Table des matières

Introduction **1**

1 Les mesures de divergences généralisées **4**
1.1 Les f-divergences de aly et silvey 4
1.2 Rappel d'une métrique 5
1.3 Divergence en variation totale 5
1.4 Divergence de Hellinger 8
1.5 Divergence de khi-deux 10
1.6 Divergence de kullback leibler 11
1.6.1 Définition 11
1.6.2 Propriétés 12
1.6.3 Duale de l'information de kullback 18
1.6.4 Lien entre divergence de kullback et maximum de vraisemblance 18

2 Les critères de sélections de modèles basés sur la J-divergence symétrique de kullback dans le cas d'un échantillon de petit taille **20**
2.1 AIC, AICc et KIC 20
2.2 KICc 24
2.3 Approximation de KICc 27

3 Le critère AIC et les divergences symétriques de kullback **28**
3.1 Le critère AIC et le mélange moyen pondéré 28
3.2 AIC , Moyenne géométrique et Moyenne harmonique 31
3.2.1 Moyenne géométrique 31
3.2.2 Moyenne harmonique 32

3.3 Rapport entre les différentes divergences symétriques de kullback 33
3.4 Exemple de nouvelle correction du biais de AIC 34

4 Applications 1 : Évaluations numériques 35
4.1 Régression polynomiale 35
4.2 Modélisations Autorégressives 37
4.3 Les prévisions dans les taux de changes 38

5 Applications 2 : Évaluations numériques : Application à la sélection de l'ordre des modèles polynomiales 42

Conclusion 48

Bibliographie 49

Rapport-gratuit.com
LE NUMERO 1 MONDIAL DU MÉMOIRES

Introduction

L'un des problèmes les plus fréquents en statistique est la détermination d'un modèle approprié pour décrire ou caractériser un ensemble de données expérimentales. Cette détermination se fait sur deux choses ; le choix d' un modèle à structure approprié et l'estimation de ses paramètres.
L'estimation des paramètres se fait en général par la méthode du maximum de vraisemblance ; le choix du modèle est facilité par des critères de sélections.Un critère de sélection est un estimateur d'une variante de la divergence entre le vrai modèle de loi inconnue et le modèle candidat de loi connue. Cette divergence évalue l'écart ou "distance" entre ces deux modèles.
Plusieurs divergences ont été proposées pour évaluer cette "distance" comme la divergence d'hellinger , la divergence variationnelle , la divergence de khi-deux , la divergence de kullback leibler dirigée ou information de kullback ou encore entropie relative etc. parmi eux certaines verifient toutes les propriétés d'une distance , d'autres ne vérifient pas toutes les propriétés d'une distance comme la distance de kullback leibler dirrigée qui est asymétrique et ne respecte pas l'inégalité triangulaire.

Dans ce papier ,des divergences symétriques seront crées à partir des divergences dirrigés et alternatives de kullback, et des critères de sélection en seront déduits . Ces critères de sélection sont des estimateurs asymptotiquement sans biais des variantes de ces divergences symétriques pour les modèles de petites tailles ,et une version corrigée de quelques de ces critères est aussi déduite en vue d'une réduction de leurs biais sous les modèles de régressions linèaires . Ces critères corrigés sont des estimateurs exactement sans biais dans le cas des modèles correctement spécifiés c'est à dire lorsque l'ensemble des modéles candidats contient le vrai modèle . Ce papier montre que le critère corrigé est plus performant que le critère non corrigé. Une comparaison de ces critères par rapport aux autres critères basés sur la divergence dirigée ou I-divergence comme le critère d'information d'akaike

AIC , le critère d'information bayienne BIC, et l'erreur de prédiction finale FPE, est aussi faite.Le critère d'information d'akaike AIC sert comme un estimateur asymptotiquement sans biais de d'une variante de la divergence dirigée de Kullback, entre le modèle générateur ou vrai modèle et le modèle d' approximation équipée ou modèle candidat sous l'hypothèse que le vrai modèle est correctement spécifié .Comme le nombre de paramètres k du modèle candidat est souvent grand par rapport à la taille n de l'échantillon (on considere k est grand par rapport à n si $\frac{n}{k} < 40$), l'AIC devient trop biaisé et estime mal les informations de Kullback-Leibler et conduit au choix de modèles surparamétrés. Une version corrigée de AIC serait alors nécessaire , notée AICc, a été initialement proposé par Suguira pour modèles de régression linéaire avec une vue vers la réduction du biais . Il a été constaté plus tard par Hurvich et Tsai de produire non seulement une réduction dramatique du biais , mais améliore fortement la sélection de modèle pour les échantillons de petites tailles .Le critère d'information bayien BIC est un critère de sélection de modèle basé sur de arguments bayiens et sur le maximum de la probalité a posteriori .Le critère d'information bayésien (BIC) ou critère Schwarz (SBIC) est un critère pour la sélection de modèles parmi un ensemble fini de modèles. Il est basé, en partie, sur la fonction de vraisemblance, et il est étroitement lié à l'information d'Akaike critère (AIC). Lorsque les modèles de montage ne sont pas adéquats, il est possible d'augmenter la probabilité en ajoutant des paramètres, mais cela peut entraîner surapprentissage. Le BIC résout ce problème en introduisant un terme de pénalité pour le nombre de paramètres dans le modèle. Le terme de pénalité est plus grand dans le BIC que dans l'AIC. Le BIC a été développé par Gideon E. Schwarz, qui a donné un argument bayésien pour l'adopter. En fait, Akaike a été tellement impressionné par le formalisme bayésien de Schwarz qu'il a développé son formalisme bayésien propres, maintenant souvent désigné comme l'ABIC. Le BIC est un résultat asymptotique obtenu sous l'hypothèse que la distribution des données est dans la famille exponentielle.

Ce mémoire est divisé en trois chapitres et deux applications. Dans le premier chapitre, nous ferons l'étude des différentes formes de mesures de divergences et faire une comparaison entre eux.
Dans le deuxième chapitre , la J-divergence symétrique de kullback est présentée et un critère d'information de kullback KIC est étudié en tant

que estimateur asymptotiquement sans biais d'une variante de la J-divergence. Un critère corrigé de KIC nommé KICc est aussi présenté en vue d'étudier sa performance par rapport à KIC et par rapport aux autres critères sous certains contextes.
Dans le troisième chapitre, on a présenté différentes formes de divergences symétriques : sous forme de moyenne arithmétique, sous forme de moyenne géométrique, sous forme de moyenne harmonique . Pour chaque divergence symétrique un critère de sélection de modèle est dérivé en tant estimateur asymptotiquement sans biais d'une variante de cette divergence. Pour la forme moyenne arithmétique un critère corrigé nommé AIC^*c est déduit en vue d'étudier sa performance par rapport au critère non corrigé et comparé éga-
lement sa performance par rapport au performance du critère $AICc$ vu dans le chapitre précédent et des autres critères dans la sélection des modèles de régressions linéaires.
La première application est une évaluation numérique du chapitre 2, dans la sélection de l'ordre des modèles polynomiales régressifs , dans les séries chronologiques de modélisation , et à la prévision du taux de change du dollar en starling avec des données réelles
La deuxième application est une évaluation numérique du chapitre 3, toujours dans la sélection de l'ordre des modèles polynomiales régressifs

Chapitre Premier

Les mesures de divergences généralisées

1.1 Les f-divergences de aly et silvey

Les f-divergences sont une famille de divergences "distance"qui évaluent l'écart entre deux distributions de probabilités elles sont introduites par Csiszàr en 1967 et par Ali et Silvey en 1966 dans leurs travaux en théorie de l'information pour des problèmes statistiques

Définition 1.1.1. *soient P , Q deux mesures de probabilités (loi) sur un espace probabilisés (E,B), une f-divergence de P par rapport à Q est définie par ;*

$$D(P,Q) = g(\int f\{\phi(dP,dQ)\}dP) \tag{1.1}$$

ϕ représente le rapport $\frac{dQ}{dP}$
g est une fonction réelle croissante sur R
f est une fonction réelle convexe et continue sur R^+,souvent de classe C^2.
on impose de plus :
$g(f(1)) = 0$ pour garantir $D(P,P) = 0$
$g'(f(1))f''(1) > 0$ pour la métrique
si $g(x) = x$ et $f(x) = -\ln(x)$ on a la divergence de kullback leibler
si $g(x) = x$ et $f(x) = \frac{1}{2}(\sqrt{x}-1)^2$ on a la divergence de hellinnger
si $g(x) = x$ et $f(x) = \frac{1}{2}|1-x|$ on a la divergence en variation totale
si $g(x) = x$ $f(x) = (x-1)^2$ on a la divergence de khi-deux

Rapport-Gratuit.com

1.2 Rappel d'une métrique

Une métrique sur un ensemble X est une fonction (appelée fonction de distance ou simplement la distance) $d : X \times X \to R$
(où R est l'ensemble des nombres réels). Pour tous x , y , z dans X , cette fonction est nécessaire pour satisfaire aux conditions suivantes :
1. $d(x, y) \geq 0$(non-négative)
2. $d(x, y) = 0$si et seulement si $x = y$ (l'identité des indiscernables . Notez que la condition 1 et 2 produisent ensemble définie positive)
3. $d(x, y) = d(y, x)$ (symétrie)
4. $d(x, z) \leq d(x, y) + d(y, z)$ (sous-additivité / inégalité du triangle).

1.3 Divergence en variation totale

Définition 1.3.1. *La divergence en variation totale ou divergence variationnelle entre deux mesures de probabilité P et Q sur (E,B) est définie par :*

$$V(P, Q) = \sup_{b \in B} |P(b) - Q(b)| \tag{1.2}$$

Officieusement, il s'agit de la plus grande différence possible entre les probabilités que les deux distributions de probabilité peut attribuer à un même événement.

Propriétés

la divergence en variation totale est une métrique car elle vérifie toutes les propriétés d'une métrique

Proposition 1.3.1. *soient P et Q deux mesures de probabilités sur (E,B), la divergence variationnelle entre P et Q est donnée par :*

$$V(P, Q) = \frac{1}{2} \int |dQ - dP| \tag{1.3}$$

NB :la divergence en variation fait partie des f-divergence, elle correspond à $g(x) = x$ et $f(x) = \frac{1}{2}|1 - x|$

Expression avec densité

Soient P et Q deux lois de probabilité absolument continues par rapport une mesureμ (dominées par μ) sur (E ,B), de densités $f = \frac{dP}{d\mu}$ et $g = \frac{dQ}{d\mu}$ respectivement.ALors (2) s'écrit

$$V(P,Q) = \frac{1}{2}\int |f-g|d\mu \tag{1.4}$$

cette expression ne dépend pas du choix deμ

Démonstration. Nous démontrons d'abord que :

$$V(P,Q) = \frac{1}{2}\int_E |f-g|d\mu.$$

Pour cela on montre en premier lieu que : $V(P,Q) \leq \frac{1}{2}\int_E |f-g|d\mu$. Soit b un élément de B. En notant $\bar{b}$ le complémentaire de b, on a :
$|P(b)-Q(b)| = |P(\bar{b})-Q(\bar{b})|$. Alors :

$$\begin{aligned} 2|P(b)-Q(b)| &= |P(b)-Q(b)| + |P(\bar{b})-Q(\bar{b})| \\ &= |\int_B (f-g)d\mu| + |\int_{\bar{B}} (f-g)d\mu| \\ &\leq \int_B |f-g|d\mu + \int_{\bar{B}} |f-g|d\mu \\ &= \int_E |f-g|d\mu. \end{aligned}$$

Ensuite on démontre que :

$$\int_{(f<g)} (g-f)d\mu = \int_{(f>g)} (f-g)d\mu = \frac{1}{2}\int_E |f-g|d\mu$$

Comme P et Q sont des lois de probabilités sur E, on a alors :

$$\int_E f d\mu = 1 = \int_E g d\mu,$$

donc on a :

$$0 = \int_E (f-g)d\mu = \int_{(f<g)} (f-g)d\mu + \int_{(f>g)} (f-g)d\mu,$$

et donc :

$$\int_{(f<g)}(g-f)d\mu = \int_{(f>g)}(f-g)d\mu.$$

Alors :

$$\begin{aligned}\int_E |f-g|d\mu &= \int_{(f<g)}(g-f)d\mu + \int_{(f>g)}(f-g)d\mu \\ &= 2\int_{(f>g)}(f-g)d\mu \\ &= 2\int_{(f<g)}(g-f)d\mu.\end{aligned}$$

Finalement on démontre que : $V(P,Q) \geq \frac{1}{2}\int_E |f-g|d\mu$.
Définissons $B_0 = \{\mathrm{f} > \mathrm{g}\}$. Alors :

$$\begin{aligned}V(P,Q) &\geq |P(B_0) - Q(B_0)| \\ &= |\int_{(f>g)}(f-g)d\mu| \\ &= \int_{(f>g)}(f-g)d\mu \\ &= \frac{1}{2}\int_E |f-g|d\mu.\end{aligned}$$

Pour finir, démontrons l'égalité suivante :

$$\int_E |f-g|d\mu = \frac{1}{2}sup\{|\int_E hdP - \int_E hdQ\ |; \|h\|_\infty \leq 1\}.$$

Soit φ une fonction mesurable définie sur Ω et à valeurs dans R.
Commençons par remarquer que, pour toute fonction h mesurable de norme infinie inférieure ou égale à 1, on a :

$$|\int_E \varphi h d\mu| \leq \int_E |\varphi|d\mu,$$

et que l'égalité a lieu pour la fonction $h = \mathbf{I}_{\varphi\geq 0} - \mathbf{I}_{\varphi<0}$. On en déduit que :

$$\int_E |\varphi|d\mu = \sup\{|\int_E \varphi h d\mu|; \|h\|_\infty \leq 1\}.$$

L'égalité cherchée est obtenue en appliquant ce résultat à la fonction $|f-g|$. □

1.4 Divergence de Hellinger

La divergence de Hellinger a été introduite en 1909 dans le cas discret. Elle joue un role crucial en inférence statistique. L'un de ses intérêts est qu'elle est souvent calculable explicitement. Elle correspond aux fonctions $g(x) = x$ et $f(x) = \frac{1}{2}(\sqrt{x} - 1)^2$ d'une f-divergence

Définition 1.4.1. *Soient P et Q deux mesures de probabilités sur (E ,B) la divergence d' Hellinger au carré entre P et Q est donnée par :*

$$H(P, Q) = \frac{1}{2} \int (\sqrt{dP} - \sqrt{dQ})^2 \tag{1.5}$$

Expression avec densités

soient P et Q deux lois de probalité absolument continues par rapport une mesureμ (dominées par μ) sur (E ,B), de densitée $f = \frac{dP}{d\mu}$ et $g = \frac{dQ}{d\mu}$ respectivement.ALors (1.5) s'écrit :

$$H(P, Q) = \frac{1}{2} \int (\sqrt{f} - \sqrt{g})^2 d\mu \tag{1.6}$$

Cette expression ne dépend pas du choix deμ

$$H(P, Q) = 1 - \int \sqrt{fg} d\mu \tag{1.7}$$

Pour la preuve il suffit de développer $(\sqrt{f} - \sqrt{g})^2$ et de tenir compte que l' intégrale de f et l'intégrale g sont égales à 1 sur E

Propritées

La divergence d'hellinger est une métrique, car elle vitrifie toute les propriétés d'une métrique

Démonstration. Soient f ,g et h les densités de P,Q et R respectivement
$0 \leq H(P, Q) \leq 1$ positive
(Inégalité de Cauchy-Schatz donne ce résultats)
$H(P, Q) = H(Q, P)$ symétrique

$$\begin{aligned} H(P,Q)+H(Q,R) &= \frac{1}{2}\int(\sqrt{f}-\sqrt{g})^2d\mu+\frac{1}{2}\int(\sqrt{g}-\sqrt{h})^2d\mu \\ &\geq \frac{1}{2}\int[|\sqrt{f}-\sqrt{g}|+|\sqrt{g}-\sqrt{h}|]^2d\mu \\ &\geq \frac{1}{2}\int[\sqrt{f}-\sqrt{g}+\sqrt{g}-\sqrt{h}]^2d\mu \\ &\geq \int[\sqrt{f}-\sqrt{h}]^2d\mu \end{aligned}$$

$H(P,Q)+H(Q,R)\geq H(P,R)$
$H(P,Q)+H(Q,R)\geq H(P,R)$ inégalité triangulaire

□

Proposition 1.4.1. *(Lien entre divergence de Hellinger et la divergence en variation totale) Soient P et Q deux mesures de probabilité sur l'espace (E,B).*

$$H(P,Q)\leq V(P,Q)\leq\sqrt{H(P,Q)}\sqrt{2-H(P,Q)}.$$

Ici on a encadrer la mesure en variation totale par la mesure de Hellinger

Démonstration. Montrons successivement les deux inégalités. Soit μ une mesure qui domine les deux mesures P et Q. Pour l'inégalité de gauche, il suffit de remarquer que :

$$\forall x,y\in R^+,(\sqrt{x}-\sqrt{y})^2\leq|x-y|,$$

pour obtenir :

$$\frac{1}{2}\int_E(\sqrt{f}-\sqrt{g})^2d\mu\leq\frac{1}{2}\int_E|f-g|d\mu$$

c'est à dire :

$$H(P,Q)\leq V(P,Q).$$

Pour l'inégalité de droite, on utilise l'inégalité de Cauchy-Schwarz :

$$\begin{aligned} V^2(P,Q) &= \frac{1}{4}(\int_E|\sqrt{f}-\sqrt{g}||\sqrt{f}+\sqrt{g}|d\mu)^2 \\ &\leq \frac{1}{4}(\int_E(\sqrt{f}-\sqrt{g})^2d\mu)(\int_E(\sqrt{f}+\sqrt{g})^2d\mu). \end{aligned}$$

Or la formule de la divergence de Hellinger donne :
$H(P,Q) = 1 - \int \sqrt{fg} d\mu$et donc :

$$\int_E (\sqrt{f} + \sqrt{g})^2 d\mu = 4 - 2H(P,Q).$$

D'où :

$$V^2(P,Q) \leq H(P,Q)(2 - H(P,Q)).$$

□

1.5 Divergence de khi-deux

Définition 1.5.1. *Soient P et Q deux mesures de probabilité sur (E ,B) telles que P soit absolument continue par rapport à Q, la divergence khi-deux entre P et Q est définie par*

$$\chi^2(P,Q) = \int \frac{(dP - dQ)^2}{dQ} \tag{1.8}$$

NB : la divergence de khi-deux correspond à une une f-divergence pour $f(x) = x$ et $C(x) = (x-1)^2$

Expression avec densités

Si P et Q ont des densités respectives f et g par rapport à une mesure μ la mesure de χdevient

$$\chi^2(P,Q) = \int \frac{(f-g)^2}{g} d\mu \tag{1.9}$$

Propriétés

la divergence de khi-deux est une métrique, car elle vérifie toute les propriétés d'une métrique

Proposition 1.5.1. *soient P et Q deux lois de probabilité telle P soit absolument continue par rapport à Q. Alors :*

1. $V^2(P,Q) \leq \frac{1}{4}\chi^2(P,Q)$
2. $H(P,Q) \leq \frac{1}{2}\chi^2(P,Q)$

Pour la démonstration de cette proposition voir la thèse de *Béatrice Lachaud* docteur de l'université Réné Décarte - Paris 5. Un des avantages de la distance du χ^2 est qu'elle se comporte bien vis à vis des produits dans le sens où la distance du χ^2 entre deux mesures produits peut s'exprimer simplement en fonction des distances du χ^2 entre les marginales

1.6 Divergence de kullback leibler

La divergence (ou "distance") de kullback leibler ou encore entropie relative ou encore information de kullback est une mesure de dissimilarité entre deux distributions de probabilités P et Q.Elle doit son nom à solomone kullback et richard leibler, deux cryptanalystes americains. Selon la NSA ,c'est durant les années 60 ,alors qu'ils travaillaient pour cette agence, que kullback et leibler ont inventé cette mesure. Elle aurait d'ailleurs servi à la NSA dans son effort de cryptanalyse pour le projet de VENONA. Cette mesure s'interprète comme la différence moyenne du nombre de bits nécessaires au codage d'échantillons de P selon que le codage est choisi optimal pour la distribution P ou Q. Typiquement, P représente les données, les observations, ou une distribution de probabilités calculée avec précision. La distribution Q représente typiquement une théorie, un modèle, une description ou une approximation de P. La divergence de Kullback-Leibler entre dans la catégorie plus large des f-divergences

1.6.1 Définition

soient P et Q deux mesures de probabilité sur (E,B),la divergence de kullback ou information de kullback ou encore entropie relative de Q par rapport à P est donnée par

$$I(P,Q) = \int dP \ln \frac{dP}{dQ} = E_P(\ln \frac{dP(X)}{dQ(X)}) \tag{1.10}$$

Cette définition n'a de sens que si P$\ll$ Q c'est à dire P est absolument continue par rapport à Q.Dans le cas, contraire, on pose $I(P,Q) = +\infty$
NB : la 'distance' de kullback est une forme de f-divergence elle correspond à $f(x) = x$ et $C(x) = -\ln(x)$

Expression avec densités

Soient P et Q deux lois de probabilité absolument continues par rapport une mesure μ (dominées par μ) sur (E ,B), de densité $f = \frac{dP}{d\mu}$ et $g = \frac{dQ}{d\mu}$ respectivement. Alors P est absolument continu par Q ,signifie que $\{x : f(x) > 0\} \subset \{x : g(x) > 0\}$. La divergence de kullback de Q par rapport à P s'écrit alors :

$$I(P,Q) = \int f \ln \frac{f}{g} d\mu \tag{1.11}$$

Cette définition est indépendante du choix de μ

1.6.2 Propriétés

soient P et Q deux mesures de probabilité sur (E,B) on a :
$I(P,Q) \geq 0$
$I(P,Q) = 0 \Leftrightarrow P = Q$
avant de passer à la preuve, nous allons faire un rappel de l'inégalité de jensen dont nous aurons besoin dans la preuve.

Inégalité de Jensen

L'inégalité de Jensen pour les fonctions convexes joue un rôle crucial dans la théorie des inégalités. Cela est du au fait que d'autres inégalités telles que celle de la moyenne arithmétique et de la moyenne géométrique, les inégalité de Hölder et de Minkowski, l'inégalité de Ky Fan etc, peuvent être considérés comme des cas particuliers de celle-ci.
Soit C une partie convexe de l'espace linéaire X et f une fonction sur C. Si $p = (p_1, ..., P_n)$ est une suite de probabilités et $x = (x_1, ..., X_n) \in C^n$, alors

$$f(\sum_{i=1}^{n} p_i x_i) \leq \sum_{i=1}^{n} p_i f(x_i),$$

bien connu dans la littérature sous le nom de : inégalité de Jensen. En 1989, C.J. Pecarè et ali ont obtenu le raffinement suivant de l'équation

précédente :

$$\begin{aligned} f(\sum_{i=1}^{n} p_i x_i) &\leq \sum_{i_1,...,i_{k+1}=1} p_{i_1}...p_{i_{k+1}} f(\frac{x_{i_1}+...+x_{i_{k+1}}}{k+1}) \\ &\leq \sum_{i_1,...,i_k=1} p_{i_1}...p_{i_k} f(\frac{x_{i_1}+...+x_{i_{k+1}}}{k}) \\ &\leq ... \leq \sum_{i=1}^{n} p_i f(x_i), \end{aligned}$$

pour $k \geq 1$,et p, x défini comme ci-dessus.
Le théorème suivant présente un autre raffinement de l'inégalité de Jensen pour les fonctions convexes définies sur les espaces linéaires.

Théorème 1.6.1. *Soit une fonction convexe f, $f : C \longrightarrow R$, définie dans un sous-espace convexe C de l'espace linéaire X, $x_i \in C$, $p_i > 0$, $i \in \{1, ..., n\}$ avec $\sum_{i=1}^{n} p_i = 1$. Alors :*

$$\begin{aligned} f(\sum_{j=1}^{n} p_j x_j) &\leq \min_{k\in(1,...,n)} [(1-p_k) f(\frac{\sum_{j=1}^{n} p_j x_j - p_k x_k}{1-p_k}) + p_k f(x_k)] \\ &\leq \frac{1}{n}[\sum_{k=1}^{n}(1-p_k) f(\frac{\sum_{j=1}^{n} p_j x_j - p_k x_k}{1-p_k}) + \sum_{k=1}^{n} p_k f(x_k)] \\ &\leq \max_{k\in(1,...,n)} [(1-p_k) f(\frac{\sum_{j=1}^{n} p_j x_j - p_k x_k}{1-p_k}) + p_k f(x_k)] \\ &\leq \sum_{j=1}^{n} p_j f(x_j). \end{aligned}$$

En particulier

$$\begin{aligned} f(\frac{1}{n}\sum_{j=1}^{n} x_j) &\leq \frac{1}{n}\min_{k\in(1,...,n)}[(n-1)f(\frac{\sum_{j=1}^{n} x_j - x_k}{n-1}) + f(x_k)] \\ &\leq \frac{1}{n^2}[(n-1)\sum_{k=1}^{n} f(\frac{\sum_{j=1}^{n} x_j - x_k}{n-1}) + \sum_{k=1}^{n} f(x_k)] \\ &\leq \frac{1}{n}\max_{k\in(1,...,n)}[(n-1)f(\frac{\sum_{j=1}^{n} x_j - x_k}{n-1}) + f(x_k)] \\ &\leq \frac{1}{n}\sum_{j=1}^{n} f(x_j). \end{aligned}$$

Démonstration. Pour tout k ∈{1,...,n}, nous avons :

$$\begin{aligned} &\sum_{j=1}^{n} p_j x_j - p_k x_k \\ &\qquad = \sum_{j=1, j\neq k}^{n} p_j x_j \\ &\qquad = \frac{\sum_{j=1,j\neq k}^{n} p_j}{\sum_{j=1,j\neq k}^{n} p_j} \sum_{j=1,j\neq k}^{n} p_j x_j \\ &\qquad = (1-p_k)\frac{1}{\sum_{j=1,j\neq k}^{n} p_j} \sum_{j=1,j\neq k}^{n} p_j x_j \end{aligned}$$

Ce qui implique que :

$$\frac{\sum_{j=1}^{n} p_j x_j - p_k x_k}{1-p_k} = \frac{1}{\sum_{j=1,j\neq k}^{n} p_j} \sum_{j=1,j\neq k}^{n} p_j x_j \in C$$

Pour chaque $k \in \{1,...,n\}$, de plus le terme à droite de l'équation précédente est une combinaison convexe des éléments $x_j \in C$, $j \in \{1,...,n\}\backslash\{k\}$. En prenant la fonction f sur l'équation précédente et en

appliquant l'inégalité de Jensen, nous obtenons successivement :

$$f(\frac{\sum_{j=1}^{n} p_j x_j - p_k x_k}{1-p_k})$$

$$\begin{aligned} &= f(\frac{1}{\sum_{j=1,j\neq k}^{n} p_j} \sum_{j=1,j\neq k}^{n} p_j x_j) \\ &\leq \frac{1}{\sum_{j=1,j\neq k}^{n} p_j} \sum_{j=1,j\neq k}^{n} p_j f(x_j) \\ &= \frac{1}{1-p_k}[\sum_{j=1}^{n} p_j f(x_j) - p_k f(x_k)] \end{aligned}$$

Pour tout $k \in \{1, ..., n\}$, ce qui implique

$$(1-p_k) f(\frac{\sum_{j=1}^{n} p_j x_j - p_k x_k}{1-p_k}) + p_k f(x_k) \leq \sum_{j=1}^{n} p_j f(x_j)$$

Pour chaque $k \in \{1, ..., n\}$. En utilisant la convexité de la fonction, nous avons aussi

$$\begin{aligned} &(1-p_k) f(\frac{\sum_{j=1}^{n} p_j x_j - p_k x_k}{1-p_k}) + p_k f(x_k) \\ &\geq f[(1-p_k).\frac{\sum_{j=1}^{n} p_j x_j - p_k x_k}{1-p_k} + p_k x_k] \\ &= f(\sum_{j=1}^{n} p_j x_j) \end{aligned}$$

Pour chaque $k \in \{1, ..., n\}$. En prenant le minimum dans l'équation qui précéde, et en utilisant le fait que

$$\min_{k\in\{1,...,n\}} \alpha_k \leq \frac{1}{n} \sum_{k=1}^{n} \alpha_k \leq \max_{k\in\{1,...,n\}} \alpha_k$$

et en utilisant les résultats passés, nous obtenons l'inégalité demandée. □

Corollaire 1.6.1. *Soit une fonction convexe f, $f : C \longrightarrow R$, définie dans un sous-espace convexe C de l'espace linéaire X, $0 \in C$, $y_i \in X$, $q_j > 0$,*

$j \in \{1,...,n\}$ *avec* $\sum_{i=1}^{n} q_i = 1$. *Si* $y_j - \sum_{l=1}^{n} q_l y_l \in C$ *pour tout* $j \in \{1,...,n\}$, *alors*

$$\begin{aligned} f(0) &\leq \min_{k\in(1,...,n)} \{(1-q_k) f[\frac{q_k}{1-q_k}(\sum_{l=1}^{n} q_l y_j - y_k)] + q_k f(y_k - \sum_{l=1}^{n} q_l y_l)\} \\ &\leq \frac{1}{n}\{\sum_{l=1}^{n}(1-q_k) f[\frac{q_k}{1-q_k}(\sum_{l=1}^{n} q_l y_j - y_k)] + \sum_{l=1}^{n} q_k f(y_k - \sum_{l=1}^{n} q_l y_l)\} \\ &\leq \max_{k\in(1,...,n)} \{(1-q_k) f[\frac{q_k}{1-q_k}(\sum_{l=1}^{n} q_l y_j - y_k)] + q_k f(y_k - \sum_{l=1}^{n} q_l y_l)\} \\ &\leq \sum_{j=1}^{n} q_j f(y_k - \sum_{l=1}^{n} q_l y_l). \end{aligned}$$

En particulier, si $y_j - \frac{1}{n}\sum_{l=1}^{n} y_l \in C$ *pour tout* $j \in \{1,...,n\}$, *alors*

$$\begin{aligned} f(0) &\leq \min_{k\in(1,...,n)} \{(n-1) f[\frac{1}{n-1}(\frac{1}{n}\sum_{l=1}^{n} y_l - y_k)] + f(y_k - \frac{1}{n}\sum_{l=1}^{n} y_l)\} \\ &\leq \frac{1}{n^2}\{(n-1)\sum_{k=1}^{n} f[\frac{1}{n-1}(\frac{1}{n}\sum_{l=1}^{n} y_l - y_k)] + \sum_{l=1}^{n} f(y_k - \frac{1}{n}\sum_{l=1}^{n} y_l)\} \\ &\leq \frac{1}{n}\max_{k\in(1,...,n)} \{(n-1) f[\frac{1}{n-1}(\frac{1}{n}\sum_{l=1}^{n} y_l - y_k)] + f(y_k - \frac{1}{n}\sum_{l=1}^{n} y_l)\} \\ &\leq \sum_{j=1}^{n} f(y_j - \frac{1}{n}\sum_{l=1}^{n} y_l). \end{aligned}$$

Preuve

rappel inégalité de jensen
f une fonction convexe f :$R^n \longrightarrow R$ et X un vecteur aléatoire en $E \subset R^n$ et P sa loi de probabilité,alors,

$$E_p(f(x)) \geq f(E_P(x))$$

si f est strictement convexe et l'inégalité de jensen se vérifie par une égalité alors X est constante c'est à dire qu'il existe un $x^* \in E$ tel que

$$P(x) = \delta(x - x*), \forall x \in E$$

preuve de la propriété

$$\begin{aligned} I(P,Q) &= E_P(\ln \frac{dP(X)}{dQ(X)}) \\ &= E_P(-ln\frac{dQ(X)}{dP(X)}) \end{aligned}$$

or ln est concave, -ln est convexe
Alors,

$$\begin{aligned} I(P,Q) &\geq -\ln E_P(\frac{dQ(X)}{dP(X)}) \\ &\geq -\ln \int \frac{dQ(x)}{dP(x)} dP(x) \\ &\geq -\ln \int dQ(x) \\ &\geq -\ln 1 \\ I(P,Q) &\geq 0 \end{aligned}$$

Supposons $I(P,Q) = 0$

$$-\ln E_P(\frac{dQ(X)}{dP(X}) = 0 \text{ et donc }, -\ln E_P(\frac{dQ(X)}{dP(X)}) = E_P\{-\ln(\frac{dQ(X)}{dP(X)})\}$$

or -ln est convexe donc d'après l'égalité de jensen , on a :

$$\frac{dQ(X)}{dP(X)} = k(\text{constante }) \text{ presque surement}$$

d'où $\frac{dQ(x)}{dP(x)} = k$ pour tout $x \in E$ donc,

$$dQ(x) = k\, dP(x)$$

,en intégrant sur E,on on obtient k=1

$$dQ(x) = dP(x) \, \forall x \in E$$

Donc, les mesures P et Q sont identiques
Si $P = Q \text{alors} I(P,Q) = 0$
la "distance" de kullback n'est pas une métrique ou distance au vrai sens du terme car elle n'est pas symétrie et ne respecte pas l'inégalité triangulaire

1.6.3 Duale de l'information de kullback

un calcul direct du langragien permet de montrer que la divergence de kullback écrite sous la forme :

$$I(h) = \int h \ln(h) dQ \tag{1.12}$$

ou $h = \frac{dP}{dQ}$ admet comme dual convexe la fonction :

$$I^*(\phi) = \ln \int e^{\phi} dQ \tag{1.13}$$

Autrement dit l'information de kullback admet la formulation variationnelle :

$$I(P, Q) = \sup_{\phi}(E_P(\phi) - \ln E_Q(e^{\phi})) \tag{1.14}$$

due à donsker-varadhan.

Proposition 1.6.1. *(comparaison de la diveregnce de kullback avec les autres mesures)*

1. $V^2(P, Q) \leq \frac{1}{2} I(P, Q)$
2. $H(P, Q) \leq 2I(P, Q)$
3. $I(P, Q) \leq \log(1 + \chi^2(P, Q)) \leq \chi^2(P, Q)$

1.6.4 Lien entre divergence de kullback et maximum de vraisemblance

Rappel : Loi forte des grands nombres

soit x_n, $n \geq 1$ une suite de variables aléatoires réelles indépendantes du second ordre

$$\text{si } E(x_n) = \mu_n \text{ avec } \lim \mu_n = \mu$$

$$\text{et} \qquad \sum_{n \geq 1} \frac{var(x_k)}{k^2} < +\infty$$

$$\text{Alors, } \frac{1}{n} \sum_{i=1}^{n} (x_i) \text{ converge presque surement vers } \mu$$

Cas particulier

soit x une variable aléatoire de loi f^*
et soit $(x_1, x_2, ..., x_n)$ un échantillon d'observations de x (variables aléatoires indépendantes de même loi que x)
si $E(x)$ existe

$$\text{Alors, } \frac{1}{n}\sum_{i=1}^{n}(x_i) \text{ converge presque surement vers } E(x)$$

Proposition 1.6.2. soit $X_N = (x_1, x_2, ..., x_N)$ un échantillon d'observations indépendantes et identiquement distribuées de loi f^* *et* $(g_\theta)_{\theta\in\Theta}$ un modèle paramétrique ou Θ désigne l'ensemble des paramètres

$$\begin{aligned} I(f^*, g_\theta) &= E_{f^*}(\ln \frac{f^*(x)}{g_\theta(x)}) \\ &= E_{f^*}(\ln f^*(x)) - E_{f^*}(\ln g_\theta(x)) \text{ ou } x \sim f^* \end{aligned}$$

l'intégrale $E_{f^*}(g_\theta(x))$ n'est pas directement évaluable car f^* est inconnue, nous faisons l'approximation suivante, jusifiée par la loi des grands nombres *:*
$\frac{1}{N}\sum_{i=1}^{n}\ln g_\theta(x_i)$ converge presque surement vers $E_{f^*}(\ln g_\theta(x))$
$\sum_{i=1}^{n}\ln g_\theta(x_i) = \ln\prod_{i=1}^{n} g_\theta(x_i) = \ln g_\theta(X_N)$ *(*logarithme du vraisemblance de l'échantillon par g_θ*)*
ces résultats nous conduisent à dire : pour déterminer la densité g_θ qui rend minimale la distance de kullback on peut déterminer la densité qui rend maximale la vraisemblance de l'échantillon par g_θ d'où l'estimation du maximum de vraisemblance de g_θ que nous noterons $g_{\hat{\theta}}$ minimise la distance de kullback

$$I(f, g_{\hat{\theta}}) \backsimeq E_{f^*}(\ln f^*(x)) - \frac{1}{N} ln g_{\hat{\theta}}(X_N)$$

Chapitre Deux

Les critères de sélections de modèles basés sur la J-divergence symétrique de kullback dans le cas d'un échantillon de petit taille

Le critère d'information de kullback ou KIC est un outil de sélection de modèle , il sert comme un estimateur asymptotiquement sans biais d'une variante de la J-divergence symétrique de kullback entre le modèle candidat et le vrai modèle.Dans cette partie , une correction du biais de KIC est dérivée pour les modèles de régressions linéaires ,KICc, il est particulièrement utile lorsque la taille de l'échantillon est petit.Dans le cas ou le vrai modèle est correctement spécifié c'est à dire il appartient à la famille des modèles candidats, KICc est un estimateur exactement sans biais par contre dans le cas ou le vrai modèle est mal spécifié KICc comme KIC reste toujours biaisé. Par ailleurs, lorsqu'il est appliqué la régression polynomiale et à la modélisation auto-régressives des séries chronologiques , KICc se trouve à estimer l'ordre du modèle avec plus de prévision que tout autre méthode.

2.1 AIC, AICc et KIC

On considère un vecteur de collection de données $X_N = (x_1, x_2, ..., x_N)$,généré selon un modèle paramétrique inconnu donc de loi inconnue $p(x/\theta_0)$.

Notre objectif est de déterminer un modèle paramétrique de loi connue qui est proche de notre modèle inconnu $p(x/\theta_0)$.

Pour se faire,on dispose plusieurs modèles et on choisit parmi eux le meilleur au sens d'une mesure de proximité,plusieurs mesures sont envisageable pour calculer cette proximité comme la divergence de hellinger, la divergence de variation totale ou la divergence de kullback . Dans ce papier nous utiliserons la "distance" kullback.

Soit $M_k = \{p(x/\theta_k), \theta_k \in \Theta_k\}$ une famille de modèles paramétriques et $dim\Theta_k = k + 1$ c'est à dire le vecteur θ_k , est composé par les k paramètres du modèle plus la variance du bruit. on estime le vecteurθ_k par$\hat{\theta}_k$ sur tout Θ_k par la méthode du maximum de vraisemblance de la fonction de vraisemblance $p(X_N/\theta_k)$ et on désigne alors par $p(x/\hat{\theta}_k)$ le modèle ajusté.Pour souci de simplicité nous supposons $k = 1, 2, ...k_{max}$ de sorte que la collection se comporte de famille imbriquée c'est à dire$\Theta_1 \subset \Theta_2.... \subset \Theta_{kmax}$. Pour déterminer le modèle candidat qui est le plus proche du modèle générateur on utilise la divergence de kullback dirrigée entre le modèle candidat ajusté $p(x/\hat{\theta}_k)$et notre modèle générateur$p(x/\theta_0)$,dirigée par ce dernier,multipliée par 2.
On a alors,

$$2I_n(\theta_0, \theta_k) = E_{\theta_0}(2 \ln \frac{p(x/\theta_0)}{p(x/\theta_k)}) \tag{2.1}$$

$2I_n(\theta_0, \theta_k)$= $E_{\theta_0}(-2 \ln p(x/\theta_k)) - E_{\theta_0}(-2 \ln p(x/\theta_0))$

$$2I_n(\theta_0, \theta_k) = d_n(\theta_0, \theta_k) - d_n(\theta_0, \theta_0) \tag{2.2}$$

avec $d_n(\theta_0, \theta_k) = E_{\theta_0}(-2 \ln p(x/\theta_k))$
et $d_n(\theta_0, \theta_0) = E_{\theta_0}(-2 \ln p(x/\theta_0))$

le terme $d_n(\theta_0, \theta_0)$ ne dépend pas de k, il dépend uniquement de θ_0 donc la variation de l'écart entre les deux modèles dépend de la variation du terme $d_n(\theta_0, \theta_k)$,donc ajuster la "distance" de kullback revient à ajuster le terme $d_n(\theta_0, \theta_k)$.
d'où on remplace θ_k par $\hat{\theta}_k$, et on obtient $d_n(\theta_0, \hat{\theta}_k)$
ce terme $d_n(\theta_0, \hat{\theta}_k)$ n'est pas directement évaluable car la densité $p(x/\theta_0)$ n'est pas connue, on le calcule donc par approximation $-2 \ln p(X_N, \hat{\theta}_k)$,

proposée par AKAIKE comme un estimateur biaisé,il en ensuite déduit un estimateur asymptotiquement sans biais :

$$AIC = -2\ln p(X_N, \hat{\theta}_k) + 2(k+1) \tag{2.3}$$

(-2×log de vraisemblance pénalisée par deux fois le nombre de paramétres). En effet, si on pose $\Delta_n(k, \theta_0) = E_{\theta_0}(d_n(\theta_0, \hat{\theta}_k))$ on obtient $\Delta_n(k, \theta_0) = E_{\theta_0}(AIC) + o(1)$ (pour la démonstration voir Seghouane and Amari : AIC criteron and symmetrizing the kullback divergence) Motivé par le fait que AIC est asymptotiquement sans biais,HURVICH propose une version corrigé du biais :

$$AICc = -2\ln p(X_N, \hat{\theta}_k) + \frac{2(k+1)n}{n-k-2} \tag{2.4}$$

comme un estimateur exactement sans biais de $d_n(\theta_0, \hat{\theta}_k)$ dans un contexte de régression linéaire. On a donc : $\Delta_n(k, \theta_0) = E_{\theta_0}(AICc)$, (pour la preuve voir toujours Seghouane and Amari : AIC criteron and symmetrizing the kullback divergence) lorsque la nombre de paramètres ajustés est un modéré à large de la taille de l'échantillon c'est dire lorsque la dimension du modèle candidat devient relativement importante comparée à la taille de l'échantillon ($n/k \leq 40$) , AIC devient une estimateur tros biaisé et conduit à des choix de modèles surparamétrés donc à un surajustement. Il serait alors utile d'utiliser le critére corrigé du biais de AIC c'est à dire AICc car ce dernier évite mieux le sur-apprentissage

J-divergence symétrique de kullback

La J-divergence symétrique de kullback est définie comme :

$$2J_n(\theta_0, \theta_k) = 2I_n(\theta_0, \theta_k) + 2I_n(\theta_k, \theta_0) \tag{2.5}$$

On voit bien que la J-divergence respecte plus les propriétés d'une distance qu'une divergence dirigée car elle symétrique et ce dernier ne l'est pas.La J-divergence apporte alors des informations supplémentaires sur la dissemblance des deux modèles qui nous permettraient de mieux refléter l'erreur due à une sou-ajustement. Nous allons déduire, à partir de la divergence symétrique , une autre critère de sélection que l'on appelle KIC , il supperforme alors AIC et nécessite moins d'apprentissage que AIC pour

la sélection de modèles de régressions linéaires et pour la sélection de modèles autorégressifs.

$$\begin{aligned} 2J_n(\theta_0,\theta_k) &= 2I_n(\theta_0,\theta_k) + 2I_n(\theta_k,\theta_0) \\ &= d_n(\theta_0,\theta_k) - d_n(\theta_0,\theta_0) + d_n(\theta_k,\theta_0) - d_n(\theta_k,\theta_k) \end{aligned}$$

Le terme $d_n(\theta_0,\theta_0)$ est indépendant de k
posons, $K_n(\theta_0,\theta_k) = d_n(\theta_0,\theta_k) + d_n(\theta_k,\theta_0) - d_n(\theta_k,\theta_k)$, on obtient :

$$2J_n(\theta_0,\theta_k) = K_n(\theta_0,\theta_k) + d_n(\theta_0,\theta_0) \tag{2.6}$$

estimer l'écart $2J_n(\theta_0,\theta_k)$, revient alors à estimer $K_n(\theta_0,\theta_k)$.
On substitue θ_k par $\hat{\theta}_k$
On ne peut toujours pas calculer $K_n(\theta_0,\hat{\theta}_k)$ car nécessite une intégration suivant la loi inconnue, CANAVAUG propose alors le KIC comme un estimateur asymptotiquement sans biais de $K_n(\theta_0,\hat{\theta}_k)$

$$KIC = -2\ln p(X_N,\hat{\theta}_k) + 3(k+1) \tag{2.7}$$

c'est dire, si on pose $\Omega_n(k,\theta_0) = E_{\theta_0}(K_n(\theta_0,\hat{\theta}_k)$ on obtient
$\Omega_n(k,\theta_0) = E_{\theta_0}(KIC) + o(1)$

Preuve :
Les résultats issus de l'approximation de taylor des fonctions $I(\theta_k)$ et $F(\theta_k,X)$ à 0(1) sont utilisés dans cette preuve pour approcher l'espérance de $K_n(\theta_0,\hat{\theta}_k)$.

$$I(\theta_k) = E_{\theta_0}(-\frac{\partial^2 lnp(X,\theta_k)}{\partial\theta_k\partial\theta_k'}) \text{ et } F(\theta_k,X) = -\frac{\partial^2 lnp(X,\theta_k)}{\partial\theta_k\partial\theta_k'}$$

$F(\theta_k,X)$ est la matrice d'information de fisher des observations
en élargissant $-2p(x,\theta_0)$ autour de $\hat{\theta}_k$, on obtient
$-2p(x/\theta_0) = -2\ln p(X,\hat{\theta}_k) + (\hat{\theta}_k-\theta_0)'F(\theta_k^*,X)(\hat{\theta}_k-\theta_0) + \epsilon_1$
ou θ_k^* est entre $\hat{\theta}_k$ et θ_0 et ϵ_1 est $o_p(1)$

$$d_n(\theta_0,\theta_0) = E_{\theta_0}\{-2\ln p(X,\hat{\theta}_k)\} + E_{\theta_0}\{(\hat{\theta}_k-\theta_0)'F(\hat{\theta}_k,X)(\hat{\theta}_k-\theta_0)\} + o(1) \tag{2.8}$$

le développement du second ordre de $d_n(\theta_0,\hat{\theta}_k)$ autour de θ_0 et les conditions régulières usuelles nous donnent :

$$d_n(\theta_0,\hat{\theta}_k) = d_n(\theta_0,\theta_0) + (\hat{\theta}_k-\theta_0)'I(\theta_0)(\hat{\theta}_k-\theta_0) + \epsilon_2$$

où ϵ_2 est $o_p(1)$
or $2I_n(\theta_0, \hat{\theta}_k) = d_n(\theta_0, \hat{\theta}_k) - d_n(\theta_0, \theta_0)$
et donc,

$$2I_n(\theta_0, \hat{\theta}_k) = (\hat{\theta}_k - \theta_0)' I(\theta_0)(\hat{\theta}_k - \theta_0) + \epsilon_2 \tag{2.9}$$

En faisant un développement du second ordre de $d_n(\hat{\theta}_k, \theta_0)$ autour de $\hat{\theta}_k$, nous obtenons :

$$d_n(\hat{\theta}_k, \theta_0) = d_n(\hat{\theta}_k, \hat{\theta}_k) + (\hat{\theta}_k - \theta_0)' I(\hat{\theta}_k)(\hat{\theta}_k - \theta_0) + \epsilon_3$$

où $\epsilon_3$est $o_p(1)$
or $2I_n(\hat{\theta}_k, \theta_0) = d_n(\hat{\theta}_k, \theta_0) - d_n(\hat{\theta}_k, \hat{\theta}_k)$
et donc,

$$2I_n(\hat{\theta}_k, \theta_0) = (\hat{\theta}_k - \theta_0)' I(\hat{\theta}_k)(\hat{\theta}_k - \theta_0) + \epsilon_3 \tag{2.10}$$

Les termes $(\hat{\theta}_k - \theta_0)' I(\hat{\theta}_k)(\hat{\theta}_k - \theta_0)$, $(\hat{\theta}_k - \theta_0)' I(\theta_0)(\hat{\theta}_k - \theta_0)$ et $(\hat{\theta}_k - \theta_0)' F(\hat{\theta}_k, X)(\hat{\theta}_k - \theta_0)$ convergent vers la loi de khi-deux à k+1 degrés de liberté , donc chacun a une espérance k+1 au limite $o(1)$
Alors,

$$\begin{aligned}
\Omega_n(k, \theta_0) &= E_{\theta_0}(K_n(\hat{\theta}_0, \theta_k)) \\
&= E_{\theta_0}(2J_n(\theta_0, \hat{\theta}_k)) + d_n(\theta_0, \theta_0) \\
&= E_{\theta_0}(2I_n(\theta_0, \hat{\theta}_k) + 2I_n(\hat{\theta}_k, \theta_0)) + d_n(\theta_0, \theta_0) \\
&= E_{\theta_0}\{(\hat{\theta}_k - \theta_0)' I(\theta_0)(\hat{\theta}_k - \theta_0)\} + E_{\theta_0}\{(\hat{\theta}_k - \theta_0)' I(\hat{\theta}_k)(\hat{\theta}_k - \theta_0)\} \\
&+ E_{\theta_0}\{-2 \ln p(X, \hat{\theta}_k)\} + E_{\theta_0}\{(\hat{\theta}_k - \theta_0)' F(\hat{\theta}_k, X)(\hat{\theta}_k - \theta_0)\} + o(1) \\
&= E_{\theta_0}\{-2lnp(X, \hat{\theta}_k)\} + (k+1) + (k+1) + (k+1) + o(1)
\end{aligned}$$

démonstration close

2.2 KICc

Ici on considère le cas des modèles de régressions linéaires .On génère nos données selon le modèle (a), et on dispose d'un modèle candidat (b)

$$y = X\beta_0 + \varepsilon \ \text{ où } \ \varepsilon \sim N(0, \sigma_0^2 I_n) \qquad \text{(a)}$$

$$y = X\beta_k + \varepsilon \ \text{ où } \ \varepsilon \sim N(0, \sigma_k^2 I_n) \qquad \text{(b)}$$

Où y est un vecteur d'observations à n composantes
Le bruit ε a n composantes

β_k etβ_0 sont des vecteurs de paramètres, chacun à k composantes
X est une matrice d'orde $n \times k$
On pose $\theta_k = (\beta_k, \sigma_k^2)^t$ le vecteur des paramètres du modèle candidat
On suppose que le vrai modèle est bien spécifié ou overfitting c'est à dire le vrai modèle se retrouve dans la famille des modèles candidats c'est dire $\theta_0 \in \Theta_k$

Proposition 2.2.1. *On propose un estimateur exactement sans biais de $K_n(\theta_0, \hat{\theta}_k)$ pour les modèles de régressions linéaires, critère d'information de kullback corrigé :*

$$KICc = -2\ln p(y, \hat{\theta}_k) + \frac{2(k+1)n}{n-k-2} - n\psi((n-k)/2) + n\ln\frac{n}{2} \qquad (2.11)$$

où ψ est le digamma ou la la fonction psi
(sous condition le modèle est bien spécifié)

$$E_{\theta_0}(KICc) = \Omega_n(k, \theta_0) \qquad (2.12)$$

Preuve : on considère les modèles (a) et (b)

$$\begin{aligned} d_n(\theta_i, \theta_j) &= E_{\theta_i}(-2\ln p(y/\theta_j)) \\ &= n\ln 2\pi + n\ln {\sigma^2}_j + n\frac{{\sigma^2}_i}{{\sigma^2}_j} + \frac{1}{{\sigma^2}_i}^t (\beta_i - \beta_j)^t XX(\beta_i - \beta_j) \end{aligned}$$

on a alors :

$$\begin{aligned} K(\theta_0, \theta_k) &= n\ln {\sigma_k}^2 + n\ln 2\pi - n\ln\frac{{\hat{\sigma}_k}^2}{{\hat{\sigma}_0}^2} - n + \frac{1}{{\sigma^2}_0}^t (\hat{\beta}_k - \beta_0)^t XX(\hat{\beta}_k - \beta_0) \\ &+ \frac{1}{{\hat{\sigma}^2}_k}^t (\hat{\beta}_k - \beta_0)^t XX(\hat{\beta}_k - \beta_0) + n\frac{{\hat{\sigma}_k}^2}{{\sigma_0}^2} + n\frac{{\sigma_0}^2}{{\hat{\sigma}_k}^2} \end{aligned}$$

or ,

$$n\frac{{\hat{\sigma}_k}^2}{{\sigma_0}^2} \sim \chi^2_{n-k}$$

si μ suit la loi de khi-deux à α degré de liberté $\frac{1}{\mu}$ suit la même loi à $\frac{1}{\alpha - 2}$ degré de liberté,alors $\frac{{\sigma^2}_0}{n\hat{\sigma^2}_k}$ suit la loi de khi-deux à $\frac{1}{n-k-2}$ degré de

Rapport-gratuit.com
LE NUMERO 1 MONDIAL DU MÉMOIRES

liberté et donc $E_{\theta_0}(n\frac{\hat{\sigma_k}^2}{\sigma_0{}^2}) = n-k$ et$E_{\theta_0}(\frac{\sigma^2{}_0}{n\hat{\sigma^2}_k}) = \frac{1}{n-k-2}$

$$\frac{1}{\sigma^2{}_0}{}^t(\hat{\beta}_k - \beta_0)^t XX(\beta_k \hat{-} \beta_0) \sim \chi^2_k$$

et donc,

$$E_{\theta_0}(\frac{1}{\sigma^2{}_0}{}^t(\hat{\beta}_k - \beta_0)^t XX(\hat{\beta}_k - \beta_0) = k$$

les variables ${}^t(\hat{\beta}_k - \beta_0)^t XX(\hat{\beta}_k - \beta_0)et\hat{\sigma}^2_k$ sont indépendantes et donc :

$$\begin{aligned} E_{\theta_0}(\frac{1}{\hat{\sigma^2}_k}{}^t(\hat{\beta}_k - \beta_0)^t XX(\hat{\beta}_k - \beta_0)) &= nE_{\theta_0}(\frac{\sigma^2{}_0}{n\hat{\sigma}_k})E_{\theta_0}({}^t(\hat{\beta}_k - \beta_0)\frac{{}^tXX}{\sigma^2{}_0}(\hat{\beta}_k - \beta_0) \\ &= \frac{nk}{n-k-2} \end{aligned}$$

$$\begin{aligned} \Omega_n(k,\theta_0) &= E_{\theta_0}(n\ln{\sigma_k}^2 + n\ln 2\pi) + \frac{n^2}{n-k-2} + k \\ &+ n-k+\frac{nk}{n-k-2} - E_{\theta_0}(n\ln\frac{\hat{\sigma^2}_k}{\sigma^2{}_0}) - n \\ &= E_{\theta_0}(n\ln{\sigma_k}^2 + n\ln 2\pi + n) + 2\frac{(k+1)n}{n-k-2} - E_{\theta_0}(n\ln\frac{\hat{\sigma}^2{}_k}{\sigma^2{}_0}) \\ &= E_{\theta_0}(-2\ln p(y,\hat{\theta}_k)) + 2\frac{(k+n)n}{n-k-2} - nE_{\theta_0}(\ln n\frac{\sigma^2{}_k}{\sigma^2{}_0}) + n\ln n \\ &= E_{\theta_0}(-2\ln p(y,\hat{\theta}_k)) + 2\frac{(k+n)n}{n-k-2} - nE_{\theta_0}(\ln n\frac{\hat{\sigma^2}_k}{\sigma^2{}_0}) + n\ln n \end{aligned}$$

or,d'apres koltz,$E_{\theta_0}(\ln n\frac{\sigma^2{}_k}{\sigma^2{}_0}) = \psi(\frac{n-k}{2}) + \ln 2$
donc, $\Omega_n(k,\theta_0) = E_{\theta_0}(-2\ln p(y,\hat{\theta}_k)) + 2\frac{(k+n)n}{n-k-2} - n\psi(\frac{n-k}{2}) + n\ln\frac{n}{2}$
$\Omega_n(k,\theta_0) = E_{\theta_0}(KICc)$ d'ou KICc est un estimateur sans biais de la variante de la divergence de kullback symétrie $K_n(\theta_0,\hat{\theta}_k)$
KIC comme AIC est un estimateur asymptotiquement sans biais, donc pour les petits échantillons ou bien lorsque le nombre de paramètres ajustés est tros grand par rapport à la taille de l'échantillon , KIC serait biaisé et risquerait toujours de créer une surajustement. Donc, il serait toujours utile

d'utiliser le critère corrigé du biais de KIC c'est à dire le critère KICc, pour les petits échantillons , cela permettrait d'améliorer fortement la sélection de modèle de régression linéaire .

2.3 Approximation de KICc

Si nous utilisons un développement du digamma, nous aurons une approximation de KICc notée AKICc
$\psi(\frac{n-k}{2}) \simeq ln\frac{n-k}{2} - \frac{1}{n-k} + o(\frac{1}{(n-k)^2})$
en faisant un développement de taylor d'ordre 2 de $\ln(n - \frac{k}{2})$ on obtient
$\psi(\frac{n-k}{2}) \simeq \ln(\frac{n}{2}) - \frac{k}{n} - \frac{1}{n-k} + o(\frac{1}{(n-k)^2}) + o((\frac{k}{n})^2)$
On a,

$$AKICc = -2\ln p(y, \hat{\theta}_k) + \frac{(k+1)(3n-k-2)}{n-k-2} + \frac{k}{n-k} \qquad (2.13)$$

Chapitre Trois

Le critère AIC et les divergences symétriques de kullback

L'AIC est un critère de sélection basé sur la divergence non symétrique de kullback.Ce manque de symétrique peut être un inconvénient dans les applications.Dans cette partie nous allons ensuite créer d'autres divergences de kullback symétriques et reconstruire d'autres critères de sélections basés sur ces divergences symétriques. Les techniques qui seront utilisées pour la création des divergences symétriques sont la moyenne arithmétique, la moyenne géométrique et la moyenne harmonique.

3.1 Le critère AIC et le mélange moyen pondéré

On considère toujours un vecteur de collection de données
$X_N = (x_1, x_2, ..., x_N)$,généré selon un modèle paramétrique inconnu donc de loi inconnue $p(x/\theta_0)$
Nous essayons de trouver un modèle paramétrique, qui fournit une approximation convenable de $p(x/\theta_0)$
On se donne alors $M_k = p(x/\theta_k), \theta_k \in \Theta_k$une famille de modèles candidats paramétriques, de dimension k ou $\theta_k$un vecteur des paramètres, composé par les k paramètres du modèle,ici y'a pas le paramètre supplémentaire. On désigne par $\hat{\theta}_k$,l'estimation du vecteurs θ_k,obtenu par la méthode du maximum de vraisemblance de la fonction$p(X_N/\theta_k)$, on désigne $p(x/\hat{\theta}_k)$ le modèle ajusté,on suppose une collection de familles $)M_k$ imbriquées

Nous considérons toujours la divergence de kullback dirigée pour mesurer la disparité entre les deux modèles, définie par

$$2I_n(\theta_0, \theta_k) = E_{\theta_0}(2ln\frac{p(x/\theta_0)}{p(x/\theta_k)}),$$

qui est un exemple de ali et silvey chapitre 1
$2I_n(\theta_0, \theta_k) = d_n(\theta_0, \theta_k) - d_n(\theta_0, \theta_0)$ d'après (2.2)
$d_n(\theta_0, \theta_0)$ est indépendant de θ_k, donc indépendant de l'écart entre les deux modèles
x On peut alors dire que : Quelque soit le classement du modèle candidat selon $2I_n(\theta_0, \theta_k)$ serait identique au classement selon $d_n(\theta_0, \theta_k)$donc suivant $D_{1,n}(\theta_0, \theta_k)$ avec

$$D_{1,n}(\theta_0, \theta_k) = 2I_n(\theta_0, \theta_k) + d_n(\theta_0, \theta_0) \tag{3.1}$$

Proposition 3.1.1.

$$AIC = -2 \ln p(X_N, \hat{\theta}_k,) + 2k \tag{3.2}$$

est un estimateur asymptotiquement sans biais de $D_{1,n}(\theta_0, \hat{\theta}_k)$, variante de la divergence de kullback dirrigé entre le modèle inconnu $p(x/\theta_0)$ et le modèle candidat équilibré $p(x/\hat{\theta}_k)$
c'est à dire, si on pose $\Delta(k, \theta_0) = E_{\theta_0}(D_{1,n}(\theta_0, \hat{\theta}_k))$ on obtient
$\Delta(k, \theta_0) = E_{\theta_0}(AIC) + o(1)$

Par ailleurs,la divergence alternative $2I_n(\theta_k, \theta_0)$ fournit des informations supplémentaires , elle nous permet de mieux refléter l'erreur due à une sou-apprentissage, qui peuvent être nécessaire de tenir en compte dans les problèmes de sélections de modèles.Par conséquent,ce qui est nécessaire, c'est une divergence symétrique, qui peut être asymptotiquement estimer pour fournir un critère d'information théorique.
Pour résoudre le problème de symétrique , différentes solutions peuvent être considérées
la moyenne pondérée définie par :

$$M_\eta(\theta_0, \theta_k) = \eta I_n(\theta_0, \theta_k) + (1 - \eta) I_n(\theta_k, \theta_0), \eta \in [0, 1]$$

Un solution simple est de générer un cas particulier de la divergence symétrique. Exemple en prenant $\eta = \frac{1}{2}$ on a le melange moyen arithmétique de divergences de kullback,qui est une divergence symétrique

$$M_n(\theta_0, \theta_k) = \frac{I_n(\theta_0, \theta_k) + I_n(\theta_k, \theta_0)}{2}$$

$M_n(\theta_0, \theta_k)$ est exemple de ali et silvey(1.1) avec $c(x) = \frac{x-1}{2} \ln x$ et $f(x) = x$ (il suffit de remplacer chaque terme par son expression)

$$M_n(\theta_0, \theta_k) = \frac{2I_n(\theta_0, \theta_k) + 2I_n(\theta_k, \theta_0)}{4}$$

$$M_n(\theta_0,\theta_k)=\frac{1}{4}(d_n(\theta_0,\theta_k)-d_n(\theta_0,\theta_0)+d_n(\theta_k,\theta_0)-d_n(\theta_k,\theta_k))$$

$d_n(\theta_0,\theta_0)$ indépendant de k
En analogie avec (3.1) , utiliser l'écart $2M_n(\theta_0,\theta_k)$pour la classement du modèle candidat équivaut à utiliser l'écart suivant :

$$D_{2,n}(\theta_0,\theta_k)=2M_n(\theta_0,\theta_k)+d_n(\theta_0,\theta_0) \tag{3.3}$$

et on a :

$$AIC=-2\ln p(X_N,\hat{\theta}_k)+2k \tag{3.4}$$

comme un estimateur asymptotiquement sans biais de $D_{2,n}(\theta_0\hat{\theta}_k)$, variante de la divergence de kullback symétrique $2M_n(\theta_0,\theta_k)$ qui correspond à une J-divergence ,traité dans le chapitre 2
On a alors, $E_{\theta_0}(D_{2,n}(\theta_0,\hat{\theta}_k))=E_{\theta_0}(AIC)+o(1)$
Pour la preuve, on procéde de manière analogue à (2.7), mais ici on remplace $K_n(\theta_0,\hat{\theta}_k)$ par $D_{2,n}(\theta_0\hat{\theta}_k)$

D'autres formes de mélange pondéré peuvent être imaginées pour créer des divergences symétriques, exemple

$$S_\eta(\theta_0,\theta_k)=\eta I_n(\theta_0,\theta_k)+(2-\eta)I_n(\theta_k,\theta_0)$$

Si $\eta=1$on retrouve la j-divergence symétrique de kullback ;

$$\begin{aligned}S_n(\theta_0,\theta_k) &= 2S_{\{}\eta=1\}(\theta_0,\theta_k)\\ &= 2J_n(\theta_0,\theta_k)\\ &= 2I_n(\theta_0,\theta_k)+2I_n(\theta_k,\theta_0)\end{aligned}$$

$$S_n(\theta_0,\theta_k)=d_n(\theta_0,\theta_k)-d_n(\theta_0,\theta_0)+d_n(\theta_k,\theta_0)-d_n(\theta_k;\theta_k)$$

Le coefficient $S_n(\theta_0,\theta_k)$ est un exemple de ali et silvey (1)dans les classe de mesure de divergence avec $C(x)=(x-1)\log x$ et $f(x)=x$
$d_n(\theta_0,\theta_0)$ ne dépend pas de k
En analogie avec (3.1) utiliser l'écart $2S_n(\theta_0,\theta_k)$ pour choisir le modèle candidat approprié équivaut à utiliser l'écart suivant

$$D^*_{2,n}(\theta_0,\theta_k)=2S_n(\theta_0,\theta_k)+d_n(\theta_0,\theta_0) \tag{3.5}$$

on a alors,

$$KIC=-2\ln p(X_N,\hat{\theta}_k)+3k \tag{3.6}$$

comme un estimateur asymptotiquement sans biais de $D^*_{2,n}(\theta_0, \hat{\theta}_k)$ écart entre le vrai modèle et le modèle candidat ajusté
(résultats déjà obtenus dans le chapitre 2)

3.2 AIC , Moyenne géométrique et Moyenne harmonique

Nous allons à partir des divergences dirigées et des divergences alternatives construire des moyennes géométriques et harmoniques pour obtenir des divergences symétriques. Pour chaque divergence symétrique nous allons en déduire des estimateurs asymptotiquement sans biais.

3.2.1 Moyenne géométrique

La forme générale de la moyenne géométrique entre la divergence dirigée et alternative est :

$$G_\eta(\theta_o, \theta_k) = I_n(\theta_0, \theta_k)^\eta I_n(\theta_k, \theta_o)^{1-\eta}$$

Le cas particulier ou $\eta = \frac{1}{2}$, donne la moyenne géométrique :

$$G_n(\theta_o, \theta_k) = I_n(\theta_0, \theta_k)^{\frac{1}{2}} I_n(\theta_k, \theta_o)^{\frac{1}{2}}$$

$\Rightarrow$

$$\ln G_n(\theta_o, \theta_k) = \frac{1}{2} \ln I_n(\theta_0, \theta_k) + \frac{1}{2} \ln I_n(\theta_k, \theta_0)$$

De la même manière que les sections précédentes, $2d_n(\theta_0, \theta_0)$ ne dépend pas e k
En analogie avec (3.1),utiliser l'écart $2G_n(\theta_o, \theta_k)$ pour choisir le bon modèle équivaut à utiliser l'écart suivant :

$$D_{3,n}(\theta_0, \theta_k) = 2G_n(\theta_o, \theta_k) + d_n(\theta_0, \theta_0) \tag{3.7}$$

on a,alors :

$$AIC = -2 \ln p(X_N, \hat{\theta}_k) + 2k \tag{3.8}$$

comme un estimateur asymptotiquement sans biais de$D_{3,n}(\theta_0, \hat{\theta}_k)$, c'est à dire : $E_{\theta_0}(D_{3,n}(\theta_0, \hat{\theta}_k)) = E_{\theta_0}(AIC) + O(n^{-\frac{1}{2}})$

Preuve : On considère toujours la même démarche que (2.7), et on a :

$$\begin{aligned} E_{\theta_0}(D_{3,n}(\theta_0,\hat{\theta}_k)) &= E_{\theta_0}\{2G_n(\theta_o,\hat{\theta_k})\} + d_n(\theta_0,\theta_0) \\ &= E_{\theta_0}\{\sqrt{2I_n(\theta_0,\hat{\theta_k})2I_n(\theta_k,\hat{\theta_0})}\} + d_n(\theta_0,\theta_0) \\ &= E_{\theta_0}\{\sqrt{(\hat{\theta}_k-\theta_0)'I(\hat{\theta}_0)(\hat{\theta}_k-\theta_0)+\epsilon_2}.\sqrt{(\hat{\theta}_k-\theta_0)'I(\hat{\theta}_k)(\hat{\theta}_k-\theta_0)+\epsilon_3}\} \\ &+ E_{\theta_0}(-2\ln p(y,\hat{\theta}_k)) + E_{\theta_0}\{(\hat{\theta}_k-\theta_0)^{'}F(\hat{\theta}_k,y)(\hat{\theta}_k-\theta_0)\} + o(1) \end{aligned}$$

En faisant le développement de taylor de $I(\hat{\theta}_k)$ au voisinage de $\theta_0$on obtient :

$$\begin{aligned} I(\hat{\theta}_k) &\backsimeq I(\theta_0) + \frac{\partial I(\theta_0)}{\partial(\theta_0)}(\hat{\theta}_k-\theta_0)^{'} \\ &= I(\theta_0) + I^{'}(\theta_0)(\hat{\theta}_k-\theta_0)^{'} \end{aligned}$$

En utilisant le développement $(1+x)^{\frac{1}{2}} \backsimeq 1+\frac{x}{2}$pour les petit x, on obtient :

$$\begin{aligned} \sqrt{\hat{\theta}_k-\theta_0)'I(\theta_0)(\hat{\theta}_k-\theta_0).(\hat{\theta}_k-\theta_0)'I(\hat{\theta}_k)(\hat{\theta}_k-\theta_0)} &\backsimeq \{(\hat{\theta}_k-\theta_0)^{'}I(\theta_0)(\hat{\theta}_k-\theta_0).(\hat{\theta}_k-\theta_0)^{'} \\ &\times \{I(\theta_0)+I(\theta_0)^{'}\hat{\theta}_k-\theta_0)^{'}\}(\hat{\theta}_k-\theta_0)^{'}\}^{\frac{1}{2}} \\ &\backsimeq (\hat{\theta}_k-\theta_0)^{'}I(\theta_0)(\hat{\theta}_k-\theta_0).\sqrt{1+\epsilon_4} \\ &\backsimeq (\hat{\theta}_k-\theta_0)^{'}I(\theta_0)(\hat{\theta}_k-\theta_0)+\epsilon_5 \end{aligned}$$

ou ϵ_4 et ϵ_5 sont des $O_p(n^{-\frac{1}{2}})$, on peut alors écrire :

$$\begin{aligned} E_{\theta_0}(D_{3,n}(\theta_0,\hat{\theta}_k)) &\backsimeq E_{\theta_0}\{(\hat{\theta}_k-\theta_0)^{'}I(\theta_0)(\hat{\theta}_k-\theta_0)\} + O(n^{-\frac{1}{2}}) + E_{\theta_0}(-2lnp(y,\hat{\theta}_k) \\ &+ E_{\theta_0}\{(\hat{\theta}_k-\theta_0)^{'}F(\hat{\theta}_k,y)(\hat{\theta}_k-\theta_0)\} \\ &= E_{\theta_0}(-2lnp(y,\hat{\theta}_k)+2k+O(n^{-\frac{1}{2}}) \end{aligned}$$

démonstration close !

3.2.2 Moyenne harmonique

La moyenne harmonique entre divergence dirigée et alternative est définie par :

$$H_n(\theta_0,\theta_k) = \frac{2}{\frac{1}{I_n(\theta_0,\theta_k)}+\frac{1}{I_n(\theta_k,\theta_0)}}$$

en analogie avec (3.1) utiliser l'écart $H_n(\theta_0,\theta_k)$pour choisir le bon modèle équivaut à utiliser l'écart suivant ;

$$D_{4,n}(\theta_0,\theta_k) = 2H_n(\theta_o,\theta_k) + d_n(\theta_0,\theta_0) \tag{3.9}$$

On ajuste la distance $D_{4,n}(\theta_0, \theta_k)$par $D_{4,n}(\theta_0, \hat{\theta}_k)$ Il n'est pas possible d'évaluer directement $D_{4,n}(\theta_0, \hat{\theta}_k)$ car θ_0 est inconnu on ne peut faire que l'estimer
le critère AIC

$$AIC = -2 \ln p(X_N, \hat{\theta}_k) + 2k \tag{3.10}$$

satisfait également,
$E_{\theta_0}(D_{4,n}(\theta_0, \hat{\theta}_k)) = E_{\theta_0}(AIC) + O(n^{-\frac{1}{2}})$
Pour la preuve je vous renvoie à l'article de Abd-Krim Seghouane, Member, IEEE, and Shun-Ichi Amari, Life Fellow, IEEE
Les moyennes géométrique et harmoniques ne sont pas des formes de ali et silvey mais En raison de leur lien direct avec la distance de kullback,ils vérifient les deux premières propriétés d'une distance(positivité, s'annule si les distributions sont identiques) de plus ils sont symétrique, ils sont donc comme des j-divergences

3.3 Rapport entre les différentes divergences symétriques de kullback

Rappel α-moyenne

La α-moyenne entre deux nombres x et $y$positifs est définie par :

$$m_\alpha(x; y) = t_\alpha[x^{\frac{1-\alpha}{2}} + y^{\frac{1-\alpha}{2}}]^{\frac{2}{1-\alpha}}$$

ou t_α est une constante qui permet de vérifier l'égalité $m_\alpha(x, x) = x$

Nous vous présentons le α représentation d'un nombre u par

$$l_\alpha(u) = \frac{2}{1-\alpha} u^{\frac{1-\alpha}{2}}, \text{si } \alpha \neq 1 \quad l_\alpha(u) = \log u, \text{sinon}$$

Proposition 3.3.1. *le α -représentation du α -moyenne de deux nombres est égal à la moyenne arithmétique des α-représentations de ces deux nombres. c'est à dire :*

$$l_\alpha(m_\alpha(x, y)) = \frac{1}{2}[l_\alpha(x) + l_\alpha(y)]$$

A partir de ces relations, on peut déduire :
si $\alpha = 3$, $m_3(x,y) = \frac{2}{\frac{1}{x}+\frac{1}{y}}$
si $\alpha = 1$, $m_1(x,y) = \sqrt{xy}$
si $\alpha = -1$, $m_{-1}(x,y) = \frac{1}{2}(x+y)$
si $\alpha = -\infty$, $m_{-\infty}(x;y) = max(x,y)$
si$\alpha = +\infty$, $m_{+\infty}(x,y) = min(x,y)$
le α-moyenne est décroissante suivant α, c'est à dire si $\alpha \geq \alpha^{'}$ alors
$m_\alpha(x,y) \leq m_{\alpha'}(x,y)$
Par conséquent,
$max[I_n(\theta_0,\theta_k), I_n(\theta_k,\theta_0)] \geq M_n(\theta_0,\theta_k) \geq G_n(\theta_0,\theta_k)$
et
$min[I_n(\theta_0,\theta_k), I_n(\theta_k,\theta_0)] \leq H_n(\theta_0,\theta_k) \leq G_n(\theta_0,\theta_k)$

3.4 Exemple de nouvelle correction du biais de AIC

Comme dans le chapitre 2. Ici,pour chaque fonction de classement, une correction du biais de l'AIC existe. Exemple pour l'AIC de la fonction de classement $D_{1,n}(\theta_0,\hat{\theta}_k)$ la version corrigée du biais est :

$$AICc = -2\ln p(X_N,\hat{\theta}_k) + \frac{(2(k+1)n)}{n-k-2}$$

comme un estimateur sans biais de $D_{1,n}(\theta_0,\hat{\theta}_k)$.
Et dans le cas des modèles de régressions linéaires univariées la version corrigé de l'AIC de la fonction de classement $D_{2,n}(\theta_0,\hat{\theta}_k)$ est :

$$AIC^*c = -2\ln p(y,\hat{\theta}_k) + \frac{(k+1)n}{n-k-2} - n\psi((n-k)/2) + nln\frac{n}{2}$$

est un estimateur sans biais de $D_{2,n}(\theta_0,\hat{\theta}_k)$ sous hypothèse que le vrai modèle est bien spécifié
Preuve(même raisonnement que le chapitre 2 pour le KICc)
Si nous remplaçons la fonction digamma par un développement , nous obtenons une approximation AIC^*c :

$$AIC^*c \simeq -2\ln p(y,\hat{\theta}_k) + \frac{(k+1)(2n-k-2)}{n-k-2} + \frac{k}{n-k}$$

On note que asymptotiquement AIC^*c converge vers AIC

Chapitre Quatre

Applications 1 : Évaluations numériques

4.1 Régression polynomiale

Pour montrer l'efficacité du critère KICc par rapport aux autres critères, nous considérons le problème classique de classement des fonctions polynomiales ajustées dans un ensemble de points sur un support compact.La motivation de cet exemple d'étude et que les polynômes créent un difficile problème de sélection de modèles qui tend à produire une sur-ajustement catastrophique, surtout lorsque la taille des données est faible.En outre,les polynômes constituent une classe intéressante de modèles linéaires univariées pour lesquels il existe des techniques efficaces pour calculer le meilleur ajustement .
On considère quatre ensembles de simulations basées sur deux différentes tailles de l'échantillon (n=25 et n=30)et 2 niveaux de bruits (SNR = 0 dB et SNR = 10 dB) On génère les données selon un modèle polynomiale d'ordre 3 avec $\beta = [1, -0.5, -5, -1.5]^T$
Les points de conceptions sont également espacés sur l'intervalle $[-x_{max}, x_{max}]$ avec $x_{max} = 3$.Données sont réalisées par adjonction de variables aléatoires gaussiennes iid de moyenne nulle et de variance ${\sigma^2}_0$ ajusté pour avoir la valeur requise de $SNR = 10log_{10}(\frac{{\sigma^2}_s}{{\sigma^2}_0})$ avec ${\sigma^2}_s = \frac{1}{2x_{max}} \int_{-x_{max}}^{x_{max}} f^2(x)dx$
on a mené une simulation de Monté Carlo de 1000 pistes pour la réalisation de chaque ensemble de données , on a aussi utilisé un algorithme des moindres carrés pour ajuster les degrés des polynômes candidats de 0 à 10. On a 4.1 et 4.2 suivants : :

n	SNR	ordre	AIC	AICc	FPE	BIC	KIC	KICc
25	0 dB	$< k_0$	0	2	0	4	2	19
25	0 dB	$= k_0$	513	855	550	784	751	**914**
25	0 dB	$> k_0$	487	143	450	212	247	67
25	10 dB	$< k_0$	0	0	0	0	0	0
25	10 dB	$= k_0$	512	856	555	782	748	**929**
25	10 dB	$> k_0$	488	144	445	218	252	71
30	0 dB	$< k_0$	0	0	0	2	1	5
30	0 dB	$= k_0$	558	839	587	837	789	**924**
30	0 dB	$> k_0$	442	160	413	161	210	71
30	10 dB	$< k_0$	0	0	0	0	0	0
30	10 dB	$= k_0$	557	834	581	441	786	**922**
30	10 dB	$> k_0$	443	166	419	159	214	78

TABLE 4.1 – Fréquence de sélection de l'ordre du modèle pour chaque critère pour 1000 simulations

les résultats du tableau 4.1 montre pour chaque niveau de bruit et pour chaque taille d'échantillon la fréquence de sélection du vrai ordre pour le critère KICc dans les 1000 simulations est supérieure aux fréquences des autres critères. Cette supériorité montre KICc est plus performant que le AIC , AICc , BIC , FPE et KIC. Cette performance est due a deux facteurs : le premier est le fait KICc est basé sur la J-divergence symétrique ,qui apporte des informations supplémentaires sur la dissemblance des deux modèles,le deuxième est le fait que KICc est un estimateur exactement sans biais de la J-divergence au lieu d'être asymptotiquement sans biais comme le KIC pour les petit échantillons

n	SNR	Ordre	AKICc	KICc
25	0 dB	$< k_0$	16	19
25	0 dB	$= k_0$	**908**	**914**
25	0 dB	$> k_0$	76	67
25	10 dB	$< k_0$	0	0
25	10 dB	$= k_0$	**919**	**929**
25	10 dB	$> k_0$	81	71
30	0 dB	$< k_0$	4	5
30	0 dB	$= k_0$	**910**	**924**
30	0 dB	$> k_0$	86	71
25	10 dB	$< k_0$	0	0
25	10 dB	$= k_0$	**918**	**922**
25	10 dB	$> k_0$	82	78

TABLE 4.2 – Fréquences de sélection du vrai ordre du modèle pour les critères AKICc et KICc pour 1000 réalisations

Ces résultats montrent que les fréquences de sélection de l'ordre du modèle sont approximativement égales pour les critères AKICc et KICc

4.2 Modélisations Autorégressives

Rappel processus AR(k) On appelle processus autorégressif d'ordre k noté AR(k) ,tout processus, qui est généré par l'équation :

$$y_t = \theta_1 y_{t-1} + \theta_2 y_{t-2} + ... + \theta_k y_{t-k} + \epsilon_t \tag{4.1}$$

où ϵ_t est un bruit blanc (variables gaussiennes iid de moyenne nulle et de variance ${\sigma^2}_0$)

On se donne un ensemble d'observation $Y_n = (y_1, y_2, ..., y_n)$ à partir d'un tel processus.Notre objectif est de déterminer un ordre approprié k de ce modèle autorégressif
Nous mettons (4.1) sous forme linéaire

$$Y = X\theta + \xi \tag{4.2}$$

où $\theta = (\theta_1, \theta_2, ..., \theta_k)^T$, $\xi = (\epsilon_1, \epsilon_2, ..., \epsilon_n)^T$ et X est une matrice d'ordre $n \times k$

$$X = \begin{bmatrix} y_0 & y_{-1} & \cdots & y_{1-k} \\ y_1 & y_0 & \cdots & y_{2-k} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ y_{n-1} & y_{n-2} & \cdots & y_{n-k} \end{bmatrix}$$

Cette forme linéaire du modèle autorégressif et le comportement asymptotique des estimations de yull waker nous permettent de déduire que KICc comme un estimateur approximativement sans biais de la divergence symétrique de kullback pour la classe des modèles AR.
Nous identifions un modèle autorégressif par deux tailles d'échantillons : n=23 et n=30
Une simulation de monté carlo pour 1000 réalisation de l'ensemble des données de tailles n ont été générés à partir des deux modèles
Modèle 1 : $y_t = 0.95y_{t-1} + \epsilon_t$
Modèle 2 $y_t = 0.99y_{t-1} - 0.8y_{t-2} + \epsilon_t$
$t = 1, ..., n$
Pour chaque ensemble de données, la méthode de levirson derbing a été utilisée pour ajuster les modèles candidats autorégressifs d' ordre de 1 à 20.Ainsi, nous obtenons les tableaux 3 et 4 suivants
Ces tableaux donnent des résultats similaires à ceux obtenus dans le cas de la sélection du modèle polynomiale .KICc montre toujours une bonne performance en comparaison aux autres critères . On a aussi l'approximation, AKICc, qui donnent des résultats très bonnes et très proches de ceux obtenus en utilisant KICc.

4.3 Les prévisions dans les taux de changes

Processus stationnaire

Un processus $(x_t, t \in Z)$ est dit stationnaire de second ordre si les trois conditions suivantes sont vérifiés
$\forall t \in Z, E(x_t^2) < \infty$ (convergence des moments d'ordre 2)
$\forall x \in Z, E(x_t) = m$ (même espérance quelque soit la date t)
$\forall (t, h) \in Z, cov(x_t, x_{t+h}) = \gamma(h)$ (indépendant du temps)

rang	M	n	ordre	AIC	AICc	FPE	BIC	KIC	KICc
1	1	23	$< k_0$	0	0	0	0	0	0
1	1	23	$= k_0$	863	932	867	949	**944**	**970**
1	1	23	$> k_0$	137	68	133	51	56	30
2	2	30	$< k_0$	0	0	0	0	0	0
2	2	30	$= k_0$	835	895	837	952	**925**	**965**
2	2	30	$> k_0$	165	105	163	48	75	35
3	1	23	$< k_0$	25	36	25	50	45	72
3	1	23	$= k_0$	820	899	824	897	**890**	**901**
3	1	23	$> k_0$	155	65	151		53	65
4	2	30	$< k_0$	4	4	4	6	6	7
4	2	30	$= k_0$	827	908	829	950	**926**	**964**
4	2	30	$> k_0$	169	88	167	44	68	29

TABLE 4.3 – fréquences de sélection de l'ordre du modèle par chaque critère pour 1000 réalisations

rang	M	n	ordre	AKICc	KICc
1	1	23	$< k_0$	0	0
1	1	23	$= k_0$	970	972
1	1	23	$> k_0$	30	28
2	1	30	$< k_0$	0	0
2	1	30	$= k_0$	962	965
2	1	30	$> k_0$	38	35
3	2	23	$< k_0$	71	72
3	2	23	$= k_0$	903	901
3	2	23	$> k_0$	26	25
4	2	30	$< k_0$	7	7
4	2	30	$= k_0$	961	964
4	2	30	$> k_0$	32	29

TABLE 4.4 – fréquences de sélection de l'ordre du modèle pour les critères AKICc et KICc pour les 1000 réalisations

Processus MA(p)

On appelle processus Moyenne mobile d'ordre p noté MA(p) tout processus $(x_t, t \in Z)$ qui est modélisé par la forme ;
$x_t = \varepsilon_t + \varphi_1\varepsilon_{t-1} + \varphi_1\varepsilon_{t-2} + ... + \varphi_k\varepsilon_{t-k}$
ou ε_t est une suite de variable aléatoire gaussienne iid centré réduite.le processus MA(p) est stationnaire

Processus ARMA(p,q)

On appelle processus autorégressif moyenne mobile d'ordre p et q ARMA(p,q) tout processus $(x_t, t \in Z)$répondant à la modélisation suivante ;
$x_t = \theta_1 x_{t-1} + \theta_2 x_{t-2} + ... + \theta_k x_{t-k} + \varepsilon_t + \varphi_1\varepsilon_{t-1} + \varphi_1\varepsilon_{t-2} + ... + \varphi_k\varepsilon_{t-k}$
Un modèle ARMA(p,q) est un jauge de modèle AR(p) et de modèle MA(q). sous certaines condition le processus ARMA est stationnaire.

Processus ARIMA(p,1,q)

On dira qu'un processus non-stationnaire x_t est intégré d'ordre 1, si en le différentiant une fois on obtient un processus stationnaire c'est à dire$y_t = x_t - x_{t-1}$ est stationnaire.
Le processus obtenu en différentiant une seul fois un processus ARMA(p,q) non stationnaire est appelé processus ARIMA(p,1q) ou processus autorégressif intégré d'ordre 1 de la moyenne mobile.La différenciation permet de retirer les tendances qui sont les signes les plus claires du non-stationnarité. Donc, un processus intégré a plus de chance d'être stationnaire.

Prévision des taux de changes

la Prévision des marchés financiers est un problème de prédiction intéressant pour lequel un grand nombre de modèles et de techniques de prévisions ont été spécialement développés. La classe d' autorégressif intégrée de la moyenne mobile (ARIMA) a été l'un des modèles de ceux au début introduite par Box-Jenkins dans le but de la prévision des séries chronologiques de données
on considère le processus ARIMA(p,1,q) pour étudier la performance du KICc sur les autres critères dans le problème de prévision du taux de

	AIC	AICc	EPF	BIC	KIC	KICc
MSE $\times 10$	1.403	1.365	1.368	1.456	1.456	**1.281**
NOS	163	165	165	159	159	**173**

TABLE 4.5 – Erreur quadratique moyenne de prédiction MSE et le nombre de sélections optimal NOS pour chaque critère

change de dollar en starling . La simulation donne le tableaux 5
Pour chaque ensemble de donner l'erreur de prédiction est calculer ;
$\varepsilon^2(m) = (x(mT) - \hat{x}^2(mT))$
où $T = 22 chantillons m = 1, ..., \frac{5192}{T}$
et $\hat{x}(mT) = x(mT-1) + \hat{y}(mT-1)$ la prévision en mT
le tableau 4.5 donne l'estimation de l'erreur quadratique moyenne de prédiction MSE pour chaque critère et l'estimation du nombre de sélection optimale NOS c'est à dire le nombre de fois qu'un critère donné produit la plus petite erreur de prédiction.
Ce tableau donne des valeurs MSE petit et des valeurs NOS grand pour KICc comparées aux valeurs des autres critères. Donc la méthode KICc évite mieux les erreurs de prédiction en comparaison aux méthodes classiques d'où il est le meilleur critères pour les problèmes prévisions.

Discussion

Les résultats montrent que KICc fonctionne comme un critère de sélection de modèle efficace dans les applications des petits échantillons . La bonne performance de KICc peut être attribuée à deux facteurs. Le premier est que KICc basé sur la J-divergence (2.5) , qui fournit des informations supplémentaires sur la dissemblance des modèles en comparaison à la divergence dirigée I,le deuxième est que KICc est un estimateur exactement sans biais(2.11) au lieu d'être asymptotiquement sans biais comme le critère KIC. tout Cela donne le critère KICc les meilleurs résultats en comparaison avec le critère KIC dans le cas des petits échantillons
Les approximations numériques faites dans le calcul de AKICc sont très raisonnables, tel que testés par des simulations. Bien que AKICc est légèrement biaisée, sa simplicité de calcul et sa performance presque égale à celle de KICc justifient l'utilisation d'une telle approximation.

Chapitre Cinq

Applications 2 : Évaluations numériques : Application à la sélection de l'ordre des modèles polynomiales

Dans cette application, nous proposons une comparaison des deux versions corrigées univariées de AIC c'est à dire AICc et AIC^*c étudiées dans le chapitre 3 et nous examinerons l'efficacité de AIC dans l'estimation des différentes fonctions de classements étudiées dans cet chapitre. Pour cela nous considérons toujours le problème classique de fonctions polynomiales ajustées dans un ensemble de point à support compact
On considère deux polynômes d'ordre 3 et 6 de paramétres
$\beta = [1, -0.5, -5, -1.5]^T$ et $\beta = [1, -0.5, -5, -1.5, 1, 2, -4]^T$, sont ulisés pour générer 8 ensembles de données.
Les tailles N des échantillons utilisées pour chaque polynômes sont 15 et 20 les bruits de niveaux 0db et 10db sont utilisées pour le polynôme d'ordre 3, et les bruits de niveaux 10db et 15db pour le polynôme d'ordre 6.
Les points de conceptions sont également répartis sur l'intervalle $[-x_{max}, +x_{max}]$ avec $x_{max} = 3$
.Données sont réalisées par adjonction de variables aléatoires gaussiennes iid de moyenne nulle et de variance ${\sigma^2}_0$ ajusté pour avoir la valeur requise de $SNR = 10log_{10}(\frac{{\sigma^2}_s}{{\sigma^2}_0})$ avec ${\sigma^2}_s = \frac{1}{2x_{max}} \int_{-x_{max}}^{x_{max}} f^2(x)dx$
Une simulation de Monte-Carlo constitué de 1000 pistes a été menée. Pour chaque réalisation de l'ensemble des données, un algorithme des moindres carrés a été utilisé pour ajuster les polynômes candidats de degré 0 à 10.
On obtient tableau suivant, qui donne une comparaison de AIC, AICc, AIC^*c,KIC et EPF

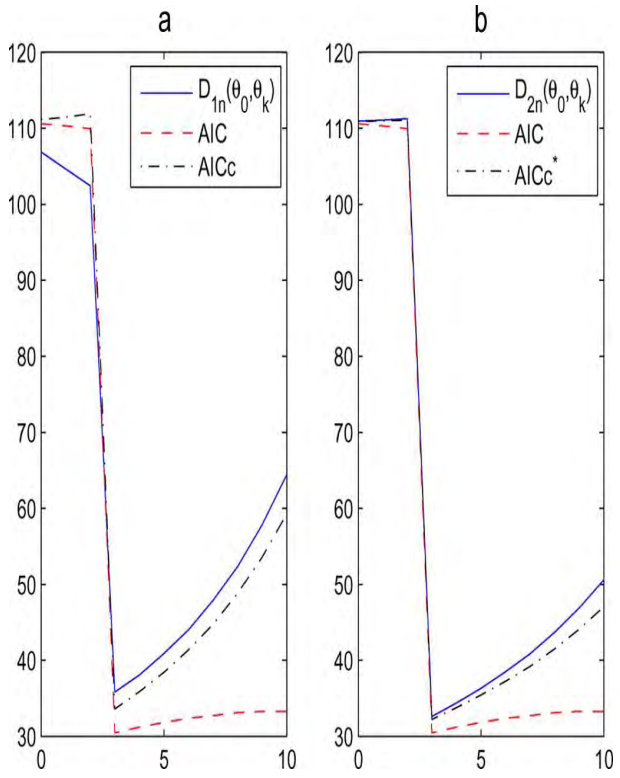
ordre du polynôme	SNR	N	ordre	AIC	AICc	AIC^*c	KIC	FPE	BIC
3	0	15	$<k_0$	12	109	70	45	15	38
3	0	15	$=k_0$	246	**800**	**749**	453	331	391
3	0	15	$>k_0$	742	91	181	502	654	571
3	10	15	$< k_0$	0	0	0	0	0	0
3	10	15	$= k_0$	264	**888**	**804**	512	359	452
3	10	15	$> k_0$	736	112	196	488	641	548
3	0	20	$< k_0$	7	35	23	33	7	33
3	0	20	$=k_0$	422	**791**	**725**	638	471	636
3	0	20	$> k_0$	571	174	252	329	522	331
3	10	20	$< k_0$	0	0	0	0	0	0
3	10	20	$= k_0$	429	**851**	**782**	692	473	690
3	10	20	$>k_0$	571	149	218	308	527	310
3	10	30	$< k_0$	0	0	0	0	0	0
3	10	30	$= k_0$	549	**819**	**766**	799	566	860
3	10	30	$>k_0$	451	181	234	201	434	140
3	10	100	$<k_0$	0	0	0	0	0	0
3	10	100	$= k_0$	688	**768**	**747**	871	690	958
3	10	100	$> k_0$	312	232	253	129	310	42
6	10	15	$< k_0$	54	683	421	152	87	120
6	10	15	$= k_0$	213	**284**	**461**	306	300	280
6	10	15	$> k_0$	733	33	118	542	613	600
6	15	15	$< k_0$	0	177	42	6	1	1
6	15	15	$=k_0$	253	**781**	**834**	415	377	378
6	15	15	$>k_0$	747	42	124	579	622	621
6	10	20	$< k_0$	70	371	211	156	88	156
6	10	20	$= k_0$	385	**559**	**606**	512	455	512
6	10	20	$> k_0$	545	70	183	332	457	332
6	15	20	$< k_0$	36	266	137	96	41	96
6	15	20	$= k_0$	399	**637**	**647**	535	460	535
6	15	20	$>k_0$	565	97	216	369	499	369

TABLE 5.1 – fréquences de selection de l'ordre du modèle par chaque critère pour 1000 réalisations

on remarque pour les quatre premiers échantillons (données) de tailles inférieures à 30 ,générés à partir de polynôme d'ordre 3, c'est AICc qui présente les plus grandes fréquences de sélection de l'ordre du modèle suivi de AIC^*c, pour le cinquième et le sixième échantillon qui sont de tailles 30 et 100 respectivement, c'est le critère BIC qui donne les plus grandes fréquences suivi du AICc ou du KIC,pour les quatre derniers échantillons générés à partir du polynôme d'ordre 6 c'est le critère AIC^*c qui présente les plus grandes fréquences suivi de AICc. Donc, on peut dire , la performance des critères corrigés de AIC c'est AICc et AIC^*c sur les autres critères dépend de l'ensemble des données. La performance de l'un des deux critères corrigés sur l'autre dépend aussi de l'ensemble des données mais aussi elle est lié sur le comportement des fonctions de classement D_1 et D_2 dont ils se dérivent. Nous pouvons remarquer à partir des simulations que AIC^*c sou-ajuste moins que AICc lorsque le polynôme est d'ordre 6, ceci en raison de la présence d'un terme supplémentaire dans D_2 ,qui mesure la dissemblance des modèles d'une manière plus sensible au sou-ajustement. C'est pourquoi AIC^*c est mieux performant que AICc dans ce cas particulier où l'ordre du polynôme est égal à 6.Donc en fonction de la situation, il y'aura des fonctions de classement qui seront plus adaptées que d'autres dans la dérivation des critères appropriés de sélection de modèle

la figure1 donne une aperçue des raisons pour lesquelles l'AICc et AIC^*c est mieux performant l'AIC dans la sélection du bon ordre du vrai modèle.on considère les données du 4ième ensemble de simulation du tableau précédent. Les valeurs simulées sont : $E_{\theta_0}(AICc)$, $E_{\theta_0}(D_{1,n}(\theta_0, \hat{\theta}_k))$, $E_{\theta_0}(AIC^*c)$, $E_{\theta_0}(D_{2,n}(\theta_0, \hat{\theta}_k))$ et $E_{\theta_0}(AIC)$ sont obtenues en faisant la moyenne de $AICc$, $D_{1,n}(\theta_0, \hat{\theta}_k)$, AIC^*c , $D_{2,n}(\theta_0, \hat{\theta}_k)$ et AIC respectivement au cours des 1000 réalisations.

Figure1 : Le polynôme est d'ordre 3,la taille de l'échantillon est égale à 20 et SNR=10db :
Fig(a) $E_{\theta_0}(AICc)$, $E_{\theta_0}(AICc)$ et $E_{\theta_0}(D_{1,n}(\theta_0, \hat{\theta}_k))$
Fig(b) $E_{\theta_0}(AIC^*c)$, $E_{\theta_0}(D_{2,n}(\theta_0, \hat{\theta}_k))$ et $E_{\theta_0}(AIC)$
et l'axe des abscisses représente l'ordre k du modèle



la fig(a) montre que $E_{\theta_0}(AICc)$ à est plus proche de $E_{\theta_0}(D_{1,n}(\theta_0, \hat{\theta}_k))$ en comparaison avec $E_{\theta_0}(AIC)$ surtout lorsque $k > 3$. ceci illustre clairement le faible biais de l'AICc par rapport à celui de l'AIC, AICc est mieux que AIC dans l'estimation de la divergence dirigée ajustée $D_{1,n}(\theta_0, \hat{\theta}_k)$ surtout lorque la taille de l'ensemble des données de simulations est petite et l'ordre du modèle augmente .d'ou AICc plus performant que AIC dans ce cas particulier.La fgure (b) donne les mêmes interprétations que celle de (a),elle approuve que AIC^*c plus performant toujours AIC pour les petits

échantillons même lorsque la fonction de classement qu'ils estiment est symétrique.

Les figures 3 et 4 suivantes sont obtenues par le 5ième et 6ième ensembles de simulation du tableau6 respectivement
La figure 3 :polynôme d'odre 3 avec n=30 et SNR=10db, l'axe des abscisses représente l'ordre k du modèle
la figure 4 : polynôme d'ordre 3 avec n=100 et SNR=10db . l'axe des abscisses représente l'ordre du modèle

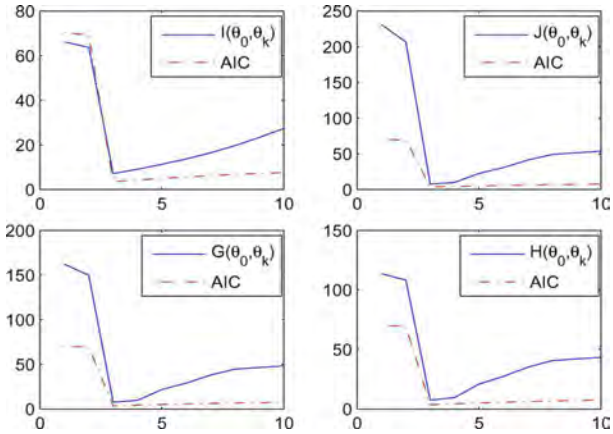


FIGURE3

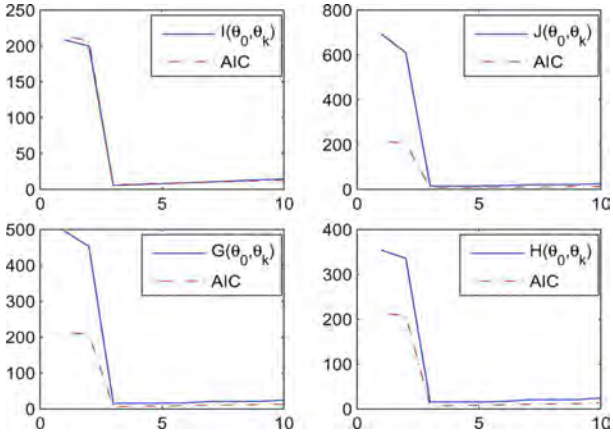


FIGURE4

Ces figures sont les courbes représentatives des moyennes des fonctions de classement D_1 ,D_2, D_3 et D_4 et à chaque courbe on a représenté également également à coté la courbe de la fonction moyenne de l'AIC correspondant.

Nous observons sur ces figures que toutes les fonctions de divergences ou fonctions de classement sont réduit à leurs minimums lorsque k=3.
La figure3 montre un écart entre les deux courbes, qui devient de plus en plus grand lorsque k augmente.Ceci illustre, pour les petits n, l' AIC est un estimateur négativement biaisé de la variante de ces différentes divergences symétriques ou asymétriques, surtout lorsque l'ordre k du modèle est grand.
La figure 4 :On constate que la taille des échatillons est grands et y'a un faible écart entre les deux courbes. Ceci illustre pour les grands échantillons $n \geq 100$, l'AIC est un bon estimateur des variantes de ces divergences symétriques ou asymétriques.

discussions

Dans le cas où l'échantillon est de taille petite et les modèles candidats se composent de familles trop paramétrées les critères AIC peuvent présenter des biais négatifs dans l'estimation des variantes des différentes fonctions classement proposées,une correction du biais donnerait des critères de sélection corrigées qui sont mieux performant que AIC, qui sont même mieux performants que les autres critère considérés dans le présent document dans la sélection des petits modèles équipés.Pour les grands modèles équipés, ce critère constitue degré insuffisant de correction du biais ; comme l'indique un résultat du tableau, il a tendance à sous-estimer la fonction de classement dont il est dérivé. Des travaux récents ont conduit à une petite réussite de raffinement des échantillons des termes de pénalités de l'AIC et KIC . Deux types de raffinements sont utilisés dans la littérature ; Le premier est basé sur l'hypothèse d'une particulière formulaire pour les familles de modèles candidats et en utilisant les caractéristique de cette classe de familles de modèles pour dériver approximation plus précise pour le terme de pénalités, le deuxième variante est basée sur l'utilisation du bootstrap pour rapprochement de la correction du biais.

Conclusion

L'objectif de ce mémoire était d'étudier différents critères de sélection de modèles basés sur les divergences dirigées et symétriques de kulback , les comparés entre eux et les comparés avec les formes corrigées dans la sélection des modèles régressifs univariés et des modèles autorégressifs des séries chronologiques .Il a été constater que les critères corrigés , basées sur les divergences symétriques sont beaucoup plus performants que les critères non corrigé dans les sélections de modèles, cette performance est dû que, les critères corrigés sont exactement sans biais au lieu d'être asymptotiquement sans biais comme les critères non corrigé et également dû par le fait qu'ils sont basés sur des divergences symétriques donc ils apportent des informations supplémentaires sur la dissemblance entre le modèle générateur et candidat. Ce mémoire nous a présenté plusieurs formes de symétrisations de la divergence de kulback ; symétrisation par moyenne arithmétique, symétrisation par moyenne géométrique, symétrisation par moyenne harmonique. L'étude des critères de sélection dans l'estimation de ces divergences symétriques, dites fonctions de classement dans ce papier est harmonieusement faite, et on a constaté que lorsque la taille de l'échantillon est faible le cirtère AIC est un estimateur tros biaisé de ces fonctions de classement surtout lorsque le nombre de paramètre du modèle est grand, une correction du biais a été utile et qui donnait un critère corrigé noté AICc qui est mieux performant l'AIC. Par contre pour les échantillons de grandes tailles on a constaté que l'AIC estime bien ces fonctions de classement. Nous nous demandons que lorsque le vrai modéle n'appartient à la famille des modèles candidats que seraient les performances entres ces différents critères d'informations et leurs versions corrigées dans l'estimation des différentes fonctions de classement ?

Bibliographie

[1] J. E. Cavanaugh, *A large-sample model selection criterion based on Kullbac's symmetric divergence*, Statist. Probability Lett., vol. 42, pp. 333 − 343, 1999.

[2] C. R. Rao and Y. Wu, *On model selection*, ISM Lecture Notes-Monograph Series, vol. 38, pp. 1 − 57, 2000.

[3] M. Wax and T. Kailath *Detection of signals by information theoriticcriteria* IEEE Trans. Acoust., Speech, Signal Processing, vol. 33, pp.387-392, 1985

[4] H. Clergeot *Filter order selection in adaptive maximum likelihood estimation* IEEE Trans. Inform. Theory, vol. IT-30, pp. 199-210, 1984.

[5] S. L. Sclove *Application of model selection criteria to some problems in multivariate analysis*, Psychometrika, vol. 52, pp. 333-343, 1987.

[6] M. Hansen and B. Yu *Wavelet thresholding via MDL for natural images*, IEEE Trans. Inform. Theory, vol. 46, pp. 1778-1788, Aug. 2000.

[7] H. Akaike *A new look at the statistical model identification* IEEE Trans. Automat. Contr., vol. AC-19, pp. 716-723, 1974.

[8] G. Schwarz *Estimating the dimension of a model Ann. Statist.*, vol. 6, pp. 461-464, 1978.

[9] J. Rissanen *Modeling by shortest data description Automatica* vol. 14, pp. 465-471, 1978.

[10] R. A. Leibler *On information and sufficiency* Ann. Math. Statist., vol. 22, pp. 76-86, 1951.

[11] N. Sugiura *Further analysis of the data by Akaike's information criterion and the finite corrections* Commun. Statist., vol. A 7, pp. 13*-26, 1987.

[12] C. M. Hurvich and C. L. Tsai *Regression and time series model selection in small samples* Biometrika, vol. 76, pp. 297-307, 1989.

[13] A. K. Seghouane and M. Bekara *A small sample model selection criterion based on the Kullback symmetric divergence* IEEE Trans. Signal Process., vol. 52, no. 12, pp. 3314-3323, Dec. 2004.

[14] H. Jeffreys *An invariant form of the prior probability in estimation problems* J. Roy. Stat. Soc., vol. A, pp. 453-469, 19.

[15] C. M. Hurvich and C. L. Tsai *Regression and time series model selection in small samples* Biometrika, vol. 76, pp. 297-307, 1989.

[16] S. M. Ali and S. D. Silvey *A class of coefficients of divergence of one distribution from another* J. Roy. Stat. Soc., ser. B, vol. 28, pp. 131-142, 1966.

[17] S. Amari and H. Nagoaka *Methods of Information Geometry* London, U.K. : Oxford Univ. Press, 1993.

[18] C. M. Hurvich and C. L. Tsai *A corrected Akaike information criterion for vector autoregressive model selection* J. Time Ser. Anal., vol. 14, pp. 271-279, 1993.

[19] E. J. Bedrik and C. L. Tsai *Model selection for multivariate regression in small smaples* Biometrics, vol. 50, pp. 226-231, 1994.

[20] A. K. Seghouane *Autoregressive model order selection from small samples usingKullback ?s symmetric divergence* IEEE Trans. Circuits Syst. I, Reg. Papers, 2006, to be published.