Analyse et modélisation de séries temporelles à l'aide de cascades. Application à l'étude du trafic Internet

Pierre CHAINAIS[1], Stéphane ROUX[1], Patrice ABRY[1] et Darryl VEITCH[2]

[1]Laboratoire de Physique, CNRS UMR 5672, École Normale Supérieure de Lyon
46, allée d'Italie 69364 Lyon Cedex 07, France.

[2]EMUlab, Dept. of E&EE, Univ. Melbourne, Australia.

Avec le soutien d'Ericsson, du MENRT/ACI "jeune chercheur" 2329 et du CNRS, programme "télécommunications" 99035.

Pierre.Chainais, Stephane.Roux, Patrice.Abry@ens-lyon.fr, d.veitch@emulab.ee.mu.au

Résumé –

Les Cascades Infiniment Divisibles (CID) constituent un modèle riche et souple pour décrire une large variété de comportements en lois d'échelle. Elles incluent et enrichissent les modèles des processus auto-similaires ou multifractals. Les analyses en ondelettes fournissent un outil idéalement adapté à l'étude des lois d'échelle. Nous nous intéressons ici aux performances statistiques d'estimateurs des paramètres définissant les CID, construits sur les coefficients d'ondelette. Nous effectuons, sur des signaux synthétiques de référence, deux types de comparaisons : d'une part, utilisation d'une transformée en ondelettes discrète (TOD) versus la méthode dite des maxima des modules (MMTO) ; d'autre part, utilisation des moments des coefficients d'ondelette versus cumulants de leur log. Nous appliquons ces estimateurs à l'étude de données de télétrafic informatique.

Abstract –

Infinitely Divisible Cascades (IDC) form a rich and versatile framework to model a large variety of scaling phenomena, including self-similar and multifractal processes. Wavelet analysis is naturally matched to scaling phenomena, yielding ideal tools for their analysis. We study the statistical performance of wavelet-based estimators for the parameters defining IDC. Two types of comparisons, based on synthetic reference signals, are performed: First, the use of the discrete wavelet transform versus that of the so called wavelet transform modulus maxima method; second, the use of the moments of the wavelet coefficients versus that of the cumulants of their logarithm. The estimators are applied to the study of Internet data.

1 Motivations

Dans l'étude de phénomènes complexes très divers (turbulence [5], géophysique, biologie, télétrafic informatique [8, 11], finance [4], ...), on est confronté à des signaux très irréguliers, qui présentent une forte variabilité à toutes les échelles. On parle alors de comportements en lois d'échelle : il n'existe pas d'échelle particulière mais plutôt une relation permettant de passer d'une échelle à une autre. La modélisation de tels phénomènes présente deux intêrets principaux : permettre une meilleure compréhension des mécanismes de production des lois d'échelle et dÕvelopper nos capacité à prédire les données.

De nombreux travaux ont montré la pertinence de l'outil ondelette pour l'analyse des phénomènes de loi d'Õchelle [1]. Nous formulerons donc modèles et analyses directement sur les coefficients d'ondelette, dont nous rappelons brièvement la définition[1]. Soit X la sÕrie temporelle ÕtudiÕe, nous noterons $\{T_X(a,t) =< X, \psi_{a,t} >\}$ ses coefficients en ondelette oÌ les $\psi_{a,t}(u) = |a|^{-1}\psi(a^{-1}(u-t))$ sont des dilatÕes [2], d'un facteur a, et translatÕes, d'un facteur t, de l'ondelette-mère ψ.

La plupart des modèles dédiés à la description de lois d'échelle reposent sur la notion d'autosimilarité, qu'elle soit globale ou locale. Parmi les processus globalement autosimilaires, les plus connus sont les mouvements browniens fractionnaires, pour lesquels on obtient :

$$\ln \mathbb{E}|T_X(a,t)|^q = c_q \exp(qH \ln(a)) \ \ \forall t > 0\ ,\ \ \forall q\ , \tag{1}$$

où le seul paramètre H, le paramètre d'autosimilarité, intimement relié à la régularité globale du signal, suffit à spécifier complètement l'évolution de tous les moments à travers les échelles. La complexité des observations expérimentales impose souvent d'avoir recours à un ensemble d'exposants $H(q)$ plutôt qu'un seul H : on rencontre alors les modèles multifractals pour lesquels

$$\ln \mathbb{E}|T_X(a,t)|^q = c_q \exp(H(q) \ln(a)) \ \ \forall t > 0\ ,\ \ \forall q\ . \tag{2}$$

L'ecart de $H(q)$ à qH rend alors compte des fluctuations de la régularité locale des processus. Ces modèles restent cependant limités dans la mesure où la dépendance en échelle a est figée en loi de puissance. Or, expérimentalement, on observe relativement systématiquement des écarts à ces lois de puissance. On peut en rendre compte en remplaçant la fonction $\ln a$ par une fonction $n(a)$:

$$\ln \mathbb{E}|T_X(a,t)|^q = c_q \exp(H(q) n(a)) \ \ \forall t > 0\ ,\ \ \forall q\ . \tag{3}$$

Ces relations peuvent être obtenues à partir des cascades infiniment divisibles (CID), modèle riche et souple qui permet de décrire dans un cadre de travail unique la plupart

[1]Pour plus de dÕtails nous renvoyons le lecteur Ĺ la rÕfÕrence [9].

[2]La normalisation en $|a|^{-1}$, inhabituelle en physique parce qu'elle ne préserve pas l'énergie (la norme L_2), est particulièrement bien adaptée à l'étude des lois d'échelles.

des modèles en loi d'échelle. Le paragraphe suivant définit précisément les CID puis propose des estimateurs des éléments qui les définissent. Ces estimateurs reposent sur les moments des coefficients d'ondelette ou sur les cumulants de leur logarithme. Les performances statistiques de ces estimateurs sont ensuite étudiées et comparées à l'aide de simulations numériques réalisées sur des signaux de synthèse. La comparaison porte également sur diverses transformées en ondelettes : la transformée en ondelettes discrète (TOD) et la méthode dite du maxima des modules de la transformée en ondelettes (MMTO). Enfin, nous appliquons ces estimateurs à l'étude de séries temporelles expérimentales correspondant à des flux d'information sur un réseau Internet.

2 Cascade Infiniment Divisible

Définition. Une CID lie la densité de probabilité du log des coefficients d'ondelette $Y(a,t) = \ln|T_X(a,t)|$ à l'échelle a à celle de l'échelle $a' \geq a$ à travers un noyau de convolution appelé propagateur de la cascade [3] :

$$p_a(Y) = \int G_{a,a'}(Y-Y')p_{a'}(Y')dY', \quad (4)$$

On impose, de plus que la transformée de Laplace $\hat{G}_{a,a'}(q)$ de $G_{a,a'}(Y)$ soit caractérisée par une séparation des variables a et q : $\ln \hat{G}_{a,a'}(q) = \hat{G}_0(q)^{(n(a)-n(a'))}$. La fonction $\hat{G}_0(q)$ se lit alors comme le pas élémentaire de la cascade, tandis que la fonction $n(a)$ décrit la façon dont la cascade se déroule le long des échelles. En utilisant le développement de la seconde fonction caractéristique en cumulants,

$$\ln \mathbb{E}|T_X(a,t)|^q = \ln \mathbb{E} \exp(q \ln|T_X(a,t)|) = \sum_l c_{a,l} q^l/l!,$$

(où les $c_{a,l}$ sont les cumulants des $Y(a,t)$), nous obtenons les relations centrales suivantes :

$$\begin{array}{rcl}
\ln \mathbb{E}|T_X(a,t)|^q &=& H(q)(n(a)-n(a')) + \ln \mathbb{E}|T_X(a',t)|^q \\
c_{a,l} &=& C_l(n(a)-n(a')) + c_{a',l} \\
\ln \mathbb{E}|T_X(a,t)|^q &=& H(q)/H(p) \ln \mathbb{E}|T_X(a,t)|^p + K_{q,p} \\
c_{a,l} &=& (C_l/C_p)c_{a,p} + \beta_{l,p} \\
H(q) &=& \sum_{l=1}^{\infty} C_l q^l/l!\ .
\end{array} \quad (5)$$

où $H(q) = \ln \hat{G}_0(q)$. Il convient de noter que les fonctions $H(q)$ et $n(a)$ sont définies respectivement à une constante multiplicative près, et à une constante multiplicative et additive près [14]. Ces Õquations nous indiquent que, quelle que soit la forme de $n(a)$, les moments se comportent en loi de puissance les uns par rapport aux autres. Lorsque $n(a) \equiv \ln a$, les relations (5) indiquent que les $\mathbb{E}|T_X(a,t)|^q$ se comportent en lois de puissance des échelles. Lorsque ces comportements existent dans la limite des petites échelles, l'analyse par CID se réduit à une analyse multifractale [12, 14]. Dans le cas encore plus spécifique où $n(a) \equiv \ln(a)$ sur toutes les échelles et où $H(q) \equiv qH$, le processus X est H-auto-similaire [1, 8].

Principe de l'estimation. A partir des relations (5) ont été développées différentes méthodes pour tester l'adéquation données-modèle, d'une part, et estimer les paramètres des CID, d'autre part. Celles-ci reposent sur l'estimation des moments des coefficients d'ondelette ou des cumulants de leur logarithme. Grâce à la stationnarité et à la faible dépendance statistique des coefficients d'ondelette on peut substituer la moyenne temporelle à la moyenne statistique. C'est-à-dire estimer pour une réalisation $\mathbb{E}|T_X(a,t)|^q$ et $\mathbb{E}(\ln T_X(a,t))^l, l \in \mathbb{Z}$, par :

$$\begin{array}{rl}
S_q(a) = & 1/n_a \sum_{k=1}^{n_a} |T_X(a,t)|^q\ , \\
M_l(a) = & 1/n_a \sum_{k=1}^{n_a} (\ln|T_X(a,t)|)^l\ .
\end{array}$$

A partir des $M_l(a)$ sont calculées les estimées $\hat{c}_{a,l}$ des cumulants.
Pour vérifier l'hypothèse de CID, la procédure consiste à tester la linéarité de la courbe $\ln S_q(a)$ vs $\ln S_p(a)$.
Pour estimer $H(q)$, on choisit une référence p, on fixe les constantes multiplicative et additive ($H(p) \equiv 1$, $\log S_p - H(p)n(a) \equiv 0$). On effectue une régression linéaire dans les diagrammes $\ln S_q(a)$ vs $\ln S_p(a)$:

$$\ln S_q(a) = \quad \hat{H}_p(q) \ln S_p(a) + \hat{K}_{q,p}\ .$$

Pour estimer $n(a)$, on utilise l'ordonnée à l'origine $\hat{K}_{q,p}$, pour écrire :

$$\hat{n}(a) \quad = \quad \left\langle \frac{1}{\hat{H}_p(q)} \left(\ln S_q(a) - \hat{K}_{q,p} \right) \right\rangle_q .$$

$< . >_q$ représente la moyenne sur les valeurs de q. Ces estimateurs ont été proposés dans [6] et exploitent la séparation des variables q et a.
Nous pouvons réécrire des estimateurs de $H(q)$ et $n(a)$ reposant sur les cumulants et leur inter-linéarité :

$$\hat{c}_{a,l} = \quad \hat{C}_p(l)\hat{c}_{a,p} + \hat{\beta}_{l,p}\ .$$

On estime $n(a)$ par

$$\begin{array}{rcl}
\hat{n}(a) &=& \left\langle \frac{1}{\hat{C}_p(l)} \left(\hat{c}_{a,l} - \hat{\beta}_{l,p} \right) \right\rangle_l , \\
\hat{H}_p(q) &=& (\sum_l \hat{C}_p(l) q^l/l!)/(\sum_l \hat{C}_p(l) p^l/l!)\ .
\end{array}$$

Transformées en Ondelettes. Il existe plusieurs formes de transformées en ondelettes : continue (TOC), discrète (TOD) et maxima des modules (MMTO). Les coefficients de la seconde constituent la partie de ceux de la première localisée sur une grille figée, dite dyadique : $d_X(j,k) = T_X(a = 2^j, t = 2^j k)$. Les coefficients de la troisième forment également une partie de la première correspondant à ses maxima locaux. Il est important de noter que, dans la MMTO, les maxima du module sont chaînés pour former des lignes de maxima. Pour le calcul de $S_q(a)$, on remplace $T_X(a,t)$ par le maximum local $\sup_{a' \leq a} T_X(a', t(a'))$ obtenu le long de la ligne de maxima à laquelle il appartient [3, 10]. Les équations (5) sont définies pour toutes les formes de TO. Il est donc possible d'envisager chacun des estimateurs précédents pour chacune des déclinaisons de la TO. Nous voulons ici les comparer.

Simulation de CID. Suivant les algorithmes proposés dans [2], nous construisons des CID sur bases d'ondelettes orthogonales. On peut écrire $\forall x \in L^2(\mathbb{R})$, $\forall j_0 \in \mathbb{Z}$,

$$x(t) = \sum_{k \in \mathbb{Z}} a_X(j_0,k)\phi_{j_0,k}(t) + \sum_{j \leq j_0} \sum_{k \in \mathbb{Z}} d_X(j,k)\psi_{j,k}(t)\ , \quad (6)$$

type	I	I	I	I	II
C_1	-0.8	-0.8	-0.2	-0.2	-0.8
C_2	0.03	0.1	0.03/4	0.1/4	0.03
type	II*	II	II	III	III
C_1	-0.8	-0.2	-0.2	-0.2	-0.8
C_2	0.1	0.03/4	0.1/4	0.03/4	0.03

où les $\{\phi_{j_0,k}, k \in \mathbb{Z}\}$ et les $\{\psi_{j,k}, k \in \mathbb{Z}, j \leq j_0\}$, dilatées et translatées de la fonction d'échelle ϕ_0 et de l'ondelette-mère ψ_0, forment une base orthonormée. On choisit les coefficients d'approximation $\{a_X(j_0,k)\}_k$ égaux à 0. Les $\{d_X(j_0,k)\}_k$ suivent une loi normale de moyenne nulle et de variance égale à 1. Les coefficients $\{d_X(j,k)\}_{j,k}$ pour $j < j_0$ sont définit récursivement de la manière suivante :

$$d_{j-1,2k} = \quad W^{(1)}_{j-1,j}\, d_{j,k}\ ,\ \ d_{j-1,2k+1} = \quad W^{(2)}_{j-1,j}\, d_{j,k}.$$

Les $W_{j-1,j}$ sont définis par $\pm\exp(\omega_{j-1,j})$ où les $\omega_{j-1,j}$ sont des variables aléatoires i.i.d. dont les cumulants satisfont les deuxième et quatrième relations de (5).

3 Comparaisons des estimateurs

Protocole. Nous allons maintenant qualifier les performances statistiques des deux estimateurs, mis en œuvre sur les TOC, TOD et MMTO. Ces comparaisons sont conduites sur des CID synthétisées numériquement, avec trois catégories de fonction : (I) $n(a) \equiv \ln a$, (II) $n(a)$ log par morceaux, a_* désigne l'échelle de raccord choisie égale à 2^7 (forme observée dans l'étude du trafic Internet [14]), (III) $n(a) = a^{-\beta}/\beta$ (modèle utilisé en turbulence développée [3]). Les $\omega_{j-1,j}$ sont des variables alétoires normales de moyenne $m_{j-1,j} = C_1(n(2^j) - n(2^{j-1}))$ et variance $\sigma^2 = C_2(n(2^j) - n(2^{j-1}))$; tous les autres cumulants sont nuls. Différentes paires de paramètres C_1, C_2 sont envisagées, correspondant à des situations de difficultés diverses (faible ou fort écart de $H(q)$ à qH, cf. tableau). La taille des signaux synthétisés varie de $n = 2^{18}$ à 2^{25}.

Résultats. Les estimateurs sont mis en œuvre sur toute la gamme d'echelles disponible, de 2^3 à 2^{10}, gamme que l'on ne cherche pas à adapter. On ne se préoccupe ici que de performances d'estimations, et non de test de modèle (i.e., les CID sont supposées toujours valides). Plutôt que de détailler les estimations relatives à chaque famille de signaux, nous présentons les figures pour une seule famille représentative (repérée par * dans le tableau). Nos conclusions reposent néanmoins sur l'analyse de l'ensemble des figures de type 1 (ces figures pour chaque CID sont disponibles à *www.ens-lyon.fr/∼sroux*). Les performances d'estimation obtenues avec la TOC sont significativement moindres que celles obtenues avec la TOD ou la MMTO, Nous avons donc choisi de l'exclure dans la présentation des résultats.
En ce qui concerne la comparaison moments vs cumulants du log, on constate systématiquement que, malgré une bonne estimation du C_1, le C_2 n'est quasiment jamais bien estimé. Cet effet résulte de la difficulté à estimer les cumulants du log de quantités centrées en 0, difficulté d'autant plus grande que l'échelle est petite. La méthode MMTO, par construction, contourne cette difficulté. On constate néanmoins que le biais d'estimation reste important (plus de 20%) et plus C_2 est faible, meilleure est l'estimation. Cependant, les intervalles de confiance montre alors qu'il est impossible de le discriminer de $C_2 \equiv 0$. i.e., de distinguer entre processus multifractal ou autosimilaire. Dans tous les cas les estimations des paramètres reposant sur les moments sont beaucoup plus précises que celles basées sur les cumulants. On note également que toutes les estimations de $H(q)$ et C_l présentent un biais résiduel (i.e., qui ne décroît pas quand on augmente la durée de l'observation) alors que la variance des estimations décroît. Expérimentalement, ce biais semble croître avec C_2/C_1, indiquant que l'estimation de $H(q)$ est d'autant plus biaisée que $H(q)$ dévie de qH : il est très difficile d'estimer $H(q)$ lorsqu'il diffère de qH !
Au niveau de la comparaison TOD et MMTO, on constate que les performances d'estimation sont comparables, que ce soit pour les cumulants ou pour les moments. Cependant, les estimées par MMTO, $\hat{H}_p(q)$, exhibent systématiquement plus de courbure (ce qui ne signifie pas moins de biais) que celles issues de la TOD.
La fonction $n(a)$ est dans tous les cas toujours mieux estimée par les moments que par les cumulants. Cette estimation permet de discriminer entre les trois types de fonction $n(a)$ retenues et donc de discriminer les processus invariants d'échelle des autres.
En conclusion, nous dirons qu'en termes d'estimation (les conclusions pourraient être différentes s'il s'agissait de faire de la validation de modèle), l'utilisation des moments est clairement préférable à celle des cumulants. La MMTO ne semble pas apporter d'amélioration substantielle en dépit d'un coût de calcul significativement supérieur à celui de la TOD. Nous préférons donc l'utilisation de cette dernière. Notons aussi que l'on trouve toujours à l'oeil une gamme d'échelles où les estimations sont très bonnes. Mais cette gamme n'est pas la même entre cumulants et moments. Le choix "humain" de la gamme d'échelles de mesure reste donc un facteur substanciel d'amélioration.

4 Trafic Internet

Nous disposons de données TCP/IP enregistrées par le groupe WAND de l'université de Waikato (Nelle-Zélande), sur un lien haut-débit ATM connectant l'université d'Auckland au reste du monde. Celles-ci, décrites en détail sur *wand.cs.waikato.nz/wand/wits/index.html*, consistent en enregistrements d'une masse considérable d'information pendant plusieurs heures dont sont extraites diverses séries temporelles telles que le flux des paquets au protocole TCP/IP, le flux de connexions actives, le taux d'extinction, d'activation de connexions, le volume en bytes de ces connexions ou le délai entre connexions successives.

L'utilisation des outils précédemment décrits sur ces données montrent que celles-ci sont bien décrites par des CID, dont la caractéristique principale réside dans leurs fonctions $n(a)$. En effet, celles-ci diffèrent significativement de $\ln a$ et ressemblent davantage à des fonctions log par morceaux (justifiant le type (II) des signaux de synthèse choisi plus tôt). Les fonctions $H(q)$ observées diffèrent faiblement du comportement linéaire qH et les résultats des simulations numériques décrites précédemment nous invite à conclure que cet écart ne peux être considéré comme significatif. La conclusion est donc que la complexité des comportements en lois d'échelle dans les données Internet tient davantage à la forme de $n(a)$ qu'à celle de $H(q)$, justifiant à posteriori l'intérêt de n'avoir pas forcé $n(a) \equiv \ln a$ à priori. Des analyses plus complètes de ces données Internet peuvent être trouvées dans [14, 13].

Références

[1] P. Abry, P. Flandrin, M.S. Taqqu and D. Veitch, Wavelets for the analysis, estimation and synthesis of scaling data, dans [11], 39-88.

[2] A. Arneodo, E. Bacry, J.F. Muzy, Random cascade on wavelet dyadic trees, *Journal of Mathematical physics*, 39(8) : 4142-4164, 1998.

[3] A. Arneodo, J.F. Muzy, S.G. Roux, Experimental analysis of self-similar random cascade processes : application to fully developped turbulence, *J. Phys. II France*, 7 :363–370, 1997.

[4] A. Arneodo, J.F. Muzy, D. Sornette, Direct causal cascade in the stock market, *Eur. Phys. J. B*, 2 :277-282 1998

[5] B. Castaing, Y. Gagne, E. Hopfinger, Velocity probability density functions of high Reynolds number turbulence, *Physica D*, 46 :177,1990.

[6] P. Chainais, P. Abry et J.F. Pinton, Intermittency and coherent structures in a turbulent flow : a wavelet analysis of joint pressure and velocity measurements, *Phys. Fluids*, 11(11) :3524-3539, 1999.

[7] A.C. Gilbert, W. Willinger, A. Feldmann, Scaling analysis of random cascades, with applications to network traffic, *IEEE Trans. on Info. Theory*, Special Issue on Multiscale Statistical Signal Analysis and its Applications, 45(3) :971–991, 1999.

[8] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, On the self-similar nature of Ethernet traffic, Extended-Version, *IEEE/ACM Trans. on Networking*, 2 :1–15, 1994.

[9] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, Boston, 1997.

[10] J.F. Muzy, E. Bacry, A. Arneodo, The multifractal formalism revisited with wavelets, *Int. Journal of Bifurcation and Chaos*, 4(2) : 245–302, 1994.

[11] *Self-Similar Network Traffic and Performance Evaluation*, K. Park and W. Willinger, eds. Wiley Interscience, 2000.

[12] R. Riedi, M.S. Crouse, V.J. Ribeiro, R.G. Baraniuk, A Multifractal Wavelet Model with Application to Network Traffic, *IEEE Trans. on Info. Theory*, Special Issue on Multiscale Statistical Signal Analysis and its Applications, 45(3) :992–1018, April, 1999.

[13] S. Roux, D. Veitch, P. Abry, L. Huang, J. Micheel and P. Flandrin, Statistical scaling analysis of TCP/IP data using cascades, *Proc. of the Int. Conf. on Acoust. speech and Sig. Proc.*, Salt-Lake City, USA, 2001.

[14] D. Veitch, P. Abry, P. Flandrin and P. Chainais, Infinetely Divisible Cascade Analysis of network Trafic Data, *Proc. of the Int. Conf. on Acoust. speech and Sig. Proc.*, Istanbul, Turkey, 2000.

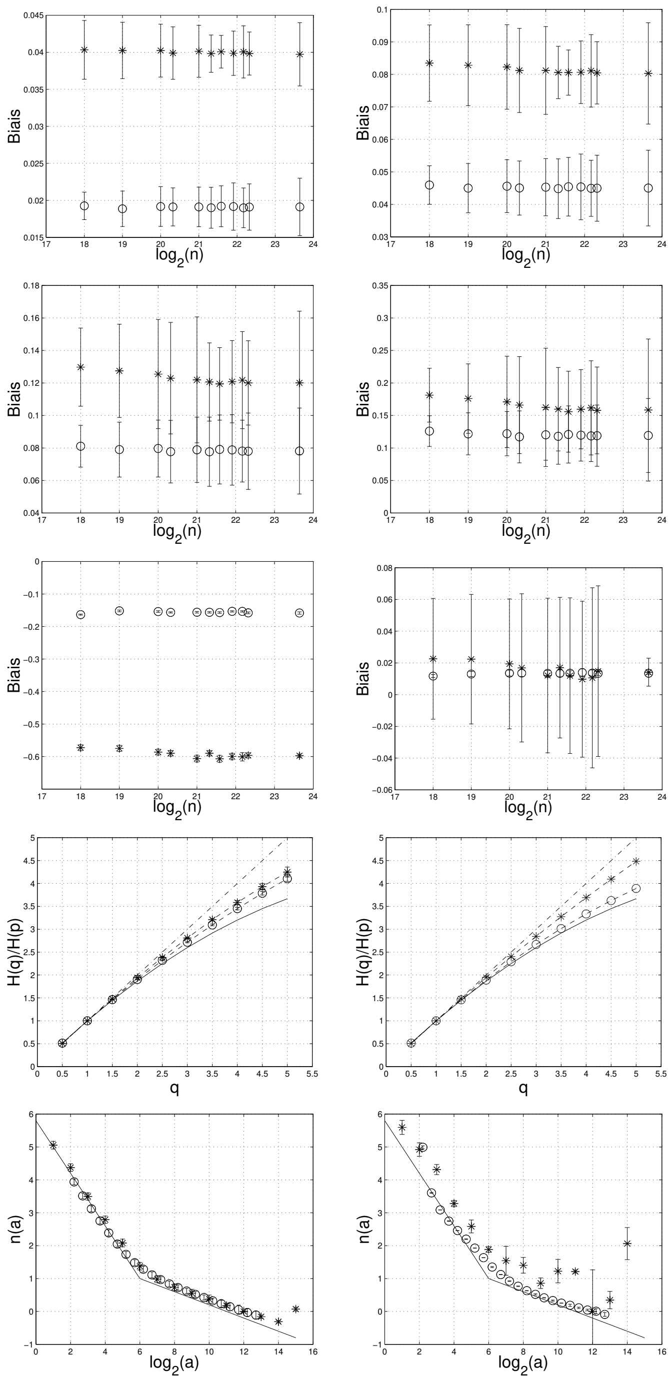

FIG. 1 – **Estimations.** Sur tous les figures, les $'*'$ correspondent à la TOD et les $'o'$ à la MMTO. Deux premiers rangs, biais relatifs de $\hat{H}_p(q)$ pour $p = 1$ et $q = 2, 3, 4, 5$; trosième rang : biais relatif de $\hat{C}_1(2)$ et absolu de $\hat{C}_1(3)$. Quatrième rang, estimées de $\hat{H}_p(q)$ par les moments (gauche) et les cumulants (droite), le trait plein indique les valeurs théoriques. Cinquième rang, estimées de $n(a)$ par les moments (gauche) et les cumulants (droite).