Mesures de pertinence par les critères du " maximum de vraisemblance " et de " BIC 1 " appliqués à l'évaluation des paramètres stochastiques de modèles de Markov cachés

Bernard Robles, Manuel Avila, Florent Duculty, Pascal Vrignat, Frédéric Kratz

Mesures de pertinence par les critères du « maximum de vraisemblance » et de « BIC [1] » appliqués à l'évaluation des paramètres stochastiques de modèles de Markov cachés

Bernard Roblès *, Manuel Avila *, Florent Duculty *, Pascal Vrignat *, Frédéric Kratz **

** Laboratoire PRISME, Equipe MCDS, IUT de l'Indre – Dpt GEII, 2 av. F. Mitterrand, 36000 Châteauroux*
*** Laboratoire PRISME, Equipe MCDS, Ecole Nationale Supérieure d'Ingénieurs, 88 boulevard Lahitolle, 18020 Bourges cedex*

Section de rattachement : 61
Secteur : secondaire

RÉSUMÉ. Les modèles de Markov cachés ou HMM [2] sont largement utilisés dans les domaines de la reconnaissance des formes, de la parole ainsi que dans la modélisation de processus complexes. Nous proposons dans cette étude d'évaluer la pertinence des paramètres de modèles de Markov cachés de façon objective sans connaissance à priori. Dans un premier temps, nous faisons un état de l'art sur les critères de sélection de modèles les plus utilisés dans la littérature. Nous présentons ensuite deux critères permettant d'évaluer la pertinence d'événements stochastiques issus de modèles de Markov cachés. Nous étayons notre étude en nous appuyant sur l'exemple concret d'un processus industriel. Nous évaluons alors les paramètres de sortie des différents modèles testés sur ce processus, pour finalement s'orienter vers le modèle le plus pertinent.

MOTS-CLÉS : modèles de Markov cachés, sélection de modèles, caractérisation, pertinence, maximum de vraisemblance, BIC.

1. Bayesian Information Criterion
2. Hidden Markov Models

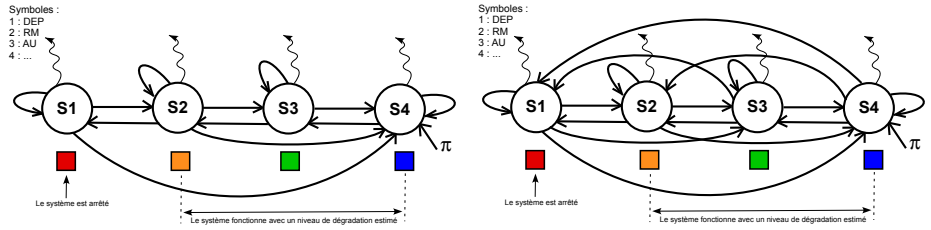
1 Introduction

Selon Lebarbier [6], la problématique de la sélection d'un modèle est basée sur la minimisation d'un critère de pénalité. Les premiers critères qui apparaissent dans la littérature sont l'*AIC* : *l'Akaike Information Criterion* [1], le *BIC* : *Bayesian Information Criterion* [12], le *MDL* : *Minimum Description Length* [10] et le *Cp* de Mallows [7]. De nombreux travaux théoriques ont été réalisés sur leurs propriétés statistiques afin de les adapter à des modèles spécifiques. Comme par exemple, les versions corrigées du critère *AIC* d'Hurvish [5] avec l'AIC_c, Sugira [13] avec le $c-AIC$ pour les petites tailles d'échantillon par rapport au nombre de paramètres à estimer.

L'objet de notre étude est de mesurer la pertinence de modèles de Markov cachés. Nous étayons notre étude par un exemple concret d'un processus industriel. Dans la section 2, nous introduisons les problèmes et les notations utilisées. Dans la section 3, nous décrivons notre méthode pour estimer la pertinence des paramètres de modèles de Markov cachés. Enfin en section 4, nous faisons l'étude d'un cas concret en donnant le modèle le plus pertinent avant de conclure et de proposer des perspectives sur la suite des travaux à mener.

2 Modèle de Markov Caché ou HMM

Un modèle de Markov caché (Rabiner [9], Fox [3]) est un automate à états cachés qui est constitué d'une variable non observable. Celle-ci peut représenter l'état du système à modéliser. Seule la variable de sortie est observable. Cela nous permet d'avoir une séquence d'observations en sortie de l'automate ; à partir de maintenant, nous parlerons de « symboles » pour représenter ces observations (voir figure 1).



(a) Modèle 1 (b) Modèle 2

Fig 1 – Modèles à topologie orientée à quatre états

2.1 Définitions formelles d'un modèle de Markov caché à observations discrètes

- Soit N, le nombre d'états cachés possibles et $S = \{S_1, S_2, \ldots, S_N\}$, l'ensemble des valeurs possibles de la variable S. On notera q_t, la valeur de cette variable à l'instant t ;
- Le processus ainsi modélisé, doit répondre à l'hypothèse Markovienne d'ordre 1 : l'état à un instant t ne dépend que de l'état à l'instant $t-1$;

– Soit T, le nombre total de symboles d'observations et nous notons $O = \{o_1, o_2, \ldots, o_T\}$, la séquence d'observations du processus modélisé ;
– Soit $A = \{a_{ij}\}$, la distribution de probabilité de la transition d'état avec :

$$a_{ij} = P(q_{t+1} = S_j | q_t = S_i) \qquad 1 \leq i, j \leq N. \tag{1}$$

– Soit $B = \{b_j(k)\}$, la distribution de probabilité des observations à l'état j, avec :

$$b_j(k) = P(O_t = o_k | q_t = S_j) \qquad 1 \leq j \leq N \qquad 1 \leq j \leq T, \tag{2}$$

avec O_t, la valeur de la variable d'observation à l'instant t.
– Soit $\pi = \{\pi_i\}$, la distribution des probabilités initiales, avec :

$$\pi = P(q_1 = S_i) \qquad 1 \leq i \leq N. \tag{3}$$

– Le modèle de Markov caché sera noté $\lambda = (A, B, \pi)$.

2.2 Hypothèse de Markov

La prédiction de l'état futur n'est pas rendue plus précise par connaissance supplémentaire d'information à priori i.e. toute l'information utile pour la prédiction de l'état suivant est contenue dans l'état présent du processus.

$$P(X_{n+1} = j | X_n = i) = P(X_1 = j | X_0 = i). \tag{4}$$

3 Méthodes de mesures de la pertinence de modèles

3.1 Maximum de vraisemblance

Pour un modèle statistique P_μ donné, et étant donné la séquence d'observations X, la probabilité de son apparition suivant P peut être mesurée par $f(X, \mu)$ qui représente la densité de X où μ apparaît. Puisque μ est inconnue, il semble alors naturel de favoriser les valeurs de μ pour lesquelles $f(X, \mu)$ est élevée : c'est la notion de la *vraisemblance* de μ pour l'observation X.

Définition 3.1.1 *Expression de la vraisemblance V :*

$$V(x_1, \ldots, x_n; \mu) = \prod_{i=1}^{n} f(x_i; \mu) \text{ ,avec } \mu \text{ l'espérance mathématique}. \tag{5}$$

Définition 3.1.2 *Expression du maximum de vraisemblance :*

$$\widehat{V}(x_1, \ldots, x_n; \mu). \tag{6}$$

Maximum de vraisemblance pour un échantillon discret $P_\mu(x_i)$ qui représente la probabilité discrète où μ apparaît :

$$\log(\widehat{V}(x_1, \ldots, x_n; \mu)) = \sum_{i=1}^{n} \log(P_\mu(x_i)). \tag{7}$$

En pratique, on **maximise le logarithme de la fonction de vraisemblance** pour comparer plusieurs modèles.

3.2 Critère de BIC

Définition 3.2.1 *BIC (Bayesian Information Criterion) :*

$$BIC = -2\ln V + k\ln n, \tag{8}$$

où k est le nombre de paramètres libres du modèle de Markov, n est le nombre de données, $k\ln n$ est le terme de pénalité.

Le modèle à retenir est celui qui montre **le*BIC* le plus faible**. Le *BIC* utilise le principe du maximum de vraisemblance (équation 7). Il pénalise les modèles comportant trop de variables, et évite le sur-apprentissage [12].

3.3 Commentaires sur les méthodes de mesures

Il faut d'abord faire attention au choix du critère de sélection du modèle. Il doit être conditionné par l'objectif de l'analyse et la connaissance des données. Les données d'apprentissage sont déterminantes dans la construction du modèle. Il n'existe pas de critère universellement meilleur. En pratique, seule une parfaite connaissance du milieu à analyser permet de donner un sens à la notion de supériorité d'un critère sur un autre. D'après [8], le principe du maximum de vraisemblance conduit en général à sur-paramétrer le modèle pour avoir de bons résultats. Une pénalisation du terme de vraisemblance peut pallier cet inconvénient. Le critère de type « log-vraisemblance » pénalisé le plus célèbre est l'*AIC* [1], même s'il n'est pas totalement satisfaisant. Il améliore le principe du maximum de vraisemblance mais conduit aussi à une sur-paramétrisation. D'autres critères désormais classiques, *BIC* et *HQC* (Hannan-Quinn information Criterion) [4], assurent une meilleure estimation en pénalisant justement le sur-dimensionnement du modèle.

4 Expérimentations

4.1 Présentation du cadre de l'étude

Cette étude porte sur des données issues d'un processus continu de production de pain de mie. Tous les processus de l'usine sont liés entre eux de manière séquentielle. L'arrêt d'un processus engendre l'arrêt des éléments en amont. Une maintenance préventive est donc indispensable. Pour ce faire, les agents de maintenance doivent consigner leurs actions ou observations dans une base de données centralisée. Les données présentées, dans cet article, sont issues d'une peseuse volumétrique sur l'une des lignes de production. Nous utilisons ensuite ces séquences de symboles (voir tableau 1) pour modéliser le niveau de dégradation du processus. Nous modélisons cette « signature » à l'aide de modèles de Markov cachés.

VEP	VEP	TEP	TEP	SEC	TEP	TEP	DEP	AU	jour/heure/…

Tableau 1 – Séquence d'un message issue des données de maintenance.

4.2 Modélisation du corpus d'apprentissage

Dans le cadre d'activités de maintenance, Vrignat et al. [14] modélise ces dysfonctionnements à l'aide de modèles de Markov. Nous rappelons dans le tableau 2, la signification des symboles choisis issus des observations. Ces symboles définissent les actions de maintenance menées sur le processus. Par exemple, le symbole DEP correspond à un dépannage avec arrêt de la production. C'est un état critique qu'il faut minimiser. L'étude de [14] considère deux modèles (voir figure 1(a) et 1(b)) différents avec deux corpus d'apprentissage (algorithme Baum–Welch et décodage variable Forward), l'un sur l'année 2005 et l'autre sur les deux années 2005–2006. Les symboles RAS sont insérés pour avoir un échantillonnage à la journée ou toutes les 6 heures. Pour la suite, nous adopterons les dénominations suivantes :

– 05M1 | 1j : Corpus d'apprentissage 2005 modèle 1 / 1 donnée par jour
– 0506M1 | 1j : Corpus 2005-2006 modèle 1 / 1 donnée par jour
– 05M2 | 1j : Corpus 2005 modèle 2 / 1 donnée par jour
– 0506M2 | 1j : Corpus 2005–2006 modèle 2 / 1 donnée par jour
– 05M1 | 6h : Corpus 2005 modèle 1 / 1 donnée toutes les 6 heures
– 0506M1 | 6h : Corpus 2005–2006 modèle 1 / 1 donnée toutes les 6 heures
– 05M2 | 6h : Corpus 2005 modèle 2 / 1 donnée toutes les 6 heures
– 0506M2 | 6h : Corpus 2005–2006 modèle 2 / 1 donnée toutes les 6 heures

Etat du processus	
MARCHE	
ARRET	

	Nature des interventions
1	DEP (Dépannage / arrêt de la production)
2	RM (Réglage Machine)
3	AU (Autre)
4	OBS (Observation)
5	TEP (Travaux Entretien Préventif pas d'arrêt de prod)
6	SEC (Sécurité)
7	RAN (Remise A Niveau / planifié)
8	NET (Nettoyage Machine)
9	VEP (Visite Entretien Préventif)
10	RAS (pas d'intervention)

Tableau 2 – Codification symbolique des interventions de maintenance (source [14]).

4.3 Description des modèles de Markov utilisés

Les modèles se présentent sous la forme d'un automate stochastique. Les états représentent les niveaux de dégradation du processus. Les symboles de l'automate représentent les observations du processus (voir figure 1). Le modèle donne la probabilité d'être dans l'un des quatre états (S1–S4) en fonction des symboles (notés « Cd » dans le tableau 3), selon les hypothèses de Markov pour les modèles d'ordre 1 (§ 2.2). Les probabilités des quatre états sont données par la variable Forward et les différents niveaux par l'algorithme de Viterbi [14].

Les séquences de symboles ainsi produites, nous donnent selon le modèle, une estimation du taux de dégradation du processus (voir figure 2). Nous pouvons alors quantifier ces

n°	DATE	Cd	Cd. Symb.	S1	S2	S3	S4	Niveau
1	09/01/2007	RAS	10	0,0%	0,0%	0,0%	100,0%	4
2	10/01/2007	RAS	10	0,0%	0,0%	56,1%	42,9%	3
3	11/01/2007	VEP	9	0,0%	66,3%	33,766,3%	0,0%	2
4	12/01/2007	TEP	5	0,0%	86,811,7%	11,7%	1,5%	2
⋮	⋮			⋮				⋮
8	16/01/2007	DEP	1	100,0%	0,0%	0,0%	0,0%	1

Tableau 3 – Séquence de symboles / niveaux de dégradation, 0506M1/1jour.

informations au moyen de différents critères et ainsi établir une évaluation des modèles, selon ces critères.

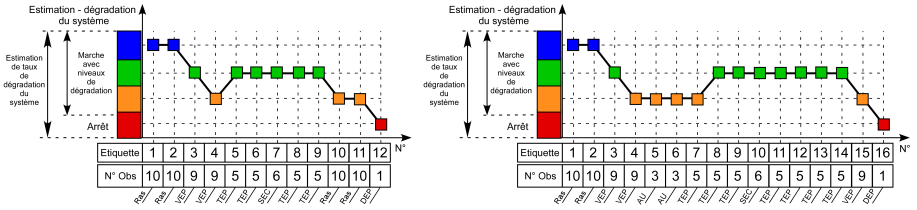


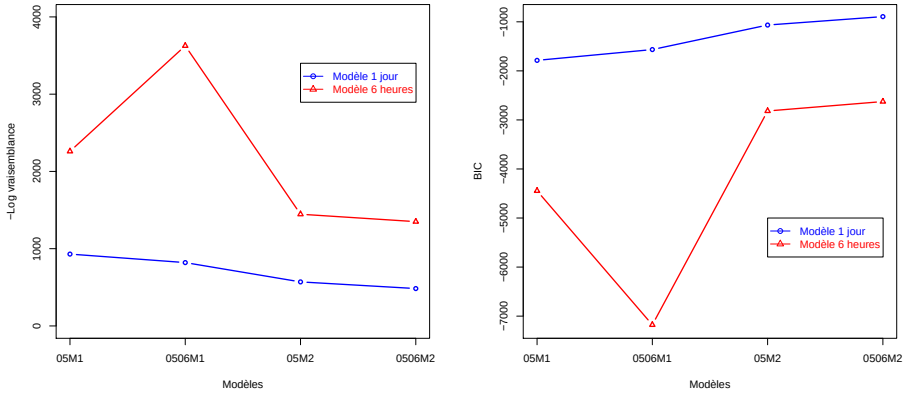
Fig 2 – Exemples de dégradations du processus, 0506M1/1jour (source [14]).

4.4 Résultats du maximum de vraisemblance

Nous utilisons ici une partie de la méthode de Bourguigon (2004) [2], qui permet de sélectionner des « modèles de Markov parcimonieux » en utilisant le principe du maximum de vraisemblance. Nous allons maximiser le logarithme de la fonction de vraisemblance i.e. pour notre cas discret, nous calculons le maximum de vraisemblance sur les probabilités de transition de chaque modèle. Nous observons un maximum pour le modèle 0506M1 | 6 heures, voir graphe 3(a). Les résultats pour les modèles de base de temps 1 jour ne nous donnent pas des résultats exploitables, la variance trop faible indique que la dispersion des résultats n'est pas satisfaisante pour tirer des conclusions pertinentes. Le principe du maximum de vraisemblance nous donne comme modèle le plus pertinent : 0506M1 | 6 heures. Les résultats concernant 0506M1 | 1 jour ne sont pas éloquents. Comme pour l'entropie [11], le principe du maximum de vraisemblance est intéressant pour la sélection de modèle dans la mesure où le nombre de données est suffisamment important.

4.5 Mesure du critère de BIC

Le critère *BIC* pénalise plus les modèles ayant un grand nombre de données. Le modèle le plus pertinent étant celui qui obtient la valeur minimale, voir figure 3(b). Nous voyons que ce critère nous amène aux mêmes conclusions que précédemment pour le modèle 0506M1 | 6 heures. C'est donc le plus pertinent au sens de *BIC*. Le critère d'*AIC* n'est pas utilisable dans notre cas car la pénalité ne prend pas en compte le nombre de données ($AIC = -2\ln V + 2k$) voir §3.3. Les résultats de [14] montrent que ce modèle fonctionne mais il est trop sensible et engendre des interventions de maintenance inutiles. Par contre il



(a) Log vraisemblance (b) Critère *BIC*

Fig 3 – Log vraisemblance et BIC

n'a pas été retenu car il est trop sensible dans la détection des pannes. En effet, le passage d'un niveau de dégradation à un autre se fait avec plusieurs rebonds.

5 Conclusions et perspectives

Dans notre étude, nous avons présenté une démarche visant à évaluer la pertinence de symboles dans des modèles de Markov cachés. Dans un précédent papier [11], nous avons appliqué l'entropie de Shannon, sur les symboles isolés, les bigrammes et trigrammes selon différents échantillonnages pour pouvoir comparer les différents découpages temporels. De la même façon, le principe du maximum de vraisemblance et l'indice de *BIC* présentés ici, ont été évalués en fonction des états des modèles afin de vérifier que le plus pertinent obtienne également un bon score de « vraisemblance » et un faible score « *BIC* ». Nous illustrons ainsi que sans connaissance à priori, le modèle 0506M1 est le plus pertinent. Nos calculs corroborent ainsi les résultats de [15] qui montrent que le modèle 0506M1 est celui qui appréhende le mieux la réalité. En ce qui concerne l'échantillonnage du modèle, Vrignat [15] montre qu'il est préférable d'utiliser **1 jour** au lieu de 6 heures. Le modèle 0506M1 | 6 heures fonctionne aussi mais est trop précis dans la détection des pannes. En effet, ce modèle est trop réactif par rapport aux contraintes fixées au départ qui sont d'avoir une certaine stabilité dans les changements d'états. Il engendre ainsi des rebonds lors du passage d'un état à un autre. La méthode par la mesure de l'entropie [11] réagit comme un filtre « anti-rebond » et donne le bon échantillonnage (1 jour). Par contre, les méthodes utilisant le principe de maximum de vraisemblance et *BIC* nous désignent le modèle le plus réactif (échantillonnage 6 heures).

Dans nos travaux futurs, nous allons tenter d'établir une classification des modèles de Markov étudiés, en utilisant la divergence de Kullback-Leibler. Notre objectif est de pouvoir valider de façon objective, un choix de modèle : topologie, ordre, symbole,... sans connaissance à priori sur les résultats.

Références

[1] AKAIKE, H. Information theory and an extension of the maximum likelihood principle. 2nd inter. symp. on information theory. *2nd Inter. Symp. on Information Theory* (1973), 267–281.

[2] BOURGUIGNON, P. Y., AND ROBELIN, D. Modèles de markov parcimonieux : sélection de modèle et estimation. *Statistique et Génome* (2004).

[3] FOX, M., GHALLAB, M., INFANTES, G., AND LONG, D. Robot introspection through learned hidden markov models. *Artif. Intell. 170*, 2 (2006), 59–113.

[4] HANNAN, E. J., AND QUINN, B. G. The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B (Methodological) 41*, 2 (1979), 190–195.

[5] HURVICH, C. M., AND TSAI, C. L. Regression and time series model selection in small samples. *Biometrika 76*, 2 (June 1989), 297–307.

[6] LEBARBIER, E., AND MARY-HUARD, T. Le critère bic : fondements théoriques et interprétation. Research Report RR-5315, INRIA, 2004.

[7] MALLOWS, J. Some comments on cp. *Echnometrics, 15* (1973), 661–675.

[8] OLIVIER, C., JOUZEL, F., EL MATOUAT, A., AND COURTELLEMONT, P. Un nouveau critère pour la sélection de l'ordre d'un modèle. *Seizième colloque Gretsi* (1997).

[9] RABINER, L. R. A tutorial on hidden markov models and selected applications in speech recognition. *Proceeding of the IEEE, 77(2) SIAM interdisciplinary journal* (1989), 257–286.

[10] RISSANEN, J. Modelling by the shortest data description. *Automatica 14* (1978), 465–471.

[11] ROBLÈS, B., AVILA, M., DUCULTY, F., VRIGNAT, P., AND KRATZ, F. Evaluation de la pertinence des paramètres stochastiques sur des modèles de markov cachés. *CNRIUT* (2010).

[12] SCHWARZ, G. Estimating the dimension of a model. *The Annals of Statistics 6* (1978), 461–464.

[13] SUGIURA, N. Further analysts of the data by akaike' s information criterion and the finite corrections – further analysts of the data by akaike' s. *Communications in Statistics - Theory and Methods 7*, 1 (1978), 13–26.

[14] VRIGNAT, P., AVILA, M., DUCULTY, F., AND KRATZ, F. Modélisation des dysfonctionnements d'un système dans le cadre d'activités de maintenance. *16ème Congrès de Maîtrise des Risques et de Sûreté de Fonctionnement, Avignon, Communication 4A-1* (2008).

[15] VRIGNAT, P., AVILA, M., DUCULTY, F., ROBLÈS, B., AND KRATZ, F. Utilisation des chaînes de markov cachées pour une évaluation des activités de maintenance dans le cadre d'un processus industriel pour l'agroalimentaire. *CNRIUT* (2009).