REVUE DE STATISTIQUE APPLIQUÉE

PIERRE HAMMAD
PAP NGOM

Test d'ajustement et test de choix fondés sur une distance informationnelle généralisée

Revue de statistique appliquée, tome 46, nº 3 (1998), p. 89-107
<http://www.numdam.org/item?id=RSA_1998__46_3_89_0>

Rev. Statistique Appliquée, 1998, **XLVI** (3), 89-107

TEST D'AJUSTEMENT ET TEST DE CHOIX FONDÉS SUR UNE DISTANCE INFORMATIONNELLE GÉNÉRALISÉE

Pierre Hammad, Pap Ngom

Laboratoire de Mathématiques Appliquées, Université d'Aix-Marseille, Espace Forbin, 15-19 allée Claude Forbin, 13627 Aix-en-Provence

RÉSUMÉ

On choisit d'appuyer ce travail sur une mesure de discrimination Δ_r, entre deux distributions de probabilité, construite à partir d'une information de type Sharma et Mittal (1977), pour développer d'abord un test d'ajustement, puis un test de choix entre deux modèles paramétriques. On tente ensuite d'apprécier le degré de performance de ces tests par rapport à ceux du khi-deux et de Kolmogorov-Smirnov dans le premier cas, et la méthode de Vuong et Wang dans le deuxième cas.

***Mots-clés** : information généralisée, estimation, test d'hypothèses, simulation par Monte Carlo.*

ABSTRACT

We choose to base this work on a discrimination measure Δ_r, between two probability distributions, obtained through information-type measure of Sharma and Mittal (1977) in order to developp at first a test for goodness of fit, afterwards a test of parametric model choice. We try to estimate the degree of performance in terms of power about these statistical tests with regards to Chi-square and Kolmogorov-Smirnov at first, and Vuong and Wang method in the second case.

***Keywords** : generalized information, estimation, hypothesis test, Monte Carlo simulation.*

Introduction

L'essentiel des tests d'adéquation d'un modèle à un échantillon se fondent sur des statistiques suivant asymptotiquement une loi du khi-deux. On sait que cette méthode, basée par Pearson sur un regroupement des données en M classes et sur un calcul d'une «distance» entre proportion empirique et fréquence théorique, a été confortée par la suite par de nombreux auteurs dont Watson (1959), Moore (1978, 1986) par exemple.

Le rapprochement entre théorie de l'information et théorie de l'estimation et tests (Hammad 1987) justifie que l'on s'inspire, par exemple, d'autres distances pour certains types de tests. On peut citer les travaux de D. Morales et Menendez (1994) qui ont proposé comme base générale une (h, ϕ)-divergence et tenté une application pour les tests d'ajustement, de prédiction et d'homogénéité.

Dans le même ordre d'idées, il nous a paru intéressant de sélectionner un type particulier de mesure de discrimination Δ_r entre deux distributions et d'utiliser son comportement asymptotique pour des tests d'adéquation dans le cadre de modèles paramétriques. Le choix de Δ_r est dicté par un certain nombre de ses propriétés et notamment la recherche du test le plus puissant en fonction du paramètre r. Une comparaison de ce test avec ceux plus classiques du khi-deux et de Kolmogorov-Smirnov souligne en outre son efficacité.

Par ailleurs, dans la recherche d'un test de choix entre distributions, il est fréquent de s'appuyer sur le critère AIC de Akaike (1973), qui consiste, dans le cas de modèles paramétriques, à choisir le modèle fournissant le maximum de la log-vraisemblance pénalisée d'une quantité égale au nombre de paramètres : $AIC = L_n(\widehat{\theta}_n) - p$, où L_n désigne la log-vraisemblance du modèle, p la dimension du vecteur de paramètres θ et $\widehat{\theta}$ l'estimateur de θ. Un handicap majeur lié à l'utilisation de ce critère est qu'il ne précise pas le seuil de confiance que l'on peut accorder au modèle retenu.

Pour tenir compte du niveau de signification inhérente à toute décision statistique, Vuong et Wang (1993) proposent l'usage d'un test asymptotiquement normal lorsque la sélection de modèle est fondée sur des statistiques de type Pearson.

De façon analogue à l'approche de Vuong et Wang, nous suggérons, dans une deuxième partie, d'appuyer le problème du test de choix entre deux modèles paramétriques sur une statistique construite à partir de la distance $\Delta_{1/2}$ et nous établissons une comparaison, selon la loi suivie par les observations, entre les deux modèles.

1. Estimateur de la mesure Δ_r

1.1. Définitions et hypothèses

C'est à Rényi (1966) que l'on doit une première généralisation des mesures d'information, introduite à partir de la mesure suivante de proximité entre deux distributions P et Q.

$$\begin{cases} D_r^1\,[P,Q] = (r-1)^{-1} \ln \left\{ \sum_i p_i^r q_i^{1-r} \right\} \\ r \neq 1, r > 0 \end{cases} \tag{1}$$

Plus tard, Sharma et Mittal (1977) ont basé cette notion sur une mesure à deux paramètres, incluant celle de Rényi comme cas limite, en posant :

$$\begin{cases} D_r^s\,[P,Q] = (s-1)^{-1} \left\{ \left\{ \sum_i p_i^r q_i^{1-r} \right\}^{\frac{s-1}{r-1}} - 1 \right\} \\ r \neq 1, r > 0, s > 0 \end{cases} \tag{2}$$

D_r^r, «information» correspondant au cas $r = s$ dans (2), a été largement étudiée par plusieurs auteurs. Pour des développements plus généraux, on pourra se référer aux contributions - entre autres - de Mathaï et Rathie (1975), Tanéja (1979) ou encore Hammad (1987).

L'habituelle symétrie recherchée dans l'utilisation de ces critères de proximité requiert que l'on travaille plutôt à partir d'une distance de type Jeffreys (Voir Rényi (1966)), ici généralisée et déduite de (2) :

$$\Delta_r[P,\ Q] = D_r^r[P,\ Q] + D_r^r[Q,\ P] \tag{3}$$

C'est autour de cette mesure $\Delta_r[P,\ Q]$, et de son estimateur (défini ultérieurement) que nous proposons une méthode de test d'ajustement d'une série d'observations à un modèle paramétrique donné.

La mise en œuvre d'une telle procédure de test passe par quelques hypothèses de base dont nous rappelons l'essentiel.

Hypothèse (A_1) :

Les observations X_i, $i = 1, 2, \dots$, sont supposées i.i.d, avec une distribution commune H. L'espace d'échantillonnage Ξ est partitionné en M classes $E_1, E_2, \dots, E_M$, deux à deux disjointes.

Considérons un modèle $H_\theta = \{H(x, \theta)\ ;\ x\ \in \Xi,\ \theta \in \Theta \subset R^k\}$ et faisons l'hypothèse (A_1) que le vecteur des probabilités associées à H est de la forme :

$$h(\theta) = (h_1(\theta), h_2(\theta), \dots, h_M(\theta))$$

avec :

$$h_i(\theta) = \int_{E_i} dH(x, \theta) \quad i = 1, 2, \dots, M \tag{4}$$

Hypothèse (A_2) :

On suppose que $h_i(\theta)$ vérifie les conditions de régularité classiques :

(i) le support de H_θ est indépendant de tout x

(ii) les dérivées partielles suivantes existent et sont finies :

$$\frac{\partial h(\theta)}{\partial \theta_i}, \qquad \frac{\partial^2 h(\theta)}{\partial \theta_i \partial \theta_j}, \qquad \frac{\partial^3 h(\theta)}{\partial \theta_i \partial \theta_j \partial \theta_k}$$

(iii) la matrice d'information de Fisher

$$I_X^h(\theta) = \Big[E\{\frac{\partial}{\partial \theta_i} \log h(\theta) . \frac{\partial}{\partial \theta_j} \log h(\theta)\}\Big]_{i,j=1,\dots,M}$$

est définie positive.

Considérons un échantillon de taille n et $E_1, E_2, \ldots, E_M$ la partition en M classes qui lui est associée. On peut calculer la probabilité observée relative à chaque classe E_i, en posant :

$$\widehat{f} = (\widehat{f}_1, \widehat{f}_2, \ldots, \widehat{f}_M) \quad \text{où} \quad \widehat{f}_i = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}_{E_i}(X_j) \quad i = 1, 2, \ldots, M \tag{5}$$

avec $$\mathbf{1}_{E_i}(X_j) = \begin{cases} 1 & \text{si } X_j \in E_i \\ 0 & \text{sinon} \end{cases}$$

Pour évaluer l'écart entre les fréquences observées et les probabilités théoriques, (le paramètre θ étant supposé inconnu), on propose d'utiliser la mesure d'information $\Delta_r[f,\ h(\theta)]$ basée sur (3), où $f = \mathbf{E}(\widehat{f})$. La statistique associée à cette mesure sera définie par $\widehat{\Delta}_r$, obtenue en remplaçant θ par l'estimateur $\widehat{\theta}$.

On posera donc :

$$\widehat{\Delta}_r = \widehat{\Delta}_r[f,\ h(\theta)] = \Delta_r[f,\ h(\widehat{\theta})] \tag{6}$$

Examinons, à présent, comment se comporte la loi régissant cet estimateur.

1.2. Comportement asymptotique de $\widehat{\Delta}_r$

On situera l'estimateur $\widehat{\theta}$ de θ dans la classe des estimateurs vérifiant traditionnellement le principe de normalité asymptotique en ce sens que :

$$\sqrt{n}(\widehat{\theta} - \theta) \longrightarrow N[0,\ \Omega(\theta)] \tag{7}$$

où $\Omega(\theta)$ est l'inverse d'une matrice inversible.

Dans le cas précis où $\widehat{\theta}$ est l'estimateur du maximum de vraisemblance de θ, on a $\Omega(\theta) = I^{-1}(\theta)$ où $I(\theta)$ désigne la matrice d'information de Fisher. La loi asymptotique de $\widehat{\Delta}_r$ sera alors donnée par le théorème qui suit :

Théorème 1 :

Soit $\widehat{\Delta}_r[f,\ h(\theta)]$ l'estimateur de $\Delta_r[f,\ h(\theta)]$ obtenu en remplaçant θ par l'estimateur $\widehat{\theta}$ vérifiant (7).

On pose $f = \{\mathbf{E}(\widehat{f}_i)\}_i \quad \forall i = 1, \ldots, M$

(i) Si $f_l = h_l(\theta),\ \ l = 1, 2, \ldots, M,$ on a :

$$\frac{n}{r}\widehat{\Delta}_r\,[f, h(\theta)] \longrightarrow \chi^2_k$$

où $k = dim\Theta$

(ii) Si $f_l \neq h_l(\theta), \quad l = 1, 2, \ldots, M,$ on a :

$$\sqrt{n}\{\widehat{\Delta}_r[f,\ h(\theta)] - \Delta_r[f,\ h(\theta)]\} \longrightarrow N[0, \Gamma^2]$$

avec pour expression de Γ^2 :

$$\Gamma^2 = \lambda^2 \sum_{i,j} \Omega_{ij}(\theta) \Big\{ \sum_{l=1}^{M} \Big((1-r)\frac{f_l^r}{h_l^r(\theta)} + r\frac{h_l^{r-1}(\theta)}{f_l^{r-1}} \Big) \Big\}^2 \frac{\partial}{\partial\theta_i} h_l(\theta) \frac{\partial}{\partial\theta_j} h_l(\theta)$$

où $\lambda = (r-1)^{-1}$ et où $\Omega(\theta)_{ij}$ représente le terme général de la matrice $\Omega(\theta)$.

La démonstration repose sur un développement limité de Taylor de la fonction $\psi(\theta) = \Delta_r[f,\ h(\theta)]$ autour de θ à l'ordre 1 pour (i) et à l'ordre 2 pour (ii) (voir D. Morales *et al.* (1994) pour une démonstration d'une version plus générale de ce théorème).

2. Application aux tests d'adéquation

Nous proposons, dans cette section, une procédure de test d'adéquation à partir de Δ_r, pour ensuite tenter de l'interpréter par rapport aux tests habituels.

2.1. Ajustement à un modèle donné

Soient $f = (f_1, f_2, \ldots, f_M)$ et $h(\theta) = (h_1(\theta), h_2(\theta), \ldots, h_M(\theta))$ les vecteurs de probabilité définis à partir de (4), (5) et correspondant respectivement aux fréquences empirique et théorique du modèle associées à la partition considérée.

Les hypothèses à tester sont formulées comme suit :

$$\begin{aligned} H_o &: f = h \\ H_1 &: f \neq h \end{aligned} \qquad (8)$$

On considère, pour résoudre ce problème de test, la statistique

$$\widehat{\Delta}_r[f, h(\theta)] = \frac{1}{r-1}\Big[\sum_{l=1}^{M}\Big\{\frac{f_l^r}{h_l^{r-1}(\widehat{\theta})} + \frac{h_l^r(\widehat{\theta})}{f_l^{r-1}}\Big\} - 2\Big] \qquad (9)$$

afin d'estimer l'écart entre la distribution empirique et la loi du modèle.

Sous H_o, $\widehat{\Delta}_r$ a tendance d'après (9) à prendre de «petites valeurs», de sorte que, si l'on se fixe un niveau de signification α, la fonction de test est définie de la manière suivante :

$$\phi_1(x_1, x_2, \ldots, x_n) = \begin{cases} 1 & \text{si} \quad \widehat{\Delta}_r > C_\alpha \\ 0 & \text{sinon} \end{cases}$$

La valeur de la constante C_α résulte du théorème 1 (précédent) en vertu duquel, sous l'hypothèse nulle, la loi de $\frac{n}{r}\widehat{\Delta}_r$ suit asymptotiquement une loi du khi-deux. On a alors :

$$C_\alpha = \frac{r}{n}\,\chi^2_k(\alpha) \tag{10}$$

où $\chi^2_k(\alpha)$ est la valeur du khi-deux pour laquelle la probabilité de dépassement est égale à α.

Sous l'hypothèse alternative, la distribution asymptotiquement normale de $\sqrt{n}\,\widehat{\Delta}_r$ permet d'exprimer la puissance sous la forme :

$$P^r_n = 1 - \phi[\frac{\sqrt{n}}{\Gamma}(C_\alpha - \Delta_r[f,\ h(\theta)])] \tag{11}$$

où $\phi(.)$ désigne la fonction de répartition de la loi normale centrée réduite.

Le test ainsi obtenu est asymptotiquement convergent, au sens de Fraser :

$$\lim_{n\to+\infty} P^r_n = 1 \tag{12}$$

Ce résultat tient au fait que, dans (10), C_α tend vers 0 si n tend vers l'infini, et qu'en outre dans (11), Δ_r étant strictement positive, $(C_\alpha - \Delta_r[f,\ h(\theta)])$ est négatif dès que n est assez élevé.

L'expression (12) est la traduction d'un risque de seconde espèce asymptotiquement nul.

On tente à présent de se faire une idée du degré de performance de ce test en le comparant, par exemple, aux tests traditionnels du khi-deux et de Kolmogorov-Smirnov.

2.2 Comparaison des propriétés des tests

Pour comparer la précision ou l'exactitude de résultats relatifs à des tests asymptotiques, on dispose en général de deux méthodes respectivement fondées sur des procédures d'approximation ou bien sur des simulations. Parce que la première méthode conduit souvent à des calculs analytiques compliqués, on choisit ici de comparer les performances de ces tests en procédant à des simulations par la méthode de Monte Carlo .

Considérons une expérience de Monte Carlo [1], dans laquelle N réalisations de $\widehat{\Delta}_r$, du khi-deux, et de Kolmogorov-Smirnov sont générées en utilisant un processus de génération des données (PGD) incluant l'hypothèse nulle comme cas particulier, et suivant par exemple une loi exponentielle $\exp(1/\theta)$, de densité $f(x,\ \theta) = \frac{1}{\theta}\exp(-\frac{x}{\theta})\mathbf{1}_{\mathbf{R}^+}(x)$.

[1] Tous les calculs dans cet article ont été faits à partir de Gauss Version 3.1

On pose :

$$\begin{cases} \theta = \theta_o + \varepsilon \\ \theta_o > 0 \quad ; \quad \varepsilon \geqslant 0 \end{cases}$$

On suppose qu'on désire tester l'hypothèse nulle : $\theta_o = 1$, et que $\widehat{\Delta}_r$ est définie par la relation (9) précédente.

Compte tenu du rôle joué par l'origine (valeur modale) et par l'unité (valeur moyenne), on partitionnera l'espace des observations en trois classes, (ce qui représente le nombre minimum de classes, puisqu'il y a un seul paramètre à estimer dans l'expression de la densité $f(x, \theta)$) ainsi délimitées :

$$C_1 = [0,\ 0.1[\ ;\ C_2 = [0.1,\ 1[\text{ et } C_3 = [1,\ \infty[$$

La statistique de Kolmogorov K_n est basée sur la distribution empirique :

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{]-\infty,\, x[}(x_i)$$

à partir de laquelle :

$$K_n(x) = \sup_x |F_n(x) - F_o(x)|$$

F_o représentant la fonction de répartition théorique de référence.

On choisit ici une taille expérimentale d'échantillon égale à 100 et un nombre de répliques N fixé à 5000. Les résultats de la simulation basée sur ces différentes statistiques sont interprétés d'abord à partir des p-valeurs, puis des valeurs de la puissance, ce pour chacun des tests considérés.

• *Comparaison des p-valeurs*

Une comparaison des probabilités de rejet sous l'hypothèse nulle, effectuée de façon traditionnelle, consiste à tabuler les résultats obtenus pour quelques valeurs standards du niveau de signification α (1 %, 5 % ou 10 %).

On notera déjà, dans ce tableau, la croissance des probabilités de rejet (sous H_o) relatives à $\widehat{\Delta}_r$ en fonction de r, probabilités dont les valeurs sont de toutes façons supérieures aux valeurs nominales de départ.

Une interprétation plus synthétique (vraie quel que soit α) peut être obtenue graphiquement en utilisant la méthode introduite par Davidson & Mackinnon (1994). Cette méthode repose principalement sur l'estimation de la fonction de répartition empirique des p-valeurs correspondant au test utilisé. Considérons, par exemple, une expérience de Monte Carlo dans laquelle N réalisations d'une statistique S sont générées par un PGD. A chacune des N répliques de la simulation, on obtient une valeur s_j $(1 \leqslant j \leqslant N)$ de S et donc une valeur p_j de la p-valeur donnée par :

$$p_j = P[S > s_j] = 1 - F_S(s_j)$$

où F_S représente la fonction de répartition asymptotique de S.

TABLEAU 1
Comparaison entre niveaux de signification nominaux et réponses obtenues dans le cas d'une loi exponentielle.

Niveau de signification nominal α	0.10	0.05	0.01
Probabilité de rejet du χ^2	0.135	0.064	0.011
Probabilité de rejet de Kolmogorov	—	0.002	0.000
Probabilité de rejet de $\widehat{\Delta}_{1/2}$	0.143	0.073	0.016
Probabilité de rejet de $\widehat{\Delta}_1$	0.143	0.073	0.016
Probabilité de rejet de $\widehat{\Delta}_2$	0.148	0.082	0.022
Probabilité de rejet de $\widehat{\Delta}_3$	0.161	0.091	0.030

L'estimateur $\widehat{F}$ de la distribution empirique F des p-valeurs, moyennes des fonctions indicatrices $\mathbf{1}_{(p_j \leqslant x)}$, s'écrit :

$$\widehat{F}(x) \equiv \frac{1}{N} \sum_{j=1}^{N} \mathbf{1}_{(p_j \leqslant x)} \tag{13}$$

pour chaque point x dans $[0,\ 1]$.

Lorsque la distribution utilisée pour déterminer les p-valeurs p_j correspond à la loi exacte de la statistique S, on a alors, en prenant l'espérance mathématique de $\widehat{F}$ dans (13) :

$$\begin{aligned}\mathbf{E}(\widehat{F}(x)) &= P(p_j < x) = P[1 - F_S(s_j) < x] \\ &= 1 - P[1 - F_S(s_j) \geqslant x] = 1 - P[F_S(s_j) \leqslant 1 - x] \\ &= 1 - F_S(F_S^{-1}(1 - x)) = x\end{aligned}$$

Dans ce contexte précis, $\widehat{F}(x) - x$ traduit la différence entre le niveau de signification estimé par $\widehat{F}(x)$ et le niveau nominal x. On peut alors tracer (figure 1) la courbe correspondante donnant $(\widehat{F}(x) - x)$ en fonction de x. Pour des raisons liées aux difficultés de calcul des fractiles de la loi de Kolmogorov, nous nous limiterons ici aux statistiques $\Delta_{1/2}$, Δ_1, Δ_2, Δ_3 et celle du khi-deux.

La figure 1 montre, comme l'on pouvait si attendre, compte tenu du tableau 1, que la distance du khi-deux fournit les résultats les plus proches des p-values nominales. On constate, par ailleurs, que l'ensemble des courbes en figure 1 se comportent globalement de façon analogue (surtout pour $\widehat{\Delta}_{1/2}$ et $\widehat{\Delta}_1$), celle du χ^2 surclassant les autres.

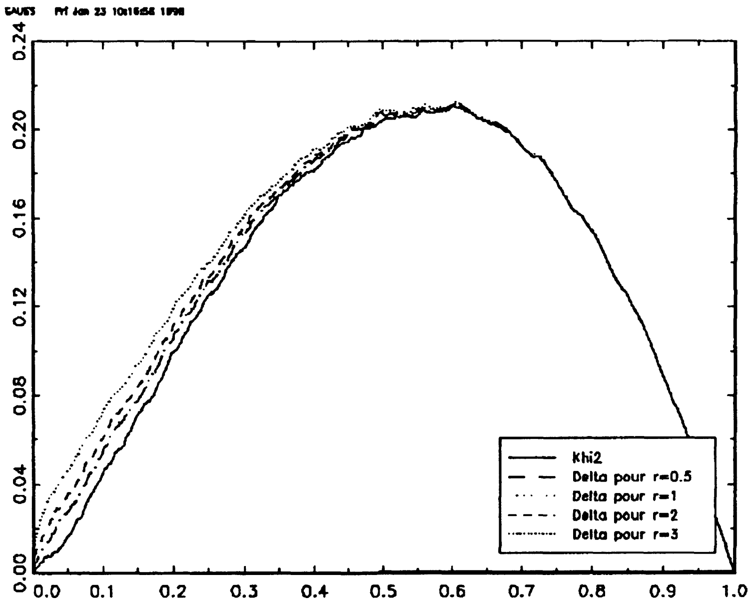

FIGURE 1
Graphe de $(\widehat{F}(x) - x)$ *en fonction de* x*, pour* $n = 100$

• *Comparaison des puissances*

Pour une meilleure appréciation des propriétés de ces statistiques, considérons leur comportement en termes de puissance; dans le cas de $\widehat{\Delta}_r$, la puissance résultera de l'expression :

$$P_n^r = 1 - \phi[\frac{\sqrt{n}}{\Gamma}(C_\alpha - \Delta_r)]$$

Le tableau qui suit établit la comparaison de ces puissances pour différents tests (basés sur $\widehat{\Delta}_r$, le χ^2 et Kolmogorov).

La puissance du test fondé sur $\widehat{\Delta}_r$ dépend évidemment de r et il apparaît intéressant de se faire une idée de son comportement par rapport à ce paramètre. C'est ce qu'un examen du tableau 2 permet de faire à travers le choix (justifié) d'un ensemble de quatre valeurs de r : $\{0.5; 1; 2; 3\}$. On note une croissance monotone de la puissance avec r. Ainsi, pour un niveau de signification de 5 % , la puissance du test du khi-deux, lorsque $\varepsilon = 1.4$, est égale à 73.10 % alors qu'elle est de 76.90 % pour $\widehat{\Delta}_{1/2}$, de 81.70 % pour $\widehat{\Delta}_2$ et 82.10 % pour $\widehat{\Delta}_3$. En comparaison, la fréquence de rejet est de 90.50 % pour le test de Kolmogorov.

Même si la puissance liée à $\widehat{\Delta}_r$ croît avec r, le cas $\widehat{\Delta}_{1/2}$ semble privilégié à plus d'un titre. C'est d'abord ce qu'illustre la figure 1 lorsque l'on s'intéresse à une comparaison des p-valeurs. Ensuite, la construction même de $\widehat{\Delta}_{1/2}$ présente

TABLEAU 2
Valeurs de la puissance en fonction du paramètre ε et de la statistique utilisée.

ε	1.1	1.2	1.3	1.4	1.5
khi-deux	0.199	0.358	0.555	0.731	0.855
Kolmogorov	0.014	0.118	0.472	0.905	0.999
$\widehat{\Delta}_{1/2}$	0.130	0.286	0.530	0.769	0.920
$\widehat{\Delta}_1$	0.157	0.331	0.581	0.809	0.940
$\widehat{\Delta}_2$	0.167	0.344	0.594	0.817	0.943
$\widehat{\Delta}_3$	0.177	0.353	0.601	0.821	0.945

l'avantage d'une propriété métrique qui en fait une vraie distance pour laquelle on constate que :

$$\begin{aligned}\Delta_{1/2}[f, h(\theta)] &= D_{1/2}^{1/2}[f,\ h(\theta)] + D_{1/2}^{1/2}[h(\theta),\ f]\\ &= 4\Big[1 - \sum_{i=1}^{M}\Big\{f_i\, h_i(\theta)\Big\}^{1/2}\Big]\end{aligned} \tag{14}$$

avec l'encadrement :

$$0 \leqslant \Delta_{1/2}[f, h(\theta)] \leqslant 4$$

Parmi les mesures connues de «type» $\Delta_{1/2}$, on peut citer par exemple, celles de Bhattacharya (1943) et de Matusita (1951, 1967), respectivement définies, pour deux distributions p et q, par :

$$B[p,\ q] = 1 - \sum_{i=1}^{n}\Big\{p_i\, q_i\Big\}^{1/2} \tag{15}$$

$$M[p,\ q] = \sum_{i=1}^{n}\Big\{p_i^{1/2} - q_i^{1/2}\Big\}^2 \tag{16}$$

$B[p,\ q]$ comme $M[p,\ q]$ possèdent les propriétés d'une vraie métrique et, de toute évidence, on a de plus :

$$\Delta_{1/2}[p,\ q] = 4B[p,\ q] = 2M[p,\ q]\ ; \tag{17}$$

ce qui entraîne par conséquent, pour les trois distances B, M et $\Delta_{1/2}$, des propriétés asymptotiques analogues.

3. Test de sélection de modèles

La recherche d'un test pour choisir une distribution parmi deux distributions s'appuie traditionnellement sur la méthode de Akaike (1973) ou celle souvent mieux adaptée de Vuong et Wang (1993) dont la base, dans le dernier cas, est la distance du khi-deux.

Par comparaison, on suggère ici une procédure pour déterminer, entre deux modèles paramétriques H_θ et G_π, celui qui s'adapte le mieux à la loi empirique d'une série d'observations donnée. On se basera pour cela sur les mesures d'information de type $\Delta_{1/2}$ servant de mesure de divergence entre le modèle H_θ ou G_π et les observations.

Les fonctions f, h et g désignent respectivement la fréquence empirique, la loi théorique du modèle H_θ et celle de G_π. Les estimateurs de θ et de π vérifient la relation (7).

Soient $\widehat{\Delta}_{1/2}[f,\ h(\theta)]$ et $\widehat{\Delta}_{1/2}[f,\ g(\pi)]$ les estimateurs respectifs de $\Delta_{1/2}[f,\ h(\theta)]$ et $\Delta_{1/2}[f,\ g(\pi)]$.

On considère les hypothèses suivantes :

(i) $\mathrm{H}_o : \ \Delta_{1/2}[f,\ h(\theta)] = \Delta_{1/2}[f,\ g(\pi)]$

(ii) $\mathrm{H}_{1,g} : \ \Delta_{1/2}[f,\ h(\theta)] > \Delta_{1/2}[f,\ g(\pi)]$

(iii) $\mathrm{H}_{1,h} : \ \Delta_{1/2}[f,\ h(\theta)] < \Delta_{1/2}[f,\ g(\pi)]$.

L'hypothèse (i) signifie que les modèles H_θ et G_π sont équivalents; (ii) traduit le fait que G_π est meilleur que H_θ et (iii) suggère de choisir H_θ plutôt que G_π.

La résolution de ce problème de choix entre H_θ et G_π sera fondée sur la statistique

$$\widehat{D}_n = D_n[h(\widehat{\theta}),\ g(\widehat{\pi})] = \sqrt{n}\{\widehat{\Delta}_{1/2}[f,\ h(\theta)] - \ \widehat{\Delta}_{1/2}[f,\ g(\pi)]\}$$

qui estime $\sqrt{n}\{\Delta_{1/2}[f,\ h(\theta)] - \ \Delta_{1/2}[f,\ g(\pi)]\}$.

Sous l'hypothèse nulle H_o, la loi asymptotique de $\widehat{D}_n$ est donnée par une version du théorème de Vuong-Wang (1993) :

Théorème 2 :

Si $\widehat{\theta}$ et $\widehat{\pi}$ représentent respectivement les $E.M.V$ de θ et π , on a (avec la notation ci-dessus de $\widehat{D}_n$) :

(1) sous $\mathrm{H}_o : \ \widehat{D}_n[h(\theta),\ g(\pi)] \longrightarrow N[0,\ \Sigma^2]$

(2) sous $\mathrm{H}_{1,g} : \ \widehat{D}_n[h(\theta),\ g(\pi)] \xrightarrow{P} +\infty$

(3) sous $\mathrm{H}_{1,h} : \ \widehat{D}_n[h(\theta),\ g(\pi)] \xrightarrow{P} -\infty$

Dans le cadre présent, l'expression de la variance Σ^2 associée à la statistique $\widehat{D}_n$, est déterminée moyennant un développement limité de Taylor à l'ordre 1 des fonctions $\widehat{\Delta}_{1/2}[f,\ h(\theta)]$ et $\widehat{\Delta}_{1/2}[f,\ g(\pi)]$.

Posons :

$$\widehat{\Delta}_{1/2}(\theta) = \widehat{\Delta}_{1/2}[f,\ h(\theta)] \quad ; \quad \widehat{\Delta}_{1/2}(\pi) = \widehat{\Delta}_{1/2}[f,\ g(\pi)]$$
$$\frac{\partial}{\partial\theta}\Delta_{1/2}(\theta) = p(\theta) \qquad \text{et} \qquad \frac{\partial}{\partial\pi}\Delta_{1/2}(\pi) = q(\pi)$$

On obtient alors :

$$\sqrt{n}\widehat{\Delta}_{1/2}(\theta) = \sqrt{n}\Delta_{1/2}(\theta) + \sqrt{n}\sum_i(\widehat{\theta}_i - \theta_i)\frac{\partial}{\partial\theta_i}\Delta_{1/2}(\theta) + R_n^1 \tag{18}$$

$$\sqrt{n}\widehat{\Delta}_{1/2}(\pi) = \sqrt{n}\Delta_{1/2}(\pi) + \sqrt{n}\sum_i(\widehat{\pi}_i - \pi_i)\frac{\partial}{\partial\pi_i}\Delta_{1/2}(\pi) + R_n^2 \tag{19}$$

Par différence des relations (18) et (19) :

$$\begin{aligned}\widehat{D}_n &= D_n + \sqrt{n}\sum_i(\widehat{\theta}_i - \theta_i)\ p_i(\theta) - \sqrt{n}\sum_i(\widehat{\pi}_i - \pi_i)\ q_i(\pi) + R_n \\ &= D_n + (p^t(\theta)\ ,\ -q^t(\pi))\begin{pmatrix}\sqrt{n}(\widehat{\theta} - \theta)\\ \sqrt{n}(\widehat{\pi} - \pi)\end{pmatrix} + R_n\end{aligned} \tag{20}$$

En posant :

$$C^t = C^t(\theta,\ \pi) = (p^t(\theta)\ ,\ -q^t(\pi)) \quad ; \quad \widehat{\eta} = \sqrt{n}(\widehat{\theta} - \theta)$$
$$\text{et} \quad \widehat{\delta} = \sqrt{n}(\widehat{\pi} - \pi)$$

on obtient :

$$\widehat{D}_n = D_n + C^t(\theta\ ,\ \pi)\begin{pmatrix}\widehat{\eta}\\ \widehat{\delta}\end{pmatrix} + R_n \tag{21}$$

et comme $R_n = R_n^1 - R_n^2 \to 0$ quand n tend vers l'infini, on en déduit que les deux variables aléatoires $\widehat{D}_n - D_n$ et $C^t(\theta\ ,\ \pi)\begin{pmatrix}\widehat{\eta}\\ \widehat{\delta}\end{pmatrix}$ ont asymptotiquement la même distribution, autrement dit :

$$\widehat{D}_n - D_n \to N[0,\ \Sigma^2]$$

avec

$$\Sigma^2 = C^t \Lambda C$$

où

$$\Lambda = \begin{bmatrix} I^{-1}(\theta) & \mathbf{E}(\widehat{\eta}\,\widehat{\delta}^t) \\ \mathbf{E}(\widehat{\delta}\,\widehat{\eta}^t) & I^{-1}(\pi) \end{bmatrix}$$

puisque $\widehat{\eta}$ et $\widehat{\delta}$ sont des variables aléatoires centrées.

3.1. Règle de décision associée à la statistique $\widehat{D}_n$

Nous allons nous appuyer sur l'inégalité triangulaire que vérifie la métrique $\Delta_{1/2}$ pour un encadrement de $\widehat{D}_n$. En effet :

$$\Delta_{1/2}[f,\ h(\theta)] \leqslant \Delta_{1/2}[f,\ g(\pi)] + \Delta_{1/2}[g(\pi),\ h(\theta)]$$

soit :

$$\Delta_{1/2}[f,\ h(\theta)] - \Delta_{1/2}[f,\ g(\pi)] \leqslant \Delta_{1/2}[g(\pi),\ h(\theta)] \qquad (22)$$

D'autre part :

$$\Delta_{1/2}[f,\ g(\pi)] \leqslant \Delta_{1/2}[f,\ h(\theta)] + \Delta_{1/2}[h(\theta),\ g(\pi)]$$

ce qui entraîne :

$$\Delta_{1/2}[f,\ g(\pi)] - \Delta_{1/2}[f,\ h(\theta)] \leqslant \Delta_{1/2}[h(\theta),\ g(\pi)] \qquad (23)$$

Posons :

$$k_n = \sqrt{n}\,\Delta_{1/2}[h(\widehat{\theta}),\ g(\widehat{\pi})]$$

En multipliant (22) par $\sqrt{n}$ et (23) par $-\sqrt{n}$, et en remplaçant ensuite θ et π par leurs estimateurs respectifs $\widehat{\theta}$ et $\widehat{\pi}$, on obtient en fait :

$$-k_n \leqslant \widehat{D}_n \leqslant k_n$$

Pour réaliser le test de choix entre h et g, on peut envisager une règle de décision définie comme suit, pour un niveau de signification supposé égal α :

– il y a équivalence entre h et g si :

$$\widehat{D}_n \in \Big[-z_{\alpha/2}\,\Sigma,\ z_{\alpha/2}\,\Sigma\Big]$$

– on décide en faveur de h lorsque :

$$\widehat{D}_n \in \Big[-k_n,\ -z_{\alpha/2}\,\Sigma\Big]$$

– on décide en faveur de g si :

$$\widehat{D}_n \in \left[z_{\alpha/2}\,\Sigma,\ k_n\right]$$

Σ^2 représentant la variance de la statistique $\widehat{D}_n$ et $z_{\alpha/2}$ le quantile $(1-\alpha/2)$ de la loi normale centrée réduite.

3.2. Exemples d'application

On propose ici une comparaison entre la statistique $\widehat{D}_n[h(\theta),\ g(\pi)]$, construite avec $\widehat{\Delta}_{1/2}$ et $\widehat{K}_n[h(\theta),\ g(\pi)]$, obtenue en fonction de la statistique de Pearson (voir Vuong et Wang 1993). Ces statistiques sont définies comme suit :

$$\widehat{D}_n[h(\theta),\ g(\pi)] = \sqrt{n}\{\widehat{\Delta}_{1/2}[f,\ h(\theta)] - \widehat{\Delta}_{1/2}[f,\ g(\pi)]\} \tag{24}$$

$$\widehat{K}_n[h(\theta),\ g(\pi)] = \frac{1}{\sqrt{n}}\{\widehat{Q}_n[f,\ h(\theta)] - \widehat{Q}_n[f,\ g(\pi)]\} \tag{25}$$

$\widehat{Q}_n[f,\ h(\theta)]$ désignant la distance du khi-deux entre la fréquence empirique f et la distribution théorique h.

A titre d'illustration, des simulations par Monte Carlo ont été mises en œuvre à partir de quelques distributions, afin de comparer la méthode de Vuong et Wang avec la procédure que nous avons proposée. On se limitera ici, à trois types de lois dont les densités de probabilité sont définies sur :

– un intervalle $[a,\ b]$,

– l'ensemble R_+

– l'ensemble R.

Le nombre de répliques utilisé pour construire les distributions empiriques est $N = 5000$ et la taille des échantillons considérés varie entre 70 et 800. Le niveau de signification retenu est de 5 %.

3.2.1. Cas de deux distributions définies sur $[0,\ 1]$

On veut sélectionner un modèle parmi deux distributions (une loi Bêta et une loi uniforme), sur la base d'une série d'observations obtenues à partir de deux processus de génération des données (PGD) Y_1 et Y_2 de densités respectives f_1 et f_2 :

$$Y_1 \sim Be(p,\ q) \qquad f_1 = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}x^{p-1}(1-x)^{q-1}\mathbf{1}_{[0,\ 1]}(x)$$

$$Y_2 \sim U_{[0,\ 1]} \qquad f_2 = \mathbf{1}_{[0,\ 1]}(x)$$

Afin d'espérer des résultats tangibles, il est nécessaire de pouvoir raisonnablement discerner les deux distributions; on choisira à cet effet les valeurs $p = 1$ et

$q = 2$ et l'on regroupera les données en trois classes $C_1 = [0,\ 0.2[$; $C_2 = [0.2,\ 0.8[$ et $C_3 = [0.8,\ 1]$.

On obtient les tableaux suivants :

TABLEAU 3
PGD : $Y_1 \sim Be(1,\ 2)$

Taille de l'échantillon		70	100	150	200
Modèle fondé sur la statistique $\widehat{K}_n$	décision : f_1	0.634	0.809	0.940	0.981
	indécision	0.366	0.191	0.060	0.019
	décision : f_2	0.000	0.000	0.000	0.000
Modèle fondé sur la statistique $\widehat{D}_n$	décision : f_1	0.605	0.806	0.946	0.987
	indécision	0.395	0.194	0.054	0.013
	décision : f_2	0.000	0.000	0.000	0.000

TABLEAU 4
PGD : $Y_2 \sim U_{[0,\ 1]}$

Taille de l'échantillon		100	150	200	300
Modèle fondé sur la statistique $\widehat{K}_n$	décision : f_1	0.000	0.000	0.000	0.000
	indécision	0.449	0.254	0.123	0.024
	décision : f_2	0.551	0.746	0.877	0.976
Modèle fondé sur la statistique $\widehat{D}_n$	décision : f_1	0.000	0.000	0.000	0.000
	indécision	0.164	0.050	0.014	0.001
	décision : f_2	0.836	0.950	0.986	0.999

3.2.2. Cas de deux distributions définies sur R_+

Envisageons le problème qui consiste à choisir, par exemple, entre une distribution exponentielle $\exp(\theta)$ de paramètre θ et une loi Gamma $\Gamma(p,\ \alpha)$ de paramètres $(p,\ \alpha)$, de densités respectives :

$$\begin{cases} f_3(x,\ \theta) = \theta \exp(-\theta x) \\ \quad x \geqslant 0 \quad \text{et} \quad \theta > 0 \end{cases} \tag{26}$$

$$\begin{cases} f_4(x,\ p,\ \alpha) = \dfrac{\alpha^p}{\Gamma(p)} x^{p-1} \exp(-\alpha x) \\ \quad x \geqslant 0, \quad p > 0 \quad et \quad \alpha > 0 \end{cases} \tag{27}$$

Nous supposerons dans ce qui suit que les estimateurs de θ et α sont obtenus par la méthode MV.

Pour des raisons de calcul, on donnera une valeur entière à p, la valeur 2 par exemple. Par ailleurs, pour espérer ici obtenir des résultats significatifs, on prendra dans (26) $\theta = 0.707$ et dans (27) $\alpha = 1$, de telle sorte que les données issues de ces deux lois conduisent à la même variance, atténuant ainsi «l'écart» entre les distributions choisies. Les observations seront réparties en trois classes :

$$C_1 = [0,\ 0.3[;\ C_2 = [0.3,\ 1.5[;\ C_3 = [1.5,\ +\infty[$$

Dans le cas présent, nous générons les échantillons à partir de deux processus de génération des données :

$Y_3 \sim \exp(0.707)$

$Y_4 \sim \Gamma[2,\ 1]$

TABLEAU 5

PGD : $Y_3 \sim \exp(0.707)$

Taille de l'échantillon		100	200	300	500
Modèle fondé sur la statistique $\widehat{K}_n$	décision : f_3	0.449	0.749	0.906	0.989
	indécision	0.551	0.251	0.094	0.011
	décision : f_4	0.000	0.000	0.000	0.000
Modèle fondé sur la statistique $\widehat{D}_n$	décision : f_3	0.578	0.859	0.961	0.997
	indécision	0.422	0.141	0.039	0.003
	décision : f_4	0.000	0.000	0.000	0.000

TABLEAU 6

PGD : $Y_4 \sim \Gamma[2,\ 1]$

Taille de l'échantillon		300	500	600	800
Modèle fondé sur la statistique $\widehat{K}_n$	décision : f_3	0.003	0.001	0.000	0.000
	indécision	0.681	0.464	0.387	0.249
	décision : f_4	0.316	0.535	0.613	0.751
Modèle fondé sur la statistique $\widehat{D}_n$	décision : f_3	0.000	0.000	0.000	0.000
	indécision	0.502	0.271	0.207	0.092
	décision : f_4	0.498	0.729	0.793	0.908

3.2.3 Cas de deux distributions définies sur R

On veut choisir entre une distribution de Laplace $\xi(\alpha,\ \lambda)$ (ou loi exponentielle double) et une loi normale $N[m,\ \sigma^2]$. On considère les PGD Y_5 et Y_6 suivants :

$$Y_5 \sim \xi(\alpha,\ \lambda) \qquad f_5(x,\ \alpha,\ \lambda) = \frac{\lambda}{2} e^{-\lambda|x-\alpha|}$$

$$Y_6 \sim N[m,\ \sigma^2] \qquad f_6(x,\ m,\ \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

Nous supposerons les observations issues de populations ayant même moyenne, $m = \alpha = 0$ et, pour simplifier, on prendra $\sigma = \lambda = 1$. On pose $C_1 =]-\infty, \ -2[$; $C_2 = [-2, \ 2[$ et $C_3 = [2, \ +\infty[$, comme partition associée aux observations.

On obtient les résultats ci-dessous :

TABLEAU 7
PGD : $Y_5 \sim \xi(0, \ 1)$

Taille de l'échantillon		100	200	300	500
Modèle fondé sur la statistique $\widehat{K}_n$	décision : f_5	0.279	0.467	0.652	0.865
	indécision	0.721	0.533	0.348	0.135
	décision : f_6	0.000	0.000	0.000	0.000
Modèle fondé sur la statistique $\widehat{D}_n$	décision : f_5	0.311	0.599	0.797	0.948
	indécision	0.687	0.401	0.203	0.052
	décision : f_6	0.002	0.000	0.000	0.000

TABLEAU 8
PGD : $Y_6 \sim N[0, \ 1]$

Taille de l'échantillon		100	200	300	500
Modèle fondé sur la statistique $\widehat{K}_n$	décision : f_5	0.017	0.010	0.003	0.000
	indécision	0.818	0.702	0.398	0.036
	décision : f_6	0.165	0.288	0.599	0.964
Modèle fondé sur la statistique $\widehat{D}_n$	décision : f_5	0.000	0.000	0.000	0.000
	indécision	0.728	0.634	0.507	0.088
	décision : f_6	0.272	0.366	0.493	0.912

Dans le tableau 3, le test fondé sur la statistique du khi-deux donne des résultats sensiblement proches de celui fondé sur la mesure D_n. D'autre part, dans les tableaux 4, 5, 6 et 7, les résultats sont nettement meilleurs lorsque l'on considère le test obtenu à partir de $\widehat{D}_n$.

En revanche, dans le tableau 8, on notera que la méthode associée à $\widehat{K}_n$ semble préférable dès que la taille de l'échantillon devient suffisamment grande . En effet, sur la base de 500 observations par exemple, la bonne décision se traduit par une probabilité d'acceptation de l'ordre de 96.40 % pour $\widehat{K}_n$ et de 91.20 % pour $\widehat{D}_n$.

Conclusion

Nous avons, dans cet article, tenté d'utiliser une distance informationnelle de type Rényi, pour des tests aussi bien d'ajustement que de choix de modèles. Pour en cerner l'efficacité, nous avons en parallèle, comparé nos résultats dans les deux situations avec ceux fournis par les tests classiques du khi-deux ou de Kolmogorov. De cette tentative informationnelle et de cette comparaison, on retiendra essentiellement ce qui suit :

- pour le test d'ajustement, à travers le critère des p-valeurs, les distances $\Delta_{1/2}$ et Δ_1 (très proches l'une de l'autre) sont, parmi les Δ_r, les plus efficaces mais s'avèrent moins performantes que le khi-deux (le test de Kolmogorov n'a pas été ici pris en compte en raison de difficultés de calcul évidentes). Avec le critère «puissance», à partir de certaines valeurs du paramètre ε, Δ_r quel que soit r est préférable au khi-deux, le test de Kolmogorov s'avérant cependant meilleur;

- pour le test de choix de modèle, on s'est limité, en le justifiant, à comparer $\widehat{\Delta}_{1/2}$ et le khi-deux au travers des statistiques $\widehat{D}_n$ et $\widehat{K}_n$ données en (24) et (25). Il apparaît, d'après les résultats obtenus, qu'aucune des deux statistiques de test considérées ici n'est systématiquement plus performante que l'autre (tableaux 3 et 8). Cependant, dans de nombreux cas, le test basé sur $\widehat{\Delta}_{1/2}$ engendre une meilleure puissance (tableaux 4, 5, 6 et 7).

On retiendra enfin que dans le cadre des petits échantillons (pour les échantillons de grande taille, ces statistiques de test sont équivalentes), les résultats obtenus, en plus de la simplicité de calcul de $\widehat{\Delta}_{1/2}$, plaident en faveur de cette distance dans plusieurs situations.

Références

[1] AKAIKE H. (1973). «Information theory and Extension of the Likelihood Ratio Principe», *Proceedings of the second International Symposium of Information theory*, ed. By. Pietrov, B.N and Csaki, F. Budapest : Akademiai Kiado, pp. 257-281.

[2] BHATTACHARYYA A. (1943). «On a measure of divergence between two statistical populations defined by their probability distributions», *Bull. Calcutta Math.Soc.*, 35, 99-109.

[3] DAVIDSON R., J.G. MACKINNON (1994). «Graphical methods for investigating the size and Power of hypothesis tests», *Documents de travail G.R.E.Q.A.M* n° 94A23 Juin.

[4] HAMMAD P. (1987). «Information, Systèmes et distributions», *Editions Cujas*, Paris.

[5] MATUSITA K. (1951). «On theory of décision functions», *Ann. Inst. Statist. Math.*, 3, 17-35.

[6] MATUSITA K. (1967). «On the notion of affinity of several distributions and some of its applications», *Ann. Inst. Statist. Math.*, 19, 181-192.

[7] MORALES D., PARDO L., Salicrù M., Menendez M.L. (1992). «A test of independance based on the (r, s)-directed divergence», *Tamkang Journal of Mathematics*, Vol. 23, n° 2, Summer.

[8] MORALES D., PARDO L., SALICRÙ M. and MENENDEZ M.L. (1994). «Asymptotic properties of divergence statistics in a stratified random sampling and its applications to test statistical hypotheses», *Journal of Statitical Planning and Inference*, 38, p. 201-222 North-Holland.

[9] MOORE D.S. (1978). «Chi-Squared Tests», *in statudies in statistics*, ed. by HoGG , R.V. Volume, The Mathematical Association of America.

[10] MOORE D.S. (1986). «Test of Chi-Squared type», *ed. D'Agostino R.B and Stephens M.A.*

[11] MATHAI A.M. and RATHIE P.N. (1975). «Basic Concepts of Information Theory and Statistics», Wiley, New York.

[12] PEARSON K. (1990). «On the criterion that a given System of deviation from the probable in the case of a correlated System of Variables is Such that it can be reasonably supposed to have Arisen from Random Sampling», *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science*, 50, pp. 157-175.

[13] A. RENYI (1966). Calcul des probabilités (avec un appendice sur la théorie de l'information)», Dunod, Paris.

[14] SHARMA B.D. and MITTAL D.P. (1977). «New nonadditive measures of entropy for discrete probability distributions», *J. Math. Sci.*, 10, 28-40.

[15] TANEJA I.J. (1979). «Some Contributions to Information Theory I (A survey) : On Measures of Information», *J. Comb., Inform. Sys. Sci.* 4(4), 253-274.

[16] VUONG Q.H. (1989). «Likelihood Ratio tests for model Selection and non-nested Hypotheses», *Econometrica*, 57, pp. 257-306.

[17] VUONG Q.H. and W. WUANG (1993). «Selecting Estimated Models using Chi-Square Statistics», *Annales d'économie et de Statistique*, 30, pp. 143-164.

[18] WATSON G.S. (1959). «Some Recent Results in Chi-Square Goodness-of-Fit Tests», *Biometrics*, 15, pp. 440-468.