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, each of them linear, are driving the series behaviour. 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 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.

2 The Method

The problem of finding the "true" number of regimes can be rewritten as a classification problem 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 $(T-p)$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 based on the following remark. 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 "squared error", which we define, in the second case, as the sum of squared residuals within each class.

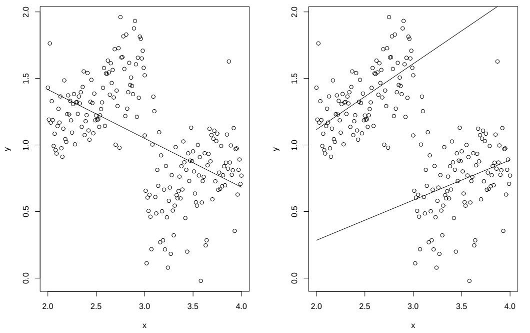

Figure 1: Fitting clustered data

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. We have chosen to use Kohonen maps to do this task because they provide homogeneous clusters and we can use the topology conservation to fasten the hierarchical classification algorithm. 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, $I_i = \sum_{j=1}^{N_i} d^2\left(x_{i,j}, g_i\right)$ 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 sum of squared errors. Obviously, in this case, the inter-class variation cannot be defined.

For a fixed number of classes, k, the total sum of squared errors is defined as $SSE_w = \sum_{l=1}^{k} SSE_{C_l}$, where the sum of squared errors for class l, $l = 1, ..., k$, is $SSE_{C_l} =$

$\sum_{t \in C_l} (y_t - \hat{y_t})^2$and $\hat{y_t}$ is the estimated value of y_t by the linear regression of order p fitted in the class C_l.

When passing from k to $k-1$ classes, if C_i and C_j were clustered, the total sum of squared errors becomes $SSE_w = \sum_{l=1, l \neq i, l \neq j}^{k} SSE_{C_l} + SSE_{C_i \cup C_j}$.

We want to minimize the increase of the total sum of squared errors, which means that we search the classes i_0 and j_0 such that :

$(i_0, j_0) = argmin_{i \neq j} \Delta_{i,j} SSE_w = argmin_{i \neq j} \left(SSE_{C_i \cup C_j} - SSE_{C_i} - SSE_{C_j} \right)$

The algorithm goes as follows :

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 :

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

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 total 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.

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 was taken from the paper of Gonzalo&Pitarkis.

$y_t = \begin{cases} -3 + 0.5 y_{t-1} - 0.9 y_{t-2} + \varepsilon_t & , y_{t-2} \leq 1.5 \\ 2 + 0.3 y_{t-1} + 0.2 y_{t-2} + \varepsilon_t & , y_{t-2} > 1.5 \end{cases}$, 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 we get good estimators for the parameters of the model when choosing two regimes as shown in Table 2.

When increasing the number of points in the sample, we will also increase the map size (we considered a 7x7 map for the 400 sample and a 9x9 for the 800). This was done in 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, but we'll skip them for parcimony purposes.

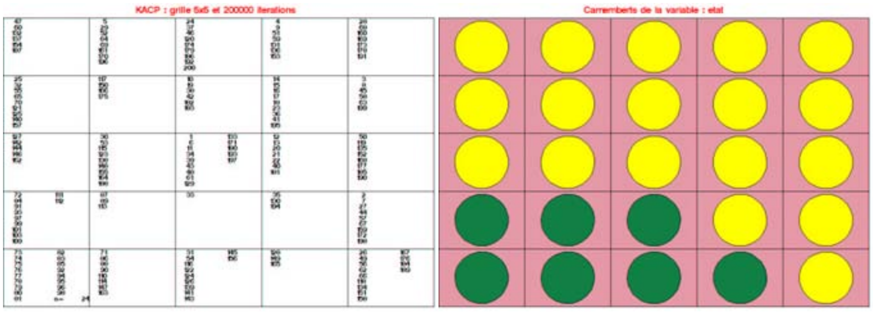
KACP : grille 5x5 et 200000 iterations
Camemberts de la variable : etat

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

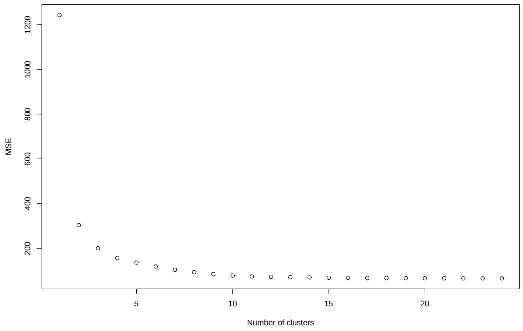

Figure 3: Squared error criterion for TAR model

3.2 A Two Regime Markov Switching Model

Let us first define an autoregressive Markov switching process. If $\{y_t\}_{t\in\mathbb{N}}$ is the observed time series, we will suppose that it follows a linear autoregressive process or order p, that we have, for example, two regimes and that the passage from one regime to the other is 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):

	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

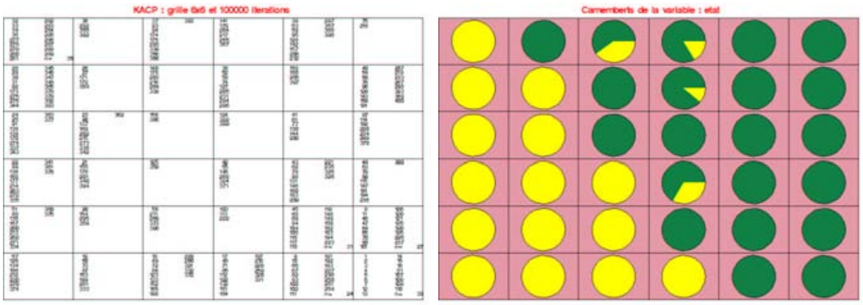

Figure 4: Initial clustering with a Kohonen map for two regimes Markov

	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

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

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).

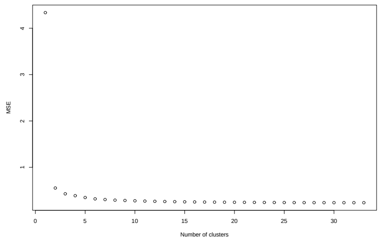

Figure 5: Squared error criterion for 2 regime Markov

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

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

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.

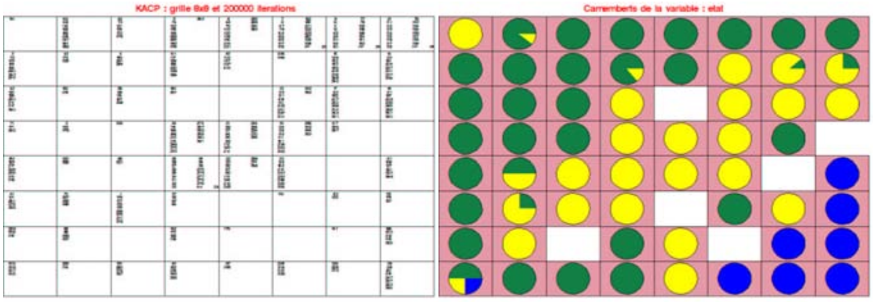

Figure 6: Initial clustering with a Kohonen map

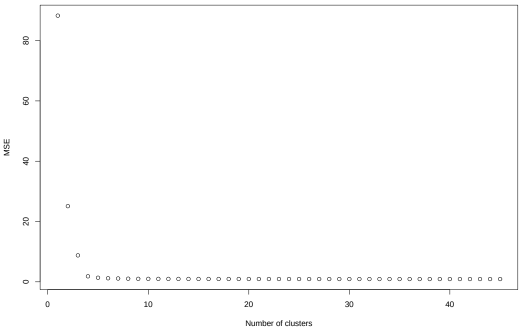

Figure 7: Squared error criterion for 3 regimes Markov

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. If we project the data on a two or three-dimensional space, the same four observations are far from the rest. Thus, we may conclude that the algorithm has identified a small class of outliers and that a three regime model would probably be the best choice.

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. Still, it is only a descriptive method and it should be used carefully. 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. As for the second question, no theoretical result is available yet, the econometricians prefering 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 a real data set. We have chosen two classical examples, the laser series from the Santa Fe time series prediction and analysis competition and the U.S. GNP (Gross National Product) series, used by Hamilton to introduce the switching Markov models.

For the laser series, a highly nonlinear data set, the algorithm selects three classes as shown in Figure 8. Ten lags of time were used and the 10000 observations were initially clustered by a 9x9 map. Once the number of regimes was fixed, the parameters were estimated using the EM algorithm as described in Rynkiewicz[5]. The prediction results are comparable with those obtained by Weigend[7] or Rynkiewicz[5] and the number of estimated parameters is smaller. Still, an adaptation of the method by replacing the linear regressions with nonlinear ones could be more interesting for this kind of data.

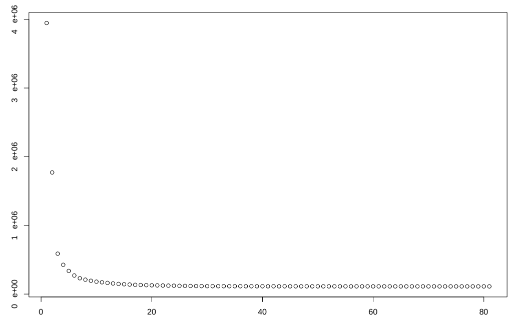

Figure 8: Squared error criterion for the laser series

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 9.

A first break appears when considering six classes instead of seven, but this can be interpreted as being due to the property of self-organizing maps of creating clusters strongly homegeneous. 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.

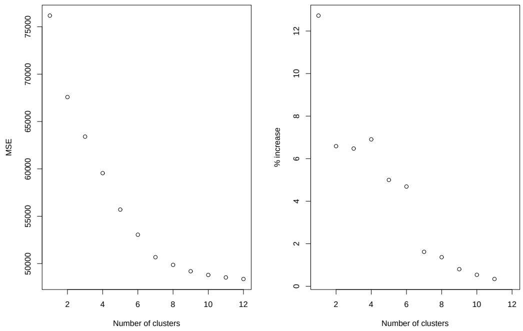

Figure 9: Squared error criterion for GNP data

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 looking for a smarter initial clustering that would avoid mixing regimes in the same cluster and replacing the linear regressors by nonlinear functions.

References

[1] 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.

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

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

[4] 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.*

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

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

[7] 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.