HAL
open science

A descriptive method to evaluate the number of regimes in a switching autoregressive model

Madalina Olteanu

To cite this version:

Madalina Olteanu. A descriptive method to evaluate the number of regimes in a switching autoregressive model. Neural Networks, 2006, 19, pp.963-972. 10.1016/j.neunet.2006.05.019 . hal-00133172

HAL Id: hal-00133172

https://hal.science/hal-00133172v1

Submitted on 1 Mar 2007

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Descriptive Method to Evaluate the Number of Regimes in a Switching Autoregressive Model

Madalina Olteanu
SAMOS-MATISSE, University of Paris I
90 Rue de Tolbiac, 75013 Paris, France
madalina.olteanu@univ-paris1.fr

Abstract - *This paper proposes a descriptive method for an open problem in time series analysis : determining the number of regimes in a switching autoregressive model. We will translate this problem into a classification one and define a criterion for clustering hierarchically different model fittings. Finally, the method will be tested on simulated examples and real-life data.*

Key words - switching autoregressive models, hierarchical clustering, Ward distance, SOM

1 Introduction

In the past few years, several nonlinear autoregressive models were proposed for time series analysis. Some of these models are based on the idea that the process is characterized not by a unique linear autoregression, but by the fact that two or more regimes are driving the series behaviour. In each regime, an autoregressive function is fitted. We are interested in the case where the autoregressive functions are linear in every regime. The most classical examples are TAR (Treshold Autoregressive) models introduced by Tong (1978) with regime switching according to the magnitude of a treshold variable, the smoothed version of TAR models (STAR), or the more recent Markov switching autoregressive models, first used by Hamilton (1989) to model the U.S. Gross National Product.
Estimating the parameters of these models is usually done by maximizing the likelihood function, but under a very strong hypothesis, a fixed number of regimes. Choosing the "true" number of regimes is still an open problem, as this is equivalent to testing with lack of identifiability under the null hypothesis. This leads to a degenerated Fisher information matrix and thus the chi-square theory and the likelihood ratio tests fail to apply. An empirical method to detect this kind of non-linearity using Kohonen maps and hierarchical clustering of linear regressions is given below. The second section describes the method, while the third provides examples on simulated and real-life data. Finally, a conclusion will follow.

2 The Method

The problem of finding the "true" number of regimes can be rewritten as a classification problem by using a sliding window as follows . Suppose that we have observed the values of

a time series $\{y_t\}_{t=\overline{1,T}}$ and we decide to fit an autoregressive model. Once the order of the model has been determined (with an AIC criterion, for example), we can consider the data set of dimension $(T-p)\mathrm{x}(p+1)$, $\{y_t, y_{t-1}, ...y_{t-p}\}_{t=\overline{p+1,T}}$. Looking for the number of regimes is actually equivalent to looking for the number of regression lines (or hyperplanes) which will best fit the data.

The idea is simple and is based on the possibility of finding patterns in data which will identify the regression hyperplanes. Given the data set in Figure 1, fitting one regression line to the data is clearly not the good choice. If we now suppose that we managed to cluster the data into two groups and we perform a regression within each of these groups, we get two lines which seem to describe better the sample. This is confirmed by the "within squared error", which is equal to the sum of squared residuals if there is only one regime and, if there are several, to the total sum of squared residuals within each group.

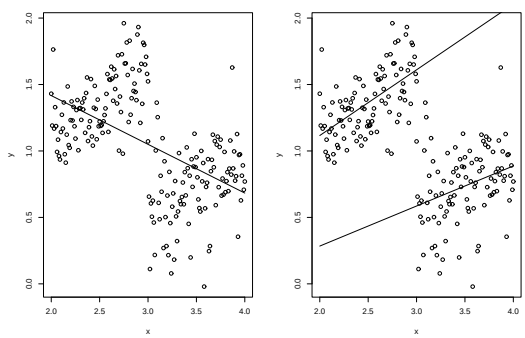

Figure 1: Fitting clustered data

In the general case, we would like to start with some "good" initial clusters which will be then classified hierarchically using some squared error criterion and we would expect to have an important break in the increasing values of this criterion, once we pass from the true number of regimes to a smaller one.A "good" cluster should contain observations belonging to the same regime and, at the same time, have enough points to estimate a regression line.

2.1 Initial clustering

For the initial clustering, self-organizing maps were employed. This choice was motivated by several properties of the Kohonen maps such as the vector quantization property (the last steps of the algorithm use "0-neighbours"and thus are equivalent to classical VQ techniques), the homogeneity of the clusters as well as the topology conservation used to fasten the hierarchical classification algorithm.

One problem arises once we get the clusters and try to fit a regression within each one : are there enough points in every cluster? No cluster will be allowed to have less points than the number of lags or regressor variables. For this, either we eliminate from the analysis those which do not verify this condition, either we force the points to move to a different cluster, by assigning them to the closest sufficiently large cluster.Since very few observations are concerned with this problem and in order to shorten the computing time, the first approach was preferred.

The assumption we make here is that the property of homogeneity of self-organizing maps manages to create clusters in which observations belong to one regime. This could be justified by the fact that the variables used for the classification contain information concerning the regime of the observation and similar profiles will belong to the same cluster. Indeed, in the simulated examples where the different regimes are known, we'll see that the map clusters are generally homogeneous from this point of view. This property will no longer apply if the regression hyperplanes are too close and the noise is important.

2.2 Hierarchical classification

As we actually need to compare different data fits, which is also equivalent to different numbers of hyperplanes, we need to adapt a hierarchical classification to our case (let us first make the convention to call "clusters" the result of the Kohonen map and "classes" two or more "clusters" joined together by the hierarchical method). We will choose a new "distance" between classes by developping a squared error criterion.
A very popular method used in classification is to minimize a within-class variation criterion, the variation within a class being defined as the sum of squared distances from the individuals to the barrycenter. By considering a data set $x_i \in \mathbb{R}^n$, $i = 1, ..., N$ and a fixed number of clusters k, the total within-class variation can be written as $I_w = \sum_{i=1}^{k} w_i I_i$, where w_i is the weight of class i, $I_i = \sum_{j=1}^{N_i} d^2 (x_{i,j}, g_i)$ being the variation within the class i and g_i is the corresponding barrycenter. We denoted by N_i the number of individuals in class i and by $x_{i,j}$ the data point indexed by j in class i.
In hierarchical classification, this principle was adapted by Ward and the algorithm consists in clustering together the individuals who minimize the increase of the within-class variation. Our idea was to build an algorithm similar to Ward's, but, as our interest is to estimate the number of hyperplanes characterizing the data, we will replace the barrycenters by regression lines and the within-class variation becomes the within sum of squared errors. Obviously, in this case, the inter-class variation cannot be defined.
For a fixed number of classes, k, the within sum of squared errors is defined as
$SSE_{w,k} = \sum_{l=1}^{k} SSE_{C_l}$,
where $SSE_{C_l} = \sum_{t \in C_l} (y_t - \hat{y}_t)^2$ is the sum of squared residuals and $\hat{y}_t$ is the predicted value of y_t by the linear regression of order p fitted in class C_l, $l = \overline{1, k}$.
Now, in the frame of hierarchical clustering, when passing from k to $k-1$ classes, if classes i and j were clustered, the within sum of squared errors becomes :
$SSE^{i,j}_{w,k-1} = \sum_{l=1, l \neq i, l \neq j}^{k} SSE_{C_l} + SSE_{C_i \cup C_j}$
Following the same principle as Ward's, we want to minimize the increase in the within inertia, which in this case is defined by the within sum of squared errors. This is equivalent to finding i and j which minimize
$\Delta S^{i,j}_{w,k,k-1} = SSE^{i,j}_{w,k-1} - SSE_{w,k} = SSE_{C_i \cup C_j} - SSE_{C_i} - SSE_{C_j}$

2.2.1 The within sum of squared errors criterion

Now, let us take a closer look at this difference and give an explicit expression as a function of the data and the residuals. The following notations are necessary :

- $Y_i = \{Y_t\}_{t \in C_i} \in \mathbb{R}^{n_i}$, $Y_j = \{Y_t\}_{t \in C_j} \in \mathbb{R}^{n_j}$, where n_i and n_i are the cardinalities of C_i and, respectively, C_j
- $X_i = \{1, Y_{t-1}, ..., Y_{t-p}\}_{t \in C_i} \in \mathbb{R}^{n_i \times (p+1)}$, $X_j = \{1, Y_{t-1}, ..., Y_{t-p}\}_{t \in C_j} \in \mathbb{R}^{n_j \times (p+1)}$.

The linear regressions fitted in classes C_i and C_j can be written as :

$Y_i = X_i \cdot \beta_i + u_i = X_i \cdot \hat{\beta}_i + e_i$

$Y_j = X_j \cdot \beta_j + u_j = X_j \cdot \hat{\beta}_j + e_j$,

where $\beta_i, \beta_j, \hat{\beta}_i, \hat{\beta}_j \in \mathbb{R}^{p+1}$, $\hat{\beta}_i$ and $\hat{\beta}_j$ are the least squares estimates of β_i and β_j, $u_i, e_i \in \mathbb{R}^{n_i}$ and $u_j, e_j \in \mathbb{R}^{n_j}$ are the error and, respectively, the residuals vectors, $u_i \sim N(0, \sigma_i^2 I_{n_i})$ and $u_j \sim N(0, \sigma_j^2 I_{n_j})$.

The linear regression fitted in the joint class $C_i \cup C_j$ is written as :

$Y = X \cdot \beta + u = X \cdot \hat{\beta} + e$, where $Y = \begin{bmatrix} Y_i \\ Y_j \end{bmatrix} \in \mathbb{R}^{n_i + n_j}$, $X = \begin{bmatrix} X_i \\ X_j \end{bmatrix} \in \mathbb{R}^{(n_i+n_j) \times (p+1)}$, $\beta, \hat{\beta} \in \mathbb{R}^{p+1}$, $\hat{\beta}$ is the least squares estimate of β, $u, e \in \mathbb{R}^{n_i+n_j}$ are the error and the residuals vector and $u \sim N(0, \Omega)$, $\Omega = \begin{pmatrix} \sigma_i^2 I_{n_i} & 0 \\ 0 & \sigma_j^2 I_{n_j} \end{pmatrix}$.

Then, we can compute $\Delta S^{i,j}_{w,k,k-1}$ as :

- $\Delta S^{i,j}_{w,k,k-1} = SSE_{C_i \cup C_j} - SSE_{C_i} - SSE_{C_j} = e\prime e - (e_i\prime e_i + e_j\prime e_j)$

Remark 1 :

Using this form, Toyoda (1974) proves that $\Delta S^{i,j}_{w,k,k-1}$ is approximately distributed as $\sigma^2 \chi^2 (p+1)$, where σ^2 is any well-chosen weighted average of σ_i^2 and σ_j^2 and p is the number of lags considered.

- $\Delta S^{i,j}_{w,k,k-1} = SSE_{C_i \cup C_j} - SSE_{C_i} - SSE_{C_j} = \left\| Y - X \cdot \hat{\beta} \right\|^2 - \left\| Y_i - X_i \cdot \hat{\beta}_i \right\|^2 - \left\| Y_j - X_j \cdot \hat{\beta}_j \right\|^2$ (1)

But $Y - X \cdot \hat{\beta} = \begin{bmatrix} Y_i - X_i \cdot \hat{\beta} \\ Y_j - X_j \cdot \hat{\beta} \end{bmatrix} = \begin{bmatrix} Y_i - X_i \cdot \hat{\beta}_i \\ Y_j - X_j \cdot \hat{\beta}_j \end{bmatrix} + \begin{bmatrix} X_i \cdot \hat{\beta_i} - X_i \cdot \hat{\beta} \\ X_j \cdot \hat{\beta}_j - X_j \cdot \hat{\beta} \end{bmatrix}$ (2)

and thus

$$\left\| Y - X \cdot \hat{\beta} \right\|^2 = \left\| \begin{matrix} Y_i - X_i \cdot \hat{\beta} \\ Y_j - X_j \cdot \hat{\beta} \end{matrix} \right\|^2 = \left\| \begin{matrix} Y_i - X_i \cdot \hat{\beta}_i \\ Y_j - X_j \cdot \hat{\beta}_j \end{matrix} \right\|^2 + \left\| \begin{matrix} X_i \cdot \hat{\beta_i} - X_i \cdot \hat{\beta} \\ X_j \cdot \hat{\beta}_j - X_j \cdot \hat{\beta} \end{matrix} \right\|^2$$

since the cross-product on the right term in (2) can easily be seen to be zero. We get that :

$$\left\| Y - X \cdot \hat{\beta} \right\|^2 = \left\| Y_i - X_i \cdot \hat{\beta}_i \right\|^2 + \left\| Y_j - X_j \cdot \hat{\beta}_j \right\|^2 + \left\| X_i \cdot \hat{\beta_i} - X_i \cdot \hat{\beta} \right\|^2 + \left\| X_j \cdot \hat{\beta}_j - X_j \cdot \hat{\beta} \right\|^2 (3)$$

and by replacing (3) in (1) :

$$\Delta S^{i,j}_{w,k,k-1} = \left\| X_i \cdot \hat{\beta_i} - X_i \cdot \hat{\beta} \right\|^2 + \left\| X_j \cdot \hat{\beta}_j - X_j \cdot \hat{\beta} \right\|^2 =$$

$$= (\hat{\beta}_i - \hat{\beta})\prime X_i\prime X_i (\hat{\beta}_i - \hat{\beta}) + (\hat{\beta}_j - \hat{\beta})\prime X_j\prime X_j (\hat{\beta}_j - \hat{\beta}) \ (4)$$

Using the form of the least squares estimators $\hat{\beta}_i$, $\hat{\beta}_j$ and $\hat{\beta}$

$$\hat{\beta}_i - \hat{\beta} = \beta_i - \beta + \left\{ \left[(X_i\prime X_i)^{-1} X_i\prime, 0 \right] - (X_i\prime X_i + X_j\prime X_j)^{-1} \left[X_i\prime, X_j\prime \right] \right\} \cdot \begin{bmatrix} e_i \\ e_j \end{bmatrix}$$

$$\hat{\beta}_j - \hat{\beta} = \beta_j - \beta + \left\{ \left[0, (X_j\prime X_j)^{-1} X_j\prime \right] - (X_i\prime X_i + X_j\prime X_j)^{-1} [X_i\prime, X_j\prime] \right\} \cdot \begin{bmatrix} e_i \\ e_j \end{bmatrix}$$

and (4) becomes

$$\Delta S^{i,j}_{w,k,k-1} = (\beta_i - \beta)\prime X_i\prime X_i(\beta_i - \beta) + (\beta_j - \beta)\prime X_j\prime X_j(\beta_j - \beta) +$$
$$+(\beta_i - \beta)\prime \left\{ X_i\prime e_i - X_i\prime X_i \left(X_i\prime X_i + X_j\prime X_j\right)^{-1} \left(X_i\prime e_i + X_j\prime e_j\right) \right\} +$$
$$+ \left\{ e_i\prime X_i - \left(e_i\prime X_i + e_j\prime X_j\right) \left(X_i\prime X_i + X_j\prime X_j\right)^{-1} X_i\prime X_i \right\} (\beta_i - \beta) +$$
$$+(\beta_j - \beta)\prime \left\{ X_j\prime e_j - X_j\prime X_j \left(X_i\prime X_i + X_j\prime X_j\right)^{-1} \left(X_i\prime e_i + X_j\prime e_j\right) \right\} +$$
$$+ \left\{ e_j\prime X_j - \left(e_i\prime X_i + e_j\prime X_j\right) \left(X_i\prime X_i + X_j\prime X_j\right)^{-1} X_j\prime X_j \right\} (\beta_j - \beta) +$$
$$+e_i\prime X_i \left(X_i\prime X_i\right)^{-1} X_i\prime e_i + e_j\prime X_j \left(X_j\prime X_j\right)^{-1} X_j\prime e_j -$$
$$- \left(e_i\prime X_i + e_j\prime X_j\right) \left(X_i\prime X_i + X_j\prime X_j\right)^{-1} \left(X_i\prime e_i + X_j\prime e_j\right)$$

If classes i and j come form the same regime, that is $\beta_i = \beta_j = \beta$, the increase in the within sum of sqared errors is only

$$\Delta S^{i,j}_{w,k,k-1} = e_i\prime X_i \left(X_i\prime X_i\right)^{-1} X_i\prime e_i + e_j\prime X_j \left(X_j\prime X_j\right)^{-1} X_j\prime e_j -$$
$$- \left(e_i\prime X_i + e_j\prime X_j\right) \left(X_i\prime X_i + X_j\prime X_j\right)^{-1} \left(X_i\prime e_i + X_j\prime e_j\right)$$

This quantity is very close to zero if the classes contain enough points. Thus, together with remark 1,we get that if the joint classes are from the same regime, the within sum of squared errors should be close to zero and if the classes are from different regimes, the larger the difference between the parameters of the two regimes, the increase in the within sum of squared errors should be larger.

2.3 The algorithm

Now, we can write the steps of the algorithm which, at the same time, clusters the data and models the dependencies within each class:
We consider k going from M to 2, where M is the number of clusters resulting from the Kohonen map.
For each k, we have the following steps :

- Compute the k regressions lines corresponding to the k classes
- Find (i_0, j_0) which minimize $\Delta_{i,j} SSE_w$
- Join these classes together, put $k = k - 1$ and go to the first step.

At the last step all points are clustered together and there is a single regression line.
The next thing to do is draw the dendrogram and look how the within sum of squared errors increases over the classification. As it was mentioned earlier, one might expect an important break when the number of classes is smaller than the real number of regimes.

	Regime1			Regime2		
	$Intercept$	y_{t-1}	y_{t-2}	$Intercept$	y_{t-1}	y_{t-2}
value	-2.82	0.59	-0.89	1.4	0.37	0.32
t-value	-19.15	13.09	15.89	8.24	5.96	4.79

Table 1: Coefficients for the TAR model

3 Examples and Results

The method was tested on several nonlinear autoregressive regime switching models. Historically, the first introduced were the treshold models (we won't speak here about the other variants of these models, smoothed etc), followed by the switching Markov and next we will consider both examples.

3.1 TAR Models

The example is a TAR of order two and the coefficients were taken from the paper of Gonzalo&Pitarkis.

$y_t = \begin{cases} -3 + 0.5y_{t-1} - 0.9y_{t-2} + \varepsilon_t & , y_{t-2} \leq 1.5 \\ 2 + 0.3y_{t-1} + 0.2y_{t-2} + \varepsilon_t & , y_{t-2} > 1.5 \end{cases}$, where $\{y_t\}$ is the observed series and ε_t is i.i.d. standard gaussian.

Three samples containing 200, 400 and 800 points, respectively, were simulated. Let us examine the 200 points sample. The self-organizing map dimension was fixed equal to 5x5 and $\{y_t, y_{t-1}, y_{t-2}\}$ were the variables used for the classification. Crossing the map with a boolean variable which distinguishes whether y_{t-2} is above or below the treshold value allows to see that the clusters are homogeneous and each of them takes values in one regime.

If the hierarchical clustering algorithm is run on the twenty-five clusters, the squared error criterion increases as in Figure 3 and suggests that a good choice would be a two-regimes model. The hierarchical classification provides also good estimators for the parameters of the model when choosing two regimes as shown in Table 1.

The threshold value was not estimated and we'll see that the decision on the type of the model (TAR, Markov switching) cannot be made on this basis only.

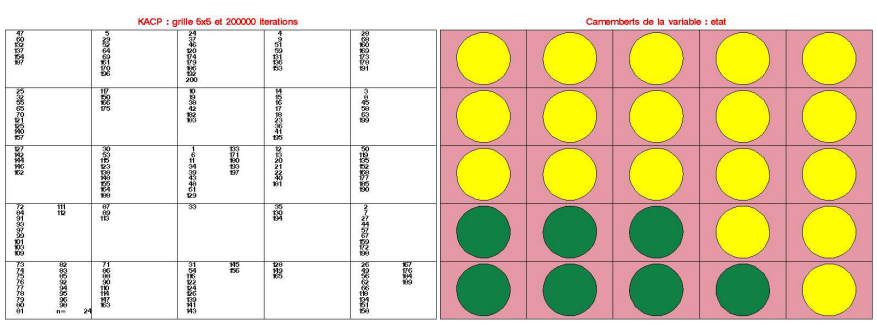

Figure 2: Initial clustering with a Kohonen map for TAR model

When increasing the number of points in the sample, we have chosen to increase also the map size (we considered a 7x7 map for the 400 sample and a 9x9 for the 800). This was done in

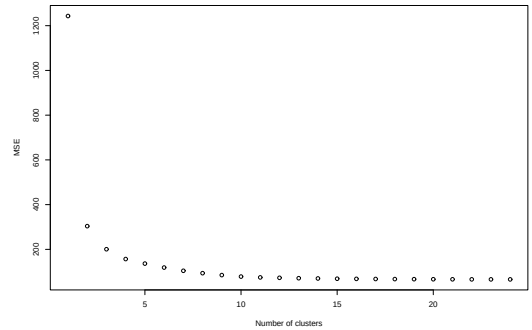

Figure 3: Squared error criterion for TAR model

the purpose of conserving clusters homogeneity, but a good criterion for choosing the size of the map is still to be defined. The results are very similar and in both cases the hierarchical algorithm suggests a two-regimes model.

3.2 A Two Regime Markov Switching Model

For this example, we will first define an autoregressive Markov switching process. If $\{y_t\}_{t\in\mathbb{N}}$ is the observed time series, let us suppose that it follows a linear autoregressive process of order p and that we have two regimes, the passage from one regime to the other being driven by an unobserved Markov chain, $\{x_t\}_{t\in\mathbb{N}}$, which has a transition probability matrix A.
For two regimes and two lags of time, we can write the model as follows : $y_t = f_{x_t}(y_{t-1}, y_{t-2}) + \sigma_{x_t}\varepsilon_t$

$f_{x_t}(y_{t-1}, y_{t-2}) \in \{f_1, f_2\}$, $\sigma_{x_t} \in \{\sigma_1, \sigma_2\}$, ε_t i.i.d. noise (usually a standard gaussian) and $A = \begin{pmatrix} p & 1-p \\ 1-q & q \end{pmatrix}$ is the transition probability matrix of x_t.
The data used here were simulated with the parameters below (a globally stationary process was chosen):

$$\begin{cases} f_1(y_{t-1}, y_{t-2}) = 0.2 + 0.5y_{t-1} + 0.1y_{t-2} \\ f_2(y_{t-1}, y_{t-2}) = 0.3 + 0.9y_{t-1} - 0.1y_{t-2} \end{cases}, \begin{cases} \sigma_1 = 0.03 \\ \sigma_2 = 0.02 \end{cases} \text{ and } A = \begin{pmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{pmatrix}$$

As for the previous example, three samples (200, 400, 800 points) were considered. We will only list the results for the 400 sample and remark that the outputs for the other two cases were very similar. The initial clustering was performed using a 6x6 Kohonen map and $\{y_t, y_{t-1}, y_{t-2}\}$as variables. Figure 4 shows that the map is well organized, the clusters are homogeneous and when crossing with the variable giving the regime, there is a good separation of them in the initial clusters.
Afterwards, from the hierarchical classification of these clusters, we get, again, a huge jump when passing from two classes to one, as shown in Figure 5.The estimated coefficients in each of the two classes are shown in Table 2 (we will also note that the two clusters are homogeneous, the pourcentage of explained variance being larger than 92% in each of them).

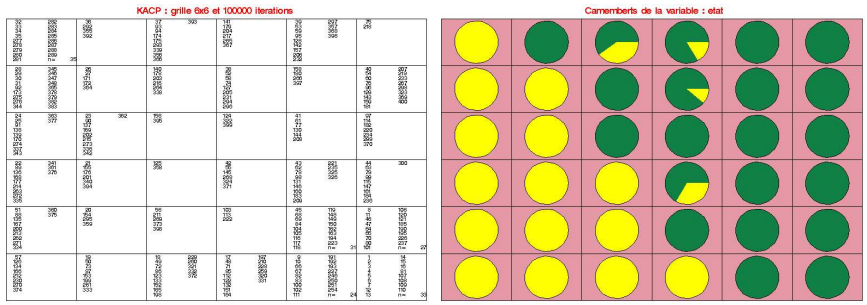

Figure 4: Initial clustering with a Kohonen map for two regimes Markov

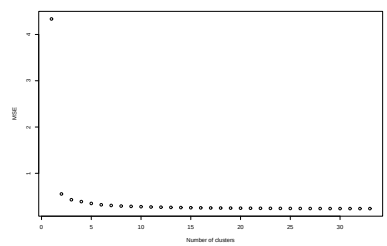

Figure 5: Squared error criterion for 2 regime Markov

	Regime1			Regime2		
	$Intercept$	y_{t-1}	y_{t-2}	$Intercept$	y_{t-1}	y_{t-2}
value	0.3	0.88	-0.09	0.18	0.39	0.24
t-value	41.56	32.89	-3.64	20.91	12.07	8.87

Table 2: Coefficients for the two regime Markov model

3.3 A Three Regime Markov Switching Model

Now let us see what happens if we add a new regime to the model, which will moreover be explosive and drive the process into a nonstationary one. The following example was considered :

$y_t = f_{x_t}(y_{t-1}, y_{t-2}) + \sigma_{x_t}\varepsilon_t$, $f_{x_t}(y_{t-1}, y_{t-2}) \in \{f_1, f_2, f_3\}$, $\sigma_{x_t} \in \{\sigma_1, \sigma_2, \sigma_3\}$, ε_t is i.i.d. standard gaussian and

$$\begin{cases} f_1(y_{t-1}, y_{t-2}) = 0.2 + 0.5y_{t-1} + 0.1y_{t-2} \\ f_2(y_{t-1}, y_{t-2}) = 0.3 + 0.9y_{t-1} - 0.1y_{t-2} \\ f_3(y_{t-1}, y_{t-2}) = 0.5 + 1.2y_{t-1} + 0.5y_{t-2} \end{cases}, \begin{cases} \sigma_1 = 0.03 \\ \sigma_2 = 0.02 \\ \sigma_1 = 0.03 \end{cases} \text{and } A = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.6 & 0.2 & 0.2 \end{pmatrix}$$

The following results are from a 400 points sample, with an initial 8x8 map. By crossing the map with the regime variable, there is a relatively good separation, although we may notice that the first two regimes seem to come closer together with respect to the third one.

Here, a first conclusion would be that there are four regimes. But let us take a closer look at the hierarchical classification. One of the four final classes contains only one cluster, one cell of the map. Moreover, this cell (the 8th) is isolated from the rest of the map and it contains only four observations with very high values. If we project the data on a two or

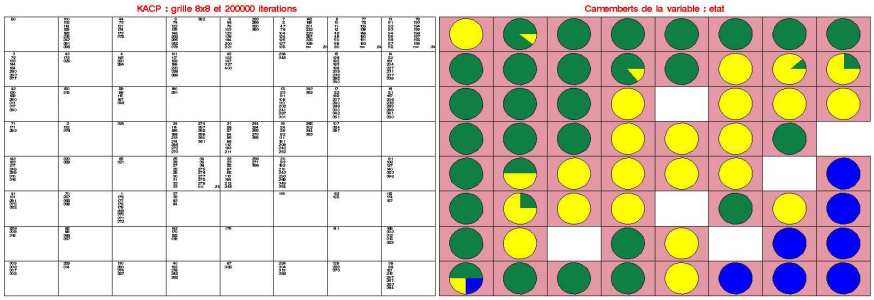

Figure 6: Initial clustering with a Kohonen map

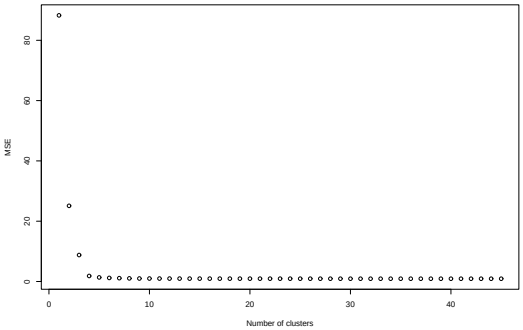

Figure 7: Squared error criterion for 3 regimes Markov

three-dimensional space, the same four observations are far from the rest. The algorithm has identified a small class of outliers which are considered as a separate regime and from this point of view the method is close to Ward's which is also sensitive to this kind of observations.

We can't continue with the examples before making an important remark. We've seen that this method of identifying the number of regression lines works quite good in the examples above. Concerning the parameter estimation and making a decision about the model (TAR or Markov switching?, for instance), the hierarchical classification provides only the estimators for the regression lines, a likelihood approach should be used instead, once we fixed the number of regimes, to estimate the rest of the parameters : threshold value, transition matrix etc. As for the second question, no theoretical result is available yet, the econometricians preferring to decide on other criteria (economic, social etc).

3.4 What about Real Life Data?

The results on the simulated examples being encouraging, we decided to run the algorithm on real data sets. Three examples were chosen, the first two are the benchmarks Old Faithful Geyser Data and Santa Fe Competition Laser Data and the third is the U.S. GNP (Gross National Product) series, used by Hamilton to introduce the switching Markov models.

3.4.1 Old Faithful Data

The first set of data is the classical Old Faithful Geyser in Yellowstone National Park, consisting of 299 pairs of measurements referring to the waiting time between two successive eruptions, w_t, and the duration of the subsequent eruption, d_t. The data was collected between August 1st and August 15th, 1985 and the two variables are recorded in minutes. Several studies of this sample are available, most authors trying to assess either the clustering of the data, either the dependency between successive events. A literature overview, as well as an analysis using time series while assuming "a priori" the existence of two patterns of dependency, can be found in Azzalini and Bowman (1990) paper. Although there are no autoregressors in this case, the problem is the same : find the number of clusters and fit a regression within each one.

Our approach would be to detect the clustering of the data and, at the same time, model the dependency within each class. The idea of addressing clustering and dependency at the same time was firstly introduced by Hennig (2000) who uses regression fixed point clusters.

Here, the duration of the eruption was modeled as a linear function of the waiting time before the eruption. While plotting the duration against the waiting time, one can see that there are at least two clusters of points, depending on the waiting time. Let us make one last remark on the data, which is that due to inexact observations during the night, there are 53 points with duration=4 (long eruption) and 20 with duration=2 (short eruption). The medium eruptions (duration=3) appear only once.

For the Kohonen classification the data set $\{w_t, d_t\}_{t=\overline{1,299}}$ was considered (no lags of time were introduced). The map was chosen to be a 6x6 grid and 1000 iterations were performed. Afterwards, an hierarchical clustering minimizing the within sum of squared errors criterion was applied to the 33 "valid" clusters (one cluster was void and two others contained only one point).

In Figure 8, the within sum of squared errors as well as the difference $\Delta SSE_{k,k-1}$ are plotted. The heterogeneity of the data is obvious and one can see that at least two clusters should be considered. On the contrary, passing from three clusters to two is less obvious and we shall see next that there's overlapping of the clusters and that the regression coefficients are very close.

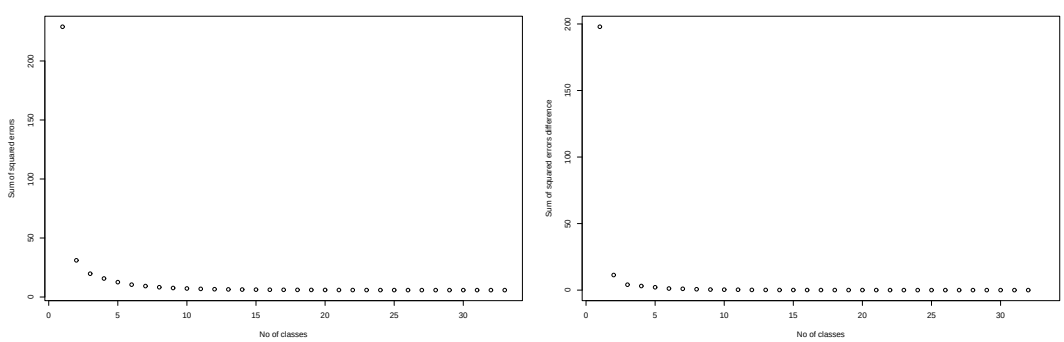

Figure 8: Squared error criterion for the old faithful data

Once we get the hierarchical classification, we get back to the self-organizing map to see how the classes spread over the grid. On the first graph in Figure 9, the tree-cut at two classes

in presented : the first class in the right upper corner in light-grey circles and the second beyond the diagonal of the grid in dark-grey circles. Clusters 7 and 28 do not contain enough points and were not considered in the classification.

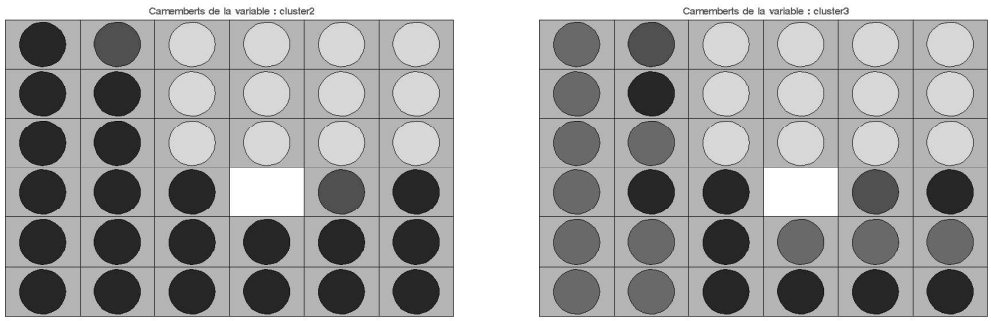

Figure 9: The self-organizing map for the old faithful data

While the 2-classes model is meaningful, in the 3-classes case things seem to be more complicated : the first class is the same, well isolated in the right upper part of the grid, while classes 2 and 3 are mixed on the grid. This is even more obvious if we plot the duration d_t against the waiting time w_t and identify the classes as shown in Figure 10.

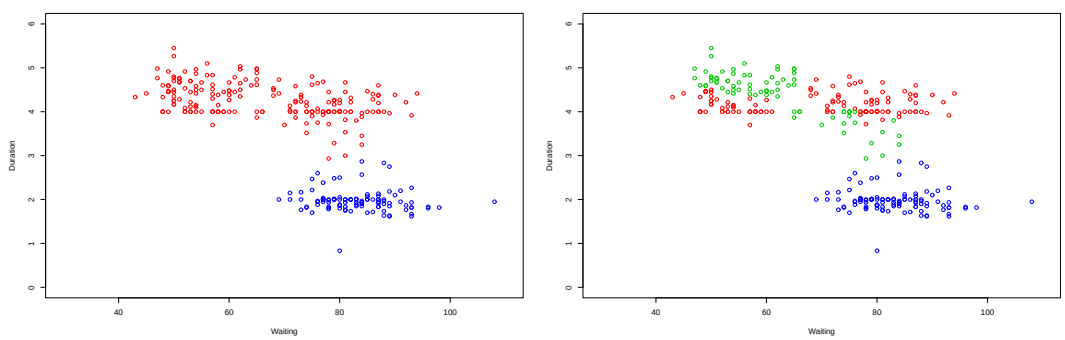

Figure 10: 2-clusters and 3-clusters of the old faithful data

The same situation as for the Kohonen map occurs. In both cases, 2-classes and 3-classes model, a first class (the right upper part of the self-organizing map) is well isolated from the rest and is concentrated around the waiting=2 points. This class corresponds to short time eruptions preceeded by rather long waiting times. The second class on the first graph containts all the long duration points including all duration=4 points. This class is splitted into two when considering the 3-classes case : one sub-class is formed by the duration=4 points and the long time erruptions preceeded by long waiting times, while the second contains the data with a moderately decreasing tendency in the duration for increasing waiting times.
Although there is an interpretation for the 3-classes model, the within squared error criterion used for the classification suggests a 2-classes model, because it corresponds to an important break in the increase of the squared error. Besides, in the discussion of Azzalini and Bowman

(1990) geological evidence for the existence of two distinct patterns of eruptions is given and thus our conclusion is enhanced.

3.4.2 Santa Fe Competition Laser Series

For the laser series, a highly nonlinear data set, the algorithm selects three classes as shown in Figure 11. In order to compare the results with the existing literature, ten lags of time were used. The 10000 observations were initially clustered on a 9x9 map. Once the number of regimes was fixed, the three regimes were supposed to be the states of a discrete Markov chain and the parameters were estimated using the EM algorithm as described in Rynkiewicz (1999). The prediction results are weaker than those obtained by Weigend (1995) or Rynkiewicz (1999), but the number of estimated parameters is much smaller. But, as the nonlinearities of the series are not entirely explained by a mixture of linear regressions, an adaptation of the method by replacing the linear functions with nonlinear ones could be more interesting for this kind of data.

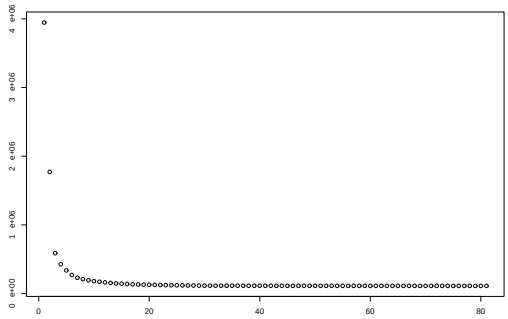

Figure 11: Squared error criterion for the laser series

3.4.3 GNP Series

Concerning the GNP series, Hamilton's approach was based on the assumption that the mean growth rate is subject to occasional, discrete shifts. We dispose of 136 trimestrial observed values of the series, from 1952 until 1984. The maximal lag to be considered was determined with the AIC criterion and was fixed at $p = 3$.

The data was initially clustered with a 4x4 map and considering $\{y_t, y_{t-1}, y_{t-2}, y_{t-3}\}$ as variables. The sixteen clusters were grouped hierarchically by the squared error criterion and the results are displayed in Figure 12.

The first graph contains the within sum of squared errors plotted against the number of classes and the sencond its percentage increase. A first break appears when considering six classes instead of seven, but this can be interpreted as being due to possible stronlgy homogeneous clusters from the same regime which get mixed. There is a second break(13%) when passing from two classes to one but this is less obvious than in previous cases and the decision to model this series by a two-regime model is questionable from our point of view.

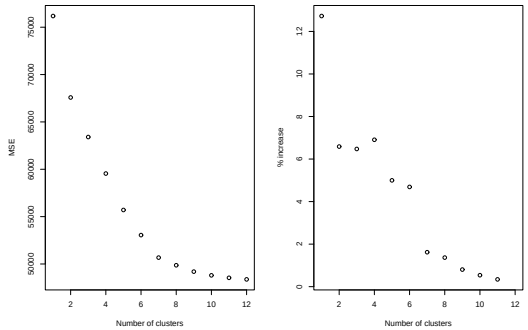

Figure 12: Squared error criterion for GNP data

4 Conclusion and Future Work

We've introduced a descriptive method to assess the presence of regime changes in nonlinear time series analysis. As there is no theoretical answer and no statistical test to solve this problem for the moment, this method may be used, but with precaution. Indeed, self organizing maps could mix the regimes if the regression hyperplanes are too close and the square error criterion seems to be sensitive to outliers.
Thus, several improvements should be made in the future, like considering a different distance that would take into account the temporal dependency of the data for the initial clustering or like looking for a smarter initial clustering that would avoid mixing regimes in the same cluster. Replacing the linear regressors by nonlinear functions could also be a possibility, but then the number of parameters becomes larger and the fiability of the estimators decreases.

References

[1] Azzalini A., Bowman A.W. (1990), A Look at Some Data on the Old Faithful Geyser, *Applied Statistics*, **vol. 39** p. 357-365.

[2] Chow G.C. (1960), Tests of Equality Between Sets of Coefficients in Two Linear Regressions, *Econometrica*, **vol. 28** p. 591-605.

[3] Gonzalo J., Pitarakis J-Y. (2002), Estimation and Model Selection Based Inference in Single and Multiple Treshold Models, *Journal of Econometrics*, **vol. 110** p. 319-352.

[4] Hamilton J.D. (1989), New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle, *Econometrica*, **vol. 57** p. 357-384.

[5] Hennig C. (2000), Regression Fixed Point Clusters : Motivation, Consistency and Simulations, *Preprint 2000-02, Fachbereich Mathematik, Universitat Hamburg*

[6] Kohonen T. (1997), *Self-Organizing Maps*, New-York, Springer-Verlag.

[7] Letremy P. (2000), Notice d'installation et d'utilisation de programmes bases sur l'algorithme de Kohonen et dedies a l'analyse des donnees, *Prepub. Samos 131.*

[8] Rynkiewicz J.(1999), Hybrid HMM/MLP Models for Time Series Prediction, *ESANN'1999 Proceedings*, p. 455-462.

[9] Tong H. (1978), On a treshold model, *Pattern Recognition and Signal Processing*, ed C.H. Chen, Amsterdam : Sijhoff&Noordhoff.

[10] Toyoda T. (1974), Use of the Chow Test under Heteroscedasticity, *Econometrica*, **vol. 42** p. 601-608.

[11] Weigend A.S., Mangeas M., Srivastava A.N.(1995), Nonlinear gated experts for time series : discovering regimes and avoiding overfitting, *International Journal of Neural Systems*, **vol. 6** p. 373-399.