Un nouveau critère pour la sélection de l'ordre d'un modèle

Christian Olivier[(1)], Frédéric Jouzel[(1)], Abdelaziz El Matouat[(2)], Pierre Courtellemont[(1)]

[(1)]PSI-La3i, Faculté des Sciences, Université de Rouen,
76821 Mont Saint Aignan cedex, France

[(2)]Ecole Normale Supérieure de Fès, BP 34A, Fès, Maroc

RÉSUMÉ

H. AKAIKE (1973) a proposé le critère AIC (Akaike Information Criterion) pour l'estimation de l'ordre k d'un modèle statistique paramétré, incluant ce terme k dans une pénalisation de la vraisemblance. Il améliore ainsi le principe du maximum de vraisemblance. Néanmoins, la sélection par ce critère conduit asymptotiquement à une surestimation stricte de l'ordre. C'est pourquoi nous proposons l'emploi d'un nouveau critère convergent pour deux modèles paramétrés : un modèle autorégressif (AR) et un modèle de Markov. De façon générale, les critères utilisés sont composés d'un terme de log-vraisemblance mesurant l'adéquation du modèle aux observations et d'un terme de pénalisation dépendant de la taille de l'échantillon et du nombre de paramètres libres du modèle. Les performances du nouveau critère sont étudiées sur des simulations, et comparées aux critères traditionnels AIC, BIC et φ.

ABSTRACT

H. AKAIKE (1973) suggested the AIC criterion for the estimation of the order k of a parametrized statistical model, including the term k as penalization of likelihood function. He improves the maximum likelihood principle. Nevertheless, selection according to this criterion leads asymptotically to a strict overestimation of the order. This is why we suggest the use of another consistent criterion for two parametrized models: an autoregressive (AR) model and a Markov model. The often used criteria are made of two components: a log-likelihood term which measures model fitting and a penalization term depending on the sample size and the free parameters number. The performance of the new criterion is analysed on simulations and compared with the traditionnal criteria AIC, BIC and φ.

1 Introduction

Considérons un modèle paramétré caractérisé par la distribution de probabilité (DDP) $f(\cdot \mid \theta^*)$ dans lequel le paramètre θ^* et sa dimension k^*, appelée ordre du modèle, ne sont pas connus. Si les techniques classiques comme le maximum de vraisemblance permettent d'estimer le paramètre θ^* lorsque l'ordre est connu, le problème de l'estimation de l'ordre reste plus difficile. En effet, le principe du maximum de vraisemblance (MV) conduit en général à surparamétrer (surdimensionner) le modèle.

Une pénalisation du terme de vraisemblance peut pallier cet inconvénient. Un des critères de type log-vraisemblance pénalisée le plus célèbre est AIC [1], même s'il n'est pas totalement satisfaisant : il améliore le principe MV mais conduit aussi à une surparamétrisation stricte de l'ordre. D'autres critères désormais classiques, BIC [10, 7] et φ [5], assurent une estimation convergente de l'ordre.

2 Position du problème

2.1 Critères usuels

Soit $X^N = X_1, ..., X_N$ une série d'observations indépendantes d'un processus aléatoire de loi g inconnue, et soit $f\left(\cdot \mid \theta_k\right)$ une DDP spécifiée par le vecteur des paramètres $\theta_k, \theta_k \in \mathbb{R}^k$, où l'indice k représente le nombre de paramètres libres du modèle considéré (nombre de degrés de liberté). On cherche à approcher $g(\cdot)$ par $f\left(\cdot \mid \theta_k\right)$. Soit $\widehat{\theta}_k$ l'estimateur au sens MV de θ_k, c'est à dire $\widehat{\theta}_k = \arg\max_{\theta_k} f\left(X^N \mid \theta_k\right)$.

Si l'on pose $L(k) = \log f(X^N \mid \widehat{\theta}_k) = \sum_{i=1}^{N} \log f(X_i \mid \widehat{\theta}_k)$, alors les différents critères de sélection de modèle peuvent s'écrire de la façon suivante (IC pour *Information Criterion*) :

$$IC(k) = -2L(k) + c_N k \qquad (1)$$

où $L(\cdot)$ désigne la log-vraisemblance et $c_N k$ le terme de pénalité. La valeur k minimisant (1) est prise comme estimateur de l'ordre du modèle. On peut remarquer que pour deux modèles possédant un nombre identique de degrés de liberté, IC est simplement l'estimateur MV. Les critères usuels se distinguent par la forme de la pénalité. A partir d'une minimisation de l'information de KULLBACK-LEIBLER entre $g(\cdot)$ et $f\left(\cdot \mid \theta_k\right)$, H. AKAIKE obtient $c_N = 2$ [1]. Notons que certains auteurs, comme par exemple R.J. BHANSALI [2], préconisent l'utilisation de $c_N = \alpha$, $\alpha > 2$ de façon à augmenter la pénalité (critère AIC_α). Ces critères sont connus pour ne pas converger. Pour pallier cet inconvénient, G. SCHWARZ propose le critère BIC, pour une famille de distribution exponentielle, par une approche bayésienne [10]. J. RISSANEN propose un critère équivalent par la théorie du codage [7], ce qui correspond au critère MDL (Minimum Description Length). Pour ce critère, $c_N = \log N$. Remarquons que ce dernier critère pénalise davantage le terme de log-vraisemblance quand N augmente. Un troisième critère est introduit par E.J. HAN-

NAN et B.G. QUINN [5] dans le cas du modèle AR. Pour ce critère, $c_N = \log\log N$. Ce critère apparait comme un compromis entre AIC et BIC. Il sera noté φ.

2.2 Comportement asymptotique

Les critères introduits, hormis AIC (et AIC_α), satisfont à l'une des conditions asymptotiques de convergence données par R. NISHII [6] qui spécifie de façon générale la forme des pénalités admissibles ainsi que la nature de la convergence correspondante :

- si $\lim_{N\to\infty} \frac{c_N}{N} = 0$ et $\lim_{N\to\infty} \frac{c_N}{\log\log N} = +\infty$ alors $\lim_{N\to\infty} \widehat{k} \overset{p.s.}{=} k^*$ (convergence presque sûre) ;
- si $\lim_{N\to\infty} \frac{c_N}{N} = 0$ et $\lim_{N\to\infty} c_N = +\infty$ alors $\lim_{N\to\infty} P\left(\widehat{k} = k^*\right) = 1$ (convergence en probabilité).

Ainsi, pour BIC, il s'agira d'une convergence presque sûre et pour φ d'une convergence en probabilité.

2.3 Généralisation du critère de HANNAN et QUINN

Nous proposons une généralisation du critère φ, obtenue à partir de la *complexité stochastique* introduite par J. RISSANEN [9]. Ce critère sera noté φ_β. Le développement théorique est exposé dans [4].

Pour une séquence d'observations X^N, on considère la *complexité stochastique* [9] :

$$C_k\left(X^N\right) = -2\sum_{i=2}^{N} \log_2 f\left(X_i \mid \widehat{\theta}_k^{(i-1)}\right) + \log^* k + \log_2 c$$

où $\log^* k = \log_2 k + \log_2\log_2 k + \log_2\log_2\log_2 k + \ldots$, $c = \sum_{k=1}^{+\infty} 2^{-\log^* k} \cong 2.865$ (cf. [8]) et $\widehat{\theta}_k^{(i-1)}$ est l'estimateur au sens MV construit à partir de X^{i-1}. Cette complexité représente la longueur du code (nombre de bits) associée à $f\left(X^N, \theta\right) = \prod_{i=1}^{N} f\left(X_i, \theta\right)$ quand la dimension du paramètre est inconnue.

En effet, la longueur nécessaire pour coder $f\left(X^N \mid \theta\right)$ est $-2\sum_{i=2}^{N} \log_2 f\left(X_i \mid \widehat{\theta}_k^{(i-1)}\right)$. Pour coder l'entier inconnu k, on choisit la DDP $P(k) = 2^{-\log^* k}/c$ sur $\mathbb{N}^*$ de telle façon que $-2\log_2 P(k) = \log^* k + \log_2 c$ corresponde à la longueur du code associée à cette représentation. Il s'agit du codage universel introduit dans [8].

La sélection de l'ordre du modèle est alors fondée sur la minimisation de $E\left[C_k\left(X^N\right)\right]$ (longueur moyenne du code). La proposition suivante est introduite dans [4] :

$$\forall \beta \in]0,1[\,,\ H\left(\theta_k\right) + (1-\varepsilon)\frac{k\log\log N}{2N} \leqslant \frac{1}{N} E\left[C_k\left(X^N\right)\right] \leqslant H\left(\theta_k\right) + (1+\varepsilon)\frac{k\log\log N}{2N^{1-\beta}}$$

où $H\left(\theta_k\right)$ est l'entropie associée à $f\left(\cdot \mid \theta_k\right)$. En considérant ensuite un majorant de $\frac{1}{N}E\left[C_k\left(X^N\right)\right]$, nous obtenons le critère suivant, noté φ_β :

$$\varphi_\beta(k) = -2L(k) + kN^\beta \log\log N,\ \beta \in]0,1[\qquad (2)$$

Comme $\varphi_\beta(k) - \varphi(k)$ tend vers zéro en probabilité quand β tend vers zéro, le critère φ apparait comme un cas limite de φ_β ($\beta = 0$). Les conditions de convergence presque sûre de φ_β sont vérifiées (cf. partie 2.2).

3 Comparaison expérimentale

Dans la suite, nous donnons deux applications des critères explicités plus haut. Le premier exemple concerne le modèle AR (partie 3.1), et le second le modèle de Markov (partie 3.2). Nous spécifions tout d'abord l'expression des critères dans chacun des cas puis nous proposons une simulation qui permet de préciser le comportement du critère φ_β. Enfin, nous discutons des qualités de φ_β par rapport aux critères traditionnels.

3.1 Modèle AR

De nombreux résultats existent en ce qui concerne les comportements relatifs des critères traditionnels AIC, BIC et φ dans le cadre du modèle AR [3, 12]. Nous proposons d'intégrer φ_β dans ces diverses études comparatives. Soit une séquence d'observations $X^N = X_1, ..., X_N$ que l'on désire modéliser par un modèle AR d'ordre k :

$$\begin{cases} X_t = -\sum_{i=1}^{k} a_i X_{t-i} + e_t \\ E(e_t) = 0,\ E(e_s e_t) = \sigma_e^2 \delta_{st} \end{cases}$$

où δ_{st} est le symbole de KRONECKER et $e^N = e_1, ..., e_N$ est un bruit blanc gaussien de variance σ_e^2.

Avec les notations de la partie 2, $\theta_k = \left(a_1 \ldots a_k\right)^T$. Le terme de vraisemblance s'écrit alors :

$$\begin{aligned} f\left(X^N \mid \widehat{\theta}_k, \widehat{\sigma}_e^2\right) &= f\left(e^N \mid \widehat{\theta}_k, \widehat{\sigma}_e^2\right) \\ &= \frac{1}{\left(2\pi\widehat{\sigma}_e^2\right)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\widehat{\sigma}_e^2}\sum_{i=1}^{N} e_i^2\right] \end{aligned}$$

En ne tenant pas compte des termes qui ne dépendent pas de k, l'expression (1) devient, dans le cas du modèle AR :

$$IC(k) = N\log\widehat{\sigma}_e^2 + c_N k$$

avec $\widehat{\sigma}_e$ l'estimateur au sens MV de σ_e. Les deux modèles choisis pour la synthèse des signaux sont les suivants :

- un modèle AR d'ordre 2 : $a_1 = -0.55$ et $a_2 = -0.05$;
- un modèle AR d'ordre 15 : $a_1 = -0.50$, $a_2 = -0.06$, $a_{15} = -0.45$ et $a_i = 0$ pour tous les autres coefficients [1].

Les tableaux 1 et 2 présentent les résultats d'estimation obtenus pour ces deux signaux synthétiques, pour différentes valeurs de N. Les ordres testés vont de 0 à 20. La variance du bruit est choisie égale à 1 [2]. L'expérience est répétée 100 fois. Les taux indiqués représentent ainsi la fréquence des ordres obtenus (en pourcentage).

[1] Les deux modèles AR considérés ici sont fréquemment utilisés dans la littérature, par exemple dans [2, 12].

[2] On constate expérimentalement qu'une augmentation de la variance du bruit altère peu le comportement des critères.

TAB. 1 — Résultats sur le modèle AR(15), $\sigma_e = 1$

$N = 1000$

ordre	AIC	BIC	φ	$\varphi_{0.2}$	$\varphi_{0.5}$
0	0	0	0	0	0
0	0	0	0	0	100
...	...	...	...	...	...
15	**50**	**82**	**49**	**84**	**0**
16	9	5	9	3	0
17	9	7	7	7	0
18	12	3	14	4	0
$\geqslant$19	20	3	21	2	0

$N = 10000$

ordre	AIC	BIC	φ	$\varphi_{0.2}$	$\varphi_{0.5}$
0	0	0	0	0	0
...	...	...	...	...	...
15	**65**	**100**	**74**	**100**	**100**
16	13	0	11	0	0
17	7	0	6	0	0
18	6	0	5	0	0
$\geqslant$19	9	0	4	0	0

TAB. 2 — Résultats sur le modèle AR(2), $\sigma_e = 1$

$N = 1000$

ordre	AIC	BIC	φ	$\varphi_{0.2}$	$\varphi_{0.5}$
0	0	0	0	0	0
1	47	90	41	91	100
2	**35**	**10**	**34**	**9**	**0**
3	8	0	10	0	0
4	5	0	5	0	0
$\geqslant$5	5	0	10	0	0

$N = 10000$

ordre	AIC	BIC	φ	$\varphi_{0.2}$	$\varphi_{0.5}$
0	0	0	0	0	0
1	0	1	0	9	100
2	**74**	**99**	**80**	**91**	**0**
3	7	0	7	0	0
4	2	0	2	0	0
$\geqslant$5	17	0	11	0	0

$N = 100000$

ordre	AIC	BIC	φ	$\varphi_{0.2}$	$\varphi_{0.5}$
0	0	0	0	0	0
1	0	0	0	0	100
2	**74**	**100**	**83**	**100**	**0**
3	10	0	8	0	0
4	7	0	6	0	0
$\geqslant$5	9	0	3	0	0

La figure ci-dessous présente la valeur moyennée sur 100 expériences des différents critères pour le modèle AR(15), avec $\sigma_e = 1$ et pour $N = 1000$.

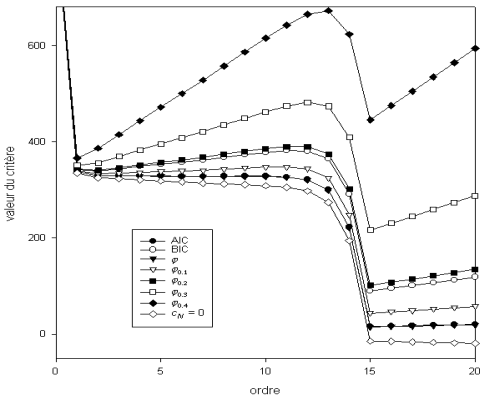

FIG. 1 — valeur moyennée des critères

3.2 Modèle de Markov

Soit une séquence d'observations $X^N = X_1, ..., X_N$ modélisée par une chaîne de Markov à temps discret, ergodique, et à valeur dans l'espace d'états E à M éléments. L'ordre optimal du modèle sera par définition le plus petit entier k vérifiant :

$$\forall t,\ P\left(X_t \mid X^{t-1}\right) = P\left(X_t \mid X_{t-k}, \ldots, X_{t-1}\right)$$

Ces probabilités conditionnelles constituent l'ensemble des paramètres θ_k, en négligeant les probabilités initiales. Sous cette hypothèse, le terme de vraisemblance s'écrit [11] :

$$P(X^N \mid \theta_k) = \prod_{i_1, i_2, \ldots, i_k, i} p_{i_1 \ldots i_k i}^{n_{i_1 \ldots i_k i}}$$

où $p_{i_1 \ldots i_k i} = P\left(X_t = i \mid X_{t-1} = i_k, \ldots, X_{t-k} = i_1\right)$ désigne la probabilité de transition d'ordre k et $n_{i_1 \ldots i_k i}$ est le nombre d'occurrences des $k+1$ états successifs $i_1, \ldots, i_k, i$ de E dans la chaîne de longueur N.

Le nombre de paramètres est ici M^{k+1} (probabilités de transition), mais étant données les M^k contraintes :

$$\forall \left(i_1, i_2, \ldots, i_k\right) \in E^k, \sum_{i \in E} p_{i_1 \ldots i_k i} = 1,$$

le nombre de paramètres libres est $M^k\,(M-1)$. Ainsi, en ne tenant pas compte des termes qui ne dépendent pas de l'ordre k, l'expression (1) devient, dans le cas du modèle markovien :

$$IC\,(k) = -2 \sum_{i_1, \ldots, i_k, i} n_{i_1 \ldots i_k i} \log \frac{n_{i_1 \ldots i_k i}}{\sum\limits_i n_{i_1 \ldots i_k i}} + c_N M^k\,(M-1)$$

Notons que $n_{i_1 \ldots i_k i} / \sum\limits_i n_{i_1 \ldots i_k i}$ est l'estimateur au sens MV de $p_{i_1 \ldots i_k i}$. Le tableau 3 présente les résultats de l'estimation de l'ordre d'un modèle markovien par les différents critères. Le modèle utilisé pour la simulation est un modèle d'ordre 2, à deux états, et dont les paramètres ont les valeurs suivantes :

$P(1 \mid 1,1) = 0.3 \quad P(1 \mid 1,0) = 0.8 \quad P(1 \mid 0,1) = 0.4$
$P(1 \mid 0,0) = 0.1 \quad P(0 \mid 1,1) = 0.7 \quad P(0 \mid 1,0) = 0.2$
$P(0 \mid 0,1) = 0.6 \quad P(0 \mid 0,0) = 0.9$

TAB. 3 — Modèle de Markov d'ordre 2, $card(E) = 2$

$N = 100$					
ordre	AIC	BIC	φ	$\varphi_{0.2}$	$\varphi_{0.5}$
0	0	0	0	0	63
1	6	35	4	19	36
2	**77**	**64**	**58**	**79**	**1**
3	17	1	30	2	0
$\geqslant 4$	0	0	8	0	0

$N = 1000$					
ordre	AIC	BIC	φ	$\varphi_{0.2}$	$\varphi_{0.5}$
0	0	0	0	0	0
1	0	0	0	0	58
2	**84**	**100**	**82**	**100**	**42**
3	14	0	14	0	0
$\geqslant 4$	2	0	4	0	0

$N = 10000$					
ordre	AIC	BIC	φ	$\varphi_{0.2}$	$\varphi_{0.5}$
0	0	0	0	0	0
1	0	0	0	0	0
2	**85**	**100**	**91**	**100**	**100**
3	12	0	9	0	0
$\geqslant 4$	3	0	0	0	0

3.3 Analyse des résultats

A la lecture des tableaux 1 à 3, nous constatons que lorsque le nombre d'observations est suffisant, le critère BIC fournit une bonne estimation de l'ordre. En revanche, il y a un risque important de sousparamétrisation quand la taille de l'échantillon est faible. Les résultats confirment la tendance à la surparamétrisation du critère AIC, contrairement aux autres critères. Le critère φ a une vitesse de convergence faible, ce qui le rend peu attractif d'un point de vue pratique.

Le critère $\varphi_\beta, 0 < \beta < 0.5$, donné en équation (2), permet une sélection correcte de l'ordre, mais un ajustement de β s'avère nécessaire relativement à la taille de l'échantillon et à la complexité du modèle. En effet, pour $\beta \geqslant 0.5$ la pénalisation est souvent trop élevée vis-à-vis du terme de vraisemblance, d'où un risque important de sousparamétrisation. A titre de comparaison, on peut constater expérimentalement que l'estimation par $MV^{(3)}$ surestime systématiquement l'ordre (quelque soit la taille de l'échantillon), ce qui se vérifie sur la figure 1. Cette figure montre la bien meilleure lisibilité de l'ordre (ici $k = 15$) par les critères φ_β. Le minimum local observé est dû au choix particulier des paramètres a_i. Enfin, l'équivalence de φ et AIC pour $N = 1000$ est due à l'approximation $\log\log 1000 \simeq 1.933 \simeq 2$.

Pour des modèles complexes (cas par exemple des processus de Markov d'ordre élevé avec un nombre d'états important), l'utilisation des critères d'informations classiques pour la sélection de l'ordre est inefficace, car conduisant à une sousparamétrisation du modèle). Une investigation sur la forme de la pénalité est nécessaire pour ce type de modèle.

[3] Ce qui reviendrait à poser $c_N = 0$ dans (1).

4 Conclusion

Dans ce papier, nous avons introduit un nouveau critère pour la sélection de l'ordre d'un modèle. Outre sa qualité de convergence forte, nous montrons son efficacité sur deux types de modélisation classiques, ainsi qu'un ajustement possible à la complexité (nombre de paramètres) de ces modèles.

Références

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and F. Csaki, editors, *Proc. 2nd Int. Symp. Information Theory*, pages 267–281. Akademia Kiado, Budapest, 1973.

[2] R.J. Bhansali and D.Y. Downham. Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterion. *Biometrika*, 64 :547–551, 1977.

[3] J.R. Dickie and A.K. Nandi. A comparative study of AR order selection methods. *Signal Processing*, 40(2) :239–256, November 1994.

[4] A. El Matouat and M. Hallin. Order selection, stochastic complexity and Kullback-Leibler information. In P.M. Robinson and M. Rosenblatt, editors, *Time Series Analysis*, volume 2, in memory of E.J. Hannan, pages 291–299. Springer Verlag, New York, 1996.

[5] E.J. Hannan and B.G. Quinn. The determination of the order of an autoregression. *Journal of the Roy. Stat. Soc. B*, 41(2) :190–195, 1979.

[6] R. Nishii. Maximum likelihood principle and model selection when the true model is unspecified. *Journal of Multivariate analysis*, 27 :392–403, 1988.

[7] J. Rissanen. Modeling by shortest data description. *Automatica*, 14 :465–471, 1978.

[8] J. Rissanen. A universal prior for the integers and estimation by MDL. *The Annals of Statistics*, 11(2) :416–431, June 1983.

[9] J. Rissanen. Stochastic complexity and modeling. *The Annals of Statistics*, 14 :1080–1100, 1986.

[10] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6 :461–464, 1978.

[11] H. Tong. Determination of the order of a markov chain by Akaike's information criterion. *Journal of Applied Probability*, 12 :488–497, 1975.

[12] T. Van Eck. On objective autoregressive model testing. *Signal Processing*, 10 :185–191, 1986.