Hidden Rust Models

Benjamin Connault*

Princeton University, Department of Economics

Job Market Paper

November 2014

Abstract

This paper shows that all-discrete dynamic choice models with hidden state provide for general, yet tractable structural models of single-agent dynamic choice with unobserved persistence. I call such models hidden Rust models and I study their econometrics. First, I prove that hidden Rust models and more general dynamic discrete models have a generic identification structure, in a strong technical sense, as well as a stable one. Generic implies that checking identification at one parameter value is enough to get identification at almost every parameter value and I provide the tools to do so. Second, I prove that hidden Rust models have good time-series asymptotic properties despite their complicated dynamic structure. Third, I show that both the maximum likelihood estimator and Bayesian posteriors can be efficiently computed by carefully exploiting the structure of the model. These results are applied in a model of dynamic financial incentives inspired by Duflo, Hanna and Ryan (2012). Several lessons carry over to other dynamic discrete models with unobserved components.

*princeton.edu/~connault and connault@princeton.edu

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 and participants at the Princeton Econometrics seminar 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. 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 persistence, regime switching, structural breaks and heterogeneity are not part of the classical Rust model world. The challenge[1] is to find a framework that would allow for these important economic features while maintaining the structural character, the flexibility and the empirical tractability of classical Rust models.

In this paper, I show that a family of all-discrete dynamic choice models with hidden states, which I call *hidden Rust models*, provides 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 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 infinite-horizon assumption is without loss of generality, as explained in section 2.3. When it comes to computing estimators in section 5, the random utility shocks are further assumed to be additively separable and independent identically Gumbel distributed, as in Rust (1987). Since the decision-making process is the same in classical dynamic discrete choice models and in hidden Rust models, the familiar conditional choice probabilities are present. However, they

[1] Working paper versions of Duflo et al. (2012) used to say: "Allowing for serial correlation and heterogeneity considerably complicates the estimation procedure, but we show that these features are very important in this application."

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. This is explained in more detail in section 2.2. Issues of identification, asymptotics and estimation must be studied anew. This is what I do in this paper.

In section 3, I examine the question of identification in hidden Rust models. It is well-known that models with unobserved variables can suffer from loss of identification, meaning that different structural parameters might generate observationally equivalent distributions for the data. I cast the question of identification in hidden Rust models in terms of multivariate systems of polynomials and I study the corresponding zero-sets from an algebro-geometric point of view. The results are motivated by hidden Rust models but valid generally for any discrete model (Theorem 1) and any dynamic discrete model (Theorem 2), respectively.

Theorem 1 says that discrete models have the same identification structure almost everywhere on the parameter space, specifically outside of the zero-set of a polynomial system. Such zero-sets are very small (Corollary 2). A particular case is when the model is globally identified for almost every parameter value; then I say that the model is *generically identified*.

A spectrum of methods can be used to compute identification in a particular model. At one extreme, there are computationally intensive algorithms to compute a minimal singular region along with the generic identification structure. On the other side of the spectrum, global identification at a randomly drawn parameter value implies generic identification.

Another identification theorem (Theorem 2) says that the identification structure of dynamic discrete models stabilizes after some finite time horizon. A corollary of Theorem 2 is that if a dynamic discrete model is identified from its infinite-horizon joint distribution, then it is identified from a finite-horizon set of marginals (Corollary 3). An example of a smooth but non-discrete dynamic model identified from its infinite-horizon joint distribution but not from any finite-horizon set of marginals (Remark 1) shows that Theorem 2 captures a phenomenon specific to dynamic discrete models.

Next, 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. The results carry over directly to a fixed number of individuals and many successive observations. Time-series asymptotics are the relevant asymptotic framework in the empirical application of this paper, inspired by Duflo et al. (2012) and presented in section 6.

Theorem 3 shows that hidden Rust models are locally asymptotically normal. This is a regularity property: smooth, independent and identically distributed models are the typical locally asymptotically normal models. This means that the time-series properties of hidden Rust models are not irregular the way unit-root asymptotics are. Theorem 4 shows that the maximum likelihood estimator is consistent and asymptotically normal. Theorem 5 is a Bernstein–von Mises theorem, a result about the asymptotic behavior of Bayesian posteriors from a frequentist point of view. Two consequences of Theorem 5 are that Bayesian posterior estimates, such as the posterior mode, are asymptotically equivalent to the maximum likelihood estimator and that confidence intervals can be obtained from the posterior variance. Theorem 5 is obtained by applying the general weakly dependent Bernstein–von Mises theorem of Connault (2014) to hidden Rust models. A direct consequence of local asymptotic normality (Theorem 3) is that the maximum likelihood estimator and Bayesian posterior estimates are statistically efficient in the strong sense of the Le Cam convolution theorem, meaning they will behave better than any other reasonable estimator under any reasonable loss function.

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 evaluated in an efficient way by building on the statistical structure of a hidden Rust model. There are 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 dynamic program parameterized in expected value function. This is often faster than the usual fixed-point algorithm and the technique automatically takes advantage of the sparsity structure commonly found in these models. In the second stage, a "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 Gaussian state-space models.

Tractable evaluation of the likelihood means both the maximum likelihood estimator and Bayesian posteriors can be computed in an efficient way. The maximum likelihood estimator can be computed by optimizing over the parameter space in an outer loop while evaluating the likelihood in an inner loop in two stages, as described above. This inner-outer algorithm is similar in spirit to Rust's original nested fixed-point algorithm — without the fixed-point part and with the discrete filter taking care of the unobserved state variable. Bayesian posteriors can be simulated via Markov chain Monte Carlo methods. This is also an inner-outer algorithm, not very different from maximum likelihood. The inner loop is the same. The outer loop "samples" from the likelihood as opposed to finding its maximum. These two estimation methods belong to the tradition of Rust's nested fixed-point algorithm. I also comment about the possibility of constrained optimization approaches, as in Su and Judd (2012), and 2-step estimators in the spirit of Hotz and Miller (1993).

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. The data 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[2] full-information[3] 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." On the identification front, the identification results of this paper have already been stated for general discrete and dynamic discrete models. On the asymptotic front, the results apply directly to other dynamic discrete models with dynamics similar to those of hidden Rust models. On the estimation front, 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 as in Su and Judd (2012) can be used.

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

[3]Duflo et al. (2012) uses an alternative model of unobserved persistence, with autoregressive random utility shocks. The model has to be estimated by simulated moments on a subset of moments.

1.2 Literature review

Some econometric aspects of dynamic discrete choice models with hidden 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.

On the identification side, a potentially useful approach to identification can be found in Hu and Shum (2012) and An et al. (2014) (see also Kasahara and Shimotsu (2009) for the particular case of heterogeneity, i.e., static mixing). Both papers give sufficient conditions for global identification that require checking the invertibility of a number (as many as the dimension of the observed state) of square[4] matrices of marginal probabilities. In the original papers, the conditions must be checked for all parameter values θ. Invoking this paper's Theorem 1, if the conditions are verified at a randomly drawn parameter $\theta^\star$, then generic identification follows.

Hu and Shum's (2012) and An et al.'s (2014) results apply only to models with some level of regularity,[5] whereas hidden Rust models can be very singular in applications, because of assumptions such as structural zero transition probabilities and conditional independences.[6] In Rust's (1987) famous bus example, 97% of the coefficients of the conditional state transition matrix are structural zeros because a bus's mileage can increase by zero, one or two brackets on each trip, out of 90 possible mileage brackets observed in the sample. Figure 9 in appendix section A14 pictures the structure of the conditional state transition matrices in the model of section 6; the sparsity is even higher at 98%.

Discrete models have generic identification features. This is a well-identified phenomenon in the algebraic statistics literature (Allman et al. (2009)). Structural assumptions can interact with these features, making one-size-fits-all sufficient conditions for identification hard to envision. This is a major motivation for developing a convenient, systematic, model-by-model

[4] In the discrete context, the theorems apply to models where the unobserved variable x_t and the observed variable y_t have equal dimension. In a hidden Rust model typically $d_x \ll d_y$, but one can try to check the sufficient conditions for an aggregate observed variable $g(y_t)$, $dim(g(y)) = dim(x)$. There is a combinatorial explosion in the possible ways of aggregating $g(y_t)$.

[5] For instance, under Hu and Shum's (2012) Theorem 1, a stationary model is identified from 4 time periods, whereas it is known that particular stationary hidden Markov models may require an arbitrary number of time periods to be identified (Gilbert, 1959).

[6] For instance hidden Rust models fall badly on the singular region of Petrie's (1969) classical generic identification theorem for hidden Markov models. One of the sufficient conditions for identification in Petrie (1969) is that the transition probabilities from non-observables to observables be non-zero for all possible pairs of values. In a hidden Rust model cast as a hidden Markov model with "unobservable" $z_t = (x_t, s_t, a_t)$ and observable $y_t = (s_t, a_t)$, the relevant transition probabilities are zero for most pairs of values.

approach to identification analysis. This paper's results are a first step in this direction.

Gilbert (1959) gives an explicit bound[7] for stable identification in the case of stationary hidden Rust models. Theorem 2 applies to much more general models, including non-stationary hidden Rust models.

On the asymptotic theory side, Baum and Petrie (1966) proved consistency and asymptotic normality of the maximum likelihood estimator for strict hidden Markov models (see Figure 1) 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 (see section 4.4 for an outline). 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. By contrast, hidden Rust models have autoregressive hidden Markov dynamics with discrete state but potentially very sparse conditional transition matrices.

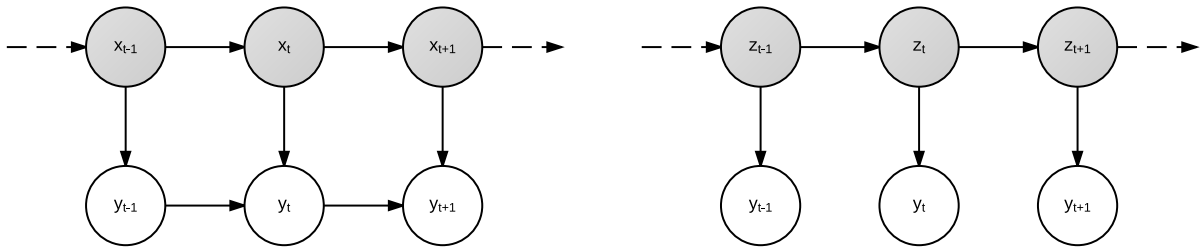


Figure 1: Autoregressive (left) and strict (right) hidden Markov dynamics. Unobserved variables are shaded.

[7] $2(d_x d_y - d_y + 1)$ where x_t is the unobserved state and y_t the observed variable (observed state and decision).

On the estimation side, Arcidiacono and Miller (2011) considers models slightly less general but very related to the models of section 5. 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 known in various fields dealing with dynamic discrete models; see, for instance, Zucchini and MacDonald (2009).

2 Hidden Rust models

The underlying model of dynamic decision making is identical in a hidden Rust model and in a classical dynamic discrete choice 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.

Section 2.1 describes hidden Rust models. Section 2.2 explains why the econometric theory of classical dynamic discrete choice models does not carry over to hidden Rust models and why they need to be reconsidered — which is what I do in this paper. Section 2.3 comments about a number of assumptions maintained throughout the paper.

2.1 Model description

An agent makes dynamic decisions following a stationary infinite-horizon[8] 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 $k_t \in \{1, \ldots, d_s\}$ 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 $x_t \in \{1, \ldots, d_x\}$. I assume autonomous Markov dynamics for the unobserved state and no direct feedback from the unobserved state to the observed state, as in Figure 2.[9] I use these specifc dynamics for the time-series asymptotic analysis in section 4 (see also section 7 for comments about asymptotic results under more general dynamics). The estimation techniques of section 5 are valid for more general dynamics where the state k_t is Markov conditionally on a_t, without further restriction. The identification results of section 3 are valid much more generally, for any dynamic discrete model.

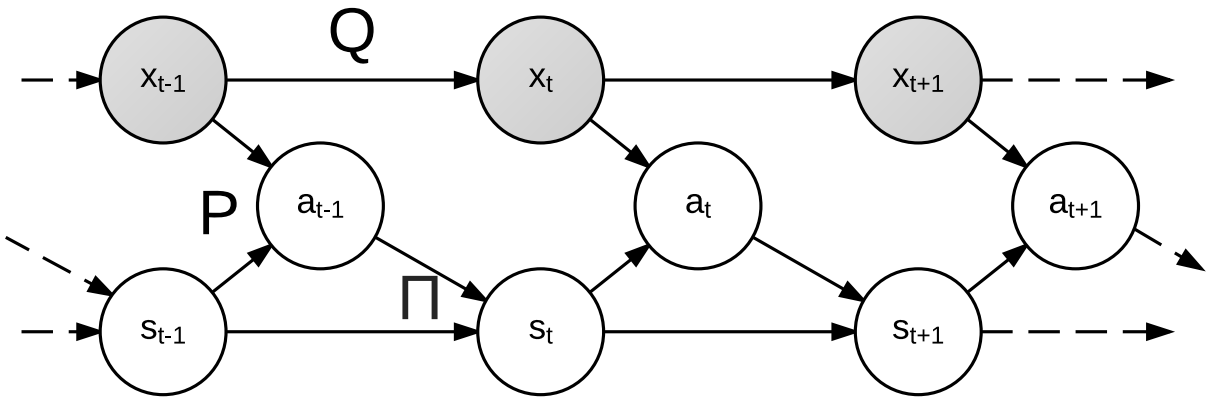


Figure 2: A hidden Rust model (unobserved variables are shaded).

The distribution of the data is fully specified by some initial distribution and the following transition matrices: the transition matrix Q for the unobserved state, the conditional state transition matrices Π_a, $\Pi_{a,ss'} = \mathbb{P}(s'|s,a)$, and the conditional choice probability matrix P, $P_{ka} = \mathbb{P}(a|k)$. The conditional choice probabilities must be compatible with rational decision-making, typically indexed by a structural utility parameter θ_u. Other transition matrices may be parameterized structurally or simply by their coefficients. All parameters

[8] Section 2.3 shows that the infinite-horizon assumption is without loss of generality.

[9] All dynamic assumptions can be stated formally in terms of conditional independence statements such as $\mathbb{P}((x,s,a)_{t+1}|(x,s,a)_{1:t}) = \mathbb{P}((x,s,a)_{t+1}|(x,s,a)_t)$. Alternatively, a graphical model such as Figure 2 specifies without ambiguity all the conditional independences. See Jordan and Weiss (2002) for an introduction to graphical models.

including θ_u are grouped in the model's structural parameter θ. The resulting two-level structure is leveraged throughout the paper:

$$\text{STRUCTURAL PARAMETERS} \rightarrow \text{TRANSITION MATRICES}$$
$$\rightarrow \text{DISTRIBUTION OF THE OBSERVABLES}$$

Although the economic variables $z_t = (x_t, s_t, a_t)$ follow a Markov chain, the observed data $y_t = (s_t, a_t)$ is not Markovian of any order. The log-likelihood can be written:

$$L_T(\theta) = L_T^{np}(P(\theta), Q(\theta), \Pi(\theta))$$

where L_{np} is a "non-parametric" log-likelihood:

$$\begin{aligned} L_T^{np}(P, Q, \Pi) &= \frac{1}{T} \log \mathbb{P}((s_t, a_t)_{2:T} | s_1, a_1; P, Q, \Pi) \\ &= \frac{1}{T} \log \sum_{x_{1:T}} \mathbb{P}(x_1 | s_1, a_1) \mathbb{P}((s, x, a)_{2:T} | (s, x, a)_1; P, Q, \Pi) \end{aligned}$$

As far as identification and asymptotics are concerned, the specific form of the mapping from structural parameters to transition matrices does not matter. When it comes to computing estimators in section 5, I will make the assumption of discounted utilities with additively separable random utility Gumbel shocks, as in Rust (1987). Since the decision-making is identical in the fully and partially observed models, standard results aboout the Rust model will apply: Bellman equation, logit form of the conditional choice probabilities, etc. To make the paper self-contained, the relevant facts will be recalled in section 5.

Example: In section 6, I consider a model of financial incentives inspired by Duflo et al. (2012). Teachers decide whether to go to work ($a_t = 1$) or not ($a_t = 2$) every day. On the last day of the month, they are paid a wage $w(s_t)$ based on the number of days they have worked during the month. On a given day, the observed state s_t includes the number of days worked so far during the month as well the number of days left in the month. As stressed in Duflo et al. (2012), the financial incentive structure does not account for the observed dynamic patterns: there seem to be important unobserved persistence factors. I use a hidden Rust model to account for the unobserved persistence. x_t will be a general willingness-to-work variable, which can capture unobserved dynamic effects at various frequencies such as health accidents, the need to travel, a dynamic type heterogeneity across teachers, etc. The structural parameters include the unobserved state transition matrix and utility parameters. We can estimate, in rupees, the value of leisure at different willingness-to-work levels.

2.2 The econometrics of hidden Rust models require new arguments

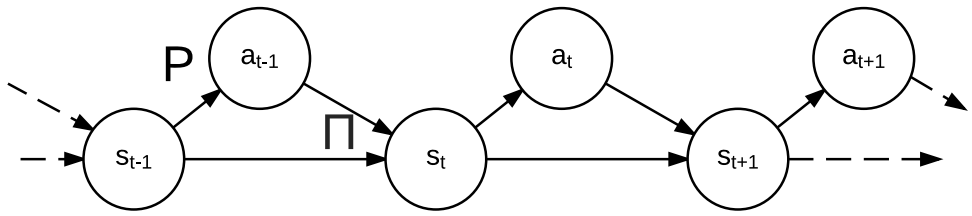


Figure 3: A classical dynamic discrete choice model.

In a classical dynamic discrete choice 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 matrices (P, Π) are identified, that is to say, as soon as the mapping $\theta \to (P, \Pi)$ 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."

Classical dynamic discrete choice models achieve a rare combination of being fully structural and very tractable. Their one major shortcoming is that all the variables relevant to decision-making must be observed; in particular, there can be no unobserved persistence or heterogeneity.

Hidden Rust models correct this shortcoming by allowing unobserved persistence patterns to be carried by the unobserved state variable. At the same time, hidden Rust models remain fully structural and this paper shows that they also have good econometric properties. However, new arguments are needed. In terms of identification, two different sets of transition matrices (P, Q, Π) could give rise to the same distribution of the observables. In terms of asymptotics, the observed data is not Markovian of any order, making the asymptotic analysis much harder. For instance, the log-likelihood cannot even be written as an ergodic sum. 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 estimates of P and Q because we do not observe transitions in data.

2.3 Discussion about some assumptions

This paper focuses on stationary infinite-horizon models without loss of generality. Under the additively separable Gumbel assumption, any finite-horizon model can be formally cast as an equivalent stationary infinite-horizon model. If you optimize over your yearly savings and consumption in this life, it does not matter if you have infinitely many similar lives waiting for you after the current one or none at all, as long as this life's choices have no impact on the next. In practice, this is done by adding time to the state variable and infinitely repeating the finite-horizon model, drawing the initial state randomly at each renewal. See appendix section A10 for an example. The resulting stationary infinite-horizon model will have very sparse transition matrices, something that the technique described in section 5.2 takes advantage of for efficient likelihood evaluation. See the discussion in section 5.2.

Another assumption is that the state variable is discrete. There are at least two reasons for making this assumption. First, discreteness is a central feature in many economic models and as such deserves to be examined carefully in and for itself. For example, state variables are often restricted to evolve from one value to only a handful of other values in one period. This results in transition matrices with many zeros, which is an opportunity when it comes to solving the dynamic program (see section 5) but a challenge for identification and asymptotics (sections 3 and 4). Second, many insights carry over from discrete models to continuous ones: discreteness allows us to focus on statistical issues, without worrying about issues of numerical approximation of continuous quantities by discrete ones in computers. On the other hand, this paper is not taking advantage of the fact that transition probabilities are likely to vary smoothly as functions of the state variables. The extension of hidden Rust models to various continuous settings is an interesting research question.

Two other assumptions are that the model is well-specified and that the number of unobserved states is known. From one point of view, the unobserved state is here to carry unobserved persistence patterns in the data and has no literal real-world counterpart; then thinking about misspecification is relevant. From an opposite point of view, consistently selecting the true dimension of the unobserved state variable matters. These two points of view are not strictly incompatible. Both directions would be interesting to pursue.

3 Identification

This section studies identification in hidden Rust models and discrete econometric models in general. The results are obtained from an algebro-geometric point of view after the question of identification in discrete models (such as hidden Rust models) is cast as a system of multivariate polynomial equations. All proofs are in appendix section A11. A fully worked out example of a generically identified model, complete with an explicit minimal singular region, is given in appendix section A11.5.

Section 3.1 casts the issue of identification in hidden Rust models as a polynomial system. Section 3.2 shows that discrete models have a generic identification structure (Theorem 1). Section 3.3 shows that dynamic discrete models have a stable identification structure (Theorem 2). Section 3.4 looks at practical issues of computing identification in particular models.

3.1 Identification in discrete models as a polynomial system

In a hidden Rust model the mapping from structural parameters θ_u to conditional choice probabilities is usually not polynomial; however, the mapping from transition matrices (P, Q, Π) to the probability distribution of the observables is polynomial in the individual coefficients of (P, Q, Π), or at least rational. 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:

$$\mathbb{P}(y_{1:T}) = \sum_{x_{1:T}} \mathbb{P}(y_{1:T}, x_{1:T}) = \sum_{x_{1:T}} \mathbb{P}(x_1, s_1, a_1) \prod_{t=1}^{T-1} Q_{x_t x_{t+1}} \Pi_{a_t, s_t s_{t+1}} P_{k_{t+1} a_{t+1}}$$

There are three cases. If the initial distribution is known, then $\mathbb{P}(y_{1:T})$ is a polynomial function of the coefficients of (P, Q, Π). If the initial distribution is estimated, then $\mathbb{P}(y_{1:T})$ is a polynomial function of the coefficients of (P, Q, Π) and the initial distribution. If the initial distribution is assumed to be the stationary distribution, then it is a rational function of the coefficients of (P, Q, Π) (as an eigenvalue) and thus $\mathbb{P}(y_{1:T})$ is a rational function of the coefficients of (P, Q, Π). In all cases, $\mathbb{P}(y_{1:T})$ is a rational mapping $\frac{f(\theta^{np}; y_{1:T})}{g(\theta^{np}; y_{1:T})}$ of a "non-parametric parameter" θ^{np}, which contains the non-zero coefficients of (P, Q, Π) and those of the initial distribution when applicable.

If Φ_T is the mapping from θ^{np} to the distribution of $Y_{1:T}$, i.e. to all d_y^T probabilities $\mathbb{P}(y_{1:T})$, then for any θ^{np}, $\bar{\theta}^{np}$:

$$\begin{aligned}
\Phi_T\left(\theta^{np}\right) = \Phi_T\left(\bar{\theta}^{np}\right) &\iff \forall y_{1:T}, \quad \frac{f\left(\theta^{np}; y_{1:T}\right)}{g\left(\theta^{np}; y_{1:T}\right)} = \frac{f\left(\bar{\theta}^{np}; y_{1:T}\right)}{g\left(\bar{\theta}^{np}; y_{1:T}\right)} \\
&\iff \forall y_{1:T}, \quad f\left(\theta^{np}; y_{1:T}\right) g\left(\bar{\theta}^{np}; y_{1:T}\right) = f\left(\bar{\theta}^{np}; y_{1:T}\right) g\left(\theta^{np}; y_{1:T}\right) \\
&\iff F_T\left(\theta^{np}, \bar{\theta}^{np}\right)
\end{aligned}$$

where F_T is a system of d_y^T multivariate polynomial equations.

A hidden Rust model is (non-parametrically) identified at $\bar{\theta}^{np}$ if:

$$\forall \bar{\theta}, \qquad \Phi_T\left(\theta^{np}\right) = \Phi_T\left(\bar{\theta}^{np}\right) \Longrightarrow \theta^{np} = \bar{\theta}^{np}$$

This means we can study non-parametric identification of a hidden Rust model by studying the set of solutions to a polynomial system, which is what I do in this paper.

Non-parametric identification is not the same as identification at the structural level. There are at least three reasons why studying identification at the non-parametric level is interesting. First, generic identification at the non-parametric level as in Theorem 1 is expected to carry back to generic identification at the structural level. Once a specific mapping from structural parameters to transition matrices is specified (for instance, with the assumption of additively separable Gumbel shocks), it is a matter of showing that the image of this mapping[10] intersects cleanly with any variety in the space of transition matrices. Second, the non-parametric level is where we have the best theoretical and computational tools to attack[11] the issue of identification. 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.

[10] Although this is not enough to conclude, note that, under the assumption of additively separable Gumbel shocks, a direct argument already shows that the mapping cannot be polynomial (suppose it is and derive a contradiction using the dynamic program equation (DPv) in section 5).

[11] It is not a coincidence that the identification results of Kasahara and Shimotsu (2009) or Hu and Shum (2012) are obtained at the non-parametric level. Although the approach is not explicitly algebro-geometric, all the sufficient conditions in these two papers can be formulated as "if $F_s(\theta^{np}) \neq 0$ then the model is identified," where F_s is a polynomial function in transition matrix coefficients. In section 3.2, it will become apparent that these are explicit but weaker (the singular region might be too large) occurences of Theorem 1 in particular discrete models.

3.2 Discrete models have a generic identification structure

Motivated by the example of hidden Rust models, I define a *discrete model* as any mapping Φ from a parameter space Θ to the distribution of some observable such that the identification equation can be written as a (not necessarily finite) system F of polynomials:

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

This definition includes the usual "discrete models," such as discrete mixtures of discrete random variables, but also many parametric families of continuous random variables.

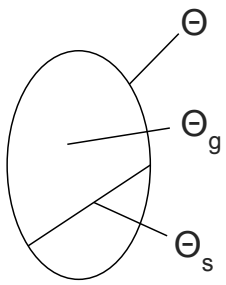
The results of this section apply to all discrete models. All proofs are in appendix section A11.

Theorem 1: Generic identification structure
Let Φ be a discrete model. There is a unique minimal singular region Θ_s *such that:*

(i) if $\theta_1^\star \notin \Theta_s$ and $\theta_2^\star \notin \Theta_s$, then $\theta_1^\star$ and $\theta_2^\star$ have the same identification structure.

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

Minimality is for the inclusion.



We call the complement of the singular region the *generic region* $\Theta_g = \Theta \setminus \Theta_s$.

The meaning of "having the same identification structure" is technical[12] and is explained in more detail in appendix section A11. The case of particular interest is when the model is generically identified. To handle cases of label-switching, suppose we know there are n_{ls} observationally equivalent parameter values, meaning $\Phi(\theta) = \Phi(\theta^\star)$ has at least n_{ls} known solutions. $n_{ls} = 1$ when there is no label-switching.

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

[12] I say that $\theta_1^\star$ and $\theta_2^\star$ have the same identification structure when the sets of θ-solutions to $\Phi(\theta) = \Phi(\theta_1^\star)$ and $\Phi(\theta) = \Phi(\theta_2^\star)$ seen as Zariski topological sets have equally many irreducible components of each topological dimension. See Definition 1 and Example 1 in appendix section A11.

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

Corollary 2:
Suppose $\Theta = [0,1]^{d_\theta}$. *Then:*

(i) Θ_g *is open dense in* Θ.

(ii) Θ_s *has Lebesgue measure zero in* Θ.

3.3 Dynamic discrete 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 structural model* as any mapping Φ from a parameter space Θ 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 discrete in the sense of section 3.2, i.e., there is a (not necessarily finite) system F_T of polynomials such that:

$$\Phi_T(\theta) = \Phi_T(\theta^\star) \qquad \Longleftrightarrow \qquad F_T(\theta, \theta^\star) = 0$$

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

$$\Phi(\theta) = \Phi(\theta^\star) \qquad \Longleftrightarrow \qquad \forall T, \quad \Phi_T(\theta) = \Phi_T(\theta^\star)$$

In particular Φ itself is a discrete model in the sense of section 3.2, with associated identification polynomial system $F = \bigcup_T F_T$. By Φ_∞ I will mean Φ.

The results of this section apply to all dynamic discrete models. All proofs are in appendix section A11.

Theorem 2: Stable identification structure
There is a smallest $T_0 < \infty$ *such that for any* $\theta^\star$, *the set of* θ*-solutions to* $\Phi_T(\theta) = \Phi_T(\theta^\star)$ *is constant for* $T_0 \leq T \leq \infty$.

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

Remark 1: Theorem 2 and Corollary 3 capture a phenomenon specific to discrete models. I give an example of a smooth dynamic model that is identified from an infinite number of marginals but not from a finite number of them. This shows that Corollary 3 cannot be generalized beyond discrete models. Consider a sequence $Y_{1:\infty}$ of 0's and 1's. Suppose $Y_{1:\infty}$ always has 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 $q_T := \mathbb{P}(Y_{1:T} = (1, \dots, 1))$. Now consider a model for $Y_{1:\infty}$ from parameter space $\Theta = [0, 1]$ as follows: $q_T(\theta) = 1$ for $0 \leq \theta \leq 1/(T+1)$, $q_T(\theta) = 0$ for $1/T \leq \theta \leq 1$ and $\theta \to q_T(\theta)$ is smooth (infinitely differentiable). Then $\theta^\star = 0$ is identified from the distribution of the whole sequence (all Y_t's are 1 with probability 1 iff $\theta = 0$) but not from any set of finite marginals (all $0 \leq \theta \leq 1/(T+1)$ are compatible with $\mathbb{P}(Y_{1:T} = (1, \dots, 1)) = 1$). □

A model can be generically identified before its identification structure stabilizes. The following fact is true for any dynamic econometric model (not only discrete ones):

Fact 1: "more marginals, more identified"
If Φ_{T_1} is globally identified at $\theta^\star$, then Φ_T is globally identified at $\theta^\star$ for every $T \geq T_1$.

Proof. A system with more equations can only have fewer solutions. □

In the particular case of dynamic discrete models, if Φ_{T_1} is generically identified, then $\Theta_g(T) \subset \Theta_g(T')$ for any $T_1 \leq T \leq T'$. If T_g is the smallest T_1 such that Φ_{T_1} is generically identified (T_g can be infinite), then T_g can be arbitrarily smaller, equal or arbitrarily bigger than T_0.

3.4 Computing identification in practice

A complete identification analysis in a given discrete model consists of computing both the minimal singular region and the generic identification structure. The algebro-geometric point of view, which was I use as a theoretical basis to obtain the identification results of sections 3.2 and 3.3, also provides computational tools. Some of these tools are surveyed in appendix section A11.4. See also appendix section A11.5, where the minimal singular region and the generic identification structure are computed in a toy model. To put it in a nutshell, there are techniques to compute exactly the minimal region, a singular region close to the minimal region, and a singular region potentially much larger than the minimal one, in decreasing order of computational complexity. These methods are very computationally intensive.

Theorem 1 suggests an alternative approach to identification analysis. If the model is identified at a randomly drawn parameter value, then it is generically identified. Symbolic or numerical methods to solve systems of polynomial equations may be used to solve the equation $\Phi(\theta) = \Phi(\theta^\star)$ for a fixed random $\theta^\star$. The computational complexity is less than in the previous paragraph, where the solution set of $\Phi(\theta) = \Phi(\theta^\star)$ must be considered jointly in θ and $\theta^\star$. Alternatively, known identification conditions, as in, e.g., Hu and Shum (2012), might be checked for a fixed random $\theta^\star$.

The polynomial structure of the model can be leveraged to check local identification. The Jacobian of the polynomial mapping from parameters to the distribution of the observables can be computed symbolically in order to check that it has full rank. However, the results of this paper do not imply that local identification implies gobal identification. What they do imply is that if $\theta^\star$ is an isolated θ-solution to $\Phi(\theta) = \Phi(\theta^\star)$, for a random draw of $\theta^\star$, then the θ-solution-set of $\Phi(\theta) = \Phi(\theta^\star)$ will have an isolated solution[13] for almost every $\theta^\star$.

4 Asymptotics

This section studies the time-series asymptotics of hidden Rust models for one individual. The results carry over easily to a fixed number of individuals and many time periods.

Section 4.1 defines the *merging time* of a Markov chain. Section 4.2 states the assumptions used in the asymptotic analysis. Section 4.3 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 A12.

4.1 Definition: Merging and the merging time

The so-called *merging* properties of Markov chains (sometimes called *weak ergodicity* properties) play an important role at various stages of the asymptotic analysis. Here I give only the definitions needed to state the assumptions of section 4.2. Appendix section A12.1 contains a motivated introduction to merging and a description of the role of merging in the asymptotic analysis of hidden Rust models.

[13] Generic local identification is not, stricly speaking, guaranteed. In extremely pathological cases it is possible that $\Phi(\theta) = \Phi(\theta^\star)$'s isolated solution may be $\theta' \neq \theta^\star$ and that $\theta^\star$ is part of a non-zero-dimensional component of the generic identification structure!

A finite-state Markov chain z is merging if there is a distribution $\mu^\diamond$, necessarily unique, such that:

$$\forall z_1, \qquad d_{TV}(\mathcal{L}(Z_t|Z_1 = z_1), \mu^\diamond) \xrightarrow[t\to\infty]{} 0$$

The ϵ-merging time $\tau_z(\epsilon)$ of a merging Markov chain is defined as follows, for $0 < \epsilon < 1$:

$$\tau_z(\epsilon) = \min\left\{t : \max_{z_1}\{d_{TV}(\mathcal{L}(Z_t|Z_1 = z_1), \mu^\diamond)\} < \epsilon\right\}$$

The absolute merging time, or simply the *merging time*, is (the definition is motivated in appendix section A12.1):

$$\tau_z = \inf_{0\leq\epsilon<1} \frac{\tau_z(\epsilon/2)}{(1-\epsilon)^2}$$

A recurrent chain is merging if and only if it is aperiodic. Remember that the recurrent aperiodic chains are the "nice" chains with fully supported unique stationary distributions and no periodic behavior at equilibrium.

4.2 Assumptions

$z_t = (x_t, s_t, a_t)$ is generated by an arbitrary initial distribution $\mu^\star$ along with transition matrices $P(\theta^\star)$, $Q(\theta^\star)$ and $\Pi(\theta^\star)$. 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 log-likelihood[14] 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 only:

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

The theorems of section 4.3 show that the model does have good time-series asymptotic properties, and section A12.3 explains how the log-likelihood can be asymptotically approximated by an ergodic sum using an "infinite-past" strategy.

[14] I use the conditional log-likelihood for convenience. This is without consequence in the time-series context.

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 $\theta \to \Pi(\theta)$ are three times continuously differentiable.

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

Assumption (A3): Identification[15]
If $\theta, \theta' \in \Theta$ induce the same joint stationary[16] 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.

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 $\theta^\star$.

(A2) together with (A1) has the important consequence that z is *uniformly* merging over Θ, in the sense that the merging time $\tau_z(\theta)$ is uniformly bounded over Θ (see appendix section A12.2). Recurrence in (A2) is not used for any other purpose than bounding uniformly $\tau_z(\theta)$. If needed, (A2) could be relaxed to z being merging without being necessarily recurrent, as long as z remains uniformly merging. It is even possible that merging and compactness are already sufficient conditions for uniform merging, without recurrence — although I do not know a proof of this.

Along with smoothness and compactness (assumption (A1)), assumption (A4) implies that the coefficients of Q are uniformly lower bounded by some $\underline{q} > 0$. This is used in the proofs of the limit theorems for the log-likelihood, score and information under stationarity (in appendix section A12.4). Assumption (A4) could probably be relaxed to weaker mixing properties with some work; see comments in section 7.

[15] In the context of section 3, assumption (A3) says that the model is generically identified (in the sense of Corollary 1) and that Θ is a subset of the generic region.

[16] (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 $\Pi(\theta)$ induces a stationary distribution on the Markov chain $Z_t = (X_t, Y_t)$. Note that the identification assumption bears on the stationary distribution, although the true distribution of Y_t is not necessarily the stationary one. This is enough, thanks to the fact that Y_t will merge to its stationary distribution as T increases.

4.3 Asymptotic theorems

I state three asymptotic theorems and I comment on their implications. Theorem 3 shows 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. Theorem 4 shows that the maximum likelihood estimator is well-behaved asymptotically. Theorem 5 shows that Bayesian posteriors are also well-behaved asymptotically, from a frequentist perspective.

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 $h_T \xrightarrow{P_{\theta^\star}} h$:

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

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:

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

Let q_T be the Bayesian posterior distribution. Under (A1) to (A5):

Theorem 5: Bernstein-von Mises theorem

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

Proof. See section 4.4 for an outline and appendix section A12 for complete proofs. □

Local asymptotic normality (Theorem 3) is before all[17] a regularity property. The typical locally asymptotically normal models are the smooth, independent and identically distributed models. In a times-series context, non-unit-root autoregressive models are locally asymptotically normal, but unit-root models are not locally asymptotically normal. An important consequence of Theorem 3 is that the maximum likelihood estimator is statistically efficient in the strong sense of the Le Cam convolution theorem, meaning that the maximum likelihood estimator is optimal against a large class of loss functions. This is much stronger than having the smallest asymptotic variance among all the asymptotically normal estimators. Theorem 3 is also a key technical step toward Theorem 4 and Theorem 5 (see the proof outline, section 4.4). See van der Vaart (1998) for more information about the local asymptotic normality property.

Bayesian posteriors can be thought of as functions from the data to the distributions on Θ, as much as standard estimators are functions from the data to the *points* in Θ. From this perspective, their asymptotic properties can be studied exactly similar to the way we study the asymptotic properties of the maximum likelihood or other classical estimators. This is what the Bernstein–von Mises theorem (Theorem 5) does. Theorem 5 says that Bayesian posteriors will be asymptotically Gaussian, centered at the maximum likelihood estimator and with a variance-covariance matrix $1/T$ times the asymptotic variance-covariance matrix of the maximum likelihood estimator. This is a well-known phenomenon in smooth independent and identically distributed models. As a consequence of Theorem 5, posterior statistics such as the mean,[18] mode or median are asympotically equivalent to the maximum likelihood estimator. Consistent confidence intervals can be obtained from computing the Bayesian posterior variance. The fact that the asymptotic behavior of Bayesian posteriors does not depend on the prior shows that the influence of the prior fades away with time.

[17] At a literal level, Theorem 3 says that the log-likelihood admits an asymptotic stochastic quadratic approximation around the true parameter value, in a suitably uniform sense apparent in the fact that the sequence h_n can be data-driven (random). Uniformity is used to show asymptotic normality of the maximum likelihood estimator. Local asymptotic normality is obtained via a Taylor expansion of the log-likelihood, similar to the smooth independent and identically distributed theory. The difficulty is to show that the relevant limit theorems hold in the time-series context of hidden Rust models.

[18] Because Θ is compact, all posterior moments exist.

4.4 Outline of the proofs

This section describes the five steps involved in proving Theorem 3, Theorem 4 and Theorem 5. The proofs themselves are given in full in appendix section A12.

First, I prove three limit theorems assuming the data is stationary but allowing for a slightly misspecified likelihood computed with a wrong initial distribution μ, instead of the stationary $\mu^{\diamond}$. The three limit theorems are a uniform law of large numbers for the log-likelihood, a central limit theorem for the score and a uniform law of large numbers for the observed information. This first step is where most of the technical work is done. The general "infinite-past" strategy I use is described in appendix section A12.3. A point of particular interest is the use of a vector ergodic theorem, valid for random vectors taking value in seperable metric spaces, to obtain a uniform law of large numbers directly rather than through pointwise limits and stochastic equicontinuity type arguments.

Second, I extend these limit theorems to the case where the data is nonstationary. I use an argument based on the merging properties of the chain. As far as I know, this argument is new. It can be used to extend stationary limit theorems to nonstationary ones in general Markovian contexts. I motivate this new argument in detail in appendix section A12.5.1.

Third, uniform local asymptotic normality of the model (Theorem 3) follows from the central limit theorem for the score and the uniform law of large numbers for the observed information.

Fourth, I show that the maximum likelihood estimator is consistent. Asymptotic normality of the maximum likelihood estimator follows from the consistency and uniform local asymptotic normality of hidden Rust models (Theorem 4).

Fifth, I prove Theorem 5 by checking that the assumptions of the general weakly dependent Bernstein–von Mises theorem from Connault (2014) are verified. The main assumptions of that theorem are local asymptotic normality[19] and the existence of a uniformly consistent estimator. Stable identification (Theorem 2) implies that hidden Rust models are identified from the marginals at some finite horizon T_0. I show that a frequency estimator is uniformly consistent for these marginals using concentration inequalities for Markov and hidden Markov chains recently derived in Paulin (2014).

[19] One major motivation in writing Connault (2014) was to obtain a Bernstein–von Mises theorem valid under little more than local asymptotic normality, in the spirit of Le Cam's (1986) theorem for smooth independent and identically distributed models (see also van der Vaart (1998)).

5 Estimation

This section shows how to efficiently compute the maximum likelihood estimator and Bayesian posteriors in hidden Rust models. I assume that choices are made based on flow utilities discounted with additively separable Gumbel shocks, as in Rust (1987). The high tractability of hidden Rust models is a major advantage compared to other models of unobserved persistence for dynamic discrete choice. The structure of the likelihood reflects the 2-level structure of hidden Rust models:

$$\text{STRUCTURAL PARAMETERS} \rightarrow \text{TRANSITION MATRICES}$$
$$\rightarrow \text{DISTRIBUTION OF THE OBSERVABLES}$$

The first level requires solving a dynamic program exactly as in a classical Rust model. At the second level, the unobserved state can be marginalized out of the likelihood efficiently, thanks to a recursive algorithm similar in spirit to the Kalman filter.

In this section, I relax the state dynamic assumption made in section 2. I allow for the general dynamics usually found in dynamic discrete choice models, namely the state $k_t = (x_t, s_t)$ is Markov conditional on choice a_t. This includes as a particular case the dynamics considered in Arcidiacono and Miller (2011). I call M_a the conditional state transition matrix. Under the dynamic assumptions of section 2, using a reverse lexicographical order on the state $k = (x, s) = (1, 1), (2, 1)$, etc., M_a has the following expression:

$$M_a = \Pi_a \otimes Q$$

The identification results of section 3 apply to the more general dynamics, but not the time-series asymptotic results of section 4, although I expect them to remain true under the more general dynamics; see comments in section 7. Of course, in the case of many independent and identically distributed individuals and a fixed time horizon, usually considered in the literature, the dynamic assumption does not matter and the maximum likelihood estimator is consistent, asymptotically normal and statistically efficient under identification.

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 follow in section 5.3. 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 5.4).

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: as explained in the introduction, from the agent's point of view, it does not matter which variables the econometrician observes. 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:

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

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:

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

Thanks to a_t and k_t being discrete, 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. Let us 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:

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

The decision-making rule (1) implies that choices a_t are made with constant conditional 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, the second stage being the marginalization of the unobserved state component. $u \to v$ is the computationallly expensive part of computing $\theta_u \to u \to v \to P$ and is usually solved by fixed-point iteration, using the fixed-point structure of the dynamic program (DPv).

There is an alternative route $\theta_u \to u \to V \to P$ to computing $\theta_u \to P$. Indeed, V is the unique solution to the following equation, which I call the dynamic program[20] parameterized in V:

$$\sum_a e^{u_a + (\beta M_a - I)V} = 1 \tag{DPV}$$

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

The conditional choice probabilities are then given by $P_a = e^{u_a + (\beta M_a - I)V}$.

[20] By abuse of language; this is not a Bellman equation.

5.2 Evaluating the likelihood

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

To solve the dynamic program, i.e., to compute P at a given value of θ_u, I recommend solving for V in (DPV) seen as a system of nonlinear equations:

$$F(V) = \sum_a e^{u_a + (\beta M_a - I)V} - 1 = 0$$

Then $P_a = e^{u_a + (\beta M_a - I)V}$. The Jacobian is easily computed. If $\text{diagm}(W)$ is the diagonal matrix whose diagonal is the vector W:

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

An off-the-shelf numerical solver can be used. Convergence will usually be faster than with the usual fixed-point iteration method, 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, has never been used in the dynamic discrete choice context. This approach can be used in classical Rust models, since the dynamic program is identical.

A major advantage of the nonlinear system point of view is that it 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) (see section 6 and Figure 9 in appendix section A14). Not only will evaluating F and $\dot{F}$ involve one sparse matrix multiplication[21], 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 ($d_k = 756$, $d_a = 2$), it takes around 10 milliseconds to solve $F(V) = 0$ with numerical precision 10^{-8}.

The nonlinear system point of view can also be applied to (DPv), although (DPV) is a sytem of size d_k and (DPv) is a system of size $d_k d_a$, or $d_k(d_a - 1)$ when solved in $\Delta v_a = v_a - v_1$, $a > 1$ (the Δv_a's are enough to compute P). A drawback of the nonlinear system point of

[21] Rust (1987) mentions, in a nested fixed-point context, the possibility of using "special band-matrix linear algebra routines" when applicable. Even greater efficiency can be expected from leveraging the specific sparsity structure, such as the band structure in Rust's (1987) example or the Kronecker structure in a hidden Rust model with $M_a = \Pi_a \otimes Q$. I recommend going beyond the widely available sparse matrix routines only if it is clear that solving the dynamic program is the computational bottleneck.

view is that convergence is not guaranteed as it is for the iterated fixed-point algorithm.

As explained in section 2.3 (see also an example in appendix section A10), finite-horizon models can be cast as stationary infinite-horizon models. For finite-horizon models, backward-solving algorithms can usually be written to compute P faster than by solving (DPV) in the equivalent stationary infinite-horizon model. In the empirical model of section 6, a backward-solving algorithm is 10 times faster. Writing backward-solving algorithms can require a substantial amount of work. There is a trade-off between using a fast custom backward solving algorithm or an off-the-shelf slightly slower solver. Again, my recommendation is to turn to backward-solving only when it is clear that the dynamic program stage is the computational bottleneck.

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:

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

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:

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

In terms of transition matrices, $H_{t+1,xx'} = M_{a_t,(x,s_t)(x',s_{t+1})} P_{(x',s_{t+1})a_{t+1}}$ under general dynamics and $H_{t+1,xx'} = Q_{xx'} \Pi_{a_t,s_t s_{t+1}} P_{(x',s_{t+1})a_{t+1}}$ under the dynamics of section 2.

Proof.

$$\begin{aligned}
\mathbb{P}((s,a)_{1:t+1}, x_{t+1}) &= \sum_{x_t} \mathbb{P}((s,a)_{1:t+1}, x_{t+1}, x_t) \\
&= \sum_{x_t} \mathbb{P}((x,s,a)_{t+1}|(s,a)_{1:t}, x_t) \mathbb{P}((s,a)_{1:t}, x_t) \\
&= \sum_{x_t} \mathbb{P}((x,s,a)_{t+1}|(x,s,a)_t) \mathbb{P}((s,a)_{1:t}, x_t) \qquad \text{by the Markov property}
\end{aligned}$$

□

See Zucchini and MacDonald (2009) for uses of the discrete filter in hidden Markov models.

π_1 is initialized according to an initial distribution assumption. As proved in section 4, it does not matter which initial distribution is used in a time-series context. In panel-data asymptotics, however, the influence of the initial distribution will not fade away with time and a misspecified initial distribution can induce a failure of consistency.

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 M 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 A13.

5.3 Tractable estimation of hidden Rust models

The maximum likelihood estimator can be computed by an inner-outer algorithm. The "inner loop" is the evaluation of the likelihood at a given value of the structural parameter θ, as described in section 5.2. The "outer loop" is optimization over the parameter θ. Gradient-free or numerical-gradient methods must be used, since the discrete filter of section 5.2 allows for efficient evaluation of the likelihood but not of its Jacobian. 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. A Markov chain Monte Carlo algorithm can be used to obtain a numerical approximation of the Bayesian posterior. For example, under the dynamic assumptions of section 2, if Π is known and Q is parameterized by its coefficients, a Gibbs sampler can be used. There are two Gibbs blocks, for each of θ_u and Q. The θ_u block can be updated by a Metropolis-within-Gibbs step with a Gaussian proposal. The Q block can be updated by a Metropolis-Hastings-within-Gibbs step with a Dirichlet proposal for each

row of Q.

5.4 Other estimation approaches for hidden Rust models

The estimators I recommend in section 5.3 belong to the tradition of Rust's nested fixed-point estimator. Other approaches have proved useful in classical Rust models. 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, is also applicable.

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

First, the essence of a 2-step approach consists of 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.

Second, the essence of a 2-step approach consists of projecting a set of observed marginals to a structural parameter space; basically, a method of moments in a parametric context. The stable identification theorem (Theorem 2) implies that there is always a finite identifying set of marginals. In theory, such an approach would be possible. 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 $L_T^{np}(M, P)$ be the "non-

parametric" likelihood parameterized in transition matrices. The following constrained optimization program computes the maximum-likelihood estimator:

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

The discrete filter is used at each iteration to evaluate $L_T^{np}(M,P)$. $F(\theta,M,P)=0$ can be any constraint that expresses the condition "where M and P are the transition matrices compatible with θ." For instance, a constraint could be used based on (DPv) or (DPV). The crucial property that $F(\theta,M,P)=0$ must have is uniqueness of the (M,P) solution for each θ. In experiments, the constrained optimization approach did not perform as well as the direct approach of section 5.3. However, see section 7 for applications in models with computationally expensive or multiple-solution dynamic programs.

Arcidiacono and Miller's (2011) section 5 suggests 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 the constrained EM algorithm is that it computes the maximum likelihood estimator when it converges. However, it is not clear that it keeps the increasing-likelihood property of the original EM algorithm. It would be interesting to study the convergence property of the algorithm. The discrete filter of section 5.2 could speed up some of the moves in the constrained EM algorithm. Similar to plain constrained optimization approaches, one must be careful to use a constraint expressing the "where M and P are the transition matrices compatible with θ" condition. If the constraint has extraneous solutions, a highest-likelihood selection mechanism will not rule out selecting transition matrices that do not belong to the model.

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 wage was changed from a flat wage of 1000 rupees[22] to a fixed plus variable structure of 500 rupees plus 50 rupees for every day of work beyond ten days. 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. Other steps in Duflo et al.'s (2012) empirical study 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, as in section 5.1, 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:

$$u(s_t, a_t) = u_l \cdot 1[a_t = 1] + u_w \cdot w(s_t, a_t) \tag{3}$$

$w(\cdot)$ is the wage, 500 rupees plus 50 rupees for each day after ten days, paid on the last day of the month. The authors 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 5 below for a 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 as in (3). The second family (models

[22] At the time of the experiment, 1000 rupees were $23 at the real exchange rate, or about $160 at purchasing power parity.

VI, VII and VIII, or "shifter" models) are models of *observed* persistence:[23] 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 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.[24] 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 9 in appendix section A14 pictures the structure of the conditional state transition matrices. Figure 4 pictures the evolution of the state in a typical month.

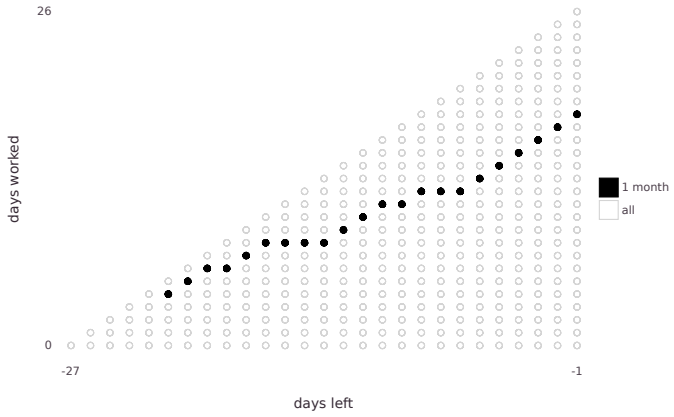


Figure 4: Teacher 22, month 16.

[23] 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.

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

I estimate hidden Rust models with $d_x = 2$, 3, 4 or 7 unobserved states. The dynamic assumptions are as in section 2.1, meaning the unobserved states have independent Markov dynamics. 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 case $d_x = 1$ is simply the baseline model (3).

Figure 5 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) and the shifter model (three parameters) fit the data better than a fixed-effects model with 108 parameters demonstrates the importance of serial correlation.

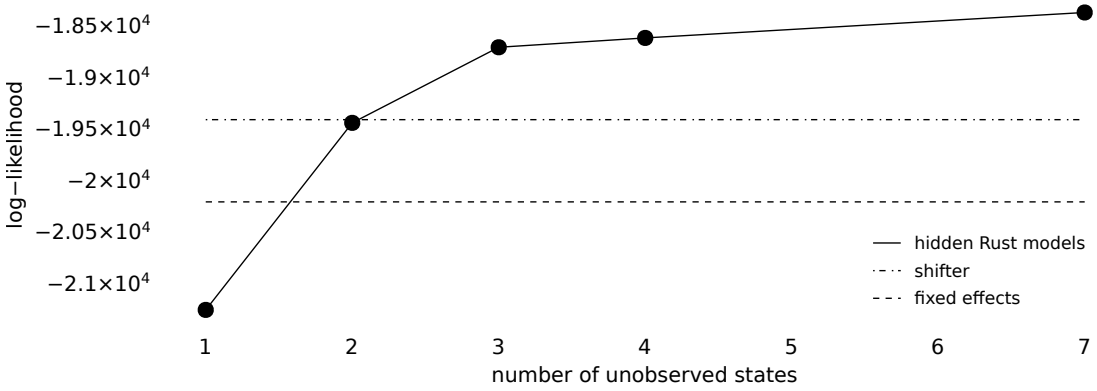


Figure 5

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

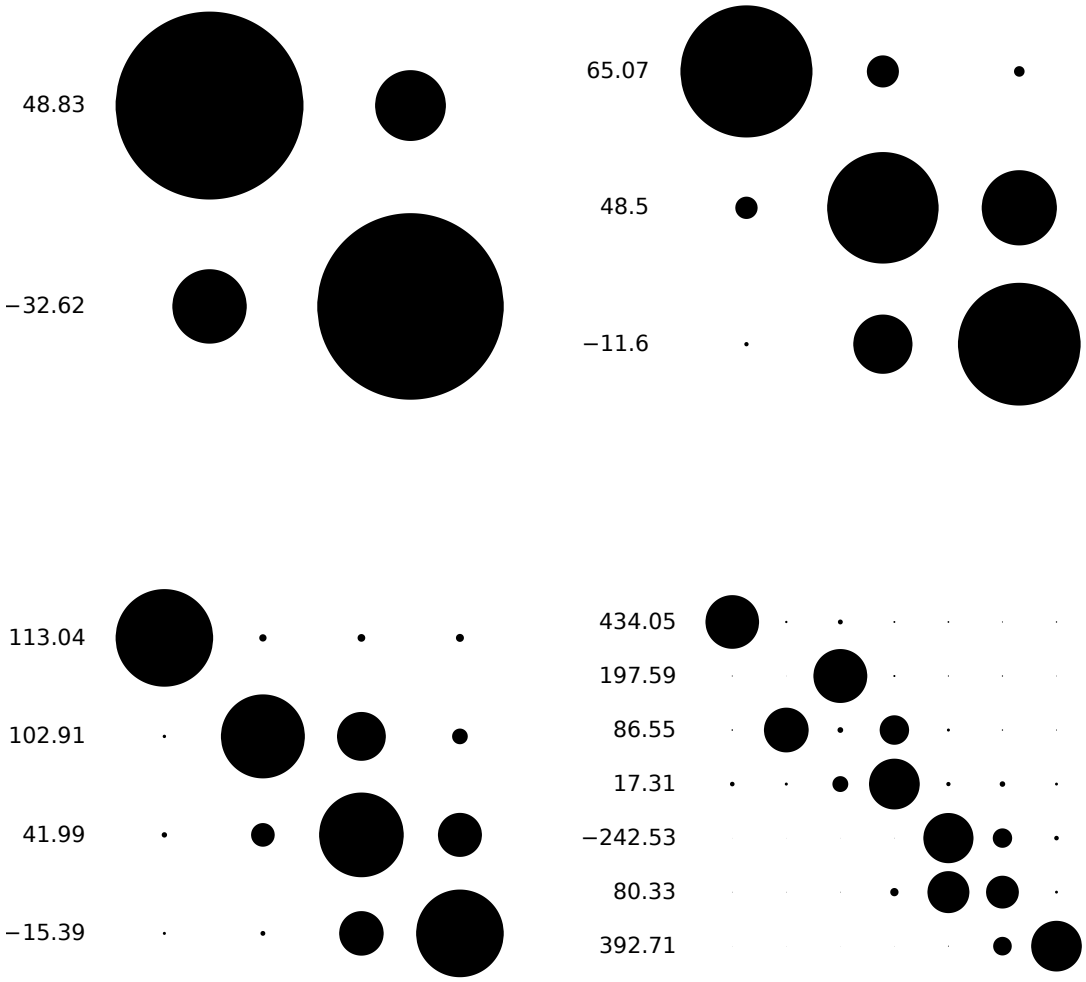


Figure 6: Hidden Rust models estimated for 2, 3, 4 and 7 unobserved states.

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 A14. 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.

The likelihood always selects a clustered dynamic heterogeneity 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 very rare transitions between clusters but relatively frequent switching within the cluster. The fact that the corresponding leisure utilities are not ranked monotonically indicates that the unobserved dynamic heterogeneity structure goes beyond unobserved persistence. The likelihood (Figure 5) also tells us that the clustered structure is *not* explained by pure static heterogeneity (fixed effects): the high and low frequency transitions are important in explaining the data.

In all cases, one unobserved state is spent on a negative leisure utility state. Negative leisure utility can be interpreted as a taste for work or fear of being fired. It is an important feature of the data. 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.

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). This is mostly due to the estimated utility of money being close to zero. This may be an artifact due to overfitting the variability of the data to increasingly open unobserved dynamics. The number of statistical parameters grows like $d_x^2 + 1$ with the number of unobserved states (d_x leisure utilities + 1 money utility + $d_x(d_x - 1)$ transition probabilities). 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 A14 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.[25]

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)	94% (0.044)	5% (0.044)	1% (0.002)
							3% (0.015)	67% (0.031)	30% (0.033)
							0% (0.0006)	19% (0.043)	81% (0.043)

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

[25] In this hidden Rust model, computing the maximum likelihood with numerical precision 10e-8 takes typically between 2 and 4 seconds on an average 2013 desktop computer.

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 2.3, 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.

Specific to hidden Rust models, a maximum likelihood path for the unobserved state variable can 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 can be 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). See how the Viterbi algorithm picks up outliers in a larger sample of 60 teachers in Figure 10, appendix section A14.

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 tractable, both in terms of maximum likelihood estimation and counterfactual compuations.

7 More general models

Remember the 2-level structure of a hidden Rust model:

$$\text{STRUCTURAL PARAMETERS} \rightarrow \text{TRANSITION MATRICES}$$
$$\rightarrow \text{DISTRIBUTION OF THE OBSERVABLES}$$

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 zero transition probabilities at the transition matrix level. I used little beyond a reasonable degree of smoothness for the mapping from 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 7.

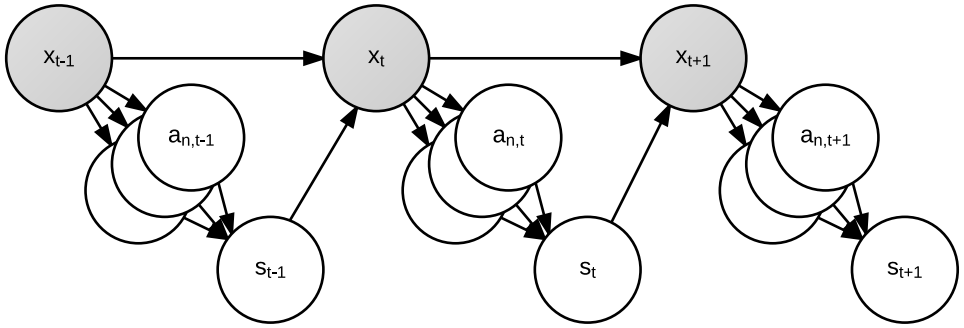


Figure 7: Graphical model for a dynamic game.

Resuls on identification, asymptotics and estimation will all apply to more general hidden structural models, to some extent.

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 for any hidden structural model with autoregressive hidden Markov dynamics, under assumption (A4) that the transition matrix for the unobserved state has full support. I expect the results to hold beyond assumption (A4) and also beyond autoregressive hidden Markov dynamics. The dynamic game model represented in Figure 7 does not have autoregressive hidden Markov dynamics. I believe the argument in appendix section A12.4 may be extended beyond assumption (A4) with a little bit of work, but the more interesting question of isolating sufficient conditions for general dynamic discrete models to have good time-series asymptotic properties will probably require new arguments

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 is specific to the economic model. Econometric tractability is an incentive for designing theoretical models with good computational properties at this level. See, for instance, the model of industry dynamics in Abbring et al. (2013).

8 Conclusion

Hidden Rust models provide a solution for adding unobserved persistence features to the dynamic discrete choice framework while maintaining structurality and tractability. 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 weak 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. For estimation, various model assumptions may be relaxed, in some cases using existing tools from the dynamic discrete choice literature. More general dynamic discrete models, such as dynamic games, may be considered. There are also interesting issues I did not touch upon. For example, it would be interesting to investigate the finite-sample statistical properties of hidden Rust models, as well as the weak identification features induced by failure of identification on the singular region of a generically identified model.

The most interesting challenges for hidden Rust models are empirical. To mention a specific example, one use of structural models is to make predictions. A hidden Rust model leaves enough room for economic agents to be rational (the observed state enters a dynamic optimization program) but not *too* rational (the unobserved state can account for a wide range of behaviors). Can this give hidden Rust models an edge over non-economic statistical models in some situations? For individuals? For firms? Should the decision-making model be modified, and, if yes, how so? These are all open questions that empirical studies will help answer.

November 18, 2014

APPENDICES

A10 Appendix for section 2: Hidden Rust models

I show how a finite-horizon Rust model can be cast as a stationary infinite-horizon Rust model in a toy example. See section 5.1 for notation.

There are two fishing spots: Close and Far. Bob the fisherman usually fishes at Close but he keeps an eye on his two private beacons, which tell him exactly how much fish there is at both spots. Most of the times there is approximately as much fish at Close as at Far ($k_t = 1$), but sometimes a particularly big shoal shows up at Far ($k_t = 2$). Bob's average utility from fishing at Close ($a_t = 1$) is normalized to zero, and he derives $u_s > 0$ utility on average from fishing at Far ($a_t = 2$) when the shoal is there. There is also a disutility $-u_f$, $u_f > 0$, for taking his boat to Far.

$$u = \begin{pmatrix} 0 & -u_f \\ 0 & u_s - u_f \end{pmatrix}$$

Even when there is no shoal at Far, there might be sufficiently more fish at Far than at Close for Bob to take his boat to Far. Reciprocally, if a small shoal shows up, Bob might want to stay at Close, especially if he is afraid of frightening the shoal away for the next time he goes fishing. In Bob's country, the fishing season lasts two months. If the shoal shows up in the first month and Bob goes to Far, there is little chance he will see the shoal again. Otherwise the chances are somewhat better.

$$M_1 = \begin{pmatrix} 2/3 & 1/3 \\ 2/3 & 1/3 \end{pmatrix} \qquad M_2 = \begin{pmatrix} 2/3 & 1/3 \\ 5/6 & 1/6 \end{pmatrix}$$

One way to compute Bob's discounted utilities is by backward solving. In the second month the discounted utilities v^2 are simply the flow utilities.

With $u_s = 1$ and $u_f = 0.2$, we find that in the first month $v^1 = \begin{pmatrix} 0.394 & 0.194 \\ 0.394 & 1.146 \end{pmatrix}$. The corresponding conditional choice probabilities in the first month and in the second month

are:

$$P^1 = \begin{pmatrix} 0.550 & 0.450 \\ 0.320 & 0.680 \end{pmatrix} \qquad P^2 = \begin{pmatrix} 0.550 & 0.450 \\ 0.310 & 0.690 \end{pmatrix}$$

There is a small dynamic effect: Bob goes after the shoal more frequently in the second month (69.0% rather than 68.0%).

Now suppose that Bob is an immortal fisherman with infinitely many fishing seasons ahead of him. At the beginning of a new fishing season, the shoal shows up with probability 1/2. Increase the state variable by including the current month so that this is now a stationary infinite-horizon Rust model with:

$$k = \begin{pmatrix} \text{first month, no shoal} \\ \text{first month, shoal} \\ \text{second month, no shoal} \\ \text{second month, shoal} \end{pmatrix} \qquad u = \begin{pmatrix} 0 & -u_f \\ 0 & u_s - u_f \\ 0 & -u_f \\ 0 & u_s - u_f \end{pmatrix}$$

$$M_1 = \begin{pmatrix} 0 & 0 & 2/3 & 1/3 \\ 0 & 0 & 2/3 & 1/3 \\ 1/2 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \end{pmatrix} \qquad M_2 = \begin{pmatrix} 0 & 0 & 2/3 & 1/3 \\ 0 & 0 & 5/6 & 1/6 \\ 1/2 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \end{pmatrix}$$

Section 5.2 explains how to compute the corresponding stationary discounted utilities. Note how M_1 and M_2 have entire blocks of zeros — the techniques of section 5.2 are particularly fast when this is the case. We find:

$$v = \begin{pmatrix} 0.816 & 0.616 \\ 0.816 & 1.568 \\ 0.842 & 0.642 \\ 0.842 & 1.642 \end{pmatrix} \qquad P = \begin{pmatrix} 0.550 & 0.450 \\ 0.320 & 0.680 \\ 0.550 & 0.450 \\ 0.310 & 0.690 \end{pmatrix}$$

While the value matrix v is different from $\begin{pmatrix} v^1 \\ v^2 \end{pmatrix}$, the conditional choice probability matrix

P is exactly $\begin{pmatrix} P^1 \\ P^2 \end{pmatrix}$. This is because the expected values are all the same for the next fishing seasons beyond the second month. When deciding a course of action, only the current season matters. Said differently, the choices made during the current fishing season have no impact on the next ones.

Of course, both models are observationally equivalent and one could compute P^1 and P^2 in the finite-horizon model by pretending that it is only the first of infinitely many fishing seasons. Although this is not the case for Bob, the flow utilities could also perfectly depend on the month. In particular, one could discount differently within a season and between seasons, or not discount within a season as is sometimes done in finite-horizon programs.

A11 Appendix for section 3: Identification

As explained in section 3.1, the question of identification can be cast as a polynomial system in discrete models. The mathematical field that studies the zero-sets of polynomials is known as algebraic geometry. I use algebraic geometry to prove the identification results of section 3. Theorem 1 is proved in section A11.2, and the other results are proved in section A11.3. Algebraic geometry also provides some computational tools that can be used to compute identification in particular models. I mention some of them in section A11.4. In section A11.5, I compute the minimal singular region and the generic identification structure in a toy discrete model, using computational algebraic geometry tools.

In section A11.1, I state the standard definitions and facts from algebraic geometry that I need in the other sections. This includes the definition of two parameter values "having the same identification structure" (Definition 1).

The fundamental reason why discrete 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).

A11.1 Relevant notions and facts from algebraic geometry

Algebraic geometry is the study of some geometric objects — such as zero-sets of polynomials — by algebraic means, as well as the study of algebraic objects — such as rings — by

geometric means. This subsection introduces the concepts and states the facts relevant to section 3.

Section A11.1.1 covers geometric aspects. The generic identification structure for econometric models, Theorem 1 and corollaries in section 3.2, is obtained mostly through geometric techniques. Section A11.1.2 covers algebraic aspects. The stable identification structure for dynamic econometric models, Theorem 2 and corollaries in section 3.3, is obtained mostly through algebraic techniques. Section A11.1.3 brings together geometric and algebraic aspects.

Appendix section A11.5 carries identification analysis in a toy econometric model. It provides a down-to-earth illustration of the concepts introduced in this section.

See Reid (1988) for an introduction to algebraic geometry. All definitions and facts in this section are standard. Those facts not proven in Reid (1988), in particular, some of the properties of the Zariski topology, can be found in the first chapter of Görtz and Wedhorn (2010).

A11.1.1 The Zariski topology

This subsection covers geometric aspects. The generic identification structure for econometric models, Theorem 1 and corollaries in section 3.2, is obtained mostly through geometric techniques. In particular, the "generic identification structure" of Theorem 1 is to be understood in the sense of Fact 2 below.

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 $\mathbb{A}$. By a Zariski topological space I mean a subset of an affine space with its induced Zariski topology.

By the dimension of a Zariski topological space, I will simply mean its topological dimen-

sion. 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 2: 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 $V_{m'}$.

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

Definition 1: $\theta_1^\star$ and $\theta_2^\star$ have the same identification structure
$\theta_1^\star$ and $\theta_2^\star$ have the same identification structure if $V_\theta\left(F\left(\cdot,\theta_1^\star\right)\right)$ and $V_\theta\left(F\left(\cdot,\theta_2^\star\right)\right)$, seen as Zariski topological spaces, have equally many irreducible components of each dimension.

When the model is identified at $\theta^\star$, $V_\theta\left(F\left(\cdot,\theta^\star\right)\right)$ has 1 "irreducible component of dimension zero" (i.e. 1 point, namely $\theta^\star$) 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 $W_i = V_\theta\left(F\left(\cdot,\theta_i^\star\right)\right)$ for $1 \leq i \leq 5$. Suppose the W_i's have the following decompositions in irreducible components (all V_j's of dimension 2):

$$W_1 = \{\theta_1^\star\} \qquad W_2 = \{\theta_2^\star\} \cup \{\theta_3^\star\} \qquad W_3 = \{\theta_2^\star\} \cup \{\theta_3^\star\} \qquad W_4 = V_4$$
$$W_5 = \{\theta_5^\star\} \cup V_5 \qquad W_6 = \{\theta_6^\star\} \cup V_6 \qquad W_7 = \{\theta_7^\star\} \cup V_7 \cup V_7' \qquad W_8 = \{\theta \neq \theta_8^\star\} \cup V_8$$

The model is (globally) identified at $\theta_1^\star$. $\theta_2^\star$ and $\theta_3^\star$ are observationally equivalent and the model is locally identified at $\theta_2^\star$ and $\theta_3^\star$. If we were anticipating a label-switching feature between $\theta_2^\star$ and $\theta_3^\star$, we could consider the model to be globally identified at $\theta_2^\star$ and $\theta_3^\star$. The model is not locally identified at $\theta_4^\star$. The model is locally identified at $\theta_5^\star$, $\theta_6^\star$ and $\theta_7^\star$, but not globally identified. $\theta_5^\star$ and $\theta_6^\star$ have the same identification structure but not $\theta_7^\star$. $\theta_8^\star$ is somewhat of a pathological case with an isolated solution that is not $\theta_8^\star$ ($\theta_8^\star$ must belong to V_8). The model is not locally identified at $\theta_8^\star$, but $\theta_8^\star$ has the same identification structure as $\theta_5^\star$ and $\theta_6^\star$. □

The Zariski topology has unusual properties. Here is one of them (compare with $]0,1[$ in $\mathbb{R}$ with the Euclidean topology):

Fact 3:
Any non-empty open set is dense in an irreducible Zariski topological space.

The following fact compares the Zariski and the Euclidean topologies if $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$.

Fact 4:
If $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$, any Zariski open (resp. closed) set is Euclidean open (resp. closed). Non-empty open sets of $\mathbb{A}^n(\mathbb{C})$ are dense in $\mathbb{C}^n$ and $\mathbb{R}^n$. Non-empty open sets of $\mathbb{A}^n(\mathbb{R})$ are dense in $\mathbb{R}^n$. Proper closed subsets of $\mathbb{A}^n(\mathbb{C})$ have Lebesgue measure 0 in $\mathbb{C}^n$ and $\mathbb{R}^n$. Proper closed subsets of $\mathbb{A}^n(\mathbb{R})$ have Lebesgue measure 0 in $\mathbb{R}^n$. On $\mathbb{R}^n$, the subspace topology inherited from $\mathbb{A}^n(\mathbb{C})$ and the $\mathbb{A}^n(\mathbb{R})$ topology coincide.

A word about terminology: Zariksi closed subsets of $\mathbb{A}$ and irreducible Zariski closed subsets of $\mathbb{A}$ are sometimes called affine algebraic sets and irreducible affine algebraic sets, respectively, or sometimes affine algebraic sets and affine varieties, respectively, or sometimes affine varieties and irreducible affine varieties, respectively.

A11.1.2 Polynomial ideals

This subsection covers algebraic aspects. The Noetherianity of $\mathbb{K}[z]$ (see below) is key to proving Theorem 2 in section 3.3, about stable identification structure for dynamic econometric models. The correspondance between geometric and algebraic objects, as detailed in the next subsection, section A11.1.3, is also needed to follow the proofs of section 3.3.

$\mathbb{K}[z]$ has a ring structure. The ideals of $\mathbb{K}[z]$ are called the polynomial ideals. Remember that a subset I of a ring R is an ideal of R if it is a subgroup for the addition and if is absorbing for the multiplication, meaning for any $f \in I$ and $g \in R$, $fg \in I$. An arbitrary family of elements $F = (f_i)_i$ generates an ideal, written $\langle F \rangle$, as follows:

$$\langle F \rangle = \{ f_{i_1} g_1 + \ldots + f_{i_M} g_M \mid M \in \mathbb{N}, g_m \in R \}$$

For instance in $\mathbb{R}[z_1]$, $\langle 1 + z_1^2 \rangle$ is all the polynomials of which $(1 + z_1^2)$ is a factor. Note that $\{0\} = \langle 0 \rangle$ and $\mathbb{K}[z] = \langle 1 \rangle$ are ideals and that for any ideal I of $\mathbb{K}[z]$, $\langle 0 \rangle \subset I \subset \langle 1 \rangle$.

The Hilbert basis theorem says that $\mathbb{K}[z]$ is a Noetherian ring. A ring R is Noetherian if and only if any ideal I of R is finitely generated, meaning there is $f_1, \ldots, f_m \in R$ such that $I = \langle f_1, \ldots, f_m \rangle$. We will use the apparently slightly stronger but in fact equivalent result:

Fact 5:
A ring R is Noetherian if and only if for any (not necessarily finite) family $f_i \in R$ there is a finite subfamily $(f_{i_1}, \ldots, f_{i_m})$ such that:

$$\langle f_i \rangle = \langle f_{i_1}, \ldots, f_{i_m} \rangle$$

Two types of ideals play an important role in the theory of polynomial rings: radical ideals and prime ideals. Being prime is a stronger property than being radical: prime ideals are radical.

The radical $\sqrt{I}$ of an ideal I is $\sqrt{I} = \{f \mid \exists n : \ f^n \in I\}$. Radicals remove multiplicities: for instance, $\sqrt{\langle (1+z_1)^2 \rangle} = \langle 1 + z_1 \rangle$. Of course, in general $\sqrt{I} \subset I$. I is radical if $\sqrt{I} = I$.

I is prime if: for any $f_1, f_2 \in R$, if $f_1 f_2 \in I$ then $f_1 \in I$ or $f_2 \in I$.

In a Noetherian ring such as $\mathbb{K}[z]$, any radical ideal I admits a unique decomposition in prime ideals (up to renumbering and under the condition that no I_m is included in another $I_{m'}$):

$$I = I_1 \cap \ldots \cap I_M$$

A11.1.3 Ideals and Zariski closed sets

We can now bring geometric aspects and algebraic aspects together.

For any set F of polynomials, the zero-set $V(F)$ of F is equal to the zero-set $V\left(\sqrt{\langle F \rangle}\right)$ of $\sqrt{\langle F \rangle}$. The fact that $V(F)$ is equal to $V(\langle F \rangle)$ is easy to see, and the passage to the radical has to do with the fact that the zero-sets do not remember multiplicities. For instance, the zero-set of $\langle (1-z_1)^2 \rangle$ in $\mathbb{R}[z_1]$ is just $\{1\}$, which is also the zero-set of $\langle 1 - z_1 \rangle$.

Reciprocally, we can introduce the vanishing ideal of an arbitrary subset S of $\mathbb{A}$, which is the set of all polynomials that cancel at every point of S (easily seen to be an ideal):

$$I(S) = \{f \in \mathbb{K}[z] \mid \forall z^\star \in S, \ f(z^\star) = 0\}$$

It can be shown that $I(S)$ is equal to the vanishing ideal of the Zariski closure of S, $I(S) = I\left(\bar{S}\right)$, and that $I(S)$ is always radical.

Thus, a strong relationship between radical ideals and Zariski closed sets is already apparent. In fact, the relationship is even stronger: the mapping $V \circ I$ is the identity mapping on the Zariski closed sets. The inclusion reversing relationship can be summarized as follows:

varieties	$\emptyset$	$\subset$	V_1	$\subset$	$V_1 \cup V_2$	$\subset$ $\mathbb{A} = \mathbb{K}^n$
$I(\cdot) \downarrow \uparrow V(\cdot)$			$\xrightarrow[\text{less polynomials}]{\text{more zeros}}$			
ideals	$\langle 1 \rangle = \mathbb{K}[z]$	$\supset$	$I(V_1)$	$\supset$	$I(V_1) \cap I(V_2)$	$\supset$ $\langle 0 \rangle = \{0\}$

If $\mathbb{K}$ is not algebraically closed, I fails to be surjective and V fails to be injective. For instance, in $\mathbb{R}[z_1]$, $V\left(\langle 1 + z_1^2 \rangle\right) = V\left(\langle 1 + z_1^4 \rangle\right)$.

When $\mathbb{K}$ is algebraically closed, Hilbert's famous Nullstellensatz implies that there is a 1-to-1 relationship between radical ideals and Zariski closed sets, given by V and its inverse I. For this reason, algebraic geometry is much nicer in $\mathbb{A}^n(\mathbb{C})$ than in $\mathbb{A}^n(\mathbb{R})$. For instance, if $\mathbb{K}$ is algebraically closed, there is a 1-to-1 relationship between the irreducible decomposition of a Zariski closed set and the prime decomposition of the corresponding radical ideal.

For this reason, Theorem 1 about the generic identification structure is a theorem over $\mathbb{C}$. However, thanks to the unusual features of the Zariski topology, the genericity carries over to, e.g., the real numbers between 0 and 1, as stated in Corollary 2.

A11.2 Proof of Theorem 1

First, we prove the existence part of Theorem 1, i.e., that there is some singular region Θ_s satisfying conditions (i) and (ii). The uniqueness part (the fact that there is a unique such singular region minimal for the inclusion) will follow easily.

We will obtain the $\bar{\theta}$-generic structure of the θ-solution set of $F\left(\theta, \bar{\theta}\right) = 0$ from the structure of the joint $\left(\theta, \bar{\theta}\right)$-solution set of $F\left(\theta, \bar{\theta}\right) = 0$.

Let $\mathbb{A} = \mathbb{A}^{2d_\theta}(\mathbb{C})$ be the joint affine space for $\left(\theta, \bar{\theta}\right)$ and let W be the zero set of $F\left(\theta, \bar{\theta}\right)$

jointly in both variables:

$$W = V(F) = \left\{ \left(\theta, \bar{\theta}\right) \in \mathbb{A} \middle| F\left(\theta, \bar{\theta}\right) = 0 \right\}$$

Let $\bar{\mathbb{A}} = \mathbb{A}^{d_\theta}(\mathbb{C})$ be the affine space for $\bar{\theta}$ only and $\pi : W \to \bar{\mathbb{A}}$ the restriction of the coordinate projection $\pi\left(\theta, \bar{\theta}\right) = \bar{\theta}$ to W. For any $\bar{\theta}^\star$, write $W\left(\bar{\theta}^\star\right)$ for the fiber of π at $\bar{\theta}^\star$:

$$W\left(\bar{\theta}^\star\right) = \pi^{-1}\left(\bar{\theta}^\star\right) = \left\{ \left(\theta, \bar{\theta}^\star\right) \in \mathbb{C}^3 \times \{\bar{\theta}^\star\} \middle| F\left(\theta, \bar{\theta}^\star\right) = 0 \right\}$$

Thus, $W\left(\bar{\theta}^\star\right)$ is the set of solutions in θ to the system of equations $F\left(\theta, \bar{\theta}^\star\right) = 0$ at a specific $\bar{\theta}^\star \in \bar{\mathbb{A}}$: this is exactly the system that needs to be solved for identification analysis.

An equivalent statement of the existence part of Theorem 1 is that $W\left(\bar{\theta}^\star\right)$ has a generic structure for $\bar{\theta}^\star \in \bar{U}$ where $\bar{U}$ is a (Zariski) open dense subset of $\bar{\mathbb{A}}$. To see that this is an equivalent statement, let $\bar{F}$ be a non-zero system of polynomials such that $\Theta_s = V\left(\bar{F}\right)$ as in Theorem 1. Remember that the zero-sets form the closed sets of the Zariski topology on $\bar{\mathbb{A}}$: saying that a property holds outside of $V\left(\bar{F}\right)$ is equivalent to saying that it holds on an open set $\bar{U}$, the complement of $V\left(\bar{F}\right)$. $\bar{U}$ is non-empty because $\bar{F}$ is not $\{0\}$. Any non-empty open set is dense in the Zariski topology. Thus, the statements are indeed equivalent.

To prove this equivalent statement we relate the generic structure of $W\left(\bar{\theta}^\star\right)$ to the generic structure of W, the "joint" zero-set. Consider the decompositions in irreducible components of W and of $W\left(\bar{\theta}^\star\right)$, for any $\bar{\theta}^\star$:

$$W = \bigcup_{m=1}^{M} V_m \qquad \text{and} \qquad W\left(\bar{\theta}^\star\right) = \bigcup_{j=1}^{J\left(\bar{\theta}^\star\right)} V_j\left(\bar{\theta}^\star\right)$$

We can use Theorem A.14.10 p. 349 from Sommese and Wampler (2005), adapted to our context:

Theorem: Theorem A.14.10 in Sommese and Wampler (2005)
There is a (Zariski) open dense set $\bar{U} \subset \bar{\mathbb{A}}$ such that for any $\bar{\theta}^\star \in \bar{U}$, and any $1 \leq m \leq M$, if V_m is an irreducible component of W of dimension d_m, then $V_m \cap W\left(\bar{\theta}^\star\right)$ is the union of a fixed number n_m (n_m independent of $\bar{\theta}^\star$) of irreducible components $V_j\left(\bar{\theta}^\star\right)$ of $W\left(\bar{\theta}^\star\right)$ of dimension $d_m - d_\theta$.

In particular, for any $\bar{\theta}^\star \in \bar{U}$, $W\left(\bar{\theta}^\star\right)$ has a fixed number of irreducible components of each dimension, which proves the existence statement of Theorem 1.

Example:
Here is a figurative example that illustrates the above. Consider:

$$W = \underbrace{V_1 \cup V_2}_{\text{dimension } d_\theta} \cup \underbrace{V_3}_{\text{dimension } d_\theta+1} \cup \underbrace{V_4}_{\text{dimension } d_\theta+2}$$

The generic structure of the fiber might be: 3 components of dimension 0 (points) and 1 component of dimension 2:

$$W\left(\bar{\theta}^\star\right) = \underbrace{V_1\left(\bar{\theta}^\star\right) \cup V_2\left(\bar{\theta}^\star\right) \cup V_3\left(\bar{\theta}^\star\right)}_{\text{dimension } 0} \cup \underbrace{V_4\left(\bar{\theta}^\star\right)}_{\text{dimension } 2}$$

Each $V_m \cap W\left(\bar{\theta}^\star\right)$ must have the same number of components $V_m\left(\bar{\theta}^\star\right)$. For instance, for some value $\bar{\theta}^\star$, we might have:

$$V_1 \cap W\left(\bar{\theta}^\star\right) = V_1\left(\bar{\theta}^\star\right) \cup V_2\left(\bar{\theta}^\star\right) \qquad V_2 \cap W\left(\bar{\theta}^\star\right) = V_3\left(\bar{\theta}^\star\right)$$
$$V_3 \cap W\left(\bar{\theta}^\star\right) = \emptyset \qquad V_4 \cap W\left(\bar{\theta}^\star\right) = V_4\left(\bar{\theta}^\star\right)$$

While for some other value $\bar{\theta}^\diamond$, we might have:

$$V_1 \cap W\left(\bar{\theta}^\diamond\right) = V_1\left(\bar{\theta}^\diamond\right) \cup V_3\left(\bar{\theta}^\diamond\right) \qquad V_2 \cap W\left(\bar{\theta}^\diamond\right) = V_2\left(\bar{\theta}^\diamond\right)$$
$$V_3 \cap W\left(\bar{\theta}^\diamond\right) = \emptyset \qquad V_4 \cap W\left(\bar{\theta}^\diamond\right) = V_4\left(\bar{\theta}^\diamond\right)$$

See section A11.5 for an actual example in a toy econometric model. □

Note that $\bar{F}$ such that $\Theta_s = V\left(\bar{F}\right)$ could include polynomials with coefficients in $\mathbb{C}$. In fact, we can do better and show that $\Theta_s \cap \mathbb{R}^{d_\theta}$ is the zero-set of a finite number of real polynomials. Indeed, $\Theta_s \cap \mathbb{R}^{d_\theta}$ is a closed subset of $\mathbb{R}^{d_\theta}$ with the subtopology inherited from $\mathbb{A}^{d_\theta}(\mathbb{C})$, which coincides with $\mathbb{A}^{d_\theta}(\mathbb{R})$ (Fact 4).

Turning to the uniqueness part of Theorem 1, consider the union Θ_g of all $\bar{U}$, such that the existence statement holds for $\bar{U}$. Θ_g is open and dense. Θ_g satisfies the existence statement and contains by definition all the open dense sets that do, i.e., Θ_g is maximal for the inclusion.

A11.3 Proof of other results

Proof of Corollary 1. This is a particular case of Theorem 1 where the generic identification structure has n_{ls} irreducible components of dimension zero and no component in other dimensions. □

Proof of Corollary 2. (i) Θ_g is Zariksi open in $\mathbb{C}^{d_\theta}$ implies Θ_g is Euclidean open in $\mathbb{C}^{d_\theta}$ (Fact 4) implies $\Theta_g \cap [0,1]^{d_\theta}$ is Euclidean open in $[0,1]^{d_\theta}$ (subspace topology). For the Euclidean topology, Θ_g is dense in $\mathbb{R}^{d_\theta}$ (Fact 4) implies Θ_g is dense in $]0,1[^{d_\theta}$ implies Θ_g is dense in $[0,1]^{d_\theta}$. Thus Θ_g is Euclidean open dense in $[0,1]^{d_\theta}$.
(ii) Θ_s has Lebesgue measure zero in $\mathbb{R}^{d_\theta}$ (Fact 4) and thus has Lebesgue measure zero in $[0,1]^{d_\theta}$. □

Proof of Theorem 2. $\langle F \rangle$ is generated by a finite number of elements of F (Fact 5). 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 $V_{\theta,\theta^\star}(F_T) = V_{\theta,\theta^\star}(F_{T_0})$ and in particular for any $\theta^\star$, $V_\theta\left(F_T\left(\cdot,\theta^\star\right)\right) = V_\theta\left(F_{T_0}\left(\cdot,\theta^\star\right)\right)$. □

Proof of Corollary 3. This is a direct consequence of Theorem 2. □

A11.4 Computing identification with computational algebraic geometry

Computational algebraic geometry is an academic field in itself. Methods for solving systems of polynomial equations play a central role, but there are also algorithms for computing radicals, primary decompositions, etc. Those can also be useful for a computational approach to identification analysis. A complete survey is beyond the scope of this paper. See Cox et al. (2005) for an entry point to computational algebraic geometry. Several specialized computer algebra systems implement algorithms for computational algebraic geometry. I use Singular (Decker et al., 2014) to compute identification in section A11.5.

A complete identification analysis in a given discrete model consists of computing both the minimal singular region as well as the generic identification structure. This means we need to look at the identification equation $F\left(\theta,\theta^\star\right) = 0$ jointly in θ and $\theta^\star$, somewhat similar to the strategy I used to prove Theorem 1 and Theorem 2. One way to compute the minimal singular region in a generically identified model is as follows:

1. Compute the prime decomposition of $\sqrt{\langle F \rangle}$. If the model is generically identified, Theorem 1 tells us it will have one irreducible component of dimension d_θ (namely

$V(\langle \theta - \theta^\star \rangle) = \{\{\theta, \theta^\star\}, \theta \in \Theta\}$ and other irreducible components V_i's of higher dimension. The singular region is (the Zariski closure of) the intersection of $W_s = \cup_i V_i$ with the $\theta^\star$ space, since for these values of $\theta^\star$, $F(\theta, \theta^\star) = 0$ will have more solutions than just $\{\theta\}$.

2. The intersection of $W_s = \cup_i V_i$ with the $\theta^\star$ space (the affine space $\bar{\mathbb{A}}$) is not Zariski closed in general, but its Zariski closure is the minimal singular region. This operation is the familiar "elimination of variables" procedure. There are algorithms to eliminate variables.

See section A11.5 for an application of this method. Unfortunately, the joint perspective in θ and $\theta^\star$ makes this approach particularly computationally intensive. Computing prime decompositions is also a computationally intensive task.

An alternative approach is to find a singular region $\tilde{\Theta}_s \supset \Theta_s$. In generically identified models, this corresponds to finding sufficient conditions for identification rather than necessary and sufficient conditions: outside of $\tilde{\Theta}_s$, the model is identified. A naïve way to do this is to look at $F(\theta, \theta^\star) = 0$ as polynomials in θ, with coefficients in $\mathbb{K}[\theta^\star]$ (the polynomials in $\theta^\star$) rather than $\mathbb{K}$. Then one can try to solve the system by any standard method, putting aside all values of $\theta^\star$ that cancel a denominator along the way. This will be much faster than the technique of the previous paragraph, but $\tilde{\Theta}_s$ might be much too big. There are better ways to use the $\mathbb{K}(\theta^\star)[\theta]$ perspective (known in computational algebraic geometry as the "parametric polynomial system" perspective) to compute a $\tilde{\Theta}_s$ that is very close to the actual Θ_s; see, e.g., Montes and Wibmer (2010).

Finally, Theorem 1 suggests the following approach to computing identification: draw a random parameter value $\theta^\star$, and check identification at $\theta^\star$ by solving $F(\theta, \theta^\star) = 0$. This gives generic identification, although not an explicit expression for the singular region: the economist must be willing to rule out all singularities beyond those she chose to include in her model to start with. Any tool for solving plain systems of polynomials can be used to solve $F(\theta, \theta^\star) = 0$.

The methods of computational algebraic geometry are all symbolic. There are other computational approaches to solving systems of polynomials that might be useful, such as the methods of semi-algebraic geometry (real solutions to polynomial systems) and homotopy methods in numerical computational algebraic geometry.

None of the methods outlined in this section takes into account the specific structure of $F(\theta, \theta^\star) = 0$. F is not *any* system of polynomial equations, but one that is generated by the transition matrices and initial distribution of the model. The specific structure might be leveraged both for theoretical and computational purposes. The symbolic computational approach is the most promising one in this respect.

A11.5 Identification analysis in a toy model

This subsection carries a detailed identification analysis in a toy model. The model retains all the interesting features of a hidden Rust model, except for the dynamic aspects. It is a good opportunity to illustrate:

- General algebraic-geometric concepts such as the decomposition in irreducible components of a Zariski topological set.
- The mechanism behind Theorem 1.
- The advantage of automatic methods to compute the generic identification structure as well as the minimal singular region.

Computations are carried in Singular (see section A11.4).

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 $\theta = (p, p_a, p_b)$ is identified from the distribution of Y given by:

$$\begin{aligned}
\mathbb{P}(Y = 0) &= (1 - p)(1 - p_a)^3 + p(1 - p_b)^3 \\
\mathbb{P}(Y = 1) &= (1 - p)3(1 - p_a)^2 p_a + p3(1 - p_b)^2 p_b \\
\mathbb{P}(Y = 2) &= (1 - p)3(1 - p_a)p_a^2 + p3(1 - p_b)p_b^2 \\
\mathbb{P}(Y = 3) &= (1 - p)p_a^3 + pp_b^3
\end{aligned}$$

With $\bar{\theta} = (q, q_a, q_b)$, the identification system can be written:

$$F\left(\theta, \bar{\theta}\right) = 0$$

$$\iff$$

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

In this simple example, we can solve the system "by hand," keeping track of the forbidden divisions by zero as we proceed. We might find, for instance, that as long as $g\left(\bar{\theta}\right) = (q_b - q_a)q(1-q)(1-2q) \neq 0$, $F\left(\theta, \bar{\theta}\right) = 0$ is equivalent to:

$$\begin{cases} p(1-p) = q(1-q) \\ p_a = q_a \frac{1-q-p}{1-2q} + q_b \frac{p-q}{1-2q} \\ p_b = q_b \frac{1-q-p}{1-2q} + q_a \frac{p-q}{1-2q} \end{cases}$$

Thus, the system is identified outside of the singular region $g\left(\bar{\theta}\right) = 0$ (up to label-switching).

Now, in order to illustrate the mechanism behind Theorem 1, we would like to compute the decomposition of $W = V(F)$, the zero-set of $F\left(\theta, \bar{\theta}\right)$ jointly in $\left(\theta, \bar{\theta}\right)$. While this geometric decomposition is hard to compute directly, we can rely on the strong correspondance between geometry and algebra that lies at the core of algebraic geometry, and carry an algebraic computation. Remember from section A11.1 that $V(F) = V\left(\sqrt{\langle F \rangle}\right)$ where $\sqrt{\langle F \rangle}$ is the radical of the ideal generated by F, and that the irreducible decomposition of W is 1-to-1 with the decomposition of $\sqrt{\langle F \rangle}$ in primes. We can compute the prime decomposition of $\sqrt{\langle F \rangle}$ with Singular (this is instantaneous). It has 11 components:

$$\begin{aligned} \sqrt{\langle F \rangle} = {} & \langle p_b - q_b, p_a - q_a, p - q \rangle \cap \langle p_b - q_a, p_a - q_b, (1-p) - q \rangle && I_1 \\ & \cap \langle p_b - q_b, p_a - q_a, q_a - q_b \rangle && I_2 \\ & \cap \begin{cases} \langle p_b - q_b, q_a - q_b, p \rangle \cap \langle p_b - q_b, p_a - q_b, q \rangle \\ \cap \langle p_a - q_b, q_a - q_b, p - 1 \rangle \cap \langle p_b - q_a, p_a - q_a, q - 1 \rangle \end{cases} && I_3 \\ & \cap \begin{cases} \langle q - 1, p, p_b - q_a \rangle \cap \langle q, p, p_b - q_b \rangle \\ \cap \langle q - 1, p - 1, p_a - q_a \rangle \cap \langle q, p - 1, p_a - q_b \rangle \end{cases} && I_4 \end{aligned}$$

Correspondingly:

$$W = \underbrace{V(I_1)}_{\text{generic region}} \cup \underbrace{V(I_2) \cup V(I_3) \cup V(I_4)}_{\text{singular region}} = V_g + V_s$$

$V(I_1)$ is the "nice" region, which contains the identified parameter values $\theta = \bar{\theta}$ as well as their label-switched versions. $V(I_2)$ contains parameter values for which coin flips happen with the same probability regardless of being in a/b or in $\theta/\bar{\theta}$. Of course, the first stage probabilities are not identified when this is the case. $V(I_3)$ contains parameter values for which in one $\theta/\bar{\theta}$ world, one of the flips is never observed, and in the other world, both flips happen but are indistinguishable due to having the same flipping probability. There are four subcomponents by symmetry and label-switching. $V(I_4)$ contains parameter values for which only one flip happens in both $\theta/\bar{\theta}$ worlds. There are also four subcomponents by symmetry and label-switching.

In this simple model, we can tell by eyeballing that $V_g = V(I_1)$ makes up the generic component of W, while $V_s = V(I_2) \cup V(I_3) \cup V(I_4)$ makes up its singular component. The singular region in $\bar{\theta}$ space is simply the intersection of $\bar{\mathbb{A}}$ with the singular points V_s of W. The geometric object $V_s \cap \bar{\mathbb{A}}$ is not necessarily a closed set, but once more, we can use algebraic methods to compute its closure. We find:

$$\Theta_s = \overline{V_s \cap \bar{\mathbb{A}}} = V(< (q_b - q_a)q(1-q) >)$$

We were not too far-off in our computation "by hand" above. The only difference is that $q = 1/2$ does not appear in the singular region — indeed if $q = 1/2$, there are as many solutions as at other points of the generic region: $(p, p_a, p_b) = (1/2, q_a, q_b)$ and $(p, p_a, p_b) = (1/2, q_b, q_a)$.

The advantage of automatic methods to compute the generic identification structure as well as the minimal singular region becomes quickly apparent in slightly more involved examples, although computational issues arise when the number of variables, as well as the number and degree of polynomials, increases.

A12 Appendix for section 4: Asymptotics

A12.1 Preamble: Merging Markov chains

This section defines *merging* and the *merging time* of a finite-state Markov chain, and states concentration inequalities for hidden Markov models from Paulin (2014).

In this paper, merging is used at various stages of the asymptotic analysis. Non-homogeneous merging of various conditional Markov chains is key in showing limit theorems for the log-likelihood, score and information under stationarity (in appendix section A12.4). Merging of the observed variables $y_t = (s_t, a_t)$ is also used to extend these stationary limit theorems to their nonstationary versions (in appendix section A12.5), and in a similar fashion to build uniformly consistent tests under non-stationarity from their stationary analogs (in appendix section A12.7). Finally, the asymptotic study of the Bayesian posterior uses concentration inequalities from Paulin (2014), valid for merging chains and stated with concentration constants proportional to the merging time ((4), (5) and (6)). See also sections 4.4 to A12.5.1 for an overview of some of these proofs.

This section does not contain original results. Most of the definitions are standard, although the terminology might vary. General facts about finite-state Markov chains, including merging properties, can be found in Seneta (2006). I define the merging time following the slighlty nonstandard definition of Paulin (2014), which is motivated by the concentration bounds (4) - (6).

The following terminology is used. A non-homogeneous Markov chain is a chain whose transition matrix changes with time. For homogeneous chains, an aperiodic chain is always irreducible and an irreducible chain is always recurrent.

A not necessarily homogeneous chain Z_t is *merging* when the total variation distance between the distributions of two independent chains started at arbitrary points goes to zero with time — the chains "merge":

$$\forall s, z_s, z'_s, \qquad d_{TV}(\mathcal{L}(Z_t|Z_s = z_s), \mathcal{L}(Z_t|Z_s = z'_s)) \xrightarrow[t\to\infty]{} 0$$

We can think of merging as the influence of the initial distribution fading away with time.

If Z_t is homogeneous and merging, there is a distribution $\mu^\diamond$, necessarily unique, such that:

$$\forall z_1, \qquad d_{TV}(\mathcal{L}(Z_t|Z_1 = z_1), \mu^\diamond) \xrightarrow[t\to\infty]{} 0$$

Furthermore $\mu^\diamond$ is the unique stationary distribution of Z_t and the convergence happens geometrically fast: there is $\rho < 1$ and $c > 0$ such that:

$$\forall z_1, \qquad d_{TV}(\mathcal{L}(Z_t|Z_1 = z_1), \mu^\diamond) \leq c\rho^t$$

The distribution of Z_t "merges" to $\mu^\diamond$ regardless of the chain's initial distribution. This is a familiar phenomenon for anyone who has seen traceplots of Markov chain Monte Carlo algorithms started far from equilibrium.

If Z_t is homogeneous and merging, under the minor additional assumption that Z_t has no transient states, or by ignoring the transient states (which correspond to zeros in $\mu^\diamond$), Z_t is also aperiodic. Conversely, an aperiodic chain is recurrent and merging.

The (joint) stationary distribution of a merging chain has very strong ergodic properties (so-called *mixing* properties), although we will not use these beyond ergodicity.

The merging time is a quantity used to measure the merging speed. The ϵ-merging time $\tau_z(\epsilon)$ of a merging Markov chain z is defined as follows, for $0 < \epsilon < 1$:

$$\tau_z(\epsilon) = \min\{t : \max_z\{d_{TV}(\mathcal{L}(Z_t|Z_1 = z_1), \mu^\diamond)\} < \epsilon\}$$

The absolute merging time, or simply *merging time*, is:

$$\tau_z = \inf_{0\leq\epsilon<1} \frac{\tau_z(\epsilon/2)}{(1-\epsilon)^2}$$

This seemingly ad-hoc definition is the convenient one for stating concentration inequalities where the concentration constant is directly proportional to the merging time.

I state McDiarmid inequalities for (not necessarily stationary) hidden Markov models from Paulin (2014) (corollary 2.14 p.12):

Lemma 1: Concentration inequalities for HMMs
For (x, y) a HMM, write τ_x for the merging time of the Markov chain x. Let f be a function

of T arguments with bounded differences:

$$f(y_1, ..., y_T) - f(y'_1, ..., y'_T) \leq \sum_{t=1}^{T} c_t 1[y_t \neq y'_t]$$

Then the following one-sided and two-sided concentration inequalities hold:

$$P\left(f < \mathbb{E}\left[f\right] - u\right) \leq \exp\left(-\frac{1}{2}\frac{u^2}{\tau_x \sum_{t=1}^{T} c_t^2}\right) \tag{4}$$

$$P\left(f > \mathbb{E}\left[f\right] + u\right) \leq \exp\left(-\frac{1}{2}\frac{u^2}{\tau_x \sum_{t=1}^{T} c_t^2}\right) \tag{5}$$

$$P\left(|f - \mathbb{E}[f]| > u\right) \leq 2\exp\left(-\frac{1}{2}\frac{u^2}{\tau_x \sum_{t=1}^{T} c_t^2}\right) \tag{6}$$

According to 4, concentration is "$4\tau_x$ times slower" for a merging Markov chain than for a sequence of independent and identically distributed random variables.

Finally, I will use the following lemma (see, e.g., theorem 4.9 p.141 in Seneta (2006)):

Lemma 2: Merging for (non-homogeneous) Markov chains under minorization
If the transition probabilities of a (not necessarily homogeneous) Markov chain Z_t are uniformly minorized in the sense that there are probability distributions ν_t and a constant $\underline{a} > 0$ such that for any values of the chain z_t and z_{t+1}:

$$P(z_{t+1}|z_t) \geq \underline{a}\nu_t(z_t + 1)$$

Then Z_t satisfies merging: for any two initial distributions μ_1 and μ_2:

$$d_{TV}\left(\mathcal{L}(Z_t|Z_0 \sim \mu_1), \mathcal{L}(Z_t|Z_0 \sim \mu_2)\right) < (1 - \underline{a})^t$$

A word of caution on terminology. Merging properties are sometimes referred to as "weak ergodicity" properties and the merging time is usually called the "mixing time." We keep "ergodicity" and "mixing" for properties that stationary distributions can have. The merging time is also more usually defined as $\tau_z(\epsilon)$ for an arbitrary value of ϵ, traditionally $1/4$. This is because in other contexts the arbitrary choice of ϵ does not matter (see Levin et al. (2009) for a textbook treatment of the theory of merging times).

A12.2 z is uniformly merging

We know that Θ is compact and that for any θ, z is recurrent and merging. We want to show that z is *uniformly merging*, in the sense that the merging time is uniformly bounded: there is $\bar{\tau}_z < \infty$ such that:

$$\forall \theta \in \Theta, \qquad \tau_z(\theta) < \bar{\tau}_z$$

It is not clear that τ_z itself is a continuous function of the transition matrix but we will use a bound from Paulin (2014).

Because z is merging, z has a unique marginal stationary distribution, which we write $\mu^\diamond(\theta)$. Because z is recurrent, $\mu^\diamond(\theta) > 0$.

Define $\lambda(\theta)$ as the second biggest eigenvalue of the multiplicative reversibilization $\tilde{P}_z$ of the transition matrix P_z of z (see Fill (1991)). $\tilde{P}_z$ is a continuous function of P_z, and $\tilde{P}_z$ is recurrent merging because P_z is.

A consequence of bound (3.12) p. 16 together with bound (2.9) p. 10 in Paulin (2014) is:

$$\tau_z(\theta) \leq 4 \frac{1 + 2\log 2 + \log 1/\mu^\diamond_{min}(\theta)}{1 - \lambda(\theta)}$$

Stationary distributions and eigenvalues are continuous functions of matrix coefficients. As a consequence, $\lambda(\theta)$ is bounded away from 1 and $\mu^\diamond(\theta)$ is bounded away from zero on Θ. The conclusion follows. We call τ_z the lowest uniform upper bound:

$$\tau_z = \sup_{\theta \in \Theta} \tau_z(\theta) < \infty$$

For the same reasons, blocks of z's are also uniformly merging. Let R be any natural number ≥ 1 and $\hat{z}_s$ be non-overlapping consecutive blocks of R z_t's: $\hat{z}_1 = (z_1, \ldots, z_R), \hat{z}_2 = (z_{R+1}, \ldots, z_{2R})$, etc. $\hat{z}$ is merging because z is merging and consequently $\hat{z}$ is uniformly merging for the same reason z is uniformly merging. We write $\tau_{\hat{z}}$ for the corresponding uniform merging time ($\tau_{\hat{z}}$ can depend on R).

A12.3 The infinite-past strategy

All three limit theorems are based on an *infinite-past* strategy where taking some conditional expectations to the infinite past plays a key role. In order for the infinite-past to make

sense, stationarity is key. The infinite-past strategy is best illustrated with the log-likelihood.

First, I write $L_T(\theta)$ as the sum of the auxiliary processes U_{1t}:

$$L_T(\theta) = \frac{1}{T}\sum_{t=1}^{T-1} U_{1t}(\theta) \qquad \text{where } U_{1t}(\theta) = \log P_\theta(Y_{t+1}|Y_{1:t})$$

Second, I take the auxiliary process $U_{mt}(\theta) = \log P_\theta(Y_{t+1}|Y_{m:t})$ to its infinite past $m \to -\infty$. I can think of the infinite-past limit $U_t(\theta) = U_{-\infty t}(\theta)$ as $\log P_\theta(Y_{t+1}|Y_{-\infty:t})$, although care is needed (see appendix section A12). To take the infinite-past limit, I show that $U_{mt}(\theta)$ is $\theta^\star$ almost surely Cauchy, uniformly in θ. The uniform lower bound $\underline{q}$ on the unobserved state transition matrix Q is used to obtain non-homogeneous merging of the conditional Markov chain $(X_t|y_{m:T})_{t\geq m}$. This is a key step. See appendix section A12.1 for an introduction to merging and an overview of how it is used at different stages of the proof.

Third, I show that $\left\|L_T(\theta) - \frac{1}{T}\sum_{t=1}^{T-1} U_t(\theta)\right\| \xrightarrow{\theta^\star \ as} 0$. This follows directly from some bounds derived to show the Cauchy property.

Fourth, I am left with showing a uniform law of large numbers for U_t, which is now an ergodic sum:

$$\left\|\frac{1}{T}\sum_{t=1}^{T-1} U_t(\theta) - L(\theta)\right\| \xrightarrow{\theta^\star \ as} 0$$

For this I use a vector ergodic theorem valid for random vectors taking value in separable metric spaces from Krengel (1985). $U_t(\theta)$ inherits continuity from $U_{mt}(\theta)$, thanks to the almost surely uniform Cauchy property. As a consequence U_t can be seen globally as a random vector taking value in $\mathcal{C}(\Theta)$, which is separable because Θ is compact.

The same general infinite-past strategy is used for the score $s_T = \nabla_{\theta^\star} L_T(\theta)$ and the observed information $h_T = \nabla^2_\theta L_T(\theta)$, although new complications arise, such as making sure that some infinite sums are almost surely summable as we take m to $-\infty$.

A12.4 Limit theorems for stationary hidden Rust models

In this section, assume (X_t, Y_t) are distributed according to the stationary distribution induced by $\theta^\star$ (recall that, by the merging assumption (A2), each θ induces a unique marginal stationary distribution $\mu^\diamond(\theta)$). Under this assumption, (X_t, Y_t) can be extended to $(X, Y)_{-\infty:+\infty}$.

The log-likelihood is:

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

$P_{\theta,\mu}$ means the value of the probability under transitions indexed by θ and initial distribution (for the earliest index appearing in $P_{\theta,\mu}(\cdot)$) μ. In particular μ is not necessarily equal to $\mu^\diamond(\theta)$; thus, the likelihood can be misspecified with respect to the initial distribution. In fact, μ plays no role in the proof and is suppressed from notation for simplicity. It is probably easier at a first reading to assume $\mu = \mu^\diamond(\theta^\star)$, in which case P_θ is well-specified.

A12.4.1 Uniform law of large numbers for the log-likelihood

This section shows a uniform law of large numbers for the log-likelihood:

$$\|L_T(\theta) - L(\theta)\| \xrightarrow{\theta^\star \ as} 0$$

where $L(\theta)$ can be thought of as $\mathbb{E}_{\theta^\star}[\log P_\theta\left(Y_1|Y_{-\infty:0}\right)]$. The precise meaning of this notation is made clear below.

The following decomposition is used:

$$\begin{cases} L_T(\theta) = \frac{1}{T}\sum_{t=1}^{T-1} U_{1t}(\theta) \\ \left\| \frac{1}{T}\sum_{t=1}^{T-1} U_{1t}(\theta) - \frac{1}{T}\sum_{t=1}^{T-1} U_t(\theta) \right\| \xrightarrow{\theta^\star \ as} 0 \\ \left\| \frac{1}{T}\sum_{t=1}^{T-1} U_t(\theta) - L(\theta) \right\| \xrightarrow{\theta^\star \ as} 0 \end{cases}$$

We define the auxiliary process $U_{mt}(\theta) = \log P_\theta\left(Y_{t+1}|Y_{m:t}\right)$. $U_t(\theta)$ will be defined as the limit of $U_{mt}(\theta)$ as $m \to -\infty$, which will be shown to exist. $U_t(\theta)$ can be thought of as $\log P_\theta\left(Y_{t+1}|Y_{-\infty:t}\right)$ both in a moral sense and in a technical sense to be made precise. This is why we speak of an "infinite-past" proof strategy. $U_t(\theta)$ will also be seen to be stationary ergodic. Of course, $L(\theta) = \mathbb{E}_{\theta^\star}[U_0(\theta)]$.

There are three steps in the proof:

1. $U_{mt}(\theta)$ is an (almost surely) Cauchy sequence in $m \to -\infty$, uniformly in θ. We call $U_t(\theta)$ its limit and we explain in which sense $U_t(\theta) = \log P_\theta\left(Y_{t+1}|Y_{-\infty:t}\right)$.
2. $\left\| \frac{1}{T}\sum_{t=1}^{T-1} U_{1t}(\theta) - \frac{1}{T}\sum_{t=1}^{T-1} U_t \right\|(\theta) \xrightarrow{\theta^\star \ as} 0$.
3. $\left\| \frac{1}{T}\sum_{t=1}^{T-1} U_t(\theta) - L(\theta) \right\| \xrightarrow{\theta^\star \ as} 0$.

Step 1: U_{mt} is $\theta^\star$ almost surely uniform Cauchy

We want to define $U_t(\theta) = U_{-\infty t}(\theta)$ and show the following $\theta^\star$ almost sure geometric bound, valid for any $-\infty \leq m < m' \leq 1$ and uniform in θ:

$$|U_{m't}(\theta) - U_{mt}(\theta)| \leq K\rho^{t-m'} \tag{7}$$

First, we show (7) for $-\infty < m < m' \leq 1$. Note that because $|\log x - \log y| \leq \frac{|x-y|}{x \wedge y}$:

$$|\log P_\theta(y_{t+1}|y_{m':t}) - \log P_\theta(y_{t+1}|y_{m:t})| \leq \frac{|P_\theta(y_{t+1}|y_{m':t}) - P_\theta(y_{t+1}|y_{m:t})|}{P_\theta(y_{t+1}|y_{m':t}) \wedge P_\theta(y_{t+1}|y_{m:t})}$$

A lower bound for the denominator is easy to find. Note that:

$$P_\theta(y_{t+1}|y_{m:t}) = \sum_{x_{t+1},x_t} P_\theta(y_{t+1}|x_{t+1},y_t)P_\theta(x_{t+1}|x_t)P_\theta(x_t|y_{m:t})$$

So that:

$$P_\theta(y_{t+1}|y_{m:t}) \geq \underline{q} \sum_{x_{t+1}} P_\theta(y_{t+1}|x_{t+1},y_t)$$

Let us turn to the numerator. By conditional independence, $(X_t|y_{m:T})_{t\geq m}$ is (non-homogeneous) Markov. We show that it satisfies a merging property uniformly in θ (see section A12.1 for an introduction about merging):

Lemma 3: Uniform merging for $(X_t|y_{m:T})_{t\geq m}$
There is $\rho < 1$ such that for any m, for any two initial distributions μ_1 and μ_2 on X_m, for any θ, the following inequality holds (for any $y_{m:T}$ with positive probability):

$$d_{TV}\left(\mathcal{L}_\theta(X_t|y_{m:T};\mu_1), \mathcal{L}_\theta(X_t|y_{m:T};\mu_2)\right) < \rho^{t-m}$$

Proof. For any $t \geq m$, any $(x_t, y_{m:T})$ with positive probability:

$$\begin{aligned}
P_\theta(x_{t+1}|x_t,y_{m:T})P_\theta(y_{t+1:T}|x_t,y_{m:t}) &= P_\theta(x_{t+1},y_{t+1:T}|x_t,y_{m:t}) \\
&= P_\theta(y_{t+1:T}|x_{t+1},x_t,y_{m:t})P_\theta(x_{t+1}|x_t,y_{m:t}) \\
&= P_\theta(y_{t+1:T}|x_{t+1},y_t)P_\theta(x_{t+1}|x_t) \\
P_\theta(x_{t+1}|x_t,y_{m:T}) &= \frac{P_\theta(y_{t+1:T}|x_{t+1},y_t)P_\theta(x_{t+1}|x_t)}{\sum_{x_{t+1}} P_\theta(y_{t+1:T}|x_{t+1},y_t)P_\theta(x_{t+1}|x_t)} \\
&\geq \underline{q}\frac{P_\theta(y_{t+1:T}|x_{t+1},y_t)}{\sum_{x_{t+1}} P_\theta(y_{t+1:T}|x_{t+1},y_t)}
\end{aligned}$$

$\nu_t(x_{t+1};\theta) := \frac{P_\theta(y_{t+1:T}|x_{t+1},y_t)}{\sum_{x_{t+1}} P_\theta(y_{t+1:T}|x_{t+1})}$ defines a probability distribution on X_{t+1}, and the above inequality is a uniform minorization of the transition probabilities by $\nu_t(\theta)$:

$$P_\theta(x_{t+1}|x_t, y_{m:T}) \geq \underline{q}\nu_t(x_{t+1};\theta)$$

By Lemma 2 in section A12.1:

$$d_{TV}\left(\mathcal{L}_\theta(X_t|y_{m:T};\mu_1), \mathcal{L}_\theta(X_t|y_{m:T};\mu_2)\right) < (1-\underline{q})^{t-m}$$

□

Coming back to bounding the numerator, remember that for any two probabilities μ_1 and μ_2 and any f, $0 \leq f \leq 1$:

$$|\mu_1 f - \mu_2 f| \leq d_{TV}(\mu_1, \mu_2) \tag{8}$$

Then, for any $y_{m:T}$ with positive probability:

$$\begin{aligned}
&|P_\theta(y_{t+1}|y_{m':t}) - P_\theta(y_{t+1}|y_{m:t})| \\
&\quad = \left| \sum_{x_{t+1},x_t} P_\theta(y_{t+1}|x_{t+1},y_t) P_\theta(x_{t+1}|x_t)(P_\theta(x_t|y_{m':t}) - P_\theta(x_t|y_{m:t})) \right| \\
&\quad \leq \sum_{x_{t+1}} P_\theta(y_{t+1}|x_{t+1},y_t) \left| \sum_{x_t} P_\theta(x_{t+1}|x_t)(P_\theta(x_t|y_{m':t}) - P_\theta(x_t|y_{m:t})) \right| \\
&\quad \leq \sum_{x_{t+1}} P_\theta(y_{t+1}|x_{t+1},y_t) \left| \sum_{x_t} P_\theta(x_{t+1}|x_t) \left(P_\theta(x_t|y_{m':t}) - \sum_{x_{m'}} P_\theta(x_t|y_{m':t},x_{m'}) P_\theta(x_{m'}|y_{m:t}) \right) \right| \\
&\quad \leq \sum_{x_{t+1}} P_\theta(y_{t+1}|x_{t+1},y_t) d_{TV}\left(\mathcal{L}_\theta(X_t|y_{m':t};x_{m'}|y_{m':t}), \mathcal{L}_\theta(X_t|y_{m':t};x_{m'}|y_{m:t})\right) \quad \text{by (8)} \\
&\quad \leq \rho^{t-m'} \sum_{x_{t+1}} P_\theta(y_{t+1}|x_{t+1},y_t) \qquad \text{by merging, lemma 3}
\end{aligned}$$

Putting bounds for the numerator and denominator together, for $-\infty < m < m' \leq 1$ we have (almost surely):

$$\left|\log P_\theta\left(Y_{t+1}|Y_{m':t}\right) - \log P_\theta\left(Y_{t+1}|Y_{m:t}\right)\right| \leq \frac{\rho^{t-1}\sum_{x_{t+1}} P_\theta\left(Y_{t+1}|X_{t+1},Y_t\right)}{\underline{q}\sum_{x_{t+1}} P_\theta\left(Y_{t+1}|X_{t+1},Y_t\right)} = \frac{\rho^{t-m'}}{\underline{q}} \tag{9}$$

Note that we can collect all the (countable) null sets of (9) so that the precise statement is that the inequality holds "almost surely: for all m and m'" and not "for each m and m': almost surely." This implies that $U_{mt}(\theta)$ is almost surely a Cauchy sequence as $m \to -\infty$.

As a consequence there is U_t such that (almost surely):

$$U_{mt}(\theta) \xrightarrow[m\to-\infty]{} U_t(\theta)$$

Note that $U_t(\theta) = \log P_\theta\left(Y_{t+1}|Y_{-\infty:t}\right)$ in the following sense: if $g(Y_{-\infty:t+1}) = \mathbb{E}_\theta[Y_{t+1}|Y_{-\infty:t}]$ is a version of Kolmogorov's conditional expectation under θ, then $U_t(\theta) = g(Y_{-\infty:t+1})$ where $Y_{-\infty:t+1}$ is distributed according to $\theta^\star$.

By continuity, (7) holds for $m = -\infty$ too.

Step 2: $\left\|\frac{1}{T}\sum_{t=1}^{T-1} U_{1t}(\theta) - \frac{1}{T}\sum_{t=1}^{T-1} U_t(\theta)\right\| \xrightarrow{\theta^\star \ as} 0$

From the (almost sure) geometric bound (7) (with $m = -\infty$ and $m' = 1$), we have (almost surely):

$$\begin{aligned}
&\left|\frac{1}{T}\sum_{t=1}^{T-1} \log P_\theta\left(Y_{t+1}|Y_{1:t}\right) - \frac{1}{T}\sum_{t=1}^{T-1} \log P_\theta\left(Y_{t+1}|Y_{-\infty:t}\right)\right| \\
&\qquad \leq \frac{1}{T}\sum_{t=1}^{T-1} \left|\log P_\theta\left(Y_{t+1}|Y_{1:t}\right) - \log P_\theta\left(Y_{t+1}|Y_{-\infty:t}\right)\right| \\
&\qquad \leq \frac{1}{T}\frac{1}{\underline{q}}\sum_{t=1}^{T-1} \rho^{t-1} \\
&\qquad \leq \frac{1}{T}\frac{1}{\underline{q}}\frac{1}{1-\rho}
\end{aligned}$$

which implies:

$$\left\|\frac{1}{T}\sum_{t=1}^{T-1} U_{1t}(\theta) - \frac{1}{T}\sum_{t=1}^{T-1} U_t(\theta)\right\| \xrightarrow{\theta^\star \ as} 0$$

Step 3: $\left\|\frac{1}{T}\sum_{t=1}^{T-1} U_t(\theta) - L(\theta)\right\| \xrightarrow{\theta^\star \ as} 0$

This is a consequence of the almost sure ergodic theorem in function space. To see this, note that $\theta \to U_{mt}(\theta)$ is (almost surely) continuous because $P_\theta\left(y_{t+1}|y_{m:t}\right)$ is a rational function of the transition matrices' coefficients. The results of the first step imply that $\theta \to U_t(\theta)$ is (almost surely) continuous as a uniform Cauchy limit. Thus we can consider all of them to be everywhere continuous without loss of generality. Now define $\mathcal{C}(\Theta)$ to be the set of

continuous functions from Θ to $\mathbb{R}$ and:

$$\begin{array}{rccll} s: & \mathcal{Y}^{\mathbb{Z}} & \longrightarrow & \mathcal{Y}^{\mathbb{Z}} & \text{(the shift operator)} \\ & (y_t)_{t\in\mathbb{Z}} & \longrightarrow & (y_{t+1})_{t\in\mathbb{Z}} & \\ l: & \mathcal{Y}^{\mathbb{Z}} & \longrightarrow & \mathcal{C}(\Theta) & \\ & (y_t)_{t\in\mathbb{Z}} & \longrightarrow & U_0(\theta) = P_\theta(y_1|y_{-\infty:0}) & \end{array}$$

Using the standard notation:

$$s^t l = l \circ \underbrace{s \circ \ldots \circ s}_{t \text{ times}}$$

we can rewrite:

$$\frac{1}{T}\sum_{t=1}^{T-1} \log P_\theta\left(Y_{t+1}|Y_{-\infty:t}\right) = \frac{1}{T}\sum_{t=1}^{T-1} s^t l\left(Y\right)$$

Y is stationary ergodic: this is exactly the setting of the ergodic theorem. In the most familiar case, l would be a measurable function from $\mathcal{Y}^{\mathbb{Z}}$ to $\mathbb{R}$, which is not the case here. However, $(\mathcal{C}(\Theta), \|\cdot\|_\infty)$ is separable because Θ is compact: a *vector* almost sure ergodic theorem holds, exactly similar to the scalar case (see theorem 2.1 p.167 in Krengel (1985)).

Thus:

$$\left\| \frac{1}{T}\sum_{t=1}^{T-1} U_t(\theta) - L(\theta) \right\| \xrightarrow{\theta^\star \ as} 0$$

where:

$$L = \mathbb{E}_{\theta^\star}\left[l\left(Y\right)\right] = \mathbb{E}_{\theta^\star}\left[\log P_\theta\left(Y_1|Y_{-\infty:0}\right)\right] \in \mathcal{C}(\Theta)$$

A12.4.2 Central limit theorem for the score

Let $s_T = \nabla_{\theta^\star} L_T(\theta)$ be the (observed) score. This section shows a central limit theorem (pointwise at $\theta^\star$) for the score: there is I such that:

$$\sqrt{T} s_T \overset{\theta^\star}{\rightsquigarrow} \mathcal{N}(0, I)$$

Similar to the proof of the uniform law of large numbers for the log-likelihood (section A12.4.1), the following decomposition is used:

$$\begin{cases} s_T = \frac{1}{T}\sum_{t=1}^{T-1} V_{1t} + o_{\theta^\star}\left(\frac{1}{\sqrt{T}}\right) \\ \left|\sqrt{T}\frac{1}{T}\sum_{t=1}^{T-1} V_{1t} - \sqrt{T}\frac{1}{T}\sum_{t=1}^{T-1} V_t\right| \xrightarrow{\theta^\star \ as} 0 \\ \sqrt{T}\frac{1}{T}\sum_{t=1}^{T-1} V_t \overset{\theta^\star}{\rightsquigarrow} \mathcal{N}(0, I) \end{cases}$$

V_t is the "infinite-past" limit of an auxiliary process V_{mt} (to be defined shortly) as $m \to -\infty$.

In order to define the auxiliary process V_{mt}, note an identity of Louis (1982) (equation (3.1) p. 227), which says in substance:

$$\nabla_{\theta^\star} \log P_\theta(Y) = \mathbb{E}_{\theta^\star}[\nabla_{\theta^\star} \log P_\theta(Y, X)|Y] \tag{10}$$

Here:

$$\begin{aligned}
s_T &= \frac{1}{T}\nabla_{\theta^\star} \log P_\theta\left(Y_{2:T}|Y_1\right) \\
&= \frac{1}{T}\mathbb{E}_{\theta^\star}\left[\nabla_{\theta^\star} \log P_\theta\left(Y_{2:T}, X_{1:T}|Y_1\right)|Y_{1:T}\right] && \text{by (10)} \\
&= \frac{1}{T}\mathbb{E}_{\theta^\star}\left[\nabla_{\theta^\star}\left(\sum_{s=1}^{T-1} \log P_\theta\left(Y_{s+1}|X_{s+1}\right) + \log P_\theta\left(X_{s+1}|X_s\right)\right)\middle| Y_{1:T}\right] \\
&\qquad + \underbrace{\frac{1}{T}\underbrace{\mathbb{E}_{\theta^\star}\left[\nabla_{\theta^\star} \log P_\theta\left(X_1|Y_1\right)|Y_{1:T}\right]}_{\xrightarrow{\theta^\star \ a.s.} \mathbb{E}_{\theta^\star}[\cdot|Y_{1:+\infty}]}}_{=o_{\theta^\star}\left(\frac{1}{\sqrt{T}}\right)} && \text{by conditional independence}
\end{aligned}$$

Now write:

$$J_s = \nabla_{\theta^\star} \log P_\theta\left(Y_{s+1}|X_{s+1}\right) + \nabla_{\theta^\star} \log P_\theta\left(X_{s+1}|X_s\right)$$

and consider the telescopic sum:

$$\begin{aligned}
\mathbb{E}_{\theta^\star}\left[\sum_{s=1}^{T-1} J_s\middle| Y_{1:T}\right] = \quad & \mathbb{E}_{\theta^\star}\left[\sum_{s=1}^{T-1} J_s\middle| Y_{1:T}\right] - \mathbb{E}_{\theta^\star}\left[\sum_{s=1}^{T-2} J_s\middle| Y_{1:T-1}\right] && (= V_{1,T-1}) \\
& + \mathbb{E}_{\theta^\star}\left[\sum_{s=1}^{T-2} J_s\middle| Y_{1:T-1}\right] - \mathbb{E}_{\theta^\star}\left[\sum_{s=1}^{T-3} J_s\middle| Y_{1:T-2}\right] && (= V_{1,T-2}) \\
& + \ldots \\
& + \mathbb{E}_{\theta^\star}\left[J_1|Y_{1:2}\right] && (= V_{1,1})
\end{aligned}$$

Finally introduce the auxiliary process:

$$\begin{aligned}
V_{mt} &= \mathbb{E}_{\theta^\star}\left[\sum_{s=m}^{t} J_s\middle| Y_{m:t+1}\right] - \mathbb{E}_{\theta^\star}\left[\sum_{s=m}^{t-1} J_s\middle| Y_{m:t}\right] \\
&= \mathbb{E}_{\theta^\star}\left[J_t|Y_{m:t+1}\right] + \sum_{s=m}^{t-1}\left(\mathbb{E}_{\theta^\star}\left[J_s|Y_{m:t+1}\right] - \mathbb{E}_{\theta^\star}\left[J_s|Y_{m:t}\right]\right)
\end{aligned}$$

As announced:

$$s_T = \frac{1}{T}\sum_{t=1}^{T-1} V_{1t} + o_{\theta^\star}\left(\frac{1}{\sqrt{T}}\right)$$

The remainder of the proof proceeds in three steps:

1. V_{mt} is a ($\theta^\star$ almost-sure) Cauchy sequence for $m \to -\infty$. We call V_t its limit.
2. $\left|\sqrt{T}\frac{1}{T}\sum_{t=1}^{T-1} V_{1t} - \sqrt{T}\frac{1}{T}\sum_{t=1}^{T-1} V_t\right| \xrightarrow{\theta^\star \ as} 0$.
3. $\sqrt{T}\frac{1}{T}\sum_{t=1}^{T-1} V_t \overset{\theta^\star}{\rightsquigarrow} \mathcal{N}(0, I)$.

Step 1: V_{mt} is $\theta^\star$ almost surely Cauchy

Similar to step 1 in section A12.4.1, we want to define $V_t = V_{-\infty t}$ and show the following $\theta^\star$ almost sure inequality for $-\infty \leq m < m' \leq 1$:

$$|V_{m't} - V_{mt}| < K\rho^{t-m'} \tag{11}$$

For now fix $-\infty < m < m' \leq 1$. We split the sum $|V_{m't} - V_{mt}|$ into four regions. Call $k = \lfloor \frac{m'+t}{2} \rfloor$.

$$\begin{aligned}
|V_{m't} - V_{mt}| = &\left|\left(\mathbb{E}_{\theta^\star}\left[\sum_{s=m'}^{t} J_s \middle| Y_{m':t+1}\right] - \mathbb{E}_{\theta^\star}\left[\sum_{s=m'}^{t-1} J_s \middle| Y_{m':t}\right]\right)\right. \\
&\left. - \left(\mathbb{E}_{\theta^\star}\left[\sum_{s=m}^{t} J_s \middle| Y_{m:t+1}\right] - \mathbb{E}_{\theta^\star}\left[\sum_{s=m}^{t-1} J_s \middle| Y_{m:t}\right]\right)\right| \\
\leq &\sum_{s=k}^{t} |\mathbb{E}_{\theta^\star}[J_s|Y_{m':t+1}] - \mathbb{E}_{\theta^\star}[J_s|Y_{m:t+1}]| + \sum_{s=k}^{t-1} |\mathbb{E}_{\theta^\star}[J_s|Y_{m':t}] - \mathbb{E}_{\theta^\star}[J_s|Y_{m:t}]| \\
&+ \sum_{s=m'}^{k-1} |\mathbb{E}_{\theta^\star}[J_s|Y_{m':t+1}] - \mathbb{E}_{\theta^\star}[J_s|Y_{m':t}]| + \sum_{s=m}^{k-1} |\mathbb{E}_{\theta^\star}[J_s|Y_{m:t+1}] - \mathbb{E}_{\theta^\star}[J_s|Y_{m:t}]|
\end{aligned}$$

The geometric bound (11) for V_{mt} is then a consequence of the following $\theta^\star$ almost sure geometric bounds, one for each region:

$$|\mathbb{E}_{\theta^\star}[J_s|Y_{m':t+1}] - \mathbb{E}_{\theta^\star}[J_s|Y_{m:t+1}]| \leq K\rho^{s-m'} \tag{12}$$

$$|\mathbb{E}_{\theta^\star}[J_s|Y_{m':t}] - \mathbb{E}_{\theta^\star}[J_s|Y_{m:t}]| \leq K\rho^{s-m'} \tag{13}$$

$$|\mathbb{E}_{\theta^\star}[J_s|Y_{m':t+1}] - \mathbb{E}_{\theta^\star}[J_s|Y_{m':t}]| \leq K\rho^{t-s} \tag{14}$$

$$|\mathbb{E}_{\theta^\star}[J_s|Y_{m:t+1}] - \mathbb{E}_{\theta^\star}[J_s|Y_{m:t}]| \leq K\rho^{t-s} \tag{15}$$

Bounds (12) and (13) are a consequence of the merging properties of $(X_t|y_{m:T})_{t\geq m}$, which were proven in lemma 3. Bounds (14) and (15) are a consequence of the merging properties of another Markov chain, namely $(X_{T-t}|y_{m:T})_{0\leq t\leq T-m}$ (note the reverse time). The merging of $(X_{T-t}|y_{m:T})_{0\leq t\leq T-m}$ as well as the bounds (12), (13), (14) and (15), is proven similarly to the corresponding results in the log-likelihood section (section A12.4.1 lemma 3 and bound (7)). The proofs are omitted for brevity.

Note that (12), (13) and (15) extend to $m=-\infty$ because:

$$\mathbb{E}_{\theta^\star}\left[J_s|Y_{m:t}\right] \xrightarrow{\theta^\star\ as} \mathbb{E}_{\theta^\star}\left[J_s|Y_{-\infty:t}\right]$$

As a consequence, $\mathbb{E}_{\theta^\star}\left[J_s|Y_{-\infty:t+1}\right]-\mathbb{E}_{\theta^\star}\left[J_s|Y_{-\infty:t}\right]$ is (almost surely) absolutely summable in $s\to-\infty$ and $\sum_{s=-\infty}^{t-1}\left(\mathbb{E}_{\theta^\star}\left[J_s|Y_{-\infty:t+1}\right]-\mathbb{E}_{\theta^\star}\left[J_s|Y_{-\infty:t}\right]\right)$ is well-defined. We can legitimately define:

$$V_t = V_{-\infty t} = \mathbb{E}_{\theta^\star}\left[J_t|Y_{-\infty:t+1}\right] + \sum_{s=-\infty}^{t-1}\left(\mathbb{E}_{\theta^\star}\left[J_s|Y_{-\infty:t+1}\right]-\mathbb{E}_{\theta^\star}\left[J_s|Y_{-\infty:t}\right]\right)$$

(In a few lines it will be shown that $V_t=V_{-\infty t}$ is the (almost sure) limit of V_{mt} as $m\to-\infty$.)

With this definition, the inequality:

$$\begin{aligned}|V_{m't}-V_{mt}| \leq &\sum_{s=k}^{t}\left|\mathbb{E}_{\theta^\star}\left[J_s|Y_{m':t+1}\right]-\mathbb{E}_{\theta^\star}\left[J_s|Y_{m:t+1}\right]\right| + \sum_{s=k}^{t-1}\left|\mathbb{E}_{\theta^\star}\left[J_s|Y_{m':t}\right]-\mathbb{E}_{\theta^\star}\left[J_s|Y_{m:t}\right]\right| \\ &+\sum_{s=m'}^{k-1}\left|\mathbb{E}_{\theta^\star}\left[J_s|Y_{m':t+1}\right]-\mathbb{E}_{\theta^\star}\left[J_s|Y_{m':t}\right]\right| + \sum_{s=m}^{k-1}\left|\mathbb{E}_{\theta^\star}\left[J_s|Y_{m:t+1}\right]-\mathbb{E}_{\theta^\star}\left[J_s|Y_{m:t}\right]\right|\end{aligned}$$

is valid for $m=-\infty$ too, and by (12), (13), (14) and (15):

$$\begin{aligned}|V_{m't}-V_{mt}| &\leq \sum_{s=k}^{t} c\rho^{s-m'} + \sum_{s=k}^{t-1} c\rho^{s-m'} + \sum_{s=m'}^{k-1} c\rho^{t-s} + \sum_{s=m}^{k-1} c\rho^{t-s} \\ &\leq c(\rho^{k-m'}+\rho^{k-m'}+\rho^{t-k}+\rho^{t-k}) \\ &\leq c\sqrt{\rho}^{t-m'}\end{aligned}$$

In particular, $V_{mt}\to V_t$ as $m\to-\infty$.

Step 2: $\left|\sqrt{T}\frac{1}{T}\sum_{t=1}^{T-1}V_{1t} - \sqrt{T}\frac{1}{T}\sum_{t=1}^{T-1}V_t\right| \xrightarrow{\theta^\star\ as} 0$

From the geometric bound (11) (with $m=-\infty$ and $m'=1$), we have (almost surely):

$$\begin{aligned}\left|\sqrt{T}\frac{1}{T}\sum_{t=1}^{T-1}V_{1t}-\sqrt{T}\frac{1}{T}\sum_{t=1}^{T-1}V_t\right| &\leq \sqrt{T}\frac{1}{T}\sum_{t=1}^{T-1}|V_{1t}-V_t| \\ &\leq K\sqrt{T}\frac{1}{T}\sum_{t=1}^{T-1}\rho^{t-1} \\ &\leq K\sqrt{T}\frac{1}{T}\frac{1}{1-\rho} \\ &\xrightarrow[T\to\infty]{} 0\end{aligned}$$

Step 3: $\frac{1}{T}\sum_{t=1}^{T-1}V_t \overset{\theta^\star}{\rightsquigarrow} \mathcal{N}(0,I)$
Remember that:

$$V_t = \mathbb{E}_{\theta^\star}\left[J_t|Y_{-\infty:t+1}\right] + \sum_{s=-\infty}^{t-1}\left(\mathbb{E}_{\theta^\star}\left[J_s|Y_{-\infty:t+1}\right] - \mathbb{E}_{\theta^\star}\left[J_s|Y_{-\infty:t}\right]\right)$$

V_t is immediately ergodic stationary because (X,Y) is. We show that it is also a martingale difference sequence with respect to the $\sigma\left(Y_{-\infty:t+1}\right)$ filtration. Note that for $m > -\infty$:

$$\begin{aligned}\mathbb{E}_{\theta^\star}\left[\sum_{s=m}^{t-1}\left(\mathbb{E}_{\theta^\star}\left[J_s|Y_{-\infty:t+1}\right] - \mathbb{E}_{\theta^\star}\left[J_s|Y_{-\infty:t}\right]\right)\middle| Y_{-\infty:t}\right] &= \sum_{s=m}^{t-1}\left(\mathbb{E}_{\theta^\star}\left[J_s|Y_{-\infty:t}\right] - \mathbb{E}_{\theta^\star}\left[J_s|Y_{-\infty:t}\right]\right) \\ &= 0\end{aligned}$$

Thus by dominated convergence:

$$\mathbb{E}_{\theta^\star}\left[\sum_{s=-\infty}^{t-1}\left(\mathbb{E}_{\theta^\star}\left[J_s|Y_{-\infty:t+1}\right] - \mathbb{E}_{\theta^\star}\left[J_s|Y_{-\infty:t}\right]\right)\middle| Y_{-\infty:t}\right] = 0$$

So that:

$$\begin{aligned}\mathbb{E}_{\theta^\star}\left[V_t|Y_{-\infty:t}\right] &= \mathbb{E}_{\theta^\star}\left[J_t|Y_{-\infty:t}\right] + 0 \\ &= \mathbb{E}_{\theta^\star}\left[\mathbb{E}_{\theta^\star}\left[d\log P\left(X_{t+1},Y_{t+1}|X_t,Y_t\right)|X_t,Y_t\right]|Y_{-\infty:t}\right] \\ &= 0 \qquad\qquad \text{(expectation of the conditional score)}\end{aligned}$$

By the central limit theorem for ergodic stationary martingale difference sequences:

$$\sqrt{T}\frac{1}{T}\sum_{t=1}^{T-1}V_t \overset{\theta^\star}{\rightsquigarrow} \mathcal{N}(0,I) \qquad \text{where } I = \mathbb{E}_{\theta^\star}[V_1V_1']$$

A12.4.3 Uniform law of large numbers for the observed information

Let $i_T(\theta) = -\nabla^2_\theta L_T(\theta)$ be the (observed) information. This section shows a uniform law of large numbers for the information:

$$\|i_T(\theta) - i(\theta)\| \xrightarrow{\theta^\star \ as} 0$$

where $i(\theta)$ will be defined below and $i(\theta^\star) = I$ is the asymptotic variance of the score (see section A12.4.2).

Similar to the proofs of the uniform law of large numbers for the log-likelihood (section A12.4.1) and of the central limit theorem for the score (section A12.4.2), the following decomposition is used:

$$\begin{cases} i_T(\theta) = \frac{1}{T}\sum_{t=1}^{T-1} W^e_{1t}(\theta) + \frac{1}{T}\sum_{t=1}^{T-1} W^v_{1t}(\theta) + \frac{1}{T}\sum_{t=1}^{T-1} W^c_{1t}(\theta) + o_{\theta^\star}(1) \\ \left\| \frac{1}{T}\sum_{t=1}^{T-1} W^e_{1t}(\theta) - \frac{1}{T}\sum_{t=1}^{T-1} W^e_{t}(\theta) \right\| \xrightarrow{\theta^\star \ as} 0 \\ \left\| \frac{1}{T}\sum_{t=1}^{T-1} W^v_{1t}(\theta) - \frac{1}{T}\sum_{t=1}^{T-1} W^v_{t}(\theta) \right\| \xrightarrow{\theta^\star \ as} 0 \\ \left\| \frac{1}{T}\sum_{t=1}^{T-1} W^c_{1t}(\theta) - \frac{1}{T}\sum_{t=1}^{T-1} W^c_{t}(\theta) \right\| \xrightarrow{\theta^\star \ as} 0 \\ \left\| \frac{1}{T}\sum_{t=1}^{T-1} W^e_{t}(\theta) - i^e(\theta) \right\| \xrightarrow{\theta^\star \ as} 0 \\ \left\| \frac{1}{T}\sum_{t=1}^{T-1} W^v_{t}(\theta) - i^v(\theta) \right\| \xrightarrow{\theta^\star \ as} 0 \\ \left\| \frac{1}{T}\sum_{t=1}^{T-1} W^c_{t}(\theta) - i^c(\theta) \right\| \xrightarrow{\theta^\star \ as} 0 \end{cases}$$

$W^e_t(\theta)$, $W^v_t(\theta)$ and $W^c_t(\theta)$ are the uniform Cauchy "infinite-past" limits of the auxiliary processes $W^e_{mt}(\theta)$, $W^v_{mt}(\theta)$ and $W^c_{mt}(\theta)$ (to be defined shortly) respectively, as $m \to -\infty$.

In order to define the auxiliary processes $W^e_{mt}(\theta)$, $W^v_{mt}(\theta)$ and $W^c_{mt}(\theta)$, note another identity of Louis (1982) (equation (3.2) p. 227), which says in substance:

$$\nabla^2_{\theta^\star} \log P_\theta(Y) = \mathbb{E}_{\theta^\star}[\nabla^2_{\theta^\star} \log P_\theta(Y, X)|Y] + V_{\theta^\star}[\nabla_{\theta^\star} \log P_\theta(Y, X)|Y] \tag{16}$$

Define:

$$J_s(\theta) = \nabla_\theta \log P_\theta\left(Y_{s+1}|X_{s+1}\right) + \nabla_\theta \log P_\theta\left(X_{s+1}|X_s\right)$$

$$H_s(\theta) = \nabla^2_\theta \log P_\theta\left(Y_{s+1}|X_{s+1}\right) + \nabla^2_\theta \log P_\theta\left(X_{s+1}|X_s\right)$$

(In particular, J_s in the previous section is $J_s(\theta^\star)$ according to this definition.)

We use the following convention as a notation for conditional expectations defined under one

measure but evaluated under a different measure. Suppose $g(Z_2) = \mathbb{E}_\theta[Z_1|Z_2]$ is a version of Kolmogorov's conditional expectation under θ. We will write $E_\theta[Z_1|Z_2]$ for $g(Z_2)$ where Z_2 can be distributed according to $\theta^\star \neq \theta$. Note that this is well-defined as long as θ and $\theta^\star$ are probabilities defined on the same background σ-field, in the sense that g's Z_2-measurability depends only on the σ-fields. We follow the same convention for conditional variances.

Now by (16) and conditional independence:

$$\begin{aligned}
i_T(\theta) &= \frac{1}{T}\nabla_\theta^2 \log P_\theta\left(Y_{2:T}|Y_1\right) \\
&= \frac{1}{T}E_\theta\left[\nabla_\theta^2 \log P_\theta\left(Y_{2:T}, X_{1:T}|Y_1\right)\middle|Y_{1:T}\right] + \frac{1}{T}V_\theta\left[\nabla_\theta \log P_\theta\left(Y_{2:T}, X_{1:T}|Y_1\right)|Y_{1:T}\right] \\
&= \frac{1}{T}E_\theta\left[\sum_{s=1}^{T-1} H_s(\theta)\middle|Y_{1:T}\right] + \frac{1}{T}V_\theta\left[\sum_{s=1}^{T-1} J_s(\theta)\middle|Y_{1:T}\right] + \frac{1}{T}E_\theta\left[\nabla_\theta^2 \log P_\theta\left(X_1|Y_1\right)\middle|Y_{1:T}\right] \\
&\qquad + \frac{1}{T}2Cov_\theta\left[\sum_{s=1}^{T-1} J_s(\theta), \nabla_\theta \log P_\theta\left(X_1|Y_1\right)\middle|Y_{1:T}\right] + \frac{1}{T}V_\theta\left[\nabla_\theta \log P_\theta\left(X_1|Y_1\right)|Y_{1:T}\right]
\end{aligned}$$

$E_\theta\left[\nabla_\theta^2 \log P_\theta\left(X_1|Y_1\right)|Y_{1:T}\right]$ and $V_\theta\left[\nabla_\theta \log P_\theta\left(X_1|Y_1\right)|Y_{1:T}\right]$ converge to some "infinite-past" limit by Cauchy-ness, using the same proof strategy we have used repeatedly, so that:

$$\frac{1}{T}E_\theta\left[\nabla_\theta^2 \log P_\theta\left(X_1|Y_1\right)\middle|Y_{1:T}\right] + \frac{1}{T}V_\theta\left[\nabla_\theta \log P_\theta\left(X_1|Y_1\right)|Y_{1:T}\right] = o_{\theta^\star}(1)$$

We introduce the auxiliary processes:

$$\begin{aligned}
W_{mt}^e(\theta) &= E_\theta\left[\sum_{s=m}^{t} H_s(\theta)\middle|Y_{m:t+1}\right] - E_\theta\left[\sum_{s=m}^{t-1} H_s(\theta)\middle|Y_{m:t}\right] \\
W_{mt}^v(\theta) &= V_\theta\left[\sum_{s=m}^{t} J_s(\theta)\middle|Y_{m:t+1}\right] - V_\theta\left[\sum_{s=m}^{t-1} J_s(\theta)\middle|Y_{m:t}\right] \\
W_{mt}^c(\theta) &= Cov_\theta\left[\sum_{s=m}^{t} J_s(\theta), \nabla_\theta \log P_\theta\left(X_1|Y_1\right)\middle|Y_{m:t+1}\right] - Cov_\theta\left[\sum_{s=m}^{t-1} J_s(\theta), \nabla_\theta \log P_\theta\left(X_1|Y_1\right)\middle|Y_{m:t}\right]
\end{aligned}$$

Using telescopic sums similar to what was done for the score in section A12.4.2, we have, as announced:

$$i_T(\theta) = \frac{1}{T}\sum_{t=1}^{T-1} W_{1t}^e(\theta) + \frac{1}{T}\sum_{t=1}^{T-1} W_{1t}^v(\theta) + \frac{1}{T}\sum_{t=1}^{T-1} W_{1t}^c(\theta) + o_{\theta^\star}(1)$$

The remainder of the proof proceeds in three steps:

1. W_{mt}^e, W_{mt}^v and W_{mt}^c are $\theta^\star$ almost sure uniform Cauchy sequences as $m \to -\infty$. We call $W_t^e(\theta)$, $W_t^v(\theta)$ and $W_t^c(\theta)$ their limits.

2. $\left\|\frac{1}{T}\sum_{t=1}^{T-1} W_{1t}^{e}(\theta) - \frac{1}{T}\sum_{t=1}^{T-1} W_{t}^{e}(\theta)\right\| \xrightarrow{\theta^\star \ as} 0$, $\left\|\frac{1}{T}\sum_{t=1}^{T-1} W_{1t}^{v}(\theta) - \frac{1}{T}\sum_{t=1}^{T-1} W_{t}^{v}(\theta)\right\| \xrightarrow{\theta^\star \ as} 0$ and $\left\|\frac{1}{T}\sum_{t=1}^{T-1} W_{1t}^{c}(\theta) - \frac{1}{T}\sum_{t=1}^{T-1} W_{t}^{c}(\theta)\right\| \xrightarrow{\theta^\star \ as} 0$.

3. $\left\|\frac{1}{T}\sum_{t=1}^{T-1} W_{t}^{e}(\theta) - i^{e}(\theta)\right\| \xrightarrow{\theta^\star \ as} 0$, $\left\|\frac{1}{T}\sum_{t=1}^{T-1} W_{t}^{v}(\theta) - i^{v}(\theta)\right\| \xrightarrow{\theta^\star \ as} 0$ and $\left\|\frac{1}{T}\sum_{t=1}^{T-1} W_{t}^{c}(\theta) - i^{c}(\theta)\right\| \xrightarrow{\theta^\star \ as} 0$.

Step 1.1: $W_{mt}^{e}(\theta)$ is $\theta^\star$ almost surely uniform Cauchy

Similar to step 1 in the log-likelihood section (section A12.4.1) and step 1 in the score section (section A12.4.2), we define $W_{t}^{e}(\theta) = W_{-\infty t}^{e}(\theta)$ and show the following $\theta^\star$ almost-sure inequality, valid for any $-\infty \leq m < m' \leq 1$ and uniform in θ:

$$|W_{m't}^{e}(\theta) - W_{mt}^{e}(\theta)| < K\rho^{t-m'} \tag{17}$$

The proof follows the same lines of step 1 in the score section (section A12.4.2). The only difference is that we cannot use $\mathbb{E}_{\theta^\star}\left[\cdot|Y_{m:t+1}\right] \xrightarrow{\theta^\star \ as} \mathbb{E}_{\theta^\star}\left[\cdot|Y_{-\infty:t+1}\right]$ when dealing with $E_{\theta}\left[\cdot|Y_{m:t+1}\right]$ instead of $\mathbb{E}_{\theta^\star}\left[\cdot|Y_{m:t+1}\right]$. We have to first use Cauchy-ness to take the limit and then extend the geometric bound of interest to the limit, exactly similar to what is done in step 1 in the log-likelihood section (section A12.4.1) and in the next step (step 1.2) for W_{mt}^{v}. The proof is omitted for brevity.

Step 1.2: $W_{mt}^{v}(\theta)$ is $\theta^\star$ almost surely uniform Cauchy

Similar to step 1 in the log-likelihood section (section A12.4.1), step 1 in the score section (section A12.4.2) and step 1.1 for $W_{mt}^{e}(\theta)$, we define $W_{t}^{v}(\theta) = W_{-\infty t}^{v}(\theta)$ and show the following $\theta^\star$ almost-sure inequality, valid for any $-\infty \leq m < m' \leq 1$ and uniform in θ:

$$|W_{m't}^{v}(\theta) - W_{mt}^{v}(\theta)| < K\rho^{t-m'} \tag{18}$$

The following $\theta^\star$ almost-sure bounds hold for any $-\infty < m \leq m' \leq r \leq s \leq t$, uniformly in θ:

$$|Cov_{\theta}\left[J_r(\theta), J_s(\theta)|Y_{m':t}\right] - Cov_{\theta}\left[J_r(\theta), J_s(\theta)|Y_{m:t}\right]| \leq K\rho^{r-m'} \tag{19}$$

$$|Cov_{\theta}\left[J_r(\theta), J_s(\theta)|Y_{m:t+1}\right] - Cov_{\theta}\left[J_r(\theta), J_s(\theta)|Y_{m:t}\right]| \leq K\rho^{t-s} \tag{20}$$

$$|Cov_{\theta}\left[J_r(\theta), J_s(\theta)|Y_{m':t}\right]| \leq K\rho^{s-r} \tag{21}$$

The proofs of the bounds (19), (20) and (21) follow from the merging properties along the same lines of bounds (12), (13), (14) and (15) in the score section (section A12.4.2) and

bound (7) in the log-likelihood section (section A12.4.1). They are omitted for brevity.

Note that for any $-\infty < m < m' \leq 1$:

$$W^v_{m't}(\theta) - W^v_{mt}(\theta) = \left(\sum_{r=m'}^{t} \sum_{s=m'}^{t} Cov_\theta \left[J_r(\theta), J_s(\theta) | Y_{m':t+1} \right] - \sum_{r=m'}^{t-1} \sum_{s=m'}^{t-1} Cov_\theta \left[J_r(\theta), J_s(\theta) | Y_{m':t} \right] \right) - \left(\sum_{r=m}^{t} \sum_{s=m}^{t} Cov_\theta \left[J_r(\theta), J_s(\theta) | Y_{m:t+1} \right] - \sum_{r=m}^{t-1} \sum_{s=m}^{t-1} Cov_\theta \left[J_r(\theta), J_s(\theta) | Y_{m:t} \right] \right) \tag{22}$$

The bounds (19), (20) or (21) apply to different ways of grouping the terms in the sum (22). The idea is to split the sums into different regions, and then apply the sharpest bound available in each region. This is what was done explictly in the log-likelihood and score sections; the "region management" becomes too cumbersome here. We use a higher level approach. Fix m,m' and t and partition $A = m \leq r \leq t, m \leq s \leq t$ as $A_1 \cup A_2 \cup A_3 \cup A_4$ (see also figure 8) where:

$$\begin{aligned}
A_1 &= \{m \leq r < m', m \leq s \leq t-1\} \cup \{m \leq r \leq t-1, m \leq s < m'\} \\
A_2 &= \{m' \leq r \leq t-1, m' \leq s \leq t-1\} \\
A_3 &= \{r = t, m' \leq s \leq t\} \cup \{m' \leq s \leq t, s = t\} \\
A_4 &= \{r = t, m \leq s < m'\} \cup \{m \leq r < m', s = t, \}
\end{aligned}$$

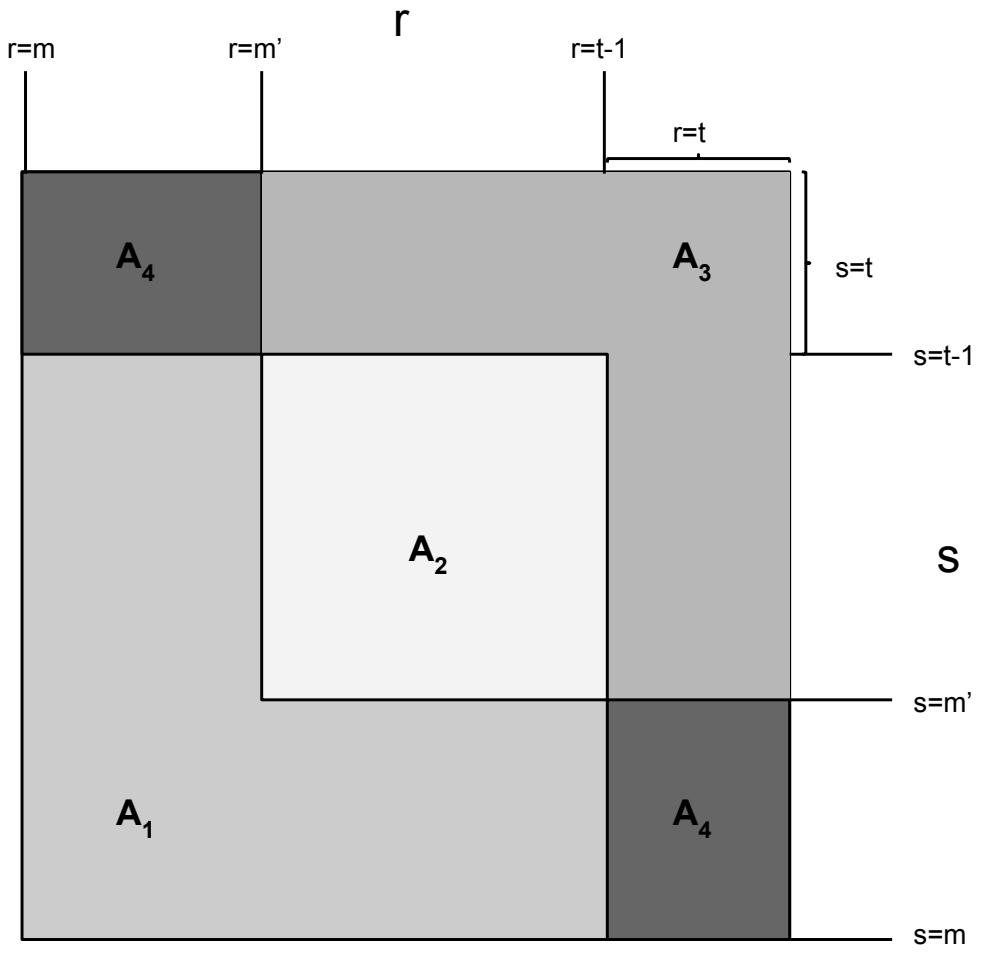


Figure 8: Partition of A

Note that:

$$\begin{aligned} W^v_{m't}(\theta) - W^v_{mt}(\theta) = & \sum_{A_1} \left(Cov_\theta \left[J_r(\theta), J_s(\theta) | Y_{m:t+1} \right] - Cov_\theta \left[J_r(\theta), J_s(\theta) | Y_{m:t} \right] \right) \\ & + \sum_{A_2} \left(\left(Cov_\theta \left[J_r(\theta), J_s(\theta) | Y_{m':t+1} \right] - Cov_\theta \left[J_r(\theta), J_s(\theta) | Y_{m':t} \right] \right) \right. \\ & \qquad \left. - \left(Cov_\theta \left[J_r(\theta), J_s(\theta) | Y_{m:t+1} \right] - Cov_\theta \left[J_r(\theta), J_s(\theta) | Y_{m:t} \right] \right) \right) \\ & + \sum_{A_3} \left(Cov_\theta \left[J_r(\theta), J_s(\theta) | Y_{m':t+1} \right] - Cov_\theta \left[J_r(\theta), J_s(\theta) | Y_{m:t+1} \right] \right) \\ & + \sum_{A_4} Cov_\theta \left[J_r(\theta), J_s(\theta) | Y_{m:t+1} \right] \end{aligned}$$

By the bounds (19), (20) and (21):

$$|W^v_{m't}(\theta) - W^v_{mt}(\theta)| \leq \sum_{A_1} K\rho^{t-r\vee s} \wedge \rho^{r\vee s - r\wedge s} + \sum_{A_2} K\rho^{t-r\vee s} \wedge \rho^{r\vee s - r\wedge s} \wedge \rho^{r\wedge s - m'} + \sum_{A_3} K\rho^{r\vee s - r\wedge s} \wedge \rho^{r\wedge s - m'} + \sum_{A_4} K\rho^{r\vee s - r\wedge s}$$

Furthermore:

- On A_1, $r \wedge s - m' \leq 0 \leq r \vee s - r \wedge s$.
- On A_3, $t - r \vee s = 0 \leq r \vee s - r \wedge s$.
- On A_4, $t - r \vee s = 0 \leq r \vee s - r \wedge s$ and $r \wedge s - m' \leq 0 \leq r \vee s - r \wedge s$.

So that:

$$\begin{aligned} |W^v_{m't}(\theta) - W^v_{mt}(\theta)| &\leq K \sum_{A_1 \cup A_2 \cup A_3 \cup A_4} \rho^{t-r\vee s} \wedge \rho^{r\vee s - r\wedge s} \wedge \rho^{r\wedge s - m'} \\ &= K \sum_{r=m}^{t} \sum_{s=m}^{t} \rho^{(t-r\vee s)\vee(r\vee s - r\wedge s)\vee(r\wedge s - m')} \end{aligned}$$

Define $n = t - m + 1$, $\rho_n = \rho^n$, $a = \frac{m'-m+1}{t-m+1}$ and the function $g(r,s) := (1 - r \vee s) \vee (r \vee s - r \wedge s) \vee (r \wedge s - a)$ for $(r,s) \in [0,1] \times [0,1]$. Then:

$$\begin{aligned} &\sum_{r=m}^{t} \sum_{s=m}^{t} \rho^{(t-r\vee s)\vee(r\vee s - r\wedge s)\vee(r\wedge s - m')} \\ &\quad = n^2 \frac{1}{n^2} \sum_{r=1}^{n} \sum_{s=1}^{n} a^{g\left(\frac{r}{n}, \frac{s}{n}\right)} \\ &\quad \leq n^2 \int_{0\leq r\leq 1} \int_{0\leq s\leq 1} a^{g(r,s)} dr ds \qquad \text{because } x \to a^x \text{ is decreasing} \\ &\quad = n^2 6 \frac{3a^{1/3} - 4a^{1/2} + a}{\log^2 a} \\ &\quad \leq n^2 6 \frac{3a^{1/3}}{\log^2 a} \qquad \text{because } a - 4\sqrt{a} < 0 \text{ when } 0 < a < 1 \\ &\quad = \frac{18K}{\rho} n^2 \frac{1}{n^2 \log^2 \rho} \rho^{n/3} \\ &\quad \leq \frac{18K}{\rho \log^2 \rho} \left(\rho^{1/3}\right)^{t-m'} \end{aligned}$$

This proves (18) for $-\infty \leq m < m' \leq 1$. As a consequence $W^v_{mt}(\theta)$ is $\theta^\star$ almost surely a uniform Cauchy sequence, and as such converges to a limit $W^v_t(\theta) = W^v_{\infty t}(\theta)$ as $m \to -\infty$.

(18) extends to $m = -\infty$ by continuity.

In order to give an explicit "infinite-past" representation for $W_t^v(\theta)$ which will be used to apply the ergodic theorem in step 3.2, note that (19) implies that:

$$Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{m:t+1}\right] \xrightarrow{\theta^\star \ as} Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{-\infty:t+1}\right]$$

Then bounds (19), (20) and (21) extend to $m = -\infty$ and imply that $Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{-\infty:t+1}\right] - Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{-\infty:t}\right]$ is (doubly) absolutely summable in $r \to -\infty$ and $s \to -\infty$, and $Cov_\theta\left[J_s(\theta), J_t(\theta)|Y_{-\infty:t+1}\right]$ is absolutely summable in $s \to -\infty$. Rewrite W_{mt}^v:

$$\begin{aligned}
W_{mt}^v(\theta) &= \sum_{r=m'}^{t}\sum_{s=m'}^{t} Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{m':t+1}\right] - \sum_{r=m'}^{t-1}\sum_{s=m'}^{t-1} Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{m':t}\right] \\
&= \sum_{s=m}^{t} Cov_\theta\left[J_s(\theta), J_t(\theta)|Y_{m:t+1}\right] + \sum_{s=m}^{t-1} Cov_\theta\left[J_s(\theta), J_t(\theta)|Y_{m:t+1}\right] \\
&\quad + \sum_{r=m}^{t-1}\sum_{s=m}^{t-1}\left(Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{m:t+1}\right] - Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{m:t}\right]\right)
\end{aligned}$$

By taking m to $-\infty$ we get the "infinite-past" representation of $W_{mt}^v(\theta)$:

$$\begin{aligned}
W_t^v(\theta) = W_{-\infty t}^v(\theta) &= \sum_{s=-\infty}^{t} Cov_\theta\left[J_s(\theta), J_t(\theta)|Y_{-\infty:t+1}\right] + \sum_{s=-\infty}^{t-1} Cov_\theta\left[J_s(\theta), J_t(\theta)|Y_{-\infty:t+1}\right] \\
&\quad + \sum_{r=-\infty}^{t-1}\sum_{s=-\infty}^{t-1}\left(Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{-\infty:t+1}\right] - Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{-\infty:t}\right]\right)
\end{aligned}$$

Step 1.3: $W_{mt}^c(\theta)$ is $\theta^\star$ almost surely uniform Cauchy

Similar to step 1 in the log-likelihood section (section A12.4.1), step 1 in the score section (section A12.4.2), step 1.1 for $W_{mt}^e(\theta)$ and step 1.2 for $W_{mt}^v(\theta)$. We define $W_t^c(\theta) = W_{-\infty t}^c(\theta)$ and show the following $\theta^\star$ almost-sure inequality, valid for any $-\infty \leq m < m' \leq 1$ and uniform in θ:

$$\left|W_{m't}^c(\theta) - W_{mt}^c(\theta)\right| < K\rho^{t-m'} \tag{23}$$

Along the same lines as steps 1.1 and 1.3 and omitted for brevity.

Step 2.1: $\left|\frac{1}{T}\sum_{t=1}^{T-1} W_{1t}^e(\theta) - \frac{1}{T}\sum_{t=1}^{T-1} W_t^e(\theta)\right| \xrightarrow{\theta^\star \ as} 0.$

Step 2.2: $\left|\frac{1}{T}\sum_{t=1}^{T-1} W_{1t}^v(\theta) - \frac{1}{T}\sum_{t=1}^{T-1} W_t^v(\theta)\right| \xrightarrow{\theta^\star \ as} 0.$

Step 2.3: $\left|\frac{1}{T}\sum_{t=1}^{T-1} W_{1t}^c(\theta) - \frac{1}{T}\sum_{t=1}^{T-1} W_t^c(\theta)\right| \xrightarrow{\theta^\star \ as} 0.$

These three steps follow from the geometric bounds (17), (18) and (23) similarly to step 2 in the log-likelihood section (section A12.4.1) and step 2 in the score section (section A12.4.2).

Step 3.1: $\left\| \frac{1}{T} \sum_{t=1}^{T-1} W_{1t}^e(\theta) - i^e(\theta) \right\| \xrightarrow{\theta^\star \ as} 0.$
Step 3.2: $\left\| \frac{1}{T} \sum_{t=1}^{T-1} W_{1t}^v(\theta) - i^v(\theta) \right\| \xrightarrow{\theta^\star \ as} 0.$
Step 3.3: $\left\| \frac{1}{T} \sum_{t=1}^{T-1} W_{1t}^c(\theta) - i^c(\theta) \right\| \xrightarrow{\theta^\star \ as} 0.$

These two steps follow by the functional ergodic theorem. Their infinite-past representations show that $W_{1t}^e(\theta)$, $W_{1t}^v(\theta)$ and $W_{1t}^c(\theta)$ are ergodic, and continuity follows from the uniform limits. The functional ergodic theorem applies exactly as in step 3 of the log-likelihood section (section A12.4.1).

Note that $i(\theta^\star) = I$, for the usual reason, namely:

$$\begin{aligned}
T i_T(\theta) &= -\nabla_\theta^2 \log P_\theta(Y_{2:T}|Y_1) \\
&= -\frac{\nabla_\theta^2 P_\theta(Y_{2:T}|Y_1)}{P_\theta^2(Y_{2:T}|Y_1)} + \frac{\nabla_\theta P_\theta(Y_{2:T}|Y_1) \nabla_\theta' P_\theta(Y_{2:T}|Y_1)}{P_\theta^2(Y_{2:T}|Y_1)} \\
T\mathbb{E}_{\theta^\star}[i_T(\theta^\star)|Y_1] &= -\underbrace{\sum_{y_{2:T}} \nabla_{\theta^\star}^2 P_{\theta^\star}(y_{2:T}|Y_1)}_{=\nabla^2 \sum = \nabla^2 1 = 0} + T^2 \mathbb{E}_{\theta^\star}\left[s_T s_T'|Y_1\right] \\
\mathbb{E}_{\theta^\star}[i_T(\theta^\star)] &\to i(\theta^\star) \\
\text{and } \mathbb{E}_{\theta^\star}[i_T(\theta^\star)] &= T\mathbb{E}_{\theta^\star}[s_T s_T'] \to I
\end{aligned}$$

A12.5 Hidden Rust models are uniformly locally asymptotically normal (proof of Theorem 3)

Under stationarity, I proved a central limit theorem for the score in appendix section A12.4.2 and a uniform law of large numbers for the observed information in appendix section A12.4.3. (Uniform) local asymptotic normality (Theorem 3) is a standard consequence of these two theorems (see van der Vaart (1998)).

By extending these two limit theorems to their non-stationary versions one gets (uniform) local asymptotic normality (Theorem 3) under non-stationarity too. In A12.5.2 I do so for the central limit theorem for the score. The uniform law of large numbers for the observed information can be extended to non-stationary data along the same lines.

A12.5.1 Limit theorems under non-stationarity

The limit theorems proved under stationarity need to be extended to non-stationarity, when the true initial distribution $\mu^\star$ is different from the stationary one $\mu^\diamond$. Let us consider the case of the central limit theorem.

For Markov chains, non-stationary central limit theorems such as:

$$\sqrt{T}\frac{1}{T}\sum_{t=1}^{T}(f(Y_t) - \mathbb{E}[f(Y_t)]) \Rightarrow \mathcal{N}(0, \omega^2)$$

are usually proven using regeneration arguments. By splitting the sequence of observations into consecutive blocks starting and ending at a distinguished value, the problem is reduced to an independent identically distributed central limit theorem for averages of blocks of observations. The blocks have random length but the length is finite with probability one.

There are two issues with this approach. First, it is not obvious how to extend it to the score, which is a non-additively-separable function of $Y_{1:T}$. Second, and more interestingly, it seems that the qualitative link between the stationary and the non-stationary distributions should be captured by the merging properties rather than regeneration or similar lower-level phenomena. After all, regeneration is one of several lower-level arguments that can be used to show merging in the first place.

In appendix section A12.5.2 I give a proof of the non-stationary central limit theorem for the score using merging and the bounded differences property of the score. Let σ_T^i be the i^{th} coefficient of the unscaled score Ts_T. σ_T^i has bounded differences uniformly in T and t, meaning that there is c_i such that for any $y_{1:T}$, any $\hat{y}_t$:

$$\sigma_T^i\left(y_1, \ldots, y_{t-1}, y_t, y_{t+1}, \ldots, y_T\right) - \sigma_T^i\left(y_1, \ldots, y_{t-1}, \hat{y}_t, y_{t+1}, \ldots, y_T\right) \leq c_i$$

This follows from almost-sure Cauchy bounds, as done repeatedly in appendix section A12.4.

Bounded differences conditions are typically[26] used as sufficient conditions for the concentration of non-additively-separable functions.

[26]Including in this paper, where I use the bounded differences property of the score in this concentration context in order to verify one of the assumptions for the Bernstein-von Mises theorem; see appendix section A12.7.4.

Heuristically, merging and bounded differences seem to be reasonable sufficient conditions to take a stationary limit theorem to its non-stationary version. Observations are distributed more and more according to the stationary distribution, and no single observation has an overwhelming influence on the score. Appendix section A12.5.2 makes this intuition precise.

The argument of appendix section A12.5.2 is of general interest. It can be used to show non-stationary central limit theorems for Markov chains or other merging processes such as hidden Markov models. One advantage is that it can handle non-additively-separable functions in addition to the more usual averages $\frac{1}{T}\sum_{t=1}^{T} f(Y_t)$.

I use a somewhat similar argument in appendix section A12.7.2 to show that a uniformly consistent estimator under stationarity is also uniformly consistent under non-stationarity in order to check one of the assumptions of the Bernstein-von Mises theorem, see appendix section A12.7.2.

A12.5.2 Non-stationary central limit theorem for the score

Let $\tilde{Y}_t$ be the sequence under stationarity and Y_t under any initial distribution. Write $Y = Y_{1:T}$, $\tilde{Y} = \tilde{Y}_{1:T}$, $Y_{-t} = Y_{1:t-1,t+1,T}$ and $\tilde{Y}_{-t} = \tilde{Y}_{1:t-1,t+1,T}$. $\tilde{Y}$ and Y do not have to live on the same probability space; this will be made more precise below. Let $s_T = \frac{1}{T}\sigma_T\left(Y_{1:T}\right)$ be the score for the non-stationary model and $\tilde{s}_T = \frac{1}{T}\sigma_T\left(\tilde{Y}_{1:T}\right)$ the score for the stationary model, both computed under the potentially misspecified assumption that the data are generated with some arbitrary initial stationary distribution μ, as in section A12.3. In section A12.4.2 I have shown that the slight misspecification does not matter in asymptotics and that the following central limit theorem for the score holds under stationarity:

$$\sqrt{T}\tilde{s}_T \Rightarrow \mathcal{N}(0, I)$$

Now I want to show:

$$\sqrt{T}s_T \Rightarrow \mathcal{N}(0, I)$$

Let $d_{TV}\left(Z_1, Z_2\right)$ be the notation for the total variation distance between the distributions of any two random variables Z_1 and Z_2.

Recall that Y satisfies merging, i.e., there is $\rho < 1$, $c > 0$ such that:

$$d_{TV}\left(Y_t, \tilde{Y}_1\right) < c\rho^t$$

Assume that s is scalar for simplicity.

For P_1 and P_2 probability distributions on $\mathbb{R}$, write $P_1 \otimes P_2$ for the space of measures on $\mathbb{R}^2$ whose marginals are P_1 and P_2 (not to be confused with the product measure $P_1 \times P_2$). Let W be the Wasserstein metric between probability distributions on $\mathbb{R}$:

$$W(P_1, P_2) = \inf_{P \in P_1 \otimes P_2} \left(\int (z_1 - z_2)^2 P(dz) \right)^{1/2}$$

It is well-known that W metrizes weak convergence.

For any two random variables Z_1 and Z_2, let $W(Z_1, Z_2)$ be the notation for $W(P_{Z_1}, P_{Z_2})$. Then:

$$W(Z_1, Z_2) = \inf_{P \in P_{Z_1} \otimes P_{Z_2}} \mathbb{E}_P \left[(Z_1 - Z_2)^2 \right]^{1/2}$$

In particular: for $a > 0$:

$$W(aZ_1, aZ_2) = \inf_{P \in P_{Z_1} \otimes P_{Z_2}} \mathbb{E}_P \left[(aZ_1 - aZ_2)^2 \right]^{1/2} = aW(Z_1, Z_2)$$

Consider the following inequality where on each line $Y_{1:t-1}, Y_t, \tilde{Y}_t, \tilde{Y}_{t+1:T}$ must have any joint distribution respecting the marginal distributions of Y and $\tilde{Y}$, but these joint distributions do not have to be compatible between lines:

$$\begin{aligned}
W\left(s_T, \tilde{s}_T\right) &= W\left(\frac{1}{T}\sigma_T\left(Y_1, \ldots, Y_T\right), \frac{1}{T}\sigma_T\left(\tilde{Y}_1, \ldots, \tilde{Y}_T\right)\right) \\
&= \frac{1}{T} W\left(\sigma_T\left(Y_1, \ldots, Y_T\right), \sigma_T\left(\tilde{Y}_1, \ldots, \tilde{Y}_T\right)\right) \\
&\leq \frac{1}{T}\Bigg(W\left(\sigma_T\left(Y_1, \ldots, Y_T\right), \sigma_T\left(Y_1, \ldots, Y_{T-1}, \tilde{Y}_T\right)\right) \\
&\quad + W\left(\sigma_T\left(Y_1, \ldots, Y_{T-1}, \tilde{Y}_T\right), \sigma_T\left(Y_1, \ldots, Y_{T-2}, \tilde{Y}_{T-1}, \tilde{Y}_T\right)\right) \\
&\quad + \ldots \\
&\quad + W\left(\sigma_T\left(Y_1, Y_2, \tilde{Y}_3, \ldots, \tilde{Y}_T\right), \sigma_T\left(Y_1, \tilde{Y}_2, \ldots, \tilde{Y}_T\right)\right) \\
&\quad + W\left(\sigma_T\left(Y_1, \tilde{Y}_2, \ldots, \tilde{Y}_T\right), \sigma_T\left(\tilde{Y}_1, \ldots, \tilde{Y}_T\right)\right) \Bigg)
\end{aligned}$$

Let us bound each term separately. Fix t and define:

$$h(y_t) = \sigma_T\left(Y_1, \ldots, Y_{t-1}, y_t, \tilde{Y}_{t+1}, \ldots, \tilde{Y}_T\right)$$

Fix $P \in P_Y \otimes P_{\tilde{Y}}$:

$$\begin{aligned}\mathbb{E}_P\left[\left(h\left(Y_t\right)-h\left(\tilde{Y}_t\right)\right)^2\right] &\leq \mathbb{E}_P\left[c^2 1\left[Y_t \neq \tilde{Y}_t\right]\right] \qquad \text{by bounded differences of the score} \\ &= c^2 P\left(Y_t \neq \tilde{Y}_t\right)\end{aligned}$$

And:

$$\begin{aligned}W\left(h\left(Y_t\right), h\left(\tilde{Y}_t\right)\right)^2 &= \inf_{P \in P_Y \otimes P_{\tilde{Y}}} \mathbb{E}_P\left[\left(h\left(Y_t\right)-h\left(\tilde{Y}_t\right)\right)^2\right] \\ &\leq \inf_{P \in P_Y \otimes P_{\tilde{Y}}} c^2 P\left(Y_t \neq \tilde{Y}_t\right)\end{aligned}$$

Looking only at the marginal at time-horizon t, it is well-known that there is $P_t^\star \in P_{Y_t} \otimes P_{\tilde{Y}_t}$ such that:

$$\inf_{P \in P_{Y_t} \otimes P_{\tilde{Y}_t}} P\left(Y_t \neq \tilde{Y}_t\right) = \min_{P \in P_{Y_t} \otimes P_{\tilde{Y}_t}} P\left(Y_t \neq \tilde{Y}_t\right) = P_t^\star\left(Y_t \neq \tilde{Y}_t\right) = d_{TV}\left(Y_t, \tilde{Y}_t\right)$$

We want to extend this property to the joint probabilities on $1:T$. Fix $P_t^\star$ as above. Let $P^\star \in P_Y \otimes P_{\tilde{Y}}$ such that $\left(Y_t^\star, \tilde{Y}_t^\star\right) \sim P_t^\star$, $Y_{-t}^\star$ and $\tilde{Y}_{-t}^\star$ are independent conditionally on $\left(Y_t^\star, \tilde{Y}_t^\star\right)$, $Y_{-t}^\star | Y_t^\star \sim P_{Y_{-t}|Y_t}$ and $\tilde{Y}_{-t}^\star | \tilde{Y}_t^\star \sim P_{\tilde{Y}_{-t}|\tilde{Y}_t}$. Then:

$$P^\star\left(Y_t \neq \tilde{Y}_t\right) = P_t^\star\left(Y_t \neq \tilde{Y}_t\right) = d_{TV}\left(Y_t, \tilde{Y}_t\right)$$

(Although we don't need it for bounding W, in fact $P^\star$ achieves $\inf_{P \in P_Y \otimes P_{\tilde{Y}}} P\left(Y_t \neq \tilde{Y}_t\right)$ because for any P, $d_{TV}\left(Y_t, \tilde{Y}_t\right) \leq P\left(Y_t \neq \tilde{Y}_t\right)$.)

As a consequence:

$$W\left(h\left(Y_t\right), h\left(\tilde{Y}_t\right)\right)^2 \leq c^2\ d_{TV}\left(Y_t, \tilde{Y}_t\right)$$

And thanks to merging:

$$W\left(h\left(Y_t\right), h\left(\tilde{Y}_t\right)\right) \leq c\sqrt{\rho}^T$$

Putting all the terms back together:

$$W\left(s_T, \tilde{s}_T\right) \leq \frac{1}{T} c(1+\sqrt{\rho}+\ldots+\sqrt{\rho}^T) \leq \frac{1}{T} c \frac{1}{1-\sqrt{\rho}}$$

So that finally:

$$W\left(\sqrt{T} s_T, \mathcal{N}(0, I)\right) \leq W\left(\sqrt{T} s_T, \sqrt{T} \tilde{s}_T\right) + W\left(\sqrt{T} \tilde{s}_T, \mathcal{N}(0, I)\right) \to 0$$

$$\text{ie } \sqrt{T} s_T \Rightarrow \mathcal{N}(0, I)$$

A12.6 Asymptotic distribution of the maximum likelihood estimator (proof of Theorem 4)

A12.6.1 Consistency of the maximum likelihood estimator

We show that $\theta^\star$ is the unique maximum of $L(\theta) = \mathbb{E}_{\theta^\star}[\log P_\theta(Y_1|Y_{-\infty:0})]$ in two steps.

Step 1: $P_\theta(Y_{1:T}|Y_{-\infty:m}) \xrightarrow{\theta^\star \ as} P_\theta(Y_{1:T})$ **when** $m \to -\infty$

Note that this is yet another "infinite-past" statement, although of a different kind compared to the limit theorems of appendix section A12.4. We use the merging property of $z_t = (x_t, y_t)$ directly:

$$\begin{aligned}
&|P_\theta(y_{1:T}) - P_\theta(y_{1:T}|y_{-\infty:m})| \\
&\quad \leq \left| \sum_{x_0,y_0} P_\theta(y_{1:T}|x_0,y_0) P_\theta(x_0,y_0) - \sum_{x_0,y_0} P_\theta(y_{1:T}|x_0,y_0) P_\theta(x_0,y_0|y_{-\infty:m}) \right| \\
&\quad \leq \sum_{x_0,y_0} P_\theta(y_{1:T}|x_0,y_0) \left| \sum_{x_{m+1},y_{m+1}} P_\theta(x_0,y_0|x_{m+1},y_{m+1}) (P_\theta(x_{m+1},y_{m+1}) - P_\theta(x_{m+1},y_{m+1}|y_{-\infty:m})) \right| \\
&\quad \leq \sum_{x_0,y_0} P_\theta(y_{1:T}|x_0,y_0) c\rho^m \qquad \text{by merging} \\
&\quad \leq d_x d_y c \rho^m \\
&\quad \xrightarrow[m\to-\infty]{} 0
\end{aligned}$$

Step 2: $\theta^\star$ **is the unique maximum of** $L(\theta)$

Let us show by contradiction that, for $\theta \neq \theta^\star$, $P_\theta(Y_1|Y_{-\infty:0})$ is not $\theta^\star$ almost surely equal to $P_{\theta^\star}(Y_1|Y_{-\infty:0})$. Suppose it is. Then by the law of iterated expectations and stationarity, $P_\theta(Y_{1:T}|Y_{-\infty:0}) = P_{\theta^\star}(Y_{1:T}|Y_{-\infty:0})$ ($\theta^\star$-as) for any $T \geq 1$; and by integration and stationarity: $P_\theta(Y_{1:T}|Y_{-\infty:m}) = P_{\theta^\star}(Y_{1:T}|Y_{-\infty:m})$ ($\theta^\star$-as) for any $T \geq 1 \geq m$. Then by step 1, $P_\theta(Y_{1:T}) = P_{\theta^\star}(Y_{1:T})$ for any $T \geq 1$, which contradicts the identification assumption (A3).

Then by the strict Jensen inequality:

$$\begin{aligned}\mathbb{E}_{\theta^\star}\left[\log\frac{P_\theta\left(Y_1|Y_{-\infty:0}\right)}{P_{\theta^\star}\left(Y_1|Y_{-\infty:0}\right)}\right] &< \log\mathbb{E}_{\theta^\star}\left[\frac{P_\theta\left(Y_1|Y_{-\infty:0}\right)}{P_{\theta^\star}\left(Y_1|Y_{-\infty:0}\right)}\right]\\ &= \log\mathbb{E}_{\theta^\star}\left[\mathbb{E}_{\theta^\star}\left[\frac{P_\theta\left(Y_1|Y_{-\infty:0}\right)}{P_{\theta^\star}\left(Y_1|Y_{-\infty:0}\right)}\right]\middle|Y_{-\infty:0}\right]\\ &= 0\end{aligned}$$

Thus:

$$L(\theta) < L\left(\theta^\star\right)$$

The continuity of L and the compactness of Θ imply that $\theta^\star$ is a well-separated maximum and the uniform law of large numbers for L_T implies consistency.

A12.6.2 Asymptotic normality of the maximum likelihood estimator

I is invertible by identification (section A12.6.1). Asymptotic normality of the maximum likelihood estimator is a standard consequence of the uniform local asymptotic normality property (Theorem 3) and consistency.

A12.7 Asymptotic distribution of the Bayesian posterior: Bernstein–von Mises theorem (proof of Theorem 5)

I apply the weakly dependent Bernstein–von Mises of Connault (2014).

In a hidden Rust model, the domination assumption (A1) is verified. Local asymptotic normality (A4) is of course Theorem 3. Assume that a prior that verifies the support assumption (A2) is used. Assumptions (A3) (uniformly consistent tests), (A5) (a local linear lower bound for the score), (A6) (a large deviation inequality for the score) and (A7) (a large deviation inequality for blocks of data) remain to be checked .

A12.7.1 Uniformly consistent estimators

Uniformly consistent estimators can be used to build uniformly consistent tests (thus checking assumption (A3)). By uniformly consistent estimators, I mean estimators $\hat{\theta}_T$ such that, for some distance d:

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

d can be any distance as long as it is locally stronger than the reference Euclidean distance around $\theta^\star$, i.e.:

$$\forall \epsilon > 0, \exists \eta > 0 : \qquad \|\theta - \theta^\star\| > \epsilon \Longrightarrow d\left(\theta, \theta^\star\right) > \eta$$

See section A12.7.2 for why we need d to be locally stronger than the Euclidean distance.

$\hat{\theta}_T$ is not an estimator that is meant to be used in practice. It is used only as a technical device in the proof of the Bernstein–von Mises theorem. It does not matter if $\hat{\theta}_T$ has terrible short-sample properties or is not statistically efficient, as long as it is uniformly consistent. An idea is to construct $\hat{\theta}_T$ using the dynamic structure of the model regardless of the economic structure. For instance, in the independent identically distributed case, one could use non-parametric estimators of the marginal distribution, or non-constructively prove that there exist uniformly consistent estimators of the marginal distribution as is done in van der Vaart (1998).

For hidden Rust models, the comments of the previous paragraph are valid but the situation is complicated by the weakly dependent dynamics. I rely on Theorem 2 of section 3 about identification. By assumption (A3), the model is identified under stationarity. By merging, this implies it is also identified under a different initial distribution (any marginal can be arbitrarily well approximated by waiting long enough). By Theorem 2, let T_0 be a time horizon such that the marginal stationary distribution of T_0 consecutive y's identify θ. Let π be the corresponding marginal, i.e., the distribution of $y_{1:T_0}$ under stationarity. Let $(\hat{x}_s, \hat{y}_s)$ be non-overlapping blocks of T_0 consecutive (x_t, y_t)'s and $S = \lfloor T/T_0 \rfloor$. Let $\hat{\theta}_T = \hat{\pi}_T$ be the empirical distribution estimator of π defined by:

$$\hat{\pi}_T\left(Y_{1:T}; \hat{y} = y_{1:T_0}\right) = \frac{1}{S}\sum_{s=1}^{S} 1\left[\hat{Y}_s = \hat{y}\right]$$

Finally let d_{TV} be the total variation distance. Because the model is identified, $d_{TV}\left(\hat{\pi}_T, \pi\right)$ is a particular choice of distance $d\left(\hat{\theta}_T, \theta\right)$. I show that $\hat{\pi}_T$ is a uniformly consistent estimator.

$$\forall \epsilon, \qquad \sup_\theta P_\theta\left(d_{TV}\left(\hat{\pi}_T, \pi\right) \geq \epsilon\right) \to 0 \tag{24}$$

The heuristic reason why we expect (24) to hold is as follows. (24) looks exactly like a uniform Glivenko-Cantelli theorem. Indeed if $\mathcal{A}$ is the underlying σ-algebra, d_{TV} can be expressed as $d_{TV}(p_1, p_2) = \sup_{A \in \mathcal{A}} |p_1(A) - p_2(A)|$. Here y_t is discrete and $\mathcal{A}$ is the set of all subsets of $\{1, \ldots, d_y\}^{T_0}$. In particular $\mathcal{A}$ has finite Vapnik-Chervonenkis dimension. There are two well-known Glivenko-Cantelli results under finite Vapnik-Chervonenkis dimension.

First, a uniform Glivenko-Cantelli theorem $\sup_P P\left(d_{TV}\left(\hat{\pi}_T, \pi\right) \geq \epsilon\right) \to 0$ holds under independent identically distributed assumption for P. Second, a universal Glivenko-Cantelli theorem $\forall P, P\left(d_{TV}\left(\hat{\pi}_T, \pi\right) \geq \epsilon\right) \xrightarrow{P} 0$ holds under suitable mixing assumptions for P. Since hidden Rust models are "uniformly mixing," we can expect the uniform statement (24) to hold.

To prove (24) we show a Dvoretzky-Kiefer-Wolfowitz type inequality, that is a quantitative bound going to zero with the time-series length: there is a sequence $\alpha(T) \xrightarrow[T\to\infty]{} 0$ such that:

$$\forall \theta, \qquad P_\theta\left(d_{TV}\left(\hat{\pi}_T, \pi\right) \geq \epsilon\right) \leq \alpha(T) \tag{25}$$

(25) follows in turn from a concentration bound for $d_{TV}\left(\hat{\pi}_T, \pi\right)$ around its expectation, and a separate bound for its expectation.

Step 1: concentration bound for $d_{TV}\left(\hat{\pi}_T, \pi\right)$.

We want to apply the concentration inequality (5). Let us check that $d_{TV}\left(\hat{\pi}_T, \pi\right)$ verifies a suitable bounded differences condition. By definition:

$$d_{TV}\left(\hat{\pi}_T, \pi\right) = \frac{1}{2} \sum_{\hat{y}} \left|\hat{\pi}_T\left(\hat{y}\right) - \pi\left(\hat{y}\right)\right|$$

For any sequence of observations $y_{1:T}$ and $\tilde{y}_{1:T}$:

$$\begin{aligned}
&d_{TV}\left(\hat{\pi}_T\left(y_{1:T}\right), \pi\right) - d_{TV}\left(\hat{\pi}_T\left(\tilde{y}_{1:T}\right), \pi\right) \\
&\quad \leq \frac{1}{2} \sum_{\hat{y}} \left|\hat{\pi}_T\left(y_{1:T}; \hat{y}\right) - \hat{\pi}_T\left(\tilde{y}_{1:T}; \hat{y}\right)\right| \qquad \text{by triangle inequality} \\
&\quad \leq \frac{1}{2} \frac{1}{S} \sum_{s=1}^{S} 1\left[\hat{y}_s \neq \hat{\tilde{y}}_s\right] \\
&\quad \leq \frac{1}{2} \frac{1}{S} \sum_{t=1}^{T_0} 1\left[y_t \neq \tilde{y}_t\right] \\
&\quad \leq \frac{1}{2} \frac{1}{T - T_0} \sum_{t=1}^{T} 1\left[y_t \neq \tilde{y}_t\right]
\end{aligned}$$

Thus $d_{TV}\left(\hat{\pi}_T, \pi\right)$ verifies a bounded differences condition with $c_t = \frac{1}{2} \frac{1}{T - T_0}$ and we can apply

(5) as announced:

$$P_{\theta,\mu}\left(d_{TV}\left(\hat{\pi}_T,\pi\right)>\mathbb{E}_{\theta,\mu}\left[d_{TV}\left(\hat{\pi}_T,\pi\right)\right]+u\right)\leq\exp\left(-\frac{1}{2}\frac{u^2}{\tau_x(\theta)\sum_{t=1}^{T}\left(\frac{1}{2}\frac{1}{T-T_0}\right)^2}\right)$$
$$\leq\exp\left(-\frac{1}{2}\frac{u^2}{\tau_x\frac{T}{4(T-T_0)^2}}\right)$$

Step 2: bounding $\mathbb{E}_{\theta,\mu}\left[d_{TV}\left(\hat{\pi}_T,\pi\right)\right]$.

To bound the expectation under an arbitrary non-stationary initial distribution I show a bound under the stationary distribution and I take it to the non-stationary case using merging properties. This is similar to the way I show a non-stationary central limit theorem for the score from a stationary central limit theorem using merging, in the local asymptotic normality section (section A12.5).

Step 2.1: bounding $\mathbb{E}_{\theta}\left[d_{TV}\left(\hat{\pi}_T,\pi\right)\right]:=\mathbb{E}_{\theta,\mu^{\diamond}(\theta)}\left[d_{TV}\left(\hat{\pi}_T,\pi\right)\right]$.

Paulin (2014) gives a bound under stationarity for the empirical distribution estimator of the one-dimensional marginal of a Markov chain (bound (3.31) p.21). We can apply this bound to the block Markov chain $(\hat{x},\hat{y})_s$. Write λ for the joint distribution of $(\hat{x},\hat{y})_1=(x,y)_{1:T_0}$ under stationarity and $\hat{\lambda}_T$ for the corresponding empirical distribution estimator. A consequence of (3.31) from Paulin (2014) is that for T big enough:

$$\mathbb{E}_{\theta}\left[d_{TV}\left(\hat{\lambda}_T,\lambda\right)\right]\leq\sqrt{\frac{1}{S\gamma_{(\hat{x},\hat{y})}(\theta)}}\sum_{\hat{y}}\sqrt{\lambda\left(\hat{y}\right)}$$

Let $\hat{d}=d_{\hat{y}}=d_y^{T_0}$. Remember the general inequality between ℓ^p norms:

$$\left\|\lambda\right\|_{1/2}\leq\hat{d}^{\frac{1}{1/2}-\frac{1}{1}}\left\|\lambda\right\|_1=\hat{d}$$

So that:

$$\mathbb{E}_{\theta}\left[d_{TV}\left(\hat{\lambda}_T,\lambda\right)\right]\leq\sqrt{\frac{\hat{d}}{(T-T_0)\gamma_{(\hat{x},\hat{y})}}}$$

Now consider the projection function $h_y\left(\hat{x},\hat{y}\right)=\hat{y}$, which takes the joint Markov chain $(\hat{x},\hat{y})_s$ to its observable component $\hat{y}_s$. Then $\hat{\pi}_T$ and π are the distributions of $h_y\left(\hat{x},\hat{y}\right)$ under $\hat{\lambda}_T$ and λ, respectively (i.e., $\hat{\pi}_T=\hat{\lambda}_T\circ h_y^{-1}$ and $\hat{\pi}_T=\hat{\lambda}_T\circ h_y^{-1}$). Now if $d_{TV}(Z_1,Z_2)$ means the total variation distance between the distributions of two arbitrary random variables Z_1 and

Z_2 and h is any function, d_{TV} satisfies $d_{TV}(h(Z_1), h(Z_2)) \leq d_{TV}(Z_1, Z_2)$. In the case at hand:

$$\mathbb{E}_\theta \left[d_{TV} \left(\hat{\pi}_T, \pi \right) \right] \leq \mathbb{E}_\theta \left[d_{TV} \left(\hat{\lambda}_T, \lambda \right) \right] \leq \sqrt{\frac{\hat{d}}{(T - T_0)\gamma_{(\hat{x},\hat{y})}}}$$

Step 2.2: bounding $\mathbb{E}_{\theta,\mu} \left[d_{TV} \left(\hat{\pi}_T, \pi \right) \right]$.

Remember that Y_t satisfies merging uniformly in θ and μ by assumption: there is $\rho < 1$, $c > 0$ such that:

$$d_{TV}(Y_t, \mu^\diamond) \leq c\rho^t$$

Somewhat similar to what I do in section A12.5 to show a nonstationary central limit theorem for the score, I show that $\mathbb{E}_{\theta,\mu} \left[d_{TV} \left(\hat{\pi}_T, \pi \right) \right] - \mathbb{E}_\theta \left[d_{TV} \left(\hat{\pi}_T, \pi \right) \right]$ goes to zero using the bounded differences property of $d_{TV} \left(\hat{\pi}_T, \pi \right)$ together with merging. Let $(Y_t)_t$ be distributed according to θ and μ (nonstationary) and $\left(\tilde{Y}_t \right)_t$ be distributed according to θ and stationary. Y and $\tilde{Y}$ do not have to live on the same probability space. Consider the following inequality where on each line $Y_{1:t-1}, Y_t, \tilde{Y}_t, \tilde{Y}_{t+1:T}$ must have any joint distribution respecting the marginal distributions of Y and $\tilde{Y}$, but these joint distributions do not have to be compatible between lines:

$$\begin{aligned}
&\left| \mathbb{E}_{\theta,\mu} \left[d_{TV} \left(\hat{\pi}_T, \pi \right) \right] - \mathbb{E}_\theta \left[d_{TV} \left(\hat{\pi}_T, \pi \right) \right] \right| \\
&\quad = \left| \mathbb{E} \left[d_{TV} \left(\hat{\pi}_T \left(Y_1, \ldots, Y_T \right), \pi \right) \right] - \mathbb{E} \left[d_{TV} \left(\hat{\pi}_T \left(\tilde{Y}_1, \ldots, \tilde{Y}_T \right), \pi \right) \right] \right| \\
&\quad \leq \left| \mathbb{E} \left[d_{TV} \left(\hat{\pi}_T \left(Y_1, \ldots, Y_T \right), \pi \right) \right] - \mathbb{E} \left[d_{TV} \left(\hat{\pi}_T \left(Y_1, \ldots, Y_{T-1}, \tilde{Y}_T \right), \pi \right) \right] \right| \\
&\qquad + \left| \mathbb{E} \left[d_{TV} \left(\hat{\pi}_T \left(Y_1, \ldots, Y_{T-1}, \tilde{Y}_T \right), \pi \right) \right] - \mathbb{E} \left[d_{TV} \left(\hat{\pi}_T \left(Y_1, \ldots, Y_{T-2}, \tilde{Y}_{T-1}, \tilde{Y}_T \right), \pi \right) \right] \right| \\
&\qquad + \ldots \\
&\qquad + \left| \mathbb{E} \left[d_{TV} \left(\hat{\pi}_T \left(Y_1, Y_2, \tilde{Y}_3, \ldots, \tilde{Y}_T \right), \pi \right) \right] - \mathbb{E} \left[d_{TV} \left(\hat{\pi}_T \left(Y_1, \tilde{Y}_2, \ldots, \tilde{Y}_T \right), \pi \right) \right] \right| \\
&\qquad + \left| \mathbb{E} \left[d_{TV} \left(\hat{\pi}_T \left(Y_1, \tilde{Y}_2, \ldots, \tilde{Y}_T \right), \pi \right) \right] - \mathbb{E} \left[d_{TV} \left(\hat{\pi}_T \left(\tilde{Y}_1, \ldots, \tilde{Y}_T \right), \pi \right) \right] \right|
\end{aligned}$$

Let us bound each term separately. Fix t, $1 \leq t \leq T$.

$$\begin{aligned}
&\left| \mathbb{E} \left[d_{TV} \left(\hat{\pi}_T \left(Y_{1:t-1}, Y_t, \tilde{Y}_{t+1:T} \right), \pi \right) \right] - \mathbb{E} \left[d_{TV} \left(\hat{\pi}_T \left(Y_{1:t-1}, \tilde{Y}_t, \tilde{Y}_{t+1:T} \right), \pi \right) \right] \right| \\
&\quad \leq \mathbb{E} \left[d_{TV} \left(\hat{\pi}_T \left(Y_{1:t-1}, Y_t, \tilde{Y}_{t+1:T} \right), \hat{\pi}_T \left(Y_{1:t-1}, \tilde{Y}_t, \tilde{Y}_{t+1:T} \right) \right) \right] \qquad \text{by triangle inequality} \\
&\quad = \mathbb{E} \left[\frac{1}{S} 21 \left[Y_t \neq \tilde{Y}_t \right] \right]
\end{aligned}$$

Exactly as in section A12.5, we can build $P^\star$ such that $\left(Y, \tilde{Y} \right) \sim P^\star$ and $P^\star \left(Y_t \neq \tilde{Y}_t \right) =$

$d_{TV}\left(Y_t, \tilde{Y}_t\right)$. Hence the bound:

$$\begin{aligned}
&\left|\mathbb{E}\left[d_{TV}\left(\hat{\pi}_T\left(Y_{1:t-1}, Y_t, \tilde{Y}_{t+1:T}\right), \pi\right)\right] - \mathbb{E}\left[d_{TV}\left(\hat{\pi}_T\left(Y_{1:t-1}, \tilde{Y}_t, \tilde{Y}_{t+1:T}\right), \pi\right)\right]\right| \\
&\quad \leq \frac{2}{T - T_0} d_{TV}\left(Y_t, \tilde{Y}_t\right) \\
&\quad \leq \frac{2c}{T - T_0} \rho^t
\end{aligned}$$

Putting back all the terms together, we get the bound:

$$\left|\mathbb{E}_{\theta,\mu}\left[d_{TV}\left(\hat{\pi}_T, \pi\right)\right] - \mathbb{E}_{\theta}\left[d_{TV}\left(\hat{\pi}_T, \pi\right)\right]\right| \leq \frac{2c}{T - T_0}(1 + \rho + \ldots + \rho^T) \leq \frac{2c}{(T - T_0)(1 - \rho)}$$

And with step 2.1:

$$\mathbb{E}_{\theta,\mu}\left[d_{TV}\left(\hat{\pi}_T, \pi\right)\right] \leq \frac{2c}{(T - T_0)(1 - \rho)} \sqrt{\frac{\hat{d}}{(T - T_0)\gamma_{(\hat{x},\hat{y})}}}$$

Finally, putting steps 1 and 2 together:

$$\begin{aligned}
P_{\theta,\mu}\left(d_{TV}\left(\hat{\pi}_T, \pi\right) \geq \epsilon\right) &= P_{\theta,\mu}\left(d_{TV}\left(\hat{\pi}_T, \pi\right) \geq \mathbb{E}\left[d_{TV}\left(\hat{\pi}_T, \pi\right)\right] + \left(\epsilon - \mathbb{E}_{\theta,\mu}\left[d_{TV}\left(\hat{\pi}_T, \pi\right)\right]\right)\right) \\
&\leq \exp\left(-\frac{1}{2} \frac{\left(\epsilon - \mathbb{E}_{\theta,\mu}\left[d_{TV}\left(\hat{\pi}_T, \pi\right)\right]\right)^2}{\tau_x \frac{T}{4(T - T_0)^2}}\right)
\end{aligned}$$

This is of the form (25) (for a fixed ϵ and T big enough) and shows that $\hat{\theta}_T = \hat{\pi}_T$ is a uniformly consistent estimator as in (24).

A12.7.2 Uniformly consistent tests (checking assumption (A3))

Let $\hat{\theta}_T$ be any uniformly consistent estimator in the sense of section A12.7.1 ($\hat{\pi}_T$ is such an estimator by section A12.7.1). Let $\epsilon > 0$. Recall that by definition of a uniformly consistent estimator (section A12.7.1), there is $\eta > 0$ such that:

$$\|\theta - \theta^\star\| > \epsilon \Longrightarrow d\left(\theta, \theta^\star\right) > \eta$$

Let us show that $\phi_T = 1\left[d\left(\hat{\theta}_T, \theta^\star\right) \geq \eta/2\right]$ is a uniformly consistent test for ϵ.

First,

$$\mathbb{E}_{\theta^\star}[\phi_T] = P_{\theta^\star}\left(d\left(\hat{\theta}_T, \theta^\star\right) \geq \eta/2\right) \to 0 \qquad \text{by consistency}$$

Second,

$$\begin{aligned}
\mathbb{E}_\theta[1-\phi_T] &= P_\theta(\phi_T = 0) \\
&= P_\theta\left(d\left(\hat{\theta}_T, \theta^\star\right) < \eta/2\right) \\
&\leq P_\theta\left(d\left(\theta, \theta^\star\right) - d\left(\hat{\theta}_T, \theta\right) < \eta/2\right) && \text{by triangle inequality} \\
&= P_\theta\left(d\left(\hat{\theta}_T, \theta\right) > d\left(\theta, \theta^\star\right) - \eta/2\right)
\end{aligned}$$

So that:

$$\begin{aligned}
\sup_{\|\theta-\theta^\star\|>\epsilon} \mathbb{E}_\theta[1-\phi_T] &\leq \sup_{\|\theta-\theta^\star\|>\epsilon} P_\theta\left(d\left(\hat{\theta}_T, \theta\right) > \eta/2\right) \\
&\to 0 && \text{by uniform consistency}
\end{aligned}$$

A12.7.3 Local linear lower bound for the score (checking assumption (A5))

We can rely on the smoothness of $\theta \to \mathbb{E}_\theta\left[s_T\right]$ on Θ. Note that as usual:

$$\begin{aligned}
\nabla_{\theta^\star}\mathbb{E}_\theta\left[s_T\right] &= \sum_{\tilde{y}_{1:T}} s_T \nabla_{\theta^\star} P_\theta(\tilde{y}_{2:T}|\tilde{y}_1) P(\tilde{y}_1) \\
&= \sum_{\tilde{y}_{1:T}} s_T \frac{\nabla_{\theta^\star} P_\theta(\tilde{y}_{2:T}|\tilde{y}_1)}{P_{\theta^\star}(\tilde{y}_{2:T}|\tilde{y}_1)} P_{\theta^\star}(\tilde{y}_{2:T}|\tilde{y}_1) P(\tilde{y}_1) \\
&= T\mathbb{E}_{\theta^\star}[s_T s_T'] \\
&= \mathbb{E}_{\theta^\star}[i_T(\theta^\star)] \qquad \text{(see the end of section A12.4.3)}
\end{aligned}$$

Write $h(\theta) = \nabla^2_\theta \mathbb{E}_\theta\left[s_T\right]$. Consider a second-order Taylor expansion around $\theta^\star$ with Lagrange remainder: there is $\bar{\theta}$, $\theta^\star_i \leq \bar{\theta}_i \leq \theta_i$, such that:

$$\mathbb{E}_\theta\left[s_T\right] = \mathbb{E}_{\theta^\star}\left[s_T\right] + \mathbb{E}_{\theta^\star}[i_T(\theta^\star)](\theta-\theta^\star) + (\theta-\theta^\star)' h\left(\bar{\theta}\right)(\theta-\theta^\star)$$

Since $\mathbb{E}_{\theta^\star}[i_T(\theta^\star)] \xrightarrow[T\to\infty]{} I$ (see section A12.4), I is invertible by assumption and h is bounded over Θ by smoothness and compactness, there is T_0, $\delta < 1$ and c such that for any $\|\theta-\theta^\star\| \leq \delta$, $T > T_0$:

$$\left\|\mathbb{E}_\theta\left[s_T\right] - \mathbb{E}_{\theta^\star}\left[s_T\right]\right\| \geq c\left\|\theta-\theta^\star\right\|$$

A12.7.4 Large deviation inequality for the score (checking assumption (A6))

Let σ_T^i be the i^{th} coefficient of the unscaled score Ts_T. The score has bounded differences; see section A12.5. We can apply the concentration inequality (6): for any θ:

$$P_\theta\left(\left|\sigma_T^i/T - \mathbb{E}_\theta\left[\sigma_T^i/T\right]\right| > u\right) \leq 2\exp\left(-\frac{1}{2}\frac{u^2}{\tau_x(\theta)\sum_{t=1}^T (c_i/T)^2}\right)$$
$$\leq 2\exp\left(-\frac{1}{2}\frac{u^2}{\tau_x c_i/T}\right)$$

To conclude, note that for a general random vector X:

$$P(\|X\|_2 > u) < d_X \max_i P\left(|X_i| > \frac{u}{\sqrt{d_X}}\right)$$

So that if $\bar{c} = \max_{1\leq i\leq d_\theta} c_i$:

$$P_\theta\left(\|s_T - \mathbb{E}_\theta\left[s_T\right]\| > u\right) \leq 2\exp\left(-\frac{1}{2}\frac{u^2}{\tau_x \bar{c}/T}\right)$$

Thus assumption (A6) holds with $c = \tau_x \bar{c}$

A12.7.5 Large deviation inequality for blocks (checking assumption (A7))

Let $R \in \mathbb{N}$ and define blocks $(\hat{x}_s, \hat{y}_s)$ to be non-overlapping blocks of R consecutive (x_t, y_t)'s. $(\hat{x}_s, \hat{y}_s)$ itself is a hidden Markov model and satisfies the concentration inequalities of Paulin (2014). In particular, for any g, $0 \leq g \leq 1$, applying the one-sided inequality (4) to $f\left(\hat{y}_1, \ldots, \hat{y}_S\right) = \frac{1}{S}\sum_{s=1}^S g_s$ where $g_s = g\left(\hat{y}_s\right)$ and we have:

$$P_\theta\left(\frac{1}{S}\sum_{s=1}^S g_s < \mathbb{E}\left[\frac{1}{S}\sum_{s=1}^S g_s\right] - u\right) \leq \exp\left(-\frac{1}{2}\frac{u^2}{\tau_{\hat{x}}(\theta)/S}\right)$$
$$\leq \exp\left(-\frac{1}{2}\frac{u^2}{\tau_{\hat{x}}/S}\right)$$

Thus, assumption (A7) holds with $c_R = \tau_{\hat{x}}$.

A13 Appendix for section 5: Estimation

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. Probabilities getting small is a classical numerical issue usually resolved by taking logarithms. Here logarithms

are not directly compatible with the linear recursive formula $\pi_{t+1} = \pi_t H_{t+1}$. The algorithm needs to be augmented with the log of a normalization factor for π_t, say ρ_t .

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

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

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

A14 Appendix for section 6: A structural model of dynamic financial incentives

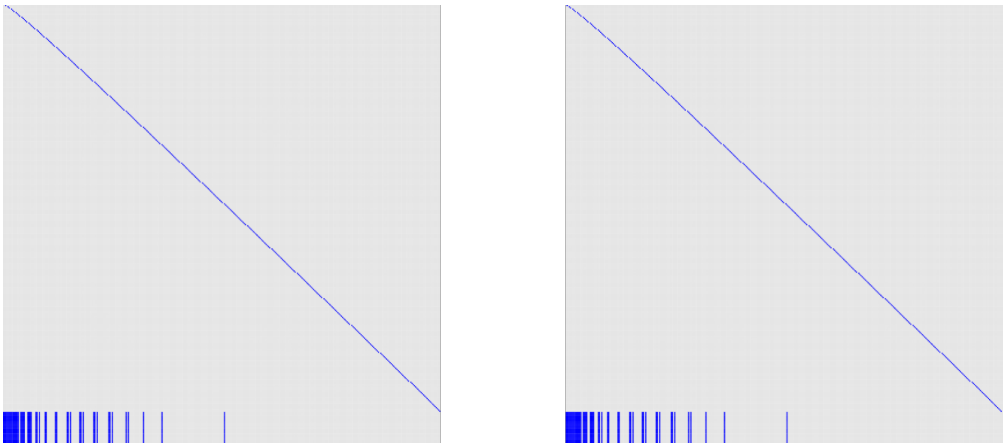


Figure 9: Conditional state transition matrices. Blue: non-zero coefficients. Light gray: zero coefficients. Sparsity: 98%.

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 6 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	$\begin{pmatrix} 98.7\% & 0.2\% & 0.8\% & 0.1\% & 0.1\% & 0.1\% & 0.0\% \\ 0.0\% & 0.0\% & 99.7\% & 0.1\% & 0.1\% & 0.1\% & 0.0\% \\ 0.0\% & 68.6\% & 1.1\% & 29.8\% & 0.3\% & 0.2\% & 0.0\% \\ 0.8\% & 0.3\% & 8.6\% & 88.4\% & 0.6\% & 1.0\% & 0.3\% \\ 0.0\% & 0.0\% & 0.0\% & 0.0\% & 86.4\% & 12.9\% & 0.7\% \\ 0.0\% & 0.0\% & 0.0\% & 2.4\% & 60.4\% & 36.9\% & 0.3\% \\ 0.0\% & 0.0\% & 0.0\% & 0.0\% & 0.0\% & 11.9\% & 88.1\% \end{pmatrix}$

A maximum likelihood path for the unobserved state variables can easily be computed at the maximum likelihood value of the structural parameters. As an illustration, I estimated a hidden Rust model with seven unobserved state variables on a larger sample of teachers that includes six additional teachers who never or almost never work, and I computed the most likely paths for the unobserved state variables. Figure 10 represents the corresponding proportion of periods spent in each of the seven unobserved states. The teachers are ranked according to their attendace rate in sample. Unobserved state specific leisure utilities, measured in rupees, are given on the right column. The likelihood spends one unobserved state (state 1, with very high leisure utility) to account for the outlier teachers.

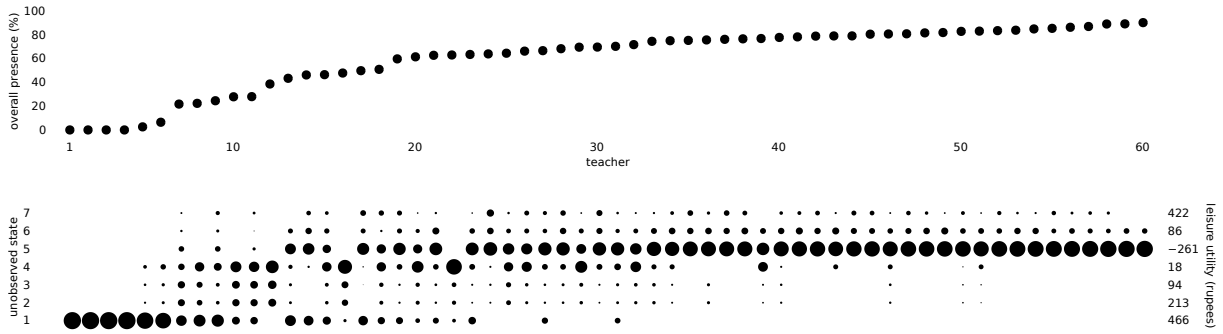


Figure 10: Most likely unobserved path.

Top: attendance rate. Bottom: periods 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.

References

Abbring, J., J. Campbell, J. Tilly, and N. Yang (2013): "Very Simple Markov-Perfect Industry Dynamics," *Working Paper*, https://www.chicagofed.org/digital_assets/publications/working_papers/2013/wp2013_20.pdf.

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.

An, Y., Y. Hu, and M. Shum (2014): "Identifiability and Inference of Hidden Markov Models," *Working Paper*.

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.

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

Connault, B. (2014): "A Weakly Dependent Bernstein–von Mises Theorem," *Working Paper*, http://www.princeton.edu/~connault/.

Cox, D., J. Little, and D. O'shea (2005): *Using Algebraic Geometry*, Springer.

Decker, W., G.-M. Greuel, G. Pfister, and H. Schönemann (2014): "Singular 4-0-1 — A computer algebra system for polynomial computations," http://www.singular.uni-kl.de.

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.

Fill, J. A. (1991): "Eigenvalue Bounds on Convergence to Stationarity for Nonreversible Markov Chains, with an Application to the Exclusion Process," *The Annals of Applied Probability*, 1, 62–87.

Gilbert, E. (1959): "On the Identifiability Problem for Functions of Finite Markov Chains," *The Annals of Mathematical Statistics*, 30, 688–697.

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.

Jordan, M. and Y. Weiss (2002): "Graphical models: Probabilistic Inference," *The Handbook of Brain Theory and Neural Networks*, 490–496.

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.

Krengel, U. (1985): *Ergodic Theorems*, Cambridge University Press.

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

Levin, D., Y. Peres, and E. L. Wilmer (2009): *Markov Chains and Mixing Times*, American Mathematical Society.

Louis, T. (1982): "Finding the Observed Information Matrix when Using the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, 44, 226–233.

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

Montes, A. and M. Wibmer (2010): "Gröbner Bases for Polynomial Systems with Parameters," *Journal of Symbolic Computation*, 45, 1391–1425.

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.

Paulin, D. (2014): "Concentration Inequalities for Markov Chains by Marton Couplings and Spectral Methods," *Working Paper*, http://arxiv.org/abs/1212.2015v3.

Petrie, T. (1969): "Probabilistic Functions of Finite State Markov Chains," *The Annals of Mathematical Statistics*, 40, 97–115.

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.

Seneta, E. (2006): *Non-Negative Matrices and Markov Chains*, Springer.

Sommese, A. and C. Wampler (2005): *The Numerical Solution of Systems of Polynomials Arising in Engineering and Science*, World Scientific.

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

van der Vaart, A. (1998): *Asymptotic Statistics*, Cambridge University Press.

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.