Hidden Rust Models

Benjamin Connault[*]

May 2016

Abstract

Hidden Rust models answer the need for convenient models of unobserved dynamics in a dynamic discrete choice context. They can be thought of as classical Rust models with an additional unobserved state variable. As opposed to their classical counterparts, hidden Rust models are not Markovian. I study their econometrics in terms of: identification, time-series asymptotics, practical estimation. I illustrate their use in a model of dynamic financial incentives inspired by Duflo, Hanna and Ryan (2012). Several lessons carry over to other structural models with unobserved dynamics.

Keywords: dynamic discrete choice, unobserved dynamics.

[*] University of Pennsylvania, Department of Economics, connault@econ.upenn.edu
This paper grew out of my thesis and I am very grateful to Bo Honoré, Ulrich Müller, Andriy Norets and Chris Sims for continuous guidance and support. I also thank Sylvain Chassang, Kirill Evdokimov, Myrto Kalouptsidi, Michal Kolesár, Eduardo Morales, Mark Watson for helpful discussion and comments.

1 Introduction

1.1 Overview

The single-agent theory of dynamic discrete choice has been consolidating around a common framework often associated with John Rust's seminal paper (Rust, 1987) and based on earlier work by Wolpin (1984), Miller (1984), Pakes (1986) and others. Classical Rust models have several advantages. They are fully structural: they are built on economically meaningful parameters and allow for simulation and counterfactual experiments. They are flexible, with a versatile state variable formalism and the possibility of both panel data and times series, as well as both finite and infinite horizons. They are also empirically tractable, and OLS-like estimation is often possible. For these reasons, Rust models have been applied in many different contexts: three examples in three different fields are Keane et al. (2011) in labor economics, Duflo et al. (2012) in development economics and Diermeier et al. (2005) in political economy. There are many more.

Unfortunately, unobserved dynamic features such as unobserved persistence, regime switching, structural breaks and heterogeneity are out of scope for classical Rust models. The challenge is to find a framework that would allow for unobserved dynamics while retaining the advantages of classical Rust models.

In this paper, I describe a family of models, which I call *hidden Rust models*, and I show that they provide such a framework.

Section 2 describes hidden Rust models. Hidden Rust models can be thought of as partially observed Rust models: agents make decisions exactly as in classical Rust models but the econometrician observes only part of the state variable. More specifically, agents make decisions a_t according to a stationary infinite-horizon dynamic discrete choice model with discrete state $k_t = (x_t, s_t)$. The econometrician observes a_t and s_t but not x_t. The unobserved state x_t carries the unobserved dynamics and does not necessarily have a structural interpretation. The infinite-horizon assumption is without loss of generality. Since the decision-making process is the same as in classical Rust models, the familiar conditional choice probabilities are present. However, they are not directly observed "in data": there is an additional stage between the conditional choice probabilities and the distribution of the observables, corresponding to the marginalization of the unobserved state variable x_t.

The observed data (s_t, a_t) is not Markovian of any order in a hidden Rust model. Because of this, the econometric theory of classical Rust models does not carry over to hidden Rust models. Issues of identification, asymptotics and practical estimation must be studied anew. This is what I do in the next three sections of the paper.

In section 3, I examine the question of identification in *algebraic models*, defined as models where the question of identification can be cast in terms of multivariate systems of polynomials. These include hidden Rust models parametrized in transition probabilities as special cases. My approach consists in studying the corresponding zero-sets from an algebro-geometric point of view. I obtain two main identification results. The *generic identification structure result* (Theorem 1) says that algebraic models have the same identification structure almost everywhere on the parameter space. As a consequence, if identification can be checked at a randomly drawn parameter value in a pre-data exercise, then the model is assured to be generically identified. The *stable identification structure result* (Theorem 2) says that the identification structure of a dynamic algebraic model stabilizes after some finite time horizon. As a consequence, if a dynamic algebraic model is identified from its infinite-horizon joint distribution, then it is identified from a finite-horizon set of marginals (Corollary 3).

In section 4, I study the asymptotics of hidden Rust models. The panel-data asymptotics with many independent and identically distributed individuals and a fixed time horizon, usually considered in the literature, are standard random sampling asymptotics. I focus on the non-Markovian time-series asymptotics, with one individual and many successive observations. I select a set of relatively weak assumptions, allowing in particular for almost arbitrary sparsity structure in the conditional transition matrices for the observed state, a common feature in economic applications. Under these assumptions I show that hidden Rust models are regular, in the local asymptotic normality sense (Theorem 3). $\sqrt{T}$-consistency and asymptotic normality of the maximum likelihood estimator follow (Theorem 4) and I also prove a Bernstein–von Mises theorem (Theorem 5). The proof of the Bernstein–von Mises theorem uses the stable identification structure of the model.

In section 5, I look at the practical issue of estimating hidden Rust models and I show that they remain very tractable. I explain how the likelihood can be efficiently evaluated in two stages. In the first stage, a dynamic program exactly similar to the dynamic program of a classical Rust model needs to be solved. I recommend using an off-the-shelf numerical solver on a system of nonlinear equations directly expressed in terms of the structural parameters and the conditional choice probabilities. This is often faster than the usual fixed-point algorithm on the conditional discounted value functions, and the technique automatically takes advantage of the sparsity structure commonly found in these models. In the second stage, a closed-form *discrete filter* can be used to integrate the contribution of the unobserved state variables out of the likelihood. This is similar in spirit to the Kalman filter for linear Gaussian state-space models. Repeated evaluations of the likelihood can be called within an optimization outer loop or a posterior simulation outer loop according to one's inclinations.

In section 6, the tools developed in this paper are applied to a structural model of financial incentives inspired by Duflo et al. (2012). Teacher attendance data was collected by the authors in a region of rural India where teacher absenteeism is a significant issue. I use the attendance data on a group of teachers treated with a progressive pay scheme. Following the authors' observation that the data has important unobserved persistence features, I set up a hidden Rust model whose unobserved state captures dynamic heterogeneity in the teachers' unobserved willingness to work. Fast[1] estimation of the unobserved state transition matrix provides interesting insights into the dynamic heterogeneity structure.

Going beyond hidden Rust models, I argue in section 7 that many of this paper's ideas can be applied to more general "hidden structural models." My identification results and point of view are readily applicable. The time-series asymptotic results will apply under some constraints on the dynamics. As for practical estimation, marginalization of the unobserved state variable via the discrete filter applies to any dynamic discrete model. If there is no tractable equivalent to "solving the dynamic program" in the hidden structural model of interest, constrained optimization of the likelihood can often be used.

[1] The MLE is computed in around three seconds on an average 2013 desktop computer.

1.2 Literature review

Some econometric aspects of dynamic discrete choice models with unobserved state have been considered previously in the literature. In terms of identification and asymptotics, there are also relevant results in the statistical literature on hidden Markov models.

Identification results in the statistics literature are designed for reduced-form statistical models and are fragile against deeper structural modelling typical in economics model, such as additional conditional independence assumptions and structural zero transition probabilities. For instance, in Rust's (1987) famous bus example, 97% of the coefficients of the conditional state transition matrix are structural zeros because a bus' mileage can increase by zero, one or two brackets on each trip, out of 90 possible mileage brackets observed in sample. If a model is a submodel of a generically identified supermodel, the generic identification structure may be lost because the submodel may live in the exceptional identification region of the supermodel. This happens for hidden Rust models with respect to fully-supported hidden Markov models. Fully-supported hidden Markov models themselves are known to be generically identified (Baum and Petrie (1966)). Call x_t the unobserved state and y_t the observed variable (observed state and decision) in a hidden Rust model. Hidden Rust models can be cast as degenerate hidden Markov models by lumping $\tilde{x}_t = (x_t, y_t)$ and $\tilde{y}_t = y_t$. However the deterministic emissions then violate the sufficient condition for identification of Baum and Petrie (1966), or those derived from tensor decomposition approaches in more recent papers such as Allman et al. (2009) or Bonhomme et al. (2016). An objective of this paper is to develop results robust against deeper structural modelling. The key advantage of Theorem 1 and Theorem 2 is that they are valid at the level of singularity structurally specified by the economist, no matter how pathological. This is needed in the context of structural economic modelling, by contrast with descriptive statistical modelling where full-support type assumptions are not an issue.

Hu and Shum (2012) provides a reduced-form approach which may be applicable to some hidden Rust models (see also Kasahara and Shimotsu (2009) for the particular case of static mixtures). Those papers give sufficient conditions for global identification at a specific parameter value λ that require checking the invertibility of a number of square matrices of marginal probabilities. When applicable, that type of results will work well with this paper's generic identification structure result (Theorem 1): if the relevant square matrices are found to be invertible at a randomly drawn parameter value $\lambda^\star$ in a pre-data exercise, then the model is assured to be globally identified at *almost every* parameter value λ. However, being reduced-form, the approach will not cover many models with additional structure. For instance, assumption 2 in Hu and Shum (2012) implies that a move from any observed state to any other observed state has positive probability.

Gilbert's (1959) results imply an explicit bound on the stabilization horizon of the identification structure of a stationary hidden Rust models. Theorem 2 applies to much more general models, including non-stationary hidden Rust models. The bound is $2(d_x d_y - d_y + 1)$.

On the asymptotic theory side, Baum and Petrie (1966) proved consistency and asymptotic normality of the maximum likelihood estimator for hidden Markov models under a uniform lower bound assumption on the transition matrix coefficients. Baum and Petrie (1966) introduced the "infinite-past" proof strategy, which is the strategy I use in this paper. More recent papers have focused on extending the results of Baum and Petrie (1966) to continuous observables. Many of those also use the "infinite-past" strategy; see, in particular, Bickel and Ritov (1996), Bickel et al. (1998), Douc et al. (2004) and Douc et al. (2011). Douc et al. (2004) in particular studies autoregressive hidden Markov dynamics with a continuous state and a uniform lower bound on the observed state's conditional transition density. Hidden Rust models also have autoregressive hidden Markov dynamics, although with discrete state but potentially very sparse conditional transition matrices.

On the estimation side, Arcidiacono and Miller (2011) considers models slightly less general but very related to hidden Rust models. Whereas I develop estimators in the Rust's (1987) nested fixed-point tradition, Arcidiacono and Miller (2011) takes a different approach, developing estimation methods where no dynamic-program solving is required. I come back to Arcidiacono and Miller's (2011) estimators in section 5. Norets (2009) focuses on the issue of computing Bayesian posteriors in a more general but less tractable model of unobserved persistence for dynamic discrete choice.

The recursive algorithm used in the discrete filter to marginalize out the unobserved state is well-known in various fields dealing with dynamic discrete models; see, for instance, Zucchini and MacDonald (2009).

2 Hidden Rust models

Section 2.1 describes hidden Rust models. Section 2.2 explains why the econometric theory of classical Rust models does not carry over to hidden Rust models.

2.1 Model description

The underlying model of dynamic decision making is identical in a hidden Rust model and in a classical Rust model. An economic agent makes repeated decisions under a changing economic environment. His choices partially influence the otherwise random evolution of the economic environment. He takes this impact into account in his rational decision-making. A hidden Rust model can be thought of as a partially observed dynamic discrete choice model: the econometrician observes the decision but only part of the state.

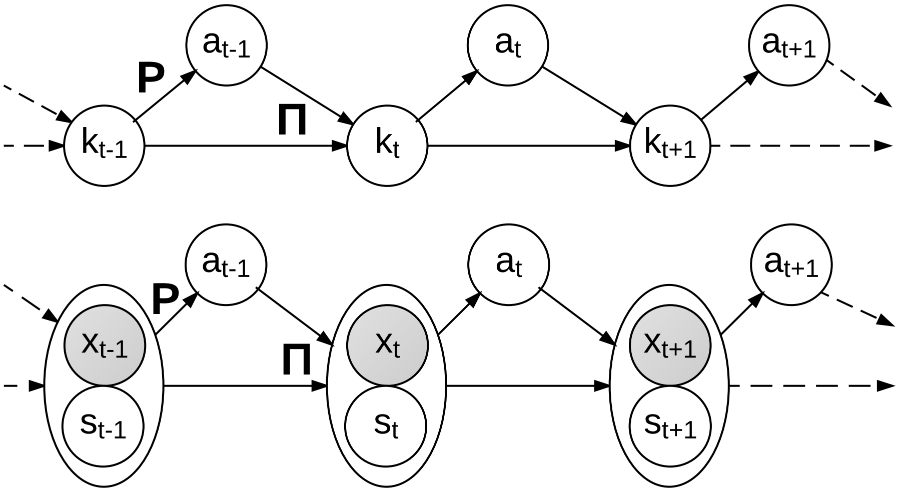


Figure 1: (top) Dynamics of the choice model. If the state k_t and the action a_t are fully observed by the econometrician, we get a classical Rust model. (bottom) A hidden Rust model. The econometrician does not observe the x_t part of the state.

In more detail, an agent makes dynamic decisions following a stationary infinite-horizon dynamic discrete choice model with discrete state. $a_t \in \{1, \ldots, d_a\}$ are the choices made by the agent and a state variable $k_t \in \{1, \ldots, d_k\}$ characterizes the economic environment. The choices are observed by the econometrician. The state $k_t = (x_t, s_t)$ has an observed component $s_t \in \{1, \ldots, d_s\}$ and an unobserved component $x_t \in \{1, \ldots, d_x\}$. The dynamics of the data are summarized in Figure 1. The distribution of the data is fully specified by some initial distribution μ on (x_1, s_1, a_1), the conditional state transition matrices Π_a, $\Pi_{a,kk'} = \mathbb{P}(k_{t+1} = k'|k_t = k, a_t = a)$, and the conditional choice probability matrix P, $P_{ka} = \mathbb{P}(a_t = a|k_t = k)$. The economics of the model typically live within the conditional choice probability matrix P: P must be compatible with some model of decision-making by the agent, indexed by a structural utility parameter θ_u. The specific model of choice $(\theta_u, \Pi) \to P$ used in Rust (1987) is popular and convenient, but for most of the paper I only need to assume some smoothness properties of the mapping, which I leave unspecified at this stage. $\Pi = (\Pi_a)_a$ is parametrized as $\Pi(\theta_\Pi)$. The initial distribution will not play a major role, and can be either known, taken to be a function of (Π, P) (such as a unique stationary distribution for the Markov chain $z_t = (x_t, s_t, a_t)$), or parametrized separately as $\mu(\theta_\mu)$. I collect all the deep structural parameters in $\theta = (\theta_u, \theta_\Pi, \theta_\mu) \in \mathbb{R}^{d_\theta}$. I also collect $M = (\Pi, P)$. M can be seen as the transition matrix for the Markov chain z_t. The resulting two-level structure of the model is useful to keep in mind throughout the paper:

$$\text{STRUCTURAL PARAMETERS } \theta \to \text{TRANSITION MATRICES } M$$
$$\to \text{DISTRIBUTION OF THE OBSERVABLES}$$

The stationary infinite-horizon set-up of this paper is without loss of generality, in the sense that a finite-horizon model can be cast as an infinite-horizon model by enlarging the state space and drawing a random new initial period after the last period. The number of unobserved states is assumed to be known. More often than not, I will view the unobserved state x_t as a technical device able to carry flexible unobserved dynamics patterns, with no necessary literal real-world counterpart.

2.2 The econometrics of hidden Rust models require new arguments

In a classical Rust model, the econometrician observes $z_t = (k_t, a_t) = (s_t, a_t)$ in full. The good econometric properties of these models are well-known; see, for instance, Aguirregabiria and Mira (2010). The fact that the observed data z_t is Markov plays a central role. In terms of identification, the model is identified as soon as the transition matrix M is identified, that is to say, as soon as the mapping $\theta \to M$ is injective. In terms of asymptotics, the time-series asymptotic properties are relatively straightforward thanks to the Markov structure. In terms of computing estimators, there are two popular methods. The maximum likelihood estimator computed with Rust's nested fixed-point algorithm relies on the fact that the probability of a sequence of observations is the product of the transition probabilities. Two-step estimators in the spirit of Hotz and Miller (1993) rely on the fact that transition probabilities can be consistently estimated by counting the transitions "in data."

In a hidden Rust model, the observed data is not Markovian of any order. Two different sets of transition matrices M could give rise to the same distribution of the observables. The asymptotic analysis is also harder: for instance, the log-likelihood cannot be written as an ergodic sum after successive conditioning, because all past variables remain in the conditioning set. In terms of computing estimators, path probabilities cannot be computed by just following the transition probabilities along an observed path, and we cannot form consistent frequency estimates of M because we do not observe transitions in data.

3 Identification

Section 3.1 defines *algebraic models*, of which hidden Rust models are special cases. Section 3.2 shows that algebraic models have a generic identification structure (Theorem 1). Section 3.3 shows that dynamic algebraic models have a stable identification structure (Theorem 2).

3.1 Algebraic models

A statistical model for an arbitrary observable Y is a mapping Φ from a parameter space Λ to the space of probability distributions for Y. I define an *algebraic model* as any statistical model for which the *identification equation* $\Phi(\lambda) = \Phi(\lambda^\star)$ can be written as a (not necessarily finite) system F of polynomials in λ and $\lambda^\star$:

$$\Phi(\lambda) = \Phi(\lambda^\star) \qquad \Longleftrightarrow \qquad F(\lambda, \lambda^\star) = 0$$

Algebraic models include many "discrete models," such as discrete mixtures of discrete random variables, but also many parametric families of continuous random variables.

Example: Consider an unobserved discrete random variable $X \in \{a, b\}$ and an observed discrete random variable $Y \in \{0, 1, 2, 3\}$. $X = b$ with probability p and $X = a$ with probability $1 - p$. If $X = a$, Y is binomial $B(3, p_a)$ — the sum of three biased coin flips — and if $X = b$, Y is binomial $B(3, p_b)$. In an identification context, we ask if the statistical parameter $\lambda = (p, p_a, p_b)$ is identified from the distribution of Y. With $\lambda^\star = (q, q_a, q_b)$, the identification system can be written:

$$F(\lambda, \lambda^\star) = 0$$

$$\Longleftrightarrow$$

$$\begin{cases} (1-p)(1-p_a)^3 + p(1-p_b)^3 = (1-q)(1-q_a)^3 + q(1-q_b)^3 \\ (1-p)3(1-p_a)^2 p_a + p3(1-p_b)^2 p_b = (1-q)3(1-q_a)^2 q_a + q3(1-q_b)^2 q_b \\ (1-p)3(1-p_a)p_a^2 + p3(1-p_b)p_b^2 = (1-q)3(1-q_a)q_a^2 + q3(1-q_b)q_b^2 \\ (1-p)p_a^3 + pp_b^3 = (1-q)q_a^3 + qq_b^3 \end{cases}$$

Thus the model is algebraic. We will come back to this example in the next section. □

In a hidden Rust model the mapping from structural parameters θ to conditional choice probabilities is usually not polynomial; however, the mapping from transition matrices M to the probability distribution of the observables is polynomial in the individual coefficients of M (or rational in the case of an initial stationary distribution). Indeed, the probability of observing a given path $y_{1:T} = (s_{1:T}, a_{1:T})$ is the sum of the probabilities of all possible paths $(x_{1:T}, y_{1:T})$, and the probability of a path $(x_{1:T}, y_{1:T})$ is the product of initial and transition probabilities along the path. Thus a hidden Rust model is an algebraic model when parametrized in the non-structurally zero transition probabilities, which we collect in the intermediate parameter λ. This means we can study the identification properties of a hidden Rust model at this intermediate λ level, using the tools developed in this section. Those properties are sometimes called the *non-parametric identification* properties of the model.

Non-parametric identification is not the same as identification at the structural level θ. There are at least three reasons why studying identification at the non-parametric level is interesting. First, there is hope to carry back generic identification at the non-parametric level as in Theorem 1 to generic identification at the structural level. Once a specific mapping from structural parameters to transition matrices is specified, it is a matter of showing that the image of this mapping intersects cleanly with any variety in the space of transition matrices. Second, the non-parametric level provides us with theoretical and computational tools to attack the identification issue. Although the approach in Kasahara and Shimotsu (2009) or Hu and Shum (2012) is not explictly algebro-geometric, their identification results are obtained at the non-parametric level. In fact the sufficient conditions in these two papers can be seen to be special instances of next section's generic identification structure result (Theorem 1), because they can be formulated as "if $F_e(\lambda) \neq 0$ then the model is identified," where F_e are determinants of square matrices of marginals, i.e. polynomials in transition matrix coefficients. Third, non-parametric identification is key from a 2-step estimation perspective, where the transition matrices are estimated in a first step and then projected on the structural parameter space. This approach is popular in the dynamic discrete choice literature (Hotz and Miller, 1993).

3.2 Algebraic models have a generic identification structure

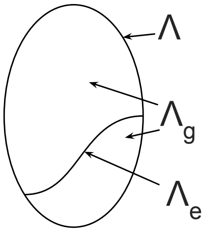
I say that two parameters $\lambda_1^\star$ and $\lambda_2^\star$ *have the same identification structure* when the sets of observationally equivalent parameters to $\lambda_1^\star$ and $\lambda_2^\star$, i.e. the solution sets of $F(\lambda_1^\star, \cdot)$ and $F(\lambda_2^\star, \cdot)$, have the same structure in a technical sense. The precise definition is given and illustrated in appendix section 9.1. It respects the intuitive notions of cardinality (for finite sets) and dimension (for continuous sets). Theorem 1 says that algebraic models have a constant identification structure outside of a "small" exceptional region:

Theorem 1: Generic identification structure
Let Φ be a algebraic model. There is a unique minimal exceptional region Λ_e *such that:*

(i) if $\lambda_1^\star \notin \Lambda_e$ and $\lambda_2^\star \notin \Lambda_e$, then $\lambda_1^\star$ and $\lambda_2^\star$ have the same identification structure.

(ii) Λ_e is the zero-set of a non-zero polynomial system.

Minimality is for the inclusion.



We call the complement of the exceptional region the *generic region* $\Lambda_g = \Lambda \setminus \Lambda_e$.

The case of particular interest is when the model is generically globally identified, i.e. when $\lambda^\star$ is generically the only solution to $F(\lambda^\star, \cdot)$. To handle cases of label-switching, suppose we know there are n_{ls} observationally equivalent parameter values, meaning $\Phi(\lambda) = \Phi(\lambda^\star)$ has at least n_{ls} known solutions in λ for each $\lambda^\star$. $n_{ls} = 1$ is the usual notion of global identification.

Corollary 1: Generically identified models
If the model is (globally) identified at any $\lambda^\star$ in the generic region, meaning $\Phi(\lambda) = \Phi(\lambda^\star)$ has exactly n_{ls} complex solutions, then it is (globally) identified everywhere in the generic region. In this case the model is said to be generically identified.

Example (continued): If $g(\lambda^\star) := q(1-q)(q_b^\star - q_a^\star) \neq 0$, then $\Phi(\lambda) = \Phi(\lambda^\star)$ has exactly two solutions: $(p, p_a, p_b) = (q, q_a, q_b)$ and $(p, p_a, p_b) = (1-q, q_b, q_a)$. The model is generically identified. This example is detailed in supplementary appendix section 1.3 with an explicit connexion with the theoretical and computational algebro-geometric viewpoint. □

The genericness statement of Theorem 1 is a very strong one: we know that the exceptional region is the zero-set of a polynomial system. This is stronger than the two most common technical definitions of genericness:

Corollary 2:
Suppose $\Lambda = [0,1]^{d_\lambda}$. Then:

(i) Λ_g *is open dense in* Λ.

(ii) Λ_e *has Lebesgue measure zero in* Λ.

All results in this section are proved in supplementary appendix section 1. The fundamental reason why algebraic models have a generic identification structure (i.e., why Theorem 1 holds) is that zero-sets of systems of polynomial equations are small in ambient space, which is not necessarily the case for systems of other types of equations (even very smooth).

3.3 Dynamic algebraic models have a stable identification structure

In this section, again motivated by hidden Rust models, I add a dynamic dimension to the set-up of the previous section. I define a *dynamic algebraic model* as any mapping Φ from a parameter space Λ to the distribution of a sequence of observables $Y_{1:\infty}$ such that, for any T, the marginal model Φ_T for $Y_{1:T}$ is algebraic i.e., there is a (not necessarily finite) system F_T of polynomials such that:

$$\Phi_T(\lambda) = \Phi_T\left(\lambda^\star\right) \qquad \iff \qquad F_T\left(\lambda, \lambda^\star\right) = 0$$

By definition of the product measure, the identification equation for Φ can be written as:

$$\Phi(\lambda) = \Phi\left(\lambda^\star\right) \qquad \iff \qquad \forall T, \quad \Phi_T(\lambda) = \Phi_T\left(\lambda^\star\right)$$

In particular Φ itself is an algebraic model in the sense of section 3.2, with associated polynomial system $F = \bigcup_T F_T$. By Φ_∞ I will mean Φ. Theorem 2 says that the identification structure of a dynamic algebraic model becomes constant after some finite horizon T_0, uniformly in the underlying $\theta^\star$:

Theorem 2: Stable identification structure
Let Φ be any dynamic algebraic model. There is a smallest $T_0 < \infty$ such that for any $\lambda^\star$, the set of λ-solutions to $\Phi_T(\lambda) = \Phi_T\left(\lambda^\star\right)$ is constant for $T_0 \leq T \leq \infty$.

The proof of Theorem 2 is concise and illustrates well the usefulness of the algebro-geometric point of view. For any set $G(X)$ of polynomials in X, write $V_X(G)$ the zero-set of G.

Proof. By Noetherianity of the ideal $\langle F \rangle$ generated by F, there a finite number of elements of F which generates $\langle F \rangle$. Fix $\tilde{F}$ such a set. There is necessarily T_0 such that $\tilde{F} \subset F_{T_0}$. Then for any $T_0 \leq T \leq \infty$, $\langle F_{T_0} \rangle \subset \langle F_T \rangle \subset \langle F \rangle = \langle F_{T_0} \rangle$. Then also $V_{\lambda,\lambda^\star}(F_T) = V_{\lambda,\lambda^\star}(F_{T_0})$ and in particular for any $\lambda^\star$, $V_\lambda\left(F_T\left(\cdot, \lambda^\star\right)\right) = V_\lambda\left(F_{T_0}\left(\cdot, \lambda^\star\right)\right)$. □

As an easy corollary of particular interest:

Corollary 3: Infinite-horizon identification implies finite-horizon identification.
If Φ is globally identified for every $\lambda^\star$ in a region $\bar{\Lambda} \subset \Lambda$, then there is $T_0 < \infty$ such that for any $T \geq T_0$, Φ_T is globally identified for every $\lambda^\star$ in $\bar{\Lambda}$.

Theorem 2 and Corollary 3 capture a phenomenon specific to algebraic models. Appendix section 9.2 gives an example of a smooth dynamic model that is identified from an infinite number of marginals but not from any finite number of them. This shows that Corollary 3 cannot be generalized beyond algebraic models.

Theorem 2 is not constructive in T_0. I illustrate how it can be used in the context of a time-series asymptotic study in the proof of the Berstein–von Mises theorem for hidden Rust model, Theorem 5 in section 4.

4 Asymptotics

This section studies the time-series asymptotics of hidden Rust models for one individual. Section 4.1 states the assumptions used in the asymptotic analysis. Section 4.2 states local asymptotic normality of the model (Theorem 3), consistency and asymptotic normality of the maximum likelihood estimator (Theorem 4) and a Bernstein–von Mises theorem for Bayesian posteriors (Theorem 5). All proofs are in appendix section 2.

4.1 Assumptions

For the purpose of the time-series asymptotic analysis, I assume that the dynamics of the model are such that the unobserved state x_t has exogenous Markov dynamics in the following sense: $\mathbb{P}(x_{t+1}, s_{t+1}|x_t, s_t, a_t) = \mathbb{P}(x_{t+1}|x_t)\mathbb{P}(s_{t+1}|s_t, a_t)$. In matrix notation, this means that Π_a factorizes as $\Pi_a = \tilde{\Pi}_a \otimes Q$, where $Q_{xx'} = \mathbb{P}(x_{t+1} = x'|x_t = x)$ and $\tilde{\Pi}_{a,ss'} = \mathbb{P}(s_{t+1} = s'|s_t = s, a_t = a)$, and a reverse lexicographical order is used on $k = (x, s)$. Thus $z_t = (x_t, s_t, a_t)$ is generated according to an arbitrary initial distribution $\mu^\star$ along with transition matrices $P(\theta^\star)$, $Q(\theta^\star)$ and $\tilde{\Pi}(\theta^\star)$. Θ is a compact subspace of $\mathbb{R}^{d_\theta}$ and $\theta^\star$ is in the interior of Θ. The econometrician observes $y_t = (s_t, a_t)$ for $1 \leq t \leq T$. There is no hope of estimating the initial distribution from just one time series, and the econometrician is not likely to know $\mu^\star$: I allow misspecification of the initial distribution. The (conditional) log-likelihood is computed under the assumption that the data are generated with some arbitrary initial distribution μ:

$$L_T(\theta) = \frac{1}{T} \log P_{\theta,\mu}(Y_{2:T}|Y_1)$$

The observed data $Y_{1:T}$ have non-Markovian dynamics. It is not clear a priori that the model and estimators have good time-series asymptotic properties. For instance, the log-likelihood cannot be written as an ergodic sum. Successive conditioning gives:

$$L_T(\theta) = \frac{1}{T} \sum_{t=1}^{T-1} \log P_{\theta,\mu}(Y_{t+1}|Y_{1:t})$$

The theorems of section 4.2 show that the model does have good time-series asymptotic properties. I make the following assumptions:

Assumption (A1): Compactness of Θ and smoothness of the model
Θ is a compact subspace of $\mathbb{R}^{d_\theta}$ and the functions $\theta \to P(\theta)$, $\theta \to Q(\theta)$ and $\theta \to \tilde{\Pi}(\theta)$ are three times continuously differentiable.

The following assumption guarantees that z will *merge* to a unique stationary distribution regardless of the initial distribution:

Assumption (A2): z is merging
For any $\theta \in \Theta$, z is recurrent and aperiodic.

(A2) implies that z has a unique marginal stationary distribution $\mu^\diamond(\theta)$ under θ. $\mu^\diamond(\theta)$ along with $P(\theta)$, $Q(\theta)$ and $\tilde{\Pi}(\theta)$ induces a stationary distribution on the Markov chain $Z_t = (X_t, Y_t)$, which can be extented to $t \in \mathbb{Z}$ in the standard fashion. The following identification assumption bears on the stationary distribution, although the true distribution of Y_t is not necessarily the stationary one.

Assumption (A3): Identification
If $\theta, \theta' \in \Theta$ induce the same joint stationary distribution on $Y_{-\infty:\infty}$, then $\theta = \theta'$.

Assumption (A4): Mixing properties of the unobserved state x
The transition matrix Q for the unobserved state has full support.

The set of assumptions I use allows for almost arbitrary structural zero patterns for Π, the conditional transition matrices for the observed state of the model. This is a key feature of many economic models. The only restriction is the merging assumption (A2).

The following standard prior mass assumption will also be used for the Bernstein–von Mises theorem:

Assumption (A5): Prior mass
The prior is absolutely continuous with respect to the Lebesgue measure in a neighborhood of $\theta^\star$, with a continuous positive density at $\theta^\star$.

4.2 Asymptotic theorems

Theorem 3 states that hidden Rust models are uniformly locally asymptotically normal, meaning the log-likelihood has a certain asymptotic stochastic quadratic approximation around the true parameter value. Local asymptotic normality is a regularity property of hidden Rust models, which holds independently of the estimation strategy. In particular, it is stated with respect to the true initial distribution $\mu^\star$ and not the econometrician's misspecified μ. Let $\ell_T(\theta) = \log P_{\theta,\mu^\star}(Y_{2:T}|Y_1)$ be the non-scaled, well-specified log-likelihood, $\sigma_T = \nabla_{\theta^\star}\ell_T(\theta)$ the non-scaled score and $\Delta_T = \sigma_T/\sqrt{T}$. Under *(A1)* to *(A4)*:

Theorem 3: Uniform local asymptotic normality
$\Delta_T \overset{\theta^\star}{\rightsquigarrow} \mathcal{N}(0, I)$ where I is invertible, and for any sequence of random variables $h_T \xrightarrow{P_{\theta^\star}} h$:

$$\ell_T\left(\theta^\star + h_T/\sqrt{T}\right) = \ell_T(\theta^\star) + h'\Delta_T - \frac{1}{2}h'Ih + o_{\theta^\star}(1)$$

Theorem 4 verifies that the maximum likelihood estimator has a well-behaved asymptotic behavior. Let $\hat{\theta}_T = \underset{\theta \in \Theta}{\operatorname{argmax}}\, L_T(\theta)$ be the maximum likelihood estimator. Under $(A1)$ to $(A4)$:

Theorem 4: Consistency and asymptotic normality of the maximum likelihood
$\hat{\theta}_T$ is strongly consistent:

$$\hat{\theta}_T \xrightarrow{\theta^\star \ as} \theta^\star$$

Furthermore, $\hat{\theta}_T$ is asymptotically normal:

$$\sqrt{T}\left(\hat{\theta}_T - \theta^\star\right) \overset{\theta^\star}{\rightsquigarrow} \mathcal{N}(0, I^{-1})$$

Theorem 5 shows that Bayesian posteriors are also well-behaved asymptotically, from a frequentist perspective. It is a standard result for smooth, independent and identically distributed models. It can be used as a basis for obtaining efficient frequentist estimators and correct condidence intervals from a Bayesian posterior. It is also a formal guarantee that the influence of the prior fades as the sample size increases. Let q_T be the Bayesian posterior distribution. Under $(A1)$ to $(A5)$:

Theorem 5: Bernstein-von Mises theorem

$$d_{TV}\left(q_T, \mathcal{N}\left(\hat{\theta}_T, I^{-1}/T\right)\right) \xrightarrow{P_{\theta^\star}} 0$$

All proofs are given in supplementary appendix section 2.

A particular step of the proof provides for a good example of an application of the stable identification structure theorem (Theorem 2). As part of the proof of the Bernstein–von Mises theorem, we want to show that there exists some uniformly consistent estimator $\hat{\theta}_T$ for the parameter θ, in the sense that:

$$\forall \epsilon, \qquad \sup_{\theta} P_\theta\left(d\left(\hat{\theta}_T, \theta\right) \geq \epsilon\right) \to 0$$

where d is any distance satisfying a mild technical condition detailed in supplementary appendix section 2. Appealing to the stable identification structure theorem (Theorem 2) together with identification of the model (assumption (A3)), there exists $T_0 < \infty$ such that the marginal distribution of T_0 consecutive y's identifies θ. Let

π be the corresponding marginal and $\hat{\theta}_T = \hat{\pi}_T$ be the empirical distribution estimator for π. In supplementary appendix section 2, I show that $\hat{\pi}_T$ is a uniformly consistent estimator with $d\left(\hat{\theta}_T, \theta\right) = d_{TV}\left(\hat{\pi}_T, \pi\right)$ as needed.

5 Estimation

This section shows how to efficiently compute the maximum likelihood estimator and Bayesian posteriors in hidden Rust models. The high tractability of hidden Rust models is a major advantage compared to other models of unobserved persistence for dynamic discrete choice.

The general idea is to evaluate the likelihood in two stages, corresponding to the two levels of a hidden Rust model: $\theta \to M \to$ *distribution of the observables*. For this section I assume a specific form for the mapping $\theta \to P$, namely that choices are made based on expected discounted utility with additively separable Gumbel shocks, as in Rust (1987).

I recall some standard facts from dynamic discrete choice theory in section 5.1. I explain how to evaluate the likelihood efficiently in section 5.2. Maximum likelihood and Bayesian estimation can then be implemented in a straightforward fashion. Two-step estimation and maximum likelihood estimation by constrained optimization are also possible, although I do not recommend their use in a typical hidden Rust model (section 5.3).

5.1 Dynamic discrete choice models with the Rust assumption

For estimation purposes, we assume the agent makes decisions as in Rust (1987). Classical results from dynamic dicrete choice theory apply. Tools developed with classical dynamic discrete choice models in mind may be used to relax the additively separable Gumbel assumption. See, for instance, Chiong et al. (2014) for the Gumbel assumption and Kristensen et al. (2014) for the additive separability assumption. In this section, I recall some relevant facts from classical dynamic discrete choice theory. All results are well-known; see, for instance, Aguirregabiria and Mira (2010).

We assume the agent derives an instantaneous payoff equal to the sum of a deterministic flow utility component u_{k_t,a_t} and a random utility shock ϵ_{t,a_t}. The agent forms discounted utilities v based on current flow utilities and expected future realizations of flow utilities and shocks. The future is discounted with a known factor β:

$$v_{k_t,a_t} = u_{k_t,a_t} + \mathbb{E}\left[\sum_{s=t+1}^{\infty} \beta^{s-t}(u_{k_s,a_s} + \epsilon_{s,a_s})\right]$$

At time t, the agent chooses an action a_t by maximizing his discounted payoff:

$$a_t = \underset{a}{\operatorname{argmax}}\{v_{k_t,a} + \epsilon_{t,a}\} \tag{1}$$

The discounted utility matrix v is a stationary, non-random quantity that can be expressed as the solution of a fixed-point equation. Indeed, it is the unique solution of the standard dynamic program:

$$v_{k,a} = u_{k,a} + \beta\mathbb{E}\left[\mathbb{E}\left[\max_{a'}\{v_{k',a'} + \epsilon_{a'}\}\middle|k'\right]\middle|k\right] \tag{2}$$

Under the assumption that the shocks are independent and identically distributed across time and choices, with a centered extreme value Gumbel distribution, the expected utility before the shocks are realized $V_k = \mathbb{E}\left[\max_a\{v_{k,a} + \epsilon_a\}\right]$ (sometimes called the ex-ante or interim value function) has a closed-form expression:

$$V_k = \mathbb{E}\left[\max_a\{v_{k,a} + \epsilon_a\}\right] = \log\left(\sum_a e^{v_{k,a}}\right)$$

The dynamic program (2) can be written in vector/matrix notation. Let v_a be the a^{th} column of v, corresponding to the choice a. We use the convention that functions are applied coefficient-wise where it makes sense, so that, for instance, e^{v_a} and $\log\left(\sum_a e^{v_a}\right)$ are $d_k \times 1$ column vectors. The dynamic program (2) is equivalent to:

$$v_a = u_a + \beta\Pi_a \log\left(\sum_{a'} e^{v_{a'}}\right) \tag{3}$$

The decision-making rule (1) implies that choices a_t are made with constant conditional choice probabilities $\mathbb{P}(a_t = a|k_t = k) = \mathbb{P}(a|k)$. The matrix P, $P_{ka} = \mathbb{P}(a|k)$, is the conditional choice probability matrix. The Gumbel distribution assumption on shocks implies the usual logit expression for the conditional choice probabilities:

$$P_a = \frac{e^{v_a}}{\sum_a e^{v_a}}$$

From an econometric point of view and given a parameterization of flow utilities, the above steps induce mappings $\theta_u \to u \to v \to P$. The image of the resulting mapping $\theta_u \to P$ contains the transition matrices compatible with rational decision-making as specified by the model. Computing $\theta_u \to P$ is the first stage of evaluating the likelihood in a hidden Rust model. This is usually done by solving solving (3) via fixed-point iterations. I will recommend an laternative approach in the next subsection.

5.2 Evaluating the likelihood

Evaluating the likelihood involves two stages: solving the dynamic program and marginalizing out the unobserved state.

To solve the dynamic program, I recommend solving for V in the following system of nonlinear equations:

$$F(V) = \sum_a e^{u_a + (\beta\Pi_a - I)V} = 1 \tag{4}$$

P is then given by $P_a = e^{u_a + (\beta\Pi_a - I)V}$. The Jacobian can easily be evaluated:

$$\dot{F}(V) = \sum_a \text{diagm}\left(e^{u_a + (\beta\Pi_a - I)V}\right)(\beta\Pi_a - I)$$

Proof. We need to verify that (4) defines V uniquely.
(i) V is a solution. $V = \log\left(\sum_a e^{v_a}\right)$ implies $e^V = \sum_a e^{v_a} = \sum_a e^{u_a + \beta\Pi_a V}$ or equivalently $\sum_a e^{u_a + (\beta\Pi_a - I)V} = 1$.
(ii) V is the unique solution. Let $\tilde{V}$ be any solution of (4). Let us show $\tilde{V} = V$. Define $\tilde{v}_a = u_a + \beta\Pi_a\tilde{V}$. $\sum_a e^{u_a + (\beta\Pi_a - I)\tilde{V}} = 1$ implies $\tilde{V} = \log\left(\sum_a e^{\tilde{v}_a}\right)$ implies $\tilde{v}_a = u_a + \beta\Pi_a \log\left(\sum_\alpha e^{\tilde{v}_\alpha}\right)$. Then $\tilde{v}_a = v_a$ because (3) has a unique solution and finally $\tilde{V} = \log\left(\sum_a e^{v_a}\right) = V$. □

There are two advantages compared to the usual fixed-point iteration method used on (3). First, convergence may require many less steps, especially for values of β close to

1. The applicability of the nonlinear system point of view for dynamic programming is well-known (Rust, 1996), but, as far as I know, rarely used the dynamic discrete choice context. Second and more importantly, the approach will automatically take advantage of the sparsity structure of M. In applications, M is often extremely sparse. For example, it is 97% sparse in Rust (1987) and 98% sparse in Duflo et al. (2012). Not only will evaluating F and $\dot{F}$ involve one sparse matrix multiplication, but the Jacobian $\dot{F}$ itself will inherit $(\beta M_a - I)$'s sparsity structure. Numerical solvers take advantage of sparse Jacobians. In the empirical model of section 6 (with $d_k = 756$, $d_a = 2$), it takes around 10 milliseconds to solve $F(V) = 0$ with numerical precision 10^{-8}. A drawback of the nonlinear system point of view is that convergence is not guaranteed like it is for the iterated fixed-point algorithm. Note that (4) does not suffer from numerical stability issues despite the exponential terms, because the exponents must be nonpositive. In the special case where the dynamics are finite-horizon, backward-solving will remain faster than solving (4).

Turning to marginalizing out the contribution of the unobserved state to the likelihood, note that a naïve approach would require computing a sum over an exponentially increasing number of paths as T increases:

$$L_T(\theta) = \frac{1}{T} \log \sum_{x_{1:T}} \mathbb{P}(x_1|s_1, a_1) \prod_{t=1}^{T-1} \mathbb{P}((s,x)_{t+1}|(s,x,a)_t; \Pi(\theta))\ \mathbb{P}(a_{t+1}|(s,x)_{t+1}; P(\theta))$$

The *discrete filter* is a recursive algorithm that brings down the computational cost to a linear function of T. It is a recursive algorithm on a particular vector of joint probabilities. π_t is the row vector whose x^{th} coordinate is the joint probability of $x_t = x$ together with the observed data $(a,s)_{1:t}$, $\pi_{t,x} = \mathbb{P}((s,a)_{1:t}, x_t = x)$. π_t obeys the following recursive formula:

$$\pi_{t+1} = \pi_t H_{t+1} \qquad \text{where } H_{t+1,xx'} = \mathbb{P}(x_{t+1} = x', (s,a)_{t+1}|x_t = x, (s,a)_t)$$

The value of $\mathbb{P}((s,a)_{1:T})$ is simply the sum of the coefficients of π_T. Thus, the log-likelihood can be computed from P and Π by doing T matrix multiplications. In practice, because the probabilities of long paths are typically very small, a variant of the algorithm must be used for numerical stability; see appendix section 9.3.

The maximum likelihood estimator can be computed by a straightforward inner-outer algorithm. The "inner loop" is the evaluation of the likelihood at a given value of the structural parameter θ, as described above. The "outer loop" is optimization over the parameter θ. Numerical or exact gradient (with a first-order version of the discrete filter) methods can be used. The resulting inner-outer algorithm is a direct analog to Rust's (1987) nested fixed-point algorithm, although it does not use a fixed-point method in its inner loop and, of course, it includes the marginalization of the unobserved state, absent from a classical Rust model.

Bayesian posteriors can also be computed. A Bayesian posterior can be used in two ways. The first way is the standard Bayesian interpretation. The second way is as a device to obtain classical estimators, by considering a posterior statistic such as the mean, mode or median. A consequence of the Bernstein–von Mises theorem (Theorem 5) is that such posterior estimators will be asymptotically equivalent to the maximum likelihood estimator. Furthermore, consistent confidence intervals can be obtained from the posterior variance. Since structural parameters are economically meaningful, priors are easily formulated. Their influence will fade away as more data come in.

5.3 Other estimation approaches for hidden Rust models

The inner-outer estimators belong to the tradition of Rust's nested fixed-point estimator. Other approaches have proved useful in the classical Rust model literature. Two-step estimators as in Hotz and Miller (1993) and constrained optimization approaches as in Su and Judd (2012) can be generalized to hidden Rust models. Arcidiacono and Miller's (2011) estimator, which combines ideas from both 2-step and constrained optimization estimation, may also be applicable.

There are at least two different ways of generalizing a 2-step approach to hidden Rust models.

A first way is to view a 2-step approach as forming "non-parametric" maximum likelihood estimates of the transition matrices in a first step, and projecting them to a structural parameter space in a least-squares way in a second step. In a classical Rust model, those non-parametric maximum likelihood estimates are available in closed form by counting transitions in data. This is not the case anymore in hidden Rust models where part of the state is unobserved. However, the likelihood can still be numerically maximized at the transition matrix level. The standard technique for this is known as the Baum-Welch algorithm, which is a combination of the EM algorithm and the discrete filter (see, e.g., Zucchini and MacDonald (2009)). This 2-step approach to hidden Rust models is identical to Arcidiacono and Miller's (2011) section 6 estimator. Two-step estimation is statistically efficient with a suitable choice of weighting matrix, but, as Arcidiacono and Miller (2011) points out, it is known to suffer from poor finite sample performances.

A second way is to view a 2-step approach as projecting a set of observed marginals to the structural parameter space. The stable identification theorem (Theorem 2) implies that there is always a finite identifying set of marginals. In practice, it is not clear how to select a good set of marginals. Furthermore, such an estimation approach would likely have problematic short-sample properties and would not be statistically efficient in general.

Su and Judd's (2012) constrained optimization approach to computing the maximum likelihood estimator has a direct analog in hidden Rust models. Let $L_T^{np}(M)$ be the "non-parametric" likelihood parameterized in transition matrices. The following constrained optimization program computes the maximum-likelihood estimator:

$$\begin{aligned}\left(\hat{\theta}, \hat{M}\right) = \underset{\theta, M}{\operatorname{argmax}} \quad & L_T^{np}\left(M\right) \\ & \text{such that: } F\left(\theta, M\right) = 0\end{aligned}$$

The discrete filter is used at each iteration to evaluate $L_T^{np}\left(M\right)$. $F\left(\theta, M\right) = 0$ can be any constraint that expresses the condition "where M is the transition matrix compatible with θ.", for instance based on (3) or (4). The crucial property that $F\left(\theta, M\right) = 0$ must have is uniqueness of the M solution for each θ. A key advantage of the constrained optimization approach is that there is no need to solve the dynamic program.

Arcidiacono and Miller's (2011) section 5 suggests yet an alternative way of computing the maximum likelihood estimator based on a constrained EM algorithm. The EM algorithm at the transition matrix level is modified to take into account a structural constraint, bringing together 2-step and constrained optimization ideas. An advantage of Arcidiacono and Miller's (2011) constrained EM algorithm is that it computes the maximum likelihood estimator when it converges. However, it does not seem to have the increasing-likelihood property of the original EM algorithm, making its convergence properties unclear. The discrete filter of section 5.2 can speed up some of the moves in Arcidiacono and Miller's (2011) constrained EM algorithm.

When the dynamic program is reasonably easy to solve, for instance in a hidden Rust model with sparse transitions matrices with the technique described in the previous section (section 5.2), then the inner-outer approach is fast, efficient and straightforward and I recommend its usage. When the dynamic program is harder to solve (with an eye towards application to dynamic games models), then approaches which do not require explicit solving of the dynamic program may be useful.

6 A structural model of dynamic financial incentives

One-teacher schools may be hard to monitor in sparsely populated regions. When this is the case, *teacher* absenteeism may be a serious issue. To study the effect of financial and monitoring incentives in this context, Duflo et al. (2012) conducted a randomized experiment in the area of Udaipur, Rajasthan, starting in the summer of 2003. Sixty teachers were drawn randomly from a population of 120. Their monthly wage was changed from a flat wage of 1000 rupees to a fixed plus variable structure of 500 rupees plus 50 rupees for every day of work beyond ten days. At the time of the experiment, 1000 rupees were $23 at the real exchange rate, or about $160 at purchasing power parity. At the same time, they were given a camera and instructed to take a picture of themselves with their students at the beginning and the end of each day of class, and to send the pictures to the NGO in charge of the schools. The camera effectively provided a presence-monitoring device.

The randomized control experiment framework cannot disentangle the monitoring effect from the financial incentive effect. A structural model is called for. This is what I focus on by estimating a hidden Rust model as an alternative to Duflo et al.'s (2012) structural model specifications. I will not talk about other steps of Duflo et al. (2012), which include a reduced-form analysis of the experiment's results as well as an examination of the experiment's effects on outcomes such as student learning. The conclusion of the paper is that incentives work and that financial incentives are able to explain most of the observed change in behavior.

Consider a baseline fully observed classical Rust model with choices $a_t = 2$ (the teacher works) or $a_t = 1$ (the teacher does not work), observed state s_t including the number of days left in the month and the number of days worked in the month so far, and flow utilities with two additively separable components for leisure and money:

$$u(s_t, a_t) = u_l \cdot 1[a_t = 1] + u_w \cdot w(s_t, a_t) \tag{5}$$

$w(\cdot)$ is the wage, paid on the last day of the month. Duflo et al.'s (2012) show that such a baseline model cannot explain the correlation patterns in the data (see p. 1259 and Appendix Table 1 there, or see Figure 3 below for an alternative likelihood-based argument). They consider two families of models that add serial correlation to this baseline model. The first family (models III, IV and V in Duflo et al. (2012), or "AR" models) are models of *unobserved* persistence. The unobserved persistence is modelled with AR(1) random utility shocks hitting flow utilities specified as in (5). The second family (models VI, VII and VIII, or "shifter" models) are models of *observed* persistence:[2] classical Rust models where the utility of leisure is higher after a previous day of leisure. Yesterday's decision enters the state, and the flow utilities are given by:

$$u\left(\tilde{s}_t = (a_{t-1}, s_t), a_t\right) = u_{l1} \cdot 1[a_t = 1] + u_{l2} \cdot 1[a_{t-1} = 1] + u_w \cdot w(s_t, a_t)$$

While the shifter models are classical Rust models for which the maximum likelihood

[2] In principle, there is a major testable difference between the unobserved-persistence and the observed-persistence models. The observed data is Markovian in the latter case but not in the former. In practice, the individual time-series lengths are too short to carry this test here. Using a hidden Rust model for unobserved persistence, I can select the unobserved persistence hypothesis over the observed persistence one by looking at the likelihood, which is impossible with an AR model whose likelihood is intractable. See below.

estimator is easily computed with a nested-fixed-point or related algorithm, the AR models are much less tractable and are estimated in Duflo et al. (2012) by a method of simulated moments, using a subset of the model's moments.

Hidden Rust models are alternative, much more tractable models of unobserved persistence. I estimate a hidden Rust model on Duflo et al.'s (2012) data.[3] There are data for 54 teachers, with between 560 and 668 days of data for each teacher. The number of observed states (ordered pairs of days left and days worked) is 378. Different months have different numbers of work days, and teachers may get worked days for free in some months. This gives a little randomness at the start of a new month; otherwise, the evolution of the state is deterministic. As a consequence, the conditional state transition matrices are 98% sparse. Figure 2 pictures the evolution of the state in a typical month.

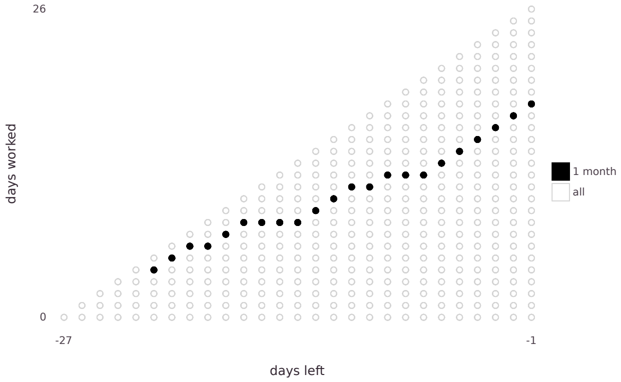


Figure 2: Teacher 22, month 16.

I estimate hidden Rust models with $d_x = 2$, 3, 4 or 7 unobserved states. The dynamic assumptions are as in section 4.1, meaning that the unobserved state has independent Markov dynamics: $\mathbb{P}(x_{t+1}, s_{t+1}|x_t, s_t, a_t) = \mathbb{P}(x_{t+1}|x_t)\mathbb{P}(s_{t+1}|s_t, a_t)$. I use a daily discount factor of $\beta = .9995$. Teachers have unobserved-state specific leisure utilities, meaning the flow utilities are as follows:

$$u(x_t, s_t, a_t) = u_{x_t} \cdot 1[a_t = 1] + u_w \cdot w(s_t, a_t)$$

The structural parameters are the u_x's, u_w and the transition matrix for the unob-

[3] The data are available at http://dspace.mit.edu/handle/1721.1/39124.

served state, estimated “non-parametrically”. There are $d_x + 1 + d_x(d_x - 1) = d_x^2 + 1$ parameters. The case $d_x = 1$ is simply the baseline model (5).

Figure 3 represents the maximized likelihood of hidden Rust models with 1 (baseline model), 2, 3, 4 or 7 unobserved states, along with the maximized likelihoods of the shifter model described above and of a fixed-effects model where the baseline model is estimated separately for each teacher. These models are not nested but they are all nested by a super-model that allows for switching among 54 types and a shifter utility component, making the likelihood comparison meaningful. A hidden Rust model with two unobserved state components already fits the data better than the fixed-effects model. The fact that a hidden Rust model with two unobserved states (five statistical parameters) fit the data better than a fixed-effects model with 108 parameters demonstrates the importance of serial correlation.

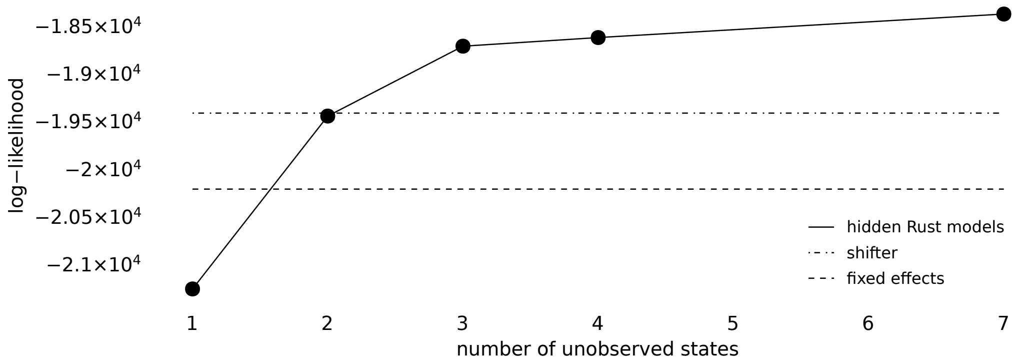


Figure 3

The estimation results for $d_x = 2$, 3, 4 and 7 are presented on the next page.

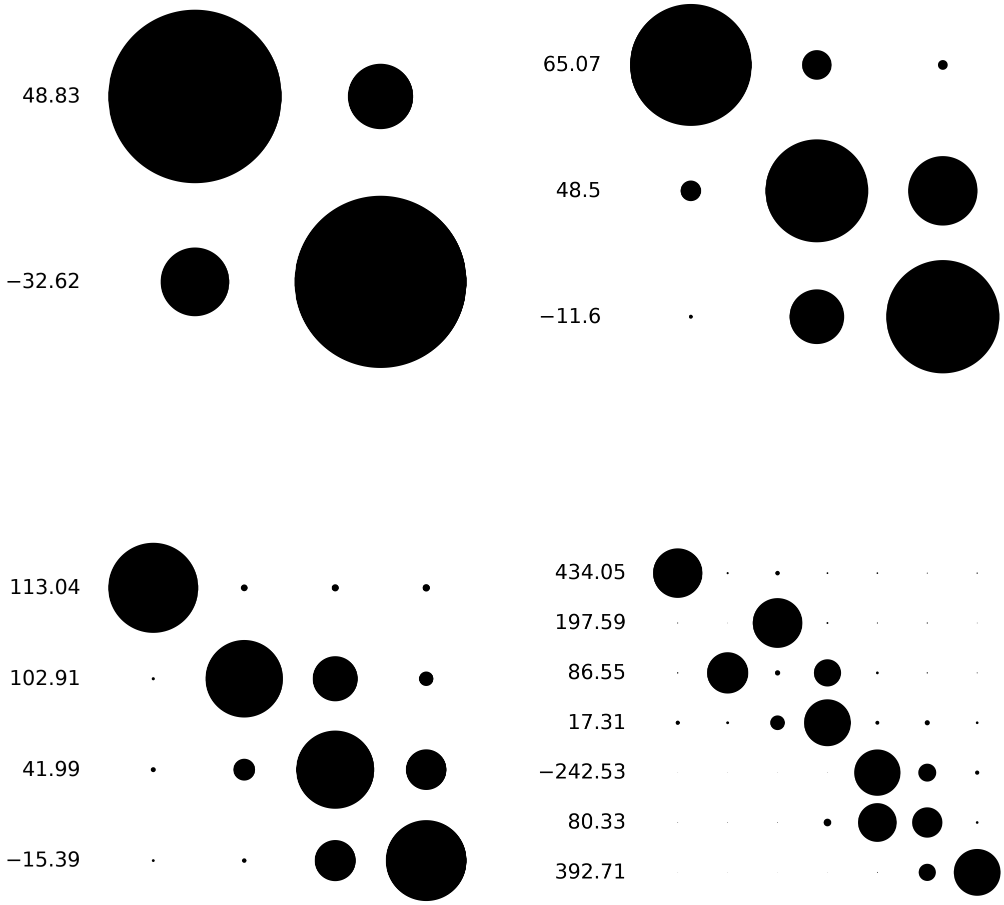


Figure 4: Hidden Rust models estimated for 2, 3, 4 and 7 unobserved states.

Graphical representation of the transition matrices for the unobserved state. The area of each circle is proportional to the corresponding transition probability. Numerical values of the transition probabilities are given in Table 4 in appendix section 9.4. Unobserved state specific leisure utilities are given on the left of the transition matrices, measured in rupees by normalizing by the estimated utilities of money.

Plots such as Figure 4 give us a window to the population's unobserved static and dynamic heterogeneity structure. Here, the likelihood selects a clustered structure. Note that the model does not restrict the shape of the transition matrix whatsoever. I have ordered the sates according to the apparent cluster structure. For instance, with 7 unobserved states, there are three clusters: $\{1\}$, $\{2,3,4\}$ and $\{5,6,7\}$, with rare transitions between clusters but frequent switching within the cluster. The likelihood (Figure 3) already told us that the unobserved structure goes beyond static heterogeneity. Figure 4 shows that it also goes beyond unobserved persistence, which would have translated into a “fat diagonal” transition matrix with monotonic leisure utilities.

The likelihood spends its first degree of freedom on a negative leisure utility state. Negative leisure utility can be interpreted as taste for work or fear of being fired. Its presence is important for the credibility of the structural model. This is necessary in order to correctly predict teacher presence in the out-of-sample control group, which does not have financial incentives. Hidden Rust models pass this model validation test; see Table 2 below. I see the fact that the likelihood reaches this conclusion without using the control group data as evidence in favor of the structural model.

As the number of unobserved states grows, so does the magnitude of the estimated leisure utilities. With seven unobserved states, one day of leisure is worth up to almost ten days of work (500 rupees), driven by the estimated utility of money being close to zero. This is due to overfitting the variability of the data to increasingly open unobserved dynamics. Overfitting is already apparent in the marginal gains in maximized likelihood in Figure 3. For this reason, my favorite specification is a hidden Rust model with $d_x = 3$ unobserved states. From now on I focus on this model. See Table 4 in appendix section 9.4 for the numerical values of the estimated transition matrices for $d_x = 2$, 4 and 7.

Table 1 presents the estimation results with three unobserved states, along with confidence intervals obtained by computing 200 parametric bootstrap draws.

TABLE 1: HIDDEN RUST MODEL WITH THREE UNOBSERVED STATES

utilities (in shock s.d.)				leisure utilities (in rupees)			transition matrix
u_1	u_2	u_3	u_w	u_1/u_w	u_2/u_w	u_3/u_w	Q
10.8 (0.34)	8.02 (0.33)	−1.92 (0.64)	0.165 (0.006)	65.4 (1.6)	48.6 (0.31)	−11.6 (4.2)	$\begin{pmatrix} \underset{(0.044)}{94\%} & \underset{(0.044)}{5\%} & \underset{(0.002)}{1\%} \\ \underset{(0.015)}{3\%} & \underset{(0.031)}{67\%} & \underset{(0.033)}{30\%} \\ \underset{(0.0006)}{0\%} & \underset{(0.043)}{19\%} & \underset{(0.043)}{81\%} \end{pmatrix}$

Two validation exercises can be carried out. A flat wage of 1000 rupees should predict behavior in the control group. Furthermore, after the experiment was over the wage structure of the treatment group was changed to 700 rupees plus 70 rupees after 12 days. Duflo et al. (2012) reports the average presence in both cases. Table 2 presents the corresponding counterfactual attendance probabilities, computed at the maximum likelihood estimate of the structural parameters. Results from Duflo et al.'s (2012) model V are given for comparison, although model selection there was based on matching these statistics.

Table 2: Model validation

		Hidden Rust model		data	Model V
	wage (rupees)	presence (% of days)	days (out of 25)	days	days
factual	$500 + 50 > 10$	68.1%	17.0	17.16	16.75
counterfactual	1000	45.8%	11.5	12.9	12.9
counterfactual	$700 + 70 > 12$	85.7%	21.4	17.39	17.77

The elasticity of labor supply can be computed with respect to a 1% increase in the bonus wage and with respect to an increase of one day in the minimum number of days before the bonus starts to apply. This is done in Table 3, along with bootstrap confidence intervals. While the signs coincide with those of Duflo et al. (2012), I estimate bigger elasticities.

Table 3: Elasticities

		Hidden Rust model		Model V
	wage (rupees)	presence (% of days)	elasticity	elasticity
factual	$500 + 50 > 10$	68.1%	–	–
counterfactual	$500 + 50.5 > 10$	68.8%	1.25% (0.39%)	0.20% (0.053%)
counterfactual	$500 + 50 > 11$	66.9%	−2.77% (1.89%)	−0.14% (0.14%)

Many other counterfactual exercises could be conveniently carried out. Counterfactual distributions are computed exactly (not simulated) by computing the stationary distribution of the hidden Rust model at the maximum likelihood value of the structural parameters. As discussed in section 2, this applies to finite as well as infinite-horizon models. For this reason, computing counterfactuals is very tractable and optimization over counterfactual policies in order to achieve a given policy objective is easily implemented.

An appealing feature of hidden Rust models is that a maximum likelihood path for the unobserved state variable can easily be computed at the maximum likelihood value of the structural parameters, providing additional insight into the cross-section and dynamic heterogeneity patterns present in the data. This is done using an efficient recursive algorithm similar to the discrete filter of section 5 and known as the Viterbi algorithm; see Zucchini and MacDonald (2009). For the sake of illustration, I carried this exercise for seven unobserved states and a larger sample of 60 teachers including six teachers who were almost always absent (Figure 5). Notice how the likelihood spends one unobserved state to explain the outliers and how the Viterbi algorithm picks them up.

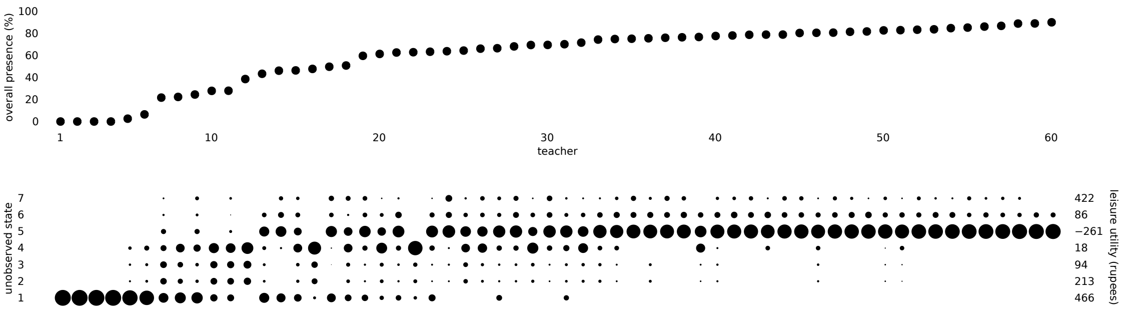


Figure 5: Most likely unobserved path.

x-axis: 60 teachers ranked by attendance rate. Top: attendance rate. Bottom: time spent in each unobserved state at the Viterbi path. The area of each circle is proportional to the proportion of periods spent in the corresponding state. Unobserved state specific leisure utilities are given on the right column, measured in rupees by normalizing by the estimated utilities of money.

In this model of dynamic financial incentives, a hidden Rust model is able to account for cross-section and dynamic heterogeneity patterns in a flexible way. At the same time, it keeps all the advantages of a fully structural model and is very easy to estimate and use for counterfactual computations. The estimate of the transition matrix for the unobserved state provides interesting insight into the unobserved dynamics.

7 More general models

Remember the 2-level structure of a hidden Rust model $\theta \rightarrow (P, M) \rightarrow$ *distribution of the observables*. Most of the results of this paper focus on the transition matrices $\rightarrow$ distribution level of the model, with a particular emphasis on accommodating structural assumptions such as structural zero transition probabilities at the transition matrix level. I used little beyond a reasonable degree of smoothness of the mapping from deeper structural parameters to transition matrices. As a consequence, most of the results of this paper hold directly or are expected to hold for more general "hidden structural models." An example of such a hidden structural model is a dynamic game with n players, unobserved state x_t (which might include both private information and public information unobserved by the econometrician) and public signal s_t, as in Figure 6.

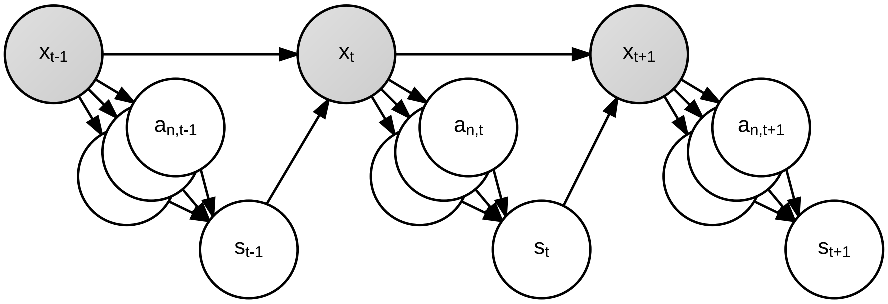


Figure 6: Graphical model for a dynamic game.

Concerning identification, I proved the identification results of section 3 under very general assumptions that any hidden structural model will verify.

The time-series asymptotic results of section 4 will hold under the assumptions stated there. For instance the dynamic game model represented in Figure 6 does not have autoregressive hidden Markov dynamics and is beyond the scope of section 4 I expect the results to hold under more general assumptions.

The estimation approach of section 5 closely matches the 2-level structure of the model. The discrete filter can marginalize out the unobserved state in any dynamic discrete model. On the other hand, how structural parameters map to transition matrices (the “dynamic program level”) is specific to the economic model at hand. In section 5.3 I explained how some cases of intractable dynamic programs can be accommodated.

8 Conclusion

Hidden Rust models provide a framework for unobserved dynamics in a dynamic discrete choice context. I studied their identification and time-series-asymptotic properties, paying particular attention to the role of structural assumptions such as zero transition probabilities. I explained how identification can be attacked at the transition matrix level, and I proved a generic identification theorem that provides a theoretical basis for a convenient model-by-model approach to identification analysis. I proved that hidden Rust models have good time-series asymptotic properties under reasonable assumptions. I explained how to compute the maximum likelihood estimator via an inner-outer algorithm in the spirit of Rust’s (1987) nested fixed-point algorithm.

This paper raises several interesting technical questions. I mentioned some of them in previous sections. Better tools for *computing* identification may be designed. The time-series asymptotic results may be extended to general dynamic discrete models with potentially sparse structure. More general dynamic discrete models, such as dynamic games, may be considered.

9 Appendix

9.1 Meaning of two parameters having the same identification structure

Let $\mathbb{K}$ be a field — such as $\mathbb{R}$ or $\mathbb{C}$ — and $\mathbb{K}[z] = \mathbb{K}[z_1, \ldots, z_n]$ the set of polynomials in n variables with coefficients in $\mathbb{K}$. For $F = (f_i)_i$ an arbitrary collection of polynomials, let $V(F)$ denote the sets of zeros of F:

$$V(F) = \{z \in \mathbb{K}^n \mid \forall i,\ f_i(z) = 0\}$$

As F varies in $\mathbb{K}[z]$, the sets of zeros $V(F)$ have all the properties of the closed sets of a topology. As such they define a topology on $\mathbb{K}^n$ called the *Zariksi topology.* $\mathbb{K}^n$ with its Zariksi topology is called the *affine space* and is written $\mathbb{A}^n(\mathbb{K})$ or simply $\mathbb{A}$. A Zariski topological space is a subset of an affine space with its induced Zariski topology. The dimension of a Zariski topological space is its topological dimension. Dimension is well-behaved for Zariski closed subsets of the affine space: $\mathbb{A}^n(\mathbb{K})$ has dimension n, points have dimension 0, the zero-set of a non-zero polynomial has dimension $n-1$ ("hypersurface"), etc. A non-empty topological space is called irreducible if it cannot be written as a union of two closed proper subsets. Among the irreducible closed subsets of the affine space, those of dimension 0 are the points. Zariski topological spaces have the following property:

Fact 1: Decomposition in irreducible components
Any Zariski topological space W can be written as a finite union of irreducible components $W = V_1 \cup \ldots \cup V_M$. This decomposition is unique up to renumbering under the condition that no V_m is included in another $V_{m'}$.

See Reid (1988) for an introduction to algebraic geometry, and the first chapter of Görtz and Wedhorn (2010) for some of the properties of the Zariski topology.

In the context of algebraic models (see section 3), I make the following definition:

Definition 1: $\lambda_1^\star$ and $\lambda_2^\star$ have the same identification structure
$\lambda_1^\star$ and $\lambda_2^\star$ have the same identification structure if $V_\lambda\left(F\left(\cdot, \lambda_1^\star\right)\right)$ and $V_\lambda\left(F\left(\cdot, \lambda_2^\star\right)\right)$, seen as Zariski topological spaces, have equally many irreducible components of each dimension.

When the model is identified at $\lambda^\star$, $V_\lambda\left(F\left(\cdot,\lambda^\star\right)\right)$ has 1 "irreducible component of dimension zero" (i.e. 1 point, namely $\lambda^\star$) and no other irreducible component, or more generally n_{ls} "irreducible components of dimension zero" when there is label-switching.

Example 1: Consider a model without label-switching. Let $W_i = V_\lambda\left(F\left(\cdot,\lambda_i^\star\right)\right)$ for some $\lambda_i^\star$, $1 \leq i \leq 5$. Suppose the W_i's have the following decompositions in irreducible components (all V_j's of dimension 2):

$$\begin{array}{llll} W_1 = \{\lambda_1^\star\} & W_2 = \{\lambda_2^\star\} \cup \{\lambda_3^\star\} & W_3 = \{\lambda_2^\star\} \cup \{\lambda_3^\star\} & W_4 = V_4 \\ W_5 = \{\lambda_5^\star\} \cup V_5 & W_6 = \{\lambda_6^\star\} \cup V_6 & W_7 = \{\lambda_7^\star\} \cup V_7 \cup V_7' & W_8 = \{\lambda \neq \lambda_8^\star\} \cup V_8 \end{array}$$

The model is (globally) identified at $\lambda_1^\star$. $\lambda_2^\star$ and $\lambda_3^\star$ are observationally equivalent and the model is locally identified at $\lambda_2^\star$ and $\lambda_3^\star$. If we were anticipating a label-switching feature between $\lambda_2^\star$ and $\lambda_3^\star$, we could consider the model to be globally identified at $\lambda_2^\star$ and $\lambda_3^\star$. The model is not locally identified at $\lambda_4^\star$. The model is locally identified at $\lambda_5^\star$, $\lambda_6^\star$ and $\lambda_7^\star$, but not globally identified. $\lambda_5^\star$ and $\lambda_6^\star$ have the same identification structure but not $\lambda_7^\star$. $\lambda_8^\star$ is somewhat of a pathological case with an isolated solution that is not $\lambda_8^\star$ ($\lambda_8^\star$ must belong to V_8). The model is not locally identified at $\lambda_8^\star$, but $\lambda_8^\star$ has the same identification structure as $\lambda_5^\star$ and $\lambda_6^\star$. □

9.2 Counterexample of a smooth model identified from its infinite collection of marginals but not from any finite number of them

Consider a sequence $Y_{1:\infty}$ of 0's and 1's with the following structure: a sequence of 1's (empty, finite or infinite) followed by a sequence of 0's (infinite, infinite or empty, respectively). The distribution of $Y_{1:\infty}$ is fully specified by the decreasing sequence of numbers $q_T := \mathbb{P}(Y_{1:T} = (1,\ldots,1))$. Now consider a model for $Y_{1:\infty}$ from parameter space $\Lambda = [0,1]$ as follows: $q_T(\lambda) = 1$ for $0 \leq \lambda \leq 1/(T+1)$, $q_T(\lambda) = 0$ for $1/T \leq \lambda \leq 1$ and $\lambda \to q_T(\lambda)$ is smooth (infinitely differentiable). Then $\lambda^\star = 0$ is identified from the distribution of the whole sequence (all Y_t's are 1 with probability 1 iff $\lambda = 0$) but not from any set of finite marginals (all $0 \leq \lambda \leq 1/(T+1)$ are compatible with $\mathbb{P}(Y_{1:T} = (1,\ldots,1)) = 1$).

9.3 Numerically stable discrete filter

The discrete filter cannot be implemented directly in practice due to numerical precision issues. The probability of a long path $(s, a)_{1:T}$ is typically very small. The algorithm needs to be augmented with the log of a normalization factor for π_t, say ρ_t .

$$\text{initialization:} \qquad \begin{cases} \tilde{\pi}_1 & = \pi_1 = \mu^\star(s_1, x_1, a_1) \\ \log \rho_1 & = 0 \end{cases}$$

$$\text{iteration:} \qquad \begin{cases} \pi_{t+1} & = \tilde{\pi}_t H_{t+1} \\ \tilde{\pi}_{t+1} & = \dfrac{\pi_{t+1}}{\|\pi_{t+1}\|_1} \\ \log \rho_{t+1} & = \log \rho_t + \log \|\pi_{t+1}\|_1 \end{cases}$$

At the end of the recursion, ρ_T is directly $\log \mathbb{P}((s, a)_{1:T})$.

9.4 Some details for the empirical application

Table 4 presents the estimated values of the transition probabilities for $d_x = 2$, 3, 4 and 7. Those are the numerical values used to create the bubble representation of the transition matrices, Figure 4 in section 6.

Table 4: Estimated transition matrices

d_x	Q
2	$\begin{pmatrix} 87.5\% & 12.5\% \\ 13.7\% & 86.3\% \end{pmatrix}$
3	$\begin{pmatrix} 93.9\% & 5.5\% & 0.6\% \\ 2.7\% & 66.9\% & 30.4\% \\ 0.1\% & 18.9\% & 81.0\% \end{pmatrix}$
4	$\begin{pmatrix} 98.2\% & 0.5\% & 0.6\% & 0.7\% \\ 0.1\% & 72.7\% & 24.6\% & 2.6\% \\ 0.0\% & 0.2\% & 20.4\% & 79.4\% \end{pmatrix}$
7	$\begin{pmatrix} 98.7\% & 0.2\% & 0.8\% & 0.1\% & 0.1\% & 0.1\% & 0.0\% \\ 0.0\% & 0.0\% & 99.7\% & 0.1\% & 0.1\% & 0.1\% & 0.0\% \\ 0.0\% & 68.6\% & 1.1\% & 29.8\% & 0.3\% & 0.2\% & 0.0\% \\ 0.8\% & 0.3\% & 8.6\% & 88.4\% & 0.6\% & 1.0\% & 0.3\% \\ 0.0\% & 0.0\% & 0.0\% & 0.0\% & 86.4\% & 12.9\% & 0.7\% \\ 0.0\% & 0.0\% & 0.0\% & 2.4\% & 60.4\% & 36.9\% & 0.3\% \\ 0.0\% & 0.0\% & 0.0\% & 0.0\% & 0.0\% & 11.9\% & 88.1\% \end{pmatrix}$

References

Aguirregabiria, V. and P. Mira (2010): "Dynamic Discrete Choice Structural Models: A Survey," *Journal of Econometrics*, 156, 38–67.

Allman, E., C. Matias, and J. Rhodes (2009): "Identifiability of Parameters in Latent Structure Models with Many Observed Variables," *The Annals of Statistics*, 37, 3099–3132.

Arcidiacono, P. and R. Miller (2011): "Conditional Choice Probability Estimation of Dynamic Discrete Choice Models with Unobserved Heterogeneity," *Econometrica*, 79, 1823–1867.

Baum, L. E. and T. Petrie (1966): "Statistical Inference for Probabilistic Functions of Finite State Markov Chains," *The Annals of Mathematical Statistics*, 37, 1554–1563.

Bickel, P. J. and Y. Ritov (1996): "Inference in Hidden Markov Models I: Local Asymptotic Normality in the Stationary Case," *Bernoulli*, 2, 199–228.

Bickel, P. J., Y. Ritov, and T. Ryden (1998): "Asymptotic Normality of the Maximum-Likelihood Estimator for General Hidden Markov Models," *The Annals of Statistics*, 26, 1614–1635.

Bonhomme, S., K. Jochmans, and J.-M. Robin (2016): "Estimating multivariate latent-structure models," *The Annals of Statistics*, 44, 540–563.

Chiong, K., A. Galichon, and M. Shum (2014): "Duality In Dynamic Discrete Choice Models," *Working Paper*.

Diermeier, D., M. Keane, and A. Merlo (2005): "A Political Economy Model of Congressional Careers," *American Economic Review*, 95, 347–373.

Douc, R., E. Moulines, J. Olsson, and R. van Handel (2011): "Consistency of the Maximum Likelihood Estimator for General Hidden Markov Models," *The Annals of Statistics*, 39, 474–513.

Douc, R., E. Moulines, and T. Ryden (2004): "Asymptotic Properties of the Maximum Likelihood Estimator in Autoregressive Models with Markov Regime," *The Annals of Statistics*, 32, 2254–2304.

Duflo, E., R. Hanna, and S. P. Ryan (2012): “Incentives Work: Getting Teachers to Come to School,” *The American Economic Review*, 102, 1241–1278.

Görtz, U. and T. Wedhorn (2010): *Algebraic Geometry*, Springer.

Hotz, V. and R. Miller (1993): “Conditional Choice Probabilities and the Estimation of Dynamic Models,” *Review of Economic Studies*, 60, 497–529.

Hu, Y. and M. Shum (2012): “Nonparametric Identification of Dynamic Models with Unobserved State Variables,” *Journal of Econometrics*, 171, 32–44.

Kasahara, H. and K. Shimotsu (2009): “Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices,” *Econometrica*, 77, 135–175.

Keane, M., P. Todd, and K. Wolpin (2011): “The Structural Estimation of Behavioral Models: Discrete Choice Dynamic Programming Methods and Applications,” *Handbook of Labor Economics*, 4, 331–461.

Kristensen, D., L. Nesheim, and A. de Paula (2014): “CCP and the Estimation of Nonseparable Dynamic Discrete Choice Models,” *Working Paper.*

Miller, R. (1984): “Job Matching and Occupational Choice,” *The Journal of Political Economy*, 92, 1086–1120.

Norets, A. (2009): “Inference in Dynamic Discrete Choice Models With Serially Correlated Unobserved State Variables,” *Econometrica*, 77, 1665–1682.

Pakes, A. (1986): “Patents as Options: Some Estimates of the Value of Holding European Patent Stocks,” *Econometrica*, 54, 755–784.

Reid, M. (1988): *Undergraduate Algebraic Geometry*, Cambridge University Press.

Rust, J. (1987): “Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher,” *Econometrica*, 55, 999–1033.

——— (1996): “Numerical Dynamic Programming in Economics,” *Handbook of Computational Economics*, 1, 619–729.

Su, C.-L. and K. Judd (2012): “Constrained Optimization Approaches to Estimation of Structural Models,” *Econometrica*, 80, 2213–2230.

Wolpin, K. (1984): “An Estimable Dynamic Stochastic Model of Fertility and Child Mortality,” *The Journal of Political Economy*, 92, 852–874.

Zucchini, W. and I. MacDonald (2009): *Hidden Markov Models for Time Series*, CRC Press.