Une introduction au critère BIC : fondements théoriques et interprétation

Tristan Mary-Huard

UNE INTRODUCTION AU CRITÈRE BIC : FONDEMENTS THÉORIQUES ET INTERPRÉTATION

Émilie LEBARBIER, Tristan MARY-HUARD *

RÉSUMÉ

Dans cet article, nous proposons une discussion sur le critère de sélection de modèles BIC (Bayesian Information Criterion). Afin de comprendre son comportement, nous décrivons les étapes de sa construction et les hypothèses nécessaires à son application en détaillant les approximations dont il découle. En s'appuyant sur la notion de quasi-vrai modèle, nous reprécisons la propriété de « consistance pour la dimension » définie pour BIC. Enfin, nous mettons en évidence les différences de fond entre le critère BIC et le critère AIC d'Akaike en comparant leurs propriétés.

Mots clés : Critère de sélection de modèles, Critère bayésien, Approximation de Laplace, Consistance pour la dimension.

ABSTRACT

In this article we propose a discussion on the Bayesian model selection criterion BIC (Bayesian Information Criterion). In order to understand its behaviour, we describe the steps of its construction as well as the hypotheses required for its application and the approximations needed. Relying on the notion of quasi-true model, we explain the « dimension-consistency » property of BIC. Finally we show the basic differences between BIC and AIC via the comparison of their respective properties.

Keywords : Model selection criterion, Bayesian criterion, Laplace approximation, Dimension-consistency.

1. Introduction

La sélection de modèles est un problème bien connu en statistique. Lorsque le modèle est fixé, la théorie de l'information fournit un cadre rigoureux pour l'élaboration d'estimateurs performants. Mais dans un grand nombre de situations, les connaissances a *priori* sur les données ne permettent pas de déterminer un unique modèle dans lequel se placer pour réaliser une inférence. C'est pourquoi, depuis la fin des années 70, les méthodes pour la sélection de modèles à partir des données ont été développées. Les exemples classiques

* INA-PG (dépt OMIP) / INRA (dépt MIA), 16 rue Claude Bernard, Paris Cedex 05. Emilie.Lebarbier@inapg.fr. Tristan.Maryhuard@inapg.fr

d'application de ces méthodes sont la sélection de variables, ou le choix du nombre de composantes d'un mélange de lois, d'un ordre d'auto-régression, ou de l'ordre d'une chaîne de Markov.

L'une des réponses apportées par les statisticiens au problème de la sélection de modèles est la minimisation d'un critère pénalisé. Les premiers critères apparaissant dans la littérature sont l'Akaike Information Criterion (AIC, Akaike 1973), le Bayesian Information Criterion (BIC, Schwarz 1978), le Minimum Description Length (MDL, Rissanen 1978) et le C_p de Mallows (Mallows 1974). Parmi ces critères, AIC et BIC ont été largement diffusés et appliqués. D'un point de vue théorique, beaucoup de travaux ont été réalisés concernant leurs propriétés statistiques et leur adaptation à des modèles spécifiques. En particulier, plusieurs versions corrigées du critère AIC ont été proposées : AICC (Hurvich et Tsai 1989) et c-AIC (Sugiura 1978) pour de petites tailles d'échantillons par rapport au nombre de paramètres à estimer ; AICR (Ronchetti) pour une régression avec erreurs non-gaussiennes ; QAIC (Burham et Anderson 2002) et c-QAIC (Shi et Tsai 1998) pour des données sur-dispersées. Il existe ainsi une littérature très fournie sur la sélection de modèles par critère pénalisé, qui se développe encore actuellement avec l'apparition d'outils sophistiqués de probabilité, comme par exemple les inégalités de concentration et de déviation, permettant à la fois la construction de critères et leur étude.

Nous nous intéressons ici au critère BIC qui se place dans un contexte bayésien de sélection de modèles. Bien que couramment utilisé par les statisticiens et largement décrit, certains points de sa construction et de son interprétation sont régulièrement omis dans les démonstrations proposées dans la littérature. Il est bien connu que le critère BIC est une approximation du calcul de la vraisemblance des données conditionnellement au modèle fixé. Cependant les résultats théoriques utilisés sont souvent peu explicités, tout comme les hypothèses nécessaires à leurs applications. Par ailleurs, l'interprétation de BIC et la notion de « consistance pour la dimension » ne sont pas toujours très claires pour les utilisateurs.

L'objectif de cet article est d'expliciter ces différents points. Dans un premier temps, nous reprenons de manière détaillée la démonstration de l'ensemble des approximations asymptotiques sur lesquelles repose la construction du critère BIC, en précisant les hypothèses et le rôle des distributions *a priori* posées sur les modèles et les paramètres des modèles (Partie 2). Dans un deuxième temps, nous explicitons le sens des notions de probabilité *a priori* et *a posteriori*. Cela permettra de préciser l'objectif du critère BIC qui est loin d'être explicite au regard de sa définition, et de discuter de l'hypothèse que le « vrai » modèle appartient aux modèles en compétition, hypothèse généralement posée par les auteurs (Partie 3). Enfin, nous présentons et commentons les méthodes de comparaison entre BIC et AIC usuellement proposées, ces deux critères étant souvent mis en concurrence dans la pratique (Partie 4).

2. Construction du critère BIC

Dans cette partie, nous présentons la construction du critère BIC. Pour cela, nous nous appuyons sur les propositions de Raftery (1995).

On dispose d'un n-échantillon $X = (X_1, \ldots, X_n)$ de variables aléatoires indépendantes de densité inconnue f et l'objectif est de l'estimer. Pour cela, on se donne une collection finie de modèles paramétrés $\{M_1, \ldots, M_m\}$. Un modèle M_i est l'ensemble des densités g_{M_i} de paramètre θ_i appartenant à l'espace vectoriel Θ_i de dimension K_i :

$$M_i = \{g_{M_i}(., \theta_i) \ ; \ \theta_i \in \Theta_i\}$$

Il s'agit de choisir un modèle parmi cette collection de modèles.

Le critère BIC se place dans un contexte bayésien : θ_i et M_i sont vus comme des variables aléatoires et sont munis d'une distribution *a priori*. Notons

- $P(M_i)$ la distribution *a priori* du modèle M_i. Elle représente le poids que l'on souhaite attribuer à ce modèle. Par exemple, à partir d'informations que peut détenir l'utilisateur, on peut suspecter que la vraie densité f est proche de certains modèles particuliers et on peut alors donner à ces modèles un poids plus important. En général cependant cette distribution *a priori* est supposée non-informative (uniforme), ne privilégiant aucun modèle :
$$P(M_1) = P(M_2) = \ldots = P(M_m) = 1/m.$$
- $P(\theta_i|M_i)$ la distribution *a priori* de θ_i sachant le modèle M_i. L'ensemble des paramètres θ_i est en effet défini pour le modèle M_i considéré. Nous verrons que cette distribution n'intervient pas dans la forme du critère BIC mais la qualité des approximations faites peut en dépendre.

BIC cherche à sélectionner le modèle M_i qui maximise la probabilité *a posteriori* $P(M_i|X)$:

$$M_{BIC} = \underset{M_i}{\operatorname{argmax}}\, P(M_i|X). \tag{1}$$

En ce sens BIC cherche à sélectionner le modèle le plus vraisemblable au vu des données. La partie 3 est plus particulièrement consacrée à l'interprétation de la probabilité *a posteriori* de M_i. D'après la formule de Bayes, $P(M_i|X)$ s'écrit

$$P(M_i|X) = \frac{P(X|M_i)P(M_i)}{P(X)}. \tag{2}$$

Nous supposons dans toute la suite que la loi *a priori* des modèles M_i est non informative. Sous cette hypothèse et d'après (1) et (2), la recherche du meilleur modèle ne nécessite que le calcul de la distribution $P(X|M_i)$. Ce calcul s'obtient par l'intégration de la distribution jointe du vecteur des paramètres θ_i et des données X conditionnellement à M_i, $P(X, \theta_i|M_i)$, sur

toutes les valeurs de θ_i :

$$P(X|M_i) = \int_{\Theta_i} P(X, \theta_i|M_i)d\theta_i = \int_{\Theta_i} g_{M_i}(X, \theta_i)P(\theta_i|M_i)d\theta_i,$$

où $g_{M_i}(X, \theta_i)$ est la vraisemblance correspondant au modèle M_i de paramètre θ_i :

$$g_{M_i}(X, \theta_i) = \prod_{k=1}^{n} g_{M_i}(X_k, \theta_i) = P(X|\theta_i, M_i).$$

On réécrit cette intégrale sous la forme

$$P(X|M_i) = \int_{\Theta_i} \mathrm{e}^{g(\theta_i)}d\theta_i, \text{ où } g(\theta_i) = \log\left(g_{M_i}(X, \theta_i)P(\theta_i|M_i)\right).$$

La probabilité $P(X|M_i)$ est appelée *vraisemblance intégrée pour le modèle* M_i. Le calcul exact de cette probabilité est rarement possible, on l'approche alors en utilisant la méthode d'approximation de Laplace :

Proposition 2.1 (Approximation de Laplace). — *Soit une fonction* $L : \mathbb{R}^d \to \mathbb{R}$ *telle que* L *est trois fois différentiable sur* $\mathbb{R}^d$ *et atteint un unique maximum sur* $\mathbb{R}^d$ *en* $u^\star$. *Alors*

$$\int_{\mathbb{R}^d} e^{nL(u)}du = e^{nL(u^\star)}\left(\frac{2\pi}{n}\right)^{\frac{d}{2}} |-L''(u^\star)|^{-\frac{1}{2}} + O(n^{-1}).$$

Nous détaillons en Annexe A la démonstration de cette approximation proposée par Tierney et Kadane (1986) et discutons de plus les hypothèses pour l'application de ce résultat à des fonctions L qui dépendent de n comme c'est le cas ici puisque nous l'appliquons à la fonction :

$$L_n(\theta_i) = \frac{g(\theta_i)}{n} = \frac{1}{n}\sum_{k=1}^{n} \log\left(g_{M_i}(X_k, \theta_i)\right) + \frac{\log\left(P(\theta_i|M_i)\right)}{n}. \tag{3}$$

Nous notons

- $\theta_i^\star = \mathrm{argmax}_{\theta_i \in \Theta_i} L_n(\theta_i)$,
- $A_{\theta_i^\star}$ l'opposé de la matrice hessienne des dérivées secondes partielles de la fonction $L_n(\theta_i)$ en θ_i :

$$A_{\theta_i^\star} = -\left[\frac{\partial^2 L_n(\theta_i)}{\partial\theta_i^j \partial\theta_i^l}\right]_{j,l} |_{\theta_i=\theta_i^\star}, \tag{4}$$

  où θ_i^j est la $j$ème composante du vecteur des paramètres θ_i.

Nous obtenons

$$P(X|M_i) = \mathrm{e}^{g(\theta_i^\star)} \left(\frac{2\pi}{n}\right)^{K_i/2} |A_{\theta_i^\star}|^{-1/2} + O_P(n^{-1}),$$

ou encore

$$\log\left(P(X|M_i)\right) = \log\left(g_{M_i}(X,\theta_i^\star)\right) + \log\left(P(\theta_i^\star|M_i)\right) - \frac{K_i}{2}\log(n) \qquad (5)$$
$$+ \frac{K_i}{2}\log(2\pi) - \frac{1}{2}\log\left(|A_{\theta_i^\star}|\right) + O_P(n^{-1}).$$

La difficulté maintenant est l'évaluation de $\theta_i^\star$ et de $A_{\theta_i^\star}$. Asymptotiquement, $\theta_i^\star$ peut être remplacé par l'estimateur du maximum de vraisemblance $\hat{\theta}_i$:

$$\hat{\theta}_i = \underset{\theta_i \in \Theta_i}{\operatorname{argmax}} \frac{1}{n} g_{M_i}(X,\theta_i),$$

et $A_{\theta_i^\star}$ remplacé par $I_{\hat{\theta}_i}$, où $I_{\hat{\theta}_i}$ est la matrice d'information de Fisher pour une observation définie par :

$$I_{\hat{\theta}_i} = -\mathbb{E}\left(\left[\frac{\partial^2 \log\left(g_{M_i}(X_1,\theta_i)\right)}{\partial\theta_i^j \partial\theta_i^l}\right]_{j,l} |_{\theta_i=\hat{\theta}_i}\right).$$

En effet, lorsque n est grand, $\log(g_{M_i}(X,\theta_i)P(\theta_i|M_i))$ se comporte comme $\log(g_{M_i}(X,\theta_i))$, qui croît avec n tandis que $\log(P(\theta_i|M_i))$ reste constant. Remplacer $\theta_i^\star$ par $\hat{\theta}_i$ et $A_{\theta_i^\star}$ par $I_{\hat{\theta}_i}$ dans (5) introduit un terme d'erreur en $n^{-1/2}$ (*cf.* Annexe B). Nous obtenons :

$$\log\left(P(X|M_i)\right) = \log\left(g_{M_i}(X,\hat{\theta}_i)\right) - \frac{K_i}{2}\log(n) \qquad (6)$$
$$+ \log\left(P(\hat{\theta}_i|M_i)\right) + \frac{K_i}{2}\log(2\pi) - \frac{1}{2}\log\left(|I_{\hat{\theta}_i}|\right) + O_P(n^{-1/2}).$$

Par continuité et par la convergence en probabilité de l'estimateur du maximum de vraisemblance $\hat{\theta}_i$, nous obtenons que le premier terme est de l'ordre de $O_P(n)$, le second de l'ordre de $O(\log(n))$ et tous les derniers termes de l'ordre de $O_P(1)$. En négligeant les termes d'erreurs $O_P(1)$ et $O_P(n^{-1/2})$, nous obtenons

$$\log\left(P(X|M_i)\right) \approx \log\left(g_{M_i}(X,\hat{\theta}_i)\right) - \frac{K_i}{2}\log(n).$$

C'est de cette approximation que le critère BIC est issu. Plus précisément, pour le modèle M_i il correspond à l'approximation de $-2\log P(X|M_i)$. BIC est donc défini par :

$$BIC_i = -2\log\left(g_{M_i}(X,\hat{\theta}_i)\right) + K_i \log(n). \qquad (7)$$

Le modèle sélectionné par ce critère est

$$M_{BIC} = \underset{M_i}{\operatorname{argmin}}\, BIC_i.$$

Remarque 1. — Nous avons fait l'hypothèse que la loi des modèles $P(M_i)$ est uniforme. La prise en compte d'une information *a priori* sur les modèles est toutefois possible, on utilise alors le critère modifié :

$$-2\log\left(g_{M_i}(X,\hat{\theta}_i)\right) + K_i \log(n) - 2\log\left(P(M_i)\right)$$

On se reportera à l'article de Kass et Wasserman 1996 pour une discussion sur la spécification d'une loi *a priori* informative sur les modèles.

Remarque 2. — L'erreur en $O_P(n^{-1/2})$ dans l'égalité (6) est négligeable lorsque n tend vers l'infini. Par contre l'erreur d'approximation en $O_P(1)$ peut perturber le choix du modèle final même si les deux premiers termes sont prépondérants quand n est grand puisqu'elle est systématique. Néanmoins, pour certaines distributions *a priori* sur les paramètres θ_i, le terme d'erreur peut être plus petit que $O_P(1)$ (Raftery 1995 ; Kass et Wasserman 1995).

3. Interprétation du critère BIC

L'une des difficultés du critère BIC est son interprétation. La question est la suivante : quel est le modèle que l'on cherche à sélectionner par le critère BIC ? À ce niveau, les notions de probabilité *a priori* ou *a posteriori* d'un modèle sont peu explicites et ne donnent pas une idée intuitive de ce que BIC considère être un « bon » modèle. Les considérations asymptotiques présentées ici vont nous permettre d'interpréter cette notion de meilleur modèle, de déterminer ce que l'on entend par probabilité *a posteriori* d'un modèle, et de préciser en quel sens BIC est un critère « consistant pour la dimension ». Cette interprétation nous permettra aussi de discuter la nécessité de l'hypothèse d'appartenance du vrai modèle à la liste des modèles considérés.

Concernant la consistance du critère BIC, les premiers travaux remontent au milieu des années 80, et visaient à établir la consistance du critère dans des cas simples, par exemple dans le cadre de familles exponentielles (Hartigan 1985 ; Haughton 1988 ; Poskitt). D'autres cas ont été ensuite étudiés, comme la consistance de BIC pour la détermination du nombre de composantes d'un modèle de mélange (Chernoff et Lander 1995 ; Dacunha-Castelle et Gassiat 1997 ; Keribin 1998) ou pour la détermination de l'ordre d'une chaîne de Markov (Csiszar et Shields 2000), et font encore actuellement l'objet de publications (Azaïs, Gassiat et Mercadier 2003). Le lecteur ne trouvera ici qu'une présentation simplifiée de la démonstration de la consistance de BIC dans un cas simple, visant à établir proprement les notions de probabilités *a priori* et *a posteriori*. Une démonstration plus rigoureuse de cette consistance pourra par exemple être trouvée dans (Dudley et Haughton 1997) ou (Dudley et Haughton 2002).

3.1. Le « quasi-vrai » modèle

Nous reprenons ici la remarquable présentation de cette notion proposée par Burnham et Anderson (2002).

Rappelons que la densité à estimer est f. Par simplicité, on se place dans le cas simple où les m modèles $M_1, ..., M_m$ sont supposés emboîtés, i.e. $\Theta_1 \subset \Theta_2 ... \subset \Theta_m$. La pseudo-distance de Kullback-Leibler (appelée dans la suite distance KL) entre deux densités f et g est définie par :

$$d_{KL}(f, g) = \int_\Omega \log\left(\frac{f(x)}{g(x)}\right) f(x)dx.$$

Par abus de notation, on définit la distance KL de f au modèle M_i par :

$$d_{KL}(f, M_i) = \inf_{\theta_i} \; d_{KL}(f, g_{M_i}(., \theta_i)). \tag{8}$$

Puisque les modèles sont emboîtés, la distance KL est une fonction décroissante de la dimension K_i. On note M_t le modèle à partir duquel cette distance ne diminue plus. Du point de vue de la distance KL, M_t doit être préféré à tous les sous-modèles M_i, $i = 1, ..., t-1$ puisqu'il est plus proche de f. Par ailleurs, M_t doit aussi être préféré à tous les modèles d'ordre supérieurs M_i, $i = t+1, ..., m$, puisqu'ils sont plus compliqués que M_t sans pour autant être plus proches de f : ces modèles sont donc surajustés. Nous allons montrer que le critère BIC est consistant pour ce modèle particulier, désigné par Burnham et Anderson comme le modèle « quasi-vrai ». Pour n supposé grand, on s'intéresse à la différence :

$$BIC_i - BIC_t, \quad i \neq t.$$

Premier cas : $i < t$

D'après (7), on a :

$$\begin{aligned} BIC_i - BIC_t &= -2\log(g_{M_i}(X, \hat{\theta}_i)) + 2\log(g_{M_t}(X, \hat{\theta}_t) + (K_i - K_t)\log(n) \\ &= 2n\left[-\frac{1}{n}\sum_{k=1}^{n}\log(g_{M_i}(x_k, \hat{\theta}_i)) + \frac{1}{n}\sum_{k=1}^{n}\log(g_{M_t}(x_k, \hat{\theta}_t))\right] \\ &\quad + (K_i - K_t)\log(n) \\ &= 2n\left[\frac{1}{n}\sum_{k=1}^{n}\log\left(\frac{f(x_k)}{g_{M_i}(x_k, \hat{\theta}_i)}\right) - \frac{1}{n}\sum_{k=1}^{n}\log\left(\frac{f(x_k)}{g_{M_t}(x_k, \hat{\theta}_t)}\right)\right] \\ &\quad + (K_i - K_t)\log(n). \end{aligned}$$

Les deux dernières sommes sont des estimateurs convergeant en probabilité des quantités $d_{KL}(f, M_i)$ et $d_{KL}(f, M_t)$, respectivement (*cf.* Ripley 1995). Pour n grand, on a donc :

$$BIC_i - BIC_t \approx 2n[d_{KL}(f, M_i) - d_{KL}(f, M_t)] + (K_i - K_t)\log(n).$$

Cette approximation, bien que déterministe, suffit à expliciter le comportement asymptotique de $BIC_i - BIC_t$: le premier terme domine et tend vers $+\infty$ avec n. On en déduit donc qu'asymptotiquement les modèles M_i, $i = 1, ..., t-1$ sont disqualifiés par le critère BIC.

Deuxième cas : $i > t$

Dans ce cas là, on reconnaît dans le terme $2\log(g_{M_i}(X, \hat{\theta}_i)) - 2\log(g_{M_t}(X, \hat{\theta}_t))$ la statistique du test du rapport de vraisemblance pour deux modèles emboîtés, qui sous l'hypothèse Ho suit asymptotiquement une loi du Chi-2 à $(K_i - K_t)$ degrés de liberté. On a donc :

$$BIC_i - BIC_t \approx -\chi^2_{(K_i - K_t)} + (K_i - K_t)\log(n).$$

C'est ici le second terme qui domine et tend vers $+\infty$ avec n, les modèles M_i, $i = t+1, ..., m$ sont eux aussi disqualifiés. Le terme en $\log(n)$ joue donc un rôle fondamental : il assure que le critère BIC permet de converger vers le quasi-vrai modèle. Cette convergence vers le quasi-vrai modèle, même s'il est emboîté dans un modèle plus complexe, est appelée consistance pour la dimension.

Il nous est maintenant possible d'interpréter clairement ce que l'on entend par probabilité *a posteriori* du modèle M_i. Elle s'estime à partir des différences $\Delta BIC_i = BIC_i - BIC_{min}$, où BIC_{min} désigne la plus petite valeur observée de BIC sur les m modèles. On a :

$$P(M_i|X) \approx \frac{exp(-\frac{1}{2}\Delta BIC_i)}{\sum_{l=1}^{m} exp(-\frac{1}{2}\Delta BIC_l)}.$$

Cette probabilité tend vers 1 pour le modèle quasi-vrai lorsque n tend vers l'infini, et vers 0 pour tous les autres. Au vu des considérations précédentes, nous pouvons définir cette probabilité comme la probabilité que M_i soit le modèle quasi-vrai de la liste considérée, sachant les données.

3.2. Le vrai modèle fait-il partie de la liste ?

La question de savoir si le vrai modèle ayant engendré les données doit apparaître ou non dans la liste des modèles considérés est longtemps demeurée en suspens dans la littérature consacrée au critère BIC. Bien que nulle part cette hypothèse apparaisse comme nécessaire dans la construction du critère BIC, les auteurs (Schwarz 1978 ; Raftery 1995) posent souvent cette hypothèse, sans toutefois préciser à quelle étape du raisonnement elle intervient. On peut alors se demander si l'hypothèse est nécessaire du point de vue théorique d'une part, pour démontrer la consistance du critère BIC, et du point de vue pratique d'autre part, pour appliquer le critère BIC.

Du point de vue théorique, nous avons vu dans la partie précédente que l'hypothèse n'est pas nécessaire pour établir la consistance vers le quasi-vrai modèle. En réalité, l'hypothèse n'est nécessaire que si l'on s'intéresse à la convergence de BIC vers le vrai modèle : elle sert alors à identifier le quasi-vrai modèle, vers lequel la convergence est toujours assurée, au vrai modèle.

Du point de vue pratique, supposer que le vrai modèle fait partie des modèles en compétition semble peu réaliste, excepté dans de rares cas où le phénomène étudié est simple et bien décrit. Cette constatation n'a aucune conséquence lorsque l'on souhaite comparer des modèles entre eux puisque l'hypothèse n'est pas nécessaire pour la dérivation du critère BIC. Remarquons toutefois que le quasi-vrai modèle de la collection peut être arbitrairement loin (au sens de la distance KL) du vrai modèle. La consistance du critère BIC ne garantit donc pas la qualité du modèle sélectionné, qui dépend fondamentalement du soin apporté par l'expérimentateur pour construire la collection des modèles envisagés.

4. Comparaison des critères AIC et BIC

Les critères AIC (Akaike 1973) et BIC ont souvent fait l'objet de comparaisons empiriques (Burnham et Anderson 2002 ; Bozdogan 1987). Dans la pratique, il a été observé que le critère BIC sélectionne des modèles de dimension plus petite que le critère AIC, ce qui n'est pas surprenant puisque BIC pénalise plus qu'AIC (dès que $n > 7$). La question qui nous intéresse ici est de savoir si l'on peut réellement comparer les performances de ces deux critères, et si oui sur quelles bases. Cette question se justifie pleinement au vu de la littérature. Bien souvent les conclusions des auteurs sur les performances d'AIC et de BIC sont plus guidées par l'idée que se font les auteurs d'un « bon critère » que par la démonstration objective de la supériorité d'un critère sur l'autre, comme l'illustre la présentation des deux critères par Burnham et Anderson (Burnham et Anderson 2002).

Nous commencerons donc par rappeler les propriétés respectives d'AIC et de BIC, avant de considérer les méthodes proposées pour leur comparaison.

4.1. Propriétés des critères

Nous avons vu que le critère BIC est consistant pour le modèle quasi-vrai. Montrons maintenant qu'AIC ne partage pas cette propriété. En effet l'objectif du critère AIC est de choisir le modèle M_i vérifiant :

$$M_{AIC} = \underset{M_i}{\operatorname{argmin}}\, \mathbb{E}\left[\int \log\left(\frac{f(x)}{g_{M_i}(x,\hat{\theta}_i)}\right) f(x)dx\right], \tag{9}$$

en minimisant le critère suivant :

$$M_{AIC} = \underset{M_i}{\operatorname{argmin}}\, -2\log(g_{M_i}(X,\hat{\theta}_i)) + 2K_i.$$

En reprenant le raisonnement asymptotique détaillé pour BIC sur l'exemple de la partie 3.1, on a :

$$\begin{aligned} AIC_i - AIC_t &\approx 2n[d_{KL}(f, M_i) - d_{KL}(f, M_t)] + 2(K_i - K_t) \qquad i < t \\ AIC_i - AIC_t &\approx -\chi^2_{(K_i - K_t)} + 2(K_i - K_t) \qquad i > t. \end{aligned}$$

Les modèles M_i, $i < t$ sont asymptotiquement disqualifiés. En revanche, la probabilité de disqualifier les modèles M_i, $i > t$ ne tend pas vers 0, puisque le terme issu des pénalités $2(K_i - K_t)$ ne diverge pas quand n tend vers l'infini. AIC n'est donc pas consistant pour le quasi-vrai modèle (une démonstration complète de ce résultat peut être trouvée dans (Hanna 1980).

Ce résultat ne démontre en rien la supériorité de BIC sur AIC, car ce dernier n'a pas été conçu pour être consistant, mais dans une optique d'efficacité prévisionnelle. En effet, l'objectif d'AIC est de choisir parmi les m modèles considérés le modèle vérifiant (9), ou de manière équivalente :

$$M_{AIC} = \underset{M_i}{\operatorname{argmin}} \left(d_{KL}(f, M_i) + \mathbb{E}\left[\int_\Omega \log\left(\frac{g_{M_i}(x, \bar{\theta}_i)}{g_{M_i}(x, \hat{\theta}_i)} \right) f(x) dx \right] \right),$$

où $\bar{\theta}_i$ est la valeur de θ_i vérifiant (8). Le premier terme mesure la distance de f au modèle M_i (biais) et le deuxième la difficulté d'estimer $\bar{\theta}_i$ dans le modèle M_i (variance). Sélectionner un modèle par AIC revient donc à chercher le modèle qui fait le meilleur compromis biais - variance pour le nombre de données n dont on dispose. La prise en compte de la taille de l'échantillon vient de ce que l'on somme sur tous les échantillons possibles la distance KL entre f et $g_{M_i}(., \bar{\theta}_i)$. Le meilleur modèle au sens AIC dépend donc de n.

L'objet de cet article n'étant pas de démontrer les propriétés d'AIC, nous nous contenterons ici de dire que dans le cadre gaussien et à nombre de modèles candidats M_i fini, AIC est efficace alors que BIC ne l'est pas (*cf.* Birgé et Massart).

4.2. Méthodes de comparaison

Maintenant qu'il est clair que la notion de meilleur modèle est différente pour AIC et BIC, nous pouvons examiner les méthodes proposées dans la littérature pour les comparer. Nous présentons ici deux points de vue usuels, basés sur des simulations, qui nous permettront de conclure plus généralement sur l'ensemble des méthodes de comparaison proposées. Chaque méthode est discutée d'un point de vue théorique, puis d'un point de vue pratique.

4.2.1. Sélection du vrai modèle

La première méthode est basée sur la simulation de données à partir d'un modèle M_t, qui fait partie de la liste des modèles $M_1 \subset ... \subset M_t \subset ... \subset M_m$ considérés par la suite. Puisque l'on connaît le vrai modèle, on regarde sur un grand nombre de simulations lequel des deux critères le retrouve le plus souvent (Bozdogan 1987). Théoriquement, on peut considérer deux situations, suivant la taille de l'échantillon :

– Lorsque n est petit, le choix optimal pour AIC n'est pas forcément M_t. Ce dernier peut être trop complexe pour la quantité de données n disponible, et il peut exister un modèle M_i de dimension plus petite réalisant un meilleur compromis biais-variance.
– Lorsque n est (très) grand, M_t est meilleur que tous ses sous-modèles, puisque la variance est négligeable devant le biais. Toutefois, un modèle légèrement sur-ajusté aura le même biais que M_t et sa variance, bien que plus grande, sera de toute façon elle aussi négligeable devant le biais. Du point de l'efficacité, les deux modèles sont donc admissibles.

Dans les deux cas, AIC choisit donc un modèle optimal (au sens biais-variance) sans pour autant choisir M_t. Du point de vue théorique, ce type de comparaison favorise donc le critère BIC, puisque lui seul a pour objectif de sélectionner le vrai modèle [1].

En pratique, les résultats obtenus sur des simulations donnent des conclusions très différentes suivant la taille de l'échantillon et la complexité du vrai modèle. Généralement les modèles simulés sont très simples. BIC sélectionne alors le vrai modèle, et AIC le vrai modèle ou un modèle plus grand, ce qui amène les auteurs à conclure que BIC est plus performant pour le choix du vrai modèle. Toutefois, lorsque le modèle est plus complexe, par exemple composé d'une multitude de « petits effets », on constate que BIC devient moins performant qu'AIC car même pour de grandes tailles d'échantillon BIC sélectionne des modèles sous-ajustés.

4.2.2. Sélection d'un modèle prédictif

La deuxième méthode est basée sur la qualité de la prédiction (Burnham et Anderson 2002). On simule des données (x_i, y_i) et l'objectif est de sélectionner un modèle de régression pour faire de la prédiction. Les données simulées sont divisées en deux échantillons X^1 et X^2, de taille respective n_1 et n_2. X^1 est utilisé pour choisir un modèle de prédiction dans la liste $M_1 \subset ... \subset M_m$, et X^2 sert pour le calcul de la performance de prédiction du modèle choisi, mesurée par :

$$MSE = \frac{1}{n_2} \sum_{i \in X^2} (\hat{y_i} - y_i)^2.$$

D'un point de vue théorique, le critère AIC est favorisé puisqu'il prend explicitement en compte la difficulté d'estimation des paramètres dans le terme de variance. Par ailleurs, dans la plupart des cas les données simulées sont gaussiennes. Les critères MSE et AIC sont alors équivalents. Ainsi, le critère gagnant est celui qui a été créé pour répondre à la question posée.

En pratique, on observe généralement de meilleures performances pour AIC que pour BIC, mais pour une raison toute autre que celle invoquée ci-dessus. La pénalité en $\log(n)$ de BIC fait que les modèles sélectionnés par ce critère sont souvent sous-ajustés. En conséquence, le biais de ces modèles est grand, les performances de prédiction ne sont pas satisfaisantes. Toutefois, ici encore

1. Dans le cas de simulations, le vrai modèle fait bien partie de la liste.

les résultats dépendent du modèle simulé et de la taille de l'échantillon. En particulier lorsque le modèle M_t est simple et n_1 grand, BIC peut montrer de meilleures performances qu'AIC.

Lorsque l'objectif de l'analyse statistique est la prédiction, une alternative possible à la sélection d'un « meilleur » modèle pour la prédiction est de considérer l'ensemble des prédictions faites à partir des différents modèles et d'en faire la synthèse, en considérant une moyenne pondérée des prédictions de chaque modèle. Cette alternative a été explorée et peut se justifier théoriquement du point de vue bayésien (Hoeting, Madigan, Raftery et Volinski 1999 ; Madigan et Raftery 1994) comme du point de vue de la théorie de l'information ((Burnham et Anderson 2002)). Dans le cadre bayésien, on parle alors de Bayesian Model Averaging (BMI), et le critère BIC peut être utilisé pour établir la pondération de chaque modèle. Les résultats obtenus en pratique sont souvent meilleurs que ceux obtenus en réalisant la prédiction à l'aide d'un unique modèle. Le lecteur intéressé se reportera avec profit à l'article de Hoeting *et al.* (1999).

4.2.3. Quel critère choisir ?

La conclusion est que le choix d'un critère de sélection de modèles doit être conditionné par l'objectif de l'analyse et la connaissance des données. De nombreux auteurs ont remarqué que BIC et AIC sont utilisés indifféremment quel que soit le problème posé, bien que n'ayant pas le même objectif (Reschenhoffer 1996). Pourtant, choisir entre ces deux critères revient à choisir entre un modèle explicatif et un modèle prédictif. Ce choix devrait donc être argumenté en fonction de l'objectif des utilisateurs. Par ailleurs, les résultats sur données simulées montrent à quel point les performances pratiques sont fonction des situations, en particulier de la complexité du vrai modèle et des modèles candidats, et de la taille de l'échantillon. Ces considérations pratiques et théoriques montrent qu'il n'existe pas de critère universellement meilleur. Seuls l'objectif de l'expérimentateur et sa connaissance des données à analyser peuvent donner un sens à la notion de supériorité d'un critère sur l'autre.

5. Conclusions

Nous avons éclairci les hypothèses, les objectifs et les propriétés du critère BIC. Les considérations de cet article ne prétendent pas être exhaustives : en particulier, nous n'avons pas présenté ici les liens entre BIC et les facteurs de Bayes (Kass et Raftery 1995), ni la place de BIC dans la théorie de la complexité stochastique développée par Rissanen (1987). Nous terminerons par quelques remarques sur les différents points abordés dans cet article.

Il est important de souligner que la construction du critère BIC réalisée en partie 2 a été obtenue dans un cadre asymptotique. En pratique, les tailles d'échantillons sont souvent trop petites pour rentrer dans ce cadre, ce qui peut poser différents problèmes. D'une part les approximations réalisées, comme la méthode de Laplace, peuvent se révéler très inexactes. D'autre part, on constate que la loi *a priori* sur les paramètres $P(\theta_i|M_i)$ n'apparaît

pas dans le critère (7). Cette absence est rassurante, puisqu'elle signifie qu'une mauvaise spécification de $P(\theta_i|M_i)$ n'aura aucun poids sur la sélection de modèles. Toutefois, cette absence ne se justifie qu'asymptotiquement, lorsque l'on remplace θ^* par $\hat{\theta}$ dans l'équation (6), autrement dit lorsque l'on peut faire l'hypothèse que l'information apportée par $P(\theta_i|M_i)$ est négligeable comparée à l'information apportée par l'échantillon. Cette hypothèse n'est pas convenable lorsque n est petit, sauf si $P(\theta_i|M_i)$ est supposée uniforme, ce qui n'est pas toujours possible. On retrouve ici la difficulté propre à l'application de critères asymptotiques à des cas concrets.

Malgré ces considérations, dans un grand nombre de cas l'application du critère BIC fournit des résultats très satisfaisants. Tout d'abord, il existe des cas pour lesquels on souhaite explicitement décrire la structure de la population étudiée. La sélection de modèles dans le cadre du modèle de mélange (Fraley et Raftery 1998) est un bon exemple : l'objectif est de trouver le nombre de composantes du mélange, qui sera ensuite interprété pour distinguer autant de sous-populations distinctes. C'est pourquoi les auteurs (Celeux et Soromenho 1995 ; MacLachlan et Peel 2000) s'accordent pour dire que BIC donne de meilleurs résultats qu'AIC : AIC est logiquement disqualifié puisqu'il n'est pas consistant. Notons toutefois que pour cet objectif, d'autres critères plus performants que BIC ont été proposés pour la sélection du nombre de composantes dans un mélange (Biernacki, Celeux et Govaert 2000).

Par ailleurs, les comparaisons entre AIC et BIC sont généralement réalisées avec une collection finie de modèles. Mais il existe des situations où le nombre de modèles à considérer augmente avec le nombre de données. On peut citer les exemples de la détection de ruptures ou de l'estimation de vraisemblance par histogramme. Pour ces situations, il a été observé que la dimension des modèles choisis avec AIC explose alors que BIC semble proche de l'efficacité (Lebarbier 2005 ; Castellan 1999). Dans ces cas précis, AIC est donc battu sur son propre terrain ! Le paradoxe n'est ici qu'apparent. Lorsque le nombre de modèles à considérer augmente plus vite que la taille de l'échantillon, Birgé et Massart (2001) ont démontré que seuls des critères ayant un terme en $\log(n)$ peuvent être efficaces. Bien que possédant un terme en $\log(n)$ pour d'autres raisons, le critère BIC est alors plus efficace qu'AIC.

Annexes

Pour simplifier l'écriture, nous notons $\theta = \theta_i$, $M_i = M$ et $P(\theta|M) = P(\theta)$. Tous les résultats démontrés dans cette partie requièrent que $P(\theta) \neq 0$ et plus particulièrement que $\log(P(\theta))$ reste borné pour tout θ. On suppose aussi que la vraisemblance et ses dérivées ne dégénèrent pas en $\theta = \hat{\theta}$.

Annexe A

Démontrons la proposition 2.1. Pour plus de simplicité, nous supposons que la fonction L est définie sur $\mathbb{R}$ mais le résultat s'étend facilement à des fonctions de $\mathbb{R}^d$. L'idée principale derrière le résultat de la proposition 2.1 est que

l'intégrale

$$\int_{\mathbb{R}} \mathrm{e}^{nL(u)} du \tag{10}$$

est concentrée autour son maximum (unique) quand n est grand. Pour obtenir l'ordre de l'erreur d'approximation, nous effectuons tout d'abord le développement de Taylor de la fonction $L(u)$ autour de son maximum $L(u^\star)$ à l'ordre 3 :

$$L(u) = L(u^\star) + (u-u^\star)L'(u^\star) + \frac{(u-u^\star)^2}{2}L''(u^\star) + \frac{(u-u^\star)^3}{6}L'''(u^\star) + O((u-u^\star)^4).$$

L'intégrale (10) devient :

$$\int_{\mathbb{R}} \mathrm{e}^{nL(u)} du = \mathrm{e}^{nL(u^\star)} \int_{\mathbb{R}} \mathrm{e}^{\frac{n(u-u^\star)^2}{2}L''(u^\star)} \mathrm{e}^{\frac{n(u-u^\star)^3}{6}L'''(u^\star)} \mathrm{e}^{O(n(u-u^\star)^4)} du \tag{11}$$

puisque par hypothèse, $L'(u^\star) = 0$. Nous cherchons à faire apparaître les moments d'une loi gaussienne. Pour cela, nous développons le second terme exponentiel sous l'intégrale en utilisant le développement de la fonction exponentielle à l'ordre 2 autour de 0 :

$$\mathrm{e}^x = 1 + x + \frac{x^2}{2} + O(x^3).$$

L'intégrale dans l'expression (11) vaut

$$\begin{aligned}
&\int_{\mathbb{R}} \mathrm{e}^{n\frac{(u-u^\star)^2}{2}L''(u^\star)} du \\
&+ \int_{\mathbb{R}} \mathrm{e}^{n\frac{(u-u^\star)^2}{2}L''(u^\star)} \left[\frac{n(u-u^\star)^3}{6}L'''(u^\star) + \frac{n^2(u-u^\star)^6}{72}L'''(u^\star)^2\right] du \\
&+ \int_{\mathbb{R}} \mathrm{e}^{n\frac{(u-u^\star)^2}{2}L''(u^\star)} \left[O(n(u-u^\star)^4) + O(n^2(u-u^\star)^7) + O(n^2(u-u^\star)^8)\right] du.
\end{aligned}$$

En posant

$$\sigma = \frac{1}{\sqrt{-nL''(u^\star)}} \quad \text{et} \quad v = \frac{(u-u^\star)}{\sigma}, \tag{12}$$

nous avons pour $i \geqslant 0$,

$$\int_{\mathbb{R}} (u-u^\star)^i \mathrm{e}^{n\frac{(u-u^\star)^2}{2}L''(u^\star)} du = \sqrt{2\pi}\sigma\sigma^i \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} v^i \mathrm{e}^{-\frac{v^2}{2}} dv.$$

On reconnaît le moment d'ordre i d'une variable aléatoire V de loi gaussienne centrée réduite à une constante près. Les moments d'ordre impair étant nuls,

nous obtenons

$$\begin{aligned}
&\int_{\mathbb{R}} \mathrm{e}^{nL(u)} du \\
= \;& \mathrm{e}^{nL(u^\star)} \sqrt{2\pi}\sigma \left[\mathbb{E}[V^0] + \frac{n^2\sigma^6}{72}\mathbb{E}[V^6] + \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} O\left(v^4 n\sigma^4\right) \mathrm{e}^{-\frac{v^2}{2}} dv \right] \\
= \;& \mathrm{e}^{nL(u^\star)} \sqrt{2\pi}\sigma \left[1 + \frac{5}{24}\frac{1}{nL''(u^\star)^3} + \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} O\left(\frac{v^4}{n}\right) \mathrm{e}^{-\frac{v^2}{2}} dv \right].
\end{aligned}$$

Le terme d'erreur est d'ordre en $1/n$ et il est facile de voir que les termes d'erreurs supérieurs sont d'ordre inférieur ou égal à $1/n$. L'intégrale (10) devient

$$\int_{\mathbb{R}} \mathrm{e}^{nL(u)} du = \mathrm{e}^{nL(u^\star)} \sqrt{-\frac{2\pi}{nL''(u^\star)}} \left[1 + O(n^{-1})\right]. \tag{13}$$

Ce qui conclut la preuve de la proposition 2.1.

Ce résultat est obtenu pour une fonction L qui ne dépend pas de n. Si ce n'est pas le cas, le résultat n'est plus si évident. En effet, en effectuant un développement de Taylor de $L(u)$ autour de $L(u^\star)$ à l'ordre 4, on obtient l'expression explicite du terme d'erreur en $O(n^{-1})$ (dans (13)) qui est :

$$\frac{1}{n}\left[\frac{5}{24}\frac{L'''(u^\star)^2}{L''(u^\star)^3} - \frac{1}{8}\frac{L''''(u^\star)}{L''(u^\star)^2}\right] + O(n^{-2}).$$

Ainsi pour que l'égalité (13) reste valable, il faut que les deux coefficients précédant $1/n$ restent bornés en n (Tierney et Kadane 1986). Il sera donc nécessaire de poser des conditions de régularité sur la fonction L qui assurent les conditions précédentes.

Ici nous cherchons à appliquer la proposition à la fonction L_n (définie par l'équation 3) qui dépend de n. Nous pouvons supposer que la convergence des fonctions d'intérêts vers des quantités possédant des bonnes propriétés suffit. Dans notre cas, on dispose de la convergence en probabilité de L_n vers une quantité qui ne dépend pas de n. En effet, on peut décomposer $L_n(\theta)$ de la manière suivante :

$$L_n(\theta) = LGn(\theta) + B_n(\theta) \quad ,$$

où

$$LG_n(\theta) = \frac{1}{n}\sum_{k=1}^{n} \log\left(g_M(X_k, \theta)\right) \quad et \quad B_n(\theta) = \frac{1}{n}\log\left(P(\theta)\right). \tag{14}$$

Sous la condition $\mathbb{E}[\log\left(g_M(X_1, \theta)\right)] < \infty$, la loi faible des grands nombres assure la convergence en probabilité de $LG_n(\theta)$ vers $\mathbb{E}[\log\left(g_M(X_1, \theta)\right)]$. De plus, $B_n(\theta) \xrightarrow{p.s} 0$. Ce qui conclut sur la convergence en probabilité de $L_n(\theta)$ vers $\mathbb{E}[\log\left(g_M(X_1, \theta)\right]$. Par le même raisonnement, on obtient facilement la convergence des dérivées de la fonction L_n.

Annexe B

L'objectif est ici de donner l'ordre des erreurs d'approximation de $\theta^\star$ par $\hat{\theta}$ et de $A_{\theta^\star}$ par $I_{\hat{\theta}}$.

B1. Approximation de $\theta^\star$ par $\hat{\theta}$

On cherche à montrer

$$\sqrt{n}(\theta^\star - \hat{\theta}) = O_P(1).$$

On décompose ce terme en la somme de deux termes

$$\sqrt{n}(\hat{\theta} - \theta_0) + \sqrt{n}(\theta_0 - \theta^\star),$$

où θ_0 est l'unique maximum de $\mathbb{E}[\log(g_M(X_1, \theta))]$ (l'unicité existe sous la condition d'identifiabilité du modèle). Il suffit alors de montrer que ces deux termes sont bornés en probabilité.

Il est bien connu que sous des conditions de régularité, l'estimateur du maximum de vraisemblance $\hat{\theta}$ satisfait $\sqrt{n}(\hat{\theta} - \theta_0) \overset{\mathcal{L}, n\to\infty}{\longrightarrow} \mathcal{N}(0, I_{\theta_0}^{-1})$. Ce qui assure

$$\sqrt{n}(\hat{\theta} - \theta_0) = O_P(1).$$

Pour le second terme, le résultat est assuré par la convergence en probabilité de $L'_n(\theta)$ vers la même quantité que $LG'_n(\theta)$ qui est $E\left[\dfrac{\partial \log(g_M(X_1, \theta))}{\partial \theta}\right]$ (démonstration similaire à celle présentée en fin de l'Annexe A). Puisque les hypothèses sur cette limite ont déjà été posées pour obtenir le résultat sur $\hat{\theta}$, on obtient la convergence en probabilité de $\theta^\star$ vers θ_0 (*cf.* Lemme 5.10 dans Van der Vaart 1998) et la normalité asymptotique (*cf.* Théorème 5.21 dans Van der Vaart 1998).

B2. Approximation de $A_{\theta^\star}$ par $I_{\hat{\theta}}$

Avant de commencer la démonstration, nous démontrons que $\sqrt{n}(A_\theta - I_\theta)$ est borné en probabilité. D'après les définitions de A_θ (4), $LG_n(\theta)$ et $B_n(\theta)$ (14), on écrit

$$A_\theta - I_\theta = LG''_n(\theta) - I_\theta + \frac{1}{n}[\log(P(\theta))]'',$$

où la dérivée signifie dérivée par rapport à θ. En notant que $\mathbb{E}[LG''_n(\theta)] = I_\theta$, sous la condition que $\mathbb{E}\left(\left[\dfrac{\partial^2 \log(g_M(X_1, \theta))}{\partial \theta^j \partial \theta^l}\right]^2_{j,l}\right) < \infty$, par le théorème central limite, on a la convergence en loi de $\sqrt{n}\,(LG''_n(\theta) - I_\theta)$, ce qui implique

$$LG''_n(\theta) - I_\theta = O_P(n^{-1/2}).$$

Comme $\dfrac{1}{\sqrt{n}}[\log(P(\theta))]''$ converge presque sûrement vers 0, on obtient pour tout θ

$$\sqrt{n}\,(A_\theta - I_\theta) = O_P(1) \quad . \tag{15}$$

Le second résultat qui va nous servir est celui démontré dans la partie précédente qui est

$$\theta^\star = \hat{\theta} + O_P(n^{-1/2}). \tag{16}$$

En effectuant un développement de Taylor de $A_{\theta^\star}$ autour de $A_{\hat{\theta}}$ à l'ordre 1, puisque $\theta^\star$ et $\hat{\theta}$ sont proches quand n est grand, on peut écrire

$$\sqrt{n}(A_{\theta^\star} - I_{\hat{\theta}}) = \sqrt{n}(A_{\hat{\theta}} - I_{\hat{\theta}}) + \sqrt{n}(\theta^\star - \hat{\theta})A'_{\hat{\theta}} + o(\sqrt{n}(\theta^\star - \hat{\theta})^2).$$

On remarque que $A'_{\hat{\theta}} = L'''_n(\hat{\theta})$ et on rappelle que la condition que cette quantité soit bornée en n est demandée pour obtenir l'ordre en $1/n$ dans l'approximation de Laplace. Le résultat ($\text{premiere}_{\text{cv}}$) reste vrai pour $\theta = \hat{\theta}$. Il en vient que le premier terme est de l'ordre de $O_P(1)$. Par (16), on a que le second terme est de l'ordre de $O_P(1)$ et que le dernier (en $o_P(n^{1/2})$) est négligeable. On obtient alors

$$\sqrt{n}(A_{\theta^\star} - I_{\hat{\theta}}) = O_P(1).$$

Références

Akaike H. (1973). Information theory as an extension of the maximum likelihood principle. In B.N. Petrov et F. Csaki (Eds.), *Second International Symposium on Information Theory*, pp. 267-281. Akademiai Kiado, Budapest.

Azaï s J.-M., E. Gassiat, et C. Mercadier (2003). Asymptotic distribution and power of the likelihood ratio test for mixtures : bounded and unbounded case. Technical report, Preprint, Orsay.

Biernacki C., G. Celeux, et G. Govaert (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on pattern analysis and machine intelligence 22* (7), 719-725.

Birgé L. et P. Massart (2001). Gaussian model selection. *J. Eur. Math. Soc.* **3**, 203-268.

Bozdogan H. (1987). Model selection and Akaike's information criterion (AIC) : the general theory and its analytical extensions. *Psychometrika* **52**, 345-370.

Burnham K.P. et D. Anderson (2002). *Model selection and multi-model inference*. Springer-Verlag New York.

Castellan G. (1999). Modified Akaike's criterion for histogram density estimation. Technical Report 61, Universite Paris XI.

Celeux G. et G. Soromenho (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Classification Journal* **13**, 195-212.

Chernoff H. et E. Lander (1995). Asymptotic distribution of the likelihood ratio test that a mixture of two binomials is a single binomial. *Journal of Statistical Planning and Inference 43* (1), 19-40.

Csiszar I. et P.C. Shields (2000). The consistency of the bic markov order estimator. Ann. Statist. 28 (6), 1601-1619.

Dacunha-Castelle D. et E. Gassiat (1997). The estimation of the order of a mixture model. *Bernoulli Journal of Mathematical Statistics and Probability* 3(3), 279-299.

DUDLEY R.M. et D. HAUGHTON (1997). Information criteria for multiple data sets and restricted parameters. *Statistica Sinica*, 265-284.

DUDLEY R.M. et D. HAUGHTON (2002). Asymptotic normality with small relative errors of posterior probabilities of half-spaces. *Ann. Statist. 30* (5), 1311-1344.

FRALEY C. et A. E. RAFTERY (1998). How many clusters ? which clustering method ? answer via model-based cluster analysis. *The Computer Journal* **41**, 578-588.

HANNAN E.J. (1980). The estimation of the order of an arma process. *Ann. Statist.* **8**, 1071-1081.

HARTIGAN (1985). A failure of likelihood ratio asymptotics for normal mixtures. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*.

HAUGHTON D.M.A. (1988). On the choice of a model to fit data from an exponentiel family. *Ann. Statist. 16* (1), 342-355.

HOETING J.A., D. MADIGAN, A.E. RAFTERY, et C.T. VOLINSKY (1999). Bayesian model averaging : A tutorial. *Statist. Science 14* (4), 382-417.

HURVICH C.M. et C.L. TSAI (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297-307.

KASS R.E. et L.A. WASSERMAN (1996). The selection of prior distributions by formal rules. *J. Amer. Statist. Assoc.* **90**, 1343-1370.

KASS R. E. et A. E. RAFTERY (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90**, 773-795.

KASS R. E. et L. WASSERMAN (1995). A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *J. Amer. Statist. Assoc.* **90(2)**, 928-934.

KERIBIN C. (1998). Consistent estimate of the order of mixture models. *Comptes Rendus de l'Académie des Sciences* **326**, 243-248.

LEBARBIER E. (2005). Detecting multiple change-points in the mean of gaussian process by model selection. *Signal processing* **85**, 717-736.

MADIGAN D. et A.E. RAFTERY (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Statist. Assoc.* **89**, 1535-1546.

MALLOWS C.L. (1974). Some comments on Cp. *Technometrics* **15**, 661-675.

MCLACHLAN G. et D. PEEL (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics.

POSKITT D. S. (1987). Precision, complexity and bayesian model determination. *J. R. Statist. Soc. B* **49** (2), 199-208.

RAFTERY A. E. (1995). Bayesian model selection in social research (with discussion). *Sociological Methodology*, 111-196.

RESCHENHOFFER E. (1996). Prediction with vague prior knowledge. *Comm. Statist.* **25**, 601-608.

RIPLEY B.D. (1995). *Pattern Recognition and Neural Networks*. Cambridge University Press.

RISSANEN J (1978). Modelling by the shortest data description. *Automatica* **14**, 465-471.

RISSANEN J. (1987). Stochastic complexity. *J. R. Statist. Soc. B* **49**, 223-239.

RONCHETTI (1985). Robust model selection in regression. *Statis. Probab. Lett.* **3**, 21-23.

SCHWARZ G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

SHI P. et C.L. TSAI (1998). A note on the unification of the Akaike information criterion. *J. R. Statist. Soc. B* **60**, 551-558.

SUGIURA (1978). Further analysis of the data by akaike's information criterion and the finite corrections. *Comm. Statist.* **A7**, 13-26.

TIERNEY L. et J.B. KADANE (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81**, 33-59.

VAN DER VAART A.W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics.