

# Evaluation of relevance of stochastic parameters on Hidden Markov Models

B. Roblès, M. Avila, F. Duculty & P. Vrignat

*PRISME Laboratory, MCDS team, University of Orleans, France*

F. Kratz

*PRISME Laboratory, MCDS team, ENSI, Bourges, France*

**ABSTRACT:** Prediction of physical particular phenomenon is based on knowledge of the phenomenon. This knowledge helps us to conceptualize this phenomenon around different models. Hidden Markov Models (HMM) can be used for modeling complex processes. This kind of models is used as tool for fault diagnosis systems. Nowadays, industrial robots living in stochastic environment need faults detection to prevent any breakdown. In this paper, we wish to evaluate relevance of Hidden Markov Models parameters, without a priori knowledges. After a brief introduction of Hidden Markov Model, we present the most used selection criteria of models in current literature and some methods to evaluate relevance of stochastic events resulting from Hidden Markov Models. We support our study by an example of simulated industrial process by using synthetic model of Vrignat's study (Vrignat 2010). Therefore, we evaluate output parameters of the various tested models on this process, for finally come up with the most relevant model.

## 1 INTRODUCTION

According to (Vrignat et al. 2010), we find two keywords in maintenance definition: *maintain* and *restore*. The first one refers to preventive action. The second refers to corrective action. Thus, maintenance optimization for reliability determines "optimal" preventive maintenance. Events preceding a problem in maintenance activities are often recurrent. Special events series should inform us on next failure. For example, in mechanical systems, noises, vibrations precede failure. The loss of performances reflects failure or technical faults. Our works (Vrignat et al. 2010) show that it is possible to model degradation levels of a process and results show that our approach combined with work of (Zille, Bérenguer, Grall, Despujols, & Lonchamp 2007) can provide decision support for industrial maintenance. We also show (Vrignat et al. 2010) that our model provides a good failure prediction. With this, we make a reference model, named *synthetic model*, which fits to real industrial processes. Our research consist here to evaluate three different Hidden Markov Models topologies, with parameters outcome from this *synthetic model*. According to (Lebarbier and Mary-Huard 2004), problems of model selection (i.e. which model gives best failure prediction) are based on the minimization of penalty criterion. First criteria which appear in

literature are the *AIC: Akaike Information Criterion* (Akaike 1973), the *BIC: Bayesian Information Criterion* (Schwarz 1978).

In this work, the emphasis is on measuring relevance of Hidden Markov Models (HMM) parameters, based on several criteria used in current literature. Then, we try to evaluate best HMM topology. The structure is as follows: in section 2, we outline hidden Markov model and define its parameters. We present criteria used to evaluate relevance of HMM parameters (Shannon's entropy (Shannon 1948), likelihood, *AIC* and *BIC*), in section 3. Finally, we use our synthetic model to compare several HMM topologies, from among a candidate set, with previous criterion and try to give the best one, in section 4.

## 2 HIDDEN MARKOV MODEL

Hidden Markov Model (Rabiner 1989), (Fox et al. 2006) is an automaton with hidden states which consists of unobservable variable. This one represents the system status to be modeled. Only output variable is observable. Then we get observations sequence from output of the automaton; from now, we rename observations sequence as *symbols*, representing these observations (see an example of model figure 1). This is precisely relevance of

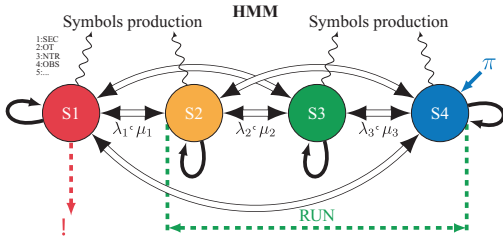


Figure 1. Four states Hidden Markov Model.

these symbols that we attempt to evaluate. Hidden Markov Model is characterized by the following:

- State number;
- Number of distinct observation symbols per state, observation symbols corresponding to the physical output of the system being modeled;
- Distribution probability of state transitions;
- Distribution probability of observation symbols;
- Initial state distribution.

### 2.1 Markov Assumption

States prediction is not made more accurate by additional priori knowledge information, i.e. all useful information for future prediction is contained in present state of the process.

$$P(X_{n+1} = j | X_0, X_1, \dots, X_n = i) = P(X_{n+1} = j | X_n = i). \quad (1)$$

### 2.2 Definitions for discrete Hidden Markov Model

Let us describe variables for HMM:

- Let  $N$ , the number of workable hidden states and  $E = \{E_1, E_2, \dots, E_N\}$ , the set of this variable. Let  $q_t$ , the value of this variable at time  $t$ ;
- Modeled process, must match to first-order Markov assumption (§2.1);
- Let  $T$ , the full number of observation symbols and let  $X = \{x_1, x_2, \dots, x_T\}$ , observations sequence of the modeled process;
- Let  $A = \{a_{ij}\}$ , distribution probability of state transitions with:

$$a_{ij} = P(q_{t+1} = E_j | q_t = E_i) \quad 1 \leq i, j \leq N, \quad (2)$$

- Let  $B = \{b_j(m)\}$ , distribution probability of observation symbol in  $j$  state, with:

$$b_j(m) = P(X_t = x_m | q_t = E_j) \quad 1 \leq j \leq N \quad 1 \leq m \leq T, \quad (3)$$

with  $X_t$ , value of observation variable at time  $t$ .

- Let  $\pi = \{\pi_i\}$ , initial state distribution with:  $\pi = P(q_1 = E_i) \quad 1 \leq i \leq N, \quad (4)$

- Hidden Markov Model will be set as:  $(A, B, \pi)$ ,
- $\lambda_i$  is failure rate and  $\mu_i$  is repair rate.

## 3 CRITERIA USED FOR EVALUATION

A lot of criteria in model selection are proposed in literature. We try to evaluate the best Hidden Markov Model topology proposed in (Vrignat 2010), by using Shannon's entropy (Shannon 1948), especially maximum entropy principle used in (Chandrasekaran et al. 2007). Calculation is made with states and observations: symbols productions of synthetic HMM §4. To emphasize our analysis, we also use some criteria which penalize likelihood value, in order to overcome over-parameterization models, like (Akaike 1973) and Bayes (Chen and Gopalakrishnan 1998) criteria.

### 3.1 Shannon's entropy

We now study notions of Shannon's entropy. It is a mathematical function which calculate the information rate contained in an information source. This source can be a text written in any language, an electrical signal or an unspecified electronic file...

#### 3.1.1 Entropy definition

Shannon's entropy is defined in (Cover and Thomas 1991) as follows:

$$H(S) = - \sum_{i=1}^n P_i \log_b P_i, \quad (5)$$

$P_i$  is the average probability to find the  $i$  symbol in  $S$ .

#### 3.1.2 Formal properties of Shannon's entropy

- Entropy value should be **minimal** if only one symbol is represented (uncertainty is null when there is only one event).

$$H(0, \dots, 0, 1, 0, 1, 0, \dots, 0) = 0. \quad (6)$$

- On the other hand, entropy value should be **maximal** if all symbols are equiprobable (uncertainty

is highest when all possible events are equipossible). For further information, reader will refer to (Beirlant et al. 1997) which present a state of the art about methods for entropy estimation and their properties.

### 3.1.3 Maximum entropy principles

The two principles of entropy's maximization in (Jaynes 1957) are the following:

- Principle of probabilities assignment to a distribution when we haven't enough informations on it;
- For all probability distributions that satisfy the constraints, we choose the one which has the maximum entropy according to Shannon.

Chandrasekaran in (Chandrasekaran et al. 2007) uses this 2nd principle for models selection, and (Arminjon and Imbault 2000) for building even more accurate models, by adding information. Our step consists in comparing the average entropies for various models. Value of average entropy would be then maximum for the most relevant model.

### 3.1.4 Entropic filter

We now introduce "Entropic Filter" concept. According to the 2nd principle of entropy stated in §3.1.3, we choose the model whose average entropy is maximum. On the other side, outliers values can generate miscalculation in real entropy value of the model. Especially *NTR* symbols (Nothing To Report) which are not useful for evaluation (entropy is maximum). *SP* (Stop Production) symbols have likewise been eliminated (entropy is null). Indeed, they are totally discriminated for *S1* state of HMM. To improve calculation of entropy, it is therefore better to eliminate these values. This approach is used through ID3 (Quinlan 1979) and C4.5 (Quinlan 1993) algorithm when creating decision tree, removing recursively attribute with zero entropy. In order to improve the calculation of entropy, we propose to eliminate discriminated symbols of zero entropy and the most representative symbols, where entropy is maximum. This operation will be named "Entropic Filter". We then calculate the average entropy of models to assess relevance of observation sequences. Best model is the one which has the best average entropy, after entropic filtering.

## 3.2 Maximum likelihood

Let us now turn to studying maximum likelihood principle. Let  $P_\alpha$  a statistical model, and  $X$ , an observation sequence, the probability to see  $X$  according to  $P$  can be measured by  $f(X, \alpha)$  function which represents the density of  $X$  when  $\alpha$  appears. Since  $\alpha$  is unknown, it seems natural to promote

values of  $\alpha$  where  $f(X, \alpha)$  is maximum: it is the notion of likelihood of  $\alpha$  for observation  $X$ .

– Expression of likelihood  $V$ :

$$V(x_1, \dots, x_n; \alpha) = \prod_{i=1}^n f(x_i; \alpha), \quad (7)$$

$\alpha$  is mathematical expectation,

A strictly increasing transformation does not change a maximum. Maximum likelihood can also be written as:

$$\log(V(x_1, \dots, x_n; \alpha)), \quad (8)$$

Then

$$\log(V(x_1, \dots, x_n; \alpha)) = \sum_{i=1}^n \log(f(x_i; \alpha)), \quad (9)$$

– For a discrete sample:

$$f(X; \alpha) = P_\alpha(X = x_i), \quad (10)$$

$P_\alpha(X = x_i)$  represents discrete probability where  $\alpha$  appears,

– Maximum likelihood for a discrete sample  $P_\alpha(x_i)$  representing the discrete probability where  $\alpha$  appears:

$$\log(V(x_1, \dots, x_n; \alpha)) = \sum_{i=1}^n \log(P_\alpha(x_i)). \quad (11)$$

Actually, we maximize the logarithm of likelihood function to compare several models. According to (Olivier et al. 1996), principle of maximum likelihood results in over-parameterization of the model to have good performances. Penalization of likelihood value can overcome this disadvantage. Most famous penalized log-likelihood criterion is the *AIC* (Akaike 1973), even if it is not completely satisfactory: it improves maximum likelihood principle but also led to an over-parameterization. Other traditional criteria, *BIC* and *HQC*, ensure a better estimation by penalizing oversizing models.

## 3.3 Akaike information criterion

According to (Ash 1990), entropy of a random variable is a regularity measurement. We can easily extend this concept to a model having several random variables. In the literature, Akaike Information Criterion (*AIC*) is often associated with another known criterion, called Bayes Information Criterion (*BIC*). In his report, (Lebarbier and

Mary-Huard 2004) describe all assumptions necessary to its implementation.

$$AIC = -2 \ln V + 2k, \tag{12}$$

$k$  is the number of free parameters,  $2k$  is the penalty,  $V$  is the likelihood.

Best model is the one which has the weakest  $AIC$ . This criterion uses maximum likelihood principle seen in (11). It penalizes models with too many variables, and avoids over-learning models.

### 3.4 Measurement of bayesian information criterion

$AIC$  criterion is often presented with Bayesian or Schwarz criterion:  $BIC$ , which more penalizes over-parameterized models.  $BIC$  criterion was introduced in (Schwarz 1978) and is different for the correction term:

$$BIC = -2 \ln V + k \ln(n), \tag{13}$$

$k$  is the number of free parameters of Markov Model (Avila 1996),  $n$  is the number of data,  $k \ln(n)$  is the penalty term.

Like  $AIC$ , best model is the one who gets the minimum value of  $BIC$ . Choosing between these two criteria is to choose between a predictive model and an explanatory model (Lebarbier and Mary-Huard 2004). It checks the validity of a particular model but it is mainly used to compare several models together.  $AIC$  criterion is less relevant than  $BIC$  for over-learning models.

## 4 SIMULATED INDUSTRIAL PROCESS

Nowadays, every industrial factory is using **preventive maintenance**. Maintenance agents can consign their actions and observations in a centralized database (see table 1). For example, symbols “PM, OT, SP, ...” could characterize maintenance activities carried out on industrial process. We recall the

Table 1. Example of recorded events.

Name	Date	Ope.	Cd	IT	N°	Code
Dupond	11/01/2007	Lubrication	PM	20	1	9
Dupond	11/01/2007	Lubrication	PM	20	2	9
Dupond	12/01/2007	Lubrication	SEC	30	3	5
Dupond	12/01/2007	Lubrication	PM	30	4	5
Dupond	13/01/2007	Padlock	PM	10	5	6
Dupond	13/01/2007	Padlock	NTR	30	6	5
Dupond	13/01/2007	Padlock	NTR	30	7	5
Dupond	16/01/2007	Lubrication	SP	90	8	1
Dupond	19/01/2007	Padlock	OT	10	9	3

meaning of selected symbols resulting from observations, in table 2. “SP” symbol corresponds to a stop of production units: process state = “STOP” in table 2. It is a critical condition that our research tries to minimize. Process state = “RUN” when production units are running without failure. We study here this kind of **maintenance** by using **synthetic model** (§4.1) to simulate real industrial environment. We choose “ $\lambda_i$ ” (failure rate) and “ $\mu_i$ ” (repair rate) of HMM parameters (Vrignat 2010), to match as possible, with maintenance consignment (table 1).

### 4.1 Synthetic model

We make our synthetic model with Matlab by using four states oriented model 2 presented in fig. 3(b). We use this model feature because it has good performance in maintenance activities (Vrignat et al. 2010). Then, we build sequences of data (also named “signature”) using this model as the reference model, by injecting “stochastic” symbols in this HMM. We use these symbols sequences as

Table 2. Symbolic coding system of maintenance interventions.

Process state	
RUN	
STOP	
Interventions type	
1	SP (Troubleshooting/Stop Production)
2	SM (Setting Machine)
3	OT (Other)
4	OBS (Observation)
5	PM (Preventive Maintenance, Production not stopped)
6	SEC (Security)
7	PUP (Planified Upgrading)
8	CM (Cleaning Machine)
9	PMV (Preventive Maintenance Visit)
10	NTR (Nothing to report)

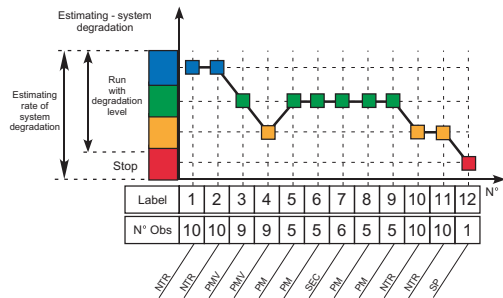


Figure 2. Degradation of process.

Markov chain (see table 3), to model degradation level of a process (example in figure 2). These simulated symbols, according to real industrial process (Vrignat et al. 2010), are obtained by using uniform and Gaussian distribution. We inject these symbols into three different HMM topologies, described in figure 3, by using two different learning and decoding algorithms:

- Baum-Welch learning (Baum et al. 1970), decoding by Forward (Rabiner 1989),
- Segmental K-means learning (Juang and Rabiner 1990), decoding by Viterbi (Viterbi 1967).

Table 3. Sequence of a message from maintenance database.

PM	PM	SEC	PM	PM	NTR	NTR	SP	...
----	----	-----	----	----	-----	-----	----	-----

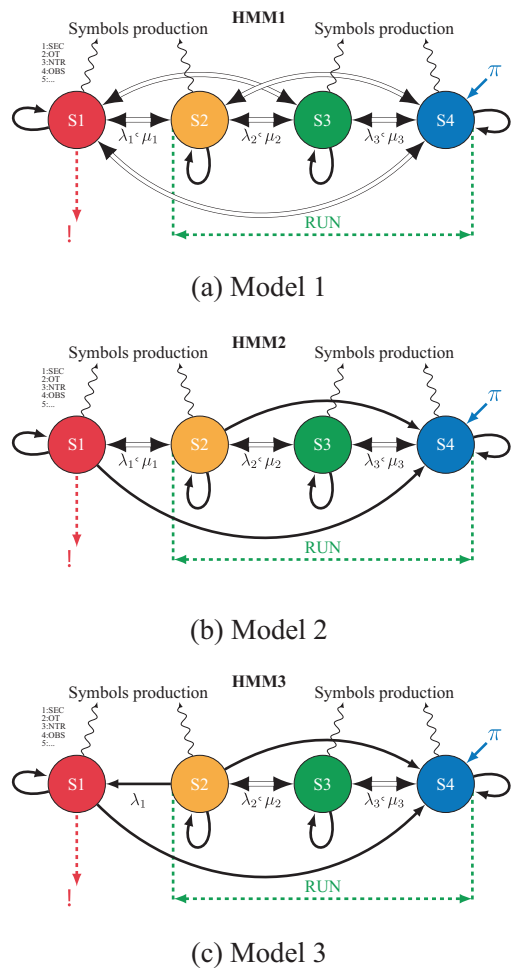


Figure 3. Hidden Markov Models.

About 1000 symbols were produced by reference model (distribution in figure 8 and 9). Each sequence ends with a stop of production (symbol SP in red) see fig. 2. We get 11 sequences in our 1000 simulated symbols. You can see distribution symbols/states for the first sequence: HMM 1, HMM 1/Baum-Welch and HMM 1/Segmental K-means algorithms, in figure 10. Finally, we obtain states sequences for each HMM outside. Later, these states are used to make comparisons between HMM, studied in section 3, see results in section 4.2. Diagram of this process is given in figure 4.

#### 4.2 Results

Without a priori knowledge we can give the most relevant model in the way of Shannon. Namely,

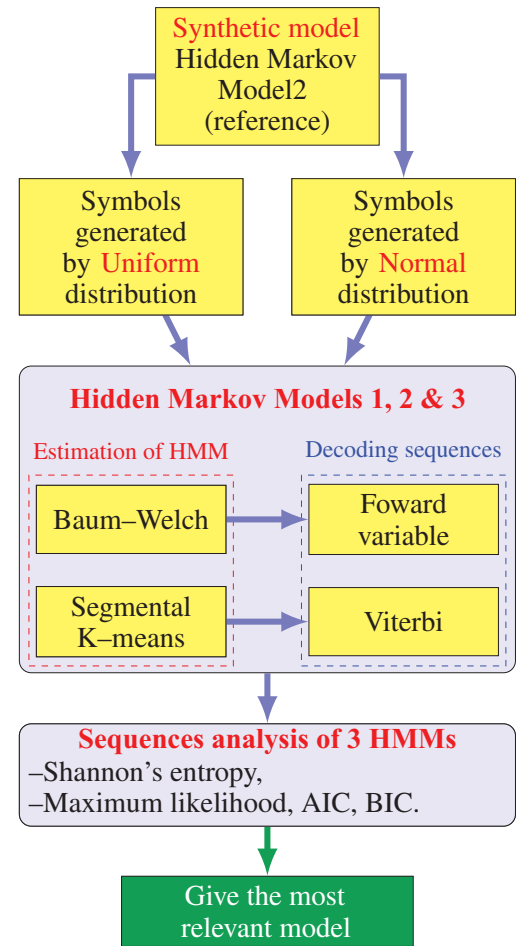


Figure 4. Matching model method, using synthetic model.

we verify that the best model (which provides the better estimation of degradation level (Vrignat et al. 2010)) obtains a good “entropic” score through entropic filter, illustrated in figure 5. The best model is model 2 with Baum–Welch learning,

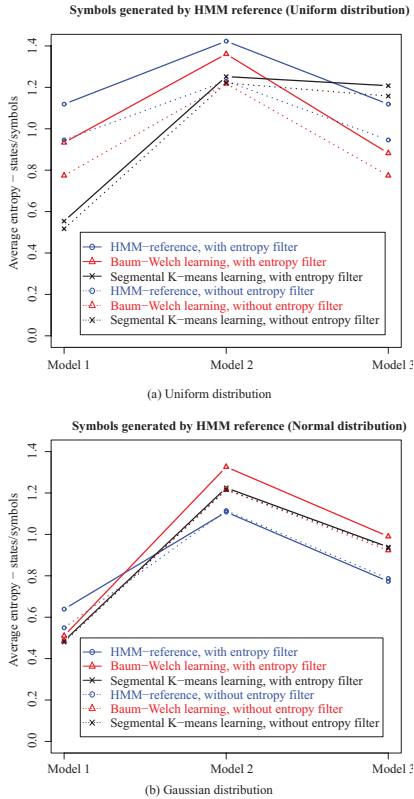


Figure 5. Average entropy of models.

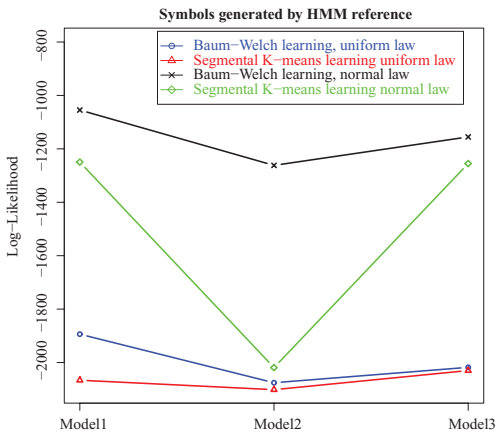


Figure 6. Log-likelihood of 2 learning algorithms.

where entropy is maximum. It also highlight the best learning algorithm recommended in (Vrignat et al. 2010): **Baum–Welch with Forward variable decoding**, whatever distribution of symbols (uniform or normal).

We evaluate likelihood (or probability) of observations sequences given by synthetic HMM. Results of maximum likelihood and *BIC* are

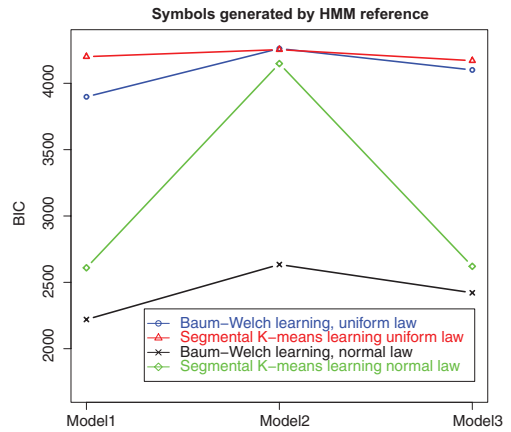


Figure 7. BIC with 2 learning algorithms.

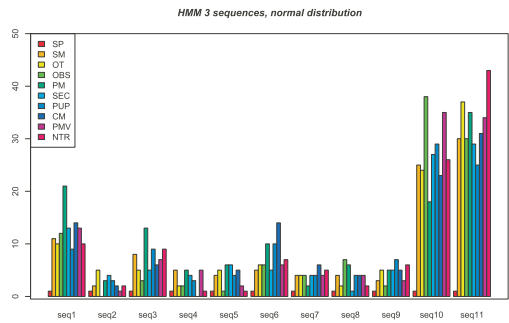


Figure 8. HMM sequences example, uniform law.

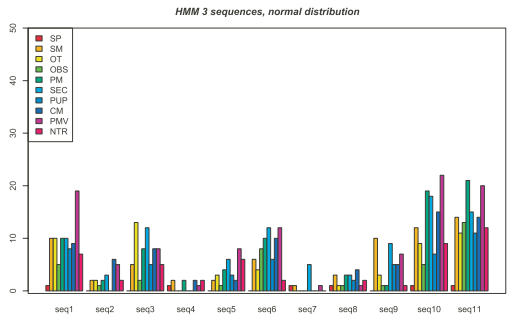


Figure 9. HMM sequences, Gaussian law.

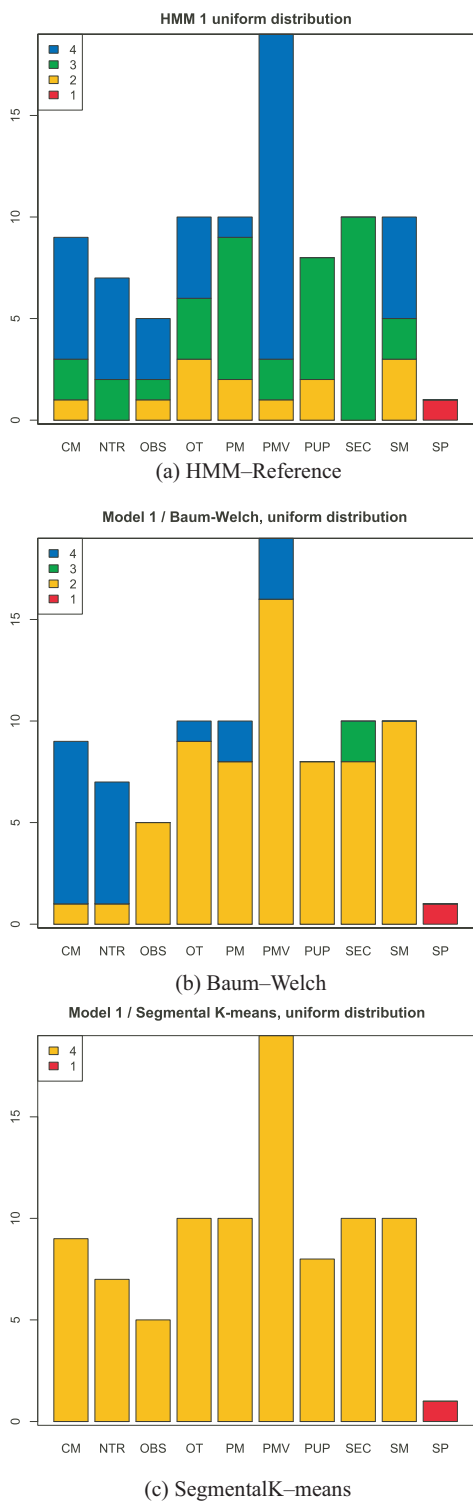


Figure 10. First sequence, using normal distribution.

presented respectively in figure 6 and 7. Our results highlight the most relevant model: **HMM 2**, fig. 3(b). That corroborate (Vrignat et al. 2010) results. On the other side, our results don't show clearly, differences between algorithms, we can not conclude for the best learning and decoding algorithm. Nevertheless, with Segmental K-means algorithm, in figure 10(c), the reader can see a bad distribution of symbols. *AIC* does not penalize our 1000 data, that's why *BIC* is more suitable, because of " $k \ln(n)$ " term of equation 13.

## 5 CONCLUSIONS

In our study, we presented a way for evaluating relevance on Hidden Markov Models based on three different criteria. We have successfully applied this method to three different models. The first one, uses Shannon's entropy and entropic filter. Given set of observations sequences simulated by our synthetic model, we verify that the most relevant model obtains a good "entropic" score. That corroborates (Vrignat et al. 2010) results which show that model 2 is the one which comes closest to real industrial process. This criterion also shows that Baum-Welch learning algorithm with Forward variable decoding gives best results. Second and third criterion (Maximum likelihood and *BIC*) emphasize that HMM 2 is the best model, whatever distribution of symbols. Unfortunately, these criteria are too near each other to make conclusions about learning algorithm.

Without a priori knowledge, we illustrated that topology model 2 (fig. 3(b)) with Baum-Welch learning algorithm and Forward variable decoding is the best one. We also show that model 2 is a good model with Log-likelihood and *AIC* criterion. Unfortunately, these methods can't give us the best learning algorithm.

In further work, we will try some statistics tests like Bartlett or Aspin-welch. We also try the Kolmogorov-Smirnov fit test of distribution of two samples. Our research, is to be able to validate a reel choice of a model: topology, symbol,...without a priori knowledge on results.

## REFERENCES

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. 2nd inter. symp. on information theory. *2nd Inter. Symp. on Information Theory*, 267–281.
- Arminjon, M. & D. Imbault (2000). Maximum entropy principle and texture formation. *Zeitschrift für angewandte Mathematik und Mechanik*, 80, Suppl. N°1, 13–16.
- Ash, R. (1990). Information theory. *Dover Publications*.
- Avila, M. (1996). *Optimisation de modèles Markoviens pour la reconnaissance de l'écrit*. Ph. D. thesis, Université de Rouen.

- Baum, L.E., T. Petrie, G. Soules, & N. Weiss (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics* 41(1), pp. 164–171.
- Beirlant, J., E.J. Dudewicz, L. Györfi, & E.C. Meulen (1997). Nonparametric entropy estimation: An overview. *International Journal of the Mathematical Statistics Sciences* 6, 17–39.
- Chandrasekaran, V., J.K. Johnson, & A.S. Willsky (2007). Maximum entropy relaxation for graphical model selection given inconsistent statistics. *Laboratory for Information and Decision Systems, Massachusetts Institute of Technology Cambridge, MA 02139*.
- Chen, S.S. & P.S. Gopalakrishnan (1998, February). Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, Virginia, USA.
- Cover, T.M. & J.A. Thomas (1991). *Elements of information theory*. New York, NY, USA: Wiley-Interscience.
- Fox, M., M. Ghallab, G. Infantes, & D. Long (2006). Robot introspection through learned hidden markov models. *Artif. In-tell.* 170(2), 59–113.
- Jaynes, E. (1957). Information theory and statistical mechanics. *Physical Review*, vol. 16, no. 4, 620–630.
- Juang, B.H. & L.R. Rabiner (1990, September). The segmental k-means algorithm for estimating parameters of hidden markov models. *Acoustics, Speech and Signal Processing, IEEE Transactions on* 38(9).
- Lebarbier, E. & T. Mary-Huard (2004). Le critère bic: fondements théoriques et interprétation. Research Report RR-5315, INRIA.
- Olivier, C., F. Jouzel, A. El Matouat, & P. Courtellemont (1996). Prediction with vague prior knowledge. *Communications in Statistics 25- Theory and Methods*, 601–608.
- Quinlan, J.R. (1979). Discovering rules by induction from large collections of examples. In D. Michie (Eds.), *Expert Systems in the Micro-Electronic Age*. Edinburgh: Edinburgh University Press., 168–201.
- Quinlan, J.R. (1993, January). C4.5: Programs for Machine Learning (*Morgan Kaufmann Series in Machine Learning*) (1 ed.). Morgan Kaufmann.
- Rabiner, L.R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceeding of the IEEE, 77(2) SIAM interdisciplinary journal*, 257–286.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* 6, 461–464.
- Shannon, C.E. (1948). A mathematical theory of communication. *Bell system technical journal* 27.
- Viterbi, A. (1967, April). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory* 13(2), 260–269.
- Vrignat, P. (2010). *Génération d'indicateurs de maintenance par une approche semiparamétrique et par une approche markovienne*. Ph. D. thesis, Université d'Orléans.
- Vrignat, P., M. Avila, F. Duculty, & F. Kratz (2010). Use of hmm for evaluation of maintenance activities. *IJAIS, International Journal of Adaptive and Innovative Systems, Vol. 1, Nos. 3/4*, 216–232.
- Zille, V., C. Bérenguer, A. Grall, A. Despujols, & J. Lonchamp (2007). Modelling and performance assessment of complex maintenance programs for multi-component systems. *ES-REDA 32nd Seminar Proceedings, Alghero, Espagne*.