



HAL
open science

Etude expérimentale de l'influence d'un échantillonnage irrégulier dans l'estimation du paramètre de Hurst

Sullivan Hidot, Christophe Saint-Jean, Jean-Yves Lafaye

► **To cite this version:**

Sullivan Hidot, Christophe Saint-Jean, Jean-Yves Lafaye. Etude expérimentale de l'influence d'un échantillonnage irrégulier dans l'estimation du paramètre de Hurst. *Journal de la Société Française de Statistique*, 2008, 149 (1), 12 p. hal-00180113

HAL Id: hal-00180113

<https://hal.science/hal-00180113>

Submitted on 6 Jan 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ÉTUDE EXPÉRIMENTALE DE L'INFLUENCE D'UN ÉCHANTILLONNAGE IRRÉGULIER DANS L'ESTIMATION DU PARAMÈTRE DE HURST

Sullivan HIDOT¹, Christophe SAINT-JEAN²,
Jean-Yves LAFAYE¹

RÉSUMÉ

Dans cet article, nous proposons d'étudier un estimateur du paramètre de Hurst pour des trajectoires browniennes fractionnaires échantillonnées irrégulièrement. Les trajectoires sont simulées à partir de l'algorithme de Cholesky et l'estimation du paramètre de Hurst est obtenue par maximum de vraisemblance, ces deux techniques étant coûteuses en temps de calcul mais bien adaptées pour ce type de données. Nous présentons des tableaux contenant l'estimation de l'indice d'autosimilarité en fonction des méthodes d'échantillonnage et des tailles des signaux analysés. L'étude des tableaux est basée sur une série de tests statistiques (Student, Fisher) permettant de comparer les résultats des différentes irrégularités considérées. Plus l'échantillonnage est erratique, plus les résultats diffèrent de ceux attendus pour un échantillonnage régulier et cette différence tend à diminuer avec la taille des signaux. Les résultats obtenus pour un échantillonnage irrégulier sont plus proches du modèle régulier lorsque l'aléatoire est uniforme.

Mots-clés : Irrégularité d'échantillonnage, Mouvement brownien fractionnaire, Estimation du paramètre de Hurst, Estimateur du Maximum de Vraisemblance.

ABSTRACT

In this article, we propose to study an estimator of the Hurst parameter for irregularly sampled fractional Brownian trajectories. Trajectories are simulated by means of Cholesky algorithm, and the Hurst parameter is estimated by maximising likelihood. Both techniques are time consuming, but prove to be well suited to this type of data. We present various tables containing the estimates of the self-similarity measure, according to several sampling procedures with several sizes of trajectories. The study of these tables is based on a series of statistical tests (Student, Fisher), making it possible to compare and analyse the differences between the sampling processes on hand. The more erratic the sampling, the greater the discrepancy between the results and those expected for a regular sampling. This discrepancy tends to decrease when the size of the signals increases. Results from random

1. Laboratoire d'Informatique, Image et Interaction (L3I)
2. Laboratoire de Mathématiques et Applications (LMA), Avenue M. Crépeau, Université de La Rochelle, France.
{shidot,csaintje,jylafaye}@univ-lr.fr

sampleings are closer to those from a regular deterministic sampling when the random sampling model is uniform.

Keywords : Irregular sampling, Fractional Brownian Motion, Hurst parameter estimation, Maximum Likelihood estimator.

1. Introduction

Le mouvement brownien fractionnaire (fBm pour fractional Brownian motion), issu des travaux de Mandelbrot et Van Ness [MN68], fait partie d'une classe de processus généralisant le mouvement brownien. Il possède la propriété d'autosimilarité, quantifiée par le paramètre de Hurst H . Le fBm trouve de nombreuses applications en modélisation de signaux : biologie [Man83], analyse du mouvement [GW03], séries économiques [CS04], traitement d'image [KS86], réseau Ethernet [VV00]... En pratique, il est difficile d'affirmer qu'un signal possède un caractère autosimilaire. Certains tests sont proposés pour vérifier la bonne adéquation modèle/donnée [eGJ96, Bar99]. Cependant, ces tests ne s'appliquent pas dans le cas de signaux obtenus par échantillonnage irrégulier. La modélisation par un fBm pour ce type de données peut alors devenir inadéquate. La plupart des méthodes d'identification et d'estimation [Coe00, eGJ01, eGL94, PLV94] de H d'un fBm supposent un échantillonnage régulier. Ces méthodes nécessitent, dans le cas contraire, une modification en profondeur des algorithmes.

L'échantillonnage irrégulier des séries temporelles a donné lieu à beaucoup de travaux, et notamment en analyse spectrale [ST00]. Dans cet article, nous proposons une démarche expérimentale pour la modélisation par fBm d'un signal irrégulier dans le temps. Il s'agit d'une part de comparer les résultats obtenus avec des études antérieures faites sur un échantillonnage régulier [eGJ01] et d'autre part de déterminer la présence éventuelle d'un biais de l'estimateur. L'estimation de H se fait par maximum de vraisemblance [Leh80, KS86] où l'on a préalablement adopté le modèle proposé par Norros [Nor94].

La section 2 est consacrée aux définitions et propriétés de base du fBm. Une méthode de simulation de trajectoires irrégulières est présentée en section 3. Nous décrivons en section 4 l'algorithme d'estimation par maximum de vraisemblance estimant le paramètre H et prenant en compte l'irrégularité d'échantillonnage. Avant de conclure, nous présentons en section 5 les résultats expérimentaux avec un ensemble d'estimations de H obtenues en simulant des trajectoires avec différents types d'irrégularité temporelle. Nous comparons les résultats obtenus et évaluons la présence du biais dans l'estimation.

2. Définition et propriétés du mouvement brownien fractionnaire

Un fBm de paramètre de Hurst $H \in]0, 1[$, noté $(Z(t))_{t \in \mathbb{R}}$, est un processus gaussien centré réel vérifiant :

$$Z(0) = 0 \quad \text{et} \quad \forall (t, s) \in \mathbb{R}^2, \quad E(|Z(t) - Z(s)|^2) = |t - s|^{2H} \quad (1)$$

On en déduit l'expression de la fonction de covariance :

$$\forall (t, s) \in \mathbb{R}^2, \quad E(Z(t)Z(s)) = \frac{1}{2} (|t|^{2H} + |s|^{2H} - |t - s|^{2H}) \quad (2)$$

Le processus $(Z(t))_{t \in \mathbb{R}}$ est autosimilaire d'ordre H et a ses accroissements stationnaires :

$$\forall (t, \alpha) \in \mathbb{R} \times \mathbb{R}^{*+}, \quad Z(\alpha t) \sim \alpha^H Z(t) \quad (3)$$

$$\forall (t, s) \in \mathbb{R}^2, \quad Z(t) - Z(s) \sim Z(t - s) \quad (4)$$

où \sim désigne l'égalité en loi.

$(Z(t))_{t \in \mathbb{R}}$ est l'unique processus gaussien centré continu autosimilaire à accroissements stationnaires [Her04]. En particulier, pour $H = 1/2$, on retrouve le mouvement brownien classique où les accroissements sont de plus indépendants. La densité spectrale $f(\omega)$ des accroissements d'un fBm se comporte comme $|\omega|^{1-2H}$ quand ω est proche de 0, ce qui est une caractéristique remarquable des processus à longue-mémoire [Ber94] (ou $1/f$).

Soit $k(., .)$ le noyau d'intégration fractionnaire de $\mathcal{L}^2(\mathbb{R})$ défini de la manière suivante :

$$\forall (t, \lambda) \in \mathbb{R}^2, \quad k(t, \lambda) = |t - \lambda|^{H-1/2} - |t|^{H-1/2} \quad (5)$$

Le fBm a une représentation dite moyenne mobile [ST94]. Chacune des variables du processus $Z(t)$ pour t fixé s'écrit comme une somme de gaussiennes centrées réduites pondérée par les coefficients du noyau $k(., .)$:

$$\forall t \in \mathbb{R}, \quad Z(t) = \int_{\mathbb{R}} k(t, \lambda) dW(\lambda) \sim \mathcal{N}(0, \|k(t, .)\|^2) \quad (6)$$

où W est une mesure brownienne, i.e. une isométrie de $\mathcal{L}^2(\mathbb{R})$ vers un espace gaussien. Le processus $(dW(\lambda))_{\lambda \in \mathbb{R}}$ est aussi appelé bruit blanc. Dans le cas d'un processus à états discrets, les variables sont également des gaussiennes et chaque trajectoire est obtenue en sommant une réalisation d'une gaussienne avec une pondération particulière, ce que nous verrons dans la prochaine section avec l'algorithme de Cholesky.

3. Simulation de trajectoires échantillonnées irrégulièrement

Il existe de nombreuses méthodes de simulation de trajectoires browniennes fractionnaires : approches stochastiques, multi-échelle, spectrales, par la matrice de covariance du processus [AS96, Coe00, DM03, eGJ01, Sel95, WC94]... La plupart de ces méthodes supposent que la période d'échantillonnage soit constante. Nous avons choisi de simuler des trajectoires à partir de la matrice de covariance du processus en utilisant la méthode de Cholesky, qui prend en compte l'irrégularité d'échantillonnage via l'équation (2). Les méthodes à base d'ondelettes, celles de type Paxson ou de simulations spectrales, fournissant de bonnes simulations, peuvent être utilisées. Une manière simple serait de

simuler une trajectoire régulière puis de la ré-échantillonner avec un pas non constant.

L'algorithme de Cholesky se formule de la manière suivante. Soient (t_2, \dots, t_N) les instants d'échantillonnage d'une trajectoire de longueur $N - 1$ et H son paramètre de Hurst. La particularité des fBm est que la matrice de covariance du processus est entièrement déterminée par les t_i et H . On calcule la matrice de covariance associée via (2) :

$$\Gamma_H(i, j) = E(Z(t_i)Z(t_j)) = \frac{1}{2} (|t_i|^{2H} + |t_j|^{2H} - |t_i - t_j|^{2H}) \text{ avec } 2 \leq i, j \leq N \quad (7)$$

La matrice Γ_H est symétrique définie positive et peut donc être décomposée par la méthode de Cholesky :

$$\Gamma_H = LL' \quad (8)$$

où L est la matrice triangulaire inférieure dans la décomposition de Cholesky. Si y est un échantillon de longueur $N - 1$ d'une variable gaussienne centrée réduite, alors l'ensemble $(0, Ly)$ est une trajectoire de longueur N d'un fBm aux instants (t_1, \dots, t_N) avec $t_1 = 0$ et d'autosimilarité H .

En général, on applique la méthode de Cholesky à la matrice de covariance, notée V , de la dérivée du processus à une certaine résolution, appelée aussi bruit gaussien fractionnaire (ou fGn pour fractional Gaussian noise). La méthode de Cholesky appliquée à V est dite exacte car la matrice de covariance calculée dans le cas discret coïncide avec le cas continu. De plus, pour gagner en temps de calcul, une alternative est d'opter pour l'algorithme de Levinson [Pel98] qui permet de reconstruire la matrice dans la décomposition de Cholesky à partir des éléments de la première ligne de V . Dans le cas d'un échantillonnage irrégulier, cet algorithme n'est plus applicable puisque R n'a plus nécessairement la propriété Toeplitz. L'irrégularité des échantillons simulés est obtenue en contrôlant les valeurs du paramètre temporel t qui constitue la base du calcul de la matrice de covariance Γ_H .

4. Estimation de l'autosimilarité

4.1. Modèle théorique

Le modèle proposé par Norros [Nor94], utilisé pour l'étude du trafic Ethernet, exprime un fBm Z avec une tendance linéaire à estimer *a posteriori* qui n'est pas à prendre en compte en raison de la méthode de simulation. Sans cette tendance, le modèle s'exprime de la manière suivante :

$$X(t) = \sqrt{a}Z(t) \quad (9)$$

Le processus Z est un mouvement brownien fractionnaire pur, i.e. un processus gaussien d'autosimilarité H , de moyenne nulle et de matrice de covariance Γ_H . La quantité a est la variance de la variable $X(1)$.

Dans ce cas, X est un fBm d'autosimilarité H et de covariance égale à :

$$E(X(t_i)X(t_j)) = aE(Z(t_i)Z(t_j)) = a\Gamma_H(i, j) \quad (10)$$

Nous avons finalement deux paramètres à considérer (a, H) que nous estimons par l'approche du maximum de vraisemblance [Leh80].

4.2. Estimateur du maximum de vraisemblance (EMV)

Cet estimateur a été adapté par Dahlhaus [Dah89] pour le calcul du coefficient d'autosimilarité d'un fBm. L'intérêt de cette méthode est qu'elle bénéficie de bonnes propriétés asymptotiques [Ber94]. En particulier, l'estimateur \hat{H}_{EMV} converge presque sûrement vers la valeur théorique H quand N grandit. Le choix de l'EMV est motivé par le fait qu'il ne nécessite pas de connaître *a priori* la densité spectrale de puissance (DSP) du signal, contrairement à l'estimateur de Whittle [FT86]. Parmi les méthodes d'estimation de la DSP, la plus simple est la méthode du périodogramme. Pour un échantillonnage arbitraire, Scargle [Sca82] en propose une version généralisée, qui possède les mêmes propriétés statistiques que le périodogramme classique. Cet estimateur est asymptotiquement non biaisé mais nécessite de recourir à des techniques de lissage local [Yue78] pour diminuer sa variance intrinsèquement élevée [BD91]. Par conséquent, la DSP est estimable quelque soit l'échantillonnage mais l'approximation de Whittle nécessite de connaître le périodogramme d'une trajectoire de l'accroissement, qui n'est pas directement observable dans le cas d'un échantillonnage non uniforme.

L'estimateur du maximum de vraisemblance se formule de la manière suivante. Soit $\mathbf{x} = (x_1, \dots, x_N)'$ une trajectoire de X aux instants d'échantillonnage $\mathbf{t} = (t_1, \dots, t_N)'$. On cherche à maximiser par rapport à (a, H) une probabilité de vraisemblance donnée par la fonction de densité multivariée :

$$L(\mathbf{x}; a, H) = (2\pi)^{-\frac{N}{2}} (a^N |\Gamma_H|)^{-\frac{1}{2}} e^{-\frac{1}{2a} \mathbf{x}' \Gamma_H^{-1} \mathbf{x}} \quad (11)$$

où $|\cdot|$ désigne le déterminant. Le logarithme de la vraisemblance est donné par :

$$\begin{aligned} \log L(\mathbf{x}; a, H) &= -\frac{N}{2} \log(2\pi) - \frac{1}{2} \log(a^N |\Gamma_H|) - \frac{1}{2a} \mathbf{x}' \Gamma_H^{-1} \mathbf{x} \\ &= -\frac{N}{2} \log(2\pi) - \frac{N}{2} \log(a) - \frac{1}{2} \log(|\Gamma_H|) - \frac{1}{2a} \mathbf{x}' \Gamma_H^{-1} \mathbf{x} \end{aligned} \quad (12)$$

Maximiser le logarithme de la vraisemblance revient à maximiser la quantité :

$$-N \log(a) - \log |\Gamma_H| - \frac{1}{a} \mathbf{x}' \Gamma_H^{-1} \mathbf{x} \quad (13)$$

L'estimé de \hat{a} est obtenu par la condition nécessaire que la dérivée de (13) par rapport à a soit nulle. On obtient :

$$\hat{a} = \frac{1}{N} \mathbf{x}' \Gamma_H^{-1} \mathbf{x} \quad (14)$$

En incorporant l'équation (14) dans (13), le problème consiste dorénavant à estimer H . Aux constantes près, cela se résume à maximiser relativement à H la fonction objectif suivante :

$$-N \log \left[\frac{1}{N} \mathbf{x}' \Gamma_H^{-1} \mathbf{x} \right] - \log |\Gamma_H| \quad (15)$$

Pour maximiser la fonction objectif, nous appliquons une méthode quasi-Newton. En toute rigueur, il faudrait démontrer la concavité de (13) considérée comme une fonction de H . Il s'agit d'une question ardue et nous nous limiterons à une vérification expérimentale.

Dans le cas d'un échantillonnage quelconque, Lai [Lai04] montre que l'estimateur \hat{H} du paramètre de Hurst d'un fBm est asymptotiquement gaussien, sans biais et d'écart-type :

$$\sigma(\hat{H}_{EMV}) = \frac{2}{\text{trace}([\Gamma_H^{-1} \Gamma'_H]^2)} \quad (16)$$

où Γ'_H est la matrice dérivée de Γ_H :

$$\Gamma'_H(i, j) = \frac{\partial \Gamma_H(i, j)}{\partial H} \quad (17)$$

5. Résultats expérimentaux

Nous proposons de considérer deux types d'irrégularité temporelle dans l'échantillonnage. Dans les deux cas, nous procédons par sous-échantillonnage dans un échantillon régulier de référence : $\{t_i\}_{i=1}^n$ avec $t_i - t_{i-1} = 1$. Le premier sous-échantillonnage noté Irr_S est systématique, et considère alternativement des instants consécutifs et des instants séparés de deux unités de temps, à savoir : $\dots, t_i, t_{i+1}, t_{i+3}, t_{i+4}, \dots$. Le second sous-échantillonnage noté Irr_U est aléatoire et s'obtient par tirage uniforme dans l'échantillon de référence.

5.1. Exemple simple

Nous commençons par un exemple simple montrant les divergences des résultats suivant l'échantillonnage. Soient deux signaux simulés de taille $N = 1024$ avec $H = 0.6$ par Cholesky avec respectivement une régularité d'échantillonnage ($t_i = i$) et une irrégularité issue de la loi uniforme (Irr_U). Ces deux signaux ont été obtenus à partir du même échantillon gaussien (noté y en section 3). Nous estimons par l'approximation de Whittle (échantillonnage supposé régulier) et par EMV les coefficients d'autosimilarité

des deux signaux. Les résultats sont présentés dans le tableau suivant :

	Signal 1 (<i>reg.</i>)	Signal 2(<i>Irr_U</i>)
Whittle	0.565	0.531
EMV	0.562	0.570

On constate sur cet exemple que l'estimateur de Whittle sous-estime la valeur théorique de H . L'estimateur proposé donne une meilleure estimation (proche de 0.6) avec une différence moins significative. Par conséquent, l'irrégularité d'échantillonnage semble jouer un rôle important, ce qui nous amène à étudier son influence dans l'estimation de H . Cet exemple illustratif ne permet pas de mesurer la réelle efficacité de l'EMV et il convient donc, pour une certaine longueur N , de procéder à une étude systématique effectuée en répétant les procédures expérimentales de simulation et d'estimation, de manière à étudier la distribution d'échantillonnage de l'estimateur de H , son biais et sa précision.

5.2. Choix des paramètres, stratégie de tests

Pour mener l'analyse, nous nous référons à l'article de Jennane et coll. [eGJ01]. La démarche est la suivante. Ayant fixé un mode d'échantillonnage, nous étudions des signaux de différentes tailles (soit N) pour différentes valeurs de l'exposant de Hurst (H). Pour un mode d'échantillonnage et pour N et H fixés, nous répétons R fois les opérations de simulation d'un fBm (sous les contraintes précédentes) et d'estimation de H . On obtient ainsi un échantillon de taille suffisante pour approcher la distribution de l'estimateur \hat{H} du maximum de vraisemblance. Nous étudions alors expérimentalement les propriétés de biais et de convergence de l'estimateur en question. Le choix de R est fondé sur le contrôle des risques et de la puissance des tests de comparaison de moyennes et de variances que nous allons mener. Conformément à l'approche de Jennane et coll., nous adoptons pour le calcul de R l'hypothèse selon laquelle la variance de l'estimateur serait égale à la borne de Cramer-Rao correspondante, dont leur article donne la valeur. Cette borne dépend de H , et l'on adopte une stratégie « pessimiste » en choisissant la valeur maximum tout H confondus pour un N fixé. En supposant que la densité de l'estimateur soit gaussienne et avec un niveau de confiance de 1%, on trouve le tableau :

N	32	64	128	256	512	1024
R	3053	1492	750	383	187	96

Dans ce qui suit, nous comparons les résultats obtenus pour chaque irrégularité d'échantillonnage avec le cas classique d'estimation (régularité) et nous étudions l'influence de l'irrégularité temporelle sur le biais de l'estimateur. Toute la démarche est résumée en figure 1.

ÉTUDE EXPÉRIMENTALE DE L'INFLUENCE D'UN ÉCHANTILLONNAGE

Afin de valider les observations, nous avons procédé à des tests de comparaison d'échantillons en nous intéressant à la précision et à la présence éventuelle d'un biais et ce, pour les deux irrégularités. Dans tous les cas, nous considérons un risque de première espèce α égal à 0.01. Nous appliquons successivement (« *Fish + Stud*₁ » fig. 1), dans le cas gaussien, le test de Fisher de comparaison des variances puis le test de Student de comparaison des moyennes [Sap90]. L'hypothèse nulle H_0 du test de Fisher est l'homogénéité des variances des populations dont sont issus les deux échantillons, l'alternative étant l'hétérogénéité des variances. Puisque ces derniers ont la même taille, la statistique de Fisher à $R - 1$ degrés de liberté est simplement le rapport entre s_1^2 et s_2^2 , estimateurs empiriques des variances :

$$F(R - 1, R - 1) = \frac{s_1^2}{s_2^2} \quad \text{avec} \quad s_1^2 > s_2^2 \quad (18)$$

Pour une irrégularité donnée, ce test est appliqué pour chacune des cases homologues des deux tableaux. Pour les cas où il n'y a pas rejet de H_0 , nous appliquons le test de Student où l'hypothèse nulle est l'égalité des moyennes $E(H_1)$ et $E(H_2)$. La statistique de test est la suivante :

$$T_{2R-2} = \sqrt{\frac{R-1}{s_1^2 + s_2^2}} (\bar{H}_1 - \bar{H}_2) \quad (19)$$

où \bar{H}_1 et \bar{H}_2 sont les moyennes empiriques de l'estimateur. Si l'hypothèse nulle des deux tests n'est pas rejetée, on conclut que l'irrégularité de l'échantillonnage n'a pas d'influence sur les propriétés de l'estimateur.

Afin de juger du biais de l'estimateur, nous calculons l'intervalle de confiance (IC) au seuil α (« *Stud*₂ » fig. 1) et vérifions qu'il contient la valeur théorique H du paramètre de Hurst. Une estimation de H étant indépendante des autres, on applique le théorème central limite sur \bar{H} et on en déduit que \bar{H} suit approximativement une loi gaussienne. Puisque l'écart-type est inconnu, la statistique $\frac{\bar{H} - H}{s/\sqrt{R-1}}$ de l'échantillon de moyenne \bar{H} et d'écart-type empirique s suit approximativement une loi de Student T_{R-1} à $R - 1$ degrés de liberté, ce qui donne l'encadrement :

$$-t_{\alpha/2} < \frac{\bar{H} - H}{s/\sqrt{R-1}} < t_{\alpha/2} \quad (20)$$

où $t_{\alpha/2}$ est $(1 - \alpha)/2$ percentile de T_{R-1} . On en déduit l'IC :

$$\bar{H} - t_{\alpha/2} \frac{s}{\sqrt{R-1}} < H < \bar{H} + t_{\alpha/2} \frac{s}{\sqrt{R-1}} \quad (21)$$

De cette façon, on obtient l'IC qui permet de vérifier l'appartenance ou non de la valeur théorique choisie pour la simulation. Le calcul de l'IC est analogue à tester l'hypothèse nulle $E(\hat{H}) = H$ avec la même statistique $\frac{\bar{H} - H}{s/\sqrt{R-1}}$.

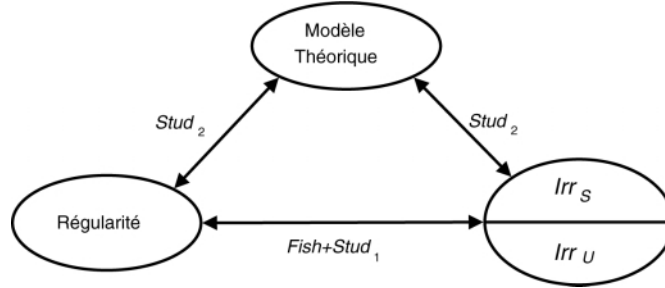


FIG 1. — Stratégie de tests pour la comparaison et l'étude du biais de l'estimateur de H par EMV. $Stud_2$: Test de Student (Intervalle de confiance); $Fish + Stud_1$: Test de Fisher + Test de Student (Comparaison d'échantillons).

5.3. Comparaison d'échantillons ($Fish + Stud_1$)

L'ensemble des calculs ont été effectués à l'aide du logiciel statistique R [R D05]. Les tableaux présentant les résultats figurent en annexe. Le tableau 1 contient les moyennes et les écart-types obtenus pour un échantillonnage régulier et croise les valeurs de H avec la taille des signaux. Les valeurs des tableaux sont arrondies mais les tests sont menés sur les valeurs exactes. On retrouve les propriétés usuelles de la méthode d'estimation : l'écart-type de l'estimateur diminue quand N grandit, illustrant ainsi le fait que l'EMV est convergent. Globalement, les moyennes obtenues sous-estiment les valeurs de H .

TABEAU 1. — Estimations par EMV avec régularité d'échantillonnage.

$H \setminus N$	32	64	128	256	512	1024
0.1	0.113±0.068	0.100±0.047	0.098±0.034	0.099±0.024	0.098±0.016	0.098±0.011
0.2	0.202±0.093	0.199±0.064	0.200±0.043	0.201±0.030	0.199±0.021	0.198±0.014
0.3	0.291±0.107	0.295±0.073	0.298±0.050	0.299±0.034	0.297±0.025	0.298±0.018
0.4	0.386±0.115	0.391±0.080	0.394±0.056	0.397±0.038	0.401±0.025	0.401±0.017
0.5	0.485±0.117	0.494±0.081	0.497±0.058	0.499±0.039	0.498±0.029	0.498±0.021
0.6	0.581±0.119	0.592±0.082	0.597±0.055	0.599±0.041	0.599±0.030	0.601±0.020
0.7	0.676±0.115	0.685±0.081	0.691±0.056	0.694±0.040	0.699±0.031	0.701±0.019
0.8	0.769±0.110	0.788±0.078	0.795±0.058	0.798±0.043	0.801±0.030	0.803±0.021
0.9	0.858±0.092	0.875±0.068	0.887±0.050	0.892±0.038	0.899±0.028	0.900±0.019

Le tableau 2 présente les mêmes estimations en considérant le cas Irr_S . Tous les écart-types obtenus sont inférieurs ou égaux à ceux du tableau 1. L'irrégularité Irr_S réduit donc l'écart-type de l'estimateur et les valeurs théoriques de H sont sous-estimées. Nous avons comparé ces résultats avec ceux obtenus dans le tableau 1 à l'aide des tests statistiques expliqués plus haut. Les cellules en gris clair (resp. foncé) correspondent au non rejet de H_0 pour le test de Fisher (resp. Fisher et Student). Les non-rejets sont comparables pour les deux tableaux à partir de la taille $N = 512$. Pour la taille maximale 1024, les résultats obtenus dans les tableaux 1 et 2 sont similaires.

ÉTUDE EXPÉRIMENTALE DE L'INFLUENCE D'UN ÉCHANTILLONNAGE

TABLEAU 2. — Estimations par EMV avec Irr_S , comparaison avec le tableau 1. Les cellules en gris clair (resp. foncé) correspondent au non rejet de H_0 pour le test de Fisher (resp. Fisher et Student).

H \ N	32	64	128	256	512	1024
0.1	0.112±0.066	0.103±0.045	0.102±0.032	0.102±0.022	0.102±0.015	0.102±0.011
0.2	0.197±0.086	0.195±0.059	0.196±0.040	0.197±0.027	0.197±0.019	0.198±0.014
0.3	0.293±0.098	0.296±0.069	0.299±0.047	0.300±0.033	0.300±0.024	0.300±0.017
0.4	0.390±0.105	0.394±0.071	0.397±0.048	0.398±0.033	0.398±0.023	0.398±0.016
0.5	0.483±0.110	0.490±0.073	0.493±0.054	0.495±0.038	0.496±0.024	0.497±0.017
0.6	0.586±0.111	0.592±0.076	0.598±0.053	0.601±0.038	0.601±0.026	0.602±0.018
0.7	0.679±0.107	0.691±0.075	0.693±0.054	0.694±0.038	0.698±0.027	0.697±0.018
0.8	0.774±0.099	0.787±0.070	0.793±0.052	0.797±0.039	0.799±0.027	0.797±0.019
0.9	0.863±0.087	0.880±0.062	0.888±0.047	0.895±0.035	0.898±0.025	0.897±0.019

Le tableau 3 contient les estimés avec l'irrégularité Irr_U . Les résultats obtenus sont nettement meilleurs qu'avec les deux cas précédents en terme de variance. Tous les écart-types sont inférieurs ou égaux à ceux des deux autres tableaux et la différence est plus forte globalement pour des valeurs de H autour de 0.5. Les non-rejets sont comparables pour les tableaux 1 et 3 à partir d'une taille comprise entre 512 et 1024. Seuls les moyennes et écart-types des échantillons pour la taille 1024 se rapprochent des résultats obtenus avec un échantillonnage régulier. On constate que pour une irrégularité périodique (Irr_S), les valeurs des estimés sont plus proches du cas régulier, les tests de comparaison d'échantillons étant beaucoup plus rejetés dans le tableau 3. En conclusion, plus l'échantillonnage est erratique, plus la convergence des estimateurs est lente et cette particularité tend à diminuer avec la taille des signaux. L'échantillonnage n'influence significativement les estimations que pour des signaux de longueur faible.

TABLEAU 3. — Estimations par EMV avec Irr_U , comparaison avec le tableau 1.

H \ N	32	64	128	256	512	1024
0.1	0.109±0.060	0.101±0.040	0.101±0.029	0.099±0.020	0.101±0.013	0.101±0.010
0.2	0.204±0.077	0.202±0.053	0.202±0.036	0.202±0.025	0.201±0.017	0.201±0.013
0.3	0.300±0.087	0.300±0.060	0.300±0.038	0.302±0.027	0.300±0.020	0.302±0.015
0.4	0.396±0.090	0.398±0.062	0.397±0.042	0.399±0.028	0.397±0.020	0.402±0.016
0.5	0.495±0.092	0.496±0.064	0.502±0.045	0.500±0.032	0.501±0.022	0.502±0.017
0.6	0.589±0.091	0.594±0.064	0.596±0.044	0.599±0.029	0.598±0.021	0.602±0.018
0.7	0.692±0.092	0.698±0.064	0.695±0.045	0.699±0.032	0.697±0.022	0.701±0.018
0.8	0.784±0.086	0.791±0.062	0.797±0.044	0.801±0.032	0.798±0.021	0.801±0.018
0.9	0.877±0.071	0.887±0.054	0.897±0.041	0.898±0.031	0.897±0.021	0.901±0.018

5.4. Influence de l'irrégularité d'échantillonnage sur le biais de l'estimateur ($Stud_2$)

5.4.1. *Test de Student*

Dans cette section, nous étudions l'importance du biais de l'estimateur en fonction du choix de l'irrégularité d'échantillonnage. Pour cela, et en complémentarité avec le calcul de l'IC, nous appliquons un test de Student d'égalité de l'espérance de l'estimateur avec la valeur théorique ($H_0 : E(\hat{H}) = H$). Les tableaux 4, 5 et 6 présentent l'ensemble des p -values du test pour chaque type d'échantillonnage. Puisque le test est bilatéral symétrique, il faut comparer les résultats obtenus avec la moitié du seuil de référence choisi initialement ($\frac{\alpha}{2} = 0.005$). Le test de Student est équivalent à déterminer si la valeur théorique appartient à l'IC mais a l'avantage de préciser l'analyse par l'intermédiaire des p -values. Pour les trois tableaux, le non-rejet du test correspond globalement aux signaux de grande longueur. Plus la taille des signaux est grande et plus la valeur théorique appartient à l'IC, d'où le non-biais asymptotique de l'estimateur. Il y a 35 non-rejets pour la régularité d'échantillonnage, 30 pour Irr_S , 42 pour Irr_U et nous obtenons donc de ce point de vue les meilleurs résultats avec Irr_U . La moyenne des p -values des non-rejets est de 0.221 pour la régularité d'échantillonnage, 0.133 pour Irr_S et 0.168 pour Irr_U . La régularité d'échantillonnage accepte moins de tests que Irr_U , mais les p -values sont globalement plus élevées. Dans un intervalle fixé, plus le nombre d'observations est grand et plus l'échantillonnage Irr_U tend vers un échantillonnage régulier. Ce qui explique que les résultats de Irr_U se rapprochent de la régularité d'échantillonnage en termes de p -values quand N grandit.

TABLEAU 4. — Test de Student ($H_0 : E(\hat{H}) = H$) : ensemble des p -values pour le cas régulier. Les cellules grisées correspondent au non-rejet du test (p -values $> \frac{\alpha}{2}$).

H\N	32	64	128	256	512	1024
0.1	$< 10^{-16}$	0.426	0.107	0.229	0.076	0.081
0.2	0.173	0.241	0.409	0.253	0.268	0.103
0.3	2.428×10^{-6}	0.005	0.108	0.246	0.081	0.141
0.4	5.211×10^{-12}	3.333×10^{-6}	0.003	0.066	0.383	0.350
0.5	1.636×10^{-13}	0.001	0.088	0.387	0.183	0.234
0.6	$< 10^{-16}$	5.158×10^{-5}	0.047	0.397	0.360	0.248
0.7	$< 10^{-16}$	5.535×10^{13}	3.269×10^{-6}	0.002	0.316	0.258
0.8	$< 10^{-16}$	4.351×10^{-9}	0.006	0.216	0.406	0.124
0.9	$< 10^{-16}$	$< 10^{-16}$	2.526×10^{-13}	8.625×10^{-6}	0.277	0.426

ÉTUDE EXPÉRIMENTALE DE L'INFLUENCE D'UN ÉCHANTILLONNAGE

TABLEAU 5. — Test de Student ($H_0 : E(\hat{H}) = H$) : ensemble des p -values pour le cas Irr_S .

H \ N	32	64	128	256	512	1024
0.1	$< 10^{-16}$	0.004	0.056	0.048	0.021	0.046
0.2	0.024	4.757×10^{-4}	0.004	0.018	0.023	0.054
0.3	6.990×10^{-5}	0.011	0.202	0.440	0.460	0.437
0.4	2.341×10^{-7}	0.001	0.060	0.082	0.141	0.137
0.5	1.110×10^{-16}	2.763×10^{-7}	1.022×10^{-4}	0.003	0.013	0.052
0.6	9.403×10^{-13}	5.379×10^{-5}	0.103	0.360	0.315	0.133
0.7	$< 10^{-16}$	8.776×10^{-7}	3.920×10^{-4}	0.001	0.122	0.077
0.8	$< 10^{-16}$	1.098×10^{-13}	2.847×10^{-4}	0.073	0.243	0.078
0.9	$< 10^{-16}$	$< 10^{-16}$	1.171×10^{-11}	0.004	0.103	0.069

TABLEAU 6. — Test de Student ($H_0 : E(\hat{H}) = H$) : ensemble des p -values pour le cas Irr_U .

H \ N	32	64	128	256	512	1024
0.1	4.440×10^{-16}	0.263	0.226	0.109	0.153	0.125
0.2	0.003	0.052	0.050	0.060	0.115	0.153
0.3	0.432	0.439	0.488	0.084	0.435	0.168
0.4	0.005	0.086	0.034	0.163	0.036	0.179
0.5	0.002	0.012	0.074	0.384	0.273	0.190
0.6	4.393×10^{-11}	6.702×10^{-5}	0.008	0.228	0.129	0.200
0.7	2.078×10^{-6}	0.097	0.003	0.317	0.059	0.211
0.8	$< 10^{-16}$	2.664×10^{-9}	0.023	0.271	0.079	0.223
0.9	$< 10^{-16}$	$< 10^{-16}$	0.018	0.107	0.030	0.253

5.4.2. Analyse des p -values

En analysant le maximum des p -values par colonne, on observe une certaine structure pour chacun des tableaux. Pour la régularité d'échantillonnage, le maximum est atteint pour $H = 0.1$, pour des signaux de longueur faible, jusqu'à $H = 0.9$ pour des signaux longs. Pour Irr_S , le maximum est atteint pour $H = 0.3$ quelque soit N et pour Irr_U , la dynamique est plus aléatoire mais concerne aussi pour la plupart des cas $H = 0.3$. L'irrégularité d'échantillonnage a donc moins d'influence pour les H proches de 0.1 car les variations des signaux sont plus grandes. Par conséquent, l'irrégularité Irr_S rajoute globalement un biais. Pour Irr_U , les valeurs obtenues avec une irrégularité uniforme sont plus proches du modèle théorique : le test de Student pour l'IC accepte 42 valeurs contre 30 pour Irr_S et 35 pour la régularité d'échantillonnage.

Nous concluons donc un biais réduit pour la régularité d'échantillonnage mais aussi pour Irr_U . L'irrégularité « déterministe » Irr_S donne significativement de moins bons résultats, ce qui coïncide avec les analyses de l'IC, impliquant un biais plus prononcé.

6. Conclusion

Nous nous sommes fixés l'objectif de mener une étude expérimentale sur l'estimation du paramètre de Hurst de mouvements browniens fractionnaires dans le cas où l'échantillonnage temporel n'est pas régulier. Notre travail présente donc trois volets. Le premier traite de la simulation de mouvements browniens fractionnaires, le second de leur identification, et le troisième de l'analyse des résultats expérimentaux. Les techniques habituelles de simulation supposent un échantillonnage temporel régulier, ce qui confère aux matrices utilisées dans les calculs des formes particulières qui ne sont pas de mise en échantillonnage irrégulier. Les techniques de simulation ont ainsi dû être contrôlées et adaptées. Les meilleurs estimateurs du paramètre H sont obtenus par maximum de vraisemblance. Dans la littérature scientifique sur le sujet, la résolution et le calcul de tels estimateurs supposent une fois encore un échantillonnage régulier. Nous avons adapté les techniques classiques au cas irrégulier tout en limitant la combinatoire. Nous avons délibérément repris les protocoles expérimentaux dans le cas d'un échantillonnage régulier, proposés dans [VV00] et [eGJ01], de manière à ce que nos résultats soient directement comparables avec les leurs sans aucun biais méthodologique.

De manière générale, nous confirmons naturellement les résultats théoriques liés aux estimateurs du maximum de vraisemblance, et en particulier, le caractère asymptotiquement sans biais et la convergence. L'intérêt majeur de notre travail est de permettre la validation des techniques de simulation et d'estimation dans le cas général. Nous fournissons des résultats quantitatifs sous forme d'intervalles de confiance, qui permettent de cerner la précision des estimateurs pour différentes tailles de signaux et différentes valeurs du paramètre de Hurst. Nous avons également envisagé plusieurs types d'irrégularités d'échantillonnage, l'une systématique avec un échantillonnage périodique, l'autre aléatoire avec un échantillonnage aléatoire uniforme. Cette dernière analyse confirme en particulier la plus grande sensibilité (biais supérieur) de l'estimation du maximum de vraisemblance dans le cas systématique par rapport au cas aléatoire.

Bibliographie

- [AS96] P. ABRY and F. SELLAN. The wavelet-based synthesis for fractional brownian motion. *Applied and Computational Harmonic Analysis*, 3(4) : 377–383, 1996.
- [Bar99] J.M. BARDET. Un test d'autosimilarité pour les processus gaussiens à accroissements stationnaires. *C.R. Acad. Sci. Paris, Série I Math.*, (6) : 521–526, 1999.
- [BD91] P.J. BROCKWELL and R.A. DAVIS. *Time Series : Theory and Methods*. Second edition, Springer, 1991.
- [Ber94] J. BERAN. *Statistics for Long-Memory Processes*. Chapman and Hall, New York, 1994.

- [Coe00] J.-F. COEURJOLLY. Simulation and identification of the fractional Brownian motion : a bibliographical and comparative study. *Journal of Statistical Software*, 5 : 1–53, 2000.
- [CS04] A. CARBONE, G. CASTELLI and H.E. STANLEY. Time-dependent hurst exponent in financial time series. *Physica A* 344, pages 267–271, 2004.
- [Dah89] R. DAHLHAUS. Efficient parameter estimation for self-similar processes. *The Annals of Statistics*, 17(4) : 1749–1766, 1989.
- [DM03] A. B. DIEKER and M. MANDJES. On spectral simulation of fractional brownian motion. *Probab. Eng. Inf. Sci.*, 17(3) : 417–434, 2003.
- [eGJ96] R. JENNANE, R. HARBA et G. JACQUET. Estimation de la qualité des méthodes de synthèse du mouvement brownien fractionnaire. *Revue Traitement du Signal*, 13(4) : 289–302, 1996.
- [eGJ01] R. JENNANE, R. HARBA et G. JACQUET. Méthodes d'analyse du mouvement brownien fractionnaire : théorie et résultats comparatifs. *Revue Traitement du Signal*, 18(5-6) : 419–436, 2001.
- [eGL94] J. ISTAS et G. LANG. Variations quadratiques et estimation de l'exposant de hölder local d'un processus gaussien. *C.R. Acad. Sci. Paris*, 319 : 201–206, 1994.
- [FT86] R. FOX and M.S. TAQQU. Large sample properties of parameter estimates for strongly dependent stationary time series. *The Annals of Statistics*, 14(2) : 517–532, 1986.
- [GW03] N. SCAFETTA, L. GRIFFIN and B.J. WEST. Hölder exponent spectra for human gait. *Physica A* 328, pages 561–583, 2003.
- [Her04] E. HERBIN. *Processus (multi-)fractionnaires à paramètres multidimensionnels et régularité höldérienne*. PhD thesis, Paris XI Orsay, 2004.
- [KS86] T. LUNDAHL, W. OHLEY, S. KAY and R. SIFFERT. Fractional brownian motion : A maximum likelihood estimator and its application to image texture. *IEEE Trans. on Medical Imaging*, 5 : 152–161, 1986.
- [Lai04] D. LAI. Estimating the hurst effect and its application in monitoring clinical trials. *Computational Statistics and Data Analysis*, 45 : 549–562, 2004.
- [Leh80] E.L. LEHMANN. Efficient likelihood estimators. *The American Statistician*, 34 : 233–235, 1980.
- [Man83] B.B. MANDELBROT. *The Fractal Geometry of Nature*. Freeman, New York, 1983.
- [MN68] B.B. MANDELBROT and J.W. VAN NESS. Fractional brownian motions, fractional noises and applications. *SIAM Review*, 10(4) : 422–437, 1968.
- [Nor94] I. NORROS. A storage model with self-similar input. *Queueing Systems*, 16 : 387–396, 1994.
- [Pel98] R.F. PELTIER. *Processus stochastiques fractals avec applications en finance*. PhD thesis, Univ. Paris VI, 1998.
- [PLV94] R.F. PELTIER and J. LEVY-VEHEL. A new method for estimating the parameter of fractional brownian motion. Research report 2396, INRIA Rocquencourt, 1994.
- [R D05] R Development Core Team. *R : A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienne, Autriche, 2005. ISBN 3-900051-07-0.
- [Sap90] G. SAPORTA. *Probabilités, analyse des données et statistiques*. Editions Technip, Paris, 1990.

ÉTUDE EXPÉRIMENTALE DE L'INFLUENCE D'UN ÉCHANTILLONNAGE

- [Sca82] J.D. SCARGLE. Studies in astronomical time series. II. Statistical aspects of spectral analysis of unevenly spaced data. *Astrophysical Journal*, 263 : 835–853, 1982.
- [Sel95] F. SELLAN. Synthèse de mouvements brownien fractionnaires à l'aide de la transformation par ondelettes. *C.R. Acad. Sci. Paris*, 321 : 351–358, 1995.
- [ST94] G. SAMORODNITSKY and M.S. TAQQU. *Stable non-Gaussian random processes : Stochastic models with infinite variance*. Chapman and Hall, New York, 1994.
- [ST00] D. POTTS, G. STEIDL and M. TASCHE. *Fast Fourier transforms for nonequispaced data : A tutorial*, chapter 12, pages 253–274. Birkhauser, 2000.
- [VV00] A. VIDACS and J.T. VIRTAMO. Parameter estimation of geometrically sampled fractional brownian traffic. *INFOCOM*, 3, 2000.
- [WC94] A. WOOD and G. CHAN. Simulation of stationary Gaussian processes in $[0, 1]^d$. *Journal of computational and graphical statistics*, 3(4) : 409–432, 1994.
- [Yue78] C.K. YUEN. Quadratic Windowing in the Segment Averaging Method for Power Spectrum Computation. *Technometrics*, 20(2) : 195–200, 1978.

