Statistics and Probability Letters 183 (2022) 109347

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

A proof of consistency of the MLE for nonlinear Markov-switching AR processes

Lisandro Fermín [a], José Marcano [b], Luis-Angel Rodríguez [a,b,*]

[a] CIMFAV, Facultad de Ingeniería, Universidad de Valparaíso, Chile
[b] Dpto. de Matemáticas, FACYT, Universidad de Carabobo, Venezuela

ARTICLE INFO

Article history:
Received 25 July 2021
Received in revised form 5 December 2021
Accepted 21 December 2021
Available online 29 December 2021

MSC:
primary 60G17
secondary 62G07

Keywords:
Nonlinear autoregressive process
Markov switching
Asymptotic normality
Consistency
Hidden Markov chain

ABSTRACT

We propose a new approach to demonstrate the consistency of the maximum likelihood estimator for nonlinear Markov-switching AR processes (abbreviated MS-NAR). We obtain a uniform exponential memory loss property for the prediction filter by approximating it by a filter with finite memory. From the α-mixing property for the MS-NAR process we obtain an ergodic theorem. Finally, we show that in the linear and Gaussian case our assumptions are fully satisfied.

1. Introduction

Markov-switching autoregressive processes can be considered a combination of hidden Markov models (HMM) and threshold regression models. They were introduced in an econometric context by Goldfeld and Quandt (1973) and have become quite popular in the literature since Hamilton (1989) used them in the analysis of the growth rate of the US GNP series in the presence of two regimes: one of contraction and the other of expansion. This family of models describes the evolution of a time series subject to discrete changes, whose transition is controlled by an HMM.

We consider a real nonlinear Markov-switching AR process $\{Y_n\}_{n\geq 0}$, defined for integers $n \geq 1$ by

$$Y_n = r(Y_{n-1}, \theta_{X_n}) + e_n, \tag{1}$$

where $\{e_n\}_{n\geq 1}$ is an i.i.d. sequence of random variables, and $\{X_n\}_{n\geq 1}$ is a homogeneous Markov chain with state space $\{1, \ldots, m\}$. Let $\mathcal{F} = \{r(\cdot, \theta) : \theta \in \Theta\}$ be a family of real valued functions defined on $\mathbb{R}^{m+1}$, indexed by a parameter $\theta = (\theta_1, \ldots, \theta_m) \in \Theta$, for Θ a compact set of $\mathbb{R}^m$. We denote by A the probability transition matrix of the Markov chain $\{X_n\}_{n\geq 1}$, i.e. $A = [a_{ij}]$, with $a_{ij} = \mathbb{P}(X_1 = j | X_0 = i)$. The parameter space is the set

$$\Psi = \left\{ \psi = (\theta, A) : \theta \in \Theta, A = [a_{ij}], a_{ij} \in [0, 1] \text{ and } \sum_{j=1}^{m} a_{ij} = 1 \right\}.$$

* Corresponding author at: Dpto. de Matemáticas, FACYT, Universidad de Carabobo, Venezuela.
E-mail address: larodri@uc.edu.ve (L.-A. Rodríguez).

https://doi.org/10.1016/j.spl.2021.109347
0167-7152/

We fix a distinguished element $\psi^* \in \Psi$. The maximum likelihood estimator (MLE) is one of the most commonly used and popular statistical methods. The MLE should in general be consistent in the sense that it converges to the true parameter value ψ^* as the number of observations tends to infinity. The consistency of the maximum likelihood estimator for the parameter ψ in the MS-NAR model is given in Krishnamurthy and Rydén (1998), while the consistency and asymptotic normality are proved in a more general context in the work of Douc et al. (2004).

Let us introduce some notation:

- $V_{1:n}$ stands for the random vector $(V_1, \ldots, V_n)$, and by $v_{1:n} = (v_1, \ldots, v_n)$ we mean a realization of the respective random vector.
- $p(V_{1:n} = v_{1:n})$ denotes the density distribution of the random vector $V_{1:n}$ evaluated at $v_{1:n}$.

We have p_ψ as a generic symbol for densities and distributions parameterized by ψ, and we define the conditional log-likelihood as $l_n(\psi) = \log p_\psi(Y_{1:n}|Y_0)$. Let us next, recall the basic approach for proving consistency of the MLE. The first step of the proof aims to establish that for any $\psi \in \Psi$, there is a constant $H(\psi^*, \psi)$ such that

$$\lim_{n\to\infty} \frac{1}{n} \log p_\psi(Y_{1:n}|Y_0) = \lim_{n\to\infty} \frac{1}{n} \mathbb{E}(\log p_\psi(Y_{1:n}|Y_0)) = H(\psi^*, \psi)$$

We define $K(\psi^*, \psi) = H(\psi^*, \psi^*) - H(\psi^*, \psi)$ as the relative entropy rate between the observation laws of the parameters $\psi*$ and ψ, respectively. The second step of the proof aims to establish identifiability, that is, that $K(\psi^*, \psi)$ is minimized only at those parameters ψ that are equivalent to ψ^*, hence, the distributions corresponding to these parameters are equivalent. Finally, the third step of the proof aims to prove that the maximizer of the conditional likelihood $\psi \to p_\psi(Y_{1:n}|Y_0)$ converges $\mathbb{P}_{\psi^*} - a.s.$ to the maximizer of $H(\psi^*, \psi)$. Let us note that one could write the likelihood as

$$\lim_{n\to\infty} \frac{1}{n} \log p_\psi(Y_{1:n}|Y_0) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \log p_\psi(Y_k|Y_{0:k-1}).$$

If the limit of $p_\psi(Y_k|Y_{-n:k})$ as $n \to \infty$ exists $\mathbb{P}_{\psi^*} - a.s.$, the existence of the relative entropy rate follows from the ergodic theorem and yields the explicit representation $H(\psi^*, \psi) = \mathbb{E}_\psi(\log p_\psi(Y_1|Y_{-\infty,0}))$.

For MS-NAR processes, an ergodic theorem of the *log*-likelihood is obtained in Rynkiewicz (2000) and Krishnamurthy (2002) using an additive function of the extended Markov chain

$$(Y_n, X_n, \mathbb{P}_\psi(X_k|Y_{0:n})).$$

In Douc et al. (2004) the law of large numbers of the *log*-likelihood follows from the uniform exponential memory loss of the initial distribution of the prediction filter.

The principal contribution of this paper is a new approach to prove the consistency of the maximum likelihood estimator for nonlinear AR processes with Markov-switching. Also, we prove an exponential uniform memory loss property for the prediction filter, approximating it by a filter with finite memory, as opposed to the classic approach that approximates by an infinite memory filter. From the α-mixing property for MS-NAR processes, we obtain an ergodic theorem. Finally, we show that, in the linear and Gaussian case, our assumptions are fully satisfied.

The paper is organized as follows. In Section 1, we set the notation and we present the general conditions of the model. In Section 2, we prove the consistency of the maximum likelihood estimator for MS-NAR. First, we prove an exponential uniform memory loss property for the prediction filter, approximating it by a filter with finite memory. Then, from α-mixing property for MS-NAR processes, we obtain an ergodic theorem.

2. Preliminaries

We review the key properties of the MS-NAR model. To achieve this goal, we will include and demonstrate some known results. The following conditions will be used throughout the article.

Assumption 1. The variable Y_0, the Markov chain $\{X_n\}_{n\geq 1}$ and the sequence $\{e_n\}_{n\geq 1}$ are mutually independent.

The process $\{X_n\}$, called regime, is not observable and its inference has to be carried out in terms of the observable process $\{Y_n\}$.

Assumption 2. The Markov chain $\{X_n\}_{n\geq 1}$ is positive recurrent and aperiodic. Hence, it has an invariant distribution that we denote by $\mu = (\mu_1, \ldots, \mu_m)$.

Assumption 3.

1. The functions $y \to r(y, \theta_i)$, for $i = 1, \ldots, m$, are continuous.
2. The random variables $\{e_n\}_{n\geq 1}$ admit a common density probability function Φ with respect to the Lebesgue measure.
3. There exists $b > 0$ and a compact set C of $\mathbb{R}$ such that $\inf_{e\in C} \Phi(e) > b$.

We recall the markovian properties of the MS-NAR model. Assumption 1 implies that $\{(Y_n, X_n)\}_{n\geq 1}$ with states space $\mathbb{R}\times\{1,\ldots,m\}$ is a Markov process. Under assumption 3.1 the Markov process $\{(Y_n, X_n)\}_{n\geq 1}$ is a Feller Markov chain and it is a strong Feller Markov chain if the assumptions 3.2 and 3.3 also hold.

Assumption 4.

1. There exist positive constants ρ_i, b_i, $i = 1,\ldots,m$, such that for $y\in\mathbb{R}$, the following inequality holds
$$|r(y,\theta_i)| \leq \rho_i|y| + b_i.$$
2. $\gamma = \sum_{i=1}^{m} \log \rho_i\mu_i < 0$.

Assumption 5. $\mathbb{E}(|e_1|^s) < \infty$, for some $s \geq 1$.

The model is called sublinear if Assumption 4 holds. For the sublinear MS-NAR model, Yao and Attali (1999) proved that there exists a unique stationary geometric ergodic solution. Moreover, under Assumption 5, if the spectral radius of the matrix $Q_s = \left(\rho_j^s a_{ij}\right)_{i,j=1\ldots m}$ is strictly less than 1, then $\mathbb{E}(|Y_n|^s) < \infty$. The Markov chain is stable under the moment condition $s \geq 1$, but for the asymptotic properties of the MLE it will be necessary to assume $s > 2$.

Assumption 6. The random variable Y_0 admits a density function $p(Y_0 = y_0)$ with respect to Lebesgue measure.

Under conditions 3.2 and 6, the random vector $(Y_{0:n}, X_{1:n})$ admits the probability density $p(Y_{0:n} = y_{0:n}, X_{1:n} = x_{1:n})$ equal to

$$\Phi(y_n - r(y_{n-1},\theta_{x_n}))\cdots\Phi(y_1 - r(y_0,\theta_{x_n}))a_{x_{n-1}x_n}\cdots a_{x_1x_2}\mu_{x_1}p(Y_0 = y_0),$$

with respect to the product measure $\lambda\otimes\mu_c$, where λ and μ_c denote Lebesgue and counting measures respectively. For a proof of this result see Fermín et al. (2017).

2.1. Mixing property

A strictly stationary stochastic process $Y = \{Y_n\}_{n\in\mathbb{Z}}$ is called strongly mixing, if

$$\alpha_n := \sup\{|\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)| : A\in\mathcal{M}_{-\infty}^0, B\in\mathcal{M}_n^\infty\} \to 0, \quad \text{as } n\to\infty, \tag{2}$$

where $\mathcal{M}_a^b$, with $a, b \in \overline{\mathbb{Z}}$, is the σ-algebra generated by $\{Y_k\}_{k=a:b}$. Under Assumptions 1–6 the process MS-NAR is α-mixing with α-mixing coefficients decreasing geometrically. For the proof of this statement see Fermín et al. (2017). In the following example we can see that if assumption 3.2 is not satisfied then we can build a non-mixing MS-AR.

Example 2.1 (*Linear Autoregressive with Non-Mixing MS-AR*)**.** In the case where $r(y,(b_i,\rho_i)^t) = \rho_i y + b_i$, the model is a MS-AR and it is defined by:

$$Y_n = \rho_{X_n}Y_{n-1} + b_{X_n} + e_n. \tag{3}$$

For each $1\leq i\leq m$, we denote $\theta_i = (b_i,\rho_i)^t$ and

$$\theta = \begin{pmatrix} b_1 & b_2 & \cdots & b_m \\ \rho_1 & \rho_2 & \cdots & \rho_m \end{pmatrix}.$$

More specifically, we consider the process MS-AR with $\theta_i = (0,\rho_i)^t$ for all $i = 1,\ldots,m$ and such that the random variable e_1 has a Bernoulli distribution with parameter q, and $Y_0 = 0$. In this case, we have

$$Y_n = \sum_{k=0}^{n-1}\rho_{X_k}\cdots\rho_{X_1}e_{k+1},$$

and we adopt the convention that $\rho_{X_k}\cdots\rho_{X_1} = 1$ for $k = 0$. This process is non α-mixing. In fact, according to Andrews (1984) if $0 < \rho_i \leq 1/2$, for $t\in\mathbb{N}$ there exist some sets $A\in\mathcal{M}_{-\infty}^0$, $B_t\in\mathcal{M}_n^\infty$, with $\mathbb{P}(A) > 0$, $\mathbb{P}(B_s)\leq c$ for some constant $c < 1$ such that $\mathbb{P}(B_t|A) = 1$, therefore

$$\alpha_t(Y) \geq \mathbb{P}(A\cap B_t) - \mathbb{P}(A)\mathbb{P}(B_t) = \mathbb{P}(A)(\mathbb{P}(B_t|A) - \mathbb{P}(B_t)) \geq \mathbb{P}(A)(1-c).$$

This implies that $\alpha_t(Y)$ does not tend to 0 as $t\to\infty$ and so Y is a non α-mixing process.

3. Maximum likelihood estimation

The aim of this section is to show consistency of the MLE for the MS-NAR model. The conditional log-likelihood is

$$\begin{aligned} l_n(\psi) &= \sum_{k=1}^{n} \log p_\psi(Y_k|Y_{0:k-1}) \\ &= \sum_{k=1}^{n} \log \iint \Phi(Y_k - r(Y_{k-1}, \theta_{x_k}))a_{x_{k-1},x_k}\mathbb{P}(x_{k-1}|Y_{l:k-1})\mu_c(dx_k)\mu_c(dx_{k-1}). \end{aligned}$$

We defined the maximum likelihood estimator (MLE) of ψ by $\hat{\psi}_n = \arg\max_\psi l_n(\psi)$. We denote by ψ^* the fixed true parameter. The MLE is consistent if $\hat{\psi}_n \to \psi^*$ as $n \to \infty$ a.s. We need to show that for each $\psi \in \Psi$, $n^{-1}l_n(\psi)$ converges uniformly to $H(\psi^*, \psi)$, which is a deterministic function with a single global maximum in ψ^*. Under continuity of the function l_n, it attains its maximum on a compact set, then for all n we can find $\psi_n \in \arg\max_\psi l_n(\psi)$.

Assumption 7. Let $\inf_{i,j=1:m} a_{ij} > \delta$.

We can express $p_\psi(Y_k|Y_{0:k-1})$ as a functional of the prediction filter $\mathbb{P}_\psi(X_k|Y_{0:n})$,

$$D^\psi_{k,l} = \log \iint \Phi(Y_k - r(Y_{k-1}, \theta_{x_k}))a_{x_{k-1},x_k}\mathbb{P}(x_{k-1}|Y_{l:k-1})\mu_c(dx_k)\mu_c(dx_{k-1}),$$

for $0 < l < k$.

The following lemma proves that the quantity $D^\psi_{k,0}$, which depends on the observations $Y_{0:k}$, can be approximated by $D^\psi_{k,l}$ which is a function of only a fixed number of observations $Y_{l:k}$.

Lemma 3.1. *Under Assumptions 1 and 7 the following inequality holds*

$$|D^\psi_{k,l} - D^\psi_{k,0}| \leq 2\delta^{-1}(1-\delta)^{k-1-l}.$$

Proof. First, we bound from below the quantities $\exp(D^\psi_{k,0})$ and $\exp(D^\psi_{k,l})$. By the Fubini Theorem we have

$$\exp(D^\psi_{k,0}) \geq \delta \int \Phi(Y_k - r(Y_{k-1}, \theta_{x_k}))\mu_c(dx_k)$$

and the same for $\exp(D^\psi_{k,l})$, thus

$$\min(\exp(D^\psi_{k,0}), \exp(D^\psi_{k,l})) \geq \delta \int \Phi(Y_k - r(Y_{k-1}, \theta_{x_k}))\mu_c(dx_k).$$

Using the inequality $|\log x - \log y| \leq |x-y|/\min(x,y)$, we obtain the estimate

$$\begin{aligned} &|D^\psi_{k,l} - D^\psi_{k,0}| \\ &\leq \frac{\iint \Phi(Y_k - r(Y_{k-1}, \theta_{x_k}))a_{x_{k-1},x_k}|\mathbb{P}_\psi(x_{k-1}|Y_{l:k-1}) - \mathbb{P}_\psi(x_{k-1}|Y_{0:k-1})|\mu_c(dx_k)\mu_c(dx_{k-1})}{\delta \int \Phi(Y_k - r(Y_{k-1}, \theta_{x_k}))\mu_c(dx_k)} \\ &\leq \frac{1}{\delta}\|\mathbb{P}_\psi(X_{k-1} \in \cdot|Y_{l:k-1}) - \mathbb{P}_\psi(X_{k-1} \in \cdot|y_{0:k-1})\|_{TV}. \end{aligned}$$

Applying the Proposition 4.3.26 (iii) in Cappe et al. (2005), pág 109, which asserts that

$$\|\mathbb{P}_\psi(X_k \in \cdot|Y_{l:k}) - \mathbb{P}_\psi(X_k \in \cdot|Y_{0:k})\|_{TV} \leq 2(1-\delta)^{k-l},$$

we conclude that $|D^\psi_{k,l} - D^\psi_{k,0}| \leq 2\delta^{-1}(1-\delta)^{k-l}$. ■

Our main result is presented in Theorem 3.1. In order to prove this theorem we take advantage of the mixing property of the MS-AR model and the bounds of the finite approximation filter.

Theorem 3.1. *Under Assumptions 1–7, suppose Ψ is a compact set and the condition*

$$\mathbb{E}_{\psi^*}\left(\sup_{\theta_i} |\log \Phi(Y_1 - r(Y_0, \theta_i))|\right) < \infty, \tag{4}$$

for $i = 1, \ldots, m$ holds. Then $l_n(\psi)$ is a continuous function and $H(\psi^*, \psi) = \lim_{n\to\infty} n^{-1}l_n(\psi)$ exists a.s for each $\psi \in \Psi$.

Proof. We note that the likelihood function can be written as

$$\frac{1}{n}l_n(\psi) = \frac{1}{n}\sum_{k=1}^{n}(D_{k,0}^{\psi} - D_{k,l}^{\psi}) + \frac{1}{n}\sum_{k=1}^{n}(D_{k,l}^{\psi} - \mathbb{E}_{\psi^*}(D_{k,l}^{\psi})) + \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}_{\psi^*}(D_{k,l}^{\psi}) = T_1 + T_2 + T_3.$$

We will prove that T_1 and T_2 converge to zero as $n \to \infty$, and the term T_3 converges to $H(\psi^*, \psi)$ a.s, for each $\psi \in \Psi$.

From Lemma 3.1

$$|D_{k+l,l}^{\psi} - D_{k+l,0}^{\psi}| \leq 2\delta^{-1}(1-\delta)^k \tag{5}$$

hence $\sup_l |D_{k+l,l}^{\psi} - D_{k+l,0}^{\psi}| \to 0$ as $k \to \infty$, this implies, according to Cesaro's Theorem, that T_1 converges to zero.

Now, taking expectation with respect to $Y_{0:k}$ in inequality (5), by Jensen inequality and the stationarity of Y we have

$$\sup_l |\mathbb{E}_{\psi^*}(D_{k+l,l}^{\psi}) - \mathbb{E}_{\psi^*}(D_{k,l}^{\psi})| = \sup_l |\mathbb{E}_{\psi^*}(D_{k,0}^{\psi}) - \mathbb{E}_{\psi^*}(D_{k,l}^{\psi})| \leq 2\delta^{-1}(1-\delta)^k$$

Thus, $\{D_{k,l}\}_{k\geq 1}$ is a Cauchy sequence and therefore convergent, i.e.

$$H(\psi^*, \psi) = \lim_{k\to\infty} \mathbb{E}_{\psi^*}(D_{k,l}^{\psi}), \text{ exists for every } \psi \in \Psi.$$

And by Cesaro's theorem, $T3$ also converges to the function $H(\psi^*, \psi)$. Finally, we proceed to demonstrate that

$$\frac{1}{n}\sum_{k=1}^{n}(D_{k,l}^{\psi} - \mathbb{E}_{\psi^*}(D_{k,l}^{\psi})) \to 0 \ a.s.$$

We first prove that $\mathbb{E}(|D_{k,l}^{\psi}|) < \infty$, indeed

$$\mathbb{E}(|D_{k,l}^{\psi}|) = \mathbb{E}\left(\left|\log\left\{\sum_{j=1}^{m}\Phi(Y_k - r(Y_{k-1}, \theta_j))\mathbb{P}_{\psi}(x_k = j|Y_{l:k-1})\right\}\right|\right),$$

and we have that

$$\log\sum_{j=1}^{m}\Phi(Y_k - r(Y_{k-1}, \theta_j))\mathbb{P}_{\psi}(x_k = j|Y_{l:k-1}) \leq \log\sup_{\theta_i}\Phi(Y_k - r(Y_{k-1}, \theta_i)).$$

On the other hand,

$$\begin{aligned}\inf_{\theta_i}\log\Phi(Y_k - r(Y_{k-1}, \theta_i)) &\geq \inf_{\theta_i}\{-|\Phi(Y_k - r(Y_{k-1}, \theta_i))|\} \\ &= -\sup_{\theta_i}|\Phi(Y_k - r(Y_{k-1}, \theta_i))|\end{aligned}$$

yields

$$\left|\log\left\{\sum_{j=1}^{m}\Phi(Y_k - r(Y_{k-1}, \theta_j))\mathbb{P}_{\psi}(x_k = j|Y_{l:k-1})\right\}\right| \leq \sup_{\theta_i}|\log\Phi(Y_k - r(Y_{k-1}, \theta_i))|$$

and by (4)

$$\begin{aligned}&\mathbb{E}\left(\left|\log\left\{\sum_{j=1}^{m}\Phi(Y_k - r(Y_{k-1}, \theta_j))\mathbb{P}_{\psi}(x_k = j|Y_{l:k-1})\right\}\right|\right) \\ &\qquad \leq \mathbb{E}_{\psi^*}\left(\sup_{\theta_i}|\log\Phi(Y_k - r(Y_{k-1}, \theta_i))|\right) < \infty.\end{aligned}$$

We have shown that the sequence $\{D_{k,l}^{\psi}\}_{k\geq 1}$ is α-mixing with geometric coefficients α_k and $\mathbb{E}(|D_{k,l}^{\psi}|) < \infty$; according to Corollary 3.1 in Rio (2000) we have

$$\frac{1}{n}\sum_{k=1}^{n}(D_{k,l}^{\psi} - \mathbb{E}_{\psi^*}(D_{k,l}^{\psi})) \to 0 \ a.s.$$

■

The next results follow from the classic approach for the consistency of the maximum likelihood estimator introduced by Wald.

Lemma 3.2. *Suppose Ψ is a compact set. Let $l_n : \Psi \to \mathbb{R}$ be a sequence of continuous functions that converges uniformly to a function $l : \Psi \to \mathbb{R}$. Then*

$$\hat{\psi}_n = \arg\max_{\psi} l_n(\psi) \to \arg\max_{\psi} H(\psi^*, \psi).$$

Proof. As a continuous function on a compact space attains its maximum, for all n we can find $\psi_n \in \arg\max_{\psi} l_n(\psi)$. Using an argument that goes to Wald (1949) we have

$$\lim_{n\to\infty} H(\psi^*, \psi_n) = \sup_{\psi\in\Psi} H(\psi^*, \psi). \tag{6}$$

Suppose that the sequence $\{\psi_n\}$ does not converge to the set

$$\{\tilde{\psi} : H(\psi^*, \tilde{\psi}) = \max_{\psi\in\Psi} H(\psi^*, \psi)\}.$$

By compactness there exists a subsequence $\{\psi_n'\} \subset \{\psi_n\}$ which converges to $\psi' \notin \{\tilde{\psi} : H(\psi^*, \tilde{\psi}) = \max_{\psi\in\Psi} H(\psi^*, \psi)\}$. But $H(\psi^*, \psi)$ is continuous, so $H(\psi^*, \psi_n') \to H(\psi^*, \psi') < \sup_{\psi\in\Psi} H(\psi^*, \psi)$ and according to (6), this is a contradiction. ■

Theorem 3.2. *Suppose Ψ is a compact set. Assume that*

1. *$\psi = \psi^*$ iff $\mathbb{P}_\psi = \mathbb{P}_{\psi^*}$.*
2. *For all $i, j \in \{1, \ldots, m\}$ and all $y, y' \in \mathbb{R} \times \mathbb{R}$ the functions $\psi \to a_{ij}$ and $\psi \to p_\psi(Y_1 = y | Y_0 = y', X_1 = i)$ are continuous.*
3. *There exists a constant $c < \infty$ such that $|D_k^\psi - D_k^{\psi'}| \le c\|\psi - \psi'\|$, for all integers $k > 1$.*

Then the maximum likelihood estimate $\hat{\psi}_n$ is consistent.

Proof. By Theorem 3.1, the Lipschitz condition 3 and compactness implies that the sequence $l_n \to H$ a.s uniformly. According to Lemma 2.2

$$\hat{\psi}_n \to \psi_* = \arg\max_{\psi} H(\psi^*, \psi),$$

and this value is unique under identifiability. ■

In the Gaussian and linear case we can prove directly identifiability and equicontinuity. This allows us to obtain the consistency of the MLE without assuming a Lipschitz condition for the parameters.

Theorem 3.3. *Suppose an MS-AR model with Gaussian innovations $\{e_n\}$. Under the Assumptions 1–7, suppose Ψ is a compact set. We assume that for the true model Ψ^* the vector components $\{(\alpha_i, b_i, \sigma_i)\}_{i=1}^m$ are different. Then the maximum likelihood estimate $\hat{\psi}_n$ is consistent.*

Proof. Consider the model defined by (3) and that the $\{e_n\}$ are Gaussian i.i.d. random variables. Our goal in this example is to check that the conditions for consistency apply in this case. In fact, under the assumption that for the true model Ψ^* the vector components $\{(\alpha_i, b_i, \sigma_i)\}_{i=1}^m$ are different for every n, there exists a point $Y_{n-1} \in \mathbb{R}$ such that the $\{(\alpha_i Y_{n-1} + b_i, \sigma_i)\}_{i=1}^m$ are different. Therefore, in agreement with the Remark 2.10 of Krishnamurthy and Yin (Krishnamurthy, 2002) the model is identifiable in the following sense: If K stands for the Kullback–Leibler divergence and $K(\psi, \psi^*) = 0$, then $\psi = \psi^*$, which proves the identifiability. On the other hand, from Lemma 4.1, in Ríos and Rodríguez (2008) it follows that $\frac{1}{n}\log p_\psi(Y_1^n | Y_0 = y_0)$ is an equicontinuous sequence a.s-$\mathbb{P}_{\psi_*}$.

Condition (4) is satisfied if Y_1 has a moment of order 2. Indeed, conditioning with respect to $Y_0 = y_0$, we have,

$$\mathbb{E}\left(\mathbb{E}_{\psi^*}\left(\sup_{\theta_i} |\log \Phi(Y_1 - y_0\rho_i - b_i)| \,|Y_0 = y_0\right)\right)$$
$$\le C + \mathbb{E}\left(\int \sup_{\rho_i, b_i} \frac{(y_1 - \rho_i y_0 - b_i)^2}{2\sigma^2} f(y_1) dy_1 | Y_0 = y_0\right) < \infty.$$

We conclude that in this case the MLE is consistent. ■

Acknowledgments

Luis-Angel Rodriguez is grateful for the partial support on the projects Anillo ACT1112 and GEMINI-CONICYT 2012 No. 32120025 CR 211291055 REXE 04464-14. Research facilities and hospitality in CIMFAV of the Universidad de Valparaíso and the sabbatical support from the Universidad de Carabobo. The authors are grateful to K. Bertin, P. Linares and S. Brassesco for their careful reading.

References

Andrews, D., 1984. Non-strong mixing autoregressive processes. J. Appl. Probab. 21, 930–934.
Cappe, O., Moulines, E., Rydén, T., 2005. Inference in Hidden Markov Models. Springer-Verlag.
Douc, R., Moulines, E., Rydén, T., 2004. Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. Ann. Statist. 32, 2254–2304.
Fermín, L., Ríos, Rodríguez, L.A., 2017. A Robbins Monro algorithm for nonparametric estimation of NAR process with Markov-Switching: consistency. J. Time Ser. Anal. 38 (6), 809–837.
Goldfeld, S.M., Quandt, R., 1973. A Markov model for switching regressions. J. Econometrics 1, 3–16.
Hamilton, J.D., 1989. A new approach to the economic analysis of non stationary time series and the business cycle. Econometrica 357–384.
Krishnamurthy, V., 2002. Recursive Algorithms for estimation of hidden Markov Models with markov regime. IEEE Trans. Inform. Theory 48 (2), 458–476.
Krishnamurthy, V., Rydén, T., 1998. Consistent estimation of linear and non-linear autoregressive models with Markov regime. J. Time Series Anal. 19, 291–307.
Rio, E., 2000. Théorie asymptotique des processus faiblement dépendents, vol. 31. Springer-SMAI, Paris.
Ríos, R., Rodríguez, L.A., 2008. Penalized estimate of the number of states in gaussian linear ar with markov regime. Electron. J. Stat. 1111–1128.
Rynkiewicz, J., 2000. Modéles hybrides intégrant des réseaux de neurones artificiels à des modeles de chaînes de markov cachee: application à la prediction de series temporelles (Ph.D. thesis). Universite Paris I.
Yao, J., Attali, J.G., 1999. On stability of nonlinear AR process with Markov switching. Adv. Appl. Probab.

Further reading

Doukhan, P., 1994. Mixing: Properties and Examples. In: Lecture Notes in Statist., vol. 85.
Handel, R.v., 2008. Hidden Markov Models. In: Lecture notes, URL https://www.princeton.edu/rvan/.
Vandekerkhove, P., 2005. Consistent and asymptotically normal parameter estimates for hidden Markov mixtures of Markov models. Bernoulli 11, 103–129.