Maximum Likelihood Estimation in Markov Regime-Switching Models with Covariate-Dependent Transition Probabilities*

Demian Pouzo,[†] Zacharias Psaradakis[‡] and Martin Sola[§]

February 18, 2022

Abstract

This paper considers maximum likelihood (ML) estimation in a large class of models with hidden Markov regimes. We investigate consistency of the ML estimator and local asymptotic normality for the models under general conditions which allow for autoregressive dynamics in the observable process, Markov regime sequences with covariate-dependent transition matrices, and possible model misspecification. A Monte Carlo study examines the finite-sample properties of the ML estimator in correctly specified and misspecified models. An empirical application is also discussed.

Key words and phrases: Autoregressive model; consistency; covariate-dependent transition probabilities; covariance matrix estimation; hidden Markov model; Markov-switching model; maximum likelihood; local asymptotic normality; misspecified models.

*We would like to thank Noureddine El Karoui, Adityanand Guntuboyina, Michael Jansson, a co-editor, and three anonymous referees for helpful comments and suggestions. All errors are our own.

†e-mail: dpouzo@berkeley.edu (corresponding author). Address: UC Berkeley, Dept. of Economics; 530 Evans Hall, Berkeley CA 94619, USA.

‡e-mail: z.psaradakis@bbk.ac.uk. Address: Birkbeck, University of London, Dept. of Economics, Mathematics & Statistics; Malet Street, London WC1E 7HX, UK

§e-mail: msola@utdt.edu. Address: Universidad Torcuato Di Tella, Dept. of Economics; Figueroa Alcorta 7350, C1428 Buenos Aires, Argentina.

1 Introduction

Stochastic models with parameters that are subject to changes driven by an unobservable Markov chain (the regime or state sequence) have attracted considerable attention in many different areas, the influential work by Hamilton [1989] being a prominent example from the econometrics literature. An important subclass of such models, the so-called hidden Markov models, in which observations are conditionally independent given the regime sequence, are also widely used in a variety of disciplines. A common assumption in these models is that the unobservable Markov chain is temporally homogeneous.

In this paper, we focus on a larger class of models in which, conditionally on the hidden regime process $(S_t)_t$, the observation process $(X_t)_t$ is a temporally inhomogenous Markov chain such that the conditional distribution of X_t, given $(S_t, X_{t-1}, S_{t-1}, X_{t-2}, S_{t-2}, \ldots)$, depends on S_t and X_{t-1}. In addition, the regime process $(S_t)_t$ is also a temporally inhomogeneous Markov chain so that the conditional distribution of S_t, given $(X_{t-1}, S_{t-1}, X_{t-2}, S_{t-2}, \ldots)$, depends not only on S_{t-1} but also on X_{t-1}. This is a useful generalization of models in which the Markovian transition function of the regime process does not depend on observable covariates and is time-invariant, and is a generalization that has found numerous applications, especially in economics and finance.[1] Statistical inference in this class of models is predominantly likelihood based, even though very little is known about the asymptotic properties of the relevant inferential procedures. In a typical application, inference is conducted on the implicit assumption that the maximum likelihood (ML) estimator of unknown parameters has its familiar properties of consistency, asymptotic normality, and asymptotic efficiency, and associated confidence sets and hypotheses tests are constructed in the usual manner. It hardly needs noting that, unless these properties of likelihood-based inferential procedures known from regular parametric estimation problems hold for Markov regime-switching models with covariate-dependent transition probabilities, inferences drawn from them cannot be justified in any meaningful way and should be interpreted very cautiously. Arguing, for instance, as is common in applied work, that an economic variable is a useful leading indicator for business-cycle phases because it appears to have a 'statistically significant' coefficient in the transition functions of a Markov-switching model for output growth is problematic when little is known about

[1] Examples include, among many others, applications to the analysis of business-cycle fluctuations (e.g., Filardo [1994], Filardo and Gordon [1998], Ravn and Sola [1999], Simpson et al. [2001], Gadea Rivas and Perez-Quiros [2015]), interest rates and yields (e.g., Gray [1996], Bekaert et al. [2001], Ang and Bekaert [2002a], Ang and Bekaert [2002b], Psaradakis and Sola [2021]), consumption growth (Whitelaw [2000]), currency crises (e.g., Martinez Peria [2002], Mouratidis [2008]), and cryptocurrency returns (Tan et al. [2021]).

the properties of the relevant estimators and related statistical tests.

The main contribution of this paper is to provide consistency and asymptotic normality results for a large class of models that are relevant in applications. Our approach allows for autoregressive dynamics in the observable process $(X_t)_t$, covariate-dependence in the transition functions of the hidden regime process $(S_t)_t$, and potential model misspecification (i.e., the model may not contain the data-generating process). To the best of our knowledge, the only asymptotic results available on ML estimation in Markov regime-switching models with covariate-dependent transition probabilities are those of Ailliot and Pène [2015], who investigate consistency of the ML estimator in a correctly specified model. Our results include both consistency of the ML estimator and local asymptotic normality (LAN) for the model, from which asymptotic normality of the ML estimator can be inferred. Unlike Ailliot and Pène [2015], who allow for a general hidden state space, we require the latter to be finite, but do not restrict the model to be correctly specified. In doing so, we also extend some results of White [1982] for independent, identically distributed (i.i.d.) data to the case of dependent observations and for classes of parametric distributions associated with dynamic models with hidden Markov regimes. As such stochastic specifications are typically highly parametric, it is important to understand the properties of likelihood-based inferential procedures in situations where the true probability structure of the data does not necessarily lie within the parametric family of distributions specified by the model. We show that the ML estimator in our setting converges to the true parameter value if the model is correctly specified and to a pseudo-true parameter set if the model is misspecified.[2] We also show that the sample log-likelihood satisfies the LAN property, establish an asymptotic linear representation for the ML estimator, obtain the asymptotic distribution of the estimator, and present results relating to consistent estimation of its asymptotic covariance matrix. These are the most general results available for Markov regime-switching models with autoregressive dynamics and covariate-dependent transition probabilities.

In related earlier work, Mevel and Finesso [2004] examine consistency and asymptotic normality of the ML estimator in misspecified hidden Markov models with a finite state space, while Douc and Moulines [2012] consider consistency under general state spaces. Bickel and Ritov [1996], Bickel et al. [1998], Jensen and Petersen [1999], Douc and Matias [2001], and Douc et al. [2011] investigate consistency and/or asymptotic normality in correctly specified hidden Markov models with regime sequences defined on either a finite or a general state space. Francq and Roussignol [1998] and Krishnamurthy and Rydén [1998] consider

[2]We refer to estimators for the various models discussed throughout as ML estimators even though they may be obtained from a pseudo-likelihood based on a misspecified model.

consistency in correctly specified autoregressive models with Markov regimes defined on a finite state space. Douc et al. [2004b] and Kasahara and Shimotsu [2019] investigate consistency and asymptotic normality in a similar autoregressive setup, but allow the regime sequence to take values in a space that is not necessarily countable. In all of these papers, the hidden regime sequence is assumed to be a temporally homogeneous Markov chain.

In the sequel, we follow Bickel et al. [1998] and Douc et al. [2004b] fairly closely in terms of the technical tools and the arguments used to establish our results, but our setup is more general in certain respects. Like Bickel et al. [1998], we consider models with a finite hidden state space, but allow for autoregressive dynamics in the observation sequence, covariate-dependence in the transition probabilities of the regime sequence, and potential model misspecification. In Douc et al. [2004b], the hidden Markov chain is allowed to take values in a compact topological space, but is restricted to be temporally homogeneous and the model is assumed to be correctly specified. The cornerstone of the methods used in these papers for establishing the asymptotic properties of the ML estimator are mixing-type results for the unobservable regime sequence conditional on the observation sequence (see also Bickel and Ritov [1996]). This is also true for our approach, although the aforementioned results cannot be invoked directly because they are established under the assumption of temporal homogeneity of the hidden Markov chain. We, therefore, extend these results to allow for more general Markov regime sequences; in particular, we establish mixing-type results for the unobservable regime sequence given the observed data, allowing for a particular form of covariate-dependence in the transition kernels. This last result is, to our knowledge, novel and may be of interest in its own right.

The remainder of the paper is organized as follows. Section 2 defines the class of models under consideration, gives sufficient conditions for stationarity and ergodicity of the observation process, and describes the estimation problem. Section 3 investigates consistency of the ML estimator in a general setting. Section 4 contains results on the LAN property of the model and the asymptotic normality of the ML estimator. Section 5 presents simulation results on the finite-sample properties of estimators based on well-specified and misspecified likelihoods. Section 6 presents an illustration using real-world data. Proofs of the main results are gathered in an Appendix.

The following notational conventions are used throughout the paper. For an infinite sequence $(V_j)_j$, $V_a^b = (V_a, \ldots, V_b)$ for any $a \leq b$; $\mathcal{P}(\mathbb{V})$ denotes the set of Borel probability measures on a Polish space $\mathbb{V}$; for a probability measure P, $E_P(\cdot)$ denotes expectation with respect to P, $o_P(\cdot)$ and $O_P(\cdot)$ indicate order in probability under P, $\Rightarrow_P$ signifies weak convergence under P, and $L^r(P)$, $1 \leq r < \infty$, denotes the class of measurable functions integrable to order r with respect to P; ∇_ϑ and ∇^2_ϑ are the gradient and Hessian operators, respectively, with respect to ϑ; $\|\cdot\|$ denotes the Euclidean norm of a vector or matrix; $1\{\cdot\}$ denotes the indicator function; $\mathbb{N}$ denotes the set of positive integers. Unless stated otherwise, limits are taken as the sample size, T, diverges to infinity.

2 Model and Estimation

2.1 Statistical Model

Let $(X_t, S_t)_{t=0}^{\infty}$ be a discrete-time stochastic process such that, for each $t \in \mathbb{N} \cup \{0\}$ and some $h \in \mathbb{N}$, $S_t \in \mathbb{S} \equiv \{s_1, \ldots, s_{|\mathbb{S}|}\} \subset \mathbb{R}$ is the unobservable state and $X_t \in \mathbb{X} \subseteq \mathbb{R}^h$ is the observable state. Moreover, for each $t \in \mathbb{N}$, the conditional distribution of X_t, given X_0^{t-1} and S_0^t, depends only on X_{t-1} and S_t, and the conditional distribution of S_t, given X_0^{t-1} and S_0^{t-1}, depends only on X_{t-1} and S_{t-1}, so that

$$\begin{aligned} X_t \mid (X_0^{t-1}, S_0^t) &\sim P_*(X_{t-1}, S_t, \cdot), \\ S_t \mid (X_0^{t-1}, S_0^{t-1}) &\sim Q_*(X_{t-1}, S_{t-1}, \cdot), \end{aligned}$$

with $(x, s) \mapsto P_*(x, s, \cdot) \in \mathcal{P}(\mathbb{X})$ and $(x, s) \mapsto Q_*(x, s, \cdot) \in \mathcal{P}(\mathbb{S})$ denoting the true transition probabilities. It is further assumed that, for each $(x, s) \in \mathbb{X} \times \mathbb{S}$, $P_*(x, s, \cdot)$ admits a density $p_*(x, s, \cdot)$ with respect to some σ-finite measure on $\mathbb{X}$. Our framework imposes no additional restrictions on this measure; for instance, it can be the Lebesgue measure (i.e., allow for continuous X_t) or a counting measure (i.e., allow for discrete X_t).

The researcher's model is given by a family of transition probabilities $(x, s) \mapsto P_\theta(x, s, \cdot) \in \mathcal{P}(\mathbb{X})$ and $(x, s) \mapsto Q_\theta(x, s, \cdot) \in \mathcal{P}(\mathbb{S})$ indexed by an (unknown) parameter $\theta \in \Theta \subseteq \mathbb{R}^q$, for some $q \in \mathbb{N}$, such that, for each $\theta \in \Theta$,

$$X_t \mid (X_0^{t-1}, S_0^t) \sim P_\theta(X_{t-1}, S_t, \cdot),$$

$$S_t \mid (X_0^{t-1}, S_0^{t-1}) \sim Q_\theta(X_{t-1}, S_{t-1}, \cdot),$$

and, for each $(x, s) \in \mathbb{X} \times \mathbb{S}$, $P_\theta(x, s, \cdot)$ admits a density $p_\theta(x, s, \cdot)$ with respect to the same measure used to define $p_*(x, s, \cdot)$.

A few remarks about this setup are worth making. First, and perhaps most importantly, the unobservable states (regimes) are a Markov chain whose transition kernel can depend on the lagged value of the observable state. This is the main departure from prior literature which, with the exception of Ailliot and Pène [2015], has focused on the case where the conditional distribution of S_t, given X_0^{t-1} and S_0^{t-1}, depends only on S_{t-1}. Second, the model $\{(P_\theta, Q_\theta) \colon \theta \in \Theta\}$ is allowed to be misspecified in the sense that $(P_*, Q_*) \notin \{(P_\theta, Q_\theta) \colon \theta \in \Theta\}$.

This setup encompasses a rich family of models that arise in econometric and statistical applications, some examples of which are given below. Note that, although the family $\{Q_\theta \colon \theta \in \Theta\}$ may or may not contain Q_*, it is defined on the same finite state space $\mathbb{S}$ as Q_*, so misspecification of the number of unobservable regimes is ruled out.[3] Finally, even though the conditional distribution of S_t, given X_0^{t-1} and S_0^{t-1}, is assumed to depend only on X_{t-1} and S_{t-1}, it is straightforward to extend all our results to situations where this distribution is dependent on additional higher-order lags of X_t and/or S_t.

Example 1 (Hidden Markov Model with Covariate-Dependent Transition Probabilities)**.** Let $x = (y, z) \in \mathbb{X} = \mathbb{R}^2$ and $\mathbb{S} = \{0, 1\}$. Let P_* be determined by the equations

$$
\begin{aligned}
Y_t &= \mu^*(S_t) + \sigma^*(S_t)U_{1,t}, \\
Z_t &= \mu_2^* + \psi^* Z_{t-1} + \sigma_2^* U_{2,t},
\end{aligned}
$$

where $(U_{1,t}, U_{2,t})_t$ are i.i.d. (independent of $(S_t)_t$) with zero mean and covariance matrix indexed by a parameter ρ^*, e.g., $\begin{bmatrix} 1 & \rho^* \\ \rho^* & 1 \end{bmatrix}$. The transition probabilities of $(S_t)_t$ are allowed to depend on Z_{t-1}; for instance, $(s, z) \mapsto Q_*(z, s, s) \equiv \Pr(S_t = s \mid Z_{t-1} = z, S_{t-1} = s) = [1 + \exp(-\alpha_s^* - \beta_s^* z)]^{-1}$ for $s \in \mathbb{S}$. The homogeneous specification with $\beta_0^* = \beta_1^* = 0$ has been used to model regime shifts in a variety of economic and financial time series, including output growth (Albert and Chib [1993], Gadea Rivas and Perez-Quiros [2015]), foreign exchange rates (Engel and Hamilton [1990], Bollen et al. [2008]), and equity returns (Rydén et al. [1998], Ang and Bekaert [2002a]). The homogeneity restriction is relaxed in Diebold et al. [1994], Engel and Hakkio [1996], and Ang and Bekaert [2002a], among others, to allow the transition probabilities to depend on Z_{t-1}. The results obtained here establish the asymptotic properties of the ML estimator in this class of models. $\triangle$

Example 2 (Markov-Switching Autoregressive Model with Covariate-Dependent Transition Probabilities)**.** A useful generalization of the previous example is one where the outcome equation is extended to

$$
Y_t = \mu^*(S_t) + \phi^* Y_{t-1} + \sigma^*(S_t)U_{1,t}.
$$

Variations of the model with $\beta_0^* = \beta_1^* = 0$ have found widespread application in economics (e.g., Hansen [1992], McCulloch and Tsay [1994], Ruge-Murcia [1995], Ang et al. [2008]) and beyond. Generalizations of the model without the restriction of time-invariant transition probabilities are also very popular and

[3] Such misspecification would introduce additional complexities (e.g., locally non-quadratic likelihood surfaces, non-identifiable parameters) which are beyond the scope of this paper.

have been used, for example, in the modeling of output growth (Gadea Rivas and Perez-Quiros [2015]), interest rates (Ang and Bekaert [2002b]), consumption growth (Whitelaw [2000]), and bond spreads (Psaradakis and Sola [2021]). The results obtained here establish the asymptotic properties of the ML estimator in this class of models. $\triangle$

Example 3 (Mixture Autoregressive Model)**.** Let $x \in \mathbb{X} = \mathbb{R}$, $\mathbb{S} = \{0, 1\}$, and for each $t \in \mathbb{N}$, $\Pr(S_t = 0 \mid X_{t-1}) = G_*(X_{t-1})$ for some $x \mapsto G_*(x) \in [0, 1]$ and $X_t \sim P_*(X_{t-1}, S_t, \cdot)$. This specification implies a conditional density for X_t, given X_{t-1}, that is a mixture of the type

$$x \mapsto p_*(x \mid x_{t-1}) = G_*(x_{t-1})p_*(x_{t-1}, 0, x) + (1 - G_*(x_{t-1}))p_*(x_{t-1}, 1, x).$$

The covariate-dependence of the transition functions is reflected by the fact that G_* depends on X_{t-1}. Models which belong to the general class of mixture autoregressive models (e.g., Dueker et al. [2007], Tadjuidje et al. [2009], Dueker et al. [2011], Kalliovirta et al. [2015]) are covered by this framework. $\triangle$

The examples above, as well as those that follow, illustrate that in many areas of application the stochastic process $(X_t, S_t)_{t=0}^{\infty}$ is typically highly complex and it is natural/desirable to allow for feedback from past realizations of the observable process $(X_t)_{t=0}^{\infty}$ to the law of the unobservable regime sequence $(S_t)_{t=0}^{\infty}$; a tractable way for modeling such feedback is to allow the transition kernel Q_θ to depend on X_{t-1}. This feature adds an additional level of complexity and with it sources of potential misspecification. For instance, a common assumption in most applications is that the transition probabilities of hidden regimes are time-invariant, an assumption which may result in misspecification of the transitions functions. In applications that allow for covariate-dependent transition functions, a somewhat more subtle and often overlooked source of misspecification, associated with endogeneity of the transition-driving covariates, may come into play. Specifically, in the context of models such as those in Examples 1 and 2, contemporaneous correlation between $U_{1,t}$ and $U_{2,t}$ is typically ignored and inference is based on the likelihood implied by the outcome equation alone; we consider this case in more detail in Sections 5.2 and 6.

Before discussing estimation of θ, we give a result regarding the mixing and ergodicity properties of $(X_t)_{t=0}^{\infty}$. To do so, let $\bar{P}_*^{\kappa}$ denote the true distribution over $(X_t)_{t=0}^{\infty}$ when the distribution of (X_0, S_0) is κ. Under the following assumptions, Lemma 1 below ensures that there exists a Borel probability measure on $\mathbb{X} \times \mathbb{S}$, denoted henceforth by ν, for which $(X_t)_{t=0}^{\infty}$ is stationary and ergodic.

Assumption 1. *There exists a continuous function* $\underline{q} : \mathbb{X} \to \mathbb{R}_+ \setminus \{0\}$ *such that, for all* $Q \in \{Q_\theta : \theta \in \Theta\} \cup Q_*$, $Q(x, s, s') \geq \underline{q}(x)$ *for all* $(s', s, x) \in \mathbb{S}^2 \times \mathbb{X}$.

Assumption 2. *There exist constants* $\lambda' \in (0,1)$, $\gamma \in (0,1)$, $b' > 0$ *and* $R > 2b'/(1-\gamma)$, *a lower semi-continuous function* $\mathcal{U} : \mathbb{X} \to [1,\infty)$, *and a measure* $\varpi \in \mathcal{P}(\mathbb{X})$ *such that, for all* $s \in \mathbb{S}$: *(i)* $\int_{\mathbb{X}} \mathcal{U}(x')P_*(x,s,dx') \leq \gamma\mathcal{U}(x) + b'1\{x \in A\}$, *with* $A \equiv \{x \in \mathbb{X} \colon \mathcal{U}(x) \leq R\}$; *(ii)* A *is bounded and* $\varpi(A) > 0$; *(iii)* $\inf_{x\in A} P_*(x,s,C) \geq \lambda'\varpi(C)$ *for any Borel set* $C \subseteq \mathbb{X}$.

The following lemma establishes stationarity, ergodicity, and β-mixing of $(X_t)_{t=0}^{\infty}$.

Lemma 1. *Suppose Assumptions 1 and 2 hold. Then, there exists a* $\nu \in \mathcal{P}(\mathbb{X} \times \mathbb{S})$ *such that, under* $\bar{P}_*^{\nu}$, $(X_t)_{t=0}^{\infty}$ *is stationary, ergodic, and* β*-mixing with mixing coefficients* $\beta_n = O(\gamma^n)$, $n \in \mathbb{N}$.

Proof. See Supplemental Material SM.1. □

The result follows in a standard manner by using Assumptions 1 and 2 to establish that the implied transition kernel of the joint process $(X_t, S_t)_{t=0}^{\infty}$ has a unique invariant distribution and also that it is Harris recurrent and aperiodic. This fact, in turn, is used to show that $(X_t)_{t=0}^{\infty}$ is stationary, ergodic, and β-mixing at a geometric rate.

Remark 1 (Discussion of Assumptions 1 and 2)**.** Assumption 1 is an extension of a common assumption in the literature (cf. Douc et al. [2004b], Ailliot and Pène [2015]) to the case where the transition kernel of $(S_t)_{t=0}^{\infty}$ depends on X_{t-1}. Allowing the lower bound $\underline{q}$ to depend on x is especially relevant when the support of X_t is unbounded because, while $\underline{q}(x) > 0$, it is allowed to converge to zero as $\|x\| \to \infty$. Although this assumption is not innocuous, we view it as mild because it accommodates the typical specifications used in the literature, where Q is parameterized by a standard Gaussian or logistic cumulative distribution function and a single index $x^{\intercal}\beta$, with β restricted to take values in a bounded subset of a finite-dimensional Euclidean space.

Assumption 2(iii) is an analogous condition for the transition kernel P_*. By inspection of the proof of Lemma 1, it is easy to see that it suffices to obtain a minorization condition for the "joint" kernel, i.e., $\inf_{x\in A} P_*(x,s',C)Q(x,s,s') \geq \lambda\widetilde{\varpi}(C,s')$ for any Borel set $C \subseteq \mathbb{X}$ and for some $\widetilde{\varpi} \in \mathcal{P}(\mathbb{X}\times\mathbb{S})$ and $\lambda \in (0,1)$. Thus, Assumptions 1(i) and 2(i) could be relaxed; e.g., the former could be relaxed to $Q(x,s,s') \geq \underline{q}(x)\varrho(s')$, where $\varrho \in \mathcal{P}(\mathbb{S})$, or the latter could be relaxed to $\inf_{x\in A} P_*(x,s',C) \geq \lambda'\widetilde{\varpi}(C,s')$, where $\widetilde{\varpi} \in \mathcal{P}(\mathbb{X}\times\mathbb{S})$.

Assumption 2(i),(ii) is a so-called Foster–Lyapunov drift condition; see Meyn and Tweedie [1993] and references therein for a discussion of the assumption. △

In view of Lemma 1, under ν, the process $(X_t)_{t=0}^{\infty}$ can be extended to a two-sided sequence $(X_t)_{t=-\infty}^{\infty}$. With a slight abuse of notation, we still use $\bar{P}_*^{\nu}$ to denote the true probability distribution over $(X_t, S_t)_{t=-\infty}^{\infty}$; $\bar{P}_\theta^{\nu}$ is defined analogously for the model $(Q_\theta, p_\theta, \nu)$.[4]

2.2 Parameter Estimation

For any $T \in \mathbb{N}$, let $\ell_T^{\nu} : \mathbb{X}^{T+1} \times \Theta \to \mathbb{R}$ be the sample criterion function given by

$$\ell_T^{\nu}(X_0^T, \theta) = T^{-1} \sum_{t=1}^{T} \log p_t^{\nu}(X_t \mid X_0^{t-1}, \theta), \tag{1}$$

where $p_t^{\nu}(X_t \mid X_0^{t-1}, \theta)$ denotes the conditional density of X_t given X_0^{t-1} for any $\theta \in \Theta$; the latter is defined recursively as follows: for any $t \geq 1$,

$$p_t^{\nu}(X_t \mid X_0^{t-1}, \theta) = \sum_{s' \in \mathbb{S}} \sum_{s \in \mathbb{S}} p_\theta(X_{t-1}, s', X_t) Q_\theta(X_{t-1}, s, s') \delta_t^{\theta,\nu}(s),$$

and $s \mapsto \delta_t^{\theta,\nu}(s) \equiv \bar{P}_\theta^{\nu}(S_{t-1} = s \mid X_0^{t-1})$. For each $t \geq 2$ and any $s \in \mathbb{S}$, $s \mapsto \delta_t^{\theta,\nu}(s)$ satisfies the recursion

$$\delta_t^{\theta,\nu}(s) = \sum_{\tilde{s} \in \mathbb{S}} \frac{Q_\theta(X_{t-1}, \tilde{s}, s) p_\theta(X_{t-2}, \tilde{s}, X_{t-1}) \delta_{t-1}^{\theta,\nu}(\tilde{s})}{\sum_{s' \in \mathbb{S}} p_\theta(X_{t-2}, s', X_{t-1}) \delta_{t-1}^{\theta,\nu}(s')},$$

with $s \mapsto \delta_1^{\theta,\nu}(s) = \sum_{\tilde{s} \in \mathbb{S}} Q_\theta(X_0, \tilde{s}, s) \nu(\tilde{s}|X_0)$, where $\nu(\cdot|\cdot)$ is the conditional density corresponding to ν.

For a given initial distribution $\kappa \in \mathcal{P}(\mathbb{X} \times \mathbb{S})$ over (X_0, S_0), we define our estimator as $\hat{\theta}_{\kappa,T}$, where

$$\ell_T^{\kappa}(X_0^T, \hat{\theta}_{\kappa,T}) \geq \sup_{\theta \in \Theta} \ell_T^{\kappa}(X_0^T, \theta) - \eta_T, \tag{2}$$

for some $\eta_T \geq 0$ and $\eta_T = o(1)$.

3 Consistency

The main result of this section establishes convergence of the estimator $\hat{\theta}_{\nu,T}$ to the set of points in Θ that are closest to the true model under the Kullback–Leibler information criterion.

[4]Throughout the text, we use $\bar{P}_\theta^{\nu}$ to denote any marginal or conditional probabilities associated with $\bar{P}_\theta^{\nu}$; the same holds for $\bar{P}_*^{\nu}$.

Let $H^* : \Theta \to \mathbb{R}_+ \cup \{\infty\}$ be the Kullback–Leibler information criterion $\theta \mapsto H^*(\theta)$, which is given by

$$H^*(\theta) = E_{\bar{P}_*^\nu}\left[\log \frac{p_*^\nu(X_0 \mid X_{-\infty}^{-1})}{p^\nu(X_0 \mid X_{-\infty}^{-1}, \theta)}\right],$$

where, for any $\theta \in \Theta$, $p^\nu(X_0 \mid X_{-\infty}^{-1}, \theta)$ denotes the conditional density of X_0 given $X_{-\infty}^{-1}$ induced by $(P_\theta, Q_\theta, \nu)$, and $p_*^\nu(X_0 \mid X_{-\infty}^{-1})$ is its counterpart induced by the true transition kernels (P_*, Q_*, ν); we refer the reader to the Supplemental Material SM.3 for details about the construction of these objects and their properties.

Our results allow for misspecified models, and thus $p_*^\nu \notin \{p^\nu(\cdot \mid \cdot, \theta) : \theta \in \Theta\}$. Hence, as in White [1982], the relevant limiting set for our estimator is

$$\Theta_* = \arg\min_{\theta \in \Theta} H^*(\theta),$$

which is the *pseudo-true parameter (set)* that minimizes the Kullback–Leibler information criterion. Under Assumption 3 below, Θ_* is non-empty and compact by the Weierstrass Theorem.

Assumption 3. *(i) Θ is compact; (ii) H^* exists and is lower semi-continuous.*

The following additional assumptions are used to establish the main result. To state these, for any $\delta > 0$ and any $\dot{\theta} \in \Theta$, let $B(\delta, \dot{\theta}) \equiv \{\theta \in \Theta \colon ||\dot{\theta} - \theta|| < \delta\}$.

Assumption 4. *(i) For any $\epsilon > 0$, there exists some $\delta > 0$ such that*

$$\max_{\dot{\theta} \in \Theta} E_{\bar{P}_*^\nu}\left[\sup_{\theta \in B(\delta, \dot{\theta})} \frac{p^\nu(X_0 \mid X_{-\infty}^{-1}, \theta)}{p^\nu(X_0 \mid X_{-\infty}^{-1}, \dot{\theta})}\right] \leq 1 + \epsilon;$$

(ii) there exists a function $(x, x') \mapsto C(x, x') \in \mathbb{R}_+$ such that $\sup_{\theta \in \Theta} \frac{\max_{s \in \mathbb{S}} p_\theta(X, s, X')}{\min_{s \in \mathbb{S}} p_\theta(X, s, X')} \leq C(X, X')$ and $\frac{\max_{s \in \mathbb{S}} p_*(X, s, X')}{\min_{s \in \mathbb{S}} p_*(X, s, X')} \leq C(X, X')$ a.s.-$\bar{P}_*^\nu$.

Assumption 5. $T^{-1} \sum_{t=1}^T \max\{1, C(X_{t-1}, X_t)\} \prod_{i=0}^{t-1}(1 - \underline{q}(X_i)) = o_{\bar{P}_\nu^*}(1)$.

Remark 2 (Discussion of Assumptions 3, 4 and 5)**.** Assumption 3(i) is standard. Assumption 3(ii) is high-level but can be obtained from lower-level conditions (e.g., by following the reasoning in Proposition 11 of Douc and Moulines [2012]).

Assumption 4(i) is a high-level condition used for establishing uniform law of large numbers results (see Lemma 3 in Section A.1). Assumption 4(ii) is akin to Assumption A4 in Bickel et al. [1998]; it essentially restricts the support of p_θ and p_* for different values of the state variable.[5]

Assumption 5 essentially requires that $\underline{q}$ is not "too close" to zero on average and that $C(X_{t-1}, X_t)$ is finite a.s.-$\bar{P}^*_\nu$. For instance, if $\underline{q}(x) \geq c$ for some $c > 0$, then Assumption 5 is automatically satisfied provided $E_{\bar{P}^*_\nu}[C(X_0, X_1)] < \infty$. Moreover, by exploiting the fact that $(X_t)_t$ is β-mixing (see Lemma 1), Lemma 13 in the Supplemental Material SM.4 provides sufficient conditions of the form $E_{\bar{P}^*_\nu}\left[\underline{q}(X_1)\right] > 0$ and $E_{\bar{P}^*_\nu}\left[C(X_1, X_0)^l\right] < \infty$ for some $l > 1$.[6] $\triangle$

We now establish consistency of the estimator defined by (2).

Theorem 1. *Suppose Assumptions 1–4 hold. Then,* $d_\Theta(\hat{\theta}_{\nu,T}, \Theta_*) = o_{\bar{P}^\nu_*}(1)$.[7]

Proof. See Appendix A.1. □

This result is analogous to Theorem 2 in Douc and Moulines [2012] but for a somewhat different setup; specifically, we allow for autoregressive dynamics and covariate-dependent transition probabilities, but restrict $\mathbb{S}$ to be finite.[8]

Clearly, if the model is correctly specified and point identified, i.e., there exists a $\theta \in \Theta$ such that $(P_*, Q_*) = (P_\theta, Q_\theta)$, then $\Theta_* = \{\theta\}$ and our estimator converges in $\bar{P}^\nu_*$-probability to this point. If, however, the model is misspecified, our estimator converges to the set of parameters that is closest to the true set, when closeness is measured by means of the Kullback–Leibler information criterion (cf. White [1982], Douc and Moulines [2012]).

To prove Theorem 1, we first show that $T^{-1}\sum_{t=1}^T \log p^\nu_t(X_t \mid X_0^{t-1}, \theta)$ is well-approximated by $T^{-1}\sum_{t=1}^T \log p^\nu(X_t \mid X_{-\infty}^{t-1}, \theta)$ (see Lemma 2 in Appendix A.1). Second, relying on ergodicity (Lemma 1) and Assumption 4, we establish a *uniform* law of large numbers for the latter quantity (see Lemma 3 in Appendix A.1). The proof of consistency then follows the standard Wald approach.

The approximation result in the first step relies on "mixing" results for the process $(S_t)_{t=-\infty}^\infty$, given $(X_t)_{t=-\infty}^\infty$. The following theorem, which might be of independent interest, establishes such a "mixing" result in our setting.

[5] The second part of Assumption 4(ii) is used to show that $p^\nu_*(\cdot|X^{-1}_{-\infty})$ integrates to one.

[6] We thank a referee and the co-editor for suggestions on how to weaken Assumption 4, which led to these sufficient conditions for Assumption 5.

[7] For any set $A \subseteq \Theta$, $d_\Theta(\theta, A) \equiv \inf_{\dot{\theta}\in A} ||\theta - \dot{\theta}||$.

[8] Finiteness of $\mathbb{S}$ simplifies the proofs as we do not have to be concerned with uniformity issues of certain quantities as functions of the hidden state. By combining techniques available in the literature (e.g., Douc and Moulines [2012]) – which essentially amount to imposing requirements like compactness of $\mathbb{S}$ and continuity – with ours, we conjecture that our results can be extended to the case where $\mathbb{S}$ is a compact set.

Theorem 2. *Take any* $(j,m) \in \mathbb{N}^2$*. Suppose that, for any* $\theta \in \Theta$*, there exist mappings* $x \mapsto \varrho(x,\cdot) \in \mathcal{P}(\mathbb{S})$ *and* $\underline{q} : \mathbb{X} \to \mathbb{R}_+$ *such that, for all* $(s,s') \in \mathbb{S}^2$*,*

$$Q_\theta(X,s,s') \geq \underline{q}(X)\varrho(X,s') \quad a.s.\text{-}\bar{P}^\nu_*. \tag{3}$$

Then,[9]

$$\max_{(b,c)\in\mathbb{S}^2} \left\| \bar{P}^\nu_\theta(S_{j+1} = \cdot | S_{-m} = b, X^j_{-m}) - \bar{P}^\nu_\theta(S_{j+1} = \cdot | S_{-m} = c, X^j_{-m}) \right\|_1 \leq \prod_{n=-m}^{j} (1 - \underline{q}(X_n)).$$

Proof. See Appendix A.2. □

This result is analogous to results in Douc et al. [2004b] (e.g., their Lemma 1 and Corollary 1) but for a more general transition function Q_θ (albeit under the requirement that $\mathbb{S}$ is finite). Specifically, we allow Q_θ to depend on X, and do not restrict the lower bound in (3) to be uniform in s'; these features are, to our knowledge, novel.

The proof of Theorem 2 relies on bounding the Dobrushin coefficient of the transition kernel $\bar{P}^\nu_\theta(S_{l+1} = \cdot | S_l = \cdot, X^j_{-m})$ by $1 - \underline{q}(X_l)$, for each $l \in \{-m, ..., j\}$. The case where $\varrho(x,\cdot)$, in condition (3), is uniformly bounded from below has been studied in the literature, and such a bound can be obtained by elementary calculations. For the general case, however, the standard technique cannot be used, and we develop a novel, to our knowledge, approach based on coupling techniques; see Appendix A.2 for details.

Remark 3. Remark 1 and the fact that Theorem 2 is established under condition (3), imply that, in regards to consistency, Assumption 1 could be replaced by the weaker condition (3). This remark, however, does not extend to the LAN results of Section 4, since we do not know whether Assumption 1 could be weakened to condition (3) in this case. △

We discuss next a canonical example that encompasses many commonly used specifications, such as those in Examples 1 and 2. The purpose of this example is to verify our assumptions under primitive and low-level conditions.

Example 4 (Canonical Example)**.** We verify the regularity conditions for a model with $\mathbb{S} = \{0,1\}$ and

$$\begin{aligned} X_t &= \mu(S_t) + \Phi^\intercal X_{t-1} + \Sigma(S_t)^{1/2}\varepsilon_t, \\ S_t &\sim Q_{\bar{\vartheta}}(X_{t-1}, S_{t-1}, \cdot), \end{aligned}$$

where $(\varepsilon_t)_t \sim$ i.i.d. $\mathcal{N}(0,I)$, I being the identity matrix, and $x \mapsto Q_{\bar{\vartheta}}(x,s,s) \equiv \Pr(S_t = s \mid X_{t-1} = x, S_{t-1} = s) = \Psi(x^\intercal\bar{\vartheta}_s)$ for $s \in \mathbb{S}$, where Ψ is a full-support,

[9]For any $P,Q \in \mathcal{P}(\mathbb{S})$, $||P - Q||_1 \equiv \sum_{s\in\mathbb{S}} |P(s) - Q(s)|$.

continuous cumulative distribution function (e.g., logistic or normal). It is assumed that the parameter set Θ is compact and that any $\theta \equiv ((\mu(s), \Sigma(s))_{s\in\mathbb{S}}, \Phi, \bar{\vartheta})$ in it is such that $\Sigma(\cdot)$ and $\Phi\Phi^{\intercal}$ have eigenvalues uniformly bounded away from zero and infinity. For simplicity, we assume that the true model is indexed by $((\mu_*(s), \Sigma_*(s))_{s\in\mathbb{S}}, \Phi_*, \bar{\vartheta}_*)$, for which analogous conditions hold.

Let $q(x) \equiv \inf_b \Psi(x^{\intercal}b)$ and note that $q(x) > 0$ for each finite x as Ψ has full support. Thus, Assumption 1 holds and $E_{\bar{P}^*_\nu}[q(X)] > 0$. The conditions of Assumption 2 follow by the results in Douc et al. [2004a]; see Lemma 14 in the Supplemental Material SM.5 for the formal argument. In Lemma 15 in the Supplemental Material SM.5, it is shown that

$$C^{-1}\underline{p}(x,y) \leq f_{\mathcal{N}}(\{y - \Phi^{\intercal}x - \mu(s)\}\Sigma(s)^{-1/2}) \leq C\overline{p}(x,y),$$

for some $C \geq 1$, where $f_{\mathcal{N}}$ is the $\mathcal{N}(0, I)$ probability density function,

$$\begin{aligned}\underline{p}(x,y) &\equiv \exp\{-0.5(a_1\|y\|^2 + a_2\|x\|^2 + 2a_3\|x\|\|y\|) - a_4\|y\| - a_5\|x\|\},\\ \overline{p}(x,y) &\equiv \exp\{-0.5(b_1\|y\|^2 + b_2\|x\|^2 - 2b_3\|x\|\|y\|) + b_4\|y\| + b_5\|x\|\},\end{aligned}$$

and a_i, b_i $(i = 1, \ldots, 5)$ are positive constants that are functions of uniform bounds of $\Sigma(\cdot)$, $\mu(\cdot)$ and Φ over Θ; they are defined in Lemmas 15 and 16 in the Supplemental Material SM.4. Therefore, by Lemma 12 in the Supplemental Material SM.3, it follows that Assumptions 3 and 4 hold with $(x, x') \mapsto C(x, x') \equiv \overline{p}(x, x')/\underline{p}(x, x')$, provided that $E_{\bar{P}^*_\nu}[\exp\{-0.5l((b_1 - a_1)\|Y\|^2 + (b_2 - a_2)\|X\|^2 - 2(b_3 + a_3)\|X\|\|Y\|) + l(b_4 + a_4)\|y\| + l(b_5 + a_5)\|X\|\}] < \infty$ for some $l \geq 1$. Lemma 16 in the Supplemental Material SM.5 shows that the latter condition holds under some restrictions on the eigenvalues of $\Sigma(\cdot)$, $\Sigma_*(\cdot)$, $\Phi\Phi^{\intercal}$, and $\Phi_*\Phi_*^{\intercal}$ (see the aforementioned lemma and its subsequent remark for the exact formulation and more details). Finally, under these conditions, Lemma 13 in the Supplemental Material SM.4 shows that Assumption 5 holds. $\triangle$

Lastly, we note that further characterization and interpretation of the pseudo-true parameter set Θ_* is not straightforward in our general setup, and we believe that progress in this direction should be made on a case-by-case basis. While pursuing this program is outside the scope of the present paper, in Section 5 we present numerical results that attempt to shed some light on the characterization of this set.[10]

[10] For a subclass of the models in Example 2, namely hidden Markov models with covariate-dependent transition probabilities, Θ_* can be characterized analytically when the (misspecified) model is a simple mixture. Specifically, it can be shown that, even though the mixture model (with i.i.d. regimes) misspecifies the dependence structure of the hidden state, and thus of the overall system, parameters related to the outcome equation are consistently estimated; details can be found in the Supplemental Material SM.6 of Pouzo et al. [2021].

4 Asymptotic Distribution Theory

In this section, we establish a LAN property ([Ibragimov and Has'minskii, 1981, Ch. II], Le Cam [1986]) for our model and an asymptotic linear representation for our estimator, from which asymptotic normality of the estimator is inferred. For this, we require the following assumptions.

Assumption 6. *(i)* $\Theta_* = \{\theta_*\} \subset \text{int}(\Theta)$*; (ii)* $\theta \mapsto p_\theta(X, s, X')$ *and* $\theta \mapsto Q_\theta(X, s, s')$ *are twice continuously differentiable a.s.-*$\bar{P}^\nu_*$ *for all* $(s, s') \in \mathbb{S}^2$*.*

Assumption 7. *For some* $\delta > 0$ *and* $a \geq 1$*, and for all* $(s', s) \in \mathbb{S}^2$*: (i)*

$$E_{\bar{P}^\nu_*}\left[\sup_{\theta \in B(\delta,\theta_*)} \left\|\nabla_\theta \log p_\theta(X, s, X')\right\|^{2a}\right] < \infty \text{ and } E_{\bar{P}^\nu_*}\left[\sup_{\theta \in B(\delta,\theta_*)} \left\|\nabla_\theta \log Q_\theta(X, s, s')\right\|^{2a}\right] < \infty;$$

(ii)

$$E_{\bar{P}^\nu_*}\left[\sup_{\theta \in B(\delta,\theta_*)} \left\|\nabla^2_\theta \log p_\theta(X, s, X')\right\|^{2a}\right] < \infty \text{ and } E_{\bar{P}^\nu_*}\left[\sup_{\theta \in B(\delta,\theta_*)} \left\|\nabla^2_\theta \log Q_\theta(X, s, s')\right\|^{2a}\right] < \infty.$$

Assumption 8. $\sum_{j=0}^{\infty} \left(E_{\bar{P}^\nu_*}\left[\prod_{i=0}^{j}(1 - \underline{q}(X_i))^{\frac{2a}{1-a}}\right]\right)^{p\left(\frac{1-a}{2a}\right)} < \infty$ *for some* $p \in (0, 2/3)$ *(and the same* $a \geq 1$ *that appears in Assumption 7).*

Remark 4 (Discussion of Assumptions 6, 7 and 8)**.** Part (i) of Assumption 6 is standard in the literature. The restriction that Θ_* is a singleton could be relaxed using the ideas of Liu and Shao [2003] for non-identified ML estimators. This extension, albeit interesting, would present nuances that are beyond the scope of the present paper. Part (ii) of Assumption 6 is also standard, and so is Assumption 7 (see Bickel et al. [1998] for a discussion). Finally, Assumption 8 is a strengthening of Assumption 5, and is required in order to establish the existence of a random sequence $(\Delta_t(\theta_*))_t$ which approximates the "score" function well (in the sense of Lemma 18 in Appendix A.3). As was the case with Assumption 5, Lemma 13 in the Supplemental Material SM.4 provides sufficient conditions of the form $E_{\bar{P}^*_\nu}\left[(1 - \underline{q}(X_1))^{2a/(1-a)}\right] < 1$. $\triangle$

The next theorem establishes a LAN-type property for the log-likelihood criterion function defined in (1).

Theorem 3. *Suppose Assumptions 1, 2, 6, 7 and 8 hold. Then, there exists a stationary and ergodic process* $(\Delta_t(\theta_*))_t$ *in* $L^2(\bar{P}^\nu_*)$*, a sequence of negative definite matrices* $(\xi_t(\theta_*))_t$*, and a compact set* $K \subseteq \Theta$ *that includes* 0*, such that,* *for any $v \in K$,*

$$\ell_T^{\nu}(X_0^T, \theta_* + v) - \ell_T^{\nu}(X_0^T, \theta_*) = v^{\intercal}\left(T^{-1}\sum_{t=0}^{T}\Delta_t(\theta_*) + o_{\bar{P}_*^{\nu}}(T^{-1/2})\right) + (1/2)v^{\intercal}\left(T^{-1}\sum_{t=0}^{T}\xi_t(\theta_*) + o_{\bar{P}_*^{\nu}}(1)\right)v + R_T(v),$$

where $v \mapsto R_T(v) \in \mathbb{R}$ is such that $\lim_{\delta\to 0}\bar{P}_*^{\nu}\left(\sup_{v\in B(\delta,0)}||v||^{-2}R_T(v) \geq \epsilon\right) = 0$ for any $\epsilon > 0$ and any $T \in \mathbb{N}$.

Proof. See Appendix A.3. □

Theorem 3 extends the results in Bickel and Ritov [1996] and Bickel et al. [1998] (see their remark on p. 1620) to a more general setup which allows for covariate-dependent transition probabilities, autoregressive dynamics, and misspecified models. The proof develops along the same lines as theirs. The main difference relates to the way in which one establishes that the "score" $\nabla_\theta \ell_T^{\nu}(\cdot, \theta_*)$ and the Hessian $\nabla_\theta^2 \ell_T^{\nu}(\cdot, \theta_*)$ can be approximated by $T^{-1}\sum_{t=0}^{T}\Delta_t(\theta_*)$ and $T^{-1}\sum_{t=0}^{T}\zeta_t(\theta_*)$, respectively (see Lemmas 7 and 8 in Appendix A.3). As mentioned earlier, these approximations rely on "mixing" properties of the *temporally inhomogeneous* hidden Markov chain; see Lemma 22 in the Supplemental Material SM.7.

Theorem 3 may be used to establish the following asymptotic linear representation for our estimator in terms of $(\Delta_t(\theta_*))_{t=0}^{\infty}$ and $(\xi_t(\theta_*))_{t=0}^{\infty}$.

Theorem 4. *Suppose Assumptions 1–8 hold and $\eta_T = o(T^{-1})$. Then,*

$$\frac{\sqrt{T}(\hat{\theta}_{\nu,T} - \theta_*)}{\sqrt{\mathrm{tr}\{\Sigma_T(\theta_*)\}}} = -\{E_{\bar{P}_*^{\nu}}[\xi_1(\theta_*)] + o_{\bar{P}_*^{\nu}}(1)\}^{-1}T^{-1/2}\sum_{t=0}^{T}\frac{\Delta_t(\theta_*)}{\sqrt{\mathrm{tr}\{\Sigma_T(\theta_*)\}}} + o_{\bar{P}_*^{\nu}}(1),$$

where $\Sigma_T(\theta_*) \equiv T^{-1}E_{\bar{P}_*^{\nu}}[\{\sum_{t=0}^{T}\Delta_t(\theta_*)\}\{\sum_{t=0}^{T}\Delta_t(\theta_*)^{\intercal}\}]$.

Proof. See Appendix A.3. □

Theorem 4 readily implies that, if $T^{-1/2}\Sigma_T(\theta_*)^{-1/2}\sum_{t=0}^{T}\Delta_t(\theta_*) \Rightarrow_{\bar{P}_*^{\nu}} \mathcal{N}(0, I)$, then

$$\sqrt{T}\Sigma_T(\theta_*)^{-1/2}E_{\bar{P}_*^{\nu}}[\xi_1(\theta_*)](\hat{\theta}_{\nu,T} - \theta_*) \Rightarrow_{\bar{P}_*^{\nu}} \mathcal{N}(0, I). \tag{4}$$

(If $\Sigma_T(\theta_*) \to \Sigma(\theta_*)$, $\Sigma_T(\theta_*)$ may be replaced by $\Sigma(\theta_*)$ in these statements). This result is akin to results in White [1982] and shares the same features, i.e., the asymptotic covariance matrix $\Omega_T(\theta_*) \equiv (E_{\bar{P}_*^{\nu}}[\xi_1(\theta_*)])^{-1}\Sigma_T(\theta_*)(E_{\bar{P}_*^{\nu}}[\xi_1(\theta_*)])^{-1}$

has a "sandwich" form and the familiar Fisher information equality does not necessarily hold (see also [White, 1994, Ch. 6]).

The next theorem complements Theorem 4 by presenting results on consistent estimation of the asymptotic covariance matrix of $\hat{\theta}_{\nu,T}$, vis., $\Omega_T(\theta_*)$, in both correctly specified and potentially misspecified models. In the latter case, the proposed estimator is of the "heteroskedasticity and autocorrelation consistent" type. As in many other results for such estimators (e.g., Newey and West [1987], [White, 1994, Ch. 8]), consistency requires some conditions over the score covariance process $(\Delta_{t+\tau}(\cdot)\Delta_t(\cdot)^\intercal)_{t=0}^T$, $\tau \in \mathbb{N}$, namely continuity and law of large numbers results (see Lemma 25 in the Supplemental Material SM.8). In what follows, the modulus of continuity is defined as

$$\ddot{\varpi}(\delta') \equiv \max_t \left\| \sup_{||\theta-\theta_*||<\delta'} \|\Delta_t(\theta)\Delta_0(\theta)^\intercal - \Delta_t(\theta_*)\Delta_0(\theta_*)^\intercal\| \right\|_{L^1(\bar{P}_*^\nu)},$$

for any $\delta' \in (0, \delta]$, where δ is as in Assumption 7.[11]

Theorem 5. *Suppose the assumptions of Theorem 4 hold.*

(a) If θ_* is such that, for any $t \geq 0$ and $T \geq 1$, $p_t^\nu(\cdot \mid X_{t-T}^{t-1}; \theta^*) = p_*^\nu(\cdot \mid X_{t-T}^{t-1})$, then

$$||\Omega_T(\theta_*) - \{-H_T^{-1}(\hat{\theta}_{\nu,T})\}|| = o_{\bar{P}_*^\nu}(1),$$

where

$$H_T(\theta) \equiv T^{-1} \sum_{t=1}^T \nabla_\theta^2 \log p_t^\nu(X_t|X_0^{t-1}, \theta).$$

(b) If, for any $l \geq 0$, $||E_{\bar{P}_*^\nu}[\Delta_l(\theta_*)\Delta_0(\theta_*)^\intercal]|| \leq \bar{\upsilon}(l)$ for some integrable function $l \mapsto \bar{\upsilon}(l) \in \mathbb{R}_+$, then

$$||\Omega_T(\theta_*) - H_T^{-1}(\hat{\theta}_{\nu,T}) J_T(\hat{\theta}_{\nu,T}) H_T^{-1}(\hat{\theta}_{\nu,T})|| = o_{\bar{P}_*^\nu}(1),$$

where

$$\begin{aligned} J_T(\theta) \equiv\ & T^{-1} \sum_{t=1}^T \nabla_\theta \log p_t^\nu(X_t|X_0^{t-1}, \theta) \nabla_\theta \log p_t^\nu(X_t|X_0^{t-1}, \theta)^\intercal \\ & + \sum_{\tau=1}^{L_T} \frac{\omega(\tau, L_T)}{T-\tau} \sum_{t=1}^{T-\tau} \nabla_\theta \log p_{t+\tau}^\nu(X_{t+\tau}|X_0^{t+\tau-1}, \theta) \nabla_\theta \log p_t^\nu(X_t|X_0^{t-1}, \theta)^\intercal \\ & + \sum_{\tau=1}^{L_T} \frac{\omega(\tau, L_T)}{T-\tau} \sum_{t=1}^{T-\tau} \nabla_\theta \log p_t^\nu(X_t|X_0^{t-1}, \theta) \nabla_\theta \log p_{t+\tau}^\nu(X_{t+\tau}|X_0^{t+\tau-1}, \theta)^\intercal, \end{aligned}$$

[11] Lemma 25 in the Supplementary Material SM.8 shows that the modulus of continuity converges to zero as δ vanishes.

$\omega(\cdot,\cdot)$ are bounded real weights with $\lim_{T\to\infty}\omega(\tau, L_T) = 1$ for all $\tau \geq 1$, and $(L_T)_{T=1}^{\infty} \subseteq \mathbb{N}$ is such that $L_T(\ddot{\varpi}(T^{-1/2}\log\log T)\log\log T + r_T + T^{-1/2}) = o(1)$, with $(r_T)_T$ being a positive sequence converging to zero.

Proof. See Supplemental Material SM.8. □

Part (a) of Theorem 5 deals with correctly specified models and makes use of the Fisher information equality. Part (b) provides a consistent covariance estimator in the general case of potentially misspecified models. The conditions for the weights $\omega(\tau, L_T)$ and the tuning parameters $(L_T)_T$ are standard for estimators of this type.[12] The terms $\ddot{\varpi}(T^{-1/2}\log\log T)\log\log T$ and r_T in the growth condition for L_T are analogous to those appearing in Newey and West [1987] and arise as "costs" of working with $\nabla_\theta \log p_t^\nu(X_t|X_0^{t-1}, \hat{\theta}_{\nu,T})$ as opposed to $\nabla_\theta \log p_t^\nu(X_t|X_0^{t-1}, \theta_*)$, and of working with sample averages as opposed to population means, respectively.[13] On the other hand, the $T^{-1/2}$ term does not appear in Newey and West [1987] and arises because we need to approximate $\Delta_t(\theta_*)$, that depends on $X_{-\infty}^t$, with the score, that depends on X_0^t.

Theorems 4 and 5, together with an asymptotic normality result such as (4), provide the means for constructing asymptotically correct confidence sets and hypotheses tests for θ_*. In correctly specified models, $(\Delta_t(\theta_*))_{t=0}^\infty$ is a martingale difference sequence, and thus result (4) can be obtained by invoking a martingale central limit theorem. In potentially misspecified models, $(\Delta_t(\theta_*))_{t=0}^\infty$ will not, in general, be a martingale difference sequence, so one should use a different approach; in some situations, a central limit theorem for β-mixing sequences can be used instead. The following corollary formalizes this discussion.

Corollary 1. *Suppose the assumptions of Theorem 4 hold.*

(a) If θ_* is such that, for any $t \geq 0$ and $T \geq 1$, $p_t^\nu(\cdot \mid X_{t-T}^{t-1}; \theta^*) = p_*^\nu(\cdot \mid X_{t-T}^{t-1})$, then

$$\sqrt{T}\{-H_T(\hat{\theta}_{\nu,T})\}^{1/2}(\hat{\theta}_{\nu,T} - \theta_*) \Rightarrow_{\bar{P}_*^\nu} \mathcal{N}(0, I).$$

(b) If there exists $\bar{L} > 0$ such that, for any $k, T > \bar{L}$, $p_k^\nu(X_k \mid X_{k-T}^{k-1}; \theta^*) = p_k^\nu(X_k \mid X_{k-\bar{L}}^{k-1}; \theta^*)$, $\liminf_{T\to\infty} e_{\min}(\Sigma_T(\theta_*)) > 0$,[14] and $E_{\bar{P}_*^\nu}[\|\Delta_1(\theta_*)\|^{4+4\delta}] < \infty$ for some $\delta > 0$, then

$$\sqrt{T}\hat{\Omega}_T(\hat{\theta}_{\nu,T})^{-1/2}(\hat{\theta}_{\nu,T} - \theta_*) \Rightarrow_{\bar{P}_*^\nu} \mathcal{N}(0, I),$$

where $\hat{\Omega}_T(\hat{\theta}_{\nu,T}) \equiv H_T^{-1}(\hat{\theta}_{\nu,T})J_T(\hat{\theta}_{\nu,T})H_T^{-1}(\hat{\theta}_{\nu,T})$.

[12] The additional condition that $\omega(\tau, L_T) = \sum_{j=1+\tau}^{L_T} c(j, L_T)c(j-\tau, L_T)$ for some constants $c(1, L_T), \ldots, c(L_T, L_T)$ guarantees that $J_T(\hat{\theta}_{\nu,T})$ is positive semidefinite. Weights obtained from the commonly used Bartlett, Parzen, and quadratic-spectral kernels satisfy this condition.

[13] The $\log\log T$ factor is used in order to avoid working with constants.

[14] For any real symmetric matrix A, $e_{\min}(A)$ denotes its minimum eigenvalue.

Proof. See Appendix A.3. □

Regarding the asymptotic normality result, both parts of the corollary rely on the structure of the process defined by the random variables $\nabla_\theta \log p_k^\nu(X_k \mid X_{-\infty}^{k-1}; \theta^*)$; this is a martingale difference sequence in part (a) and a geometrically β-mixing sequence in part (b). The latter result is a consequence of the β-mixing structure of $(X_t)_{t=-\infty}^{\infty}$ and of the fact that the conditional density at θ_* depends on a finite number of lags. Examples of models for which such properties hold true are mixture models and mixture autoregressive models (cf. Example 3).[15]

Regarding the covariance estimators used in parts (a) and (b) of the corollary, these rely on the corresponding parts of Theorem 5. The result in part (b) is established by exploiting the β-mixing structure of the score process $(\Delta_k(\theta_*))_{k=-\infty}^{\infty}$ and the finiteness of its $4+4\delta$ moments (under $\bar{P}_*^\nu$) for some $\delta > 0$. Lemma 9 in Appendix A.3 shows that $r_T = L_T T^{-1/2}$, and, since $\theta \mapsto \log p_\theta$ and $\theta \mapsto \log Q_\theta$ are smooth (see Assumptions 6 and 7), it follows that $\delta' \mapsto \ddot{\varpi}(\delta') = C\delta'$ for some finite constant C. Hence, the growth condition on $(L_T)_T$ translates into $L_T T^{-1/4} \log\log T = o(1)$, which is analogous to that in Newey and West [1987] (apart from the $\log\log T$ factor, which was introduced in Theorem 5 for convenience).

The next example verifies the assumptions in the context of the models considered in Example 4.

Example 5. In view of the results in Example 4, we only need to verify Assumptions 6–8. Part (i) of Assumption 6 is standard and is directly imposed, while part (ii) follows from the setup of the example. Assumption 7 follows by the continuity of the derivatives. Finally, Lemma 13 in the Supplemental Material SM.2 implies that Assumption 8 holds.

Thus, Theorem 4 holds for the class of models considered in Example 4. In particular, in the correctly specified case, Corollary 1(a) guarantees asymptotic normality of the studentized ML estimator of θ_*, thereby providing the basis for inference. These results are, to our knowledge, new in the context of Markov-switching autoregressive models with covariate-dependent transition probabilities. △

[15] At this level of generality, we cannot establish an asymptotic normality result in the general case where the model is misspecified and $\nabla_\theta \log p_k^\nu(X_k \mid X_{-\infty}^{k-1}; \theta^*)$ depends on the entire $X_{-\infty}^{k-1}$. The reason is that, although the process $X_{-\infty}^{\infty}$ is β-mixing, there is no guarantee that $\nabla_\theta \log p_k^\nu(X_k \mid X_{-\infty}^{k-1}; \theta^*)$ inherits these mixing properties, or that it is a martingale difference sequence.

5 Monte Carlo Simulations

The objective in this section is twofold. First, to assess the quality of approximations provided by our asymptotic results by examining the finite-sample properties of the ML estimator and related statistics in a correctly specified Markov-switching autoregressive model with covariate-dependent transition probabilities. Second, to explore the effects of a type of empirically relevant misspecification which involves the use of an incomplete approximation to the likelihood function that ignores potential contemporaneous correlation between the observation variable (Y_t) and the variable (Z_t) upon the lagged value of which the transition probabilities depend.

Monte Carlo experiments are based on artificial data $(X_t = (Y_t, Z_t))_t$ generated according to the equations

$$Y_t = \mu_0(1 - S_t) + \mu_1 S_t + \phi Y_{t-1} + [\sigma_0(1 - S_t) + \sigma_1 S_t]U_{1,t}, \tag{5}$$
$$Z_t = \mu_2 + \psi Z_{t-1} + \sigma_2 U_{2,t}, \tag{6}$$

for $t \in \mathbb{N}$, with $X_0 = (0.5, 0.4)$, $\mu_0 = -\mu_1 = 1$, $\phi = 0.9$, $\sigma_0 = \sigma_1 = 1$, $\mu_2 = 0.2$, $\psi = 0.8$, $\sigma_2 = 0.6$, and $(U_{1,t}, U_{2,t})_t \sim$ i.i.d. $\mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$ with $\rho \in \{0, 0.8\}$. The regimes $(S_t)_t$ are a Markov chain on $\{0, 1\}$, independent of $(U_{1,t}, U_{2,t})_t$, with transition probabilities

$$Q_\theta(z, s, s) \equiv \Pr(S_t = s \mid Z_{t-1} = z, S_{t-1} = s) = [1 + \exp(-\alpha_s - \beta_s z)]^{-1}, \tag{7}$$

for $s \in \{0, 1\}$, where $\alpha_0 = \alpha_1 = 2$ and $\beta_0 = -\beta_1 = -0.5$. The model defined by (5)–(7) is a prototypical Markov-switching autoregressive model with covariate-dependent transition probabilities (cf. Example 2). In each of 1000 independent Monte Carlo replications, $100 + T$ data points for $(X_t)_t$ are generated, with $T \in \{200, 800, 1600, 3200\}$, and the last T points are used to compute estimates of the parameters of interest. In order to conserve space, only a selection of the results are reported (the full set of results is available upon request).

5.1 Correct Specification

In the first set of experiments, we consider estimation of the parameters of the model in (5)–(7) using the likelihood function based on the conditional distribution of X_t given X_0^{t-1}. Table 1 reports the deviation of the mean of the finite-sample distributions of the ML estimators of the elements of $\vartheta = (\mu_0, \mu_1, \phi, \sigma_0, \sigma_1, \alpha_0, \beta_0, \alpha_1, \beta_1)$ from the corresponding true parameter values (bias) when $\rho = 0.8$. We also report the ratio of the sampling standard deviation of the estimators to the estimated standard errors (averaged across replications);

Table 1: Bias and Standard Deviation of ML Estimators ($\rho = 0.8$)

T	μ_0	μ_1	α_1	β_1	α_0	β_0	σ_0	σ_1	ϕ
	Bias								
200	-0.007	0.023	0.004	0.126	0.043	0.209	-0.011	-0.005	-0.005
800	-0.002	0.004	-0.002	0.023	0.002	0.026	-0.001	0.000	-0.001
1600	-0.001	0.002	-0.002	0.019	0.007	0.019	0.001	0.000	-0.001
3200	0.001	0.002	0.002	-0.005	0.002	0.011	0.001	0.000	0.000
	Ratio of sampling standard deviation to estimated standard error								
200	1.085	1.016	1.084	1.078	1.132	1.178	1.066	1.043	1.072
800	0.983	0.991	0.980	1.024	1.044	1.018	1.012	1.010	1.045
1600	1.021	1.002	0.964	0.955	1.033	1.042	1.000	0.999	0.997
3200	1.026	0.979	1.010	1.017	0.990	0.955	1.012	1.020	1.020

the latter are computed using the Hessian estimator (cf. Theorem 5(a)).[16] Although the estimators of β_0 and β_1 are somewhat biased when $T = 200$, bias is insignificant in the rest of the cases. Estimated standard errors are somewhat downwards biased in most cases but, unless T is small, the bias is not generally substantial and decreases as T increases.

We also examine conventional hypothesis tests for ϑ. Table 2 reports the rejection frequencies of: (i) a t-type test of $\mathcal{H}_0 : \vartheta_j = \vartheta_j^*$ versus $\mathcal{H}_1 : \vartheta_j \neq \vartheta_j^*$, where ϑ_j is the j-th element of ϑ and ϑ_j^* is its true value; (ii) a t-type test of $\mathcal{H}_0 : \vartheta_j = 0$ versus $\mathcal{H}_1 : \vartheta_j \neq 0$. These rejection frequencies are referred to as "size" and "power", respectively, and are computed using the 0.975 standard-normal quantile as critical value.[17] Tests tend to have Type I error probabilities which are generally close to the nominal 0.05 level, especially for $T > 200$. Tests are also powerful enough to reject the hypothesis of a zero parameter value, except in the case of β_0 and β_1 with $T = 200$. The distributions of studentized statistics associated with the elements of the ML estimator of ϑ (ratio of estimation error to corresponding estimated standard error) generally tend to have mean and variance (not shown) that do not differ substantially from zero and one, respectively, and Gaussianity is never rejected for $T > 200$.[18]

[16] Results for $\rho = 0$ are not reported since they are very similar to those for $\rho = 0.8$.

[17] Results should be interpreted with caution in the case of $\mathcal{H}_0 : \sigma_i = 0$, $i \in \{0, 1\}$, because the null value of σ_i is on the boundary of the maintained hypothesis. Our asymptotic theory does not allow for parameters that may lie on the boundary of the parameter space.

[18] Statistics associated with ϕ, σ_0 and σ_1 appear to fare somewhat worse than others when $T = 200$, a finding similar to that reported in Psaradakis and Sola [1998] for models with a time-invariant transition mechanism.

Table 2: Size and Power of t-Type Tests ($\rho = 0.8$)

T	μ_0	μ_1	α_1	β_1	α_0	β_0	σ_0	σ_1	ϕ
	Size								
200	0.063	0.076	0.069	0.069	0.081	0.043	0.084	0.088	0.108
800	0.053	0.063	0.061	0.051	0.074	0.049	0.065	0.065	0.073
1600	0.065	0.061	0.047	0.043	0.060	0.048	0.052	0.047	0.068
3200	0.049	0.058	0.053	0.058	0.054	0.042	0.051	0.061	0.053
	Power								
200	0.999	1.000	0.990	0.261	0.934	0.207	1.000	1.000	1.000
800	1.000	1.000	1.000	0.812	0.999	0.732	1.000	1.000	1.000
1600	1.000	1.000	1.000	0.988	1.000	0.966	1.000	1.000	1.000
3200	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

5.2 Misspecification

In the second set of experiments, we consider estimation of the parameter $\vartheta = (\mu_0, \mu_1, \phi, \sigma_0, \sigma_1, \alpha_0, \beta_0, \alpha_1, \beta_1)$ using the partial likelihood function based on the conditional distribution of Y_t given (Y_{t-1}, S_t). Inference in models like (5)–(7) is predominantly based on such a partial likelihood that ignores the equation for Z_t (see, e.g., Diebold et al. [1994], Filardo [1994]). Formally, the misspecified model may be viewed as defined by equations (5)–(7), with the additional assumption that $\rho = 0$. Under this assumption, estimation of ϑ is based on (5) alone, the (potentially incorrect) rational behind this approach being that, since the conditional distribution of Y_t given (Y_{t-1}, S_t) and the transition probabilities of $(S_t)_t$ depend only on Z_{t-1}, (5) may be analyzed, without loss of relevant information, independently of $(Z_t)_t$. Even though such an approach may be appealing because of its relative simplicity, it is far from obvious that it provides a valid way for conducting inference on ϑ; it is unclear, for example, what the limit point of the ML estimator based on the partial likelihood might be when $\rho \neq 0$. In earlier sections, we considered a theoretical framework that acknowledges this source of misspecification (among others) and provided tools for asymptotically valid inference. We now quantify the implications of this misspecification in finite samples. For brevity, we refer to the maximizer of the partial likelihood function associated with the conditional distribution of Y_t given (Y_{t-1}, S_t) as the 'partial ML' estimator to distinguish it from the 'joint ML' estimator based on the joint model for the conditional distribution of X_t given X_0^{t-1} (cf. Section 5.1).

Table 3 shows the estimated bias of the partial ML estimators of the elements of ϑ and the ratio of the sampling standard deviation of the estimators to the estimated standard errors (averaged across replications) when $\rho = 0.8$.

Table 3: Bias and Standard Deviation of Partial ML Estimators ($\rho = 0.8$)

T	μ_0	μ_1	α_1	β_1	α_0	β_0	σ_0	σ_1	ϕ
	Bias								
200	0.019	0.046	-0.172	0.898	0.397	1.363	-0.056	-0.039	-0.001
800	0.017	0.017	-0.156	0.485	0.216	0.838	-0.025	-0.024	0.002
1600	0.008	0.011	-0.150	0.446	0.211	0.824	-0.018	-0.021	0.003
3200	0.012	0.008	-0.149	0.444	0.196	0.773	-0.020	-0.019	0.003
	Ratio of sampling standard deviation to estimated standard error								
200	1.244	1.135	1.395	1.585	1.349	1.406	1.161	1.062	1.250
800	1.027	1.041	1.153	1.170	1.067	1.119	0.988	1.020	1.080
1600	0.991	1.014	1.070	1.075	1.072	1.161	1.006	0.959	1.054
3200	1.047	1.027	1.050	1.063	1.034	1.088	1.021	1.004	1.056

To reflect what is common practice in applied research, standard errors are computed using the Hessian estimator (which relies on the assumption of a correctly specified likelihood) instead of a "sandwich" estimator (which allows for mispecification). It is immediately apparent that the partial ML estimator of most of the parameters is considerably more biased than the joint ML estimator. The differences between the two estimators are more pronounced for parameters associated with the transition probabilities (α_0, β_0, α_1, β_1), the partial ML estimators of which are significantly biased even for $T = 3200$. This suggests that the bias of the partial ML estimator when $\rho \neq 0$ is not associated only with small samples, a finding that is consistent with our asymptotic results. Regarding the accuracy of estimated standard errors, the latter are downwards biased in most cases, the bias being somewhat larger than it is for joint ML estimators. However, unless T is small, this bias is not generally substantial and declines as T increases, despite the fact that standard errors are obtained from the Hessian.[19]

We note that hypothesis tests based on studentized statistics analogous to those considered in Section 5.1 (not shown) are unreliable when partial ML estimates are used. This is especially true in the case of parameters associated with the transition probabilities, the corresponding tests being either excessively conservative or excessively liberal. Although tests of this type are extensively used in applied work, they should be interpreted with caution since the statistics on which they are based have an asymptotically normal null distribution only when $\rho = 0$. We also note that using the "sandwich" estimator of The-

[19]Results for $\rho = 0$ (not shown) are not substantially different from those obtained from the joint ML procedure. This is not surprising since the joint and partial ML estimators are both consistent for the true parameter value when $\rho = 0$.

orem 5(b) instead of the estimator based on the observed information matrix is not without difficulty when $\rho \neq 0$. As Freedman [2006] points out, the use of such an estimator for inference is unlikely to produce results that are any less misleading under misspecification since the problem of bias/inconsistency of the ML estimator for the true parameter value remains. It is indeed clear from the results in Table 3 that the bias of the partial ML estimator presents a much more serious problem in our setting than the inaccuracy of conventionally computed standard errors.

6 Empirical Illustration

In this section, we present an empirical illustration based on a regime-switching model of a type that is commonly used in economics. Specifically, we investigate the potential contribution of the interest rate spread and the growth in tax revenues in predicting regime changes in U.S. real output growth. The model is a variant of the specification used in the simulations and is given by

$$Y_t = \mu_0(1 - S_t) + \mu_1 S_t + \sum_{i=1}^{h_1} \phi_i Y_{t-i} + \sigma_1 U_{1,t}, \tag{8}$$

$$Z_t = \mu_2 + \sum_{i=1}^{h_2} \psi_i Z_{t-i} + \sigma_2 U_{2,t}, \tag{9}$$

for some $h_1, h_2 \in \mathbb{N}$, with the hidden, two-state Markov chain $(S_t)_t$ being governed by the transition probabilities

$$Q_\theta(z, s, s) \equiv \Pr(S_t = s \mid Z_{t-1} = z, S_{t-1} = s) = [1 + \exp(-\alpha_s - \beta_s z)]^{-1}, \tag{10}$$

with $s \in \{0, 1\}$, and $(U_{1,t}, U_{2,t})_t$ postulated to be i.i.d. $\mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$ and independent of $(S_t)_t$. In (8)–(10), Y_t stands for the growth rate of real gross domestic product and Z_t is either the spread between the 10-year Treasury note rate and the 3-month Treasury bill rate or the growth rate of real government receipts of direct and indirect taxes.[20] The data are quarterly and span the period 1954:3–2009:2.[21]

[20] The model could be generalized to allow for Markov changes in all the parameters. However, since Z_t is thought of here as a potential leading indicator for business-cycle phases, it does not seem sensible to allow the parameters in both (8) and (9) to be subject to changes driven by $(S_t)_t$. Modeling regime changes in $(Y_t)_t$ and $(Z_t)_t$ as being driven by two independent Markov processes is more attractive, but we choose to abstract from this as it is not directly related to the main problem under study.

[21] Interest rate data are taken from the FRED database; output and tax data are taken from Auerbach and Gorodnichenko [2012]. The likelihood ratio test of Hansen [1992] rejects the hypothesis that $\mu_0 = \mu_1$ in (8).

Table 4: ML Estimates (Output Growth, Interest Rate Spread)

	Partial ML		Joint ML		
μ_0	0.0091 (0.0012)	μ_0	0.0092 (0.0012)		
μ_1	0.0001 (0.0013)	μ_1	0.0004 (0.0013)	μ_2	0.2956 (0.0984)
ϕ_1	0.1543 (0.0761)	ϕ_1	0.1449 (0.0763)	ψ_1	0.8098 (0.0012)
ϕ_2	0.0620 (0.0855)	ϕ_2	0.0595 (0.0852)	ψ_2	-0.1065 (0.1251)
ϕ_3	-0.0468 (0.0867)	ϕ_3	-0.0435 (0.0759)	ψ_3	0.3368 (0.1747)
ϕ_4	-0.0136 (0.0929)	ϕ_4	-0.0183 (0.0914)	ψ_4	-0.2495 (0.0859)
α_0	-1.3367 (3.8866)	α_0	-1.5210 (2.6935)	σ_2	0.7053 (0.1091)
β_0	8.8363 (9.4778)	β_0	9.1169 (6.4769)	ρ	-0.0849 (0.1198)
α_1	3.1927 (0.7646)	α_1	3.1218 (0.7688)		
β_1	-1.0779 (0.4304)	β_1	-1.0338 (0.4293)		
σ_1	0.0077 (0.0006)	σ_1	0.0078 (0.0006)		

Since the aim is not only to assess the predictive ability of the interest rate spread and tax revenues for regime changes in output growth but also to examine whether treating these variables as exogenous yields results which are different from those obtained from a joint model, we compute two sets of estimates: partial ML estimates based on (8) alone and joint ML estimates based on the system (8)–(9). We note that in econometric models of the business cycle such as (8)–(10), it is common to rely on partial ML estimation (see, e.g., Filardo [1994]). Parameter estimates are reported in Tables 4 and 5, with estimated standard errors given in parentheses; the latter are obtained from the "sandwich" estimator of Theorem 5(b).[22] On the basis of t-type tests based on joint ML estimates, at least one of the parameters (β_0, β_1) is significantly different from zero, indicating that the spread and tax revenues contain significant information about the probability of switching between the two regimes.

Regarding the implications of treating Z_t as exogenous, the differences between partial and joint ML estimates are substantial in the model with tax revenues (especially for autoregressive coefficients and the parameters associated with the transition probabilities) but much less so in the model with the interest rate spread. This is not entirely unexpected in view of the fact that the estimated value of the conditional correlation ρ is relatively large (0.6034) in the former model but much smaller (-0.0849), and insignificantly different from zero, in the latter. Such findings are in line with the analytical and sim-

[22]The weights $\omega(\cdot, L_T)$ are obtained from the Parzen kernel and L_T is determined using the automatic plug-in method of Andrews [1991].

ulation results presented in previous sections. The relatively large estimate of ρ in Table 5 also suggests that inference based on the partial ML estimator is potentially misleading because of the likely bias of the estimator.

Table 5: ML Estimates (Output Growth, Growth in Taxes)

	Partial ML		Joint ML		
μ_0	0.0071 (0.0014)	μ_0	0.0081 (0.0014)		
μ_1	-0.0119 (0.0034)	μ_1	-0.0100 (0.0075)	μ_2	0.5455 (0.2633)
ϕ_1	0.2076 (0.0679)	ϕ_1	0.0987 (0.0665)	ψ_1	0.1723 (0.1220)
ϕ_2	0.0709 (0.0953)	ϕ_2	0.0754 (0.0833)	σ_2	3.0458 (0.3082)
ϕ_3	-0.0530 (0.0754)	ϕ_3	-0.1168 (0.0610)	ρ	0.6034 (0.0673)
ϕ_4	-0.0291 (0.0885)	ϕ_4	-0.0345 (0.0721)		
α_0	3.4588 (0.6379)	α_0	3.8835 (1.3467)		
β_0	0.2754 (0.1237)	β_0	0.4061 (0.1379)		
α_1	0.3852 (0.8704)	α_1	-2.5349 (4.7906)		
β_1	0.2558 (0.1053)	β_1	0.0579 (0.1713)		
σ_1	0.0076 (0.0006)	σ_1	0.0084 (0.0009)		

A Appendix: Proofs

A.1 Consistency

In order to prove Theorem 1, we need two lemmas (the proofs of which are relegated to the Supplemental Material SM.2). The first lemma shows that the log-likelihood function $\ell_T^\nu(X_0^T,\cdot)$ can be approximated by the sample average of $(\log p^\nu(X_t \mid X_{-\infty}^{t-1},\cdot))_{t\in\mathbb{N}}$; this function is used to construct the function H^* that defines the pseudo-true parameter set. The result relies on "mixing" properties established in Theorem 2 (see Lemma 11 in the Supplemental Material SM.2).

Lemma 2. *Suppose Assumptions 1, 4(ii) and 5 hold. Then,*

$$\sup_{\theta\in\Theta}\left|T^{-1}\sum_{t=1}^{T}\left(\log p_t^\nu(X_t \mid X_0^{t-1},\theta) - \log p^\nu(X_t \mid X_{-\infty}^{t-1},\theta)\right)\right| = o_{\bar{P}_*^\nu}(1).$$

The second lemma essentially establishes a uniform law of large numbers for the sample average of $(\log p^\nu(X_t \mid X_{-\infty}^{t-1},\cdot))_{t\in\mathbb{N}}$.

Lemma 3. *Suppose Assumptions 1, 2, 3 and 4(i) hold. Then: (i) For any compact $K\subseteq\Theta$ and any $\epsilon>0$, there exists $T(\epsilon)\in\mathbb{N}$ such that*

$$\bar{P}_*^\nu\left(\sup_{\theta\in K}T^{-1}\sum_{t=1}^{T}\left(\log p^\nu(X_t \mid X_{-\infty}^{t-1},\theta) - E_{\bar{P}_*^\nu}\left[\log p^\nu(X_t \mid X_{-\infty}^{t-1},\theta)\right]\right) > \epsilon\right) \leq \epsilon, \tag{11}$$

for all $T\geq T(\epsilon)$.

(ii) For any $\theta_*\in\Theta_*$,

$$\left|T^{-1}\sum_{t=1}^{T}\left(\log p^\nu(X_t \mid X_{-\infty}^{t-1},\theta_0) - E_{\bar{P}_*^\nu}\left[\log p^\nu(X_t \mid X_{-\infty}^{t-1},\theta_0)\right]\right)\right| = o_{\bar{P}_*^\nu}(1).$$

Proof of Theorem 1. For simplicity, we set $\eta_T=0$ throughout the proof. Formally, we wish to establish that, for all $\epsilon>0$, there exists $T(\epsilon)\in\mathbb{N}$ such that

$$\bar{P}_*^\nu\left(d_\Theta(\hat{\theta}_{\nu,T},\Theta_*)\geq\epsilon\right) < \epsilon,$$

for all $t\geq T(\epsilon)$. For this, it suffices to show that there exists a $\theta_0\in\Theta_*$ such that, for any $\epsilon>0$, there exists a $T(\theta_0,\epsilon)$ such that

$$\bar{P}_*^\nu\left(\sup_{\theta\in\Theta\setminus\Theta_*^\epsilon}\ell_T^\nu(X_0^T,\theta)\geq\ell_T^\nu(X_0^T,\theta_0)\right) < \epsilon,$$

for all $T \geq T(\theta_0, \epsilon)$, where $\Theta_*^\epsilon = \{\theta \in \Theta \colon d_\Theta(\theta, \Theta_*) < \varepsilon\}$. Since, by Lemma 2, $\ell_T^\nu(X_0^T, \cdot)$ is well approximated by $\ell_T^\nu(X_{-\infty}^T, \cdot) \equiv T^{-1} \sum_{t=1}^T \log p^\nu(X_t \mid X_{-\infty}^{t-1}, \cdot)$, it suffices to work with the latter function.

Let $A_T(\delta) = \left\{X_{-\infty}^\infty \colon \sup_{\theta \in \Theta \setminus \Theta_*^\epsilon} T^{-1} \sum_{t=1}^T \left(\log p^\nu(X_t \mid X_{-\infty}^{t-1}, \theta) - E_{\bar{P}_*^\nu}\left[\log p^\nu(X_t \mid X_{-\infty}^{t-1}, \theta)\right]\right) \leq \delta\right\}$ and $B_T(\delta) = \left\{X_{-\infty}^\infty \colon \left|T^{-1} \sum_{t=1}^T \left(-\log p^\nu(X_t \mid X_{-\infty}^{t-1}, \theta_0) + E_{\bar{P}_*^\nu}\left[\log p^\nu(X_t \mid X_{-\infty}^{t-1}, \theta_0)\right]\right)\right| \leq \delta\right\}$, for any $\delta > 0$ and any $\theta_0 \in \Theta_*$. Observe that

$$
\begin{aligned}
&\bar{P}_*^\nu\left(\sup_{\theta \in \Theta \setminus \Theta_*^\epsilon} \ell_T^\nu(X_{-\infty}^T, \theta) \geq \ell_T^\nu(X_{-\infty}^T, \theta_0)\right) \\
\leq &\bar{P}_*^\nu\left(\sup_{\theta \in \Theta \setminus \Theta_*^\epsilon} \ell_T^\nu(X_{-\infty}^T, \theta) \geq \ell_T^\nu(X_{-\infty}^T, \theta_0) \cap A_T(\delta) \cap B_T(\delta)\right) + \bar{P}_*^\nu\left(A_T(\delta)^c\right) + \bar{P}_*^\nu\left(B_T(\delta)^c\right) \\
\leq &\bar{P}_*^\nu\left(\sup_{\theta \in \Theta \setminus \Theta_*^\epsilon} T^{-1} \sum_{t=1}^T E_{\bar{P}_*^\nu}\left[\log \frac{p^\nu(X_t \mid X_{-\infty}^{t-1}, \theta)}{p_*^\nu(X_t \mid X_{-\infty}^{t-1})}\right] \geq T^{-1} \sum_{t=1}^T E_{\bar{P}_*^\nu}\left[\log \frac{p^\nu(X_t \mid X_{-\infty}^{t-1}, \theta_0)}{p_*^\nu(X_t \mid X_{-\infty}^{t-1})}\right] - 2\delta\right) \\
&+ \bar{P}_*^\nu\left(A_T(\delta)^c\right) + \bar{P}_*^\nu\left(B_T(\delta)^c\right) \\
\leq &\bar{P}_*^\nu\left(\inf_{\theta \in \Theta \setminus \Theta_*^\epsilon} H^*(\theta) \leq H^*(\theta_0) + 2\delta\right) + \bar{P}_*^\nu\left(A_T(\delta)^c\right) + \bar{P}_*^\nu\left(B_T(\delta)^c\right),
\end{aligned}
$$

where the last line follows from the stationarity of $X_{-\infty}^\infty$ and the definition of H^*. By Assumption 3 and the fact that, for any $\theta \in \Theta \setminus \Theta_*^\epsilon$, $H^*(\theta) > H^*(\theta_0)$ (otherwise, θ would belong to Θ_*), it follows that $\inf_{\theta \in \Theta \setminus \Theta_*^\epsilon} H^*(\theta) - H^*(\theta_0) \equiv \Delta > 0$. Hence, choosing $\delta < 0.5\Delta$, the first term of the right-hand side (RHS) vanishes. By Assumption 3(i), $\Theta \setminus \Theta_*^\epsilon$ is compact; hence, by Lemma 3, there exists a T' (which may depend on ϵ and θ_0) such that $\bar{P}_*^\nu\left(A_T(\delta)^c\right) + \bar{P}_*^\nu\left(B_T(\delta)^c\right) \leq \epsilon$ for any $\delta \leq 0.5\epsilon$ and all $T \geq T'$, and thus the desired result follows. □

A.2 Mixing Results

Throughout, fix m and j as in the statement of Theorem 2. For any n, n' such that $-m \leq n, n' \leq j + 1$, we denote the Dobrushin coefficient of $\bar{P}_\theta^\nu(S_{n'} = \cdot \mid S_n = \cdot, X_{-m}^j)$ as

$$
\alpha_{\theta,n',n}(X_{-m}^j) \equiv \tfrac{1}{2} \max_{(a,b) \in \mathbb{S}^2} \left\| \bar{P}_\theta^\nu(S_{n'} = \cdot \mid S_n = a, X_{-m}^j) - \bar{P}_\theta^\nu(S_{n'} = \cdot \mid S_n = b, X_{-m}^j) \right\|_1. \tag{12}
$$

Since $\alpha_{\theta,j+1,-m}(X_{-m}^j) \leq \prod_{n=-m}^j \alpha_{\theta,n+1,n}(X_{-m}^j)$ (e.g., Dobrushin [1956], Sethuraman and Varadhan [2005]), to prove Theorem 2, it suffices to show the following.

Lemma 4. *For any $l \in \{-m, \ldots, j\}$ and any $\theta \in \Theta$, $\alpha_{\theta,l+1,l}(X_{-m}^j) \leq 1 - \underline{q}(X_l)$ a.s.-$\bar{P}_*^\nu$.*

Lemma 4 follows immediately from Lemmas 5 and 6 below. To state these lemmas, we construct the following processes that will be used for coupling. For any $i \in \{1, 2\}$ and any $\theta \in \Theta$, let $(X_{i,t}, \eta_{i,t}, \upsilon_{i,t})_{t=-m}^{\infty}$, with $(X_{i,t}, \eta_{i,t}, \upsilon_{i,t}) \in \mathbb{X} \times \mathbb{S} \times \{0, 1\}$, be defined as follows: $(X_{i,-m}, \eta_{i,-m}) \sim \nu$; given $(X_{i,t})_{-m}^{\infty}$, $(\upsilon_{i,t})_{t=-m}^{\infty}$ is i.i.d. with $\Pr(\upsilon_{i,t} = 1 \mid X_{i,-m}^{\infty}) = \Pr(\upsilon_{i,t} = 1 \mid X_{i,-m}^{t}) \equiv \underline{q}(X_{i,t})$; for each $t \geq -m$, $\eta_{i,t+1} \sim \varrho(X_{i,t}, \cdot)$ if $\upsilon_{i,t} = 1$, and $\eta_{i,t+1} \sim \frac{Q_\theta(X_{i,t},\eta_{i,t},\cdot) - \underline{q}(X_{i,t})\varrho(X_{i,t},\cdot)}{1-\underline{q}(X_{i,t})}$ if $\upsilon_{i,t} = 0$ (the last quotient expression is a valid transition kernel under condition (3) in Theorem 2); finally, $X_{i,t+1} \sim p_\theta(X_{i,t}, \eta_{i,t+1}, \cdot)$.

This construction implies that the transition kernel of $(\eta_{i,t})_t$ is given by

$$\begin{aligned}
\Pr_\theta(\eta_{i,t+1} = \cdot \mid \eta_{i,t}, X_{i,t}) =& \underline{q}(X_{i,t})\varrho(X_{i,t}, \cdot) + (1 - \underline{q}(X_{i,t}))\frac{Q_\theta(X_{i,t}, \eta_{i,t}, \cdot) - \underline{q}(X_{i,t})\varrho(X_{i,t}, \cdot)}{1 - \underline{q}(X_{i,t})} \\
=& Q_\theta(X_{i,t}, \eta_{i,t}, \cdot),
\end{aligned}$$

and since the transition for $X_{i,t+1}$ given $(X_{i,t}, \eta_{i,t+1})$ is governed by p_θ, the following result holds (its proof is relegated to the Supplemental Material SM.6).

Lemma 5. *For any $l \in \{-m, \ldots, j\}$ and any $\theta \in \Theta$,*

$$\bar{P}_\theta^\nu(S_{l+1} = \cdot \mid S_l = \cdot, X_{-m}^{j}) = \Pr_\theta(\eta_{i,l+1} = \cdot \mid \eta_{i,l} = \cdot, X_{-m}^{j}), \ \forall i \in \{1, 2\},$$

a.s.-$\bar{P}_*^\nu$.

Furthermore, since $\eta_{i,t+1}$ becomes independent of its own past whenever $\upsilon_{i,t} = 1$, the following result can be established (its proof is relegated to the Supplemental Material SM.6).

Lemma 6. *For any $l \in \{-m, \ldots, j\}$ and any $\theta \in \Theta$,*

$$\frac{1}{2}\max_{(a,b)\in\mathbb{S}^2} \left\| \Pr_\theta(\eta_{1,l+1} = \cdot \mid \eta_{1,l} = a, X_{-m}^{j}) - \Pr_\theta(\eta_{2,l+1} = \cdot \mid \eta_{2,l} = b, X_{-m}^{j}) \right\|_1 \leq 1 - \underline{q}(X_l),$$

a.s.-$\bar{P}_*^\nu$.

It is easy to see that Lemma 4 (and thus Theorem 2) follows from the last two lemmas.

A.3 Asymptotic Distribution Theory

The next two lemmas are used to prove Theorems 3 and 4 (their proofs are relegated to the Supplemental Material SM.7.1). In what follows, for vector/matrix-valued functions $X \mapsto f(X)$, $\|f\|_{L^r(P)}$ is short-hand notation for the $L^r(P)$-norm of $x \mapsto \|f(x)\|$, where $\|\cdot\|$ denotes the Euclidean/dual norm of f.

Lemma 7. *Suppose Assumptions 1, 2, 6, 7(i) and 8 hold. Then, there exists a stationary and ergodic (under $\bar{P}_*^\nu$) process $(\Delta_t(\theta_*))_{t=-\infty}^{\infty}$ in $L^2(\bar{P}_*^\nu)$ such that*

$$\lim_{T\to\infty}\left\|T^{-1/2}\sum_{t=0}^{T}\{\nabla_\theta \log p_t^\nu(\cdot|\cdot,\theta_*)-\Delta_t(\theta_*)\}\right\|_{L^2(\bar{P}_*^\nu)}=0.$$

Lemma 8. *Suppose Assumptions 1, 2, 6, 7 and 8 hold. Then, there exists a sequence of $\mathbb{R}^{q\times q}$-valued continuous functions $(\theta\mapsto\xi_t(\theta))_t$ such that $\xi_t(\theta)$ is negative definite for all t and*

$$\lim_{T\to\infty}\left\|\sup_{\theta\in B(\delta,\theta_*)}||T^{-1}\sum_{t=0}^{T}\{\nabla_\theta^2\log p_t^\nu(\cdot\mid\cdot,\theta)-\xi_t(\theta)\}||\right\|_{L^1(\bar{P}_*^\nu)}=0,$$

where $\delta>0$ is the same as in Assumption 7.

Proof of Theorem 3. Choose K compact such that, for any $v\in K$, $||v||\leq\delta$ for $\delta>0$ as in Lemma 8. For any $v\in K$, by Assumption 6,

$$\ell_T^\nu(X_0^T,\theta_*+v)-\ell_T^\nu(X_0^T,\theta_*)=v^\intercal\nabla_\theta\ell_T^\nu(X_0^T,\theta_*)+0.5v^\intercal\left(\int_0^1\nabla_\theta^2\ell_T^\nu(X_0^T,\theta_*+wv)dw\right)v.$$

By Lemmas 7 and 8, and the fact that $(\theta_*+wv)\in B(v,\theta_*)$,

$$\begin{aligned}\ell_T^\nu(X_0^T,\theta_*+v)-\ell_T^\nu(X_0^T,\theta_*)=&v^\intercal\left(T^{-1}\sum_{t=0}^{T}\Delta_t(\theta_*)+o_{\bar{P}_*^\nu}(T^{-1/2})\right)\\&+0.5v^\intercal\left(\int_0^1T^{-1}\sum_{t=0}^{T}\xi_t(\theta_*+wv)dw+o_{\bar{P}_*^\nu}(1)\right)v.\end{aligned}$$

Now, let $R_T(v)\equiv v^\intercal\left(T^{-1}\sum_{t=0}^T\int_0^1\{\xi_t(\theta_*+wv)-\xi_t(\theta_*)\}dw\right)v$. Observe that $||v||^{-2}|R_T(v)|\leq\int_0^1\left\|T^{-1}\sum_{t=0}^T\{\xi_t(\theta_*+wv)-\xi_t(\theta_*)\}\right\|dw$, so, for any $\delta>0$,

$$\begin{aligned}\bar{P}_*^\nu\left(\sup_{v\in B(\delta,0)}\frac{|R_T(v)|}{||v||^2}\geq\epsilon\right)\leq&\bar{P}_*^\nu\left(\sup_{v\in B(\delta,0)}\int_0^1\left\|T^{-1}\sum_{t=0}^{T}\{\xi_t(\theta_*+wv)-\xi_t(\theta_*)\}\right\|dw\geq\epsilon\right)\\\leq&\epsilon^{-1}E_{\bar{P}_*^\nu}\left[\sup_{v\in B(\delta,0)}\int_0^1\|\xi_1(\theta_*+wv)-\xi_1(\theta_*)\|\,dw\right],\end{aligned}$$

where the second line follows by the Markov inequality and stationarity. The desired result then follows by the continuity of ξ_1 (see Lemma 8) and the same arguments as in [Bickel et al., 1998, p. 1634]. □

Proof of Theorem 4. Henceforth, let $\bar{\Delta}_T \equiv T^{-1}\sum_{t=0}^{T}\Delta_t(\theta_*) + o_{\bar{P}_*^\nu}(T^{-1/2})$. By Theorem 1, $\hat{\theta}_{\nu,T} - \theta_*$ converges to zero with probability approaching one (w.p.a.1). Thus, $R_T(v_T) = o_{\bar{P}_*^\nu}(||v_T||^2)$ and, by Theorem 3,

$$\begin{aligned}\ell_T^\nu(X_0^T, \hat{\theta}_{\nu,T}) - \ell_T^\nu(X_0^T, \theta_*) =& (\hat{\theta}_{\nu,T} - \theta_*)^\intercal \bar{\Delta}_T \\ &+ 0.5(\hat{\theta}_{\nu,T} - \theta_*)^\intercal \left(T^{-1}\sum_{t=0}^{T}\xi_t(\theta_*) + o_{\bar{P}_*^\nu}(1)\right)(\hat{\theta}_{\nu,T} - \theta_*).\end{aligned}$$

Ergodicity of $X_{-\infty}^{\infty}$ (Lemma 1) implies ergodicity of $(\xi_t(\theta_*))_{t=-\infty}^{\infty}$; therefore, by Lemma 8 and Birkhoff's ergodic theorem,

$$\begin{aligned}\ell_T^\nu(X_0^T, \hat{\theta}_{\nu,T}) - \ell_T^\nu(X_0^T, \theta_*) =& (\hat{\theta}_{\nu,T} - \theta_*)^\intercal \bar{\Delta}_T \\ &+ 0.5(\hat{\theta}_{\nu,T} - \theta_*)^\intercal \left(E_{\bar{P}_*^\nu}[\xi_1(\theta_*)] + o_{\bar{P}_*^\nu}(1)\right)(\hat{\theta}_{\nu,T} - \theta_*),\end{aligned} \tag{13}$$

and $E_{\bar{P}_*^\nu}[\xi_1(\theta_*)]$ is non-singular. The rest of the proof proceeds in two steps.

Step 1. Let $r_T \equiv \min\{o_{\bar{P}_*^\nu}(1), o(T^{-1}) + E_{\bar{P}_*^\nu}\left[(\bar{\Delta}_T)^\intercal(\bar{\Delta}_T)\right]\}$. We first establish that $||\hat{\theta}_{\nu,T} - \theta_*|| = O_{\bar{P}_*^\nu}(\sqrt{r_T})$; by Theorem 1, the $o_{\bar{P}_*^\nu}(1)$ part of r_T has been established.

By (13) and the fact that $\hat{\theta}_{\nu,T}$ is an (approximate) maximizer of the likelihood function,

$$-2\eta_T \leq 2(\hat{\theta}_{\nu,T} - \theta_*)^\intercal \bar{\Delta}_T - (\hat{\theta}_{\nu,T} - \theta_*)^\intercal A(\theta_*)(\hat{\theta}_{\nu,T} - \theta_*),$$

with $A(\theta_*) \equiv \left(-E_{\bar{P}_*^\nu}[\xi_1(\theta_*)] + o_{\bar{P}_*^\nu}(1)\right)$. Simple algebra yields

$$-2\eta_T \leq -\left\|(\hat{\theta}_{\nu,T} - \theta_*)^\intercal A(\theta_*)^{1/2} - \bar{\Delta}_T^\intercal A(\theta_*)^{-1/2}\right\|^2 + \bar{\Delta}_T^\intercal A(\theta_*)^{-1}\bar{\Delta}_T.$$

Moreover, by simple algebra and the Markov inequality,

$$\left\|(\hat{\theta}_{\nu,T} - \theta_*)^\intercal A(\theta_*)^{1/2}\right\| = O_{\bar{P}_*^\nu}\left(\sqrt{\eta_T} + \sqrt{E_{\bar{P}_*^\nu}\left[(\bar{\Delta}_T)^\intercal(E_{\bar{P}_*^\nu}[\xi_1(\theta_*)])^{-1}(\bar{\Delta}_T)\right]}\right).$$

This expression, the fact that $E_{\bar{P}_*^\nu}[\xi_1(\theta_*)]$ is non-singular, and $\eta_T = o(T^{-1})$, imply the desired result.

Step 2. We now show that, for any $\epsilon > 0$,

$$\bar{P}_*^\nu\left(r_T^{-1/2}\left\|(\hat{\theta}_{\nu,T} - \theta_*) - (-E_{\bar{P}_*^\nu}[\xi_1(\theta_*)] + o_{\bar{P}_*^\nu}(1))^{-1}\bar{\Delta}_T\right\| \geq \epsilon\right) \to 0.$$

Since, by Step 1, $\left\|\hat{\theta}_{\nu,T} - \theta_*\right\| = O_{\bar{P}^\nu_*}(\sqrt{r_T})$, it suffices to show that

$$\bar{P}^\nu_*\left(\left\{r_T^{-1/2}\left\|(\hat{\theta}_{\nu,T} - \theta_*) - (-E_{\bar{P}^\nu_*}[\xi_1(\theta_*)] + o_{\bar{P}^\nu_*}(1))^{-1}\bar{\Delta}_T\right\| \geq \epsilon\right\} \cap \{\left\|\hat{\theta}_{\nu,T} - \theta_*\right\| \leq \sqrt{r_T}M\}\right) \to 0, \tag{14}$$

where $M > 0$. To this end, note that, by Theorem 3 and the fact that $T^{-1}\sum_{t=1}^T \xi_t(\theta_*) = E_{\bar{P}^\nu_*}[\xi_1(\theta_*)] + o_{\bar{P}^\nu_*}(1)$, it follows that

$$\ell^\nu_T(X_0^T, \theta_* + v) - \ell^\nu_T(X_0^T, \theta_*) = (\bar{\Delta}_T)^\intercal v - 0.5v^\intercal(-E_{\bar{P}^\nu_*}[\xi_1(\theta_*)] + o_{\bar{P}^\nu_*}(1))v + R_T(v),$$

for any $v \in K$. Letting $\Lambda_T(v) \equiv \ell^\nu_T(X_0^T, \theta_* + v) - \ell^\nu_T(X_0^T, \theta_*)$ and $Q_T(v) \equiv (\bar{\Delta}_T)^\intercal v - 0.5v^\intercal(-E_{\bar{P}^\nu_*}[\xi_1(\theta_*)] + o_{\bar{P}^\nu_*}(1))v$, we show that $\sup_{v\in\{v\colon ||v||\leq\sqrt{r_T}M\}} r_T^{-1}|\Lambda_T(v) - Q_T(v)| = o_{\bar{P}^\nu_*}(1)$. To do so, it suffices to prove that $\sup_{v\in\{v\colon ||v||\leq\sqrt{r_T}M\}}|R_T(v)| = o_{\bar{P}^\nu_*}(r_T)$. But this follows from Theorem 3 and the fact that $\sqrt{r_T} = o_{\bar{P}^\nu_*}(1)$. Since $(\hat{\theta}_{\nu,T} - \theta_*) \in \{v\colon ||v|| \leq \sqrt{r_T}M\}$ and maximizes $\Lambda_T(\cdot)$ (within a η_T margin), the previous result implies that

$$\begin{aligned}\hat{\theta}_{\nu,T} - \theta_* &= \arg\max_{v\in\{v\colon ||v||\leq T^{-1/2}M\}} Q_T(v) + o_{\bar{P}^\nu_*}(\sqrt{r_T}) + \eta_T\\ &= (-E_{\bar{P}^\nu_*}[\xi_1(\theta_*)] + o_{\bar{P}^\nu_*}(1))^{-1}\bar{\Delta}_T + o_{\bar{P}^\nu_*}(\sqrt{r_T}),\end{aligned}$$

and thus (14) follows. □

The proof of Corollary 1 uses the following lemma (whose proof is relegated to the Supplemental Material SM.7.2).

Lemma 9. *Suppose there exists* $\bar{L} \in \mathbb{N}$ *such that* $\nabla_\theta \log p^\nu_t(X_t|X_0^{t-1}, \theta) = \nabla_\theta \log p^\nu_t(X_t|X_{t-\bar{L}}^{t-1}, \theta)$ *for all* $t \geq 0$. *Then:*

(a) For all $t \geq 0$, $\Delta_t(\theta_*) \equiv \Delta_{t,-\infty}(\theta_*)(X_{-\infty}^t) = \nabla_\theta \log p^\nu_t(X_t|X_{t-\bar{L}}^{t-1}, \theta_*)$.

(b) If, in addition, there exists $\delta > 0$ *such that* $E_{\bar{P}^\nu_*}[||\nabla_\theta \log p^\nu_1(X_1|X_{1-\bar{L}}^0, \theta_*)||^{4+4\delta}] < \infty$, *then, for any* $L \in \mathbb{N}$,

$$\max_{j\in\{0,\ldots,L\}} ||T^{-1}\sum_{t=1}^T \Delta_{t+j,-\infty}(\theta_*)\Delta_{t,-\infty}(\theta_*)^\intercal - E_{\bar{P}^\nu_*}[\Delta_{j,-\infty}(\theta_*)\Delta_{0,-\infty}(\theta_*)^\intercal]|| = O_{\bar{P}^\nu_*}\left(\frac{L}{\sqrt{T}}\right).$$

Proof of Corollary 1. Throughout the proof, we use C to denote a universal constant that can take different values in different places. Also, for any $k, T \geq 0$ and any $\theta \in \Theta$, we write $\Delta_{k,k-T}(\theta) \equiv \Delta_{k,k-T}(\theta)(X_{k-T}^k) = \nabla_\theta \log p^\nu_k(X_k|X_{k-T}^{k-1}; \theta)$ (for the last equality, see Lemma 17 in the Supplemental Material SM.7.1) and $\Delta_k(\theta) \equiv \Delta_{k,-\infty}(\theta)$.

We first show that, under the conditions of part (a), $(\Delta_t(\theta_*))_t$ is a martingale difference sequence (MDS) with respect to the natural filtration of $(X_t)_t$. To establish this, observe that

$$E_{\bar{P}_*^\nu}\left[\Delta_{k,k-T}(\theta_*) \mid X_{-\infty}^{k-1}\right] = E_{\bar{P}_*^\nu}\left[\Delta_{k,k-T}(\theta_*) \mid X_{k-T}^{k-1}\right] = \int \frac{\nabla_\theta p_k^\nu(x_k \mid X_{k-T}^{k-1}; \theta^*)}{p_k^\nu(x_k \mid X_{k-T}^{k-1}; \theta^*)} p_*^\nu(x_k \mid X_{k-T}^{k-1}) dx_k.$$

By assumption, $p_k^\nu(x_k \mid X_{k-T}^{k-1}; \theta^*) = p_*^\nu(x_k \mid X_{k-T}^{k-1})$. Moreover, by using the representation of $\Delta_{k,k-T}(\cdot)$ in the Supplemental Material SM.7.1 and Assumption 7, it can be shown that $\nabla_\theta \log p_k^\nu(\cdot \mid X_{k-T}^{k-1}; \theta)$ is uniformly bounded for all $\theta \in B(\delta, \theta_*)$ (where $\delta > 0$ is as in Assumption 7). This result allow us to use standard "Fisher information equality" calculations and thus deduce that $E_{\bar{P}_*^\nu}\left[\Delta_{k,k-T}(\theta_*) \mid X_{-\infty}^{k-1}\right] = 0$.

Next, we show that, for any $k \geq 0$, $A_k \equiv E_{\bar{P}_*^\nu}\left[\Delta_{k,-\infty}(\theta_*) \mid X_{-\infty}^{k-1}\right] = 0$. To do so, observe that, for any $k, T \geq 0$,

$$\begin{aligned}
\|A_k\|_{L^2(\bar{P}_*^\nu)}^2 =& E_{\bar{P}_*^\nu}\left[\left(E_{\bar{P}_*^\nu}\left[\Delta_{k,-\infty}(\theta_*) \mid X_{-\infty}^{k-1}\right] - E_{\bar{P}_*^\nu}\left[\Delta_{k,k-T}(\theta_*) \mid X_{-\infty}^{k-1}\right]\right)^2\right] \\
\leq & \|\Delta_{0,-\infty}(\theta_*) - \Delta_{0,-T}(\theta_*)\|_{L^2(\bar{P}_*^\nu)}^2 \\
\leq & C\left(\max\{\sum_{j=[-T/2]}^{-1} \varrho(j,-T), \sum_{j=-T}^{[-T/2]-1} \varrho(-1,j)\}\right),
\end{aligned}$$

where the second line follows from the Jensen inequality and stationarity, and the third line follows from Lemma 18(i). Now, recall that, for any $j \geq k$, $\varrho(j,k) \equiv \left(E_{\bar{P}_*^\nu}\left[\prod_{i=k}^{j}(1-\underline{q}(X_i))^{\frac{2a}{1-a}}\right]\right)^{\frac{1-a}{2a}} = \left(E_{\bar{P}_*^\nu}\left[\prod_{i=0}^{j-k}(1-\underline{q}(X_i))^{\frac{2a}{1-a}}\right]\right)^{\frac{1-a}{2a}} = \varrho(j-k,0)$. Moreover, by Assumption 8, $(\varrho(j,0))_j$ is p-summable with $p < 2/3$, and thus $\lim_{j\to\infty} \varrho(j,0)^p j = 0$ (if not, then $\varrho(j,0) > c/j^{1/p}$ for some $c > 0$ and all j above certain point, and this violates the assumption). Hence, for sufficiently large T,

$$\begin{aligned}
\|A_k\|_{L^2(\bar{P}_*^\nu)}^2 \leq & C\left(\max\{\sum_{j=[-T/2]}^{-1} \varrho(j+T,0), \sum_{j=-T}^{[-T/2]-1} \varrho(-1-j,0)\}\right) \\
\leq & C \sum_{l=[T/2]}^{T-1} l^{-(1/p)}.
\end{aligned}$$

As $1/p > 1$, it follows that, as T diverges, the RHS converges to zero, and thus $\|A_k\|_{L^2(\bar{P}_*^\nu)}^2 = 0$. Since $\Delta_k(\theta_*) = \Delta_{k,-\infty}(\theta_*)$ for any k, the desired result follows.

By Theorem 4 and the central limit theorem for MDS, we have, therefore,

$$\sqrt{T}(\hat{\theta}_{\nu,T} - \theta_*) \Rightarrow_{\bar{P}^\nu_*} \mathcal{N}(0, (E_{\bar{P}^\nu_*}[\xi_1(\theta_*)])^{-1}\Sigma(\theta_*)(E_{\bar{P}^\nu_*}[\xi_1(\theta_*)])^{-1}),$$

where $\Sigma(\theta_*) \equiv \lim_{T\to\infty} \Sigma_T(\theta_*)$, with $\Sigma(\theta_*) = E_{\bar{P}^\nu_*}[(\Delta_1(\theta_*))(\Delta_1(\theta_*))^\intercal]$ since $(\Delta_t(\theta_*))_t$ is a stationary MDS. Moreover, it can be shown that $E_{\bar{P}^\nu_*}[\xi_1(\theta_*)] = -E_{\bar{P}^\nu_*}[(\Delta_1(\theta_*))(\Delta_1(\theta_*))^\intercal]$. This follows by standard "Fisher information equality" calculations and derivations analogous to those above (so they are omitted). Hence, the proof of part (a) is complete once we show that $-H_T^{-1}(\hat{\theta}_{\nu,T})$ converges in $\bar{P}^\nu_*$-probability to $(E_{\bar{P}^\nu_*}[\xi_1(\theta_*)])^{-1}\Sigma(\theta_*)(E_{\bar{P}^\nu_*}[\xi_1(\theta_*)])^{-1}$. But this follows from Theorem 5(a).

Under the conditions of part (b), and in view of Lemma 9(a), $\Delta_k(\theta_*) = \nabla_\theta \log p^\nu_k(X_k|X^{k-1}_{k-\bar{L}};\theta_*)$ for all $k \geq 0$. This result and the fact that $(X_k)^\infty_{k=-\infty}$ is β-mixing with mixing coefficients $\beta_n = O(\gamma^n)$ (see Lemma 1) imply that $(\Delta_k(\theta_*))^\infty_{k=-\infty}$ is also β-mixing with mixing coefficients of the same order. Hence, to establish that $T^{-1/2}\sum_{t=0}^T \Delta_t(\theta_*) \Rightarrow_{\bar{P}^\nu_*} \mathcal{N}(0,\Sigma(\theta_*))$, it is enough to verify that, for some $\delta > 0$, $E_{\bar{P}^\nu_*}[||\Delta_1(\theta_*)||^{2+\delta}] < \infty$ and $\sum_{n=1}^\infty \alpha_n^{\delta/(2+\delta)} < \infty$, where $(\alpha_n)_n$ are the α-mixing coefficients of $(\Delta_k(\theta_*))^\infty_{k=-\infty}$ (see, e.g., Doukhan et al. [1994]). But, the summability condition on the α-mixing coefficients is satisfied, because the β-mixing coefficients of the process decay at rate $O(\gamma^n)$, and the moment condition is directly assumed.

The proof of part (b) is completed by showing that $\hat{\Omega}_T(\hat{\theta}_{\nu,T})$ converges in $\bar{P}^\nu_*$-probability to $(E_{\bar{P}^\nu_*}[\xi_1(\theta_*)])^{-1}\Sigma(\theta_*)(E_{\bar{P}^\nu_*}[\xi_1(\theta_*)])^{-1}$, which is non-singular by assumption. We do so by invoking Theorem 5(b) and verifying its conditions. As stated above, $\alpha_n = O(\gamma^n)$. Moreover, by Corollary 6.17 in White [2001], there exists $C < \infty$ such that, for any $l > 0$,

$$||E_{\bar{P}^\nu_*}[\Delta_l(\theta_*)\Delta_0(\theta_*)^\intercal]|| \leq C(\alpha_l)^{\frac{2}{2+2\delta}}\sqrt{E_{\bar{P}^\nu_*}[||\Delta_l(\theta_*)||^2]}(E_{\bar{P}^\nu_*}[||\Delta_l(\theta_*)||^{2+2\delta}])^{\frac{1}{2+2\delta}}.$$

Under our assumptions, the RHS equals $C(\alpha_l)^{\frac{2}{2+2\delta}}$ for some $C < \infty$. Hence, one can set $\bar{v}(l) \equiv C(\gamma^{\frac{2}{2+2\delta}})^l$. Since $\gamma < 1$, the function $l \mapsto \bar{v}(l)$ is integrable. Also, by Lemma 9(b), it follows that $r_T = L_T T^{-1/2}$.

Next, we show that $\ddot{\varpi}(\delta') = C\delta'$ for some $C < \infty$ and any $\delta' \leq \delta$, where $\delta > 0$ is as in Assumption 7. To do so, we note that for any θ and any $t \in \mathbb{N}$,

$$\begin{aligned}||\Delta_t(\theta)\Delta_0(\theta)^\intercal - \Delta_t(\theta_*)\Delta_0(\theta_*)^\intercal|| \leq &||\Delta_0(\theta)|| \times ||\Delta_t(\theta) - \Delta_t(\theta_*)|| \\ &+ ||\Delta_t(\theta_*)|| \times ||\Delta_0(\theta) - \Delta_0(\theta_*)||.\end{aligned}$$

By the calculations in the proof of Lemma 18 in the Supplemental Material SM.7, $||\Delta_t(\theta)||$ is a linear combination of terms involving $\Gamma(\cdot,\theta)$ and $\Lambda(\cdot,\theta)$.

Since $\Delta_t(\theta)$ only depends on $X_{t-\bar{L}}^t$ for some finite $\bar{L}$, there are only finitely many terms in this linear combination. Moreover, by Assumption 7, there exists $C < \infty$ such that $E_{\bar{P}_*^\nu}[\sup_{\theta \in B(\delta,\theta_*)} ||\Delta_t(\theta)||^2] \leq C$, where $\delta > 0$ is as in Assumption 7. This bound is also uniform over t. Hence, on account of this result and stationarity, there exists $C < \infty$ such that

$$E_{\bar{P}_*^\nu}\left[\sup_{||\theta-\theta_*|| \leq \delta'} \|\Delta_t(\theta)\Delta_0(\theta)^\intercal - \Delta_t(\theta_*)\Delta_0(\theta_*)^\intercal\|\right] \leq C\left(E_{\bar{P}_*^\nu}\left[\sup_{||\theta-\theta_*|| \leq \delta'} \|\Delta_0(\theta) - \Delta_0(\theta_*)\|^2\right]\right)^{1/2},$$

for any $\delta' \leq \delta$ and any $t \in \mathbb{N}$. It remains to show that the RHS is of the form $C\delta'$ for some finite constant C. By the mean value theorem, $\|\Delta_0(\theta) - \Delta_0(\theta_*)\| \leq \sup_{\theta \in B(\delta,\theta_*)} ||\nabla_\theta \Delta_0(\theta)|| \times ||\theta - \theta_*||$, so it suffices to show that $E_{\bar{P}_*^\nu}[\sup_{\theta \in B(\delta,\theta_*)} ||\nabla_\theta \Delta_0(\theta)||^2] \leq C$ for some $C < \infty$. Since $\Delta_0(\theta) = \nabla_\theta \log p_0^\nu(X_0 \mid X_{-\bar{L}}^{-1})$, this condition is equivalent to $E_{\bar{P}_*^\nu}[\sup_{\theta \in B(\delta,\theta_*)} ||\nabla_\theta^2 \log p_0^\nu(X_0 \mid X_{-\bar{L}}^{-1}, \theta)||^2] \leq C$. By the calculations in [Bickel et al., 1998, pp. 1627–1628], $\nabla_\theta^2 \log p_0^\nu(X_0 \mid X_{-\bar{L}}^0, \theta)$ is a linear combination of finitely many terms involving $\Gamma(\cdot, \theta)$, $\Lambda(\cdot, \theta)$, and their derivatives. By Assumption 7 and calculations analogous to those in the proof of Lemma 18, it then follows that $E_{\bar{P}_*^\nu}[\sup_{\theta \in B(\delta,\theta_*)} ||\nabla_\theta^2 \log p_0^\nu(X_0 \mid X_{-\bar{L}}^{-1}, \theta)||^2] \leq C$. Thus, there exists $C < \infty$ such that $\ddot{\varpi}(\delta') \leq C\delta'$ for any $\delta' \leq \delta$. Therefore, all the conditions of Theorem 5(b) are satisfied.

□

References

P. Ailliot and F. Pène. Consistency of the maximum likelihood estimate for non-homogeneous Markov-switching models. *ESAIM: Probability and Statistics*, 19:268–292, 2015.

J. H. Albert and S. Chib. Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *Journal of Business and Economic Statistics*, 11:1–15, 1993.

D. W. K. Andrews. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59:817–858, 1991.

A. Ang and G. Bekaert. International asset allocation with regime shifts. *Review of Financial Studies*, 15:1137–1187, 2002a.

A. Ang and G. Bekaert. Regime switches in interest rates. *Journal of Business and Economic Statistics*, 20:163–182, 2002b.

A. Ang, G. Bekaert, and M. Wei. The term structure of real rates and expected inflation. *Journal of Finance*, 63:797–848, 2008.

A. J. Auerbach and Y. Gorodnichenko. Measuring the output responses to fiscal policy. *American Economic Journal: Economic Policy*, 4:1–27, 2012.

G. Bekaert, R. J. Hodrick, and D. A. Marshall. Peso problem explanations for term structure anomalies. *Journal of Monetary Economics*, 48:241–270, 2001.

P. J. Bickel and Y. Ritov. Inference in hidden Markov models I: Local asymptotic normality in the stationary case. *Bernoulli*, 2:199–228, 1996.

P. J. Bickel, Y. Ritov, and T. Rydén. Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *Annals of Statistics*, 26:1614–1635, 1998.

N. P. B. Bollen, S. F. Gray, and R. E. Whaley. Regime switching in foreign exchange rates: Evidence from currency option prices. *Journal of Econometrics*, 94:239–276, 2008.

F. X. Diebold, J.-H. Lee, and G. C. Weinbach. Regime switching with time-varying transition probabilities. In C. P. Hargreaves, editor, *Nonstationary Time Series Analysis and Cointegration*, pages 283–302. Oxford University Press, Oxford, 1994.

R. L. Dobrushin. Central limit theorem for nonstationary Markov chains. II. *Theory of Probability and Its Applications*, 1:329–383, 1956.

R. Douc and C. Matias. Asymptotics of the maximum likelihood estimator for general hidden Markov models. *Bernoulli*, 7:381–420, 2001.

R. Douc and E. Moulines. Asymptotic properties of the maximum likelihood estimation in misspecified hidden Markov models. *Annals of Statistics*, 40:2697–2732, 2012.

R. Douc, G. Fort, E. Moulines, and P. Soulier. Practical drift conditions for subgeometric rates of convergence. *Annals of Applied Probability*, 14:1353–1377, 2004a.

R. Douc, É. Moulines, and T. Rydén. Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *Annals of Statistics*, 32: 2254–2304, 2004b.

R. Douc, E. Moulines, J. Olsson, and R. van Handel. Consistency of the maximum likelihood estimator for general hidden Markov models. *Annals of Statistics*, 39: 474–513, 2011.

P. Doukhan, P. Massart, and E. Rio. The functional central limit theorem for strongly mixing processes. *Annales de l'Institut Henri Poicaré, Section B*, 30:63–82, 1994.

M. J. Dueker, M. Sola, and F. Spagnolo. Contemporaneous threshold autoregressive models: estimation, testing and forecasting. *Journal of Econometrics*, 141:517–547, 2007.

M. J. Dueker, Z. Psaradakis, M. Sola, and F. Spagnolo. Multivariate contemporaneous threshold autoregressive models. *Journal of Econometrics*, 160:311–325, 2011.

C. Engel and C. S. Hakkio. The distribution of the exchange rate in the EMS. *International Journal of Finance and Economics*, 1:55–67, 1996.

C. Engel and J. D. Hamilton. Long swings in the Dollar: Are they in the data and do markets know it? *American Economic Review*, 80:689–713, 1990.

A. J. Filardo. Business-cycle phases and their transitional dynamics. *Journal of Business and Economic Statistics*, 12:299–308, 1994.

A. J. Filardo and S. F. Gordon. Business cycle durations. *Journal of Econometrics*, 85:99–123, 1998.

C. Francq and M. Roussignol. Ergodicity of autoregressive processes with Markov-switching and consistency of the maximum-likelihood estimator. *Statistics*, 32:151–173, 1998.

D. A. Freedman. On the so-called "Huber sandwich estimator" and "robust standard errors". *American Statistician*, 60:299–302, 2006.

M. D. Gadea Rivas and G. Perez-Quiros. The failure to predict the Great Recession–A view through the role of credit. *Journal of the European Economic Association*, 13: 534–559, 2015.

S. F. Gray. Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics*, 42:27–62, 1996.

J. D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57:357–384, 1989.

B. E. Hansen. The likelihood ratio test under nonstandard conditions: Testing the Markov switching model of GNP. *Journal of Applied Econometrics*, 7:S61–S82 (Erratum: 11, 195–198), 1992.

I. A. Ibragimov and R. Z. Has'minskii. *Statistical Estimation: Asymptotic Theory.* Springer-Verlag, New York, 1981.

J. L. Jensen and N. V. Petersen. Asymptotic normality of the maximum likelihood estimator in state space models. *Annals of Statistics*, 27:514–535, 1999.

L. Kalliovirta, M. Meitz, and P. Saikkonen. A Gaussian mixture autoregressive model for univariate time series. *Journal of Time Series Analysis*, 36:247–266, 2015.

H. Kasahara and K. Shimotsu. Asymptotic properties of the maximum likelihood estimator in regime switching econometric models. *Journal of Econometrics*, 208: 442–467, 2019.

V. Krishnamurthy and T. Rydén. Consistent estimation of linear and non-linear autoregressive models with Markov regime. *Journal of Time Series Analysis*, 19:291–307, 1998.

L. Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, New York, 1986.

X. Liu and Y. Shao. Asymptotics for likelihood ratio tests under loss of identifiability. *Annals of Statistics*, 31:807–832, 2003.

M. S. Martinez Peria. A regime-switching aproach to the study of speculative attacks: A focus on the EMS crisis. *Empirical Economics*, 27:299–334, 2002.

R. E. McCulloch and R. S. Tsay. Statistical analysis of economic time series via Markov switching models. *Journal of Time Series Analysis*, 15:523–539, 1994.

L. Mevel and L. Finesso. Asymptotical statistics of misspecified hidden Markov models. *IEEE Transactions on Automatic Control*, 49:1123–1132, 2004.

S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, London, 1993.

K. Mouratidis. Evaluating currency crises: A Bayesian Markov switching approach. *Journal of Macroeconomics*, 30:1688–1711, 2008.

W. K. Newey and K. D. West. A simple, positive-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55:703–708, 1987.

D. Pouzo, Z. Psaradakis, and M. Sola. Maximum likelihood estimation in Markov regime-switching models with covariate-dependent transition probabilities. arXiv: 1612.04932v3 [math.ST], 2021.

Z. Psaradakis and M. Sola. Finite-sample properties of the maximum likelihood estimator in autoregressive models with Markov switching. *Journal of Econometrics*, 86:369–386, 1998.

Z. Psaradakis and M. Sola. Markov-switching models with state-dependent time-varying transition probabilities. *Econometrics and Statistics*, 2021. in press (https://doi.org/10.1016/j.ecosta.2021.04.007).

M. O. Ravn and M. Sola. Business cycle dynamics: Predicting transitions with macrovariables. In P. Rothman, editor, *Nonlinear Time Series Analysis of Economic and Financial Data*, pages 231–265. Kluwer Academic Publishers, Dordrecht, 1999.

F. J. Ruge-Murcia. Credibility and changes in policy regime. *Journal of Political Economy*, 103:176–208, 1995.

T. Rydén, T. Teräsvirta, and S. Åsbrink. Stylized facts of daily returns series and the hidden Markov model. *Journal of Applied Econometrics*, 13:217–244, 1998.

S. Sethuraman and S. R. S. Varadhan. A martingale proof of Dobrushin's theorem for non-homogeneous Markov chains. *Electronic Journal of Probability*, 10:1221–1235, 2005.

P. W. Simpson, D. R. Osborn, and M. Sensier. Modelling business cycle movements in the uk economy. *Economica*, 68:243–267, 2001.

K. J. Tadjuidje, H. Ombao, and R. A. Davis. Autoregressive processes with data-driven regime switching. *Journal of Time Series Analysis*, 30:505–533, 2009.

C.-Y. Tan, Y.-B. Koh, K.-H. Ng, and K.-H. Ng. Dynamic volatility modelling of Bitcoin using time-varying transition probability Markov-switching GARCH model. *North American Journal of Economics and Finance*, 56:Article 101377, 2021.

H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50: 1–25, 1982.

H. White. *Estimation, Inference and Specification Analysis.* Cambridge University Press, Cambridge, 1994.

H. White. *Asymptotic Theory for Econometricians.* Revised Edition, Academic Press, London, 2001.

R. F. Whitelaw. Stock market risk and return: An equilibrium approach. *Review of Financial Studies*, 13:521–547, 2000.