Parameter estimation of switching piecewise linear system

J. Ragot, G. Mourot, D. Maquin

Centre de Recherche en Automatique de Nancy, UMR CNRS 7039
Institut National Polytechnique de Nancy
2, avenue de la forêt de Haye. F 54 516 Vandoeuvre les Nancy
{jragot, gmourot,dmaquin@ensem.inpl-nancy.fr}

Abstract

During the last years, a number of methodological papers on models with discrete parameter shifts have revived interest in the so-called regime switching models. Piecewise linear models are attractive when modelling a wide range of nonlinear systems and determining simultaneously 1) the data partition 2) the time instant of change 3) the parameter values of the different local models. This is a difficult problem for which no solution exists in the general case and we show here some aspects and particular results concerning the problem of off line learning of switching time series. We propose a method for identifying the parameters of the local models when choosing an adapted weighting function, this function allowing to select the data for which each local model is active. Indeed the proposed method is able to solve simultaneously the data allocation and the parameter estimation. The feasibility and the performance of the procedure is demonstrated using several academic examples.

Keywords : piecewise system, switching, parameter estimation, data classification, signal segmentation.

1. Introduction

For the identification of nonlinear systems, there has been a large activity during the past years. In particular many interesting results have been reported in connection with multi-model [1] and/or multiple models, [2], hinging hyperplanes [3], [4], hidden Markov models [5], mixture of regressions [6], segmented curves [7]. Most of these works refer to quasistationnary or locally stationary systems characterised by abrupt changes between stationary segments with different statistical properties. Many formulations of this problem also appear in the field of fuzzy systems [8], [9].

In the following we focus the attention on PieceWise Auto Regressive Exogeneous models (PWARX). As it will be pointed out latter, if the partition of piecewise mapping is known, the problem of identification can easily be solved by using standard techniques of estimation. However, when the partition is unknown the problem becomes much more difficult. Thus, there are two possibilities. Either a partitioning, defining the local domains in which the system is constant, is a priori defined or the partitioning has to be estimated along with the local models.

Our contribution is to illustrate this problem in the case where the structure and the number of the local models are known. Thus, we restrict the estimation problem to (1) the estimation of switching between the local models, (2) the estimation of the parameters of the local models. Summarising, the main ideas of our contribution deal with the use of adapted weights allowing a powerful classification of the data and a sequential estimation of the different local model parameters.

This paper is organised as follows : section 2 explains, through a simple example, what is the problem to solve and the foregoing difficulties. Section 3 constitutes the contribution of the paper and is followed by a conclusion. Some simulation examples provide an illustration of the proposed algorithms both in section 4.

2. Model description

To begin with, let us consider systems in regression form

$$y_k = \theta_k^T \theta_j \qquad j = 1..s \tag{1a}$$

$$\text{if} \quad H_j^T \theta_k \neq 0 \tag{1b}$$

where $\varphi_k \in \mathbb{R}^p$ is a regression vector $\theta_j \in \mathbb{R}^r$ the parameter vector associated with the jth local model, and $H_j \in \mathbb{R}^r$ are unknown parameters. We then consider that the observations are generated by switching among s different AR models of orders p and parameters θ_j. Further we will use also the notation :

$$y_k = \varphi_k^T \theta(v) \tag{2}$$

where v is a key vector describing in what mode the system is for the time being ; v can be a function of (k,u,y) or some external input and takes its values in a finite set $I_s = \{1,\ldots,s\}$. Thus the time series is generated by the combination of s functions $\varphi_k^T \theta(v)$.

This is not the only way to describe switching system and the reader should refer to [4], [10], [11], [12], [13], [14] for

other formulations using mixture of models, endogenous switching, structural break models, self exciting threshold autoregressions (SETAR), smooth transition autoregressive model (STAR), neural network [15], and at last hybrid systems [16].

The regression vector φ could consist of old inputs and outputs. The sets $\Omega_j = \{H_j \varphi_k \varphi \}$, $j = 1..s$ are polyhedral partitions of the $\varphi \varphi$ *space* .
Our problem, when we are given y_k and φ_k, $k = 1..N$, consists in finding the PWARX model that best matches the given data, the number s being generally unknown. The model (1) can be identified by minimising the optimisation criterion :

$$J = \sum_{j=1}^{s} \sum_{k=1}^{N} \left(y_k - \varphi_k^T \theta_j \right)^2 \omega_j(\varphi_k) \qquad (3)$$

subject to :

$$\omega_j(\varphi_k) = \begin{cases} 1 & \text{if } H_j^T \varphi_k \geq 0 \\ 0 & \text{otherwise} \end{cases} \qquad (4)$$

where θ_j and H_j, $j = 1..s$, are the unknown.
An important problem is that of change-point detection, namely, detecting when the time-series has switched in some manner (eq. 1). At time k, for a given regression vector φ_k, only one condition (1b) is satisfied and exactly one of the functions $\omega_j(\varphi_k)$ equals 1. This means that all the terms in (3) are zero excepted one, those corresponding to the active local model. Consequently, if we known that a set of data is belonging to a particular local model, minimising (3) is straightforward. However, in general, the data partitioning is a priori unknown and the parameter estimation problem becomes difficult.

In our case, we limit the estimation problem to the one of θ_j ; however, we need to simultaneously estimate the values of the function ω_j in order to know the time switching i.e. the data useful to estimate the parameters of the jth local model. In fact we are not involved with the explanation of the switching, i.e. the estimation of the H_j parameters.

3. The main algorithm

Here we present our main contribution. Let y_k represents the output measurements of the underlying system and $y_{k,j}$ the output of the jth local model. To fit the local model to the data, we attempt to minimise the error function :

$$J = \sum_{k=1}^{N} \sum_{j=1}^{s} \left(y_{k,j} - y_k \right)^2 p_{k,j} \qquad (5)$$
$$y_{k,j} = \varphi_k^T \theta_j$$

where the weights $p_{k,j}$ have to be designed such that the local model j is adapted only with the input-output data for which it is concerned. It can be seen that the cost function (5) represents a trade-off between local and global learning. Indeed, when the model output $y_{k,j}$ is closed to the measurement y_k then model j matches the measurements and $p_{k,j}$ must be smaller than $p_{k,l}$, $\forall l \neq j$. In general, this performance index has to be minimised with respect to the parameter vectors θ_j to all possible disjoint partitions of the measurement set and to all possible numbers of submodels. Here we restrict the identification problem, the number of submodels being a priori chosen. Obviously, the key point is the design of these weights. In the following a non parametric estimation is used because there is no need to parameterize the weighting functions, only their values being useful to separate the data according to the s local models. The ideal situation deals with the knowledge of the partition of the data into s groups, the first one gathering the data in accordance with the first model, and similarly for the other groups. These s sets are noted S_j , $j = 1..s$:

$$S_j = \left\{ (x_k, y_k), k = 1..N / (x_k, y_k) \ \textit{satisfy model} \ j \right\}$$

Thus, the optimal weights are defined by :

$$p_{k,j} = \begin{cases} 1 & \text{if } (x_k, y_k) \in S_j \\ 0 & \text{if } (x_k, y_k) \notin S_j \end{cases} \qquad k = 1..N \qquad j = 1..s \qquad (6)$$

In fact our algorithm try to adapt the weights as closed as possible to the optimal ones.

The complete iterative algorithm is now described. Each iteration consists of two steps. The first one (step 1) is to determine an estimation of the weighting functions $p_{k,j}$ given the local models. The second step (step 2) is to identify the local models given the weights. Note that in [17] a similar model description is used in the context of weighted combination of local linear state-space systems but using a different approach based on an extended Kalman smoother allowing to estimate changes in the weights.
It should be noted that the proposed algorithm estimates sequentially these local models (and use a serial data allocation). More precisely, the algorithm estimates the first local model, the second local model explains the residuals of the first local model and so on. The algorithm uses an adapted weighting function allowing the clustering of the data automatically.

Algorithm : sequential estimation

Step 0. Initialisation
Select s the number of local models
Select a set of weighting matrices W_j for the s local models.
Define the matrices :

$$X = \left[\varphi_1 \ \cdots \ \varphi_N \right]^T \qquad \varepsilon_0 = y \qquad \hat{\theta}_0 = 0$$
$$y = \left[y_1 \ \cdots \ y_N \right]$$

Set $r = 0$

Step 1. Parameter computation
For $j = 1..s$:

residual regression : $\Delta\hat{\theta}_j = \left(X^T W_j X\right)^{-1} X^T W_j \Delta y_{j-1}$

residual estimation : $\Delta\hat{y}_{j-1} = X\Delta\hat{\theta}_j$

residual : $\Delta y_j = \Delta y_{j-1} - \Delta\hat{y}_{j-1}$

local model parameters : $\hat{\theta}_j = \hat{\theta}_{j-1} + \Delta\hat{\theta}_j$

residual criterion $\varepsilon_j = \left\|\Delta y_j\right\|^2_{W_j}$

Step 2. Weight computation
For $j = 1..s$

weights : $p_j = \prod_{\substack{q=1 \\ q \neq j}}^{s} \left(\Delta y_q\right)^{2r}$

normalised weights : $p_j = p_j / \sum_{j=1}^{s} p_j$

$W_j^{(r)} = diag\left(p_j\right),$

The operator $\otimes$ is used for evaluating the Hadamard product of vectors. The / operator allows to divide two vectors component by component. The "diag" operator allows to construct a diagonal matrix from a vector.

Step 3. Convergence test
Check for termination in some convenient matrix norm. If $\left\|W^{(r)} - W^{(r-1)}\right\| < \delta$ go to step 4, otherwie set $r = r + 1$ and return to step 1.

Step 4 Classification
The fuzzy data allocation is naturally given by the values of the weights. It is also possible to transform these weights into a binary representation involving only the values 0 and 1.

Remarks
- For the implementation issues, in step 2, the coefficient r enforced the weight and in our experience, it must be chosen between 2 and 4.
- The preceding algorithm supposes that matrices $X^T W_j X$ are regular. Indeed, it rarely occurs in practice that particular data and weights will cause singularity of $X^T W_j X$.
- The convergence of the algorithm is not discussed here and the reader may refer to [18] in which the use of EM and FCRM algorithms encounter the same convergence problem. As a evidence, the initialisation is a key point. It is important to note that, generally speaking, classification algorithms may terminate at extrema different from the true value. In our case, the proposed algorithm is quite enough insensitive to the initialisation used. However it is always possible to generate particular data for which the algorithm will trap at a local solution.

3. Examples

To verify the validity of the proposed algorithm and test its performance, we conducted several Monte-Carlo experiments with simulated data most of them being collected from systems used as benchmark in the literature. Here some results are presented.

Example 1

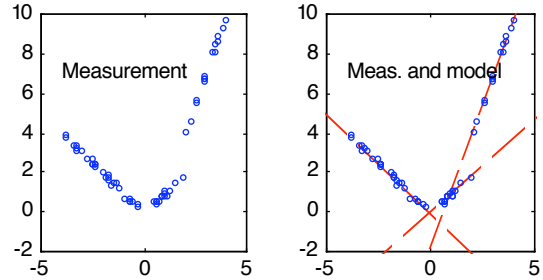
The data have been generated by the PWARX system (this structure has been proposed and analysed in [19] :

$$y(k+1) = au(k) + b + e(k) \qquad (7)$$
$$\begin{cases} a = -1 \quad b = 0 & if \quad u(k) \in [-4, 0] \\ a = 1 \quad b = 0 & if \quad u(k) \in [0, 2] \\ a = 3 \quad b = -2 & if \quad u(k) \in [2, 4] \end{cases}$$

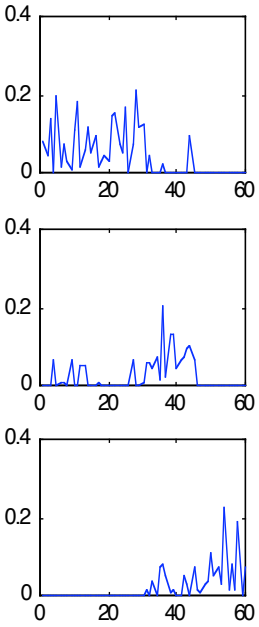
where the input $u(k) \in$ is a random sequence with uniform distribution on $[-4, 4]$ and where the noise $e(k)$ is a random sequence with standard deviation $\sigma = 0.1$. In that example, the three clusters have respectively, 30, 15 and 15 data. Ideally, cluster 1, 2 and 3 (corresponding to models 1, 2 and 3)) would respectively contain indices 1 to 30, indices 31 to 45 and indices 46 to 50. We have applied the algorithm 1 to these data when the number of local models is fixed to 3 (which corresponds to the exact number of clusters in the data). The identified parameters are :

$$\begin{array}{ll} & 1 : \ a = -0.999 \quad b = 0.011 \\ Model & 2 : \ a = 0.905 \quad b = 0.133 \\ & 3 : \ a = 2.917 \quad b = -1.754 \end{array}$$

Figure 1a present the data of the simulated system in the plan $\{y(k),\ u(k)\}$. Figure 1b simultaneously displays the data and the estimated model for each class. The vicinity of the data in regard to the local models provides a good approximation of the PWARX system (7). As an example of classification performance, fig. 1c (left) show the estimation errors and fig. 1d (right) the weighting functions for local models 1, 2 and 3 at iteration 7 (after convergence of the algorithm). These weights allow to perform the classification of the data which is given here as membership coefficient normalised between 0 and 1 (it would be also possible to define a boolean classification). On the same figure 1d, the true data allocation is indicated by shading areas ; in the computation, of course, the class labels of each data point are not known to the algorithm which "see" all the data simply as points $\{y_{k+1}, u_k\} \in$.



Figures 1a and 1b. The PWARX data sets in the plane $\{y(k+1),\ u(k)\}$.



Figures 1c. Estimation erro versus time

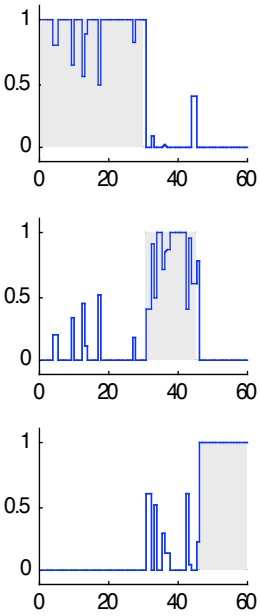


Figure 1d. Weights (data classification) versus time

Example 2

The time series $y(k)$ is defined in [11]

$$
\begin{aligned}
&x(k+1) = ax(k) + be(k)\\
&y(k) = x(k) + \varepsilon(k)\\
&a = \begin{cases} 0.9 & \text{if } |y(k)| \geq 0.7\\ -0.9 & \text{otherwise} \end{cases}\\
&b = 0.2
\end{aligned}
\qquad (8)
$$

The random term $e(k)$ is a normally distributed white noise process with zero mean and variance 0.0625 ; the noise $\varepsilon(k)$ is also normally distributed with variance 0.02. Figure 2 shows the data and the result. Rows 1 and 2 present the output evolution and those of the parameter a only taking one of the two values -0.9 or 0.9. Row 3 at left shows the data in the plane $\{y(k), y(k-1)\}$ while the right part compare the data with the obtained model. The normalised and rounded weights (taking only the values 0 and 1) given by the algorithm are drawn on row 4 and perfectly agree with the evolution of the a parameter ; these weights may be used to allocate the data to the two local models. The final clusters represented are correctly defined. Note that there is no hyperplan that separate the data sets (in the coordinate space $y(k)/y(k-1)$) because the clusters are not convex. In our approach, the clusters are defined through the weights that have been estimated simultaneously with the model parameters. Row 5 of figure 2 allows to compare the measured and the reconstructed output using the estimated parameters and time switching. Excepted in the vicinity of time origin for which bad initial conditions justify discrepancy, reconstructed state agrees with the true one.

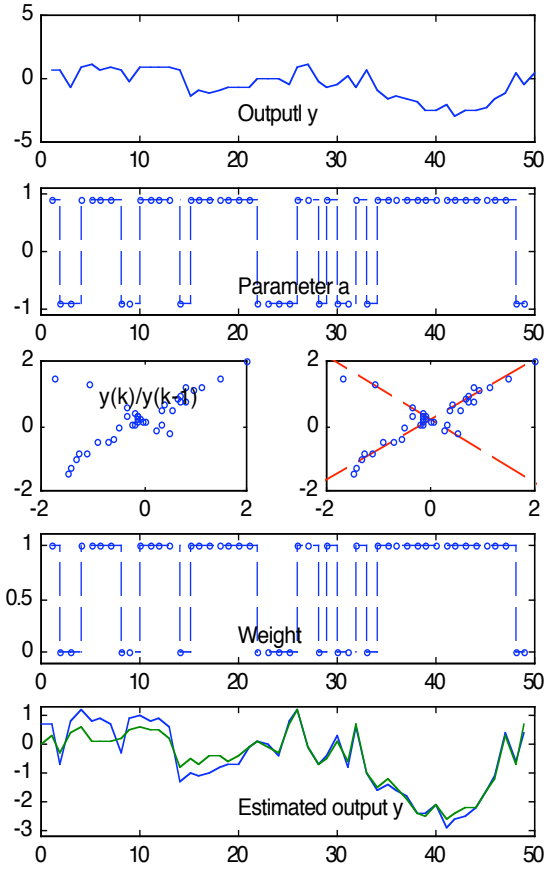


Figure 2. Data and estimations.

Example 3. A hybrid tank system

The identification procedure is now applied to the simulation data of a tank system shown in figure 3a. The valve V can be continuously manipulated whereas the flow output only linearly depends on the level in the tank. The system's hybrid nature results from the interaction of the continuous dynamics and the discrete event dynamics and vice versa. The continuous dynamic depends on the liquid level in the tank while the dynamics switches if the levels rise above of fall beneath the height h_s for which the section of the tank is changing.

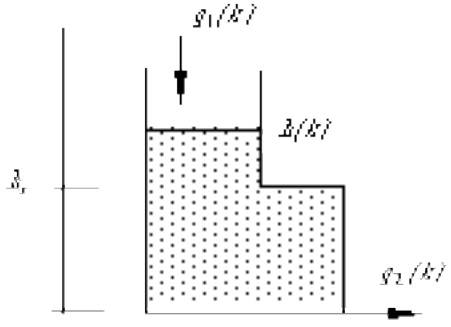


Figure 3a : An "hybrid" tank

The model of the system is piecewise linear :

$$
\begin{aligned}
&h(k+1) = h(k) + S(v(k))\left(q_1(k) - q_2(k)\right)\\
&q_2(k) = Kh(k)\\
&v(k) = \begin{cases} 0 & \text{if } h(k) > h_s \\ 1 & \text{if } h(k) < h_s \end{cases}\\
&S(v(k)) = S_1 + (S_2 - S_1)v(k)
\end{aligned}
\qquad (9)
$$

Figure 3b gathers the data and the results. Rows 1 to 3 respectively indicate the input, the level and the output of the process while rows 4 and 5 respectively present the switching according to the section modification and the estimated switching ; excepted for one value at time 36, the estimated switching perfectly agrees with the true ones.

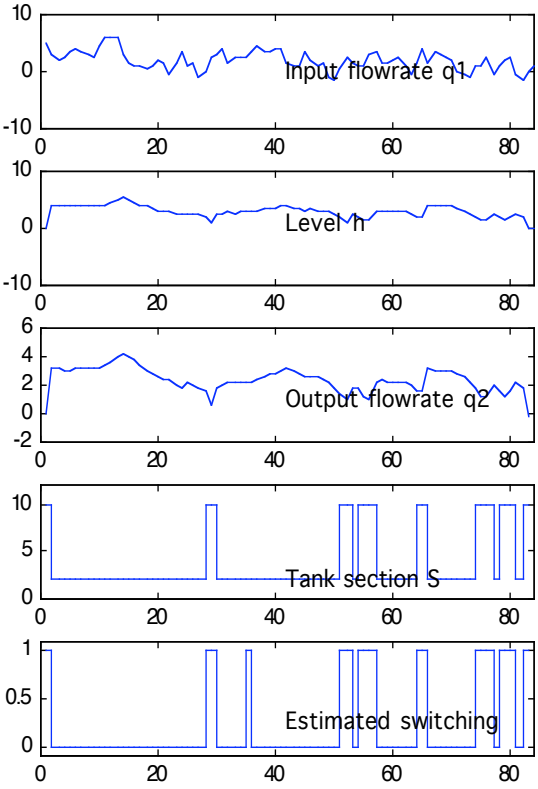


Figure 3b. Data and estimation versus time.

Example 4. System with unknown a priori number of models

All the previous examples use an a priori knowledge about the number of local models. Here, the given example shows how a false hypothesis concerning the number of local models influences the identification results. The simulated system is described by a first order model taking three different sets of parameters. The measurements $y(k)$ are corrupt by a white noise process $\varepsilon(k)$ with zero mean and standard deviation 0.1 :

$$
\begin{aligned}
&x(k+1) = ax(k) + bu(k)\\
&y(k) = x(k) + \varepsilon(k)
\end{aligned}
\qquad (10)
$$

$$
\begin{cases} a = 1 & b = 2 \\ a = 2 & b = 8 \\ a = -1 & b = 3 \end{cases}
$$

The procedure is performed with 4 local models. Figure 4a shows the measurements (left) and the obtained models superposed to the measurements (right). The data classification is given in figure 4b which also shows the true allocation (grey boxes). The estimated parameters are collected in table 2.

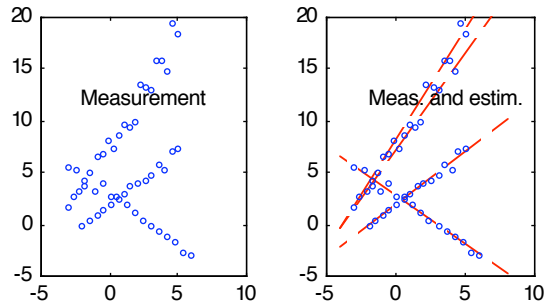


Figure 4a. Data and estimated model in plane $\{y(k+1) / u(k),\ y(t) / u(k)\}$

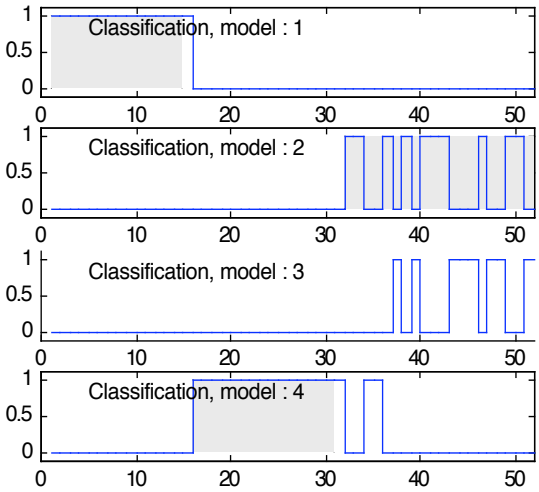


Figure 4b. Classification versus time

	model 1	model 2	model 3	model 4
a	0.994	2.143	2.413	-0.960
b	2.075	8.094	8.260	2.961

Table 2. Estimated parameters

Analysing the results clearly points out that the structure of the model is not optimal. There are several ways to analyse the problem of structure optimisation of the complete model [1], [20], [21]. The general ideal is to detect the presence of neighbouring local models or the presence of local model with a very few number of associated data. In the present situation, according to the model coefficients in table 2, two local models 2 and 3 have to be merged. For that purpose the merging may be performed by constraining models 2 and 3 to have the same behaviour, i.e. the same parameter vector. The new results are not presented, but they confirm the optimal structure with only three local models.

Conclusion

The proposed approach combines the identification of the parameters of a piecewise linear or affine form and the clustering of the data. This allows to identify both the affine local models and the partition of the domain in which each local model is valid ; in other words we have solved the data allocation task which consists in discovering that several local models exist and separated the data into groups corresponding to each model. We have successfully applied the proposed approach to static and dynamic switching regressions.
Further investigations will firstly focus on the performances of the method namely in the case of noisy measurement. Secondly, the order selection of the local models together with their number needs to be further analysed and optimised. Thirdly, so far all the data sets we have deal with a relatively low dimension, a large scale simulation will provide more insight into the robustness of this approach.

References

[1] Gasso G., Mourot G. Ragot J. Structure identification in multiple model representation : elimination and merging of local models. IEEE Conference on Decision and Control, Orlando, Floride, 2001.

[2] Mihaylova L., Lampaert V., Bruyninckx H., Sweters J. Hysteresis functions identification by multiple model approach. International Conference MFI, Baden-Baden, 2001.

[3] Pucar P., Sjoberg J. On the hinge-finding algorithm for hinging hyperplanes. IEEE Transactions of Information Theory, 4 (3), p. 1310-1319, 1998.

[4] Breiman L. Hinging hyperplans for regression, classification and function approximation. IEEE Transactions on Information Theory, 39 (3), p. 999-1013, 1993.

[5] Ding Z., Hong L. An interactive multiple model algorithm with a switching markov chain. Math. Comput. Modelling, 25 (1), p. 1-9, 1997.

[6] Quandt R.E. The estimation of the parameters of a linear regression system obeying two separate regimes. Journal of the American Statistical Association, p. 873-880, 1958.

[7] Hudson D.J. Finding segmented curves whose join points have to be estimated. Journal of the American Statistical Association, 61, p. 1097-1129, 1966.

[8] Kim E., Park M., Seunghwan L., Park M. A new approach to fuzzy modeling. IEEE Transactions on Fuzzy Systems, 5 (3), p. 328-337, 1997.

[9] Yu J.R., Tzeng G.H., Li H.L. General fuzzy piecewise regression analysis with automatic change point detection. Fuzzy Sets and Systems, 119, p. 247-257, 2001.

[10] Krolzig H.M. Markov-switching vector autoregressive modellin, statistical inference and application to business and analysis. Lecture notes in Economics and Mathematical Systems, 454, Springer.

[11] Medeiros M., Veiga A., Resenda M.G.C. A combinatorial approach to piecewise linear time series analysis. Journal of Computational and Graphical Statistics, 11, March, No. 1, pages 236-258, 2000.

[12] Roll J. Robust verification and identification of piecewise affine systems. Thesis 899, Linköping, 2001.

[13] Chua L.O., Kang S.M. Section-wise piecewise linear functions : canonical representation, properties and applications. Proceedings of IEEE, 65, p. 915-929, 1977.

[14] Fontaine L., Mourot G., Ragot J. Segmentation d'électrocardiogrammes par réseau de modèles locaux. Troisième Conférence Internationale sur l'Automatisation Industrielle, Montréal, Canada, 7-9 juin 1999.

[15] Kehagias A., Petridis V. Predictive modular neural networks for unsupervised segmentation of switching time series : the data allocation problem. submitted paper, 2001 (http://users.auth.gr/~kehagiat/kehPub/journal/2001DatAlloc.PDF)

[16] Münz E., Krebs V. Identification of hybrid systems using a priori knowledge. Proceedings of the 15th IFAC World Congress, Barcelone, 2002.

[17] Verdult V., Verhaegen M. Identification of a weighted combination of multivariable local linear state-space systems from input and output data. Proceedings of the 40th IEEE Conference on Decision and Control, p. 4760-4765, 2001.

[18] Hathaway R.J., Bezdek C. Switching regression models and fuzzy clustering. IEEE Transactions on Fuzzy Systems, 1 (3), p. 195-204, 1993.

[19] Ferrari-Trecate G., Muselli M., Liberati D., Morari M. Identification of piecewise affine and hybrid systems. Proceedings of the American Control Conference, p. 3521-3526, 2001.

[20] Rao A.V., Miller J., Rose K., Gersho A. A deterministic annealing approach for parcimonious design of piecewise regression models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 21 (2), p. 159-173, 1999.

[21] Strikholm B., Teräsvirta T. Determining the number of regimes in a threshold autoregressive model using smooth transition autoregression. 13th model selection and evaluation EC2 conference, Bologna, 2002