

Modélisation et détection de ruptures des signaux physiologiques issus de compétitions d'endurance

Imen Kammoun

▶ **To cite this version:**

Imen Kammoun. Modélisation et détection de ruptures des signaux physiologiques issus de compétitions d'endurance. Mathématiques [math]. Université Panthéon-Sorbonne - Paris I, 2007. Français. NNT : . tel-00200441

HAL Id: tel-00200441

https://theses.hal.science/tel-00200441

Submitted on 20 Dec 2007

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.

UNIVERSITÉ PARIS I - PANTHÉON SORBONNE

U.F.R. DE MATHEMATIQUES ET INFORMATIQUE

THÈSE DE DOCTORAT

Présentée par
Imen KAMMOUN

Pour obtenir le grade de
Docteur en sciences

Spécialité:
Mathématiques appliquées

MODÉLISATION ET DÉTECTION DE RUPTURES DES SIGNAUX PHYSIOLOGIQUES ISSUS DE COMPÉTITIONS D'ENDURANCE

Sous la direction de
Jean Marc BARDET

Composition du Jury

Jean Marc BARDET	Professeur à l'Université Panthéon-Sorbonne	**Directeur**
Pierre BERTRAND	Professeur à l'Université Clermont-Ferrand 2	**Examinateur**
Véronique BILLAT	Professeur à l'Université Evry Val d'Essonne	**Examinateur**
Paul DOUKHAN	Professeur à l'ENSAE	**Examinateur**
Remigijus LEIPUS	Professeur à l'Université de Vilnius	**Rapporteur**
Jean Michel POGGI	Professeur à l'Université Orsay	**Rapporteur**
Philippe SOULIER	Professeur à l'Université Nanterre	**Examinateur**
Gilles TEYSSIÈRE	Professeur à l'ENSAI	**Examinateur**

Décembre 2007

Avant-propos

Mes premiers remerciements vont à mon directeur de thèse, Jean Marc Bardet. Je lui suis extrêmement reconnaissante d'avoir été présent jusqu'au bout et de m'avoir permis de terminer et concrétiser ces années de thèse. Je remercie Véronique Billat pour m'avoir permis de travailler sur les données physiologiques et pour la confiance qu'elle a montrée envers mon travail. Je remercie également Jean Pierre Koralsztein pour ses conseils.
Je remercie Marie Cottrell pour m'avoir permis d'effectuer ma thèse au sein du laboratoire SAMOS de Paris I et pour son éternel soutien à tous les thésards.

Je souhaite exprimer toute ma gratitude au rapporteurs de ma thèse Remigijus Leipus et Jean Michel Poggi qui ont bien voulu consacrer à ma thèse une partie de leur temps. Je remercie chaleureusement Pierre Bertrand, Paul Doukhan, Marc Lavielle, Philippe Soulier et Gilles Teyssière pour l'honneur qu'ils m'ont fait en acceptant de participer à mon jury de thèse.

Je tiens à remercier particulièrement Madalina pour son soutien et pour toutes ces discussions qui remontent le moral. Un grand merci également aux membres du laboratoire SAMOS pour cette bonne humeur qui a régné et particulièrement Sandie, Olivier, Hatem, Cécile, Vincent, Omar, Patrice, Charles, Ciprian, Patrick, Joseph, Corinne, Xavier, Annie, Marie K., ... et les ex-membres Riadh et Catherine. Je remercie toutes les personnes rencontrées au laboratoire tout au long de ma thèse.

A ma famille et ma belle famille, je leur écris un énorme merci en particulier mon père, ma mère, mon frère et ma soeur. Merci à mes amis de toujours : Nadra et Slim.
Il m'est bien évidemment impossible de ne pas citer la personne qui m'a soutenue et m'a encouragée de continuer mon cursus universitaire en France : Habib Bouchriha.

Avec beaucoup de tendresse, merci Skander pour ton soutien et ta patience.

Enfin, je voudrais dédier cette thèse à la mémoire de ma grand-mère.

Résumé

Ce travail porte sur la modélisation et l'estimation de paramètres pertinents pour les signaux de fréquences cardiaques instantanées recueillies par le LEPHE (Laboratoire d'Etude de la PHysiologie de l'Exercice) auprès d'athlètes courant le marathon de Paris 2004. Nous avons choisi de nous intéresser plus particulièrement à un paramètre que l'on pourrait appeler grossièrement "fractal", qui témoigne d'une certaine manière de la régularité locale de la trajectoire et de la dépendance entre les données. L'évolution d'un tel paramètre au fil de la course pourrait permettre de nouvelles interprétations et explications des comportements physiologiques pendant l'effort physique.

Dans un cadre de processus stationnaires (ce que l'on a pu à peu près obtenir après "nettoyage" des données de fréquences cardiaques et découpage des séries en plusieurs phases, début, milieu et fin de course, par détection automatique des ruptures, voir le Chapitre 2), ce paramètre se rapproche du paramètre de Hurst, utilisé pour les processus à longue mémoire. La plupart des estimateurs du paramètre de Hurst, appelé parfois coefficient de la loi de puissance, consiste à réaliser un ajustement par régression linéaire d'une quantité en fonction d'échelles dans une représentation logarithmique. Ceci inclut la méthode d'analyse des fluctuations redressées (Detrended Fluctuation Analysis, DFA) et la méthode d'analyse par ondelettes. La méthode DFA, qui consiste en une version de la méthode des variances agrégées pour les séries temporelles avec tendance, est étudiée en détail dans le Chapitre 3. En particulier, les propriétés asymptotiques de la fonction DFA et de l'estimateur de H, qui s'en déduit, sont étudiées pour le bruit gaussien fractionnaire et plus généralement pour une classe semi-paramétrique de processus stationnaires à longue mémoire avec ou sans tendance. Finalement, le chapitre 3 montre l'absence d'intérêt à utiliser la méthode DFA, que ce soit dans le cas stationnaire longue mémoire, où des méthodes comme celles du log-périodogramme ou d'analyse par ondelettes, donnent de bien meilleurs résultats, ou même dans le cas de processus composés d'un bruit longue mémoire et d'une tendance.

Si la méthode DFA n'est pas du tout robuste dans le cas de tendances polynomiales, ce n'est pas le cas de la méthode d'analyse par ondelettes, qui est également utilisable pour

des processus plus généraux. Dans le Chapitre 4, on propose la modélisation des séries de fréquences cardiaques par une généralisation du bruit gaussien fractionnaire, appelée bruit gaussien localement fractionnaire. Un tel processus stationnaire est construit à partir du paramètre dit de fractalité locale, qui est une sorte de paramètre de Hurst (pouvant prendre cependant des valeurs dans $\mathbb{R}$ et non seulement dans $(0,1)$) mais pour une bande de fréquence restreinte. L'estimation de ce paramètre de fractalité locale peut également se faire à partir d'une analyse par ondelettes, tout comme la construction d'un test d'adéquation. On montre ainsi la pertinence du modèle et une évolution du paramètre pendant la course, confirmant des résultats obtenus par d'autres auteurs dans leur étude permettant de distinguer les sujets sains de ceux ayant des dysfonctionnements cardiaques (voir Peng *et al.*).

Ces conclusions nous amènent à déduire qu'une éventuelle détection des changements de ce paramètre pourrait être extrêmement plus appropriée pour expliquer les éventuels changements physiologiques dans les fréquences cardiaques de l'athlète. Le Chapitre 5 propose une méthode de détection de multiples ruptures des paramètres de longue mémoire (respectivement d'autosimilarité, de fractalité locale) à partir d'un échantillon d'une série gaussienne stationnaire (respectivement séries chronologiques, processus à temps continu à accroissements stationnaires). A partir de méthodes d'analyse par ondelettes, un estimateur des m points de changement (m est supposé connu) est construit et on montre qu'il vérifie un théorème limite. Un théorème de la limite centrale est également établi pour l'estimateur de chaque paramètre et un test d'ajustement est mis en place dans chaque zone où le paramètre est inchangé. Enfin, on montre la même évolution du paramètre de fractalité locale sur les séries de fréquences cardiaques.

Abstract

This work focuses on the modeling and the estimation of relevant parameters characterizing instantaneous heart rate signals recorded in LEPHE (Laboratoire d'Etude de la PHysiologie de l'Exercise) laboratory from athletes during the marathon of Paris 2004. We choose to focus especially in an exponent that can be called "Fractal", which indicates the local regularity of the path and the dependency between data. The evolution of such parameter throughout the race could allow new interpretations and explanations of physiological behaviour during physical exercise.

In the case of stationary processes (what we have got after "cleaning" HR data and cutting series in different race phases - beginning, Middle and end - by an automatic detection of changes, see Chapter 2), this parameter is close to the Hurst parameter, defined for long range dependent processes. The most common estimators of the Hurst parameter, sometimes called scaling behaviour exponents, consist in performing a linear regression fit of a scale dependent quantity versus the scale in a logarithmic representation. This includes the Detrended Fluctuation Analysis (DFA) method and wavelet analysis method. The DFA which is a version, for time series with trend, of the aggregated variance method is studied in details in Chapter 3. In particular, the asymptotic properties of the DFA function and the deduced estimator of H are studied in the case of fractional Gaussian noise and extended to a general class of stationary semi-parametric long-range dependent processes with or without trend. Finally, Chapter 3 shows that DFA method is inappropriate for stationary long memory processes, where methods such as Log-periodogram or wavelet analysis provide best results, and even for processes composed by trended long memory noise.

If the DFA method is not at all robust in the case of polynomial trends, this is not the case of the wavelet analysis method used for more general models. In Chapter 4, we propose the modelling of heart rate data with a generalization of fractional Gaussian noise, called locally fractional Gaussian noise. Such stationary process is built from a parameter called of local fractality which is a kind of Hurst parameter (that may take values in $\mathbb{R}$ and not only in $(0, 1)$) in restricted band frequency. The estimation of local

fractality parameter and also the construction of goodness-of-fit test can be made with wavelet analysis. We also show the relevance of model and an evolution of the parameter during the race, which confirms results obtained by other authors in their study concerning the distinguish of healthy from pathologic data(see Peng *et al.*).

These last conclusions lead us to deduce that potential detection of changes in this parameter can be extremely meaningful for explaining the probable physiological changes of athlete's HR. In Chapter 5, a method detecting multiple abrupt changes of long memory parameter (respectively self-similarity, local fractality) from a sample of a Gaussian stationary times series (respectively time series, continuous-time process having stationary increments) is presented. From a wavelet analysis, an estimator of the m change instants (the number m is supposed to be known) is proved to satisfy a limit theorem. A central limit theorem is established for the estimator of each parameter and a goodness-of-fit test is also built in each zona where the parameter does not change. Finally, we show the same evolution of local fractality parameter relating to HR time series.

Si tu veux courir, cours un kilomètre, si tu veux changer ta vie, cours un marathon
Emil Zatopek[a]

[a] Célèbre coureur de fond du vingtième siècle

Chapitre 1

Introduction

1.1 Motivations

Ces dernières décennies, le dopage s'est professionnalisé et généralisé dans plusieurs sports. En athlétisme, Ben Johnson, le sprinter canadien recordman du monde de 100 mètres, a été contrôlé positif aux anabolisants en 1988. Carl Lewis, le plus titré des athlètes américains, avait été testé positif à trois reprises. L'américaine Kelly White, championne du monde du 100 et du 200 mètres est reconnue coupable de dopage en 2003 ainsi que la triple championne olympique Marion Jones en 2004 et après aveu en 2007. En cyclisme, Lance Armstrong est contrôlé positif lors de sa première victoire dans le Tour de France 1999 ou encore Tylor Hamilton en 2004 et Floyd Landis en 2006. En football, Diego Maradona contrôlé positif à la cocaïne en plusieurs reprises. Zineddine Zidane et Didier Deshamps qui reconnaissent avoir pris de la créatine en 2002. Plus proche du sujet traité ici, le Français Benoît Zwierzchiewski, co-détenteur du record d'Europe du marathon en 2h06mn35s a été aussi soupçonné de dopage.
Mais peut-on réaliser des performances et battre des records sans se doper ?

Autre question : le sport, souvent associé à l'acquisition d'une bonne santé, se trouve souvent responsable d'accidents ou même de décès. En particulier, les marathons, qui nécessitent d'énormes efforts physiques, sont malheureusement le lieu récurrent d'accidents cardio-vasculaires (le dernier Marathon de Chicago 2007 couru sous une forte chaleur a dû être arrêté au bout de 4h après avoir enregistré un très grand nombre de malaises et un décès...). De nombreuses raisons, en mettant de côté celles liées au dopage, peuvent expliquer ce paradoxe d'un sport synonyme de danger pour la santé : sur-entraînements, défis démesurés, manque de précautions médicales, manque de témoins permettant de dire "stop" au cours de l'effort...

Là encore, on peut se demander s'il est possible de pratiquer un sport demandant un gros investissement physique et physiologique sans se mettre en danger ?

La réponse à ces deux questions est oui. D'ailleurs ce qui motive le travail de l'équipe du Laboratoire d'Etude de la PHysiologie de l'Exercice (LEPHE, rattaché récemment à l'INSERM et dirigé par le Professeur Véronique Billat), est le développement de méthodes scientifiques d'entraînement permettant des alternatives efficaces au dopage et la prévention des accidents physiologiques. Ce travail de thèse, qui s'appuie essentiellement sur des données d'inter-durées entre battements de coeur recueillies auprès de 50 athlètes courant un marathon (Marathon de Paris 2004) par le LEPHE, a aussi pour ambition de développer des méthodes d'étude de signaux physiologiques lors d'un exercice d'endurance. Ceci devrait contribuer à l'amélioration d'outil de suivi médical ou d'aide au diagnostic de dysfonctionnement cardiaque pour les athlètes.

Les techniques employées par le LEPHE (mais également par d'autres laboratoires) pour l'amélioration de l'entraînement des athlètes, commencent par un test d'effort qui consiste à enregistrer certains témoins du fonctionnement de l'organisme soumis à une activité physique. Outre les indicateurs de consommation d'oxygène, d'émission de gaz carbonique, etc..., prélevés lors de tests destinés plutôt à des athlètes de haut niveau, l'étude des enregistrements de fréquences cardiaques peut être pratiquée sur un grand public surtout avec le développement des appareils de mesure portatifs. De plus, et contrairement à d'autres paramètres physiologiques, ce paramètre biologique peut être quotidiennement utilisé comme contrôle à l'entraînement. L'athlète dispose alors d'indications sur ses capacités physiques et sur les fréquences cardiaques qu'il peut soutenir sans danger.

Dans notre cas d'étude, on s'intéresse donc aux inter-durées entre battements de coeur (ce qui peut également se traduire en terme de fréquences cardiaques instantanées) recueillies auprès d'athlètes courant le Marathon de Paris 2004[1]. Les 50 athlètes suivis pratiquent régulièrement la course à pied et participent régulièrement à ce genre de compétitions. La durée entre deux battements de coeur successifs est relevée instantanément sur des cardiofréquencemétres. La fiabilité des cardiofréquencemétres semble bonne à partir du moment où ils sont posés correctement, mais le marathon étant une épreuve particulièrement longue (42,195km), de nombreux enregistrements de coureurs n'ont pas pu être pris en compte. Finalement, seuls neuf athlètes, pour lesquels les enregistrements permettent de disposer d'une base de données de bonne qualité, ont été retenus. Pour ces athlètes, le parcours a été réalisé en moyenne en 3h14mn. Si les données relevées sont les durées entre les pulsations successives (en ms), on travaille

[1] Cette expérimentation a été conduite par l'équipe de Pr. Véronique Billat du Laboratoire d'Etude de la Physiologie de l'Exercice (LEPHE).

plutôt à partir des fréquences cardiaques instantanées (FC) mesurées en nombre de battements par minute (BPM). La moyenne de la FC pour l'ensemble de l'échantillon est de 162 BPM.

En plus des FC, d'autres variables physiologiques sont mesurées et étudiées (le lecteur est renvoyé aux articles de recherches et de presse de V. Billat[2]) tel que le débit d'oxygène consommé lors de l'effort. La variable d'intérêt étant le VO2max c'est-à-dire le volume maximal d'oxygène prélevé au niveau des poumons et utilisé par les muscles par unité de temps. Cette variable est exprimée en litres par minutes ($l.mn^{-1}$) ou en $l.mn^{-1}.kg^{-1}$ (c'est-à-dire ramenée au poids du sujet). La variable VCO2 représentant le volume de gaz carbonique éjecté peut aussi être mesurée. En effet, des techniques de récupération des gaz permettent de mesurer les échanges au cours de l'exercice.

Des appareils de mesure de la vitesse de la course sont aussi employés tels que les GPS ou les accéléromètres. Des variables sont aussi déduites les unes des autres telle la VMA (Vitesse Maximale Aérobie), qui représente la vitesse de course à laquelle le coureur atteint sa consommation maximale d'oxygène, ou plus récemment, le débit sanguin, qui est le produit de la FC et du VES (volume d'éjection systolique) qui représente la quantité de sang éjectée à chaque battement[3]...

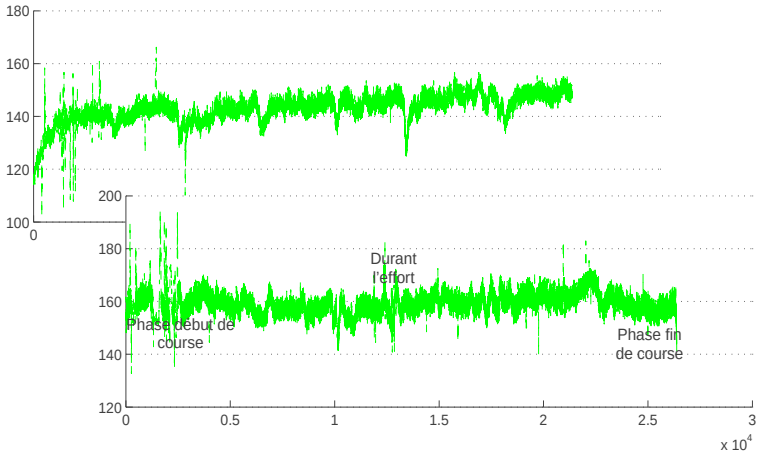


FIG. 1.1 – Evolution de la FC chez deux athlètes et les phases qui peuvent êtres observées durant l'effort

Si l'on se restreint aux données de battement cardiaque durant un marathon, on remarque que la fréquence cardiaque en moyenne n'augmente pas significativement et que le coureur se base sur environ 85% de sa fréquence cardiaque maximale (FCM)[4].

[2] http ://www.billat.net/

[3] Une première réalisée durant le marathon de Paris 2007 par V. Billat pour essayer de comprendre comment le marathon peut entraîner des troubles de fonctionnement du coeur.

[4] La FCM théorique étant souvent calculée par la formule "grossière" : $FCM = 220 - \text{âge du sujet}$.

Ces observations sont confirmées par d'autres données enregistrées sur des coureurs (voir Billat, 2006), qui montrent en revanche que c'est le rapport entre la fréquence cardiaque et la vitesse de la course qui augmente[5] (ce qui peut être traduit par une chute de vitesse).

1.2 Modélisation des signaux physiologiques

Une caractéristique des systèmes physiologiques est leur grande complexité. L'approche traditionnelle qui prétend qu'un système biologique est autorégulé et s'équilibre après une quelconque perturbation est discutée par des études théoriques qui indiquent que les signaux générés par les organismes biologiques tendent à montrer une certaine tendance ou non-stationnarité et des fluctuations complexes même à l'état de repos. Des méthodes mathématiques et de traitement du signal ont relevé aussi la présence de corrélation à long terme dans les séries temporelles qui décrivent les fluctuations au cours du temps des systèmes physiologiques (Golberger, 2001), (Goldberger *et al.*, 2002).

Une fois la base de données, contenant les relevés de FC pour 9 athlètes, "nettoyée" (ce que l'on pourra retrouver dans le chapitre qui suit), le premier but que nous nous sommes fixé a été de tenter de modéliser ces séries temporelles.

En premier lieu, on peut chercher à déterminer s'il existe une tendance, ou tout au moins des variations en moyenne, en variance ou en saisonnalité, pour chacune des séries chronologiques. Si de tels motifs sont présents alors ce sont des signes de non-stationnarité. En fait, les séries étudiées présentent des ruptures (ou des changements) en moyenne et des ruptures en variances qui se traduisent par des fluctuations verticales pour la série. Mais ces deux phénomènes distinguent une phase de début de course, une phase de milieu de course et une phase de fin de course (voir Figure 1.1 et Chapitre 2 pour plus de détails). Donc, il peut être possible d'envisager une stationnarité "par morceaux", c'est-à-dire que le processus serait à peu près stationnaire sur des intervalles de temps fixés. Il ne reste plus alors qu'à modéliser cette composante stationnaire.

Une autre piste possible consisterait à travailler non pas directement sur les séries de FC, mais sur la série constituée par le cumul de ces données (une fois celles-ci recentrées). Si on perd clairement la stationnarité par un tel procédé, des études précédentes (Bassingthwaighte *et al.*, 1994), (Goldberger, 1996) ont montré que la série ainsi obtenue forme un graphe semblant présenter le même type de distribution quelque soit le

[5] ce qui se traduit par le coût cardiaque défini comme le nombre de battements nécessaires pour parcourir chaque mètre

"zoom" que l'on peut faire sur ce graphe ; cette invariance par changement d'échelle est également appelée autosimilarité. Ceci ouvre donc d'autres voies de modélisation pour les données de FC.

Avant de dire plus, il nous semble nécessaire maintenant de donner quelques éléments sur les objets et propriétés mathématiques que nous venons d'évoquer et que nous verrons souvent revenir dans l'ensemble de ce travail.

1.3 Préambule mathématique

Voyons donc maintenant les bases fondamentales utilisées par la suite lors de notre travail de modélisation.

1.3.1 La notion de stationnarité

Une propriété qui constitue une hypothèse importante à prendre en compte, que ce soit pour la synthèse ou l'analyse de données des séries temporelles, est la stationnarité du processus. Son importance réside dans le fait que les conditions de moyenne constante, variance constante et de covariance ne dépendant que du "lag", sont essentielles pour estimer les paramètres et identifier les modèles que décrivent les données.

Un processus aléatoire est stationnaire si la loi de probabilité des variables aléatoires le constituant n'évolue pas au cours du temps. D'où cette définition de la stricte stationnarité (ou stationnarité au sens fort) :

Définition 1.3.1. *Un processus $X = \{X_t : t \in T\}$ est dit stationnaire au sens fort si pour tout $n \in \mathbb{N}^*$, tout $(t_1, \ldots, t_n) \in T^n$, et tout $t \in T$, $\left(X_{t_1}, \ldots, X_{t_n}\right)$ a la même loi que $\left(X_{t_1+t}, \ldots, X_{t_n+t}\right)$.*

Une définition moins stricte de la stationnarité (cependant équivalente à la stationnarité stricte pour les processus gaussiens) est plus souvent utilisée pour les processus possédant des moments d'ordre 2 :

Définition 1.3.2. *Un processus $X = \{X_t : t \in T\}$ tel que $\mathbb{E}X_t^2 < \infty$ pour tout $t \in T$, est dit stationnaire au sens faible si pour tout $(t, t') \in T^2$, $\mathbb{E}(X_t) = \mathbb{E}(X_{t'})$ et $Cov(X_t, X_{t'}) = r(|t' - t|)$ avec r une fonction appelée autocovariance de X.*

Parmi les processus stationnaires que nous envisagerons en vue de modéliser les séries temporelles de FC, une attention toute particulière sera portée à celles dites de longue mémoire.

1.3.2 Processus à longue mémoire

Le comportement de mémoire longue (appelée encore longue dépendance) est un phénomène souvent observé sur des données diverses. Ce phénomène a été développé dans divers travaux à commencer par les données hydrologiques (Hurst, 1951), climatologiques, les trafics informatiques et récemment en économie et finance. Plusieurs définitions de la longue mémoire existent qui ne sont pas toujours équivalentes. Nous allons nous limiter aux notions de longue mémoire les plus courantes.

Le comportement de longue mémoire peut se traduire par des comportements de persistance au niveau des données, par la décroissance lente de la fonction d'autocorrélation (dans un domaine temporel) ou dans un domaine fréquentiel par la divergence de la densité spectrale à l'origine.

Définition 1.3.3. *Soit un processus stationnaire $X = \{X_t : t \in \mathbb{N}\}$ tel que $\mathbb{E}X_t^2 < \infty$ pour tout $t \in \mathbb{N}$ et soit $r(\cdot)$ la fonction d'autocovariance de X_t. Le processus X_t est dit à longue mémoire si $\sum_{-\infty}^{+\infty} \mid r(k) \mid = \infty$.*

A la différence de ce qui se passe en courte mémoire (que l'on pourrait définir par $\sum_{-\infty}^{+\infty} \mid r(k) \mid < \infty$) en présence de longue mémoire, les données restent corrélées longtemps. La longue mémoire est souvent caractérisée par le paramètre H appelé paramètre de Hurst tel que :

Définition 1.3.4. *Soit un processus stationnaire $X = \{X_t : t \in \mathbb{N}\}$ tel que $\mathbb{E}X_t^2 < \infty$ pour tout $t \in \mathbb{N}$. X est un processus à mémoire longue s'il existe un réel $H \in (1/2, 1)$ et une fonction $L(k)$ à variations lentes en ∞ (c'est-à-dire telle que $\forall t > 0$, $L(xt)/L(x) \to 1$ quand $x \to \infty$) vérifiant*

$$r(k) = L(k)k^{-(2-2H)}.$$

Dans le domaine spectral, si on peut associer à un processus stationnaire X_t ayant des moments d'ordre 2 une fonction d'autocovariance $r(\cdot)$, alors d'après le théorème

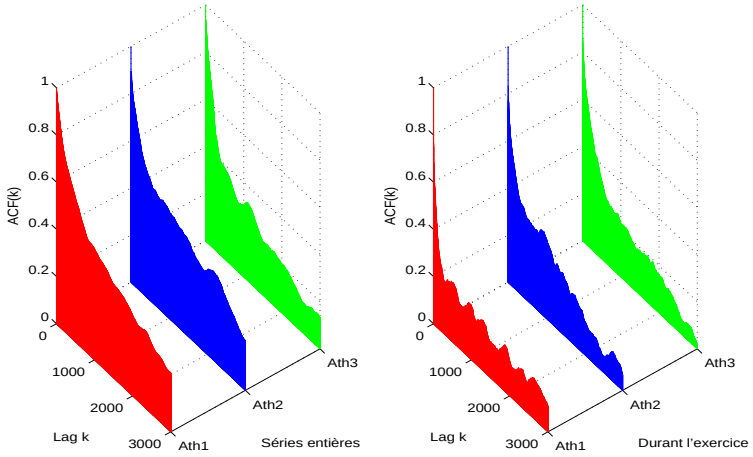


FIG. 1.2 – Représentation de la fonction d'autocorrélation empirique (ACF) pour la série des FC d'un athlète durant tout un marathon, et pour les données enregistrées seulement "en milieu" d'exercice

Wiener-Khinchin, la densité spectrale f d'un processus stationnaire, si elle existe, est définie comme étant la transformée de Fourier de la fonction d'autocovariance, soit

Définition 1.3.5. *Soit un processus stationnaire* $X = \{X_t : t \in \mathbb{N}\}$ *tel que* $\mathbb{E}X_t^2 < \infty$ *pour tout* $t \in \mathbb{N}$. *Pour* $\lambda \in [-\pi, \pi]$, *si elle existe, on appelle densité spectrale* f *en la fréquence* λ, $f(\lambda) = \dfrac{1}{2\pi} \sum_{-\infty}^{\infty} r(k) e^{-i\lambda k}$.

On peut aussi définir la longue mémoire à partir de la densité spectrale :

Définition 1.3.6. *Soit un processus stationnaire* $X = \{X_t : t \in \mathbb{N}\}$ *tel que* $\mathbb{E}X_t^2 < \infty$ *pour tout* $t \in \mathbb{N}$. X_t *est un processus à mémoire longue s'il existe un réel* $D \in (0, 1)$ *et une fonction* $L(\lambda)$ *à variations lentes en* 0 *si pour tout* $\lambda \in [-\pi, 0) \cup (0, \pi]$

$$f(\lambda) = L(\lambda) \mid \lambda \mid^{-D} .$$

Les bruits gaussiens fractionnaires (FGN) et les processus FARIMA (fractionally autoregressive integrated moving average) sont les exemples les plus souvent évoqués de processus longue mémoire :

Exemple de processus à longue mémoire : FARIMA(p, d, q)

Les processus FARIMA (Granger et Joyeux, 1980), (Hosking, 1981) représentent une généralisation des processus ARIMA(p, d, q) (Box et Jenkins, 1970) en introduisant un degré de différenciation d non entier.

Définition 1.3.7. *Soit $(\varepsilon_t)_{t\in\mathbb{N}}$ une suite de variables aléatoires indépendantes identiquement distribuées centrées et de variance σ_ε^2. Un processus $X = \{X_t, t \in \mathbb{Z}\}$ est un processus FARIMA(p, d, q) stationnaire et inversible s'il vérifie l'équation suivante :*

$$\phi(B)(1-B)^d(X_t - \mu) = \theta(B)\varepsilon_t$$

avec $d \in (-1/2, 1/2)$, μ la moyenne du processus, $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$ et $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q$ qui sont deux polynômes à coefficients réels n'ayant pas de racines communes et ayant leurs racines en dehors du cercle unité, B étant l'opérateur de retard tel que $BX_t = X_{t-1}$ pour $t \in \mathbb{N}$.

Pour $0 < d < 1/2$, un processus FARIMA(p, d, q) présente une longue mémoire. En effet, à partir de l'expression de la densité spectrale f d'un processus FARIMA(p, d, q), on montre qu'avec $D = 2d$, $f(\lambda) \sim C \mid \lambda \mid^{-D}$ pour $\lambda \to 0$. D'une certaine manière, plus d est proche de $1/2$ plus le processus a une mémoire qui croît, la valeur $1/2$ indiquant un passage de la stationnarité à la non-stationnarité.

Cas particulier : le processus FARIMA$(0, d, 0)$. Ce processus stationnaire possède une représentation moyenne mobile infinie :

$$X_t = \sum_{k=0}^{\infty} c_k \varepsilon_{t-k}$$

avec $c(k) = \frac{\Gamma(k+d)}{\Gamma(d)\Gamma(k+1)}$. Avec $H = d + 1/2$, comme dans le cas du FGN, la fonction d'autocovariance se comporte asymptotiquement comme une loi de puissance de $2d-1$:

$$\mathrm{Cov}(X_t, X_{t+h}) \sim C_d h^{2d-1} \quad \text{quand} \quad h \to \infty$$

avec $C_d = \frac{\Gamma(1-2d)}{\pi} \sin$. La densité spectrale de ce processus est relativement simple :

$$f(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi}(2\sin(\lambda/2))^{-2d} \sim \frac{\sigma_\varepsilon^2}{2\pi} \mid \lambda \mid^{-2d} \quad \text{quand} \quad \lambda \to 0.$$

Exemple de processus à longue mémoire : FGN

Le bruit gaussien fractionnaire (Mandelbrot et Van Ness, 1968) est un exemple de processus gaussien stationnaire à longue mémoire. Plus précisément, $\{X_t^H, t \in \mathbb{N}\}$ est un FGN avec une fonction d'autocovariance qui s'écrit

$$r_{X^H}(k) = \frac{\sigma^2}{2}(|k+1|^{2H} + |k-1|^{2H} - 2|k|^{2H}) \qquad \forall k \in \mathbb{N}$$

avec $H \in (0,1)$ et $\sigma^2 > 0$. (voir Samorodnitsky et Taqqu, 1994, pour plus de détails). On a aussi pour $1/2 < H < 1$,

$$r_{X^H}(k) \sim H(2H-1)|k|^{2H-2} \quad \text{quand } k \to \infty;$$

dans ce cas, X_t^H est bien un processus à longue mémoire. Sa densité spectrale s'écrit (voir Sinai, 1976) :

$$f(\lambda) = C_H(2\sin(\lambda/2))^2 \sum_{k=-\infty}^{\infty} \frac{1}{|\lambda + 2\pi k|^{2H+1}} \sim C_H \mid \lambda \mid^{1-2H} \quad \text{quand} \quad \lambda \to 0.$$

Notons que le processus agrégé déduit à partir d'un bruit gaussien fractionnaire centré correspond au mouvement brownien fractionnaire (voir ci-dessous).

1.3.3 Propriété d'autosimilarité

Le phénomène de longue mémoire est étroitement lié à celui de l'autosimilarité (Mandelbrot et Van Ness, 1968), (Mandelbrot et Wallis, 1969). L'autosimilarité se définit par le fait qu'un objet peut être décomposé en sous-unités, puis en sous-sous-unités, etc..., qui toutes, ressemblent à la structure de l'objet global et possèdent les mêmes propriétés statistiques que celui-ci ; on pourrait donc dire que la distribution d'un processus autosimilaire est invariante par changement d'échelles.

Définition 1.3.8. *Un processus $\{Y_t, t \geq 0\}$ est autosimilaire de paramètre d'autosimilarité $H > 0$ si pour tout $c > 0$ le processus $\{Y_{ct}, t \geq 0\}$ et $\{c^H Y_t, t \geq 0\}$ ont la même distribution.*

Notons que dans certaines applications, il est nécessaire d'imposer des bornes supérieures et inférieures aux nombres de décompositions (en sous-unités) pour pouvoir appliquer l'autosimilarité.
Par ailleurs, si Y^H est un processus autosimilaire à accroissements stationnaires de paramètre H et tel que $\mathbb{E}(Y_t^H)^2 < \infty$ pour tout t, alors le processus des accroissements X^H défini par $X^H = \{Y_{t+1}^H - Y_t^H, t \geq 0\}$ est stationnaire, centré, et à longue mémoire dès que $H \in (1/2, 1)$, puisque

$$\begin{aligned} \mathrm{Cov}(X_k^H, X_0^H) &= \frac{\mathbb{E}(Y_1^H)^2}{2}\left(|k+1|^{2H} + |k-1|^{2H} - 2\,|k|^{2H}\right) \\ &\sim H(2H-1)\mathbb{E}(Y_1^H)^2\, k^{-(2-2H)} \quad \text{quand } k \to \infty. \end{aligned}$$

Exemple de processus autosimilaire gaussien : FBM

L'exemple classique de processus autosimilaire est le mouvement brownien fractionnaire (FBM) de paramètre d'autosimilarité H (Mandelbrot et Van Ness, 1968).

Définition 1.3.9. *Un FBM $Y^H = \{Y_t^H, t \geq 0\}$ est un processus gaussien centré à accroissements stationnaires tel que pour tout $(s,t) \in \mathbb{R}^2$: $\mathbb{E}|Y_s^H - Y_t^H| = \sigma^2|t-s|^{2H}$ et $Y_0^H = 0$. Il est indexé par un paramètre réel $0 < H < 1$, son exposant de Hurst, et $\sigma^2 > 0$.*

On peut facilement montrer que :

Propriété 1.3.10. *Y^H est un FBM $\Longleftrightarrow$ Y^H est un processus gaussien autosimilaire à accroissements stationnaires.*

Notons par exemple que le mouvement brownien bifractionnaire (voir Houdré et Villa, 2003) est un exemple de processus autosimilaire gaussien n'ayant pas des accroissements stationnaires.

Exemple de processus autosimilaire non-gaussien : processus de Rosenblatt

Les processus de Rosenblatt d'ordre $m \in \mathbb{N}^*$ (voir Taqqu, 1979), sont une généralisation du FBM, définis par :

$$Z_{m,H}(t) = \int_{\mathbb{R}^m} \frac{e^{it(u_1+\cdots+u_m)} - 1}{i(u_1+\cdots+u_m)} |u_1 u_2 \cdots u_m|^{(H-1)/2} \, \widehat{dB}(u_1) \cdots \widehat{dB}(u_m),$$

où dB est une mesure brownienne et $\widehat{dB}$ sa transformée de Fourier. Les processus de Rosenblatt sont des processus ayant un moment d'ordre 2, mH-autosimilaires avec $mH < 1$, à accroissements stationnaires et non gaussiens pour $m \geq 2$ (pour $m = 1$, on retrouve le mouvement brownien fractionnaire). On peut aussi citer certains processus α-stables autosimilaires, tels que les processus linéaires fractionnaires stables qui eux n'ont pas de moments d'ordre 2 (voir encore Samorodnitsky et Taqqu, 1994).

1.3.4 Généralisations gaussiennes du FBM

Le paramètre H caractérise l'invariance par changement d'échelle du mouvement brownien fractionnaire. Durant les années 70 et 80, l'étude du FBM a été développée

et appliquée dans différents domaines (voir par exemple chapitre 14 du livre édité par Samorodnitsky et Taqqu, 1994). Plusieurs auteurs ont étudié des modèles plus généraux, pour lesquels le paramètre H est remplacé par une fonction qui dépend de t. En effet, dans de nombreuses applications, il semble intéressant de modéliser les données avec cette généralisation du FBM (pensons par exemple à une modélisation des profils d'irrégularité de montagnes, pouvant changer de régularité, et donc de paramètre H, en fonction de l'âge des montagnes). Peltier et Lévy-Véhel (1996) (ainsi que Benassi *et al.* (1997) ont défini et étudié un nouveau processus : le mouvement brownien multifractionnaire (MBM) qui est un processus généralisant le FBM, le paramètre de Hurst H est remplacé par une fonction continu (et même hölderienne) qui dépend du temps (ce qui exclu toute forme de stationnarité pour un tel processus). Une représentation du MBM peut se déduire de la représentation harmonisable du MBF (Kolmogorov, 1940) définie comme intégrale stochastique à partir d'une mesure de Wiener :

$$Y^H(t) = \int_{\mathbb{R}} \frac{e^{it\xi} - 1}{\mid \xi \mid^{H+1/2}} \widehat{dB}(\xi).$$

Une représentation possible pour le MBM normalisée est alors la suivante :

$$X_t \;=\; g(H(t)) \int_{\mathbb{R}} \frac{e^{ßtx} - 1}{|x|^{H(t)+1/2}} \widehat{dB}(x),$$

$t \mapsto H(t)$ étant une fonction telle qu'il existe $C_H > 0$, $\beta \in (0,1]$, $H_{\sup}$ et $H_{\inf}$ vérifiant

$$\sup_{0\le t<s\le 1} |H(t)-H(s)|/|t-s|^{\beta} \le C_H \quad \text{et} \quad H_{\sup} := \sup_{t\in[0,1]} H(t) < 1, \quad H_{\inf} := \inf_{t\in[0,1]} H(t) > 0.$$

Enfin $g(H(t))$ est une fonction de normalisation telle que $\mathbb{E}X_t^2 = 1$, *i.e.* pour $H \in (0,1)$,

$$g(H) = \Big(\frac{\sin(\pi H)\Gamma(2H+1)}{2\pi}\Big)^{1/2}. \tag{1.1}$$

Ayache et Lévy-Véhel (2000) ont introduit alors un processus gaussien généralisant le MBM dont le paramètre de Hurst est remplacé par une fonction qui peut être très irrégulière.

Dans d'autres cas, le paramètre de Hurst semble être une fonction en escalier du temps. Benassi *et al.* (2000) ont proposé le SFBM (step fractional Brownian motion) où l'exposant de Hurst est remplacé par une fonction $t \mapsto H(t)$ dans la représentation en série d'ondelettes du MBF (Benassi, Jaffard et Roux, 1997) ; ce processus a aussi été étudié par Ayache *et al.* (2006) et peut se présenter sous la forme suivante :

$$B_H(t) = \sum_{j\in\mathbb{N},k\in\mathbb{Z}} \Big[\int_{\mathbb{R}} \frac{e^{it\xi} - 1}{\mid \xi \mid^{H+1/2}} \overline{\widehat{\psi}}_{j,k}(\xi)(d\xi) \Big] \zeta_{j,k}$$

avec ψ une ondelette, et $\zeta_{j,k}$ une famille i.i.d de v.a. gaussiennes standardes.

Autre généralisation possible, celle pour laquelle le paramètre H est une fonction de la fréquence ξ (dans la représentation harmonisable de FBM évoquée un peu plus haut). Un intérêt d'une telle généralisation est qu'elle permet, contrairement aux MBM, de définir un processus à accroissements stationnaires. Collins et De Lucas (1993), puis Benassi et Duguy (1999) ont proposé un modèle avec deux paramètres de Hurst pour les basses et les hautes fréquences. Dans Bardet et Bertrand (2007), on étudie un processus gaussien à accroissements stationnaires dont le paramètre H est une fonction par morceaux des fréquences $\xi \mapsto H(\xi)$ qui est le mouvement brownien multiéchelle.

$$Y(t) = 2\sum_{j=0}^{K} \int_{\omega_j}^{\omega_{j+1}} \sigma_j \frac{e^{it\xi}-1}{\mid \xi \mid^{H_j+1/2}} \widehat{dB}(\xi)$$

avec pour $i = 0, 1, \ldots, K$, $\sigma_i > 0$, $H_i \in \mathbb{R}$ (sauf $H_0 < 1$ et $H_K > 0$) et $\omega_0 = 0 < \omega_1 < \ldots < \omega_K < \omega_{K+1} = \infty$. Ce processus est aussi étudié dans le Chapitre 4 et plus généralement dans le Chapitre 5 où on définit le processus gaussien localement fractionnaire par morceaux :

$$Y(t) := \int_{\mathbb{R}} \frac{e^{it\xi}-1}{\rho_j(\xi)} \widehat{dB}(\xi) \quad \text{pour} \quad t \in [\tau_j\,,\,\tau_{j+1})$$

pour $0 = \tau_0 < \tau_1 < \ldots < \tau_m < \tau_{m+1}$ et pour $j = 0, \ldots, m$, $\rho_j : \mathbb{R} \to [0, \infty)$ et que

- $\rho_j(\xi) = \dfrac{1}{\sigma_j^*} \mid\xi\mid^{H_j+1/2}$ for $\mid\xi\mid \in [f_{min}\,,\,f_{max}]$ avec $H_j \in \mathbb{R}$, $\sigma_j > 0$;
- $\displaystyle\int_{\mathbb{R}} \left(1 \wedge |\xi|^2\right) \frac{1}{\rho_j^2(\xi)}\, d\xi < \infty.$

1.3.5 Techniques d'estimation du paramètre de Hurst

En général, l'existence de tendance, de périodicité ou d'autres sources de perturbations a un certain effet sur l'estimation du paramètre de longue mémoire ou du paramètre d'auto-similarité. Ici nous nous restreindrons au cadre de séries chronologiques stationnaires fortement dépendantes. Pour la suite on considère donc $X = \{X_t, t \in \mathbb{N}\}$ le processus étudié et $(X_1, \ldots, X_N)$ une trajectoire connue.

En premier lieu, citons les méthodes paramétriques : méthode du maximum de vraisemblance (que nous ne présenterons pas), et surtout, car plus générale et bien plus intéressante numériquement, la méthode du minimum de contraste de Whittle. Des méthodes d'estimation semi-paramétriques ont également été développées, telles que la méthode R/S (rescaled adjusted range), la méthode des variances agrégées, la ou plutôt les méthodes du log-périodogramme, la méthode d'analyse par ondelettes, etc... (on trouvera beaucoup de détails théoriques et appliquées sur tout ce qui suit dans le livre édité par Doukhan *et al.*, 2003).

Une méthode paramétrique : la méthode du maximum de vraisemblance approché de Whittle

La méthode proposée par Whittle (1951) est construite à partir du périodogramme et elle nécessite la connaissance de la fonction explicite de la densité spectrale sans connaître les valeurs exactes des paramètres, supposés en nombre fini, la composant. L'estimateur de Whittle (après renormalisation des paramètres) représente la valeur de η qui minimise :

$$Q(\eta) = \int_{-\pi}^{\pi} \frac{I(\lambda)}{f(\lambda,\eta)} d\lambda$$

où $I(\lambda)$ est le périodogramme qui est défini par :

$$I(\lambda) = \frac{1}{2\pi N} \Big| \sum_{t=1}^{N} X_t e^{ij\lambda} \Big|^2$$

pour $\lambda \in [-\pi, \pi]$. $I(\lambda)$ est un estimateur (non convergent ponctuellement) de la densité spectrale et $f(\lambda, \eta)$. Dans le cas, par exemple, du FGN ou du processus FARIMA$(0, d, 0)$, η représente le paramètre H ou d respectivement, ainsi que σ^2.
Pour le modèle FARIMA(p, d, q), le vecteur η comprend aussi les coefficients des parties MA et AR du modèle. Dans les applications, on remplace l'intégrale par la somme de Riemann correspondante pour des fréquences de la forme $\lambda_j = 2\pi j/N$ avec $j = 1, 2, \ldots, (N-1)/2$ et donc la fonction à minimiser est :

$$Q^*(\eta) = \sum_{j=1}^{(N-1)/2} \frac{I(\lambda_j)}{f^*(\lambda_j, \eta)}$$

$f^* = \beta \cdot f$ tel que $\int_{-\pi}^{\pi} f^*(\lambda, \eta) d\lambda = 0$.

Présentons maintenant les estimateurs semi-paramétriques les plus utilisés.

La méthode de Hurst : La statistique R/S

Historiquement, la première méthode mise en place pour estimer le paramètre H a été introduite par Hurst lui-même en 1951 lors de ses études hydrologiques. Cette méthode a ensuite été étudiée par Mandelbrot et Wallis (1969). On construit $Y(n) = \sum_{i=1}^{n} X_i$ et $S^2(n) = (1/n) \sum_{i=1}^{n} X_i^2 - (1/n)^2 Y^2(n)$ la variance empirique. La statistique R/S est donnée par :

$$\frac{R}{S}(n) = \frac{1}{S(n)} \Big[\max_{0 \le t \le n} \Big(Y(t) - \frac{t}{n} Y(n) \Big) - \min_{0 \le t \le n} \Big(Y(t) - \frac{t}{n} Y(n) \Big) \Big]$$

Pour estimer H, on divise la trajectoire de la série (de taille N) en K segments de taille N/K. Pour chaque fenêtre de taille n, on calcule $R/S(n)$ pour toutes les séries déduites qui commencent aux points $k_i = iN/K + 1$ pour $i = 0, 1, \ldots, N-1$ et tel que $k_i + n \leq N$. Dans ce cas un certain nombre estimant $R/S(n)$ est obtenu pour chaque pas n. En choisissant des valeurs de n espacées logarithmiquement, et en traçant le graphe $\log(R/S(n))$ en fonction de $\log(n)$, et la droite ajustant ce nuage de points par la droite des moindres carrés ordinaires, on peut déterminer une estimation du coefficient de Hurst.

Notons que pour un FGN ou un processus FARIMA, $\mathbb{E}(R/S(n)) \sim C_H n^H$ quand $n \to \infty$, $C_H > 0$ ne dépend que de H.

L'avantage de cette méthode est qu'elle permet d'obtenir un estimateur qui possède des bonnes propriétés de robustesse (voir Mandelbrot et Taqqu, 1979). Pour ce qui est des inconvénients, la distribution exacte de la statistique R/S est difficile à déterminer.

Lo (1991) a modifié la statistique R/S en utilisant une somme pondérée de la fonction d'autocovariance comme normalisation au lieu de la variance empirique. Cette méthode permet de tester l'hypothèse nulle d'absence de dépendance de long terme contre l'hypothèse alternative de dépendance de long terme.

La méthode de variance agrégée

Considérant la série agrégée obtenue en divisant la série $(X_1, \ldots, X_N)$ en m segments et en calculant la moyenne dans chacune de ces fenêtres :

$$X^{(m)}(k) = \frac{1}{m} \sum_{i=(k-1)m+1}^{km} X_i \quad , \qquad k = 1, 2, \ldots, [N/m].$$

La variance empirique des $X^{(m)}(k)$ est :

$$\widehat{\mathrm{Var} X^{(m)}} = \frac{1}{N/m} \sum_{k=1}^{N/m} (X^{(m)}(k) - \overline{X})^2$$

On répète cette procédure pour différentes valeurs de m et on trace le logarithme de ces variances empiriques par rapport à $\log m$. La pente de la droite des moindres carrés ordinaires ajustant ce nuage de points étant $2H - 2$ ce qui permet de déterminer un estimateur de H.

La méthode de Geweke et Porter-Hudak dite méthode du log-périodogramme

Le principe de cette méthode est le suivant : Pour une série présentant une dépendance longue mémoire, la densité spectrale est de la forme :

$$f(\lambda) = L(\lambda)|\lambda|^{-D},$$

au voisinage de 0. Sous certaines conditions, le périodogramme défini ci-dessus constitue un estimateur de la densité spectrale. Donc, une régression du logarithme du périodogramme sur $\log \lambda$ pour des fréquences très proches de l'origine ($\lambda_j = 2\pi j/N$), nous donne une estimation du coefficient D appelée estimateur GPH (méthode introduite par Geweke et Porter-Hudak en 1983).

Différentes versions et généralisations (locales et globales, adaptatives) de cet estimateur ont été développées (on trouvera un résumé assez exhaustif de ces différentes méthodes dans Moulines et Soulier, 2003).

La méthode des fluctuations redressées (DFA)

Dans une première étape, la série originale est "intégrée" en remplaçant chaque donnée par la somme cumulée des écarts à la moyenne :

$$\tilde{Y}(k) = \sum_{i=1}^{k} (X(i) - \overline{X}) \ , \quad \text{pour} \quad k \in \{1, \ldots, N\}$$

La série Y est ensuite divisée en fenêtres de même taille n. Dans chaque fenêtre la droite des moindres carrés est estimée pour être soustraite de $Y(k)$. La fonction DFA (Detrended Fluctuations Analysis) représente alors l'écart-type des résidus de cette regression pour toute la série :

$$\tilde{F}(n) = \sqrt{\frac{1}{n \cdot [N/n]} \sum_{k=1}^{n \cdot [N/n]} \left(\tilde{Y}(k) - \widehat{Y}_n(k) \right)^2}.$$

Ce calcul est répété pour différentes tailles de fenêtre n_i et on montre que

$$F(n) \simeq C_H \cdot n^H. \tag{1.2}$$

La représentation de $\log F(n_i)$ par rapport à $\log n_i$ nous donne un estimateur de H.

Estimateur basé sur les ondelettes

Par analogie avec le périodogramme, qui utilise le carré de la transformée discrète d'ondelette des données, l'analyse basée sur les ondelettes fait intervenir le carré de la transformée discrète d'ondelettes. Soit $\psi : \mathbb{R} \to \mathbb{R}$ une fonction appelée ondelette mère. Soit $(a, b) \in \mathbb{R}_+^* \times \mathbb{R}$ et $\lambda = (a, b)$. On calcule la transformée en ondelettes qui utilise des translations et des dilatations de la fonction fixe ψ

$$\psi_\lambda(t) = \frac{1}{\sqrt{a}} \psi \left(\frac{t}{a} - b \right)$$

a étant l'échelle et b le paramètre de translation. Notons $d_X(a, b)$ les coefficients d'ondelettes relatives au processus X,

$$d_X(a, b) = \frac{1}{\sqrt{a}} \int_{\mathbb{R}} \psi(\frac{t}{a} - b) X(t) dt.$$

Pour un échantillon de série temporelle, une somme de Riemann peut remplacer l'intégrale précédente pour définir des coefficients d'ondelettes discrétisés $e_X(a, b)$. En supposant que la fonction ψ vérifie $\int_{\mathbb{R}} t^m \psi(t) dt = 0$ pour tout $m \in \{0, 1, \ldots, M\}$, (Flandrin, 1992) et (Abry, Veich et Flandrin, 1998) ont montré que pour un processus X stationnaire et à longue mémoire, la variance de ces coefficients d'ondelettes s'écrivent :

$$\mathbb{E}(d_X^2(a, b)) = \mathrm{Var}(d_X(a, b)) \sim C(\psi, H) a^{2H-1} \quad \text{quand } a \to \infty$$

et pour un processus X auto-similaire à accroissements stationnaires,

$$\mathbb{E}(d_X^2(a, b)) = \mathrm{Var}(d_X(a, b)) \sim K(\psi, H) a^{2H+1} \quad \text{pour tout } a > 0.$$

$C(\psi, H)$ et $K(\psi, H)$ des constantes positives. Cette variance peut être approchée par la moyenne des carrés des coefficients d'ondelettes et une log-log régression de ces quantités par rapport à différentes échelles choisies permet d'estimer H.

1.3.6 Détection de ruptures

La détection de ruptures ou de changements dans les caractéristiques de signal est un thème largement étudié avec les approches différentes (voir par exemple Basseville, 1988 ou Basseville et Nikoforov, 1993). L'idée est de trouver les instants de rupture dans des séries chronologiques afin de développer des modèles pour les différents segments durant lesquelles les caractéristiques du signal restent inchangées. Différentes méthodes ont été appliquées aux données financières, économiques, hydrologiques ou encore physiologiques. Le principe d'un problème de localisation de ruptures réside dans le choix

d'un critère à optimiser, qui entraîne également un choix de l'algorithme d'estimation ainsi que celui d'une possible statistique de test. Quand le nombre de ruptures est inconnu, le nombre de ruptures peut également être estimé en utilisant un critère de choix de modèle ou de pénalisation.

De nombreux problèmes de détection de ruptures ont été étudiés dans le cas de processus indépendants (Csörő et Horváth, 1988, 1997), (Brodsky et Darkhovsky, 1993) faiblement dépendants (Kokoszka et Leipus 1999, 2000), (Horváth, Kokoszka et Teyssière, 2001) et fortement dépendants (Giraitis, Leipus et Surgailis 1996), (Lavielle, 1999), (Kokoszka et Leipus, 2003).

Des différentes approches, certaines sont paramétriques. Le principe est alors de considérer un vecteur de paramètres θ qui décrit les propriétés des observations. Avant un instant t_0, le paramètre θ est égal à θ_0 alors qu'au delà de l'instant de rupture, il est égal à $\theta_1 \neq \theta_0$. C'est l'exemple de détection de rupture selon la moyenne et/ou la variance. Dans d'autres cadres non-paramétriques, la distribution ou la densité spectrale peuvent être choisies comme critère de détection de changements.

Le problème de détection de rupture a été étudié dans le cadre d'une seule rupture et généralisé en de multiples ruptures qui peuvent être détectées successivement, avec, par exemple, la méthode de segmentation binaire (Vostrikova, 1981) qui permet de trouver tous les points de changements possibles ou simultanément (Lavielle et Teyssière 2005,2006) avec un critère de choix de modèle (pénalisation) permettant d'estimer le nombre de changements adapté.

Dans la littérature de nombreuses statistiques de tests de détection de ruptures ont été étudiées telles les statistiques basées sur la moyenne (Sen et Srivastava, 1975), celles basées sur le maximum de vraisemblance (Sen et Srivastava, 1975), (Cobb, 1978) ou encore (Csörő et Horváth, 1988) qui ont proposé des méthodes basées sur les rangs de la fonction de répartition empirique. Pour déterminer la loi de la statistique de test, Deshayes et Picard (1986) ont proposé des méthodes asymptotiques.

Plus proche des processus considérés dans notre travail, un test construit à partir d'un théorème de la limite centrale pour les formes quadratiques a aussi été développé dans (Beran et Terrin, 1996), (Horváth et Shao, 1999). Dans ces dernières références, comme dans (Ray et Tsay, 2002), on décide si le paramètre de dépendance longue mémoire caractéristique de la série change ou ne change pas en fonction du temps. En effet, pour certains processus fortement dépendants, le paramètre de longue mémoire peut varier en fonction du temps et une légère variation peut avoir un grand impact dans les caractéristiques statistiques de tels processus comme cela a été démontré par Beran

et Terrin (1996). Ayache *et al.* (2006) ont également proposé dans le cas du processus SFBM des estimateurs de ces points de ruptures. Ces points sont estimés séparément dès qu'ils sont assez espacés. Dans le cas d'un seul point de rupture, l'estimateur est obtenu comme le premier instant où une fonction dépendant du temps dépasse un certain seuil, cette fonction étant construite à partir de variations quadratiques généralisées.

1.4 Organisation de la thèse

A travers cette thèse, on essaye de comprendre l'évolution des signaux relatifs aux fréquences cardiaques, de détecter leurs changements de comportements et de les modéliser.

La première étape qui fait l'objet du deuxième chapitre, présente un pré-traitement des signaux physiologiques. En effet, le signal brut relatif à la fréquence cardiaque de chaque athlète peut présenter différents régimes : celui avant le début de la course, la transition (enregistrée entre le début de la course et le palier de FC atteint durant l'effort), la phase de FC au cours de l'exercice, la phase d'arrivée jusqu'à la fin de course ainsi que la phase de récupération. Le premier traitement sur le signal consiste à détecter le début et la fin de la course. A cet effet nous avons eu recours à la méthode de détection de ruptures développée par Lavielle (1999). D'autres étapes de pré-traitements ont été nécessaires telles que la détection et la correction des données aberrantes moyennant un lissage de Kalman par intervalles, réalisé récursivement. Afin de détecter un possible changement de comportement du rythme cardiaque durant les trois phases caractéristiques du marathon, une première modélisation par un bruit gaussien fractionnaire est proposée de paramètre long mémoire H. Ce paramètre pourrait être une nouvelle manière d'interpréter et d'expliquer les comportements physiologiques.

La plupart des estimateurs de ce paramètre, appelé également par les physiologues, coefficient de la loi de puissance, consiste à réaliser un ajustement par régression linéaire d'une quantité en fonction d'échelles dans une représentation logarithmique. Ceci inclut la méthode d'analyse des fluctuations redressées (DFA) et la méthode d'analyse par ondelettes. Durant ces dernières années, un certain nombre d'auteurs travaillant sur des signaux physiques ou biologiques ont fréquemment utilisé la méthode DFA, en particulier lorsque ces signaux semblent ne pas être stationnaires. Le troisième chapitre présente une analyse des propriétés asymptotiques de la fonction DFA dans le cas du bruit gaussien fractionnaire, ce qui nous a permis, sous certaines conditions, de prouver la convergence de l'estimateur du paramètre de Hurst. Ces résultats ont été aussi généralisés à une classe semi-paramétrique de processus stationnaires longue

mémoire gaussiens. Les propriétés de la fonction DFA ont été aussi étudiées dans différents cas particuliers de processus avec tendances, et il s'avère que la méthode n'est pas robuste dans de tels cas. Pour aller un peu plus loin, on peut dire que le chapitre 3 montre l'absence d'intérêt d'utiliser la méthode DFA, que ce soit dans le cas stationnaire longue mémoire, où des méthodes comme celles du log-périodogramme donnent de bien meilleurs résultats, ou même dans le cas de processus composés d'un bruit longue mémoire et d'une tendance.

Ceci n'est pas le cas de la méthode d'analyse par ondelettes. En effet, au niveau du quatrième chapitre, on montre que cette méthode fournit des résultats bien plus robustes et la possibilité de traiter des modèles plus généraux. Elle permet la construction de processus semi-paramétriques (les bruits gaussiens localement fractionnaires) qui se révèlent être très pertinents pour modéliser les données relatives aux fréquences cardiaques. L'analyse par ondelettes montre également une évolution du paramètre de fractalité locale (une sorte de paramètre de Hurst pour certaines fréquences) pendant la course, ce qui confirme des résultats obtenus par Peng *et al.* dans leur étude portant sur des enregistrements de fréquences cardiaques durant un exercice pour des personnes saines (où le paramètre observé est proche de celui estimé en début de course) et pour des personnes ayant une insuffisance cardiaque (où le paramètre observé est proche de celui estimé en fin de course). Cette évolution, qui ne peut pas être observée par la méthode DFA, peut être associée à l'effet de fatigue qui apparaît durant la phase finale du marathon.

Ces résultats ont été aussi confirmés lors du cinquième chapitre qui étudie la détection de possibles changements du paramètre de longue mémoire, d'autosimilarité et de fractalité locale. Dans ce chapitre, on considère le cadre d'un échantillon d'un processus gaussien stationnaire (respectivement séries temporelles, processus continu ayant des accroissements stationnaires) dont le paramètre de longue mémoire (respectivement d'autosimilarité et de régularité locale) évolue par paliers en fonction du temps. On s'intéresse alors à l'estimation de ces points de ruptures et à étudier les propriétés asymptotiques de ces estimateurs. On obtient alors un théorème limite pour l'estimateur de ces points de ruptures, un théorème de la limite centrale est aussi établi pour l'estimateur des paramètres. Enfin, un test d'ajustement est construit dans chaque zone où le paramètre est inchangé. Ceci est particulièrement intéressant dans le cadre des données de fréquences cardiaques. En effet, dans les chapitre 2 et 4, nous avons utilisé des détections de ruptures selon la moyenne et la variance afin de déduire les phases caractéristiques du marathon (au nombre de 3 le plus souvent : début, milieu et fin de course). Cependant une détection de ruptures selon le paramètre de fractalité locale H serait plus appropriée car c'est essentiellement l'évolution de ce paramètre qui nous intéresse. Nous verrons alors que l'on retrouve les résultats du chapitre 4, à savoir une

augmentation de H au cours de la course.

Chapitre 2

Data processing and modeling

2.1 Introduction

Fifty athletes were followed during a marathon (Paris Marathon 2004) and for each one, various physiological parameters are measured. Each heart rates (HR) signal, recorded instantaneously on cardio-frequency meter (CFM), corresponds to an endurance type effort observed on a course of 42km realized, on average, in 3h14mn. For each runner, the periods (in ms) between the successive pulsations (see Fig. 2.1) are recorded. The HR signal in number of beats per minute (bpm) is then deduced (the HR average for the whole sample is of 162 bpm).

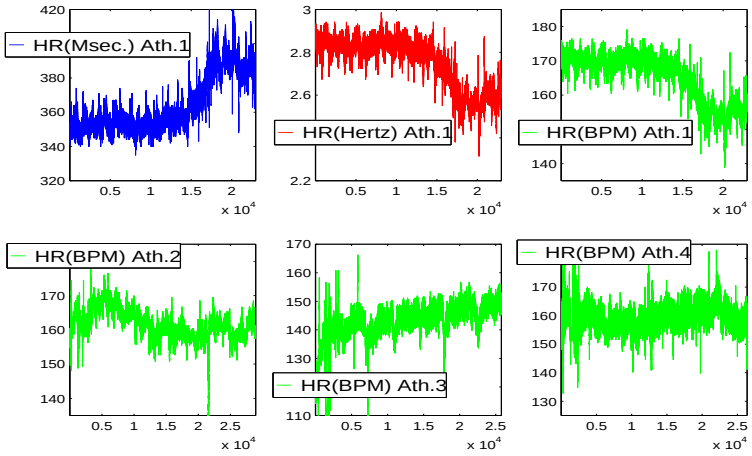


FIG. 2.1 – Heat rate signals of Athlete 1 in ms, Hertz and BPM (up), of Athletes 2, 3 and 4 in BPM (down)

During effort, one or more phases can be observed, which evolve and change differently from an athlete to another. Moreover, data depend in particular on the installation of the CFM. So, as a first step of this study, a pretreatment of these data is proposed for "cleaning" them of outliers and detecting different significative stages during the race. In Section 2.3, two methods are presented for estimating the regularity parameter : the DFA method and wavelet analysis which is developed for a more general models. The last of section is devoted to applications of both methods to generated data and HR data.

2.2 Data processing

2.2.1 Abrupt change detection

HR data of each athlete may show various modes : before the race beginning, during a transition step (recorded between the race beginning and the stage of HR reached during the effort), the main stage during the exercise, an arrival phase until the race end and sometimes a recovery phase. For distinguishing these different steps, a method of change points detection developed by Lavielle (see for instance (63)) is adapted and applied.

To begin with, a first treatment consists in detecting the beginning and the end of the race. The main idea is to consider that the signal distribution depends on a vector of unknown characteristic parameters in each stage. The different stages (before, during and after the race) and therefore the different vectors of parameters, change at two unknown instants (here the number of change points is known, but the method can be also used even if its number is unknown by adding a penalization term, see above). For instance and it will be our choice, changes in mean and variance can be detected.

2.2.1.1 General principle of the method of change detection

Assume that a sample of a time series $(Y(i), i = 1, \ldots, n)$ is observed. Assume also that it exists $\tau = (\tau_1, \tau_2, \ldots, \tau_{K-1})$ with $0 = \tau_0 < \tau_1 < \tau_2 < \ldots < \tau_{K-1} < n = \tau_K$ and such that for each $j \in \{1, 2, \ldots, K\}$, the distribution law of $Y(i)$ is depending on a parameter $\theta_j \in \Theta \subset \mathbb{R}^d$ (with $d \in \mathbb{N}$) for all $\tau_{j-1} < i \leq \tau_j$. Therefore, K is the number of segments to be deduced starting from the series and $\tau = (\tau_1, \tau_2, \ldots, \tau_{K-1})$ is the ordered change instants.

Now, define a contrast function

$$U_\theta\big(Y(\tau_j+1), Y(\tau_j+2), \ldots, Y(\tau_{j+1})\big),$$

of $\theta \in \mathbb{R}^d$ applied on each vector $\big(Y(\tau_j+1), Y(\tau_j+2), \ldots, Y(\tau_{j+1})\big)$ for all $j \in \{0, 2, \ldots, K-1\}$. A general example of such a contrast function is

$$U_\theta\big(Y(\tau_j+1), Y(\tau_j+2), \ldots, Y(\tau_{j+1})\big) = -2\log L_\theta\big(Y(\tau_j+1), Y(\tau_j+2), \ldots, Y(\tau_{j+1})\big),$$

where L_θ is the likelihood. Then, for all $j \in \{0, 2, \ldots, K-1\}$, define :

$$\widehat{\theta}_j = \underset{\theta\in\Theta}{\text{Argmin}}\ U_\theta\big(Y(\tau_j+1), Y(\tau_j+2), \ldots, Y(\tau_{j+1})\big).$$

Now, set :

$$\widehat{G}(\tau_1, \ldots, \tau_{K-1}) = \sum_{j=0}^{K-1} U_{\widehat{\theta}_j}\big(Y(\tau_j+1), Y(\tau_j+2), \ldots, Y(\tau_{j+1})\big)$$

As a consequence, an estimator $(\widehat{\tau}_1, \ldots, \widehat{\tau}_{K-1})$ can be defined as :

$$(\widehat{\tau}_1, \ldots, \widehat{\tau}_{K-1}) = \underset{0<\tau_1<\tau_2<\ldots<\tau_{K-1}<n}{\text{Argmin}}\ \widehat{G}(\tau_1, \ldots, \tau_{K-1}). \tag{2.1}$$

The principle of such method of estimation is very general (it can be also devoted to estimate abrupt change in polynomial trends) and different asymptotic behavior of the estimator $(\widehat{\tau}_1, \ldots, \widehat{\tau}_{K-1})$ can be deduced under general assumption on the time series Y (see for instance Bai Perron, Lavielle Moulines and Lavielle). For HR data, it is obvious that the beginning and the end of the race implies respectively an increasing (respectively decreasing) of the mean of HR. However, for avoiding all confusion linked for instance to the stress of the runner or other harmful noises, it was chosen to detect a change in mean and variance.

2.2.1.2 Change detection in mean and variance

Therefore, for all $j \in \{0, 1, \ldots, K-1\}$, consider the following general model :

$$Y(i) = \mu_j + \sigma_j \varepsilon_i \ \text{ for all } i \in \{\tau_j+1, \ldots, \tau_{j+1}\},$$

where $\theta_j = (m_j, \sigma_j) \in \mathbb{R} \times (0, \infty)$ and (ε_i) is a sequence of zero-mean random variables with unit variance.
In the case of changes in both mean and variance, and it is such a framework we consider for the heart rates series, a "natural" contrast function is defined by :

$$U_{\theta_j}\big(Y(\tau_j+1), \ldots, Y(\tau_{j+1})\big) = \sum_{\ell=\tau_j+1}^{\tau_{j+1}} \frac{(Y(\ell)-m_j)^2}{\sigma_j^2},$$

and therefore the well-known estimator of θ_j is :

$$\widehat{\theta}_j = (\widehat{m}_j, \widehat{\sigma}_j) = \Big(\frac{1}{\tau_{j+1} - \tau_j} \sum_{\ell=\tau_j+1}^{\tau_{j+1}} Y(\ell), \frac{1}{\tau_{j+1} - \tau_j} \sum_{\ell=\tau_j+1}^{\tau_{j+1}} \big(Y(\ell) - \widehat{m}_j\big)^2 \Big).$$

Now, the estimator $(\widehat{\tau}_1, \ldots, \widehat{\tau}_{K-1})$ can be deduced from (2.1).

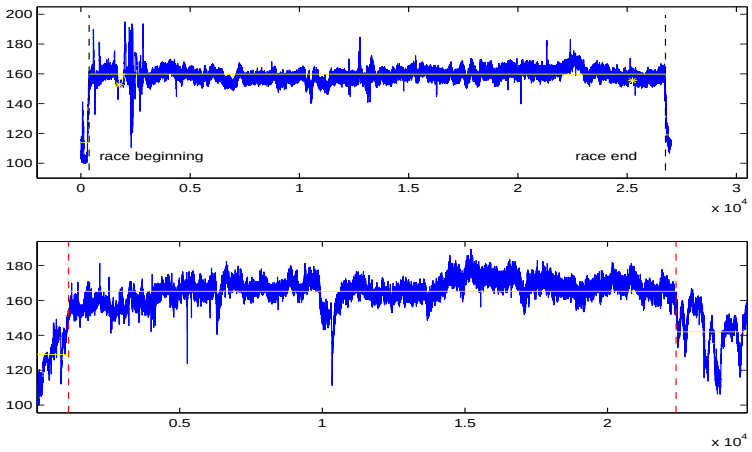


FIG. 2.2 – Detection of the race beginning and end from HR data (in BPM)

This method was applied to the different HR data for detecting the beginning and the end of the race. For avoiding the possibility of estimate an abrupt change during the race (explained for instance by a stop for drinking and eating), both $\widehat{\tau}_1$ (instant of the beginning) and $\widehat{\tau}_2$ were assumed to satisfied $\widehat{\tau}_1 \leq 1500$ and $n - \widehat{\tau}_2 \leq 1500$, nearly corresponding to less than 10mn after the beginning of data record and before the end of data record. The Fig. 2.2 exhibits an example of an application of the method to HR data.
The final race time of each athlete (that is known) can be compared to the aggregation of beats periods between $\widehat{\tau}_1$ and $\widehat{\tau}_2$. For all athletes, the difference between those two ways of measuring the same time is very often smaller than $1mn$. However, for several athletes, an important difference appears corresponding to truncated HR series or athlete forgetting to start their ECF at time.

2.2.2 Data smoothing

After this first step of detection of race beginning and end, a correction of aberrant data (due very often to a temporal bad contact between the athlete skin and the ECF) was needed. During exercise, the variation between two successive beats should not

exceed ±10% (see for instance ((7), (80))). This can be also justified by observing the empirical distribution of HR data (see histograms in Fig. 2.3). Thus, for such an histogram, its form should not be spread out on both sides of ±10%. The detection of aberrant data consists to observe the increments series $(C(i))_{1\leq i<n}$ of the signal $(Y(i))_{1\leq i\leq n}$ as well as the decrements series $(C'(i))_{1<i\leq n}$, with :

$$C(i) = \frac{Y(i+1) - Y(i)}{Y(i)}, \qquad C'(i) = \frac{Y(i-1) - Y(i)}{Y(i)},$$

and to find the observation of which the relative increments exceed ±10%. For example (see Fig. 2.3) for a HR series of 26380 observations, it was found 32 observations which have to be corrected (186 was found for another HR series with 27077 observations).

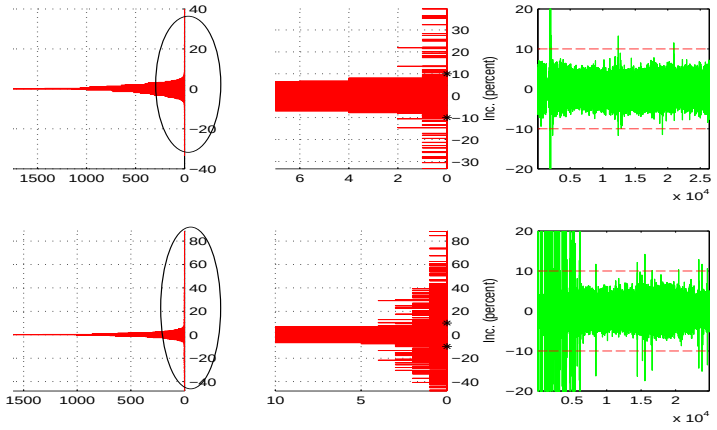


FIG. 2.3 – Plot of the increments of observed HR series for Ath1 (top) and Ath2 (bottom)

For "cleaning" HR data, abnormal observations have to be replaced by suitable others. For determining these new values, various procedures were applied.
First, an exponential smoothing is applied. It consists in replacing an abnormal observation by a linear combination of all the past observations affected by decreasing weights. However, this method is not able to correct every abnormal observation and there remain always some increments of frequencies which exceed ±10% (see Fig. 2.4).

For improving these results, a recursive method was considered : the Kalman smoothing (see for instance (67)). It is a problem of smoothing on a fixed time interval where one seeks to calculate the optimal approximation of a series value knowing the observations in the selected interval. This problem like the filtering and forecast ones is solved recursively (49).

2.2.3 Fixed-interval Kalman smoothing

Considers the general state space model :

$$\begin{cases} Z(t+1) &= A(t) \cdot Z(t) + \varepsilon(t) \\ Y(t) &= C(t) \cdot Z(t) + \eta(t) \quad , t > 0 \end{cases}$$

where $Y(1), \ldots, Y(T)$ the observations, $A(t)$ and $C(t)$ are deterministic. We have to calculate an approximation of $Z(t)$:

$$\widehat{Z}^T(t) = \mathbb{E}(Z(t)/Y(1), \ldots, Y(T))$$

and $\Sigma^T(t) = \mathrm{Var}\big(Z(t) - \widehat{Z}^T(t)\big)$. It is a problem of smoothing on a fixed interval $\{1, \ldots, T\}$. It is solved recursively.

Indeed, for $t = 1, \ldots, T-1$, we have :

$$\widehat{Z}^T(t) = \widehat{Z}^t(t) + F(t) \cdot \big(\widehat{Z}^T(t+1) - \widehat{Z}^t(t+1)\big)$$

where $F(t) = \Sigma^t(t) \cdot A(t)' \cdot \big(\Sigma^t(t+1)\big)^{-1}$ and

$$\Sigma^T(t) = \Sigma^t(t) + F(t) \cdot \big(\Sigma^T(t+1) - \Sigma^t(t+1)\big) \cdot F'(t).$$

In practice, quantities $\widehat{Z}^T(t)$ and $\Sigma^T(t)$ are calculated recursively by going up time with initial conditions $\widehat{Z}^T(T)$ and $\Sigma^T(T)$ for $t = T-1$. These values like all the quantities $\widehat{Z}^t(t)$, the forecast $\widehat{Z}^t(t+1)$, the mean square error of filtering $\Sigma^t(t)$ and that of forecast $\Sigma^t(t+1)$ for $t = T-1, \ldots, 1$ are updated during the phase of filtering.

Indeed, for $t > 0$, $\widehat{Z}^t(t) = \mathbb{E}(Z(t)/Y(1), \ldots, Y(t))$ is calculated as follows :

$$\widehat{Z}^t(t) = \widehat{Z}^{t-1}(t) + K(t) \cdot \big(Y(t) - C(t) \cdot \widehat{Z}^{t-1}(t)\big)$$

where $K(t) = \Sigma^{t-1}(t) \cdot C'(t) \cdot \big(C(t) \cdot \Sigma^{t-1}(t) \cdot C'(t) + R\big)^{-1}$ with $R = \mathrm{Var}(\eta(t))$. The mean square error of filtering on $Z(t)$ at t is :

$$\Sigma^t(t) = (Id - K(t) \cdot C(t)) \cdot \Sigma^{t-1}(t)$$

The forecast of $Z(t+1)$ made at instant t is such as :

$$\widehat{Z}^t(t+1) = A(t) \cdot \widehat{Z}^t(t)$$

The corresponding mean square error of forecast :

$$\Sigma^t(t+1) = A(t) \cdot \Sigma^t(t) \cdot A'(t) + Q \quad \text{where} \quad Q = \mathrm{Var}(\varepsilon(t)).$$

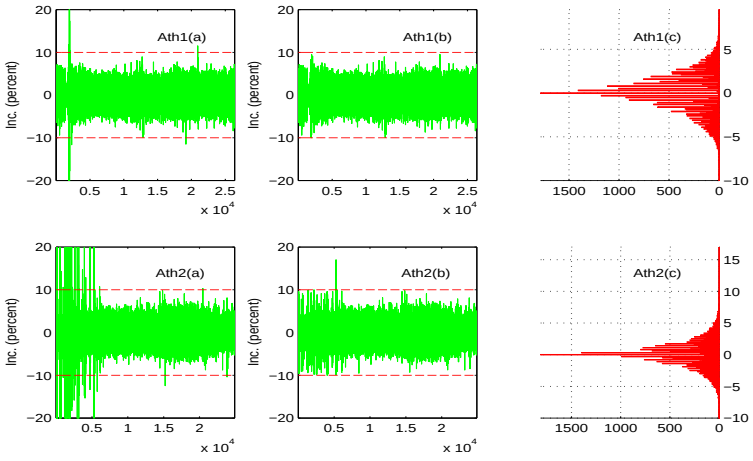


FIG. 2.4 – (a) Increments in HR time series after exponential smoothing (b) Increments in HR time series after Kalman smoothing (c) Histogram of increments after processing

Contrary to a simple exponential smoothing, the Kalman smoothing, applied in this example on an interval of 260 observations, presents a clear improvement of the results. Often, only one iteration of the Kalman smoothing was needed to correct the whole signal (in the sense that increments do not exceed $\pm 10\%$). In the other cases, the procedure is repeated for different selected intervals. For example (see Fig. 2.4), one iteration was needed to correct 31 observations and a second one for the remaining observation. In a second HR series, after the first iteration, there remain 10 observations to be corrected and it was done after 6 other iterations.

2.2.4 Detection of the different stages of a race

It is also interesting to distinguish the different stages during the race in order to unveil if a change of behavior was happened. These stages can be detected using the previous method of change points detection (see Fig. 2.5). The procedure is exactly the same except that the number of changes is unknown and can be also estimated. Thus, a new contrast V is built by adding to the previous contrast U an increasing function depending on the change number K, *i.e.* more precisely,

$$\widehat{V}(\tau_1, \ldots, \tau_{K-1}, K) = \widehat{G}(\tau_1, \ldots, \tau_{K-1}) + \beta \times \text{pen}(K),$$

with $\beta > 0$. As a consequence, by minimizing V in $\tau_1, \ldots, \tau_{K-1}, K$, an estimator $\widehat{K}$ is obtained which varies with the penalization parameter β.

For HR data, the choice of pen(K) was K. Let $\widehat{G}_K = \widehat{G}(\widehat{\tau}_1, \ldots, \widehat{\tau}_{K-1})$, for $K =$

$K_1, \ldots, K_{MAX}$ we define

$$\beta_i = \frac{\widehat{G}_{K_i} - \widehat{G}_{K_{i+1}}}{K_{i+1} - K_i} \quad \text{and } l_i = \beta_i - \beta_{i+1} \quad \text{with } i \geq 1.$$

Then the retained K is the greatest value of K_i such that $l_i >> l_j$ for $j > i$.
Applied to the whole set HR data, the number of abrupt changes is estimated at 4 or 3. Three phases were selected to be studied, which are located in the beginning of the race, in the middle and in the end. However for certain recorded signals the first or the last phase can not be distinguished probably for measurements reasons.

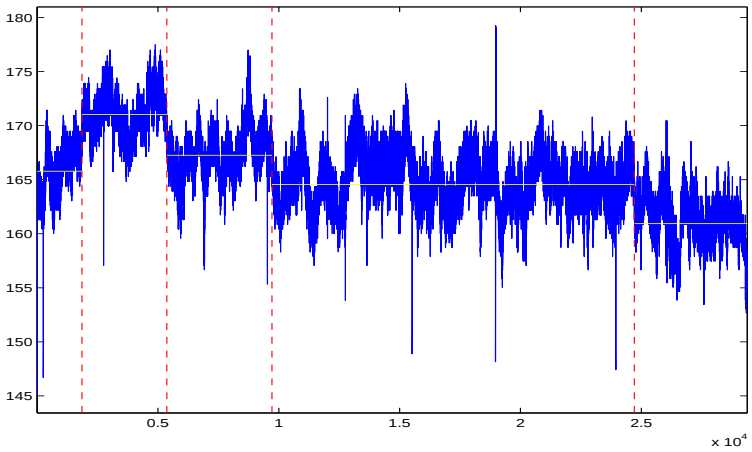


FIG. 2.5 – The estimated configuration of changes in a HR time series of an athlete

2.3 HR data modeling with a long range dependent process

In this section, a first model is proposed for modeling HR data. After a statistic study showing a badness-of-fit of this model to the data, a more suitable model is defined. Then, using a wavelet based procedure, some physiologic conclusions can be obtained from HR data.

When we observe entire or partial (during the three phases) HR time series, we remark that it exhibits a certain persistence and the related correlations decays very slowly with time what characterizes trajectories of a long memory Gaussian noise (see Figure 1.2 in Chapter 1). Also, the distribution of data recorded during the phases leads as to suspect a Gaussian behaviour in these data. Of course this is only an assumption and we can check it with tests considered for long range dependent processes. But in our case we will try to test whether a Gaussian process could model these data.

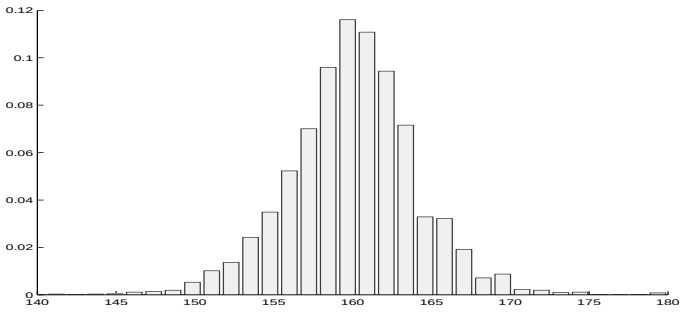


FIG. 2.6 – Distribution of data during the race recorded for one athlete which seems to be gaussian

Moreover, the aggregated signals (see for example Fig. 2.7) present a certain regularity very close to that of fractional Brownian motion simulated trajectories with a parameter close to 1 (Fig. 2.9). So, one first model which could correspond to our data is the fractional Gaussian noise.

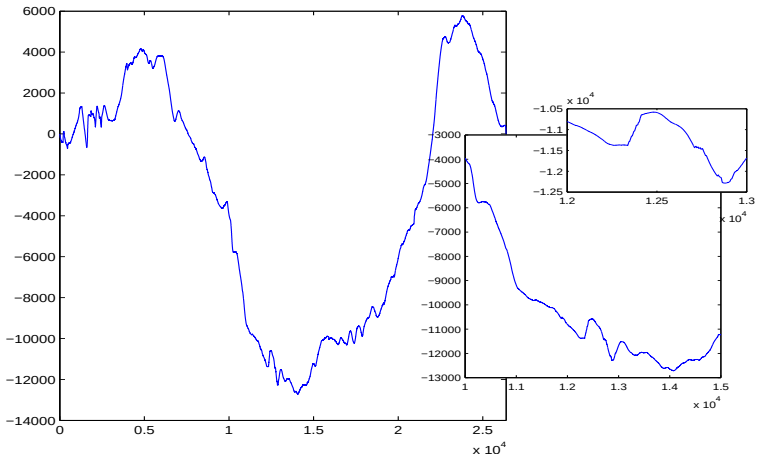


FIG. 2.7 – The self-similarity of the aggregated HR signals (representation of the aggregated HR fluctuations at 3 different time resolutions)

The following Figure 2.8 presents a comparison between the graphs of HR data during a stage (detected previously) and a fractional Gaussian noise (FGN in the sequel) with parameter $H = 0.99$ (see the definition above). Before using statistical tools for testing the similarities of both these graphs, let us remind some elements concerning the FGN.

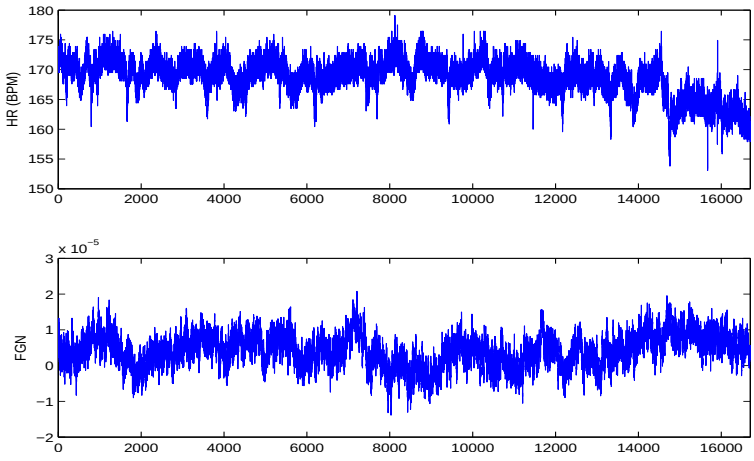


FIG. 2.8 – Comparison of HR data in the middle of race (Ath4) and generated FGN(H=0.99) trajectories

2.3.1 A first model : the fractional Gaussian noise

The FGN is one of the most famous example of stationary long range dependent (LRD in the sequel) process. The LRD phenomenon was observed in many fields including telecommunication, hydrology, biomechanic, economy... A stationary second order process $Y = \{Y(k),\, k \in \mathbb{N}\}$ is said to be a LRD process if :

$$\sum_{k\in\mathbb{N}} |r_Y(k)| = \infty \quad \text{with} \quad r_Y(k) = \mathbb{E}\big[Y(0)Y(k)\big].$$

Thus $Y(k)$ is depending on $Y(0)$ even if k is a very large lag. Another way for writing the LRD property is the following :

$$r_Y(k) \sim k^{2H-2}L(k) \text{ , as } k \to \infty,$$

with $L(k)$ a slowly varying function (*i.e.* $\forall t > 0$, $L(xt)/L(x) \to 1$ when $x \to \infty$) and the Hurst parameter $H \in (\frac{1}{2}, 1)$.
The LRD is closely related to the self-similarity concept. A process $X = \{X(t),\, t \geq 0\}$ is so called a self-similar process with self-similarity exponent H, if $\forall c > 0$:

$$\big(X(ct)\big)_t \overset{\mathcal{L}}{=} c^H \big(X(t)\big)_t.$$

Now, if we consider the aggregated process $\{X(t),\, t \geq 0\}$ defined by $X(k) = \sum_{i=1}^{k} Y(i)$ with Y a LRD process, then under weak conditions (for instance Y is a Gaussian or a causal process), it can be proved that, roughly speaking, for $k \to \infty$, the law of $\{X(t), t \geq k\}$ is a self-similar law (see Doukhan *et al.*, 2003, for more details).
The FGN is an example of a LRD Gaussian process. More precisely, $Y^H = \{Y^H(k),\, k \in$

$\mathbb{N}\}$ is a FGN,

$$r_{Y^H}(k) = \frac{\sigma^2}{2}(|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}) \quad \forall k \in \mathbb{N},$$

with $H \in (0,1)$ and $\sigma^2 > 0$ (it can be proved that such a Gaussian time series exists, *i.e.* all covariance matrix of any vector is a Toeplitz positive definite matrix, see for instance more details in Samorodnitsky and Taqqu, 1994). As a consequence, for $H \in (\frac{1}{2}, 1)$, a usual Taylor formula implies

$$r_{Y^H}(k) \sim \sigma^2 H(2H-1)k^{2H-2} \text{ , when } k \to \infty.$$

For a zero-mean FGN, the corresponding aggregated process, denoted here X^H, is so-called the fractional Brownian motion (FBM) and X^H is a self-similar Gaussian process with self-similar parameter H and therefore satisfies,

$$\mathrm{Var}(X^H(k)) = \sigma^2|k|^{2H} \quad \forall k \in \mathbb{N}$$

(it can be even proved that X^H is the only Gaussian self-similar process with stationary increments). It is obvious that $Y^H(k) = X^H(k) - X^H(k-1)$, the sequence of the increments of a FBM, is a FGN.

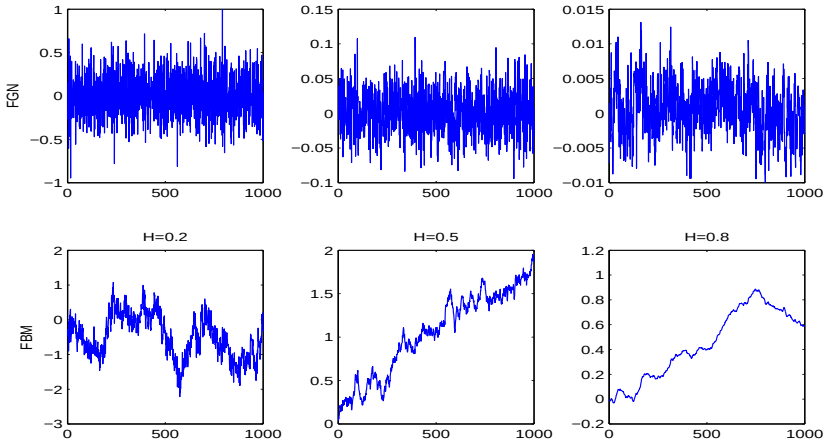


FIG. 2.9 – Generated FGN trajectories and corresponding aggregated series (FBM) for $H = 0.2 < 0.5$ anti-persistant noise (left), $H = 0.5$ white noise (center) and $H = 0.8 > 0.5$ LRD process (right)

Several generated trajectories of FGN and corresponding FBM are presented in Fig. 2.9 for different values of H.

2.3.2 Methods of estimations of the Hurst parameter

For testing if a HR path can be suitably model by a FGN, a first step consists in estimating H. Here we chose to use two estimators (but there exist many else, see for instance Doukhan *et al.*, 2003) that are known to be unchanged to the presence of a possible trend.

So, several common estimators of this parameter, so-called scaling behavior exponents, consist in performing a linear regression fit of a scale-dependent quantity versus the scale in a logarithmic representation. This includes the Detrended Fluctuation Analysis (DFA) method which is frequently used in the case of physiological data processing and wavelet analysis method. However, in Chapter 3, even if the DFA estimator of Hurst parameter is proved to be convergent with a reasonable convergence rate for LRD stationary Gaussian processes, it is not at all a robust method in case of trend. The wavelet based method, presented in Chapter 4, provides a more precise and robust estimator of the Hurst parameter. Thus, the results obtained from this wavelet estimator seem to be more valid.

Chapitre 3

Asymptotic Properties of the Detrended Fluctuation Analysis of Long Range Dependent Processes

3.1 Introduction

In the past few years, numerous methods of analysis of a trended long range process have been proposed. One of these methods is the Detrended Fluctuation Analysis (DFA), frequently used in the case of physiological data processing in particular the heartbeat signals recorded on healthy or sick subjects (see for instance (55), (59), (75), (76) and (77)). Indeed, it can be interesting to find some constants among the fluctuations of physiological data. The parameter of long-range dependence (also called the Hurst parameter) of the original signal, or the self-similarity parameter of the aggregated signal could be a new way of interpreting and explaining a physiological behavior.

The DFA method is a version for time series with trend of the method of aggregated variance used for a long-memory stationary process (see for instance (59)). It consists in 1. aggregating the process by windows with fixed length, 2. detrending the process from a linear regression in each window, 3. computing the standard deviation of the residual errors (the DFA function) for all data, 4. estimating the coefficient of the power law from a log-log regression of the DFA function on the length of the chosen window. After the first stage, the process is supposed to behave like a self-similar process with stationary increments added to a trend. The second stage is supposed to remove the trend. Finally, the third and fourth stages are identical to those of the aggregated me-

thod (for zero-mean stationary process).

The processing of experimental data, and in particular physiological data, exhibits a major problem which is the non-stationarity of the signal. Hu *et al.* (2001) have studied different types of non-stationarities associated with examples of trends (linear, sinusoidal and power-law trends) and deduced their effect on an added noise and the kind of competition which exists between this two signals. They have also explained (see Chen *et al.*, 2002) the effects of three other types of non-stationarities, which are often encountered in real data. The DFA method was applied to signals with having some segments, with random spikes or with different local behavior. The results were compared with the case of stationary correlated signals.

In Taqqu *et al.* (1995), the case of the fractional Gaussian noise (FGN) is studied. A theoretical proof to the power law followed by the expectation of the DFA function of this process is established. This is an important first step in order to prove the convergence of the estimator of the Hurst parameter. The study we propose here constitutes an accomplishment of this work. Indeed the convergence rate of the Hurst parameter estimator is obtained, in a semi-parametric frame.

The paper is organized as follows. In Section 3.2, the DFA method is presented and two general properties are proved. Section 3.3 is devoted to providing asymptotic properties (illustrated beforehand by simulations) of the DFA function in the case of the FGN. Section 3.4 contains an extension of these results for a general class of stationary long-range dependence processes. Finally, in Section 5.2.1, the method is proved not to be robust in different particular cases of trended processes. Indeed the trend is dominant in the case of power law and polynomial trends where the slope of the DFA log-log regression line for trended processes is always close to 2 or in the case of a piecewise constant trend where the slope is estimated at $\frac{3}{2}$, which dominates the Hurst exponent. The proofs of the different results are in Section 3.7.

3.2 Definitions and first properties of the DFA method

The DFA method was introduced in (76). The aim of this method is to highlight the self-similarity of a time series with trend. Let $(Y(1), \ldots, Y(N))$ be a sample of a time series $(Y(n))_{n\in\mathbb{N}}$.

1. The first step of the DFA method is a "discrete integration" of the sample, *i.e.*

the calculation of $(X(1),\dots,X(N))$ where

$$X(k)=\sum_{i=1}^{k}Y(i) \quad \text{for } k\in\{1,\dots,N\}. \tag{3.1}$$

2. The second step is a division of $\{1,\dots,N\}$ in $[N/n]$ windows of length n (for $x\in\mathbb{R}$, $[x]$ is the integer part of x). In each window, the least-square regression line is computed, which represents the linear trend of the process in the window. Then, we denote by $\widehat{X}_n(k)$ for $k=1,\dots,N$ the process formed by this piecewise linear interpolation. Then the DFA function is the standard deviation of the residuals obtained from the difference between $X(k)$ and $\widehat{X}_n(k)$, therefore,

$$F(n)=\sqrt{\frac{1}{n\cdot[N/n]}\sum_{k=1}^{n\cdot[N/n]}\Big(X(k)-\widehat{X}_n(k)\Big)^2}$$

3. The third step consists in a repetition of the second step with different values $(n_1,\dots,n_m)$ of the window's length. Then the graph of $\log F(n_i)$ by $\log n_i$ is drawn. The slope of the least-square regression line of this graph provides an estimation of the self-similarity parameter of the $(X(k))_{k\in\mathbb{N}}$ process or the Hurst parameter of the $(Y(n))_{n\in\mathbb{N}}$ process (see above for the explanations).

From the construction of the DFA method, it is interesting to define the restriction of the DFA function in a window. Thus, for $n\in\{1,\dots,N\}$, one defines the partial DFA function computed in the j-th window, *i.e.*

$$F_j^2(n)=\frac{1}{n}\sum_{i=n(j-1)+1}^{nj}(X(i)-\widehat{X}_n(i))^2 \tag{3.2}$$

for $j\in\{1,\dots,[N/n]\}$. Then, it is obvious that

$$F^2(n)=\frac{1}{[N/n]}\sum_{j=1}^{[N/n]}F_j^2(n). \tag{3.3}$$

Remark : In Hu *et al.* and Kantelhardt *et al.* papers (for details see (55), (58) and (59)), the definition of the time series $(X(n))_{n\in\mathbb{N}}$ computed from $(Y(n))_{n\in\mathbb{N}}$ is different from (3.1), *i.e.*

$$\tilde{X}(k)=\sum_{i=1}^{k}(Y(i)-\overline{Y}_N)\ , \quad \text{for} \quad k\in\{1,\dots,N\}$$

with $\overline{Y}_N=\dfrac{1}{N}\displaystyle\sum_{j=1}^{N}Y(j)$.

It is obvious that in both definitions, $(X(k)-\widehat{X}_n(k))$ is the same and therefore the value of $F(n)$ is the same.

Lemma 3.2.1. *With the previous notations, let $\tilde{F}(n)$ be the DFA function built from $(\tilde{X}(k))$, i.e.*

$$\tilde{F}(n) = \sqrt{\frac{1}{n \cdot [N/n]} \sum_{k=1}^{n \cdot [N/n]} \left(\tilde{X}(k) - \widehat{\tilde{X}}_n(k) \right)^2}.$$

Then for $n \in \{1, \ldots, N\}$, $F(n) = \tilde{F}(n)$.

Proof : Consider the j-th window, $j \in \{1, \ldots, [N/n]\}$ and define the vectors $X^{(j)} = (X(1 + n(j-1)), \ldots, X(nj))'$ and $\tilde{X}^{(j)} = (\tilde{X}(1 + n(j-1)), \ldots, \tilde{X}(nj))' = X^{(j)} - (1 + n(j-1), \ldots, nj)' \cdot \overline{Y}_N$. In this j-th window, define E_j the vector subspace of $\mathbb{R}^n$ generated by the two vectors of $\mathbb{R}^n$, $(1, \ldots, 1)'$ and $\big((j-1)n+1, (j-1)n+2, \ldots, nj\big)'$. It is well known that if P_A is the matrix of the orthogonal projection on a vector subspace A of $\mathbb{R}^n$, then

$$F_j^2(n) = \frac{1}{n} \left(P_{E_j^\perp} \cdot X^{(j)} \right)' \cdot P_{E_j^\perp} \cdot X^{(j)} \quad \text{and} \quad \tilde{F}_j^2(n) = \frac{1}{n} \left(P_{E_j^\perp} \cdot \tilde{X}^{(j)} \right)' \cdot P_{E_j^\perp} \cdot \tilde{X}^{(j)},$$

where $E_j^\perp$ is the orthogonal vector subspace of E_j.
But $(1 + n(j-1), \ldots, nj)' \cdot \overline{Y}_N \in E_j$, and therefore

$$P_{E_j^\perp} \cdot \tilde{X}^{(j)} \; = P_{E_j^\perp} \cdot X^{(j)} - P_{E_j^\perp} \cdot (1 + n(j-1), \ldots, nj)' \, \overline{Y}_N = P_{E_j^\perp} \cdot X^{(j)},$$

and thus, $F_j^2(n) = \tilde{F}_j^2(n)$, this implies that $F(n) = \tilde{F}(n)$. □

In order to simplify the following proofs, the case of the DFA method applied to a stationary process $\{Y(t),\, t \geq 0\}$ can be considered. The following lemma shows that the law of $F_j^2(n)$ does not depend on j and that the application of DFA to a stationary process yields a stationary process again.

Lemma 3.2.2. *Let $\{Y(t),\, t \geq 0\}$ be a stationary process. Then, with $X(k) = \sum_{i=1}^{k} Y(i)$ for $k \in \{1, \ldots, N\}$, for any $n \in \{1, \ldots, N\}$, the time series $(F_j^2(n))_{1 \leq j \leq [N/n]}$ is a stationary process.*

Proof : Set $j \in \{1, \ldots, [N/n]\}$ and define the vector $X^{(j)} = (X(1+n(j-1)), \ldots, X(nj))'$. Then,

$$X^{(j)} - X(n(j-1)+1) \cdot \; (1, \ldots, 1)' \stackrel{\mathcal{L}}{=} X^{(1)} - X(1) \cdot (1, \ldots, 1)'. \tag{3.4}$$

Indeed

$$X^{(j)} - X(n(j-1)+1) \cdot (1, \ldots, 1)'$$
$$= \Big(0, Y(2 + n(j-1)), \ldots, \sum_{k=2}^{n-1} Y(k + n(j-1)), \sum_{k=2}^{n} Y(k + n(j-1))\Big)$$

$$\text{and} \quad X^{(1)} - X(1)\cdot(1,\dots,1)' = \Big(0, Y(2), \dots, \sum_{k=2}^{n-1} Y(k), \sum_{k=2}^{n} Y(k)\Big)$$

We have $(Y(2),\dots,Y(n)) \overset{\mathcal{L}}{=} (Y(2+(j-1)n),\dots,Y(jn))$ because $\{Y(t), t \geq 0\}$ is a stationary process. Then with $g : \mathbb{R}^{n-1} \to \mathbb{R}^{n-1}$ a Borelian function defined by $g(y_2,\dots,y_n) = (y_2,\dots,\sum_{k=2}^{n-1} y_k, \sum_{k=2}^{n} y_k)$, it is clear that $g(Y(2),\dots,Y(n)) \overset{\mathcal{L}}{=} g(Y(2+(j-1)n),\dots,Y(jn))$ and therefore (3.4) is true.

Now, in each window j, and with the same definition of the vector subspace E_j as in the proof of Lemma 5.4.1,

$$\begin{aligned} F_j^2(n) &= \frac{1}{n}\left(P_{E_j^\perp}\cdot X^{(j)}\right)'\cdot P_{E_j^\perp}\cdot X^{(j)} \\ &= \frac{1}{n}\big(X^{(j)} - X(n(j-1)+1)\cdot(1,\dots,1)'\big)'\cdot P_{E_j^\perp}\cdot\big(X^{(j)} - X(n(j-1)+1)\cdot(1,\dots,1)'\big) \end{aligned}$$

with $P_{E_j^\perp}\cdot(1,\dots,1)' = (0,\dots,0)'$. But $E_1 = E_j$ and thus $E_j^\perp = E_1^\perp$. Therefore, with (3.4), we obtain $F_j^2(n) \overset{\mathcal{L}}{=} F_1^2(n)$ for all $j \in \{1,\dots,[N/n]\}$.

Moreover, for all $m \in \mathbb{N}^*$, $(j_1,\dots,j_m) \in \{1,\dots,[N/n]\}^m$ and $t \in \mathbb{N}^*$, the same reasoning can be used again for the case of vectors $(F_{j_1}^2(n),\dots,F_{j_m}^2(n))$ and $(F_{j_1+t}^2(n),\dots,F_{j_m+t}^2(n))$. Indeed,

$$\begin{aligned} &\Big(X^{(j_1)'} - X(n(j_1-1)+1)\cdot(1,\dots,1),\dots,X^{(j_m)'} - X(n(j_m-1)+1)\cdot(1,\dots,1)\Big)' \\ &\overset{\mathcal{L}}{=} \Big(X^{(j_1+t)'} - X(n((j_1+t)-1)+1)\cdot(1,\dots,1),.. \\ &\qquad\qquad .,X^{(j_m+t)'} - X(n((j_m+t)-1)+1)\cdot(1,\dots,1)\Big)' \end{aligned}$$

and $P_{E_{j_1}} = \cdots = P_{E_{j_m}} = P_{E_{j_1+t}} = \cdots = P_{E_{j_m+t}}$. This achieves the proof. □

3.3 Asymptotic properties of the DFA function for a FGN

In this section, we study the asymptotic behavior (both the sample size N and the length of window n increase to ∞) of the DFA when $(Y(n))_{n\in\mathbb{N}}$ is a stationary Gaussian process called a fractional Gaussian noise (FGN), *i.e.* $(X_1,\dots,X_N)$ is a Gaussian process having stationary increments and called a fractional Brownian motion (FBM). First, let us remind some definitions and properties of both these processes.

3.3.1 Definition and first properties of the FBM and the FGN

Let $\{X^H(t),\, t \geq 0\}$ be a fractional Brownian motion (FBM) with parameters $H \in]0,1[$ and $\sigma^2 > 0$, *i.e.* a real zero mean Gaussian process satisfying

1. $X^H(0) = 0$ a.s.
2. $E[(X^H(t) - X^H(s))^2] = \sigma^2 |t-s|^{2H} \quad \forall (t,s) \in \mathbb{R}_+^2$.

Here are some properties of a FBM $\{X^H(t),\, t \geq 0\}$ (see more details in Samorodnitsky and Taqqu, 1994)

- The process $\{X^H(t),\, t \geq 0\}$ has stationary increments. As a consequence, if we denote $\{Y^H(t),\, t \geq 0\}$ the process defined by $Y^H(t) = X^H(t+1) - X^H(t)$ for $t \geq 0$, then $\{Y^H(t),\, t \geq 0\}$ is a zero-mean stationary Gaussian process also called a fractional Gaussian noise (FGN).
- $\{X^H(t),\, t \geq 0\}$ is a self-similar process satisfying $\forall c > 0$, $X^H(ct) \overset{\mathcal{L}}{=} c^H X^H(t)$ and H is also called the exponent of self-similarity.
- The covariance function of a FBM $\{X^H(t),\, t \in \mathbb{R}\}$, for all $(s,t) \in \mathbb{R}^2$ is

$$\mathrm{Cov}(X^H(t), X^H(s)) = \frac{\sigma^2}{2}(|s|^{2H} + |t|^{2H} - |t-s|^{2H}). \tag{3.5}$$

- The covariance function of a FGN $\{Y^H(t),\, t \in \mathbb{R}\}$, for all $(s,t) \in \mathbb{R}^2$ is

$$\mathrm{Cov}(Y^H(t), Y^H(s)) = \frac{\sigma^2}{2}(|t-s+1|^{2H} + |t-s-1|^{2H} - 2|t-s|^{2H}). \tag{3.6}$$

  Therefore, $\mathrm{Cov}(Y^H(t), Y^H(s)) \sim H(2H-1)|t-s|^{2H-2}$ when $|t-s| \to \infty$: when $1/2 < H < 1$, Y^H is a long memory process (see also (3.11) below) and H is the Hurst (or long range dependent) parameter of Y^H.

3.3.2 Some numerical results of the DFA applied to the FGN

The following Figures 3.1 and 3.2 show an example of the DFA method applied to a FGN with different values of H ($H = 0.6$ in the first figure and $H = 0.2$, 0.4, 0.5, 0.7, 0.8 in the second one, with $N = 10000$ in both cases). Such a sample path is generated with a circulant matrix algorithm (see for instance Bardet *et al.*, 2002). Let us remark that if $(Y(n))_{n \in \mathbb{N}}$ is a sample path of a discretized FGN, then $(X(1), \ldots, X(N))$ is a sample path of the associated discretized FBM.

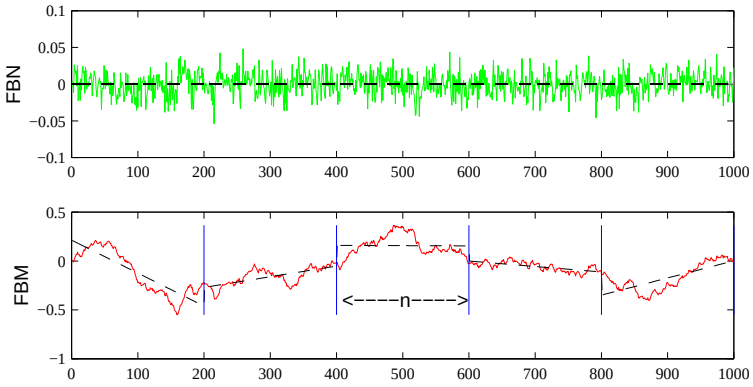


FIG. 3.1 – The two first steps of the DFA method applied to a path of a discretized FGN (with $H = 0.6$ and $N = 10000$)

On the right of Figure 3.2 appear the different estimations of H computed from the DFA method. Those values have to be compared with theoretical ones. The results seem to be quite good and it seems that, under certain conditions, the asymptotic behavior of the DFA function $F(n)$ can be written as

$$F(n) \simeq c(\sigma, H) \cdot n^H, \tag{3.7}$$

where c is a positive function depending only on σ and H (see its expression above). The approximation (3.7) explains that the slope of the least-square regression line of $(\log F(n_i))$ by $\log(n_i)$ for different values of n_i provides an estimation of H. We now provide a mathematical proof of this result.

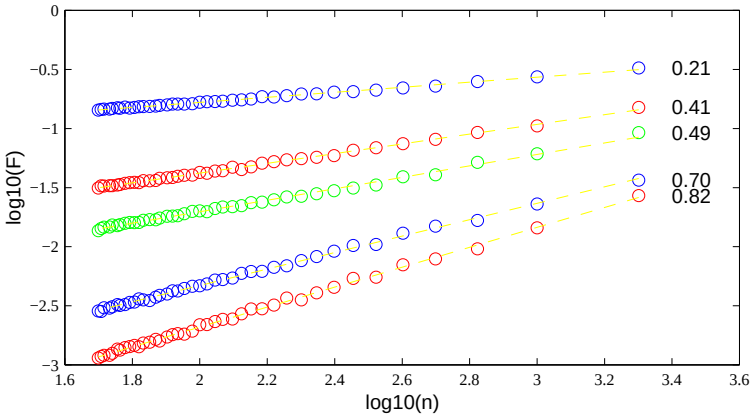


FIG. 3.2 – Results of the DFA method applied to a path of a discretized FGN for different values of $H = (0.2, 0.4, 0.5, 0.7, 0.8)$ (also with $N = 10000$)

Let $\{X^H(t),\, t \geq 0\}$ be a FBM, built as a cumulated sum of a FGN $\{Y^H(t),\, t \geq 0\}$. We first give some asymptotic properties of $F_1^2(n)$.

Property 3.3.1. *Let $\{X^H(t), t \geq 0\}$ be a FBM with parameters $0 < H < 1$ and $\sigma^2 > 0$. Then, for n and j large enough,*

$$\begin{aligned}
1.\ \ \mathbb{E}(F_1^2(n)) &= \sigma^2 f(H) \cdot n^{2H}\Big(1 + O(\frac{1}{n})\Big),\\
2.\ \ Var\big(F_1^2(n)\big) &= \sigma^4 g(H) \cdot n^{4H}\Big(1 + O(\frac{1}{n})\Big),\\
3.\ \ Cov(F_1^2(n), F_j^2(n)) &= \sigma^4 h(H) \cdot n^{4H} \cdot j^{2H-3} \cdot \Big(1 + O(\frac{1}{n}) + O(\frac{1}{j})\Big),
\end{aligned}$$

with $f(H) = \dfrac{(1-H)}{(2H+1)(H+1)(H+2)}$ *, g depending only on H, see (3.21), and* $h(H) = \dfrac{H^2(H-1)(2H-1)^2}{48(H+1)(2H+1)(2H+3)}$.

The proofs of these results (and of the others) are provided in Section 3.7.
In order to obtain a central limit theorem for the logarithm of the DFA function, one considers normalized DFA function

$$\tilde{S}_j(n) = \frac{F_j^2(n)}{n^{2H}\sigma^2 f(H)} \quad \text{and} \quad \tilde{S}(n) = \frac{F^2(n)}{n^{2H}\sigma^2 f(H)} \tag{3.8}$$

for $n \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, [N/n]\}$.
As a consequence, for $n \in \{1, \ldots, N\}$, the stationary time series $(\tilde{S}_j(n))_{1 \leq j \leq [N/n]}$ satisfy

$$\begin{cases}
\mathbb{E}(\tilde{S}_j(n)) &= 1 + O(\frac{1}{n})\\
\mathrm{Var}(\tilde{S}_j(n)) &= \dfrac{g(H)}{f(H)^2}\Big(1 + O(\frac{1}{n})\Big)\\
\mathrm{Cov}(\tilde{S}_1(n), \tilde{S}_j(n)) &= \dfrac{h(H)}{f(H)^2} \cdot \dfrac{1}{j^{3-2H}} \cdot \Big(1 + O(\frac{1}{n}) + O(\frac{1}{j})\Big)
\end{cases} \tag{3.9}$$

Under certain conditions on the asymptotic length n of the windows, one proves a central limit theorem satisfied by the logarithm of the empirical mean $\tilde{S}(n)$ of the random variables $(\tilde{S}_j(n))_{1 \leq j \leq [N/n]}$.

Property 3.3.2. *Under the previous assumptions and notations, let $n \in \{1, \ldots, N\}$ be such that $N/n \to \infty$ and $N/n^3 \to 0$ when $N \to \infty$. Then*

$$\sqrt{[\frac{N}{n}]} \cdot \log(\tilde{S}(n)) \underset{\substack{n\to\infty \\ N\to\infty}}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0, \gamma^2(H))),$$

where $\gamma^2(H) > 0$ depends only on H.

This result can be obtained for different lengths of windows satisfying the conditions $N/n \to \infty$ and $N/n^3 \to 0$. Let $(n_1, \ldots, n_m)$ be such different window lengths. Then,

one can write for N and n_i large enough

$$\log(\tilde{S}(n_i)) \simeq \frac{1}{\sqrt{[N/n_i]}} \cdot \varepsilon_i \quad \Longrightarrow$$
$$\log(F(n_i)) \simeq H \cdot \log(n_i) + \frac{1}{2}\log(\sigma^2 f(H)) + \frac{1}{\sqrt{[N/n_i]}} \cdot \varepsilon_i,$$

with $\varepsilon_i \sim \mathcal{N}(0, \gamma^2(H))$. As a consequence, a linear regression of $\log(F(n_i))$ on $\log(n_i)$ provides an estimation of H. More precisely,

Proposition 3.3.3. *Under the previous assumptions and notations, let $n \in \{1, \ldots, N\}$, $m \in \mathbb{N}^* \setminus \{1\}$, $r_i \in \{1, \ldots, [N/n]\}$ for each i with $r_1 < \cdots < r_m$ and $n_i = r_i n$ be such that $N/n \to \infty$ and $N/n^3 \to 0$ when $N \to \infty$. Let $\widehat{H}$ be the estimator of H from the linear regression of $\log(F(r_i \cdot n))$ on $\log(r_i \cdot n)$, i.e.*

$$\widehat{H} = \frac{\sum_{i=1}^m (\log(F(r_i \cdot n)) - \overline{\log(F)})(\log(r_i \cdot n) - \overline{\log(n)})}{\sum_{i=1}^m (\log(r_i \cdot n) - \overline{\log(n)})^2}.$$

Then $\widehat{H}$ is a consistant estimator of H such that

$$\mathbb{E}[(\widehat{H} - H)^2] \leq C(H, m, r_1, \ldots, r_m)\frac{1}{[N/n]} \tag{3.10}$$

with $C(H, m) > 0$.

Remark 3.3.4. *More precisely, it could be possible to show a central limit theorem for $\widehat{H}$, with a convergence rate of $\sqrt{[N/n]}$. Unfortunately, the proof of such a result requires the asymptotic development of $Cov(\tilde{S}_i(n_k), \tilde{S}_j(n_\ell))$, which is more than complicated, for obtaining a multidimensional central limit theorem for $(\log(\tilde{S}(n_1)), \ldots, \log(\tilde{S}(n_m))$.*

3.4 Extension of the results for a general class of a long-range dependent process

Let $\{Y(k), k \in \mathbb{N}\}$ be a stationary zero mean long-range dependent process with a Hurst parameter $H \in]\frac{1}{2}, 1[$. More precisely, let $r_Y(k)$ be the autocorrelation function of this process and let us assume that there exists a slowly varying function $L(k)$ such that :

$$r_Y(k) \sim k^{2H-2}L(k) \text{ , as } k \to \infty. \tag{3.11}$$

Under different additional assumptions on Y, Davydov (1970), Taqqu (1975), Dobrushin and Major (1979), Giraitis and Surgailis (1985) and other authors have studied the

asymptotic behavior of the Donsker line and obtained the following convergence,

$$\big(L(n)^{-\frac{1}{2}} n^{-H} \sum_{i=1}^{[nt]} Y(i)\big)_{t>0} \xrightarrow[N\to\infty]{\mathcal{D}} \big(\sigma \cdot B_H(t)\big)_{t>0}, \tag{3.12}$$

with $\sigma > 0$ and B_H a fractional Brownian motion. Remind that $Z = \{Z(k),\, k \in \mathbb{N}\}$ is a linear process when

$$Z(k) = \sum_{i=-\infty}^{\infty} a_i\, \xi_{k-i} \quad \text{for } k \in \mathbb{N},$$

with (a_k) a sequence of real numbers and (ξ_n) a sequence of zero mean i.i.d.r.v. Then,

Theorem 3.4.1. (Davydov, Taqqu) *Let $Y = \{Y(k),\, k \in \mathbb{N}\}$ be a stationary zero mean long-range dependent process satisfying assumption (3.11). Then, if :*

- *Y is a linear process,*
- *or Y is a function of a Gaussian process with Hermite rank $r = 1$,*

then (3.12) holds, and the convergence takes place in the Skorohod space.

A limit theorem is also obtained by Dobrushin and Major (1979), Giraitis and Surgailis (1989) and Ho and Hsing (1997) for sums of polynomials of linear (or moving average) process with slowly decreasing coefficients a_i. It is obtained under the hypothesis that (ξ_n) are i.i.d standard normal random variable and that the Polynomial Hermit rank satisfies $2 \cdot r < (1-H)^{-1}$.

So, in this case of general class of LRD process, the aggregated process $(X(k))$ has nearly the same behavior as a fractional Brownian motion and the previous asymptotic results of the DFA method can be applied. But Property 3.3.2 and Proposition 3.3.3 cannot be proved under so general assumptions. Indeed, the proofs of such results use a very precise expression of the covariance and a stronger version of assumption (3.11) is necessary. Hence, the covariance r_Y of the stationary process Y is now supposed to satisfy $r_Y \in \mathcal{H}(H, \beta, C)$ with

$$\mathcal{H}(H,\beta,C) = \Big\{r,\ r(k) = C \cdot k^{2H-2}\big(1 + O(1/k^{\beta})\big) \text{ when } k \to \infty\Big\}, \tag{3.13}$$

with $1/2 < H < 1$, $C > 0$ and $\beta > 0$. In such a semi-parametric frame, the previous proofs are still valid and :

Theorem 3.4.2. *Let $Y = \{Y(k),\, k \in \mathbb{N}\}$ be a Gaussian stationary zero mean long-range dependent process with covariance $r_Y \in \mathcal{H}(H, \beta, C)$. Then, Property 3.3.1 holds with the addition of $O(1/n^{\beta})$ in each expansion. Moreover, if $N = o\big(n^{\max(2\beta+1,3)}\big)$, Property 3.3.2 and Proposition 3.3.3 hold.*

As a consequence of this theorem, if $0 < \beta \leq 1$, the DFA method provides a semi-parametric estimator of H with the well-known minimax rate of convergence for the Hurst parameter in this semi-parametric setting (see for instance Giraitis *et al.*, 1997), *i.e.*

$$\limsup_{N\to\infty} \sup_{r_Y \in \mathcal{H}(H,\beta,C)} N^{2\beta/(1+2\beta)} \mathbb{E}[(\widehat{H} - H)^2] < +\infty.$$

However, if $\beta \geq 1$, this result is replaced with $\limsup_{N\to\infty} \sup_{r_Y \in \mathcal{H}(H,\beta,C)} N^{2/3} \mathbb{E}[(\widehat{H} - H)^2] < +\infty$ (it is so in the case of FGN or Gaussian FARIMA (p,d,q)). Thus, the DFA estimator is not rate optimal for all $\beta > 0$ like local Whittle, local log-periodogram or wavelet based estimators are (see respectively (81), (45) and (73)).

3.5 Cases of particular trended long-range dependent processes

In this Section, two general examples of trended long-range dependent processes are considered and it is proved that the DFA method in such cases yields a biased and unusable estimation of the Hurst parameter. In order to consider trended processes, the following lemma for two independent processes could be considered :

Lemma 3.5.1. *Let $Y = \{Y(k),\, k \in \mathbb{N}\}$ and $Y' = \{Y'(k),\, k \in \mathbb{N}\}$ be two independent processes, with $\mathbb{E}(Y(k)) = 0$ for all $k \in \mathbb{N}$, and let us denote respectively F^2_Y, $F^2_{Y'}$ and $F^2_{Y+Y'}$ the DFA functions associated to Y, Y' and $Y + Y'$. Then, for $n \in \{1, \cdots, N\}$,*

$$\mathbb{E}(F^2_{Y+Y'}(n)) = \mathbb{E}(F^2_Y(n)) + \mathbb{E}(F^2_{Y'}(n)).$$

Proof : With X and X' the aggregated processes associated to Y and Y', it is obvious that

$$\begin{aligned}
&\mathbb{E}(F^2_{Y+Y'}(n)) \\
&= \frac{1}{n \cdot [N/n]} \sum_{k=1}^{n\cdot[N/n]} \mathbb{E}\Big(\Big(X(k) + X'(k) - \widehat{X}_n(k) - \widehat{X}'_n(k)\Big)^2\Big) \\
&= \mathbb{E}(F^2_Y(n)) + \mathbb{E}(F^2_{Y'}(n)) + \frac{2}{n \cdot [N/n]} \cdot \sum_{k=1}^{n\cdot[N/n]} \mathbb{E}\Big(\big(X(k) - \widehat{X}_n(k)\big)\big(X'(k) - \widehat{X}'_n(k)\big)\Big).
\end{aligned}$$

From the independence of X and X' and thanks to the assumption $\mathbb{E}(Y(k)) = 0$ for all $k \in \mathbb{N}$ which implies $\mathbb{E}(X(k)) = 0$ and $\mathbb{E}(\widehat{X}(k)) = 0$ for all $k \in \mathbb{N}$, we deduce that $\mathbb{E}\Big(\big(X(k) - \widehat{X}_n(k)\big)\big(X'(k) - \widehat{X}'_n(k)\big)\Big) = 0$. □

Let $Y = \{Y(k), k \in \mathbb{N}\}$ be a Gaussian stationary zero mean long-range dependent process satisfying assumption (3.13) (for instance, Y is a FGN) and let $f : \mathbb{R} \mapsto \mathbb{R}$ be a deterministic function. From Lemma 3.5.1, it is obvious that $n \in \{1, \cdots, N\}$,

$$\mathbb{E}(F^2_{Y+f}(n)) = \mathbb{E}(F^2_Y(n)) + \mathbb{E}(F^2_f(n)). \tag{3.14}$$

Moreover, let us denote respectively $F^2_{Y,j}$ and $F^2_{f,j}$ the DFA function of Y and f relating to window $j \in \{1, \ldots, [\frac{N}{n}]\}$. Then, with few changes in the proof of Lemma 3.5.1,

$$\mathbb{E}(F^2_{Y+f,j}(n)) = \mathbb{E}(F^2_{Y,j}(n)) + \mathbb{E}(F^2_{f,j}(n)). \tag{3.15}$$

3.5.1 Case of power law and polynomial trends

First, let us assume that there exists $\lambda > 0$ and $a \in \mathbb{R}$ such that

$$f(t) = a(t^{\lambda+1} - (t-1)^{\lambda+1}), \quad \text{for} \quad t \geq 1.$$

Then, the associated integrated function is $g(k) = \sum_{i=1}^{k} f(i) = ak^{\lambda+1}$. For this kind of trend,

Property 3.5.1. *For $f(t) = a(t^{\lambda+1} - (t-1)^{\lambda+1})$, with $\gamma(a, N, \lambda)$ a real number depending only on a, N and λ, $\log F_f(n) \simeq 2 \log n + \gamma(a, N, \lambda)$ for $n \to \infty$.*

Thus, it appears that a linear regression of $\log F_f(n_i)$ and $\log(n_i)$ for different values of n_i will provide a slope 2 for any $\lambda > 0$.

Proof : In the j-th window, with $j \in \{1, \ldots, [N/n]\}$, let us consider E_j the vector subspace defined above and define the vector $G^{(j)} = a((1 + n(j-1))^{\lambda+1}, \ldots, (nj)^{\lambda+1})'$. We have

$$F^2_{f,j}(n) = \frac{1}{n}\left(G^{(j)\prime} \cdot G^{(j)} - G^{(j)\prime} \cdot P_{E_j} \cdot G^{(j)}\right)$$

An explicit asymptotic expansion (in n and N) of this partial DFA function can be obtained by approximating sums by integrals. Then,

$$F^2_{f,j}(n) = a^2 n^{2\lambda+2}\Big(1 + O(\frac{1}{n})\Big)\Big(\int_0^1 \int_0^1 (x + j - 1)^{2\lambda+2} - (4 - 6(x+y) + 12xy) \cdot \\ \cdot (x + j - 1)^{\lambda+1}(y + j - 1)^{\lambda+1} dxdy\Big)$$

Moreover, using Taylor expansion in j up to order 3, one obtains

$$F^2_{f,j}(n) = \alpha(a, \lambda) \cdot n^{2\lambda+2} j^{2\lambda-2}\Big(1 + O(\frac{1}{n}) + O(\frac{1}{j})\Big), \tag{3.16}$$

and it implies that the DFA function relating to f can be written as

$$\begin{aligned} F_f^2(n) &= \frac{1}{[N/n]} \sum_{j=1}^{[N/n]} F_{f,j}^2(n) \\ &= \beta(a,\lambda) \cdot n^4 N^{2\lambda-2}\Big(1 + O\big(\frac{1}{n}\big) + O\big(\frac{1}{[\frac{N}{n}]}\big)\Big), \end{aligned}$$

with $\alpha(a,\lambda)$, $\beta(a,\lambda)$ two positive numbers depending only on a and λ. □

For illustrating this result (see Figure 3.3), several simulations have been made for various values of $\lambda > 0$, a and $(n_1, \ldots, n_m)$. The presented results exhibit the relation between $\log F_f(n_i)$ and $\log(n_i)$, that is nearly linear with a slope of the adjustment linear line estimated at 2 like it was theoretically proved.

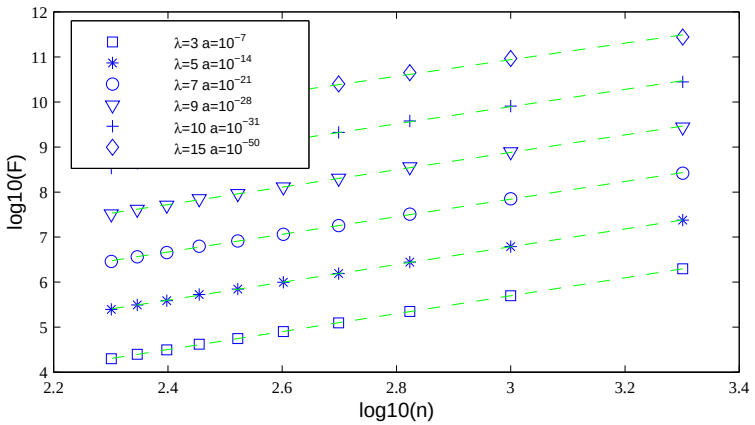


FIG. 3.3 – Relation between $\log F_f(n_i)$ and $\log n_i$ in the case of power law trend

This result can also be used to deduce similar results for polynomial trends.

Property 3.5.2. *Let us assume that there exists $p \in \mathbb{N}^*$ and a family $(a_j)_{0\le j\le p}$ with $a_p \neq 0$ such that for $k \in \mathbb{N}$, $f(k) = a_p k^p + \cdots + a_0$. Then,*

$$\Longrightarrow \quad \log F_{a_p k^p + \cdots + a_0}(n) \simeq 2\log n + \gamma(a_p, N, p) \quad \textit{for } n \to \infty.$$

Proof : Indeed, ,

$$f(k) = a_p k^p + \cdots + a_0 \quad \Longrightarrow \quad g(k) = \sum_{i=1}^{k} f(i) = b_{p+1} k^{p+1} + \cdots + b_0,$$

with $b_{p+1} \neq 0$, *i.e.* the associated integrated function is also a polynomial function. From the expression of the partial DFA function and with the asymptotic expansion

(3.16) depending on the degree λ, for large enough n and N,

$$F^2_{a_pk^p+\cdots+a_0,j}(n) = F^2_{a_pk^p,j}(n)\Big(1+O(\frac{1}{n})+O(\frac{1}{j})\Big)$$

(the power of n in the partial DFA function relating to a_pk^p is greater than the ones in the partial DFA function relating to the other monomes). This approximation leads to the following expression of the DFA function of a polynomial function,

$$F^2_{a_pk^p+\cdots+a_0}(n) = \beta(b_{p+1})\cdot n^4N^{2\lambda-2}\Big(1+O(\frac{1}{n})+O(\frac{1}{[\frac{N}{n}]})\Big).\square$$

Using relations (3.14) and (3.15), the previous results for trends can be used for deducing the behavior of the DFA function of trended long range dependent processes. Hence, in both the previous cases of trends, there exists $C>0$ such that

$$\begin{aligned}\mathbb{E}(F^2_{Y+f}(n)) &= C\cdot n^4N^{2\lambda-2}\Big(1+O(\frac{1}{n})+O(\frac{1}{[\frac{N}{n}]})\Big)+\sigma'^2f(H)\cdot n^{2H}\Big(1+O(\frac{1}{n^{\min(1,\beta)}})\Big)\\ &\simeq C\cdot n^4N^{2\lambda-2}.\end{aligned}$$

Hence, it is clear that the trend is dominant for a large n and the graph tracing the relation between $\log F_{Y+f}(n_i)$ and $\log n_i$ for different power law trends and different coefficients H confirms this (the estimated slope is always close to 2).

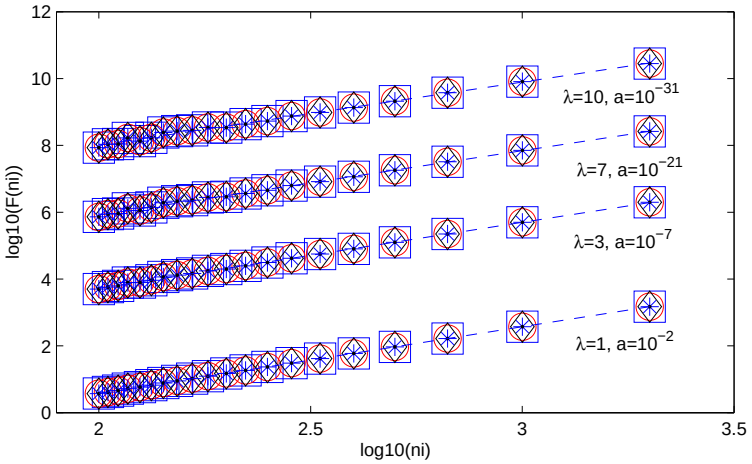


FIG. 3.4 – Relation between $\log F_{Y+f}(n_i)$ and $\log n_i$ in the case of a power law trend ($N=10000$, $H=0.2$ ($\square$), $H=0.4$ ($\bigcirc$), $H=0.5$ ($\Diamond$), $H=0.7$ ($*$) and $H=0.8$ ($\cdot$))

3.5.2 Case of a piecewise constant trend

Let us assume now that f is a step function of the form $f(t) = \sum_{i=0}^{m-1} a_i \mathbb{1}_{]t_i,t_{i+1}]}$ with $t_0 = 0$, $t_m = N$ and $m \in \mathbb{N}^*$. The associated integrated series is

$$g(k) = \sum_{i=0}^{m-1} \big(\sum_{s=0}^{i} (a_{s-1} - a_s)t_s + a_i k \big) \mathbb{1}_{]t_i,t_{i+1}]} \text{ with } a_{-1} = 0$$

For $j \in \{1, \dots, [N/n]\}$, the partial DFA function $F^2_{fj}(n)$ is null unless there exist i_p with $p \in \{1, \dots, r\}$ and $(r, i_r) \in \{1, \dots, m-1\}^2$ such that $t_{i_p} \in [(j_p - 1)n + \tau n, j_p n - \tau n]$ with $\tau \in]0, \frac{1}{2}[$. In such a case, we calculate the partial DFA function :

$$\begin{aligned} F^2_{f,j_p}(n) &= \frac{1}{n} \sum_{k=1}^{n} (g(k + (j_p - 1)n) - \widehat{g}_n(k + (j_p - 1)n))^2 \\ &= \frac{1}{n} \Big(G^{(j_p)'} \cdot P_{E^{\perp}_{j_p}} \cdot G^{(j_p)} \Big) \end{aligned}$$

If we consider the first window, the partial DFA function can be estimated from below :

$$F^2_{f,1}(n) \geq \frac{1}{n} \Big(\sum_{k=1}^{\tau n} (g(k) - \widehat{g}_n(k))^2 + \sum_{k=n-\tau n}^{n} (g(k) - \widehat{g}_n(k))^2 \Big)$$

where the $n \times 1$ vector $\big(g(k) - \widehat{g}_n(k)\big)_{1 \leq k \leq n} = P_{E_1^{\perp}} \cdot G^{(1)}$ with :

$$G^{(1)} = \big(a_0 \cdot 1, \dots, a_0 \cdot t_1, (a_0 - a_1)t_1 + a_1 \cdot (t_1 + 1), \dots, (a_0 - a_1)t_1 + a_1 \cdot n\big)'$$

Then,

$$\sum_{k=1}^{\tau n} (g(k) - \widehat{g}_n(k))^2 = \Big(J_{\tau n} \cdot P_{E_1^{\perp}} \cdot G^{(1)} \Big)' \cdot \Big(J_{\tau n} \cdot P_{E_1^{\perp}} \cdot G^{(1)} \Big)$$

where $J_{\tau n}$ is a square matrix of order n with ones in the τn first terms of the diagonal and zeros elsewhere. When we approximate sums by integrals, this expression can be written as follows :

$$\sum_{k=1}^{\tau n} (g(k) - \widehat{g}_n(k))^2 = n^3 \Big(1 + O\big(\frac{1}{n}\big) \Big) \cdot$$

$$\Big(\int_0^{\tau} \Big(\int_0^1 a_0 y - \big(a_0 x \cdot \mathbb{1}_{x \leq \frac{t_1}{n}} + \big(a_1 x + (a_0 - a_1)\frac{t_1}{n} \big) \mathbb{1}_{x > \frac{t_1}{n}} \big)(4 - 6(x + y) + 12xy) dx \Big)^2 dy \Big)$$

For $\tau \in]0, \frac{1}{2}[$, the second term can be developed in the same way by replacing $J_{\tau n}$ by $J_{n-\tau n}$ which is $Id - J_{\tau n}$. Then, this term can be approximated by :

$$\begin{aligned} \sum_{k=n-\tau n}^{n} (g(k) - \widehat{g}_n(k))^2 = n^3 \Big(1 + O\big(\frac{1}{n}\big) \Big) \Big(\int_{1-\tau}^1 \Big(\int_0^1 (a_0 - a_1)\frac{t_1}{n} + a_1 y \\ - \big(a_0 x \cdot \mathbb{1}_{x \leq \frac{t_1}{n}} + \big(a_1 x + (a_0 - a_1)\frac{t_1}{n} \big) \mathbb{1}_{x > \frac{t_1}{n}} \big)(4 - 6(x + y) + 12xy) dx \Big)^2 dy \Big) \end{aligned}$$

Then after developing of the two terms, we deduce that there exists a positive number $c(a_0, \ldots, a_{i_p}, t_{i_p}, \tau)$ such that the partial DFA function in the j_p-th window where $t_{i_p} \in [(j_p - 1)n + \tau n, j_p n - \tau n]$, for $p \in \{1, \ldots, r\}$ and n large ebough, can be written as :

$$F^2_{f,j_p}(n) \geq c(a_0, \ldots, a_{i_p}, t_{i_p}, \tau) n^2. \tag{3.17}$$

Then if we suppose that there exists only one change point or a definite number of windows $j_1, \ldots, j_r$, there exists $c'(a_0, \ldots, a_{i_r}, t_{i_1}, \ldots, t_{i_r}, \tau) > 0$ such that the DFA function relating to f is :

$$\begin{aligned} F^2_f(n) &= \frac{1}{[\frac{N}{n}]} \sum_{j=j_1}^{j_r} F^2_{f,j}(n) \\ &\geq c'(a_0, \ldots, a_{i_r}, t_{i_1}, \ldots, t_{i_r}, \tau) n^3 N^{-1} \Big(1 + O(\frac{1}{n})\Big) \end{aligned}$$

Then, for different values $(n_1, \ldots, n_m)$, the graph tracing the relation between $\log F_f(n_i)$ and $\log(n_i)$, shows a slope estimated at $\frac{3}{2}$ (see Figure 3.5).

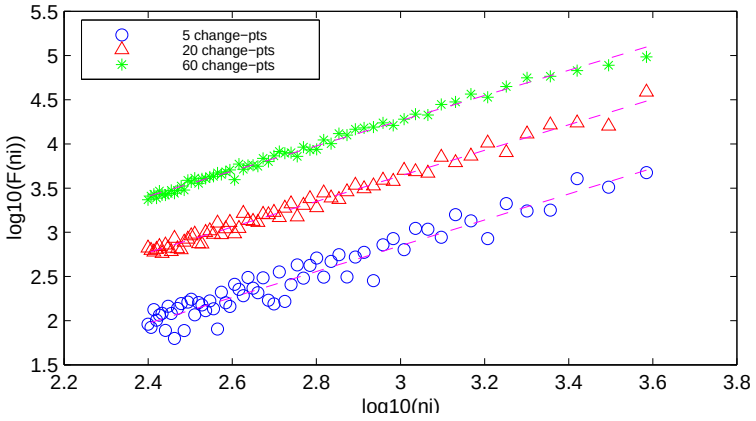


FIG. 3.5 – Relation between $\log F_f(n_i)$ and $\log n_i$ in the case of trend with change points

If we consider the signal formed by the superposition between the trend and a long range dependent process, we point out that $\mathbb{E}(F^2_Y(n)) = \sigma^{\cdot 2} f(H) \cdot n^{2H} \Big(1 + O\big(\frac{1}{n^{\min(1,\beta)}}\big)\Big)$, we can deduce, from the previous conditions on n and N ($N/n \to \infty$ and $N = o(n^{\min 3, 2\beta+1})$, that the trend is dominant for large n.

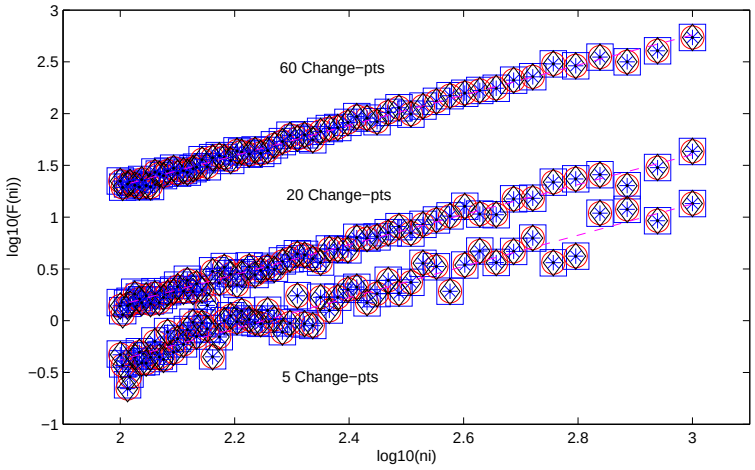


FIG. 3.6 – Relation between $\log F_{f+Y}(n_i)$ and $\log n_i$ in the case of a trend with change points ($N = 20000$, $H = 0.2$ ($\square$), $H = 0.4$ ($\bigcirc$), $H = 0.5$ ($\Diamond$), $H = 0.7$ ($*$) and $H = 0.8$ ($\cdot$))

3.6 Conclusion

In the semi-parametric frame of a long-memory stationary process, we have shown, using the DFA method, that the estimator of the long range dependence parameter is convergent with a reasonable convergence rate. However, in numerous cases of trended long-range dependent process (with perhaps the only exception of a constant trend), this estimator does not converge. Indeed the Hurst exponent is dominated by the trend in the case of power law and polynomial trends where the slope for the trended processes of a DFA log-log regression is always close to 2 and in the case of a piecewise constant trend where the slope is estimated at $\frac{3}{2}$. The DFA method is therefore not robust at all and should not be applied for trended processes.

The wavelet-based method provides a more efficient and robust estimator of the Hurst parameter especially when a polynomial trended LRD (or self-similar) process is considered. Indeed, Abry *et al.* (1998) remarked that all polynomial trend of degree M is without effects on the estimator of the Hurst parameter as soon as the mother wavelet has its M first vanishing moments. Therefore, the larger M, the more robust estimator is. Moreover, in the semi-parametric frame of general class of stationary Gaussian LRD processes, it was established by Moulines *et al.* (2007) that the estimator of the Hurst parameter converges with an optimal convergence rate (following the minimax criteria) when an optimal length of windows is known. Bardet *et al.* (2007) proposed an adaptive estimator and obtained an optimal convergence rate up to logarithmic factor.

Finally, the wavelet based estimator can be computed by Mallat's fast cascade algorithm which is a very fast algorithm (the equivalent for wavelet transform of FFT for Fourier transform) for computing wavelet coefficients. Thus, computing time of wavelet based estimator is smaller than DFA estimator one.

3.7 Proofs

Proof of Property 3.3.1 : 1. From the proof of Lemma 3.2.2 and with its notations, one obtains

$$\begin{aligned} F_1^2(n) &= \frac{1}{n}\,(X^{(1)} - P_{E_1}\cdot X^{(1)})'\cdot(X^{(1)} - P_{E_1}\cdot X^{(1)}) \\ &= \frac{1}{n}\Big(X^{(1)'}\cdot X^{(1)} - X^{(1)'}\cdot P_{E_1}\cdot X^{(1)}\Big). \end{aligned}$$

As a consequence,

$$\mathbb{E}(F_1^2(n)) = \frac{1}{n}\Big(\mathrm{trace}(\Sigma_n) - \mathrm{trace}(P_{E_1}\cdot\Sigma_n)\Big),$$

where Σ_n is the covariance matrix of $X^{(1)}$ and is such that

$$\Sigma_n = \mathrm{Cov}(X_i, X_j)_{1\le i,j\le n} = \frac{\sigma^2}{2}\big(|i|^{2H} + |j|^{2H} - |i-j|^{2H}\big)_{1\le i,j\le n}$$

But, $\mathrm{trace}(\Sigma_n) = \sigma^2\sum_{i=1}^{n}|i|^{2H} = \sigma^2 n^{2H+1}\big(\frac{1}{n}\sum_{i=1}^{n}\big|\frac{i}{n}\big|^{2H}\big) = \sigma^2 n^{2H+1}\big(\int_0^1 x^{2H}dx + O(\frac{1}{n})\big)$. Therefore, on the one hand,

$$\mathrm{trace}(\Sigma_n) = \frac{\sigma^2}{2H+1}\,n^{2H+1}\cdot\Big(1 + O\big(\frac{1}{n}\big)\Big). \tag{3.18}$$

and on the other hand, it is well known that P_{E_1} is a $(n\times n)$ square matrix such that

$$P_{E_1} = \frac{2}{n(n-1)}\Big((2n+1) - 3(i+j) + 6\,\frac{i\cdot j}{1+n}\Big)_{1\le i,j\le n}$$

Then, after some straightforward computations, we obtain the formula

$$\begin{aligned} \mathrm{trace}(P_{E_1}\cdot\Sigma_n) = \frac{\sigma^2 n^{2H+1} n^2}{n(n-1)}\cdot \\ \sum_{p=1}^{n}\sum_{q=1}^{n}\Big[\frac{1}{n^2}\Big((2+\frac{1}{n}) - 3\cdot\frac{p+q}{n} + \frac{6p\cdot q}{n(1+n)}\Big)\Big(\big|\frac{q}{n}\big|^{2H} + \big|\frac{p}{n}\big|^{2H} - \big|\frac{q-p}{n}\big|^{2H}\Big)\Big] \end{aligned}$$

In order to clarify the formula, we approximate these sums by integrals

$$\begin{aligned}&\mathrm{trace}(P_{E_1}\cdot\Sigma_n)\\&=\sigma^2 n^{2H+1}\Big(1+O(\frac{1}{n})\Big)\cdot\int_0^1\int_0^1\Big[\big(2-3(x+y)+6xy\big)\big(x^{2H}+y^{2H}-|x-y|^{2H}\big)\Big]dxdy\end{aligned}$$

After the calculation of this integral and a simplification with formula (3.18), we get the result

$$\mathrm{trace}(\Sigma_n)-\mathrm{trace}(P_{E_1}\cdot\Sigma_n)=\sigma^2 f(H)\cdot n^{2H+1}\cdot\Big(1+O(\frac{1}{n})\Big)$$

and therefore the formula of $\mathbb{E}(F_1^2(n))$.

2. From the previous notations and the property of the trace of a product of matrices,

$$\begin{aligned}\mathrm{Var}(F_1^2(n)) &= \frac{1}{n^2}\Big[\mathbb{E}(X^{(1)'}\cdot P_{E_1^\perp}\cdot X^{(1)}\cdot X^{(1)'}\cdot P_{E_1^\perp}\cdot X^{(1)})-\Big(\mathbb{E}(X^{(1)'}\cdot P_{E_1^\perp}\cdot X^{(1)})\Big)^2\Big]\\&= \frac{1}{n^2}\Big[\mathrm{trace}(\Sigma_n\cdot\Sigma_n)-\mathrm{trace}(P_{E_1}\cdot\Sigma_n\cdot\Sigma_n)\Big]\end{aligned}\tag{3.19}$$

The development of the first term provides the following asymptotic expansion

$$\begin{aligned}\mathrm{trace}(\Sigma_n\cdot\Sigma_n) &= \frac{\sigma^4}{4}\sum_{i=1}^n\sum_{p=1}^n\big(|i|^{2H}+|p|^{2H}-|i-p|^{2H}\big)^2\\&= \frac{\sigma^4}{4}n^{4H+2}\Big(1+O(\frac{1}{n})\Big)\int_0^1\int_0^1\big(|x|^{2H}+|y|^{2H}-|x-y|^{2H}\big)^2dxdy\end{aligned}$$

The calculation of these integrals provides the following simplified expression

$$\begin{aligned}\mathrm{trace}(\Sigma_n\cdot\Sigma_n)=\frac{\sigma^4}{4}n^{4H+2}\Big(1+O(\frac{1}{n})\Big)\cdot\\\Big[\frac{1}{4H+1}+\frac{1}{(4H+1)(4H+2)}-2\frac{(\Gamma(2H+1))^2}{\Gamma(4H+3)}\Big]\end{aligned}\tag{3.20}$$

The same development can be made for the second term

$$\begin{aligned}&\mathrm{trace}(P_{E_1}\cdot\Sigma_n\cdot\Sigma_n)=\frac{\sigma^4}{2}\,n^{4H+2}\Big(1+O(\frac{1}{n})\Big)\cdot\\&\int_0^1\int_0^1\int_0^1\Big(|x|^{2H}+|y|^{2H}-|y-x|^{2H}\Big)\Big(|x|^{2H}+|z|^{2H}-|x-z|^{2H}\Big)\Big(2-3(y+z)+6yz\Big)dxdydz\end{aligned}$$

After the computation of this last integral, and using relations (3.19) and (3.20)

$$\Big[\mathrm{trace}(\Sigma_n\cdot\Sigma_n)-\mathrm{trace}(P_{E_1}\cdot\Sigma_n\cdot\Sigma_n)\Big]=\sigma^4\cdot g(H)\,n^{4H+2}\Big(1+O(\frac{1}{n})\Big)$$

$$\begin{aligned}\text{with,}g(H)=\frac{1}{2}\Big(&-\frac{(16H^2+24H+17)(\Gamma(2H+1))^2}{(4H+5)\Gamma(4H+4)}+\frac{3(4H+3)}{2(2H+1)^2(H+1)^2(4H+5)}+\\&\frac{7H+3}{2(2H+1)^2(H+1)}-\frac{3}{2(H+1)^2}+\frac{H+1}{(2H+1)(4H+1)}-\frac{4}{(2H+1)^2(4H+3)}\Big).\end{aligned}\tag{3.21}$$

Then, using the relations (3.19), one obtains $\mathrm{Var}\big(F_1^2(n)\big) = \sigma^4 \cdot g(H) \cdot n^{4H}\Big(1 + O(\frac{1}{n})\Big)$.

3. An asymptotic expansion of the covariance between two DFA functions in two sufficiently far windows can be provided. Indeed

$$\begin{aligned}\mathrm{Cov}(F_1^2(n), F_j^2(n)) &= \frac{1}{n^2}\mathrm{Cov}\Big((X^{(1)} - \widehat{X}^{(1)})' \cdot (X^{(1)} - \widehat{X}^{(1)}) \cdot (X^{(j)} - \widehat{X}^{(j)})' \cdot (X^{(j)} - \widehat{X}^{(j)})\Big) \\ &= \frac{1}{n^2}\left(\mathrm{trace}\Big(\Sigma^{(1,j)} \cdot \Sigma^{(1,j)}\Big) - \mathrm{trace}\Big(P_{E_1} \cdot \Sigma^{(1,j)} \cdot \Sigma^{(1,j)}\Big)\right),\end{aligned}$$

because $P_{E_1^\perp} = P_{E_j^\perp}$ and with $\Sigma^{(1,j)}$ the covariance matrix $\mathbb{E}(X^{(1)} \cdot X^{(j),}) = (\sigma_{k,k'}^{(1,j)})_{1 \le k,k' \le n}$. As usual, this formula can be developed

$$\mathrm{Cov}(F_1^2(n), F_j^2(n)) = \frac{1}{n^2}\left(\sum_{k=1}^{n}\sum_{k'=1}^{n} \sigma_{k,k'}^{(1,j)} \cdot \sigma_{k',k}^{(1,j)} - \sum_{i=1}^{n}\sum_{k'=1}^{n}\sum_{k=1}^{n} p_{i,k} \cdot \sigma_{k,k'}^{(1,j)} \cdot \sigma_{k',i}^{(1,j)}\right),$$

with

$$\sigma_{k,k'}^{(1,j)} = \frac{\sigma^2}{2}\big(|k + nj|^{2H} + |k'|^{2H} - |k - k' + nj|^{2H}\big)_{1 \le k,k' \le n}$$

and with $P_{E_1} = (p_{i,j})_{1 \le i,j \le n}$ such that

$$p_{i,j} = \frac{2}{n(n-1)}\Big((2n+1) - 3(i+j) + 6\,\frac{i \cdot j}{1+n}\Big).$$

Now, one considers the asymptotic expansion of this formula when n is large enough

$$\begin{aligned}\mathrm{Cov}(F_1^2(n), F_j^2(n)) = \frac{\sigma^4}{4}\, n^{4H}\Big(1 + O(\frac{1}{n})\Big) \cdot \Big(&\int_0^1\!\!\int_0^1 \big(|x+j|^{2H} + y^{2H} - |x-y+j|^{2H}\big) \cdot \\ &\big(|y+j|^{2H} + x^{2H} - |y-x+j|^{2H}\big)dxdy - \int_0^1\!\!\int_0^1\!\!\int_0^1 (4 - 6(x+z) + 12xz) \cdot \\ &\cdot \big(|x+j|^{2H} + y^{2H} - |x-y+j|^{2H}\big)\big(|y+j|^{2H} + z^{2H} - |y-z+j|^{2H}\big)dxdydz\Big).\end{aligned}$$

In order to obtain an asymptotic expansion of this formula when j is large enough (*i.e.* both windows are taken away from one another), a Taylor expansion in j up to order 3 is necessary : $(x+j)^{2H} = j^{2H} + 2Hj^{2H-1}x + H(2H-1)j^{2H-2}x^2 + \frac{2H(2H-1)(H-1)}{3}j^{2H-3}(x^3 + \varepsilon(x^3))$ with $\lim_{n\to\infty} \varepsilon(x) = 0$. After calculating the integrals and simplifying their expressions, we get the result. □

Proof of Property 3.3.2 : We divide the proof into 3 steps :

- Step 1 : one proves that $[N/n] \cdot \mathrm{Var}(\tilde{S}(n)) \to \gamma^2(H)$, where $\gamma^2(H)$ depends only

on H, when $[N/n] \to \infty$. Indeed,

$$\begin{aligned}\mathrm{Var}(\tilde{S}(n)) &= \frac{1}{[\frac{N}{n}]^2}\sum_{j=1}^{[\frac{N}{n}]}\sum_{j'=1}^{[\frac{N}{n}]}\mathrm{Cov}(\tilde{S}_j(n),\tilde{S}_{j'}(n)) \\ &= \frac{1}{[\frac{N}{n}]}\mathrm{Var}(\tilde{S}_j(n)) + \frac{2}{[\frac{N}{n}]^2}\sum_{j=1}^{[\frac{N}{n}]}\big([\frac{N}{n}]-j\big)\mathrm{Cov}(\tilde{S}_1(n),\tilde{S}_j(n))\end{aligned}$$

due to stationarity.

However, with properties (3.9), one deduces that when $[N/n] \to \infty$, $\sum_{j=1}^{[\frac{N}{n}]}\mathrm{Cov}(\tilde{S}_1(n),\tilde{S}_j(n))$ and $\sum_{j=1}^{[\frac{N}{n}]} j\cdot\mathrm{Cov}(\tilde{S}_1(n),\tilde{S}_j(n))$ converge, because there exists $C \geq 0$ such that $\big|\mathrm{Cov}(\tilde{S}_1(n),\tilde{S}_j(n))\big| \leq C\cdot j^{2H-3}$ and $0 < H < 1$.

Therefore, there exists $\gamma^2(H)$ depending only on H such that

$$\lim_{[N/n]\to\infty}[N/n]\cdot\mathrm{Var}(\tilde{S}(n)) = \gamma^2(H). \tag{3.22}$$

• Step 2 : the proof of a central limit theorem for $\tilde{S}(n)$ when $[N/n] \to \infty$ can be obtained from the same method as in the proof of Proposition 2.1 in Bardet (2000) (Theorem 3 in Soulier (2000) leads to the same result).

Indeed, $\tilde{S}(n) = \dfrac{1}{n^{2H+1}\sigma^2 f(H)\cdot[N/n]}\displaystyle\sum_{i=1}^{n\cdot[N/n]} Z_i^2$, where the zero-mean Gaussian vector $Z = (Z_1,\ldots,Z_{n\cdot[N/n]})$ has the covariance matrix $P\cdot\Sigma\cdot P$, where P is a diagonal block matrix with each block consisting of (n,n) matrix $P_{E_1^\perp}$ and Σ is the covariance matrix of an FBM time series (each (n,n) block is $\Sigma^{(i,j)}$ with the previous notations). Using a Lindeberg condition, $\tilde{S}(n)$ satisfies the following central limit theorem

$$\sqrt{[N/n]}\cdot\Big(\tilde{S}(n) - \mathbb{E}(S(n))\Big) \xrightarrow[[N/n]\to\infty]{\mathcal{L}} \mathcal{N}(0,\gamma^2(H))), \tag{3.23}$$

if $\lambda = \|P\cdot\Sigma\cdot P\|$, the supremum of the eigenvalues of the symmetrical matrix $P\cdot\Sigma\cdot P$, is such that

$$\lambda = o\Big(\frac{1}{\sqrt{[N/n]}}\Big). \tag{3.24}$$

Lemma 3.7.1. (Bardet (2000)) *If $M = (m_{ij})_{1\leq i,j\leq n}$ is a symmetric matrix of real numbers, then*

$$\sup_{\lambda\in Sp(M)} \mid\lambda\mid \leq \sup_{1\leq i\leq n}\sum_{j=1}^{n}\mid m_{ij}\mid .$$

Using above Lemma and following the proof of Proposition 2.1 in Bardet (2000),

$$\begin{aligned}
\lambda &\leq \frac{1}{n^{2H+1}\sigma^2 f(H)\cdot[N/n]} \max_{i\in\{1,\dots,n\cdot[N/n]\}} \left(\sum_{j=1}^{n\cdot[N/n]} |\mathrm{Cov}(Z_i,Z_j)|\right)\\
\lambda &\leq \frac{1}{\sqrt{2}[N/n]} \max_{i\in\{1,\dots,n\cdot[N/n]\}} \left(\sum_{j=1}^{n\cdot[N/n]} \frac{\sqrt{|\mathrm{Cov}(Z_i^2,Z_j^2)|}}{n^{2H+1}\sigma^2 f(H)}\right)\\
&\leq \frac{1}{\sqrt{2}[N/n]} \max_{i\in\{1,\dots,[N/n]\}} \left(\sum_{j=1}^{[N/n]} \sqrt{\mathrm{Cov}\big(\tilde S_i(n),\tilde S_j(n)\big)}\right)\\
&\leq \frac{\sqrt{2}}{[N/n]} \left(\sum_{j=1}^{[N/n]} \sqrt{\mathrm{Cov}\big(\tilde S_1(n),\tilde S_j(n)\big)}\right)
\end{aligned}$$

So, there exists $C(H)>0$ depending only on H such that

$$\begin{aligned}
\lambda &\leq C(H)\cdot\frac{1}{[N/n]}\sum_{j=1}^{[N/n]}\left(\sqrt{j^{2H-3}}+\frac{c}{n}\right) \text{ third line of (3.9)}\\
&\leq C'(H)\cdot\Big([N/n]^{H-3/2}+\frac{1}{n}\Big). \qquad (3.25)
\end{aligned}$$

Therefore if $\frac{1}{n}=o\Big(\frac{1}{\sqrt{[N/n]}}\Big)$ (*i.e.* $N=o(n^3)$), (3.24) and (3.23) are proved.

• Step 3 : Now, $\mathbb{E}(\tilde S(n)\Big)=1+O\big(\frac{1}{n}\big)$ for n large enough. Then, if $\sqrt{[N/n]}\cdot\frac{1}{n}\to 0$, that is $N/n^3\to 0$,

$$\sqrt{[N/n]}\cdot\Big(\tilde S(n)-1\Big)\xrightarrow[[N/n]\to\infty]{\mathcal{L}}\mathcal{N}(0,\gamma^2(H))).$$

The classical Delta method allows the passage between a central limit theorem for $\tilde S(n)$ and a central limit theorem for $\log(\tilde S(n))$ (thanks to the regularity properties of the function logarithm). □

Proof of Proposition 3.3.3 : It is possible to write $\widehat H=(1,0)\cdot(Z'\cdot Z)^{-1}\cdot Z'\cdot F$, where Z is the $(m,2)$ matrix such that $Z=\begin{pmatrix}\log(r_1\cdot n) & 1\\ \vdots & \vdots\\ \log(r_m\cdot n) & 1\end{pmatrix}$ and $F=\begin{pmatrix}\log(F(r_1\cdot n))\\ \vdots\\ \log(F(r_m\cdot n))\end{pmatrix}$.

Then

$$\begin{aligned}
\mathrm{Var}(\widehat H) &= (1,0)\cdot(Z'\cdot Z)^{-1}\cdot Z'\cdot\mathrm{Cov}(F,F)\cdot Z\cdot(Z'\cdot Z)^{-1}\cdot(1,0)'\\
&\leq \|(1,0)\cdot(Z'\cdot Z)^{-1}\cdot Z'\|^2\cdot\|\mathrm{Cov}(F,F)\|\\
&\leq \|(1,0)\cdot(Z'\cdot Z)^{-1}\cdot Z'\|^2\cdot 2m\cdot\prod_{i=1}^m r_i\cdot\gamma^2(H)\cdot\frac{1}{[N/n]}.
\end{aligned}$$

Since $\|(1,0)\cdot(Z'\cdot Z)^{-1}\cdot Z'\|$ only depends on $r_1,\dots,r_m$, the proof of Proposition 3.3.3 is completed. □

Proof of Theorem 3.4.2 : From the assumptions on Y and r_Y, if $i\geq j\geq 1$,

$$\begin{aligned}\mathrm{Cov}(X(i),X(j)) &= \sum_{k=1}^{i}\sum_{\ell=1}^{j}\mathrm{Cov}(Y(k),Y(\ell))\\ &= \sum_{k=1}^{i}(i-k)r_Y(k)+\sum_{k=1}^{j}(j-k)r_Y(k)-\sum_{k=1}^{i-j}(i-j-k)r_Y(k).\end{aligned}$$

As a consequence, for all $(i,j)\in\{1,\dots,n\}^2$,

$$\mathrm{Cov}(X(i),X(j)) = C\cdot\Big(\int_0^1(1-u)u^{2H-2}du\Big)\cdot$$
$$\Big(i^{2H}\big(1+O(\tfrac{1}{i^{\min(\beta,1)}})\big)+j^{2H}\big(1+O(\tfrac{1}{j^{\min(\beta,1)}})\big)-|i-j|^{2H}\big(1+O(\tfrac{1}{(1+|i-j|)^{\min(\beta,1)}})\big)\Big)\cdot$$

Now, this covariance can be used for all proofs, replacing the previous ones. This implies

1. $\mathbb{E}(F_1^2(n)) = \sigma'^2 f(H)\cdot n^{2H}\Big(1+O(\frac{1}{n^{\min(\beta,1)}})\Big)$,
2. $\mathrm{Var}\big(F_1^2(n)\big) = \sigma'^4 g(H)\cdot n^{4H}\Big(1+O(\frac{1}{n^{\min(\beta,1)}})\Big)$,
3. $\mathrm{Cov}(F_1^2(n),F_j^2(n)) = \sigma'^4 h(H)\cdot n^{4H}\cdot j^{2H-3}\Big(1+O(\frac{1}{n^{\min(\beta,1)}})+O(\frac{1}{j})\Big)$,

with $\sigma'^2 = 2C\cdot\Big(\int_0^1(1-u)u^{2H-2}du\Big)$. The proofs of property 3.3.2 is the same as in the case of the FGN except that in (3.25),

$$\begin{aligned}\lambda &\leq C(H)\cdot\frac{1}{[N/n]}\sum_{j=1}^{[N/n]}\Big(\sqrt{j^{2H-3}}+\frac{c}{n}+\frac{c}{n^\beta}\Big)\\ &\leq C'(H)\cdot\Big([N/n]^{H-3/2}+\frac{1}{n}+\frac{1}{n^\beta}\Big).\end{aligned}$$

So, if $\frac{1}{n}+\frac{1}{n^\beta}=o\Big(\frac{1}{\sqrt{[N/n]}}\Big)$ therefore $N=o\big(n^{\max(2\beta+1,3)}\big)$, the central limit theorem is proved as well as Proposition 3.3.3 following the same proof as in the case of the FGN. □

Chapitre 4

Comparison of DFA vs wavelet analysis for estimation of regularity of HR series during the marathon

4.1 Introduction

The wavelet analysis method has been introduced by Flandrin (1992) and was developed by Abry *et al.* (2002) and Bardet *et al.* (2000). It provides more robust results than Detrended Fluctuation Analysis method and can be applied to more general models. Then, it permits the construction of the semi-parametric process which could be more relevant than other for modeling HR data. It also shows an evolution of the Hurst parameter during the race, what confirms results obtained by Peng *et al.*. According to the conclusions found, we can deduce that the increase of the LRD parameter values through the race phases, which can not be observed with DFA, may be associated with fatigue appearing during the last phase of the marathon.

4.2 Wavelet based estimator of the Hurst parameter

In (15), the asymptotic properties of the DFA function in case of a FGN path $(Y(1), \ldots, Y(N))$ are studied. In such a case the estimator $\widehat{H}_{DFA}$ converges to H with a non-optimal convergence rate ($N^{1/3}$ instead of $N^{1/2}$ reached for instance by maximum likelihood estimator). An extension of these results for a general class of stationary

Gaussian LRD processes is also established. In this semi-parametric frame, we showed that the estimator $\widehat{H}_{DFA}$ converges to H with an optimal convergence rate (following the minimax criteria) when an optimal length of windows is known.

The processing of experimental data, and in particular physiological data, exhibits a major problem that is the non-stationarity of the signal. Hu, Chen, Ivanov and stanley (2001) have studied different types of non-stationarity associated with examples of trends and deduced their effect on an added noise and the kind of competition who exists between this two signals. They have also explained (2002) the effects of three other types of non-stationarity, which are often encountered in real data. In (15), we proved that $\widehat{H}_{DFA}$ does not converge to H when a polynomial trend (with degree greater or equal to 1) or a piecewise constant trend is added to a LRD process : the DFA method is clearly a non robust estimation of the Hurst parameter in case of trend.

For improving this estimation at least for polynomial trended LRD process, a wavelet based estimator is now considered. This method has been introduced by Flandrin (1992) and was developed by Abry *et al.* (2002) and Bardet *et al.* (2000). In Wesfreid *et al.* (2005), a multifractal analysis of HR time series is presented for trying to unveil their scaling law behavior using the Wavelet Transform Modulus Maxima (WTMM) method.

4.2.1 Wavelet analysis

Let $\psi : \mathbb{R} \to \mathbb{R}$ a function so-called the mother wavelet. Let $(a, b) \in \mathbb{R}_+^* \times \mathbb{R}$ and denote $\lambda = (a, b)$. Then define the family of functions ψ_λ by

$$\psi_\lambda(t) = \frac{1}{\sqrt{a}}\,\psi\left(\frac{t}{a} - b\right)$$

Parameters a and b are so-called the scale and the shift of the wavelet transform. Let us underline that we consider a continuous wavelet transform. Let $d_Z(a, b)$ be the wavelet coefficient of the process $Z = \{Z(t),\, t \in \mathbb{R}\}$ for the scale a and the shift b, with

$$d_Z(a, b) = \frac{1}{\sqrt{a}} \int_{\mathbb{R}} \psi(\frac{t}{a} - b) Z(t) dt = < \psi_\lambda, Z >_{L^2(\mathbb{R})} .$$

For a time series instead of a continuous time process, a Riemann sum can replace the previous integral for providing a discretized wavelet coefficient $e_Z(a, b)$. The function ψ is supposed to be a function such that it exists $M \in \mathbb{N}^*$ satisfying ,

$$\int_{\mathbb{R}} t^m \psi(t) dt = 0 \quad \text{for all } m \in \{0, 1, \ldots, M\}. \tag{4.1}$$

Therefore, ψ has its M first vanishing moments. Note that it is not necessary to choose ψ to be a *"mother"* wavelet associated to a multiresolution analysis of $\mathbb{L}^2(\mathbb{R})$. The whole theory can be developed without resorting to this assumption. The choice of ψ is then very large.

The wavelet based method can be applied to LRD or self-similar processes for respectively estimating the Hurst or the self-similarity parameter. This method is based on the following properties : for Z a stationary LRD process or a self-similar process having stationary increments, for all $a > 0$, $(d_Z(a,b))_{b\in\mathbb{R}}$ is a zero-mean stationary process and

– If Z is a stationary LRD process,

$$\mathbb{E}(d_Z^2(a,b)) = \mathrm{Var}(d_Z(a,b)) \sim C(\psi,H)a^{2H-1} \quad \text{when } a \to \infty$$

– If Z is a self-similar process having stationary increments,

$$\mathbb{E}(d_Z^2(a,b)) = \mathrm{Var}(d_Z(a,b)) \sim K(\psi,H)a^{2H+1} \quad \text{for all } a > 0$$

with $C(\psi,H)$ and $K(\psi,H)$ two positive constants depending only on ψ and H (those results are proved in Flandrin, 1992, and Abry *et al.*, 1998). Therefore, in both these cases, the variance of wavelet coefficients is a power law of a, and a log-log regression provides an estimator of H. From a path $(Z(1),\ldots,Z(N))$, the estimator will be deduced from the log-log regression of the "natural" sample variance of discretized wavelet coefficients, *i.e.*,

$$S_N(a) = \frac{1}{[N/a]} \sum_{i=1}^{[N/a]} e_Z^2(a,i). \tag{4.2}$$

A graph $(\log a_i, \log S_N(a_i))_{1\le i\le \ell}$ is drawn from a priori family of scales and the slope of the least square regression line provides the estimator $\widehat{H}_{WAV}$ of H. For a FGN (respectively a FBM), Bardet *et al.* (2000) (respectively Bardet, 2002) proved that the $\widehat{H}_{WAV}$ converges to H with a non-optimal convergence rate ($N^{1/3}$ instead of $N^{1/2}$ reached for instance by maximum likelihood estimator). In the semi-parametric frame of a general class of stationary Gaussian LRD processes, it was established by Moulines *et al.* (2006) that the estimator $\widehat{H}_{WAV}$ converges to H with an optimal convergence rate (following the minimax criteria) when an optimal length of windows is known. The theoretical asymptotic behavior of $\widehat{H}_{DFA}$ and $\widehat{H}_{WAV}$ are thus comparable for a Gaussian LRD process.

This is not true any more when a polynomial trended LRD (or self-similar) processes is considered. Indeed, Abry *et al.* (1998) remarked that every degree M polynomial trend is without effects on $\widehat{H}_{WAV}$ since ψ has its M first vanishing moments. Therefore, the larger M, the more robust $\widehat{H}_{WAV}$ is.

Finally, Bardet (2002) established a Khi-squared goodness-of-fit test for a path of FBM (therefore for aggregated FGN) using wavelet analysis. This test is based on a (penalized) distance between the points $(\log a_i, \log S_N(a_i))_{1\leq i\leq \ell}$ and a pseudo-generalized least square regression line (here the scales a_i are selected to behave as $N^{1/3}$).

4.2.2 Application of both estimators to FGN

An example of the DFA and wavelet analysis methods applied to a path of a FGN with different values of H is shown in Fig. 4.1.

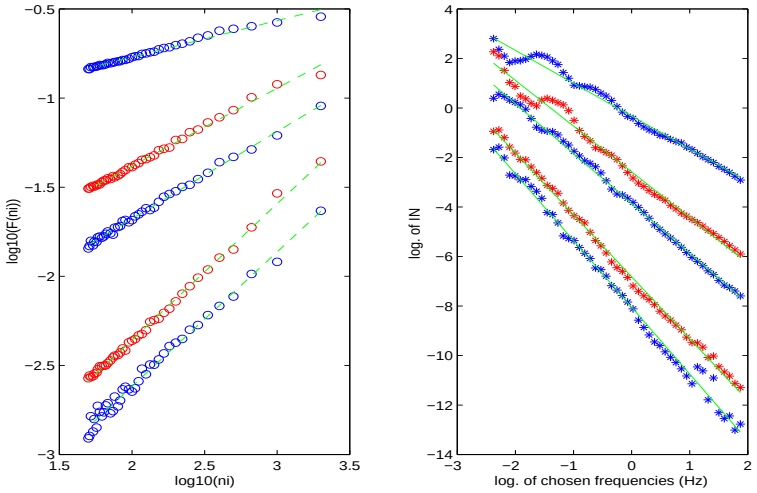


FIG. 4.1 – Results of the DFA method and wavelet analysis applied to a path of a discretized FGN for different values of $H = 0.2, 0.4, 0.5, 0.7, 0.8$, with $N = 10000$

					p-val		$\sqrt{MSE}$	
H_{fGn}	$\widehat{H}_{DFA}$	$\widehat{H}_{WAV}$	$\widehat{\sigma}_{DFA}$	$\widehat{\sigma}_{WAV}$	*DFA*	*WAV*	*DFA*	*WAV*
0.50	0.4936	0.5071	0.0263	0.0427	0.0152	0.0983	0.0187	0.0301
0.60	0.5908	0.6009	0.0290	0.0405	0.0017	0.8289	0.0204	0.0286
0.70	0.6859	0.6985	0.0317	0.0436	0.00001	0.7342	0.0223	0.0304
0.80	0.7875	0.8050	0.0326	0.0388	0.0002	0.1978	0.0230	0.0273
0.90	0.8821	0.8938	0.0366	0.0444	0.000002	0.1030	0.0258	0.0697

TAB. 4.1 – Comparison of the two samples of estimations of H with 100 realizations of fGn path (N=10000) with DFA and wavelets methods

In the Table 4.1 appear the different estimations of H computed from the DFA and wavelet analysis methods for 100 realizations of FGN paths with $N = 10000$. We choose for these simulations the concrete procedure of wavelet analysis developed by Abry *et al.* (2002) (a Daubechies wavelet is chosen and a Mallat's fast pyramidal algorithm is used to compute wavelet coefficients).

In one hand, the wavelets method appear slightly more effective than DFA method considering the p-value which is very low for the sample of the DFA estimations compared to wavelet analysis estimations. This is essentially due to the estimator bias which is more important in the case of DFA. In the other hand, if we consider the root of MSE which is the sum of the squared bias and the variance, the DFA estimator seems to be slightly more effective. Note that for FGN processes (without trend), the Whittle maximum likelihood estimator of H gives a "better" results (see for instance Taqqu *et al.*, 1999).

4.3 Application of both estimators to HR data

Both estimators of H can also be applied to the HR time series of the 9 athletes. The following figures 4.2 and 4.3 exhibit examples of applications of both the estimation method to HR data.

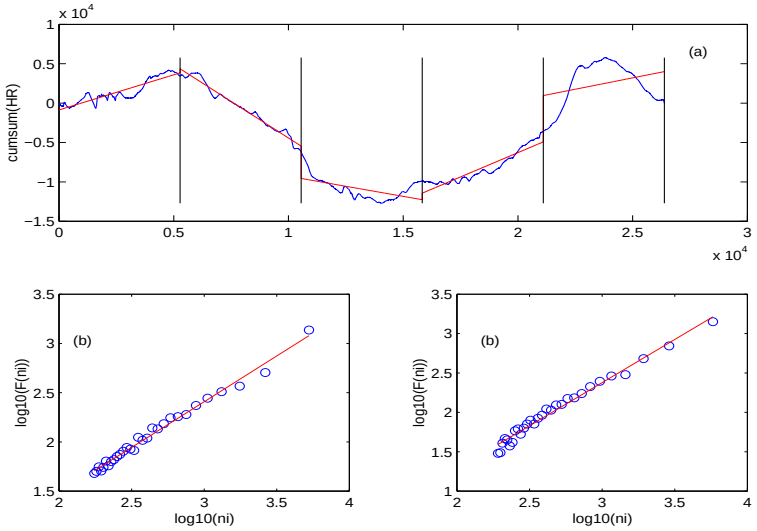


FIG. 4.2 – Two first steps of the DFA method applied to a HR series (up) and results of the DFA method applied to HR series for two different athletes (down)

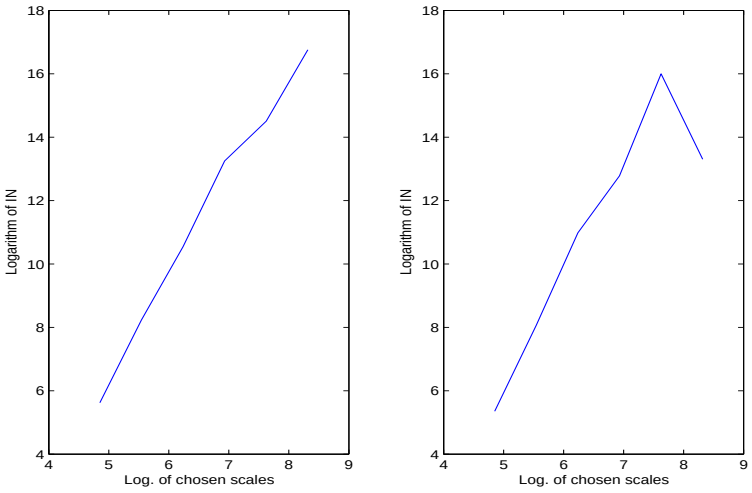


FIG. 4.3 – The log-log graph of the variance of wavelet coefficients relating to the HR series observed during the race and in the end of race (Ath2)

For each athlete, it was first done to the whole time series, and then to the different phases of the race (as it was obtained from the detection of abrupt changes, see Section 2.2). The estimation results of H, for the different signals observed during the three phases of the race, are recapitulated in the Table 4.2 using wavelets method and in Table 4.3 using DFA method.

		Phases		
	HR series	Beginning	Middle	Race end
Ath1	0.8931	1.1268	1.1064	1.2773
Ath2	1.1174	0.7871	1.0916	0.8472
Ath3	1.0208	1.0315	1.1797	-
Ath4	0.9273	-	1.0407	0.7925
Ath5	1.0986	1.3110	1.0113	1.3952
Ath6	1.0769	1.5020	1.1597	1.3673
Ath7	1.0654	1.4237	1.1766	1.0151
Ath8	0.9568	1.6600	0.9699	1.1948
Ath9	0.9379	1.5791	0.9877	0.7263

TAB. 4.2 – Estimated H with wavelets methods for HR series of different athletes

Two main problems resort from these different estimations. First, $\widehat{H}_{DFA}$ and $\widehat{H}_{WAV}$ are often larger than 1. However, the FGN is only defined for $H \in (0,1)$. For defining a process allowing $H > 1$, three main assumptions of FGN have to be changed :

1. the assumption that the process is a stationary process ;
2. the assumption that the process is a Gaussian process ;
3. the assumption that only two parameters (H and σ^2) are sufficient to define the process.

In the sequel (see above), a new model is proposed. Both the first assumptions are still satisfied and the third one is replaced by a semi-parametric assumption.

The second problem is implied by the results of the goodness-of-fit test (for wavelet analysis method). Indeed, this test is never accepted as well for the whole time series as for the partial times series. An explanation of such a phenomenon can be deduced from Figure 4.3 : for the wavelet analysis, the points $(\log a_i, \log S_N(a_i))_{1\leq i\leq \ell}$ are clearly lined for $a_i \leq a_m$, but not exactly lined for $a_i \geq a_m$. Thus the HR time series seems to nearly behave like a FGN for "small" scales (or high frequencies), but not for "large" scales (or small frequencies). A process following this conclusion can not be the better fit of HR time series...

Remark : this last conclusion leads also to a clear advantage of wavelet based over DFA estimator. Indeed, the DFA algorithm measures only one exponent characterizing the entire signal. Then, this method corresponds rather to the study of "monofractal" signals such as FGN. At the contrary, the wavelet method provides the graph $(\log a_i, \log S_N(a_i))_{1\leq i\leq \ell}$ which can be very interesting for analyze the "fractal" behavior of data (see also Billat *et al.*, 2005).

4.4 A second model : a locally fractional Gaussian noise

In Bardet and Bertrand (2007), a generalization of the FBM, so-called the (M_K)-multiscale FBM, was introduced. The (M_0)-FBM is a FBM with self-similarity parameter H_0. Roughly speaking, the (M_K)-FBM has the same harmonizable representation (and therefore quite the same behavior as the FBM) than a FBM with self-similarity parameter H_i for frequencies $|\xi| \in [\omega_i, \omega_{i+1}[$ for all $i = 0, \ldots, K$ ($K \in \mathbb{N}$). For instance, a (M_1)-FBM behaves as a FBM with self-similarity parameter H_0 for small frequencies and as a FBM with self-similarity parameter H_1 for high frequencies. Such a model was fruitfully used for modeling biomechanical signals (position of the center of pressure on a force platform during quiet postural stance measured at a frequency of 100 Hz for the one minute period).

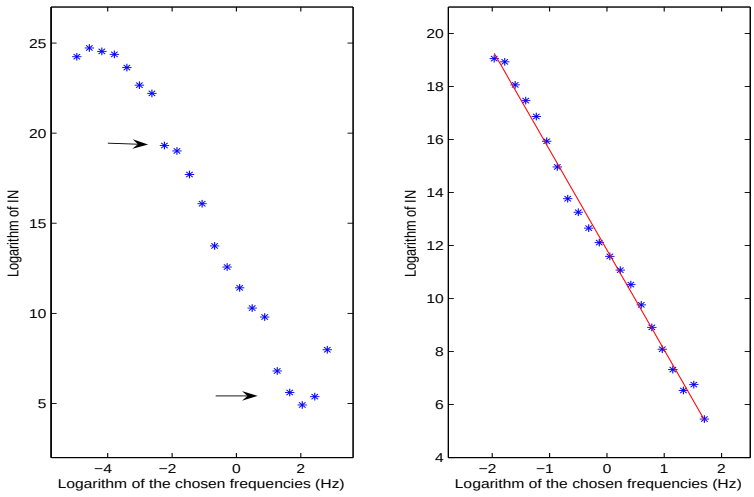


FIG. 4.4 – The log-log graph of the variance of wavelet coefficients relating to the HR series observed during the arrival phase (Ath6) with a frequency band of [0.01 12](right) and of [0.2 4](left).

Here, Fig. 4.4 suggests than a fitted model for aggregated HR data should behave like a FBM with self-similarity parameter H for low frequencies and differently for high frequencies (and not necessary like a FBM). Thus define a locally fractional Brownian motion $X_\rho = \{X_\rho(t), t \in \mathbb{R}\}$ as the process such that :

$$X_\rho(t) = \int_{\mathbb{R}} \frac{e^{it\xi} - 1}{\rho(\xi)} \widehat{W}(d\xi)$$

where the function $\rho : \mathbb{R} \to [0, \infty)$ is an even continuous function such that :

- $\rho(\xi) = \dfrac{1}{\sigma} |\xi|^{H+1/2}$ for $|\xi| \in [\omega_0, \omega_1]$ with $H \in \mathbb{R}$, $\sigma > 0$ and $0 < \omega_0 < \omega_1$
- $\displaystyle\int_{\mathbb{R}} \left(1 \wedge |\xi|^2\right) \frac{1}{\rho^2(\xi)}\, d\xi < \infty.$

and $W(d\xi)$ is a Brownian measure and $\widehat{W}(d\xi)$ its Fourier transform in the distribution meaning. Cramér and Leadbetter (1967) proved the existence of such Gaussian process with stationary increments. The main advantages of such process compared to usual FBM are the following :

1. X_ρ "behaves" like a FBM only for local band of frequencies ;
2. In this band, the parameter H is not restricted to be in $(0, 1)$: it is in $\mathbb{R}$.

From this definition, one deduces a possible model for HR data :

$$Y_\rho(t) = X_\rho(t+1) - X_\rho(t) = 2 \cdot \mathcal{R}e\Big(\int_{\mathbb{R}} \frac{e^{it\xi} \sin(\xi/2)}{\rho(\xi)} \widehat{W}(d\xi) \Big) \quad \text{for } t \in \mathbb{R}.$$

Note that $Y_\rho = \{Y_\rho(t), t \in \mathbb{R}\}$ is a stationary Gaussian process and the function $2\sin(\xi/2)\rho^{-1}(\xi)$ is so-called the spectral density of Y_ρ.

Let $\Delta_N \to 0$ and $N\Delta_N \to \infty$ when $N \to \infty$. The wavelet based estimator can provide a convergent estimation of H when a path

$$(Y_\rho(\Delta_N), Y_\rho(2\Delta_N), \ldots, Y_\rho(N\Delta_N))$$

and therefore a path $(X_\rho(\Delta_N), \ldots, X_\rho(N\Delta_N))$ is observed. Indeed, consider a "mother" wavelet ψ such that $\psi : \mathbb{R} \mapsto \mathbb{R}$ is a $\mathcal{C}^\infty$ function satisfying :

- for all $s \geq 0$, $\displaystyle\int_{\mathbb{R}} |t^s \psi(t)|\, dt < \infty$;
- its Fourier transform $\widehat{\psi}(\xi)$ is an even function compactly supported on $[-\beta, -\alpha] \cup [\alpha, \beta]$ with $0 < \alpha < \beta$.

Then, using results of Bardet and Bertrand (2007), for all $a > 0$ such that $[\frac{\alpha}{a}, \frac{\beta}{a}] \subset [\omega_0, \omega_1]$, *i.e.* $a \in [\frac{\beta}{\omega_1}, \frac{\alpha}{\omega_0}]$, $(d_{X_\rho}(a,b))_{b\in\mathbb{R}}$ is a stationary Gaussian process and

$$\mathbb{E}\big(d^2_{X_\rho}(a,.)\big) = K(\psi, H, \sigma) \cdot a^{2H+1},$$

with $K(\psi, H, \sigma) > 0$ only depending on ψ, H and σ. However this property is checked if and only if the function ψ is chosen such that :

$$\frac{\beta}{\alpha} < \frac{\omega_1}{\omega_0}.$$

Moreover, for $a \in [\frac{\beta}{\omega_1}, \frac{\alpha}{\omega_0}]$, the sample variance $S_N(a)$ defined in (4.2) and computed from a path $(X_\rho(\Delta_N), \ldots, X_\rho(N\Delta_N))$ converges to $\mathbb{E}\big(d^2_{X_\rho}(a,.)\big)$ and satisfies a central limit theorem with convergence rate $\sqrt{N\Delta_N}$. Thus, with fixed scales $(a_1, \ldots, a_\ell) \in [\frac{\beta}{\omega_1}, \frac{\alpha}{\omega_0}]^\ell$, a log-log-regression of $(a_i, S_N(a_i))_{1\leq i\leq \ell}$ provides an estimation of H (and a central limit theorem with convergence rate $N\Delta_N$ satisfied by $\widehat{H}_{WAV}$ can also be established). As previously, we consider also Khi-squared goodness-of-fit test based on the wavelet analysis and defined as a weighted distance between points $(\log(a_i), \log(S_N(a_i)))_{1\leq i\leq \ell}$ and a pseudo-generalized regression line.

Remark : The main problem with these estimator and test is the localization of the suitable frequency band $[\omega_0, \omega_1]$ (ω_0 and ω_1 are assumed to be unknown parameters). A solution consists in selecting a very large band of scales and determining then graphically the "most" linear part of the set of points $(\log(a_i), \log(S_N(a_i)))_{1\leq i\leq \ell}$. Another possible way may be to compute an adaptive estimator of this band using a quadratic criterion (following a similar procedure than in Bardet and Bertrand, 2007). Here, like 9 different paths of HR data are observed, a common frequency band $[\overline{\omega_0}, \overline{\omega_1}]$ can be graphically obtained and used for whole HR data (see above).

4.5 Application to HR data

First, one considers that a HR time series $(Y(1), \ldots, Y(n))$ can be written like $(Y_\rho(\Delta_n), Y_\rho(2\Delta_n), \ldots, Y_\rho(n\Delta_n))$, $Y_\rho = \{Y_\rho(t), t \in \mathbb{R}\}$ a process defined as previously. Secondly, the wavelet analysis is applied to the 9 (whole or partial) HR time series (the chosen "mother" wavelet is a kind of Lemarié-Meyer wavelet such that $\beta = 2\alpha$). Using first a very large band of scales for all HR time series (for example [0.01, 12] in Fig. 4.5), one estimation of frequency band is deduced : $[\overline{\omega_0}, \overline{\omega_1}] = [0.2, 4]$ is the chosen frequency band for the whole and partial signals.

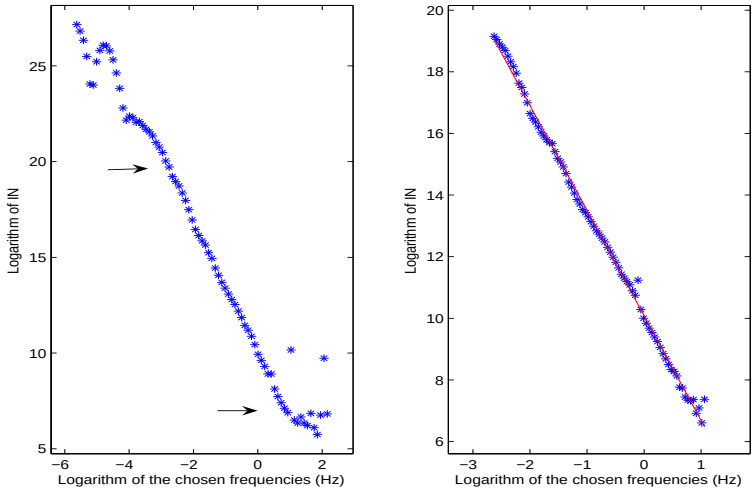


FIG. 4.5 – The log-log graph of the variance of wavelet coefficients relating to the HR series observed in the middle of the exercise (Ath5)

The estimation results of H, for the different signals observed during the three phases of the race, are recapitulated in the Table 4.3.

Both DFA and wavelet analysis methods provide estimations of Hurst exponent which reflect the possible modeling of HR data with long range dependence time series.
We also note that with a p-value of 0.64, both the samples $(\widehat{H}_{DFA})_{1,\ldots,9}$ and $(\widehat{H}_{WAV})_{1,\ldots,9}$ obtained from all HR time series are significantly close.

The same comparison can also be done when the three characteristic stages of the race (beginning, middle and end of the race) are distinguished. The result is different. Indeed, the corresponding p-values between $(\widehat{H}_{DFA})_{1,\ldots,9}$ and $(\widehat{H}_{WAV})_{1,\ldots,9}$ are significatively different in the middle part of the race (and relatively different in the stage of race end).

	HR series		Race beginning		During the race		End of race	
	$\widehat{H}_{DFA}$	$\widehat{H}_{WAV}$	$\widehat{H}_{DFA}$	$\widehat{H}_{WAV}$	$\widehat{H}_{DFA}$	$\widehat{H}_{WAV}$	$\widehat{H}_{DFA}$	$\widehat{H}_{WAV}$
Ath1	0.928	1.288*	1.032	1.192	1.060	1.214	0.429	1.400
Ath2	1.095	1.268*	0.905	0.973	1.126	1.108	1.240	1.452*
Ath3	1.163	1.048	0.553	0.898	1.130	1.172	-	-
Ath4	1.193	0.916*	-	-	1.098	1.249*	1.172	1.260
Ath5	1.239	1.110	1.267	1.117*	1.133	1.205	1.273	1.348*
Ath6	1.247	1.084*	1.237	1.106	1.091	1.172	1.436	1.338
Ath7	1.155	1.095	0.850	1.295	1.182	1.186*	1.129	1.209
Ath8	1.258	1.011	1.304	1.128*	0.995	1.134	1.122	1.247
Ath9	1.243	1.429*	0.820	1.019	1.127	1.535*	1.250	1.238*
p-value	0.6414		0.3723		0.0225		0.1260	
F-stat	0.23		0.85		6.38		2.65	

TAB. 4.3 – Estimated $\widehat{H}$, with DFA and wavelets methods, for HR series of different athletes (*) The series for which the test is rejected. Comparison of the two samples $(\widehat{H}_{DFA})_{1,\dots,9}$ and $(\widehat{H}_{WAV})_{1,\dots,9}$ for whole and partial series (p-value).

In spite of values relating to the estimator of H for all the athletes in the different phases which are relatively large, the DFA has sometimes tendency to under estimating this parameter like in the race beginning (Ath3) and the end of race (Ath1). Indeed, these value are clearly due to a certain trend supports by the fact that data points in log-log plot (Fig. 4.6) have not a straight line form, and we have proved in (15) that the DFA method is not robust in the case of trended long range dependent process. However in both the cases, the wavelets method is more effective since it removes sufficiently this kind of trend.

For HR data and when the goodness-of-fit test is accepted, the wavelet method shows a fractal parameter H close to 1. According to the different studies (using DFA method) about physiologic time series for distinguishing healthy from pathologic data sets (see (48), (78), (79)), an exponent $H \simeq 1$ indicate a healthy cardiac HR time series. Indeed, for the study concerning a 24 hours recorded interbeat time series during the exercise for healthy adults and heart failure adults, the following results are obtained : for healthy subjects, $H = 1.01 \pm 0.16$, for the group of heart failure subjects $H = 1.24 \pm 0.22$. During the different stages of the marathon race, a small increase of the fractal parameter H is observed especially at the end of races. This behavior and this evolution may be associated with fatigue appearing during the last phase of the marathon. This evolution

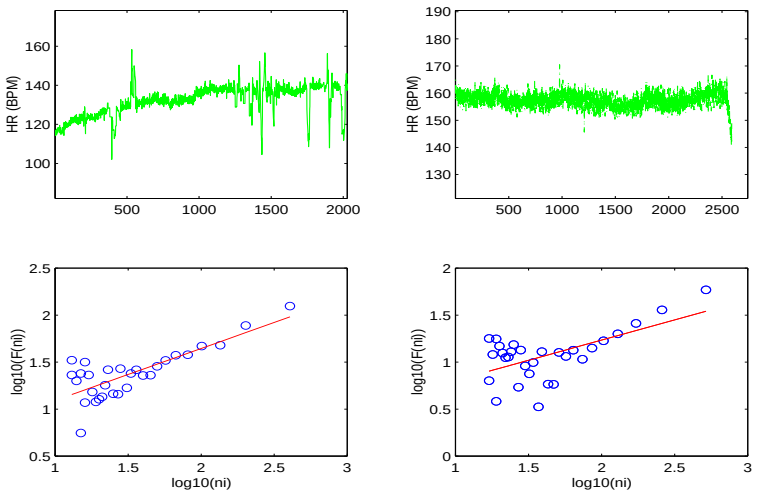


FIG. 4.6 – The results of the DFA method applied to records for race beginning (Ath3) (left) and for end of race (Ath1) (right)

can not be observed with DFA method. Indeed, in one hand, when we observe the three 9-samples of wavelet estimators (related to the 3 phases of the race), the p-value (see Fig. 4.7) indicates a significantly difference due precisely to this evolution of the fractal parameter. On the other hand, a large p-value (0.85) is obtained for the same test using DFA estimation.

	$\widehat{H}_{DFA}$	$\widehat{H}_{WAV}$
p-value	0.8570	0.0158
F-stat	0.16	5.27

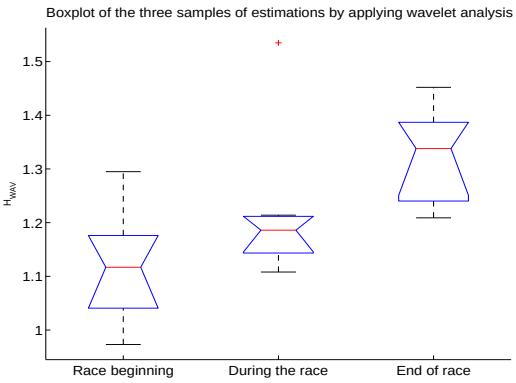


FIG. 4.7 – Comparison of the three samples constituting by estimations in the beginning of race, during the race and then in the race end by the DFA and wavelet methods

The representation given by Fig. 4.7, highlight a difference in the behaviors of HR series in the beginning of the race and in the end of race. Indeed, the dispersions in the first and last sample are more important than in the middle of race and it seems that each athlete starts and finished the race at his own rhythm but in the middle athletes seems to have the same rate.

4.6 Conclusion

As indicated in the beginning of the last section, our main goal is to see whether the heart rate time series during the race have specific properties that of scaling law behavior. The wavelet analysis and the DFA methods are applied to 9 HR time series during the whole and also the different three phases of the race (beginning, middle and end of race) obtained by an automatic procedure. Even if their results are not exactly the same, both methods provide Hurst exponents which reflect the possible modeling of HR data by a LRD time series. However, in (15), even if the DFA estimator of Hurst parameter is proved to be convergent with a reasonable convergence rate for LRD stationary Gaussian processes, it is not at all a robust method in case of trend. The wavelet based method provides a more precise and robust estimator of the Hurst parameter. Thus, the results obtained from this wavelet estimator seem to be more valid.

Moreover, a Khi-squared goodness-of-fit test can also be deduced from this method. It seems to show that a classical LRD stationary Gaussian process is not exactly a suitable model for HR data. Graphs obtained with wavelet analysis also show that a locally fractional Gaussian noise, a semi-parametric process defined in Section 2.3 could be more relevant for modeling these data. A Khi-squared test confirms the goodness-of-fit of such a model. Thus, using the wavelet estimation of a fractal parameter in a specific frequency band, one obtains a conclusion relatively close to those obtained by other studies (conclusion which can not be detected with DFA method) : these fractal parameters increase through the race phases, what may be explained with fatigue appearing during the last phase of the marathon. Thus this fractal parameter may be a relevant factor to detect a change during a long-distance race.

Finally, for the 9 athletes and as the test is validated with significance level around 0.65, we can estimate $\widehat{H}_{beginning}$ at 1.1, the $\widehat{H}_{middle}$ at 1.2 and $\widehat{H}_{end}$ at 1.3 with a larger confidence interval at the beginning and the end of the race. This behavior could bring a new way of understanding what is happening during a race.

Chapitre 5

Detecting changes in the long-range dependence or the self-similarity or the local fractality of a Gaussian process

5.1 Introduction

The content of this paper was motivated by a general study of physiological signals of runners recorded during endurance races as marathons. More precisely, after different signal procedures for "cleaning" data, one considers the time series resulting of the evolution of heart rate (HR data in the sequel) during the race. The following figure provides several examples of such data. For each runner, the periods (in ms) between the successive pulsations (see Fig. 5.1) are recorded. The HR signal in number of beats per minute (bpm) is then deduced (the HR average for the whole sample is of 162 bpm).

Numerous authors have studied such data (see for instance (78), (79) or (5)). A model proposed to fit these data is a trended long memory process with an estimated Hurst parameter close to 1 (and sometimes more than 1). In (57) three improvements have been proposed to such result : 1/ data are stepped in three different stages which are detected using a change point's detection method (see for instance (63), or (66)). The main idea of the detection's method is to consider that the signal distribution depends on a vector of unknown characteristic parameters constituted by the mean and the variance. The different stages (beginning, middle and end of the race) and therefore the different vectors of parameters, which change at two unknown instants, are estimated. 2/ during each stage, a time-continuous Gaussian process is proposed for modelling the

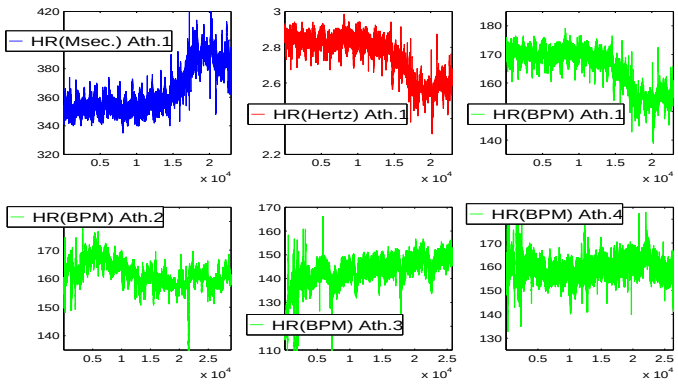


FIG. 5.1 – Heat rate signals of Athlete 1 in ms, Hertz and BPM (up), of Athletes 2, 3 and 4 in BPM (down)

detrended time series. This process is a generalization of fractional Gaussian noise also called locally fractional Gaussian noise such that roughly speaking there exists a local-fractality parameter $H \in \mathbb{R}$ only for frequencies $|\xi| \in [f_{min}\,,\, f_{max}]$ with $0 < f_{min} < f_{max}$ (see more details below). 3/ this parameter H which is very interesting for interpreting and explaining the physiological signal behaviours, is estimating from a wavelet analysis. Rigorous results are also proved providing a central limit theorem satisfied by the estimator.

In order to improve this study of HR data and since the eventual changes of H values are extremely meaningful for explaining the eventual physiological changes of athlete's HR during the race, the detection of abrupt change of H values is the aim of this paper. By this way the different stages detected during the race will be more relevant for explaining the physiological status of the athlete than stages detected from changes in mean or variance. For instance, the HR of a runner could decrease in mean even if the regularity of the HR does not change.

In this paper, an estimator of m instants ($m \in \mathbb{N}^*$) of abrupt changes of long-range dependence, self-similarity or local-fractality (more details about these terms will be provided below) is developed for a sample of a Gaussian process. Roughly speaking, the principle of such estimator is the following : in each time's domain without change, the parameter of long-range dependence (or self-similarity or local self-fractality) can be estimated from a log-log regression of wavelet coefficients' variance onto several chosen scales. Then a contrast defined by the sum on every $m+1$ possible zones of square distances between points and regressions lines is minimized providing an estimator of the m instants of change. Under general assumptions, a limit theorem with a convergence rate satisfied by such an estimator is established in Theorem 5.2.1.

Moreover, in each estimated no-change zone, parameters of long-range dependence (or self-similarity or local self-similarity) can be estimated, first with a ordinary least square (OLS) regression, secondly with a feasible generalized least square (FGLS) regression. Central limit theorems are established for both these estimators (see Theorem 5.2.2 and Proposition 5.2.3 below) and confidence intervals can therefore be computed. The FGLS estimator provides two advantages : from the one hand, its asymptotic variance is smaller than OLS estimator one. From the other hand, it allows to construct a very simple (Khi-square) goodness-of-fit test based on a squared distance between points and FGLS regression line. The asymptotic behavior of this test is provided in Theorem 5.2.4.

Different particular cases of Gaussian processes are studied :

1. long-range dependent processes with abrupt changes of values of LRD parameters. In such time series case, a semi-parametric frame is supposed including fractional Gaussian noises (FGN) and Gaussian FARIMA processes and assumptions of limit theorems are always satisfied with interesting convergence rates (see Corollary 5.3.2).
2. self-similar time series with abrupt changes of values of self-similarity parameters. In such case, fractional Brownian motions (FBM) are only considered. Surprisingly, convergence of estimators is only established when the maximum of differences between self-similarity parameters is sufficiently small. Simulations exhibit a non convergence of the estimator of instant change when a difference between two parameters is too large (see Corollary 5.3.4).
3. locally fractional Gaussian processes with abrupt changes of values of local-fractality parameters. In such a continuous time processes' case, a semi-parametric frame is supposed (including multiscale fractional Brownian motions) and assumptions of limit theorems are always satisfied with interesting convergence rates (see Corollary 5.3.6).

The problem of change-point detection using a contrast minimization was first studied in the case of independent processes (see for instance Bai and Perron (9)), then for weakly dependent processes (see for instance Bai (8), Lavielle (63) or Lavielle and Moulines (64)) and since middle of 90's in the case of processes which exhibit long-range dependance (see for instance Giraitis *et al.* (43), Kokoszka and Leipus (61) or Lavielle and Teyssière (66)). Of the various approaches, some were associated with a parametric framework for a change points detection in mean and/or variance and others where associated with a non-parametric framework (typically like detecting changes in distribution or spectrum). To our knowledge, the semi-parametric case of abrupt change detection for long-range dependent or self-similarity parameter is treated here for the first time.

However, in the literature different authors have proposed test statistics for testing

the no-change null hypothesis against the alternative that the long-memory parameter changes somewhere in the observed time series. Beran and Terrin (26) proposed an approach based on the Whittle estimator, Horváth and Shao (53) obtained limit distribution of the test statistic based on quadratic forms and Horváth (52) suggested another test based on quadratic forms of Whittle estimator of long-memory parameter. The goodness-of-fit test presented below and which satisfies the limit theorem 5.2.4 also allows to test if the long-range memory (or self-similarity or local-fractality) parameter changes somewhere in the time series.

Our approach is based on the wavelet analysis. This method applied to LRD or self-similar processes for respectively estimating the Hurst or self-similarity parameter was introduced by Flandrin (41) and was developed by Abry, Veitch and Flandrin (3) and Bardet, Lang, Moulines, Soulier (18). The convergence of wavelet analysis estimator was studied in the case of a sample of FBM in (10), and in a semi-parametric frame of a general class of stationary Gaussian LRD processes by Moulines*et al.* (73) and Bardet *et al.* (18).

A method based on wavelet analysis was also developed by Bardet and Bertrand (12) in the case of multiscale FBM (a generalization of the FBM for which the Hurst parameter depends on the frequency as a piecewise constant function) providing statistics for the identification (estimation and goodness-of-fit test) of such a process. Such a process was used for modelling biomechanical signals. In the same way, the locally fractional Gaussian process (a generalization of the FBM for which the Hurst parameter, called the local-fractality parameter, is constant in a given domain of frequencies) was studied in (57) for modelling HR data during the three characteristics stages of the race. An increasing evolution of the local-fractality parameter during the race was generally showed for any runner from this method. Using the method of abrupt change detection of local-fractality parameter H developed in Corollary 5.3.6, this result is confirmed by estimations of H for each runner even if the change's instants seem to vary a lot depending on the fatigue of the runner (see the application to HR's time series in Section 5.3).

The paper is organized as follows. In Section 5.2, notations, assumptions and limit theorems are provided in a general frame. In Section 5.3, applications of the limit theorems to three kind of "piecewise" Gaussian process are presented with also simulations. The case of HR data is also treated. Section 5.4 is devoted to the proofs.

5.2 Main results

5.2.1 Notations and assumptions

First, a general and formal frame can be proposed. Let $(X_t)_{t\in T}$ be a zero-mean Gaussian process with $T = \mathbb{N}$ or $T = \mathbb{R}$ and assume that

$$\big(X_0, X_{\delta_N}, X_{2\delta_N}, \ldots, X_{N\delta_N}\big) \text{ is known with } \delta_N = 1 \text{ or } \delta_N \underset{N\to\infty}{\longrightarrow} 0,$$

following data are modeled with a time series ($T = \mathbb{N}$) or a continuous time process $T = \mathbb{R}$. In the different proposed examples X could be a stationary long memory time series or a self-similar or locally fractional process having stationary increments.
For estimations using a wavelet based analysis, consider $\psi : \mathbb{R} \to \mathbb{R}$ a function called "the mother wavelet". In applications, ψ is a function with a compact (for instance Daubeshies wavelets) or an essentially compact support (for instance Lemarié-Meyer wavelets). For $(X_t)_{t\in T}$ and $(a, b) \in \mathbb{R}_+^* \times \mathbb{R}$, the wavelet coefficient of X for the scale a and the shift b is

$$d_X(a, b) := \frac{1}{\sqrt{a}} \int_{\mathbb{R}} \psi(\frac{t}{a} - b)X(t)dt.$$

When only a discretized path of X is available (or when $T = \mathbb{N}$), approximations $e_X(a, b)$ of $d_X(a, b)$ are only computable. We have chosen to consider for $(a, b) \in \mathbb{R}_+^* \times \mathbb{N}$,

$$e_X(a, b) := \frac{\delta_n}{\sqrt{a}} \sum_{p=1}^{N} \psi\big(\frac{p}{a} - b\big)X_{p\,\delta_N}, \tag{5.1}$$

which is the formula of wavelet coefficients computed from Mallat's algorithm for compactly supported discrete ($a \in 2^{\mathbb{N}}$) wavelet transform (for instance Daubeshies wavelets) when N is large enough and nearly this formula for discrete wavelet transform with an essentially compact support (for instance Lemarié-Meyer wavelets). Now assume that there exist $m \in \mathbb{N}$ (the number of abrupt changes) and

- $0 = \tau_0^* < \tau_1^* < \ldots < \tau_m^* < \tau_{m+1}^* = 1$ (unknown parameters) ;
- two families $(\alpha_j^*)_{0\le j\le m} \in \mathbb{R}^{m+1}$ and $(\beta_j^*)_{0\le j\le m} \in (0, \infty)^{m+1}$ (unknown parameters) ;
- a sequence of "scales" $(a_n)_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$ (known sequence) satisfying $a_n \ge a_{min}$ for all $n \in \mathbb{N}$, with $a_{min} > 0$,

such that for all $j = 0, 1, \ldots, m$ and $k \in D_N^*(j) \subset \big[\big[N\delta_N\tau_j^*\big], \big[N\delta_N\tau_{j+1}^*\big]\big]$,

$$\mathbb{E}\big[e_X^2(a_N, k)\big] \sim \beta_j^* \cdot \big(a_N\big)^{\alpha_j^*} \quad \text{when} \quad N \to \infty \text{ and } N\delta_N \to \infty. \tag{5.2}$$

Roughly speaking the variance of wavelet coefficients follows a power law of the scale, and this power law is piecewise varying following the shift. Thus piecewise sample variances can be appropriated estimators of parameters of these power laws. Hence let us define

$$S_k^{k'}(a_N) := \frac{a_N}{k'-k} \sum_{p=[k/a_N]}^{[k'/a_N]-1} e_X^2(a_N, a_N\, p) \quad \text{for } 0 \le k < k' \le N\delta_N. \tag{5.3}$$

Now set $0 < r_1 < \ldots < r_\ell$ with $\ell \in \mathbb{N}^*$ and let us suppose that a multidimensional central limit theorem can also be established for $\left(S_k^{k'}(r_i\, a_N)\right)_{1\le i\le \ell}$, *i.e.*

$$\left(S_k^{k'}(r_i\, a_N)\right)_{1\le i\le \ell} = \left(\beta_j^* \cdot \left(r_i\, a_N\right)^{\alpha_j^*}\right)_{1\le i\le \ell} + \left(a_N\right)^{\alpha_j^*} \times \sqrt{\frac{a_N}{k'-k}} \left(\varepsilon_i^{(N)}(k,k')\right)_{1\le i\le \ell}, \tag{5.4}$$

with $[N\delta_N\tau_j^*] \le k < k' \le [N\delta_N\tau_{j+1}^*]$ and it exists $\Gamma^{(j)}(\alpha_j^*, r_1, \ldots, r_\ell) = \left(\gamma_{pq}^{(j)}\right)_{1\le p,q\le \ell}$ a $(\ell \times \ell)$ matrix not depending on N such that $\alpha \mapsto \Gamma^{(j)}(\alpha, r_1, \ldots, r_\ell)$ is a continuous function, a positive matrix for all α and

$$\left(\varepsilon_i^{(N)}(k,k')\right)_{1\le i\le \ell} \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \Gamma^{(j)}(\alpha_j^*, r_1, \ldots, r_\ell)\right) \quad \text{when} \quad k'-k \to \infty. \tag{5.5}$$

With the usual Delta-Method, relation (5.4) implies that for $1 \le i \le \ell$,

$$\log\left(S_k^{k'}(r_i\, a_N)\right) = \log(\beta_j^*) + \alpha_j^* \log\left(r_i\, a_N\right) + \sqrt{\frac{a_N}{k'-k}}\, \varepsilon_i^{(N)}(k,k'), \tag{5.6}$$

for $[N\delta_N\tau_j^*] \le k < k' \le [N\delta_N\tau_{j+1}^*]$ and the limit theorem (5.5) also holds. This is a linear model and therefore a log-log regression of $\left(S_k^{k'}(r_i\, a_N)\right)_i$ onto $\left(r_i\, a_N\right)_i$ provides an estimator of α_j^* and $\log(\beta_j^*)$.

The first aim of this paper is the estimation of the unknown parameters $(\tau_j^*)_j$, $(\alpha_j^*)_j$ and $(\beta_j^*)_j$. Therefore, define a contrast function

$$U_N\left((\alpha_j)_{0\le j\le m}, (\beta_j)_{0\le j\le m}, (k_j)_{1\le j\le m}\right) = \sum_{j=0}^{m} \sum_{i=1}^{\ell} \left(\log\left(S_{k_j}^{k_{j+1}}(r_i\, a_N)\right) - \left(\alpha_j \log(r_i\, a_N) + \log\beta_j\right)\right)^2$$

$$\text{with} \quad \begin{cases} \bullet\ (\alpha_j)_{0\le j\le m} \in A^{m+1} \subset \mathbb{R}^{m+1} \\ \bullet\ (\beta_j)_{0\le j\le m} \in B^{m+1} \subset (0,\infty)^{m+1} \\ \bullet\ 0 = k_0 < k_1 < \ldots < k_m < k_{m+1} = N\delta_N, (k_j)_{1\le j\le m} \in K_m(N) \subset \mathbb{R}^m \end{cases}.$$

The vector of estimated parameters $\widehat{\alpha}_j$, $\widehat{\beta}_j$ and $\widehat{k}_j$ (and therefore $\widehat{\tau}_j$) is the vector which minimizes this contrast function, *i.e.*,

$$\left((\widehat{\alpha}_j)_{0\le j\le m}, (\widehat{\beta}_j)_{0\le j\le m}, (\widehat{k}_j)_{1\le j\le m}\right) := \text{Argmin}\left\{U_N\left((\alpha_j)_{0\le j\le m}, (\beta_j)_{0\le j\le m}, (k_j)_{1\le j\le m}\right)\right\} \quad \text{in} \quad A^{m+1} \times B^{m+1} \times K_m(N) \tag{5.7}$$

$$\widehat{\tau}_j := \widehat{k}_j / N(\delta_N) \quad \text{for} \quad 1 \le j \le m. \tag{5.8}$$

For a given $(k_j)_{1\leq j\leq m}$, it is obvious that $(\widehat{\alpha}_j)_{0\leq j\leq m}$ and $(\log\widehat{\beta}_j)_{0\leq j\leq m}$ are obtained from a log-log regression of $\big(S_{k_j}^{k_{j+1}}(r_i\,a_N)\big)_i$ onto $\big(r_i\,a_N\big)_i$, *i.e.*

$$\begin{pmatrix} \widehat{\alpha}_j \\ \log\widehat{\beta}_j \end{pmatrix} = \big(L_1'\cdot L_1\big)^{-1}L_1'\cdot Y_{k_j}^{k_{j+1}}$$

with $Y_{k_j}^{k_{j+1}} := \big(\log\big(S_{k_j}^{k_{j+1}}(r_i\cdot a_N)\big)\big)_{1\leq i\leq \ell}$ and $L_{a_N} := \begin{pmatrix} \log(r_1\,a_N) & 1 \\ \vdots & \vdots \\ \log(r_\ell\,a_N) & 1 \end{pmatrix}$. Therefore the estimator of the vector $(k_j)_{1\leq j\leq m}$ is obtained from the minimization of the contrast

$$G_N(k_1,k_2,\ldots,k_m) \;:=\; U_N\big((\widehat{\alpha}_j)_{0\leq j\leq m},\,(\widehat{\beta}_j)_{0\leq j\leq m},\,(k_j)_{1\leq j\leq m}\big) \tag{5.9}$$

$$\Longrightarrow\quad (\widehat{k}_j)_{1\leq j\leq m} = \mathrm{Argmin}\Big\{G_N(k_1,k_2,\ldots,k_m),\quad (k_j)_{1\leq j\leq m}\in K_m(N)\Big\}. \tag{5.10}$$

5.2.2 Estimation of abrupt change time-instants $(\tau_j^*)_{1\leq j\leq m}$

In this paper, parameters (α_j^*) are supposed to satisfied abrupt changes. Such an hypothesis is provided by the following assumption :

Assumption C : Parameters (α_j^*) are such that $|\alpha_{j+1}^*-\alpha_j^*|\neq 0$ for all $j=0,1,\ldots,m-1$.

Now let us define :

$$\underline{\tau}^* := (\tau_1^*,\ldots,\tau_m^*),\quad \underline{\widehat{\tau}} := (\widehat{\tau}_1,\ldots,\widehat{\tau}_m)\quad\text{and}\quad \|\underline{\tau}\|_m := \max\big(|\tau_1|,\ldots,|\tau_m|\big).$$

Then $\underline{\widehat{\tau}}$ converges in probability to $\underline{\tau}^*$ and more precisely,

Theorem 5.2.1. *Let $\ell\in\mathbb{N}\setminus\{0,1,2\}$. If Assumption C and relations (5.4), (5.5) and (5.6) hold with $(\alpha_j^*)_{0\leq j\leq m}$ such that $\alpha_j^*\in[a\,,\,a']$ and $a<a'$ for all $j=0,\ldots,m$, then if $a_N^{1+2(a'-a)}(N\,\delta_N)^{-1}\underset{N\to\infty}{\longrightarrow}0$, for all $(v_n)_n$ satisfying $v_N\cdot a_N^{1+2(a'-a)}(N\,\delta_N)^{-1}\underset{N\to\infty}{\longrightarrow}0$,*

$$\mathbb{P}\Big(v_N\|\underline{\tau}^*-\underline{\widehat{\tau}}\|_m\geq\eta\Big)\underset{N\to\infty}{\longrightarrow}0\quad \textit{for all } \eta>0. \tag{5.11}$$

Several examples of applications of this theorem will be seen in Section 5.3.

5.2.3 Estimation of parameters $(\alpha_j^*)_{0\le j\le m}$ and $(\beta_j^*)_{0\le j\le m}$

For $j=0,1,\ldots,m$, the log-log regression of $\big(S_{\widehat{k}_j}^{\widehat{k}_{j+1}}(r_i a_N)\big)_{1\le i\le \ell}$ onto $(r_i a_N)_{1\le i\le \ell}$ provides the estimators of α_j^* and β_j^*. However, even if τ_j converges to τ_j^*, $\widehat{k}_j = N\delta_N\cdot\widehat{\tau}_j$ does not converge to k_j^* (except if $N=o(v_N)$ which is quite impossible), and therefore $\mathbb{P}\big([\widehat{k}_j,\widehat{k}_{j+1}]\subset[k_j^*,k_{j+1}^*]\big)$ does not tend to 1. So, for $j=0,1,\ldots,m$, define $\tilde{k}_j$ and $\tilde{k}'_j$ such that

$$\tilde{k}_j=\widehat{k}_j+\frac{N\delta_N}{v_N}\quad\text{and}\quad\tilde{k}'_j=\widehat{k}_{j+1}-\frac{N\delta_N}{v_N}\quad\Longrightarrow\quad\mathbb{P}\big([\tilde{k}_j,\tilde{k}'_j]\subset[k_j^*,k_{j+1}^*]\big)\underset{N\to\infty}{\longrightarrow}1,$$

from (5.11) with $\eta=1/2$. Let $\Theta_j^*:=\begin{pmatrix}\alpha_j^*\\ \log\beta_j^*\end{pmatrix}$ and $\tilde{\Theta}_j:=(L_1'\cdot L_1)^{-1}L_1'\cdot Y_{\tilde{k}_j}^{\tilde{k}'_j}:=\begin{pmatrix}\tilde{\alpha}_j\\ \log\tilde{\beta}_j\end{pmatrix}$. Thus, estimators $(\tilde{\alpha}_j)_{0\le j\le m}$ and $(\tilde{\beta}_j)_{0\le j\le m}$ satisfy

Theorem 5.2.2. *Under the same assumptions as in Theorem 5.2.1, for $j=0,\ldots,m$*

$$\sqrt{\frac{\delta_N\,N\big(\tau_{j+1}^*-\tau_j^*\big)}{a_N}}\Big(\tilde{\Theta}_j-\Theta_j^*\Big)\underset{N\to\infty}{\overset{\mathcal{L}}{\longrightarrow}}\mathcal{N}\big(0,\Sigma^{(j)}(\alpha_j^*,r_1,\ldots,r_\ell)\big)\tag{5.12}$$

with $\Sigma^{(j)}(\alpha_j^*,r_1,\ldots,r_\ell):=(L_1'\cdot L_1)^{-1}L_1'\cdot\Gamma^{(j)}(\alpha_j^*,r_1,\ldots,r_\ell)\cdot L_1\cdot(L_1'\cdot L_1)^{-1}$.

A second estimator of Θ_j^* can be obtained with feasible generalized least squares (FGLS) technique. Indeed, the asymptotic covariance matrix $\Gamma^{(j)}(\alpha_j^*,r_1,\ldots,r_\ell)$ can be estimated with the matrix $\tilde{\Gamma}^{(j)}:=\Gamma^{(j)}(\tilde{\alpha}_j,r_1,\ldots,r_\ell)$ and $\tilde{\Gamma}^{(j)}\underset{N\to\infty}{\overset{\mathcal{P}}{\longrightarrow}}\Gamma^{(j)}(\alpha_j^*,r_1,\ldots,r_\ell)$, because $\alpha\mapsto\Gamma^{(j)}(\alpha,r_1,\ldots,r_\ell)$ is supposed to be a continuous function and $\tilde{\alpha}_j\underset{N\to\infty}{\overset{\mathcal{P}}{\longrightarrow}}\alpha_j^*$. Since also $\alpha\mapsto\Gamma^{(j)}(\alpha,r_1,\ldots,r_\ell)$ is supposed to be a positive matrix for all α then

$$\Big(\tilde{\Gamma}^{(j)}\Big)^{-1}\underset{N\to\infty}{\overset{\mathcal{P}}{\longrightarrow}}\Big(\Gamma^{(j)}(\alpha_j^*,r_1,\ldots,r_\ell)\Big)^{-1}.$$

Then, the FGLS estimator $\overline{\Theta}_j$ of Θ_j^* is defined from the minimization for all Θ of the following criterion

$$\|\,Y_{\tilde{k}_j}^{\tilde{k}'_j}-L_{a_N}\cdot\Theta\,\|^2_{\tilde{\Gamma}^{(j)}}=\big(Y_{\tilde{k}_j}^{\tilde{k}'_j}-L_{a_N}\cdot\Theta\big)'\cdot\big(\tilde{\Gamma}^{(j)}\big)^{-1}\cdot\big(Y_{\tilde{k}_j}^{\tilde{k}'_j}-L_{a_N}\cdot\Theta\big).$$

and therefore

$$\overline{\Theta}_j=\big(L_1'\cdot\big(\tilde{\Gamma}^{(j)}\big)^{-1}\cdot L_1\big)^{-1}\cdot L_1'\cdot\big(\tilde{\Gamma}^{(j)}\big)^{-1}\cdot Y_{\tilde{k}_j}^{\tilde{k}'_j}.$$

Proposition 5.2.3. *Under the same assumptions as in Theorem 5.2.2, for* $j = 0, \dots, m$

$$\sqrt{\frac{\delta_N\, N\big(\tau^*_{j+1} - \tau^*_j\big)}{a_N}}\Big(\overline{\Theta}_j - \Theta^*_j\Big) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}\big(0, M^{(j)}(\alpha^*_j, r_1, \dots, r_\ell)\big) \tag{5.13}$$

with $M^{(j)}(\alpha^*_j, r_1, \dots, r_\ell) := \big(L_1' \cdot \big(\Gamma^{(j)}(\alpha^*_j, r_1, \dots, r_\ell)\big)^{-1} \cdot L_1\big)^{-1} \leq \Sigma^{(j)}(\alpha^*_j, r_1, \dots, r_\ell)$ *(with order's relation between positive symmetric matrix).*

Therefore, the estimator $\overline{\Theta}_j$ converges asymptotically faster than $\tilde{\Theta}_j$; $\overline{\alpha}_j$ is more interesting than $\tilde{\alpha}_j$ for estimating α^*_j when N is large enough. Moreover, confidence intervals can be easily deduced for both the estimators of Θ^*_j.

5.2.4 Goodness-of-fit test

For $j = 0, \dots, m$, let $T^{(j)}$ be the FGLS distance between both the estimators of $L_{a_N} \cdot \Theta^*_j$, *i.e.* the FGLS distance between points $\Big(\log(r_i\, a_N), \log\big(S^{\tilde{k}'_j}_{\tilde{k}_j}\big)\Big)_{1\leq i\leq \ell}$ and the FGLS regression line. The following limit theorem can be established :

Theorem 5.2.4. *Under the same assumptions as in Theorem 5.2.1, for* $j = 0, \dots, m$

$$T^{(j)} = \frac{\delta_N\, N\big(\tau^*_{j+1} - \tau^*_j\big)}{a_N} \parallel Y^{\tilde{k}'_j}_{\tilde{k}_j} - L_{a_N} \cdot \overline{\Theta}_j \parallel^2_{\tilde{\Gamma}^{(j)}} \xrightarrow[N\to\infty]{\mathcal{L}} \chi^2(\ell - 2). \tag{5.14}$$

Mutatis mutandis, proofs of Proposition 5.2.3 and Theorem 5.2.4 are the same as the proof of Proposition 5 in (12). This test can be applied to each segment $[\tilde{k}_j, \tilde{k}'_j[$. However, under the assumptions, it is not possible to prove that a test based on the sum of $T^{(j)}$ for $j = 0, \dots, m$ converges to a $\chi^2\big((m+1)(\ell-2)\big)$ distribution (indeed, nothing is known about the eventual correlation of $\big(Y^{\tilde{k}'_j}_{\tilde{k}_j}\big)_{0\leq j\leq m}$).

5.2.5 Cases of polynomial trended processes

Wavelet based estimators are also known to be robust to smooth trends (see for instance (**?**)). More precisely, assume now that one considers the process $Y = \{Y_t, t \in T\}$ satisfying $Y_t = X_t + P(t)$ for all $t \in T$ where P is an unknown polynomial function of degree $p \in \mathbb{N}$. Then,

Corollary 5.2.1. *Under the same assumptions as in Theorem 5.2.1 for the process* X*, and if the mother wavelet* ψ *is such that* $\int t^r\, \psi(t)dt = 0$ *for* $r = 0, 1, \dots, p$*, then limit theorems (5.4), (5.5) and (5.6) hold for* X *and for* Y*.*

Let us remark that Lemarié-Meyer wavelet is such that $\int t^r \psi(t)dt = 0$ for all $r \in \mathbb{N}$. Therefore, even if the degree p is unknown, Corollary 5.2.1 can be applied. It is such the case for locally fractional Brownian motions and applications to heartbeat time series.

5.3 Applications

In this section, applications of the limit theorems to three kinds of piecewise Gaussian processes and HR data are studied. Several simulations for each kind of process are presented. In each case estimators $(\widehat{\tau}_j)_j$ and $(\tilde{\alpha}_j)_j$ are computed. To avoid an overload of results, FGLS estimators $(\overline{\alpha}_j)_j$ which are proved to be a little more accurate than $(\tilde{\alpha}_j)_j$ are only presented in one case (see Table 5.2) because the results for $(\overline{\alpha}_j)_j$ are very similar to $(\tilde{\alpha}_j)_j$ ones but are much more time consuming. For the choice of the number of scales ℓ, we have chosen a number proportional to the length of data (0.15 percent of N which seems to be optimal from numerical simulations) except in two cases (the case of goodness-of-fit test simulations for piecewise fractional Gaussian noise and the case of HR data, for which the length of data and the employed wavelet are too much time consuming).

5.3.1 Detection of change for Gaussian piecewise long memory processes

In the sequel the process X is supposed to be a piecewise long range dependence time series (and therefore $\delta_N = 1$ for all $N \in \mathbb{N}$). First, some notations have to be provided. For $Y = (Y_t)_{t\in\mathbb{N}}$ a Gaussian zero mean stationary process, with $r(t) = \mathbb{E}(Y_0 \cdot Y_t)$ for $t \in \mathbb{N}$, denote the spectral density f of Y by

$$f(\lambda) = \frac{1}{2\pi} \cdot \sum_{k\in\mathbb{Z}} r(k) \cdot e^{-ik\lambda} \quad \text{for } \lambda \in \mathbb{R}.$$

In the sequel, the spectral density of Y is supposed to satisfy the asymptotic property,

$$f(\lambda) \sim C \cdot \frac{1}{\lambda^D} \quad \text{when } \lambda \to 0,$$

with $C > 0$ and $D \in (0, 1)$. Then the process Y is said to be a long memory process and its Hurst parameter is $H = (1 + D)/2$. More precisely the following semi-parametric framework will be considered :

Assumption LRD(D) **:** Y is a zero mean stationary Gaussian process with spectral density satisfying

$$f(\lambda) = |\lambda|^{-D} \cdot f^*(\lambda) \quad \text{for all} \quad \lambda \in [-\pi, 0[\cup]0, \pi],$$

with $f^*(0) > 0$ and f^* is such that $|f^*(\lambda) - f^*(0)| \leq C_2 \cdot |\lambda|^2$ for all $\lambda \in [-\pi, \pi]$ with $C_2 > 0$.

Such assumption has been considered in numerous previous works concerning the estimation of the long range parameter in a semi-parametric framework (see for instance Robinson, 1995, Giraitis *et al.*, 1997, Moulines and Soulier, 2003). First and famous examples of processes satisfying Assumption LRD(D) are fractional Gaussian noises (FGN) constituted by the increments of the fractional Brownian motion process (FBM) and the fractionally autoregressive integrated moving average FARIMA$[p, d, q]$ (see more details and examples in Doukhan *et al.* (40)).

In this section, $X = (X_t)_{t\in\mathbb{N}}$ is supposed to be a Gaussian piecewise long-range dependent process, *i.e.*

- there exists a family $(D_j^*)_{0\leq j\leq m} \in (0,1)^{m+1}$;
- for all $j = 0, \ldots, m$, for all $k \in \left\{[N\tau_j^*], [N\tau_j^*]+1, \ldots, [N\tau_{j+1}^*]-1\right\}$, $X_k = X^{(j)}_{k-[N\tau_j^*]}$ and $X^{(j)} = (X_t^{(j)})_{t\in\mathbb{N}}$ satisfies Assumption LRD(D_j^*).

Several authors have studied the semi-parametric estimation of the parameter D using a wavelet analysis. This method has been numerically developed by Abry *et al.* (1998, 2003) and Veitch *et al.* (2004) and asymptotic results are provided in Bardet *et al.* (2000) and recently in Moulines *et al.* (2007) and Bardet *et al.* (2007). The following results have been developed in this last paper. The "mother" wavelet ψ is supposed to satisfy the following assumption : first ψ is included in a Sobolev space and secondly ψ satisfies the admissibility condition.

Assumption W_1 **:** $\psi : \mathbb{R} \mapsto \mathbb{R}$ with $[0,1]$-support with $\psi(0) = \psi(1) = 0$ and $\int_0^1 \psi(t)\, dt = 0$ and such that there exists sequence $(\psi_\ell)_{\ell\in\mathbb{Z}}$ such that
$\psi(\lambda) = \sum_{\ell\in\mathbb{Z}} \psi_\ell e^{2\pi i\ell\lambda} \in \mathbb{L}^2([0,1])$ and $\sum_{\ell\in\mathbb{Z}} (1+|\ell|)^{5/2}|\psi_\ell| < \infty$.

For ease of writing, ψ is supposed to be supported in $[0,1]$. By an easy extension the following propositions are still true for any compactly supported wavelets. For instance, ψ can be a dilated Daubechies "mother" wavelet of order d with $d \geq 6$ to ensure the smoothness of the function ψ. However, the following proposition could also be extended for "essentially" compactly supported "mother" wavelet like Lemarié-Meyer wavelet. Remark that it is not necessary to choose ψ being a "mother" wavelet associated to a multi-resolution analysis of $\mathbb{L}^2(\mathbb{R})$ like in the recent paper of Moulines *et al.*

(2007). The whole theory can be developed without resorting to this assumption. The choice of ψ is then very large. Then, in Bardet it et al. (2007), it was established :

Proposition 5.3.1. *Let X be a Gaussian piecewise long-range dependent process defined as above and $(a_n)_{n\in\mathbb{N}}$ be such that $N/a_N \underset{N\to\infty}{\longrightarrow} \infty$ and $a_N \cdot N^{-1/5} \underset{N\to\infty}{\longrightarrow} \infty$. Under Assumption W_1, limit theorems (5.4), (5.5) and (5.6) hold with $\alpha_j^* = D_j^*$ and $\beta_j^* = \log\left(f_j^*(0)\int_{-\infty}^{\infty}|\widehat{\psi}(u)|^2\cdot|u|^{-D}du\right)$ for all $j = 0, 1, \ldots, m$ and with $d_{pq} = GCD(r_p\,,\, r_q)$ for all $(p,q)\in\{1,\ldots,\ell\}$,*

$$\gamma_{pq}^{(j)} = \frac{2(r_p r_q)^{2-D_j^*}}{d_{pq}} \sum_{m=-\infty}^{\infty}\left(\frac{\int_0^\infty \widehat{\psi}(ur_p)\overline{\widehat{\psi}}(ur_q)\,u^{-D_j^*}\,\cos(u\,d_{pq}m)\,du}{\int_0^\infty|\widehat{\psi}(u)|^2\cdot|u|^{-D_j^*}du}\right)^2.$$

As a consequence, the results of Section 5.2 can be applied to Gaussian piecewise long-range dependent processes :

Corollary 5.3.2. *Under assumptions of Proposition 5.3.1 and Assumption C, for all $0<\kappa<2/15$, if $a_N = N^{\kappa+1/5}$ and $v_N = N^{2/5-3\kappa}$ then (5.11), (5.12), (5.13) and (5.14) hold.*

Thus, the rate of convergence of $\widehat{\underline{\tau}}$ to $\underline{\tau}^*$ (in probability) is $N^{2/5-3\kappa}$ for $0<\kappa$ as small as one wants. Estimators $\tilde{D}_j$ and $\overline{D}_j$ converge to the parameters D_j^* following a central limit theorem with a rate of convergence $N^{2/5-\kappa/2}$ for $0<\kappa$ as small as one wants.

Results of simulations : The following Table 5.1 represents the change point and parameter estimations in the case of a piecewise FGN with one abrupt change point. We observe the good consistence property of the estimators. Kolmogorov-Smirnov tests applied to the sample of estimated parameters lead to the following results :

1. the estimator $\widehat{\tau}_1$ can not be modeled with a Gaussian distribution ;
2. the estimator $\widehat{H}_j$ seems to follow a Gaussian distribution.

The distribution of the test statistics $T^{(0)}$ and $T^{(1)}$ (in this case $\ell = 20$ and $N = 20000$ and 50 realizations) are compared with a Chi-squared-distribution with eighteen degrees of freedom. The goodness-of-fit Kolmogorov-Smirnov test for $T^{(j)}$ to the $\chi^2(18)$-distribution is accepted (with $p = 0.3459$ for the sample of $T^{(0)}$ and $p = 0.2461$ for $T^{(1)}$). In this case and for the same parameters as in Table 5.1, the estimator $\overline{D}_j$ seems to be a little more accurate than $\tilde{D}_j$ (see Table 5.2).

Simulations are also applied to a piecewise simulated FARIMA(0,d_j,0) processes and

$N = 20000$, $\tau_1 = 0.75$, $D_0 = 0.2$ and $D_1 = 0.8$								
$\widehat{\tau}_1$	$\widehat{\sigma}_{\tau_1}$	$\sqrt{MSE}$	$\tilde{D}_0$	$\widehat{\sigma}_{D_0}$	$\sqrt{MSE}$	$\tilde{D}_1$	$\widehat{\sigma}_{D_1}$	$\sqrt{MSE}$
0.7605	0.0437	0.0450	0.2131	0.0513	0.0529	0.7884	0.0866	0.0874

TAB. 5.1 – Estimation of τ_1, D_0 and D_1 in the case of a piecewise FGN ($H_0 = 0.6$ and $H_1 = 0.9$) with one change point when $N = 20000$ and $\ell = 30$ (50 realizations)

$\widehat{\tau}_1$	$\widehat{\sigma}_{\tau_1}$	$\sqrt{MSE}$	$\overline{D}_0$	$\widehat{\sigma}_{D_0}$	$\sqrt{MSE}$	$\overline{D}_1$	$\widehat{\sigma}_{D_1}$	$\sqrt{MSE}$
0.7652	0.0492	0.0515	0.1815	0.0452	0.0488	0.8019	0.0721	0.0722

TAB. 5.2 – Estimation of D_0 and D_1 in the case of a piecewise FGN ($D_0 = 0.2$ and $D_1 = 0.8$) with one change point when $N = 20000$ and $\ell = 20$ (50 realizations)

results are similar (see Table 5.3). The following Figure 5.2 represents the change point instant and its estimation for such a process with one abrupt change point.

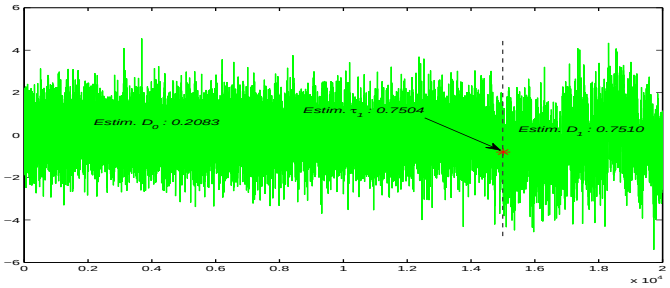


FIG. 5.2 – Detection of the change point in piecewise FARIMA(0,d_j,0) (for the first segment $d_0 = 0.1$ ($D_0 = 0.2$) for the second $d_1 = 0.4$ ($D_1 = 0.8$)). Estimated parameters $\tilde{D}_0 = 0.2083$, $\tilde{D}_1 = 0.7510$ and $\widehat{\tau}_1 = 0.7504$.

$N = 20000$, $\tau_1 = 0.75$, $D_0 = 0.2$ and $D_1 = 0.8$								
$\widehat{\tau}_1$	$\widehat{\sigma}_{\tau_1}$	$\sqrt{MSE}$	$\tilde{D}_0$	$\widehat{\sigma}_{D_0}$	$\sqrt{MSE}$	$\tilde{D}_1$	$\widehat{\sigma}_{D_1}$	$\sqrt{MSE}$
0.7540	0.0215	0.0218	0.1902	0.0489	0.0499	0.7926	0.0761	0.0764

TAB. 5.3 – Estimation of τ_1, D_0 and D_1 in the case of piecewise FARIMA(0,d_j,0) ($d_0 = 0.1$ and $d_1 = 0.4$) with one change point when $N = 20000$ and $\ell = 30$ (50 realizations)

5.3.2 Detection of abrupt change for piecewise Gaussian self-similar or locally fractional processes having stationary increments

In this sequel section the process X is supposed to be exactly or "nearly" a piecewise self-similar Gaussian process. Two cases are studied.

Case of a piecewise fractional Brownian motion

Let us recall that $B^H = (B_t^H)_{t\in\mathbb{R}}$ is a fractional Brownian motion (FBM) with two parameters $H \in (0,1)$ and $\sigma^2 > 0$ when B_H is a Gaussian process having stationary increments and such as

$$\mathrm{Var}(B_t^H)) = \sigma^2|t|^{2H} \quad \forall t \in \mathbb{R}.$$

It can be proved that B_H is the only Gaussian self-similar process having stationary increments and its self-similar parameter is H (a process $Y = (Y_t)_{t\in E}$ is said to be a H_s-self-similar process if for all $c > 0$ and for all $(t_1, \ldots, t_k) \in E^k$ where $k \in \mathbb{N}^*$, the vector $\big(Y_{ct_1}, \ldots, Y_{ct_k}\big)$ has the same distribution than the vector $c^{H_s}\big(Y_{t_1}, \ldots, Y_{t_k}\big)$).

Now, X will be called a piecewise fractional Brownian motion if :

- there exist two families of parameters $(H_j^*)_{0\le j\le m} \in (0,1)^{m+1}$ and $(\sigma_j^{*2})_{0\le j\le m} \in (0,\infty)^{m+1}$;
- for all $j = 0, \ldots, m$, for all $t \in \big[[N\tau_j^*], [N\tau_j^*]+1, \ldots, [N\tau_{j+1}^*]-1\big]$, $X_t = X^{(j)}_{t-[N\tau_j^*]}$ and $X^{(j)} = (X_t^{(j)})_{t\in\mathbb{R}}$ is a FBM with parameters H_j^* and σ_j^{*2}.

The wavelet analysis of FBM has been first studied by Flandrin (1992) and developed by Abry (1998) and Bardet (2002). Following this last paper, the mother wavelet ψ is supposed to satisfy :

Assumption W_2 : $\psi : \mathbb{R} \to \mathbb{R}$ is a piecewise continuous and left (or right)-differentiable in $[0,1]$, such that $|\psi'(t^-)|$ is Riemann integrable in $[0,1]$ with $\psi'(t^-)$ the left-derivative of ψ in t, with support included in $[0,1]$ and $\int_{\mathbb{R}} t^p\psi(t)\,dt = \int_0^1 t^p\psi(t)\,dt = 0$ for $p = 0, 1$.

As in Assumption W_1, ψ is supposed to be supported in $[0,1]$ but the following propositions are still true for any compactly supported wavelets. Assumption W_2 is clearly weaker than Assumption W_1 concerning the regularity of the mother wavelet. For instance, ψ can be a Daubechies wavelet of order d with $d \ge 3$ (the Haar wavelet, *i.e.* $d = 2$, does not satisfy $\int_0^1 t\,\psi(t)\,dt = 0$). Another choice could be infinite support wa-

velets with compact effective support (it is such the case with Meyer or Mexican Hat wavelets) but the proof of the following property has to be completed.

Proposition 5.3.3. *Assume that X is a piecewise FBM as it is defined above and let $(X_1, X_2, \ldots, X_N)$ be a sample of a path of X (therefore $\delta_N = 1$). Under Assumption W_2, if $(a_n)_{n\in\mathbb{N}}$ is such that $N/a_N \underset{N\to\infty}{\longrightarrow} \infty$ and $a_N \cdot N^{-1/3} \underset{N\to\infty}{\longrightarrow} \infty$, then limit theorems (5.4), (5.5) and (5.6) hold with $\alpha_j^* = 2H_j^* + 1$ and $\beta_j^* = \log\left(-\frac{\sigma_j^{*2}}{2}\int_0^1\int_0^1 \psi(t)\psi(t')|t-t'|^{2H_j^*}\,dt\,dt'\right)$ for all $j = 0, 1, \ldots, m$ and with $d_{pq} = GCD(r_p\,,\, r_q)$ for all $(p,q) \in \{1,\ldots,\ell\}$,*

$$\gamma_{pq}^{(j)} = \frac{2d_{pq}}{r_p^{2H_j^*+1/2}\, r_q^{2H_j^*+1/2}} \sum_{k=-\infty}^{\infty} \left(\frac{\int_0^1\int_0^1 \psi(t)\psi(t')\,|k\,d_{pq} + r_p t - r_q t'|^{2H_j^*}\,dt\,dt'}{\int_0^1\int_0^1 \psi(t)\psi(t')|t-t'|^{2H_j^*}\,dt\,dt'} \right)^2 .$$

Then, Theorem 5.2.1 can be applied to piecewise FBM but $2(a'-a)+1 = 2(\sup_j \alpha_j^* - \inf_j \alpha_j^*) + 1$ has to be smaller than 3 since $a_N \cdot N^{-1/3} \underset{N\to\infty}{\longrightarrow} \infty$. Thus,

Corollary 5.3.4. *Let $A := \left|\sup_j H_j^* - \inf_j H_j^*\right|$. If $A < 1/2$, under assumptions of Proposition 5.3.3 and Assumption C, for all $0 < \kappa < \frac{1}{1+4A} - \frac{1}{3}$, if $a_N = N^{1/3+\kappa}$ and $v_N = N^{2/3(1-2A)-\kappa(2+4A)}$ then (5.11), (5.12), (5.13) and (5.14) hold.*

Thus, the rate of convergence of $\widehat{\underline{\tau}}$ to $\underline{\tau}^*$ (in probability) can be $N^{2/3(1-2A)-\kappa'}$ for $0 < \kappa'$ as small as one wants when $a_N = N^{1/3+\kappa'/(2+4A)}$.

Remark : This result of Corollary 5.3.4 is quite surprising : the smaller A, *i.e.* the smaller the differences between the parameters H_j, the faster the convergence rates of estimators $\widehat{\tau}_j$ to τ_j^*. And if the difference between two successive parameters H_j is too large, the estimators $\widehat{\tau}_j$ do not seem to converge. Following simulations in Table 5.5 will exhibit this paroxysm. This induces a limitation of the estimators' using especially for applying them to real data (for which a priori knowledge is not available about the values of H_j^*).

Estimators $\tilde{H}_j$ and $\overline{H}_j$ converge to the parameters H_j^* following a central limit theorem with a rate of convergence $N^{1/3-\kappa/2}$ for $0 < \kappa$ as small as one wants.

Results of simulations : The following Table 5.4 represent the change point and parameter estimations in the case of piecewise FBM with one abrupt change point. Estimators of the change points and parameters seem to converge since their mean square errors clearly decrease when we double the number of observations. For testing if the estimated parameters follow a Gaussian distribution, Kolmogorov-Smirnov goodness-of-fit tests (in the case with $N = 10000$ and 50 replications) are applied :

	$N = 5000$			$N = 10000$		
τ_1	$\widehat{\tau}_1$	$\widehat{\sigma}_{\tau_1}$	$\sqrt{MSE}$	$\widehat{\tau}_1$	$\widehat{\sigma}_{\tau_1}$	$\sqrt{MSE}$
0.4	0.4467	0.0701	0.0843	0.4368	0.0319	0.0487
H_0	$\tilde{H}_0$	$\widehat{\sigma}_{H_0}$	$\sqrt{MSE}$	$\tilde{H}_0$	$\widehat{\sigma}_{H_0}$	$\sqrt{MSE}$
0.4	0.3147	0.0404	0.0943	0.3761	0.0452	0.0511
H_1	$\tilde{H}_1$	$\widehat{\sigma}_{H_1}$	$\sqrt{MSE}$	$\tilde{H}_1$	$\widehat{\sigma}_{H_1}$	$\sqrt{MSE}$
0.8	0.7637	0.0534	0.0645	0.7928	0.0329	0.0337

TAB. 5.4 – Estimation of τ_1, H_0 and H_1 in the case of piecewise fBm (H_j) with one change point when N=5000 (100 realizations) and N=10000 (50 realizations)

1. this test for $\tilde{H}_0$ is accepted as well as for $\tilde{H}_1$ and the following Figure 5.3 represents the relating distribution.
2. this is not such the case for the change point estimator $\widehat{\tau}_1$ for which the hypothesis of a possible fit with a Gaussian distribution is rejected ($KS_{test} = 0.2409$) as showed in the Figure 5.4 below which represents the empirical distribution function with the correspondant Gaussian cumulative distribution function.

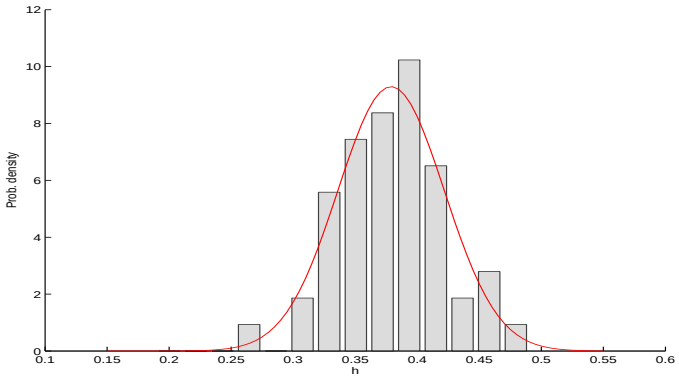


FIG. 5.3 – Modeling of $\widehat{H}_0$ sample estimations with normal distribution.

From the following example in Table 5.5, we remark that the estimated parameters seem to be non convergent when the difference between the parameters H_j is too large.

Simulations for goodness-of-fit tests $T^{(j)}$ provide the following results : when $N = 5000$, the drawn distributions of the computed test statistics (see Figure 5.5) exhibit a Khi-square distributed values ($\chi^2(5)$ since $\ell = 7$) and 95% of the 100 of the values of $T^{(0)}$ and

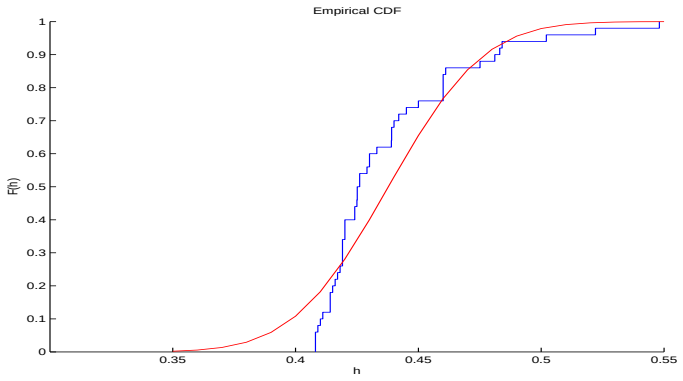


FIG. 5.4 – Comparison of the generated empirical cumulative distribution for $\widehat{\tau}_1$ (when N=10000) and the theoretical normal distribution.

$N = 5000$, $\tau_1 = 0.6$, $H_0 = 0.1$ and $H_1 = 0.9$								
$\widehat{\tau}_1$	$\widehat{\sigma}_{\tau_1}$	$\sqrt{MSE}$	$\tilde{H}_0$	$\widehat{\sigma}_{H_0}$	$\sqrt{MSE}$	$\tilde{H}_1$	$\widehat{\sigma}_{H_1}$	$\sqrt{MSE}$
0.5950	0.1866	0.1866	-0.1335	0.0226	0.2346	0.6268	0.4061	0.4894

TAB. 5.5 – Estimation of τ_1, H_0 and H_1 (when $H_1 - H_0 = 0.8 > 1/2$) in the case of piecewise FBM with one change point when $N = 5000$ (50 realizations)

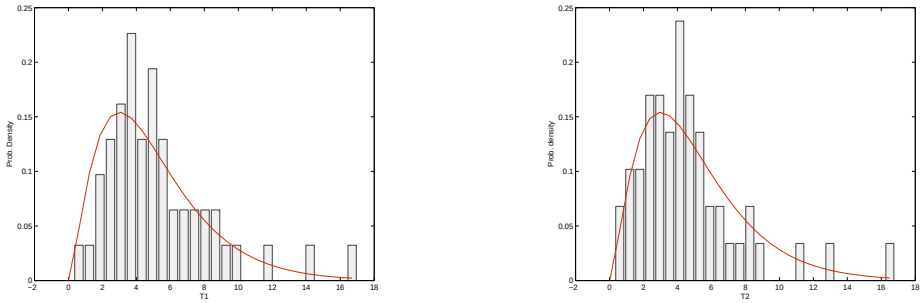


FIG. 5.5 – Testing for $\chi^2(5)$ distribution in the first detected zone (left) and the second detected zone (right) (50 realizations when $N = 5000$)

$T^{(1)}$ do not exceed $\chi^2_{95\%}(5) = 11.0705$. These results are also validated with Kolmogorov-Smirnov tests. when $N = 10000$, the drawn distributions of the computed test statistics (see Figure 5.6) exhibit a Khi-square distributed values ($\chi^2(13)$ since $\ell = 15$) and 95% of the 100 of the values of $T^{(0)}$ and $T^{(1)}$ do not exceed $\chi^2_{95\%}(13) = 22.3620$. These results are also validated with Kolmogorov-Smirnov tests.

The results below in Table 5.6 are obtained with piecewise fractional Brownian motion when two change points are considered. As previously, the KS_{test} tests for deciding whether or not the samples of estimated change points are consistent with a Gaussian

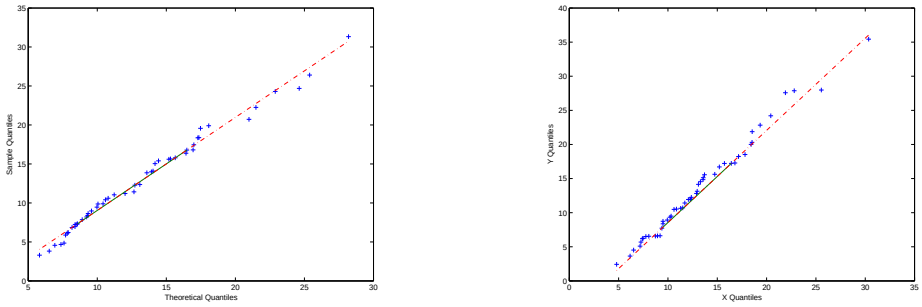


FIG. 5.6 – $\chi^2(13)$ QQ-plot for testing distribution in the first detected zone (left) and the second detected zone (right) (50 realizations when $N = 10000$)

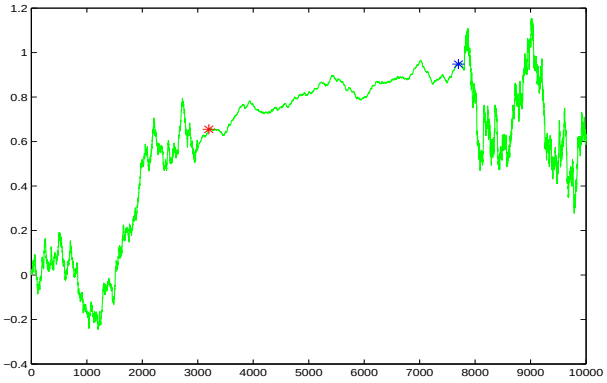
distributions are rejected. However, such a test is accepted for $\widehat{H}_j$ samples.

A graphical representation of the change point detection method applied to a piecewise FBM is given in Figure 5.7.

	$N = 5000$			$N = 10000$		
τ_1	$\widehat{\tau}_1$	$\widehat{\sigma}_{\tau_1}$	$\sqrt{MSE}$	$\widehat{\tau}_1$	$\widehat{\sigma}_{\tau_1}$	$\sqrt{MSE}$
0.3	0.3465	0.1212	0.1298	0.3086	0.0893	0.0897
τ_2	$\widehat{\tau}_2$	$\widehat{\sigma}_{\tau_2}$	$\sqrt{MSE}$	$\widehat{\tau}_2$	$\widehat{\sigma}_{\tau_2}$	$\sqrt{MSE}$
0.78	0.7942	0.1322	0.1330	0.7669	0.0675	0.0687
H_0	$\tilde{H}_0$	$\widehat{\sigma}_{H_0}$	$\sqrt{MSE}$	$\tilde{H}_0$	$\widehat{\sigma}_{H_0}$	$\sqrt{MSE}$
0.6	0.5578	0.0595	0.0730	0.5597	0.0449	0.0604
H_1	$\tilde{H}_1$	$\widehat{\sigma}_{H_1}$	$\sqrt{MSE}$	$\tilde{H}_1$	$\widehat{\sigma}_{H_1}$	$\sqrt{MSE}$
0.8	0.7272	0.0837	0.1110	0.7633	0.0813	0.0892
H_2	$\tilde{H}_2$	$\widehat{\sigma}_{H_2}$	$\sqrt{MSE}$	$\tilde{H}_2$	$\widehat{\sigma}_{H_2}$	$\sqrt{MSE}$
0.5	0.4395	0.0643	0.0883	0.4993	0.0780	0.0780

TAB. 5.6 – Estimation of τ_1, τ_2, H_0, H_1 and H_2 in the case of piecewise FBM with two change points when $N = 5000$ and $N = 10000$ (50 realizations)

The distribution of the test statistics $T^{(0)}$, $T^{(1)}$ and $T^{(2)}$ (in this case $\ell = 10$ and $N = 10000$ and 50 realizations) are compared with a Chi-squared-distribution with eight degrees of freedom. The goodness-of-fit Kolmogorov-Smirnov test for $T^{(j)}$ to the $\chi^2(8)$-distribution is accepted (with $p = 0.4073$ for the sample of $T^{(0)}$, $p = 0.2823$ for $T^{(1)}$ and $p = 0.0619$ for $T^{(2)}$).



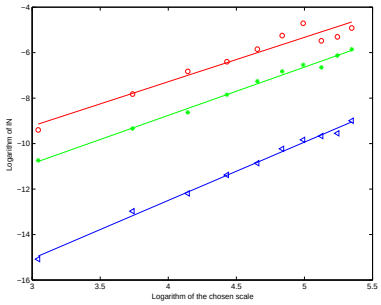


FIG. 5.7 – (left)Detection of the change point in piecewice FBM(H_j) ($\tau_1 = 0.3$, $\tau_2 = 0.78$, $H_0 = 0.6$, $H_1 = 0.8$ and $H_2 = 0.5$). The change points estimators are $\widehat{\tau}_1 = 0.32$ and $\widehat{\tau}_2 = 0.77$. (right) Representation of log-log regression of the variance of wavelet coefficients on the chosen scales for the three segments ($\tilde{H}_0 = 0.5608$ (*), $\tilde{H}_1 = 0.7814$ ($\lhd$) and $\tilde{H}_2 = 0.4751$ (o))

Case of a piecewise locally fractional Gaussian process

In this section, a continuous-time process X is supposed to model data. Therefore assume that $(X_{\delta_N}, X_{2\delta_N}, \ldots, X_{N\,\delta_N})$ is known, with $\delta_N \underset{N\to\infty}{\longrightarrow} 0$ and $N\,\delta_N \underset{N\to\infty}{\longrightarrow} \infty$. A piecewise locally fractional Gaussian process $X = (X_t)_{t\in\mathbb{R}_+}$ is defined by

$$X_t := \int_{\mathbb{R}} \frac{e^{it\xi} - 1}{\rho_j(\xi)} \widehat{W}(d\xi) \quad \text{for} \quad t \in [\tau_j^* N\,\delta_N\,,\, \tau_{j+1}^* N\,\delta_N) \tag{5.15}$$

where the functions $\rho_j : \mathbb{R} \to [0,\infty)$ are even Borelian functions such that for all $j = 0, 1, \ldots, m,$:

- $\rho_j(\xi) = \dfrac{1}{\sigma_j^*}\,|\xi|^{H_j^*+1/2}$ for $|\xi| \in [f_{min}\,,\, f_{max}]$ with $H_j^* \in \mathbb{R},\ \sigma_j^* > 0\,;$
- $\displaystyle\int_{\mathbb{R}} \left(1 \wedge |\xi|^2\right) \frac{1}{\rho_j^2(\xi)}\, d\xi < \infty$

and $W(dx)$ is a Brownian measure and $\widehat{W}(d\xi)$ its Fourier transform in the distribution meaning. Remark that parameters H_j^*, called local-fractality parameters, can be supposed to be included in $\mathbb{R}$ instead the usual interval $(0,1)$. Here $0 < f_{min} < f_{max}$ are supposed to be known parameters. Roughly speaking, a locally fractional Gaussian process is nearly a self-similar Gaussian process for scales (or frequencies) included in a band of scales (frequencies).

For locally fractional Gaussian process already studied in Bardet and Bertrand (2007) and Kammoun *et al.* (2007), the mother wavelet is supposed to satisfy

Assumption W_3 : $\psi : \mathbb{R} \mapsto \mathbb{R}$ is a $\mathcal{C}^\infty(\mathbb{R})$ function such that for all $m \in \mathbb{N}$, $\int_{\mathbb{R}} |t^m \psi(t)|\, dt < \infty$ and the Fourier transform $\widehat{\psi}$ of ψ is an even function compactly supported on $[-\mu, -\lambda] \cup [\lambda, \mu]$ with $0 < \lambda < \mu$.

These conditions are sufficiently mild and are satisfied in particular by the Lemarié-Meyer "mother" wavelet. The admissibility property, i.e. $\int_{\mathbb{R}} \psi(t)dt = 0$, is a consequence of the second one and more generally, for all $m \in \mathbb{N}$, $\int_{\mathbb{R}} t^m \psi(t)dt = 0$.

Since the function ψ is not a compactly supported mother wavelet, wavelet coefficients $d_X(a,b)$ can not be well approximated by $e_X(a,b)$ when the shift b is close to 0 or $N\,\delta_N$. Then, a restriction $\tilde{S}_k^{k'}(a_N)$ of sample wavelet coefficient's variance $S_k^{k'}(a_N)$ has to be defined :

$$\tilde{S}_k^{k'}(a_N) := \frac{a_N}{(1-2w)(k'-k)} \sum_{p=[(k+w(k'-k))/a_N]+w}^{[(k'-w(k'-k))/a_N]-1} e_X^2(a_N, a_N\, p) \quad \text{with } 0 < w < 1/2.$$

Proposition 5.3.5. *Assume that X is a piecewise locally fractional Gaussian process as it is defined above and $(X_{\delta_N}, X_{2\delta_N}, \ldots, X_{N\,\delta_N})$ is known, with $N(\delta_N)^2 \underset{N\to\infty}{\longrightarrow} 0$ and $N\,\delta_N \underset{N\to\infty}{\longrightarrow} \infty$. Under Assumptions W_3 and C, using $\tilde{S}_k^{k'}(a_N)$ instead of $S_k^{k'}(a_N)$, if $\frac{\mu}{\lambda} < \frac{f_{max}}{f_{min}}$ and $r_i = \frac{f_{min}}{\lambda} + \frac{i}{\ell}\big(\frac{f_{max}}{\mu} - \frac{f_{min}}{\lambda}\big)$ for $i = 1, \ldots, \ell$ with $a_N = 1$ for all $N \in \mathbb{N}$, then limit theorems (5.4), (5.5) and (5.6) hold with $\alpha_j^* = 2H_j^* + 1$ and $\beta_j^* = \log\big(-\frac{\sigma_j^{*2}}{2}\int_{\mathbb{R}} \big|\widehat{\psi}(u)\big|^2 |u|^{-1-2H_j^*} du\big)$ for all $j = 0, 1, \ldots, m$, for all $(p,q) \in \{1, \ldots, \ell\}$,*

$$\gamma_{pq}^{(j)} = \frac{2}{(1-2w)\,(r_p\, r_q)^{2H_j^*}} \int_{\mathbb{R}} \Big(\frac{\int_{\mathbb{R}} \overline{\widehat{\psi}}(r_p\,\xi)\widehat{\psi}(r_q\,\xi)\,|\xi|^{-1-2H_j^*} e^{-iu\xi} d\xi}{\int_{\mathbb{R}} \big|\widehat{\psi}(u)\big|^2 |u|^{-1-2H_j^*} du} \Big)^2 du. \tag{5.16}$$

Theorem 5.2.1 can be applied to a piecewise locally fractional Gaussian process without conditions on parameters H_j^*. Thus,

Corollary 5.3.6. *Under assumptions of Proposition 5.3.5 and Assumption C, then for all $0 < \kappa < \frac{1}{2}$, if $\delta_N = N^{-1/2-\kappa}$ and $v_N = N^{1/2-\kappa}$ then (5.11), (5.12), (5.13) and (5.14) hold.*

Therefore the convergence rate of $\widehat{\underline{\tau}}$ to $\underline{\tau}^*$ (in probability) is as well close to $N^{1/2}$ as one wants. Estimators $\tilde{H}_j$ and $\overline{H}_j$ converge to the parameters H_j^* following a central limit theorem with a rate of convergence $N^{1/4-\kappa/2}$ for $0 < \kappa$ as small as one wants.

5.3.3 Application to heart rate's time series

The study of regularity of physiological data and in particular the heartbeat signals have received much attention by several authors (see for instance (78), (79) or (5)). They studied HR series for healthy subjects and subjects with heart disease. In (57), a piecewise locally fractional Brownian motion is studied for modeling the cumulative HR data during three typical phases (estimated from Lavielle's algorithm) of the race (beginning, middle and end). The local-fractality parameters are estimated with wavelet analysis. The conclusions obtained are relatively close to those obtained by Peng. *et al.*. Indeed we remarked that the local-fractality parameter increases thought the race phases which may be explained with fatigue appearing during the last phase of the marathon. In this paper, we tray to unveil in which instants the behaviour of HR data changes. The following Table 5.7 presents the results for the detection of one change point.

	$\widehat{\tau}_1$	$\tilde{H}_0$	$\tilde{H}_1$	$T^{(0)}$	$T^{(1)}$
Ath1	0.0510	0.7880	1.2376	1.0184	1.0562
Ath2	0.4430	1.3470	1.4368	5.0644	1.5268
Ath3	0.6697	0.9542	1.2182	0.7836	0.9948
Ath4	0.4856	1.1883	1.2200	2.8966	1.2774
Ath5	0.8715	1.1512	1.3014	0.7838	0.8748
Ath6	0.5738	1.1333	1.1941	2.2042	0.7464
Ath7	0.3423	1.1905	1.1829	0.4120	1.5598
Ath8	0.8476	1.0222	1.2663	3.1704	0.5150
Ath9	0.7631	1.4388	1.3845	9.6574	0.5714

TAB. 5.7 – Estimated change points τ_1, parameters H_0, H_1 and goodness-of-fit test statistics ($T^{(0)}$ for the first zone and $T^{(1)}$ for the second) in the case of one change point observed in HR series of different athletes.

It is noticed that the estimator of the Hurst parameter is generally larger on the second zone than on the first although the detected change point differs from an athlete to another (only the case of Athlete 1 seems not to be relevant). This result is very interesting and confirms our conclusions in (57). Although this increase of Hurst parameter value was observed throughout time series and whatever is the position of change point, the estimation is larger in the second segment than in the first segment. Also the Hurst parameter value recorded in first zone increase in function of its size (see the example of HR data recorded for one athlete in Figure 5.8).

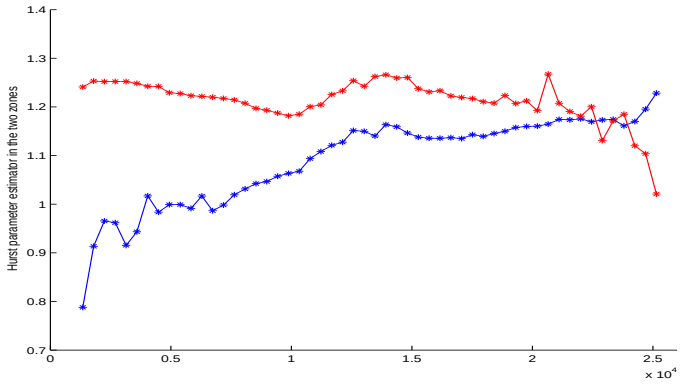


FIG. 5.8 – Evolution of Hurst parameter estimator (observed for HR series of one athlete) in the two zones when the change point varies in time (the curve of the first zone is usually under that of the second zone)

In general, the goodness-of-fit tests, with values $T^{(0)}$ and $T^{(1)}$, are less than $\chi^2_{95\%}(4) = 9.4877$ (except $T^{(0)}$ for Ath9) when $\ell = 6$. So, the HR data trajectory in the both zones seems to be correctly modeled with a stationary locally fractional Gaussian trajectory.

The detection of multiple change point could be more interesting but in this context and since the HR data are modeled by a piecewise locally fractional gaussian noise, simulations need a long time to run which is not easily realizable even for two change points.

5.4 Proofs

Before establishing the proof of Theorem 5.2.1 an important lemma can be stated :

Lemma 5.4.1. *Let $k \in \mathbb{N} \setminus \{0,1\}$, $(\gamma_i)_{1\leq i\leq k} \in (0,\infty)^k$ and $\alpha_1 > \alpha_2 > \cdots > \alpha_k$ be k ordered real numbers. For $(\alpha,\beta) \in \mathbb{R}^2$, consider the function $f_{\alpha,\beta} : x \in \mathbb{R} \mapsto \mathbb{R}$ such that*

$$f_{\alpha,\beta}(x) := \alpha\, x + \beta - \log\Big(\sum_{q=1}^{k} \gamma_q \exp\big(\alpha_q\, x\big)\Big) \quad \textit{for } x \in \mathbb{R}.$$

Let $0 < t_1 < \cdots < t_\ell$ with $\ell \in \mathbb{N} \setminus \{0,1,2\}$ and $(u_n)_{n\in\mathbb{N}}$ be a sequence of real numbers such that there exists $m \in \mathbb{R}$ satisfying $u_n \geq m$ for all $n \in \mathbb{N}$. Then there exists $C > 0$ not depending on n such that

$$\inf_{(\alpha,\beta)\in\mathbb{R}^2} \sum_{i=1}^{\ell} \big|f_{\alpha,\beta}\big(\log(u_n) + t_i\big)\big|^2 \geq C\, \min\big(1, |u_n|^{2(\alpha_2-\alpha_1)}\big).$$

Proof of Lemma 5.4.1 : For all $(\alpha,\beta)\in\mathbb{R}^2$, the function $f_{\alpha,\beta}$ is a $\mathcal{C}^\infty(\mathbb{R})$ function and

$$\frac{\partial^2}{\partial x^2} f_{\alpha,\beta}(x) = -\frac{\sum_{q=1}^{k-1}\gamma_q\gamma_{q+1}(\alpha_q-\alpha_{q+1})^2\exp\big((\alpha_q+\alpha_{q+1})\,x\big)}{\Big(\sum_{q=1}^{k}\gamma_q\exp\big(\alpha_q\,x\big)\Big)^2} < 0.$$

Therefore the function $f_{\alpha,\beta}$ is a concave function such that $\sup_{(\alpha,\beta)\in\mathbb{R}^2}\frac{\partial^2}{\partial x^2}f_{\alpha,\beta}(x)<0$ (not depending on α and β) and for all $(\alpha,\beta)\in\mathbb{R}^2$, $f_{\alpha,\beta}$ vanishes in 2 points at most. Thus, since $\ell\geq 3$ and $\big(x+t_i\big)_i$ are distinct points, for all $x\in\mathbb{R}$, it exists $C(x)>0$ not depending on α and β such that

$$\inf_{(\alpha,\beta)\in\mathbb{R}^2}\sum_{i=1}^{\ell}\big|f_{\alpha,\beta}\big(x+t_i\big)\big|^2\geq C(x).$$

Therefore, since for all $M\geq 0$,

$$\inf_{x\in[-M,M]}\left\{\inf_{(\alpha,\beta)\in\mathbb{R}^2}\sum_{i=1}^{\ell}\big|f_{\alpha,\beta}\big(x+t_i\big)\big|^2\right\}\geq\inf_{x\in[-M,M]}\big\{C(x)\big\}>0. \tag{5.17}$$

Moreover, if $u_n\to+\infty$,

$$\begin{aligned}\log\Big(\sum_{q=1}^{k}\gamma_q\exp\big(\alpha_q\,\log(u_n)\big)\Big)&\\ &=\log\Big(\gamma_1\exp\big(\alpha_1\,\log(u_n)\big)+\gamma_2\exp\big(\alpha_2\,\log(u_n)\big)\big(1+o(1)\big)\Big)\\ &=\log(\gamma_1)+\alpha_1\log(u_n)+\gamma_2\exp\big((\alpha_2-\alpha_1)\,\log(u_n)\big)\big(1+o(1)\big).\end{aligned}$$

Thus, for n large enough,

$$\frac{1}{2}\,\gamma_2\,u_n^{\alpha_2-\alpha_1}\leq\Big|\log\Big(\sum_{q=1}^{k}\gamma_q\exp\big(\alpha_q\,\log(u_n)\big)\Big)-\log(\gamma_1)+\alpha_1\log(u_n)\Big|\leq 2\,\gamma_2\,u_n^{\alpha_2-\alpha_1} \tag{5.18}$$

Therefore, for all $(\alpha,\beta)\in\mathbb{R}^2$,

$$\begin{aligned}\big|f_{\alpha,\beta}\big(\log(u_n)+t_i\big)\big|^2=\big|f_{\alpha_1,\log(\gamma_1)}\big(\log(u_n)+t_i\big)\big|^2+\Big|(\log(\gamma_1)-\beta)+(\alpha_1-\alpha)(\log(u_n)+t_i)\Big|^2&\\ -2f_{\alpha_1,\log(\gamma_1)}\big(\log(u_n)+t_i\big)\times\Big((\log(\gamma_1)-\beta)+(\alpha_1-\alpha)(\log(u_n)+t_i)\Big).&\end{aligned}$$

Using inequalities (5.18), $\frac{1}{4}\,\gamma_2^2\,u_n^{2(\alpha_2-\alpha_1)}\leq\big|f_{\alpha_1,\log(\gamma_1)}\big(\log(u_n)+t_i\big)\big|^2\leq 4\,\gamma_2^2\,u_n^{2(\alpha_2-\alpha_1)}$ and for all $(\alpha,\beta)\in\mathbb{R}^2$, $\lim_{n\to\infty}f_{\alpha_1,\log(\gamma_1)}\big(\log(u_n)+t_i\big)\times\Big((\log(\gamma_1)-\beta)+(\alpha_1-\alpha)(\log(u_n)+t_i)\Big)=$

0. Then, for all $(\alpha,\beta)\neq(\alpha_1,\log(\gamma_1))$, $\lim_{n\to\infty}\big|f_{\alpha,\beta}\big(\log(u_n)+t_i\big)\big|^2=\infty$. Consequently, for n large enough,

$$\begin{aligned}\inf_{(\alpha,\beta)\in\mathbb{R}^2}\sum_{i=1}^{\ell}\big|f_{\alpha,\beta}\big(\log(u_n)+t_i\big)\big|^2 &\geq \frac{1}{2}\sum_{i=1}^{\ell}\big|f_{\alpha_1,\log(\gamma_1)}\big(\log(u_n)+t_i\big)\big|^2\\ &\geq \frac{1}{8}\gamma_2^2\sum_{i=1}^{\ell}(u_n+t_i)^{2(\alpha_2-\alpha_1)}\\ &\geq C\,u_n^{2(\alpha_2-\alpha_1)},\end{aligned}$$

which combined with (5.17) achieves the proof. □

Proof of Theorem 5.2.1 : Let $w_N=\dfrac{N\delta_N}{v_N}$, $k_j^*=[N\delta_N\tau_j^*]$ for $j=1,\dots,m$ and

$$V_{\eta\,w_N}=\{(k_j)_{1\leq j\leq m},\ \max_{j\in1,\dots,m}|k_j-k_j^*|\geq\eta\,w_N\}.$$

Then, for $N\delta_N$ large enough,

$$\begin{aligned}&\mathbb{P}\Big(\frac{N\delta_N}{w_N}\,\|\underline{\tau}^*-\widehat{\underline{\tau}}\|_m\geq\eta\Big)\\ &\qquad\simeq\mathbb{P}\big(\max_{j\in1,\dots,m}|\widehat{k}_j-k_j^*|\geq\eta\,w_N\big)\\ &\qquad=\mathbb{P}\Big(\min_{(k_j)_{1\leq j\leq m}\in V_{\eta\,w_N}}G_N\big((k_j)_{1\leq j\leq m}\big)\leq\min_{(k_j)_{1\leq j\leq m}\notin V_{\eta\,w_N}}G_N\big((k_j)_{1\leq j\leq m}\big)\Big)\\ &\qquad\leq\mathbb{P}\Big(\min_{(k_j)_{1\leq j\leq m}\in V_{\eta\,w_N}}G_N\big((k_j)_{1\leq j\leq m}\big)\leq G_N\big((k_j^*)_{1\leq j\leq m}\big)\Big).\end{aligned}$$

For $j=\{0,\dots,m\}$ and $0=k_0<k_1<\ldots<k_m<k_{m+1}=N\delta_N$, let

- $Y_{k_j}^{k_{j+1}}:=\big(\log\big(S_{k_j}^{k_{j+1}}(r_i\cdot a_N)\big)\big)_{1\leq i\leq\ell}$,
- $\Theta_{k_j}^{k_{j+1}}=\begin{pmatrix}\alpha_j\\ \log\beta_j\end{pmatrix}$, $\widehat{\Theta}_{k_j}^{k_{j+1}}=\begin{pmatrix}\widehat{\alpha}_j\\ \log\widehat{\beta}_j\end{pmatrix}$ and $\Theta_j^*=\begin{pmatrix}\alpha_j^*\\ \log\beta_j^*\end{pmatrix}$.

1/ Using these notations, $G_N\big((k_j)_{1\leq j\leq m}\big)=\sum_{j=0}^{m}\|\,Y_{k_j}^{k_{j+1}}-L_{a_N}\cdot\widehat{\Theta}_{k_j}^{k_{j+1}}\,\|^2$, where $\|\cdot\|$

denotes the usual Euclidean norm in $\mathbb{R}^\ell$. Then, with I_ℓ the $(\ell \times \ell)$-identity matrix

$$\begin{aligned} G_N\big((k_j^*)_{1\le j\le m}\big) &= \sum_{j=0}^{m} \| \, Y_{k_j^*}^{k_{j+1}^*} - L_{a_N} \cdot \Theta_j^* \, \|^2 \\ &= \sum_{j=0}^{m} \Big\| (I_\ell - P_{L_{a_N}}) \cdot Y_{k_j^*}^{k_{j+1}^*} \Big\|^2 \quad \text{with } P_{L_{a_N}} = L_{a_N} \cdot (L'_{a_N} \cdot L_{a_N})^{-1} \cdot L'_{a_N} \\ &= \sum_{j=0}^{m} \frac{a_N}{k_{j+1}^* - k_j^*} \Big\| (I_\ell - P_{L_{a_N}}) \cdot \big(\varepsilon_i^{(N)}(k_j^*, k_{j+1}^*)\big)_{1\le i\le \ell} \Big\|^2 \quad \text{from (5.6)} \\ &\le \frac{1}{\min_{0\le j\le m}(\tau_{j+1}^* - \tau_j^*)} \cdot \frac{a_N}{N\,\delta_N} \sum_{j=0}^{m} \Big\| \big(\varepsilon_i^{(N)}(k_j^*, k_{j+1}^*)\big)_{1\le i\le \ell} \Big\|^2. \end{aligned}$$

Now, using the limit theorem (5.5), $\Big\| \big(\varepsilon_i^{(N)}(k_j^*, k_{j+1}^*)\big)_{1\le i\le \ell} \Big\|^2 \xrightarrow[N\to\infty]{\mathcal{L}} \Big\| \mathcal{N}(0, \Gamma(r_1, \ldots, r_\ell)) \Big\|^2$ since $k_{j+1}^* - k_j^* \sim N\delta_N(\tau_{i+1}^* - \tau_i^*) \underset{N\to\infty}{\longrightarrow} \infty$, and thus

$$G_N\big((k_j^*)_{1\le j\le m}\big) = O_P\Big(\frac{a_N}{N\,\delta_N}\Big), \tag{5.19}$$

where $\xi_N = O_P(\psi_N)$ as $N \to \infty$ is written, if for all $\rho > 0$, there exists $c > 0$, such as $P\Big(|\xi_N| \le c \cdot \psi_N\Big) \ge 1 - \rho$ for all sufficiently large N.

2/ Now, set $(k_j)_{1\le j\le m} \in V_{\eta\, w_N}$. Therefore, for N and $N\delta_N$ large enough, there exists $j_0 \in \{1, \ldots, m\}$ and $(j_1, j_2) \in \{1, \ldots, m\}^2$ with $j_1 \le j_2$ such that $k_{j_0} \le k_{j_1}^* - \eta\, w_N$ and $k_{j_0+1} \ge k_{j_2}^* + \eta\, w_N$. Thus,

$$G_N\big((k_j)_{1\le j\le m}\big) \ge \big\| Y_{k_{j_0}}^{k_{j_0+1}} - L_{a_N} \widehat{\Theta}_{k_{j_0}}^{k_{j_0+1}} \big\|^2.$$

Let $\Omega^* := \big(\Omega_i^*\big)_{1\le i\le \ell}$ be the vector such that

$$\begin{aligned} \Omega_i^* \;:=\; & \frac{k_{j_1}^* - k_{j_0}}{k_{j_0+1} - k_{j_0}}\, \beta_{j_1-1}^* \exp\big(\alpha_{j_1-1}^* \log(r_i\, a_N)\big) + \\ + \; & \sum_{j=j_1}^{j_2-1} \frac{k_{j+1}^* - k_{j_1}^*}{k_{j_0+1} - k_{j_0}} \beta_j^* \exp\big(\alpha_j^* \log(r_i\, a_N)\big) + \frac{k_{j_0+1} - k_{j_2}^*}{k_{j_0+1} - k_{j_0}} \beta_{j_2}^* \exp\big(\alpha_{j_2}^* \log(r_i\, a_N)\big). \end{aligned}$$

Then,

$$G_N\big((k_j)_{1\le j\le m}\big) \ge \| \, Y_{k_{j_0}}^{k_{j_0+1}} - \big(\log \Omega_i^*\big)_{1\le i\le \ell} \, \|^2 + \| \, \big(\log \Omega_i^*\big)_{1\le i\le \ell} - L_{a_N} \cdot \widehat{\Theta}_{k_{j_0}}^{k_{j_0+1}} \, \|^2 + 2\, Q, \tag{5.20}$$

with $Q = \big(Y_{k_{j_0}}^{k_{j_0+1}} - \big(\log \Omega_i^*\big)_{1\le i\le \ell}\big)' \cdot \Big(\big(\log \Omega_i^*\big)_{1\le i\le \ell} - L_{a_N} \cdot \widehat{\Theta}_{k_{j_0}}^{k_{j_0+1}}\Big)$.

In the one hand, with $S_k^{k'}(\cdot)$ defined in (5.3),

$$S_{k_{j_0}}^{k_{j_0+1}}(r_i\, a_N)$$
$$= \frac{k^*_{j_1}-k_{j_0}}{k_{j_0+1}-k_{j_0}}\, S_{k_{j_0}}^{k^*_{j_1}}(r_i\, a_N) + \sum_{j=j_1}^{j_2-1} \frac{k^*_{j+1}-k^*_{j_1}}{k_{j_0+1}-k_{j_0}} S_{k^*_j}^{k^*_{j+1}}(r_i\, a_N) + \frac{k_{j_0+1}-k^*_{j_2}}{k_{j_0+1}-k_{j_0}} S_{k_{j_2}}^{k^*_{j_0+1}}(r_i\, a_N).$$

Using the central limit theorems (5.6), for N and $N\delta_N$ large enough,

$$\begin{aligned}
\mathbb{E}\Big[\big(S_{k_{j_0}}^{k_{j_0+1}}(r_i\, a_N) - \Omega_i^*\big)^2\Big] \;\le\; & m\,\Bigg(\Big(\frac{k^*_{j_1}-k_{j_0}}{k_{j_0+1}-k_{j_0}}\Big)^2\, \mathbb{E}\Big[\big(S_{k_{j_0}}^{k^*_{j_1}}(r_i\, a_N) - \beta^*_{j_1-1}\big(r_i\, a_N\big)^{\alpha^*_{j_1-1}}\big)^2\Big] \\
& + \sum_{j=j_1}^{j_2-1} \Big(\frac{k^*_{j+1}-k^*_{j_1}}{k_{j_0+1}-k_{j_0}}\Big)^2\, \mathbb{E}\Big[\big(S_{k^*_j}^{k^*_{j+1}}(r_i\, a_N) - \beta^*_j\big(r_i\, a_N\big)^{\alpha^*_j}\big)^2\Big] \\
& + \Big(\frac{k_{j_0+1}-k^*_{j_2}}{k_{j_0+1}-k_{j_0}}\Big)^2\, \mathbb{E}\Big[\big(S_{k_{j_2}}^{k^*_{j_0+1}}(r_i\, a_N) - \beta^*_{j_2}\big(r_i\, a_N\big)^{\alpha^*_{j_2}}\big)^2\Big]\Bigg)
\end{aligned}$$

$$\begin{aligned}
\Longrightarrow \quad \mathbb{E}\Big[\big(\frac{S_{k_{j_0}}^{k_{j_0+1}}(r_i\, a_N)}{\Omega_i^*} - 1\big)^2\Big] \;&\le\; \frac{m\,\gamma^2}{\Omega_i^*}\, a_N \Big(\frac{1}{k^*_{j_1}-k_{j_0}} + \sum_{j=j_1}^{j_2-1} \frac{1}{k^*_{j+1}-k^*_{j_1}} + \frac{1}{k^*_{j_0+1}-k_{j_2^*}}\Big) \\
&\le\; C\,\frac{a_N}{\eta\, w_N},
\end{aligned}$$

with $\gamma^2 = \max_{i,j}\{\gamma_{ii}^{(j)}\}$ (where $(\gamma_{pq}^{(j)})$ is the asymptotic covariance of vector $\varepsilon_p^{(N)}(k,k')$ and $\varepsilon_q^{(N)}(k,k')$) and $C>0$ not depending on N. Therefore, for N large enough, for all $i=1,\dots,\ell$,

$$\mathbb{E}\Big[\big(\log(S_{k_{j_0}}^{k_{j_0+1}}(r_i\, a_N)) - \log(\Omega_i^*)\big)^2\Big] \le 2\,C\,\frac{a_N}{\eta\, w_N}.$$

Then we deduce with Markov Inequality that

$$\| \, Y_{k_{j_0}}^{k_{j_0+1}} - \big(\log \Omega_i^*\big)_{1\le i\le \ell} \, \|^2 = O_P\Big(\frac{a_N}{\eta\, w_N}\Big). \tag{5.21}$$

From the other hand,

$$\| \, \big(\log \Omega_i^*\big)_{1\le i\le \ell} - L_{a_N}\cdot \widehat{\Theta}_{k_{j_0}}^{k_{j_0+1}} \, \|^2 = \sum_{i=1}^{\ell} \Big(\big(\widehat{\alpha}_{j_0}\log(r_i\, a_N) + \log\widehat{\beta}_{j_0}\big) - \log\Omega_i^*\Big)^2.$$

Define $\gamma_1 := \dfrac{k^*_{j_1-1}-k_{j_0}}{k_{j_0+1}-k_{j_0}}\cdot\beta^*_{j_1-1}$, for all $p\in\{0,1,\dots,j_2-j_1-1\}$, $\gamma_p := \dfrac{k^*_{j_1+p}-k^*_{j_1+p-1}}{k_{j_0+1}-k_{j_0}}\cdot\beta^*_{j_1+p-1}$ and $\gamma_{j_2-j_1+1} := \dfrac{k_{j_0+1}-k^*_{j_2}}{k_{j_0+1}-k_{j_0}}\cdot\beta^*_{j_2}$. Then, using Lemma 5.4.1, one obtains

$$\inf_{\alpha,\beta}\Big\{\sum_{i=1}^{\ell}\Big(\big(\alpha\log(r_i\, a_N)+\log\beta\big) - \log\Omega_i^*\Big)^2\Big\} \ge C\,\min\big(1,,\, |a_N|^{2(\alpha^*_{(2)}-\alpha^*_{(1)})}\big),$$

where $C > 0$ and $\alpha^*_{(1)} = \max_{j=j_1-1,\dots,j_2} \alpha^*_j$, $\alpha^*_{(2)} = \max_{j=j_1-1,\dots,j_2,\ j\neq(1)} \alpha^*_j$. As a consequence, for satisfying all possible cases of j_0, j_1 and j_2, one obtains

$$\| \left(\log \Omega^*_i\right)_{1\leq i\leq \ell} - L_{a_N} \cdot \widehat{\Theta}^{k_{j_0+1}}_{k_{j_0}} \|^2 \geq C\, |a_N|^{2(\min_i \alpha^*_i - \max_i \alpha^*_i)}. \tag{5.22}$$

Finally, using Cauchy-Schwarz Inequality,

$$Q \ \leq \ \left(\left\| Y^{k_{j_0+1}}_{k_{j_0}} - \left(\log \Omega^*_i\right)_{1\leq i\leq \ell} \right\|^2 \cdot \left\| \left(\log \Omega^*_i\right)_{1\leq i\leq \ell} - L_{a_N} \cdot \widehat{\Theta}^{k_{j_0+1}}_{k_{j_0}} \right|^2 \right)^{1/2}$$

Therefore, using (5.21) and (5.22), since under assumptions of Theorem 5.2.1,

$$\frac{a_N}{\eta\, w_N} = o\Big(|a_N|^{2(\min_i \alpha^*_i - \max_i \alpha^*_i)} \Big),$$

then

$$Q = o_P\Big(\| \left(\log \Omega^*_i\right)_{1\leq i\leq \ell} - L_{a_N} \cdot \widehat{\Theta}^{k_{j_0+1}}_{k_{j_0}} \|^2 \Big). \tag{5.23}$$

We deduce from relations (5.20), (5.21), (5.22) and (5.23) that

$$\mathbb{P}\Big(\min_{(k_j)_{1\leq j\leq m} \in V_{\eta\, w_N}} G_N\big((k_j)_{1\leq j\leq m}\big) \geq \frac{C}{2}\, |a_N|^{2(\min_i \alpha^*_i - \max_i \alpha^*_i)} \Big) \underset{N\to\infty}{\longrightarrow} 1.$$

□

Proof of Theorem 5.2.2 : From Theorem 5.2.1, it is clear that

$$\mathbb{P}\big([\tilde{k}_j, \tilde{k}'_j] \subset [k^*_j, k^*_{j+1}]\big) \underset{N\to\infty}{\longrightarrow} 1 \quad \text{and} \quad \frac{\tilde{k}'_j - \tilde{k}_j}{N\delta_N(\tau^*_{j+1} - \tau^*_j)} \underset{N\to\infty}{\overset{\mathcal{P}}{\longrightarrow}} 1. \tag{5.24}$$

Now, for $j = 0, \dots, m$, $(x_i)_{1\leq i\leq \ell} \in \mathbb{R}^\ell$ and $0 < \varepsilon < 1$, let A_j and B_j be the events such that

$$A_j := \Big\{ [\tilde{k}_j, \tilde{k}'_j] \subset [k^*_j, k^*_{j+1}] \Big\} \bigcap \left\{ \left| \frac{\tilde{k}'_j - \tilde{k}_j}{N\delta_N(\tau^*_{j+1} - \tau^*_j)} - 1 \right| \leq \varepsilon \right\}$$

$$\text{and} \quad B_j := \left\{ \sqrt{\frac{\tilde{k}'_j - \tilde{k}_j}{a_N}} \Big(Y^{\tilde{k}'_j}_{\tilde{k}_j} - L_{a_N} \cdot \Theta^*_j \Big) \in \prod_{i=1}^{\ell} (-\infty, x_i] \right\}$$

First, it is obvious that

$$\mathbb{P}(A_j)\mathbb{P}(B_j \mid A_j) \leq \mathbb{P}(B_j) \leq \mathbb{P}(B_j \mid A_j) + 1 - \mathbb{P}(A_j). \tag{5.25}$$

Moreover, from (5.4),

$$\begin{aligned} \mathbb{P}(B_j \mid A_j) \quad &= \quad \mathbb{P}\Big(\big(\varepsilon^{(N)}_i(\tilde{k}_j, \tilde{k}'_j)\big)_{1\leq i\leq \ell} \in \prod_{i=1}^{\ell} (-\infty, x_i] \mid A_j \Big) \\ &\underset{N\to\infty}{\longrightarrow} \quad \mathbb{P}\Big(\mathcal{N}\big(0, \Gamma^{(j)}(\alpha^*_j, r_1, \dots, r_\ell)\big) \in \prod_{i=1}^{\ell} (-\infty, x_i] \Big). \end{aligned}$$

Using (5.24), it is straightforward that $\mathbb{P}(A_j) \underset{N\to\infty}{\longrightarrow} 1$. Consequently,

$$\mathbb{P}(B_j) \underset{N\to\infty}{\longrightarrow} \mathbb{P}\Big(\mathcal{N}\big(0, \Gamma^{(j)}(\alpha_j^*, r_1, \ldots, r_\ell)\big) \in \prod_{i=1}^{\ell} (-\infty, x_i]\Big)$$

and therefore $\sqrt{\frac{\tilde{k}'_j - \tilde{k}_j}{a_N}}\Big(Y_{\tilde{k}_j}^{\tilde{k}'_j} - L_{a_N} \cdot \Theta_j^*\Big) \underset{N\to\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}\big(0, \Gamma^{(j)}(\alpha_j^*, r_1, \ldots, r_\ell)\big)$. Now using again (5.24) and Slutsky's Lemma one deduces

$$\sqrt{\frac{\delta_N\left(N(\tau_{j+1}^* - \tau_j^*)\right)}{a_N}}\Big(Y_{\tilde{k}_j}^{\tilde{k}'_j} - L_{a_N} \cdot \Theta_j^*\Big) \underset{N\to\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}\big(0, \Gamma^{(j)}(\alpha_j^*, r_1, \ldots, r_\ell)\big).$$

Using the expression of $\tilde{\Theta}_j$ as a linear application of $Y_{\tilde{k}_j}^{\tilde{k}'_j}$, this achieves the proof of Theorem 5.2.2. □

Chapitre 6

Conclusion et perspectives

Cette étude nous a permis de mieux comprendre le comportement des signaux relatifs aux fréquences cardiaques enregistrées chez des athlètes au cours d'un effort d'endurance. Elle nous a permis de modéliser les données dans les différentes phases observées du marathon et de construire une méthode de détection des points de changements d'un paramètre caractéristique des fluctuations physiologiques.

Dans un premier temps, nous avons commencé par décrire les propriétés asymptotiques de la fonction DFA et de l'estimateur du paramètre de longue mémoire obtenu par la méthode DFA, méthode qui est largement utilisée pour l'étude des signaux physiologiques. Cette analyse nous a permis de démontrer un théorème de la limite centrale justifiant la régression log-log utilisée pour estimer l'exposant de Hurst H.

Dans le cadre semi-paramétrique d'un processus stationnaire à longue mémoire nous avons montré la convergence du paramètre avec une vitesse de convergence raisonnable (mais un peu moins bonne que celle obtenue avec des méthodes de log-périodogramme ou d'analyse par ondelettes). Toutefois, dans de nombreux cas d'existence d'une tendance, cet estimateur, qui ne prend en compte que celle-ci, n'est pas convergent. La méthode DFA n'est donc pas robuste et ne doit pas être appliquée dans le cas de processus à tendances.

La méthode d'estimation par analyse par ondelettes s'avère être plus robuste et fournit un estimateur du paramètre de Hurst plus performant surtout dans le cas particulier de processus longue mémoire à tendance polynomiale (des résultats peuvent également être obtenus pour des processus autosimilaires). En effet, Abry *et al.* (1998) ont prouvé que ce type de tendance est sans effet sur l'estimateur du paramètre Hurst dès que l'ondelette mère choisie présente un nombre approprié de moments nuls. Dans le cadre

d'une classe semi-paramétrique de processus gaussiens stationnaires à longue mémoire, il a été établi par Moulines *et al.* (2007) que l'estimateur du paramètre Hurst converge avec une vitesse de convergence optimale (suivant le critère minimax) pour une taille de fenêtre optimale choisie.

Cette méthode basée sur les ondelettes permet aussi la construction d'un test d'ajustement qui permet de valider la modélisation des données par un processus semi-paramétrique qui est le bruit gaussien localement fractionnaire. Ainsi, sur une bande de fréquences choisie, on obtient des conclusions relativement proches de celles obtenues par d'autres études (conclusions qui ne peuvent être décelées avec la méthode DFA). On remarque une évolution du paramètre de fractalité locale au cours des différentes phases de la course, ce qui peut s'expliquer par la fatigue qui apparaît en dernière phase du marathon. Ainsi ce paramètre s'avère être un facteur pertinent pour la détection de changements de comportement au cours d'une course d'endurance.

Une détection de rupture selon ce paramètre semble alors intéressante à effectuer. On a ainsi développé un estimateur des m points de changements du paramètre de longue mémoire, d'autosimilarité ou de fractalité locale pour des processus gaussiens. Le principe de cet estimateur est le suivant : dans chaque zone où le paramètre est constant, il est estimé à partir de la log-log régression de la variance des coefficients d'ondelettes par rapport aux différentes échelles choisies. Une fonction de contraste définie par la somme des carrés des distances entre ces points et les droites d'ajustement, dans les $m+1$ zones possibles détectées, est minimisée, donnant un estimateur des points de ruptures. Sous certaines hypothèses, on démontre un théorème limite vérifé par cet estimateur avec une vitesse de convergence explicite. Dans chaque zone détectée, on établit également un théorème de la limite centrale pour les estimateurs des paramètres. Enfin, on construit un test d'ajustement à partir d'une distance entre les points (d'abscisse, le logarithme une échelle choisie et d'ordonnée, le logarithme de la variance empirique des coefficients d'ondelettes pour cette échelle) et les droites de regression généralisée correspondantes. Des simulations ont été effectuées dans le cas de processus à longue mémoire et de processus autosimilaires, montrent l'intérêt de la méthode. Cette méthode est aussi appliquée aux données des fréquences cardiaques, que l'on modélise par un bruit gaussien localement fractionnaire. Les résulats confirment ceux obtenus précédemment notamment l'augmentation de la valeur du paramètre au cours de la course ; en revanche, les points de changements varient d'un athlète à un autre.

Cependant, dans cette approche, le nombre de points de changements m du paramètre H est supposé fixé par avance et dans le cas, par exemple, de notre application sur les données physiologiques, on aurait aimé savoir le nombre de régimes réellement existant durant la course. Donc, une extension intéressante de cette problématique pourrait être

l'estimation du nombre de ruptures inconnu, et l'adaptation de notre méthode en conséquence. On pourrait notament utiliser une fonction de contraste pénalisée par un terme qui augmente avec le nombre de ruptures et qui peut être spécifié dans ce contexte.
On pourrait également essayer de détecter des points de changements séquentiels (ou instantanés) en ne disposant que des observations passées pour détecter l'arrivée d'un éventuel point de rupture. Dans le cas des données physiologiques, ceci serait d'autant plus intéressant que cela permettrait de décider d'un état critique de fatigue, qui pourrait indiquer au coureur la nécessité de ralentir. Ceci peut se traduire mathématiquement en posant un seuil limitant un certain critère...

D'un point de vue mathématique, il pourait être également intéressant d'étendre l'ensemble des modèles et résultats obtenus pour des processus gaussiens, à des processus non gaussiens.

Comme nous l'avons indiqué au début de ce document, en plus des FC, d'autres variables physiologiques sont mesurées. Ces variables évoluent de manière dépendante les unes des autres. Il est donc très intéressant de voir si les changements de comportement de l'athlète par rapport aux autres variables surviennent aux mêmes moments de la courses. Ils permetteront d'expliquer les ruptures constatées et l'évolution enregistrée du paramètre caractéristique de la régularité au niveau des fréquences cardiaques.

Bibliographie

[1] Abry P., Flandrin P., Taqqu M.S., Veitch D., *Self-similarity and long-range dependence through the wavelet lens,* In P. Doukhan, G. Oppenheim and M.S. Taqqu editors, *Long-range Dependence : Theory and Applications*, Birkhäuser, 2003.

[2] Abry P., Flandrin P., Taqqu M., Veitch D., *Wavelets for the analysis, estimation and synthesis of scaling data*, Self-similar Network Traffic and Performance Evaluation (K. Park and W. Willinger, eds.), 39-87, Wiley, New York, 2000.

[3] Abry P., Veitch D., Flandrin P., *Long-range dependence : revisiting aggregation with wavelets.*, J. Time Ser. Anal. 19, no. 3, 253-266, 1998.

[4] Ayache A., Bertrand P., Lévy Véhel J., *A central limit theorem for the quadratic variations of the step fractional Brownian motion*, Statistical Inference for Stochastic Processes 10, 1-27, 2007.

[5] Absil P.A., Sepulchre R., Bilge A., Gérard P., *Nonlinear analysis of cardiac rhythm fluctuations using DFA method*, J. Physica A : Statistical mechanics and its applications, 235-244, 1999.

[6] Arcones M., *Limit theorems for nonlinear functionals of a stationary Gaussian sequence of vectors.*, Ann. Probab. 22, 2242-2274, 1994.

[7] Bagger M., Petersen P.H., Pedersen P.K., *Biological Variation in variables associated with exercise training*, Int. J. Sports Med, 433-440, 2003.

[8] Bai J. *Least squares estimation of a shift in linear processes*, J. of Time Series Anal. 5, 453-472, 1998.

[9] Bai J. and Perron P. *Estimating and testing linear models with multiple structural changes*, Econometrica 66, 47-78, 1998.

[10] Bardet J.M., *Statistical study of the wavelet analysis of fractional Brownian motion*, IEEE Trans. Inform. Theory. 48, 991-999, 2002.

[11] Bardet J.M., *Testing for the presence of self-similarity of Gaussian time series having stationary increments*, J. of Time Series Anal. 21, 497-516, 2000.

[12] Bardet J.M., Bertrand P., *Identification of the multiscale fractional Brownian motion with biomechanical applications*, Journal of Time Series Analysis, 1-52, 2007.

[13] Bardet J.M., Bertrand P., *Definition, properties and wavelet analysis of the multiscale fractional Brownian motion*, Fractals, 15, 73-87, 2007.

[14] Bardet J.M., Bibi H., Jouini A. *Adaptive wavelet based estimator of the memory parameter for stationary Gaussian processes*, To appear in Bernoulli, 2007.

[15] Bardet J.M., Kammoun I., *Asymptotic Properties of the Detrended Fluctuation Analysis of Long Range Dependence Processes*, Accepted in IEEE Transactions and Information Theory, 2007.

[16] Bardet J.M., Kammoun I., *Detecting abrupt changes of the long-range dependence or the self-similarity of a Gaussian process*, Preprint HAL, 2007.

[17] Bardet J.M., Kammoun I., *Detecting changes in the fluctuations of a Gaussian process and an application to heartbeat time series*, Preprint HAL, 2007.

[18] Bardet J.M., Lang G., Moulines E. and Soulier P. *Wavelet estimator of long-range dependent processes*, Statistical Inference for Stochastic Processes 3, 85-99, 2000.

[19] Bardet J.M., Lang G., Oppenheim G., Philippe A., Taqqu M.S. *Generators of long-range dependent processes : A survey. In Long-range Dependence : Theory and Applications*, Birkhauser, 2002.

[20] Basseville M., Nikiforov N., *The Detection of Abrupt Changes - Theory and Applications*, Prentice-Hall : Information and System Sciences Series, 1993.

[21] Bassingthwaighte J.B., Liebovitch L.S., West B.J, *Fractal Physiology* , Oxford Univ. Press, New York, 1994.

[22] Benassi A., Bertrand P., Cohen S., Istas J., *Identification of the Hurst index of a Step Fractional Brownian Motion*, Statistical Inference for Stochastic Processes, Vol. 3, Issue 1/2, p.101-111, 2000.

[33] Benassi A., Deguy S., *Multi-scale fractional Brownian motion : definition and identification*, Preprint LAIC, 1999.

[24] Benassi A., Jaffard S., Roux D., *Elliptic Gaussian random processes*, Rev. Mathemàtica Iberoamericana, vol. 13, No 1, p. 19-90, 1997.

[25] Beran J., *Statistics for long memory processes*, Monographs on Statist. and Appl. Probab. 61. Chapman & Hall, 315, 1994.

[26] Beran J., Terrin N., *Testing for a change of the long-memory parameter*, Biometrika, 83, 627-638, 1996.

[27] Bertrand P., Bardet J.M., *Some generalization of fractional Brownian motion and Control*, Optimal Control and Partial Differential Equations, J.L. Menaldi, E. Rofman and A. Sulem editors, 221-230, IOS Press, 2001.

[28] Box G.E.P., Jenkins G.M., *Time Series Analysis, Forecasting and Control*, Holden Day, S. Francisco, 1970.

[29] Billat V. Véronique Billat's web site : http ://www.billat.net/.

[30] Billat V., Demarle A., Paiva M., Koralsztein J.P., *Effect of Training on the Physiological Factors of Performance in Elite Marathon Runners (Males and Females)*, Int J Sports Med., 23(5) :336-41, 2002.

[31] Billat V., Demarle A., Slawinski J., Paiva M., Koralsztein J.P., *Physical and training characteristics of top-class marathon runners*, Med Sci Sports Exerc., 33(12) :2089-97, 2001.

[32] Billat V., Wesfreid E., Cottin F., Kapfer C., Koralztein J.P., *Fractal analysis of speed and physiological oscillations in long- and middle-distance running : Effect of training*, International Journal of Computer Science in Sport, 16-30, 2003.

[33] Brockwell P.J., Davis R.A., *Time series : Theory and methods*, Springer Series in Statistics, Second edition, 1990.

[34] Chen Z., Ivanov P.C., Hu K., Stanley H.E, *Effect of nonstationarities on detrended fluctuation analysis*, Physical Review E, Vol. 65, 041107, 2002.

[35] Cramér H., Leadbetter M.R., *Stationary and related stochastic processes. Sample function properties and their applications*, Wiley and Sons, 1967.

[36] Csörő M., Horváth L., *Limit Theorems in Change Point Analysis*, Wiley, 1997.

[37] Davydov Y.A., *The invariance principal for stationry processes*, Theory Probab. Appl. 15, 487-498, 1970.

[38] Delignières D., *L'analyse des processus stochastiques*, "Sport, Performance, Santé", EA 2991, Université Montpellier I, janvier 2001.

[39] Dobrushin R.L., Major P., *Non-central limit theorems for nonlinear functionals of Gaussian fields*, Z. Wahrsch. Verw. Gebiete 50, 27-52, 1979.

[40] Doukhan P., Openheim G., Taqqu M.S. (Editors), Theory and applications of long-range dependence, Birkhäuser, Boston, 2003.

[41] Flandrin P., *Wavelet analysis and synthesis of fractional Brownian motion*, IEEE Trans. on Inform. Theory 38, 910-917, 1992.

[42] Gao J.B., Cao Y., Lee J.M, *Principal component analysis of* $1/f^{\alpha}$ *noise*, Physics Letters, A 314, 392-400, 2003.

[43] Giraitis L., Leipus R., Surgailis D., *The change-point problem for dependent observations*, Journal of Statistical Planning and Inference, 53, 297-310, 1996.

[44] Giraitis L., Robinson, P. and Samarov, A., *Rate optimal semi-parametric estimation of the memory parameter of the Gaussian time series with long range dependence*, J. Time Ser. Anal., 18, 49-61, 1997.

[45] Giraitis L., Robinson, P. and Samarov, A., *Adaptive Semiparametric Estimation of the Memory Parameter*, Journal of Multivariate Analysis, Vol. 72, Issue 2, 183-207, 2000.

[46] Giraitis L., Surgailis D., *CLT and other limit theorems for functionals of Gaussian processes*, Z. Wahrsch. Verw. Gebiete 70, 191-212, 1985.

[48] Goldberger A.L., *Heartbeats, hormones, and health : is variability the spice of life ?*, Am J Respir Crit Care Med, 163, 1289-1290, 2001.

[48] Goldberger A.L., Amaral L.A.N., Hausdorff J.M., Ivanov P.C., Peng C.K., Stanley H.E., *Fractal Dynamics in Physiology : Alterations with Disease and Aging*, PNAS Vol. 99, No. 35, 2466-2472, 2002.

[49] Gourieroux C., Monfort A., *Séries temporelles et modèles dynamiques*, Ed. Economica, 1995.

[50] Guégan D., *How can we define the concept of long memory ?* An Econometric survey, Discussion Paper in Economics, Finance and Internationals Competitiveness, 178, 1-40, School of Economics and Finance, QUT, Brisbane, Australia, 2004.

[51] Ho H.C., Hsing T., *Limit theorems for functionals of moving averages*, The Annals of Probability, Vol. 25, No. 4, 1636-1669, 1997.

[52] Horváth L., *Change-Point Detection in Long-Memory Processes*, Journal of Multivariate Analysis, 78, 218-134, 2001.

[53] Horváth L., Shao Q.M., *Limit theorems for quadratic forms with applications to Whittle's estimate*, The Annals of Applied Probability, 9, 146-187, 1999.

[54] Hurst H.E., *Long-term storage capacity of reservoirs*, Transactions of the American Society of Civil Engineers, 770-808, 1951.

[55] Hu K., Ivanov P.C., Chen Z., Carpena P., Stanley H.E, *Effect of trends on detrended fluctuation analysis*, Physical Review E, Vol. 64, 011114, 2001.

[56] Hwa R.C., Ferree T.C.,*Fluctuation analysis of human electroencephalogram*, Nonlinear phenomena in complex systems, 302-307, 2002.

[57] Kammoun I., Billat V., Bardet J.M., *Comparison of DFA vs wavelet analysis for estimation of regularity of HR series during the marathon*, Preprint SAMOS, 2007.

[58] Kantelhardt J.W., Ashkenazy Y., Ivanov P.C., Bunde A., Havlin S., Penzel T., Peter J.H., Stanley H.E., *Characterization of sleep stages by correlations in the magnitude and sign of heartbeat increments*, Physical Review E, Vol. 65, 051908, 2002.

[59] Kantelhardt J.W., Koscielny-Bunde E., Rego H.A.H., Havlin S., Bunde A., *Detecting Long- range Correlations with Detrended Fluctuation Analysis*, Physica A, 295, 441-454, 2001.

[60] Karasik R., Sapir N., Ashkenazy Y., Ivanov P.C., Dvir I., Lavie P., Havlin S., *Correlation differences in heartbeat fluctuations during rest and exercise*, Physical Review E 66, 062902, 2002.

[61] Kokoszka P.S., Leipus R., *Detection and estimation of changes in regime*, In P. Doukhan, G. Oppenheim and M.S. Taqqu editors, *Long-range Dependence : Theory and Applications*, Birkhäuser, 325-337, 2003.

[62] Kolmogorov, *Wienersche Spiralen und einige andere interessante Kurven in Hilbertchen Raume*, Doklady, 26, 115-118, 1940.

[63] Lavielle M., *Detection of multiple changes in a sequence of random variables*, Stoch. Process Appl, 79-102, 1999. http ://www.math.u-psud.fr/ lavielle/.

[64] Lavielle M., Moulines E., *Least-squares estimation of an unknown number of shifts in a time series*, J. of Time Series Anal. 21, 33-59, 2000.

[65] Lavielle M., Teyssière G., *Adaptive detection of multiple change points in asset price volatility, dans : G. Teyssière et A. Kirman (Editeurs.)*, Long Memory in Economics, 129-156, Springer Verlag, Berlin, 2005.

[66] Lavielle M., Teyssière G. *Detecting Multiple Change-Points in Multivariate Time Series*, Lithuanian Mathematical Journal 46, 351-376, 2006.

[67] Lemoine M., Pelgrin F., *Introduction aux modèles espace-état et au filtre de Kalman*, Revue OFCE, 2003.

[68] Mandelbrot B.B., Taqqu M.S., *Robust R/S analysis of long-run serial correlation*, Bulletin of the International Statistical Institute, 48, 69-99, 1979.

[69] Mandelbrot B., Van Ness J.M., *Fractional Brownian motions, fractional noises and applications*, SIAM Review, vol. 10, 422-437, 1968.

[70] Mandelbrot B., Wallis J.R., *Computer experiments with fractional gaussian noises*, Water Resources Research, 5 : 228-267, 1969.

[71] Martinis M., Knezevic A., Krstacic G., Vargovic E., *Changes in the Hurst exponent of heartbeat intervals during physical activities*, Physics 0212029, 2002.

[72] Masugi M., *Detrended fluctuation analysis of IP-network traffic using a two-dimensional topology map*, Phys. A 337, no. 3-4, 664–678, 2004.

[73] Moulines E., Roueff F., Taqqu M.S., *On the Spectral Density of the Wavelet Coefficients of Long-Memory Time Series with Application to the Log-Regression Estimation of the Memory Parameter*, JTSA, Vol. 28, issue 2, 155-187, 2007.

[74] Moulines E., Soulier P., *Semiparametric spectral estimation for fractional processes*, In P. Doukhan, G. Openheim and M.S. Taqqu editors, *Theory and applications of long-range dependence*, 251-301, Birkhäuser, Boston, 2003.

[75] Nagarajan R., Kavasseri R.G, *Minimizing the effect of sinusoidal trends in detrended fluctuation analysis*, (in press) International Journal of Bifurcations and Chaos, Vol. 15, No. 2, 2005.

[76] Peng C.K., Buldyrev S.V., Havlin S., Simons M., Stanley H.E., Goldberger A.L., *Mosaic organization of DNA nucleotides*, Physical Review E, Vol. 49, 1685-1689, 1994.

[77] Peng C.K., Hausdorff J.M., Goldberger A.L., *Fractal mechanisms in neural control : Human heartbeat and gait dynamics in health and disease*. In : Walleczek J, ed. Nonlinear Dynamics, Self-Organization, and Biomedicine. Cambridge University Press, 1999.

[78] Peng C.K., Havlin S., Stanley H.E., Goldberger A.L., *Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series*, Chaos 5, 82, 1995.

[79] Peng C.K., Mietus J., Hausdorff J., Havlin S., Stanley H.E., Goldberger A.L., *Long-Range Anticorrelations and Non-Gaussian Behavior of the Heartbeat*, Phys. Rev. Lett. 70, 1343-1346, 1993.

[80] Pikkujämsä S.M., Mäkikallio T.H., Airaksinen K.E., Huikuri H.V. *Determinants and interindividual variation of R-R interval dynamics in healthy middle-aged subjects.* Am. J. Physiol. Heart Circ. Physiol. 280(3) : H1400-H1406, 2001.

[81] Robinson P.M.*Gaussian semiparametric estimation of long range dependence*, The Annals of statistics, 23 :1630-1661, 1995.

[82] Samorodnitsky G., Taqqu M.S., *Stable non-Gaussian Random Processes*, Chapman and Hall, 1994.

[83] Soulier P., *Moment bounds and central limit theorem for functions of Gaussian vectors*, Statist. Probab. Lett. 54, 193-203, 2001.

[84] Stoev S., Taqqu M.S., Park C., Marron J.S., *On the wavelet spectrum diagnostic for Hurst parameter estimation in the analysis of Internet traffic*, Elsevier, Vol. 49, 423-445, 2005.

[85] Taqqu M.S., Taqqu's Homepage : math.bu.edu/individual/murad/home.html.

[86] Taqqu M.S., *Weak Convergence to Fractional Brownian Motion and to Rosenblatt Process*, Zeit. Wahr. verw. Geb., 31, 287-302, 1975.

[87] Taqqu M.S., Teverovsky V., *Robustness of whittle type estimators for time series with long range dependence*, J. Stochastic Model, 13, 723-757, 1997.

[88] Taqqu M.S., Teverovsky V., Willinger W., *Estimators for long-range dependence : an empirical study*, Fractals, Vol. 3, No. 4, 785-788, 1995.

[89] Veitch D., Abry P., *A wavelet-based joint estimator of the parameters of long-range dependence.* IEEE Trans. Inform. Theory, Vol. 45, No. 3, 878–897, 1999.

[90] Wesfreid E., Billat V., Meyer Y.,*Multifractal analysis of heartbeat time series in human races*, Elsevier Appl. Comput. Harmon. Anal. 18, 329-335, 2005.

[91] Wood A., Chan G. *Simulation of stationary Gaussian processes in* $[0,1]^d$, Journal of Computational and Graphical statistics, 3, 409-432, 1994.

Table des figures

1.1 Evolution de la FC chez deux athlètes et les phases qui peuvent êtres observées durant l'effort 3

1.2 Représentation de la fonction d'autocorrélation empirique (ACF) pour la série des FC d'un athlète durant tout un marathon, et pour les données enregistrées seulement "en milieu" d'exercice 7

2.1 Heat rate signals of Athlete 1 in ms, Hertz and BPM (up), of Athletes 2, 3 and 4 in BPM (down) 21

2.2 Detection of the race beginning and end from HR data (in BPM) . . . 24

2.3 Plot of the increments of observed HR series for Ath1 (top) and Ath2 (bottom) 25

2.4 (a) Increments in HR time series after exponential smoothing (b) Increments in HR time series after Kalman smoothing (c) Histogram of increments after processing 27

2.5 The estimated configuration of changes in a HR time series of an athlete 28

2.6 Distribution of data during the race recorded for one athlete which seems to be gaussian 29

2.7 The self-similarity of the aggregated HR signals (representation of the aggregated HR fluctuations at 3 different time resolutions) 29

2.8 Comparison of HR data in the middle of race (Ath4) and generated FGN(H=0.99) trajectories 30

2.9 Generated FGN trajectories and corresponding aggregated series (FBM) for $H = 0.2 < 0.5$ anti-persistant noise (left), $H = 0.5$ white noise (center) and $H = 0.8 > 0.5$ LRD process (right) 31

3.1 The two first steps of the DFA method applied to a path of a discretized FGN (with $H = 0.6$ and $N = 10000$) 39

3.2 Results of the DFA method applied to a path of a discretized FGN for different values of $H = (0.2, 0.4, 0.5, 0.7, 0.8)$ (also with $N = 10000$) . . 39

3.3 Relation between $\log F_f(n_i)$ and $\log n_i$ in the case of power law trend . 45

3.4 Relation between $\log F_{Y+f}(n_i)$ and $\log n_i$ in the case of a power law trend ($N = 10000$, $H = 0.2$ ($\square$), $H = 0.4$ ($\bigcirc$), $H = 0.5$ ($\Diamond$), $H = 0.7$ ($*$) and $H = 0.8$ ($\cdot$)) 46

3.5 Relation between $\log F_f(n_i)$ and $\log n_i$ in the case of trend with change points 48

3.6 Relation between $\log F_{f+Y}(n_i)$ and $\log n_i$ in the case of a trend with change points ($N = 20000$, $H = 0.2$ ($\square$), $H = 0.4$ ($\bigcirc$), $H = 0.5$ ($\Diamond$), $H = 0.7$ ($*$) and $H = 0.8$ ($\cdot$)) 49

4.1 Results of the DFA method and wavelet analysis applied to a path of a discretized FGN for different values of $H = 0.2, 0.4, 0.5, 0.7, 0.8$, with $N = 10000$ 60

4.2 Two first steps of the DFA method applied to a HR series (up) and results of the DFA method applied to HR series for two different athletes (down) 61

4.3 The log-log graph of the variance of wavelet coefficients relating to the HR series observed during the race and in the end of race (Ath2) 62

4.4 The log-log graph of the variance of wavelet coefficients relating to the HR series observed during the arrival phase (Ath6) with a frequency band of [0.01 12](right) and of [0.2 4](left). 64

4.5 The log-log graph of the variance of wavelet coefficients relating to the HR series observed in the middle of the exercise (Ath5) 66

4.6 The results of the DFA method applied to records for race beginning (Ath3) (left) and for end of race (Ath1) (right) 68

4.7 Comparison of the three samples constituting by estimations in the beginning of race, during the race and then in the race end by the DFA and wavelet methods 68

5.1 Heat rate signals of Athlete 1 in ms, Hertz and BPM (up), of Athletes 2, 3 and 4 in BPM (down) 72

5.2 Detection of the change point in piecewise FARIMA(0,d_j,0) (for the first segment $d_0 = 0.1$ ($D_0 = 0.2$) for the second $d_1 = 0.4$ ($D_1 = 0.8$)). Estimated parameters $\tilde{D}_0 = 0.2083$, $\tilde{D}_1 = 0.7510$ and $\widehat{\tau}_1 = 0.7504$. . . . 83

5.3 Modeling of $\widehat{H}_0$ sample estimations with normal distribution. 86

5.4 Comparison of the generated empirical cumulative distribution for $\widehat{\tau}_1$ (when N=10000) and the theoretical normal distribution. 87

5.5 Testing for $\chi^2(5)$ distribution in the first detected zone (left) and the second detected zone (right) (50 realizations when $N = 5000$) 87

5.6 $\chi^2(13)$ QQ-plot for testing distribution in the first detected zone (left) and the second detected zone (right) (50 realizations when $N = 10000$) 88

5.7 (left)Detection of the change point in piecewice FBM(H_j) ($\tau_1 = 0.3$, $\tau_2 = 0.78$, $H_0 = 0.6$, $H_1 = 0.8$ and $H_2 = 0.5$). The change points estimators are $\widehat{\tau}_1 = 0.32$ and $\widehat{\tau}_2 = 0.77$. (right) Representation of log-log regression of the variance of wavelet coefficients on the chosen scales for the three segments ($\tilde{H}_0 = 0.5608$ (*), $\tilde{H}_1 = 0.7814$ ($\lhd$) and $\tilde{H}_2 = 0.4751$ (o)) 89

5.8 Evolution of Hurst parameter estimator (observed for HR series of one athlete) in the two zones when the change point varies in time (the curve of the first zone is usually under that of the second zone) 92

Liste des tableaux

4.1 Comparison of the two samples of estimations of H with 100 realizations of fGn path (N=10000) with DFA and wavelets methods 60

4.2 Estimated H with wavelets methods for HR series of different athletes . 62

4.3 Estimated $\widehat{H}$, with DFA and wavelets methods, for HR series of different athletes (*) The series for which the test is rejected. Comparison of the two samples $(\widehat{H}_{DFA})_{1,\dots,9}$ and $(\widehat{H}_{WAV})_{1,\dots,9}$ for whole and partial series (p-value). 67

5.1 Estimation of τ_1, D_0 and D_1 in the case of a piecewise FGN ($H_0 = 0.6$ and $H_1 = 0.9$) with one change point when $N = 20000$ and $\ell = 30$ (50 realizations) 83

5.2 Estimation of D_0 and D_1 in the case of a piecewise FGN ($D_0 = 0.2$ and $D_1 = 0.8$) with one change point when $N = 20000$ and $\ell = 20$ (50 realizations) 83

5.3 Estimation of τ_1, D_0 and D_1 in the case of piecewise FARIMA(0,d_j,0) ($d_0 = 0.1$ and $d_1 = 0.4$) with one change point when $N = 20000$ and $\ell = 30$ (50 realizations) 83

5.4 Estimation of τ_1, H_0 and H_1 in the case of piecewise fBm (H_j) with one change point when N=5000 (100 realizations) and N=10000 (50 realizations) 86

5.5 Estimation of τ_1, H_0 and H_1 (when $H_1 - H_0 = 0.8 > 1/2$) in the case of piecewise FBM with one change point when $N = 5000$ (50 realizations) 87

5.6 Estimation of τ_1, τ_2, H_0, H_1 and H_2 in the case of piecewise FBM with two change points when $N = 5000$ and $N = 10000$ (50 realizations) . . 88

5.7 Estimated change points τ_1, parameters H_0, H_1 and goodness-of-fit test statistics ($T^{(0)}$ for the first zone and $T^{(1)}$ for the second) in the case of one change point observed in HR series of different athletes. 91

Table des matières

Avant-Propos iii

Résumé v

Abstract vii

1 Introduction 1

1.1 Motivations . . . 1
1.2 Modélisation des signaux physiologiques . . . 4
1.3 Préambule mathématique . . . 5
1.3.1 La notion de stationnarité . . . 5
1.3.2 Processus à longue mémoire . . . 6
1.3.3 Propriété d'autosimilarité . . . 9
1.3.4 Généralisations gaussiennes du FBM . . . 10
1.3.5 Techniques d'estimation du paramètre de Hurst . . . 12
1.3.6 Détection de ruptures . . . 16
1.4 Organisation de la thèse . . . 18

2 Data processing and modeling 21

2.1 Introduction . . . 21
2.2 Data processing . . . 22
2.2.1 Abrupt change detection . . . 22
2.2.1.1 General principle of the method of change detection . . . 22
2.2.1.2 Change detection in mean and variance . . . 23
2.2.2 Data smoothing . . . 24
2.2.3 Fixed-interval Kalman smoothing . . . 26
2.2.4 Detection of the different stages of a race . . . 27
2.3 HR data modeling with a long range dependent process . . . 28

2.3.1 A first model : the fractional Gaussian noise 30
2.3.2 Methods of estimations of the Hurst parameter 32

3 Asymptotic Properties of the DFA of LRD Processes 33
3.1 Introduction 33
3.2 Definitions and first properties of the DFA method 34
3.3 Asymptotic properties of the DFA function for a FGN 37
3.3.1 Definition and first properties of the FBM and the FGN 38
3.3.2 Some numerical results of the DFA applied to the FGN 38
3.4 Extension of the results for a general class of a long-range dependent process 41
3.5 Cases of particular trended long-range dependent processes 43
3.5.1 Case of power law and polynomial trends 44
3.5.2 Case of a piecewise constant trend 47
3.6 Conclusion 49
3.7 Proofs 50

4 Wavelet analysis for estimation of regularity of HR series 57
4.1 Introduction 57
4.2 Wavelet based estimator of the Hurst parameter 57
4.2.1 Wavelet analysis 58
4.2.2 Application of both estimators to FGN 60
4.3 Application of both estimators to HR data 61
4.4 A second model : a locally fractional Gaussian noise 63
4.5 Application to HR data 66
4.6 Conclusion 69

5 Changes in the LRD or the self-similarity or the local fractality 71
5.1 Introduction 71
5.2 Main results 75
5.2.1 Notations and assumptions 75
5.2.2 Estimation of abrupt change time-instants $(\tau_j^*)_{1\leq j\leq m}$ 77
5.2.3 Estimation of parameters $(\alpha_j^*)_{0\leq j\leq m}$ and $(\beta_j^*)_{0\leq j\leq m}$ 78
5.2.4 Goodness-of-fit test 79
5.2.5 Cases of polynomial trended processes 79
5.3 Applications 80

5.3.1 Detection of change for Gaussian piecewise long memory processes 80
5.3.2 Detection of abrupt change for piecewise Gaussian self-similar or locally fractional processes having stationary increments 84
5.3.3 Application to heart rate's time series 91
5.4 Proofs 92

6 Conclusion et perspectives **99**

Bibliographie **102**

Table des figures **113**

Liste des tableaux **115**

Table des matières **119**