# ASYMPTOTIC PROPERTIES OF AUTOREGRESSIVE REGIME-SWITCHING MODELS

M. OLTEANU AND J. RYNKIEWICZ
*SAMOS-MATISSE CES UNIVERSITE PARIS 1*
*ESAIM: PS, 16 (2012) 25-47*

Abstract. The statistical properties of the likelihood ratio test statistic (LRTS) for autoregressive regime-switching models are addressed in this paper. This question is particularly important for estimating the number of regimes in the model. Our purpose is to extend the existing results for mixtures (Liu and Shao, 2003) and hidden Markov chains (Gassiat, 2002). First, we study the case of mixtures of autoregressive models (i.e. independent regime switches). In this framework, we give sufficient conditions to keep the LRTS tight and compute its the asymptotic distribution. Second, we consider the extension of the ideas in Gassiat (2002) to autoregressive models with regimes switches according to a Markov chain. In this case, it is shown that the marginal likelihood is no longer a contrast function and cannot be used to select the number of regimes. Some numerical examples illustrate the results and their convergence properties.

## 1. Introduction

Autoregressive regime-switching models are being widely used in modelling financial and economic time series such as business cycles (Hamilton, 1989; Lam, 1990), exchange rates (Engle and Hamilton, 1990), financial panics (Schwert, 1989) or stock prices (Wong and Li, 2000).

When the number of regimes is fixed, the statistical inference is relatively straightforward (Hamilton, 1990) and the asymptotic properties of the estimates have already been established (Francq and Roussignol, 1998; Krishnamurthy and Ryden, 1998; Douc R., Moulines E. and Rydén T., 2004). However, the problem of selecting the number of regimes is far less obvious and hasn't been completely answered yet. When the number of regimes is unknown, identifiability problems arise and the likelihood ratio test statistic (LRTS hereafter) is no longer convergent to a $\chi^2$-distribution. Some partial answers were proposed by Hansen (1992,1996a, 1996b) and Garcia (1998). Hansen derived an asymptotic bound for the distribution of the LRTS based on empirical processes techniques, while Garcia obtained the asymptotic distribution of the LRTS, but under some very restrictive hypothesis. Let us also mention that the consistency of the estimate of the number of regimes was proven recently in a Bayesian framework (Rios, 2008).

In the particular case of mixture models, several ideas and methods were proposed to estimate the number of components: non-parametric techniques as in Henna (1985), Roeder (1994) or Izenman and Sommer (1998), moment techniques in Lindsay (1983) or Dacunha-Castelle and Gassiat (1997) and penalized maximum-likelihood in Leroux(1992a), Keribin (2000) and Liu and Shao (2003). Furthermore, Gassiat (2002) proved that in the case of hidden Markov models, the number of regimes can be estimated using a marginal penalized-likelihood estimate. The aim of this paper is to extend the existing results for mixtures and hidden Markov models to the case where the mean of the observed process is replaced by a regression function.

In Section 2, the results on the LRTS for mixture models are extended to autoregressive regime-switching models with independent regime switches. We give sufficient conditions for the tightness of the LRTS and compute its asymptotic distribution. Section 3 is devoted to verifying the result of the previous section in the case where the noise is Gaussian and the regression functions are linear. The last section handles the case where regime switches are Markovian. Once the result in the independent case was established, it seemed natural to generalize it by using a cost function close to the marginal likelihood, as defined in Gassiat (2002). Yet, it can be seen right away that this is no longer a contrast function and the convergence is achieved only in the particular cases of constant autoregressive functions (hidden Markov models) or independent regime switches (autoregressive mixture models).

## 2. LRTS for autoregressive mixture models

2.1. **The observations.** Let us briefly recall the definition of strong mixing processes which will be needed hereafter. For a more detailed review, refer to Doukhan (1995) and Bradley (2005).

Let $(Y_k)_{k\in\mathbb{Z}}$ be a strictly stationary sequence of random variables defined on a probability space $(\Omega, \mathcal{K}, \mathbb{P})$. For every $n \geq 1$, define the $\beta$-mixing coefficients

$$\beta_n = \beta\left(\mathcal{F}_{-\infty}^{0},\, \mathcal{F}_{n}^{\infty}\right)$$

where $\mathcal{F}_{-\infty}^{0} = \sigma\left(Y_k,\, k \leq 0\right)$, $\mathcal{F}_{n}^{\infty} = \sigma\left(Y_k,\, k \geq n\right)$, as

$$\beta\left(\mathcal{A}, \mathcal{B}\right) = \frac{1}{2} \sup_{(A_i)_{i\in I}, (B_j)_{j\in J}} \sum_{(i,j)\in I\times J} \left|\mathbb{P}\left(A_i \cap B_j\right) - \mathbb{P}\left(A_i\right)\mathbb{P}\left(B_j\right)\right|$$

where $(A_i)_{i\in I}$ (resp. $(B_j)_{j\in J}$) ranges over the set of $\mathcal{A}$(resp. $\mathcal{B}$) measurable partitions.

The sequence $(Y_k)_{k\in\mathbb{Z}}$ is called $\beta$-mixing if $\lim_{n\to\infty}\beta_n = 0$.

Throughout the rest of the paper, we will assume that the observations are a realization of a stationnary process $(Y_k)$. Moreover, $(Y_k)$ will be assumed to be geometrically $\beta$-mixing. This assumption may seem strong, but actually it is fulfilled by a wide class of processes.

Finally, let us denote by $\mu$ the stationnary measure of the vector $(Y_{k+1}, Y_k)$.

2.2. **The model.** Let $\mathcal{P} = \{g_\theta, \theta \in \Theta\}$ be a set of densities with respect to some positive measure $\nu$, where $\Theta$ is a compact finite-dimensional set.

Let us consider an observed sample $\{y_1, ..., y_n\}$ of the series $Y_k$. For every $y_k$, the true density conditionally to $y_{k-1}$ is

$$g^0\left(y_k \mid y_{k-1}\right) = \sum_{i=1}^{p_0} \pi_i^0 g_{\theta_i^0}\left(y_k \mid y_{k-1}\right),$$

where $g_{\theta_i^0} \in \mathcal{P}$, $\pi_i^0 > 0$ and $\sum_{i=1}^{p_0} \pi_i^0 = 1$.

This model is a generalization of mixture models. Several regression models can be written in this way, for example mixtures of linear regressions with Gaussian noise, which are particularly important in econometrics (see Hansen 1996) and will be studied in Section 3 :

(1) $$Y_t = a^0_{X_t} Y_{t-1} + b^0_{X_t} + \sigma^0_{X_t} \varepsilon_t \,,$$

where

- $X_t$ is an i.i.d. sequence of random variables valued in a finite space $\{1, ..., p_0\}$ and with probability distribution $\pi^0 = \left(\pi^0_1, \cdots, \pi^0_{p_0}\right)$
- $\left(a^0_1, \cdots, a^0_{p_0}, b^0_1, \cdots, b^0_{p_0}\right)$ are real numbers
- $\left(\sigma^0_1, \cdots, \sigma^0_{p_0}\right)$ are strictly positive real numbers
- $\varepsilon_t$ is an i.i.d. noise $\mathcal{N}(0,1)$, independent of $(Y_{t-k})_{k\geq 1}$.

Let us remark that if $\left(a^0_1, \cdots, a^0_{p_0}\right)$ are all zero, the model is a simple Gaussian mixture.

2.3. **Approximation of the LRTS.** Let $\mathcal{G}$ be the set of possible conditional densities:

$$\mathcal{G} = \left\{ g\left(y_k \mid y_{k-1}\right) = \sum_{i=1}^{p} \pi_i g_{\theta_i}\left(y_k \mid y_{k-1}\right), \pi_i \in [0;1], \sum_{i=1}^{p} \pi_i = 1, g_{\theta_i} \in \mathcal{P}, p \in \mathbb{N}^\star \right\}$$

If $p \leq p_0$, there are no identification issues. Therefore, we will assume that $p > p_0$ in the sequel.

Let

$$l_n\left(g\right) = \sum_{k=2}^{n} \ln\, g\left(y_k \mid y_{k-1}\right)$$

be the log-likelihood function of $(y_1, \cdots, y_n)$, conditionally to $y_1$.

The LRTS is defined as:

(2) $$2\lambda_n = 2\left(\sup_{g\in\mathcal{G}} \ln(g) - \ln(g^0)\right) = 2 \sup_{g\in\mathcal{G}} \frac{\sum_{k=2}^{n}\sum_{i=1}^{p} \pi_i g_{\theta_i}\left(y_k \mid y_{k-1}\right)}{\sum_{k=2}^{n}\sum_{i=1}^{p_0} \pi^0_i g_{\theta^0_i}\left(y_k \mid y_{k-1}\right)}$$

We establish a theorem giving an approximation of the LRTS. Some notations and definitions are needed first:

- For an $\eta > 0$, denote

(3) $$\mathcal{G}_\eta := \left\{g \in \mathcal{G}, \|g - g^0\|_{L^2(\mu)} \leq \eta\right\}.$$

The extended set of score-functions $\mathcal{S}_\eta$ is defined as:

(4) $$\mathcal{S}_\eta = \left\{ s_g = \frac{\frac{g}{g^0} - 1}{\left\|\frac{g}{g^0} - 1\right\|_{L^2(\mu)}}, g \in \mathcal{G}_\eta \right\}.$$

- Let us define the limit-set of scores $\mathcal{D}$

$$\mathcal{D} = \left\{ d \in \mathbb{L}^2(\mu) \mid \exists (g_n) \in \mathcal{G}, \| \frac{g_n - g^0}{g^0} \|_{\mathbb{L}^2(\mu)} \xrightarrow[n\to\infty]{} 0, \|d - s_{g_n}\|_{\mathbb{L}^2(\mu)} \xrightarrow[n\to\infty]{} 0 \right\}.$$

  By putting $g_t = g_n$ for $t \in [0,1]$ and $n \leq \frac{1}{t} < n+1$, we obtain that, for all $d \in \mathcal{D}$, there exists a parametric path $(g_t)_{0\leq t\leq 1}$ such that $\forall t \in [0,1]$, $g_t \in \mathcal{G}$, $t \to \| \frac{g_t - g^0}{g^0} \|_{\mathbb{L}^2(\mu)}$ is continuous, $\| \frac{g_t - g^0}{g^0} \|_{\mathbb{L}^2(\mu)} \xrightarrow[t\to 0]{} 0$ and $\|d - s_{g_t}\|_{\mathbb{L}^2(\mu)} \xrightarrow[t\to 0]{} 0$.
- We recall the definition of the $\mathcal{L}_{2,\beta}(\mathbb{P})$-space and the notion of bracketing entropy. Consider $Z_k$ a strictly stationary sequence, $\beta$-mixing and such that $\sum_{n\geq 1} \beta_n < \infty$. The $\mathcal{L}_{2,\beta}(\mathbb{P})$-space is defined as

$$\mathcal{L}_{2,\beta}(\mathbb{P}) = \left\{ f, \ \|f\|_{2,\beta} < \infty \right\}, \ \ \|f\|_{2,\beta} = \sqrt{\int_0^1 \beta^{-1}(u) \left[ Q_f(u) \right]^2 du}$$

  where
  - $\beta(u)$ is the càdlàg extension of $\beta_n$ by considering $\beta(u) = \beta_{[u]}$ and $\beta_0 = 1$
  - $\varphi^{-1}(u) = \inf\{t \in \mathbb{R}, \varphi(t) \leq u\}$, if $\varphi$ is a non-increasing function
  - $Q_f$ is the quantile function of $|f(Z_0)|$, that is the inverse of $t \to \mathbb{P}(|f(Z_0)| > t)$

  Consider the extended set of score-functions $\mathcal{S}_\eta$ endowed with the norm $\|\cdot\|_{2,\beta}$. For every $\varepsilon > 0$, we define an $\varepsilon$-bracket by $[l, u] = \{f \in \mathcal{F}, l \leq f \leq u\}$ such that $\|u - l\|_{2,\beta} < \varepsilon$. The $\varepsilon$-bracketing entropy is

$$\mathcal{H}_{[\cdot]}\left(\varepsilon, \mathcal{S}_\eta, \|\cdot\|_{2,\beta}\right) = \ln\left(\mathcal{N}_{[\cdot]}\left(\varepsilon, \mathcal{S}_\eta, \|\cdot\|_{2,\beta}\right)\right),$$

  where $\mathcal{N}_{[\cdot]}\left(\varepsilon, \mathcal{S}_\eta, \|\cdot\|_{2,\beta}\right)$ is the minimum number of $\varepsilon$-brackets necessary to cover $\mathcal{S}_\eta$.

With the previous notations, we introduce the following assumption **(B)**: Assume that $\mathcal{G}$ is Glivenko-Cantelli and that there exists $\eta > 0$ such that

$$\int_0^1 \sqrt{\mathcal{H}_{[\cdot]}\left(\varepsilon, \mathcal{S}_\eta, \|\cdot\|_{2,\beta}\right)} d\varepsilon < \infty.$$

Then, according to Doukhan (1995), the set $\mathcal{S}_\eta$ is Donsker under **(B)**.

Now, let us state the following theorem which generalizes the result of Gassiat (2002). The proof is given in the Appendix.

***Theorem 1***:
*Under the assumption* **(B)**,

$$2\lambda_n = \sup_{d\in\mathcal{D}} \left( \max\left\{ \frac{1}{\sqrt{n}} \sum_{i=2}^n d(Y_i, Y_{i-1}); 0 \right\} \right)^2 + o_P(1)$$

Although this result may be applied to more general models, this paper is restricted to autoregressive mixture models.

2.4. **Asymptotic law of the LRTS.** This section is a direct application of Theorem 1.

We give sufficient conditions for which the Donsker assumption **(B)** holds in the case of autoregressive mixture models.

Usually, for parametric models, a Lipschitz condition on $\theta$ is sufficient to show that $\mathcal{S}$ is Donsker. However, if $g$ depends on the parameter $\theta$, the score-function $\theta \mapsto s_g = \frac{\frac{g}{g^0}-1}{\left\|\frac{g}{g^0}-1\right\|_{L^2(\mu)}}$ may not be continuous, thus not Lipschitz, in $\theta_0$.

The following theorem shows that assumption **(B)** holds for autoregressive mixture models under some general hypothesis. Furthermore, we prove that the limit set of scores $\mathcal{D}$ is complete and has continuous parametric paths. Hence, the asymptotic behavior of the LRTS may be completely described.

**Assumptions for the tightness of LRTS.**

**H-1:** The set $\mathcal{G}$ is Glivenko-Cantelli and the set of possible parameters:

$$\{\pi_1, \cdots, \pi_p \in [0,1], \theta_1, \cdots, \theta_p \in \Theta\}$$

contains a neighborhood of the parameters defining the true conditional density $g^0$.

**H-2:** There exists $\eta > 0$ such that for all $g \in \mathcal{G}$ with $\|g - g^0\|_{L^2(\mu)} \leq \eta$, $\left\|\frac{g}{g^0} - 1\right\|_{L^2(\mu)} < \infty$

**H-3:** By denoting $l_{\theta_i} := \frac{g_{\theta_i}}{g^0}$ and, with a slight abuse of notation, $\frac{\partial^q}{\partial \theta_j^q}$ the derivative of order $q$ with respect to all components of $\theta_j$, we assume the existence of a square-integrable function $h$ and of a neighborhood $\mathcal{N}$ of $(\theta_1^0, \cdots, \theta_{p^0}^0)$ such that, for all $(\theta_1, \cdots, \theta_{p^0}) \in \mathcal{N}$,

$$\left|\frac{\partial l_{\theta_j}}{\partial \theta_j}(\theta_j)\right| \leq h, \ \left|\frac{\partial^2 l_{\theta_j}}{\partial \theta_j^2}(\theta_j)\right| \leq h \text{ and } \left|\frac{\partial^3 l_{\theta_j}}{\partial \theta_j^3}(\theta_j)\right| \leq h.$$

**H-4:** With the following notations:

$$l'_j := \frac{\partial l_{\theta_j}}{\partial \theta_j}\left(\theta_j^0\right), \ l'' := \frac{\partial^2 l_{\theta_j}}{\partial \theta_j^2}\left(\theta_j^0\right)$$

we assume that for distinct $(\theta_i)_{1 \leq i \leq p}$

$$\left\{(l_{\theta_i})_{1 \leq i \leq p}, (l'_i)_{1 \leq i \leq p^0}, (l''_i)_{1 \leq i \leq p^0}\right\}$$

are linearly independent in the Hilbert space $L^2(\mu)$.

Let us define $\Omega : L^2(P) \to L^2(\mu)$ by $\Omega(g) = \frac{g}{\|g\|_2}$, for $g \neq 0$.

Now, we can state the following theorem, which generalizes theorem 4.1 of Liu and Shao (2003) :

***Theorem 2***:
*Let $d$ be the parametric dimension of the regression functions. Under the assumptions H-1, H-2, H-3 and H-4, there exists a centered Gaussian process $\{W_S, S \in \mathbb{F}\}$ with continuous sample path and covariance kernel $P(W_{S_1} W_{S_2}) = P(S_1 S_2)$ such that*

$$\lim_{n \to \infty} 2\lambda_n = \sup_{S \in \mathbb{F}} \left(\max(W_S, 0)\right)^2.$$

*The index set $\mathbb{F}$ is defined as $\mathbb{F} = \cup_t \mathbb{F}_t$, with the union running over $t = (t_0, \cdots, t_{p_0}) \in \mathbb{N}^{p_0+1}$ with $0 = t_0 < t_1 < \cdots < t_{p_0} \leq p$ and*

$$\mathbb{F}_t = \Big\{\Omega\left(\textstyle\sum_{i=1}^{p_0} \zeta_i l_{\theta_i^0} + \sum_{i=p_0+1}^{p} \zeta_i l_{\theta_i} + \sum_{i=1}^{p_0} \lambda_i^T l'_i + \delta \sum_{i=1}^{p_0} \sum_{j=t_{i-1}+1}^{t_i} \gamma_j^T l''_i \gamma_j\right),$$
$$\lambda_1, \cdots, \lambda_{p_0}, \gamma_1, \cdots, \gamma_{t_{p_0}} \in \mathbb{R}^d \ ; \zeta_1, \cdots, \zeta_p \in \mathbb{R}, \theta_{t_{p_0}+1}, \cdots, \theta_p \in \Theta - \left\{\theta_1^0, \cdots, \theta_{p_0}^0\right\}\Big\}$$

*where* $\delta = 1$ *if there exists a vector* $\mathbf{q}$ *such that:*
$q_j \leq 0$, $\sum_{j=t_{i-1}+1}^{t_i} q_j = 1$, $\sum_{j=t_{i-1}+1}^{t_i} \sqrt{q_j}\gamma_j^t = 0$ *for* $i = 1, \cdots, p_0$*;*
*and* $\delta = 0$ *otherwise.*

Note that the asymptotic law of the LRTS depends on the true parameters of the model. The next two sections illustrate important consequences of this theorem.

2.5. **Penalized-likelihood estimate for the number of regimes.** For $p \in \mathbb{N}^\star$, let us denote

$$\mathcal{G}_p = \left\{ g\left(y_k \mid y_{k-1}\right) = \sum_{i=1}^{p} \pi_i g_{\theta_i}\left(y_k \mid y_{k-1}\right), \pi_i \in [0;1], \sum_{i=1}^{p} \pi_i = 1, g_{\theta_i} \in \mathcal{P} \right\}.$$

For some fixed $P \in \mathbb{N}^\star$ sufficiently large, we shall consider the following class of functions

$$\mathcal{G}_P = \bigcup_{p=1}^{P} \mathcal{G}_p$$

For every $g \in \mathcal{G}_P$ we define the number of regimes as

$$p\left(g\right) = \min\left\{p \in \{1, ..., P\}, \ g \in \mathcal{G}_p\right\}.$$

With this definition, $p_0 = p\left(g^0\right)$ is the number of regimes of the true model.

The estimate of the number of regimes $\hat{p}$ can now be defined as $p \in \{1, ..., P\}$ maximizing the penalized criterion:

$$T_n\left(p\right) = \sup_{g \in \mathcal{G}_p} l_n\left(g\right) - a_n\left(p\right) \tag{5}$$

where

$$l_n\left(g\right) = \sum_{k=2}^{n} \ln \, g\left(y_k \mid y_{k-1}\right)$$

is the conditional log-likelihood with respect to $y_1$ and $a_n\left(p\right)$ is a penalty term.

With the previous definitions, the following result can be stated:

***Corollary 1***
*Suppose the following assumptions are true:*

- *Assumptions* **H-1**, **H-2**, **H-3** *and* **H-4** *are true.*
- ***(A)*** $a_n\left(\cdot\right)$ *is an increasing function of* $p$, $a_n\left(p_1\right) - a_n\left(p_2\right) \xrightarrow[n\to\infty]{} \infty$ *for every* $p_1 > p_2$ *and* $\frac{a_n(p)}{n} \xrightarrow[n\to\infty]{} 0$ *for every* $p$.

*Then,* $\hat{p}$ *maximizing the penalized criterion defined by (5) converges in probability,* $\hat{p} \xrightarrow[n\to\infty]{\mathbb{P}} p_0$.

***Proof***

The result and its proof are inspired by Gassiat (2002) and Keribin (2000).

First, let us show that $\widehat{p}$ does not overestimate $p_0$.

$$\begin{array}{rcl}\mathbb{P}\left(\hat{p}>p_0\right) & \leq & \sum_{p=p_0+1}^{P}\mathbb{P}\left(T_n\left(p\right)>T_n\left(p_0\right)\right)\\ & = & \sum_{p=p_0+1}^{P}\mathbb{P}\left(\sup_{g\in\mathcal{G}_p} l_n\left(g\right)-a_n\left(p\right)>l_n\left(g^0\right)-a_n\left(p_0\right)\right)\\ & \leq & \sum_{p=p_0+1}^{P}\mathbb{P}\left(\lambda_n>a_n\left(p\right)-a_n\left(p_0\right)\right)\end{array}$$

Since $\lambda_n$ is tight (Theorem 1) and according to assumption **(A)**,

$$\mathbb{P}\left(\lambda_n>a_n\left(p\right)-a_n\left(p_0\right)\right)\rightarrow 0$$

Thus, $\mathbb{P}\left(\hat{p}>p_0\right)\rightarrow 0$.

Let us now prove that $\hat{p}$ does not underestimate $p_0$ :

$$\begin{array}{rcl}\mathbb{P}\left(\hat{p}<p_0\right) & \leq & \sum_{p=1}^{p_0-1}\mathbb{P}\left(T_n\left(p\right)>T_n\left(p_0\right)\right)\\ & = & \sum_{p=1}^{p_0-1}\mathbb{P}\left(\frac{1}{n-1}\sup_{g\in\mathcal{G}_p}\left(l_n\left(g\right)-l_n\left(g^0\right)\right)>\frac{a_n(p)-a_n(p_0)}{n-1}\right)\end{array}$$

For all $p<p_0$, we shall prove that $\frac{1}{n-1}\sup_{g\in\mathcal{G}_p}\left(l_n\left(g\right)-l_n\left(g^0\right)\right)$ converges in probability to a strictly negative value. Then, according to the hypothesis **(A)**, the proof will be complete.

Hypothesis **(H-1)** ensures that $\mathbb{E}_\mu\left(\ln g\right)<\infty$ for all $g\in\mathcal{G}_p$.

Let us define

$$K\left(g^0,\mathcal{G}_p\right)=\inf_{g\in\mathcal{G}_p}K\left(g^0,g\right),$$

where $K\left(g^0,g\right)=\mathbb{E}_\mu\left(\ln\left(\frac{g^0}{g}\right)\right)$.

Since the set of parameters is compact and $K\left(g^0,g\right)$ is continuous with respect to the parameters, $K\left(g^0,\mathcal{G}_p\right)$ attains its infimum for some $\tilde{g}\in\mathcal{G}_p$ and, according to the hypothesis **(H-4)**, $K\left(g^0,\tilde{g}\right)>0$.

By the definition of the maximum likelihood,

$$\frac{1}{n-1}\sup_{g\in\mathcal{G}_p}\left(l_n\left(g\right)-l_n\left(g^0\right)\right)\geq\frac{1}{n-1}\left(l_n\left(\tilde{g}\right)-l_n\left(g^0\right)\right)$$

Since $\left(Y_k,Y_{k-1}\right)$ is strictly stationary and geometrically ergodic,

$$\frac{1}{n-1}\left(l_n\left(\tilde{g}\right)-l_n\left(g^0\right)\right)\xrightarrow[n\rightarrow\infty]{\mathbb{P}}-\mathbb{E}_\mu\left(\ln\left(\frac{g^0}{\tilde{g}}\right)\right)=-K\left(g^0,\tilde{g}\right)$$

Thus,

$$\lim\inf_{n\rightarrow\infty}\frac{1}{n-1}\sup_{g\in\mathcal{G}_p}\left(l_n\left(g\right)-l_n\left(g^0\right)\right)\geq-K\left(g^0,\tilde{g}\right)$$

It remains to prove that

$$\lim\sup_{n\rightarrow\infty}\frac{1}{n-1}\sup_{g\in\mathcal{G}_p}\left(l_n\left(g\right)-l_n\left(g^0\right)\right)\leq-K\left(g^0,\tilde{g}\right)$$

Since the set of parameters is compact, for all $\eta>0$, it may be covered by a finite number of balls $N_\eta$, centered in $c_i=\left(\pi_i,\theta_i\right)_p$, $i=1,...,N_\eta$ and with radius $\frac{\eta}{2}$. Let us now define

$$m_\eta\left(\left(y_1,y_2\right),\mathcal{G}_p\right)=\sup_{d\left(\left(\pi_1,\theta_1\right)_p,\left(\pi_2,\theta_2\right)_p\right)\leq\eta}\left|l_n\left(g_{\left(\pi_1,\theta_1\right)}\left(y_1,y_2\right)\right)-l_n\left(g_{\left(\pi_2,\theta_2\right)}\left(y_1,y_2\right)\right)\right|$$

Now, we can write

$$\begin{aligned}
&\tfrac{1}{n-1}\sup_{g\in\mathcal{G}_p} l_n(g) - l_n\left(g^0\right) = \\
&= \sup_{g\in\mathcal{G}_p} \tfrac{1}{n-1}\sum_{k=2}^{n}\left(l_n\left(g\left(Y_k,Y_{k-1}\right)\right) - l_n\left(g^0\left(Y_k,Y_{k-1}\right)\right)\right) \\
&\leq \sup_{i=1,\dots,N_\eta}\left(l_n\left(g_{c_i}\right) - l_n\left(g^0\right)\right) + \tfrac{1}{n-1}\sum_{k=2}^{n} m_\eta\left(\left(Y_k,Y_{k-1}\right),\mathcal{G}_p\right) \\
&\xrightarrow{n\to\infty} \sup_{i=1,\dots,N_\eta}\left(-K\left(g^0,g_{c_i}\right)\right) + \mathbb{E}_\mu\left(m_\eta\left(\left(Y_k,Y_{k-1}\right),\mathcal{G}_p\right)\right)
\end{aligned}$$

On the one hand,

$$\sup_{i=1,\dots,N_\eta}\left(-K\left(g^0,g_{c_i}\right)\right) = -\inf_{i=1,\dots,N_\eta} K\left(g^0,g_{c_i}\right) \leq -K\left(g^0,\tilde{g}\right)$$

on the other hand, if $\eta\to 0$, $m_\eta\left(\left(y_1,y_2\right),\mathcal{G}_p\right)\to 0$ and

$$\mathbb{E}_\mu\left(m_\eta\left(\left(Y_k,Y_{k-1}\right),\mathcal{G}_p\right)\right)\to 0$$

Thus

$$\limsup_{n\to\infty}\frac{1}{n-1}\sup_{g\in\mathcal{G}_p}\left(l_n(g) - l_n\left(g^0\right)\right) \leq -K\left(g^0,\tilde{g}\right)$$

■

## 3. Application to linear autoregressive models with Gaussian noise

In this section we are interested in checking whether the assumptions (H1)-(H4) in subsection 2.4 hold in the case of a very popular autoregressive regime-switching model. We shall consider that the process $(X_t, Y_t)$ follows the true model

$$Y_t = a^0_{X_t} Y_{t-1} + b^0_{X_t} + \sigma^0_{X_t}\varepsilon_t \tag{4}$$

where

- $X_t$ is an i.i.d. sequence of random variables valued in a finite space $\{1,...,p_0\}$ and with probability distribution $\pi^0 := \left(\pi^0_1,\cdots,\pi^0_{p_0}\right)$
- for every $i\in\{1,...,p_0\}$, $a^0_i, b^0_i, \sigma^0_i$ are real numbers with $\left|a^0_i\right| < 1$ and $\sigma^0_i > 0$
- $\left(\varepsilon_t\right)_{t\in\mathbb{N}}$ is a sequence of i.i.d. standard Gaussian variables, independent of $\left(Y_{t-k}\right)_{k\geq 1}$.

This model is obviously a special case of the more general model in section 2.2.

The following result which ensures strict stationarity and ergodicity can be stated. The proof may be found in the Appendix.

***Proposition 1***:
*If $\left|a^0_i\right| < 1$ for every $i\in\{1,...,p_0\}$, $(X_t, Y_t)$ is strictly stationary, geometrically ergodic and, in particular, geometrically $\beta$-mixing. Moreover, there exists $\delta > 0$ such that $E_\mu\left(e^{\delta Y_t^2}\right) < \infty$.*

The set of possible conditional densities is the following :

$$\begin{aligned}
\mathcal{G} = \Bigg\{ & g \mid g\left(y_2 \mid y_1\right) = \textstyle\sum_{i=1}^{p}\pi_i \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{1}{2\sigma_i^2}\left(y_2-\left(a_i y_1 + b_i\right)\right)^2}, p = 1,...,P, \\
& \left(a_i, b_i, \sigma_i^2\right)\in\Theta\subset\mathbb{R}\times\mathbb{R}\times\mathbb{R}_+^\star, \pi_i\in[0,1], \textstyle\sum_{i=1}^{p}\pi_i = 1\Big\},
\end{aligned}$$

with $\Theta$ a compact set.

Within this framework, $\mathcal{G}$ is Glivenko-Cantelli and the assumption H-1 is true. Moreover, in a neighborhood of the true parameter $\theta_i^0$, the second and third derivatives of

$$l_{\theta_i}(y_1,y_2)=\frac{\sqrt{2\pi{\sigma_i^0}^2}}{\sqrt{2\pi\sigma_i^2}}e^{-\frac{1}{2\sigma_i^2}(y_2-(a_iy_1+b_i))^2+\frac{1}{2{\sigma_i^0}^2}\left(y_2-\left(a_i^0y_1+b_i^0\right)\right)^2}$$

exist and are dominated by a square integrable function, hence the assumption H-3 also holds.

Next, we check whether the generalized score functions are well defined (assumption H-2):

$$\left\|\frac{g}{g^0}-1\right\|_{L^2(\mu)}<\infty,\forall g \text{ such that } \|g-g^0\|\leq\eta.$$

**Conditions for the existence of the extended score functions**

Consider the true conditional distribution

$$g^0\left(y_2\mid y_1\right)=\sum_{j=1}^{p_0}\pi_j^0 f_{\theta_j^0}\left(y_2-F_{\theta_j^0}\left(y_1\right)\right)$$

and let the possible conditional distributions be

$$g\left(y_2\mid y_1\right)=\sum_{i=1}^{p}\pi_i f_{\theta_i}\left(y_2-F_{\theta_i}\left(y_1\right)\right).$$

One can prove by direct computations that

***Proposition 2*** (the proof is available in the Appendix)

$\left\|\frac{g}{g^0}-1\right\|_{L^2(\mu)}<\infty$ *if for every* $i\in\{1,...,p\}$*, there exists* $k\in\{1,...,p_0\}$*such that* $\sigma_i^2<2\left(\sigma_k^0\right)^2$ *and* $\left|a_i-a_k^0\right|<\sqrt{\delta\left(2\left(\sigma_k^0\right)^2-\sigma_i^2\right)}$ *for* $\delta>0$ *verifying* $E\left(e^{\delta Y_t^2}\right)<\infty$.

This sufficient condition states that the possible models should not be too different from the real one so that the convergence holds.

The consequences of this condition will be discussed later. For the moment, we will assume it is fulfilled.

Finally, let us check the assumption H-4:

***Lemma 1*** (the proof is straightforward and will be omitted)

The functions

$$\left\{g_{\theta_i}, i=1,\cdots,p,\frac{\partial g_{\theta_i^0}}{\partial a_i},\frac{\partial g_{\theta_i^0}}{\partial b_i},\frac{1}{\sigma_i^0}\frac{\partial g_{\theta_i^0}}{\partial\sigma_i}+\frac{\partial^2 g_{\theta_i^0}}{\partial b_i^2},\frac{\partial^2 g_{\theta_i^0}}{\partial a_i^2},\right.$$
$$\left.\frac{\partial^2 g_{\theta_i^0}}{\partial\sigma_i^2},\frac{\partial^2 g_{\theta_i^0}}{\partial a_i\partial\sigma_i},\frac{\partial^2 g_{\theta_i^0}}{\partial a_i\partial b_i},\frac{\partial^2 g_{\theta_i^0}}{\partial b_i\partial\sigma_i}, i=1,...,p_0\right\}$$

are linearly independent.

Hence, the assumptions H-1, H-2, H-3 and H-4 are fulfilled if the possible parameters of the regression function are not too far from the true ones. Since the true regression function is not known, it seems very difficult to assume such hypothesis. If the parameter set is not restricted, we will see in the next section that the LRTS will be divergent.

3.1. **Simple regression mixture example.** Let $\mathcal{G}$ be the set of possible conditional densities:

$$\mathcal{G} = \left\{ g\left(y_k \mid y_{k-1}\right) = \pi g_\theta\left(y_k \mid y_{k-1}\right) + (1-\pi) g^0\left(y_k \mid y_{k-1}\right), \pi \in [0;1], g_\theta \in \mathcal{P} \right\}$$

with $\mathcal{P} = \left\{ g_\theta(y_k|y_{k-1}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_k-\theta y_{k-1})^2}, \theta \in \Theta \subset \mathbb{R} \right\}$ the set of conditional densities and $g^0\left(y_k \mid y_{k-1}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y_k^2}$. This model is clearly a particular case of the general mixture of expert model and is a simple example of mixture of regressions with Gaussian noise. Let

$$l_n\left(g\right) = \textstyle\sum_{k=2}^n \ln\ g\left(y_k \mid y_{k-1}\right)$$

be the conditional log-likelihood function of $(y_1, \cdots, y_n)$. We want to know whether the true model is really a mixture regression model (i.e. $\theta \neq 0$ and $\pi \neq 0$) or the observations are independent ($\theta = 0$ or $\forall x, \pi = 0$). The LRTS is defined as:
(6)

$$2\lambda_n = 2\left(\sup_{g\in\mathcal{G}} \ln(g) - \ln(g^0)\right) = 2\sup_{g\in\mathcal{G}} \sum_{k=1}^n \ln \frac{\pi g_\theta\left(y_k \mid y_{k-1}\right) + (1-\pi) g^0\left(y_k \mid y_{k-1}\right)}{g^0\left(y_k \mid y_{k-1}\right)}$$

In order to derive the behaviour of the LRTS, two cases have to be analyzed. The first one is if $\pi$ can be close to 0. The second one is when $\exists\delta > 0$ such that $\pi \geq \delta$.

**Divergence of LRTS.** The LRTS can be divergent if $\pi$ is not constraint. Indeed, for such sequence we can have $E_\mu\left(\ln(g) - \ln(g^0)\right) \to 0$ with $\theta \neq 0$. The score functions are well defined if the quantity
$\left\| \frac{g_\theta(Y_2|Y_1)}{g^0(Y_2|Y_1)} - 1 \right\|_{L^2(\mu)} = \left\| \exp\left(-\frac{\theta^2}{2}Y_1^2 + \theta Y_2 Y_1\right) - 1 \right\|_{L^2(\mu)}$ is finite. So,

$$\begin{array}{l}
\left\| \exp\left(-\frac{\theta^2}{2}Y_1^2 + \theta Y_2 Y_1\right) - 1 \right\|_{L^2(\mu)}^2 = \\
\frac{1}{2\pi} \int\int \left(\exp\left(-\frac{\theta^2}{2}y_1^2 + \theta y_2 y_1\right) - 1\right)^2 \exp\left(-\frac{1}{2}y_1^2\right) \exp\left(-\frac{1}{2}y_2^2\right) dy_1 dy_2 = \\
\frac{1}{2\pi} \int\int \left(\exp\left(-\theta^2 y_1^2 + 2\theta y_2 y_1\right) - 2\exp\left(-\frac{\theta^2}{2}y_1^2 + \theta y_2 y_1\right) + 1\right) \\
\exp\left(-\frac{1}{2}y_1^2\right) \exp\left(-\frac{1}{2}y_2^2\right) dy_1 dy_2
\end{array}$$

The integral of the dominant term (the first) is:

$$\begin{array}{rcl}
I\left(\theta\right) & = & \frac{1}{2\pi} \int\int \exp\left(-\theta^2 y_1^2 + 2\theta y_2 y_1\right) \exp\left(-\frac{1}{2}y_1^2\right) \exp\left(-\frac{1}{2}y_2^2\right) dy_1 dy_2 \\
& = & \frac{1}{2\pi} \int\int \exp\left(-\left(\theta^2 + \frac{1}{2}\right) y_1^2 + 2\theta y_1 y_2 - \frac{1}{2}y_2^2\right) dy_1 dy_2 \\
& = & \frac{1}{2\pi} \int\int \exp\left(-\left(\sqrt{\theta^2 + \frac{1}{2}}y_1 - \frac{\theta}{\sqrt{\theta^2+\frac{1}{2}}}y_2\right)^2 - \left(\frac{1}{2} - \frac{\theta^2}{\theta^2+\frac{1}{2}}\right) y_2^2\right) dy_1 dy_2 \\
& = & \frac{\sqrt{2\theta^2+1}}{\sqrt{2\pi}} \int \exp\left(-\left(\frac{1}{2} - \frac{\theta^2}{\theta^2+\frac{1}{2}}\right)_2^2\right) dy_2
\end{array}$$

Finally for $-\frac{1}{\sqrt{2}} < \theta < \frac{1}{\sqrt{2}}$,

$$\left\| \exp\left(-\frac{\theta^2}{2}Y_1^2 + \theta Y_2 Y_1\right) - 1 \right\|_{L^2(\mu)} < +\infty$$

and the score function is well defined.

Note that the distribution of the LRTS $2\lambda_n$ for a finite number of possible parameters $\theta_1, \cdots, \theta_m$ will always converge to the square of a $m$-dimensional normal distribution with covariance $\left(E\left(\nu_{\theta_i}(Y_1, Y_2)\nu_{\theta_j}(Y_1, Y_2)\right)\right)_{1\leq i,j\leq m}$. Suppose that an arbitrary number of"almost" uncorrelated random variables can be found, then $\lambda_n$ can take an arbitrarily large value since the maximum of $m$ independent samples from standard normal distribution is approximately $\sqrt{2\log m}$. Hence, Fukumizu (2003) has shown that if a sequence $\theta_1, \cdots, \theta_m, \cdots$ exists so that

$$\lim_{m\to\infty} \nu_{\theta_m}(Y_1, Y_2) \xrightarrow{P} 0$$

then the likelihood ratio $T_n$ diverges to infinite. Here, we get

$$\lim_{\theta\to\frac{1}{\sqrt{2}}, \theta<\frac{1}{\sqrt{2}}} \left\| \exp\left(-\frac{\theta^2}{2}Y_1^2 + \theta Y_2 Y_1\right) - 1\right\|_{L^2(\mu)} = +\infty$$

So, for each sphere $B$ of $\mathbb{R}^2$, centered on the origin, if $(Y_1, Y_2) \in B$:

$$\lim_{\theta\to\frac{1}{\sqrt{2}}, \theta<\frac{1}{\sqrt{2}}} \frac{\exp\left(-\frac{\theta^2}{2}Y_1^2+\theta Y_2 Y_1\right)-1}{\left\|\exp\left(-\frac{\theta^2}{2}Y_1^2+\theta Y_2 Y_1\right)-1\right\|_{L^2(\mu)}} = 0$$

and $\frac{\exp\left(-\frac{\theta^2}{2}Y_1^2+\theta Y_2 Y_1\right)-1}{\left\|\exp\left(-\frac{\theta^2}{2}Y_1^2+\theta Y_2 Y_1\right)-1\right\|_{L^2(\mu)}}$ converges to 0 in probability for $\theta \to \frac{1}{\sqrt{2}}, \theta < \frac{1}{\sqrt{2}}$. With the choice $\theta_m = \frac{1}{\sqrt{2}} - \frac{1}{m}$, we get $\lim_{m\to\infty} \nu_{\theta_m}(Y_1, Y_2) \xrightarrow{P} 0$ and the LRTS is divergent.

**Convergence of LRTS.** If $\pi$ is greater or equal than a $\delta > 0$ then, necessary, the maximum likelihood estimator $\hat{\theta}$ converges to $\theta_0 = 0$, otherwise $\lim_{n\to\infty}\lambda_n = \sup_{g\in\mathcal{G}} E_\mu\left(\ln(g) - \ln(g^0)\right)$ can not be close to 0. Thus, the model is identifiable in $\theta$ and unidentifiable in $\pi$. Since $\frac{\partial}{\partial\theta}g_\theta(y_2 \mid y_1) = y_1(y_2 - \theta y_1)g_\theta$, we have the following Taylor expansion around $\theta_0 = 0$:

$$\begin{aligned} s_g &= \frac{\frac{g_\theta}{g^0}-1}{\|\frac{g_\theta}{g^0}-1\|_{L^2}} \\ &= \frac{(\theta-\theta_0)\frac{\partial}{\partial\theta}\frac{g_\theta}{g^0}(\theta_0)+o(|\theta-\theta_0|)}{\|(\theta-\theta_0)\frac{\partial}{\partial\theta}\frac{g_\theta}{g^0}(\theta_0)+o(|\theta-\theta_0|)\|_{L^2}} \\ &= \frac{Y_1Y_2+o(1)}{\|Y_1Y_2+o(1)\|_{L^2}} \end{aligned}$$

Hence, the LRTS converges to the square of the maximum of a Gaussian process which convariance function is identically equal to 1 i.e. the classical $\chi^2$ law.

In conclusion, if the mixture weights can be as small as possible, the likelihood ratio tends to infinity and in order to avoid this divergence, it is required to constraint the parameters in a neighborhood of the true walue, which does not make much sense if the model is unknown. But if the mixture weights are bounded by below, then all parameters of regression density $g_\theta$ converge to some true one. It is thus possible, in the analysis of the asymptotics of the criterion, to restrict the set of scores as in proposition 2 and then apply theorem 2 with restricted set of score functions as well as corollary 1.

## 4. Generalization to autoregressive Markov-switching models?

The aims of the section is to study the generalization of the previous results to Markov switching models.

Let us consider the more general case where the process $(X_t, Y_t)$ follows the true model

$$Y_t = F_{\theta^0_{X_t}}(Y_{t-1}) + \varepsilon_{\theta^0_{X_t}}(t) \tag{5}$$

where

- $X_t$ is a homogeneous Markov chain, irreducible and aperiodic, with finite state-space $\{1, ..., p_0\}$ and stationary probability distribution $\pi^0 := (\pi^0_1, ..., \pi^0_{p_0})$
- for every $i \in \{1, ..., p_0\}$, $F_{\theta^0_i}(y) \in \mathcal{F}$, where $\mathcal{F} = \{F_\theta,\, \theta \in \Theta,\, \Theta \subset \mathbb{R}^d \,\text{compact set}\}$ is the family of possible regression functions. We suppose throughout the rest of this section that $F_{\theta^0_i}$ are sublinear, that is they are continuous and $\exists \left(a^0_i, b^0_i\right) \in \mathbb{R}^2_+$ such that $\left|F_{\theta^0_i}(y)\right| \le a^0_i |y| + b^0_i$, $(\forall)\ y \in \mathbb{R}$ ;
- for every $i \in \{1, ..., p_0\}$, $(\varepsilon_{\theta_i}(t))_t$ is an i.i.d. noise so that $\varepsilon_{\theta_i}(t)$ is independent of $(Y_{t-k})_{k\ge1}$. Moreover, $\varepsilon_{\theta^0_i}(t)$ has density $g^0_i \in \mathcal{P}$ , where $\mathcal{P} = \{g_\theta,\, \theta \in \Theta,\, \Theta \subset \mathbb{R}^l\}$ is a family of strictly positive densities with respect to the Lebesgue measure.

According to Yao and Attali (2000), a unique strictly-stationary and geometrically-ergodic solution $(X_t, Y_t)$ exists under the hypothesis

**(HS)** $(\exists)$ $s \ge 1$ so that $\forall i \in \{1, \cdots, p_0\}$, $E\left|\varepsilon_{\theta^0_i}\right|^s < \infty$ and the spectral radius $\rho(Q_s) < 1$, with

$$Q_s = \begin{pmatrix} \left(a^0_1\right)^s \pi^0_{11} & \cdots & \left(a^0_{p_0}\right)^s \pi^0_{1p_0} \\ \vdots & \ddots & \vdots \\ \left(a^0_1\right)^s \pi^0_{p_01} & \cdots & \left(a^0_{p_0}\right)^s \pi^0_{p_0p_0} \end{pmatrix}$$

where $a^0_i$ are the leading coefficients in the linear functions dominating $F_{\theta^0_i}$ and $\pi^0_{ij}$ are the entries of the transition matrix of $X_t$, $i, j \in \{1, ..., p_0\}$. The hypothesis (HS) is clearly verified whenever $a^0_i < 1$, for all $i \in \{1, ..., p_0\}$.

Considering an observed $n$-sample of $Y_t$, one would naturally attempt to extend the methods in the previous sections to the case where the invariant measure of the hidden Markov chain is lower bounded by a strictly positive constant. Several problems arise: on the one hand, the non-identifiability issue and on the other hand, the dependence structure of $X_t$. This dependence will not allow an explicit form for the conditional density, marginally in $X_t$:

$$g^0(y_k \mid y_{k-1}, ..., y_0) = \sum_{i=1}^{p_0} \mathbb{P}(X_k = i \mid y_{k-1}, ..., y_0)\, g_{\theta^0_i}\left(y_k - F_{\theta^0_i}(y_{k-1})\right)$$

since $\mathbb{P}(X_k = i \mid y_{k-1}, ..., y_0)$ has to be computed recursively. However, since $X_t$ is stationary and following the same idea as Gassiat (2002), a cost function which involves the invariant probability measure of the hidden Markov chain can be defined.

The class of possible mixture densities is:

$$\mathcal{G}=\left\{g \mid g\left(y_1, y_2\right)=\sum_{i=1}^{p} \pi_i g_{\theta_i}\left(y_2-F_{\theta_i}\left(y_1\right)\right), \theta_i \in \Theta\right\}$$

where $\Theta$ is a compact set.

The cost function is defined as

$$C_n(g)=\frac{1}{n} \sum_{k=2}^{n} \ln\, g\left(y_k \mid y_{k-1}\right)=\frac{1}{n-1} \sum_{k=2}^{n} \ln\left(\sum_{i=1}^{p} \pi_i g_{\theta_i}\left(y_2-F_{\theta_i}\left(y_1\right)\right)\right).$$

One may notice that $C_n(g)$ is similar to the conditional likelihood marginal in $X_t$ and may expect it to be maximized by $g=g^0$, where "the true conditional density" is now written as

$$g^0\left(y_k \mid y_{k-1}\right)=\sum_{i=1}^{p_0} \pi_i^0 g_{\theta_i^0}\left(y_k-F_{\theta_i^0}\left(y_{k-1}\right)\right),$$

where $\pi_i^0$ is the expectation of the hidden state $i$ under the true invariant distribution.

Let us check if $C_n(g)$ is a contrast function with the maximum reached at $g^0$. Let $(X, Y_2, Y_1)$ be a generic variable having the stationary measure of the extended Markov-chain $(X_k, Y_k, Y_{k-1})$ as distribution. Since $C_n(g)$ is an additive function of the Markov chain $(X_k, Y_k, Y_{k-1})_{1 \leq k \leq n}$ and $\sum_{i=1}^{p_0} 1_{\{X=i\}}(X)=1$, we have

$$\begin{aligned}
&C_n(g)=\tfrac{1}{n} \sum_{k=2}^{n} \ln\left(\sum_{j=1}^{p} \pi_j g_{\theta_j}\left(y_k-F_{\theta_j}\left(y_{k-1}\right)\right)\right) \stackrel{a.s.}{\rightarrow} \\
&E\left(\ln\left(\sum_{j=1}^{p} \pi_j g_{\theta_j}\left(Y_2-F_{\theta_j}\left(Y_1\right)\right)\right)\right)= \\
&E\left(\sum_{i=1}^{p_0} 1_{\{X=i\}}(X) \ln\left(\sum_{j=1}^{p} \pi_j g_{\theta_j}\left(Y_2-F_{\theta_j}\left(Y_1\right)\right)\right)\right)= \\
&\sum_{i=1}^{p_0} \pi_i^0 \int_{\mathbb{R}^2} \ln\left(\sum_{j=1}^{p} \pi_j g_{\theta_j}\left(y_2-F_{\theta_j}\left(y_1\right)\right)\right) g_{\theta_i^0}\left(y_2-F_{\theta_i^0}\left(y_1\right)\right) \lambda_i\left(y_1\right) dy_1 dy_2
\end{aligned}$$

where $\lambda_i(y_1)$ is the stationary measure of $Y_1$ conditionally to $X=i$.

Then,

$$\begin{aligned}
&E\left[\ln(g)-\ln(g^0)\right]=\sum_{i=1}^{p_0} \mathbb{P}(X=i)\, E\left[\ln \tfrac{g}{g^0} \mid X=i\right]= \\
&=\sum_{i=1}^{p_0} \pi_i^0 \int \ln\left(\frac{\sum_{j=1}^{p} \pi_j g_{\theta_j}\left(y_2-F_{\theta_j}\left(y_1\right)\right)}{\sum_{j=1}^{p_0} \pi_j^0 g_{\theta_j}^0\left(y_2-F_{\theta_j}^0\left(y_1\right)\right)}\right) g_{\theta_i^0}\left(y_2-F_{\theta_i^0}\left(y_1\right)\right) \lambda_i\left(y_1\right) dy_1 dy_2
\end{aligned}$$

and, by Fubini's theorem,

$$\begin{aligned}
&E\left[\ln(g)-\ln(g^0)\right]= \\
&=\int \ln\left(\frac{\sum_{j=1}^{p} \pi_j f_{\theta_j}\left(y_2-F_{\theta_j}\left(y_1\right)\right)}{\sum_{j=1}^{p_0} \pi_j^0 f_{\theta_j}^0\left(y_2-F_{\theta_j}^0\left(y_1\right)\right)}\right) \sum_{i=1}^{p_0} \pi_i^0 f_{\theta_i^0}\left(y_2-F_{\theta_i^0}\left(y_1\right)\right) \lambda_i\left(y_1\right) dy_1 dy_2
\end{aligned}$$

The last term can be proven immediately to be negative in either of the following cases:

- $\lambda_i(y_1)=\lambda(y_1)$ for all $i \in \{1, ..., p_0\}$ which leads to autoregressive mixture models already considered in Section 2.

- $F_{\theta_j}(y_1)$ and $F_{\theta_i^0}(y_1)$ are constant functions for $j \in \{1, ..., p\}$, $i \in \{1, ..., p_0\}$, but this corresponds to hidden Markov chains already studied in Gassiat (2002). Note that Theorem 2 applies in this case and it gives the probability distribution of the marginal likelihood ratio which is exactly the same as for mixture models (see Liu and Shao, 2003).

In the general case, however, there is no reason for the last integral to be negative. Some simulation results are presented for illustrating this last assertion.

**Simulation results**

Several two-regime models were considered, with transition matrices:
$M_1 = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$, $M_2 = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}$ and $M_3 = \begin{pmatrix} 0.9 & 0.5 \\ 0.1 & 0.5 \end{pmatrix}$. The first transition matrix corresponds to independent regime switches. The regression functions are either linear, or constant. The latter cases correspond to hidden Markov chains. The noise was considered normally distributed $N\left(0, (0.5)^2\right)$ and the likelihood was penalized according to the BIC criterion. For every model, several sample sizes were considered (from 200 up to 2000 input values) and for each model and sample size, twenty different samples were simulated. In each case, Tables 1 and 2 contain the estimated number of regimes (the maximum was fixed at three).

Simulation results prove that the penalized estimate $\hat{p}$ diverges when the true model is, for instance, a two-regime autoregressive Markov-switching model. This means that the cost function that was considered as a generalization of the "marginal likelihood" does not have the right properties to be a contrast function and the problem of estimating $p_0$ remains open in the general case of autoregressive Markov switching models.

| | | $M_1$ | | | $M_2$ | | | $M_3$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $n$ | $\hat{p}=1$ | $\hat{p}=2$ | $\hat{p}=3$ | $\hat{p}=1$ | $\hat{p}=2$ | $\hat{p}=3$ | $\hat{p}=1$ | $\hat{p}=2$ | $\hat{p}=3$ |
| $F_1^0(y) = 0.8y - 1$ | 200 | 0 | 20 | 0 | **0** | **15** | **5** | **0** | **17** | **3** |
| $F_2^0(y) = 0.3y + 1$ | 500 | 0 | 20 | 0 | **0** | **17** | **3** | **0** | **8** | **12** |
| | 1000 | 0 | 20 | 0 | **0** | **6** | **14** | **0** | **4** | **16** |
| | 1500 | 0 | 20 | 0 | **0** | **1** | **19** | **0** | **5** | **15** |
| | 2000 | 0 | 20 | 0 | **0** | **1** | **19** | **0** | **5** | **15** |
| $F_1^0(y) = -1$ | 200 | 0 | 20 | 0 | 0 | 20 | 0 | 0 | 20 | 0 |
| $F_2^0(y) = 1$ | 500 | 0 | 20 | 0 | 0 | 20 | 0 | 0 | 20 | 0 |
| | 1000 | 0 | 20 | 0 | 0 | 20 | 0 | 0 | 20 | 0 |
| | 1500 | 0 | 20 | 0 | 0 | 20 | 0 | 0 | 20 | 0 |
| | 2000 | 0 | 20 | 0 | 0 | 20 | 0 | 0 | 20 | 0 |

TABLE 1. Results for the "marginal-loglikelihood" BIC-penalized cost-function

| | | $M_2$ | | | $M_3$ | | |
|---|---|---|---|---|---|---|---|
| | $n$ | $\hat{p}=1$ | $\hat{p}=2$ | $\hat{p}=3$ | $\hat{p}=1$ | $\hat{p}=2$ | $\hat{p}=3$ |
| $F_1^0(y)=0.8y-1$ | 200 | 0 | 16 | 4 | 0 | 15 | 5 |
| $F_2^0(y)=0.3y+1$ | 500 | 0 | 16 | 4 | 0 | 19 | 1 |
| | 1000 | 0 | 17 | 3 | 0 | 19 | 1 |
| | 1500 | 0 | 18 | 2 | 0 | 19 | 1 |
| | 2000 | 0 | 19 | 1 | 0 | 20 | 0 |

TABLE 2. Results for the exact-loglikelihood BIC-penalized cost-function

## Appendix

4.1. **Proof of theorem 1.** Denote by $\hat{g}_n$ the functions $g$ maximizing the likelihood. Since the set $\mathcal{G}$ is Glivenko-Cantelli, for all $\eta > 0$ and for $n$ large enough: $\hat{g}_n \in \mathcal{G}_\eta$, where $\mathcal{G}_\eta$ is defined by equation (3). Now, using Theorem 1 of Doukhan (1995), under **(B)**

$$\sup_{s\in\mathcal{S}_\eta} \frac{1}{n-1}\left(\sum_{k=2}^{n} s\left(Y_{k-1}, Y_k\right)\right)^2 = \mathcal{O}_{\mathbb{P}}\left(1\right) \tag{7}$$

with $\mathcal{S}_\eta$ defined by equation (4). Moreover, $\mathcal{S}_\eta \subset \mathcal{L}_2\left(\mu\right)$, thus $\mathcal{S}_\eta^2 \subset \mathcal{L}_1\left(\mu\right)$ and using the $\mathcal{L}_2$-entropy condition $\mathcal{S}_-^2 = \left\{(s)_-^2,\ g \in \mathcal{G}_\eta\right\}$, with $(s)_-\left(y_{k-1}, y_k\right) = \min\left(0,\ s\left(y_{k-1}, y_k\right)\right)$, is Glivenko-Cantelli. Since $\left(Y_{k-1}, Y_k\right)$ is ergodic and strictly stationary, we obtain the following uniform convergence in probability:

$$\inf_{s\in\mathcal{S}_\eta} \frac{1}{n-1}\sum_{k=2}^{n} (s)_-^2\left(Y_{k-1}, Y_k\right) \longrightarrow_{n\to\infty} \inf_{s\in\mathcal{S}_\eta} \left\|(s)_-\right\|_2^2$$

.

The following lemma is a straightforward adaptation of the inequality 1.1 in Gassiat (2002).

***Lemma 2.***

Under **(B)**

$$\sup_{g\in\mathcal{G}_\eta:\ln(g)-\ln(g^0)\geq 0} \|\frac{g-g^0}{g^0}\|_2 \leq 2\sup_{g\in\mathcal{G}_\eta} \frac{\sum_{t=2}^{n} s_g\left(Y_{t-1}, Y_t\right)}{\sum_{t=2}^{n}(s_g)_-^2\left(Y_{t-1}, Y_t\right)} \tag{8}$$

One may apply this inequality to obtain

$$\sup_{g\in\mathcal{G}_\eta:\ln(g)-\ln(g^0)\geq 0} \|\frac{g-g^0}{g^0}\|_2 = O_{\mathbb{P}}\left(n^{-1/2}\right) \tag{9}$$

Taylor expansion gives that $\ln(1+u) = u - \frac{u^2}{2} + u^2R(u)$, with $\lim_{u\leftarrow 0} R(u) = 0$. Thus, for any $g$,

$$\ln(g) - \ln(g^0) = \|\tfrac{g-g^0}{g^0}\|_2 \textstyle\sum_{t=2}^{n} s_g\left(Y_{t-1}, Y_t\right) - \frac{1}{2}\|\frac{g-g^0}{g^0}\|_2^2 \sum_{t=2}^{n}\left(s_g\left(Y_{t-1}, Y_t\right)\right)^2$$
$$+\|\tfrac{g-g^0}{g^0}\|_2^2 \textstyle\sum_{t=2}^{n}\left(s_g\left(Y_{t-1}, Y_t\right)\right)^2 R\left(\|\frac{g-g^0}{g^0}\|_2 \sum_{t=2}^{n} s_g\left(Y_{t-1}, Y_t\right)\right)$$

By **(B)**, $\frac{1}{n-1}\sum_{t=2}^{n}\left(s_g\left(Y_{t-1}, Y_t\right)\right)^2 = O_{\mathbb{P}}(1)$.

Now, we have the following lemma:

***Lemma 3***

Let $(F(X_1), \cdots, F(X_n))$ be stationnary sequence of real random variables in $\mathbb{L}^2$ then

$$\max_{i\in\{1,\cdots,n\}}(F(X_i)) = o_P(\sqrt{n})$$

***Proof of lemma 3***

Let us show that

$$\forall\varepsilon > 0, \lim_{n\to\infty} P(|\max_{i\in\{1,\cdots,n\}}(F(X_i))| > \varepsilon\sqrt{n}) = 0$$

We have

$$P(|\max_{i\in\{1,\cdots,n\}}(F(X_i))| > \varepsilon\sqrt{n}) \leq$$
$$P\left(\{|F(X_1)| > \varepsilon\sqrt{n}\} \cup \cdots \cup \{|F(X_n)| > \varepsilon\sqrt{n}\}\right) \leq nP\left(|F(X_1)| > \varepsilon\sqrt{n}\right)$$

Now, since $F(X_1) \in \mathbb{L}^2$,

$$\lim_{n\to\infty} \int_{\varepsilon\sqrt{n}}^{\infty} F(x)^2 dP(x) + \int_{-\infty}^{\varepsilon\sqrt{n}} F(x)^2 dP(x) = 0$$

Hence

$$\lim_{n\to\infty} n \times P(|F(X_1)| > \varepsilon\sqrt{n}) \leq \lim_{n\to\infty} \frac{1}{\varepsilon^2} \left( \int_{\varepsilon\sqrt{n}}^{\infty} F(x)^2 dP(x) + \int_{-\infty}^{\varepsilon\sqrt{n}} F(x)^2 dP(x) \right) = 0$$

■

Furthermore, since $\mathcal{S}_\eta$ admits a square integrable envelop function $F$ and using (9) we have:

$$\sup_{g\in\mathcal{G}_\eta:\ln(g)-\ln(g^0)\geq 0} \| \frac{g-g^0}{g^0} \|_2^2 \sum_{t=2}^{n} \left(s_g\left(Y_{t-1}, Y_t\right)\right)^2 R\left( \| \frac{g-g^0}{g^0} \|_2 \sum_{t=2}^{n} s_g\left(Y_{t-1}, Y_t\right) \right) = o_{\mathbb{P}}(1).$$

Thus,

$$\begin{aligned}&\sup_{g\in\mathcal{G}_\eta} \left(\ln(g) - \ln(g^0)\right) = \\ &\sup_{g\in\mathcal{G}_\eta} \left\{ \| \tfrac{g-g^0}{g^0} \|_2 \sum_{t=2}^n s_g\left(Y_{t-1}, Y_t\right) - \tfrac{1}{2} \| \tfrac{g-g^0}{g^0} \|_2^2 \sum_{t=2}^n \left(s_g\left(Y_{t-1}, Y_t\right)\right)^2 \right\} + o_{\mathbb{P}}(1),\end{aligned}$$

which implies that

$$\sup_{g\in\mathcal{G}_\eta} \left(\ln(g) - \ln(g^0)\right) \leq \sup_{g\in\mathcal{G}_\eta:\ln(g)-\ln(g^0)\geq 0} \frac{\left( \max\left\{ \frac{\sum_{t=2}^n s_g(Y_{t-1}, Y_t)}{\sqrt{n}}; 0 \right\} \right)^2}{\frac{\sum_{t=2}^n (s_g(Y_{t-1}, Y_t))^2}{n}} + o_{\mathbb{P}}(1).$$

Since ${\mathcal{S}_\eta}^2$ is Glivenko-Cantelli:

$$\sup_{g\in\mathcal{G}_\eta} \left| \frac{\sum_{t=2}^n \left(s_g\left(Y_{t-1}, Y_t\right)\right)^2}{n} - 1 \right| = o_{\mathbb{P}}(1),$$

and

$$2 \sup_{g\in\mathcal{G}_\eta} \left(\ln(g) - \ln(g^0)\right) \leq \sup_{g\in\mathcal{G}_\eta:\ln(g)-\ln(g^0)\geq 0} \left( \max\left\{ \frac{\sum_{t=2}^n s_g\left(Y_{t-1}, Y_t\right)}{\sqrt{n}}; 0 \right\} \right)^2 + o_{\mathbb{P}}(1).$$

Let $\mathcal{G}_{\eta_n} = \left\{ g \in \mathcal{G}_\eta : \| \frac{g-g^0}{g^0} \| \leq n^{-1/4} \right\}$. Using (9), we obtain that

$$2 \sup_{g\in\mathcal{G}_\eta} \left(\ln(g) - \ln(g^0)\right) \leq \sup_{g\in\mathcal{G}_{\eta_n}} \left( \max\left\{ \frac{\sum_{t=2}^n s_g\left(Y_{t-1}, Y_t\right)}{\sqrt{n}}; 0 \right\} \right)^2 + o_{\mathbb{P}}(1).$$

Now, $\sup_{g\in\mathcal{G}_{\eta_n}} \|s_g - \mathcal{D}\|_2 \xrightarrow[n\to\infty]{} 0$, thus for a sequence $u_n$ decreasing to 0, and with

$$\Delta_n = \left\{ s_g - d \ : \ g \in \mathcal{G}_{\eta_n},\ d \in \mathcal{D}, \|s_g - d\|_2 \leq u_n \right\},$$

we obtain that

$$\begin{aligned}&2 \sup_{g\in\mathcal{G}_\eta} \left(\ln(g) - \ln(g^0)\right) \leq \\ &\left( \max\left\{ \sup_{d\in\mathcal{D}} \tfrac{\sum_{t=2}^n d(Y_{t-1}, Y_t)}{\sqrt{n}} + \sup_{\delta\in\Delta_n} \tfrac{\sum_{t=2}^n \delta(Y_{t-1}, Y_t)}{\sqrt{n}}; 0 \right\} \right)^2 + o_{\mathbb{P}}(1).\end{aligned}$$

Under **(B)**, thanks to Theorem 3 of Doukhan (1995) the empirical process indexed by $\mathcal{S}_\eta$ has the property of asymptotic stochastic equicontinuity, so:

$$\sup_{\delta\in\Delta_n} \frac{\sum_{t=2}^n \delta\left(Y_{t-1}, Y_t\right)}{\sqrt{n}} = o_{\mathbb{P}}(1),$$

and

$$2 \sup_{g\in\mathcal{G}_\eta} \left(\ln(g) - \ln(g^0)\right) \leq \sup_{d\in\mathcal{D}} \left( \max\left\{ \frac{\sum_{t=2}^n d\left(Y_{t-1}, Y_t\right)}{\sqrt{n}}; 0 \right\} \right)^2 + o_{\mathbb{P}}(1).$$

Moreover, using classical normal asymptotic properties along the parametric paths, one obtains that, for a sequence of finite subsets $\mathcal{D}_k$ increasing to $\mathcal{D}$,

$$2 \sup_{g \in \mathcal{G}_\eta} (\ln(g) - \ln(f)) \geq \sup_{d \in \mathcal{D}_k} \left( \max \left\{ \frac{\sum_{t=2}^n d(Y_{t-1}, Y_t)}{\sqrt{n}}; 0 \right\} \right)^2 + o_{\mathbb{P}}(1).$$

for any $k$. Therefore, Theorem 1 is true. ■

4.2. **Proof of Theorem 2.** Let $\eta > 0$ be a real number. Consider $\hat{\mathcal{G}}_n \neq \emptyset$ the set of functions which maximize the log-likelihood. Since, under H-1, $\mathcal{G}$ is Glivenko-Cantelli, for $n$ large enough, $\|g - g^0\|_{L^2(\mu)} < \eta$ for $g \in \hat{\mathcal{G}}_n$ so $\hat{\mathcal{G}}_n \subset \mathcal{G}_\eta$. Let us remark that, under assumption H-2, the score function $s_g \in \mathcal{S}_\eta$ is well defined in a compact neighborhood of the true density function $g^0$.

Proving that for an $\eta > 0$, a parametric family like $\mathcal{S}_\eta$ is Donsker is not so easy. The problems arise when $g \to g^0$ and the limits of $s_g$ in $L^2(\mu)$ have to be computed. To achieve our proof, let us split $\mathcal{S}$ into two classes of functions.

For a sufficiently small $\varepsilon > 0$, we consider $\mathcal{F}_0 \subset \mathcal{G}_\eta$, a neighborhood of $g^0$,

$\mathcal{F}_0 = \left\{ g \in G, \left\| \frac{g}{g^0} - 1 \right\|_{L^2(\mu)} \leq \varepsilon, \, g \neq g^0 \right\}$. $\mathcal{S}$ is splitted into $\mathcal{S}_0 = \{s_g, \, g \in \mathcal{F}_0\}$ and $\mathcal{S}_\eta \setminus \mathcal{S}_0$.

On $\mathcal{S}_\eta \setminus \mathcal{S}_0$, it can be easily seen that

$$\left\| \frac{\frac{g_1}{g^0} - 1}{\left\| \frac{g_1}{g^0} - 1 \right\|_{L^2(\mu)}} - \frac{\frac{g_2}{g^0} - 1}{\left\| \frac{g_2}{g^0} - 1 \right\|_{L^2(\mu)}} \right\|_{L^2(\mu)} \leq 2 \frac{\left\| \frac{g_1}{g^0} - \frac{g_2}{g^0} \right\|_{L^2(\mu)}}{\left\| \frac{g_1}{g^0} - 1 \right\|_{L^2(\mu)}}$$

for every $g_1, g_2 \in \mathcal{G}_\eta \setminus \mathcal{F}_0$ and, moreover, by the definition of $\mathcal{S}_0$,

$$\left\| \frac{\frac{g_1}{g^0} - 1}{\left\| \frac{g_1}{g^0} - 1 \right\|_{L^2(\mu)}} - \frac{\frac{g_2}{g^0} - 1}{\left\| \frac{g_2}{g^0} - 1 \right\|_{L^2(\mu)}} \right\|_{L^2(\mu)} \leq \frac{2}{\varepsilon} \left\| \frac{g_1}{g^0} - \frac{g_2}{g^0} \right\|_{L^2(\mu)}$$

On the other hand, by the assumption **H-3**, $\frac{g}{g^0}$ has square-integrable partial-derivatives of order one and, using the result 19.7 on parametric classes of functions in Van der Vaart (2000), we get:

$$\mathcal{N}_{[\cdot]}\left(\varepsilon, \mathcal{S} \setminus \mathcal{S}_0, \|\cdot\|_2\right) = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)^D,$$

where $D$ is the number of parameters in the model.

It remains to prove that the bracketing number is a polynom of $(\frac{1}{\varepsilon})$ for $\mathcal{S}_0$. The idea is to reparameterize the model in a convenient manner which will allow a Taylor expansion around the identifiable part of the true value of the parameters.

Let us recall that it is assumed that $p_0 < p$.

When $\frac{g}{g^0} - 1 = 0$, by the linear independence of the functions $g_{\theta_j}$, a vector of positive integers $t = (t_i)_{0 \leq i \leq p_0}$, $t_0 = 0$ exists so that:

$$\theta_{t_{i-1}+1} = ... = \theta_{t_i} = \theta_i^0, \quad \sum_{j=t_{i-1}+1}^{t_i} \pi_j = \pi_i^0, \; i \in \{1, ..., p_0\}$$

With this remark, one can define in the general case $s = (s_i)_{1\leq i\leq p_0}$ and $q = (q_j)_{1\leq j\leq p}$ so that, for every $i \in \{1, ..., p_0\}$ , $j \in \{t_{i-1}+1, ..., t_i\}$,

$$s_i = \sum_{j=t_{i-1}+1}^{t_i} \pi_j - \pi_i^0,\ q_j = \frac{\pi_j}{\sum_{l=t_{i-1}+1}^{t_i} \pi_l}$$

and a new parameterization will be

$$\Theta_t = (\phi_t, \psi_t)\,,\ \phi_t = \Big((\theta_j)_{1\leq j\leq t_{p_0}}\,, (s_i)_{1\leq i\leq p_0-1}\,, (\pi_j)_{j=t_{p_0}+1}^{p}\Big)\,,$$
$$\psi_t = \Big((q_j)_{1\leq j\leq p}\,, (\theta_j)_{j=t_{p_0}+1}^{p}\Big)$$

with $\phi_t$ containing all the identifiable parameters of the model and $\psi_t$ the non-identifiable ones. Then, for $g = g^0$, we will have:

$$\phi_t^0 = (\underbrace{\theta_1^0, ..., \theta_1^0}_{t_1}\ , ...,\ \underbrace{\theta_{p_0}^0, ..., \theta_{p_0}^0}_{t_{p_0}-t_{p_0-1}},\ \underbrace{0, ..., 0}_{p_0-1}\ \underbrace{0, ..., 0}_{p-t_{p_0}}\ )^T$$

This reparameterization allows to write a second-order Taylor expansion of $\frac{g}{g^0} - 1$ at $\phi_t^0$.

With the notations introduced in assumptions **H**, the density ratio becomes:

$$\frac{g}{g^0} - 1 = \sum_{i=1}^{p_0} \left(s_i + \pi_i^0\right) \sum_{j=t_{i-1}+1}^{t_i} q_j l_{\theta_j} + \sum_{j=t_{p_0}+1}^{p} \pi_j l_{\theta_j} - 1$$

and since $s_{p_0} = -\sum_{i=1}^{p_0-1} s_i$,

$$\frac{g}{g^0} - 1 = \sum_{i=1}^{p_0-1} \left(s_i + \pi_i^0\right) \sum_{j=t_{i-1}+1}^{t_i} q_j l_{\theta_j} + \left(\pi_{p_0}^0 - \sum_{i=1}^{p_0-1} s_i\right) \sum_{j=t_{p_0-1}+1}^{t_{p_0}} q_j l_{\theta_j} + \sum_{j=t_{p_0}+1}^{p} \pi_j l_{\theta_j} - 1$$

By remarking that when $\phi_t = \phi_t^0$, $\frac{g}{g^0}$ does not vary with $\psi_t$, we will study the variation of this ratio in a neighborhood of $\phi_t^0$ and for fixed $\psi_t$.

We can state the following result:

***Proposition 3***

*Let us denote* $D(\phi_t, \psi_t) = \left\|\frac{g(\phi_t, \psi_t)}{g^0} - 1\right\|_{L^2(\mu)}$. *With the notations of assumptions H-3 and H-4, for any fixed* $\psi_t$*, the second-order Taylor expansion at* $\phi_t^0$ *exists such as*

$$\frac{g}{g^0} - 1 = \left(\phi_t - \phi_t^0\right)^T l'_{(\phi_t^0, \psi_t)} + \frac{1}{2}\left(\phi_t - \phi_t^0\right)^T l''_{(\phi_t^0, \psi_t)} \left(\phi_t - \phi_t^0\right) - 1 + o\left(D(\phi_t, \psi_t)\right)$$

*with*

$$\left(\phi_t - \phi_t^0\right)^T l'_{(\phi_t^0, \psi_t)} = \sum_{i=1}^{p_0} \pi_i^0 \left(\sum_{j=t_{i-1}+1}^{t_i} q_j\theta_j - \theta_i^0\right)^T l'_i + \sum_{i=1}^{p_0} s_i l_{\theta_i^0}$$

$$+\sum_{j=t_{p_0}+1}^{p}\pi_j l_{\theta_j}$$

*and*

$$\left(\phi_t-\phi_t^0\right)^T l''_{\left(\phi_t^0,\psi_t\right)}\left(\phi_t-\phi_t^0\right)=\sum_{i=1}^{p_0}\left[2s_i\left(\sum_{j=t_{i-1}+1}^{t_i}q_j\theta_j-\theta_i^0\right)^T l'_i+\right.$$

$$\left.+\pi_i^0\sum_{j=t_{i-1}+1}^{t_i}q_j\left(\theta_j-\theta_i^0\right)^T l''_i\left(\theta_j-\theta_i^0\right)\right]$$

*Moreover,*

$$\left(\phi_t-\phi_t^0\right)^T l'_{\left(\phi_t^0,\psi_t\right)}+\frac{1}{2}\left(\phi_t-\phi_t^0\right)^T l''_{\left(\phi_t^0,\psi_t\right)}\left(\phi_t-\phi_t^0\right)=0\Leftrightarrow\phi_t=\phi_t^0$$

***Proof of Proposition 3***

The first term in the development can be computed easily by remarking that the gradient of $\frac{g}{g^0}-1$ at $\left(\phi_t^0,\psi_t\right)$ is:

- for $i\in\{1,...,p_0\}$ and $j\in\{t_{i-1}+1,...,t_i\}$, $\frac{\partial\left(\frac{g}{g^0}-1\right)}{\partial\theta_j}\left(\phi_t^0,\psi_t\right)=\pi_i^0 q_j l'_i$
- for $i\in\{1,...,p_0-1\}$,
  $\frac{\partial\left(\frac{g}{g^0}-1\right)}{\partial s_i}\left(\phi_t^0,\psi_t\right)=\sum_{j=t_{i-1}+1}^{t_i}q_j l_{\theta_i^0}-\sum_{j=t_{p_0-1}+1}^{t_{p_0}}q_j l_{\theta_{p_0}^0}=l_{\theta_i^0}-l_{\theta_{p_0}^0}$
- for $j\in\{t_{p_0}+1,...,p\}$, $\frac{\partial\left(\frac{g}{g^0}-1\right)}{\partial\pi_j}\left(\phi_t^0,\psi_t\right)=l_j$

The term of second order can be obtained by direct computations once the Hessian is computed at $\left(\phi_t^0,\psi_t\right)$:

- $\frac{\partial^2\left(\frac{g}{g^0}-1\right)}{\partial\theta_j^2}\left(\phi_t^0,\psi_t\right)=\pi_i^0 q_j l''_i$ , $i=1,...,p_0$ and $j=t_{i-1}+1,...,t_i$
- $\frac{\partial^2\left(\frac{g}{g^0}-1\right)}{\partial\theta_j\partial\theta_l}\left(\phi_t^0,\psi_t\right)=0$ , $j,l=1,...,p$ and $j\neq l$
- $\frac{\partial^2\left(\frac{g}{g^0}-1\right)}{\partial s_i\partial s_k}\left(\phi_t^0,\psi_t\right)=0$ , $i,k=1,...,p_0-1$
- $\frac{\partial^2\left(\frac{g}{g^0}-1\right)}{\partial s_i\partial\theta_j}\left(\phi_t^0,\psi_t\right)=q_j l'_i$ , $i=1,...,p_0-1$ and $j=t_{i-1}+1,...,t_i$
- $\frac{\partial^2\left(\frac{g}{g^0}-1\right)}{\partial s_i\partial\theta_j}\left(\phi_t^0,\psi_t\right)=-q_j l'_{p_0}$ , $i=1,...,p_0-1$ and $j=t_{p_0-1}+1,...,t_{p_0}$
- the other crossed derivatives of $s_i$ and $\theta_j$ are zero

It still has to be proven that the rest is $o\left(\left\|\phi_t-\phi_t^0\right\|\right)$. As it can be easily seen that the third derivative of $\frac{g}{g^0}-1$ can be expressed in terms of partial derivatives of order two and three of $l_{\theta_j^0}$, $j=1,...,p_0$, the result follows from the assumption H-3 and the linear independence in H-4. ■

Using the Taylor expansion above, we can now show that $\mathcal{S}_0\setminus\left\{g^0\right\}$ is a Donsker class, using the next result:

***Proposition 4***

*Let $d$ be the dimension of the parameter indexing the functions $g_\theta$. The number of $\varepsilon$-brackets $\mathcal{N}_{[\cdot]}\left(\varepsilon, \mathcal{S}_0, \|\cdot\|_2\right)$ covering $\mathcal{S}_0$ is $\mathcal{O}\left(\frac{1}{\varepsilon}\right)^{p_0\times(2d)+p}$.*

***Proof of Proposition 4***

The idea of this proof is to bound $\mathcal{N}_{[\cdot]}\left(\varepsilon, \mathcal{S}_0, \|\cdot\|_2\right)$ by the number of $\varepsilon$-brackets covering a wider class of functions. For every $g \in \mathcal{F}_0$, we will consider the reparameterization $\Phi = (\phi_t, \psi_t)$ which allows to write a second-order development of the density ratio:

$$\frac{g_{(\phi_t,\psi_t)}}{g^0} - 1 = \left(\phi_t - \phi_t^0\right)^T l'_{(\phi_t^0,\psi_t)} + \frac{1}{2}\left(\phi_t - \phi_t^0\right)^T l''_{(\phi_t^0,\psi_t)}\left(\phi_t - \phi_t^0\right) + o\left(D\left(\phi_t, \psi_t\right)\right)$$

Then, by remarking that the first two terms in the Taylor expansion are linear combinations of $l_{\theta_i^0}$, $l'_i$, $l''_i$, $i = 1, ..., p_0$ and $l_{\theta_j}, j = t_{p_0}+1, \cdots, p$, the density ratio can be written also as:

$$\frac{g_{(\phi_t,\psi_t)}}{g^0} - 1 = \sum_{i=1}^{p_0} \alpha_i l_{\theta_i^0} + \sum_{j=t_{p_0}+1}^{p} \alpha_j l_{\theta_j} + \sum_{i=1}^{p_0} \beta_i^T l'_i + \sum_{i=1}^{p_0} \gamma_i^T l''_i \gamma_i + o\left(D\left(\phi_t, \psi_t\right)\right)$$

where $(\alpha_i)_{1\leq i\leq p} \in \mathbb{R}$, $(\beta_i)_{1\leq i\leq p_0}$ and $(\gamma_i)_{1\leq i\leq p_0} \in \mathbb{R}^d$.

Now, using the linear independence, $\exists m > 0$ ,so that, for every

$$\left(\alpha_j, j = 1, \cdots, p, \beta_i, \gamma_i\gamma_i^T, i = 1, \cdots, p_0\right)$$

of norm 1,

$$\left\|\sum_{i=1}^{p_0} \alpha_i l_{\theta_i^0} + \sum_{j=t_{p_0}+1}^{p} \alpha_j l_{\theta_j} + \sum_{i=1}^{p_0} \beta_i^T l'_i + \sum_{i=1}^{p_0} \gamma_i^T l''_i \gamma_i\right\|_{L^2(\mu)} \geq m.$$

At the same time, since

$$\left\|\frac{\frac{g_{(\phi_t,\psi_t)}}{g^0} - 1}{\left\|\frac{g_{(\phi_t,\psi_t)}}{g^0} - 1\right\|_{L^2(\mu)}}\right\|_{L^2(\mu)} = 1$$

we will obtain that the Euclidean norm of the coefficients in the second-order development of $\frac{\frac{g_{(\phi_t,\psi_t)}}{g^0} - 1}{\left\|\frac{g_{(\phi_t,\psi_t)}}{g^0} - 1\right\|_{L^2(\mu)}}$ is upper bounded by $\frac{1}{m}$. This fact implies that $\mathcal{S}_0$ can be included in

$$\mathcal{H} = \begin{cases} \sum_{i=1}^{p_0}\left(\alpha_i l_{\theta_i^0} + \beta_i^T l'_i + \gamma_i^T l''_i \gamma_i\right) + \sum_{j=t_{p_0}+1}^{p} \alpha_j l_{\theta_j} + o(1), \\ \left\|\left(\alpha_j, j = t_{p_0}+1, \cdots, p, \beta_i, \gamma_i\gamma_i^T, i = 1, \cdots, p_0\right)\right\| \leq \frac{1}{m} \end{cases}$$

and then obviously $\mathcal{N}_{[\cdot]}\left(\varepsilon, \mathcal{H}, \|\cdot\|_2\right) = \mathcal{O}\left(\frac{1}{\varepsilon}\right)^{p_0\times 2d+p+1}$. ■

Since the set $\mathcal{S}_\eta$ was proved to be Donsker, it remains to identify the asymptotic index set of score functions.

**Asymptotic index set.** The set of limit score functions $\mathbb{F}$ is defined as the set of functions $d$ so that one can find a sequence $g_n$ satisfying $\|\frac{g_n-f}{f}\|_2 \to 0$ and $\|d - s_{g_n}\|_2 \to 0$.

Let us define the two principal behaviors for the sequences $g_n$ which influence the form of functions $d$ :

- If the second order term is negligible with respect to the first one :

$$\frac{g_n}{g^0} - 1 = (\Phi_n - \Phi^0)^T l'_{(\Phi_t^0,\psi_n)} + o(D(\Phi_n, \psi_n)).$$

- If the second order term is not negligible with respect to the first one :

$$\begin{array}{l} \frac{g_n}{g^0} - 1 = (\Phi_n - \Phi^0)^T l'_{(\Phi_t^0,\psi_n)} + \\ 0.5(\Phi_n - \Phi^0)^T l''_{(\Phi^0,\psi_n)}(\Phi_n - \Phi^0) + o(D(\Phi_n, \psi_n)). \end{array}$$

In the first case, a set $t = (t_0, \cdots, t_{p_0})$ exists so that the limit function of $s_{g_n}$ will be in the set:

$$\begin{array}{rl} \mathbb{D}_1^t = & \Big\{\Omega\left(\sum_{i=1}^{p_0} \zeta_i^T l_{\theta_i^0} + \sum_{i=p_0+1}^{p} \zeta_i^T l_{\theta_i} + \sum_{i=1}^{p_0} \lambda_i^T l_i'\right) \\ & \lambda_1, \cdots, \lambda_{p_0} \in \mathbb{R}^d ; \zeta_1, \cdots, \zeta_p \in \mathbb{R} \\ & \theta_{t_{p_0}+1}, \cdots, \theta_p \in \Theta - \left\{\theta_1^0, \cdots, \theta_{p_0}^0\right\}\Big\} \end{array}$$

In the second case, an index $i$ exists so that :

$$\sum_{j=t_{i-1}+1}^{t_i} q_j(\theta_j - \theta_i^0) = 0,$$

Otherwise, the second order term will be negligible compared to the first one, so

$$\sum_{j=t_{i-1}+1}^{t_i} \sqrt{q_j} \times \sqrt{q_j}(\theta_j - \theta_i^0) = 0.$$

Hence, a set a set $t = (t_0, \cdots, t_{p_0})$ exists so that the set of functions $d$ will be:

$$\begin{array}{l} \Big\{\Omega\Big(\sum_{i=1}^{p_0} \zeta_i l_{\theta_i^0} + \sum_{i=p_0+1}^{p} \zeta_i l_{\theta_i} + \sum_{i=1}^{p_0} \lambda_i^T l_i' \\ +\delta \sum_{i=1}^{p_0} \sum_{j=t_{i-1}+1}^{t_i} \gamma_j^T l_i'' \gamma_j\Big) \\ \lambda_1, \cdots, \lambda_{p_0}, \gamma_1, \cdots, \gamma_{t_{p_0}} \in \mathbb{R}^d ; \zeta_1, \cdots, \zeta_p \in \mathbb{R} \\ \theta_{t_{p_0}+1}, \cdots, \theta_p \in \Theta - \left\{\theta_1^0, \cdots, \theta_{p_0}^0\right\}\Big\} \end{array}$$

where $\delta = 1$ if there exists a vector $\mathbf{q}$ exists so that:
$q_j \le 0$, $\sum_{j=t_{i-1}+1}^{t_i} q_j = 1$, $\sum_{j=t_{i-1}+1}^{t_i} \sqrt{q_j}\gamma_j^t = 0$ for $i = 1, \cdots, p_0$; and $\delta = 0$ otherwise.

So, the limit functions will belong to $\mathbb{F}$. Conversely, let $d$ be an element of $\mathbb{F}$, as functions $d$ belong to the Hilbert sphere, one of their components is not equal to 0. Let us assume that this component is $\zeta_1$, but the proof would be similar with any other component. The norm of $d$ is 1, so any component of $d$ is determined by the ratio: $\frac{\zeta_2}{\zeta_1}, \cdots, \frac{1}{\zeta_1}\gamma_{p_0}$.

Then, by assumption H-1, the set of possible parameters contains a neighborhood of the parameters realizing the true conditional density function $g^0$, we can chose the parameters of $g_n$ so that:

$$\begin{array}{ll}\forall i \in \{2, \cdots, p_0\} : & \frac{\sum_{j=t_{i-1}+1}^{t_i} \pi_j^n - \pi_i^0}{\sum_{j=1}^{t_1} \pi_j^n - \pi_1^0} \overset{n\to\infty}{\longrightarrow} \frac{\zeta_i}{\zeta_1}, \\ \forall i \in \{1, \cdots, p_0\} : & \frac{\sum_{j=t_{i-1}+1}^{t_i} q_j^n \left(\theta_j^n - \theta_i^0\right)}{\sum_{j=1}^{t_1} \pi_j^n - \pi_1^0} \overset{n\to\infty}{\longrightarrow} \frac{1}{\zeta_1}\lambda_i, \\ \forall j \in \{1, \cdots, t_{p_0}\} : & \frac{\sqrt{q_j^n}}{\sum_{j=1}^{t_1} \pi_j^n - \pi_1^0} \left(\theta_j^n - \theta_i^0\right) \overset{n\to\infty}{\longrightarrow} \frac{1}{\zeta_1}\gamma_j, \\ \forall i \in \{p_0+1, \cdots, p\} : & \frac{\pi_i^n}{\sum_{j=1}^{t_1} \pi_j^n - \pi_1^0} \overset{n\to\infty}{\longrightarrow} \frac{1}{\zeta_1}\zeta_i. \end{array}$$

■

***Proof of Proposition 1***

Since the noise is Gaussian and $\left|a_i^0\right| < 1$ for every $i \in \{1, ..., p_0\}$, by Yao and Attali (2000), there exists a unique strictly stationary and geometrically ergodic solution, which in particular will be geometrically $\beta$-mixing.

On the other hand, the Gaussian noise implies the existence of moments of any order. Now let us prove the existence of an exponential moment for $Y_t$. By denoting $\sigma = \max_{i=1,\ldots,p_0} \sigma_i^0$ , $\rho = \max_{i=1,\ldots,p_0} \left|a_i^0\right| < 1$, $b = \max_{i=1,\ldots,p_0} \left|b_i^0\right|$ and for $s \in \mathbb{N}^\star$, one has :

$$|Y_t|^{2s} = \left|F_{\theta_{X_t}}^0 (Y_{t-1}) + \varepsilon_{\theta_{X_t}} (t)\right|^{2s} \le (\rho |Y_{t-1}| + b + \sigma |\varepsilon_t|)^{2s} \le ... \le$$

$$\le \left(b + \sigma |\varepsilon_t| + \sum_{k=1}^{\infty} \rho^k (b + \sigma |\varepsilon_{t-k}|)\right)^{2s} = \left(\sum_{k=0}^{\infty} \rho^k (b + \sigma |\varepsilon_{t-k}|)\right)^{2s}$$

By taking the expectation,

$$E\left(|Y_t|^{2s}\right)^{\frac{1}{2s}} \le E\left(\left(\sum_{k=0}^{\infty} \rho^k (b + \sigma |\varepsilon_{t-k}|)\right)^{2s}\right)^{\frac{1}{2s}} \le \sum_{k=0}^{\infty} \rho^k \left(b + \sigma E\left(|\varepsilon_{t-k}|^{2s}\right)^{\frac{1}{2s}}\right)$$

Since $\rho < 1$ and $E\left(|\varepsilon_t|^{2s}\right) \ge E\left(\varepsilon_t^2\right) = 1$, we finally obtain

$$E\left(|Y_t|^{2s}\right)^{\frac{1}{2s}} \le \frac{b + \sigma E\left(|\varepsilon_t|^{2s}\right)^{\frac{1}{2s}}}{1-\rho} \le \frac{b+\sigma}{1-\rho} E\left(|\varepsilon_t|^{2s}\right)^{\frac{1}{2s}}$$

The exponential moment can be computed then by

$$E\left(e^{\delta Y_t^2}\right) = \sum_{k=0}^{\infty} \frac{E\,|Y_t|^{2k}}{k!} \delta^k \le \sum_{k=0}^{\infty} \frac{E\,|\varepsilon_t|^{2k}}{k!} \left[\delta \left(\frac{b+\sigma}{1-\rho}\right)^2\right]^k$$

The last term being the moment generating function of a $\chi^2(1)$-distribution, it will be finite for any $\delta$ such that $0 < \delta < \frac{1}{2}\left(\frac{1-\rho}{b+\sigma}\right)^2$.

■

***Proof of Proposition 2***

The norm of the generalized score function is

$$\left\| \frac{g}{g^0} - 1 \right\|_{L^2(\mu)} = \int \frac{g^2(y_1, y_2)}{g^0(y_1, y_2)} dy_2 d\lambda(y_1) - 1 =$$

$$= \int \frac{\left(\sum_{i=1}^{p} \pi_i f_{\theta_i}(y_2 - F_{\theta_i}(y_1))\right)^2}{\sum_{j=1}^{p_0} \pi_j^0 f_{\theta_j}^0 \left(y_2 - F_{\theta_j}^0(y_1)\right)} dy_2 d\lambda(y_1) - 1$$

and by the inequality $\left(\sum_{i=1}^{p} \pi_i f_{\theta_i}(y_2 - F_{\theta_i}(y_1))\right)^2 \leq \sum_{i=1}^{p} \pi_i f_{\theta_i}^2(y_2 - F_{\theta_i}(y_1))$, the integral will be finite if

$$\int \frac{f_{\theta_i}^2(y_2 - F_{\theta_i}(y_1))}{\sum_{j=1}^{p_0} \pi_j^0 f_{\theta_j}^0 \left(y_2 - F_{\theta_j}^0(y_1)\right)} dy_2 d\lambda(y_1) < \infty$$

for all $i \in \{1, ..., p\}$. On the other hand, since $\sum_{j=1}^{p_0} \pi_j^0 f_{\theta_j}^0 \left(y_2 - F_{\theta_j}^0(y_1)\right) \geq \pi_k^0 f_{\theta_k}^0 \left(y_2 - F_{\theta_k}^0(y_1)\right)$ for every $k \in \{1, ..., p_0\}$, the generalized score function is well defined if for every $i \in \{1, ..., p\}$, there exists $k \in \{1, ..., p_0\}$ such that

$$\int \frac{f_{\theta_i}^2(y_2 - F_{\theta_i}(y_1))}{f_{\theta_k}^0 \left(y_2 - F_{\theta_k}^0(y_1)\right)} dy_2 d\lambda(y_1) < \infty$$

Next, replace $f_{\theta_i}$ and $f_{\theta_k}^0$ by centered Gaussian densities with standard errors $\sigma_i$, $\sigma_k^0$, respectively, and consider also $F_{\theta_i}(y) = a_i y + b_i$ and $F_{\theta_k}^0(y) = a_k^0 y + b_k^0$.

Then, each of the integrals above becomes:

$$\int \frac{f_{\theta_i}^2(y_2 - F_{\theta_i}(y_1))}{f_{\theta_k}^0 \left(y_2 - F_{\theta_k}^0(y_1)\right)} dy_2 d\lambda(y_1) = \int \left( \int \frac{\sigma_k^0}{\sqrt{2\pi}\sigma_i^2} \cdot \right.$$

$$\left. \exp\left\{ -\left( \frac{1}{\sigma_i^2} - \frac{1}{2\left(\sigma_k^0\right)^2} \right) (y_2 - m(y_1))^2 \right\} dy_2 \right)$$

$$\exp\left\{ \frac{\left(F_{\theta_i}(y_1) - F_{\theta_k}^0(y_1)\right)^2}{2\left(\sigma_k^0\right)^2 - \sigma_i^2} \right\} d\lambda(y_1)$$

where $m(y_1) = \frac{2\left(\sigma_k^0\right)^2 F_{\theta_i}(y_1) - \sigma_i^2 F_{\theta_k}^0(y_1)}{2\left(\sigma_k^0\right)^2 - \sigma_i^2}$

To have a sufficient condition, the integral in $y_2$ is finite if $\sigma_i^2 < 2\left(\sigma_k^0\right)^2$, and the integral in $y_1$ is finite if $\frac{\left(a_i - a_k^0\right)^2}{2\left(\sigma_k^0\right)^2 - \sigma_i^2} < \delta$.

■

## References


[1] Bradley R.C. (2005) Basic properties of strong mixing conditions. A survey and some open questions, *Probability Surveys*, **2**, 107-144

[2] Dacunha-Castelle D., Gassiat E. (1997a) The estimation of the order of a mixture model, *Bernoulli*, **3**, 279-299

[3] Dacunha-Castelle D., Gassiat E. (1997b) Testing in locally conic models, *ESAIM Prob. and Stat.*, **1**, 285-317

[4] Dacunha-Castelle D., Gassiat E. (1999) Testing the order of a model using locally conic parametrization: population mixtures and stationary ARMA processes, *The Annals of Statistics*, **27(4)** , 1178-1209

[5] Douc R., Moulines E., Rydén T. (2004) Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime, *The Annals of Statistics*, **32(5)**, 2254-2304

[6] Doukhan P. (1995) *Mixing: properties and examples*, Springer-Verlag, New York

[7] Doukhan P., Massart P., Rio E. (1995) Invariance principles for absolutely regular empirical processes, *Ann. Inst. Henri Poincaré (B) Probabilités et Statistiques*, **31(2)**, 393-427

[8] Engel Ch., Hamilton J.D. (1990) Long swings in the dollar: are they in the data and do markets know it?, *American Economic Review*, **80(4)**, 689-713

[9] Francq C., Roussignol M. (1998) Ergodicity of autoregressive processes with Markov-switching and consistency of the maximum likelihood estimator, *Statistics*, **32(2)**, 151-173

[10] Fukumizu K. (2003) Likelihood ratio of unidentifiable models and multilayer neural networks, *Ann. Statist.* , **31**, 833-851

[11] Garcia R. (1998) Asymptotic null distribution of the likelihood ratio test in Markov switching models, *International Economic Review*, **39(3)**, 763-788

[12] Gassiat E., Keribin C. (2000) The likelihood ratio test for the number of components in a mixture with Markov regime, *ESAIM P&S*, **4**, 25-52

[13] Gassiat E. (2002) Likelihood ratio inequalities with applications to various mixtures, *Ann. Inst. Henri Poincaré*, **38**, 897-906

[14] Hamilton J. D. (1989) A new approach to the economic analysis of nonstationary time series and the business cycle, *Econometrica*, **57(2)**, 357-384

[15] Hamilton J. D. (1990) Analysis of time series subject to changes in regime, *Journal of Econometrics*, **64**, 307-333

[16] Hansen B.E. (1992) The likelihood ratio test under nonstandard conditions: testing the Markov switching model of GNP, *Journal of Applied Econometrics*, **7**, 61-82

[17] Hansen B.E. (1996a) Erratum: the likelihood ratio test under nonstandard conditions: testing the Markov switching model of GNP, *Journal of Applied Econometrics*, **11**, 195-198

[18] Hansen B.E. (1996b) Inference when a nuisance parameter is not identified under the null hypothesis, *Econometrica*, **64(2)**, 413-430

[19] Henna J. (1985) On estimating the number of constituents of a finite mixture of continuous distributions, *Ann. Inst. Statist. Math.*, **37**, 235-240

[20] Izenman A.J., Sommer C. (1988) Philatelic mixtures and multivariate densities, *Journal of the American Stat. Assoc.*, **83**, 941-953

[21] Keribin C. (2000) Consistent estimation of the order of mixture models, *Sankhya: The Indian Journal of Statistics*, **62**, 49-66

[22] Krishnamurthy V., Rydén T. (1998) Consistent estimation of linear and non-linear autoregressive models with Markov regime,*Journal of Time Series Analysis*, **19**, 291-307

[23] Lam P.-S. (1990), The Hamilton model with a general autoregressive component: estimation and comparison with other models of economic time series, *Journal of Monetary Economics*, **26**, 409-432

[24] Leroux B.G. (1992a) Maximum penalized likelihood estimation for independent and Markov-dependent mixture models, *Biometrics*, **48**, 545-558

[25] Leroux B.G. (1992b) Consistent estimation of a mixing distribution, *The Annals of Statistics,* **20**, 1350-1360

[26] Lindsay B.G. (1983) Moment matrices: application in mixtures, *The Annals of Statistics*, **17**, 722-740

[27] Liu X., Shao Y. (2003) Asymptotics for likelihood ratio tests under loss of identifiability, *The Annals of Statistics*, **31(3)**, 807-832

[28] Rios R., Rodriguez L.A. (2008) Penalized estimate of the number of states in Gaussian linear AR with Markov regime, *Electronic Journal of Statistics*, **2**, 1111-1128

[29] Roeder K. (1994) A graphical technique for determining the number of components in a mixture of normals, *Journal of the American Stat. Assoc.*, **89**, 487-495

[30] Ryden T. (1995) Estimating the order of hidden Markov models, *Statistics*, **26**, 345-354

[31] Schwert G.W. (1989) Business cycles, financial crises and stock volatility, *Carnegie-Rochester Conference Series on Public Policy*, **31(1)**, 83-125

[32] Van der Vaart A.W. (2000) *Asymptotic Statistics*, Cambridge University Press

[33] Wong C.S., Li W.K. (2000) On a mixture autoregressive model, *J. Royal Statist. Soc, Series B*, **62(1)** 95-115

[34] Yao J.F., Attali J.G. (2000) On stability of nonlinear AR processes with Markov switching, *Advances in Applied Probability*, **32(2)**, 394-407