Estimating the number of regimes in a switching autoregressive model

Madalina Olteanu, Joseph Rynkiewicz

ESTIMATING THE NUMBER OF REGIMES IN A SWITCHING AUTOREGRESSIVE MODEL

M.OLTEANU AND J. RYNKIEWICZ
SAMOS-MATISSE CES UNIVERSITÉ PARIS 1

ABSTRACT. In this paper we are interested in estimating the number of regimes in a switching autoregressive model. The penalized marginal-likelihood criterion for mixture models and hidden Markov models introduced by Keribin (2000) and, respectively, Gassiat (2002) is extended to autoregressive models with independent regime changes for which a penalized-likelihood criterion is proposed. We prove the consistency of the estimate under some hypothesis which involve essentially the bracketing entropy of the generalized score-functions class and we verify these hypothesis in the Gaussian case by reparameterizing the model to avoid non-identifiability problems. Some numerical examples illustrate the result and its convergence properties. Finally, we prove that a direct generalization of the "marginal likelihood" criterion to switching Markov models is not possible.

1. INTRODUCTION

This paper addresses the problem of estimating the true number of regimes in a switching autoregressive model. We suppose that a n-sample $(Y_1, ..., Y_n)$ of a time series is observed, with Y_t depending on the past Y_{t-1} and on some hidden discrete valued process X_t which can be an iid sequence or a Markov chain. If the number of states of X_t is not known or fixed in advance, this model is a typical example in non-identifiability problems. In these cases, the Fisher information matrix is degenerate, the usual regularity conditions do not hold and the classical theory for convergence of the maximum-likelihood estimate does not apply.

However, several ideas and methods were proposed to estimate the dimension of the state-space of X_t in the particular case of mixture models : various 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 Gassiat (2002). Moreover, Gassiat (2002) proved that if X_t is a Markov chain, the penalized-likelihood estimate converges in probability to the true number of regimes. In Section 2, we extend the latter result to a penalized-likelihood criterion applied to mixtures of autoregressive processes with independent regime changes. The convergence is proven under some assumptions of which the most delicate to verify will be the Donsker property for the class of generalized score functions $\mathcal{S}$, defined as the normalized density ratio.

In Section 3, we verify these hypothesis in the case of a Gaussian noise. Since $\mathcal{S}$, the class of score functions, is parametric, one would expect it to be Donsker under some good regularity conditions, as in Van der Vaart (2000). But here is where the non-identifiability problem arises and complicates the task by breaking the regularity assumptions. The solution is to reparameterize the model in such a

Date: 11th October 2006.

way that the identifiable and non-identifiable parameters are well separated and a second-order Taylor expansion can be done in a neighbourhood of the identifiable parameter. To deal with our case, we adapt the reparameterization proposed by Liu and Shao (2003), which seems to be more convenient than the "locally conic parameterization" in Dacunha-Castelle and Gassiat (1997, 1999). The last part of Section 3 provides some simulation results illustrating the empirical speed of convergence, as well as the stability of the estimate.

The last section handles the case when regime changes are Markovian. Once we have the result in the independent case, it seems natural to generalize it by using the marginal likelihood as defined in Gassiat (2002). Yet, it can be seen right away that this likelihood is no longer a constrast function and the convergence is achieved only in the particular case of constant autoregressive functions. This proves that the marginal likelihood is not the right tool in the general case and that the true likelihood should be studied instead.

2. Main result in the independent regime-switching case

2.1. **The model.** Throughout this paper, we will study autoregressive models with one lag of time, the extension to a finite number of lags k being immediate. Let us consider the real-valued time series Y_t which verifies the following model

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

where

- X_t is a sequence of independent identically distributed variables with values in a finite space $\{1, ..., p_0\}$ and probability distribution π^0
- for every $i \in \{1, ..., p_0\}$, $F^0_i(y)$ is a parametric function depending on θ^0_i and it describes the autoregressive model in each of the p_0 regimes. We suppose throughout the rest of the paper that F^0_i are sublinear, that is they are continous and there exist (a^0_i, b^0_i) positive real numbers such that $\left|F^0_i(y)\right| \leq a^0_i |y| + b^0_i$, $y \in \mathbb{R}$, for all $i = 1, ..., p_0$.
- for every $i \in \{1, ..., p_0\}$, $\varepsilon_i(t)$ is an independent identically distributed noise with density f^0_i strictly positive with respect to the Lebesgue measure and depending on the parameter θ^0_i

The estimate for the number of regimes will be defined in the case of strict stationarity for which we need to assume sufficient conditions. Yao and Attali (2000) gave such conditions for an autoregressive model with Markov switching between the regimes. Their result can be adapted easily by replacing the invariant distribution of the hidden Markov chain with the probability distribution of the independent regime changes X_t.

Let us then introduce the next hypothesis

(HS) $(\exists)$ $s \geq 1$ such that $E|\varepsilon_1|^s < \infty$ and $\sum_{i=1}^{p_0} \pi^0_i \left(a^0_i\right)^s < 1$

Following Yao and Attali (2000) argument, one can prove that under (HS) there exists a unique strictly-stationary solution Y_t, geometrically-ergodic and with invariant probability measure admiting s-order moments. One may remark also that if the model in every regime is strictly stationary, that is $\left|a^0_i\right| < 1$ for every $i \in \{1, ..., p_0\}$, then Y_t is globally stationary.

2.2. **Penalized-likelihood estimate for the number of regimes.** Let us consider an observed sample $\{y_1, ..., y_n\}$ of the time series Y_k. Then, for every observation y_k, the conditional density with respect to the previous y_{k-1} and marginally in X_k is

$$f(y_k \mid y_{k-1}) = \sum_{i=1}^{p_0} \pi_i^0 f_i^0 \left(y_k - F_i^0 (y_{k-1})\right)$$

As the goal is to estimate p_0, the number of regimes of the model, let us consider all possible conditional densities up to a maximal number of regimes P, a fixed positive integer. We shall consider the class of functions

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

$$\mathcal{G}_p = \left\{ g \mid g(y_1, y_2) = \sum_{i=1}^{p} \pi_i f_i (y_2 - F_i (y_1)),\ \pi_i \geq 0,\ \sum_{i=1}^{p} \pi_i = 1 \right\}$$

where, for all $i = 1, ..., p$

- F_i is a parametric function depending on θ_i
- f_i is a strictly positive density with respect to the Lebesgue measure depending on θ_i

(HC) We shall assume throughout the following that the parameters $\{(\pi_i, \theta_i), i = 1, ..., p\}$ belong to a compact set.

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

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

and let $p_0 = p(f)$ be the true number of regimes.

We can now define the estimate $\hat{p}$ as the argument $p \in \{1, ..., P\}$ maximizing the penalized criterion

$$(2) \qquad T_n(p) = sup_{g \in \mathcal{G}_p} l_n(g) - a_n(p)$$

where

$$l_n(g) = \sum_{k=2}^{n} log\, g(y_{k-1}, y_k)$$

is the log-likelihood marginal in X_k and $a_n(p)$ is a penalty term.

Before stating the result on the convergence of $\hat{p}$, we need the following likelihood ratio inequality which is an extension of Gassiat (2002) to multivariate dependent data and since the proof is identical, we will skip it.

Proposition 1

Let $\mathcal{G} \subset \mathcal{G}_P$ be a parametric family of conditional densities containing the true model f and let us define the generalized score function

$$s_g(y_1, y_2) = \frac{\frac{g(y_1,y_2)}{f(y_1,y_2)} - 1}{\left\| \frac{g}{f} - 1 \right\|_{L^2(\mu)}}$$

where μ is the stationary measure of (Y_{k-1}, Y_k). Then,

$$sup_{g \in \mathcal{G}} \left(l_n(g) - l_n(f) \right) \leq \frac{1}{2} sup_{g \in \mathcal{G}} \frac{\left(\sum_{k=2}^{n} s_g(y_{k-1}, y_k) \right)^2}{\sum_{k=2}^{n} (s_g)_-^2 (y_{k-1}, y_k)}$$

with $(s_g)_-(y_{k-1}, y_k) = min(0, s_g(y_{k-1}, y_k))$.

Since the sample observations are not independent, let us recall some notions on dependent data, such as β-mixing properties and a functional central limit theorem that we will need to prove the main result.

If $(Z_k)_{k \in \mathbb{Z}}$ is a strictly stationary sequence of random variables defined on a probability space $(\Omega, \mathcal{K}, \mathbb{P})$, we consider, for every $n \geq 1$, the β-mixing coefficients

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

where $\mathcal{F}_{-\infty}^0 = \sigma(Z_k, k \leq 0)$, $\mathcal{F}_n^\infty = \sigma(Z_k, k \geq n)$ and

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

By definition, the sequence Z_k is called β-mixing if $lim_{n \to \infty} \beta_n = 0$.

Now, if the strictly stationary sequence Z_k is β-mixing and moreover $\sum_{n \geq 1} \beta_n < \infty$, one can define the $\mathcal{L}_{2,\beta}(\mathbb{P})$-space by

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

where

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

Now, let us consider $\mathcal{F}$ a set of functions on some space endowed with a norm $\|\cdot\|$. For every $\varepsilon > 0$, we define an ε-bracket by $[l, u] = \{f \in \mathcal{F}, l \leq f \leq u\}$ such that $\|u - l\| < \varepsilon$. The ε-bracketing entropy is then

$$\mathcal{H}_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|) = log\left(\mathcal{N}_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|) \right)$$

where $\mathcal{N}_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)$ is the minimum number of ε-brackets necessary to cover $\mathcal{F}$.

Doukhan, Massart and Rio (1995) proved that if the series $(Z_k)_{k \in \mathbb{Z}}$ is strictly stationary, β-mixing and $\sum_{n \geq 1} \beta_n < \infty$, then it converges in probability to a Gaussian process, uniformly over any set of functions $\mathcal{F}$ such that $\mathcal{F} \subset \mathcal{L}_{2,\beta}$ and

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

With the previous definitions, we can state the following theorem, proven in the Appendix :

Theorem 1

Consider the model (Y_k, X_k) *defined by (1) and the penalized-likelihood criterion introduced in (2). Let us introduce the next assumptions :*

- ***(A1)*** $a_n(\cdot)$ *is an increasing function of* p, $a_n(p_1) - a_n(p_2) \to \infty$ *when* $n \to \infty$ *for every* $p_1 > p_2$ *and* $\frac{a_n(p)}{n} \to 0$ *when* $n \to \infty$ *for every* p
- ***(A2)*** *the model* (Y_k, X_k) *verifies the weak identifiability assumption* ***(HI)***

$$\sum_{i=1}^{p} \pi_i f_i\left(y_2 - F_i(y_1)\right) = \sum_{i=1}^{p_0} \pi_i^0 f_i^0\left(y_2 - F_i^0(y_1)\right) \Leftrightarrow \sum_{i=1}^{p} \pi_i \delta_{\theta_i} = \sum_{i=1}^{p_0} \pi_i^0 \delta_{\theta_i^0}$$

- ***(A3)*** *the parameterization* $\theta_i \to f_i\left(y_2 - F_i(y_1)\right)$ *is continuous for every* (y_1, y_2) *and there exists* $m(y_1, y_2)$ *an integrable map with respect to the stationary measure of* (Y_k, Y_{k-1}) *such that* $|log(g)| < m$
- ***(A4)*** Y_k *satisfies the hypothesis* ***(HS)*** *and the family of generalized score functions associated to* $\mathcal{G}_P$

$$\mathcal{S} = \left\{ s_g,\ s_g(y_1, y_2) = \frac{\frac{g(y_1,y_2)}{f(y_1,y_2)} - 1}{\left\| \frac{g}{f} - 1 \right\|_{L^2(\mu)}},\ g \in \mathcal{G}_P,\ g \neq f \right\} \subset \mathcal{L}_2(\mu)$$

and for every $\varepsilon > 0$

$$\mathcal{H}_{[]}(\varepsilon, \mathcal{S}, \|\cdot\|_2) = \mathcal{O}(|log\,\varepsilon|)$$

Then, under the hypothesis (A1)-(A4) and (HC), $\hat{p} \to p_0$ *in probability.*

3. Application for linear regressions and Gaussian noise

3.1. **The model.** In this section we are interested whether the theorem above can be used in applications on simulated and real-life data. We have chosen to study the case of a gaussian noise and for this we need to verify that hypothesis **(HC)** and **(A1)-(A4)** are fulfilled. We shall consider that the process (X_t, Y_t) follows the true model

$$Y_t = F_{X_t}^0(Y_{t-1}) + \varepsilon_{X_t}(t) \qquad (3)$$

where

- X_t is a sequence of independent identically distributed variables with values in a finite space $\{1, ..., p_0\}$ and probability distribution π^0
- for every $i \in \{1, ..., p_0\}$, $F_i^0(y) = a_i^0 y + b_i^0$ describes a linear autoregressive model in each of the p_0 regimes
- for every $i \in \{1, ..., p_0\}$, $\varepsilon_i(t)$ is an independent identically distributed noise following a centered gaussian density $f_i^0 \sim \mathcal{N}\left(0, \left(\sigma_i^0\right)^2\right)$

The noise can be written then in a simpler manner as

$$\varepsilon_{X_t}(t) = \sigma^0_{X_t}\varepsilon_t$$

where $\sigma^0_{X_t} \in \{\sigma^0_1, ..., \sigma^0_{p_0}\}$ and $\varepsilon_t \sim \mathcal{N}(0,1)$.

We can then state the following result which ensures the strict stationarity and ergodicity :

Proposition 2

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

Now, let us consider a maximum number of regimes $P > 0$ and the class of all possible conditional densities of Y_t marginal in X_t :

$$\mathcal{G}_P = \bigcup_{p=1}^{P} \mathcal{G}_p\,,\ \mathcal{G}_p = \left\{ g \mid g(y_1, y_2) = \sum_{i=1}^{p} \pi_i f_i(y_2 - F_i(y_1)) \right\}$$

where

- $\sum_{i=1}^{p} \pi_i = 1$ and, with no loss of generality, we suppose that for every $i \in \{1, ..., p\}$, $\pi_i \geq \eta > 0$
- for every $i \in \{1, ..., p\}$, $F_i(y) = a_i y + b_i$, $f_i \sim \mathcal{N}(0, \sigma_i^2)$ and $\theta_i = (a_i,\, b_i,\, \sigma_i)$ belongs to a compact set

Then, the estimate $\hat{p}$ is defined as the maximizer of (2) and it converges in probability to the true number of regimes if the assumptions of Theorem 1 hold. The key hypothesis is that the class of generalized score functions

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

is Donsker. First, we shall verify that this class is well defined, that is $\left\| \frac{g}{f} - 1 \right\|_{L^2(\mu)} < \infty$, for all $g \in \mathcal{G}_P$.

3.2. **Existence of the score functions.** We shall start with the simple case of one true regime against two possible regimes, $p_0 = 1$ and $p = 2$. In this case, the true distribution will be

$$f(y_1, y_2) = f^0\left(y_2 - F^0(y_1)\right)$$

and the possible density

$$g(y_1, y_2) = \pi f_1(y_2 - F_1(y_1)) + (1 - \pi) f_2(y_2 - F_2(y_1))$$

One can prove then, by direct computations, that

Proposition 3

$\left\|\frac{g}{f}-1\right\|_{L^2(\mu)}<\infty$ *if* $\sigma_i^2<2\left(\sigma^0\right)^2$, $\left|a_i-a^0\right|<\sqrt{\delta\left(2\left(\sigma^0\right)^2-\sigma_i^2\right)}$ *for* $i\in\{1,2\}$ *and* $\delta>0$ *such that* $E\left(e^{\delta Y_t^2}\right)<\infty$.

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

The one-against-two regimes case can be easily generalized to the situation p, $p_0\in\mathbb{N}$:

Proposition 4

$\left\|\frac{g}{f}-1\right\|_{L^2(\mu)}<\infty$ *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* $E\left(e^{\delta Y_t^2}\right)<\infty$.

According to Teicher (1963), the weak identifiability hypothesis **(A2)** is verified for mixtures of gaussian densities. Moreover, since, by assumption, for every $i\in\{1,...,p\}$, $\pi_i\geq\eta>0$, the estimates $\hat{\theta}_n=\left(\hat{\theta}_{1,n},...,\hat{\theta}_{p,n}\right)$ are consistent and the sufficient conditions in Proposition 4 are verified immediately for n sufficiently large.

Next, let us prove that $\mathcal{S}$ is Donsker and that $\mathcal{H}_{[]}\left(\varepsilon,\mathcal{S},\|\cdot\|_2\right)=\mathcal{O}\left(|log\,\varepsilon|\right)$ for all $\varepsilon>0$.

3.3. **Donsker property for the class of generalized score functions $\mathcal{S}$.** For $g\in\mathcal{G}_P$, let us denote $\theta=(\theta_1,...,\theta_p)$ and $\pi=(\pi_1,...,\pi_p)$, so that the global parameter will be $\Phi=(\theta,\pi)$ and the associated generalized score function

$$s_\Phi:=s_g=\frac{\frac{g}{f}-1}{\left\|\frac{g}{f}-1\right\|_{L^2(\mu)}}$$

Proving that a parametric family like $\mathcal{S}$ is a Donsker class is usually immediate under good regularity conditions (see, for instance, Van der Vaart (2000)). In this particular case, the problems arise when $g\to f$ and the limits in $L^2(\mu)$ of s_g have to be computed. To achieve our proof, let us then split $\mathcal{S}$ into two classes of function.

We shall consider $\mathcal{F}_0\subset\mathcal{G}_P$ a neighbourhood of f such that it exists $\delta>0$ verifying $\mathcal{F}_0=\left\{g\in G_p,\ \left\|\frac{g}{f}-1\right\|_{L^2(\mu)}\leq\delta,\ g\neq f\right\}$ and let $\mathcal{S}_0=\{s_g,\ g\in\mathcal{F}_0\}$.

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

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

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

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

On the other hand, under the assumptions in Proposition 4, $\frac{g}{f}$ has square integrable partial derivatives of order one and, using the result on parametric classes of functions in Van der Vaart (2000), we get that

$$\mathcal{N}_{[]}\left(\varepsilon, \mathcal{S} \setminus \mathcal{S}_0, \|\cdot\|_2\right) = \mathcal{O}\left(\frac{1}{\delta\varepsilon}\right)^{4P}$$

It remains to prove that $\mathcal{S}_0$ is Donsker. The guiding idea is to reparameterize the model in a convenient manner which will allow a Taylor expansion around the identifiable part of the true value. We shall use a slight modification of the method proposed by Liu and Shao (2003).

In the following we will make the additional assumption $p_0 < p$.

Let us remark that when $\frac{g}{f} - 1 = 0$, the weak identifiability hypothesis **(A2)** and the fact that for every $i \in \{1, ..., p\}$, $\pi_i \geq \eta > 0$, implies that there exists a vector $t = (t_i)_{0 \leq i \leq p_0}$ such that $0 = t_0 < t_1 < ... < t_{p_0} = p$ and, modulo a permutation, Φ can be rewritten as follows :

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

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 the new parameterization will be

$$\Theta_t = (\phi_t, \psi_t), \; \phi_t = \left((\theta_j)_{1 \leq j \leq p}, (s_i)_{1 \leq i \leq p_0 - 1}\right), \; \psi_t = (q_j)_{1 \leq j \leq p}$$

with ϕ_t containing all the identifiable parameters of the model and ψ_t the non-identifiable ones. Then, for $g = f$, we will have that

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

This reparameterization allows to write a second-order Taylor expansion of $\frac{g}{f} - 1$ at ϕ_t^0. For ease of writing, we shall first denote

$$g_j(y_1, y_2) = g_{\theta_j}(y_1, y_2) = \frac{f_j(y_2 - F_j(y_1))}{\sum_{i=1}^{p_0} \pi_i^0 f_i^0(y_2 - F_i^0(y_1))} - 1$$

Then, the density ratio becomes :

$$\frac{g}{f} - 1 = \sum_{i=1}^{p_0} (s_i + \pi_i^0) \sum_{j=t_{i-1}+1}^{t_i} q_j g_j$$

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

$$\frac{g}{f}-1=\sum_{i=1}^{p_0-1}\left(s_i+\pi_i^0\right)\sum_{j=t_{i-1}+1}^{t_i}q_jg_j+\left(\pi_{p_0}^0-\sum_{i=1}^{p_0-1}s_i\right)\sum_{j=t_{p_0-1}+1}^{t_{p_0}}q_jg_j$$

By remarking that when $\phi_t=\phi_t^0$, $\frac{g}{f}$ does not vary with ψ_t, we will study the variation of this ratio in a neighbourhood of ϕ_t^0 and for fixed ψ_t. First, let us introduce the following notations of the ϕ_t-derivatives of g_j computed at ϕ_t^0:

$$g_j':=\frac{\partial g_j}{\partial\theta_j}\left(\phi_t^0,\psi_t\right),\ g_j'':=\frac{\partial^2 g_j}{\partial\theta_j^2}\left(\phi_t^0,\psi_t\right),\ g_j'''':=\frac{\partial^3 g_j}{\partial\theta_j^3}\left(\phi_t^0,\psi_t\right)$$

With these notations we can state the following result :

Proposition 5

Let us denote $D\left(\phi_t,\psi_t\right)=\left\|\frac{g_{(\phi_t,\psi_t)}}{f}-1\right\|_{L^2(\mu)}$. *For any fixed* ψ_t, *there exists the second-order Taylor expansion at* ϕ_t^0 :

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

with

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

and

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

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

Moreover,

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

Using the Taylor expansion above, we can now show that $\mathcal{S}_{\mathcal{F}_0}=\{s_g,\ g\in\mathcal{F}_0,\ g\neq f\}$ is a Donsker class. With the next result, hypothesis **(A4)** is directly verified :

Proposition 6

The number of ε*-brackets* $\mathcal{N}_{[]}\left(\varepsilon,\mathcal{S}_0,\|\cdot\|_2\right)$ *covering* $\mathcal{S}_0$ *is* $\mathcal{O}\left(\frac{1}{\varepsilon}\right)^{9p_0}$.

With this last assertion, it is proven that Theorem 1 applies in the Gaussian case and that the only constraints are the stationarity of each autoregressive model in

the mixture and the choice of a penalty term according to hypothesis **(A1)**. The next section contains some numerical examples on simulated data which verify the convergence and the stability properties of the estimate $\hat{p}$.

3.4. **Numerical examples.** Once the theoretical result is verified in the Gaussian case, let us give some numerical experiments to illustrate it. Three things will be interesting to study: the speed of convergence, since we do not have it theoretically, the stability and the influence of the penalty term. For parcimony purposes and because of the important computation time, only the BIC penalty term was considered here, other possibilities are to be studied later. The examples are mixtures of two autoregressive models in which we vary the leading coefficients and the weights of the discrete mixing distribution. For each of them, we simulate 20 samples of lenghts $n = 200, 500, 1000, 1500, 2000$ and we fix $P = 3$ the upper bound for the number of regimes.

The likelihood is maximized via an EM algorithm (see, for instance, Dempster, Laird and Rubin (1977) or Redner and Walker (1984)). To avoid local maxima, the procedure is initialized several times with different starting values : in our case, ten different initializations provided good results. The stopping criteria applies when either there is no improvement in the likelihood value, either a maximum number of iterations, fixed at 200 here for reasonable computation time, is reached.

The results are summarized in Tables 1 and 2 at the end of this paper. The true conditional density is

$$f(y_1, y_2) = \pi_1^0 f_1^0\left(y_2 - F_1^0(y_1)\right) + \left(1 - \pi_1^0\right) f_2^0\left(y_2 - F_2^0(y_1)\right)$$

with $F_i^0(y_1) = a_i^0 y_1 + b_i^0$ and $f_i^0 \sim \mathcal{N}\left(0, \left(\sigma_i^0\right)^2\right)$ for $i \in \{1, 2\}$. For every example, we pick equal standard errors $\sigma_1^0 = \sigma_2^0 = 0.5$ and let vary the rest of the coefficients: $\pi_1^0 \in \{0.5, 0.7, 0.9\}$, $a_1^0, a_2^0 \in \{0.1, 0.5, 0.9\}$, $b_1^0 \in \{1, 0.5\}$ and $b_2^0 \in \{-1, -0.5\}$. In Table 1, the convergence is reached rapidly for a small number of observations, while in Table 2 this is less obvious, since the two components are chosen closer. However, in most of the examples, 2000 sample points are enough to obtain a good estimate of the number of regimes.

4. Is it possible to generalize to Markov-switching regimes?

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

$$Y_t = F_{X_t}^0\left(Y_{t-1}\right) + \varepsilon_{X_t}(t) \qquad (4)$$

where

- X_t is a homogeneous Markov chain, irreducible and aperiodic, with finite state-space $\{1, ..., p_0\}$ and π^0 is the stationary probability measure
- for every $i \in \{1, ..., p_0\}$, $F_i^0(y)$ and $\varepsilon_i(t)$ have the same properties of sublinearity and, respectively, existence of a strictly positive density as in Section 2

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

(HS) $(\exists)$ $s \geq 1$ such that $E|\varepsilon_1|^s < \infty$ and $\rho(Q_s) < 1$,

$$Q_s = \begin{pmatrix} \left(a_1^0\right)^s \pi_{11}^0 & \cdots & \left(a_{p_0}^0\right)^s \pi_{1p_0}^0 \\ \vdots & \ddots & \vdots \\ \left(a_1^0\right)^s \pi_{p_0 1}^0 & \cdots & \left(a_{p_0}^0\right)^s \pi_{p_0 p_0}^0 \end{pmatrix}$$

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

Considering an observed n-sample of Y_t, one would attempt to naturally extend the method of constructing the estimate $\hat{p}$ in the previous section. Several problems arise : on one hand, the non-identifiability which can be managed by reparameterizing the model and, on the other hand, X_t which is an unobserved and dependent process. This dependence will not allow an explicit form for the conditional density, marginal in X_t :

$$f\left(y_k \mid y_{k-1}, ..., y_0\right) = \sum_{i=1}^{p_0} \mathbb{P}\left(X_k = i \mid y_{k-1}, ..., y_0\right) f_i^0\left(y_k - F_i^0\left(y_{k-1}\right)\right)$$

since $\mathbb{P}\left(X_k = i \mid y_{k-1}, ..., y_0\right)$ has to be computed recursively. However, since X_t is stationary, we can define a new cost function which will involve the invariant probability measure.

As in section 2, for $P > 0$ a fixed integer, we consider the class of all possible mixture densities

$$\mathcal{G}_P = \bigcup_{p=1}^{P} \mathcal{G}_p,\ \mathcal{G}_p = \left\{ g \mid g\left(y_1, y_2\right) = \sum_{i=1}^{p} \pi_i f_i\left(y_2 - F_i\left(y_1\right)\right) \right\}$$

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

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

Let us define the new cost function

$$l_n\left(g\right) = \sum_{k=2}^{n} log\, g\left(y_{k-1}, y_k\right) = \sum_{k=2}^{n} log\left(\sum_{i=1}^{p} \pi_i f_i\left(y_2 - F_i\left(y_1\right)\right)\right)$$

One may notice that $l_n\left(g\right)$ resembles to the conditional likelihood marginal in X_t and may expect it to be maximized for $g = f$, where "the true value" is now written as

$$f\left(y_k \mid y_{k-1}, ..., y_0\right) = \sum_{i=1}^{p_0} \pi_i^0 f_i^0\left(y_k - F_i^0\left(y_{k-1}\right)\right)$$

Let us now verify if $l_n\left(g\right)$ is a contrast function and the maximum is reached at f. If (X, Y_2, Y_1) is a generic variable having as distribution the stationary mesure of the extended Markov chain (X_k, Y_k, Y_{k-1})

$$E\left[ln(g) - ln(f)\right] = \sum_{i=1}^{p_0} \mathbb{P}\left(X = i\right) E\left[ln\frac{g}{f} \mid X = i\right] =$$

$$= \sum_{i=1}^{p_0} \pi_i^0 \int_{y_1, y_2 \in \mathbb{R}} ln \left(\frac{\sum_{j=1}^{p} \pi_j f_j \left(y_2 - F_j \left(y_1 \right) \right)}{\sum_{j=1}^{p_0} \pi_j^0 f_j^0 \left(y_2 - F_j^0 \left(y_1 \right) \right)} \right) f_i^0 \left(y_2 - F_i^0 \left(y_1 \right) \right) \mu_i \left(y_1 \right) dy_1 dy_2$$

where $\mu_i \left(y_1 \right)$ is the stationary measure of Y_1, conditionally to $X = i$ and finally, by Fubini,

$$E \left[ln(g) - ln(f) \right] =$$

$$= \int_{y_1, y_2 \in \mathbb{R}} ln \left(\frac{\sum_{j=1}^{p} \pi_j f_j \left(y_2 - F_j \left(y_1 \right) \right)}{\sum_{j=1}^{p_0} \pi_j^0 f_j^0 \left(y_2 - F_j^0 \left(y_1 \right) \right)} \right) \sum_{i=1}^{p_0} \pi_i^0 f_i^0 \left(y_2 - F_i^0 \left(y_1 \right) \right) \mu_i \left(y_1 \right) dy_1 dy_2$$

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

- $\mu_i \left(y_1 \right) = \mu \left(y_1 \right)$ for all $i \in \{1, ..., p_0\}$ which leads to mixtures of autoregressive models treated in Section 2.
- $F_j \left(y_1 \right)$ and $F_i^0 \left(y_1 \right)$ are constant for $j \in \{1, ..., p\}$, $i \in \{1, ..., p_0\}$, but this corresponds to hidden Markov models studied in Gassiat (2002).

In the general case, however, there is no reason for the last integral to be negative. Simulation results proved that the penalized estimate $\hat{p}$ diverges when the true model is, for instance, a two-regime Markov-switching autoregressive model. This means that the cost function considered as a generalization of the "marginal likelihood" does not have the good properties to be a contrast and the problem of estimating p_0 remains open in the general case of autoregressive Markov switching models.

Using the exact likelihood could be a possible direction to follow. Gassiat and Keribin (2000) proved the divergence of the likelihood ratio test statistic in the particular case of mixtures with Markov regime, but the consistency of some penalized-likelihood estimate should be obtained under suitable assumptions. Leroux (1992b), Ryden (1995) and Francq, Roussignol and Zakoian (2001) have already shown that the penalized-likelihood criterion does not underestimate the true number of regimes in the case of mixtures, hidden Markov models and, respectively, GARCH models with the coefficients depending on the state of an unobserved Markov chain. Extending their result to autoregressive models with Markov switching is immediate, the difficult part which remains open being to prove that the penalized-likelihood estimate does not overestimate the number of regimes.

References

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

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

[3] 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

[4] Dempster A.P., Laird N.M., Rubin D.B (1977) Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statist. Soc. (B)*, **39(1)**, 1-38

[5] 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

[6] Francq C., Roussignol M., Zakoian J-M. (2001) Conditional heteroskedasticity driven by hidden Markov chains, *Journal of Time Series Analysis*, **22**, 197-220

[7] 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
[8] Gassiat E. (2002) Likelihood ratio inequalities with applications to various mixtures, *Ann. Inst. Henri Poincaré*, **38**, 897-906
[9] Henna J. (1985) On estimating the number of constituents of a finite mixture of continuous distributions, *Ann. Inst. Statist. Math.*, **37**, 235-240
[10] Izenman A.J., Sommer C. (1988) Philatelic mixtures and multivariate densities, *Journal of the American Stat. Assoc.*, **83**, 941-953
[11] Keribin C. (2000) Consistent estimation of the order of mixture models, *Sankhya : The Indian Journal of Statistics*, **62**, 49-66
[12] Leroux B.G. (1992a) Maximum penalized likelihood estimation for independent and Markov-dependent mixture models, *Biometrics*, **48**, 545-558
[13] Leroux B.G. (1992b) Consistent estimation of a mixing distribution, *The Annals of Statistics,* **20**, 1350-1360
[14] Lindsay B.G. (1983) Moment matrices : application in mixtures, *The Annals of Statistics*, **17**, 722-740
[15] Liu X., Shao Y. (2003) Asymptotics for likelihood ratio tests under loss of identifiability, *The Annals of Statistics*, **31(3)**, 807-832
[16] Redner R.A., Walker H.F. (1984) Mixture densities, maximum likelihood and the EM algorithm, *SIAM Review*, **26(2)**, 195-239
[17] 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
[18] Rydén T. (1995) Estimating the order of hidden Markov models, *Statistics*, **26**, 345-354
[19] Teicher H. (1963) Identifiability of finite mixtures, *Ann. Math. Statist.*, **34(2)**, 1265-1269
[20] Van der Vaart A.W. (2000) *Asymptotic Statistics*, Cambridge University Press
[21] Yao J.F., Attali J.G. (2000) On stability of nonlinear AR processes with Markov switching, *Advances in Applied Probability*, **32 (2),** 394-407

	π_1^0	0.5			0.7			0.9		
	n	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$
$a_1^0=0.1$	200	0	20	0	0	20	0	0	18	2
$a_2^0=0.1$	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
$a_1^0=0.1$	200	0	20	0	0	19	1	1	19	0
$a_2^0=0.5$	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
$a_1^0=0.1$	200	0	20	0	0	20	0	4	16	0
$a_2^0=0.9$	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
$a_1^0=0.5$	200	0	19	1	0	18	2	0	20	0
$a_2^0=0.5$	500	0	20	0	0	20	0	0	18	2
	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
$a_1^0=0.5$	200	0	19	1	0	20	0	11	9	0
$a_2^0=0.9$	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
$a_1^0=0.9$	200	0	20	0	0	20	0	0	16	4
$a_2^0=0.9$	500	0	20	0	0	20	0	0	20	0
	1000	0	19	1	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 $b_1^0=1$, $b_2^0=-1$, $\sigma_1^0=\sigma_2^0=0.5$

APPENDIX

Proof of Theorem 1. The proof is an extension of Gassiat (2002). First, let us prove that $\hat{p}$ does not overestimate p_0 :

$$\mathbb{P}\left(\hat{p}>p_0\right)\leq\sum_{p=p_0+1}^{P}\mathbb{P}\left(T_n\left(p\right)>T_n\left(p_0\right)\right)=$$

$$=\sum_{p=p_0+1}^{P}\mathbb{P}\left(sup_{g\in\mathcal{G}_p}l_n\left(g\right)-a_n\left(p\right)>sup_{g\in\mathcal{G}_{p_0}}l_n\left(g\right)-a_n\left(p_0\right)\right)\leq$$

$$\leq\sum_{p=p_0+1}^{P}\mathbb{P}\left(\frac{1}{2}sup_{g\in\mathcal{G}_p}\frac{\left(\sum_{k=2}^{n}s_g\left(Y_{k-1},Y_k\right)\right)^2}{\sum_{k=2}^{n}\left(s_g\right)_{-}^{2}\left(Y_{k-1},Y_k\right)}>a_n\left(p\right)-a_n\left(p_0\right)\right)$$

	π_1^0	0.5			0.7			0.9		
	n	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$
$a_1^0=0.1$	200	20	0	0	20	0	0	20	0	0
$a_2^0=0.1$	500	18	2	0	18	2	0	20	0	0
	1000	14	6	0	9	11	0	11	9	0
	1500	6	14	0	4	16	0	5	15	0
	2000	5	15	0	0	20	0	1	19	0
$a_1^0=0.1$	200	12	8	0	13	7	0	20	0	0
$a_2^0=0.5$	500	11	19	0	6	14	0	18	2	0
	1000	0	20	0	1	19	0	14	6	0
	1500	0	20	0	0	20	0	8	12	0
	2000	0	20	0	0	20	0	7	13	0
$a_1^0=0.1$	200	0	20	0	4	16	0	17	3	0
$a_2^0=0.9$	500	0	20	0	0	20	0	9	11	0
	1000	0	20	0	0	20	0	9	11	0
	1500	0	20	0	0	20	0	4	16	0
	2000	0	20	0	0	20	0	0	20	0
$a_1^0=0.5$	200	18	2	0	20	0	0	19	1	0
$a_2^0=0.5$	500	20	0	0	19	1	0	19	1	0
	1000	14	6	0	13	7	0	10	10	0
	1500	9	11	0	5	15	0	5	15	0
	2000	3	17	0	0	20	0	3	17	0
$a_1^0=0.5$	200	9	11	0	11	9	0	20	0	0
$a_2^0=0.9$	500	0	20	0	7	13	0	19	1	0
	1000	0	20	0	0	20	0	19	1	0
	1500	0	20	0	0	20	0	18	2	0
	2000	0	20	0	0	20	0	14	6	0
$a_1^0=0.9$	200	20	0	0	19	1	0	19	1	0
$a_2^0=0.9$	500	20	0	0	18	2	0	17	3	0
	1000	14	6	0	7	13	0	11	9	0
	1500	7	13	0	5	15	0	3	17	0
	2000	6	14	0	0	20	0	0	20	0

TABLE 2. Results for $b_1^0=0.5$, $b_2^0=-0.5$, $\sigma_1^0=\sigma_2^0=0.5$

Under the hypothesis **(HS)**, there exists a unique strictly stationary solution Y_k which is also geometrically ergodic and this implies that Y_k is in particular geometrically β-mixing. Then, by remarking that

$$\beta_n^{(Y_{k-1},Y_k)}=\beta_{n-1}^{Y_k}$$

we obtain that the bivariate series (Y_{k-1},Y_k) is also strictly stationary and geometrically β-mixing.

This fact, together with the assumption on the ε-bracketing entropy of $\mathcal{S}$ with respect to the $\|\cdot\|_{L^2(\mu)}$ norm and the condition that $\mathcal{S}\subset\mathcal{L}_2(\mu)$ ensures that Theorem 4 in Doukan, Massart and Rio (1995) holds and

$$\left\{\frac{1}{\sqrt{n-1}}\sum_{k=2}^{n}s_g(Y_{k-1},Y_k)\mid g\in\mathcal{G}_p\right\}$$

is uniformly tight and verifies a functional central limit theorem. Then,

$$sup_{g\in\mathcal{G}_p}\frac{1}{n-1}\left(\sum_{k=2}^{n}s_g\left(Y_{k-1},Y_k\right)\right)^2=\mathcal{O}_{\mathbb{P}}\left(1\right)$$

On the other hand, $\mathcal{S}\subset\mathcal{L}_2\left(\mu\right)$, thus $\mathcal{S}^2\subset\mathcal{L}_1\left(\mu\right)$ and using the $\mathcal{L}_2$-entropy condition $\mathcal{S}_-^2=\left\{\left(s_g\right)_-^2,\ g\in\mathcal{G}_p\right\}$ is Glivenko-Cantelli. Since (Y_{k-1},Y_k) is ergodic and strictly stationary, we obtain the following uniform convergence in probability :

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

To finish the first part, let us prove that

$$inf_{g\in\mathcal{G}_p}\left\|\left(s_g\right)_-\right\|_2>0$$

If we suppose, on the contrary, that $inf_{g\in\mathcal{G}_p}\left\|\left(s_g\right)_-\right\|_2=0$, then there exists a sequence of functions $\left(s_{g_n}\right)_{n\geq1}$, $g_n\in\mathcal{G}_p$ such that $\left\|\left(s_{g_n}\right)_-\right\|_2\to0$. The L_2-convergence implies that $\left(s_{g_n}\right)_-\to0$ in L_1 and a.s. for a subsequence $s_{g_{n,k}}$. Since $\int s_{g_n}d\mu=0$ and $s_{g_n}=\left(s_{g_n}\right)_-+\left(s_{g_n}\right)_+$, where $\left(s_{g_n}\right)_+=max\left(0,\,s_{g_n}\right)$, we obtain that $\int\left(s_{g_n}\right)_+d\mu=-\int\left(s_{g_n}\right)_-d\mu=\int\left|\left(s_{g_n}\right)_-\right|d\mu$ and thus $\left(s_{g_n}\right)_+\to0$ in L_1 and a.s. for a subsequence $s_{g_{n,k'}}$. The hypothesis **(A4)** ensures the existence of a square-integrable dominating function for $\mathcal{S}$ and, finally, we get that a subsequence of s_{g_n} coverges to 0 a.s. and in L_2, which contradicts the fact that $\int s_g^2d\mu=1$ for every $g\in\mathcal{G}_p$, so that :

$$sup_{g\in\mathcal{G}_p}\frac{\left(\sum_{k=2}^{n}s_g\left(Y_{k-1},Y_k\right)\right)^2}{\sum_{k=2}^{n}\left(s_g\right)_-^2\left(Y_{k-1},Y_k\right)}=\mathcal{O}_{\mathbb{P}}\left(1\right)$$

Then, by the uniform tightness above and the hypothesis **(A1)**,

$$\mathbb{P}\left(\hat{p}>p_0\right)\longrightarrow_{n\to\infty}0$$

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

$$\mathbb{P}\left(\hat{p}<p_0\right)\leq\sum_{p=1}^{p_0-1}\mathbb{P}\left(T_n\left(p\right)>T_n\left(p_0\right)\right)\leq$$

$$\leq\sum_{p=1}^{p_0-1}\mathbb{P}\left(\frac{sup_{g\in\mathcal{G}_p}\left(l_n\left(g\right)-l_n\left(f\right)\right)}{n-1}>\frac{a_n\left(p\right)-a_n\left(p_0\right)}{n-1}\right)$$

Now, $l_n\left(g\right)-l_n\left(f\right)=\sum_{k=2}^{n}log\left(\frac{g\left(Y_{k-1},Y_k\right)}{f\left(Y_{k-1},Y_k\right)}\right)$ and under the hypothesis **(A3)**, the class of functions $\left\{log\frac{g}{f},\ g\in\mathcal{G}_p\right\}$ is $\mathbb{P}$-Glivenko-Cantelli (the general proof for a parametric family can be found in Van der Vaart (2000)) and since (Y_{k-1},Y_k) is ergodic and strictly stationary, we obtain the following uniform convergence in probability :

$$\frac{1}{n-1} sup_{g \in \mathcal{G}_p} \left(l_n(g) - l_n(f) \right) \longrightarrow sup_{g \in \mathcal{G}_p} \int log \frac{g}{f} f d\mu$$

Since $p < p_0$ and using assumption **(A2)**, the limit is negative. By hypothesis **(A1)**, $\frac{a_n(p) - a_n(p_0)}{n-1}$ converges to 0 when $n \to \infty$, so we finally have that $\mathbb{P}(\hat{p} < p_0) \to 0$ and the proof is done.

■

Proof of Proposition 2

Since the noise is gaussian and $\left|a_i^0\right| < 1$ for every $i \in \{1, ..., p_0\}$, the hypothesis **(HS)** is verified and, by Yao and Attali (2000), there exists a unique strictly stationary and geometrically ergodic solution, which in particular will be geometrically β-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 $\sigma = max_{i=1,...,} \sigma_i^0$, $\rho = max_{i=\overline{1,p_0}} \left|a_i^0\right| < 1$, $b = max_{i=\overline{1,p_0}} \left|b_i^0\right|$ and for $s \in \mathbb{N}^\star$, one has :

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

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

By taking the expectation,

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

and since $\rho < 1$ and the L_2-norm is dominated by the L_{2s}, we finally obtain

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

The exponential moment can be computed then by

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

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

■

Proof of Proposition 3

$$\left\| \frac{g}{f} - 1 \right\|_{L^2(\mu)} = \int_{y_1, y_2 \in \mathbb{R}^2} \left(\frac{g(y_1, y_2)}{f(y_1, y_2)} - 1 \right)^2 f(y_1, y_2) \, dy_2 d\mu(y_1) =$$

$$= \int_{y_1, y_2 \in \mathbb{R}^2} \frac{g^2(y_1, y_2)}{f(y_1, y_2)} dy_2 d\mu(y_1) - 1$$

By replacing $g(y_1, y_2) = \pi f_1(y_2 - F_1(y_1)) + (1 - \pi) f_2(y_2 - F_2(y_1))$ and using

$$2 f_1(y_2 - F_1(y_1)) f_2(y_2 - F_2(y_1)) \leq f_1^2(y_2 - F_1(y_1)) + f_2^2(y_2 - F_2(y_1))$$

one gets finally

$$\left\| \frac{g}{f} - 1 \right\|_{L^2(\mu)} \leq \int_{y_1, y_2 \in \mathbb{R}^2} \frac{\pi f_1^2(y_2 - F_1(y_1)) + (1 - \pi) f_2^2(y_2 - F_2(y_1))}{f(y_1, y_2)} dy_2 d\mu(y_1) - 1$$

The last term will be finite if, for instance,

$$\begin{cases} \int_{y_1, y_2 \in \mathbb{R}^2} \frac{f_1^2(y_2 - F_1(y_1))}{f(y_1, y_2)} dy_2 d\mu(y_1) < \infty \\ \int_{y_1, y_2 \in \mathbb{R}^2} \frac{f_2^2(y_2 - F_2(y_1))}{f(y_1, y_2)} dy_2 d\mu(y_1) < \infty \end{cases}$$

The last step is to replace f_1, f_2 and f^0 by centered gaussian densities with standard errors σ_1, σ_2 and, respectively, σ^0 and consider also $F_1(y) = a_1 y + b_1$, $F_2(y) = a_2 y + b_2$ and $F^0(y) = a^0 y + b^0$.

For $i \in \{1, 2\}$, each of the integrals above becomes :

$$\int_{y_1, y_2 \in \mathbb{R}^2} \frac{f_i^2(y_2 - F_i(y_1))}{f(y_1, y_2)} dy_2 d\mu(y_1) = \int_{y_1 \in \mathbb{R}} \left(\int_{y_2 \in \mathbb{R}} \frac{\sigma^0}{\sqrt{2\pi}\sigma_i^2} \cdot \right.$$

$$exp\left\{ - \left(\frac{1}{\sigma_i^2} - \frac{1}{2(\sigma^0)^2} \right) (y_2 - m(y_1))^2 \right\} dy_2 \left. \right) exp\left\{ \frac{(F_i(y_1) - F^0(y_1))^2}{2(\sigma^0)^2 - \sigma_i^2} \right\} d\mu(y_1)$$

where $m(y_1) = \frac{2(\sigma^0)^2 F_i(y_1) - \sigma_i^2 F^0(y_1)}{2(\sigma^0)^2 - \sigma_i^2}$

To have a sufficient condition, the integral in y_2 is finite if $\sigma_i^2 < 2(\sigma^0)^2$, as for the integral in y_1, using the existence of an exponential moment for Y_t, it is enough to have $\frac{(a_i - a^0)^2}{2(\sigma^0)^2 - \sigma_i^2} < \delta$.

■

Proof of Proposition 4

In the general case, the true and the possible conditional densities are

$$f(y_1, y_2) = \sum_{j=1}^{p_0} \pi_j^0 f_j^0 \left(y_2 - F_j^0(y_1) \right), \; g(y_1, y_2) = \sum_{i=1}^{p} \pi_i f_i(y_2 - F_i(y_1))$$

Then, the norm of the generalized score function can be written as

$$\left\| \frac{g}{f} - 1 \right\|_{L^2(\mu)} = \int_{y_1, y_2 \in \mathbb{R}^2} \frac{g^2(y_1, y_2)}{f(y_1, y_2)} dy_2 d\mu(y_1) - 1 =$$

$$= \int_{y_1, y_2 \in \mathbb{R}^2} \frac{\left(\sum_{i=1}^{p} \pi_i f_i(y_2 - F_i(y_1))\right)^2}{\sum_{j=1}^{p_0} \pi_j^0 f_j^0\left(y_2 - F_j^0(y_1)\right)} dy_2 d\mu(y_1) - 1$$

and by the inequality $\left(\sum_{i=1}^{p} \pi_i f_i(y_2 - F_i(y_1))\right)^2 \leq \sum_{i=1}^{p} \pi_i f_i^2(y_2 - F_i(y_1))$, we will obtain that the integral is finite if

$$\int_{y_1, y_2 \in \mathbb{R}^2} \frac{f_i^2(y_2 - F_i(y_1))}{\sum_{j=1}^{p_0} \pi_j^0 f_j^0\left(y_2 - F_j^0(y_1)\right)} dy_2 d\mu(y_1) < \infty$$

for all $i \in \{1, ..., p\}$. On the other hand, since $\sum_{j=1}^{p_0} \pi_j^0 f_j^0\left(y_2 - F_j^0(y_1)\right) \geq \pi_k^0 f_k^0\left(y_2 - F_k^0(y_1)\right)$ for every $k \in \{1, ..., p_0\}$, the condition will become similar to that in Proposition 3. We will have that the generalized score function is well defined if for every $i \in \{1, ..., p\}$, there exists $k \in \{1, ..., p_0\}$ such that

$$\int_{y_1, y_2 \in \mathbb{R}^2} \frac{f_i^2(y_2 - F_i(y_1))}{f_k^0(y_2 - F_k^0(y_1))} dy_2 d\mu(y_1) < \infty$$

which is verified if $\sigma_i^2 < 2\left(\sigma_k^0\right)^2$ and $|a_i - a_k^0| < \sqrt{\delta\left(2\left(\sigma_k^0\right)^2 - \sigma_i^2\right)}$.

■

Proof of Proposition 5

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

- for $i \in \{1, ..., p_0\}$ and $j \in \{t_{i-1} + 1, ..., t_i\}$, $\frac{\partial\left(\frac{g}{f} - 1\right)}{\partial \theta_j}\left(\phi_t^0, \psi_t\right) = \pi_i^0 q_j g_i'$
- for $i \in \{1, ..., p_0 - 1\}$,

$$\frac{\partial\left(\frac{g}{f} - 1\right)}{\partial s_i}\left(\phi_t^0, \psi_t\right) = \sum_{j=t_{i-1}+1}^{t_i} q_j g_{\theta_i^0} - \sum_{j=t_{p_0-1}+1}^{t_{p_0}} q_j g_{\theta_{p_0}^0} = g_{\theta_i^0} - g_{\theta_{p_0}^0}$$

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

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

It remains to prove that the rest is $o\left(\left\|\phi_t-\phi_t^0\right\|\right)$ and this will follow if the third-order derivative is uniformly bounded in Φ by a map with finite integral and using the linear independence in Lemma 1. As it can be seen easily that this derivative can be expressed in terms of g_j, $j=1,...,p$ and their partial derivatives of order one, two and three in θ_j , a sufficient condition to verify is that latter are uniformly bounded in Φ by an integrable map.

Let us check the last assertion. Since $\theta_j=(a_j,b_j,\sigma_j)$, the partial derivatives of order one of g_j, $j=1,...,p$ are :

$$\frac{\partial g_j}{\partial\theta_j}(y_1,y_2)=\frac{1}{\sqrt{2\pi}f(y_1,y_2)}\left(\sqrt{2\pi}\frac{\partial f_j}{\partial a_j}(y_1,y_2),\sqrt{2\pi}\frac{\partial f_j}{\partial b_j}(y_1,y_2),\sqrt{2\pi}\frac{\partial f_j}{\partial\sigma_j}(y_1,y_2)\right)$$

where

$$\sqrt{2\pi}\frac{\partial f_j}{\partial a_j}(y_1,y_2)=\frac{y_1}{\sigma_j^3}(y_2-F_j(y_1))\,e^{-\frac{1}{2\sigma_j^2}(y_2-F_j(y_1))^2}$$

$$\sqrt{2\pi}\frac{\partial f_j}{\partial b_j}(y_1,y_2)=\frac{1}{\sigma_j^3}(y_2-F_j(y_1))\,e^{-\frac{1}{2\sigma_j^2}(y_2-F_j(y_1))^2}$$

$$\sqrt{2\pi}\frac{\partial f_j}{\partial\sigma_j}(y_1,y_2)=\left[\frac{(y_2-F_j(y_1))^2}{\sigma_j^4}-\frac{1}{\sigma_j^2}\right]e^{-\frac{1}{2\sigma_j^2}(y_2-F_j(y_1))^2}$$

The Hessians of g_j, $j=1,...,p$ can be written as :

$$\frac{\partial^2 g_j}{\partial\theta_j^2}=\frac{1}{\sqrt{2\pi}f(y_1,y_2)}\begin{pmatrix}\sqrt{2\pi}\frac{\partial^2 f_j}{\partial a_j^2}(y_1,y_2) & \sqrt{2\pi}\frac{\partial^2 f_j}{\partial a_j\partial b_j}(y_1,y_2) & \sqrt{2\pi}\frac{\partial^2 f_j}{\partial a_j\partial\sigma_j}(y_1,y_2)\\ \sqrt{2\pi}\frac{\partial^2 f_j}{\partial a_j\partial b_j}(y_1,y_2) & \sqrt{2\pi}\frac{\partial^2 f_j}{\partial b_j^2}(y_1,y_2) & \sqrt{2\pi}\frac{\partial^2 f_j}{\partial b_j\partial\sigma_j}(y_1,y_2)\\ \sqrt{2\pi}\frac{\partial^2 f_j}{\partial a_j\partial\sigma_j}(y_1,y_2) & \sqrt{2\pi}\frac{\partial^2 f_j}{\partial b_j\partial\sigma_j}(y_1,y_2) & \sqrt{2\pi}\frac{\partial^2 f_j}{\partial\sigma_j^2}(y_1,y_2)\end{pmatrix}$$

with

$$\sqrt{2\pi}\frac{\partial^2 f_j}{\partial a_j^2}(y_1,y_2)=\left[-\frac{y_1^2}{\sigma_j^3}+\frac{y_1^2}{\sigma_j^5}(y_2-F_j(y_1))^2\right]e^{-\frac{1}{2\sigma_j^2}(y_2-F_j(y_1))^2}$$

$$\sqrt{2\pi}\frac{\partial^2 f_j}{\partial a_j\partial b_j}(y_1,y_2)=\left[-\frac{y_1}{\sigma_j^3}+\frac{y_1}{\sigma_j^5}(y_2-F_j(y_1))^2\right]e^{-\frac{1}{2\sigma_j^2}(y_2-F_j(y_1))^2}$$

$$\sqrt{2\pi}\frac{\partial^2 f_j}{\partial a_j\partial\sigma_j}(y_1,y_2)=\left[-\frac{3y_1}{\sigma_j^4}(y_2-F_j(y_1))+\frac{y_1}{\sigma_j^6}(y_2-F_j(y_1))^3\right]e^{-\frac{1}{2\sigma_j^2}(y_2-F_j(y_1))^2}$$

$$\sqrt{2\pi}\frac{\partial^2 f_j}{\partial b_j^2}(y_1,y_2)=\left[-\frac{1}{\sigma_j^3}+\frac{1}{\sigma_j^5}(y_2-F_j(y_1))^2\right]e^{-\frac{1}{2\sigma_j^2}(y_2-F_j(y_1))^2}$$

$$\sqrt{2\pi}\frac{\partial^2 f_j}{\partial b_j\partial\sigma_j}(y_1,y_2)=\left[-\frac{3}{\sigma_j^4}(y_2-F_j(y_1))+\frac{1}{\sigma_j^6}(y_2-F_j(y_1))^3\right]e^{-\frac{1}{2\sigma_j^2}(y_2-F_j(y_1))^2}$$

$$\sqrt{2\pi}\frac{\partial^2 f_j}{\partial \sigma_j^2}(y_1,y_2)=\left[\frac{1}{\sigma_j^7}(y_2-F_j(y_1))^4-\frac{5}{\sigma_j^5}(y_2-F_j(y_1))^2+\frac{2}{\sigma_j^3}\right]e^{-\frac{1}{2\sigma_j^2}(y_2-F_j(y_1))^2}$$

Concerning the third-order partial derivatives of g_j, $j=1,...,p$, we shall remark, using the expressions above, that they will be written as linear combinations of powers of $\frac{1}{\sigma_j}$, y_1 and $y_2-F_j(y_1)$, multiplied by the exponential term $e^{-\frac{1}{2\sigma_j^2}(y_2-F_j(y_1))^2}$. This remark, together with the assumptions in Section 3.2, imply the existence of a finite integral function which dominates g_j, $\frac{\partial g_j}{\partial \theta_j}$, $\frac{\partial^2 g_j}{\partial \theta_j^2}$, $\frac{\partial^3 g_j}{\partial \theta_j^3}$, $j=1,...,p$ uniformly in θ_j.

The last thing we need to prove, the inverse implication being obvious, is

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

First, let us state and prove the following lemma :

Lemma 1

The family of functions

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

is linearly independent.

Proof of Lemma 1

To prove the linear independence, we need the following equivalence which holds whenever $\theta_i=(a_i,b_i,\sigma_i)$, $i=1,...,p_0$ are distinct :

$$\sum_{i=1}^{p_0}P_i(y_1,y_2)\,e^{-\frac{1}{2\sigma_i^2}(y_2-F_i(y_1))^2}=0,\ (\forall)\,y_1,y_2\Leftrightarrow P_i(y_1,y_2)=0,\ i=1,...,p_0$$

where $P_i(y_1,y_2)$ are polynomials of y_1 and y_2 and $F_i(y_1)=a_iy_1+b_i$ for $i=1,...,p_0$.

Now, let us consider the linear combination

$$\textstyle\sum_{i=1}^{p_0}\alpha_i g_{\theta_i^0}+\sum_{i=1}^{p_0}\beta_i^T g_i'+\sum_{i=1}^{p_0}\gamma_i^T g_i''$$

with $g_i'=\left(\frac{\partial g_{\theta_i^0}}{\partial a_i},\frac{\partial g_{\theta_i^0}}{\partial b_i},\frac{\partial g_{\theta_i^0}}{\partial \sigma_i}\right)^T$, $g_i''=\left(\frac{\partial^2 g_{\theta_i^0}}{\partial a_i^2},\frac{\partial^2 g_{\theta_i^0}}{\partial b_i^2},\frac{\partial^2 g_{\theta_i^0}}{\partial \sigma_i^2},\frac{\partial^2 g_{\theta_i^0}}{\partial a_i\partial b_i},\frac{\partial^2 g_{\theta_i^0}}{\partial a_i\partial \sigma_i},\frac{\partial^2 g_{\theta_i^0}}{\partial b_i\partial \sigma_i}\right)^T$, $\beta_i^T=(\beta_{i,1},\beta_{i,2},\beta_{i,3})$ and $\gamma_i^T=(\gamma_{i,1},\gamma_{i,2},\gamma_{i,3},\gamma_{i,4},\gamma_{i,5},\gamma_{i,6})$.

Then,

$$\sum_{i=1}^{p_0}\alpha_i g_{\theta_i^0}+\sum_{i=1}^{p_0}\beta_i^T g_i'+\sum_{i=1}^{p_0}\gamma_i^T g_i''=0\Leftrightarrow\sum_{i=1}^{p_0}P_i(y_1,y_2)\,e^{-\frac{1}{2\sigma_i^2}(y_2-F_i(y_1))^2}=\sum_{i=1}^{p_0}\alpha_i$$

where

$$P_i(y_1,y_2)=\frac{1}{f(y_1,y_2)}\frac{1}{\sqrt{2\pi}\sigma_i}\left[\alpha_i+\beta_{i,1}\frac{y_1}{\sigma_i^2}(y_2-F_i(y_1))+\beta_{i,2}\frac{1}{\sigma_i^2}(y_2-F_i(y_1))+\right.$$

$$+\beta_{i,3}\left(\frac{1}{\sigma_i^3}(y_2-F_i(y_1))^2-\frac{1}{\sigma_i}\right)+\gamma_{i,1}\left(-\frac{y_1^2}{\sigma_i^2}+\frac{y_1^2}{\sigma_i^4}(y_2-F_i(y_1))^2\right)+$$

$$+\gamma_{i,2}\left(-\frac{1}{\sigma_i^2}+\frac{1}{\sigma_i^4}(y_2-F_i(y_1))^2\right)+\gamma_{i,3}\left(\frac{1}{\sigma_i^6}(y_2-F_i(y_1))^4-\frac{5}{\sigma_i^4}(y_2-F_i(y_1))^2+\frac{2}{\sigma_i^2}\right)+$$

$$+\gamma_{i,4}\left(-\frac{y_1}{\sigma_i^2}+\frac{y_1}{\sigma_i^4}(y_2-F_i(y_1))^2\right)+\gamma_{i,5}\left(-\frac{3y_1}{\sigma_i^3}(y_2-F_i(y_1))+\frac{y_1}{\sigma_i^5}(y_2-F_i(y_1))^3\right)+$$

$$\left.+\gamma_{i,6}\left(-\frac{3}{\sigma_i^3}(y_2-F_i(y_1))+\frac{1}{\sigma_i^5}(y_2-F_i(y_1))^3\right)\right]$$

If now $y_2\to\infty$, the left term in the equality vanishes and then $\sum_{i=1}^{p_0}\alpha_i=0$ and by the remark in the beginning of the proof, we obtain that $P_i(y_1,y_2)=0$ for $i=1,...,p_0$. By coefficient identification for y_2^4, $y_1y_2^3$, y_2^3, $y_1^2y_2^2$ and $y_1y_2^2$ we obtain immediately that $\gamma_{i,3}=\gamma_{i,5}=\gamma_{i,6}=\gamma_{i,1}=\gamma_{i,4}=0$ and it remains that

$$\alpha_i+\beta_{i,1}\frac{y_1}{\sigma_i^2}(y_2-F_i(y_1))+\beta_{i,2}\frac{1}{\sigma_i^2}(y_2-F_i(y_1))+\beta_{i,3}\left(\frac{1}{\sigma_i^3}(y_2-F_i(y_1))^2-\frac{1}{\sigma_i}\right)+$$

$$+\gamma_{i,2}\left(-\frac{1}{\sigma_i^2}+\frac{1}{\sigma_i^4}(y_2-F_i(y_1))^2\right)=0,\ (\forall)\,y_1,y_2$$

Here again, by coefficient identification we will have that $\gamma_{i,2}=-\frac{\beta_{i,3}}{\sigma_i}$ and $\beta_{i,1}=\beta_{i,2}=\alpha_i=0$ and the proof is done.

■

Let us go back now to what remains to prove in Proposition 5. For ψ_t fixed, let us consider ϕ_t verifying

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

The two terms can be replaced by their expressions in Proposition 5 and we obtain :

$$\sum_{i=1}^{p_0}\pi_i^0\left(\sum_{j=t_{i-1}+1}^{t_i}q_j\theta_j-\theta_i^0\right)^T g_i'+\sum_{i=1}^{p_0}s_i g_{\theta_i^0}+$$

$$+\sum_{i=1}^{p_0}\left[2s_i\left(\sum_{j=t_{i-1}+1}^{t_i}q_j\theta_j-\theta_i^0\right)^T g_i'+\pi_i^0\sum_{j=t_{i-1}+1}^{t_i}q_j\left(\theta_j-\theta_i^0\right)^T g_i''\left(\theta_j-\theta_i^0\right)\right]=0$$

With lemma 1, the equality above holds iff :

- the coefficients of $g_{\theta_i^0}$ are zero, which implies $s_i=0$, $i=1,...,p_0-1$
- the coefficients of $\frac{\partial^2 g_{\theta_i^0}}{\partial a_i^2}$ and $\frac{\partial^2 g_{\theta_i^0}}{\partial \sigma_i^2}$ are zero, then for all $i=1,...,p_0$

$$\pi_i^0 \sum_{j=t_{i-1}+1}^{t_i} q_j \left(a_j - a_i^0\right)^2 = \pi_i^0 \sum_{j=t_{i-1}+1}^{t_i} q_j \left(\sigma_j - \sigma_i^0\right)^2 = 0$$

and since π_i^0 and $q_j = \frac{\pi_j}{\sum_{l=t_{i-1}+1}^{t_i} \pi_l}$ were supposed stricly positive, we will have $a_j = a_i^0$ and $\sigma_j = \sigma_i^0$ for all $i = 1, ..., p_0$, $j = t_{i-1} + 1, ..., t_i$.
When replacing s_i, a_j and σ_j, the equality becomes

$$\sum_{i=1}^{p_0} \pi_i^0 \left(\sum_{j=t_{i-1}+1}^{t_i} q_j b_j - b_i^0 \right) \frac{\partial g_{\theta_i^0}}{\partial b_i} + \sum_{i=1}^{p_0} \pi_i^0 \sum_{j=t_{i-1}+1}^{t_i} q_j \left(b_j - b_i^0\right)^2 \frac{\partial^2 g_{\theta_i^0}}{\partial b_i^2} = 0$$

and by the linear independence $b_j = b_i^0$ for all $i = 1, ..., p_0$, $j = t_{i-1} + 1, ..., t_i$.

■

Proof of Proposition 6

The idea of this proof is to bound $\mathcal{N}_{[]}\left(\varepsilon, \mathcal{S}_0, \|\cdot\|_2\right)$ by the number of ε-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 developpement of the density ratio :

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

Then, by remarking that the first two terms in the Taylor expansion are linear combinations of $g_{\theta_i^0}$, g'_i, g''_i, $i = 1, ..., p_0$, the density ratio can be written also as :

$$\frac{g_{(\phi_t,\psi_t)}}{f} - 1 = \sum_{i=1}^{p_0} \alpha_i g_{\theta_i^0} + \sum_{i=1}^{p_0} \beta_i^T g_{i,1} + \sum_{i=1}^{p_0} \gamma_i^T g_{i,2} + o\left(D\left(\phi_t, \psi_t\right)\right)$$

where $\beta_i^T = (\beta_{i,1}, \beta_{i,2}, \beta_{i,3})$, $\gamma_i^T = (\gamma_{i,1}, \gamma_{i,2}, \gamma_{i,3}, \gamma_{i,4}, \gamma_{i,5})$, $g_{i,1} = \left(\frac{\partial g_{\theta_i^0}}{\partial a_i}, \frac{\partial g_{\theta_i^0}}{\partial b_i}, \frac{\partial^2 g_{\theta_i^0}}{\partial b_i^2} + \frac{1}{\sigma_i}\frac{\partial g_{\theta_i^0}}{\partial \sigma_i}\right)^T$ and $g_{i,2} = \left(\frac{\partial^2 g_{\theta_i^0}}{\partial a_i^2}, \frac{\partial^2 g_{\theta_i^0}}{\partial \sigma_i^2}, \frac{\partial^2 g_{\theta_i^0}}{\partial a_i \partial b_i}, \frac{\partial^2 g_{\theta_i^0}}{\partial a_i \partial \sigma_i}, \frac{\partial^2 g_{\theta_i^0}}{\partial b_i \partial \sigma_i}\right)^T$.

Now, using the linear independence in Lemma 1, there exists $m > 0$ such that for every $\left(\alpha_i, \beta_i^T, \gamma_i^T, i = \overline{1, p_0}\right)$ of norm 1,

$$\left\| \sum_{i=1}^{p_0} \alpha_i g_{\theta_i^0} + \sum_{i=1}^{p_0} \beta_i^T g_{i,1} + \sum_{i=1}^{p_0} \gamma_i^T g_{i,2} \right\|_{L^2(\mu)} \geq m$$

At the same time, since

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

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

$$\mathcal{H}=\left\{\sum_{i=1}^{p_0}\left(\alpha_i g_{\theta_i^0}+\beta_i^T g_i'+\gamma_i^T g_i''\right)+o\left(1\right),\left\|\left(\alpha_i,\beta_i^T,\gamma_i^T,i=\overline{1,p_0}\right)\right\|\leq\frac{1}{m}\right\}$$

and then obviously $\mathcal{N}_{[]}\left(\varepsilon,\mathcal{H},\|\cdot\|_2\right)=\mathcal{O}\left(\frac{1}{\varepsilon}\right)^{9p_0}$.

■