Parameter estimation of switching systems

José RAGOT [1], Abdelfettah HOCINE[1], and Didier MAQUIN[1]
[1] Centre de Recherche en Automatique de Nancy, Institut National Polytechnique de Lorraine,
UMR CNRS 7039. 2, avenue de la forêt de Haye, F 54 516 Vandoeuvre-les-Nancy
{jragot,ahocine,dmaquin}@ensem.inpl-nancy.fr

Abstract

Since several decades, researchers have been interested in various types of generalized regression models which admit changing parameter values at different time periods. The so-called regime switching models have given a lot of application in the fields of modelling of complex systems, robust identification, detection of behavior change and more generally in process diagnosis. Here, we examine the case where the change in regime cannot be directly observed but may be estimated from observed variables (the input and the output of the process). For that purpose, the well known EM (expectation-maximisation) approach may be applied; to take into account the switches between the regimes, new variables (generally known as hidden or missing) are introduced in order to construct a complete data likelihood function. In this paper, we show (i) how to directly formulate the estimation problem without introducing new variables, (ii) a natural way to solve the obtained equations using hierarchical calculus. An example is given to illustrate how to use the proposed approach.

Keywords

switching system, data classification, clustering, multi-model, segmentation, identification, expectation-maximisation.

1 Introduction

Many physical systems undergo episodes in which their behavior seems to be characterized by important changes. In this respect, one may define changing as a switch from one regime to another ; this idea was first introduced by [11] in the case of independent switches in a regression model. On a practical point of view, switching may be related to local modeling. When it is not possible to describe a process on a large domain, it is a natural way to examine the construction of local models only valid on a particular range of operation. Generally local modeling is a simple task because, locally, there are only a few number of phenomena and thus a few number of parameters to deal with. The modeling framework, that is based on using a local model for each predefined operating region, is called operating regime based model as introduced in [9]. For the identification of such systems, there has been a large activity during the past years. In particular, many interesting results have been reported in connection with multi-model [6] and/or multiple models [7], hybrid systems [1], hinging hyperplanes [10], [2], hidden Markov models [4].

In the following, we focus the attention on piecewise linear models. 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 estimation techniques. However, when the partition is unknown, the problem becomes much more difficult (see for example [6] in the field of multi-models). 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. In the first case, the number of local domains has to be chosen very large. If the amount of input-output data is sufficient in each domain, the parameter estimation of local model is generally easy ; otherwise, problems of ill conditioning often occurs. In the second case, a few number of local domains are used, but the simultaneous estimation of their number and of the parameters of the local models generally leads to potentially many local minima. Although the number of local models may be fixed by the user based on additional knowledge on the system, the major difficulty deals with the data partitioning. For a given number of local models, all possible disjoint partitions of the data have to be known which generally forms an infinite set. However, for discrete data, an enumerable subset may be found but the number of partitions increases with the number of data. Instead of discontinuous transients between the local models there is a temptation to use smooth transition ; that allows to describe the system under consideration with a continuously differentiable approximation [5], [8]. Our contribution is to illustrate this problem in the case where the structure and the number of 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. In section 2, some representations of switching systems are given, section 3 is devoted to our approach and section 4 presents some numerical results.

2 Switching models

The basic idea of regime-switching models is that the process is time invariant conditional on a regime variable indicating the regime prevailing at time t. Often, regime-switching model characterizes a non-linear process as piecewise linear by restricting the process to be linear in each regime:

$$x = \phi^T a_j + e \tag{1}$$

where x is the variable to be explain, $\phi \in \mathcal{R}^d$ the regression vector, $a_j \in \mathcal{R}^d$ the parameter vector for the particular regime j and e an error term. The error term will be consider as a random variable with a pdf represented by a family of finite linear gaussian mixture of the form :

$$p(e \mid \sigma) = \sum_{j=1}^{M} \alpha_j p_j(e \mid \sigma_j) \tag{2a}$$

$$p_j(e \mid \sigma_j) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{e^2}{2\sigma_j^2}\right) \quad j = 1..M \tag{2b}$$

$$\sum_{j=1}^{M} \alpha_j = 1 \tag{2c}$$

where $\sigma = \left(\sigma_1 \ldots \sigma_M\right)^T$ and where M is a known positive integer. In the following, we will use the vectors $\alpha = \begin{pmatrix}\alpha_1 \dots \alpha_M\end{pmatrix}$ and $a = \begin{pmatrix}a_1^T \dots a_M^T\end{pmatrix}^T$. Let us notice, that from (1) and (2) we have the pdf:

$$p_j(x \mid \theta_j) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{1}{2}\left(\frac{x - \phi^T a_j}{\sigma_j}\right)^2\right) \tag{3a}$$

$$\theta_j = \begin{pmatrix}a_j & \sigma_j\end{pmatrix}^T \tag{3b}$$

$$p(x \mid \Theta) = \sum_{j=1}^{M} \alpha_j p_j(x \mid \theta_j) \tag{3c}$$

$$\Theta = \begin{pmatrix}\alpha^T & a^T & \sigma^T\end{pmatrix}^T \tag{3d}$$

Then θ_j collects the parameters of the jth local model while Θ collects the whole set of parameters including the mixture parameters α_j. The problem to be solved can be stated as follows : given the number M of models, their respective orders and a set of observations, we have to determine first the clusters of data associated with each regime of functionning, second the parameters of each model.

3 Finding model parameters via EM approach

Our objective in this section is to find the likelihood function for the unknown parameters. The mixture density estimation problem for a given M sources involves fitting the component density σ_j and mixing coefficients α_j. For example, if we consider a switching time series generated by a combination of two sources each source being activated for a particular time interval, the problem to be solved consists in finding the models of the two local series. More clearly, there are two major steps in the procedure. The first one is the data allocation, i.e. the separation of the observed data into two groups, one group corresponding to each active source. Data allocation may be considered as a problem of classification and to perform this classification one needs to have the models of the time series. The second step of the procedure deals with the identification of the parameters of these models ; for that purpose we need to know what are the data to be used for each model. It appears that we have entered a vicious circle. Now we try to explain this point and we propose a natural way to solve the mentioned difficulty. We will suppose that data vectors are independent and identically normally distributed. That allows to express the resulting density for the whole sequence X of observations $x_i, i = 1 \dots N$:

$$p(X \mid \Theta) = \prod_{i=1}^{N} p(x_i \mid \Theta) \tag{4}$$

where $p(x_i \mid \Theta)$ is deduced from (3c) using the particular measurement x_i. Thus, using (3c) and (4) allows to built the likelihood function of the parameters Θ:

$$\mathcal{L}(\Theta) = \prod_{i=1}^{N} \sum_{j=1}^{M} \alpha_j p_j(x_i \mid \theta_j) \tag{5}$$

where $p_j(x_i \mid \theta_j)$ is derived from (3a) with the particular observation x_i. In the maximum likelihood problem, our goal is to find Θ that maximises $\mathcal{L}(\Theta)$ or equivalently its logarithm. That is, we wish to find $\hat{\theta}$ where :

$$\hat{\Theta} = arg \max_{\Theta} \sum_{i=1}^{N} \log \sum_{j=1}^{M} \alpha_j p_j(x_i \mid \theta_j) \tag{6}$$

Obtaining the estimation (6) is not straightforward and most of the time there does not exist explicit solutions. So far, the EM algorithm proposed in [3] has been the most widely and iterative approach to the estimation of mixture distribution parameters. When optimizing the likelihood function is analytically intractable, the underlying idea is to assume the existence of values for additional but missing (or hidden) variables. The hidden variables can, for example, be discrete component labels which represent class labels for the observed data thus allowing their allocation to specific local models. Assuming a joint relationship between the missing (Y) and observed variables (X) allows to define a new likelihood function $\mathcal{L}(\Theta \mid X, Y)$ called the complete likelihood. In the paper, we wish to examine the optimisation problem defined in (4) without using the EM formulation and without defined missing variables. The proposed procedure is the following.

- First we find the expression of the mixing parameters α_j and for that we introduce the Lagrange parameter λ for taking into account the normalization (3c) upon the mixing parameters α_j. Let us define the Lagrange function:

  $$\Phi = \sum_{i=1}^{N} \log \sum_{j=1}^{M} \alpha_j p_j(x_i \mid \theta_j) - \lambda \left(\sum_{j=1}^{M} \alpha_j - 1 \right) \tag{7}$$

  The optimum value of α_j has to satisfy $\partial \Phi / \partial \alpha_j = 0 \quad j = 1 \ldots M$ that gives:

  $$\sum_{i=1}^{N} \left(\frac{p_j(x_i \mid \theta_j)}{\sum_{j=1}^{M} \alpha_j p_j(x_i \mid \theta_j)} \right) - \lambda = 0 \quad j = 1 \ldots M \tag{8}$$

  Multiplying both sides of (8) by α_j and summing for j leads to express system (8) as:

  $$\sum_{i=1}^{N} \left(\frac{p_j(x_i \mid \theta_j)}{\sum_{j=1}^{M} \alpha_j p_j(x_i \mid \theta_j)} \right) = N \quad j = 1 \ldots M \tag{9}$$

  Thus, (9) describes a system of M equations with M unknown parameters α_j. It should be noted that resolution of (9) needs to know $p_j(x_i \mid \theta_j)$ and therefore the parameters θ_j. However, when $p_j(x_i \mid \theta_j)$ are known, and althougt (8) appears to be non linear in regard to the parameters, the resolution is not difficult. The structure of (9) allows to use efficient standard procedures for solving non linear equations ; with little more attention the particular structure of (9) may be used to develop specific schemes of resolution. In particular a powerful iterative estimation scheme, with guaranteed properties of convergence, may be proposed.

- The second step of the estimation procedure consists in deriving the parameters σ_j. Let us note that Φ, with (3a) may be written:

$$\Phi = \sum_{i=1}^{N} \log \sum_{j=1}^{M} \left(\alpha_j \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{1}{2} (\frac{x_i - \phi^T a_j}{\sigma_j})^2 \right) \right) - \lambda \left(\sum_{j=1}^{M} \alpha_j - 1 \right) \tag{10}$$

  Derivating (10) respect to σ_j allows to express the optimal value :

$$\sigma_j = \sqrt{\frac{1}{N} \sum_{i=1}^{M} \frac{p_j(x_i \mid \theta_j)}{\sum_{j=1}^{M} \alpha_j p_j(x_i \mid \theta_j)} (x_i - \phi_i^T a_j)^2} \quad j = 1 \dots M \tag{11}$$

- The third step is concerned with the estimation of the local model parameters a_j. Derivation of 10) in respect to the a_j yields :

$$\sum_{i=1}^{N} \frac{\alpha_j p_j(x_i \mid \theta_j)}{\sigma_j^2} \frac{\phi_i (x_i - \phi_i^T a_j)}{\sum_{j=1}^{M} \alpha_j p_j(x_i \mid \theta_j)} = 0 \tag{12}$$

  With the following definitions :

$$W_j(\theta) = diag \left(\frac{\alpha_j p_j(x_i \mid \theta_j)}{\sum_{j=1}^{M} \alpha_j p_j(x_i \mid \theta_j)} \right) \quad i = 1 \dots N \tag{13}$$

  we deduce from (12) :

$$a_j = (H^T W_j(\theta) H)^{-1} H^T W_j(\theta) x \tag{14}$$

  with $H = \begin{pmatrix} \phi_1^T \\ \dots \\ \phi_N^T \end{pmatrix}$ and $x = \begin{pmatrix} x_1 \\ \dots \\ x_N \end{pmatrix}$

Summarizing the three previous steps, the whole estimation (a_j, σ_j, α_j) is solved through the following system:

$$\sum_{i=1}^{N} \left(\frac{p_j(x_i \mid \theta_j)}{\sum_{j=1}^{M} \alpha_j p_j(x_i \mid \theta_j)} \right) = N \quad j = 1 \dots M \tag{15a}$$

$$\sigma_j = \sqrt{\frac{1}{N} \sum_{i=1}^{M} \frac{p_j(x_i \mid \theta_j)}{\sum_{j=1}^{M} \alpha_j p_j(x_i \mid \theta_j)} (x_i - \phi_i^T a_j)^2} \quad j = 1 \dots M \tag{15b}$$

$$a_j = (H^T W_j(\theta) H)^{-1} H^T W_j(\theta) x \tag{15c}$$

with the definitions:

$$p(x \mid \theta_j) = \frac{1}{\sqrt{2\pi}\sigma_j} exp\left(-\frac{1}{2}\left(\frac{x-\phi^T a_j}{\sigma_j}\right)^2\right) \tag{16a}$$

$$\theta_j = \begin{pmatrix} a_j & \sigma_j \end{pmatrix} \tag{16b}$$

$$W(\theta) = \begin{matrix} \text{diag} \\ i=1..N \end{matrix} \left(\frac{\alpha_j p_j(x_i \mid \theta_j)}{\sum_{j=1}^{M} \alpha_j p_j(x_i \mid \theta_j)}\right) \tag{16c}$$

Thus, despite the appparent complexity of the obtained system, we get a set of analytical implicit equation allowing to estimate the parameters of the local models a_j, σ_j and the mixing proportions α_j. Solving (16) may be refered to global numerical analysis techniques ; however, we suggest a hierachical procedure allowing to estimate sequentially the different parameters. This hierarchical procedure uses the principle of direct iteration on the different variables to be estimated. After an initialisation step (give initial values to $\theta_j = (a_j, \sigma_j)$), a computation cycle is made up with successive steps :

- from (16a), knowing (θ_j), calculate $p_i(x_i \mid \theta_j)$
- from (15a), knowing $p_i(x_i \mid \theta_j)$ deduce α_j
- from (16c), knowing α_j and $p_i(x_i \mid \theta_j)$, compute the weights W
- from (15b), knowing W_j and the measurementx_i, update the model parameters a_j
- from (15c), knowing α_j, $p_j(x_i \mid \theta_j)$ and a_j, update σ_j.

This cycle is started again unless a satisfaction terminal criterion has been reached i.e. the classification of all the data does not change significantly.

4 Results

To verify the proposed algorithm and test its ability to identify switching systems, we conducted several Monte-Carlo experiments with simulated data. Here, we present only one result obtained for one data set. We generated a process of 300 points composed of two segments of ARX processes with switches at time 40, 90, 150 and 240. The two processes have the respective sets of parameters indicate in table 1 (line 2, true parameters). The whole process is thus described by :

$$y_{t+1} = a_j y_t + b_j u_j \left\{ \begin{array}{lll} j=1 & if & t \in [0,40] \bigcup [90,150] \bigcup [240,300] \\ j=2 & if & 41 \le t \in [41,89] \bigcup [151,239] \end{array} \right. \tag{17}$$

The input and output data of the process are shown on figure (1); the vertical lines on the output graph indicate the switching times. Another way to describe the evolution consists in using the coordinates y_{t+1}/u_t and y_t/u_t. With this change, the model is then represented by two clusters of data belonging to the three lines :

$$y_{t+1}/u_t = a_j y_t/u_t + b_j \quad j = 1, 2, 3 \tag{18}$$

Figure (2) shows the shape of these clusters. The left part of the figure shows the points whose coordinate are y_{t+1}/u_t and y_t/u_t and we notice that it is somewhat difficult to separate the two regime if the noise measurement is important. The right part of the figure shows the preceding points and the true model. According to the algorithm derived in section 3, we have to fix the number and the structure of local models ; here, no optimization of these parameters has been done, and we used 2 local of first order. Moreover, initial values of the model parameters have to be chosen ; here, random generator has been used to select initial values (see numerical values in table 1, line 3), while $\sigma_j = 1, \ \ j = 1, 2$.

For that example, the data have been generated with noise and consequently the estimated parameter are in the vicinity of the true ones. The estimated parameters are presented in table 1 (line 4) and they provide a good approximation of the true system (see estimated model on figure (3) which looks like the true system of figure (2); so far we have obtained an estimate of the affine local models, each model characterizing a particular regime of the system.

The final step is to look to the shape of the regions in which each local system is working. The direct procedure has the drawback to present some time fluctuations of W (16c), although the corresponding data are issued from the same regime of functionning. To avoid or to reduce this effect, which causes instantaneous transitions among the local models, we suggest to filter the $p_j(x_i \mid \theta_j)$. There are several techniques that can be used and among them moving average, winsorizing moving average and boolean comparison of the probabilities values.

- Moving average is proceeded on a window of length $2L+1$ according to the following definition :

  $$\tilde{p}_j(x_i \mid \theta_j) = \frac{\sum_{k=i-L}^{i+L} p_j(x_k \mid \theta_j)}{2L+1} \quad j = 1 \dots 2 \tag{19}$$

  The user has to adjust the length $2L+1$ of the window in order to get a compromise between the quality of the data classification and the ability of the algorithm to avoid spikes in the transitions.

- Winsorizing moving average computes the mean of $2L+1$ values after the s smallest values are replaced by the $(s+1)$st smallest value and the s largest values are replaced by the $(s+1)$st largest value (i.e. the values are winsorized at each end):

  $$\tilde{p}_j(x_i \mid \theta_j) = \frac{1}{L}\left(s p_j(x_{(i-L+s)} \mid \theta_j) + \sum_{k=i-L+s}^{i+L-s} p_j(x_{(k)} \mid \theta_j) + s p_j(x_{(i+L-s)} \mid \theta_j) \right) \tag{20}$$

  where $p_j(x_{(i)} \mid \theta_j)$ represents the probability values, on the $(2L+1)$ length window, sorted in ascending order.

par.	a_1	b_1	a_2	b_2
true	0.900	0.150	0.700	0.300
initial	-0.208	1.208	-0.098	1.098
estimated	0.891	0.156	0.706	0.299

Table 1: Parameters

- Comparison, at each time, of the probabilities $p_1(x_i, \theta_1)$ and $p_2(x_i, \theta_2)$ allows to select the most important probability to perform a boolean classification of the data x_i to one cluster. Thus the classification rule become:

$$\left.\begin{array}{l}\tilde{p}_1(x_i \mid \theta_1) = 1 \\ \tilde{p}_2(x_i \mid \theta_1) = 0\end{array}\right\} \text{if } p_1(x_i, \theta_1) > p_2(x_i, \theta_2) \tag{21}$$

$$\left.\begin{array}{l}\tilde{p}_1(x_i \mid \theta_1) = 0 \\ \tilde{p}_2(x_i \mid \theta_1) = 1\end{array}\right\} \text{if } p_1(x_i, \theta_1) < p_2(x_i, \theta_2) \tag{22}$$

The filtering effect allows to better classify the data in regard to the two cluster and therefore to express what data are to identify the parameters of each model. It is also possible to combine the effects of the different filters (19), (20), (21), (22).

After the algorithm has converged (for the given example, some 30 iterations), figure 4 shows the probabilities $p_1(x_i, \theta_1)$ and $p_2(x_i, \theta_2)$. As indicated, filtering allows to suppress fluctuations and has only be used to present the final classification (figures 5, 6 and 7). The vertical lines at 40, 90, 150 and 240 indicate the switching of the system; a good approximation of the duration regime has been obtained excepted at the switching time where some confusion appears (explained by the fact that the data may belong either to the first regime or to the second).

5 Conclusion

This paper has presented an approach for data and signal segmentation and ARX modeling of switching system. Assuming that the number of local models and their orders is known, the estimation problem consists in finding the allocation of the measurement data to each local model and finding the best local model for each cluster of data. The combined estimation of the likelihood function is performed using a hierarchical iterative resolution. In fact, the proposed procedure may be linked to the EM family algorithm ; however, the EM formulation involving hidden or missing variables is not used. The extension of the proposed procedure to an unknown number of local models will be investigated. Further works will also be concerned with the optimal model orders selection. At last, the effects of outliers on parameter estimates and on inference in linear models has been studied quite extensively ; for nonlinear models, this is not the case and it would be interesting to extend the definition of additive and innovation outliers to switching models.

References

[1] A. Bemporad, J. Roll and L. Ljung, “Identification of hybrid systems via mixed-integer programming”, 40th IEEE Conference on Decision and Control, pp. 786-792, Orlando, Florida, USA, December 4-7, 2001.

[2] L. Breiman, “Hinging hyperplanes for regression, classification and function approximation”, *IEEE Transactions on Information Theory*, 39 (3), pp. 999-1013, 1993.3.

[3] A.P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm”, *J. Royal Statistical Society*, Series B 39, pp. 1-38, 1977.

[4] Z. Ding and L. Hong, “An interactive multiple model algorithm with a switching markov chain”, *Math. Comput. Modelling*, 25 (1), pp. 1-9, 1997.

[5] L. Fontaine, G. Mourot and J. Ragot, “Segmentation d’électrocardiogrammes par réseau de modèles locaux”, Troisième Conférence Internationale sur l’Automatisation Industrielle, Montréal, Canada, 7-9 juin 1999 (in French).

[6] G. Gasso, G. Mourot and J. Ragot, “Structure identification in multiple model representation: elimination and merging of local models”, 40th IEEE Conference on Decision and Control, pp. 2992-2997, Orlando, Florida, USA, December 4-7, 2001.

[7] L. Mihaylova, V. Lampaert, H. Bruyninckx and J. Sweters, “Hysteresis functions identification by multiple model approach”, International Conference on Multisensor Fusion and Integration for Intelligent Systems, MFI’2001, Baden-Baden, Germany, August 20-21, 2001.

[8] E. Münz, T. Hodrus and V. Krebs, “Top-down identification of hybrid characteristic map”, Conference on Analysis and Design of Hybrid Systems, ADHS’03, pp. 34-39, Saint-Malo, France, June 16-18, 2003.

[9] R. Murray-Smith and T.A. Johansen, *Multiple model approach in modelling and control*, Taylor and Francis, 1997.

[10] P. Pucar and J. Sjöberg, “On the hinge finding algorithm for higing hyperplanes”, *IEEE Transactions of Information Theory*, 44 (3), pp. 1310-1319, 1998.

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

[12] J. Ragot, D. Maquin and M. Pekpe, “Signal segmentation and data classification”, 10th International Workshop on Systems, Signals and Image Processing, IWSSIP’03, Prague, Czech Republic, September 10-11, 2003.Prague, 2003.

[13] J. Ragot, D. Maquin and E. Domlan, “Switching time estimation of piecewise linear systems. Application to diagnosis”, 5th IFAC Symposium Fault Detection, Supervision and Safety of Technical Processes, Safeprocess’2003, Washington D.C., USA, June 9-11, 2003.

[14] J. Roll, “Robust verification and identification of piecewise affine systems”, Technical Report Licentiate Thesis no. 899, Department of Electrical Engineering, Linköping University, SE-581 83 Linköping, Sweden, September 2001.

[15] V. Verdult and M. Verhaegen, “Identification of a weighted combination of multivariable local linear state-space sysems from input and output data”, 40th IEEE Conference on Decision and Control, pp. 4760-4765, Orlando, Florida, USA, December 4-7, 2001.

[16] J.R. Yu, G.H. Tzeng, H.L. Li. “General fuzzy piecewise regression analysis with automatic change point detection”. Fuzzy sets and systems, 119, pp. 247-257, 2001.

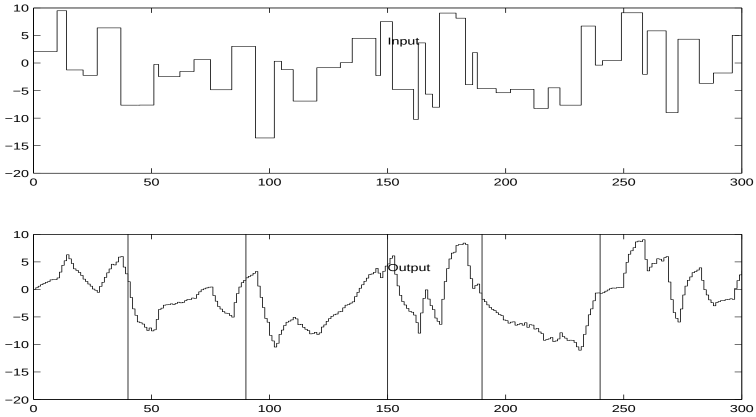

Figure 1: Input and output measurement

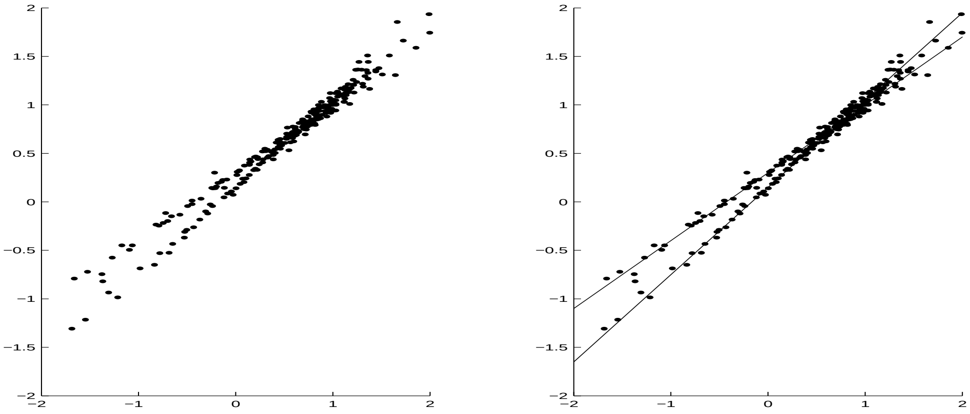

Figure 2: Input and output measurement

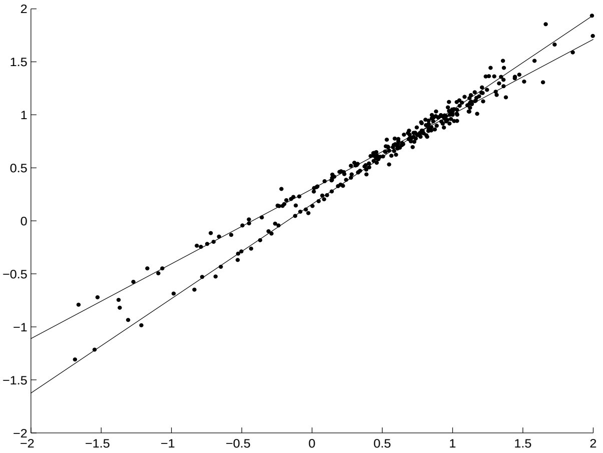

Figure 3: Input and output measurement

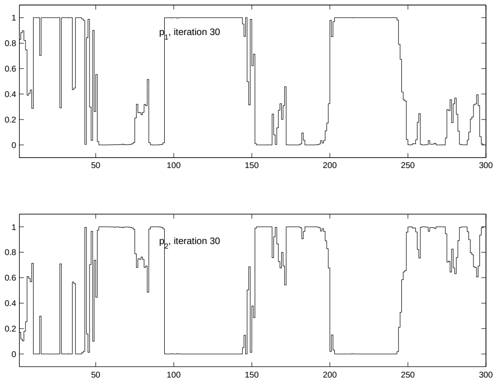

Figure 4: Model probabilities. Iteration 30

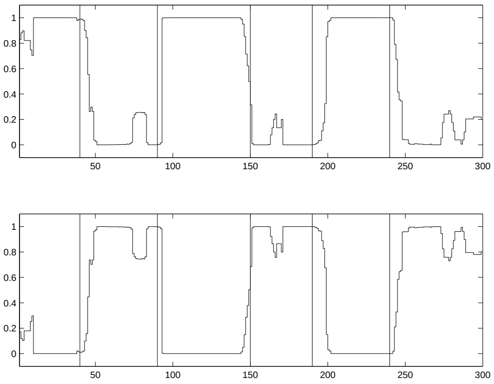

Figure 5: Model probabilities. Iteration 30, filtering with winsorizing.

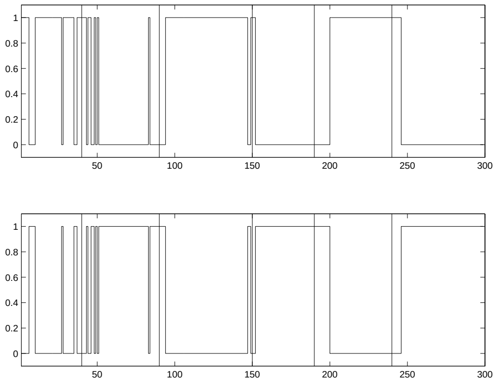

Figure 6: Model probabilities. Iteration 30, boolean filtering

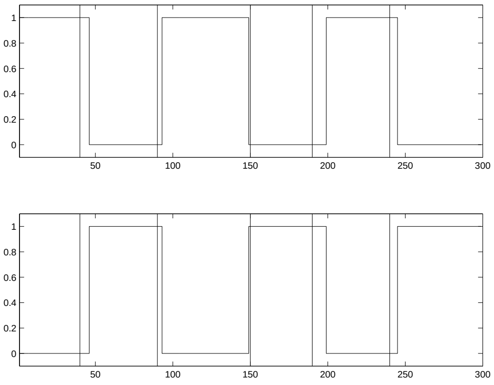

Figure 7: Model probabilities. Iteration 30, boolean-winsorizing filtering