ESTIMATING THE NUMBER OF REGIMES OF NON-LINEAR AUTOREGRESSIVE MODELS.

J. Rynkiewicz
SAMM Université Paris 1
90 Rue de Tolbiac, 75013 Paris, France

ABSTRACT

Autoregressive regime-switching models are being widely used in modelling financial and economic time series. When the number of regimes is fixed statistical inference is relatively straightforward and asymptotic properties of the estimates may be established. However, the problem of selecting 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, for example, likelihood ratio test statistic is no longer convergent to a χ^2-distribution. The problem we address in this paper is how to select number of regimes without knowing the form of the noise. One possible method to answer this problem is to consider penalized criteria. Recently, consistency of a modified BIC criterion was recently proven in the framework of likelihood criterion for linear switching models. We extend these results to mixtures of nonlinear autoregressive models with mean square error criterion and prove consistency of a penalized estimate for number of regimes under some regularity conditions. As an illustration, we use this theoretical result to propose and compare effective criteria to find the true number of regimes on a simple simulation.

KEYWORDS: time series, switching regimes, mean square error, asymptotic statistic, models selection.
MSC: 62F12, 62M10.

RESUMEN

Los modelos autorregresivos con cambio de régimen son ampliamente utilizados en las series temporales de modelos financieros y económicos. Cuando se fija el número de regímenes la inferencia estadística es relativamente sencilla y las propiedades asintóticas de los estimadores pueden ser establecidas. El problema de seleccionar el número de regímenes es mucho menos obvio y no ha sido completamente resuelto todavía. Cuando el número de regímenes es desconocido, surgen problemas de identificabilidad y, por ejemplo, el estadígrafo de la prueba del cociente de verosimilitud ya no converge a la distribución Chi-cuadrado. El problema abordado en este artículo es la selección del número de regímenes sin conocer la forma del ruido. Un método posible para resolver este problema consiste en considerar criterios penalizados. La consistencia de un criterio BIC modificado fue demostrado recientemente en el marco del criterio de verosimilitud para modelos lineales con cambio de régimen. En este trabajo extendemos estos resultados a las mezclas de modelos autorregresivos no lineales con criterio de error cuadrático medio y demostramos la consistencia de un estimador penalizado para el número de regímenes bajo algunas condiciones de regularidad. A modo de ejemplo, utilizamos este resultado teórico para proponer y comparar criterios eficaces para encontrar el verdadero número de regímenes en una simulación simple.

1. INTRODUCTION

A time series is a sequence of data points, measured typically at successive time instants spaced at uniform time intervals. The future behavior of the time series is partly determined by its past values, throughout

this paper, we shall consider that number of lags for the past is known and, for ease of writing, we shall set number of lags equal to one, the extension to l time-lags being immediate. Here, we assume that behavior of time series is not always the same but switches between different dynamics called "regimes" as in Hamilton (1989)[5]. Let X_t be a discrete random variable representative of the regimes. We consider the real-valued time series Y_t, $t \in \mathbb{Z}$, which matches the following model

$$(1) \qquad Y_t = F_{\theta^0_{X_t}}(Y_{t-1}) + \varepsilon_t,$$

where

- $(X_t)_{t\in\mathbb{Z}}$ is an iid sequence of random variables valued in a finite space $\{1, ..., p_0\}$ and with probability distribution π^0 ;
- for every $i \in \{1, ..., p_0\}$, $F_{\theta^0_i}(y) \in \mathcal{F}$ and
  $\mathcal{F} = \left\{F_\theta,\ \theta \in \Theta,\ \Theta \subset \mathbb{R}^l\right\}$, where Θ is a compact set, is the family of possible regression functions.
  We suppose throughout the rest of the paper that $F_{\theta^0_i}$ are sublinear, that is they are continuous and there exist $\left(a^0_i, b^0_i\right) \in \mathbb{R}^2_+$ such that $\forall y \in \mathbb{R}$, $\left|F_{\theta^0_i}(y)\right| \leq a^0_i |y| + b^0_i$;
- $(\varepsilon_t)_{t\in\mathbb{Z}}$ is an identically distributed and independent sequence of centered noise.

We need the following technical hypothesis which implies, according to Yao and Attali (2000)[6], strict stationarity geometric ergodicity and β-mixing for Y_t:

$$\textbf{(HS)} \quad \sum_{i=1}^{p_0} \pi^0_i \left|a^0_i\right|^s < 1$$

Let us remark that hypothesis **(HS)** does not request every component to be stationary and that it allows non-stationary "regimes" as long as they do not appear too often.
Let us briefly recall the definition of β-mixing processes which will be needed hereafter. For a more detailed review, refer to Doukhan (1995) [2] and Bradley (2005)[1].
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 β-mixing coefficients

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

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

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

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 β-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 stationary, β-mixing, process (Y_k).

2. ESTIMATION OF THE NUMBER OF REGIMES

An important question is whether there are switches in the time series and if it is the case, how many. This question is very similar to selection of the dimension of models and a rather natural idea seems to use information criteria as usual for models selection. However, justification of such technic is not obvious because if the number of regimes is over-estimated the model is no more identifiable (see for example Oltéanu and Rynkiewicz (2012) [7]) and asymptotic behavior of estimator is very different than in the identifiable case.
Let us consider an observed sample $\{y_1, ..., y_n\}$ of the time series $(Y_t)_{t\in\mathbb{Z}}$. Then, for every variable Y_t, the conditional expectation with respect to the previous observation y_{t-1} and marginally in X_t is

$$E\left(Y_t \mid y_{t-1}\right) = \sum_{i=1}^{p_0} \pi_i^0 F_{\theta_i^0}\left(y_{t-1}\right) := g^0\left(y_{t-1}\right)$$

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

$$\mathcal{G}_P = \bigcup_{p=1}^{P} \mathcal{G}_p,\ \mathcal{G}_p = \left\{ g \mid g\left(y\right) = \sum_{i=1}^{p} \pi_i F_{\theta_i}\left(y\right), \pi_i \geq 0, \sum_{i=1}^{p} \pi_i = 1 \text{ and } \theta_i \in \Theta \right\},$$

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

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

and let $p_0 = p\left(g^0\right)$ be the true number of regimes. In order to construct a criterion to guess p_0, let us define estimate $\hat{p}$ as the argument $p \in \{1, ..., P\}$ minimizing penalized criterion

$$T_n\left(p\right) = \inf_{g\in\mathcal{G}_p} E_n\left(g\right) - a_n\left(p\right) \tag{2}$$

where

$$E_n\left(g\right) = \frac{1}{2}\sum_{t=2}^{n}\left(y_t - g(y_{t-1})\right)^2$$

with $a_n\left(p\right)$ a penalty term. We will show that, although the behavior of this estimator is not entirely determined, it can be upper bounded and penalized criterion will nevertheless select asymptotically the true number of regimes p_0.

3. CONVERGENCE OF THE PENALIZED MEAN SQUARE ESTIMATE

First, let us set some definitions.

Generalized score function. For $\lambda > 0$, let us define the generalized derivative function :

$$d_g^\lambda(Y_t, Y_{t-1}) = \frac{\frac{e^{-\lambda(Y_t - g(Y_{t-1}))^2} - e^{-\lambda(Y_t - g^0(Y_{t-1}))^2}}{e^{-\lambda(Y_t - g^0(Y_{t-1}))^2}}}{\|\frac{e^{-\lambda(Y_t - g(Y_{t-1}))^2} - e^{-\lambda(Y_t - g(Y_{t-1}))^2}}{e^{-\lambda(Y_t - g^0(Y_{t-1}))^2}}\|_2}$$
$$= \frac{e^{-\lambda\left((Y_t - g(Y_{t-1}))^2 - (Y_t - g^0(Y_{t-1}))^2\right)} - 1}{\|e^{-\lambda\left((Y_t - g(Y_{t-1}))^2 - (Y_t - g^0(Y_{t-1}))^2\right)} - 1\|_2}$$

and let us define $\left(d_g^\lambda\right)_-(Y_t, Y_{t-1}) = \min\left\{0, d_g^\lambda(Y_t, Y_{t-1})\right\}$.
For now, let us assume that d_g^λ is well defined, this point will be discuss later. We can state the following inequality:
Inequality:
for $\lambda > 0$,

$$\sup_{g \in \mathcal{G}_p} n \cdot \left(E_n(g^0) - E_n(g)\right) \leq \frac{1}{2\lambda} \sup_{g \in \mathcal{G}_p} \frac{\sum_{t=1}^n d_g^\lambda(y_t, y_{t-1})}{\sum_{t=1}^n \left(d_g^\lambda\right)_-^2 (y_t, y_{t-1})} \tag{3.1}$$

Proof:
We have

$$\begin{aligned}
&n \cdot \left(E_n(g^0) - E_n(g)\right) = \\
&\tfrac{1}{\lambda} \textstyle\sum_{t=1}^n \log\left(1 + \|\frac{e^{-\lambda(Y_t - g(Y_{t-1}))^2} - e^{-\lambda(Y_t - g^0(Y_{t-1}))^2}}{e^{-\lambda(Y_t - g^0(Y_{t-1}))^2}}\|_2 d_g^\lambda(y_t, y_{t-1})\right) \\
&\leq \sup_{0 \leq p \leq \|\frac{e^{-\lambda(Y_t - g(Y_{t-1}))^2} - e^{-\lambda(Y_t - g^0(Y_{t-1}))^2}}{e^{-\lambda(Y_t - g^0(Y_{t-1}))^2}}\|_2} \tfrac{1}{\lambda} \textstyle\sum_{t=1}^n \log\left(1 + p d_g^\lambda(y_t, y_{t-1})\right) \\
&\leq \sup_{p \geq 0} \tfrac{1}{\lambda}\left(p \textstyle\sum_{t=1}^n d_g^\lambda(y_t, y_{t-1}) - \frac{p^2}{2} \sum_{t=1}^n \left(d_g^\lambda\right)_-^2 (y_t, y_{t-1})\right).
\end{aligned}$$

Since for any real number u, $\log(1+u) \leq u - \frac{1}{2}u_-^2$. Finally, replacing p by the optimal value, we found

$$n \cdot \left(E_n(g^0) - E_n(g)\right) \leq \tfrac{1}{2\lambda} \frac{\sum_{t=1}^n d_g^\lambda(y_t, y_{t-1})}{\sum_{t=1}^n \left(d_g^\lambda\right)_-^2 (y_t, y_{t-1})}$$

■

This inequality allows to prove the tightness of $n \cdot \left(E_n(g^0) - E_n(g)\right)$ under simple assumptions, but now let us recall a definition needed later.

Donsker class

- 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, β-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 c?dl?g extension of β_n by considering $\beta(u) = \beta_{[u]}$ and $\beta_0 = 1$
  - $\varphi^{-1}(u) = \inf\{t \in \mathbb{R},\ \varphi(t) \leq u\}$, if φ 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 a set of functions $\mathcal{S}$ endowed with the norm $\|\cdot\|_{2,\beta}$. For every $\varepsilon > 0$, we define an ε-bracket by $[l, u] = \{f \in \mathcal{F}, l \leq f \leq u\}$ such that $\|u - l\|_{2,\beta} < \varepsilon$. The ε-bracketing entropy is

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

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

With the previous notations, we introduce the following assumption:

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

Then, according to Doukhan, Massart and Rio (1995)[3], the set $\mathcal{S}$ is Donsker.

Main result

The consistency of $\widehat{p}$ is given by the next result, which in an extension of a result of Gassiat (2002) [4]:

Theorem *: Consider the model (Y_t, Y_{t-1}) defined by (1) and the estimator $\hat{p}$ minimizing the penalized 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 verifies the weak identifiability assumption:*

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

(A3) *It exists $\lambda > 0$ so that $\left\{d_g^\lambda, g \in \mathcal{G}_P\right\}$ is a Donsker class.*

Then, under hypothesis ***(HS)*** *and* ***(A1)-(A3)****, $\hat{p} \to p_0$ in probability.*

Proof:

By applying the inequality,

$$\begin{array}{l} P(\hat{p} > p^0) \leq \sum_{p=p^0+1}^{P} P\left(T_n(p) \leq T_n(p^0)\right) = \\ \sum_{p=p^0+1}^{P} P\left(n\left(\inf_{g\in\mathcal{G}_{p^0}} E_n(g) - \inf_{g\in\mathcal{G}_p} E_n(g)\right) \leq n\left(a_n(p) - a_n(p^0)\right)\right) \leq \\ \sum_{p=p^0+1}^{P} P\left(\frac{1}{\lambda}\sup_{g\in\mathcal{G}_p} \frac{\sum_{i=1}^n d_g^\lambda(x_i,y_i)}{\sum_{i=1}^n \left(d_g^\lambda\right)_-^2(x_i,y_i)} \geq n\left(a_n(p) - a_n(p^0)\right)\right) \end{array}$$

Now, under **(A3)**

$$sup_{g\in\mathcal{G}_p} \frac{1}{n}\left(\sum_{i=1}^{n} d_g^\lambda(x_i, y_i)\right)^2 = O_P(1)$$

where, $O_p(1)$ means bounded in probability. Moreover, under **(A3)** the set $\left\{\left(d_g^\lambda(x_i,y_i)\right)^2\right\}$ is Glivenko-Cantelli (the set admits an uniform law of large numbers see van der Vaart (2000) [8]). Hence

$$\inf_{g\in\mathcal{G}_p}\frac{1}{n}\sum_{i=1}^{n}\left(d_g^\lambda(x_i,y_i)\right)_-^2 \overset{n\to\infty}{\longrightarrow} \inf_{g\in\mathcal{G}_p}\parallel\left(d_g^\lambda(X,Y)\right)_-\parallel_2^2$$

But $\inf_{g\in\mathcal{G}_p}\parallel\left(d_g^\lambda(X,Y)\right)_-\parallel_2>0$, since the random variable $d_g^\lambda(X,Y)$ is centered and $\|d_g^\lambda(X,Y)\|_2=1$. Then, we get :

$$\frac{1}{\lambda}\sup_{g\in\mathcal{G}_p}\frac{\sum_{i=1}^n d_g^\lambda(x_i,y_i)}{\sum_{i=1}^n\left(d_g^\lambda\right)_-^2(x_i,y_i)}=O_P(1)$$

and $P(\hat{p}>p^0)$ tends to 0 as n tends to infinity. Finally,

$$P(\hat{p}<p^0)\leq\sum_{p=1}^{p^0-1}P\left(\inf_{g\in\mathcal{G}_p}\frac{E_n(g)-E_n(g^0)}{n}\leq\frac{a_n(p)-a_n(p^0)}{n}\right)$$

and $\inf_{g\in\mathcal{G}_p}\frac{E_n(g)-E_n(g^0)}{n}$ converges in probability to

$$\inf_{g\in\mathcal{G}_p}E\left(E_n(g)-E_n(g^0)\right)>0$$

since $p<p^0$, so $\hat{p}\overset{P}{\longrightarrow}p^0$ ■

The assumption **(A1)** is fairly standard for models selection. BIC criterion, for example, fulfills this condition. Note that weak identification assumption **(A2)** does not allowed to use linear regression because the regression functions have to be linearly independents. Finally, assumption **(A3)** is more difficult to check. First we note:

$$\left(e^{-\lambda\left((Y_t-g(Y_{t-1}))^2-(Y_t-g^0(Y_{t-1}))^2\right)}-1\right)^2=$$
$$e^{-2\lambda\left((Y_t-g(Y_{t-1}))^2-(Y_t-g^0(Y_{t-1}))^2\right)}-2e^{-\lambda\left((Y_t-g(Y_{t-1}))^2-(Y_t-g^0(Y_{t-1}))^2\right)}+1$$

So, d_g^λ is well defined if $E\left[e^{-2\lambda\left((Y_t-g(Y_{t-1}))^2-(Y_t-g^0(Y_{t-1}))^2\right)}\right]<\infty$, but

$$(Y_t-g(Y_{t-1}))^2-(Y_t-g^0(Y_{t-1}))^2=$$
$$(Y_t-g^0(Y_{t-1})+g^0(Y_{t-1})-g(Y_{t-1}))^2-(Y-g^0(Y_{t-1}))^2=$$
$$2\varepsilon(g^0(Y_{t-1})-g(Y_{t-1})+(g^0(Y_{t-1})-g(Y_{t-1}))^2$$

where $\varepsilon=Y_t-g^0(Y_{t-1})$ is the noise of the model. Hence, if regression functions are bounded, d_g^λ is well defined if $\lambda>0$ exists such that $e^{\lambda|\varepsilon|}<\infty$ (i.e. ε admits exponential moments). Finally, using the same techniques of reparameterization as in Liu and Shao (2003)[9] or Olteanu and Rynkiewicz (2012)[7], assumption **(A3)** can be shown to be true for mixture of MLP regression models.

4. A LITTLE EXPERIMENT

The theoretical penalization terms of the previous section can be chosen among a wide range of functions (see condition **(A1)**). In the sequel, a little experiment is conducted to assess right rate of penalization

to guess the "true" number of regimes of a model. First, let us recall the definition of special case of parametric non-linear function called multilayer perceptron (MLP). Let $y = (y(1), \cdots, y(d))^T \in \mathbb{R}^d$ be vector of inputs and $w_i := (w_{i1}, \cdots, w_{id})^T \in \mathbb{R}^d$ be a parameter vector for the hidden unit i. The MLP function with k hidden units can be written :

$$F_\theta(y) = \beta + \sum_{i=1}^{k} a_i \phi\left(w_i^T y + b_i\right),$$

with $\theta = (\beta, a_1, \cdots, a_k, b_1, \cdots, b_k, w_{11}, \cdots, w_{1d}, \cdots, w_{k1}, \cdots, w_{kd})$ the parameter vector of the model and ϕ a bounded transfer function, usually a sigmoïdal function like the hyperbolic tangent function.
Let a simulated model be:

$$Y_{t+1} = F_{\theta^0_{X_t}}(Y_t) + \varepsilon_t, t = 1, \cdots, n$$

with $Y_1 = 0$, $(\varepsilon_1, \cdots, \varepsilon_n)$ i.i.d., $\varepsilon_t \sim \mathcal{N}(0, 1)$, X_t a random variable on $\{1, 2, 3\}$ such that

$$P(X_t = 1) = P(X_t = 2) = P(X_t = 3) = \frac{1}{3}$$

and $\left(F_{\theta^0_i}\right)_{i\in\{1,2,3\}}$ three MLP functions with 4 hidden units (13 weights) and different parameters. Note that the three MLP functions are linearly independent.
We try to guess the true number of regimes (3 regimes) of the time series. Since we are only interested in guessing number of regimes, the architecture of the MLP regression model is assumed to be known (in real application such information will not be available) and the number of regimes is assumed to be between 1 and 10. Three penalized criteria are assessed:

- AIC like: $\frac{E_n(g)}{n} + \frac{2\sigma^2 C \times p}{n}$ (theoretically not consistent).
- BIC like: $\frac{E_n(g)}{n} + \frac{\sigma^2 C \log(n) \times p}{n}$ (theoretically consistent).
- SP (Strong Penalization): $\frac{E_n(g)}{n} + \frac{\sigma^2 C \times p\sqrt{n}}{n}$ (theoretically consistent).

We simulate $n = 100$, $n = 500$ and $n = 1000$ data, for each n the experiment is repeated 100 times. The constant C, proportional to the number of parameters of the regression function, is fixed to the true dimension of the regression models (13 parameters for each model). The parameters of models are always estimated by minimizing the mean square error: $\frac{E_n(g)}{n}$. Note, that these criteria involve the knowledge of the variance of the noise σ^2. In a first experiment we will use the true variance of the noise, then this problem will be addressed in the following section.

4.1. Models selection with σ^2 known

Results with $n = 100$. The following numbers of regimes are chosen by the criteria:

- AIC like:

nb of regimes	$1, 2, 3, 4$	5	6	7	8	9	10
models selected	0	3	7	18	17	23	32

- BIC like:

nb of regimes	$1, 2$	3	$4, 5, 6, 7, 8, 9, 10$
models selected	0	100	0

- SP (Strong Penalization) :

nb of regimes	$1, 2$	3	$4, 5, 6, 7, 8, 9, 10$
models selected	0	100	0

Results with $n = 500$. The following numbers of regimes are chosen by the criteria:

- AIC like:

nb of regimes	$1, 2, 3, 4, 5, 6, 7$	8	9	10
models selected	0	12	19	69

- BIC like:

nb of regimes	$1, 2$	3	$4, 5, 6, 7, 8, 9, 10$
models selected	0	100	0

- SP (Strong Penalization):

nb of regimes	$1, 2$	3	$4, 5, 6, 7, 8, 9, 10$
models selected	0	100	0

Results with $n = 1000$. The following numbers of regimes are chosen by the criteria:

- AIC like:

nb of regimes	$1, 2, 3, 4, 5, 6, 7$	8	9	10
models selected	0	7	24	69

- BIC like:

nb of regimes	$1, 2$	3	$4, 5, 6, 7, 8, 9, 10$
models selected	0	100	0

- SP (Strong Penalization):

nb of regimes	$1, 2$	3	$4, 5, 6, 7, 8, 9, 10$
models selected	0	100	0

The BIC like and the Strong Penalization criteria chose always the true number of regimes whatever the number of data. According to the theory, AIC like criterion is not consistent (see condition **(A1)**) and the number of regimes estimated is always too large. These good results assume that the true variance of the noise is known, but for regression models this is never the case. A straightforward idea is to replace the unknown variance by the estimated one, this is done in the next section.

4.2. Models selection using estimated variance $\hat{\sigma}^2$.

The estimated variance $\hat{\sigma}^2$ is the mean square error of the model:

$$\hat{\sigma}^2 := \frac{E_n(g)}{n}$$

computed for the least square estimator of parameters. Hence, the comparison is done with the penalized criteria :

- AIC like: $\frac{E_n(g)}{n} + \frac{2\hat{\sigma}^2 C \times p}{n}$ (theoretically not consistent).
- BIC like: $\frac{E_n(g)}{n} + \frac{\hat{\sigma}^2 C \log(n) \times p}{n}$ (theoretically consistent).
- SP (Strong Penalization): $\frac{E_n(g)}{n} + \frac{\hat{\sigma}^2 C \times p\sqrt{n}}{n}$ (theoretically consistent).

Results with $n = 100$. The following models are chosen by the criteria:

- AIC like:

nb of regimes	$1,2,3,4,5,6,7$	8	9	10
models selected	0	6	16	78

- BIC like:

nb of regimes	$1,2$	3	4	5	6	7	8	9	10
models selected	0	26	6	3	2	5	9	18	31

- SP (Strong Penalization):

nb of regimes	$1,2$	3	4	5	6	7	8	9	10
models selected	0	94	2	0	0	0	1	3	0

Results with $n = 500$. The following models are chosen by the criteria:

- AIC like:

nb of regimes	$1,2,3,4,5,6,7$	8	9	10
models selected	0	2	24	74

- BIC like:

nb of regimes	$1,2$	3	$4,5,6,7,8,9,10$
models selected	0	100	0

- SP (Strong Penalization):

nb of regimes	$1,2$	3	$4,5,6,7,8,9,10$
models selected	0	100	0

Results with $n = 1000$. The following models are chosen by the criteria:

- AIC like:

nb of regimes	$1, 2, 3, 4, 5, 6$	7	8	9	10
models selected	0	1	6	27	66

- BIC like:

nb of regimes	$1, 2$	3	$4, 5, 6, 7, 8, 9, 10$
models selected	0	100	0

- SP (Strong Penalization):

nb of regimes	$1, 2$	3	$4, 5, 6, 7, 8, 9, 10$
models selected	0	100	0

As usual the AIC like criterion misbehaves like in the previous section. But, for a small number of data ($n = 100$) the use of an estimation of the variance of the noise instead of the true one leads to overestimation of the number of regimes for the BIC like criterion. The explanation is that the variance of the noise is underestimated for large number of regimes and so the penalized criterion. This drawback disappears for larger number of data ($n = 500$ and $n = 1000$) because estimation of the variance becomes better. The Strong Penalized criterion seems better to guess the true number of regimes whatever the number of data.

5. CONCLUSION

We have proven the consistency of penalized criteria for estimating the number of regimes in a mixture of non-linear regression. This result can be shown without knowing the form of density function of the noise. Note that the weak identifiability assumption excludes linear regression functions. This result comes from an inequality showing that overfitting of the model is moderate if the noise admits exponential moments and the parameters of the model are assumed to be bounded. This bound justifies use of penalized criteria in order to fit model dimension. Hence, The user can select the true number of regimes thanks to penalized criteria, of the form

$$E_n(g) + a_n(p)$$

Hence, if the penalization term $a_n(p)$ is well calibrated the true number of regimes will be automatically selected if n is large enough. A little experiment suggests that a good choice of penalization seems to be the middle of the possible range: $a_n(p) = C \cdot p\sqrt{n}$. A further question could be to know if this empirical finding for the tuning of penalization term can be justified theoretically. Note that, this paper was only concerned with the identification of the true number of regimes. The point is more to get an idea of the complexity of the model than to have a predictive model. However, if there are enough data, the true model will also be the best predictive model.

RECEIVED OCTOBER, 2012
REVISED DECEMBER, 2012

REFERENCES

[1] BRADLEY, R. (2005): Basic properties of strong mixing conditions. a survey and some open questions **Probability Surveys**, 2:107–144.

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

[3] DOUKHAN, P., MASSART, P., and RIO, E. (1995): Invariance principles for absolutely regular empirical processes **Ann. Inst. Henri Poincaré (B) Probabilités et Statistiques**, 31:393–427.

[4] GASSIAT, E. (2002): Likelihood ratio inequalities with applications to various mixtures **Ann. Inst. Henri Poincaré**, 38:897–906.

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

[6] J.F., Y. and Attali, J. (2000): On stability of nonlinear ar processes with markov switching **Advances in Applied Probability**, 32(2):394–407.

[7] OLTEANU, M. and RYNKIEWICZ, J. (2012): Asymptotic properties of autoregressive regime-switching models **ESAIM P.S.**, 16:25–47.

[8] VAN DER VAART, A. (2000): **Asymptotic Statistics** Cambridge University Press.

[9] X., L. and SHAO, Y. (2003): Asymptotics for likelihood ratio tests under loss of identifiability **The Annals of Statistics**, 31(3):807–832.