j) \tag{A.19}$$

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:

$$\begin{aligned} \ln \mathbb{P}\left(\widetilde{\mathbf{y}}|\boldsymbol{\theta}\right) &= \ln p_{\boldsymbol{\theta}}(y_0) + \sum_{j=1}^{J} T^{(j)}\ \left[\widetilde{\alpha}_{\boldsymbol{\theta}}(j) + \widetilde{\sigma}_{\boldsymbol{\theta}}(j)\right] + \sum_{j=1}^{J} \Delta^{(j)}\ \left[-\widetilde{\sigma}_{\boldsymbol{\theta}}(j)\right] \\ &+ \sum_{j\neq 0}\sum_{k\neq 0} D^{(j,k)}\ \widetilde{\beta}_y(j,k) \end{aligned} \tag{A.20}$$

with $\widetilde{\sigma}_{\boldsymbol{\theta}}(j) \equiv \sigma_{\boldsymbol{\theta}}(j) - \sigma_{\boldsymbol{\theta}}(0)$. Note that we have omitted the term $(T-1)\,\sigma_{\boldsymbol{\theta}}(0)$ because it does not vary over the different histories. ■

Proof of Proposition 9. For this model, the log probability of a choice history is $\ln p_{\boldsymbol{\theta}}(y_0, d_1) + \sum_{j=1}^{J}\sum_{t=1}^{T} 1\{y_t = j\}\ \widetilde{\alpha}_{\boldsymbol{\theta}}(j) + \sum_{j=1}^{J}\sum_{k\neq\{0,j\}}\sum_{t=1}^{T} 1\{y_{t-1} = j,\ y_t = k\}\ \widetilde{\beta}_y(k,j) + \sum_{j=1}^{J}\sum_{d\geq 1}\sum_{t=1}^{T} 1\{y_{t-1} = y_t = j,\ d_t = d\}\ \widetilde{\beta}_d(j,d) + \sum_{t=1}^{T} 1\{y_{t-1} = 0\}\ \sigma_{\boldsymbol{\theta}}(0) + \sum_{j=1}^{J}\sum_{d\geq 1}\sum_{t=1}^{T} 1\{y_{t-1} = j,\ d_t = d\}\ \sigma_{\boldsymbol{\theta}}(j,d)$. Note that $1\{y_{t-1} = 0\} = 1 - \sum_{j=1}^{J}\sum_{d\geq 1} 1\{y_{t-1} = j,\ d_t = d\}$, such that the last two terms can be written as $T\ \sigma_{\boldsymbol{\theta}}(0) + \sum_{j=1}^{J}\sum_{d\geq 1}\sum_{t=1}^{T} 1\{y_{t-1} = j,\ d_t = d\}\ \widetilde{\sigma}_{\boldsymbol{\theta}}(j,d)$, with $\widetilde{\sigma}_{\boldsymbol{\theta}}(j,d) = \sigma_{\boldsymbol{\theta}}(j,d) - \sigma_{\boldsymbol{\theta}}(0)$. Using the definition of the statistics in Table 1, we can write this log-probability as follows:

$$\begin{aligned} \ln \mathbb{P}\left(\widetilde{\mathbf{y}}|\boldsymbol{\theta}\right) &= \ln p_{\boldsymbol{\theta}}(y_0, d_1) + \sum_{j=1}^{J} T^{(j)}\ \widetilde{\alpha}_{\boldsymbol{\theta}}(j) + \sum_{j=1}^{J}\sum_{d\geq 1} H^{(j)}(d)\ \widetilde{\sigma}_{\theta}(j,d) \\ &+ \sum_{j=1}^{J}\sum_{k\neq\{0,j\}} D^{(j,k)}\ \widetilde{\beta}_y(j,k) + \sum_{j=1}^{J}\sum_{d\geq 1} X^{(j)}(d)\ \widetilde{\beta}_d(j,d). \end{aligned} \tag{A.21}$$

By Lemma 3(iii), $T^{(j)} = \Delta^{(j)} + \sum_{d\geq 1} H^{(j)}(d)$. And by Lemma 3(v), we have that $X^{(j)}(d) = H^{(j)}(d+1) - \Delta^{(j)}(d+1)$. Plugging this expressions into Eq. (A.21), we have that:

$$\begin{aligned} \ln \mathbb{P}\left(\widetilde{\mathbf{y}}|\boldsymbol{\theta}\right) &= \ln p_{\boldsymbol{\theta}}(y_0, d_1) + \sum_{j=1}^{J}\sum_{d\geq 1} H^{(j)}(d)\ \left[\widetilde{\alpha}_{\boldsymbol{\theta}}(j) + \widetilde{\sigma}_{\theta}(j,d)\right] \\ &+ \sum_{j=1}^{J} \left[1\{y_T = j\} - 1\{y_0 = j\}\right] \widetilde{\alpha}_{\boldsymbol{\theta}}(j) \\ &+ \sum_{j=1}^{J}\sum_{k\neq\{0,j\}} D^{(j,k)} \widetilde{\beta}_y(j,k) + \sum_{j=1}^{J}\sum_{d\geq 1} \left[H^{(j)}(d+1) + \Delta^{(j)}(d+1)\right] \widetilde{\beta}_d(j,d), \end{aligned} \tag{A.22}$$

or

$$\begin{aligned} \ln \mathbb{P}\left(\widetilde{\mathbf{y}}|\boldsymbol{\theta}\right) &= \ln p_{\boldsymbol{\theta}}(y_0, d_1) + \sum_{j=1}^{J}\sum_{d\geq 1} H^{(j)}(d)\ \left[\widetilde{\alpha}_{\boldsymbol{\theta}}(j) + \widetilde{\sigma}_{\theta}(j,d) + \widetilde{\beta}_d(j,d-1)\right] \\ &+ \sum_{j=1}^{J} \left[1\{y_T = j\} - 1\{y_0 = j\}\right] \widetilde{\alpha}_{\boldsymbol{\theta}}(j) \\ &+ \sum_{j=1}^{J}\sum_{k\neq\{0,j\}} D^{(j,k)}\ \widetilde{\beta}_y(j,k) + \sum_{j=1}^{J}\sum_{d\geq 1} \Delta^{(j)}(d)\ \widetilde{\beta}_d(j,d-1), \end{aligned} \tag{A.23}$$

where we adopt the notation $\widetilde{\beta}_d(j,0) = 0$. This expression implies that $\ln \mathbb{P}\left(\widetilde{\mathbf{y}}\,|\boldsymbol{\theta}, \boldsymbol{\beta}\right) = U' g_{\boldsymbol{\theta}} + S' \boldsymbol{\beta}^*$, with $U = [d_1, y_0, y_T, \{H^{(j)}(d) : j \geq 1,\ d \geq 1\}]$, $S = [D^{(j,k)} : j,k \geq 1,\ j \neq k;\ \Delta^{(j)}(d) : j \geq 1;\ d \geq 2]$, and $\boldsymbol{\beta}^* = [\widetilde{\beta}_y(k,j) : j,k \geq 1,\ j \neq k;\ \widetilde{\beta}_d(j,d) : j \geq 1;\ d \geq 1]$. ■

Proof of Proposition 10. The expression of the log-probability is similar as in Proposition 9 but now we have the additional term $\sum_{t=1}^{T} \widetilde{v}_{\boldsymbol{\theta}}(y_t, d_{t+1})$ that can be written as $\sum_{j=1}^{J}\sum_{d\geq 1}\sum_{t=1}^{T} 1\{y_t = j, d_{t+1} = d\}\ \widetilde{v}_{\boldsymbol{\theta}}(j,d)$. Note that the statistic

$\sum_{t=1}^{T} 1\{y_t = j, d_{t+1} = d\}$ can be written as $H^{(j)}(d) + \Delta^{(j)}(d)$, such that $\sum_{t=1}^{T} \widetilde{v}_{\theta}(y_t, d_{t+1}) = \sum_{j=1}^{J} \sum_{d\geq 1} [H^{(j)}(d) + \Delta^{(j)}(d)]$ $\widetilde{v}_{\theta}(j, d)$. Using Eq. (A.23) from the proof of Proposition 9, and adding this additional term associated to the continuation values, we have

$$\begin{aligned} \ln \mathbb{P}(\widetilde{\mathbf{y}}|\boldsymbol{\theta}) &= \ln p_{\theta}(y_0, d_1) + \sum_{j=1}^{J} \sum_{d\geq 1} H^{(j)}(d) \left[\widetilde{\alpha}_{\theta}(j) + \widetilde{\sigma}_{\theta}(j, d) + \widetilde{v}_{\theta}(j, d) + \widetilde{\beta}_{d}(j, d-1)\right] \\ &+ \sum_{j=1}^{J} [1\{y_T = j\} - 1\{y_0 = j\}] \widetilde{\alpha}_{\theta}(j) \\ &+ \sum_{j=1}^{J} \sum_{k\neq\{0,j\}} D^{(j,k)} \widetilde{\beta}_{y}(j, k) + \sum_{j=1}^{J} \sum_{d\geq 1} \Delta^{(j)}(d) \left[\widetilde{\beta}_{d}(j, d-1) + \widetilde{v}_{\theta}(j, d)\right]. \end{aligned} \tag{A.24}$$

By Lemma 3(vi), $1\{y_T = j\} - 1\{y_0 = j\} = \sum_{d\geq 1} \Delta^{(j)}(d)$, and we have

$$\begin{aligned} \ln \mathbb{P}(\widetilde{\mathbf{y}}|\boldsymbol{\theta}) &= \ln p_{\theta}(y_0, d_1) + \sum_{j=1}^{J} \sum_{d\geq 1} H^{(j)}(d) \left[g_{\theta,1}(j, d) + \widetilde{\beta}_{d}(j, d-1)\right] \\ &+ \sum_{j=1}^{J} \sum_{d\geq 1} \Delta^{(j)}(d) \left[g_{\theta,2}(j, d) + \widetilde{\beta}_{d}(j, d-1)\right] \\ &+ \sum_{j=1}^{J} \sum_{k\neq\{0,j\}} D^{(j,k)} \widetilde{\beta}_{y}(j, k) \end{aligned} \tag{A.25}$$

with $g_{\theta,1}(j, d) \equiv \widetilde{\alpha}_{\theta}(j) + \widetilde{\sigma}_{\theta}(j, d) + \widetilde{v}_{\theta}(j, d)$, and $g_{\theta,2}(j, d) \equiv \widetilde{\alpha}_{\theta}(j) + \widetilde{v}_{\theta}(j, d)$. This expression implies that the vector of sufficient statistics is $U = [d_1, y_0, y_T, \{H^{(j)}(d), \Delta^{(j)}(d) : j \geq 1, d \geq 1\}]$, the vector of identifying statistics is $S = [D^{(j,k)} : j, k \geq 1, j \neq k]$, and the vector of identified parameters is $\boldsymbol{\beta}^* = [\widetilde{\beta}_{y}(k, j) : j, k \geq 1, j \neq k]$. ■

Proof of Proposition 11. Let $g_{\theta,1}(j, d)$ and $g_{\theta,2}(j, d)$ be the functions defined in the proof of Proposition 10. Under Assumption 2, we have that $\widetilde{v}_{\theta}(j, d) = \widetilde{v}_{\theta}(j, d_j^*)$ and $\widetilde{\sigma}_{\theta}(j, d) = \widetilde{\sigma}_{\theta}(j, d_j^*)$ for any $d \geq d_j^*$. This implies that $g_{\theta,1}(j, d) = g_{\theta,1}(j, d_j^*)$ and $g_{\theta,2}(j, d) = g_{\theta,2}(j, d_j^*)$ for any $d \geq d_j^*$. Under Assumption 2, we can re-write Eq. (A.25) as:

$$\begin{aligned} \ln \mathbb{P}(\widetilde{\mathbf{y}}|\boldsymbol{\theta}) &= \ln p_{\theta}(y_0, d_1) + \sum_{j=1}^{J} \sum_{1\leq d\leq d_j^*-1} H^{(j)}(d)\, g_{\theta,1}(j, d) + \sum_{j=1}^{J} \left[\sum_{d\geq d_j^*} H^{(j)}(d)\right] g_{\theta,1}(j, d_j^*) \\ &+ \sum_{j=1}^{J} \sum_{1\leq d\leq d_j^*-1}^{(j)} \Delta^{(j)}(d)\, g_{\theta,2}(j, d) + \sum_{j=1}^{J} \left[\sum_{d\geq d_j^*} \Delta^{(j)}(d)\right] g_{\theta,2}(j, d_j^*) \\ &+ \sum_{j=1}^{J} \sum_{d\geq 1} \left[H^{(j)}(d) + \Delta^{(j)}(d)\right] \widetilde{\beta}_{d}(j, d-1) + \sum_{j=1}^{J} \sum_{k\neq\{0,j\}} D^{(j,k)} \widetilde{\beta}_{y}(j, k). \end{aligned} \tag{A.26}$$

Under Assumption 2, we have that $\widetilde{\beta}_{d}(j, d-1) = \widetilde{\beta}_{d}(j, d_j^*)$ for any $d \geq d_j^* + 1$. This implies that we can represent $\sum_{d\geq 1} [H^{(j)}(d) + \Delta^{(j)}(d)] \widetilde{\beta}_{d}(j, d-1)$ as the sum of three terms:

$$\begin{aligned} \sum_{d\geq 1} \left[H^{(j)}(d) + \Delta^{(j)}(d)\right] \widetilde{\beta}_{d}(j, d-1) &= \sum_{1\leq d\leq d_j^*-1} \left[H^{(j)}(d) + \Delta^{(j)}(d)\right] \widetilde{\beta}_{d}(j, d-1) \\ &+ \left[H^{(j)}(d_j^*) + \Delta^{(j)}(d_j^*)\right] \left[\widetilde{\beta}_{d}(j, d_j^*-1) - \widetilde{\beta}_{d}(j, d_j^*)\right] \\ &+ \left[\sum_{d\geq d_j^*} H^{(j)}(d) + \Delta^{(j)}(d)\right] \widetilde{\beta}_{d}(j, d_j^*). \end{aligned} \tag{A.27}$$

Plugging Eq. (A.27) into (A.26), we get:

$$\begin{aligned}\ln \mathbb{P}\left(\tilde{\mathbf{y}}|\boldsymbol{\theta}\right) &= \ln p_{\boldsymbol{\theta}}(y_0, d_1) \\ &+ \sum_{j=1}^{J} \sum_{1\le d\le d_j^*-1} H^{(j)}(d)\left[g_{\boldsymbol{\theta},1}(j,d) + \widetilde{\beta}_d(j,d-1)\right] + \sum_{j=1}^{J}\left[\sum_{d\ge d_j^*} H^{(j)}(d)\right]\left[g_{\boldsymbol{\theta},1}(j,d_j^*) + \widetilde{\beta}_d(j,d_j^*)\right] \\ &+ \sum_{j=1}^{J} \sum_{1\le d\le d_j^*-1} \Delta^{(j)}(d)\left[g_{\boldsymbol{\theta},2}(j,d) + \widetilde{\beta}_d(j,d-1)\right] + \sum_{j=1}^{J}\left[\sum_{d\ge d_j^*} \Delta^{(j)}(d)\right]\left[g_{\boldsymbol{\theta},2}(j,d_j^*) + \widetilde{\beta}_d(j,d_j^*)\right] \\ &+ \left[H^{(j)}(d_j^*) + \Delta^{(j)}(d_j^*)\right]\left[\widetilde{\beta}_d(j,d_j^*-1) - \widetilde{\beta}_d(j,d_j^*)\right] + \sum_{j=1}^{J}\sum_{k\ne\{0,j\}} D^{(j,k)}\,\widetilde{\beta}_y(j,k).\end{aligned} \tag{A.28}$$

Eq. (A.28) implies that the vector of sufficient statistics U is $[d_1, y_0, y_T, \{H^{(j)}(d), \Delta^{(j)}(d) : j \ge 1, 1 \le d \le d_j^*-1\}, \sum_{d\ge d_j^*} H^{(j)}(d), \sum_{d\ge d_j^*}^{(j)} \Delta^{(j)}(d)]$, the vector of identifying statistics is $S = [D^{(j,k)} : j,k \ge 1, j \ne k;\ H^{(j)}(d_j^*) + \Delta^{(j)}(d_j^*)]$, and the vector of identified parameters is $\boldsymbol{\beta}^* = [\widetilde{\beta}_y(k,j) : j,k \ge 1, j \ne k;\ \widetilde{\beta}_d(j,d_j^*-1) - \widetilde{\beta}_d(j,d_j^*)]$. ■

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

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

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:

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

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 $n \ge d_0^*$, $N^{-1}\ell_N(n) \to_p \ell_0(d_0^*)$ 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 C such that $cN \leq 2\left[\ell_N(d_0^*) - \ell_N(d_0^* - 1)\right] \leq CN$; 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 B. 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 are *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 natural numbers, $\{0, 1, 2, \ldots\}$, and its transition rule is $x_{t+1}^d = 1\{y_t > 0\}[\ 1\left\{y_t = x_t^y\right\}\ x_t^d + 1 + \xi_{t+1}^d]$, where ξ_{t+1}^d has support $\{0, 1, 2, \ldots\}$, 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. 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′ 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 $\Pi_t(j) = \alpha_{\boldsymbol{\theta}}(j) + 1\{j = x_t^y\}\ \beta_d\left(j, x_t^d\right) + 1\{j \neq x_t^y\}\ \beta_y\left(j, x_t^y\right) + \varepsilon_t(j)$. (C) $\beta_y(j, j) = 0$, $\beta_d\left(0, x^d\right) = 0$. (D) $\{\varepsilon_t(j) : j \in \mathcal{Y}\}$ are *i.i.d.* over (i, t, j) with an extreme value type I distribution. (E) $\mathbf{z}_t$ follows a time-homogeneous Markov process. (F) The probability distribution of $\boldsymbol{\theta}$ conditional on $\{\mathbf{z}_t, \mathbf{x}_t : t = 1, 2, \ldots\}$ 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 $\mathbf{x}_t$; $x_t^d \in \{0, 1, \ldots, \infty\}$, and $x_{t+1}^d = 1\{y_t > 0\}[\ 1\left\{y_t = x_t^y\right\}\ x_t^d + 1 + \xi_{t+1}^d]$, where ξ_{t+1}^dhas support $\{0, 1, \ldots, \infty\}$, and it is i.i.d. over time and independent of $\mathbf{x}_t$. ■

The model has the following integrated Bellman equation:

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

where $\mathbb{E}_{\xi_{t+1}}(.)$ is the expectation over the distribution of $(\xi_{t+1}^y,\ \xi_{t+1}^d)$. Let $v_{\boldsymbol{\theta}}(j, \mathbf{x}_t)$ be the continuation value function $\delta\ \mathbb{E}_{\xi_{t+1}}[V_{\boldsymbol{\theta}}\left(f_y(j, \xi_{t+1}^y),\ 1\left\{j = x_t^y\right\} x_t^d +\ 1 + \xi_{t+1}^d\right)]$. 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, the model also has Property 1 and Property 2.

Property 1. In a model without duration dependence (i.e., $\beta_d = 0$), the continuation value of choosing alternative j becomes $v_{\theta_i}(j)$, which does not depend on the state variable, x_t^y. Note that the continuation value function becomes $v_{\boldsymbol{\theta}}(j) = \delta\ \mathbb{E}_{\xi_{t+1}}[V_{\boldsymbol{\theta}}(f_y(j, \xi_{t+1}^y))]$.

Property 2. In the model with duration dependence, if $y_t = j = x_t^y$, then the continuation value becomes $v_{\boldsymbol{\theta}}(j, x_t^d + 1)$. Under Assumption 2, if $y_t = j = x_t^y$ the continuation value function is such that $v_{\boldsymbol{\theta}}(j, x_t^d + 1) = v_{\boldsymbol{\theta}}(j, d_j^*)$ for any $x_t^d \geq d_j^* - 1$.

Appendix C. Monte Carlo experiments for DGPs 2, 3, and 4

Table A.1 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 A.2 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 A.3. 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 A.1
Monte Carlo experiments with DGP 2 (Two types: RC = 4.5, 9).

Estimator of β	**Sample A** ($t = 1$ to 7)			**Sample B** ($t = 1$ to 14)			**Sample C** ($t = 8$ to 21)		
	Estimate[(1)]			Estimate[(1)]			Estimate[(1)]		
	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 null hypothesis	Frequency of Ho rejection with significance level			Frequency of Ho rejection with significance level			Frequency of Ho rejection with significance level		
	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 1000 replications.

Table A.2
Monte Carlo experiments with DGP 3 (Two types: RC = 8, 9).

Estimator of β	**Sample A** ($t = 1$ to 7)			**Sample B** ($t = 1$ to 14)			**Sample C** ($t = 8$ to 21)		
	Estimate[(1)]			Estimate[(1)]			Estimate[(1)]		
	Mean	Median	St. dev.	Mean	Median	St. dev.	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			Frequency of Ho rejection with significance level			Frequency of Ho rejection with significance level		
	1%	5%	10%	1%	5%	10%	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 1000 replications.

Table A.3
Monte Carlo experiments with DGP 4 (No UH, RC = 8).

Estimator of β	**Sample A** ($t = 1$ to 7)			**Sample B** ($t = 1$ to 14)			**Sample C** ($t = 8$ to 21)		
	Estimate[(1)]			Estimate[(1)]			Estimate[(1)]		
	Mean	Median	St. dev.	Mean	Median	St. dev.	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			Frequency of Ho rejection with significance level			Frequency of Ho rejection with significance level		
	1%	5%	10%	1%	5%	10%	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 1000 replications.

References

Aguirregabiria, V., 1999. The dynamics of markups and inventories in retailing firms. Rev. Econom. Stud. 66, 275–308.
Aguirregabiria, V., Mira, P., 2007. Sequential estimation of dynamic discrete games. Econometrica 75, 1–53.
Andersen, E., 1970. Asymptotic properties of conditional maximum likelihood estimators. J. R. Stat. Soc. Ser. B Stat. Methodol. 32, 283–301.
Arcidiacono, P., Miller, R., 2011. Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity. Econometrica 79 (6), 1823–1867.
Arellano, M., Bonhomme, S., 2012. Nonlinear panel data analysis. Annu. Rev. Econ. 3, 395–424.
Arellano, M., Bonhomme, S., 2017. Nonlinear panel data methods for dynamic heterogeneous agent models. Annu. Rev. Econ. 9, 471–496.
Arellano, M., Honoré, B., 2001. Panel data models: Some recent developments. In: Heckman, J.J., Leamer, E. (Eds.), Handbook of Econometrics, vol. 5. North-Holland, Amsterdam, pp. 3229–3296.
Blackwell, D., 1965. Discounted dynamic programming. Ann. Math. Stat. 36, 226–235.
Bonhomme, S., 2012. Functional differencing. Econometrica 80, 1337–1385.
Browning, M., Carro, J., 2010. Heterogeneity in dynamic discrete choice models. Econom. J. 13, 1–39.

Browning, M., Carro, J., 2014. Dynamic binary outcome models with maximal heterogeneity. J. Econometrics 178, 805–823.
Caliendo, L., Dvorkin, M., Parro, F., 2019. Trade and labor market dynamics: General equilibrium analysis of the China trade shock. Econometrica 87, 741–835.
Chamberlain, G., 1980. Analysis of covariance with qualitative data. Rev. Econom. Stud. 47, 225–238.
Chamberlain, G., 1985. Heterogeneity, omitted variable bias, and duration dependence. In: Heckman, J.J., Singer, B. (Eds.), Longitudinal Analysis of Labor Market Data. Cambridge University Press, Cambridge.
Chamberlain, G., 2010. Binary response models for panel data: Identification and information. Econometrica 78, 159–168.
Chernozhukov, V., Fernandez-Val, I., Hahn, J., Newey, W., 2013. Average and quantile effects in nonseparable panel models. Econometrica 81, 535–580.
Cox, D., 1958. The regression analysis of binary sequences. J. R. Stat. Soc. Ser. B Stat. Methodol. 20, 215–242.
D'Addio, A., Honoré, B., 2010. Duration dependence and timevarying variables in discrete time duration models. Braz. Rev. Econom. 30, 487–527.
Das, M., 1992. A micro-econometric model of capital utilization and retirement: The case of the cement industry. Rev. Econom. Stud. 59, 277–297.
Dunne, T., Klimek, S., Roberts, M., Xu, D., 2013. Entry, exit and the determinants of market structure. Rand J. Econ. 44, 462–487.
Erdem, T., Imai, S., Keane, M., 2003. Brand and quantity choice dynamics under price uncertainty. Quant. Mark. Econ. 1, 5–64.
Erdem, T., Keane, M., Sun, B., 2008. A dynamic model of brand choice when price and advertising signal product quality. Mark. Sci. 27, 1111–1125.
Frederiksen, A., Honoré, B., Hu, L., 2007. Discrete time duration models with group-level heterogeneity. J. Econometrics 141, 1014–1043.
Hahn, J., Kuersteiner, G., 2011. Bias reduction for dynamic nonlinear panel models with fixed effects. Econometric Theory 27, 1152–1191.
Hahn, J., Newey, W., 2004. Jacknife and analytical bias reduction for nonlinear panel data models. Econometrica 72, 1295–1319.
Heckman, J., 1981. The incidental parameters problem and the problem of initial conditions in estimating a discrete time - discrete data stochastic process. In: Manski, C., McFadden, D. (Eds.), Structural Analysis of Discrete Data with Econometric Applications. MIT Press, Cambridge.
Hendel, I., Nevo, A., 2006. Measuring the implications of sales and consumer inventory behavior. Econometrica 74, 1637–1674.
Honoré, B., Kyriazidou, E., 2000. Panel data discrete choice models with lagged dependent variables. Econometrica 68, 839–874.
Honoré, B., Tamer, E., 2006. Bounds on parameters in panel dynamic discrete choice models. Econometrica 74, 611–629.
Kalouptsidi, M., 2014. Time to build and fluctuations in bulk shipping. Amer. Econ. Rev. 104, 564–608.
Kano, K., 2013. Menu costs and dynamic duopoly. Int. J. Ind. Organ. 31, 102–118.
Kasahara, H., 2009. Temporary increases in tariffs and investment: The Chilean case. J. Bus. Econom. Statist. 27, 113–127.
Kasahara, H., Shimotsu, K., 2009. Nonparametric identification of finite mixture models of dynamic discrete choices. Econometrica 77, 135–175.
Keane, M., Wolpin, K., 1997. The career decisions of young men. J. Political Econ. 105, 473–522.
Kennet, M., 1993. Did deregulation affect aircraft engine maintenance? An empirical policy analysis. Rand J. Econ. 24, 542–558.
Lancaster, T., 2000. The incidental parameter problem since 1948. J. Econometrics 95, 391–413.
Magnac, T., 2004. Panel binary variables and sufficiency: Generalizing conditional logit. Econometrica 72, 1859–1876.
Manski, C., 1987. Semiparametric analysis of random effects linear models from binary panel data. Econometrica 55, 357–362.
Miller, R., 1984. Job matching and occupational choice. J. Political Econ. 92, 1086–1120.
Newey, W., McFadden, D., 1994. Large sample estimation and hypothesis testing. In: Engle, R.E., McFadden (Eds.), Handbook of Econometrics, vol. 4. North-Holland, Amsterdam, pp. 2111–2245.
Neyman, J., Scott, E., 1948. Consistent estimates based on partially consistent observations. Econometrica 16, 1–32.
Pötscher, B., 1991. Effects of model selection on inference. Econometric Theory 7, 163–185.
Rasch, G., 1961. On general laws and the meaning of measurement in psychology. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, vol. 4. University of California Press, Berkeley, pp. 321–333.
Roberts, M., Tybout, J., 1997. The decision to export in Colombia: An empirical model of entry with sunk costs. Amer. Econ. Rev. 87, 545–564.
Rust, J., 1987. Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. Econometrica 55, 999–1033.
Rust, J., 1994. Structural estimation of Markov decision processes. In: Engle, R.E., McFadden (Eds.), Handbook of Econometrics, vol. 4. North-Holland, Amsterdam.
Ryan, S., 2013. The costs of environmental regulation in a concentrated industry. Econometrica 80, 1019–1061.
Slade, M., 1998. Optimal pricing with costly adjustment: Evidence from retail grocery stores. Rev. Econom. Stud. 65, 87–108.
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, 1763–1803.
Willis, J., 2006. Magazine prices revisited. J. Appl. Econometrics 21, 337–344.