Détection automatique des anomalies pour l'analyse multifractale

Merlin DUMEUR[1, 2, 3] Philippe CIUCIU[1, 2]

[1]CEA, DRF, Joliot, NeuroSpin, Paris-Saclay University

[2]Inria MIND team, Paris-Saclay University

[3]Department of Neuroscience and Biomedical Engineering, Aalto University

Résumé – L'analyse multifractale étudie les propriétés d'invariance d'échelle de séries temporelles et a trouvé des applications dans un large éventail de domaines. Nous motivons la recherche de valeurs aberrantes pour l'analyse multifractale (MFA) en montrant qu'en neurosciences, où l'adoption de la MFA est à la traîne, il existe de nombreuses sources potentielles de bruit impulsif qui perturbent significativement les estimations. Nous introduisons une méthode itérative pour identifier les valeurs aberrantes dans les séries temporelles, basée sur l'hypothèse que le processus sous-jacent est multifractal, et qui s'appuie sur le formalisme wavelet p-leader. Nous montrons qu'elle améliore l'estimation des paramètres multifractaux de signaux bruités, ainsi que sur un enregistrement EEG intracrânien.

Abstract – Multifractal analysis investigates the scale invariance properties of time series, and has seen applications in a wide range of domains. We motivate finding outliers in time series for multifractal analysis (MFA) by showing that in neuroscience, where adoption of MFA is lagging, there are many potential sources of impulsive noise which significantly perturb the estimates. We propose an iterative method for identifying outliers in time series based on the assumption that the underlying process is multifractal, and which leverages the wavelet p-leader framework. We show it improves the estimation of the multifractal parameters of noisy signals as well as on an intracranial EEG recording.

1 Introduction

En une dimension, l'analyse multifractale cherche à déterminer le spectre de régularité d'une série temporelle. Pour cela il n'est pas possible d'estimer point par point la régularité locale et à la place, on utilise des structures multi-résolutions intermédiaires à partir desquelles est déterminé le spectre multifractal.

Dans certain cas, le signal d'intérêt, qui est supposé invariant par changement d'échelle et potentiellement multifractal, est corrompu par certaines sources de bruit. On définit ici par bruit tout signal qui n'est pas pertinent par rapport à l'objet étudié, ce qui n'est pas toujours du bruit de mesure mais peut être généré par un processus différent.

De par la nature de l'analyse multifractale, les transitoires dans le signal ont un impact démesuré sur l'estimation des paramètres. Dans le cas de bruits impulsifs, même une faible couverture du signal par le bruit suffit à rendre les résultats aberrants et donc inutilisables.

La Figure 1 compare un signal monofractal simulé, avec sa version bruitée dans laquelle on a introduit 3 impulsions de bruit. Même si le bruit concerne seulement 3% du support temporel du signal, l'analyse multifractale qui en découle est complètement perturbée et les coefficients que l'on obtient n'ont pas de sens.

Exemples d'anomalies en neurosciences. Les différentes modalités d'enregistrement de l'activité cérébrale (EEG, MEG, SEEG, IRMf, etc.) présentent des sources de bruits transitoires, qui apparaissent comme des singularités pour l'analyse multifractale. Dans ce cadre, les capteurs peuvent évidemment être sources de bruit : par exemple en MEG les capteurs SQUID peuvent soudainement changer de point de fonctionnement [13] – le voltage change soudainement, et cela apparaît comme un saut dans le signal.

Il est également possible d'avoir des sources de bruit physiologiques, comme les artefacts musculaires dus au mouvements du visage et de la mâchoire, mais ils peuvent tout aussi provenir directement du cerveau, sous la forme de bursts d'activité rythmique très basse fréquence (<0.1 Hz) [3] par exemple.

Objectif. Pour que l'analyse multifractale devienne un outil pratique en neuroscience, il doit être possible de s'en servir sur différents types d'enregistrements de l'activité cérébrale et sans nécessiter de concevoir une chaîne de traitement du signal dédiée. En effet, bien que les méthodes classiques de traitement des signaux en neuroscience soient capables d'éliminer les artefacts pouvant corrompre les analyses classiques qui étudient la réponse du cerveau à des stimuli, il n'en va pas de même pour une analyse fine de leur dynamique temporelle comme celle requise par l'analyse multifractale.

Dans le premier cas il s'agit de découper le signal en sections puis de moyenner l'activité cérébrale sur des répétitions de même nature afin d'identifier des contrastes ; ce faisant les différentes sources de bruit s'annulent car elles sont décorrélées des stimuli et des réponses cérébrales induites. Dans le deuxième cas, et *a fortiori* lorsqu'il s'agit d'analyses en basse fréquence, il n'est pas envisageable de découper le signal de la même manière.

Dans cette contribution nous présentons une méthode pour détecter les anomalies dans le signal brut enregistré typiquement en MEG, EEG ou SEEG dans le but d'effectuer une analyse multifractale, qui fonctionne indépendamment de la nature des données et qui formule des hypothèses minimales sur la nature des enregistrements.

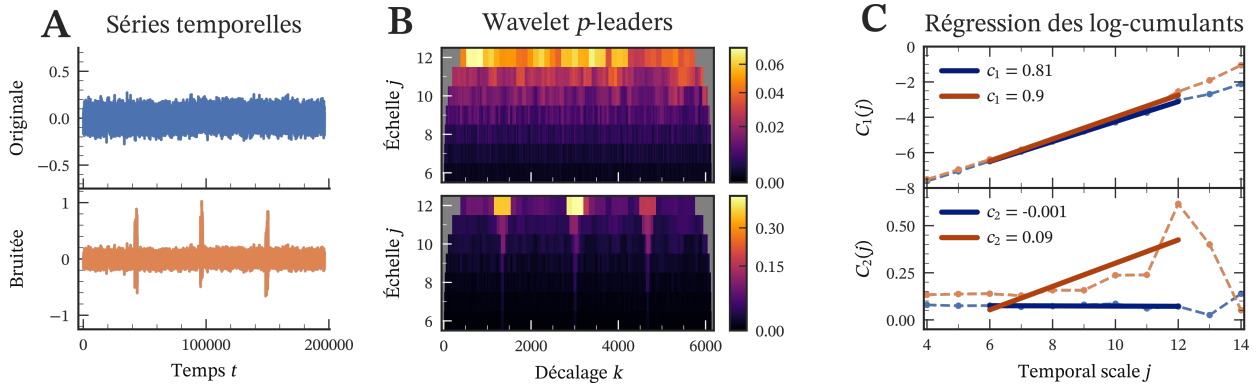

FIGURE 1 : Schéma expérimental. (A) série temporelle originelle (haut) et bruitée (bas). Ce ne sont pas des marches aléatoires mais des bruits gaussiens fractionnaires : l'ajout de bruit n'entraîne pas de discontinuité. (B) wavelet p-leaders correspondant aux signaux en (A). (C) comparaison des log-cumulants originaux (bleu) et bruités (orange), estimés à partir des wavelet p-leaders.

2 Méthode

Il existe plusieurs approches pour déterminer le spectre multifractal, dont la *Multifractal Detrendeded Fluctuation Analysis* (MF-DFA) [9] ou encore le *Wavelet Modulus Maxima* (WTMM) [4]. Nous choisissons ici plutôt d'utiliser le formalisme wavelet p-leader [10] qui permet de caractériser une notion de régularité locale du signal par les p-exposants [11] qui est utilisable sur des fonctions non-bornées.

2.1 Wavelet p-leaders

Étant donnés la représentation d'un signal en coefficients ondelettes $(d_{j,k})_{j,k}$ normalisés par la norme ℓ_1, on introduit la quantité multi-résolution $(s^{(p)}_{j,k})_{j,k}$ afin de simplifier les notations par la suite :

$$s^{(p)}_{j,k} = \sum_{k'=k-1}^{k+1} |d_{j,k'}|^p \tag{1}$$

Selon [10, Eq.(8)], la puissance p-ième des wavelet p-leaders $\left(\ell^{(p)}_{j,k}\right)_{j,k\in[\![1,\bar{j}]\!]\times[\![1,N_j-1]\!]}$ suit la relation de récurrence suivante (pour $p < +\infty$) :

$$\left(\ell^{(p)}_{j,k}\right)^p = s^{(p)}_{j,k} + \frac{1}{2}\left(\left(\ell^{(p)}_{j-1,2k-1}\right)^p + \left(\ell^{(p)}_{j-1,2k+2}\right)^p\right) \tag{2}$$

où l'on fait correspondre le wavelet p-leader $\ell^{(p)}_{j,k}$ avec le coefficient d'ondelette $d_{j,k}$ sur lequel il est centré. On note dans le reste l'article les wavelet p-leaders $(\ell_{j,k})_{j,k}$, sans ambiguïté car on travaille uniquement dans le cas $p < +\infty$.

Formalisme multifractal p-leader. Pour déterminer le spectre multifractal à partir des p-leaders, on peut obtenir un majorant du spectre à partir de la fonction d'échelle $\zeta(q)$, calculée à partir des fonctions de structure. Il est possible d'exprimer $\zeta(q)$ par son expansion en cumulants :

$$\zeta(q) = c_1 q + c_2 \frac{q^2}{2} + c_3 \frac{q^3}{3!} + \cdots \tag{3}$$

Dans notre cas, il suffit de s'intéresser aux log-cumulants $(c_m)_m$ dont les valeurs permettent de déterminer les propriétés du spectre multifractal.

On détermine les $(c_m)_m$ à partir d'une régression linéaire sur les cumulants d'ordre m, $C_m(j)$, des log p-leaders :

$$C_m(j) = \mathrm{Cum}_m\left[(\log \ell_{j,k})_k\right] = c^0_m + j c_m \tag{4}$$

Nous présentons ensuite deux approches pour estimer les paramètres multifractaux de manière robuste : les cumulants robustes (c.f. section 2.2), et une approche itérative (c.f. section 2.3.).

2.2 Cumulants robustes

La première stratégie consiste à remplacer les estimateurs empiriques des cumulants par des estimateurs robustes dont notamment ceux avec une fonction d'influence qui annule l'impact des valeurs aberrantes.

Il existe deux grandes familles d'estimateurs robustes : tout d'abord, les L-estimateurs qui se basent sur une combinaison linéaire des statistiques d'ordre des données observées, par exemple la médiane et l'écart inter-quartiles. Alternativement, les M-estimateurs se basent sur la minimisation d'une fonction d'estimation comme l'anti log-vraisemblance.

Dans ce cadre, nous avons remplacé l'estimateur de la variance empirique par le L-estimateur du Q_n [5], et l'estimateur de la moyenne empirique par le M-estimateur du *Tukey biweight* [12]. Ces choix d'estimateurs sont motivés par le fait que la contribution des observations extrêmes à l'estimateur final est nulle. Pour les cumulants d'ordre supérieur nous nous servons d'estimateurs de moments d'ordre supérieur robustes [14], à partir desquels on calcule les cumulants.

2.3 Algorithme itératif

Soit $L_{j,k}$ la variable aléatoire à l'instant k et l'échelle j dont la réalisation est le log p-leader $\ell_{j,k}$. On munit $\log L_{j,k}$ d'une loi de probabilité paramétrisée par $\boldsymbol{\theta}_j$: $\forall k, \log L_{j,k} \sim \mathcal{D}(\boldsymbol{\theta}_j)$. En pratique, on observe qu'il est nécéssaire d'utiliser une loi paramétrisant la kurtosis (statistique d'ordre 4) : nous avons choisi pour $\mathcal{D}$ la loi gaussienne généralisée $\mathcal{N}(\mu, \sigma, \kappa)$ [1].

Principe d'échantillonnage D'après (2), on remarque que les wavelet p-leaders à l'échelle j ne dépendent que des $(d_{j,k})_k$ et des $(\ell_{j-1,k})_k$. Par conséquent, conditionnellement aux p-leaders à l'échelle $j-1$, les p-leaders à l'échelle j sont indépendants des p-leaders aux échelles $j-2, ..., 1$.

Pour simuler un coefficient à l'échelle j, l'approche naïve consiste à simuler d'abord les $2^{\bar{j}-1}$ p-leaders aux échelles $j-1,\dots,1$ dont il dépend. Cela n'est pas possible en pratique, et l'on choisit donc de simuler échelle par échelle selon la loi jointe par acceptation-rejet, en partant de l'échelle maximale $\bar{j}$.

La méthode consiste en les étapes suivantes :

1. Calcul des p-leaders et estimation initiale des log-cumulants robustes $(\widehat{c}_m)_m$;
2. Rejet des p-leaders aberrants et de leurs « descendants » ;
   (a) déterminer les paramètres $(\mu_j,\sigma_j,\kappa_j)_j$ à partir des $(\widehat{c}_m)_m$ via l'Eq. (4) la méthode des moments[1].
   (b) Simuler $2N$ variables aléatoires $\log L_{j,k'} \sim \mathcal{N}(\mu_j,\sigma_j,\kappa_j)$ [8], partant de l'échelle maximale $\bar{j}$ et sous la contrainte issue de l'Eq. (2) (acceptation-rejet) :
   $$2|L_{j+1,k'}|^p \geqslant |L_{j,2k'}|^p + |L_{j,2k'+1}|^p$$
   (c) Calculer les intervalles de confiance empiriques IC^j_α aux centiles $(\frac{\alpha}{2}, 1-\frac{\alpha}{2})$ des $\left(|L_{j,k}|^p - 1/2(|L_{j-1,2k}|^p + |L_{j-1,2k+1}|^p)\right)_{j,k}$.
   (d) Rejeter les p-leaders $\ell_{j,k}$ où $s_{j,k} \notin IC^j_\alpha$ pour vérifier l'Eq. (2), ainsi que les p-leaders aux échelles supérieures $j' > j$ qui en dépendent.
3. Réestimer les $(\widehat{c}_m)_m$ après rejet. Fin de l'algorithme si convergence ou nombre maximal d'itérations atteint. Sinon retour à l'étape 2.

L'estimation du paramètre κ peut être complexe pour des échantillons de tailles finies, particulièrement concernant les larges échelles. Suivant l'hypothèse d'invariance d'échelle, l'algorithme permet d'estimer les $(\mu_j,\sigma_j,\kappa_j)_j$ à partir de seulement 6 degrés de liberté $(c_1^0, c_2^0, c_4^0, c_1, c_2, c_4)$, indépendamment du nombre d'échelles et du nombre d'échantillons par échelle.

Pour fonctionner correctement, l'algorithme nécessite des estimations suffisamment robustes des $(\hat{c}_m)_m$. Si les cumulants robustes ne sont pas utilisés dans l'algorithme, la convergence n'est pas possible.

3 Résultats

3.1 Simulations

Nous simulons des séries temporelles bruitées par du bruit impulsif sous la forme de *bursts* d'un mouvement brownien fractionnaire (fBM) avec paramètre $H = 0.9$. Nous choisissons 3 impulsions de bruit, couvrant chacune 1% du support temporel du signal, et augmentons progressivement le rapport de puissance bruit/signal de zéro à 2^5. Nous aurions pu également choisir d'augmenter le nombre d'impulsions, avec les mêmes conclusions mais des estimations variant avec l'alignement des intervalles dyadiques des quantités multi-résolution avec le bruit.

Les signaux d'intérêt sont simulés par des fBM avec paramètre $H = 0.8$ pour le cas de signaux monofractaux ainsi que des cascades multiplicatives (MRW) [7] avec paramètres

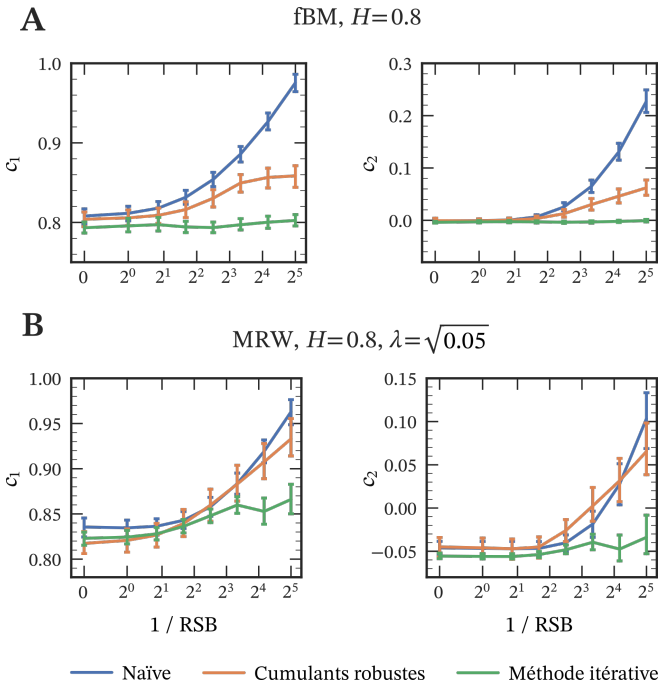

FIGURE 2 : Estimation des c_1 et c_2 moyennés sur 20 simulations, pour différents niveaux de bruits impulsifs. Le bruit est formé de 3 impulsions de la même manière que dans la Figure 1. (A) Valeurs estimées sur un fBM bruité. (B) Valeurs estimées sur une MRW bruitée. On compare les 3 approches, naïve (bleu), cumulants robustes (orange) et la méthode itérative (vert).

$H = 0.8$, $\lambda = \sqrt{0.05}$ (valeur théorique $c_2^{\text{th}} = -\lambda^2$) pour simuler des signaux multifractaux.

La Figure 2 montre que la méthode naïve pour l'estimation des log-cumulants est très facilement perturbée par le bruit impulsif, donnant des valeurs de $c_2 > 0$ sortant du formalisme multifractal. Les cumulants robustes (en orange) réduisent le biais pour estimer les paramètres d'un fBM mais ne fonctionnent pas pour estimer les paramètres de MRW. En revanche, la méthode itérative de rejet des p-leaders (en vert) reste stable même pour des rapport signal à bruit faibles à la fois pour le fBM et la MRW.

3.2 Enregistrement SEEG

Afin de vérifier que la méthode itérative peut également fonctionner sur données réelles, on l'applique à un enregistrement d'EEG intracrânien, qui présente typiquement des bruits impulsifs.

Les résultats présentés Figure 3 montrent que la méthode naïve donne un $c_2 > 0$ n'ayant pas de sens physique, et que la méthode itérative de rejet des wavelet p-leader a convergé vers l'identification des p-leaders qui n'appartiennent pas à la même distribution que le reste du signal. Par conséquent, le c_2 final est inférieur à zéro et compatible avec les valeurs typiques attendues dans le cerveau.

4 Discussion

La technique itérative présentée montre qu'il est possible de se servir du formalisme multifractal comme *a priori* pour

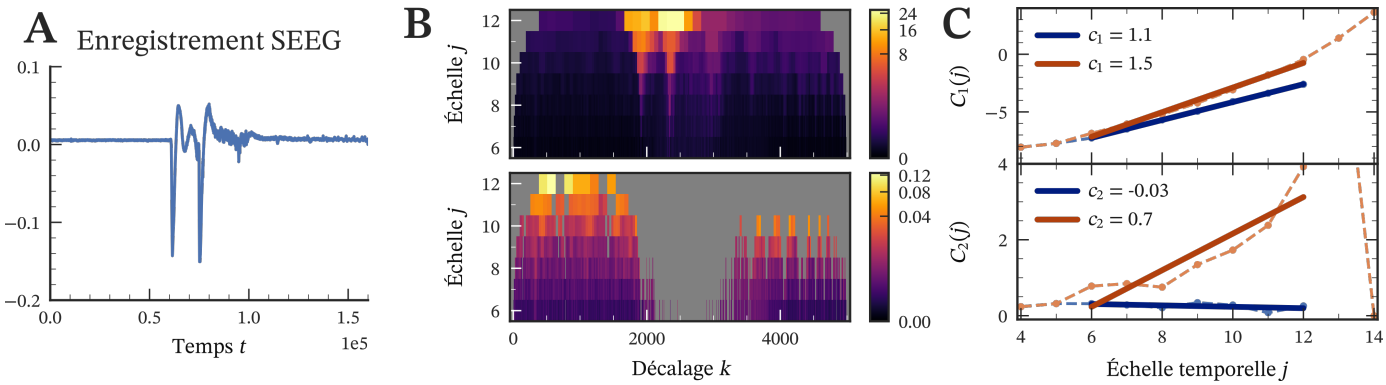

FIGURE 3 : Détection des valeurs anormales sur un signal SEEG bruité. (A) Enregistrement SEEG d'une électrode d'un patient épileptique [6] ; le signal contient deux décharges épileptiformes suivies d'activité épileptique. (B) Wavelet p-leaders, pour le signal d'origine (haut), après masquage (bas) ; la différence de couleur est due à l'échelle qui est différente après rejet. Presque la moitié des valeurs est affectée. (C) Comparaison des log-cumulants avant (orange) et après (bleu) rejet des p-leaders aberrants.

rejeter les wavelet p-leaders qui ont des valeurs aberrantes. Cependant on remarque que les estimations résultantes ne sont pas parfaites : dans le cas où le bruit est nul, l'approche itérative est biaisée par rapport à l'approche naïve pour le c_1 du fBM et le c_1 et c_2 de la MRW.

Analyse multifractale bayésienne. Bien que la première approche présentée ici dépende de la méthode classique d'estimation des paramètres multifractaux, il est logique de l'étendre à l'approche bayésienne, qui a été introduite précédemment [2]. En effet, dans le cadre bayésien, les p-leaders sont naturellement dotés d'une loi de probabilité que l'on peut alors utiliser pour simuler et effectuer l'étape de rejet.

Perspectives. Dans l'immédiat, il est possible de jouer sur les différents paramètres de cette approche (loi paramétrique des p-leaders, seuil α) ainsi que sur la possibilité de modifier α au cours des itérations afin d'améliorer la convergence de l'algorithme. De plus, il serait utile d'accélérer le processus, car actuellement il inflige un coût de calcul plus de 100 fois supérieur à l'approche naïve. À plus long terme, il serait judicieux de formuler cette approche sous la forme d'un algorithme EM pour lequel on pourrait déterminer des critères de convergence et d'optimalité qui manquent actuellement.

Références

[1] Mahesh K. Varanasi Behnaam AAZHANG : Parametric generalized gaussian density estimation. *J. Acoust. Soc. Am.*, 1989.

[2] Herwig Wendt Nicolas Dobigeon Jean Yves Tourneret Patrice ABRY : Bayesian estimation for the multifractality parameter. IEEE, 2013.

[3] M Steriade A Nuiiez F AMZICA : A novel slow (<1 hz) oscillation neocortical neurons in viva : Depolarizing hyperpolarizing components. *J. Neurosci.*, 1993.

[4] J. F. Muzy E. Bacry A. ARNEODO : Multifractal formalism for fractal signals : The structure-function approach versus the wavelet-transform modulus-maxima method. *Phys. Rev. E*, 1993.

[5] Peter J. Rousseeuw Christophe CROUX : Alternatives to the median absolute deviation. *J. Am. Stat. Assoc.*, 1993.

[6] John M. Bernabei et AL. : "hup ieeg epilepsy dataset", 2023.

[7] E. Bacry J. Delour J. F. MUZY : Multifractal rom walk. *Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Related Interdisciplinary Topics*, 2001.

[8] Martina Nardon Paolo PIANCA : Simul. techniques for generalized gaussian densities. *J. Stat. Comput. Simul.*, 2009.

[9] Jan W. Kantelhardt Stephan A. Zschiegner Eva Koscielny-Bunde Shlomo Havlin Armin Bunde H.Eugene STANLEY : Multifractal detrended fluctuation analysis nonstationary time series. *Physica A*, 2002.

[10] R. Leonarduzzi H. Wendt P. Abry S. Jaffard C. Melot S.G. Roux M.E. TORRES : p-exponent p-leaders, part ii : Multifractal analysis. relations to detrended fluctuation analysis. *Physica A*, 2016.

[11] S. Jaffard C. Melot R. Leonarduzzi H. Wendt P. Abry S. G. Roux M. E. TORRES : P-exponent p-leaders, part i : Negative pointwise regularity. *Physica A*, 2016.

[12] Albert E. Beaton John W. TUKEY : The fitting power series, meaning polynomials, illustrated on b-spectroscopic data. *Technometrics*, 1974.

[13] D. Prele M. Piat L. Sipile F. VOISIN : Operating point flux jumps a squid in flux-locked loop. *IEEE Transactions on Applied Superconductivity*, 2016.

[14] Max WELLING : Robust higher order statistics. PMLR, 2005.