UNIVERSITÉ DE LYON

ENS DE LYON

THÈSE
en vue de l'obtention du grade de Docteur, délivré par
l'ÉCOLE NORMALE SUPÉRIEURE DE LYON

École Doctorale N°52
Physique et Astrophysique de Lyon (PHAST)

Discipline : Physique

Soutenue publiquement le 19/10/2023, par :

Charles-Gérard LUCAS

Autosimilarité multivariée : estimation des exposants d'autosimilarité, tests bootstrap d'égalité entre exposants et applications

---

Devant le jury composé de :

Marianne Clausel Professeure des universités, Université de Lorraine	**Rapporteur**
Abdourrahmane Atto Maître de conférences, Université Savoie Mont Blanc	**Rapporteur**
Pierre-Olivier Amblard Directeur de recherche CNRS, Université Grenoble Alpes	**Examinateur**
Béatrice Vedel Maîtresse de conférences, Université de Bretagne Sud	**Examinatrice**
Herwig Wendt Directeur de recherche CNRS, Université de Toulouse	**Examinateur**
Patrice Abry Directeur de recherche CNRS, ENS de Lyon	**Directeur de thèse**

Remerciements

Avant toute chose, mon doctorat a été rendu possible par l'appui financier, bourse de thèse n°01D20019023, de l'Agence de l'innovation de défense de la Direction générale de l'armement, que je remercie chaleureusement.

Mais, au-delà de l'aspect pécuniaire, je dois la réalisation de ce doctorat, pour l'essentiel, à ceux qui l'ont encadré, Patrice Abry et Herwig Wendt. Les trois années qui ont abouti à l'ensemble des travaux présentés dans ce manuscrit ont été rythmées par leur présence constante et leur exigence élevée. Et cette exigence qu'ils ont eu, tout du long, vis-à-vis de moi a doucement mué en une exigence personnelle, exaltant toute ma gratitude.

Bien sûr, ce doctorat n'aurait pu se conclure sans l'évaluation de mes travaux de thèse par Marianne Clausel, Abdourrahmane Atto, Pierre-Oliver Amblard et Béatrice Védel, qui m'ont fait la faveur d'une lecture minutieuse de mon manuscrit, de questions et remarques pointues et pertinentes lors de ma soutenance. Des retours bien utiles pour parfaire cette thèse, retours pour lesquels j'ai donc une grande reconnaissance.

Évidemment, puisque ce doctorat a été mené en France, nombre de questions administratives ont été soulevées au cours de celui-ci, et je remercie les membres du secrétariat du Laboratoire de Physique de l'ENS de Lyon de les avoir résolues les unes après les autres avec une efficacité inébranlable.

Mes remerciements ne s'arrêtent cependant pas ici. Hormis les aspects scientifiques, financiers et administratifs, il y a une composante sociale importante dans la réalisation d'un doctorat. C'est à ce titre que je tiens à remercier l'ensemble de ceux qui ont été membres du Laboratoire de Physique de l'ENS de Lyon durant les années que j'y ai passé. L'ambiance joyeuse et chaleureuse qu'il y ont fait régner était un cadre idéal pour découvrir le monde de la recherche et y faire mes premières armes.

Et si je garde un souvenir si allègre de mes années de doctorat, c'est sûrement grâce à une poignée de fidèles camarades. Je remercie particulièrement Thomas Basset avec qui j'ai partagé le plus de rires. Pour nos longues discussions et leurs précieuses digressions, parfois intellectuellement riches, parfois sans queue ni tête. Pour les cruelles défaites qu'il m'a infligées à tant de jeux. Pour mes overdoses de houmous. Je remercie également Léo Mangeolle pour son humour acéré, ses imitations à la pointe et ses remarques percutantes. Je remercie Hubert Souquet-Basiège d'être un si bon confident et pour sa bonne humeur égayante. Je remercie Geoffroy Heaseler pour sa gentillesse sans égale et ses conseils esthétiques indispensables pour les illustrations scientifiques. Je remercie Janka Lengyel d'avoir été une compagne de voyage

idéale et pour sa présence réconfortante. Je remercie Nicolas Perez pour sa tranquillité et sa sympathie inextinguibles.

Merci à mes amis, à ma famille, à ma bien-aimée.

À mon père, qui aurait sacrifié des nuits pour lire ce pavé.

Table des matières

Liste des notations et abréviations

Notations

$\mathbb{N}, \mathbb{Z}, \mathbb{R}, \mathbb{C}$	ensembles des entiers naturels, des entiers relatifs, des réels et des complexes, respectivement
$\mathcal{L}^2(\mathbb{R})$	espace des fonctions de carré intégrable au sens de Lebesgue
$\mathcal{S}(M, \mathbb{R})$	ensemble des matrices symétriques réelles de taille $M \times M$
$\mathcal{S}_{\geq 0}(M, \mathbb{R})$	sous-ensemble des matrices semi-définies positives de $\mathcal{S}(M, \mathbb{R})$
$\mathcal{S}_{>0}(M, \mathbb{R})$	sous-ensemble des matrices définies positives de $\mathcal{S}(M, \mathbb{R})$
$\langle \cdot, \cdot \rangle$	produit scalaire usuel sur $\mathcal{L}^2(\mathbb{R})$
$\Re(A), \Im(A)$	parties réelle et imaginaire de la matrice A, respectivement
$\mathbb{1}_{\{x \in S\}}$	fonction indicatrice de l'ensemble S évaluée en x
$\mathrm{Card}(S)$	cardinal de l'ensemble S
$\mathcal{N}(\mu, \sigma)$	loi normale d'espérance μ et d'écart-type σ
$\mathbb{E}[X], \mathrm{Var}(X)$	espérance et variance de la variable aléatoire X, respectivement
$\{X(t)\}_{t \in T}$	processus stochastique indexé par T à trajectoires réelles
$\mathbb{I}$	matrice identité de taille donnée par le contexte
A^T, A^*	transposée et matrice adjointe de A, respectivement
$\odot$	produit matriciel de Hadamard
$\xrightarrow{\mathbb{P}}$, $\overset{\mathbb{P}}{\sim}$	limite et équivalence en probabilité, respectivement
$\overset{d}{=}$, $\xrightarrow{d}$	égalité et limite en loi, respectivement
$\underset{h \to l}{\longrightarrow}$, $\underset{h \to l}{\sim}$	limite et équivalence quand h tend vers $l \in \mathbb{R} \cup \{+\infty\}$, respectivement
$:=$	égalité par définition
$\overset{fdd}{=}$	égalité en distributions de dimensions finies
$\arg\min_{x \in S} f(x)$	argument du minimum de la fonction f sur l'ensemble S

Abréviations

mBf	mouvement brownien fractionnaire
mBof	mouvement brownien opérateur-fractionnaire
M-mBf	mouvement brownien fractionnaire M-varié

Introduction

Invariance d'échelle

Alors qu'il est courant de déterminer une ou des échelles caractéristiques pour étudier un système, l'invariance d'échelle renverse le bien fondé d'une telle approche dans de nombreux domaines du monde réel : un grand intervalle d'échelles contribue à la dynamique temporelle ou spatiale. Par exemple, alors que les signaux d'activité cérébrale sont souvent analysés en termes d'activité rythmique, notamment à travers la puissance dans des bandes de fréquences bien répertoriées par la littérature (Buzsaki, 2006), la dynamique invariante d'échelle y est aussi source d'information (Ciuciu et collab., 2014; He, 2014), notamment dans l'activité cérébrale lente (inférieure à 0.1Hz).

Dans un système invariant d'échelle, le comportement de l'ensemble se retrouve dans toutes ses parties, du moins sur une large gamme d'échelles, si bien qu'il est impossible de distinguer l'échelle d'observation. La caractérisation de la relation entre les différentes échelles, appelée loi d'échelle, s'impose alors. D'une manière plus intuitive, un signal, ou une image, invariant d'échelle s'obtient par répétition d'une de ses parties dilatée et translatée avec des aléas, et c'est l'opération en jeu qu'il convient d'identifier.

L'invariance d'échelle se manifeste dans des systèmes de natures très différentes. En physique et géophysique, elle intervient dans la turbulence (Frisch, 1995; Mandelbrot, 1975), la répartition des galaxies (Peebles, 1989), le fond diffus cosmologique (Davis et collab., 1992) ou les chutes de pluie (Olsson et collab., 1993). En biologie, elle s'observe entre autres dans la fréquence cardiaque (Ivanov, 2007; Kiyono et collab., 2006), la structure de plantes (Palmer, 1988) et la réplication de l'ADN (Takahashi, 1989). Parmi les phénomènes naturels où l'invariance d'échelle est en jeu, on peut également citer les reliefs et les séismes en géologie (Turcotte, 1990). L'activité humaine est aussi au cœur de systèmes décrits par des signaux ou images invariants d'échelle, tels que le trafic Internet (Abry et collab., 2002), les cours boursiers (Mandelbrot, 1999), la croissance urbaine (Batty et collab., 1989) et la texture de peintures (Abry et collab., 2015, 2013).

Cette large palette d'applications motive la modélisation et l'analyse de l'invariance d'échelle. À ce titre, dans les dernières décennies, nombre de modèles stochastiques et outils d'analyse multi-échelle ont été élaborés pour rendre compte de dynamiques invariantes d'échelle (Abry et collab., 2019). Le travail présenté dans ce manuscrit se focalise sur l'étude des signaux décrivant des systèmes invariants d'échelle.

Autosimilarité univariée

Une série temporelle est dite autosimilaire si sa loi est covariante par changement d'échelle des temps : toute partie d'une série temporelle autosimilaire est statistiquement indistinguable de la série temporelle d'origine après une dilatation appropriée de son amplitude. Plus précisément, les statistiques d'une série temporelle autosimilaire se comportent comme des lois de puissance par rapport à l'échelle et l'exposant de ces lois de puissance est contrôlé par l'exposant d'autosimilarité H, variant entre 0 et 1. L'autosimilarité constitue ainsi un formalisme pertinent pour modéliser l'invariance d'échelle. Le mouvement brownien fractionnaire (mBf) (Mandelbrot et Ness, 1968; Pipiras et Taqqu, 2017; Samorodnitsky et collab., 1996), le seul processus gaussien centré autosimilaire et à accroissements stationnaires, a été utilisé de manière presque exclusive comme modèle de référence pour les signaux invariants d'échelle. D'autres modèles reposant sur des processus non-gaussiens existent aussi (voir par exemple Boniece et collab. (2019); Clausel et collab. (2014)). L'autosimilarité peut également être formalisée pour des champs aléatoires, et des modèles pour les images tenant compte de leur possible anisotropie ont également vu le jour (Clausel et Vedel, 2011; Didier et collab., 2018).

L'analyse de l'autosimilarité d'une série temporelle mesurée consiste en l'estimation de l'exposant d'autosimilarité H à partir d'une observation de taille finie. Celle-ci a ainsi reçu une attention considérable, largement relatée par Bardet et collab. (2003); Pipiras et Taqqu (2017); Taqqu et Teverovsky (1998). En particulier, les représentations multi-échelles (transformées en ondelettes) ont permis l'élaboration de procédures d'estimation précises, robustes et rapides de l'exposant d'autosimilarité H (Atto et Berthoumieu, 2011; Atto et collab., 2010; Clausel et collab., 2012; Flandrin, 1992; Veitch et Abry, 1999). Ces procédures reposent essentiellement sur le fait que la variance des coefficients d'ondelettes, appelée spectre d'ondelettes, se comporte comme une loi de puissance par rapport à l'échelle dont l'exposant est contrôlé par l'exposant d'autosimilarité H. Par ailleurs, les outils d'analyse multi-échelle de l'autosimilarité ont également été étendus aux images (Atto et collab., 2013; Clausel et Vedel, 2013; Roux et collab., 2013).

Principalement par manque de modèles et d'outils adaptés, la modélisation et l'analyse de l'invariance d'échelle se sont essentiellement restreintes à un cadre univarié : différents signaux ou images issus d'un même système sont analysés indépendamment. Pourtant, dans la plupart des applications récentes et modernes, un même système est surveillé par plusieurs outils de mesures, ce qui implique naturellement l'analyse conjointe, dite multivariée, de la collection de séries temporelles qui en résulte. À titre d'exemple, dans les applications biomédicales, de multiples signaux d'activité cérébrale provenant de différentes régions du cerveau d'un même patient sont souvent enregistrés simultanément et l'analyse multivariée de telles données est usuelle (McIntosh et Mišić, 2013).

Autosimilarité multivariée

Pour tenir compte de la nature multivariée des données du monde réel, une extension multivariée du mBf a été introduite : le mouvement brownien opérateur-fractionnaire (mBof) (Didier et Pipiras, 2011, 2012). Le mBof vérifie une relation d'autosimilarité multivariée : les statistiques d'une série temporelle autosimilaire M-variée se comportent comme des mélanges de lois de puissance par rapport à l'échelle dont les exposants sont contrôlés par M exposants d'autosimilarité $H_1, \ldots, H_M$ variant entre 0 et 1. Ainsi, chaque composante d'un mBof M-varié est caractérisé par l'intégralité d'un vecteur de M exposants d'autosimilarité $\underline{H} = (H_1, \ldots, H_M)$. Pour l'analyse de données du monde réel, un cas particulier de mBof a été proposé par Amblard et Coeurjolly (2011); Amblard et collab. (2012). Celui-ci consiste en des collections

de M mBf corrélés, chacun caractérisé par un exposant d'autosimilarité H_m. Ce dernier modèle est cependant assez restrictif car chaque composante du processus en question est autosimilaire d'exposant une entrée de $\underline{H}$, signifiant que la relation d'autosimilarité multivariée en jeu se réduit simplement à M relations d'autosimilarité univariée.

De façon analogue à l'analyse d'un mBf, des représentations multi-échelles multivariées permettent l'estimation du vecteur des exposants d'autosimilarité $\underline{H} = (H_1, \ldots, H_M)$ (Coeurjolly et collab., 2013; Wendt et collab., 2017), exploitant les comportements en loi de puissance des spectres d'ondelettes de chaque composante. Cependant, de telles approches ne sont pertinentes que lorsqu'il y a association biunivoque entre composante et exposant d'autosimilarité. Plus récemment, une procédure d'estimation à partir des valeurs propres de la matrice de covariance des coefficients d'ondelettes, appelée spectre d'ondelettes multivarié, a été proposée et théoriquement discutée par Abry et Didier (2018a,b) dans la limite asymptotique de grandes tailles d'échantillon. Il s'avère que, dans un cadre assez général, chacune des statistiques des M valeurs propres du spectre d'ondelettes multivarié d'un mBof se comporte asymptotiquement comme une loi de puissance par rapport à l'échelle d'exposant contrôlé par un exposant d'autosimilarité H_m. Ces comportements suggèrent d'estimer les exposants d'autosimilarité par des régressions linéaires sur les logarithmes d'estimées des valeurs propres au travers des échelles. Cependant, pour des séries temporelles de taille finie, cette approche se heurte à l'*effet de répulsion*, un écart entre les valeurs propres estimées plus grand que l'écart entre les valeurs propres exactes à une échelle donnée (Anderson et collab., 2010; Tao, 2012; Yao et collab., 2015). Puisque le nombre de coefficients d'ondelettes intervenant dans le calcul du spectre d'ondelettes multivarié diffère à chaque échelle, il en va de même pour le biais des valeurs propres estimées, ce qui implique un biais dans les régressions linéaires et donc dans l'estimation de $\underline{H}$ (Wendt et collab., 2019). Proposer un estimateur de $\underline{H}$ tenant compte de ce biais de taille finie pour l'étude de signaux du monde réel est l'un des sujets auxquels cette thèse s'intéresse.

Dans la pratique, il est également important de détecter combien d'exposants d'autosimilarité différents sont réellement à l'origine de la dynamique conjointe des observations, et d'estimer la proportion de chacune de ces valeurs dans $\underline{H}$. En effet, l'analyse de l'autosimilarité multivariée fournit une collection d'estimées des exposants d'autosimilarité $H_1, \ldots, H_M$, mais la fluctuation de l'estimateur conduit à des estimées distinctes pour des exposants d'autosimilarité égaux. Une étude pertinente d'une collection d'observations nécessite d'identifier des groupes d'exposants d'autosimilarité en fait égaux à partir de leurs estimées. Dans un cadre bivarié, une stratégie bootstrap par blocs dans le domaine des ondelettes a été conçue par Wendt et collab. (2018) pour tester si les deux exposants d'autosimilarité H_1 et H_2 en jeu sont égaux ou non à partir de l'estimateur par valeurs propres proposé par Abry et Didier (2018a). Ces travaux ouvrent la voie au dénombrement et au regroupement d'exposants d'autosimilarité $(H_1, \ldots, H_M)$ pour un nombre de composantes M supérieur à deux : dénombrer les exposants d'autosimilarité signifie compter le nombre de valeurs distinctes présentes dans $\underline{H}$ tandis que les regrouper revient à compter le nombre d'exposants d'autosimilarité dans $\underline{H}$ prenant chacune de ces valeurs.

Par ailleurs, les applications du monde réel font parfois intervenir une très grande quantité de capteurs. Dans le domaine des neurosciences, quelques centaines de séries temporelles d'activité cérébrale peuvent être obtenues par magnétoencéphalographie et plusieurs dizaines de milliers par imagerie par résonance magnétique fonctionnelle (Ciuciu et collab., 2012). Alors que le nombre de séries temporelles mesurées est grand, leur taille reste limitée, pour des raisons de stockage par exemple. Pour être réaliste en ce qui concerne les applications, il est important d'étudier également l'autosimilarité multivariée en grande dimension, où le nombre M de séries temporelles n'est plus fixe mais croît avec leur taille. Dans ce contexte, Orejola et collab. (2022) a tout d'abord proposé une procédure pour détecter la présence d'une unique valeur dans

$\underline{H}$. Cette procédure exploite les propriétés de la distribution des valeurs propres des grandes matrices de covariance aléatoires (Bai et Silverstein, 2010; Tao et Vu, 2012). Cependant, le seuil de rejet du test proposé est défini à partir de mBof synthétiques, indépendants des données. Concevoir une procédure de test à partir d'une unique observation de données multivariées, sans étalonnage préalable, reste un enjeu important. Dans le cas où différentes valeurs sont présentes dans $\underline{H}$, l'enjeu de compter ces valeurs et estimer leur proportion n'a encore jamais été traité.

Organisation et contributions

Le chapitre 1 rapporte les outils essentiels à l'analyse de l'autosimilarité multivariée, modélisée par le mouvement brownien opérateur-fractionnaire (mBof), par représentation multi-échelle (ondelettes). Par souci pédagogique, des éléments d'analyse d'autosimilarité univariée sont d'abord rappelés : le mouvement brownien fractionaire (mBf) et l'analyse en ondelettes univariée discrète. Une description du mBof et de ses propriétés, comme extension du mBf, ainsi que différentes procédures d'estimation pour le vecteur des exposants d'autosimilarité $\underline{H}$ sont au cœur de ce chapitre. En particulier, deux procédures sont présentées : la première, proposée par Wendt et collab. (2017), consiste à considérer les différentes composantes du mBof de façon indépendante, constituant ainsi une analyse univariée ; la seconde, proposée par Abry et Didier (2018a,b), exploite les valeurs propres du spectre d'ondelettes multivarié pour tenir compte de la nature multivariée du mBof. Les performances asymptotiques théoriques des estimateurs, dans la limite asymptotique de grandes tailles d'échantillon, sont énoncées et les limites pratiques de ces estimateurs sont exposées (Abry et Didier, 2018a,b).

Le chapitre 2 relate la première contribution de ce travail. La procédure d'estimation par ondelettes initialement proposée par Abry et Didier (2018a,b) est étendue pour faire face au biais d'estimation de taille finie induit par l'effet de répulsion. L'approche proposée s'inspire de la procédure d'origine qui exploite les valeurs propres du spectre d'ondelettes multivarié à différentes échelles, mais celles-ci sont ici estimées de telle sorte que l'effet de répulsion entre elles soit du même ordre au travers des échelles. Des résultats montrent que cette procédure est théoriquement bien fondée. Notamment, l'estimateur s'avère être asymptotiquement consistant, gaussien et décorrélé sous des hypothèses faibles. Pour une étude numérique fouillée, le M-mBf, un cas particulier de mBof, est proposé et détaillé. Le M-mBf est constitué de M mélanges linéaires de M mBf corrélés, modèle plus riche que celui proposé par Amblard et collab. (2012) ne permettant pas les mélanges. Il s'avère être un modèle pratique, simple et pertinent : ses paramètres sont des quantités mesurables en pratique et sa synthèse numérique est facile à mettre en œuvre. Des simulations de Monte Carlo menées sur des M-mBf synthétiques ont ainsi permis une étude fouillée des performances d'estimation, assurant que la procédure proposée est pratiquement robuste, efficace et opérationnelle. L'influence des paramètres du modèle d'autosimilarité multivariée sur les performances d'estimation est particulièrement étudiée et illustrée. Est également réalisée une comparaison approfondie entre l'estimateur univarié ramenant à analyser des signaux indépendamment, l'estimateur multivarié proposé par Abry et Didier (2018b) et l'estimateur multivarié corrigé proposé dans ce travail. Les résultats font l'objet d'un article en cours de rédaction [7].

Le chapitre 3 présente la seconde contribution de ce travail : le développement de procédures de dénombrement et regroupement des M exposants d'autosimilarité $H_1, \ldots, H_M$ caractérisant une observation de séries temporelles autosimilaires multivariées de taille finie. Différentes procédures de test ont en effet été élaborées à partir de l'estimateur multivarié corrigé proposé dans le chapitre 2. Pour définir les seuils de rejet des différents tests proposés à partir d'une unique observation, une contribution de ce chapitre est d'exploiter une procédure de ré-échantillonnage bootstrap proposée dans Wendt et collab. (2018), conçue pour préserver la structure de dépendance

conjointe (multivariée) des coefficients d'ondelettes et donc les statistiques conjointes des estimées de $H_1, \ldots, H_M$. En découlent les procédures suivantes. En premier lieu, un test du χ^2 pour tester l'égalité de l'ensemble des exposants d'autosimilarité, i.e. l'hypothèse $H_1 = \ldots = H_M$, a mené à une communication à la conférence internationale EUSIPCO [1]. Dans le cas où l'égalité est rejetée, plusieurs valeurs distinctes sont présentes dans $\underline{H}$ et des procédures de tests par paires d'exposants s'imposent. Une procédures ordonnant les exposants estimés, $\hat{H}_1 \leq \ldots \leq \hat{H}_M$, et testant l'égalité des $M-1$ paires d'exposants successifs a fait l'objet de deux communications, la première à la conférence internationale ICASSP [2] et la seconde à la conférence nationale GRETSI [4], présentant des méthodes d'estimation des paramètres de test différentes. A ensuite été proposée une autre procédure consistant à tester l'égalité des $M(M-1)/2$ paires d'exposants disponibles et à employer une méthode de partitionnement d'un graphe des exposants traitant l'information de ces tests. Cette procédure est l'objet d'un article en cours de rédaction [8]. Les différentes procédures sont évaluées numériquement à partir de simulations de Monte Carlo sur des M-mBf synthétiques pour différentes tailles d'échantillon N, et une comparaison pour différents nombres de composantes M conclut le chapitre. Bien que les trois procédures de tests par paires montrent des performances satisfaisantes, la troisième s'avère la plus robuste aux faibles tailles d'échantillon N et grands nombres de composantes M.

Le chapitre 4 aborde le cadre de la grande dimension, où le nombre de composantes M tend vers l'infini avec la taille d'échantillon N. Tout d'abord, des simulations de Monte Carlo réalisées sur des M-mBf synthétiques montrent que la normalité multivariée de l'estimateur multivarié corrigé proposé dans le chapitre 2 peut être mise en défaut pour de grands nombres de composantes M et des tailles d'échantillons limitées N, motivant l'étude en grande dimension. En conséquence, les propriétés asymptotiques de l'estimateur multivarié corrigé sont étudiées lorsque le nombre de composantes M, la taille d'échantillon N et les échelles d'analyse tendent conjointement vers l'infini. Il s'avère que le nombre de modes de la distribution des estimées $\hat{H}_1, \ldots, \hat{H}_M$ des exposants d'autosimilarité tend asymptotiquement vers le nombre de valeurs dans $\underline{H}$. S'appuyant sur ces propriétés, une nouvelle procédure pour tester la présence d'une unique valeur dans $\underline{H}$ a été mise au point à partir de la procédure de ré-échantillonnage bootstrap de Wendt et collab. (2018). En cas de rejet, une procédure complète pour compter les valeurs distinctes présentes dans $\underline{H}$ et leur proportion a été élaborée à l'aide de tests de multimodalité. Les simulations de Monte Carlo permettent de montrer les bonnes performances de ces procédures. Ces résultats ont été rapportés dans un article soumis à la conférence nationale GRETSI [6].

Enfin, le chapitre 5 rapporte l'utilisation des outils d'estimation de $\underline{H}$ dans le cadre d'applications biomédicales. En effet, l'analyse de l'invariance d'échelle a souvent été utilisée pour étudier des signaux physiologiques, mais a jusqu'alors été essentiellement menée dans un cadre univarié (Ahn et collab., 2016; Cahyadi et collab., 2019; Domingues et collab., 2019; Gadhoumi et collab., 2015). Pourtant, la fréquente multiplicité des séries temporelles enregistrées conjointement pour décrire un même mécanisme biologique suggère le recours à une analyse multivariée. Deux applications ont ainsi été réalisées sur ce type de données. La première application est la détection de la somnolence à partir de données de polysomnographie, sujet d'une communication à la conférence internationale EMBC [3]. La seconde application est la prédiction de crises d'épilepsie à partir de signaux d'activité cérébrale, objet d'une communication à la conférence internationale EUSIPCO [5]. Des méthodes de classification ont ainsi été mises en place pour réaliser ces deux tâches à partir des estimateurs univarié et multivarié. L'étude menée montre l'intérêt de l'approche multivariée comparée à une approche univariée, la première permettant de surpasser les performances de classification de la seconde pour effectuer les tâches escomptées.

Pour répondre à l'ambition d'une science ouverte, l'ensemble des routines Matlab dévelop-

pées pour l'estimation, le dénombrement et le regroupement de $\underline{H}$ est disponible publiquement à https://github.com/charlesglucas/ofbm_tools.

Publications et communications

[1] Lucas, C.-G., P. Abry, H. Wendt et G. Didier. 2021, «Bootstrap for testing the equality of selfsimilarity exponents across multivariate time series», dans *2021 29th European Signal Processing Conference (EUSIPCO)*, IEEE, p. 1960–1964.

[2] Lucas, C.-G., P. Abry, H. Wendt et G. Didier. 2022, «Counting the number of different scaling exponents in multivariate scale-free dynamics Clustering by bootstrap in the wavelet domain», dans *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, p. 5513–5517.

[3] Lucas, C.-G., P. Abry, H. Wendt et G. Didier. 2022, «Drowsiness detection from polysomnographic data using multivariate selfsimilarity and eigen-wavelet analysis», dans *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, p. 2949–2952.

[4] Lucas, C.-G., H. Wendt, P. Abry et G. Didier. 2022, «Multivariate time-scale bootstrap for testing the equality of selfsimilarity parameters», dans *XXVIIIème Colloque Francophone de Traitement du Signal et des Images (GRETSI 2022)*.

[5] Lucas, C.-G., P. Abry, H. Wendt et G. Didier. 2023, «Epileptic seizure prediction from eigen-wavelet multivariate selfsimilarity analysis of multi-channel EEG signals», dans *2023 European Signal Processing Conference (EUSIPCO)*, IEEE.

[6] Lucas, C.-G., P. Abry, H. Wendt, G. Didier et O. Orejola. 2023, «Bootstrap based test for the unimodality of estimated Hurst exponents. Performance assessment in a high-dimensional analysis setting», dans *XXVIVème Colloque Francophone de Traitement du Signal et des Images (GRETSI 2023)*.

[7] Lucas, C.-G., G. Didier, H. Wendt et P. Abry. 2023, «Multivariate self-similarity Multiscale eigen structures for the estimation of Hurst exponents», en cours de rédaction.

[8] Lucas, C.-G., P. Abry, H. Wendt et G. Didier. 2023, «Multivariate self-similarity parameter counting Spectral clustering using wavelet-domain bootstrap», en cours de rédaction.

Chapitre 1

Autosimilarité multivariée et estimation des exposants d'autosimilarité

Sommaire

1.1 Introduction

Ce chapitre expose les outils classiques de traitement statistique du signal pour l'analyse de l'autosimilarité multivariée. Plus précisément, l'objectif du présent chapitre est de présenter les procédures existantes d'estimation du vecteur des exposants d'autosimilarité $\underline{H} = (H_1, \ldots, H_M)$, d'en détailler les performances théoriques et d'en étudier les limites. À cette fin, la section 1.2 introduit tout d'abord l'autosimilarité dans un cadre univarié avec le modèle de mouvement brownien fractionnaire (mBf). Les outils d'analyse en ondelettes univariée pour l'estimation de l'exposant d'autosimilarité H du mBf y sont aussi présentés. Ensuite, la section 1.3 introduit le mouvement brownien opérateur-fractionnaire (mBof). Deux procédures d'estimation du vecteur des exposants d'autosimilarité $\underline{H} = (H_1, \ldots, H_M)$ caractérisant le mBof sont introduites et étudiées théoriquement dans la section 1.4. Ces procédures d'estimation reposent toutes deux sur une analyse en ondelettes multivariée. Le premier estimateur étudié dans cette section exploite le comportement du spectre d'ondelettes de chaque composante m, qui suit une loi de puissance d'exposant H_m dans le cas particulier où chaque composante est caractérisée par un exposant H_m. Le second estimateur exploite les valeurs propres du spectre d'ondelettes multivarié, qui suivent chacune des lois de puissance d'exposant H_m. Ce dernier estimateur est valable dans un cadre plus général où chaque composante est caractérisée par l'ensemble des entrées du vecteur $\underline{H} = (H_1, \ldots, H_M)$.

1.2 Mouvement brownien fractionnaire (mBf)

1.2.1 Définition

Un mouvement brownien fractionnaire (mBf, Mandelbrot et Ness (1968)) d'exposant d'autosimilarité H, tel que $0 < H < 1$, est un processus stochastique gaussien $\{\mathcal{B}_H(t)\}_{t\in\mathbb{R}}$ vérifiant :

$$\begin{aligned} &(i)\ \mathcal{B}_H(0) = 0, \\ &(ii)\ \forall (t,s) \in \mathbb{R}^2, \quad \mathcal{B}_H(t) - \mathcal{B}_H(s) \sim \mathcal{N}\left(0, C_H|t-s|^{2H}\right), \end{aligned} \tag{1.1}$$

avec $C_H \in \mathbb{R}$. En d'autres termes, le mBf est un processus dont les accroissements sont gaussiens, stationnaires, centrés et de variance se comportant comme une loi de puissance par rapport au temps d'exposant $2H$. Le mBf est une généralisation du *mouvement brownien ordinaire*, obtenu lorsque $H = \frac{1}{2}$.

Le mBf est en fait l'unique processus gaussien centré à accroissements stationnaires dont la fonction de covariance s'écrit

$$\forall (t,s) \in \mathbb{R}^2, \quad \mathbb{E}\left[\mathcal{B}_H(s)\mathcal{B}_H(t)\right] = \frac{C_H}{2}\left(|t|^{2H} + |s|^{2H} - |t-s|^{2H}\right). \tag{1.2}$$

Le mouvement brownien ordinaire correspond ainsi à des accroissements indépendants.

Par ailleurs, le mBf peut également être défini à partir de sa représentation harmonisable (Samorodnitsky et collab., 1996),

$$\forall t \in \mathbb{R}, \quad \mathcal{B}_H(t) := \int_{\mathbb{R}} \frac{e^{itf} - 1}{if} a|f|^{-\left(H-\frac{1}{2}\right)} \tilde{B}(\mathrm{d}f), \tag{1.3}$$

où $\tilde{B}$ est une mesure aléatoire gaussienne hermitienne à valeurs complexes telle que $\mathbb{E}[\tilde{B}(\mathrm{d}f)\tilde{B}(\mathrm{d}f)^*] = \mathrm{d}f$, où $\tilde{B}(\mathrm{d}f)^*$ désigne le conjugué de $\tilde{B}(\mathrm{d}f)$, et $a \in \mathbb{C}$.

1.2.2 Trajectoires

Le mBf $\mathcal{B}_H$ n'est pas stationnaire mais ses accroissements sont stationnaires,

$$\{\mathcal{B}_H(t) - \mathcal{B}_H(s)\}_{(t,s)\in\mathbb{R}^2} \overset{fdd}{=} \{\mathcal{B}_H(t-s)\}_{(t,s)\in\mathbb{R}^2}\,, \tag{1.4}$$

où $\overset{fdd}{=}$ désigne l'égalité en distributions de dimensions finies.

De plus, les trajectoires d'un mBf sont continues mais nulle part dérivables. Des exemples de trajectoire de mBf sont données par la figure 1.1 pour différents exposants d'autosimilarité H. L'exposant d'autosimilarité contrôle en fait la rugosité de la trajectoire : plus H est proche de 1, plus la trajectoire est lisse.

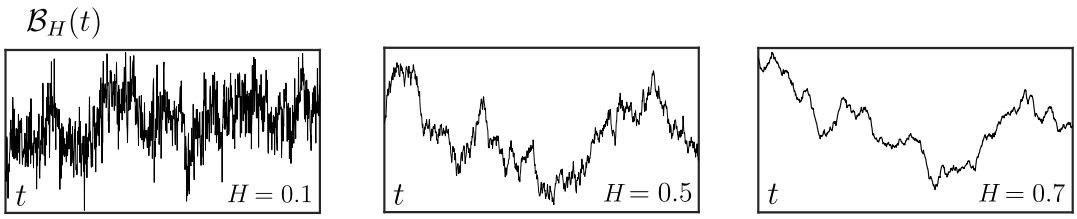


FIGURE 1.1 – **Exemples de trajectoires d'un mouvement brownien fractionnaire.**

1.2.3 Autosimilarité univariée

Le mBf $\mathcal{B}_H$ satisfait la relation d'*autosimilarité* (TAQQU, 2003) suivante :

$$\forall a > 0, \quad \{\mathcal{B}_H(t)\}_{t\in\mathbb{R}} \overset{fdd}{=} \left\{a^H \mathcal{B}_H\left(\frac{t}{a}\right)\right\}_{t\in\mathbb{R}}. \tag{1.5}$$

En d'autres termes, les propriétés statistiques du processus $\mathcal{B}_H$ restent inchangées après changement d'échelle $t \to t/a$ avec un changement d'amplitude $\mathcal{B} \to a^H\mathcal{B}$, et ce pour tout facteur de dilatation $a > 0$. Cette relation est illustrée par la figure 1.2. Il s'avère que le mBf est l'unique processus gaussien centré et continu presque sûrement qui soit autosimilaire et à accroissements stationnaires.

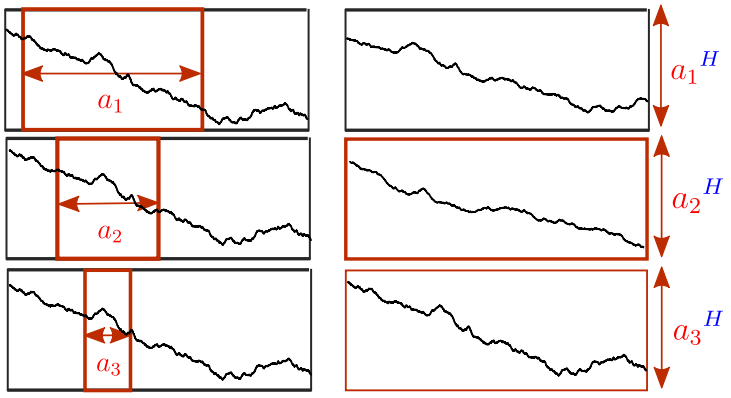


FIGURE 1.2 – **Illustration de l'autosimilarité.** À gauche, un même signal est observé à des échelles a différentes. À droite, les signaux obtenus par dilatation en temps d'un facteur $a > 0$ et en amplitude d'un facteur a^H, où $0 < H < 1$, ont les mêmes propriétés statistiques.

L'exposant d'autosimilarité H, contrôlant le comportement du mBf au travers des échelles,

est l'unique paramètre du mBf. Son estimation est donc cruciale dans le cadre d'applications.

1.2.4 Estimation par ondelettes de l'exposant d'autosimilarité

Diverses méthodes existent pour estimer l'exposant d'autosimilarité H (voir Bardet et collab. (2003); Pipiras et Taqqu (2017) pour un état de l'art). Le travail présenté dans ce manuscrit se concentre sur l'analyse en ondelettes, par définition analyse temps-échelle, qui s'avère être un outil naturel pour l'étude du mBf, processus à la fois non-stationnaire et autosimilaire. En particulier, une décomposition en ondelettes discrètes et orthogonales permet d'estimer l'exposant d'autosimilarité H (Abry et Veitch, 1998; Flandrin, 1992; Veitch et Abry, 1999).

1.2.4.1 Base d'ondelettes

Soit $\psi_0 \in \mathcal{L}^2(\mathbb{R})$ une fonction de référence oscillante à valeurs réelles, appelée *ondelette mère*, caractérisée par son nombre de moments nuls $N_\psi \geq 2$ défini par

$$\begin{cases} \forall k \in \{0, \dots, N_\psi - 1\}, & \displaystyle\int_{\mathbb{R}} t^k \, \psi_0(t) \, \mathrm{d}t = 0, \\ \displaystyle\int_{\mathbb{R}} t^{N_\psi} \, \psi_0(t) \, \mathrm{d}t \neq 0. \end{cases} \tag{1.6}$$

Autrement dit, l'ondelette mère ψ_0 est orthogonale aux polynômes de degré inférieur ou égal à $N_\psi - 1$. Dans l'ensemble du manuscrit, les hypothèses suivantes sur l'ondelette mère ψ_0 sont considérées.

Hypothèse 1. *L'ondelette mère $\psi_0 \in \mathcal{L}^2(\mathbb{R})$ vérifie l'équation (1.6) et*

$$\int_{\mathbb{R}} \psi_0(t)^2 \, \mathrm{d}t = 1. \tag{1.7}$$

Hypothèse 2. *L'ondelette mère ψ_0 est à support compact.*

Hypothèse 3. *Il existe $\alpha > 1$ tel que*

$$\sup_{x \in \mathbb{R}} \left| \hat{\psi}_0(x) \right| (1 + |x|)^\alpha < +\infty, \tag{1.8}$$

où $\hat{\psi}_0$ est la transformée de Fourier de ψ_0.

Sous les hypothèses (1-3), l'ondelette mère ψ_0 est continue et unitaire, et sa transformée de Fourier $\hat{\psi}_0$ est partout différentiable. De plus, par l'équation (1.6), $\hat{\psi}_0$ a ses $N_\psi - 1$ premières dérivées nulles en 0.

On construit alors une famille d'ondelettes $\psi_{j,k}$ par dilatation d'un facteur d'échelle 2^{-j} et translation d'un facteur k de ψ_0,

$$\forall (j,k) \in \mathbb{N} \times \mathbb{Z}, \quad \psi_{j,k}(t) := 2^{-\frac{j}{2}} \psi_0(2^{-j}t - k). \tag{1.9}$$

Le facteur $2^{-\frac{j}{2}}$ est un facteur de normalisation. La famille d'ondelettes $\psi_{j,k}$ forme ainsi une famille orthonormale de $\mathcal{L}^2(\mathbb{R})$.

1.2.4.2 Transformée en ondelettes univariée discrète

La transformée en ondelettes discrète d'un processus stochastique $\{X(t)\}_{t\in\mathbb{R}}$ à trajectoires réelles est définie par

$$\forall (j,k) \in \mathbb{N}\times\mathbb{Z}, \quad D_X(2^j,k) := \langle X, \psi_{j,k}\rangle = \int_{\mathbb{R}} X(t)\,\psi_{j,k}(t)\,\mathrm{d}t, \tag{1.10}$$

où $\langle\cdot,\cdot\rangle$ désigne le produit scalaire usuel sur $\mathcal{L}^2(\mathbb{R})$. Les coefficients $D_X(2^j,k)$ sont appelés *coefficients d'ondelettes* à l'*échelle* 2^j et à l'*instant* k, et l'entier j est appelé *octave*. La transformée en ondelettes fournit donc une série temporelle $\{D_X(2^j,k)\}_{k\in\mathbb{Z}}$ à chaque échelle 2^j. Une introduction détaillée aux transformées en ondelettes est donnée par Mallat (2008).

En pratique, seulement une série temporelle de taille finie $\{X(t)\}_{t\in\{1,\dots,N\}}$ est disponible, ce qui est donc également le cas des coefficients d'ondelettes $\{D_X(2^j,k)\}_{k\in\{1,\dots,n_j\}}$ à chaque échelle 2^j, avec $j \in \{0,\dots,\log_2 N\}$. En particulier, pour une *taille d'échantillon* N grande, le nombre de coefficients d'ondelettes disponibles à chaque échelle 2^j vaut approximativement $n_j \approx N/2^j$. Ceci signifie qu'à grande échelle 2^j, peu de coefficients d'ondelettes sont disponibles. La transformée en ondelettes discrète est illustrée par la figure 1.3.

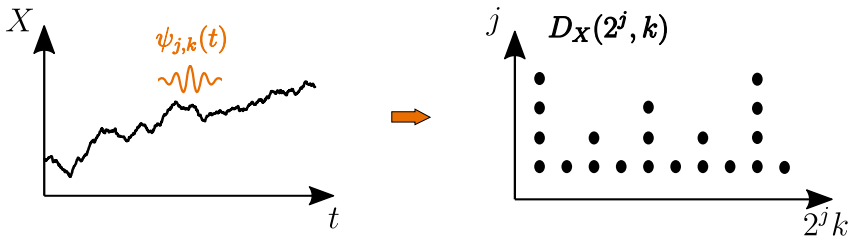


Figure 1.3 – **Illustration de la transformée en ondelettes discrète.** L'ondelette mère ψ_0 est translatée en temps d'un facteur k et dilatée en amplitude d'un facteur 2^j. À chaque octave j et chaque instant k, le produit scalaire du signal X avec l'ondelette mère dilatée et translatée $\psi_{j,k}$ donne un coefficient d'ondelette $D_X(2^j,k)$. Pour un signal X de taille finie, à chaque octave j, deux fois plus de coefficients sont disponibles qu'à l'octave $j+1$.

1.2.4.3 Spectre d'ondelettes

À chaque échelle 2^j, pour tout $j \in \mathbb{N}$, les coefficients d'ondelettes $\{D_{\mathcal{B}_H}(2^j,k)\}_{k\in\mathbb{Z}}$ d'un mBf $\mathcal{B}_H$ ont l'avantage d'être centrés, c'est-à-dire

$$\forall k \in \mathbb{Z}, \quad \mathbb{E}[D_{\mathcal{B}_H}(2^j,k)] = 0, \tag{1.11}$$

et stationnaires (cf. Eq. (1.4)). En particulier, à chaque échelle 2^j, leur variance est constante et s'écrit

$$\forall (j,k) \in \mathbb{N}\times\mathbb{Z}, \quad \mathbb{E}\left[D_{\mathcal{B}_H}(2^j,k)^2\right] = \mathbb{E}\left[D_{\mathcal{B}_H}(2^j,1)^2\right]. \tag{1.12}$$

La fonction

$$2^j \to \mathbb{E}\left[D_{\mathcal{B}_H}(2^j,1)^2\right] \tag{1.13}$$

est appelée *spectre d'ondelettes* de $\mathcal{B}_H$.

Pour une taille d'échantillon N donnée, on définit alors le *spectre d'ondelettes empirique* comme la variance (en temps) empirique des coefficients d'ondelettes $\{D_{\mathcal{B}_H}(2^j,k)\}_{k\in\{1,\dots,n_j\}}$ au

travers des échelles 2^j, c'est-à-dire

$$\forall j \in \{0, \dots, \log_2 N\}, \quad S(2^j) := \frac{1}{n_j} \sum_{k=1}^{n_j} D_{\mathcal{B}_H}(2^j, k)^2, \tag{1.14}$$

où n_j est le nombre de coefficients d'ondelettes disponibles à l'octave j. Le spectre d'ondelettes s'écrit alors $\mathbb{E}[D_{\mathcal{B}_H}(2^j, 1)^2] = \mathbb{E}[S(2^j)]$. Voir Flandrin (1992) pour plus de détails sur la transformée en ondelettes du mBf.

1.2.4.4 Propriétés du spectre d'ondelettes

L'estimation de l'exposant d'autosimilarité H repose sur les propriétés des séries temporelles $\{D_{\mathcal{B}_H}(2^j, k)\}_{k\in\mathbb{Z}}$, issues de la transformée en ondelettes du mBf $\mathcal{B}_H$, au travers des échelles 2^j. En effet, comme montré par Flandrin et Abry (1999), en passant à la transformée en ondelettes dans la relation d'autosimilarité (1.5) vérifiée par le mBf $\mathcal{B}_H$ avec $a = 2^j$, on obtient, pour tout $j \in \mathbb{N}$,

$$\begin{aligned}
\{D_{\mathcal{B}_H}(2^j, k)\}_{k\in\mathbb{Z}} = \left\{ \int_{\mathbb{R}} \mathcal{B}_H(t) \psi_{j,k}(t)\, \mathrm{d}t \right\}_{k\in\mathbb{Z}} &\overset{fdd}{=} \left\{ \int_{\mathbb{R}} 2^{jH} \mathcal{B}_H\left(\frac{t}{2^j}\right) \psi_{j,k}(t)\, \mathrm{d}t \right\}_{k\in\mathbb{Z}} \\
&\overset{fdd}{=} \left\{ \int_{\mathbb{R}} 2^{j(H+1)} \mathcal{B}_H(t)\, \psi_{j,k}(2^j t)\, \mathrm{d}t \right\}_{k\in\mathbb{Z}} \\
&\overset{fdd}{=} \left\{ \int_{\mathbb{R}} 2^{j(H+1)} \mathcal{B}_H(t)\, 2^{-\frac{j}{2}} \psi_0(2^{-j} 2^j t - k)\, \mathrm{d}t \right\}_{k\in\mathbb{Z}} \\
&\overset{fdd}{=} \left\{ \int_{\mathbb{R}} 2^{j(H+\frac{1}{2})}\, \mathcal{B}_H(t)\, \psi_0(t - k)\, \mathrm{d}t \right\}_{k\in\mathbb{Z}},
\end{aligned} \tag{1.15}$$

où la deuxième ligne résulte du changement de variable $t \to 2^j t$ dans l'intégration. D'où l'égalité en distributions de dimension finie suivante :

$$\forall j \in \mathbb{N}, \quad \left\{ D_{\mathcal{B}_H}(2^j, k) \right\}_{k\in\mathbb{Z}} \overset{fdd}{=} \left\{ 2^{j(H+\frac{1}{2})} D_{\mathcal{B}_H}(2^0, k) \right\}_{k\in\mathbb{Z}}. \tag{1.16}$$

Cette égalité signifie que les propriétés statistiques des séries temporelles $\{D_{\mathcal{B}_H}(2^j, k)\}_{k\in\mathbb{Z}}$ sont invariantes par changement d'échelle $2^j \to 2^{j+j'}$ avec un changement d'amplitude $D \to 2^{j'(H+\frac{1}{2})} D$ pour tout $j' \in \mathbb{Z}$. Ainsi, pour tout $k \in \mathbb{Z}$, les fonctions $2^j \to D_{\mathcal{B}_H}(2^j, k)$ satisfont également une relation d'autosimilarité, pouvant être exploitée pour l'estimation de l'exposant d'autosimilarité H.

Étant donnée la relation d'autosimilarité des coefficients d'ondelettes (1.16), le spectre d'ondelettes empirique $S(2^j)$ se comporte comme une loi de puissance par rapport à l'échelle 2^j d'exposant $2H + 1$,

$$\forall j \in \{0, \dots, \log_2 N\}, \quad S(2^j) \overset{d}{=} 2^{j(2H+1)} S(2^0), \tag{1.17}$$

où $\overset{d}{=}$ désigne l'égalité en loi et N une taille d'échantillon donnée.

Finalement, la relation de loi de puissance (1.17) suggère une estimation de l'exposant d'autosimilarité H par régression linéaire sur le logarithme du spectre d'ondelettes empirique $S(2^j)$ au travers des échelles $2^{j_1} \leq 2^j \leq 2^{j_2}$, avec $\{j_1, \dots, j_2\} \subset \{0, \dots, \log_2 N\}$,

$$\hat{H} := \frac{1}{2} \left(\sum_{j=j_1}^{j_2} w_j \log_2 S(2^j) - 1 \right), \tag{1.18}$$

où les poids w_j de la régression linéaire vérifient $\sum j w_j = 1$ et $\sum w_j = 0$ (Veitch et Abry, 1999). Cet estimateur a montré des performances précises et robustes (Veitch et Abry, 1999).

1.3 Mouvement brownien opérateur-fractionnaire (mBof)

1.3.1 Définition

Le mouvement brownien opérateur-fractionnaire (mBof), proposé par Didier et Pipiras (2011, 2012), est une extension du mBf au cadre multivarié. Il est défini comme un processus gaussien M-varié à partir de sa représentation harmonisable,

$$\mathcal{B}_{\underline{\underline{H}},A}(t) := \int_{\mathbb{R}} \frac{e^{itf}-1}{if} \left(f_+^{-(\underline{\underline{H}}-\frac{1}{2}\mathbb{I})} A + f_-^{-(\underline{\underline{H}}-\frac{1}{2}\mathbb{I})} \overline{A} \right) \tilde{B}(\mathrm{d}f), \tag{1.19}$$

où $\mathbb{I}$ est la matrice identité de taille $M \times M$, $f_+ = \max(f,0)$, $f_- = \min(f,0)$, $\tilde{B}$ est une mesure aléatoire gaussienne hermitienne à valeurs complexes telle que $\mathbb{E}[\tilde{B}(\mathrm{d}f)\tilde{B}(\mathrm{d}f)^*] = \mathrm{d}f$ où $*$ est la transposition hermitienne, A est une matrice de taille $M \times M$ à coefficients complexes, $\underline{\underline{H}}$ est une matrice de taille $M \times M$ à coefficients complexes dont les parties réelles des valeurs propres sont strictement comprises entre 0 et 1, $\overline{A}$ désigne la matrice conjuguée de A, et $f^{\underline{\underline{H}}}$ est une matrice de taille $M \times M$ définie par

$$f^{\underline{\underline{H}}} := \sum_{k>0} \frac{\ln(f)^k}{k!} \underline{\underline{H}}^k. \tag{1.20}$$

La matrice $\underline{\underline{H}}$ est appelée *matrice de Hurst*.

1.3.2 Hypothèses et propriétés

Le mBof $\mathcal{B}_{\underline{\underline{H}},A}$ satisfait la relation d'autosimilarité multivariée (Didier et Pipiras, 2011),

$$\forall a > 0, \quad \left\{ \mathcal{B}_{\underline{\underline{H}},A}(t) \right\}_{t\in\mathbb{R}} \overset{fdd}{=} \left\{ a^{\underline{\underline{H}}} \mathcal{B}_{\underline{\underline{H}},A}\left(\frac{t}{a}\right) \right\}_{t\in\mathbb{R}}, \tag{1.21}$$

où $a^{\underline{\underline{H}}}$ est la matrice de taille $M \times M$ définie selon l'équation (1.20). Cette relation signifie que les propriétés statistiques du processus $\mathcal{B}_{\underline{\underline{H}},A}$ sont invariantes par changement d'échelle $t \to t/a$, pour tout facteur de dilatation $a > 0$, avec cette fois-ci un changement d'amplitude matriciel $\mathcal{B} \to a^{\underline{\underline{H}}}\mathcal{B}$, contrairement à la relation d'autosimilarité univariée (1.5). Le comportement du mBof à travers les échelles est donc entièrement contrôlé par la matrice de Hurst $\underline{\underline{H}}$.

De plus, la représentation harmonisable (1.19) du mBof $\mathcal{B}_{\underline{\underline{H}},A}$ permet d'écrire sa fonction de covariance,

$$\begin{aligned} \forall (t,s) \in \mathbb{R}^2, \quad \mathbb{E}\left[\mathcal{B}_{\underline{\underline{H}},A}(t) \mathcal{B}_{\underline{\underline{H}},A}(s)^* \right] = \int_{\mathbb{R}} & \frac{(e^{itf}-1)(e^{-isf}-1)}{|if|^2} \\ & \left(f_+^{-(\underline{\underline{H}}-\frac{1}{2}\mathbb{I})} A A^* f_+^{-(\underline{\underline{H}}^*-\frac{1}{2}\mathbb{I})} \right. \\ & \left. + f_-^{-(\underline{\underline{H}}-\frac{1}{2}\mathbb{I})} \overline{(AA^*)} f_-^{-(\underline{\underline{H}}^*-\frac{1}{2}\mathbb{I})} \right) \mathrm{d}f. \end{aligned} \tag{1.22}$$

La forme de cette fonction de covariance implique que, à l'instar du mBf, le mBof n'est pas stationnaire mais a des incréments stationnaires (cf. Eq. (1.4)).

Hypothèse 4 (OFBM1)**.** *La matrice de Hurst a la forme*

$$\underline{\underline{H}} = W \ \mathrm{diag}(\underline{H}) \ W^{-1}, \tag{1.23}$$

où W est une matrice réelle inversible de taille $M \times M$ et $\mathrm{diag}(\underline{H})$ *est la matrice diagonale de taille $M \times M$ de coefficients réels $H_1, \ldots, H_M$,*

$$\mathrm{diag}(\underline{H}) = \begin{pmatrix} H_1 & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & H_M \end{pmatrix}, \tag{1.24}$$

avec

$$0 < H_1 \leq \ldots \leq H_M < 1. \tag{1.25}$$

L'hypothèse (OFBM1) signifie que la matrice de Hurst $\underline{\underline{H}}$ qui contrôle l'autosimilarité multivariée est réelle et diagonalisable à valeurs propres réelles. La diagonalisation de $\underline{\underline{H}}$ permet de ramener l'analyse de l'autosimilarité multivariée à l'estimation du *vecteur des exposants d'autosimilarité* $\underline{H} = (H_1, \ldots, H_M)$.

Hypothèse 5 (OFBM2)**.**

$$\det \Re(AA^*) > 0. \tag{1.26}$$

Hypothèse 6 (OFBM3)**.**

$$\Im(AA^*) = 0. \tag{1.27}$$

Sous les conditions (OFBM2-3), le mBof $\mathcal{B}_{\underline{\underline{H}},A}$ est identifiable et réversible en temps, c'est-à-dire invariant par changement $t \to -t$, comme stipulé par les propositions suivantes, démontrées dans Didier et Pipiras (2011).

Proposition 1.1 (Identifiabilité)**.** *Si la partie réelle de AA^* est une matrice semi-définie positive alors le mBof $\{\mathcal{B}_{\underline{\underline{H}},A}(t)\}_{t\in\mathbb{R}}$ est un processus identifiable.*

Proposition 1.2 (Réversibilité en temps)**.** *En notant $A = A_1 + iA_2$, le mBof $\{\mathcal{B}_{\underline{\underline{H}},A}(t)\}_{t\in\mathbb{R}}$ est réversible en temps, c'est-à-dire*

$$\left\{\mathcal{B}_{\underline{\underline{H}},A}(t)\right\}_{t\in\mathbb{R}} \overset{fdd}{=} \left\{\mathcal{B}_{\underline{\underline{H}},A}(-t)\right\}_{t\in\mathbb{R}}, \tag{1.28}$$

si et seulement si

$$AA^* = \overline{AA^*} \quad \textit{ou} \quad A_2A_1^* = A_1A_2^*. \tag{1.29}$$

1.3.3 Cas d'une matrice de Hurst diagonale

Un cas particulier d'intérêt est celui où la matrice de Hurst $\underline{\underline{H}}$ est diagonale de taille $M \times M$ de coefficients $H_1, \ldots, H_M$,

$$\underline{\underline{H}} = \begin{pmatrix} H_1 & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & H_M \end{pmatrix}. \tag{1.30}$$

En effet, l'autosimilarité à M variables donnée par l'équation (1.21) se simplifie alors en une collection de M relations d'autosimilarité à une variable données par l'équation (1.5),

$$\forall a > 0, \quad \{\mathcal{B}_1(t), \ldots, \mathcal{B}_M(t)\}_{t\in\mathbb{R}} \overset{fdd}{=} \left\{ a^{H_1}\mathcal{B}_1\left(\frac{t}{a}\right), \ldots, a^{H_M}\mathcal{B}_M\left(\frac{t}{a}\right) \right\}_{t\in\mathbb{R}}, \tag{1.31}$$

où $\mathcal{B}_m$ est la m-ième composante de $\mathcal{B}_{\underline{\underline{H}},A}$. Chaque composante $\mathcal{B}_m$ est autosimilaire et ne dépend alors que d'une seule entrée H_m du vecteur $\underline{H} = (H_1, \ldots, H_M)$.

1.4 Estimation par ondelettes des exposants d'autosimilarité

1.4.1 Analyse en ondelettes multivariée

Comme dans le cadre univarié, l'estimation du vecteur des exposants d'autosimilarité $\underline{H}$ se fait par une représentation multi-échelle. L'analyse en ondelettes multivariée du mBof a été développée dans Abry et Didier (2018b); Abry et collab. (2019).

1.4.1.1 Transformée en ondelettes multivariée discrète

La transformée en ondelettes multivariée d'un processus M-varié $\{Y(t)\}_{t\in\mathbb{R}}$ consiste en une concaténation des transformées en ondelettes des composantes Y_m de Y réalisées avec la même ondelette mère ψ_0,

$$\forall (j,k) \in \mathbb{N}\times\mathbb{Z}, \quad D_Y(2^j,k) := (D_{Y_1}(2^j,k), \ldots, D_{Y_M}(2^j,k)), \tag{1.32}$$

où, pour tout $m \in \{1,\ldots,M\}$, D_{Y_m} désigne la transformée en ondelettes de Y_m donnée par l'équation (1.10). Cette transformée est illustrée par la figure 1.4.

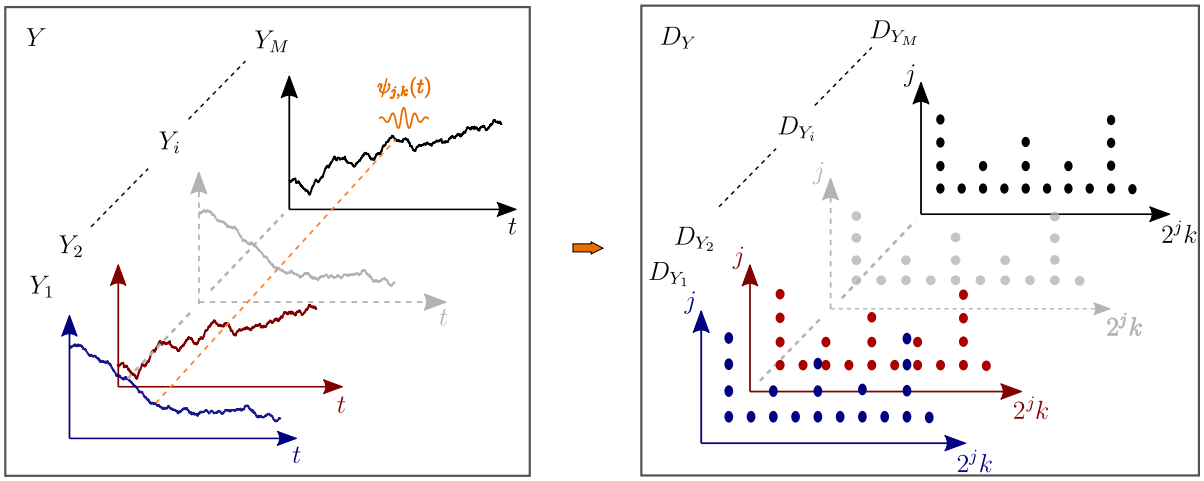


Figure 1.4 – **Illustration de la transformée en ondelettes multivariée.** Pour chaque signal Y_m, des coefficients d'ondelettes $D_{Y_m}(2^j,k)$ sont calculés à chaque octave j et chaque instant k à partir d'une même ondelette mère ψ_0 (dilatée et translatée pour donner $\psi_{j,k}$), comme dans l'illustration 1.3.

1.4.1.2 Spectre d'ondelettes multivarié

Par souci de simplicité, notons $Y = \mathcal{B}_{\underline{\underline{H}},A}$ un mBof. Comme pour la transformée en ondelettes univariée (1.10), à chaque échelle 2^j, pour tout $j \in \mathbb{N}$, les coefficients d'ondelettes multivariés

$\{D_Y(2^j,k)\}_{k\in\mathbb{Z}}$ d'un mBof Y ont l'avantage d'être centrés, c'est-à-dire

$$\forall k\in\mathbb{Z},\quad \mathbb{E}\left[D_{\mathcal{B}_{\underline{\underline{H}},A}}(2^j,k)\right]=0, \tag{1.33}$$

et stationnaires (cf. Eq. (1.4)). En particulier, à chaque échelle 2^j, leur covariance est constante et s'écrit

$$\forall (j,k)\in\mathbb{N}\times\mathbb{Z},\quad \mathbb{E}\left[D_Y(2^j,k)D_Y(2^j,k)^T\right]=\mathbb{E}\left[D_Y(2^j,1)D_Y(2^j,1)^T\right], \tag{1.34}$$

La fonction

$$2^j\to\mathbb{E}\left[D_Y(2^j,1)D_Y(2^j,1)^T\right] \tag{1.35}$$

est appelée *spectre d'ondelettes multivarié* de Y.

Pour une taille d'échantillon N donnée, le *spectre d'ondelettes multivarié empirique* est défini comme la collection de matrices de covariance (en temps) empiriques de taille $M\times M$ des coefficients d'ondelettes $\{D_Y(2^j,k)\}_{k\in\{1,\dots,n_j\}}$ à différentes échelles 2^j, c'est-à-dire

$$\forall j\in\{0,\dots,\log_2 N\},\quad S(2^j):=\frac{1}{n_j}\sum_{k=1}^{n_j}D_Y(2^j,k)D_Y(2^j,k)^T, \tag{1.36}$$

où n_j est le nombre de coefficients d'ondelettes disponibles à chaque octave j. Le spectre d'ondelettes s'écrit alors $\mathbb{E}[D_Y(2^j,1)D_Y(2^j,1)^T]=\mathbb{E}[S(2^j)]$. Ceci est une généralisation du cas univarié, décrit dans la section 1.2, où le spectre d'ondelettes, donné par l'équation (1.14), se résume à un réel.

1.4.1.3 Propriétés du spectre d'ondelettes multivarié

De par la relation d'autosimilarité multivariée (1.21) vérifiée par le mBof Y, des calculs analogues à l'équation (1.15) montrent que la transformée en ondelettes multivariée $\{D_Y(2^j,k)\}_{k\in\mathbb{Z}}$ de Y vérifie une loi de puissance à chaque échelle 2^j, comme suit :

$$\forall j\in\mathbb{N},\quad \left\{D_Y(2^j,k)\right\}_{k\in\mathbb{Z}}\overset{fdd}{=}\left\{2^{j(\underline{\underline{H}}+\frac{1}{2}\mathbb{I})}D_Y(2^0,k)\right\}_{k\in\mathbb{Z}}. \tag{1.37}$$

où $2^{j(\underline{\underline{H}}+\frac{1}{2}\mathbb{I})}$ désigne la matrice de taille $M\times M$ définie selon l'équation (1.20). Cette égalité signifie que les propriétés des statistiques $\{D_Y(2^j,k)\}_{k\in\mathbb{Z}}$ sont invariantes par changement d'échelle $2^j\to 2^{j+j'}$ avec un changement d'amplitude matriciel $D\to 2^{j'(\underline{\underline{H}}+\frac{1}{2}\mathbb{I})}D$ pour tout $j'\in\mathbb{Z}$. Ainsi, les transformées en ondelettes multivariées $\{D_Y(2^j,k)\}_{k\in\mathbb{Z}}$ satisfont également une relation d'autosimilarité à chaque échelle 2^j.

Le spectre d'ondelettes multivarié empirique $S(2^j)$, pour une taille d'échantillon N donnée, vérifie alors l'égalité en loi suivante :

$$\forall j\in\{0,\dots,\log_2 N\},\quad S(2^j)\overset{d}{=}2^{j(\underline{\underline{H}}+\frac{1}{2}\mathbb{I})}\,S(2^0)\,2^{j(\underline{\underline{H}}^T+\frac{1}{2}\mathbb{I})}. \tag{1.38}$$

Sous l'hypothèse 4 (OFBM1), chaque entrée $\mathbb{E}[S_{m,m'}(2^j)]$ du spectre d'ondelettes multivarié $\mathbb{E}[S(2^j)]$ se comporte alors comme un mélange de lois de puissance par rapport à l'échelle 2^j dont les exposants dépendent des exposants d'autosimilarité $H_1,\dots,H_M$, comme suit :

$$\mathbb{E}[S_{m,m'}(2^j)]=\sum_{1\le k\le k'\le M}\alpha_{k,k'}^{(m,m')}\,2^{j(H_k+H_{k'}+1)}, \tag{1.39}$$

pour tous $m, m' \in \{1, \dots, M\}$ et $j \in \{0, \dots, \log_2 N\}$, où les $\alpha_{k,k'}^{(m,m')} \in \mathbb{R}$ dépendent uniquement de $\mathbb{E}[S(2^0)]$.

1.4.2 Estimation univariée

On s'intéresse premièrement au cas d'un mBof $\mathcal{B}_{\underline{\underline{H}},A}$ avec une matrice de Hurst $\underline{\underline{H}}$ diagonale, c'est-à-dire le cas où $W = \mathbb{I}$ dans l'hypothèse 4 (OFBM1). Dans ce cas, comme décrit dans la section 1.3.3, chaque composante du mBof $\mathcal{B}_{\underline{\underline{H}},A}$ vérifie alors une relation d'autosimilarité. Ce cas a été étudié dans Wendt et collab. (2017).

1.4.2.1 Définition

Par souci de simplicité, notons $Y = \mathcal{B}_{\underline{\underline{H}},A}$ un mBof à M composantes. Lorsque $W = \mathbb{I}$, la matrice de Hurst $\underline{\underline{H}}$ est diagonale d'entrées $H_1, \dots, H_M$. Dans ce cas, la relation d'autosimilarité (1.21) vérifiée par Y est alors donnée par l'équation (1.31).

Comme dans l'équation (1.15), par passage à la transformée en ondelettes dans la relation d'autosimilarité (1.31), on obtient une relation de loi de puissance pour les transformées en ondelettes $\{D_{Y_m}(2^j, k)\}_{k\in\mathbb{Z}}$ de chaque composante Y_m de Y à chaque échelle 2^j,

$$\forall m \in \{1, \dots, M\},\ \forall j \in \mathbb{N}, \qquad \left\{D_{Y_m}(2^j, k)\right\}_{k\in\mathbb{Z}} \overset{fdd}{=} \left\{2^{j(H_m+\frac{1}{2})} D_{Y_m}(2^0, k)\right\}_{k\in\mathbb{Z}}. \tag{1.40}$$

Ainsi, les transformées en ondelettes des composantes Y_m de Y vérifient chacune la relation d'autosimilarité univariée (1.5) n'impliquant qu'un seul exposant d'autosimilarité H_m.

Les coefficients diagonaux $S_{m,m}(2^j)$ du spectre d'ondelettes empirique $S(2^j)$ de Y se comportent alors comme des lois de puissance par rapport à l'échelle 2^j d'exposant $2H_m + 1$, c'est-à-dire,

$$S_{m,m}(2^j) \overset{d}{=} S_{m,m}(2^0)\ 2^{j(2H_m+1)}, \tag{1.41}$$

pour tous $m \in \{1, \dots, M\}$ et $j \in \{0, \dots, \log_2 N\}$, avec N une taille d'échantillon donnée. Un estimateur de H_m est alors obtenu par régression linéaire sur les coefficients diagonaux $S_{m,m}(2^j)$ au travers des échelles 2^j, pour chaque $m \in \{1, \dots, M\}$,

$$\hat{H}_m^{(\mathrm{U})} := \frac{1}{2}\left(\sum_{j=j_1}^{j_2} w_j \log_2 S_{m,m}(2^j) - 1\right), \tag{1.42}$$

où $\{j_1, \dots, j_2\} \subset \{0, \dots, \log_2 N\}$ désigne l'intervalle d'octaves impliquées et les poids w_j de la régression linéaire vérifient $\sum j w_j = 1$ et $\sum w_j = 0$ (Veitch et Abry, 1999). Cette approche revient à estimer l'exposant H_m de chaque composante Y_m indépendamment par l'équation (1.18), c'est-à-dire par une analyse univariée.

1.4.2.2 Étude théorique des performances asymptotiques

Les propriétés de l'estimateur univarié (1.42) sont détaillées dans Wendt et collab. (2017), où est notamment énoncé le théorème suivant (Théorème 3.1).

Théorème 1.1 (Consistance et normalité asymptotique)**.** *Supposons que les hypothèses (OFBM1-3) sont vérifiées et que $W = \mathbb{I}$ dans (OFBM1). Alors, lorsque N tend vers $+\infty$,*

$$\left\{\sqrt{N}\left(\hat{H}_m^{(\mathrm{U})} - H_m\right)\right\}_{m\in\{1,\dots,M\}} \xrightarrow{d} \mathcal{N}(0, \Sigma_B), \tag{1.43}$$

où $\Sigma_B \in \mathcal{S}_{\geq 0}(M, \mathbb{R})$.

Le théorème 1.1 assure que, lorsque $W = \mathbb{I}$, l'estimateur univarié $\underline{\hat{H}}^{(\mathrm{U})} := (\hat{H}_1^{(\mathrm{U})}, \ldots, \hat{H}_M^{(\mathrm{U})})$ est consistant et asymptotiquement gaussien lorsque $W = \mathbb{I}$, c'est-à-dire pour des composantes $\mathcal{B}_m$ de mBof vérifiant chacune indépendamment une relation d'autosimilarité caractérisée par l'exposant d'autosimilarité H_m.

Par ailleurs, dans ce cas-ci, des approximations sur la variance et la corrélation de l'estimateur univarié (1.42) sont données par Wendt et collab. (2017).

Théorème 1.2 (Approximation de la covariance)**.** *Supposons que les hypothèses (OFBM1-3) sont vérifiées et que* $W = \mathbb{I}$*. Alors, pour tous* $m, m' \in \{1, \ldots, M\}$*,*

$$\mathrm{Cov}\left(\hat{H}_m^{(\mathrm{U})}, \hat{H}_{m'}^{(\mathrm{U})}\right) \underset{N\to+\infty}{\sim} c_{m,m'}^2 \sum_{j=j_1}^{j_2} \frac{w_j^2}{n_j}, \tag{1.44}$$

où $c_{m,m'} \in \mathbb{R}$*, et, en particulier,*

$$\mathrm{Var}\left(\hat{H}_m^{(\mathrm{U})}\right) \underset{N\to+\infty}{\sim} \frac{1}{2}(\log_2 e)^2 \sum_{j=j_1}^{j_2} \frac{w_j^2}{n_j}. \tag{1.45}$$

Le théorème 1.2 assure que la corrélation entre les estimées univariées $\hat{H}_m^{(\mathrm{U})}$ de H_m et leur variance décroissent en $1/N$ lorsque $W = \mathbb{I}$.

1.4.2.3 Limites de l'estimateur univarié

Lorsque $W \neq \mathbb{I}$, les mélanges de lois de puissance des coefficients diagonaux $\mathbb{E}[S_{m,m}(2^j)]$ du spectre d'ondelettes $\mathbb{E}[S(2^j)]$ à chaque échelle 2^j, donnés par l'équation (1.39), sont tous dominés par la loi de puissance d'exposant $2H_M + 1$,

$$\mathbb{E}[S_{m,m}(2^j)] \underset{j\to+\infty}{\sim} \alpha_{M,M}^{(m,m)}\, 2^{j(2H_M+1)}, \tag{1.46}$$

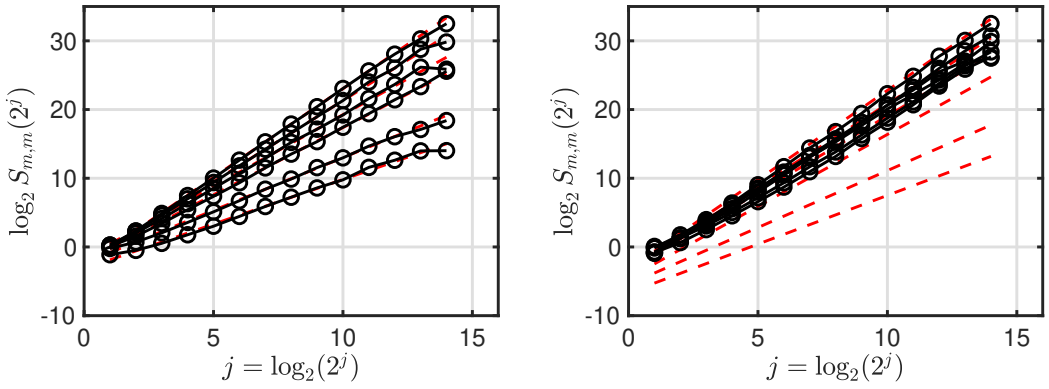
pour tous $m \in \{1, \ldots, M\}$ et $j \in \{0, \ldots, \log_2 N\}$, et les estimées $\hat{H}_m^{(\mathrm{U})}$ approximent alors davantage l'exposant dominant H_M. L'estimateur $\underline{\hat{H}}^{(\mathrm{M})}$ devient donc biaisé pour des exposants H_m différents, comme illustré par la figure 1.5.

Toutefois, lorsque $W \neq \mathbb{I}$ et tous les exposants d'autosimilarité $H_1, \ldots, H_M$ sont égaux, le mélange de lois de puissance (1.39) des $\mathbb{E}[S_{m,m}(2^j)]$ devient, pour tous $m \in \{1, \ldots, M\}$ et $j \in \{0, \ldots, \log_2 N\}$,

$$\mathbb{E}[S_{m,m}(2^j)] = \left(\sum_{1\leq k\leq k'\leq M} \alpha_{k,k'}^{(m,m)}\right) 2^{j(2H_m+1)}. \tag{1.47}$$

Ainsi, les entrées $\mathbb{E}[S_{m,m}(2^j)]$ se comportent comme des lois de puissance par rapport à l'échelle 2^j d'exposant $2H_m+1$ et l'estimateur univarié $\underline{\hat{H}}^{(\mathrm{U})}$ n'est donc pas biaisé dans ce cas particulier.

En pratique, ne considérer que les entrées diagonales $S_{m,m}(2^j)$ du spectre d'ondelettes empirique $S(2^j)$ revient en fait à négliger les dépendances temporelles entre les séries temporelles observées. Pour tenir compte de ces dépendances temporelles, une approche est de considérer les entrées non diagonales du spectre d'ondelettes empirique (Wendt et collab., 2017). Pour une



(a) $W = \mathbb{I}$ (b) $W \neq \mathbb{I}$

FIGURE 1.5 – **Illustration de l'estimateur univarié.** L'estimation univariée est réalisée par des régressions linéaires sur les logarithmes $\log_2 S_{m,m}(2^j)$ des coefficients diagonaux du spectre d'ondelettes $S(2^j)$. Lorsque la matrice de Hurst $\underline{\underline{H}}$ est diagonale (à gauche), les pentes des $\log_2 \mathbb{E}[S_{m,m}(2^j)]$ (en lignes rouges pointillées) sont bien estimées par les $\log_2 S_{m,m}(2^j)$ (en noir) tandis que, lorsque la matrice de Hurst $\underline{\underline{H}}$ n'est pas diagonale (à droite), les pentes des $\log_2 S_{m,m}(2^j)$ sont similaires alors que les H_m sont tous différents.

matrice de Hurst $\underline{\underline{H}}$ diagonale ($W = \mathbb{I}$), les entrées non diagonales $S_{m,m'}(2^j)$ de $S_Y(2^j)$ se comportent asymptotiquement comme des lois de puissance, avec un exposant d'échelle $2H_{m,m'} + 1$, où $H_{m,m'} = (H_m + H_{m'})/2$, à chaque échelle 2^j. Ceci conduit naturellement à estimer $H_{m,m'}$ par régression linéaire au travers des échelles $2^{j_1} \leq 2^j \leq 2^{j_2}$, pour tous $m, m' \in \{1, \ldots, M\}$,

$$\hat{H}^{(\mathrm{U})}_{m,m'} = \frac{1}{2}\left(\sum_{j=j_1}^{j_2} w_j \log_2 |S_{m,m'}(2^j)|\right) - \frac{1}{2}, \tag{1.48}$$

où $\{j_1, \ldots, j_2\} \subset \{0, \ldots, \log_2 N\}$ désigne l'intervalle d'octaves impliquées, et où les poids w_j des régressions linéaires satisfont $\sum j w_j = 1$ et $\sum w_j = 0$ (VEITCH et ABRY, 1999). Cette approche multivariée est étudiée dans le cadre d'applications dans le chapitre 5.

1.4.3 Estimation multivariée

1.4.3.1 Définition

On s'intéresse à présent au cas plus général d'un mBof $Y = \mathcal{B}_{\underline{\underline{H}},A}$ dont les exposants d'autosimilarité peuvent être mélangés, i.e. la matrice W peut être telle que $W \neq \mathbb{I}$ dans l'hypothèse 4 (OFBM1). Autrement dit, la matrice de Hurst $\underline{\underline{H}}$ du mBof est diagonalisable mais pas nécessairement diagonale. Dans ce cas, un estimateur construit à partir des valeurs propres du spectre d'ondelettes empirique a été proposé dans ABRY et DIDIER (2018a).

L'estimateur multivarié repose sur la propriété suivante des valeurs propres $\lambda_1(2^j) \leq \ldots \leq \lambda_M(2^j)$ du spectre d'ondelettes $\mathbb{E}[S(2^j)]$ d'un mBof Y : chaque valeur propre $\lambda_m(2^j)$ se comporte asymptotiquement comme une loi de puissance par rapport à l'échelle 2^j d'exposant $2H_m + 1$. Cette propriété, démontrée dans ABRY et DIDIER (2018a) (Théorème 2), s'énonce de la façon suivante.

Théorème 1.3. *Soit Y un mBof vérifiant les conditions (OFBM1-3). Alors les valeurs propres*

$\lambda_m(2^j)$ du spectre d'ondelettes $\mathbb{E}[S(2^j)]$ de Y vérifient, pour tout $m \in \{1,\dots,M\}$,

$$\lambda_m(2^j) \underset{j\to+\infty}{\sim} \xi_m(2^0)\, 2^{j(2H_m+1)}, \tag{1.49}$$

où $\xi_m(2^0) > 0$.

Notons $\hat{\lambda}_1(2^j),\dots,\hat{\lambda}_M(2^j)$ les valeurs propres du spectre d'ondelettes empirique $S(2^j)$ de Y à l'octave $j \in \{0,\dots,\log_2 N\}$, pour une taille d'échantillon N donnée. D'après le résultat précédent, les logarithmes $\log_2 \lambda_m(2^j)$ des valeurs propres $\lambda_m(2^j)$ du spectre d'ondelettes $\mathbb{E}[S(2^j)]$ de Y sont asymptotiquement affines de pente $2H_m+1$ à grande échelle 2^j. Cela suggère naturellement une procédure d'estimation de H_m reposant sur des régressions linéaires sur les logarithme des valeurs propres estimées $\hat{\lambda}_m(2^j)$ au travers d'échelles $2^{j_1} \le 2^j \le 2^{j_2}$, pour tout $m \in \{1,\dots,M\}$, comme suit :

$$\hat{H}_m^{(\mathrm{M})} = \frac{1}{2}\left(\sum_{j=j_1}^{j_2} w_j \log_2 \hat{\lambda}_m(2^j) - 1\right), \tag{1.50}$$

où $\{j_1,\dots,j_2\} \subset \{0,\dots,\log_2 N\}$ désigne l'intervalle d'octaves impliquées, et où les poids w_j des régressions linéaires satisfont $\sum j w_j = 1$ et $\sum w_j = 0$ (Veitch et Abry, 1999). Les performances théoriques de cet estimateur ont été étudiées dans Abry et Didier (2018a,b).

1.4.3.2 Condition pour l'étude asymptotique

La condition suivante joue un rôle essentiel dans ce manuscrit. Elle est en effet impliquée dans certains théorèmes de convergence.

Hypothèse 7 (Condition (C0))**.** *Pour tous $\ell_1,\ell_2 \in \{1,\dots,M\}$,*

$$\text{soit } H_{\ell_1} \neq H_{m_2} \text{ avec } m_1 \neq m_2, \quad \text{soit } \xi_{\ell_1}(2^0) \neq \xi_{\ell_2}(2^0). \tag{1.51}$$

où $\xi_1(2^0),\dots,\xi_M(2^0)$ sont définis dans le théorème 1.3.

Au vu des comportements asymptotiques des valeurs propres $\lambda_m(2^j)$ de $\mathbb{E}[S(2^j)]$ donnés par l'équation (1.49), la condition (C0) signifie que les valeurs propres $\lambda_m(2^j)$ sont simples pour j grand. Sous cette condition, il est montré par la suite que, dans un certain cadre, l'estimateur multivarié $\underline{\hat{H}}^{(\mathrm{M})}$ est asymptotiquement normal. Cette condition est également considérée dans le chapitre 2 pour démontrer des propriétés importantes sur l'estimateur qui y est introduit, telles que sa normalité et sa décorrélation asymptotiques.

1.4.3.3 Cadre asymptotique d'étude

L'étude théorique de l'estimateur $\underline{\hat{H}}^{(\mathrm{M})} := (\hat{H}_1^{(\mathrm{M})},\dots,\hat{H}_M^{(\mathrm{M})})$ est réalisée pour un mBof Y et une taille d'échantillon N qui tend vers l'infini. Étant donné le caractère asymptotique des relations de loi de puissance vérifiées par les valeurs propres $\lambda_m(2^j)$ (cf. Eq. (1.49)), l'estimation est étudiée dans le cas où les échelles d'analyse $2^{j_1(N)},\dots,2^{j_2(N)}$ impliquées dans les régressions linéaires croissent avec la taille d'échantillon N. En effet, plus la taille d'échantillon N est grande, plus les échelles d'analyse $1 \le 2^{j_1(N)} \le 2^j \le 2^{j_2(N)} \le N$ peuvent être choisies grandes.

Précisément, les échelles d'analyses sont définies comme suit. Soient $j_1^0, j_2^0 \in \mathbb{N}$ et $a(N)$ une

suite dyadique de facteurs d'échelle telle que

$$\begin{cases} \forall N \in \mathbb{N},\ a(N)2^{j_2^0} \leq N, \\ \dfrac{a(N)}{N} + \dfrac{N}{a(N)^{2\varpi+1}} \underset{N\to\infty}{\longrightarrow} 0, \end{cases} \tag{1.52}$$

où ϖ est le paramètre de régularité,

$$\varpi = \min\left\{ \min_{\{i\in\{2,\dots,M\}|\ H_i - H_{i-1} > 0\}} (H_i - H_{i-1}), \frac{1}{2}H_1 + \frac{1}{4} \right\}. \tag{1.53}$$

Pour une taille d'échantillon observé N, la gamme d'échelles sur laquelle sont effectuées les régressions linéaires est donnée par $\{2^{j_1(N)}, \dots, 2^{j_2(N)}\} := \{a(N)2^{j_1^0}, \dots, a(N)2^{j_2^0}\}$.

Autrement dit, les plus petite et grande octaves utilisées à une taille d'échantillon N sont respectivement données par $j_1(N) = j_1^0 + \log_2 a(N)$ et $j_2(N) = j_2^0 + \log_2 a(N)$, et le nombre d'octaves impliquées dans l'analyse est donc constant, donné par $j_2^0 - j_1^0 + 1$. Ainsi, les régressions linéaires sont réalisées sur les échelles $2^{j_1^0}, \dots, 2^{j_2^0}$ à une taille d'échantillon N_0 telle que $a(N_0) = 1$. Les octaves j_1^0 et j_2^0 doivent donc être choisies de sorte que $0 \leq j_1^0 \leq j_2^0 \leq \log_2 N_0$.

Avec ces définitions, le spectre d'ondelettes empirique $S(a(N)2^j)$ de Y est calculé à partir d'un nombre de coefficients d'ondelettes $n_{a,j}$ à l'échelle $a(N)2^j$, pour tout $j \in \{j_1^0, \dots, j_2^0\}$, défini par

$$n_{a,j} = \frac{N}{a(N)2^j}. \tag{1.54}$$

Le théorème 1.3 s'écrit alors, pour tous $m \in \{1, \dots, M\}$ et $j \in \{j_1^0, \dots, j_2^0\}$,

$$\frac{\lambda_m(a(N)2^j)}{(a(N)2^j)^{2H_m+1}} \underset{N\to+\infty}{\longrightarrow} \xi_m(2^0), \tag{1.55}$$

où les $\lambda_m(a(N)2^j)$ sont les valeurs propres du spectre d'ondelettes empirique $\mathbb{E}[S(a(N)2^j)]$ de Y et $\xi_m(2^0) > 0$ (Abry et Didier, 2018a). L'estimateur multivarié (1.50) s'écrit alors simplement, pour tout $m \in \{1, \dots, M\}$,

$$\hat{H}_m^{(\mathrm{M})} = \frac{1}{2}\left(\sum_{j=j_1^0}^{j_2^0} w_j \log_2 \hat{\lambda}_m(a(N)2^j) - 1 \right). \tag{1.56}$$

1.4.3.4 Étude théorique des performances asymptotiques

Le comportement en loi de puissance asymptotique des logarithmes des valeurs propres estimées $\log_2 \hat{\lambda}_m(a(N)2^j)$ est assuré par le théorème suivant, démontré dans Abry et collab. (2022) (Théorème 3.1).

Théorème 1.4. *Soit Y un mBof vérifiant les conditions (OFBM1-3). Alors les valeurs propres $\hat{\lambda}_m(a(N)2^j)$ du spectre d'ondelettes empirique $S(a(N)2^j)$ de Y vérifient, pour tous $m \in \{1, \dots, M\}$ et $j \in \{j_1^0, \dots, j_2^0\}$, lorsque N tend vers $+\infty$,*

$$\frac{\hat{\lambda}_m(a(N)2^j)}{(a(N)2^j)^{2H_m+1}} \xrightarrow{\mathbb{P}} \xi_m(2^0), \tag{1.57}$$

où $\xi_m(2^0) > 0$ est défini dans le théorème 1.3.

Cela signifie que les $\log_2 \hat{\lambda}_m(a(N)2^j)$ ont asymptotiquement le même comportement que les $\log_2 \lambda_m(a(N)2^j)$: les $\log_2 \hat{\lambda}_m(a(N)2^j)$ sont asymptotiquement affines d'exposant $2H_m + 1$. Ce résultat assure la consistance des estimateurs $\hat{H}_m^{(\mathrm{M})}$ définis par l'équation (1.50) dans un cadre assez général, énoncée par le théorème suivant.

Théorème 1.5 (Consistance)**.** *Supposons que les conditions (OFBM1-3) sont vérifiées. Alors, pour tout $m \in \{1, \dots, M\}$, lorsque N tend vers $+\infty$,*

$$\hat{H}_m^{(\mathrm{M})} \xrightarrow{\mathbb{P}} H_m. \tag{1.58}$$

Il est également démontré dans Abry et collab. (2022) (Théorème 3.2) que la normalité conjointe asymptotique des logarithmes des valeurs propres estimées $\log_2 \hat{\lambda}_1(a(N)2^j)$, …, $\log_2 \hat{\lambda}_M(a(N)2^j)$ est assurée pour chaque octave $j \in \{j_1^0, \dots, j_2^0\}$, mais sous l'hypothèse supplémentaire que la condition (C0) est vérifiée.

Théorème 1.6. *Soit Y un mBof vérifiant les conditions OFBM1-3) et (C0). Alors les valeurs propres $\hat{\lambda}_m(a(N)2^j)$ du spectre d'ondelettes empirique $S(a(N)2^j)$ de Y vérifient, pour tous $m \in \{1, \dots, M\}$ et $j \in \{j_1^0, \dots, j_2^0\}$, lorsque N tend vers $+\infty$,*

$$\left\{ \sqrt{n_{a,j}} \left(\log_2 \hat{\lambda}_m(a(N)2^j) - \log_2 \lambda_m(a(N)2^j) \right) \right\}_{m \in \{1,\dots,M\}}^{j \in \{j_1^0, \dots, j_2^0\}} \xrightarrow{d} \mathcal{N}(0, \Sigma_\lambda(2^j)), \tag{1.59}$$

où $\Sigma_\lambda(2^j) = \frac{1}{n_{a,j}} \Sigma_\lambda(2^0) \in \mathcal{S}_{\geq 0}(M(j_2^0 - j_1^0 + 1), \mathbb{R})$.

Ce théorème signifie que le vecteur $(\log_2 \hat{\lambda}_m(a(N)2^j))_{1 \leq m \leq M,\, j_1^0 \leq j \leq j_2^0}$ est asymptotiquement gaussien si les valeurs propres $\lambda_1(a(N)2^j), \dots, \lambda_M(a(N)2^j)$ sont asymptotiquement distinctes à chaque échelle $a(N)2^j \in \{a(N)2^{j_1^0}, \dots, a(N)2^{j_2^0}\}$. La normalité multivariée asymptotique de l'estimateur $\underline{\hat{H}}^{(\mathrm{M})} := (\hat{H}_1^{(\mathrm{M})}, \dots, \hat{H}_M^{(\mathrm{M})})$ est donc assurée sous cette hypothèse, comme énoncé par le théorème suivant.

Théorème 1.7 (Normalité asymptotique)**.** *Supposons que les conditions (OFBM1-3) et (C0) sont vérifiées. Alors, lorsque N tend vers $+\infty$,*

$$\left\{ \sqrt{\frac{N}{a(N)}} \left(\hat{H}_m^{(\mathrm{M})} - H_m \right) \right\}_{m \in \{1,\dots,M\}} \xrightarrow{d} \mathcal{N}(0, \Sigma_B), \tag{1.60}$$

avec $\Sigma_B \in \mathcal{S}_{\geq 0}(M, \mathbb{R})$.

1.5 Conclusion

Dans ce chapitre, ont été introduits les éléments essentiels à l'analyse de l'autosimilarité multivariée, modélisée de façon pertinente par un mBof et entièrement caractérisée par la matrice de Hurst $\underline{\underline{H}}$. La matrice de Hurst $\underline{\underline{H}}$ est supposée diagonalisable, ce qui ramène l'analyse de l'autosimilarité multivariée à une estimation d'un vecteur d'exposants d'autosimilarité $\underline{H} = (H_1, \dots, H_M)$ où M est le nombre de composantes du mBof. Pour l'estimation, des outils d'analyse en ondelettes multivariée ont été présentés et deux approches ont été introduites.

Le premier estimateur, dit univarié, est adapté au cas d'un mBof dont chaque entrée est autosimilaire, c'est-à-dire caractérisé par une matrice de Hurst $\underline{\underline{H}}$ diagonale. Cet estimateur montre des performances théoriques tout à fait satisfaisantes dans ce cas-ci, puisqu'il est consistant et

asymptotiquement normal. En revanche, il montre des limites importante lorsque $\underline{\underline{H}}$ n'est pas diagonale. Le second estimateur, dit multivarié, s'adapte au cas plus général de l'autosimilarité multivariée. Cet estimateur est également consistant et asymptotiquement normal sous des hypothèses faibles, et ce pour une matrice de Hurst $\underline{\underline{H}}$ diagonalisable mais pas nécessairement diagonale.

Chapitre 2

Étude et correction de l'estimateur multivarié des exposants d'autosimilarité

Sommaire

2.1 Introduction

La procédure d'estimation multivariée du vecteur des exposants d'autosimilarité $\underline{H} = (H_1, \ldots, H_M)$ décrite dans la section 1.4.3 présente des performances pratiques très satisfaisantes mais peut souffrir d'un biais d'estimation de taille finie dans certaines situations. En effet, les estimées multivariées $\hat{H}_m^{(\mathrm{M})}$ des H_m souffrent d'un biais lié à l'*effet de repulsion*, bien documenté dans la littérature sur les matrices aléatoires (Anderson et collab., 2010), qui provient du calcul des valeurs propres $\hat{\lambda}_m(2^j)$ du spectre d'ondelettes multivarié empirique $S(2^j)$. Précisément, les biais des valeurs propres estimées $\hat{\lambda}_m(2^j)$ dépendent de l'échelle 2^j, induisant des biais dans les régressions linéaires permettant d'obtenir les estimées $\hat{H}_m^{(\mathrm{M})}$. Pour atténuer ce biais, ce chapitre propose une toute nouvelle procédure d'estimation et s'intéresse théoriquement et numériquement à ses propriétés en les comparant à celles des estimateurs univarié et multivarié introduits dans le chapitre 1.

En premier lieu, la procédure d'estimation proposée est détaillée dans la section 2.2. Les performances asymptotiques théoriques de cet estimateur sont ensuite énoncées dans la section 2.3. Ensuite, ce chapitre introduit un modèle issu du mouvement brownien opérateur-fractionnaire (mBof), plus simple et adapté à des applications pratiques, nommé mouvement brownien fractionnaire multivarié (M-mBf), dans la section 2.4. Ce modèle permet notamment une étude numérique approfondie, menée dans la section 2.5 sur des simulations de Monte Carlo recourant à des M-mBf synthétiques. L'évaluation des performances asymptotiques empiriques de l'estimateur proposé et la comparaison de celles-ci avec celles des estimateurs univarié et multivarié sont abordées. En particulier, les impacts des paramètres du modèle sur les performances d'estimation sont étudiés et illustrés. Les résultats montrent l'efficacité pratique de la procédure originale d'estimation proposée.

Les résultats présentés dans ce chapitre sont l'objet d'un article en cours d'écriture intitulé « Multivariate self-similarity: Multiscale eigen structures for the estimation of Hurst exponents », par C.-G. Lucas, G. Didier, H. Wendt et P. Abry, prévu pour être soumis à une revue internationale.

2.2 Correction du biais de taille finie de l'estimateur multivarié

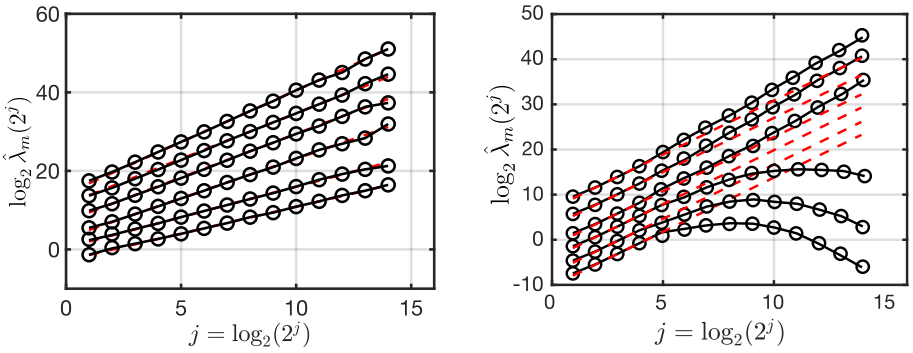
Dans cette section, on considère un mBof $Y = \mathcal{B}_{\underline{\underline{H}},A}$ à M composantes, comme défini dans la section 1.3.1, et une taille d'échantillon N. On suppose que Y vérifie les hypothèses de travail (OFBM1-3) définies dans la section 1.3.2, et est donc caractérisé par un vecteur d'exposants d'autosimilarité $\underline{H} = (H_1, \ldots, H_M)$. On note $D_Y(2^j, k)$ le coefficient d'ondelettes multivarié associé à Y à l'échelle 2^j et à l'instant k donné par l'équation (1.32) et $S(2^j)$ le spectre d'ondelettes multivarié empirique résultant défini par l'équation (1.36).

L'estimateur multivarié $\hat{\underline{H}}^{(\mathrm{M})} = (\hat{H}_1^{(\mathrm{M})}, \ldots, \hat{H}_M^{(\mathrm{M})})$ du vecteur des exposants d'autosimilarité $\underline{H}$ défini par l'équation (1.50) souffre d'un biais de taille finie dû à l'effet de répulsion, expliqué ici. La présente section propose de modifier la procédure d'estimation introduite dans la section 1.4.3 pour atténuer ce biais.

2.2.1 Effet de répulsion

La construction de l'estimateur multivarié $\hat{\underline{H}}^{(\mathrm{M})}$ du vecteur des exposants d'autosimilarité $\underline{H}$ repose sur la propriété des M valeurs propres $\lambda_m(2^j)$ du spectre d'ondelettes multivarié $\mathbb{E}[S(2^j)]$

donnée par le théorème 1.3 : chaque valeur propre $\lambda_m(2^j)$ suit une loi de puissance par rapport à l'échelle 2^j d'exposant $2H_m + 1$. Les M estimées multivariées $\hat{H}_m^{(\mathrm{M})}$ sont ainsi obtenues par des régressions linéaires sur les logarithmes des valeurs propres $\hat{\lambda}_m(2^j)$ du spectre d'ondelettes empirique $S(2^j)$ au travers des échelles 2^j. Or, chaque matrice de covariance $S(2^j)$ est calculée à partir d'un nombre $n_j \approx N/2^j$ de coefficients d'ondelettes $D_Y(2^j, k)$ qui varie au travers des échelles 2^j, ce qui produit ainsi des biais dépendants de l'échelle dans $\hat{\lambda}_m(2^j)$. En effet, les valeurs propres $\hat{\lambda}_m(2^j)$ souffrent de l'effet de répulsion : l'écart entre elles est plus grand que les valeurs propres exactes $\lambda_m(2^j)$ (Tao, 2012), et ce de façon d'autant plus importante que le rapport M/n_j est grand (cf. Yao et collab. (2015)). De plus, ce biais est accentué pour des valeurs propres $\lambda_m(2^j)$ proches, voire égales. Ceci implique nécessairement un biais de taille finie dans $\hat{H}_m^{(\mathrm{M})}$, notamment pour des exposants H_m égaux. Cet effet est illustré par la figure 2.1.



(a) H_m tous différents (b) H_m tous égaux

FIGURE 2.1 – **Illustration du biais de l'estimateur multivarié.** L'estimation $M = 6$-variée est réalisée par des régressions linéaires sur les logarithmes $\log_2 \lambda_m(2^j)$ des valeurs propres $\lambda_m(2^j)$ du spectre d'ondelettes $S(2^j)$. Lorsque les H_m sont tous différents (à gauche), les pentes des $\log_2 \hat{\lambda}_1(2^j), \ldots, \log_2 \hat{\lambda}_M(2^j)$ (en noir) sont peu affectées par l'effet de répulsion croissant, se superposant alors bien avec les $\log_2 \lambda_1(2^j), \ldots, \log_2 \lambda_M(2^j)$ (en lignes rouges pointillées). En revanche, lorsque les H_m sont tous égaux (à droite), les $\log_2 \hat{\lambda}_1(2^j), \ldots, \log_2 \hat{\lambda}_M(2^j)$ s'écartent de plus en plus les uns des autres et ne forment alors plus des droites.

2.2.2 Spectres d'ondelettes multivariés empiriques de fenêtres

Pour réduire le biais induit par l'effet de répulsion, une contribution originale du présent travail est d'estimer des matrices de covariance à partir d'un même nombre de coefficients d'ondelettes à chaque échelle, sans pour autant augmenter la variance d'estimation, de la façon suivante. On considère la plus grande octave $j_2 \in \{1, \ldots, \log_2 N\}$ qui servira dans les régressions linéaires pour l'estimation, comme dans l'équation (1.50) définissant l'estimateur multivarié $\underline{\hat{H}}^{(\mathrm{M})}$. À chaque échelle 2^j, une collection de 2^{j_2-j} matrices de covariance $S^{(w)}(2^j)$ est estimée à partir de fenêtres temporelles F_w de coefficients d'ondelettes $\{D_Y(2^j, k)\}_{k \in F_w}$ ne se chevauchant pas de tailles égales $\mathrm{Card}(F_w) = n_{j_2}$, pour tous $w \in \{1, \ldots, 2^{j_2-j}\}$ et $j \in \{1, \ldots, j_2\}$,

$$S^{(w)}(2^j) := \frac{1}{n_{j_2}} \sum_{k \in F_w} D_Y(2^j, k) D_Y(2^j, k)^T, \tag{2.1}$$

où $F_w = \{1 + (w-1)\, n_{j_2}, \ldots, w\, n_{j_2}\}$ et n_{j_2} est le nombre de coefficients d'ondelettes disponibles à l'échelle 2^{j_2}. Cette procédure est illustrée par la figure 2.2.

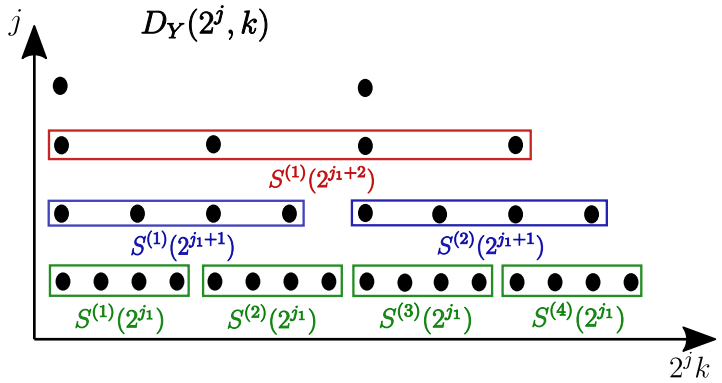


FIGURE 2.2 – **Illustration de la stratégie de correction du biais.** À l'échelle $2^{j_2} = 2^{j_1+2}$, le spectre d'ondelettes multivarié empirique $S^{(1)}(2^{j_1+2})$ est calculé à partir de tous les coefficients d'ondelettes $D_Y(2^j, \cdot)$ disponibles, soit $n_{j_1+2} = 4$ coefficients ici. À l'échelle inférieure 2^{j_1+1}, deux spectres d'ondelettes multivariés empiriques $S^{(1)}(2^{j_1+1})$ et $S^{(2)}(2^{j_1+1})$ sont calculés à partir de deux groupes de coefficients $D_Y(2^j, \cdot)$ de même taille $n_{j_1+1}/2 = n_{j_1+2} = 4$. À l'échelle inférieure 2^{j_1+2}, quatre spectres d'ondelettes empiriques $S^{(1)}(2^{j_1+1}), \ldots, S^{(4)}(2^{j_1+1})$, sont calculés à partir de quatre groupes de coefficients $D_Y(2^j, \cdot)$ de même taille $n_{j_1}/4 = n_{j_1+2} = 4$.

Les valeurs propres $\hat{\lambda}_1^{(w)}(2^j), \ldots, \hat{\lambda}_M^{(w)}(2^j)$ des $S^{(w)}(2^j)$ devraient ainsi être affectées par des effets de répulsion similaires à toutes les octaves $j \in \{1, \ldots, j_2\}$ et toutes les fenêtres $w \in \{1, \ldots, 2^{j_2-j}\}$, contrairement aux valeurs propres $\hat{\lambda}_1(2^j), \ldots, \hat{\lambda}_M(2^j)$ de $S(2^j)$.

2.2.3 Estimateur multivarié corrigé

Pour obtenir des estimées corrigées $\underline{\hat{H}}^{(\mathrm{M,bc})} = (\hat{H}_1^{(\mathrm{M,bc})}, \ldots, \hat{H}_M^{(\mathrm{M,bc})})$ de $\underline{H} = (H_1, \ldots, H_M)$, les régressions linéaires sont à présent effectuées sur les logarithmes des valeurs propres $\hat{\lambda}_1^{(w)}(2^j)$, $\ldots$, $\hat{\lambda}_M^{(w)}(2^j)$ des spectres d'ondelettes $S^{(w)}(2^j)$ moyennés au travers des fenêtres $w \in \{1, \ldots, 2^{j_2-j}\}$, notés

$$\log_2 \bar{\lambda}_m(2^j) := \frac{1}{2^{j_2-j}} \sum_{w=1}^{2^{j_2-j}} \log_2 \hat{\lambda}_m^{(w)}(2^j), \tag{2.2}$$

contre des échelles d'analyse $2^{j_1} \leq 2^j \leq 2^{j_2}$,

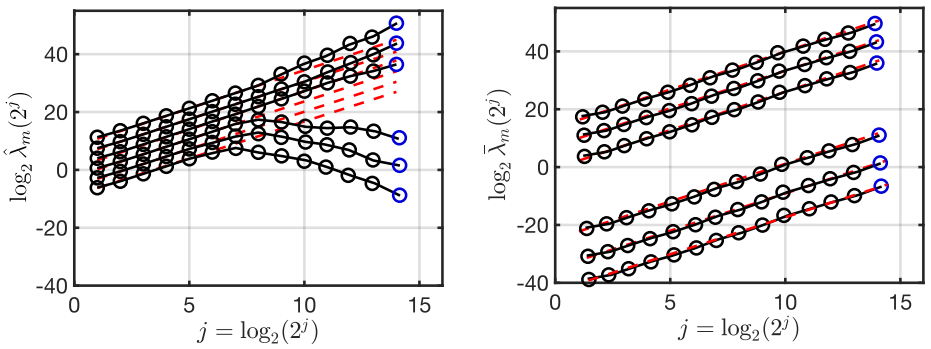
$$\hat{H}_m^{(\mathrm{M,bc})} = \frac{1}{2} \left(\sum_{j=j_1}^{j_2} w_j \log_2 \bar{\lambda}_m(2^j) - 1 \right), \tag{2.3}$$

pour tout $m \in \{1, \ldots, M\}$, où $\{j_1, \ldots, j_2\} \subset \{1, \ldots, \log_2 N\}$ désigne l'intervalle d'octaves impliquées et les poids w_j des régressions linéaires satisfont $\sum j w_j = 1$ et $\sum w_j = 0$ (VEITCH et ABRY, 1999).

Les comportements des fonctions de structure $\log_2 \hat{\lambda}_m(2^j)$ et $\log_2 \bar{\lambda}_m(2^j)$ associées respectivement à l'estimateur multivarié et sa correction sont illustrés par la figure 2.3. Des fonctions de structures issues de simulations numériques sont étudiées dans la section 2.5 portant sur les performances empiriques des estimateurs.

Comme on le remarque sur cette illustration, les biais des $\log_2 \hat{\lambda}_m(2^j)$ et des $\log_2 \bar{\lambda}_m(2^j)$ sont les mêmes à l'échelle 2^{j_2} mais sont différents aux échelles inférieures. L'effet de répulsion entre les $\log_2 \bar{\lambda}_1(2^j)$, $\ldots$, $\log_2 \bar{\lambda}_M(2^j)$ est en effet reproduit à chaque échelle 2^j, de sorte que les $\log_2 \bar{\lambda}_m(2^j)$ en fonction de 2^j soient bien approximés par des droites. Cela implique que les biais

des $\log_2 \bar{\lambda}_m(2^j)$ sont plus importants que les biais des $\log_2 \bar{\lambda}_m(2^j)$ à chaque échelle $2^j < 2^{j_2}$ et ce d'autant plus que l'échelle 2^j est petite. En particulier, les ordonnées à l'origine des régressions linéaires réalisées sur les échelles $2^{j_1} \leq 2^j \leq 2^{j_2}$ ne sont pas les mêmes avant et après correction. La procédure d'estimation corrigée permet donc de réduire le biais d'estimation des H_m mais accentue le biais d'estimation des $\log_2 \xi_m(2^0)$, où $\xi_m(2^0)$ est la constante de proportionnalité dans la loi de puissance des valeurs propres $\lambda_m(2^j)$ donnée par l'équation (1.49). Ces constantes peuvent cependant être assez bien estimées par les $\log_2 \hat{\lambda}_m(2^1)$.



(a) Estimateur multivarié (b) Estimateur multivarié corrigé

FIGURE 2.3 – **Illustration de la correction de l'estimateur multivarié.** L'estimation $M = 6$ variée (à gauche) est réalisée par des régressions linéaires sur les $\log_2 \hat{\lambda}_m(2^j)$ (en noir) qui ont des biais différents aux différentes échelles 2^j et dévient donc des droites de pente H_m (lignes rouges pointillées). L'estimation multivariée corrigée (à droite) se fait à partir des $\log_2 \bar{\lambda}_m(2^j)$ (en noir) qui ont des effets de répulsion similaires à toutes les échelles 2^j, par reproduction du biais des $\log_2 \hat{\lambda}_m(2^{j_2})$ avec $j_2 = 15$ (en bleu), formant ainsi davantage des droites.

2.3 Performances théoriques des estimateurs multivariés

Dans le cadre d'une collaboration avec Gustavo DIDIER, maître de conférences au département de mathématiques de l'université Tulane, des discussions ont été menées autour du comportement théorique de l'estimateur multivarié corrigé donné par l'équation (2.3). Gustavo DIDIER a ainsi pu démontrer les résultats théoriques énoncés dans cette section.

2.3.1 Cadre asymptotique de l'étude de l'estimation

À l'instar du chapitre 1, les estimateurs sont étudiés ici pour un mBof à M composantes vérifiant les hypothèses (OFBM1-3) décrites dans la section 1.3.2 et une taille d'échantillon N qui tend vers l'infini.

Les propriétés des estimateurs sont étudiées à $M \in \mathbb{N}$ fixé en fonction de la taille d'échantillon N. Les régressions linéaires pour obtenir les estimateurs multivariés $\underline{\hat{H}}^{(\mathrm{M})}$ et $\underline{\hat{H}}^{(\mathrm{M,bc})}$ définis respectivement par les équations (1.50) et (2.3) sont effectuées sur des échelles d'analyse dépendant de la taille d'échantillon N. Plus précisément, les échelles d'analyse vont de $2^{j_1(N)} = a(N)2^{j_1^0}$ à $2^{j_2(N)} = a(N)2^{j_2^0}$, où $a(N)$ est une suite dyadique définie dans la section 1.4.3.3 et $\{j_1^0, \ldots, j_2^0\}$ est la gamme d'échelles d'analyse à une taille d'échantillon N_0 telle que $a(N_0) = 1$. La taille d'échantillon effective à l'échelle $a(N)2^j$ est notée $n_{a,j} = N/(a(N)2^j)$. Ce cadre d'étude est le même que celui défini dans la section 1.4.3.3.

Avec ces définitions, $a(N)2^{j_2^0}/a(N)2^j = 2^{j_2^0 - j}$ spectres d'ondelettes $S^{(w)}(a(N)2^j)$ sont calculés à chaque échelle $a(N)2^j \in \{a(N)2^{j_1^0}, \ldots, a(N)2^{j_2^0}\}$ à partir de fenêtres de coefficients d'on-

delettes de taille $n_{j_2(N)} = n_{a,j_2^0}$ pour obtenir les estimées $\log_2 \bar{\lambda}_m(a(N)2^j)$. Et les estimateurs multivarié (1.50) et multivarié corrigé (2.3) s'écrivent respectivement, pour tout $m \in \{1, \dots, M\}$,

$$\hat{H}_m^{(\mathrm{M})} = \frac{1}{2}\left(\sum_{j=j_1^0}^{j_2^0} w_j \log_2 \hat{\lambda}_m(a(N)2^j) - 1 \right), \tag{2.4}$$

$$\hat{H}_m^{(\mathrm{M,bc})} = \frac{1}{2}\left(\sum_{j=j_1^0}^{j_2^0} w_j \log_2 \bar{\lambda}_m(a(N)2^j) - 1 \right). \tag{2.5}$$

2.3.2 Consistance de l'estimateur multivarié corrigé

Un premier résultat stipule que les valeurs propres $\hat{\lambda}_m^{(w)}(a(N)2^j)$ des spectres d'ondelettes empiriques $S^{(w)}(a(N)2^j)$ se comportent asymptotiquement comme des lois de puissance par rapport à l'échelle $a(N)2^j$ d'exposant $2H_m + 1$

Théorème 2.1. *Soit Y un mBof vérifiant les conditions (OFBM1-3). Alors les valeurs propres $\hat{\lambda}_m^{(w)}(a(N)2^j)$ des spectres d'ondelettes empiriques $S^{(w)}(a(N)2^j)$ de Y vérifient, pour tous $m \in \{1, \dots, M\}$, $w \in \{1, \dots, 2^{j_2^0 - j}\}$ et $j \in \{j_1^0, \dots, j_2^0\}$, lorsque N tend vers $+\infty$,*

$$\frac{\hat{\lambda}_m^{(w)}(a(N)2^j)}{(a(N)2^j)^{2H_m+1}} \xrightarrow{\mathbb{P}} \xi_m(2^0), \tag{2.6}$$

où $\xi_m(2^0) > 0$ est défini dans le théorème 1.3.

Démonstration. La démonstration est rapportée à l'annexe A.1. □

Ce résultat est analogue aux théorèmes 1.3 et 1.4 sur lesquelles repose la consistance de l'estimateur $\underline{\hat{H}}^{(\mathrm{M})}$. Les comportements en loi de puissance asymptotiques des $\hat{\lambda}_m^{(w)}(a(N)2^j)$ signifient que, pour toutes les fenêtres w, chaque estimée $\log_2 \hat{\lambda}_m^{(w)}(a(N)2^j)$ est asymptotiquement affine de pente $2H_m + 1$, justifiant d'effectuer des régressions linéaires sur la moyenne de leurs logarithmes notée $\log_2 \bar{\lambda}_m(a(N)2^j)$ et donnée par l'équation (2.2). Il en découle la consistance de l'estimateur $\underline{\hat{H}}^{(\mathrm{M,bc})}$ dans une grande généralité, énoncée par le théorème suivant.

Théorème 2.2 (Consistance)**.** *Supposons que les conditions (OFBM1−3) sont vérifiées. Alors, pour tout $m \in \{1, \dots, M\}$, lorsque N tend vers $+\infty$,*

$$\hat{H}_m^{(\mathrm{M,bc})} \xrightarrow{\mathbb{P}} H_m. \tag{2.7}$$

Démonstration. Voir l'annexe A.2. □

2.3.3 Normalité asymptotique de l'estimateur multivarié corrigé

De façon analogue au théorème 1.6, il peut être montré que les logarithmes des valeurs propres estimées $(\log_2 \bar{\lambda}_m(a(N)2^j))_{1 \le m \le M,\, j_1^0 \le j \le j_2^0}$ forment asymptotiqement un vecteur gaussien sous la condition (C0), introduite dans la section 1.4.3.2. Le théorème en question et sa démonstration sont donnés dans l'annexe A.3. En résulte la normalité multivariée asymptotique de l'estimateur $\underline{\hat{H}}^{(\mathrm{M,bc})}$, sous des conditions plus restrictives que celles apparaissant dans le

théorème 2.2 sur sa consistance puisque la condition (C0) est nécessaire, comme c'était déjà le cas pour l'estimateur multivarié $\hat{H}_m^{(\mathrm{M})}$. Ce comportement est établi dans le théorème suivant.

Théorème 2.3 (Normalité asymptotique)**.** *Supposons les conditions (OFBM1−3) et (C0) vérifiées. Alors, lorsque N tend vers $+\infty$,*

$$\left\{ \sqrt{\frac{N}{a(N)}} \left(\hat{H}_m^{(\mathrm{M,bc})} - H_m \right) \right\}_{m \in \{1,\dots,M\}} \xrightarrow{d} \mathcal{N}(0, \Sigma_B), \tag{2.8}$$

avec $\Sigma_B \in \mathcal{S}_{\geq 0}(M, \mathbb{R})$.

Démonstration. Voir l'annexe A.3. □

Sous la condition (C0), les valeurs propres $\lambda_m(a(N)2^j)$ sont simples pour N grand, comme expliqué dans la section 1.4.3.2. Ainsi, l'estimateur multivarié corrigé $\underline{\hat{H}}^{(\mathrm{M,bc})} = (\hat{H}_1^{(\mathrm{M,bc})}, \dots, \hat{H}_M^{(\mathrm{M,bc})})$ est un vecteur gaussien si les valeurs propres $\lambda_1(a(N)2^j), \dots, \lambda_M(a(N)2^j)$ sont distinctes sur les échelles d'analyse $a(N)2^{j_1^0} \leq a(N)2^j \leq a(N)2^{j_2^0}$ pour N grand.

2.3.4 Covariance des estimateurs multivariés

Dans cette section, sont proposées des appoximations de la variance et de la corrélation des estimateurs multivariés (1.50) et (2.3) sous les conditions du théorème 2.3, et sous des hypothèses supplémentaires sur l'indépendance inter-échelle des valeurs propres estimées $\hat{\lambda}_m(2^j)$ et $\bar{\lambda}_m(2^j)$ et la décorrélation entre les coefficients d'ondelettes à toutes les échelles d'analyse $a(N)2^{j_1^0} \leq a(N)2^j \leq a(N)2^{j_2^0}$. Les différents calculs permettant d'aboutir à ces approximations sont rapportés dans l'annexe A.4.

Une première approximation est la décorrélation des estimées multivariées à de grandes tailles d'échantillon N, qui s'écrit, pour tous $m, m' \in \{1, \dots, M\}$ avec $m \neq m'$,

$$\mathrm{Cov}\left(\hat{H}_m^{(\mathrm{M})}, \hat{H}_{m'}^{(\mathrm{M})} \right) \approx 0, \tag{2.9}$$

$$\mathrm{Cov}\left(\hat{H}_m^{(\mathrm{M,bc})}, \hat{H}_{m'}^{(\mathrm{M,bc})} \right) \approx 0. \tag{2.10}$$

Il est également possible d'approximer la variance des estimées multivariées à de grandes tailles d'échantillon N, comme suit :

$$\mathrm{Var}\left(\hat{H}_m^{(\mathrm{M})} \right) \approx \frac{1}{2} (\log_2 e)^2 \sum_{j=j_1^0}^{j_2^0} \frac{w_j^2}{n_{a,j}}, \tag{2.11}$$

$$\mathrm{Var}\left(\hat{H}_m^{(\mathrm{M,bc})} \right) \approx \frac{1}{2} (\log_2 e)^2 \sum_{j=j_1^0}^{j_2^0} \frac{w_j^2}{n_{a,j}}, \tag{2.12}$$

pour tout $m \in \{1, \dots, M\}$.

Autrement dit, les matrices Σ_B des théorèmes 2.3 et 1.7 sont approximativement diagonales et identiques, et de coefficients diagonaux pouvant être approximés par des fonctions qui décroissent en $a(N)/N$.

Pour une taille d'échantillon N donnée, ces approximations nécessitent

(i) la validité de la condition ($C0$), c'est-à-dire l'égalité entre des valeurs propres $\lambda_m(a(N)2^j)$ sur les échelles d'analyse $a(N)2^{j_1^0} \leq a(N)2^j \leq a(N)2^{j_2^0}$;

(ii) la décorrélation des coefficients d'ondelettes $\{D(a(N)2^j,k)\}_{k\in\{1,\dots,n_j\}}$ aux différentes échelles d'analyse $a(N)2^{j_1^0} \leq a(N)2^j \leq a(N)2^{j_2^0}$;

(iii) et l'indépendance des valeurs propres estimées $\hat{\lambda}_m(a(N)2^j)$ ou $\bar{\lambda}_m(a(N)2^j)$ entre les échelles d'analyse $a(N)2^{j_1^0} \leq a(N)2^j \leq a(N)2^{j_2^0}$.

En pratique, les coefficients d'ondelettes ne sont pas décorrélés. Cependant, par la proposition 3.2 de Abry et Didier (2018b), si la différence $|a(N)2^j k - a(N)2^{j'}k'|$, avec $j, j', k, k' \in \mathbb{N}$, est suffisamment grande, alors

$$\left\|\mathbb{E}[D(a(N)2^j,k)D(a(N)2^{j'},k')^T]\right\| \leq C(N_\psi,j,j')\frac{\left|\ln^2|a(N)2^jk - a(N)2^{j'}k'|\right|^2}{|a(N)2^jk - a(N)2^{j'}k'|^{2(N_\psi - \max\limits_{m\in\{1,\dots,M\}} H_m)}}, \tag{2.13}$$

où $C(N_\psi,j,j') = C(N_\psi)2^{(j+j')(N_\psi+\frac{1}{2})}$ pour une certaine constante $C(N_\psi) > 0$ qui dépend du nombre de moments nuls N_ψ de l'ondelette mère ψ_0 utilisée. Cela signifie que les coefficients d'ondelettes sont faiblement corrélés pour de larges valeurs de $|2^jk - 2^{j'}k'|$.

Ainsi, si les valeurs propres $\lambda_m(a(N)2^j)$ sont distinctes pour les échelles d'analyse $a(N)2^{j_1^0} \leq a(N)2^j \leq a(N)2^{j_2^0}$, les estimateurs multivariés $\hat{\underline{H}}^{(\mathrm{M})}$ et $\hat{\underline{H}}^{(\mathrm{M,bc})}$ sont approximativement asymptotiquement décorrélés, et de variances approximativement asymptotiquement identiques, décroissantes avec la taille d'échantillon N et ne dépendant ni de $\underline{H}$, ni de W, ni de Σ.

2.4 Mouvement brownien fractionnaire multivarié (M-mBf)

Jusqu'ici, l'estimation des exposants d'autosimilarité a été étudiée pour un mBof assez général. Ce modèle fait cependant appel à des paramètres ne pouvant être estimés à partir de données du monde réel et sa synthèse numérique est difficile à mettre en œuvre. Une contribution du présent travail est d'introduire un processus stochastique autosimilaire, construit à partir de mouvements browniens fractionnaires, adapté à des applications pratiques. Ce modèle est un cas particulier de mBof aux propriétés adaptées à des données du monde réel, telles que la réversibilité en temps et l'identifiabilité, et dont l'analyse en ondelettes se lit plus simplement. Celui-ci permet également une étude numérique approfondie des performances des différents estimateurs étudiés dans ce manuscrit, étude réalisée dans la section suivante.

2.4.1 Définition

Le cas d'un mouvement brownien opérateur-fractionnaire (mBof), défini dans la section 1.3.1, avec une matrice de Hurst $\underline{H}$ diagonale a été examiné dans la section 1.3.3 : chaque composante d'un tel mBof est autosimilaire. Ce cas suggère la définition suivante du mouvement brownien fractionnaire multivarié (M-mBf). Soit une collection de M mBf $\{\mathcal{B}_m(t)\}_{t\in\mathbb{R}}$ définis selon la section 1.2), caractérisés par des exposants d'autosimilarité éventuellement différents H_m, avec $0 < H_m < 1$, chacun de variance $C_{H_m} := \sigma_m^2$ et corrélés par une matrice $\underline{\underline{\rho}}$ de taille $M \times M$. Le M-mBf est défini comme un mélange linéaire de ces mBf à l'aide d'une matrice réelle inversible W de taille $M \times M$, comme suit :

$$\{\mathcal{B}_{\Sigma,W,\underline{H}}(t)\}_{t\in\mathbb{R}} := W\left\{(\mathcal{B}_1(t),\dots,\mathcal{B}_M(t))^T\right\}_{t\in\mathbb{R}}, \tag{2.14}$$

où $\underline{H} = (H_1, \ldots, H_M)$ et Σ est la matrice de covariance de taille $M \times M$ du processus M-varié $(\mathcal{B}_1, \ldots, \mathcal{B}_M)$, c'est-à-dire que les entrées de Σ sont telles que

$$\Sigma_{m,m'} := \mathbb{E}[\mathcal{B}_m(1)\mathcal{B}_{m'}(1)] = \rho_{m,m'}\sigma_m\sigma_{m'}, \tag{2.15}$$

pour tous $m, m' \in \{1, \ldots, M\}$. Le M-mBf est illustré par la figure 2.4. Contrairement au cas d'une matrice de Hurst $\underline{H}$ diagonale, chaque composante $\mathcal{B}_m$ du M-mBf n'est pas caractérisé par un unique exposant d'autosimilarité H_m mais par l'ensemble du vecteur des exposants d'autosimilarité $\underline{H}$.

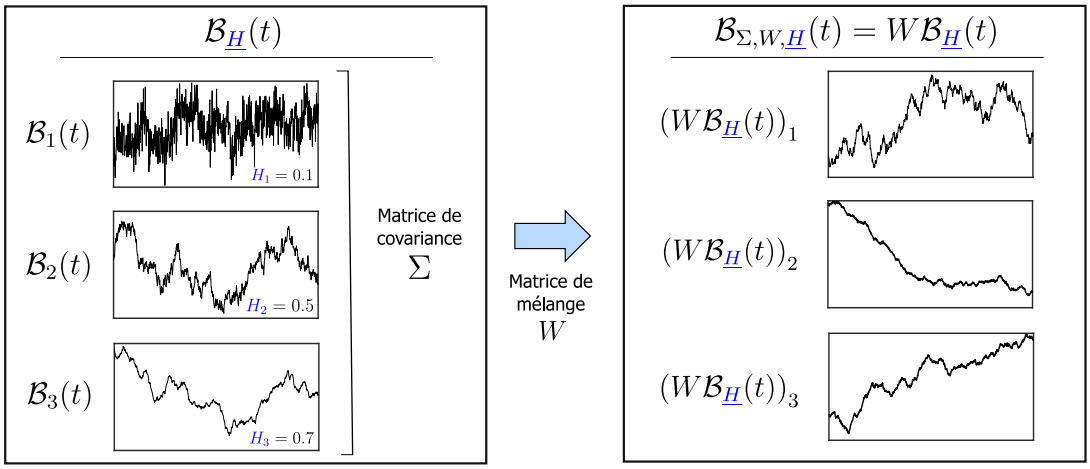


FIGURE 2.4 – **Illustration du mouvement brownien fractionnaire multivarié.**

2.4.2 Relation avec le mBof

Le théorème suivant montre que le M-mBf forme en fait un cas spécifique du mBof.

Théorème 2.4. *Le M-mBf $\mathcal{B}_{\Sigma,W,\underline{H}}$ est un mBof $\mathcal{B}_{\underline{\underline{H}},A}$ tel que*

$$AA^* := W\,(G \odot \Sigma)\,W^T, \tag{2.16}$$

$$\underline{\underline{H}} := W\,\mathrm{diag}(\underline{H})\,W^{-1}, \tag{2.17}$$

où $\mathrm{diag}(\underline{H})$ *est la matrice diagonale dont les entrées sont les exposants d'autosimilarité* $H_1, \ldots, H_M$,

$$\mathrm{diag}(\underline{H}) = \begin{pmatrix} H_1 & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & H_M \end{pmatrix}, \tag{2.18}$$

et G consiste en une matrice réelle de taille $M \times M$ avec pour entrées

$$G_{m,m'} := \frac{1}{2\pi}\Gamma(H_m + H_{m'} + 1)\sin\left((H_m + H_{m'})\frac{\pi}{2}\right), \tag{2.19}$$

où Γ désigne la fonction gamma et $\odot$ le produit matriciel de Hadamard (ou produit point par point).

Démonstration. Puisque les processus $\mathcal{B}_{\underline{\underline{H}},A}$ et $\mathcal{B}_{\Sigma,W,\underline{H}}$ sont gaussiens M-variés centrés, il suffit de

montrer que leurs fonctions de covariance sont égales. Les calculs sont donnés dans l'annexe A.5. □

Il en résulte que le M-mBf satisfait la relation d'autosimilarité multivariée (1.21). Le M-mBf fournit ainsi un modèle d'autosimilarité multivariée, adapté à une utilisation pratique en traitement du signal, défini à partir d'un vecteur d'exposants d'autosimilarité $\underline{H} = (H_1, \ldots, H_M)$, d'une matrice de covariance intrinsèque Σ et d'une matrice de mélange W. De plus, la définition du M-mBf permet de lire simplement sa matrice de covariance,

$$\mathbb{E}\left[\mathcal{B}_{\Sigma,W,\underline{H}}(1)\mathcal{B}_{\Sigma,W,\underline{H}}(1)^T\right] = W\,\Sigma\,W^T. \tag{2.20}$$

2.4.3 Propriétés

Avec les définitions du théorème 2.4, le M-mBf $\mathcal{B}_{\Sigma,W,\underline{H}}$ est identifiable si et seulement si $G \odot \Sigma$ est une matrice semi-définie positive. En effet, la proposition 1.1 assure que $\mathcal{B}_{\underline{\underline{H}},A} = \mathcal{B}_{\Sigma,W,\underline{H}}$ est identifiable si et seulement si AA^* est semi-définie positive, c'est-à-dire si et seulement si $G \odot \Sigma$ est semi-définie positive. Ceci implique que le vecteur des exposants d'autosimilarité $\underline{H}$ et la matrice de covariance Σ ne peuvent être choisis indépendants. Ceci avait été montré par Amblard et Coeurjolly (2011) dans le cas particulier $W = \mathbb{I}$, et s'avère donc encore valable pour un mélange W quelconque.

De plus, la proposition 1.2 assure la réversibilité en temps de $\mathcal{B}_{\underline{\underline{H}},A}$, c'est-à-dire la covariance de ses statistiques sous un changement $t \to -t$, lorsque $\Im(AA^*) = 0$. Puisque la matrice AA^* donnée par l'équation (2.16) est réelle, le M-mBf $\mathcal{B}_{\Sigma,W,\underline{H}}$ est réversible en temps.

Le modèle d'autosimilarité multivariée proposé est donc une version spécifique du mBof $\mathcal{B}_{\underline{\underline{H}},A}$ mieux adaptée aux objectifs du traitement du signal car ses paramètres sont des quantités qui peuvent être estimées à partir de données du monde réel.

De plus, on peut énoncer le résultat suivant.

Théorème 2.5. *Le M-mBf $\mathcal{B}_{\Sigma,W,\underline{H}}$ est l'unique processus gaussien centré de fonction de covariance, pour tous $t, s \in \mathbb{R}$,*

$$\mathbb{E}\left[\mathcal{B}_{\Sigma,W,\underline{H}}(t)\mathcal{B}_{\Sigma,W,\underline{H}}(s)^T\right] = W\,\mathbb{E}\left[\mathcal{B}(t)\mathcal{B}(s)^T\right]\,W^T \tag{2.21}$$

où, pour tous $m, m' \in \{1, \ldots, M\}$,

$$\mathbb{E}\left[\mathcal{B}_m(t)\mathcal{B}_{m'}(s)\right] = \frac{\sigma_m\sigma_{m'}}{2}\rho_{m,m'}\left(|t|^{H_m+H_{m'}} + |s|^{H_m+H_{m'}} - |t-s|^{H_m+H_{m'}}\right). \tag{2.22}$$

Démonstration. Le résultat découle des équations (A.36) et (A.39) données à l'annexe A.5. □

L'équation (2.21) peut s'écrire sous forme matricielle comme suit, pour tous $t, s \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}\left[\mathcal{B}_{\Sigma,W,\underline{H}}(t)\mathcal{B}_{\Sigma,W,\underline{H}}(s)^T\right] = \frac{1}{2}W\Big(&|t|^{\mathrm{diag}(\underline{H})}\,\Sigma\,|t|^{\mathrm{diag}(\underline{H})} + |s|^{\mathrm{diag}(\underline{H})}\,\Sigma\,|s|^{\mathrm{diag}(\underline{H})} \\ &-|t-s|^{\mathrm{diag}(\underline{H})}\,\Sigma\,|t-s|^{\mathrm{diag}(\underline{H})}\Big)\,W^T.\end{aligned} \tag{2.23}$$

où la matrice $|t|^{\mathrm{diag}(\underline{H})}$ est définie par

$$|t|^{\mathrm{diag}(\underline{H})} := \begin{pmatrix} |t|^{H_1} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & |t|^{H_M} \end{pmatrix}. \tag{2.24}$$

Cette écriture est pratique pour certains calculs, notamment pour l'analyse en ondelettes réalisée dans la section 2.4.5.

2.4.4 Synthèse numérique

Les accroissements d'un M-mBf,

$$\forall \delta > 0, \quad \{\mathcal{G}_{\Sigma,W,\underline{H},\delta}(t)\}_{t\in\mathbb{R}} = \left\{ \frac{1}{\delta} \left(\mathcal{B}_{\Sigma,W,\underline{H}}(t+\delta) - \mathcal{B}_{\Sigma,W,\underline{H}}(t) \right) \right\}_{t\in\mathbb{R}}, \tag{2.25}$$

forment un *bruit gaussien fractionnaire multivarié*, c'est-à-dire un processus gaussien multivarié dont la matrice de covariance découle de l'équation (2.21). Or, de par sa stationnarité, un bruit gaussien fractionnaire multivarié de fonction de covariance connue peut être synthétisé numériquement par plongement circulant (Helgason et collab., 2011). Un M-mBf de taille finie $\{\mathcal{B}_{\Sigma,W,\underline{H}}(n)\}_{n\in\{1,\dots,N\}}$ peut donc être synthétisé numériquement à partir de la synthèse de ses incréments $\{\mathcal{G}_{\Sigma,W,\underline{H},1}(n)\}_{n\in\{1,\dots,N\}}$. Une telle synthèse numérique permet l'étude numérique des performances d'estimateurs du vecteur des exposants d'autosimilarité $\underline{H}$ pour des M-mBf de taille finie.

2.4.5 Analyse en ondelettes

Le M-mBf est un mBof qui vérifie les hypothèses (OFBM1-2), et qui vérifie également l'hypothèse (OFBM3) s'il est identifiable. À ce titre, tous les théorèmes sur les estimateurs par ondelettes présentés dans le chapitre 1 sont valables pour un M-mBf identifiable.

Par ailleurs, les propriétés du spectre d'ondelettes multivarié d'un mBof décrites dans la section 1.4.1.3 s'écrivent plus simplement pour un M-mBf. Cela permet d'exhiber simplement certains cas où la condition (C0) (cf. Hypothèse 7) est nécessaire dans différents théorèmes de convergence des sections 1.4 et 2.3, n'est pas valide. L'analyse en ondelettes pour un M-mBf se simplifie comme suit.

Par souci de simplicité, notons $X = \mathcal{B}_{\Sigma,I_n,\underline{H}}$ et $Y = WX = \mathcal{B}_{\Sigma,W,\underline{H}}$. Par définition de la transformée en ondelettes multivariée donnée dans la section 1.4.1.1, les coefficients d'ondelettes $D_Y(2^j,k)$ associés à Y s'écrivent comme des mélanges linéaires des coefficients d'ondelettes $D_X(2^j,k)$ associés à X à partir de la matrice W, comme suit :

$$\forall (j,k) \in \mathbb{N}\times\mathbb{Z}, \quad D_Y(2^j,k) = \int_{\mathbb{R}} WX(t)\psi_{j,k}(t)\mathrm{d}t = WD_X(2^j,k). \tag{2.26}$$

La relation d'autosimilarité (1.37) vérifiée par les transformées en ondelettes $\{D_Y(2^j,k)\}_{k\in\mathbb{Z}}$ à chaque échelle 2^j s'écrit alors

$$\{D_Y(2^j,k)\}_{k\in\mathbb{Z}} \overset{fdd}{=} \left\{ 2^{j(\underline{H}+\frac{1}{2})} WD_X(2^0,k) \right\}_{k\in\mathbb{Z}}. \tag{2.27}$$

Ainsi, pour une taille d'échantillon N, la relation entre les spectres d'ondelettes multivariés empiriques $S_X(2^j)$ et $S_Y(2^j)$ de X et Y, respectivement, à chaque échelle 2^j est la suivante :

$$\forall j \in \{0, \dots, \log_2 N\}, \quad S_Y(2^j) \overset{d}{=} 2^{j(\underline{H}+\frac{1}{2})}\, W\, S_X(2^0)\, W^T\, 2^{j(\underline{H}+\frac{1}{2})}, \tag{2.28}$$

où la matrice diagonale $2^{j(\underline{H}+\frac{1}{2})}$ est donnée par

$$2^{j(\underline{H}+\frac{1}{2})} := \begin{pmatrix} 2^{j(H_1+\frac{1}{2})} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 2^{j(H_M+\frac{1}{2})} \end{pmatrix}. \tag{2.29}$$

Les constantes de proportionnalité $\xi_1(2^0), \dots, \xi_M(2^0)$ dans les lois de puissance asymptotiques des valeurs propres $\lambda_1(2^j), \dots, \lambda_M(2^j)$ de $\mathbb{E}[S_Y(2^j)]$ données par le théorème 1.3,

$$\lambda_m(2^j) \underset{j\to+\infty}{\sim} \xi_m(2^0)\, 2^{j(2H_m+1)}, \tag{2.30}$$

dépendent alors uniquement de $W\mathbb{E}[S_X(2^0)]W^T$ (voir la démonstration de la proposition 5.1 de Abry et collab. (2022)).

Or, par définition du spectre d'ondelettes multivarié, on a :

$$\forall j \in \mathbb{N}, \quad W\, \mathbb{E}[S_X(2^0)]\, W^T = W\, \mathbb{E}\left[D_X(2^0,1) D_X(2^0,1)^T\right] W^T \tag{2.31}$$

$$= W \left(\int_{\mathbb{R}} \int_{\mathbb{R}} \mathbb{E}\left[X(t)X(s)^T\right] \psi_0(t)\psi_0(s)\, \mathrm{d}t\, \mathrm{d}s \right) W^T \tag{2.32}$$

$$= -\int_{\mathbb{R}} \int_{\mathbb{R}} |t-s|^{\mathrm{diag}(\underline{H})}\, W^T\, \Sigma\, W\, |t-s|^{\mathrm{diag}(\underline{H})} \psi_0(t)\psi_0(s)\, \mathrm{d}t\, \mathrm{d}s. \tag{2.33}$$

Ainsi, les matrices $W\mathbb{E}[S_X(2^0)]W^T$ et $W\Sigma W^T$ ont les mêmes valeurs propres. En particulier, lorsque toutes les valeurs propres de $W\Sigma W^T$ sont égales, on a $\xi_1(2^0) = \dots = \xi_M(2^0)$ et la condition (C0) n'est donc plus vérifiée si $H_1 = \dots = H_M$.

2.5 Performances empiriques des estimateurs

Les performances des différents estimateurs du vecteur des exposants d'autosimilarité $\underline{H}$ sont étudiées numériquement sur des M-mBf synthétiques de taille finie N à partir de simulations de Monte Carlo pour différents ensembles de paramètres $(W, \Sigma, \underline{H})$.

2.5.1 Simulations de Monte Carlo

Pour l'évaluation des performances, $N_{\mathrm{MC}} = 1000$ réalisations de $M = 6$-mBf synthétiques de taille $N \in \{2^{13}, 2^{14}, 2^{15}, 2^{16}, 2^{17}, 2^{18}\}$ sont considérées avec les différents choix suivants pour les paramètres $\underline{H}$, Σ et W :

- deux vecteurs d'exposants d'autosimilarité $\underline{H}$ sont étudiés :
  - soit les H_m sont tous égaux : $\forall m \in \{1, \dots, M\},\ H_m = 0.7$;
  - soit les H_m sont tous différents : $\underline{H} = (0.4, 0.5, 0.6, 0.65, 0.7, 0.8)$;
- deux matrices de covariance Σ sont étudiées :

- soit les mBf ne sont pas corrélés : $\Sigma = \mathbb{I}$;
- soit les mBf sont corrélés : $\forall 1 \leq m < m' \leq M,\ \Sigma_{m,m} = 1,\ \Sigma_{m,m'} = 0.7^{|m-m'|}$;

• deux matrices de mélange W sont étudiées :

- soit les mBf ne sont pas mélangés : $W = \mathbb{I}$;
- soit les mBf sont mélangés : W est une matrice réelle non orthogonale choisie aléatoirement et identique pour toutes les expériences.

Formulaire récapitulatif des scénarios étudiés

Pour simplifier la lecture, les configurations principalement étudiées ici sont données dans le tableau suivant. Ces configurations sont expliquées en détail dès la section suivante. Pour rappel, la condition (C0) est décrite dans la section 1.4.3.2.

Nom	Mélange	Corrélation	Exposants d'autosimilarité
Multivarié	$W \neq \mathbb{I}$	$\Sigma \neq \mathbb{I}$	H_m tous différents
Effet de répulsion croissant	$W \neq \mathbb{I}$	$\Sigma \neq \mathbb{I}$	H_m tous égaux
Sans mélange	$W = \mathbb{I}$	$\Sigma \neq \mathbb{I}$	H_m tous différents
Condition (C0) invalide	$W = \mathbb{I}$	$\Sigma = \mathbb{I}$	H_m tous égaux
Changement d'ordre	$W = \mathbb{I}$	$\Sigma = \mathbb{I}$	H_m tous différents

La transformée en ondelettes donnée par l'équation (1.32) est réalisée avec l'ondelette mère ψ_0 de Daubechies la moins asymétrique à $N_\psi = 2$ moments nuls (Daubechies, 1992). Les régressions linéaires pour calculer les différents estimateurs sont réalisées avec des poids w_j de type gaussien au travers des échelles $2^{j_1(N)} = a(N)2^{j_1^0} = a(N)2^6$ à $2^{j_2(N)} = a(N)2^{j_2^0} = a(N)2^9$ à la taille d'échantillon N, où la suite dyadique $a(N)$ est définie par $a(N) = 2^{\lfloor \beta \log_2(N/N_0) \rfloor}$ avec $\beta = 0.9$ et $N_0 = 2^{13}$.

Pour rappel, pour une taille d'échantillon N, les différents estimateurs $\underline{\hat{H}}^{(\mathrm{U})}$, $\underline{\hat{H}}^{(\mathrm{M})}$ et $\underline{\hat{H}}^{(\mathrm{M,bc})}$ s'écrivent sous la forme

$$\hat{H}_m := \frac{1}{2}\left(\sum_{j=j_1^0}^{j_2^0} w_j\, y_m(a(N)2^j) - 1 \right), \tag{2.34}$$

pour tout $m \in \{1, \ldots, M\}$, où y_m correspond à la fonction de structure associée à l'estimateur $\hat{H}_m$, c'est-à-dire respectivement $\log_2 S_{m,m}$, $\log_2 \hat{\lambda}_m$ et $\log_2 \bar{\lambda}_m$ pour $\underline{\hat{H}}^{(\mathrm{U})}$, $\underline{\hat{H}}^{(\mathrm{M})}$ et $\underline{\hat{H}}^{(\mathrm{M,bc})}$.

2.5.2 Comportement des fonctions de structure

Cette section vise à justifier le choix des échelles d'analyse et présenter les différentes configurations, c'est-à-dire les différents triplets $(\underline{H}, W, \Sigma)$, considérées pour l'étude des performances. Les fonctions de structures $\log_2 S_{m,m}(2^j)$, $\log_2 \hat{\lambda}_m(2^j)$ et $\log_2 \bar{\lambda}_m(2^j)$ associées respectivement aux estimateurs univarié (1.42), multivarié (1.50) et multivarié corrigé (2.3) sont ainsi étudiées pour la plus petite taille d'échantillon $N_0 = 2^{13}$, pour laquelle $a(N_0) = 1$. La gamme d'octaves d'analyse est donc donnée par $\{j_1(N_0), \ldots, j_2(N_0)\} = \{j_1^0, \ldots, j_2^0\}$. Pour alléger les notations, les plus petite et grande octaves d'analyse sont simplement notées j_1 et j_2 dans cette section.

Échelles d'analyse

La figure 2.5 rapporte les fonctions de structure $\log_2 S_{m,m}(2^j)$, $\log_2 \hat{\lambda}_m(2^j)$ et $\log_2 \bar{\lambda}_m(2^j)$ associées aux différents estimateurs, superposées au travers des composantes $m = 1, \ldots, M$, en fonction des échelles 2^j pour des exposants H_m tous égaux en l'absence de corrélation ($\Sigma = \mathbb{I}$) et de mélange ($W = \mathbb{I}$) et une taille d'échantillon $N = 2^{13}$. D'une part, aux petites échelles 2^j, les fonctions de structure théoriques $\log_2 \mathbb{E}[S_{m,m}(2^j)]$ et $\log_2 \lambda_m(2^j)$ se croisent, c'est-à-dire que l'ordre des composantes $m = 1, \ldots, M$ change au travers des échelles. Or, les valeurs propres $\hat{\lambda}_m(2^j)$ et $\bar{\lambda}_m(2^j)$ sont triées par ordre croissant, ne pouvant être associées à des composantes du M-mBf, ce qui mène à des fonctions de structure $\log_2 \hat{\lambda}_m(2^j)$ et $\log_2 \bar{\lambda}_m(2^j)$ non monotones aux petites échelles 2^j sur la figure 2.5. Les régressions linéaires sur ces fonctions de structure sont donc biaisées pour des échelles d'analyse trop petites et j_1 doit être choisi suffisamment grand. D'autre part, il est nécessaire que $M/n_j < 1$ pour que le spectre d'ondelettes multivarié empirique $S(2^j)$ soit de rang plein. La plus grande octave j_2 doit donc être telle que $M < N_0/2^{j_2}$. Par ailleurs, les $\log_2 \bar{\lambda}_m(2^j)$ ne sont définies que de 2^1 à 2^{j_2} (avec en l'occurrence $j_2 = 9$) dans l'équation (2.2), mais sont définies dans la figure 2.5 et les suivantes par $\log_2 \bar{\lambda}_m(2^j) = \log_2 \hat{\lambda}_m(2^j)$ pour $2^j > 2^{j_2}$. L'effet de répulsion entre les $\log_2 \bar{\lambda}_m(2^j)$ n'est donc reproduit que de 2^1 à 2^{j_2} et les $\log_2 \bar{\lambda}_m(2^j)$ ne sont bien approximées par des fonctions affines que de 2^1 à $2^{j_2} = 2^9$.

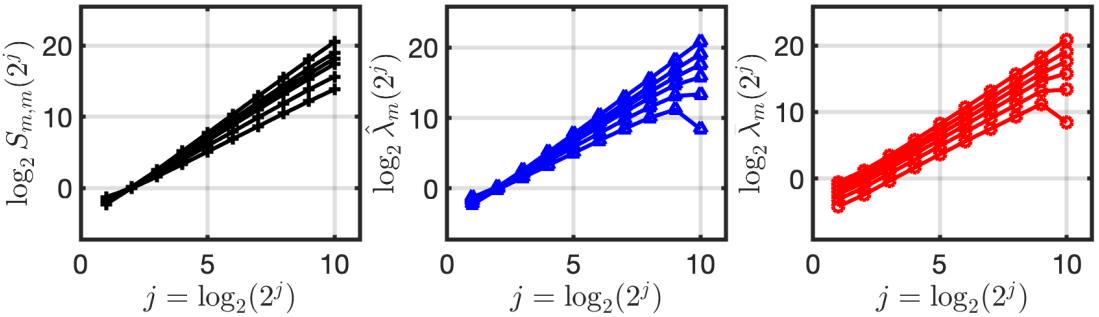
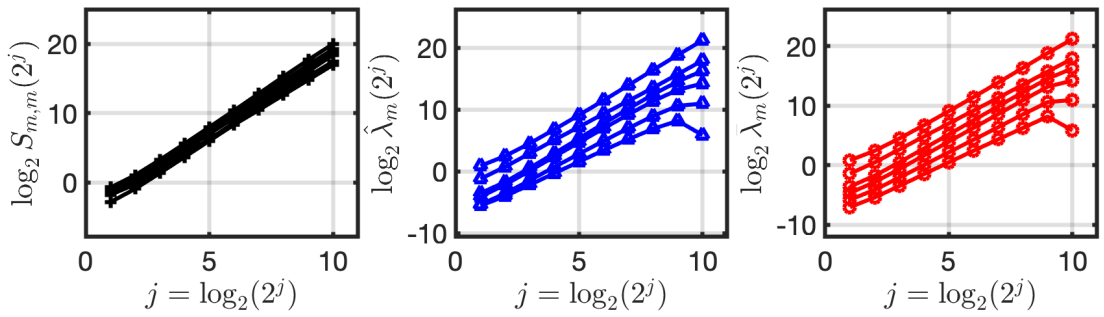


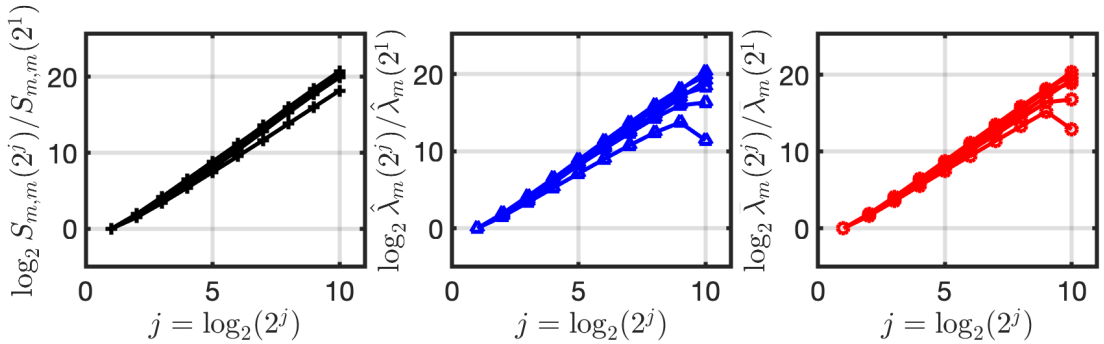
FIGURE 2.5 – **Fonctions de structure avec changement d'ordre.** Logarithmes des (en noir) entrées diagonales $S_{m,m}(2^j)$, (en bleu) valeurs propres $\hat{\lambda}_m(2^j)$ et (en rouge) valeurs propres corrigées $\bar{\lambda}_m(2^j)$ du spectre d'ondelettes empirique $S(2^j)$ moyennés au travers des réalisations de Monte Carlo en fonction des octaves j pour des mBf non corrélés ($\Sigma = \mathbb{I}$) et non mélangés ($W = \mathbb{I}$) avec des exposants H_m tous différents et une taille d'échantillon $N = 2^{13}$.

Scénario multivarié

Le cas multivarié correspond à des mBf possiblement corrélés et mélangés et caractérisés par des exposants H_m non nécessairement égaux. Ainsi, la figure 2.6 rapporte les fonctions de structure $\log_2 S_{m,m}(2^j)$, $\log_2 \hat{\lambda}_m(2^j)$ et $\log_2 \bar{\lambda}_m(2^j)$ associées aux différents estimateurs, superposées au travers des composantes $m = 1, \ldots, M$, en fonction des échelles 2^j pour des mBf corrélés ($\Sigma \neq \mathbb{I}$) et mélangés ($W \neq \mathbb{I}$) avec des exposants H_m tous différents et une taille d'échantillon $N = 2^{13}$. Dans ce cas, l'estimateur univarié n'est pas en son avantage car adapté à l'absence de mélange ($W = \mathbb{I}$). En effet, les fonctions de structure $\log_2 S_{m,m}(2^j)$ ont pour la plupart des pentes similaires car, lorsque $W \neq \mathbb{I}$, les entrées diagonales $\mathbb{E}[S_{m,m}(2^j)]$ du spectre d'ondelettes $\mathbb{E}[S(2^j)]$ se comportent comme des mélanges de lois de puissance (cf. Eq. (1.39)), tous bien approximées par une même loi de puissance d'exposant $2H_M + 1$ (cf. Eq. (1.46)). En revanche, les fonctions de structure $\log_2 \hat{\lambda}_m(2^j)$ et $\log_2 \bar{\lambda}_m(2^j)$ associées aux estimateurs multivariés ont des pentes différentes pour les différentes composantes $m = 1, \ldots, M$, ce qui est en accord avec le comportement en loi de puissance asymptotique des valeurs propres $\lambda_m(2^j)$ (cf. Eq. (1.49)).



(a) Fonctions de structure

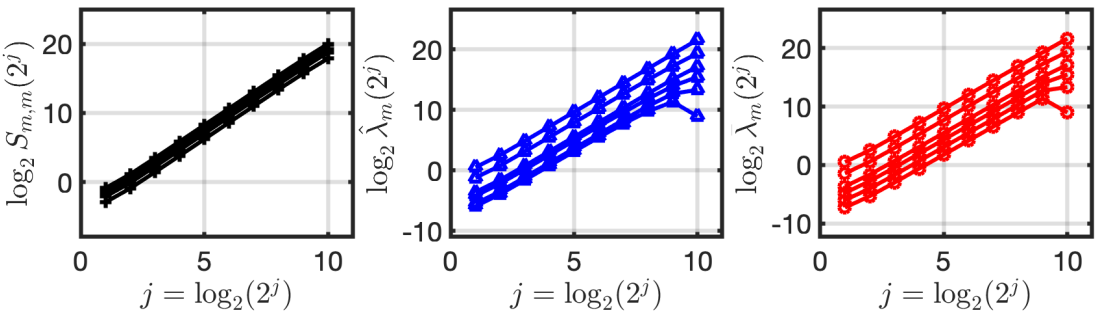


(b) Fonctions de structure ramenées à l'origine

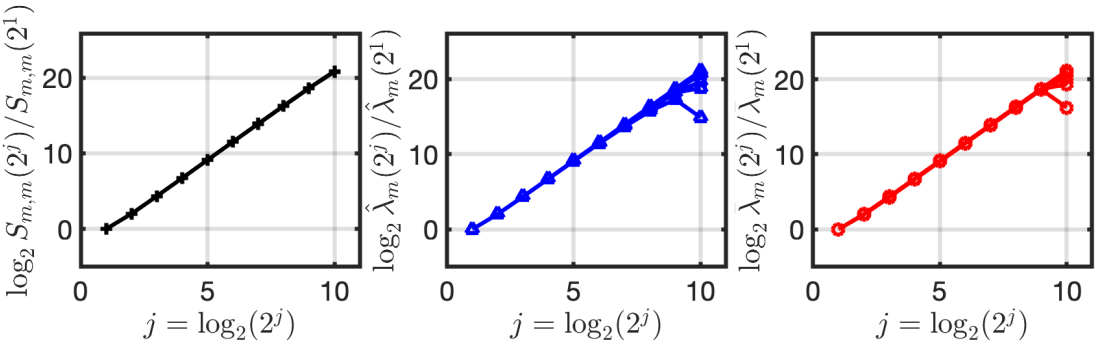
FIGURE 2.6 – **Fonctions de structure dans un cadre multivarié.** (en haut) Logarithmes des (en noir) entrées diagonales $S_{m,m}(2^j)$, (en bleu) valeurs propres $\hat{\lambda}_m(2^j)$ et (en rouge) valeurs propres corrigées $\bar{\lambda}_m(2^j)$ du spectre d'ondelettes empirique $S(2^j)$ moyennés au travers des réalisations de Monte Carlo en fonction des octaves j pour des mBf corrélés ($\Sigma \neq \mathbb{I}$) et mélangés ($W \neq \mathbb{I}$) avec des exposants H_m tous différents et une taille d'échantillon $N = 2^{13}$. (en bas) Les fonctions de structure sont ramenées à l'origine.

Scénario avec effet de répulsion croissant

L'effet de répulsion entre les valeurs propres estimées $\log_2 \hat{\lambda}_m(2^j)$ croît au travers des échelles 2^j, en particulier lorsque les exposants H_m égaux car cela implique des valeurs propres $\lambda_m(2^j)$ proches. On s'intéresse donc au cas où les H_m sont tous égaux dans un cadre multivarié. Ainsi, la figure 2.7 rapporte les fonctions de structure $\log_2 S_{m,m}(2^j)$, $\log_2 \hat{\lambda}_m(2^j)$ et $\log_2 \bar{\lambda}_m(2^j)$ associées aux différents estimateurs, superposées au travers des composantes $m = 1, \ldots, M$, en fonction des échelles 2^j pour des mBf corrélés ($\Sigma \neq \mathbb{I}$) et mélangés ($W \neq \mathbb{I}$) avec des exposants H_m tous égaux et une taille d'échantillon $N = 2^{13}$. Dans ce cas, les pentes des fonctions de structure $\log_2 S_{m,m}(2^j)$ sont similaires, comme attendu car, lorsque les H_m sont tous égaux, les mélanges de lois de puissance des $\mathbb{E}[S_{m,m}(2^j)]$ (cf. Eq. (1.39)) deviennent des lois de puissance d'exposant $2H_m + 1$ (cf. Eq. (1.47)). En revanche, les fonctions de structure $\log_2 \hat{\lambda}_m(2^j)$ s'écartent les unes des autres à mesure que 2^j croît, en raison de l'effet de répulsion. L'estimateur multivarié corrigé réduit cet écart en forçant les fonctions de structure $\log_2 \bar{\lambda}_m(2^j)$ à avoir des pentes similaires grâce à la reproduction d'un effet de répulsion similaire à toutes les échelles : les $\log_2 \bar{\lambda}_m(2^j)$ ont des biais similaires aux différentes octaves $j = 1, \ldots, j_2$ avec $j_2 = 9$, .



(a) Fonctions de structure

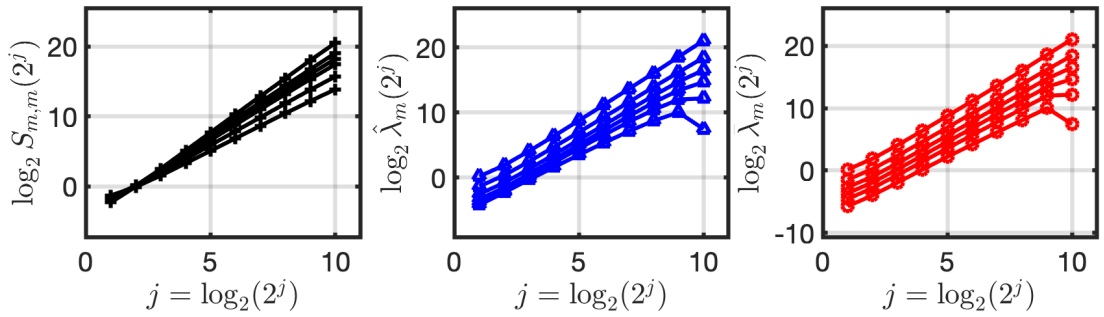


(b) Fonctions de structure ramenées à l'origine

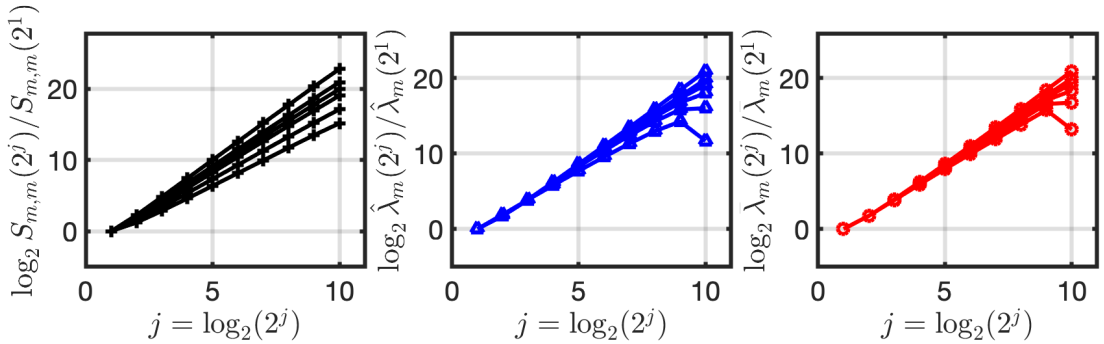
FIGURE 2.7 – **Fonctions de structure avec effet de répulsion croissant.** (en haut) Logarithmes des (en noir) entrées diagonales $S_{m,m}(2^j)$, (en bleu) valeurs propres $\hat{\lambda}_m(2^j)$ et (en rouge) valeurs propres corrigées $\bar{\lambda}_m(2^j)$ du spectre d'ondelettes empirique $S(2^j)$ moyennés au travers des réalisations de Monte Carlo en fonction des octaves j pour des mBf corrélés ($\Sigma \neq \mathbb{I}$) et mélangés ($W \neq \mathbb{I}$) avec des exposants H_m tous égaux et une taille d'échantillon $N = 2^{13}$. (en bas) Les fonctions de structure sont ramenées à l'origine.

Scénario sans mélange

Un cas particulier d'intérêt est ce lui d'absence de mélange ($W = \mathbb{I}$) des mBf. Ainsi, la figure 2.8 rapporte les fonctions de structure $\log_2 S_{m,m}(2^j)$, $\log_2 \hat{\lambda}_m(2^j)$ et $\log_2 \bar{\lambda}_m(2^j)$ associées aux différents estimateurs, superposées au travers des composantes $m = 1, \ldots, M$, en fonction des échelles 2^j pour des mBf corrélés ($\Sigma \neq \mathbb{I}$) mais non mélangés ($W = \mathbb{I}$) avec des exposants H_m tous différents et une taille d'échantillon $N = 2^{13}$. Les pentes des fonctions de structure $\log_2 S_{m,m}(2^j)$ sont très nettement différentes alors que les pentes des $\log_2 \hat{\lambda}_m(2^j)$ et $\log_2 \bar{\lambda}_m(2^j)$ sont beaucoup plus proches les unes des autres. En effet, en l'absence de mélange ($W = \mathbb{I}$), les entrées $S_{m,m}(2^j)$ se comportent comme des lois de puissance d'exposants $2H_m+1$ (cf. Eq. (1.41)), ce qui n'est le cas des valeurs propres $\lambda_m(2^j)$ que de façon asymptotique (cf. Eq. (1.49)). Les pentes des $\log_2 S_{m,m}(2^j)$ sont ainsi proches des $2H_m+1$, même aux petites échelles, contrairement aux pentes des $\log_2 \hat{\lambda}_m(2^j)$ et $\log_2 \bar{\lambda}_m(2^j)$.



(a) Fonctions de structure

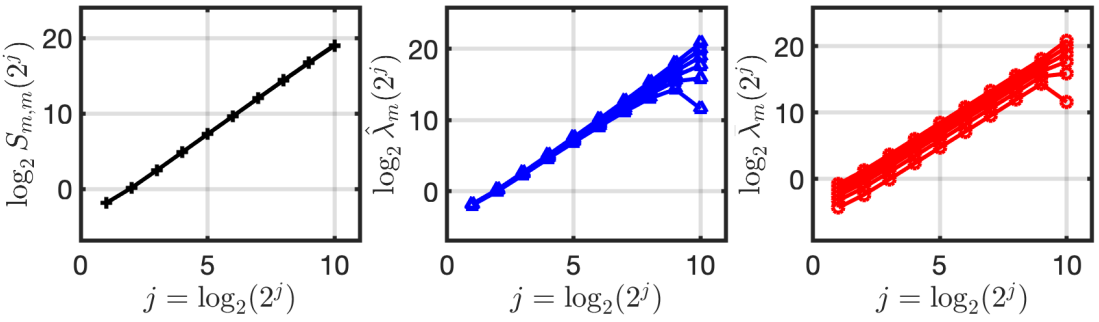


(b) Fonctions de structure ramenées à l'origine

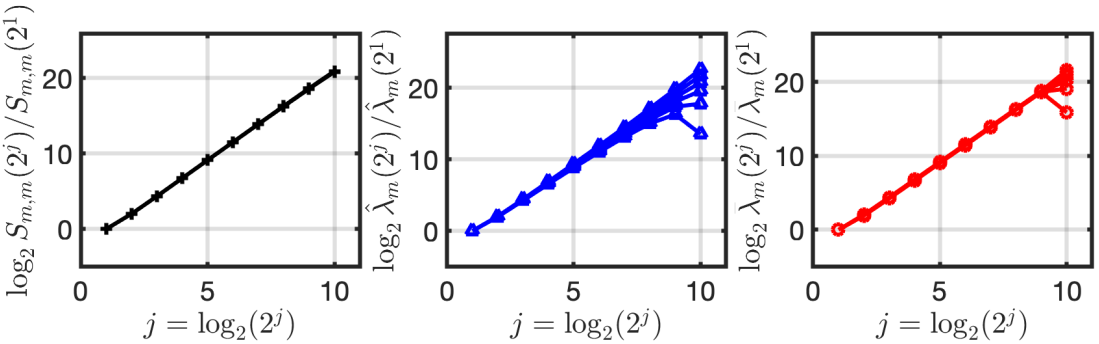
FIGURE 2.8 – **Fonctions de structure dans un cadre sans mélange.** (en haut) Logarithmes des (en noir) entrées diagonales $S_{m,m}(2^j)$, (en bleu) valeurs propres $\hat{\lambda}_m(2^j)$ et (en rouge) valeurs propres corrigées $\bar{\lambda}_m(2^j)$ du spectre d'ondelettes empirique $S(2^j)$ moyennés au travers des réalisations de Monte Carlo en fonction des octaves j pour des mBf corrélés ($\Sigma \neq \mathbb{I}$) et non mélangés ($W = \mathbb{I}$) avec des exposants H_m tous différents et une taille d'échantillon $N = 2^{13}$. (en bas) Les fonctions de structure sont ramenées à l'origine.

Scénario avec condition (C0) invalide

Un dernier cas d'intérêt est celui où les mBf ne sont ni corrélés ($\Sigma = \mathbb{I}$) ni mélangés ($W = \mathbb{I}$) et où tous les exposants H_m sont égaux. Selon la section 2.4.5, ce cas mène à des valeurs propres $\lambda_m(2^j)$ toutes égales à chaque échelle 2^j, ce qui implique que la condition (C0) (cf. Hypothèse 7) n'est pas vérifiée (cf. Section 2.4.5). Les hypothèses des théorèmes sur la normalité asymptotique et les approximations de la covariance des estimateurs multivariés sont alors en défaut. La figure 2.9 rapporte les fonctions de structure $\log_2 S_{m,m}(2^j)$, $\log_2 \hat{\lambda}_m(2^j)$ et $\log_2 \bar{\lambda}_m(2^j)$ associées aux différents estimateurs, superposées au travers des composantes $m = 1, \ldots, M$, en fonction des échelles 2^j pour des mBf ni corrélés ($\Sigma = \mathbb{I}$) ni mélangés ($W = \mathbb{I}$) avec des exposants H_m tous égaux et une taille d'échantillon $N = 2^{13}$. Puisque les exposants H_m sont tous égaux, les fonctions de structure présentent des comportements similaires au scénario avec effet de répulsion croissant présenté ci-dessus : les fonctions de structure $\log_2 S_{m,m}(2^j)$ se superposent parfaitement et l'écart entre les fonctions de structure $\log_2 \hat{\lambda}_m(2^j)$ se creuse pour de grandes échelles 2^j. L'effet de répulsion entre les $\log_2 \hat{\lambda}_m(2^j)$ est ici plus important en raison de l'égalité entre les valeurs propres $\lambda_m(2^j)$. En revanche, l'effet de répulsion est constant pour l'estimateur multivarié corrigé, puisque les fonctions de structure $\log_2 \bar{\lambda}_m(2^j)$ sont pareillement espacées à toutes les échelles 2^j.



(a) Fonctions de structure



(b) Fonctions de structure ramenées à l'origine

FIGURE 2.9 – **Fonctions de structure lorsque la condition (C0) est invalide.** (en haut) Logarithmes des (en noir) entrées diagonales $S_{m,m}(2^j)$, (en bleu) valeurs propres $\hat{\lambda}_m(2^j)$ et (en rouge) valeurs propres corrigées $\bar{\lambda}_m(2^j)$ du spectre d'ondelettes empirique $S(2^j)$ moyennés au travers des réalisations de Monte Carlo en fonction des octaves j pour des mBf non corrélés ($\Sigma = \mathbb{I}$) et non mélangés ($W = \mathbb{I}$) avec des exposants H_m tous égaux et une taille d'échantillon $N = 2^{13}$. (en bas) Les fonctions de structure sont ramenées à l'origine.

2.5.3 Comportement des estimateurs

Les objectifs principaux de cette section sont les suivants :

(*i*) comparer les performances des estimateurs multivarié $\underline{\hat{H}}^{(\mathrm{M})}$ et multivarié corrigé $\underline{\hat{H}}^{(\mathrm{M,bc})}$ dans les différentes scénarios présentés dans la section précédente ;

(*ii*) comparer les performances des estimateurs multivariés $\underline{\hat{H}}^{(\mathrm{M})}$ et $\underline{\hat{H}}^{(\mathrm{M,bc})}$ à l'estimateur univarié $\underline{\hat{H}}^{(\mathrm{U})}$ dans le scénario sans mélange, c'est-à-dire lorsque $W = \mathbb{I}$;

(*iii*) analyser le comportement de la normalité asymptotique, et plus précisément le cadre de validité d'une approximation par une loi normale multivariée ;

(*iv*) étudier le comportement de la variance et de la corrélation des différents estimateurs, et plus précisément les cadres de validité des approximations sur la variance et la corrélation.

Formulaire récapitulatif sur les estimateurs

Les hypothèses des théorèmes sur les différents estimateurs sont rappelées dans le tableau suivant. Pour rappel, sous la condition (C0), décrite dans la section 1.4.3.2, les valeurs propre $\lambda_1(2^j), \ldots, \lambda_M(2^j)$ du spectre d'ondelettes multivarié $\mathbb{E}[S(2^j)]$ sont distinctes aux grandes échelles 2^j.

Estimateur	Univarié	Multivarié	Multivarié corrigé
Notation	$\hat{H}_m^{(\mathrm{U})}$	$\hat{H}_m^{(\mathrm{M})}$	$\hat{H}_m^{(\mathrm{M,bc})}$
Fonction de structure	$\log_2 S_{m,m}(2^j)$	$\log_2 \hat{\lambda}_m(2^j)$	$\log_2 \bar{\lambda}_m(2^j)$
Consistance	Sous $W = \mathbb{I}$	Toujours	
Normalité asymptotique	Sous $W = \mathbb{I}$	Sous (C0)	
Approximation de la covariance	Sous $W = \mathbb{I}$	Sous (C0), décorrélation inter-échelle des $\lambda_m(2^j)$ et décorrélation des coefficients d'ondelettes	

2.5.3.1 Performances : biais, covariance et erreur quadratique

Définition des mesures de performance

Les performances des estimateurs univarié $\underline{\hat{H}}^{(\mathrm{U})}$, multivarié $\underline{\hat{H}}^{(\mathrm{M})}$ et multivarié corrigé $\underline{\hat{H}}^{(\mathrm{M,bc})}$ sont mesurées en termes de biais, de covariance et d'erreur quadratique moyenne. Des matrices de biais, de covariance et d'erreur quadratique moyenne sont ainsi définies respectivement par

$$\mathrm{Bias}(\underline{\hat{H}}) := \left(\mathbb{E}[\underline{\hat{H}}] - \underline{H}\right)\left(\mathbb{E}[\underline{\hat{H}}] - \underline{H}\right)^T, \tag{2.35}$$

$$\mathrm{Cov}(\underline{\hat{H}}) := \mathbb{E}\left[\left(\mathbb{E}[\underline{\hat{H}}] - \underline{\hat{H}}\right)\left(\mathbb{E}[\underline{\hat{H}}] - \underline{\hat{H}}\right)^T\right], \tag{2.36}$$

$$\mathrm{MSE}(\underline{\hat{H}}) := \mathbb{E}\left[\left(\underline{\hat{H}} - \underline{H}\right)\left(\underline{\hat{H}} - \underline{H}\right)^T\right] = \mathrm{Bias}(\underline{\hat{H}}) + \mathrm{Cov}(\underline{\hat{H}}), \tag{2.37}$$

pour les différents estimateurs $\underline{\hat{H}}$ de $\underline{H}$. Pour l'évaluation des performances, ce sont les normes spectrales $\|\mathrm{Bias}(\underline{\hat{H}})\|_2$, $\|\mathrm{Cov}(\underline{\hat{H}})\|_2$ et $\|\mathrm{MSE}(\underline{\hat{H}})\|_2$ de ces différentes matrices qui sont étudiées.

Scénario multivarié

La figure 2.10 rapporte les performances en termes de biais, variance et erreur quadratique moyenne des différents estimateurs en fonction de la taille d'échantillon N pour des réalisations de mBf corrélés ($\Sigma \neq \mathbb{I}$) et mélangés ($W \neq \mathbb{I}$) avec des exposants H_m tous différents. Cette figure montre les deux points suivants :

- (i) les estimateurs multivariés $\underline{\hat{H}}^{(\mathrm{M})}$ et $\underline{\hat{H}}^{(\mathrm{M,bc})}$ ont des performances similaires, puisque l'effet de répulsion est négligeable aux grandes échelles (et donc aux échelles d'analyse considérées) lorsque tous les H_m sont différents ;
- (ii) les performances des estimateurs multivariés $\underline{\hat{H}}^{(\mathrm{M})}$ et $\underline{\hat{H}}^{(\mathrm{M,bc})}$ surpassent, autant en termes de biais que de covariance, celles de l'estimateur univarié $\underline{\hat{H}}^{(\mathrm{U})}$, qui n'est pas adapté à la présence de mélange ($W \neq \mathbb{I}$), particularité de la configuration multivariée.

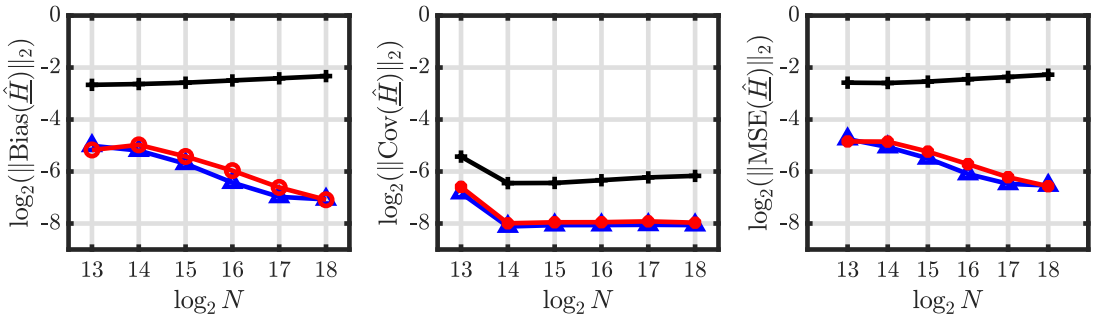


FIGURE 2.10 – **Performances des estimateurs dans un cadre multivarié.** Logarithmes des normes spectrales des biais, variances et erreurs quadratiques moyennes des estimateurs (en noir) univarié $\underline{\hat{H}}^{(\mathrm{U})}$, (en bleu) multivarié $\underline{\hat{H}}^{(\mathrm{M})}$ et (en rouge) multivarié corrigé $\underline{\hat{H}}^{(\mathrm{M,bc})}$ en fonction de la taille d'échantillon N pour des mBf corrélés ($\Sigma \neq \mathbb{I}$) et mélangés ($W \neq \mathbb{I}$) avec des exposants H_m tous différents.

On peut remarquer sur la figure 2.10 que la covariance des différents estimateurs est bien plus élevées à $N = 2^{13}$. Ceci est dû au fait que les échelles d'analyse sont grandes comparées à la taille d'échantillon N, et donc les nombres de coefficients d'ondelettes utilisés pour le calcul des fonctions de structure sont plus petits que pour les autres tailles d'échantillon N, impliquant une plus forte variance des estimateurs.

En complément, la figure 2.11 rapporte les biais relatifs à chaque entrée $m = 1, \ldots, M$ des différents vecteurs d'estimées $\underline{\hat{H}}^{(\mathrm{U})}$, $\underline{\hat{H}}^{(\mathrm{M})}$ et $\underline{\hat{H}}^{(\mathrm{M,bc})}$ pour des mBf corrélés ($\Sigma \neq \mathbb{I}$) et mélangés ($W \neq \mathbb{I}$) avec des exposants H_m tous différents. Les estimées univariées $\underline{\hat{H}}^{(\mathrm{U})}$ ont des biais différents car elles ont toutes tendance à estimer la composante dominante H_M de $\underline{H}$, si bien que les estimées $\hat{H}_1^{(\mathrm{U})}$ et $\hat{H}_2^{(\mathrm{U})}$ sont particulièrement biaisées. En revanche, les estimées multivariées $\underline{\hat{H}}^{(\mathrm{M})}$ et $\underline{\hat{H}}^{(\mathrm{M,bc})}$ ont des biais similaires au travers des composantes $m = 1, \ldots, M$.

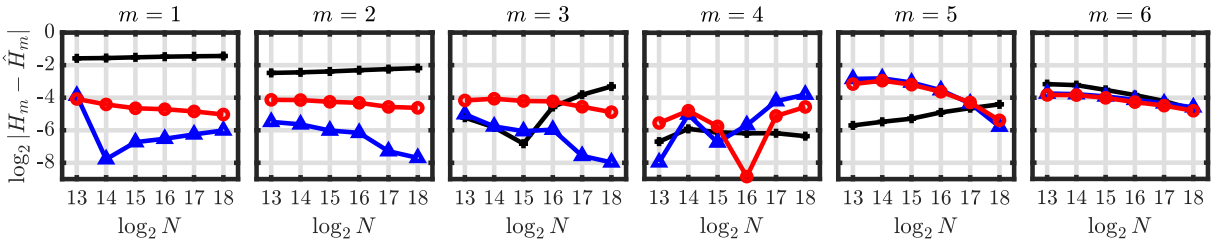


FIGURE 2.11 – **Biais des estimateurs par composante dans un cadre multivarié.** Logarithmes des biais en valeur absolue des estimées (en noir) univariées $\hat{H}_m^{(\mathrm{U})}$, (en bleu) multivariées $\hat{H}_m^{(\mathrm{M})}$ et (en rouge) multivariées corrigées $\hat{H}_m^{(\mathrm{M,bc})}$ pour (de gauche à droite) $m = 1, \ldots, 6$ en fonction de la taille d'échantillon N pour des mBf corrélés ($\Sigma \neq \mathbb{I}$) et mélangés ($W \neq \mathbb{I}$) avec des exposants H_m tous différents.

L'estimateur corrigé $\underline{\hat{H}}^{(\mathrm{M,bc})}$ n'a aucun avantage ni désavantage par rapport à l'estimateur $\underline{\hat{H}}^{(\mathrm{M})}$ dans cette configuration. La différence entre ces deux estimateurs s'opère dans un cas sujet à un effet de répulsion croissant, en particulier lorsque tous les H_m sont égaux.

Scénario avec effet de répulsion croissant

La figure 2.12 rapporte les performances en termes de biais, variance et erreur quadratique moyenne des différents estimateurs en fonction de la taille d'échantillon N pour des réalisations de mBf corrélés ($\Sigma \neq \mathbb{I}$) et mélangés ($W \neq \mathbb{I}$) avec des exposants H_m tous égaux. Ces résultats montrent que

(*i*) les performances en termes de biais de l'estimateur multivarié corrigé $\underline{\hat{H}}^{(\mathrm{M,bc})}$ surpassent celles de l'estimateur multivarié $\underline{\hat{H}}^{(\mathrm{M})}$, comme attendu puisque l'effet de répulsion croît de façon importante au travers des échelles pour l'estimateur non corrigé, et leurs performances en termes de covariance sont similaires ;

(*ii*) l'estimateur univarié $\underline{\hat{H}}^{(\mathrm{U})}$ est moins biaisé que l'estimateur multivarié $\underline{\hat{H}}^{(\mathrm{M})}$, comme attendu car l'estimateur univarié estime l'exposant dominant H_M et tous les exposants sont égaux ($H_1 = \ldots = H_M$), mais sa covariance est plus affectée que celle des estimateurs multivariés ;

(*iii*) les performances de l'estimateur multivarié corrigé $\underline{\hat{H}}^{(\mathrm{M,bc})}$ sont meilleures que celles de l'estimateur univarié $\underline{\hat{H}}^{(\mathrm{U})}$ à la fois en termes de biais et de covariance.

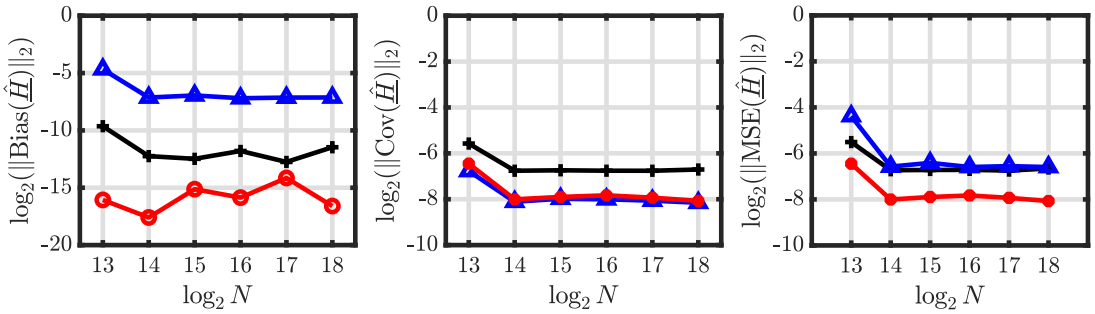


FIGURE 2.12 – **Performances des estimateurs dans un cadre avec effet de répulsion croissant.** Logarithmes des normes spectrales des biais, variances et erreurs quadratiques moyennes des estimateurs (en noir) univarié $\underline{\hat{H}}^{(\mathrm{U})}$, (en bleu) multivarié $\underline{\hat{H}}^{(\mathrm{M})}$ et (en rouge) multivarié corrigé $\underline{\hat{H}}^{(\mathrm{M,bc})}$ en fonction de la taille d'échantillon N pour des mBf corrélés ($\Sigma \neq \mathbb{I}$) et mélangés ($W \neq \mathbb{I}$) avec des exposants H_m tous égaux.

Pour une étude plus précise du comportement des estimées, la figure 2.13 montre les biais des différentes estimées $\hat{H}_m^{(\mathrm{U})}$, $\hat{H}_m^{(\mathrm{M})}$ et $\hat{H}_m^{(\mathrm{M,bc})}$ de H_m pour les composantes $m = 1, \ldots, M$ dans le cas de mBf corrélés ($\Sigma \neq \mathbb{I}$) et mélangés ($W \neq \mathbb{I}$) avec des exposants H_m tous égaux. On peut observer que les biais des estimées univariées $\hat{H}_m^{(\mathrm{U})}$ et multivariées corrigées $\hat{H}_m$, en plus d'être satisfaisants, varient peu entre les composantes $m = 1, \ldots, M$. Les biais des estimées multivariées $\hat{H}_1^{(\mathrm{M})}, \ldots, \hat{H}_M^{(\mathrm{M})}$ diffèrent en revanche, rendant compte de la croissance de l'effet de répulsion entre les fonctions de structure $\log_2 \hat{\lambda}_1(2^j), \ldots, \log_2 \hat{\lambda}_M(2^j)$ au travers des échelles d'analyse $2^{j_1(N)} \leq 2^j \leq 2^{j_2(N)}$.

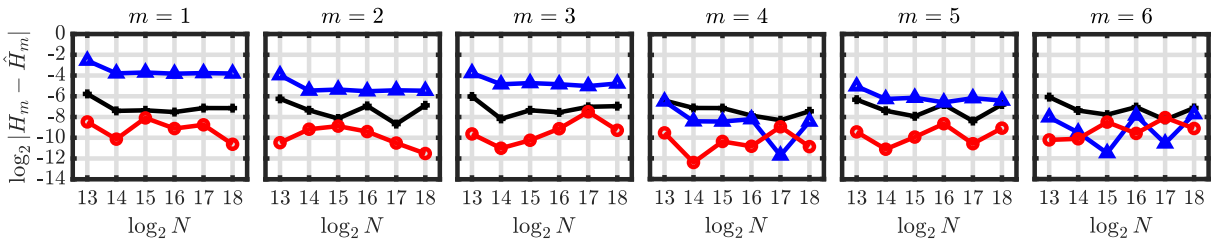


FIGURE 2.13 – **Biais des estimateurs par composante dans un cadre avec effet de répulsion croissant.** Logarithmes des biais en valeur absolue des estimées (en noir) univariées $\hat{H}_m^{(\mathrm{U})}$, (en bleu) multivariées $\hat{H}_m^{(\mathrm{M})}$ et (en rouge) multivariées corrigées $\hat{H}_m^{(\mathrm{M,bc})}$ pour (de gauche à droite) $m = 1, \ldots, 6$ en fonction de la taille d'échantillon N pour des mBf corrélés ($\Sigma \neq \mathbb{I}$) et mélangés ($W \neq \mathbb{I}$) avec des exposants H_m tous égaux, pour différentes tailles d'échantillon N.

Scénario sans mélange

La figure 2.14 rapporte les performances en termes de biais, variance et erreur quadratique moyenne des différents estimateurs en fonction de la taille d'échantillon N pour des réalisations de mBf corrélés ($\Sigma \neq \mathbb{I}$) et non mélangés ($W = \mathbb{I}$) avec des exposants H_m tous différents. Ces résultats montrent que l'estimateur univarié $\underline{\hat{H}}^{(\mathrm{U})}$ atteint des performances meilleures en termes de biais que celles des estimateurs multivariés $\underline{\hat{H}}^{(\mathrm{M})}$ et $\underline{\hat{H}}^{(\mathrm{M,bc})}$, comme attendu puisque l'absence de mélange ($W = \mathbb{I}$) correspond à une configuration univariée. Néanmoins, la covariance des estimateurs multivariés $\underline{\hat{H}}^{(\mathrm{M})}$ et $\underline{\hat{H}}^{(\mathrm{M,bc})}$ est moins affectée que celle de l'estimateur univarié $\underline{\hat{H}}^{(\mathrm{U})}$. Les performances en termes d'erreur quadratique moyenne sont ainsi équivalentes pour les différents estimateurs : il existe un compromis biais-covariance entre l'estimateur univarié $\underline{\hat{H}}^{(\mathrm{U})}$ et les estimateurs multivariés $\underline{\hat{H}}^{(\mathrm{M})}$ et $\underline{\hat{H}}^{(\mathrm{M,bc})}$. Cependant, les performances en termes de biais, et donc d'erreur quadratique moyenne, de l'estimateur multivarié corrigé $\underline{\hat{H}}^{(\mathrm{M,bc})}$ s'améliorent rapidement avec la taille d'échantillon N, semblant rattraper les performances, plus constantes, de l'estimateur univarié.

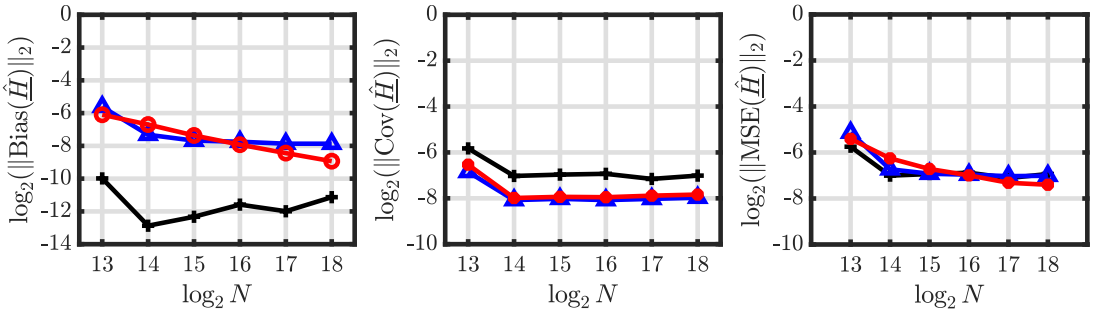


FIGURE 2.14 – **Performances des estimateurs dans un cadre sans mélange.** Logarithmes des normes spectrales des biais, variances et erreurs quadratiques moyennes des estimateurs (en noir) univarié $\underline{\hat{H}}^{(\mathrm{U})}$, (en bleu) multivarié $\underline{\hat{H}}^{(\mathrm{M})}$ et (en rouge) multivarié corrigé $\underline{\hat{H}}^{(\mathrm{M,bc})}$ en fonction de la taille d'échantillon N pour des mBf corrélés ($\Sigma \neq \mathbb{I}$) mais non mélangés ($W = \mathbb{I}$) avec des exposants H_m tous différents.

Enfin, la figure 2.15 rapporte les biais des estimées $\hat{H}_m^{(\mathrm{U})}$, $\hat{H}_m^{(\mathrm{M})}$ et $\hat{H}_m^{(\mathrm{M,bc})}$ de H_m pour les différentes composantes $m = 1, \ldots, M$ dans le cas de mBf corrélés ($\Sigma \neq \mathbb{I}$) et non mélangés ($W = \mathbb{I}$) avec des exposants H_m tous différents. Les biais des estimées univariées $\hat{H}_m^{(\mathrm{U})}$ varient peu au travers des composantes m, mais ceux des estimées multivariées $\hat{H}_m^{(\mathrm{M})}$ varient plus, quoique légèrement. En revanche, les biais des estimées multivariées corrigées $\hat{H}_m^{(\mathrm{M,bc})}$ dépendent davantage de m. Ceci est dû au fait que les écarts entre exposants $\Delta H_m = H_{m+1} - H_m$ sont plus faibles pour $m = 3$ et $m = 4$ ($\Delta H_m = 0.05$) que pour $m \in \{1, 2, 5\}$ ($\Delta H_m = 0.1$) et que l'effet de répulsion entre des fonctions de structure $\log_2 \hat{\lambda}_m(2^j)$ et $\log_2 \hat{\lambda}_{m'}(2^j)$ est plus important pour des exposants H_m et $H_{m'}$ proches.

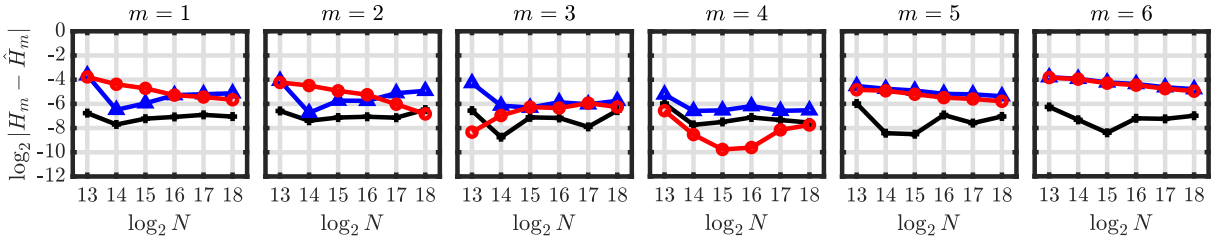


FIGURE 2.15 – **Biais des estimateurs par composante dans un cadre sans mélange.** Logarithmes des biais en valeur absolue des estimées (en noir) univariées $\hat{H}_m^{(\mathrm{U})}$, (en bleu) multivariées $\hat{H}_m^{(\mathrm{M})}$ et (en rouge) multivariées corrigées $\hat{H}_m^{(\mathrm{M,bc})}$ pour (de gauche à droite) $m = 1, \ldots, 6$ en fonction de la taille d'échantillon N pour des mBf corrélés ($\Sigma \neq \mathbb{I}$) et non mélangés ($W = \mathbb{I}$) avec des exposants H_m tous différents.

2.5.3.2 Normalité asymptotique

Les différents estimateurs étudiés sont asymptotiquement gaussiens sous des hypothèses faibles. Les théorèmes 1.7 et 2.3 assurent la normalité asymptotique des estimateurs multivariés $\underline{\hat{H}}^{(\mathrm{M})}$ et $\underline{\hat{H}}^{(\mathrm{M,bc})}$ sous la condition (C0), introduite dans la section 1.4.3.2. Quant à la normalité asymptotique de l'estimateur univarié $\underline{\hat{H}}^{(\mathrm{U})}$, elle est assurée par le théorème 1.1 en l'absence de mélange $W = \mathbb{I}$. On cherche à évaluer numériquement la vitesse de convergence vers la normalité asymptotique des différents estimateurs.

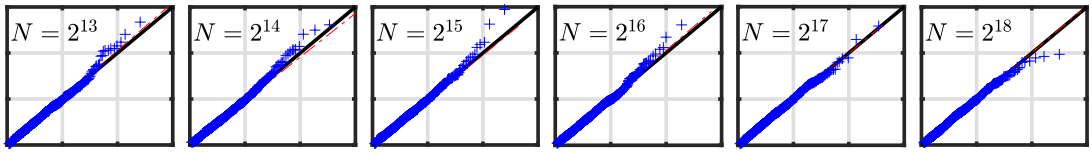
Plus précisément, l'approximation des distributions des différents estimateurs par une loi normale multivariée est étudiée pour différentes tailles d'échantillon finies N. Pour ce, on utilise le fait qu'un estimateur $\underline{\hat{H}}$ de $\underline{H}$ suit une loi normale multivariée si et seulement si le carré de la distance de Mahalanobis du vecteur $\underline{\hat{H}}$,

$$T := (\underline{\hat{H}} - \mathbb{E}[\underline{\hat{H}}])\ \mathrm{Var}(\underline{\hat{H}})^{-1}\ (\underline{\hat{H}} - \mathbb{E}[\underline{\hat{H}}])^T, \tag{2.38}$$

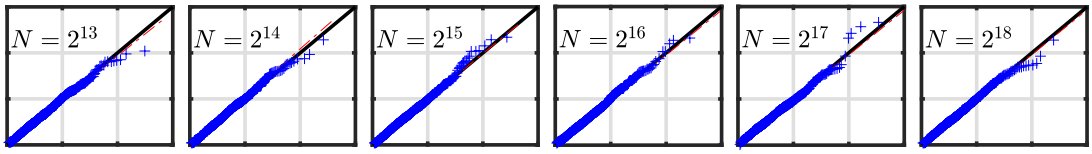
suit une loi du χ^2 à M degrés de liberté.

En premier lieu, on s'intéresse au cadre multivarié. La figure 2.16 rapporte la distribution de T au travers des réalisations de Monte Carlo contre une distribution théorique du χ^2 pour différentes tailles d'échantillon N et les différents estimateurs $\underline{\hat{H}}^{(\mathrm{U})}$, $\underline{\hat{H}}^{(\mathrm{M})}$ et $\underline{\hat{H}}^{(\mathrm{M,bc})}$ pour des mBf corrélés ($\Sigma \neq \mathbb{I}$) et mélangés ($W \neq \mathbb{I}$) avec des exposants H_m tous différents. On observe que, même pour une taille d'échantillon faible ($N = 2^{13}$), les distributions des différents estimateurs multivariés $\underline{\hat{H}}^{(\mathrm{M})}$ et $\underline{\hat{H}}^{(\mathrm{M,bc})}$ sont bien approximées par une loi normale multivariée dans ce scénario, où la condition (C0) est vérifiée. De plus, malgré la présence de mélange, la distribution de l'estimateur univarié $\underline{\hat{H}}^{(\mathrm{U})}$ est également bien approximée par une loi normale multivariée pour les différentes tailles d'échantillon N.

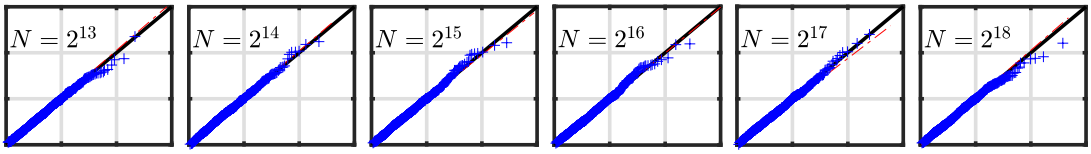
En pratique, on peut s'attendre à un écart à la loi normale multivariée des distributions des estimateurs multivariés lorsque la condition (C0) n'est pas vérifiée, autrement dit lorsque les valeurs propres $\lambda_1(2^j), \ldots, \lambda_M(2^j)$ sont proches à différentes échelles d'analyse $2^{j_1(N)} \leq 2^j \leq 2^{j_2(N)}$ (voir la section 1.4.3.2 pour plus de détails). Pour étudier ce cas limite, la figure 2.17 rapporte la distribution de T au travers des réalisations de Monte Carlo contre une distribution théorique du χ^2 à M degrés de liberté pour différentes tailles d'échantillon N et les différents estimateurs $\underline{\hat{H}}^{(\mathrm{U})}$, $\underline{\hat{H}}^{(\mathrm{M})}$ et $\underline{\hat{H}}^{(\mathrm{M,bc})}$ pour des mBf ni corrélés ($\Sigma = \mathbb{I}$) ni mélangés ($W = \mathbb{I}$) avec des exposants H_m tous égaux. Les distributions des différents estimateurs sont bien approximées par des distributions gaussiennes multivariées pour les différentes tailles d'échantillon N.



(a) Estimateur univarié $\underline{\hat{H}}^{(\mathrm{U})}$

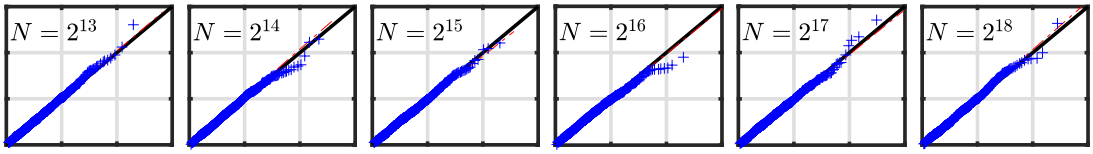


(b) Estimateur multivarié $\underline{\hat{H}}^{(\mathrm{M})}$

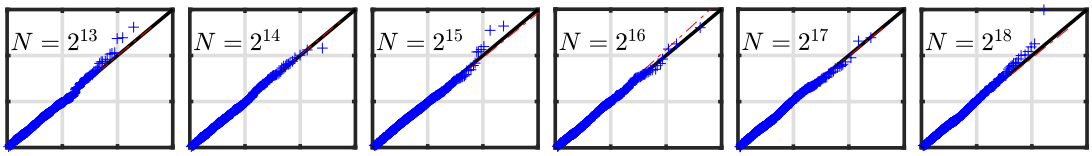


(c) Estimateur multivarié corrigé $\underline{\hat{H}}^{(\mathrm{M,bc})}$

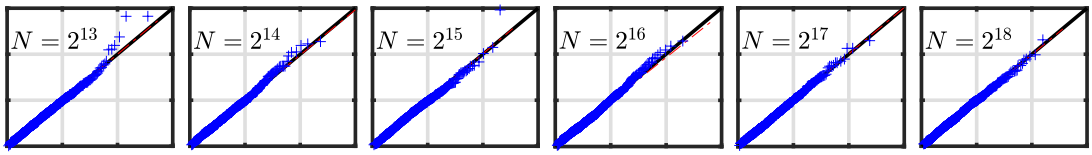
FIGURE 2.16 – **Normalité asymptotique dans un cadre multivarié.** Diagrammes quantile-quantile des distributions empiriques du carré de la distance de Mahalanobis des différents estimateurs (de haut en bas) contre une distribution du χ^2 à M degrés de liberté pour des mBf corrélés ($\Sigma \neq \mathbb{I}$) et mélangés ($W \neq \mathbb{I}$) avec des exposants H_m tous différents (condition (C0) valide), pour différentes tailles d'échantillons N (de gauche à droite).



(a) Estimateur univarié $\underline{\hat{H}}^{(\mathrm{U})}$



(b) Estimateur multivarié $\underline{\hat{H}}^{(\mathrm{M})}$



(c) Estimateur multivarié corrigé $\underline{\hat{H}}^{(\mathrm{M,bc})}$

FIGURE 2.17 – **Normalité asymptotique lorsque la condition (C0) est invalide.** Diagrammes quantile-quantile des distributions empiriques du carré de la distance de Mahalanobis des différents estimateurs (de haut en bas) contre une distribution du χ^2 théorique pour des mBf ni corrélés ($\Sigma = \mathbb{I}$) ni mélangés ($W = \mathbb{I}$) avec des exposants H_m tous égaux (condition (C0) invalide), pour différentes tailles d'échantillons N (de gauche à droite).

L'approximation par une loi gaussienne multivariée est donc valable pour les estimateurs multivariés $\underline{\hat{H}}^{(\mathrm{M})}$ et $\underline{\hat{H}}^{(\mathrm{M,bc})}$ dans un cadre assez large en pratique : à taille d'échantillon N faible et pour des fonctions de structure $\log_2 \hat{\lambda}_m(2^j)$ et $\log_2 \bar{\lambda}_m(2^j)$ (respectivement) proches les unes des autres (i.e. entre composantes m) au travers des échelles d'analyse $2^{j_1(N)} \leq 2^j \leq 2^{j_2(N)}$. Quant à l'estimateur univarié $\underline{\hat{H}}^{(\mathrm{U})}$, l'approximation de sa distribution par une loi gaussienne multivariée est également valable en pratique en présence de mélange ($W \neq \mathbb{I}$) et pour de faibles tailles d'échantillon N.

2.5.3.3 Décorrélation asymptotique

On se propose à présent d'étudier le comportement de la corrélation des estimées pour divers tailles d'échantillon et dans différents scénarios. Théoriquement, l'estimateur univarié $\underline{\hat{H}}^{(\mathrm{U})}$ est asymptotiquement décorrélé dans le cadre sans mélange ($W = \mathbb{I}$), selon le théorème 1.2, tandis que les estimateurs multivariés $\underline{\hat{H}}^{(\mathrm{M})}$ et $\underline{\hat{H}}^{(\mathrm{M,bc})}$ sont asymptotiquement décorrélés sous plusieurs hypothèses restrictives dont la condition (C0), introduite dans la section 1.4.3.2, selon la section 2.3.4. La validité de ces approximations est éprouvée dans cette section.

Scénario multivarié

La figure 2.18 présente les corrélations des différents estimateurs $\underline{\hat{H}}^{(\mathrm{U})}$, $\underline{\hat{H}}^{(\mathrm{M})}$ et $\underline{\hat{H}}^{(\mathrm{M,bc})}$ en fonction de la taille d'échantillon N pour des réalisations de mBf corrélés ($\Sigma \neq \mathbb{I}$) et mélangés ($W \neq \mathbb{I}$) avec des exposants H_m tous différents. Les corrélations des estimateurs multivariés $\underline{\hat{H}}^{(\mathrm{M})}$ et $\underline{\hat{H}}^{(\mathrm{M,bc})}$ tendent rapidement vers 0 lorsque N croît, corroborant les approximations de la section 2.3.4. En effet, même à taille d'échantillon faible $N = 2^{13}$, toutes les paires de composantes des estimateurs multivariés $\underline{\hat{H}}^{(\mathrm{M})}$ et $\underline{\hat{H}}^{(\mathrm{M,bc})}$ sont peu corrélées. En revanche, pour l'estimateur univarié $\underline{\hat{H}}^{(\mathrm{U})}$, les corrélations ne sont pas nécessairement nulles, même à grande taille d'échantillon N, et peuvent même être importantes. Ceci appuie la nécessité de l'hypothèse $W = \mathbb{I}$ dans le théorème 1.2 sur la décorrélation asymptotique de l'estimateur univarié $\underline{\hat{H}}^{(\mathrm{U})}$.

Scénario avec effet de répulsion croissant

La convergence de la corrélation des estimateurs multivariées $\underline{\hat{H}}^{(\mathrm{M})}$ et $\underline{\hat{H}}^{(\mathrm{M,bc})}$ vers 0 n'est vérifiée que si la condition (C0) est vérifiée, c'est-à-dire lorsque les valeurs propres théoriques $\lambda_1(2^j), \ldots, \lambda_M(2^j)$ sont distinctes pour toutes les échelles d'analyse $2^{j_1(N)} \leq 2^j \leq 2^{j_2(N)}$. Lorsque les exposants H_m sont égaux, les valeurs propres $\lambda_1(2^j), \ldots, \lambda_M(2^j)$ ne s'éloignent pas les unes des autres au travers des échelles d'analyse $2^{j_1(N)} \leq 2^j \leq 2^{j_2(N)}$, ce qui peut donner une convergence plus lente de la décorrélation des estimées multivariées. Ainsi, la figure 2.19 rapporte les corrélations entre les composantes des différents vecteurs d'estimées $\underline{\hat{H}}^{(\mathrm{U})}$, $\underline{\hat{H}}^{(\mathrm{M})}$ et $\underline{\hat{H}}^{(\mathrm{M,bc})}$ en fonction de la taille d'échantillon N pour des réalisations de mBf corrélés ($\Sigma \neq \mathbb{I}$) et mélangés ($W \neq \mathbb{I}$) avec des exposants H_m tous égaux. On peut observer que la corrélation entre des estimées multivariées $\hat{H}_m^{(\mathrm{M})}$ et $\hat{H}_{m'}^{(\mathrm{M})}$ et entre des estimées multivariées corrigées $\hat{H}_m^{(\mathrm{M,bc})}$ et $\hat{H}_{m'}^{(\mathrm{M,bc})}$, associées à des valeurs propres $\lambda_m(2^j)$ et $\lambda_{m'}(2^j)$ proches, en l'occurrence pour $(m, m') = (1, 2)$ et $(m, m') = (3, 4)$, n'est pas négligeable mais converge tout de même vers 0. En outre, dans ces cas-ci, la corrélation est légèrement plus grande pour l'estimateur multivarié corrigé $\underline{\hat{H}}^{(\mathrm{M,bc})}$ que pour l'estimateur multivarié $\underline{\hat{H}}^{(\mathrm{M})}$.

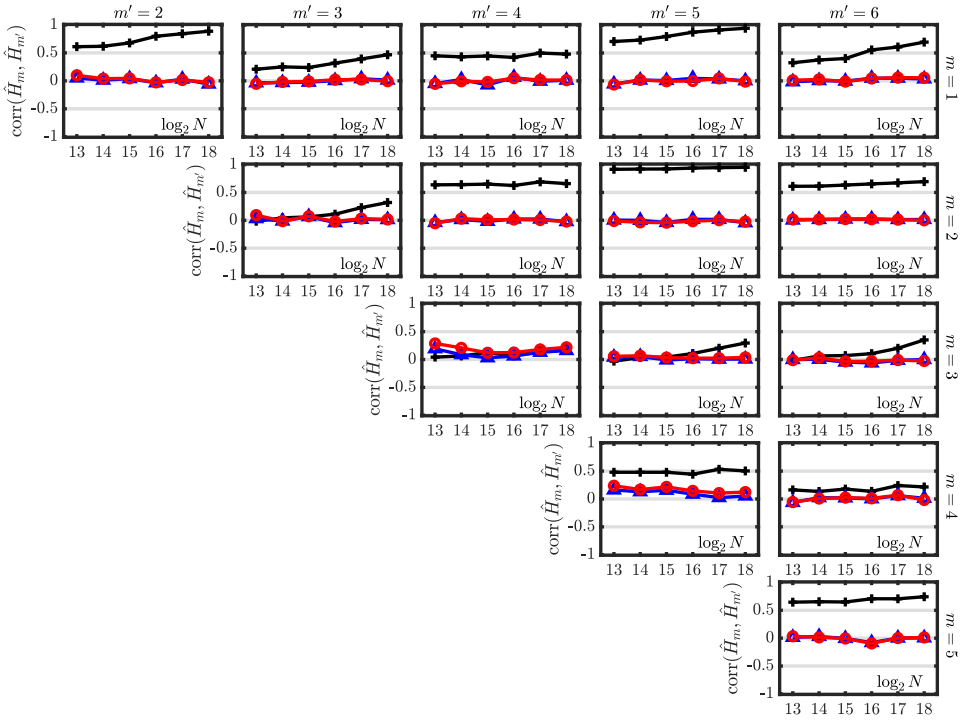


FIGURE 2.18 – **Corrélations des estimateurs dans un cadre multivarié.** Corrélations entre les différentes entrées des estimateurs (en noir) univarié $\underline{\hat{H}}^{(\mathrm{U})}$, (en bleu) multivarié $\underline{\hat{H}}^{(\mathrm{M})}$ et (en rouge) multivarié corrigé $\underline{\hat{H}}^{(\mathrm{M,bc})}$ en fonction de la taille d'échantillon N pour des mBf corrélés ($\Sigma \neq \mathbb{I}$) et mélangés ($W \neq \mathbb{I}$) avec des exposants H_m tous différents (condition (C0) valide).

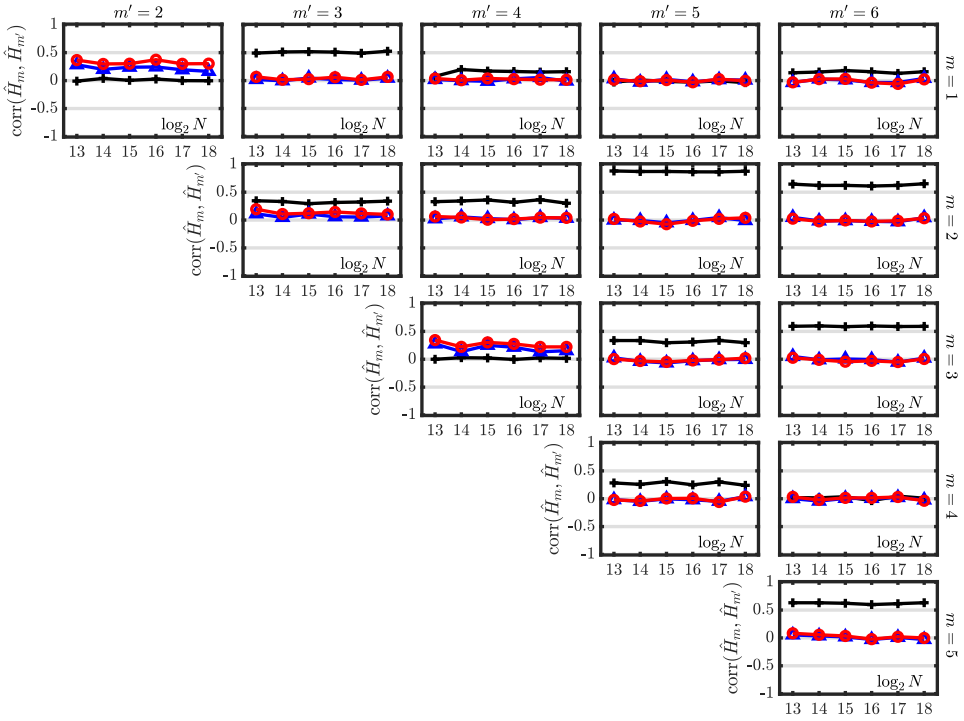


FIGURE 2.19 – **Corrélations des estimateurs dans un cadre avec effet de répulsion croissant.** Corrélations entre les différentes composantes des estimateurs (en noir) univarié $\underline{\hat{H}}^{(\mathrm{U})}$, (en bleu) multivarié $\underline{\hat{H}}^{(\mathrm{M})}$ et (en rouge) multivarié corrigé $\underline{\hat{H}}^{(\mathrm{M,bc})}$ en fonction de la taille d'échantillon N pour des mBf corrélés ($\Sigma \neq \mathbb{I}$) et mélangés ($W \neq \mathbb{I}$) avec des exposants H_m tous égaux (condition (C0) valide).

Scénario sans mélange

La décorrélation des estimateurs multivariés $\underline{\hat{H}}^{(\mathrm{M})}$ et $\underline{\hat{H}}^{(\mathrm{M,bc})}$ est encore valable en l'absence de mélange ($W = \mathbb{I}$), mais on se propose d'étudier le comportement de la corrélation de l'estimateur univarié $\underline{\hat{H}}^{(\mathrm{U})}$ dans ce cas. La figure 2.20 rapporte la corrélation des différents estimateurs $\underline{\hat{H}}^{(\mathrm{U})}$, $\underline{\hat{H}}^{(\mathrm{M})}$ et $\underline{\hat{H}}^{(\mathrm{M,bc})}$ en fonction de la taille d'échantillon N pour des réalisations de mBf corrélés ($\Sigma \neq \mathbb{I}$) mais non mélangés ($W = \mathbb{I}$) avec des exposants H_m tous différents. Comme attendu, les estimateurs multivariés $\underline{\hat{H}}^{(\mathrm{M})}$ et $\underline{\hat{H}}^{(\mathrm{M,bc})}$ sont bien tous deux décorrélés : la corrélation des estimateurs multivariés est aussi faible qu'en présence de mélange. En revanche, les estimées univariées $\hat{H}_m^{(\mathrm{U})}$ et $\hat{H}_{m'}^{(\mathrm{U})}$ sont d'autant plus corrélées que des exposants H_m et $H_{m'}$ sont proches. Cette observation confirme le compromis biais-covariance entre les estimateurs multivarié et univarié dans un cadre sans mélange ($W = \mathbb{I}$) observé dans la figure 2.14.

Nécessité de la condition (C0)

On se propose d'étudier les estimateurs dans un cas critique où la condition (C0) n'est pas vérifiée, c'est-à-dire lorsque les valeurs propres $\lambda_m(2^j)$ ne sont pas distinctes à toutes les échelles d'analyse $2^{j_1(N)} \leq 2^j \leq 2^{j_2(N)}$. Ce cas peut être rencontré pour des mBf non corrélés ($\Sigma = \mathbb{I}$) et non mélangés ($W = \mathbb{I}$) avec des exposants H_m tous égaux. La figure 2.21 rapporte dans ce cas-ci les corrélations entre des estimées $\hat{H}_m$ et $\hat{H}_{m'}$, avec $1 \leq m < m' \leq M$, issues des différents estimateurs en fonction de la taille d'échantillon N. Les estimateurs multivariés $\underline{\hat{H}}^{(\mathrm{M})}$ et $\underline{\hat{H}}^{(\mathrm{M,bc})}$ sont très corrélés, notamment pour les paires (m, m') associées à des valeurs propres $\lambda_m(2^j)$ et $\lambda_{m'}(2^j)$ proches, la corrélation atteignant 0.5 pour $m' = m + 1$ (compte tenu de l'ordre des valeurs propres $\lambda_1(2^j) \leq \ldots \leq \lambda_M(2^j)$ à une échelle 2^j fixée). De plus, cette corrélation est constante au travers des tailles d'échantillon N. La décorrélation asymptotique n'est donc plus valable pour des valeurs propres $\lambda_1(2^j), \ldots, \lambda_M(2^j)$ égales. En revanche, comme attendu en l'absence de mélange ($W = \mathbb{I}$), l'estimateur univarié $\underline{\hat{H}}^{(\mathrm{U})}$ est décorrélé à toute taille d'échantillon N.

2.5.3.4 Approximation de la variance

On souhaite enfin étudier numériquement pour des tailles d'échantillon finies les différentes approximations sur la variance des estimateurs données par le théorème 1.2 et la section 2.3.4, à savoir, pour tout $m \in \{1, \ldots, M\}$,

$$\mathrm{Var}(\hat{H}_m^{(\mathrm{U})}) \approx \mathrm{Var}^{\mathrm{th}}, \quad \mathrm{Var}(\hat{H}_m^{(\mathrm{M})}) \approx \mathrm{Var}^{\mathrm{th}}, \quad \mathrm{Var}(\hat{H}_m^{(\mathrm{M,bc})}) \approx \mathrm{Var}^{\mathrm{th}}, \tag{2.39}$$

où

$$\mathrm{Var}^{\mathrm{th}} \approx \frac{(\log_2 e)^2}{2} \sum_{j=j_1^0}^{j_2^0} \frac{w_j^2}{n_{a,j}}, \tag{2.40}$$

de sorte que $\mathrm{Var}(\hat{H}_m^{(\mathrm{U})})$, $\mathrm{Var}(\hat{H}_m^{(\mathrm{M})})$ et $\mathrm{Var}(\hat{H}_m^{(\mathrm{M,bc})})$ ne dépendent ni de $\underline{H}$, ni de W, ni de Σ.

Pour l'estimateur univarié $\underline{\hat{H}}^{(\mathrm{U})}$, les approximations sont valides asymptotiquement pour un M-mBf non mélangé ($W = \mathbb{I}$) selon le théorème 1.2. Pour les estimateurs multivariés $\underline{\hat{H}}^{(\mathrm{M})}$ et $\underline{\hat{H}}^{(\mathrm{M,bc})}$, les approximations sont valides asymptotiquement en présence de mélange ($W \neq \mathbb{I}$) mais sous plusieurs hypothèses restrictives, selon la section 2.3.4.

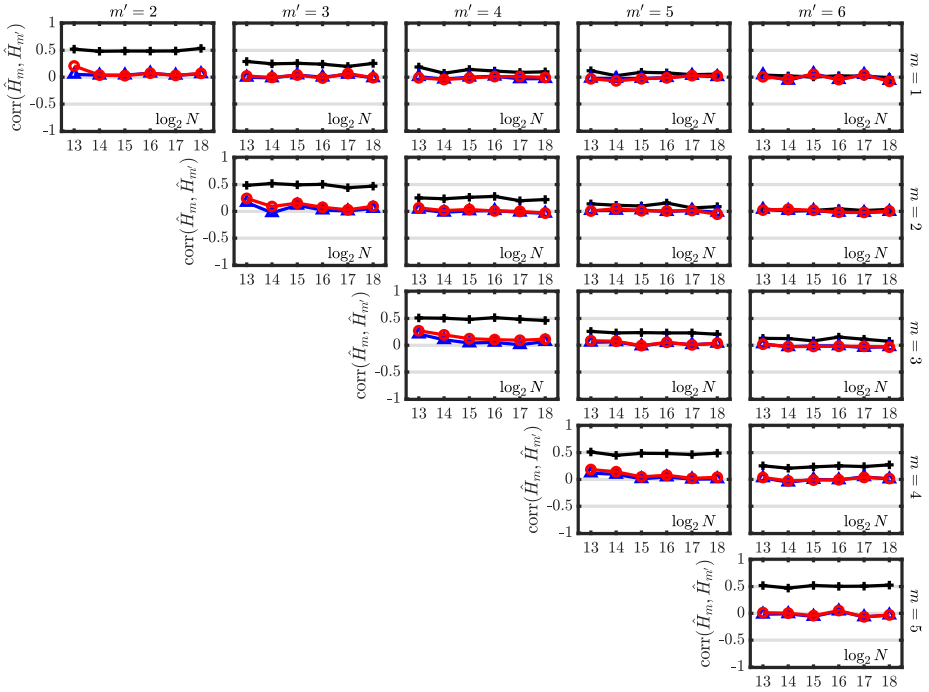


FIGURE 2.20 – **Corrélations des estimateurs dans un cadre sans mélange.** Corrélations entre les différentes composantes des estimateurs (en noir) univarié $\underline{\hat{H}}^{(\mathrm{U})}$, (en bleu) multivarié $\underline{\hat{H}}^{(\mathrm{M})}$ et (en rouge) multivarié corrigé $\underline{\hat{H}}^{(\mathrm{M,bc})}$ en fonction de la taille d'échantillon N pour des mBf corrélés ($\Sigma \neq \mathbb{I}$) et mélangés ($W \neq \mathbb{I}$) avec des exposants H_m tous différents (condition (C0) valide).

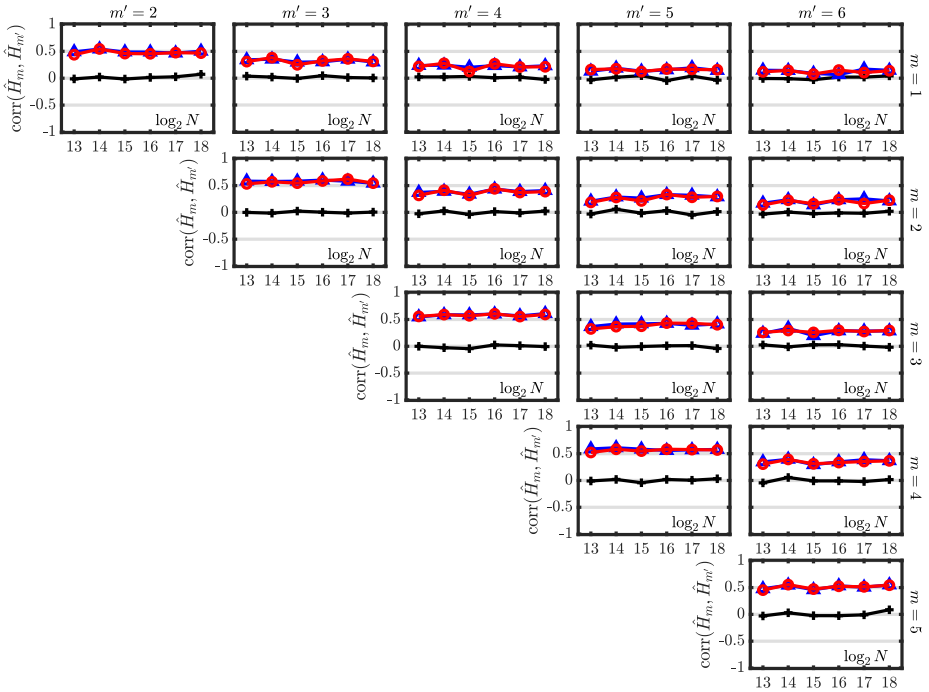


FIGURE 2.21 – **Corrélations des estimateurs lorsque la condition (C0) est invalide.** Corrélations entre les différentes composantes des estimateurs (en noir) univarié $\underline{\hat{H}}^{(\mathrm{U})}$, (en bleu) multivarié $\underline{\hat{H}}^{(\mathrm{M})}$ et (en rouge) multivarié corrigé $\underline{\hat{H}}^{(\mathrm{M,bc})}$ en fonction de la taille d'échantillon N pour des mBf non corrélés ($\Sigma = \mathbb{I}$) et non mélangés ($W = \mathbb{I}$) avec des exposants H_m tous égaux (condition (C0) invalide).

La figure 2.22 rapporte les erreurs relatives,

$$\mathrm{ER}(\mathrm{Var}^{\mathrm{obs}}) = \frac{\mathrm{Var}^{\mathrm{obs}} - \mathrm{Var}^{\mathrm{th}}}{\mathrm{Var}^{\mathrm{th}}}, \tag{2.41}$$

des variances empiriques $\mathrm{Var}^{\mathrm{obs}}$ des estimateurs par rapport à l'approximation $\mathrm{Var}^{\mathrm{th}}$ en fonction de la taille d'échantillon N pour différents ensembles de paramètres $(\underline{H}, W, \Sigma)$. On constate les points suivants :

(*i*) La figure 2.22 montre que, pour un vecteur $\underline{H}$ donné, $\mathrm{Var}(\hat{H}_m^{(\mathrm{U})})$ dépend peu de H_m. Ceci est également vrai pour $\mathrm{Var}(\hat{H}_m^{(\mathrm{M})})$ asymptotiquement, tandis que pour $\mathrm{Var}(\hat{H}_m^{(\mathrm{M,bc})})$, cette convergence est moins rapide ;

(*ii*) Les figures 2.22 (a) et 2.22 (b) montrent que $\mathrm{Var}(\hat{H}_m^{(\mathrm{U})})$ et $\mathrm{Var}(\hat{H}_m^{(\mathrm{M})})$ dépendent peu de $\underline{H}$ et que $\mathrm{Var}(\hat{H}_m^{(\mathrm{M,bc})})$ en dépend faiblement ;

(*iii*) Les figures 2.22 (a) et 2.22 (c) montrent que $\mathrm{Var}(\hat{H}_m^{(\mathrm{U})})$, $\mathrm{Var}(\hat{H}_m^{(\mathrm{M})})$ et $\mathrm{Var}(\hat{H}_m^{(\mathrm{M,bc})})$ dépendent peu de W ;

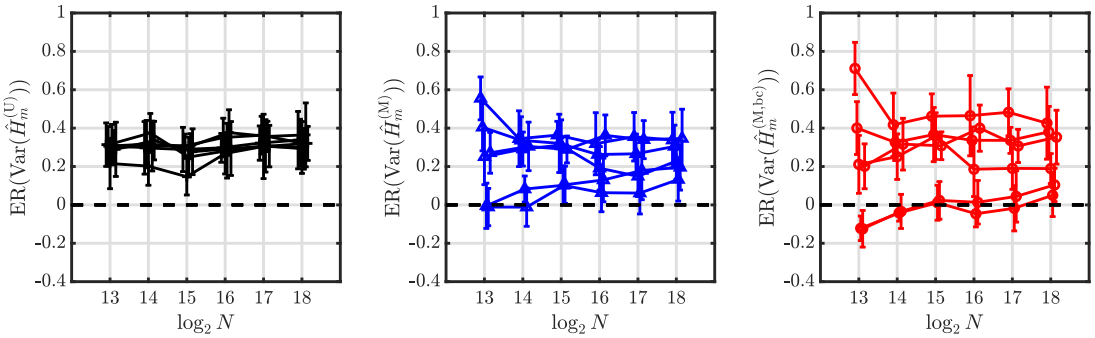
(*iv*) Les figures 2.22 (a) et 2.22 (d) montrent que $\mathrm{Var}(\hat{H}_m^{(\mathrm{U})})$ dépend peu de Σ et que $\mathrm{Var}(\hat{H}_m^{(\mathrm{M})})$ et $\mathrm{Var}(\hat{H}_m^{(\mathrm{M,bc})})$ en dépendent faiblement.

En conclusions, les approximations de la variance sont satisfaisantes asymptotiquement pour les estimateurs multivariés $\underline{\hat{H}}^{(\mathrm{M})}$ et $\underline{\hat{H}}^{(\mathrm{M,bc})}$, et dépendent peu de la configuration en jeu tant que la condition (C0) est vérifiée, mais la convergence vers l'approximation est moins rapide pour l'estimateur corrigé $\underline{\hat{H}}^{(\mathrm{M,bc})}$. De plus, en comparaison à l'estimateur univarié $\underline{\hat{H}}^{(\mathrm{U})}$, il s'agit davantage d'un résultat asymptotique pour les estimateurs multivariés $\underline{\hat{H}}^{(\mathrm{M})}$ et $\underline{\hat{H}}^{(\mathrm{M,bc})}$.

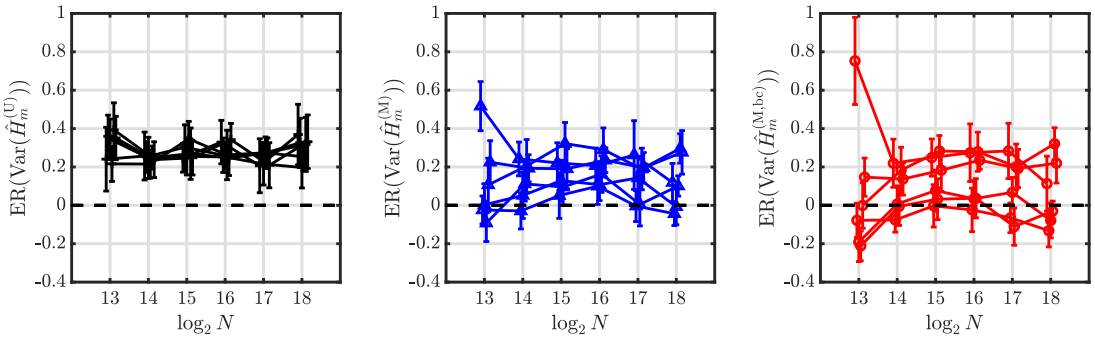
2.6 Conclusion

Dans ce chapitre, un nouvel estimateur du vecteur des exposants d'autosimilarité $\underline{H}$ régissant l'autosimilarité multivariée a été proposé et étudié théoriquement et numériquement. Cet estimateur réduit le biais de taille finie de l'estimateur multivarié $\underline{\hat{H}}^{(\mathrm{M})}$ introduit dans le chapitre 1, dû à un effet de répulsion entre les valeurs propres estimées $\hat{\lambda}_m(2^j)$ qui croît au travers des échelles 2^j. La procédure d'estimation proposée calcule plusieurs spectres d'ondelettes à partir du même nombre de coefficients d'ondelettes à chaque échelle 2^j pour reproduire le même effet de répulsion entre les valeurs propres estimées à toutes les échelles.

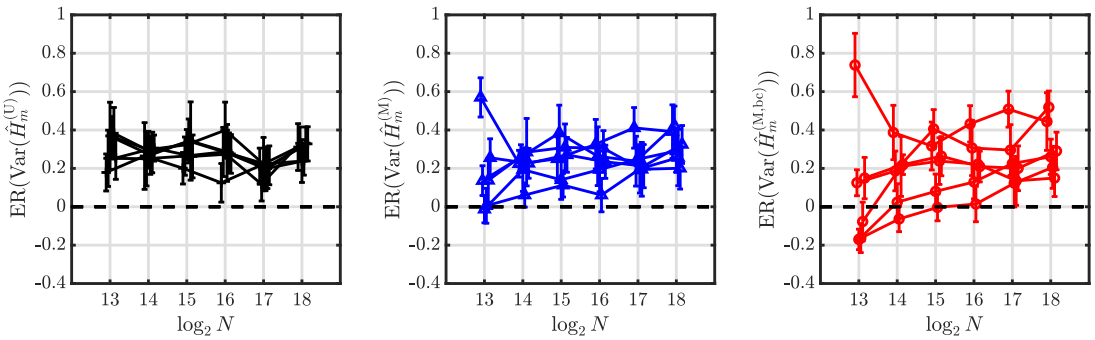
L'estimateur multivarié corrigé présente des propriétés théoriques importantes. En particulier, sa consistance et sa normalité asymptotique sont assurées sous des conditions peu restrictives. De plus, une comparaison numérique des estimateurs univarié et multivariés a été réalisé sur des M-mBf synthétiques, cas particuliers de mBof plus adaptés aux applications pratiques, pour différentes tailles d'échantillon et différentes configurations permettant d'évaluer l'impact du mélange W, de la corrélation Σ et de l'égalité des exposants d'autosimilarité H_m des mBf dont est issu le M-mBf. Au vu des résultats théoriques et numériques, l'estimateur multivarié corrigé $\underline{\hat{H}}^{(\mathrm{M,bc})}$ atteint des performances très satisfaisantes dans un cadre multivarié, notamment robustes à l'égalité entre certains exposants d'autosimilarité contrairement à l'estimateur multivarié $\underline{\hat{H}}^{(\mathrm{M})}$. En outre, les estimées multivariées corrigées $\hat{H}_m^{(\mathrm{M,bc})}$ ont l'avantage de voir leur variance décroître rapidement avec la taille d'échantillon N et d'être quasi-décorrélées dans des configurations assez générales et même à faible taille d'échantillon N. Plus particulièrement, un grand avantage de l'estimateur multivarié corrigé $\underline{\hat{H}}^{(\mathrm{M,bc})}$ est que sa covariance est similaire



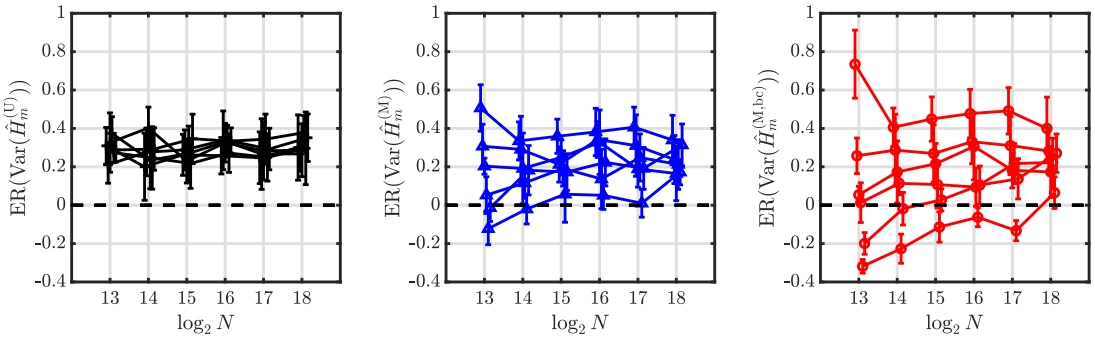
(a) H_m tous différents, $\Sigma \neq \mathbb{I}$ et $W \neq \mathbb{I}$



(b) H_m tous égaux, $\Sigma \neq \mathbb{I}$ et $W \neq \mathbb{I}$



(c) H_m tous différents, $\Sigma \neq \mathbb{I}$ et $W = \mathbb{I}$



(d) H_m tous différents, $\Sigma = \mathbb{I}$ et $W \neq \mathbb{I}$

FIGURE 2.22 – **Approximation de la variance.** Erreurs relatives entre les variances empiriques et théoriques des estimées (à gauche en noir) univariées $\hat{H}_m^{(\mathrm{U})}$, (au milieu en bleu) multivariées $\hat{H}_m^{(\mathrm{M})}$ et (à droite en rouge) multivariées corrigées $\hat{H}_m^{(\mathrm{M,bc})}$ avec superposition des composantes $m = 1, \ldots, 6$ en fonction de la taille d'échantillon N, pour (de haut en bas) différentes configurations.

à celle de l'estimateur multivarié $\underline{\hat{H}}^{(\mathrm{M})}$, n'induisant pas de compromis biais-covariance entre l'utilisation des deux estimateurs.

Cependant, la procédure d'estimation multivariée corrigée peut se heurter à un croisement des fonctions de structure : le changement d'ordre des valeurs propres est perdu lors de leur estimation si bien que les logarithmes de leurs estimées ne sont pas linéaires sur une certaine gamme d'échelle où a lieu un tel changement d'ordre. Cette procédure nécessite donc un choix judicieux des échelles d'analyse qui permette à la fois d'éviter les croisements des fonctions de structure, de maximiser le nombre de coefficients d'ondelettes disponibles et de maximiser le nombre d'échelles d'analyse.

Chapitre 3

Dénombrement et regroupement d'exposants d'autosimilarité

Sommaire

3.1 Introduction

Dans le chapitre 2, un estimateur robuste des exposants d'autosimilarité $\underline{H} = (H_1, \ldots, H_M)$ a été proposé : l'estimateur multivarié corrigé $\hat{\underline{H}}^{(\mathrm{M,bc})} = (\hat{H}_1^{(\mathrm{M,bc})}, \ldots, \hat{H}_M^{(\mathrm{M,bc})})$ est donné par l'équation (2.3) et est simplement noté $\hat{\underline{H}} = (\hat{H}_1, \ldots, \hat{H}_M)$ dans l'ensemble du présent chapitre. En pratique, il peut être utile de savoir combien de valeurs distinctes d'exposants d'autosimilarité sont en jeu dans le modèle d'autosimilarité multivariée et combien d'exposants prennent chacune de ces valeurs. Or, les estimées obtenues $\hat{H}_1, \ldots, \hat{H}_M$ diffèrent quand bien même certains des exposants d'autosimilarité $H_1, \ldots, H_M$ sont égaux. Il est donc important de mesurer cette différence qui peut être significative ou simplement une conséquence des fluctuations de l'estimateur. Dans cette optique, différents tests d'hypothèse paramétriques, exploitant la normalité asymptotique de l'estimateur multivarié corrigé étudiée dans le chapitre 2, sont proposés dans ce chapitre. Ceux-ci sont conçus pour être mis en œuvre à partir d'une seule observation de données multivariées. L'estimation des paramètres des tests fait ainsi appel à une procédure bootstrap (Zoubir et Iskander, 2004), dans l'esprit de celles développées par Wendt et collab. (2018, 2019, 2007), décrite dans la section 3.2.

Une première contribution de ce chapitre consiste à tester si les exposants sont tous égaux ou non, c'est-à-dire l'hypothèse nulle $H_1 = \ldots = H_M$. Cette procédure de test est décrite en section 3.3. Le rejet de l'hypothèse nulle indique que plusieurs valeurs sont potentiellement présentes dans $\underline{H}$. Un tel test ne permet cependant pas de connaître le nombre de valeurs distinctes. D'autres stratégies consistant à tester des paires d'exposants sont donc étudiées. Une première stratégie de partitionnement détaillée en section 3.4 consiste à tester l'égalité entre $M-1$ paires d'exposants ordonnés successifs $H_m = H_{m+1}$, avec $H_1 \leq \ldots \leq H_M$. Cette approche repose sur des statistiques de test dont la loi exacte est inconnue et est donc approchée. La seconde stratégie de partitionnement, détaillée en section 3.5, consiste à tester l'égalité entre toutes les paires d'exposants $H_m = H_{m'}$, avec $1 \leq m < m' \leq M$, en réalisant ainsi $M(M-1)/2$ tests d'hypothèse, plutôt que $M-1$, à partir de statistiques de test dont les distributions asymptotiques sont bien connues grâce aux résultats du chapitre 2. La difficulté de cette stratégie réside dans la multiplicité des hypothèses de test. Le dénombrement des exposants est alors réalisé par une méthode de partitionnement d'un graphe construit à partir de ces tests. La pertinence et la robustesse des différentes procédures sont évaluées numériquement sur des M-mBf synthétiques pour différentes tailles d'échantillon N et différents nombres de composantes M. En complément, la section 3.6 présente une comparaison des différentes procédures sur un nombre de composantes plus élevé ($M = 20$) et en illustre les avantages et inconvénients.

3.2 Bootstrap dans le domaine des ondelettes.

Pour construire un test d'hypothèse, il est nécessaire de définir un seuil de rejet, nécessitant la connaissance du comportement de la statistique de test choisie. Dans les sections suivantes, différentes statistiques de tests sont construites à partir de l'estimateur $\hat{\underline{H}}$. On détaille dans cette section une procédure de ré-échantillonnage bootstrap permettant d'estimer des paramètres de la distribution de $\hat{\underline{H}}$, tels que sa covariance. Pour tirer parti des dépendances temporelles à court terme des coefficients d'ondelettes, une procédure bootstrap est mise en œuvre dans le domaine des ondelettes plutôt que dans le domaine temporel.

Les coefficients vectoriels d'ondelettes $D(2^j, k)$, pour tous $k \in \{1, \ldots, n_j\}$ et $j \in \{j_1, \ldots, j_2\}$, sont ré-échantillonnés conjointement (c'est-à-dire de manière multivariée) par une procédure bootstrap par blocs qui préserve leur structure de dépendance en temps et en composantes, par opposition à un ré-échantillonnage pour chaque composante indépendamment, qui ne préserve-

rait pas les dépendances entre composantes (Lahiri, 2003). Techniquement, pour chaque échelle 2^j, R blocs de ré-échantillons bootstrap,

$$\forall r \in \{1, \dots, R\}, \quad D^{*(r,j)} := \left(D^{*(r)}(2^j, 1), \dots, D^{*(r)}(2^j, n_j) \right), \tag{3.1}$$

sont tirés avec remise parmi $\lceil n_j / L_B \rceil$ blocs de coefficients d'ondelettes de taille L_B,

$$\forall k \in \{1, \dots, n_j\}, \quad \left(D(2^j, k), \dots, D(2^j, k + L_B - 1) \right), \tag{3.2}$$

où n_j est le nombre de coefficients d'ondelettes disponibles à l'échelle 2^j. Cette procédure est illustrée par la figure 3.1.

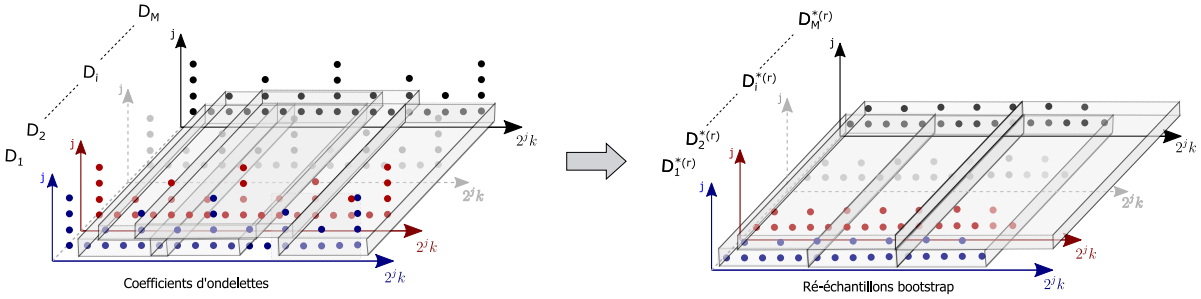


Figure 3.1 – **Illustration de la procédure bootstrap par blocs d'ondelettes multivariés.** À chaque échelle 2^j, des blocs temps-composantes de coefficients d'ondelettes $D(2^j, \cdot)$ de même taille L_B en temps, éventuellement superposés, sont tirés uniformément avec remise (à gauche) pour obtenir un ré-échantillon bootstrap $D^{*(r,j)}$ (à droite).

Les échantillons bootstrap de coefficients d'ondelettes $D^{*(r,j)}$ ainsi obtenus sont donc issus de la distribution empirique des coefficients d'ondelettes observés $D(2^j, 1), \dots, D(2^j, n_j)$. La procédure d'estimation introduite dans le chapitre 2 est alors appliquée à chaque ré-échantillon bootstrap $D^{*(r,j)}$: des estimées bootstrap $S^{*(r,w)}(2^j)$, pour tout $w \in \{1, \dots, 2^{j_2 - j}\}$, puis $\log_2 \bar{\lambda}_m^{*(r)}(2^j)$ et enfin $\hat{H}_m^{*(r)}$ sont successivement produites à l'aide des équations (2.1)-(2.3). Le principe de la stratégie bootstrap est illustré par la figure 3.2.

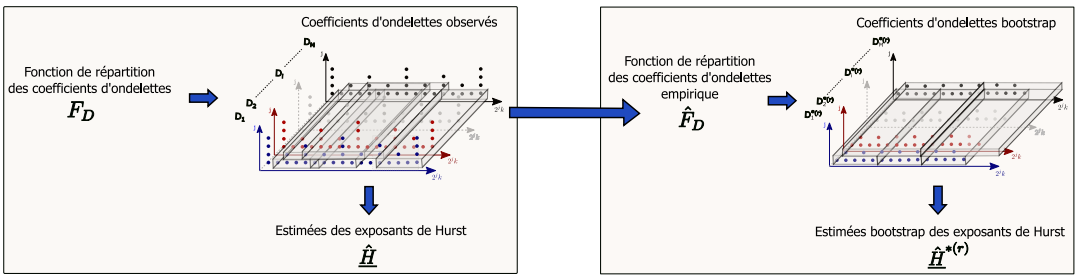


Figure 3.2 – **Illustration de l'estimateur bootstrap des exposants d'autosimilarité.**

Le choix de la taille L_B des blocs a été étudié par Wendt et collab. (2007). La taille optimale des blocs dépend du nombres de coefficients d'ondelettes disponibles n_j et varie donc à travers les échelles 2^j. Des simulations numériques menées par Wendt et collab. (2007) montrent que fixer une même taille de bloc L_B pour toutes les échelles 2^j a peu d'impact sur les performances de l'estimation bootstrap tant que $L_B \geq M_\psi$, où M_ψ est la taille du support de l'ondelette mère ψ_0 utilisée pour la transformée en ondelettes.

Dans les sections suivantes, les estimées bootstrap $\hat{H}_m^{*(r)}$ permettent notamment d'approximer la covariance de $\underline{\hat{H}}$ et la variance des différences entre estimées $\hat{H}_{m'} - \hat{H}_m$, avec $m' \neq m$.

3.3 Test d'égalité entre les exposants d'autosimilarité

Cette section est une version étendue de l'article suivant : C.-G. Lucas, P. Abry, H. Wendt et G. Didier, « Bootstrap for testing the equality of selfsimilarity exponents across multivariate time series », dans *2021 29th European Signal Processing Conference (EUSIPCO)*, IEEE, pp. 1960–1964.

L'objectif de cette section est de tester l'égalité de l'ensemble des exposants d'autosimilarité $H_1, \ldots, H_M$. L'hypothèse nulle s'écrit donc

$$\mathcal{H}_0 : \quad H_1 = H_2 = \ldots = H_M. \tag{3.3}$$

L'hypothèse alternative est notée

$$\mathcal{H}_1 : \quad \mathcal{H}_0 \text{ n'est pas vraie.} \tag{3.4}$$

3.3.1 Statistique du χ^2

Comme montré dans le chapitre 2, l'estimateur multivarié corrigé $\underline{\hat{H}} = (\hat{H}_1, \ldots, \hat{H}_M)$ de $\underline{H} = (H_1, \ldots, H_M)$ suit asymptotiquement une loi gaussienne multivariée. Ainsi une statistique classique (Lehmann et collab., 2005) pour tester l'hypothèse nulle $\mathcal{H}_0$ est

$$T := (\underline{\hat{H}} - \langle \underline{\hat{H}} \rangle \mathbf{1}_M)^T \, \Sigma_{\underline{\hat{H}}}^{-1} \, (\underline{\hat{H}} - \langle \underline{\hat{H}} \rangle \mathbf{1}_M) \tag{3.5}$$

où $\Sigma_{\underline{\hat{H}}}$ désigne la matrice de covariance de $\underline{\hat{H}}$ de taille $M \times M$, tandis que $\langle \underline{\hat{H}} \rangle$ correspond à la moyenne des estimées à travers les composantes,

$$\langle \underline{\hat{H}} \rangle := \frac{1}{M} \sum_{m=1}^{M} \hat{H}_m, \tag{3.6}$$

et $\mathbf{1}_M = (1, \ldots, 1)$. La normalité multivariée asymptotique de $\underline{\hat{H}} = (\hat{H}_1, \ldots, \hat{H}_M)$ implique que la statistique de test T suit asymptotiquement une loi du χ^2 à $M - 1$ degrés de liberté sous l'hypothèse nulle $\mathcal{H}_0$ (Cochran, 1952).

Le problème majeur est que la distribution de la statistique T est inconnue puisque la matrice $\Sigma_{\underline{\hat{H}}}$ est a priori inconnue. La matrice $\Sigma_{\underline{\hat{H}}}$ nécessite donc d'être remplacée par une estimée. Puisqu'il est ici prévu que le test proposé puisse être mis en œuvre à partir d'une seule observation de données multivariées, la matrice $\Sigma_{\underline{\hat{H}}}$ ne peut être estimée en moyennant au travers de réalisations, d'où le recours à la procédure bootstrap décrite dans la section 3.2. À partir des R échantillons bootstrap vectoriels $\underline{\hat{H}}^{*(r)} = (\hat{H}_1^{*(r)}, \ldots, \hat{H}_M^{*(r)})$, est estimée une matrice de covariance

$$\hat{\Sigma}^*_{\underline{\hat{H}}} = \mathrm{Var}^*(\underline{\hat{H}}^*), \tag{3.7}$$

qui devrait se rapprocher de la véritable matrice de covariance de $\underline{\hat{H}} = (\hat{H}_1, \ldots, \hat{H}_M)$.

3.3.2 Formulation du test bootstrap

À partir des estimées $\underline{\hat{H}}$ et de la matrice de covariance bootstrap $\hat{\Sigma}^*_{\underline{\hat{H}}}$, la statistique de test suivante est construite :

$$T^* := (\underline{\hat{H}} - \langle \underline{\hat{H}} \rangle \mathbf{1}_M)^T \, (\hat{\Sigma}^*_{\underline{\hat{H}}})^{-1} \, (\underline{\hat{H}} - \langle \underline{\hat{H}} \rangle \mathbf{1}_M). \tag{3.8}$$

La construction de cette statistique est résumée par la figure 3.3.

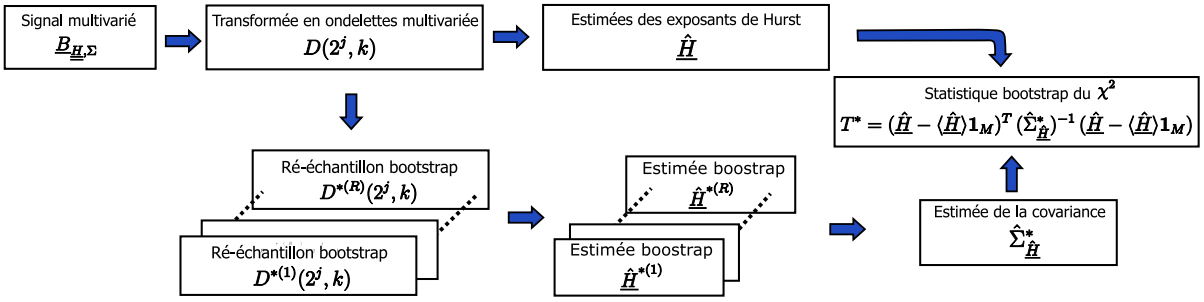


FIGURE 3.3 – **Organigramme de la procédure de test bootstrap du χ^2.**

La distribution de la statistique bootstrap T^* est inconnue mais devrait être bien approximée par une distribution du χ^2 pour une taille d'échantillon N suffisamment grande, comme vérifié par des résultats numériques dans la section 3.3.5.2, ce qui donne la décision de test suivante :

$$\begin{cases} d_\alpha = 1 \quad (\mathcal{H}_0 \text{ rejetée}) & \text{si } T^* \geq F^{-1}_{\chi^2_{M-1}}(1-\alpha), \\ d_\alpha = 0 \quad (\mathcal{H}_0 \text{ non rejetée}) & \text{sinon,} \end{cases} \tag{3.9}$$

avec $F_{\chi^2(M-1)}$ la fonction de répartition du χ^2 à $M-1$ degrés de liberté et $\alpha \in (0,1)$ le niveau de confiance.

3.3.3 Estimation de la puissance du test bootstrap

Sous une hypothèse alternative $\mathcal{H}_1$, la statistique T suit asymptotiquement une loi du χ^2 non centrée à $M-1$ degrés de liberté de paramètre de non-centralité

$$\theta := \mathbb{E}[\underline{\hat{H}} - \langle \underline{\hat{H}} \rangle] \, \Sigma^{-1}_{\underline{\hat{H}}} \, \mathbb{E}[\underline{\hat{H}} - \langle \underline{\hat{H}} \rangle]^T. \tag{3.10}$$

Ainsi, la puissance théorique du test (non bootstrap) du χ^2 s'écrit

$$\pi(\theta) := \mathbb{P}\left(T > F^{-1}_{\chi^2_{M-1}}(1-\alpha) \,|\, \mathcal{H}_1 \right) \approx 1 - F_{\chi^2_{M-1}(\theta)}\left(F^{-1}_{\chi^2_{M-1}}(1-\alpha) \right), \tag{3.11}$$

pour une taille d'échantillon N grande, avec $F_{\chi^2_{M-1}(\theta)}$ la fonction de répartition du χ^2 à $M-1$ degrés de liberté de paramètre de non-centralité θ et α le niveau de confiance.

Étant donné que la matrice $\Sigma_{\underline{\hat{H}}}$ est inconnue, le paramètre θ l'est également. Celui-ci peut cependant être estimé par bootstrap. En effet, puisque $\mathbb{E}[T] = M-1+\theta$, la statistique $T-(M-1)$ est un estimateur non biaisé de θ. Sous l'hypothèse que la statistique T est bien approximée par la statistique bootstrap T^*, un estimateur bootstrap non biaisé de θ est alors

$$\hat{\theta}^* := \max(0, T^* - (M-1)). \tag{3.12}$$

Il en découle l'estimateur bootstrap $\pi(\hat{\theta}^*)$ de la puissance $\pi(\theta)$ du test du χ^2. Voir Beran (1986) pour plus de détails sur l'estimation de puissances de test.

3.3.4 Synthèse des notations et formules

La figure 3.4 illustre la stratégie de test et résume différentes notations importantes.

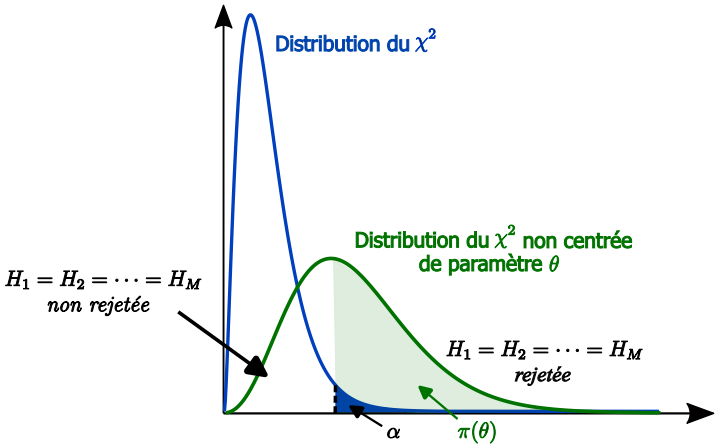


Figure 3.4 – **Illustration du test du χ^2.** La statistique T suit une loi du χ^2 non centrée. L'hypothèse nulle $\mathcal{H}_0$ est rejetée lorsque la statistique observée est supérieure à un seuil qui dépend de α, car la statistique observée appartient alors aux valeurs les moins probables de la distribution du χ^2 centrée (en bleu). Plus la distribution sous une hypothèse alternative (en vert) s'en écarte, c'est-à dire a un paramètre θ grand, plus la probabilité de rejeter l'hypothèse nulle $H_1 = \ldots = H_M$ est forte.

Formulaire récapitulatif

Les différentes quantités relatives au test de l'hypothèse $\mathcal{H}_0$ $(H_1 = \ldots = H_M)$ sont rappelées dans le tableau suivant.

Définition	Estimateur bootstrap
Statistique de test	
$T = (\underline{\hat{H}} - \langle \underline{\hat{H}} \rangle \mathbf{1}_M)^T \, \Sigma_{\underline{\hat{H}}}^{-1} \, (\underline{\hat{H}} - \langle \underline{\hat{H}} \rangle \mathbf{1}_M)$	$T^* = (\underline{\hat{H}} - \langle \underline{\hat{H}} \rangle \mathbf{1}_M)^T \, (\hat{\Sigma}^*_{\underline{\hat{H}}})^{-1} \, (\underline{\hat{H}} - \langle \underline{\hat{H}} \rangle \mathbf{1}_M)$
Décision de rejet du test	
Inconnue	d_α
Paramètre de non centralité	
$\theta = \mathbb{E}[\underline{\hat{H}} - \langle \underline{\hat{H}} \rangle] \, \Sigma_{\underline{\hat{H}}}^{-1} \, \mathbb{E}[\underline{\hat{H}} - \langle \underline{\hat{H}} \rangle]^T$	$\hat{\theta}^* = \max(0, T^* - (M-1))$
Puissance	
$\pi(\theta) = 1 - F_{\chi^2_{M-1}(\theta)} \left(F^{-1}_{\chi^2_{M-1}}(1-\alpha) \right)$	$\pi(\hat{\theta}^*) = 1 - F_{\chi^2_{M-1}(\hat{\theta}^*)} \left(F^{-1}_{\chi^2_{M-1}}(1-\alpha) \right)$

3.3.5 Évaluation des performances des estimateurs et du test

3.3.5.1 Simulations de Monte Carlo

Pour évaluer la reproduction du niveau de confiance ciblé du test sous l'hypothèse nulle et la puissance du test sous plusieurs hypothèses alternatives représentatives, des simulations de Monte Carlo sont effectuées avec des données synthétiques. Les simulations numériques reposent sur $N_{\mathrm{MC}} = 1000$ réalisations indépendantes de M-mBf synthétiques (cf. Section 2.4) avec $M = 4$ ou $M = 10$ composantes et pour trois tailles d'échantillon différentes $N \in \{2^{14}, 2^{16}, 2^{18}\}$. La matrice de covariance Σ des M-mBf est définie comme une matrice de Toeplitz, avec des entrées hors diagonale toutes fixées à $r = 0.5$. La matrice de mélange W des M-mBf est choisie aléatoirement dans l'espace des matrices inversibles de taille $M \times M$. Les deux matrices sont maintenues fixes pour toutes les expériences.

Sous l'hypothèse nulle $\mathcal{H}_0$, les exposants d'autosimilarité sont fixés à $H_1 = \ldots = H_M = 0.7$ et on étudie plusieurs niveaux de confiance $\alpha \in \{0.005, 0.01, 0.05, 0.1, 0.15\}$. Pour l'hypothèse alternative $\mathcal{H}_1$, le niveau de confiance est fixé à $\alpha = 0.05$ et deux scénarios sont testés :

(i) Dans le Scenario1, il y a seulement deux valeurs différentes dans $\underline{H}$, notées H_1 et H_2, parmi les M composantes. Dans le Scenario1a, les deux groupes de composantes correspondant à H_1 et H_2 sont de taille égale $M_1 = M_2 = M/2$; dans le Scenario1b, ils ont des tailles déséquilibrées $(M_1, M_2) = (1, 3)$ pour $M = 4$ et $(M_1, M_2) = (2, 8)$ pour $M = 10$. En outre, $H_1 = 0.5$ est fixé et H_2 varie dans l'intervalle $[0.5, 0.9]$.

(ii) Dans le Scenario2, il y a quatre groupes de composantes de tailles $(M_1, M_2, M_3, M_4) = (2, 3, 2, 3)$ pour $M = 10$ et $(M_1, M_2, M_3, M_4) = (1, 1, 1, 1)$ pour $M = 4$, avec des exposants d'autosimilarité $(H_1 = 0.5, H_2 = H_1 + \Delta H, H_3 = H_1 + 2\Delta H, H_4 = H_1 + 3\Delta H)$, où $\Delta H \in [0, 0.13]$.

Pour l'estimation, la transformée en ondelettes multivariée est calculée avec l'ondelette mère de Daubechies la moins asymétrique à $N_\psi = 3$ moments nuls (Daubechies, 1992) et les octaves d'estimation sont fixées à $j_1 = 6$ et $j_2 = \log_2(N) - 5$. Pour le test, on utilise $R = 500$ ré-échantillons bootstrap avec une taille de bloc $L_B = 6$ correspondant à la taille du support temporel de l'ondelette mère de Daubechies 3.

3.3.5.2 Comportement du test sous l'hypothèse nulle

Le test du χ^2 est construit à partir de la loi gaussienne multivariée asymptotique des estimées $\hat{H}_1, \ldots, \hat{H}_M$ qui permet de supposer que la statistique T suit une loi du χ^2 sous l'hypothèse nulle $\mathcal{H}_0$. La figure 3.5 rapporte les diagrammes quantile-quantile de la distribution empirique de T obtenue à partir des N_{MC} réalisations de Monte Carlo par rapport à la distribution théorique du χ^2 à $M - 1$ degrés de liberté. Ils montrent, pour trois tailles d'échantillon différentes N et deux nombres de composantes différents M, un excellent accord entre les deux distributions, validant ainsi l'hypothèse de la section 3.3 sur la distribution de la statistique de test T sous $\mathcal{H}_0$.

Ensuite, la figure 3.6 rapporte les diagrammes quantile-quantile de la distribution empirique de la statistique bootstrap T^* par rapport à la distribution empirique de la statistique T, toutes deux obtenues à partir des N_{MC} réalisations de Monte Carlo. Pour les trois tailles d'échantillon différentes N et les deux nombres différents de composantes M, la statistique T est reproduite de façon satisfaisante par la statistique bootstrap T^*, suggérant le bon comportement de l'estimateur bootstrap de la covariance $\hat{\Sigma}^*_{\underline{\hat{H}}}$. En outre, pour confirmer que la statistique T^* est bien adaptée à un test du χ^2, la figure 3.7 montre que les distributions empiriques des p-valeurs correspondantes sont approximativement uniformes sous l'hypothèse nulle $\mathcal{H}_0$, comme cela doit théoriquement être le cas.

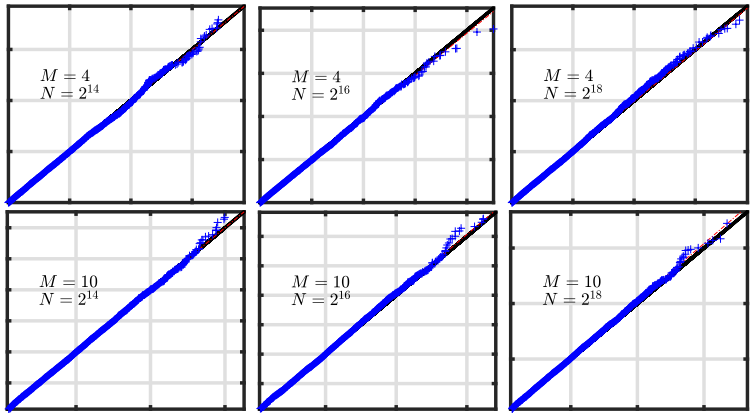


FIGURE 3.5 – **Distribution de T sous $\mathcal{H}_0$.** Diagrammes quantile-quantile de la distribution empirique de T sous $\mathcal{H}_0$ contre une loi du χ^2 à $M-1$ degrés de liberté pour plusieurs nombres de composantes M (de haut en bas) et différentes tailles d'échantillon N (de gauche à droite).

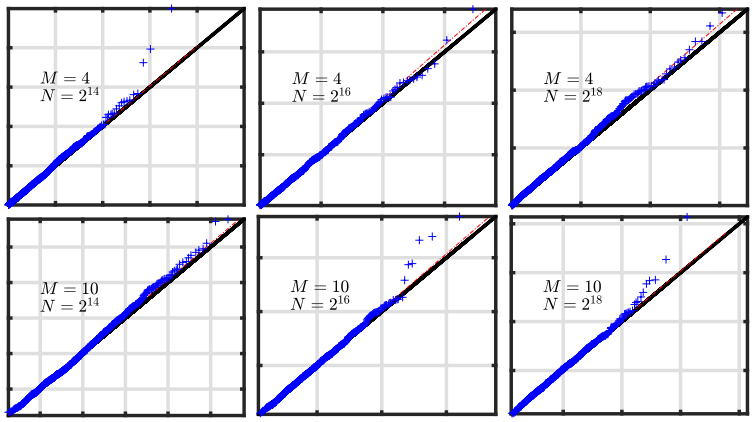


FIGURE 3.6 – **Distribution de T^* sous $\mathcal{H}_0$.** Diagrammes quantile-quantile de la distribution empirique de T^* contre la distribution empirique de T sous $\mathcal{H}_0$ pour plusieurs nombres de composantes M (de haut en bas) et différentes tailles d'échantillon N (de gauche à droite).

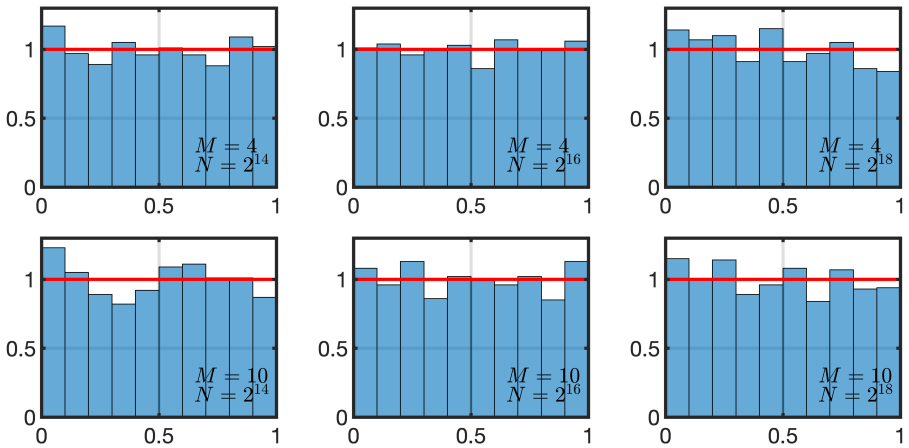


FIGURE 3.7 – **Distributions des p-valeurs sous $\mathcal{H}_0$.** Distributions empiriques des p-valeurs du test bootstrap proposé sous $\mathcal{H}_0$ comparées à la distribution uniforme (ligne en rouge) pour différents nombres de composantes M (de haut en bas) et différentes tailles d'échantillon N (de gauche à droite).

Enfin, la figure 3.8 montre, pour les trois tailles d'échantillon différentes N et les deux nombres différents de composantes M, que les niveaux de confiance atteints, obtenus comme moyennes des décisions binaires d_α du test bootstrap sur l'ensemble des réalisations de Monte Carlo, reproduisent de manière très satisfaisante les niveaux de confiance ciblé α du test.

Dans l'ensemble, ces résultats suggèrent que, pour de larges plages de tailles d'échantillon N et de nombres de composantes M, la statistique bootstrap T^* reproduit bien la distribution de la statistique T et suit bien une distribution du χ^2 à $M-1$ degrés de liberté, et que la procédure de test bootstrap proposée est efficace pour contrôler le niveau de confiance (erreur de type I) du test.

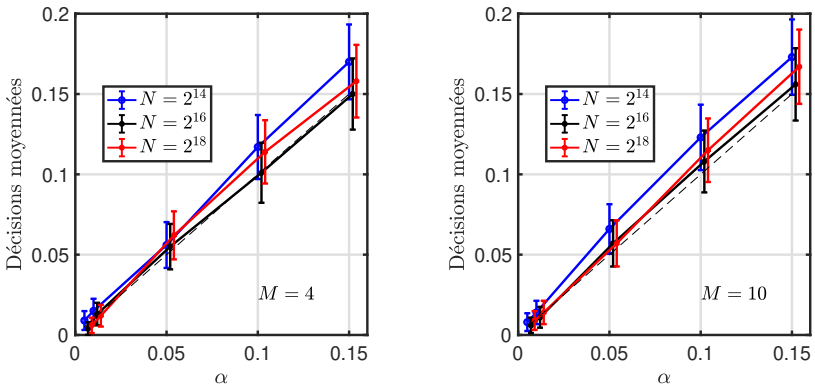


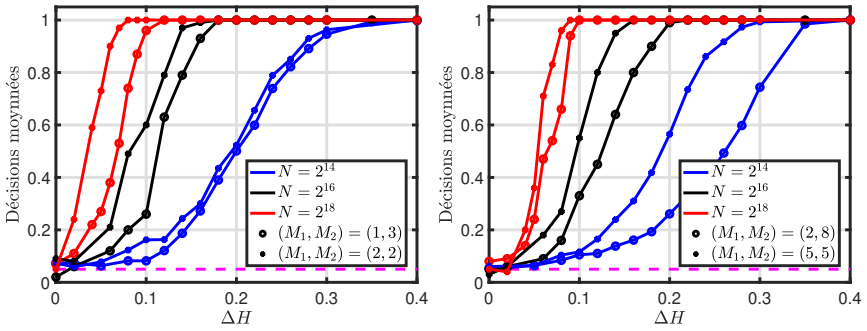
FIGURE 3.8 – **Niveaux de confiance.** Niveaux de confiance empiriques (obtenus en tant que moyennes des décisions du test bootstrap au travers des réalisations de Monte Carlo) en fonction des niveaux de confiance ciblés α sous $\mathcal{H}_0$ pour plusieurs tailles d'échantillons N et nombres de composantes M. Les lignes noires pointillées indiquent les valeurs théoriques attendues, c'est-à-dire la première bissectrice.

3.3.5.3 Puissance empirique du test

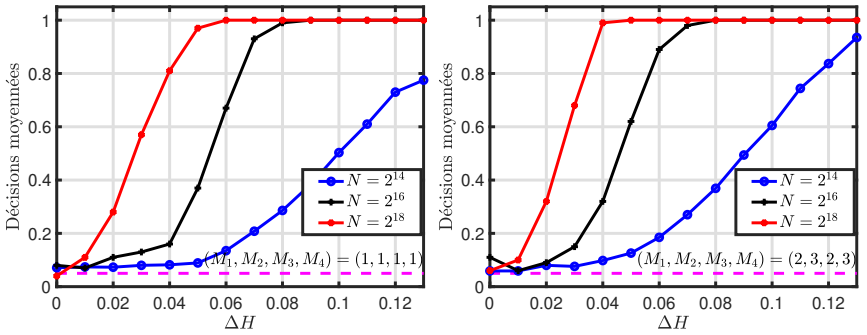
La figure 3.9 présente l'évaluation empirique (par moyenne sur des réalisations de Monte Carlo indépendantes des décisions de test) de la puissance du test bootstrap du χ^2 proposé en fonction de ΔH pour les scénarios 1 et 2. Dans les deux scénarios, et comme prévu, la puissance du test augmente avec la taille d'échantillon N et avec l'écart ΔH.

De plus, la figure 3.9(a) montre que, pour un scénario avec deux valeurs d'exposants différentes et un ΔH fixé, la puissance du test diminue lorsque le nombre de composantes M augmente. Par ailleurs, et de manière intéressante, la comparaison des résultats des Scenario1a et Scenario1b montre que, pour un ΔH fixé, la puissance de test est plus grande lorsque les groupes de composantes égales ont des tailles égales, par rapport au cas où les groupes de composantes ont des tailles déséquilibrées, ce qui peut être expliqué comme suit. Pour le scénario 1, un calcul détaillé dans l'annexe B.1 montre que le paramètre de non-centralité θ (cf. Eq. (3.10)) est essentiellement proportionnel à $(M_1 \times M_2)/(M_1 + M_2)$ et diminue donc lorsque $|M_2 - M_1|$ augmente, de même que la puissance du test.

La figure 3.9(b) montre enfin que le test a également une puissance significative lorsqu'il y a plus de deux valeurs différentes dans $\underline{H}$.



(a) Scénario 1



(b) Scénario 2

FIGURE 3.9 – **Puissance du test.** Évaluation empirique des puissances de test, obtenues comme moyenne des décisions du test bootstrap, en fonction de ΔH, pour $M = 4$ (gauche) et $M = 10$ (droite) composantes, pour différentes tailles d'échantillon N, et avec des tailles égales ou différentes de groupes de composantes égales, pour Scenario1 (haut) et Scenario2 (bas). Le niveau de confiance est fixé à $\alpha = 0.05$ (ligne pointillée magenta).

3.3.5.4 Estimation du paramètre de non-centralité

Pour étudier le comportement de l'estimateur du paramètre de non-centralité θ, il est d'abord nécessaire de vérifier le comportement de la statistique du test sous l'hypothèse alternative $\mathcal{H}_1$. La figure 3.10 rapporte les diagrammes quantile-quantile de la distribution de la statistique T estimée à partir des N_{MC} réalisations de Monte Carlo contre une loi du χ^2 non centrée de paramètre θ, défini par l'équation (3.10), estimé au travers des réalisations de Monte Carlo, sous l'hypothèse alternative donnée par le scénario 2 pour $M = 10$ composantes. Pour différentes tailles d'échantillon N et différents écarts ΔH, la distribution de la statistique T est correctement approximée par une distribution du χ^2 non centrée.

Par ailleurs, la figure 3.11 rapporte les diagrammes quantile-quantile de la distribution de la statistique T^* contre la distribution de la statistique T, toutes deux estimées à partir des N_{MC} réalisations de Monte Carlo, sous le scénario 2. Pour $M = 10$ composantes, différentes tailles d'échantillon N et différents écarts ΔH, la distribution de la statistique T^* approxime de façon satisfaisante la distribution de la statistique T pour un écart ΔH petit. La qualité de l'approximation diminue quand l'écart ΔH augmente, particulièrement pour une grande taille d'échantillon $N = 2^{18}$. Ce comportement assure le bon comportement de l'estimateur θ^*, défini par l'équation (3.12), du paramètre de non-centralité θ de la statistique T pour de faibles valeurs de ΔH.

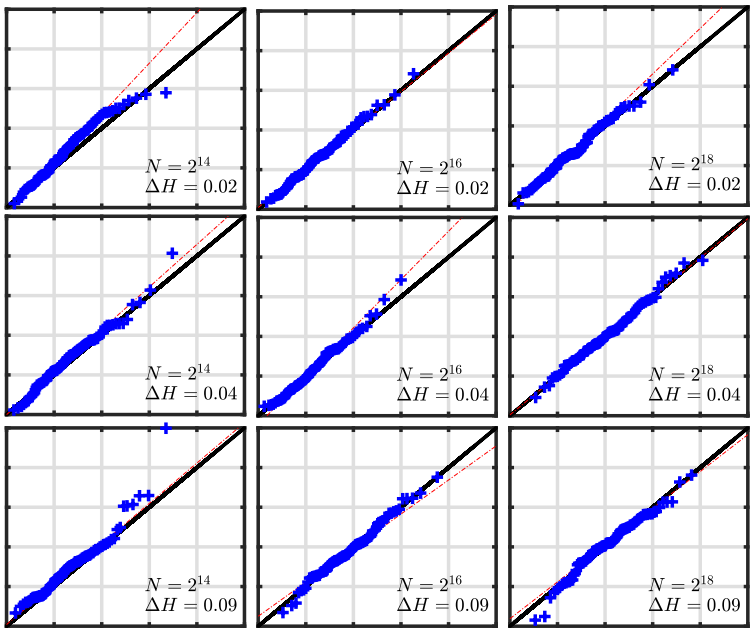


FIGURE 3.10 – **Distribution du χ^2 non centrée de T.** Diagramme quantile-quantile de la distribution empirique de T contre une distribution du χ^2 non centrée de paramètre θ estimé par Monte Carlo pour $M = 10$ composantes, différentes tailles d'échantillon N (de gauche à droite) et différents écarts ΔH (de haut en bas), sous le scénario 2.

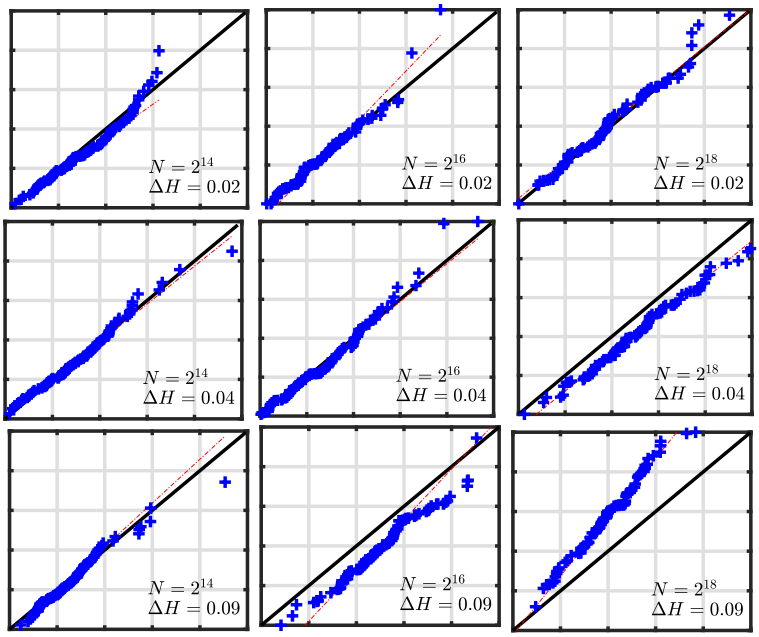
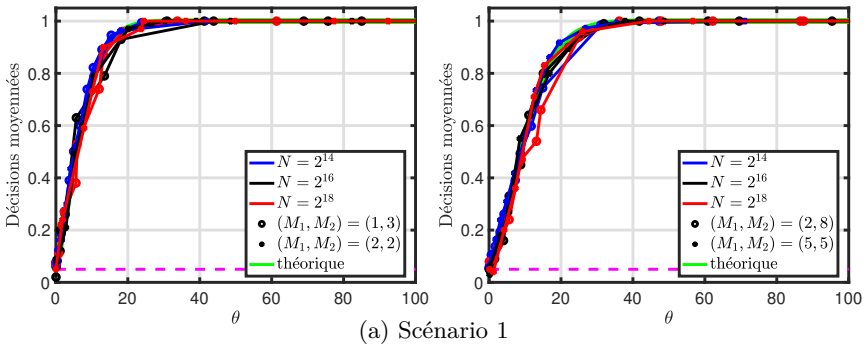


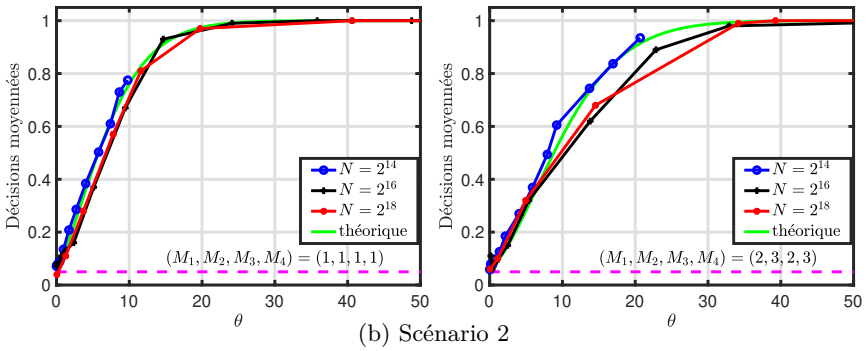
FIGURE 3.11 – **Reproduction de T par T^*.** Diagramme quantile-quantile de la distribution empirique de T^* contre la distribution empirique de T pour $M = 10$ composantes, différentes tailles d'échantillon N (de gauche à droite) et différents écarts ΔH (de haut en bas), sous le scénario 2.

3.3.5.5 Estimation de la puissance du test

Tout d'abord, la puissance empirique du test peut être étudiée en fonction du paramètre de non-centralité θ de la distribution χ^2 de la statistique T et ainsi être comparée à la puissance théorique $\pi(\theta)$ du test, définie par l'équation (3.11). La figure 3.12 rapporte la puissance empirique du test, calculée en moyennant les décisions du test au travers des réalisations de Monte Carlo, en fonction du paramètre de non-centralité θ (Eq. (3.10)) estimé par Monte Carlo, sous les différents scénarios. Quels que soient le nombre de composantes M et la taille d'échantillon N, la puissance empirique se superpose de manière satisfaisante à la puissance théorique $\pi(\theta)$ dans les différents scénarios.



(a) Scénario 1



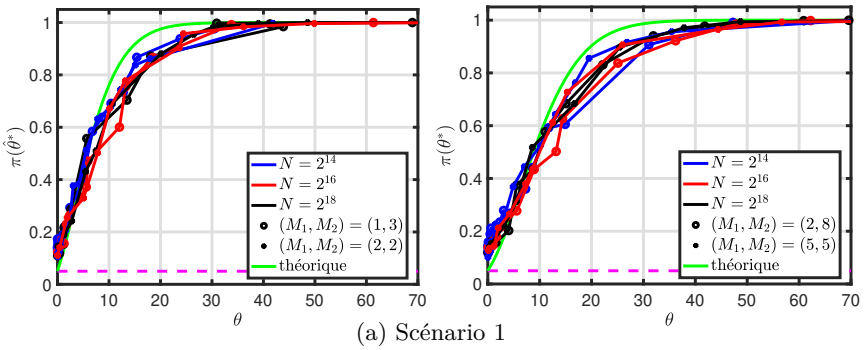
(b) Scénario 2

FIGURE 3.12 – **Puissance empirique et paramètre de non centralité.** Puissance théorique $\pi(\theta)$ (cf. Eq. (3.11)) et puissance empirique obtenue comme moyenne des décisions du test du χ^2 au travers des réalisations de Monte Carlo en fonction du paramètre de non-centralité θ (cf. Eq. (3.10)) estimé par Monte Carlo pour les différents scénarios et différentes tailles d'échantillon N.

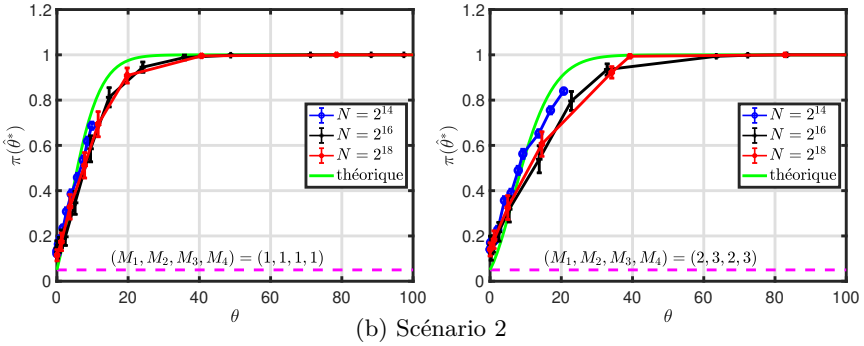
Enfin, la figure 3.13 montre la puissance du test estimée par bootstrap $\pi(\hat{\theta}^*)$ en fonction du paramètre de non-centralité θ estimé par Monte Carlo dans les différents scénarios considérés. L'estimée $\pi(\hat{\theta}^*)$ est une approximation satisfaisante de la puissance théorique $\pi(\hat{\theta})$ du test pour différentes tailles d'échantillon N et différents nombres de composantes M. On observe toutefois une légère sous-estimation de $\pi(\hat{\theta})$ qui s'explique par l'écart de T^* à T lorsque ΔH augmente observé sur la figure 3.11, suggérant un biais dans l'estimateur θ^*.

3.3.6 Conclusions

Dans cette section, a été évaluée la procédure bootstrap dans le domaine des ondelettes multivariées qui permet de tester, à partir d'une observation unique de taille finie d'une série temporelle multivariée, si tous les exposants d'autosimilarité sont égaux ou non. Des simulations de Monte Carlo, fondées sur des M-mBf synthétiques, ont permis de montrer que la procédure est efficace pour contrôler un niveau de confiance prescrit a priori et que le test proposé bénéficie d'une puissance importante sous plusieurs hypothèses alternatives. En outre, cette puissance peut être fidèlement estimée par bootstrap. Le test proposé est prêt à être appliqué aux séries temporelles multivariées du monde réel.



(a) Scénario 1



(b) Scénario 2

FIGURE 3.13 – **Estimée bootstrap de la puissance.** Estimée bootstrap $\pi(\hat{\theta}^*)$ (cf. Eq. (3.12)) de la puissance théorique $\pi(\theta)$ (cf. Eq. (3.11)) moyennée au travers des réalisations de Monte Carlo en fonction du paramètre de non-centralité θ (cf. Eq. (3.10)) estimé par Monte Carlo pour les différents scénarios et différentes tailles d'échantillon N.

3.4 Tests d'égalité par paires d'exposants successifs

Lorsque l'hypothèse nulle

$$\mathcal{H}_0 : \quad H_1 = H_2 = \ldots = H_M \tag{3.13}$$

est rejetée par la procédure précédente, l'enjeu devient de compter le nombre de valeurs distinctes dans $\underline{H} = (H_1, \ldots, H_M)$. Ceci n'est pas possible avec le test du χ^2, suggérant le recours à des tests par paires.

3.4.1 Formulation des tests

Par souci de simplicité, supposons que le vecteur $\underline{H} = (H_1, \ldots, H_M)$ est trié : $H_1 \leq \ldots \leq H_M$. Une stratégie pour estimer les groupes d'exposants d'autosimilarité égaux parmi $\underline{H}$ consiste à tester l'égalité des $M-1$ paires de composantes successives (H_m, H_{m+1}), avec $m \in \{1, \ldots, M-1\}$, dans $\underline{H}$, avec correction pour les décisions d'un test à hypothèses multiples. Les positions des non-égalités détectées parmi ces $M-1$ paires d'exposants d'autosimilarité ordonnés, s'il y en a, définissent alors les limites des groupes d'exposants d'autosimilarité distincts.

Les hypothèses nulles pour ces $M-1$ tests par paires sont donc définies comme suit :

$$\forall m \in \{1, \ldots, M-1\}, \quad \mathcal{H}_0^{(m)} : H_{m+1} = H_m. \tag{3.14}$$

Les hypothèses alternatives associées sont notées, pour tout $m \in \{1, \ldots, M-1\}$,

$$\mathcal{H}_1^{(m)} : \quad \mathcal{H}_0^{(m)} \text{ n'est pas vraie.} \tag{3.15}$$

3.4.2 Statistiques des tests

Pour construire un test pour chaque hypothèse $\mathcal{H}_0^{(m)}$, on calcule d'abord, à partir d'une observation unique de données M-variées de taille finie, le vecteur des M estimées $\underline{\hat{H}} = (\hat{H}_1, \ldots, \hat{H}_M)$, puis on trie ses entrées par ordre croissant, $\underline{\hat{H}}_\tau = (\hat{H}_{\tau(1)}, \ldots, \hat{H}_{\tau(M)})$ avec

$$\forall m \in \{1, \ldots, M-1\}, \quad \hat{H}_{\tau(m+1)} \geq \hat{H}_{\tau(m)}. \tag{3.16}$$

On calcule ensuite les $M-1$ statistiques de test définies par

$$\tilde{\delta}_m := \hat{H}_{\tau(m+1)} - \hat{H}_{\tau(m)}. \tag{3.17}$$

L'étude théorique et numérique menée dans le chapitre 2 assure que les entrées de $\underline{\hat{H}}$ sont asymptotiquement gaussiennes et faiblement dépendantes. Dans une configuration bivariée ($M = 2$), cela implique que $\tilde{\delta}_1$ suit asymptotiquement une loi normale repliée $\mathcal{FN}_{\tilde{\mu}_1, \tilde{\sigma}_1}$ de paramètres $\tilde{\mu}_1 = |\mathbb{E}[\hat{H}_2] - \mathbb{E}[\hat{H}_1]|$ et $\tilde{\sigma}_1 = \sqrt{\text{Var}\,(\hat{H}_1) + \text{Var}\,(\hat{H}_2)}$. Sous $\mathcal{H}_0^{(1)}$, la distribution de $\tilde{\delta}_1$ se résume à une loi demi-normale, c'est-à-dire $\tilde{\mu}_1 = 0$. La loi normale repliée est détaillée en annexe B.2. Quand $M > 2$, les statistiques $\tilde{\delta}_m$ ne sont pas asymptotiquement normales repliées, mais les simulations de Monte Carlo sur des M-mBf synthétiques rapportées dans la section 3.4.7 montrent que, au moins pour un faible nombre de composantes, c'est-à-dire $M \approx 6$, les distributions de $\tilde{\delta}_1, \ldots, \tilde{\delta}_{M-1}$, sont bien approximées par des distributions normales repliées $\mathcal{FN}_{\tilde{\mu}_m, \tilde{\sigma}_m}$, et donc, sous $\mathcal{H}_0^{(m)}$, par des distributions demi-normales $\mathcal{HN}_{\tilde{\sigma}_m} := \mathcal{FN}_{0, \tilde{\sigma}_m}$. Les paramètres $\tilde{\mu}_m$ et $\tilde{\sigma}_m$ sont respectivement appelés *paramètre de position* et *paramètre d'échelle*.

Les tests pour $\mathcal{H}_0^{(m)}$, pour tout $m \in \{1, \ldots, M-1\}$, peuvent être construits comme suit :

$$\text{rejeter } \mathcal{H}_0^{(m)} \text{ si } \tilde{\delta}_m > \gamma_m, \tag{3.18}$$

où chaque $\gamma_m > 0$ est un seuil de rejet. De façon équivalente, on peut écrire les tests à partir des p-valeurs $\tilde{p}_m$ des statistiques associées au rejet de $\mathcal{H}_0^{(m)}$, pour tout $m \in \{1, \ldots, M-1\}$,

$$\text{rejeter } \mathcal{H}_0^{(m)} \text{ si } \tilde{p}_m < \alpha_m, \tag{3.19}$$

où $\alpha_m > 0$ est un seuil de confiance et les p-valeurs $\tilde{p}_m$ peuvent être approximées par

$$\tilde{p}_m \approx 1 - F_{\mathcal{FN}(0, \tilde{\sigma}_m)}(\tilde{\delta}_m), \tag{3.20}$$

avec $F_{\mathcal{FN}(0, \tilde{\sigma}_m)}$ la fonction de répartition de la loi demi-normale de paramètre d'échelle $\tilde{\sigma}_m$. La procédure de test unilatéral fondée sur l'approximation des statistiques de test par une loi normale repliée est illustrée par la figure 3.14.

Pour réaliser les tests (3.19) (resp. les tests (3.18)), il faut estimer les paramètres d'échelle $\tilde{\sigma}_m$ des statistiques $\tilde{\delta}_m$ sous les hypothèses nulles $\mathcal{H}_0^{(m)}$ pour approximer les p-valeurs $\tilde{p}_m$ (resp. les seuils de rejet γ_m). Cette estimation est réalisée à partir de ré-échantillons bootstrap $\hat{H}_m^{*(1)}, \ldots, \hat{H}_m^{*(R)}$, obtenus selon la procédure bootstrap décrite dans la section 3.2. Cette procédure d'estimation du test par bootstrap est décrite dans la section suivante.

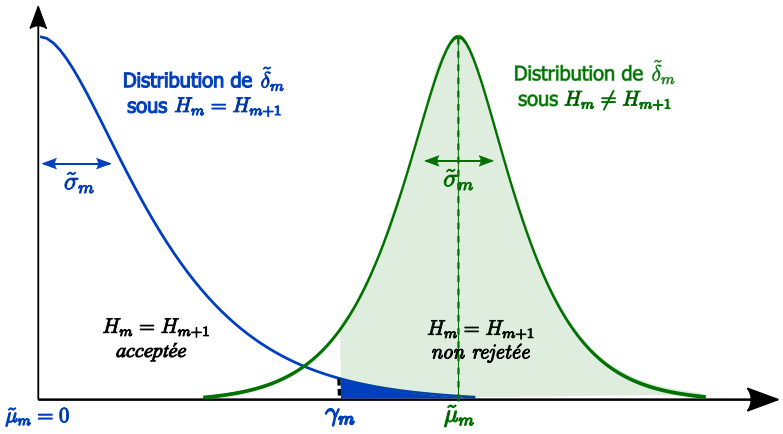


FIGURE 3.14 – **Illustration du test demi-normal.** La statistique $\tilde{\delta}_m$ est approximée par une loi demi-normale de paramètres $\tilde{\sigma}_m$ et $\tilde{\mu}_m$. L'hypothèse nulle $\mathcal{H}_0^{(m)}$ est rejetée lorsque la statistique observée est au-dessus d'un seuil γ_m, car la statistique observée appartient alors aux valeurs les moins probables pour la distribution demi-normale (en bleu). Plus la distribution normale repliée (en vert), sous une hypothèse alternative donnée, s'en écarte, plus la probabilité de rejeter l'hypothèse nulle $H_m = H_{m+1}$ est forte.

3.4.3 Estimation des p-valeurs par bootstrap

Pour estimer les paramètres d'échelle $\tilde{\sigma}_m$ de $\tilde{\delta}_m$ sous $\mathcal{H}_0^{(m)}$, pour $m \in \{1, \ldots, M-1\}$, on propose de reproduire cette statistique par bootstrap. Une première approche est de reproduire les statistiques $\tilde{\delta}_m$ sous $\mathcal{H}_0$. Cependant, sous $\mathcal{H}_0^{(m)}$, le paramètre $\tilde{\sigma}_m$ dépend des autres statistiques $\tilde{\delta}_{m'}$, $m' \neq m$, et donc du fait que les autres hypothèses $\mathcal{H}_0^{(m')}$ soient vraies ou non. Ainsi, une estimation de $\tilde{\sigma}_m$ sous $\mathcal{H}_0$ n'approxime pas toujours correctement le paramètre $\tilde{\sigma}_m$ lorsque l'hypothèse nulle $\mathcal{H}_0^{(m)}$ est vraie mais l'hypothèse nulle $\mathcal{H}_0^{(m')}$ n'est pas vraie pour un certain $m' \neq m$. Une seconde approche consiste alors à estimer les paramètres $\tilde{\sigma}_m$ et $\tilde{\mu}_m$ de la statistique $\tilde{\delta}_m$ observée ($\mathcal{H}_0^{(m)}$ vraie ou non) pour $m \in \{1, \ldots, M-1\}$. Les p-valeurs peuvent être estimées à partir de l'a priori $\tilde{\mu}_m = 0$ sous $\mathcal{H}_0^{(m)}$ et de l'estimée de $\tilde{\sigma}_m$ sous l'hypothèse observée. Ces deux approches sont décrites ci-après.

3.4.3.1 Reproduction de l'hypothèse nulle

Cette approche est l'objet de l'article suivant : C.-G. LUCAS, P. ABRY, H. WENDT et G. DIDIER, « Counting the number of different scaling exponents in multivariate scale-free dynamics: Clustering by bootstrap in the wavelet domain », dans *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 5513–5517.

On cherche ici à obtenir une approximation des statistiques $\tilde{\delta}_1, \ldots, \tilde{\delta}_{M-1}$ sous l'hypothèse nulle $\mathcal{H}_0$, qu'on utilisera pour estimer les paramètres $\tilde{\sigma}_1, \ldots, \tilde{\sigma}_{M-1}$. Premièrement, pour reproduire l'hypothèse nulle, la valeur moyenne est soustraite de chacune des composantes du ré-échantillon bootstrap,

$$\forall m \in \{1, \ldots, M\},\ \forall r \in \{1, \ldots, R\}, \quad \bar{H}_m^{*(r)} := \hat{H}_m^{*(r)} - \langle \hat{H}_m^* \rangle, \tag{3.21}$$

où $\langle \hat{H}_m^* \rangle$ correspond à la moyenne des estimées bootstrap à travers les réalisations bootstrap,

$$\langle \hat{H}_m^* \rangle := \frac{1}{R} \sum_{r=1}^{R} \hat{H}_m^{*(r)}. \tag{3.22}$$

Les moyennes des estimées $\bar{H}^*_m$ sont ainsi identiques, tout comme les moyennes des estimées $\hat{H}_m$ sous $\mathcal{H}_0$ asymptotiquement. Ensuite, les composantes $\bar{H}^{*(r)}_m$ sont ordonnées pour chaque ré-échantillon bootstrap individuellement,

$$\underline{\bar{H}}^{*(r)}_{\tau^*} = (\bar{H}^{*(r)}_{\tau^*(r,1)}, \ldots, \bar{H}^{*(r)}_{\tau^*(r,M)}), \tag{3.23}$$

où $\tau^*(r,.)$ est une permutation qui ordonne les estimées bootstrap, $\bar{H}^{*(r)}_{\tau^*(r,1)} \leq \ldots \leq \bar{H}^{*(r)}_{\tau^*(r,M)}$, pour chaque échantillon bootstrap $r \in \{1, \ldots, R\}$. Ensuite, les statistiques bootstrap,

$$\forall m \in \{1, \ldots, M-1\},\ \forall r \in \{1, \ldots, R\}, \quad \bar{\delta}^{*(r)}_m := \bar{H}^{*(r)}_{\tau^*(r,m+1)} - \bar{H}^{*(r)}_{\tau^*(r,m)}, \tag{3.24}$$

sont calculées. Les simulations de Monte Carlo, effectuées sur des M-mBf, rapportées dans la section 3.4.7, montrent que les statistiques bootstrap $\bar{\delta}^{*(r)}_m$ sont bien approximées par des distributions demi-normales $\mathcal{FN}_{0,\bar{\sigma}^*_m}$ et reproduisent de manière satisfaisante les statistiques de test $\tilde{\delta}_m$ sous l'hypothèse nulle $\mathcal{H}_0$. Les estimées $\bar{\sigma}^*_m$ des paramètres d'échelle $\tilde{\sigma}_m$ sont obtenues à partir de la variance bootstrap Var^* des statistiques bootstrap, pour tout $m \in \{1, \ldots, M-1\}$, de la façon suivante :

$$\bar{\sigma}^{*2}_m := \frac{\mathrm{Var}^*(\bar{\delta}^*_m)}{1 - \frac{2}{\pi}}. \tag{3.25}$$

Cette relation découle de la relation entre le paramètre d'échelle et la variance d'une statistique suivant une loi demi-normale (cf. Annexe B.2).

Enfin, les estimées bootstrap des p-valeurs $\tilde{p}_m$ associées au rejet de $\mathcal{H}_0^{(m)}$ sont calculées, pour tout $m \in \{1, \ldots, M-1\}$, comme suit :

$$\bar{p}^*_m := 1 - F_{\mathcal{FN}(0,\bar{\sigma}^*_m)}(\tilde{\delta}_m). \tag{3.26}$$

Cet estimateur est biaisé puisque deux statistiques $\tilde{\delta}_m$ et $\tilde{\delta}_{m'}$, avec $m \neq m'$, ne sont pas indépendantes, et $\tilde{\delta}_m$ n'a donc pas la même distribution sous $\mathcal{H}_0^{(m')}$ que sous $\mathcal{H}_1^{(m')}$, ce qui est mis en lumière sur les simulations de Monte Carlo décrites dans la section 3.4.7.

3.4.3.2 Reproduction de l'hypothèse observée

Cette approche est l'objet de l'article suivant : C.-G. Lucas, H. Wendt, P. Abry et G. Didier, « Multivariate time-scale bootstrap for testing the equality of selfsimilarity parameters », dans *XXVIIIème Colloque Francophone de Traitement du Signal et des Images (GRETSI 2022)*.

Pour approximer les statistiques $\tilde{\delta}_m$ sous une hypothèse observée ($\mathcal{H}_0^{(m)}$ ou $\mathcal{H}_1^{(m)}$), pour tout $m \in \{1, \ldots, M-1\}$, les estimées bootstrap $\hat{H}^{*(r)}_m$ sont ordonnées sans en soustraire la moyenne,

$$\underline{\hat{H}}^{*(r)}_{\tau^*} = (\hat{H}^{*(r)}_{\tau^*(r,1)}, \ldots, \hat{H}^{*(r)}_{\tau^*(r,M)}), \tag{3.27}$$

où $\tau^*(r,.)$ est une permutation qui ordonne les estimées bootstrap, $\hat{H}^{*(r)}_{\tau^*(r,1)} \leq \ldots \leq \hat{H}^{*(r)}_{\tau^*(r,M)}$, pour chaque échantillon bootstrap $r \in \{1, \ldots, R\}$. Les simulations de Monte Carlo effectuées sur des M-mBf, rapportées dans la section 3.4.7, montrent que les statistiques de test bootstrap,

$$\forall m \in \{1, \ldots, M-1\},\ \forall r \in \{1, \ldots, R\}, \quad \tilde{\delta}^{*(r)}_m := \hat{H}^{*(r)}_{\tau^*(r,m+1)} - \hat{H}^{*(r)}_{\tau^*(r,m)}, \tag{3.28}$$

sont bien approximées par des distributions normales repliées $\mathcal{FN}_{\tilde{\mu}_m^*,\tilde{\sigma}_m^*}$ et reproduisent de manière satisfaisante les statistiques de test $\tilde{\delta}_m$. Les paramètres $\tilde{\mu}_m^*$ et $\tilde{\sigma}_m^*$ sont estimés à partir des $\tilde{\delta}_m^{*(r)}$ en utilisant un estimateur classique de maximum de vraisemblance, comme décrit dans Tsagris et collab. (2014). Il consiste à résoudre, à partir d'échantillons indépendants $\{x_1, \ldots, x_n\}$, le système d'équations d'inconnu le couple $(\tilde{\mu}, \tilde{\sigma})$, avec $\tilde{\mu} > 0$ et $\tilde{\sigma} > 0$, suivant :

$$\sum_{i=1}^{n} \frac{1 - e^{\frac{2\tilde{\mu}x_i}{\tilde{\sigma}^2}}}{1 + e^{\frac{2\tilde{\mu}x_i}{\tilde{\sigma}^2}}} x_i + n\frac{\tilde{\mu}}{2} = 0, \quad \tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \tilde{\mu}^2. \tag{3.29}$$

Le couple $(\tilde{\mu}_m^*, \tilde{\sigma}_m^*)$ désigne désormais la solution de l'équation (3.29) obtenue à partir de l'échantillon bootstrap $\tilde{\delta}_m^{*(1)}, \ldots, \tilde{\delta}_m^{*(R)}$.

Sous $\mathcal{H}_0^{(m)}$, le paramètre bootstrap $\tilde{\sigma}_m^*$ devrait être un estimateur non biaisé de $\tilde{\sigma}_m$, que les autres hypothèses nulles $\mathcal{H}_0^{(m')}$ soient vraies ou non, pour tout $m' \in \{1, \ldots, M-1\}$. Les p-valeurs $\tilde{p}_m$ associées au rejet de $\mathcal{H}_0^{(m)}$ peuvent alors être estimées, pour tout $m \in \{1, \ldots, M-1\}$, par

$$\tilde{p}_m^* := 1 - F_{\mathcal{FN}(0,\tilde{\sigma}_m^*)}(\tilde{\delta}_m). \tag{3.30}$$

En outre, et c'est un avantage de cette approche, les puissances des tests pour $\mathcal{H}_0^{(m)}$ avec un niveau de confiance α_m pour tout $m \in \{1, \ldots, M-1\}$,

$$\mathbb{P}(\tilde{p}_m^* < \alpha_m \,|\, \mathcal{H}_1^{(m)}) = \mathbb{P}\left(\tilde{\delta}_m > F^{-1}_{\mathcal{FN}(0,\tilde{\sigma}_m^*)}(1-\alpha_m) \,|\, \mathcal{H}_1^{(m)}\right),$$

peuvent également être estimées à partir d'une seule observation de taille finie par

$$\tilde{\pi}(\tilde{\mu}_m^*, \tilde{\sigma}_m^*) := 1 - \mathrm{F}_{\mathcal{FN}(\tilde{\mu}_m^*,\tilde{\sigma}_m^*)}\left(F^{-1}_{\mathcal{FN}(0,\tilde{\sigma}_m^*)}(1-\alpha_m)\right). \tag{3.31}$$

3.4.4 Décisions des tests

On s'intéresse à présent à la sélection des seuils de confiance α_m pour effectuer le test à $M-1$ hypothèses donné par l'équation (3.19). En notant les décisions $d_\alpha^{(m)}$ des tests (3.19), on définit la décision pour $\mathcal{H}_0$ par

$$\begin{cases} d_\alpha = 1 \quad (\mathcal{H}_0 \text{ rejetée}) & \text{si } \exists m \in \{1, \ldots, M-1\} |\ d_\alpha^{(m)} = 1, \\ d_\alpha = 0 \quad (\mathcal{H}_0 \text{ non rejetée}) & \text{si } \forall m \in \{1, \ldots, M-1\},\ d_\alpha^{(m)} = 0. \end{cases} \tag{3.32}$$

Ainsi, l'hypothèse nulle $\mathcal{H}_0$ est rejetée si au moins l'une des hypothèses nulles $\mathcal{H}_0^{(m)}$, pour $m \in \{1, \ldots, M-1\}$, est rejetée. Pour fixer des seuils de confiance α_m de sorte que la probabilité de rejeter $\mathcal{H}_0$ à tort soit inférieure à un niveau de confiance α, i.e. $\mathbb{P}\left(d_\alpha = 1 \,\middle|\, \mathcal{H}_0\right) \leq \alpha$, trois procédures classiques de correction d'un test à hypothèses multiples sont considérées.

La correction de Bonferroni (Bonferroni, 1936) contrôle le taux d'erreur de la famille de tests (probabilité qu'au moins une hypothèse nulle $\mathcal{H}_0^{(m)}$ soit rejetée lorsque $\mathcal{H}_0$ est vraie), le forçant à être inférieur ou égal à un niveau de confiance prédéfini α, en adaptant le niveau de confiance de chaque test par paire comme suit :

$$\begin{cases} d_\alpha^{(m)} = 1 \quad (\mathcal{H}_0^{(m)} \text{ rejetée}) & \text{si } p_m < \frac{\alpha}{M-1}, \\ d_\alpha^{(m)} = 0 \quad (\mathcal{H}_0^{(m)} \text{ non rejetée}) & \text{sinon,} \end{cases} \tag{3.33}$$

pour tout $m \in \{1, \dots, M-1\}$, où p_m est une p-valeur du test pour $H_{m+1} = H_m$, en l'occurrence $p_m = \bar{p}_m^*$ ou $p_m = \tilde{p}_m^*$. Si les p-valeurs sont indépendantes, le procédure de Bonferroni reconstruit le niveau de confiance α, i.e. $\mathbb{P}\left(d_\alpha = 1 \,\middle|\, \mathcal{H}_0\right) = \alpha$, mais peut être conservative sinon, i.e. $\mathbb{P}\left(d_\alpha = 1 \,\middle|\, \mathcal{H}_0\right) < \alpha$. De plus, cette correction augmente la probabilité de ne pas rejeter l'hypothèse nulle $\mathcal{H}_0$ lorsqu'elle n'est pas vraie, ce qui réduit la puissance du test.

Les corrections du taux de fausses découvertes (*false discovery rate*) contrôlent la probabilité qu'il y ait au moins une fausse découverte (hypothèse nulle $\mathcal{H}_0^{(m)}$ rejetée à tort) parmi les décisions de rejet de $\mathcal{H}_0$, ce qui permet un test plus puissant. Ces méthodes de correction adaptent le seuil des p-valeurs comme suit :

$$\begin{cases} d_\alpha^{\pi(k)} = 1 \quad (\mathcal{H}_0^{\pi(k)} \text{ rejetée}) & \text{si } k \leq \underset{j \in \{1,\dots,M-1\}}{\arg\max} \left(j \mathbb{1}_{p_{\pi(j)} < \frac{\alpha}{M-1} c(j)}\right), \\ d_\alpha^{\pi(k)} = 0 \quad (\mathcal{H}_0^{\pi(k)} \text{ non rejetée}) & \text{sinon,} \end{cases} \tag{3.34}$$

où π désigne la permutation qui ordonne les p-valeurs, $p_{\pi(1)} < \ldots < p_{\pi(M-1)}$, et la fonction c est définie soit suivant la correction de Benjamini-Hochberg (Benjamini et Hochberg, 1995), qui est adaptée à une corrélation positive ou nulle, soit suivant la correction de Benjamini-Yekutieli (Yekutieli et Benjamini, 1999),

$$c(j) = \begin{cases} j & \text{suivant Benjamini-Hochberg,} \\ j \Big/ \displaystyle\sum_{j=1}^{J} \frac{1}{j} & \text{suivant Benjamini-Yekutieli,} \end{cases} \tag{3.35}$$

pour tout $j \in \{1, \dots, J\}$ avec J le nombre d'hypothèses, en l'occurrence $J = M - 1$. Pour des statistiques de test continues et indépendantes, la procédure de Benjamini-Hochberg reconstruit le niveau de confiance α.

Dans la suite, on notera $\bar{d}_\alpha^{(m)}$ et $\tilde{d}_\alpha^{(m)}$ les décisions corrigées issues respectivement des p-valeurs $\bar{p}_m^*$ et $\tilde{p}_m^*$ définies par les équations (3.26) et (3.30) pour tout $m \in \{1, \dots, M-1\}$. Enfin, on note $\bar{d}_\alpha$ et $\tilde{d}_\alpha$ les décisions pour $\mathcal{H}_0$ correspondantes, respectivement.

3.4.5 Stratégie de partitionnement

Le vecteur $\left(d_\alpha^{(1)}, \dots, d_\alpha^{(M-1)}\right)$ des $M-1$ décisions pour $H_m = H_{m+1}$, avec $m \in \{1, \dots, M-1\}$, est un vecteur indicateur des limites des partitions dans $\underline{H}$: s'il y a P hypothèses nulles $\mathcal{H}_0^{(m)}$ rejetées, il y a $P+1$ partitions distinctes d'exposants d'autosimilarité. Les étiquettes des partitions détectées pour H_m sont définies comme suit :

$$C_\alpha(m) := \sum_{k=1}^{m} D_\alpha(k) \tag{3.36}$$

où $D_\alpha = (1, d_\alpha^{(1)}, \dots, d_\alpha^{(M-1)})$ avec $d_\alpha^{(m)} = \bar{d}_\alpha^{(m)}$ ou $d_\alpha^{(m)} = \tilde{d}_\alpha^{(m)}$, pour tout $m \in \{1, \dots, M-1\}$. Le partitionnement à partir des décisions est illustré par la figure 3.15.

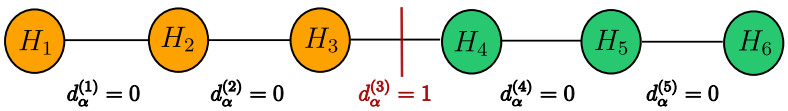


FIGURE 3.15 – **Illustration de la stratégie de partitionnement.** Les décisions de rejet $d_\alpha^{(m)} = 1$ (en rouge) affectent les exposantsH_m et H_{m+1} dans des partitions différentes (en orange et verte).

3.4.6 Synthèse des notations et formules

L'ensemble de la stratégie de partitionnement est résumée par la figure 3.16.

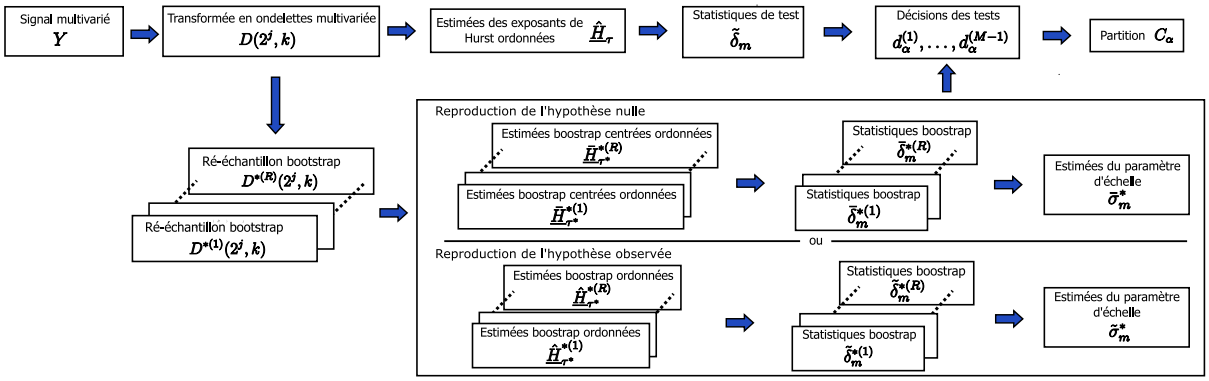


FIGURE 3.16 – **Organigramme de la procédure de partitionnement à partir de $M-1$ tests d'égalité par paires d'exposants d'autosimilarité successifs.**

Formulaire récapitulatif

Les différentes quantités relatives aux tests des hypothèses $\mathcal{H}_0^{(m)}$ $(H_m = H_{m+1})$ pour $m \in \{1, \ldots, M-1\}$ sont rappelées dans le tableau suivant.

Définition	Estimateur bootstrap par reproduction de	
	l'hypothèse nulle $\mathcal{H}_0$	l'hypothèse observée
Statistiques des tests par paires		
$\tilde{\delta}_m$	$\bar{\delta}_m^{*(1)}, \ldots, \bar{\delta}_m^{*(R)}$	$\tilde{\delta}_m^{*(1)}, \ldots, \tilde{\delta}_m^{*(R)}$
Paramètres de position des statistiques		
$\tilde{\mu}_m$	Aucun	$\tilde{\mu}_m^*$ (résoudre Eq. (3.29))
Paramètres d'échelle des statistiques		
$\tilde{\sigma}_m$	$\bar{\sigma}_m^* = \sqrt{\frac{\mathrm{Var}^*(\bar{\delta}_m^*)}{1-\frac{2}{\pi}}}$	$\tilde{\sigma}_m^*$ (résoudre Eq. (3.29))
p-valeurs des tests par paires		
$\tilde{p}_m$	$\bar{p}_m^* = 1 - F_{\mathcal{FN}(0,\bar{\sigma}_m^*)}(\tilde{\delta}_m)$	$\tilde{p}_m^* = 1 - F_{\mathcal{FN}(0,\tilde{\sigma}_m^*)}(\tilde{\delta}_m)$
Puissance des tests par paires		
$\tilde{\pi}(\tilde{\mu}_m, \tilde{\sigma}_m)$	Aucun	$\tilde{\pi}(\tilde{\mu}_m^*, \tilde{\sigma}_m^*)$ (Eq. (3.31))
Décisions de rejet des tests par paires		
Inconnue	$\bar{d}_\alpha^{(m)}$	$\tilde{d}_\alpha^{(m)}$
Décision de rejet du test pour $\mathcal{H}_0$		
Inconnue	$\bar{d}_\alpha$	$\tilde{d}_\alpha$

3.4.7 Performances des estimateurs, des tests et du partitionnement

3.4.7.1 Simulations de Monte Carlo

Afin d'évaluer la pertinence des procédures de tests par paires et de regroupement proposées et de quantifier leurs performances, des simulations de Monte Carlo sont effectuées, en utilisant $N_{\mathrm{MC}} = 1000$ copies indépendantes de $M = 6$-mBf synthétiques (cf. Section 2.4) avec différentes tailles d'échantillon $N \in \{2^{16}, 2^{17}, 2^{18}\}$. Quatre scénarios sont considérés pour le vecteur des exposants d'autosimilarité $\underline{H}$:

(*i*) le Scenario1 correspond à $H_1 = \ldots = H_M = 0.8$ (1 partition) ;

(*ii*) le Scenario2 consiste en 2 groupes de taille 3 avec des valeurs égales à 0.6 et 0.8, tels que $\mathcal{H}_0^{(3)}$ n'est pas vraie ;

(*iii*) le Scenario3 consiste en 3 partitions de taille 2 avec des valeurs égales à 0.4, 0.6 et 0.8, telles que $\mathcal{H}_0^{(2)}$ et $\mathcal{H}_0^{(4)}$ ne sont pas vraies ;

(*iv*) le Scenario4 consiste en 3 groupes de tailles différentes 1, 3 et 2 avec des valeurs égales à 0.4, 0.6 et 0.8, tels que $\mathcal{H}_0^{(1)}$ et $\mathcal{H}_0^{(4)}$ ne sont pas vraies.

La matrice de covariance Σ des M-mBf est choisie de telle sorte que toutes ses entrées diagonales soient fixées à 1 et que toutes ses entrées non diagonales soient fixées à $r = 0.5$. La matrice de mélange inversible W des M-mBf est choisie au hasard et maintenue fixe pour toutes les expériences.

L'analyse en ondelettes est effectuée avec l'ondelette de Daubechies la moins asymétrique à $N_\psi = 3$ moments nuls (Daubechies, 1992) sur les échelles $2^{j_1} = 2^8$ à $2^{j_2} = 2^{11}$. Pour la procédure de tests par paires, $R = 500$ ré-échantillons bootstrap sont tirés de blocs de coefficients d'ondelettes se chevauchant de taille $L_B = 6$ (correspondant à la taille du support temporel de l'ondelette mère de Daubechies 3).

3.4.7.2 Propriétés des statistiques de test

Les tests présentés ci-dessus sont construits à partir d'approximation sur le comportement des $M-1$ statistiques de test : chaque statistique $\tilde{\delta}_m$ est supposée suivre une loi normale repliée de paramètre de position $\tilde{\mu}_m = 0$ sous hypothèse nulle $\mathcal{H}_0^{(m)}$, pour $m \in \{1, \ldots, M-1\}$.

Approximation par une loi normale repliée

Pour vérifier le comportement adéquat des statistique $\tilde{\delta}_m$, la figure 3.17 présente, pour les différents scénarios et pour les cinq paires d'exposants d'autosimilarité (H_m, H_{m+1}), $m = 1, \ldots, M-1$, les diagrammes quantile-quantile des distributions empiriques de $\tilde{\delta}_m$ par rapport aux échantillons tirés selon une loi normale repliée $\mathcal{FN}(\tilde{\mu}_m, \tilde{\sigma}_m)$, avec les paramètres $\tilde{\mu}_m$ et $\tilde{\sigma}_m$ estimés au travers des réalisations de Monte Carlo à l'aide de l'équation (3.29), pour les différents scénarios et une taille d'échantillon $N = 2^{16}$. La figure 3.17 montre tout d'abord que les distributions normales repliées sont des approximations pertinentes des distributions des $\tilde{\delta}_m$ à la fois sous des hypothèses nulles $\mathcal{H}_0^{(m)}$ et sous des hypothèses alternatives $\mathcal{H}_1^{(m)}$.

Cependant, l'approximation par une loi normale repliée est moins adaptée sous $\mathcal{H}_0$ que sous $\mathcal{H}_1$. En effet, une statistique $\tilde{\delta}_m$ suit asymptotiquement une loi normale repliée si les estimées ordonnées $\hat{H}_{\tau(m)}$ et $\hat{H}_{\tau(m+1)}$ ne sont issues que de deux estimées gaussiennes $\hat{H}_{m+1}$ et $\hat{H}_m$. Cette approximation est donc altérée lorsque les distributions de ces estimées et les distributions d'autres estimées se chevauchent. Or, comme l'illustre la figure 3.18, qui montre les histogrammes des $\hat{H}_m$ pour $m = 1, \ldots, M$, les différents scénarios et une taille d'échantillon $N = 2^{16}$, les

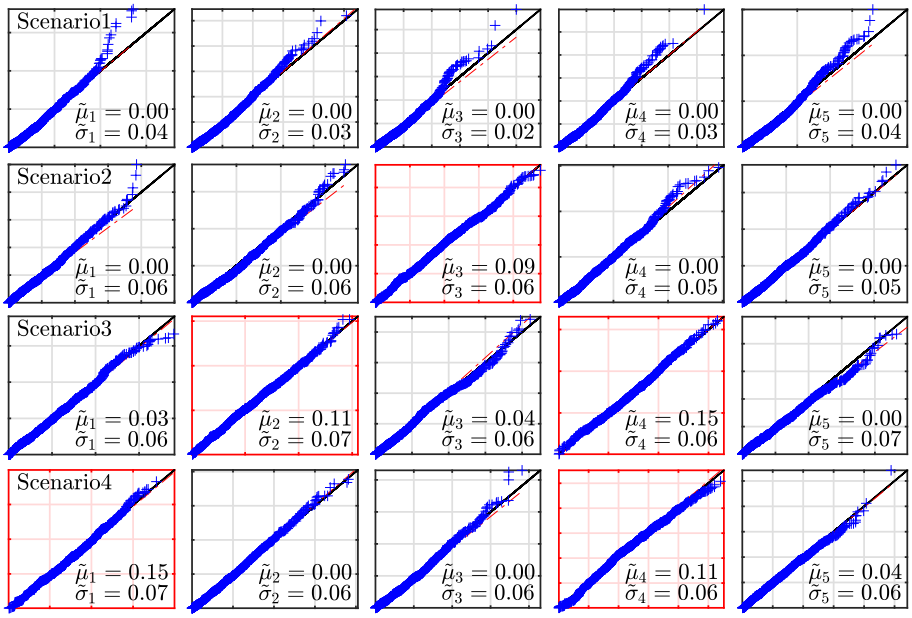


FIGURE 3.17 – **Distribution des statistiques $\tilde{\delta}_m$.** Diagrammes quantile-quantile de $\tilde{\delta}_m$ par rapport aux échantillons tirés selon une loi normale repliée $\mathcal{FN}_{\tilde{\mu}_m,\tilde{\sigma}_m}$, avec des paramètres $\tilde{\mu}_m$ et $\tilde{\sigma}_m$ estimés à l'aide de l'équation (3.29) au travers des réalisations de Monte Carlo, pour $m = 1, \ldots, 5$ (de gauche à droite), les scénarios 1, 2, 3 et 4 (de haut en bas) et une taille d'échantillon $N = 2^{16}$. Les cases rouges correspondent aux écarts par rapport à $\mathcal{H}_0^{(m)}$. *Les distributions des statistiques $\tilde{\delta}_m$ sont bien approximées par des distributions normales repliées.*

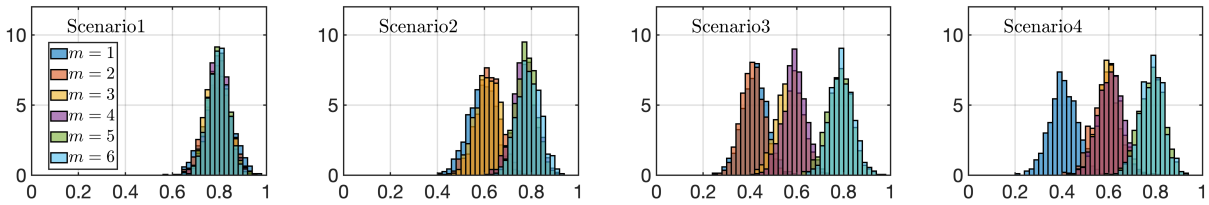


FIGURE 3.18 – **Distributions des $\hat{H}_m$.** Histogrammes des $\hat{H}_m$ au travers des réalisations de Monte Carlo, pour $m = 1, \ldots, 6$ (de différentes couleurs), les scénarios 1, 2, 3 et 4 (de gauche à droite) et une taille d'échantillon $N = 2^{16}$.

distributions gaussiennes des estimées $\hat{H}_m$ sont toutes proches sous $\mathcal{H}_0$ (Scenario1) alors que, sous $\mathcal{H}_1$, moins de distributions $\hat{H}_m$ sont superposées et elles se chevauchent peu.

Ces simulations de Monte Carlo montrent également que, dans le cadre du scénario 1, bien que les hypothèses nulles $\mathcal{H}_0^{(m)}$ soient toutes vraies, on observe que le paramètre d'échelle $\tilde{\sigma}_m$ dépend de m : $\tilde{\sigma}_m$ est plus grand pour $m = 1$ et $m = 5$, c'est-à-dire pour les premières et dernières estimées $\hat{H}_m$, et plus petit pour les estimées centrales, ce qui peut être interprété comme une conséquence de l'opération de tri appliquée aux éléments de $\underline{\hat{H}}$. Ces simulations montrent également que, sous $\mathcal{H}_0^{(m)}$, les valeurs de $\tilde{\sigma}_m$ pour un même m diffèrent entre les scénarios. Ces observations constituent les premiers résultats principaux et intéressants : la distribution de $\tilde{\delta}_m$ sous $\mathcal{H}_0^{(m)}$ dépend à la fois du rang des paires triées (c'est-à-dire de m) et du scénario, principalement à travers $\tilde{\sigma}_m$. Ainsi, un estimateur de $\tilde{\sigma}_m$ sous $\mathcal{H}_0$ est un estimateur biaisé de $\tilde{\sigma}_m$ sous $\mathcal{H}_0^{(m)}$ lorsque $\mathcal{H}_0$ n'est pas vraie. Cela appuie l'intérêt d'explorer les deux approches d'estimation bootstrap des $\tilde{\sigma}_m$ présentées dans la section 3.4.3.

Approximation sous $\mathcal{H}_0^{(m)}$

On peut tout d'abord observer sur la figure 3.17 que sous $\mathcal{H}_0^{(m)}$, on a $\tilde{\mu}_m \simeq 0$, ce qui appuie l'approximation de la distribution des $\tilde{\delta}_m$ par une distribution demi-normale. De plus, sous $\mathcal{H}_1^{(m)}$, les paramètres de position $\tilde{\mu}_m$ sont éloignés de 0, ce qui signifie que la loi de $\tilde{\delta}_m$ sous $\mathcal{H}_1^{(m)}$ s'éloigne de la loi de $\tilde{\delta}_m$ sous $\mathcal{H}_0^{(m)}$.

Pour vérifier le comportement adéquat des statistiques $\tilde{\delta}_m$ sous $\mathcal{H}_0^{(m)}$ pour $m = 1, \ldots, M-1$, la figure 3.19 présente les diagrammes quantile-quantile des distributions des $\tilde{\delta}_m$ par rapport aux échantillons tirés selon des lois demi-normales $\mathcal{FN}_{0,\bar{\sigma}_m}$ de paramètres d'échelle $\bar{\sigma}_m = \sqrt{\mathrm{Var}(\tilde{\delta}_m)/(1-2/\pi)}$ estimés par Monte Carlo pour les différents scénarios, les cinq paires d'exposants d'autosimilarité (H_m, H_{m+1}) et une taille d'échantillon $N = 2^{16}$. Ces résultats confirment la distribution approximativement demi-normale de $\tilde{\delta}_m$ sous $\mathcal{H}_0^{(m)}$ et montrent également l'écart de $\tilde{\delta}_m$ à une distribution demi-normale sous $\mathcal{H}_1^{(m)}$, confirmant la pertinence d'un test demi-normal à partir de cette statistique.

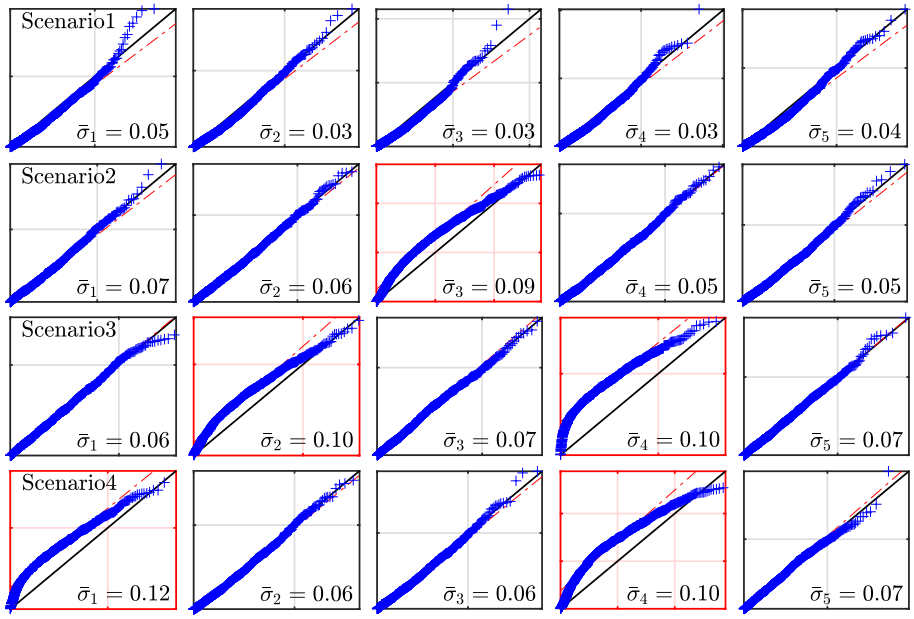


FIGURE 3.19 – **Distribution des statistiques $\tilde{\delta}_m$ sous $\mathcal{H}_0^{(m)}$.** Diagrammes quantile-quantile de $\tilde{\delta}_m$ par rapport aux échantillons tirés selon une loi demi-normale $\mathcal{FN}_{0,\bar{\sigma}_m}$ de paramètre d'échelle $\bar{\sigma}_m = \sqrt{\mathrm{Var}(\tilde{\delta}_m)/(1-2/\pi)}$ estimé par Monte Carlo, pour $m = 1, \ldots, 5$ (de gauche à droite), les scénarios 1, 2, 3 et 4 (de haut en bas) et une taille d'échantillon $N = 2^{16}$. Les cases rouges correspondent aux écarts par rapport à $\mathcal{H}_0^{(m)}$. *Les distributions des statistiques $\tilde{\delta}_m$ sont bien approximées par des distributions demi-normales sous hypothèse nulle $\mathcal{H}_0^{(m)}$ et s'en écartent sous des hypothèses alternatives $\mathcal{H}_1^{(m)}$.*

3.4.7.3 Propriétés des statistiques bootstrap

Dans cette section, sont étudiées les deux statistiques bootstrap $\bar{\delta}_m^*$ (cf. Eq. (3.24)) et $\hat{\delta}_m^*$ (cf. Eq. (3.28)), conçues pour reproduire le comportement des statistique de test $\tilde{\delta}_m$ sous l'hypothèse nulle $\mathcal{H}_0$ et sous hypothèse observée, respectivement, pour $m \in \{1, \ldots, M-1\}$.

Propriétés des statistiques bootstrap $\bar{\delta}^*_m$

Sous $\mathcal{H}_0^{(m)}$, les statistiques $\tilde{\delta}_m$ suivent approximativement des lois demi-normales $\mathcal{FN}_{0,\tilde{\sigma}_m}$, comportement que doivent donc reproduire les statistiques bootstrap $\bar{\delta}^*_m$. La figure 3.20 présente les diagrammes quantile-quantile de la distribution de $\bar{\delta}^*_m$ pour une réalisation de Monte Carlo par rapport à une distribution demi-normale de paramètre $\bar{\sigma}^*_m$ estimé selon l'équation (3.25) pour $m = 1, \ldots, M-1$, les différents scénarios et une taille d'échantillon $N = 2^{16}$. Les résultats indiquent clairement que dans tous les cas ($\mathcal{H}_0^{(m)}$ vraie ou non), les distributions des $\bar{\delta}^*_m$ sont en bon accord avec la distribution demi-normale. On constate cependant un certain écart de la distribution de $\bar{\delta}^*_m$ à une distribution demi-normale sous $\mathcal{H}_0$, plus accentué que celui observé pour la distribution de la statistique $\tilde{\delta}_m$ sur la figure 3.17 suggérant que les distributions des ré-échantillons bootstrap $\hat{H}^*_m$ se chevauchent plus que les estimées $\hat{H}_m$, altérant ainsi l'approximation demi-normale, comme expliqué plus tôt et illustré par la figure 3.18. Ces premiers résultats assurent la construction appropriée de l'estimateur bootstrap $\bar{\sigma}^*_m$.

Pour quantifier la qualité de l'estimateur bootstrap $\bar{\sigma}^*_m$, le tableau 3.1 présente les estimées de Monte Carlo des paramètres d'échelle $\bar{\sigma}_m = \sqrt{\mathrm{Var}(\tilde{\delta}_m)/(1-2/\pi)}$, tels que $\tilde{\sigma}_m = \bar{\sigma}_m$ sous l'hypothèse $\tilde{\mu}_m = 0$, ainsi que les moyennes de Monte Carlo avec intervalle de confiance à 95% de leurs estimées bootstrap $\bar{\sigma}^*_m$ pour $m = 1, \ldots, M-1$, tous les scénarios, et une taille d'échantillon $N = 2^{16}$. Les résultats confirment que les estimées bootstrap $\bar{\sigma}^*_m$ sont en excellent accord avec $\bar{\sigma}_m$ sous $\mathcal{H}_0$, mais sont de mauvaises approximations de $\bar{\sigma}_m$ sous $\mathcal{H}_1$. Cette observation confirme que $\bar{\delta}^*_m$ ne reproduit le comportement de $\tilde{\delta}_m$ que sous $\mathcal{H}_0$, et non lorsque $\mathcal{H}_1$ et $\mathcal{H}_0^{(m)}$ sont toutes deux vraies. Cependant, on peut remarquer que, sous $\mathcal{H}_1$, les paramètres boostrap $\bar{\sigma}^*_m$ s'écartent davantage de $\bar{\sigma}_m$ sous $\mathcal{H}_1^{(m)}$ que sous $\mathcal{H}_0^{(m)}$, présageant de bonnes puissances de test.

Propriétés des statistiques bootstrap $\tilde{\delta}^*_m$

Les statistiques $\tilde{\delta}_m$ suivent approximativement des lois normales repliées $\mathcal{FN}_{\tilde{\mu}_m,\tilde{\sigma}_m}$, comportement que doivent donc reproduire les statistiques bootstrap $\tilde{\delta}^*_m$. La figure 3.21 présente, pour les différents scénarios, les cinq paires $m = 1, \ldots, M-1$ et une taille d'échantillon $N = 2^{16}$, les diagrammes quantile-quantile des distributions de $\tilde{\delta}^*_m$ pour une réalisation de Monte Carlo par rapport aux échantillons tirés selon une loi normale repliée $\mathcal{FN}_{\tilde{\mu}^*_m,\tilde{\sigma}^*_m}$, avec $\tilde{\mu}^*_m$ et $\tilde{\sigma}^*_m$ estimés à l'aide de l'équation (3.29). La figure 3.21 montre que les $\tilde{\delta}^*_m$ sont bien approximées par des distributions normales repliées sous n'importe quelle hypothèse ($\mathcal{H}_0^{(m)}$ vraie ou non) et pour tous les scénarios. Ces observations assurent la bonne construction de l'estimateur $\tilde{\sigma}^*_m$.

Ensuite, le tableau 3.2 rapporte les valeurs des paramètres $\tilde{\mu}_m$ et $\tilde{\sigma}_m$ estimé par Monte Carlo et leurs estimées bootstrap $\tilde{\mu}^*_m$ et $\tilde{\sigma}^*_m$ moyennées sur les simulations de Monte Carlo (avec intervalle de confiance à 95%), pour les différents scénarios, les cinq paires $m = 1, \ldots, M-1$ et une taille d'échantillon $N = 2^{16}$. Les estimées bootstrap $\tilde{\sigma}^*_m$ approximent de façon satisfaisante les paramètres de test $\tilde{\sigma}_m$ sous n'importe quelle hypothèse ($\mathcal{H}_0^{(m)}$ vraie ou non), comme attenduEn revanche, les estimées bootstrap $\tilde{\mu}^*_m$ approximent bien les paramètres de position $\tilde{\mu}_m$ sous $\mathcal{H}_1^{(m)}$ mais pas sous $\mathcal{H}_0^{(m)}$, où l'on observe une surestimation.

Par ailleurs, le tableau 3.2 permet de vérifier que $\tilde{\mu}_m$ s'écarte de 0 sous $\mathcal{H}_1^{(m)}$ tandis que $\tilde{\sigma}_m$ varie peu entre les différents scénarios et les différentes hypothèses. Ceci laisse présager de bonnes puissances de test dépendant principalement $\tilde{\mu}_m$.

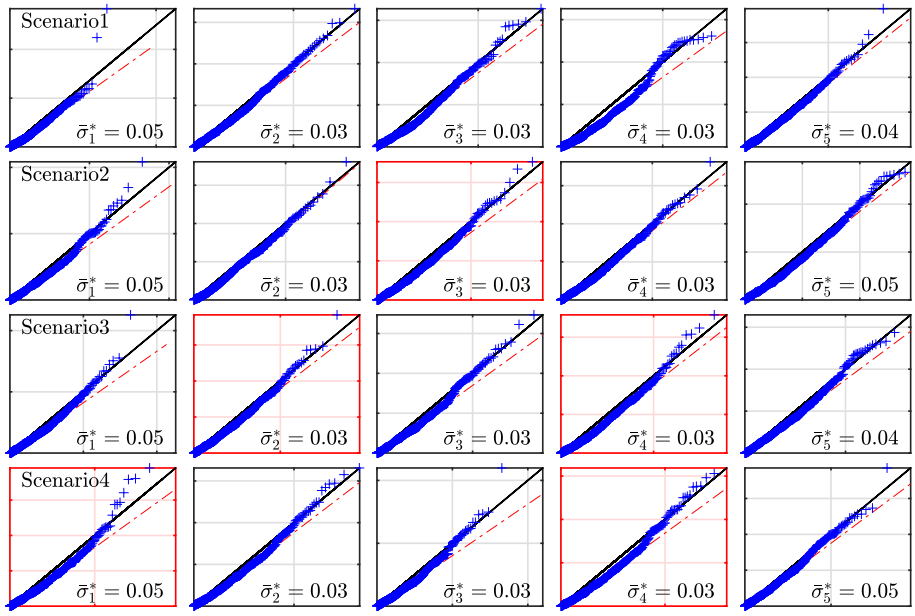


FIGURE 3.20 – **Distributions des statistiques bootstrap $\bar{\delta}_m^*$.** Diagrammes quantile-quantile des statistiques bootstrap $\bar{\delta}_m^*$ pour une réalisation de Monte Carlo choisie arbitrairement par rapport à une distribution demi-normale de paramètre $\bar{\sigma}_m^*$ estimé par l'équation (3.25) pour $m = 1, \ldots, 5$ (de gauche à droite), les différents scénarios et une taille d'échantillon $N = 2^{16}$. Les cases rouges correspondent aux écarts par rapport à $\mathcal{H}_0^{(m)}$. *Les distributions des statistiques bootstrap $\bar{\delta}_m^*$ sont bien approximées par des distributions demi-normales.*

TABLEAU 3.1 – **Estimation des paramètres des tests demi-normaux.** Estimées bootstrap $\bar{\sigma}_m^*$ (moyenne de Monte Carlo ± intervalle de confiance à 95%), calculées selon l'équation (3.25), des paramètres $\bar{\sigma}_m = \sqrt{\mathrm{Var}(\tilde{\delta}_m)/(1-2/\pi)}$ (i.e. les paramètres $\tilde{\sigma}_m$ des statistiques de Monte Carlo $\tilde{\delta}_m$ sous l'hypothèse $\tilde{\mu}_m = 0$) pour $m = 1, \ldots, 5$ et une taille d'échantillon $N = 2^{16}$. Les cases rouges correspondent aux écarts par rapport à $\mathcal{H}_0^{(m)}$.

$\times 10^2$		$m=1$	$m=2$	$m=3$	$m=4$	$m=5$		$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
$\bar{\sigma}_m$	Scenario1	4.62	3.01	2.72	2.97	4.29	Scenario2	6.57	5.88	8.56	4.85	5.24
$\bar{\sigma}_m^*$		4.74 ±0.59	3.08 ±0.25	2.73 ±0.19	2.93 ±0.20	4.12 ±0.31		4.89 ±0.55	3.24 ±0.25	2.87 ±0.22	3.07 ±0.21	4.26 ±0.30
$\bar{\sigma}_m$	Scenario3	6.24	10.24	7.16	9.80	7.07	Scenario4	11.74	6.16	5.93	9.55	6.86
$\bar{\sigma}_m^*$		4.87 ±0.56	3.30 ±0.27	2.95 ±0.22	3.13 ±0.23	4.24 ±0.30		4.99 ±0.58	3.35 ±0.26	3.00 ±0.22	3.19 ±0.22	4.34 ±0.31

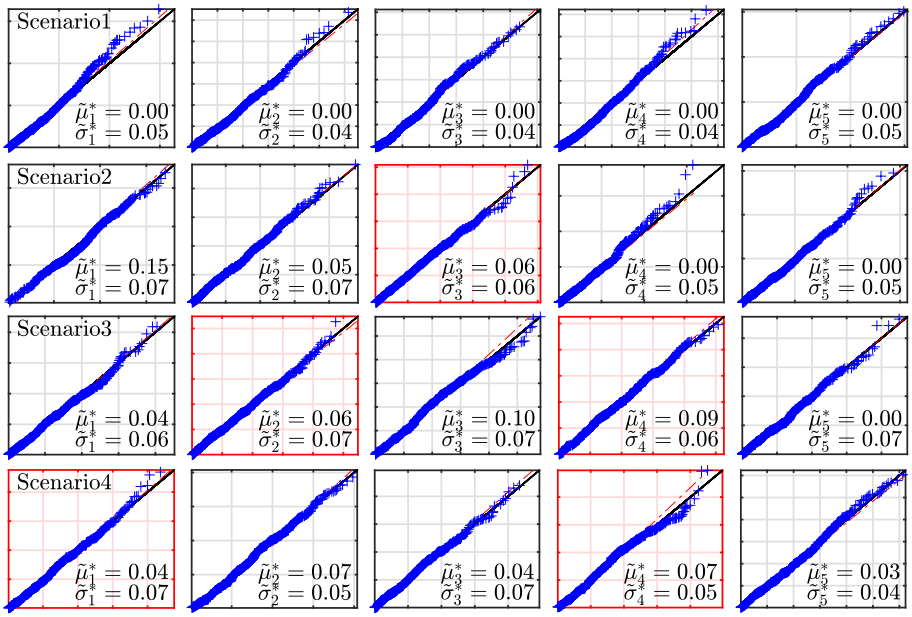


FIGURE 3.21 – **Distributions des statistiques bootstrap $\tilde{\delta}_m^*$.** Diagrammes quantile-quantile de $\tilde{\delta}_m^*$ pour une réalisation de Monte Carlo par rapport aux échantillons tirés pour une loi normale repliée $\mathcal{FN}_{\tilde{\mu}_m^*,\tilde{\sigma}_m^*}$ de paramètres $\tilde{\mu}_m^*$ et $\tilde{\sigma}_m^*$ estimés à l'aide de l'équation (3.29), pour $m = 1, \ldots, 5$ (de gauche à droite), les scénarios 1, 2, 3 et 4 (de haut en bas) et une taille d'échantillon $N = 2^{16}$. Les cases rouges correspondent aux écarts par rapport à $\mathcal{H}_0^{(m)}$. *Les distributions des statistiques bootstrap $\tilde{\delta}_m^*$ sont bien approximées par des distributions normales repliées.*

TABLEAU 3.2 – **Estimation des paramètres des tests normaux repliés.** Estimées $\tilde{\mu}_m^*$ et $\tilde{\sigma}_m^*$ (moyennes de Monte Carlo ± intervalles de confiance à 95%) des paramètres $\tilde{\mu}_m$ et $\tilde{\sigma}_m$ de loi normale repliée etimées par Monte Carlo suivant l'équation (3.29) pour $m = 1, \ldots, 5$, une taille d'échantillon $N = 2^{16}$. Les cases rouges correspondent aux écarts par rapport à $\mathcal{H}_0^{(m)}$.

$\times 10^2$		$m=1$	$m=2$	$m=3$	$m=4$	$m=5$		$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
$\tilde{\mu}_m$	Scenario1	0.00	0.00	0.00	0.00	0.00	Scenario2	0.00	0.00	8.75	0.00	0.00
$\tilde{\sigma}_m$		4.25	2.78	2.47	2.74	3.93		6.36	5.74	5.62	4.65	4.99
$\tilde{\mu}_m^*$		0.94 ±0.15	0.13 ±0.04	0.05 ±0.03	0.09 ±0.04	0.58 ±0.09		3.03 ±0.26	2.80 ±0.20	7.94 ±0.30	1.73 ±0.16	2.64 ±0.22
$\tilde{\sigma}_m^*$		5.67 ±0.07	3.93 ±0.05	3.43 ±0.04	3.57 ±0.04	4.62 ±0.05		6.99 ±0.07	6.11 ±0.05	5.74 ±0.04	5.13 ±0.05	5.27 ±0.05
$\tilde{\mu}_m$	Scenario3	2.94	10.82	3.75	15.06	0.00	Scenario4	14.70	0.00	0.00	11.00	3.70
$\tilde{\sigma}_m$		5.53	6.68	6.15	5.96	6.98		7.32	5.86	5.66	6.07	5.84
$\tilde{\mu}_m^*$		3.12 ±0.23	10.35 ±0.38	5.42 ±0.27	14.26 ±0.35	5.38 ±0.27		14.43 ±0.46	2.95 ±0.23	2.72 ±0.19	10.39 ±0.33	5.27 ±0.24
$\tilde{\sigma}_m^*$		6.87 ±0.08	6.35 ±0.05	6.43 ±0.06	6.07 ±0.04	6.14 ±0.05		7.41 ±0.07	6.27 ±0.06	5.85 ±0.05	5.84 ±0.04	6.08 ±0.05

3.4.7.4 Comportement des tests par paires

Les p-valeurs $\tilde{p}_m$ des tests pour les hypothèses nulles $\mathcal{H}_0^{(m)}$ sont approximées soit par les p-valeurs bootstrap $\bar{p}_m^*$ (cf. Eq. (3.26)) issues des statistiques bootstrap $\bar{\delta}_m^*$, soit par les p-valeurs bootstrap $\tilde{p}_m^*$ (cf. Eq. (3.30)) issues des statistiques bootstrap $\tilde{\delta}_m^*$, pour $m = 1, \ldots, M-1$. Le comportement de ces deux p-valeurs, censées suivre une loi uniforme théoriquement, et les puissances empiriques (des tests donnés par l'équation (3.19)) qui en découlent sont à présent étudiés, pour différentes tailles d'échantillon N.

Tests par paires issus des p-valeurs $\bar{p}_m^*$

En premier lieu, la figure 3.22 montre les diagrammes quantile-quantile de $\bar{p}_m^*$ au travers des réalisations de Monte Carlo par rapport à une distribution uniforme pour $m = 1, \ldots, M-1$ et différentes tailles d'échantillon $N = 2^{16}$, 2^{17}, 2^{18}. Celles-ci sont assez bien approximées par une loi uniforme sous l'hypothèse nulle $\mathcal{H}_0$ pour différentes tailles d'échantillon N, comme attendu puisque les statistiques bootstrap $\bar{\delta}_m^*$ reproduisent le comportement des statistique $\tilde{\delta}_m$ sous $\mathcal{H}_0$. En revanche, sous une hypothèse alternative $\mathcal{H}_1$, les distributions des p-valeurs $\bar{p}_m^*$ s'éloignent significativement de distributions uniformes sous les hypothèse nulles par paires $\mathcal{H}_0^{(m)}$, quelle que soit la taille d'échantillon N, car le paramètre d'échelle $\tilde{\sigma}_m$ de la loi de la statistique $\tilde{\delta}_m$ dépend du scénario.

En particulier, la figure 3.22 permet d'observer que les distributions des p-valeurs $\bar{p}_m^*$ sont davantage uniformes sous $\mathcal{H}_0^{(m)}$ pour les paires m éloignées de paires m' pour lesquelles $\mathcal{H}_0^{(m')}$ n'est pas vraie, par exemple $m = 1$ et $m = 5$ sous le scénario 2 où $m' = 3$, que pour des paires plus proches de celles-ci, par exemple $m = 2$ et $m = 4$ sous le scénario 2. Cependant, sous des hypothèses alternatives par paires $\mathcal{H}_1^{(m)}$, les distributions des p-valeurs $\bar{p}_m^*$ sont beaucoup plus éloignées de distributions uniformes, d'autant plus que N est grand.

En complément, le tableau 3.3 rapporte les puissances empiriques des tests par paires (obtenues comme moyennes des décisions de rejet $\bar{p}_m^* < \alpha_m$ à travers les réalisations de Monte Carlo) avec des niveaux de confiance $\alpha_m = 0.05$ pour $m = 1, \ldots, M-1$, les différents scénarios et différentes tailles d'échantillon N. Ces résultats confirment le comportement des p-valeurs, puisque le niveau de confiance ciblé α_m est bien reconstruit sous l'hypothèse nulle $\mathcal{H}_0$, mais pas lorsque $\mathcal{H}_0^{(m)}$ et $\mathcal{H}_1$ sont simultanément vraies, cas où les hypothèses nulles par paires $\mathcal{H}_0^{(m)}$ sont excessivement rejetées. En revanche, sous des hypothèses alternatives $\mathcal{H}_1^{(m)}$, les tests sont puissants, même à faible taille d'échantillon $N = 2^{16}$, et les puissances croissent avec la taille d'échantillon N.

Ces résultats suggèrent que la procédure de test reposant sur les p-valeurs $\bar{p}_m^*$ reproduit bien l'hypothèse nulle $\mathcal{H}_0$ et est puissante sous $\mathcal{H}_1$, mais ne reproduit pas les hypothèses nulles par paires $\mathcal{H}_0^{(m)}$.

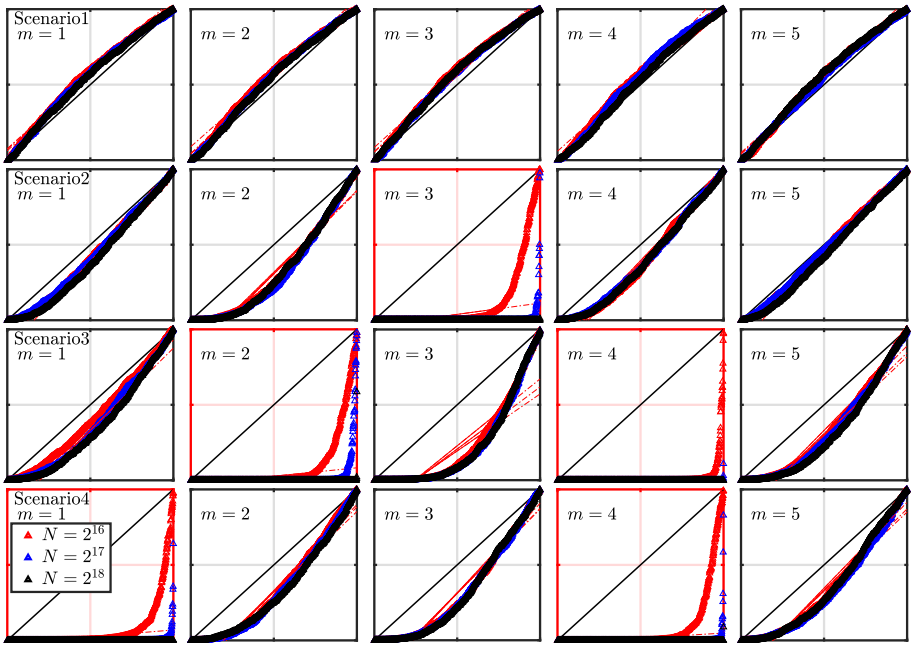


FIGURE 3.22 – **Estimées bootstrap des p-valeurs du test $\bar{p}^*_m$.** Diagrammes quantile-quantile des p-valeurs $\bar{p}^*_m$ (cf. Eq. (3.26)) pour $m = 1, \ldots, 5$ (de gauche à droite), les scénarios 1, 2, 3 et 4 (de haut en bas) et différentes tailles d'échantillon N. Les cases rouges correspondent aux écarts par rapport à $\mathcal{H}_0^{(m)}$. *Les distributions des p-valeurs $\bar{p}^*_m$ sont bien approximées par des distributions uniformes sous des hypothèses nulles par paires $\mathcal{H}_0^{(m)}$ lorsque l'hypothèse nulle $\mathcal{H}_0$ est vraie. Cette approximation est moins bonne sous des hypothèses nulles par paires $\mathcal{H}_0^{(m)}$ lorsque l'hypothèse nulle $\mathcal{H}_0$ n'est pas vraie*

TABLEAU 3.3 – **Puissances empiriques des tests issus des p-valeurs $\bar{p}^*_m$.** Puissances empiriques obtenues comme moyennes des décisions de rejet non corrigées $\bar{p}^*_m < \alpha_m$ à travers les réalisations de Monte Carlo pour $m = 1, \ldots, 5$, un niveau de confiance $\alpha_m = 0.05$ et différentes tailles d'échantillon N. Les cases rouges correspondent aux écarts par rapport à $\mathcal{H}_0^{(m)}$.

$\times 10^2$		$m=1$	$m=2$	$m=3$	$m=4$	$m=5$		$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
2^{16}	Scenario1	0.04	0.04	0.04	0.05	0.07	Scenario2	0.14	0.26	0.71	0.20	0.10
2^{17}		0.04	0.05	0.06	0.06	0.05		0.14	0.29	0.98	0.18	0.10
2^{18}		0.05	0.06	0.04	0.06	0.05		0.18	0.29	1.00	0.20	0.14
2^{16}	Scenario3	0.14	0.74	0.43	0.94	0.24	Scenario4	0.76	0.25	0.28	0.78	0.23
2^{17}		0.20	0.94	0.45	1.00	0.26		0.98	0.27	0.29	0.99	0.24
2^{18}		0.23	1.00	0.44	1.00	0.28		1.00	0.24	0.30	1.00	0.23

Tests par paires issus des p-valeurs $\tilde{p}^*_m$

En premier lieu, la figure 3.23 rapporte les diagrammes quantile-quantile des distributions empiriques des $\tilde{p}^*_m$ par rapport à une distribution uniforme pour différentes tailles d'échantillon $N = 2^{16}, 2^{17}, 2^{18}$. Ces diagrammes permettent de vérifier que les estimées bootstrap $\tilde{p}^*_m$ des p-valeurs suivent approximativement une distribution uniforme sous $\mathcal{H}_0^{(m)}$ lorsque $\mathcal{H}_0$ n'est pas vraie. Cependant, sous $\mathcal{H}_0$, la distribution des p-valeurs $\tilde{p}^*_m$ est moins bien approximée par une loi uniforme en raison de la moins bonne approximation de la distribution de $\tilde{\delta}_m$ par une loi normale repliée, mise en lumière précédemment par la figure 3.17. Enfin, sous $\mathcal{H}_1^{(m)}$, les $\tilde{p}^*_m$ s'écartent d'une distribution uniforme comme prévu, d'autant plus que la taille d'échantillon N est grande. Cet écart est cependant moins important que celui observé pour les p-valeurs $\bar{p}^*_m$ sur la figure 3.22.

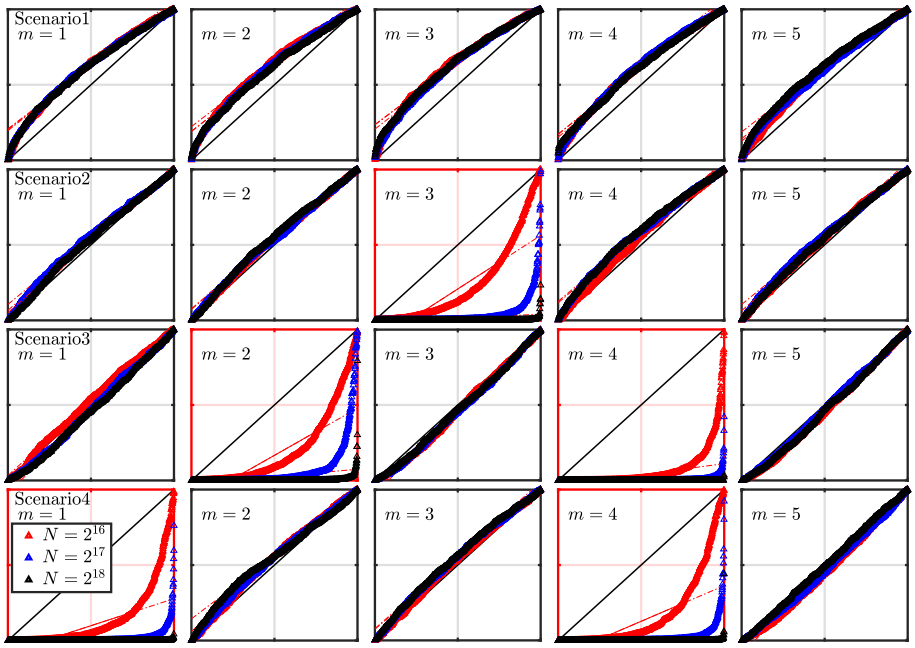


FIGURE 3.23 – **Estimées bootstrap des p-valeurs du test $\tilde{p}^*_m$.** Diagrammes quantile-quantile des p-valeurs $\tilde{p}^*_m$ (cf. Eq. (3.30)) pour $m = 1, \ldots, 5$ (de gauche à droite), les scénarios 1, 2, 3 et 4 (de haut en bas) et différentes tailles d'échantillon N. Les cases rouges correspondent aux écarts par rapport à $\mathcal{H}_0^{(m)}$. *Les distributions des p-valeurs $\tilde{p}^*_m$ sont bien approximées par des distributions uniformes sous les hypothèse nulles par paires $\mathcal{H}_0^{(m)}$, $\mathcal{H}_0$ n'est pas vraie et cette approximation est moins bonne lorsque $\mathcal{H}_0$ est vraie.*

Enfin, le tableau 3.4 compare, pour des niveaux de confiance prédéfinis $\alpha_m = 0.05$, les puissances de test empiriques, correspondant au pourcentage de rejets $\tilde{p}^*_m < \alpha_m$ au travers des réalisations de Monte Carlo, et les puissances de test estimées par bootstrap $\tilde{\pi}(\tilde{\mu}^*_m, \tilde{\sigma}^*_m)$ selon l'équation (3.31). Les résultats indiquent que, quel que soit $m \in \{1, \ldots, M-1\}$,

(*i*) le test reconstruit fidèlement le niveau de confiance prédéfini α_m sous les hypothèses nulles par paires $\mathcal{H}_0^{(m)}$ lorsque $\mathcal{H}_0$ n'est pas vraie, mais rejette peu $\mathcal{H}_0^{(m)}$ lorsque $\mathcal{H}_0$ est vraie ;

(*ii*) le test a une puissance assez faible sous $\mathcal{H}_1^{(m)}$ à faible taille d'échantillon N, atteignant jusqu'à 33% dans le scénario 2, mais qui croît de façon importante avec N ;

(*iii*) et la puissance bootstrap $\tilde{\pi}(\tilde{\mu}^*_m, \tilde{\sigma}^*_m)$ estime fidèlement le pourcentage de rejets dans les différents scénarios sous $\mathcal{H}_1^{(m)}$ mais le surestime sous $\mathcal{H}_0^{(m)}$, ce qui est en accord avec le comportement de l'estimateur $\tilde{\mu}^*_m$ du paramètre de position $\tilde{\mu}_m$ observé dans le tableau 3.2.

TABLEAU 3.4 – **Estimation des puissances des tests issus des p-valeurs $\tilde{p}^*_m$.** Estimées bootstrap $\tilde{\pi}(\tilde{\mu}^*_m, \tilde{\sigma}^*_m)$ (moyenne de Monte Carlo et intervalle interquartile) des puissances de test empiriques (i.e. le pourcentage de rejets $\tilde{p}^*_m < \alpha_m$) obtenues comme moyennes des décisions non corrigées des tests sur les réalisations de Monte Carlo pour des niveaux de confiance $\alpha_m = 0.05$, $m = 1, \ldots, 5$ et différentes tailles d'échantillon N. Les cases rouges indiquent les écarts à $\mathcal{H}_0^{(m)}$.

N			$m=1$	$m=2$	$m=3$	$m=4$	$m=5$		$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
2^{16}	$\langle \mathbb{1}_{\tilde{p}^*_m<0.05} \rangle$	Scenario1	0.02	0.01	0.02	0.01	0.02	Scenario2	0.04	0.04	0.33	0.04	0.04
	$\tilde{\pi}(\tilde{\mu}^*_m, \tilde{\sigma}^*_m)$		0.07 0.05 – 0.05	0.05 0.05 – 0.05	0.05 0.05 – 0.05	0.05 0.05 – 0.05	0.06 0.05 – 0.05		0.11 0.05 – 0.12	0.11 0.05 – 0.13	0.34 0.13 – 0.51	0.09 0.05 – 0.10	0.12 0.05 – 0.13
2^{17}	$\langle \mathbb{1}_{\tilde{p}^*_m<0.05} \rangle$		0.01	0.01	0.01	0.01	0.01		0.03	0.03	0.82	0.02	0.03
	$\tilde{\pi}(\tilde{\mu}^*_m, \tilde{\sigma}^*_m)$		0.06 0.05 – 0.05	0.05 0.05 – 0.05	0.05 0.05 – 0.05	0.05 0.05 – 0.05	0.06 0.05 – 0.05		0.11 0.05 – 0.13	0.11 0.05 – 0.13	0.67 0.05 – 0.87	0.09 0.05 – 0.09	0.11 0.05 – 0.11
2^{18}	$\langle \mathbb{1}_{\tilde{p}^*_m<0.05} \rangle$		0.01	0.01	0.00	0.00	0.01		0.05	0.05	0.99	0.02	0.03
	$\tilde{\pi}(\tilde{\mu}^*_m, \tilde{\sigma}^*_m)$		0.06 0.05 – 0.05	0.05 0.05 – 0.05	0.05 0.05 – 0.05	0.05 0.05 – 0.05	0.06 0.05 – 0.05		0.13 0.05 – 0.14	0.13 0.05 – 0.14	0.95 0.94 – 1.00	0.09 0.05 – 0.09	0.12 0.05 – 0.13
2^{16}	$\langle \mathbb{1}_{\tilde{p}^*_m<0.05} \rangle$	Scenario3	0.06	0.40	0.09	0.69	0.09	Scenario4	0.51	0.05	0.05	0.49	0.09
	$\tilde{\pi}(\tilde{\mu}^*_m, \tilde{\sigma}^*_m)$		0.12 0.05 – 0.14	0.42 0.17 – 0.65	0.19 0.07 – 0.24	0.62 0.41 – 0.86	0.19 0.07 – 0.24		0.51 0.26 – 0.77	0.11 0.05 – 0.13	0.11 0.05 – 0.13	0.46 0.21 – 0.69	0.19 0.08 – 0.23
2^{17}	$\langle \mathbb{1}_{\tilde{p}^*_m<0.05} \rangle$		0.07	0.75	0.09	0.97	0.07		0.92	0.04	0.04	0.89	0.07
	$\tilde{\pi}(\tilde{\mu}^*_m, \tilde{\sigma}^*_m)$		0.15 0.06 – 0.18	0.64 0.45 – 0.88	0.20 0.08 – 0.25	0.89 0.85 – 0.99	0.19 0.07 – 0.24		0.83 0.73 – 0.98	0.12 0.05 – 0.13	0.11 0.05 – 0.12	0.76 0.64 – 0.94	0.18 0.08 – 0.22
2^{18}	$\langle \mathbb{1}_{\tilde{p}^*_m<0.05} \rangle$		0.08	0.98	0.07	1.00	0.07		1.00	0.03	0.04	1.00	0.06
	$\tilde{\pi}(\tilde{\mu}^*_m, \tilde{\sigma}^*_m)$		0.16 0.07 – 0.21	0.92 0.91 – 1.00	0.19 0.07 – 0.24	0.99 1.00 – 1.00	0.19 0.07 – 0.25		0.99 0.99 – 1.00	0.11 0.05 – 0.12	0.11 0.05 – 0.13	0.97 0.98 – 1.00	0.18 0.07 – 0.21

3.4.7.5 Comportement de la correction pour des hypothèses multiples

Les décisions de rejets pour les hypothèses nulles $\mathcal{H}_0^{(m)}$ sont prises en comparant des p-valeurs $\bar{p}_m^*$ (cf. Eq. (3.26)) ou $\tilde{p}_m^*$ (cf. Eq. (3.30)) à des seuils définis à partir d'une procédure de correction, pour $m = 1, \ldots, M-1$. Les trois procédures de correction présentées dans la section 3.4.4 sont comparées ici à partir du comportement des décisions de rejet d_α résultantes ($\bar{d}_\alpha$ ou $\tilde{d}_\alpha$) de l'hypothèse nulle $\mathcal{H}_0$.

Test issu des p-valeurs $\bar{p}_m^*$

Sur la figure 3.24, sont rapportées les décisions de rejet $\bar{d}_\alpha$ (Eq. (3.32)) de l'hypothèse nulle $\mathcal{H}_0$ associées aux p-valeurs $\bar{p}_m^*$ (Eq. (3.26)) calculées suivant les différentes procédures de correction et moyennées sur les réalisations de Monte Carlo en fonction du niveau de confiance ciblé α sous $\mathcal{H}_0$ (Scenario1) pour différentes tailles d'échantillon $N = 2^{16}, 2^{17}, 2^{18}$. Le seuil de confiance ciblé α est reconstruit avec les procédures de Bonferroni et Benjamini-Hochberg, mais pas avec celle de Benjamini-Yekutieli, quelle que soit la taille d'échantillon N.

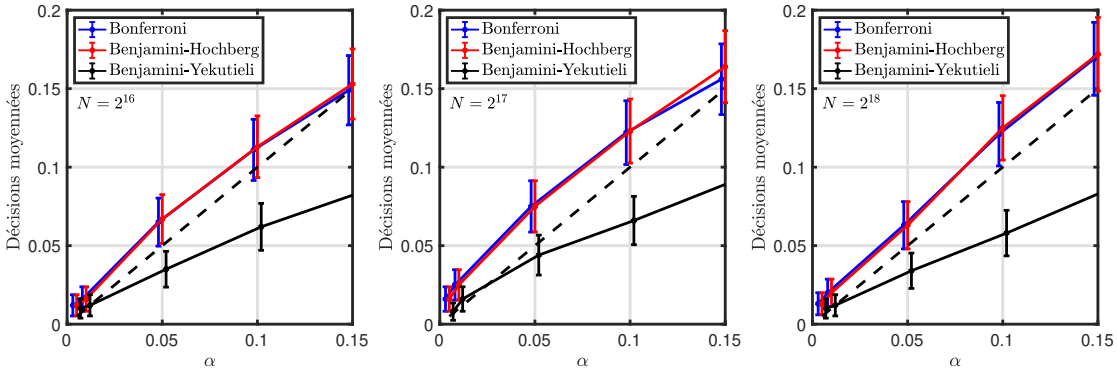


FIGURE 3.24 – **Reproduction de $\mathcal{H}_0$ à partir des p-valeurs $\bar{p}_m^*$.** Décisions $\bar{d}_\alpha$ moyennées sur les réalisations de Monte Carlo en fonction du niveau de confiance ciblé α sous $\mathcal{H}_0$ (Scenario1), différentes corrections (cf. Section 3.4.4) et différentes tailles d'échantillon N (de gauche à droite). Les lignes noires pointillées indiquent les valeurs théoriques attendues, c'est-à-dire la première bissectrice. *Les procédures de Bonferroni et Benjamini-Hochberg assurent une bonne reconstruction du niveau de confiance ciblé α.*

Test issu des p-valeurs $\tilde{p}_m^*$

La figure 3.25 rapporte les décisions de rejet $\tilde{d}_\alpha$ (Eq. (3.32)) de l'hypothèse nulle $\mathcal{H}_0$ associées aux p-valeurs $\tilde{p}_m^*$ (Eq. (3.30)) pour différentes corrections et moyennées sur les réalisations de Monte Carlo en fonction du niveau de confiance ciblé α pour différentes tailles d'échantillon $N = 2^{16}, 2^{17}, 2^{18}$. Comme attendu au vu du comportement des p-valeurs $\tilde{p}_m^*$ observé sur la figure 3.23, le test de rejet de l'hypothèse nulle $\mathcal{H}_0$ est très conservatif, et ce quelle que soit la correction utilisée. Ce comportement s'accentue lorsque la taille d'échantillon N augmente. Cependant, les procédures de Bonferroni et Benjamini-Hochberg permettent une reconstruction légèrement meilleure du seuil de confiance ciblé α que la procédure de Benjamini-Yekutieli.

Conclusions

Ces résultats confirment que les p-valeurs bootstrap $\bar{p}_m^*$ sont plus adaptées à un test de rejet de l'hypothèse nulle $\mathcal{H}_0$ que les p-valeurs bootstrap $\tilde{p}_m^*$, en raison de la mauvaise approximation des distributions des $\tilde{\delta}_m$ par des distributions normales repliées. Cependant, les p-valeurs bootstrap $\tilde{p}_m^*$ permettent une meilleure reproduction des hypothèses nulles par paires $\mathcal{H}_0^{(m)}$ sous $\mathcal{H}_1$ comparées aux p-valeurs $\bar{p}_m^*$, suggérant tout de même d'exploiter les décisions corrigées $\tilde{d}_\alpha^{(m)}$ issues des p-valeurs $\tilde{p}_m^*$ pour le partitionnement.

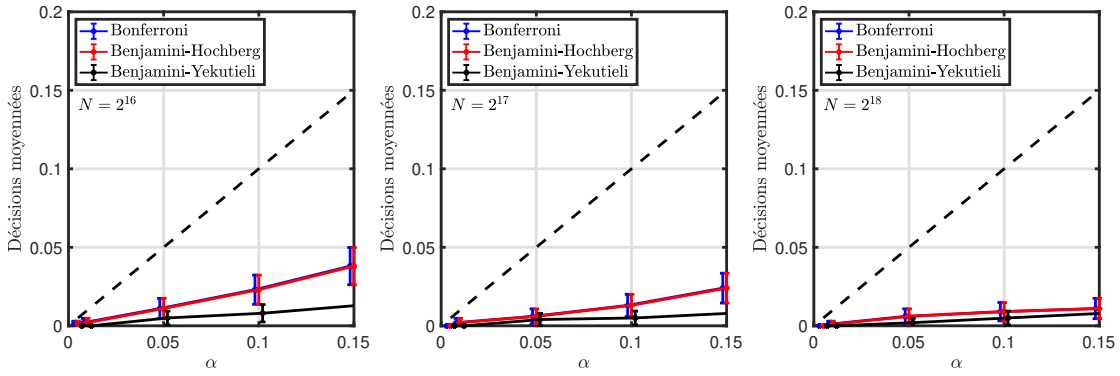


FIGURE 3.25 – **Reproduction de $\mathcal{H}_0$ à partir des p-valeurs $\tilde{p}^*_m$.** Décisions $\tilde{d}_\alpha$ moyennées sur les réalisations de Monte Carlo en fonction du niveau de confiance ciblé α sous $\mathcal{H}_0$ (Scenario1), différentes corrections (cf. Section 3.4.4) et différentes tailles d'échantillon N (de gauche à droite). Les lignes noires pointillées indiquent les valeurs théoriques attendues, c'est-à-dire la première bissectrice. *Toutes les procédures mènent à un test de rejet de l'hypothèse nulle $\mathcal{H}_0$ trop conservatif.*

3.4.7.6 Performances de la stratégie de partitionnement

La stratégie de partitionnement détaillée dans la section 3.4.5 peut être réalisée en utilisant soit les $M-1$ décisions de rejet $\bar{d}^{(m)}_\alpha$ issues des p-valeurs $\bar{p}^*_m$, soit les $M-1$ décisions de rejet $\tilde{d}^{(m)}_\alpha$ issues des p-valeurs $\tilde{p}^*_m$, calculées suivant une des procédures de correction présentées dans la section 3.4.4. Au vu des résultats précédents, la correction de Benjamini-Yekutieli est écartée de l'analyse, et seules les procédures de Bonferroni et de Benjamini-Hochberg sont considérées. On se propose de comparer ces différentes approches, paramétrées par le niveau de confiance α.

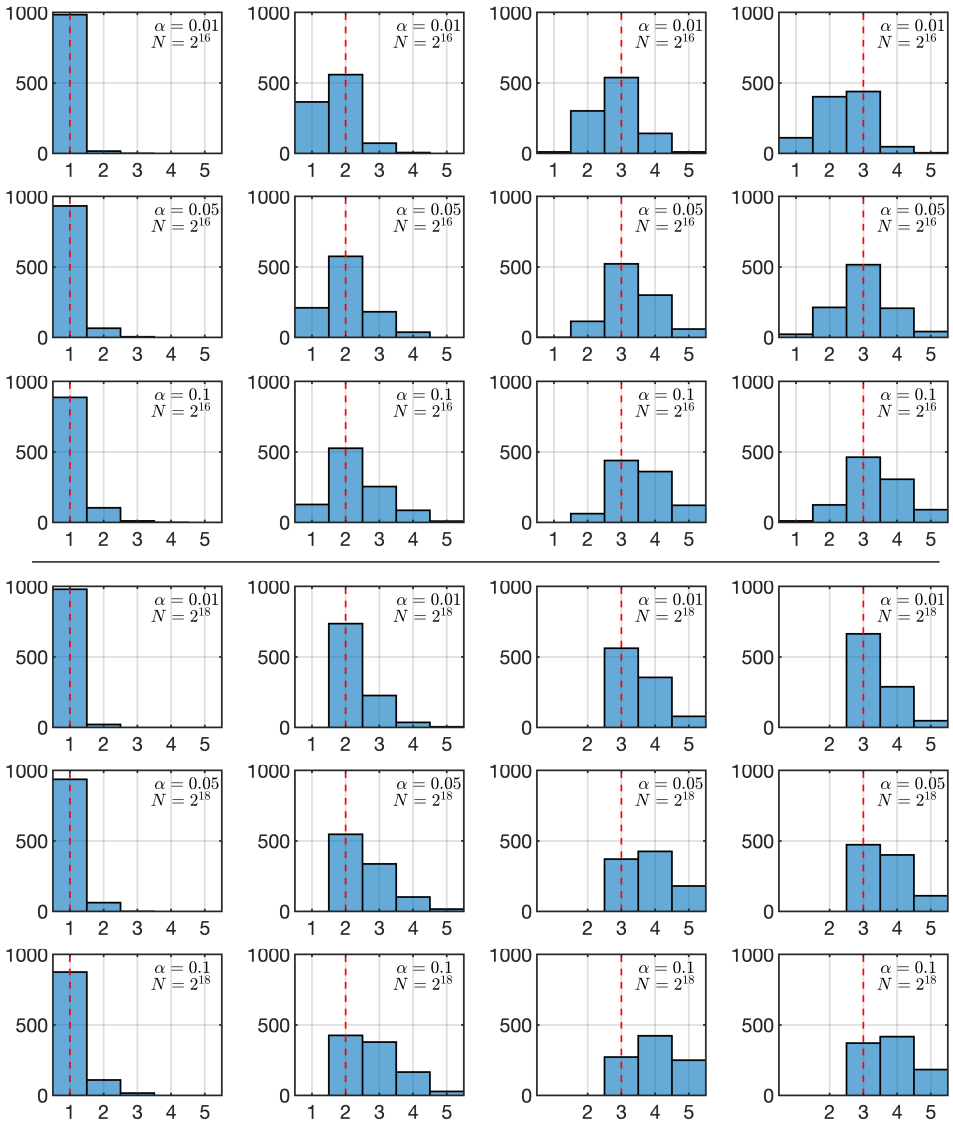
Mesures de performances

L'indice de Rand ajusté (ARI) et l'information mutuelle normalisée (NMI) (VINH et collab., 2010) sont utilisés pour quantifier les performances du partitionnement, c'est-à-dire la façon dont les éléments ayant la même valeur H sont regroupés et les éléments ayant des valeurs H différentes sont séparés. L'ARI mesure le nombre de paires d'éléments qui sont correctement regroupés ou séparés, tandis que la NMI mesure l'entropie conjointe des distributions des partitionnements estimé et correct. La définition de ces mesures de performances est donnée dans l'annexe C.

Partitionnement fondé sur les p-valeurs $\bar{p}^*_m$

En premier lieu, la figure 3.26 présente les histogrammes du nombre estimé de partitions obtenu à partir des tests par paires construits suivant la procédure de Benjamini-Hochberg pour trois niveaux différents de taux de fausses découvertes $\alpha = (0.01,\ 0.05,\ 0.10)$, les différents scénarios et différentes tailles d'échantillon $N = 2^{16}, 2^{18}$. Pour le scénario 1, dans lequel les $M-1=5$ hypothèses nulles $\mathcal{H}^{(m)}_0$ sont vraies, on obtient des taux de fausses découvertes réels de $(0.06,\ 0.07,\ 0.11)$ pour $N = 2^{16}$ et $(0.02,\ 0.06,\ 0.13)$ pour $N = 2^{18}$, respectivement, ce qui est en bon accord avec les valeurs prédéfinies. Cela corrobore davantage l'analyse numérique ci-dessus sur la pertinence de la procédure de test et de détection de l'hypothèses nulle $\mathcal{H}_0$ à partir des p-valeurs $\bar{p}^*_m$. Pour les scénarios 2, 3 et 4, dans lesquels plus d'une partition doit être détectée, la procédure proposée détecte le nombre correct de partitions dans la majorité des cas pour les différentes tailles d'échantillon $N = 2^{16}, 2^{18}$. Comme prévu, le taux de surestimation (resp. sous-estimation) du nombre de groupes augmente (resp. diminue) avec l'augmentation (resp. la diminution) du taux de fausses découvertes prédéfini α. On observe cependant que le nombre de partitions est souvent surestimé dans les différents scénarios, d'autant plus que la

taille d'échantillon N est grande. Ceci est dû à la mauvaise reconstruction des hypothèses nulles par paires $\mathcal{H}_0^{(m,m')}$, comme quantifié par les puissances de test dans le tableau 3.4.



(a) Scenario1 (b) Scenario2 (c) Scenario3 (d) Scenario4

FIGURE 3.26 – **Nombre de partitions estimé à partir des p-valeurs $\bar{p}_m^*$.** Histogrammes du nombre estimé de partitions pour (a) 1 partition, (b) 2 partitions et (c-d) 3 partitions avec $M = 6$ composantes en utilisant la stratégie de partitionnement, définie dans la section 3.4.5, à partir des p-valeurs $\bar{p}_m^*$ en utilisant la procédure de correction de Benjamini-Hochberg pour (de haut en bas) différents taux de fausses découvertes α et différentes tailles d'échantillon N. Les lignes rouges pointillées indiquent le nombre exact de partitions.

Le tableau 3.5 quantifie davantage cette analyse de performance de partitionnement et montre les valeurs moyennes des NMI et ARI avec un intervalle de confiance à 95% (à travers les réalisations de Monte Carlo) pour les quatre scénarios, différentes tailles d'échantillon $N = 2^{16}, 2^{17}, 2^{18}$ et différentes procédures de correction avec un niveau de confiance prédéfini $\alpha = 0.05$ pour le partitionnement. Les résultats indiquent que la procédure proposée conduit globalement à des performances satisfaisantes en pratique, avec des valeurs de NMI allant jusqu'à 0.95 et des valeurs d'ARI allant jusqu'à 0.88 pour les scénarios à plusieurs partitions (i.e. les scénarios 2, 3 et 4) et une taille d'échantillon $N = 2^{18}$. Par ailleurs, la procédure atteint

des performances légèrement meilleures avec la correction de Bonferroni qu'avec la correction Benjamini-Hochberg à grande taille d'échantillon N. Sous le scénario 1, correspondant à l'hypothèse nulle, les performances sont similaires entre les corrections de Bonferroni et Benjamini-Hochberg, en accord avec la reproduction de l'hypothèse nulle observée sur la figure 3.24. Il est intéressant de noter que, pour une taille d'échantillon $N = 2^{16}$, le scénario 4 est plus difficile à partitionner que le scénario 3, qui contient le même nombre de partitions et les mêmes valeurs pour H mais a des partitions de tailles égales.

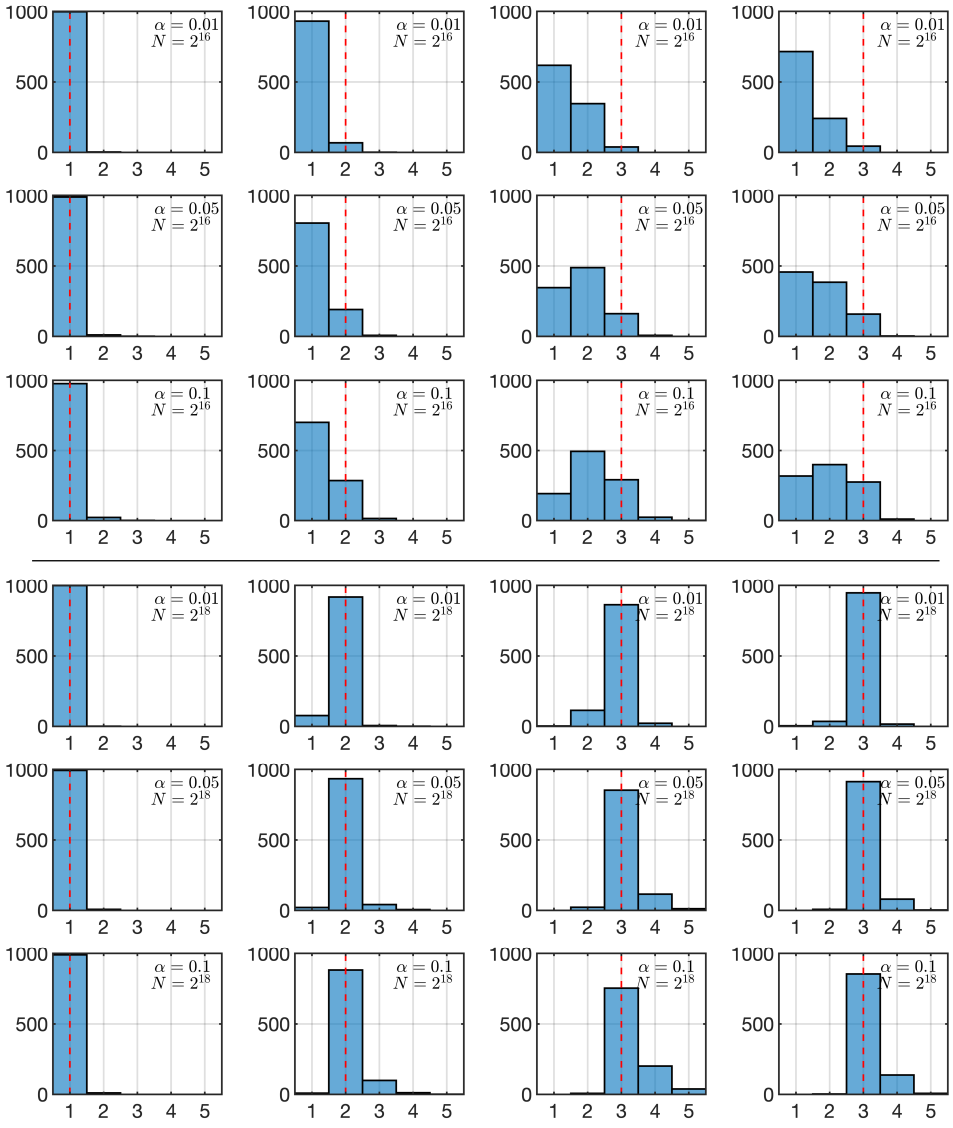
TABLEAU 3.5 – **Performances du partitionnement fondé sur les p-valeurs $\bar{p}^*_m$.** Valeurs des ARI et NMI (moyenne de Monte Carlo ± intervalle de confiance à 95%) atteintes par la stratégie de partitionnement (cf. Section 3.4.5) fondée sur les p-valeurs $\bar{p}^*_m$ pour différents scénarios, différentes procédures de correction avec un niveau de confiance prédéfini $\alpha = 0.05$ et différentes tailles d'échantillon N.

	N		Scenario1	Scenario2	Scenario3	Scenario4
Bonferroni	2^{16}	NMI	n/a	0.66 ± 0.03	0.86 ± 0.01	0.77 ± 0.01
		ARI	0.94 ± 0.02	0.61 ± 0.03	0.69 ± 0.02	0.58 ± 0.02
	2^{17}	NMI	n/a	0.91 ± 0.01	0.93 ± 0.01	0.93 ± 0.01
		ARI	0.93 ± 0.02	0.87 ± 0.01	0.83 ± 0.01	0.81 ± 0.02
	2^{18}	NMI	n/a	0.93 ± 0.01	0.95 ± 0.00	0.95 ± 0.00
		ARI	0.94 ± 0.02	0.88 ± 0.01	0.85 ± 0.01	0.85 ± 0.01
Benjamini-Hochberg	2^{16}	NMI	n/a	0.67 ± 0.02	0.87 ± 0.01	0.79 ± 0.01
		ARI	0.93 ± 0.02	0.60 ± 0.03	0.69 ± 0.02	0.58 ± 0.02
	2^{17}	NMI	n/a	0.90 ± 0.01	0.92 ± 0.01	0.92 ± 0.01
		ARI	0.93 ± 0.02	0.84 ± 0.01	0.77 ± 0.02	0.75 ± 0.02
	2^{18}	NMI	n/a	0.91 ± 0.01	0.93 ± 0.00	0.93 ± 0.00
		ARI	0.94 ± 0.02	0.83 ± 0.01	0.77 ± 0.01	0.77 ± 0.02

Sous le scénario 1, constitué d'une seule partition, l'ARI est contrôlé par le niveau de confiance réel $\hat{\alpha}$ (c'est-à-dire la moyenne des décisions de rejet au travers des réalisations de Monte Carlo, $\hat{\alpha} = \langle \bar{d}_\alpha \rangle$) par la relation $ARI = 1 - \hat{\alpha}$. Comme observé précédemment sur la figure 3.24, le niveau de confiance ciblé α est bien reconstruit par $\hat{\alpha}$.

Partitionnement fondé sur les p-valeurs $\tilde{p}^*_m$

Tout d'abord, la figure 3.27 présente les histogrammes du nombre estimé de partitions pour les différents scénarios, trois niveaux différents de taux de fausses découvertes $\alpha = (0.01, 0.05, 0.10)$ et différentes tailles d'échantillon $N = 2^{16}, 2^{18}$. Pour le scénario 1, dans lequel les $M - 1 = 5$ hypothèses nulles sont vraies, nous obtenons des taux de fausses découvertes réels de $(0.00, 0.01, 0.02)$ pour $N = 2^{16}$ et de $(0.00, 0.01, 0.01)$ pour $N = 2^{18}$. Ces taux sont inférieurs aux valeurs prédéfinies, comme attendu au vu des résultats donnés par la figure 3.25. Pour les scénarios 2, 3 et 4, dans lesquels plus d'une partition doit être détectée, la procédure proposée détecte le nombre correct de partitions dans la majorité des cas pour $N = 2^{18}$ mais sous-estime généralement ce nombre pour $N = 2^{16}$. Ce résultat est en accord avec le tableau 3.3, puisque les puissances des tests par paires sont très faibles à $N = 2^{16}$ mais augmentent rapidement avec la taille d'échantillon N.



(a) Scenario1 (b) Scenario2 (c) Scenario3 (d) Scenario4

FIGURE 3.27 – **Nombre de partitions estimé à partir des p-valeurs $\tilde{p}^*_m$.** Histogrammes du nombre estimé de partitions pour (a) 1 partition, (b) 2 partitions et (c-d) 3 partitions avec $M = 6$ composantes en utilisant la stratégie de partitionnement, définie dans la section 3.4.5, à partir des p-valeurs $\tilde{p}^*_m$ en utilisant la procédure de correction de Benjamini-Hochberg pour (de haut en bas) différents taux de fausses découvertes α et différentes tailles d'échantillon N. Les lignes rouges pointillées indiquent le nombre exact de partitions.

Le tableau 3.6 quantifie la stratégie de partitionnement en montrant les valeurs moyennes des NMI et ARI avec un intervalle de confiance à 95% (calculées à travers les réalisations de Monte Carlo) pour les quatre scénarios, différentes tailles d'échantillon $N = 2^{16}, 2^{17}, 2^{18}$ et différentes procédures de correction avec un niveau de confiance prédéfini $\alpha = 0.05$. Les résultats indiquent que la procédure proposée conduit à de très faibles performances de partitionnement pour la faible taille d'échantillon $N = 2^{16}$, avec notamment des valeurs de NMI et ARI respectives de 0.17 et 0.16 dans le scénario 2, en raison de la sous-estimation du nombre de partitions N_C. En comparaison, pour le partitionnement à partir des p-valeurs $\bar{p}^*_m$, les valeurs de NMI et ARI ne descendent pas en dessous de 0.66 et 0.60, respectivement (cf. Tableau 3.5).

TABLEAU 3.6 – **Performances du partitionnement fondé sur les p-valeurs $\tilde{p}^*_m$.** Valeurs des ARI et NMI (moyenne de Monte Carlo ± intervalle de confiance à 95%) atteintes par la stratégie de partitionnement (cf. Section 3.4.5) fondée les p-valeurs $\tilde{p}^*_m$ pour différents scénarios, différentes procédures de correction avec un niveau de confiance prédéfini $\alpha = 0.05$ et différentes tailles d'échantillon N.

	N		Scenario1	Scenario2	Scenario3	Scenario4
Bonferroni	2^{16}	NMI	n/a	0.17 ± 0.02	0.49 ± 0.02	0.39 ± 0.02
		ARI	0.99 ± 0.01	0.16 ± 0.02	0.31 ± 0.02	0.26 ± 0.02
	2^{17}	NMI	n/a	0.62 ± 0.03	0.84 ± 0.01	0.83 ± 0.02
		ARI	0.99 ± 0.00	0.62 ± 0.03	0.69 ± 0.02	0.73 ± 0.02
	2^{18}	NMI	n/a	0.97 ± 0.01	0.98 ± 0.00	0.99 ± 0.02
		ARI	0.99 ± 0.00	0.97 ± 0.01	0.95 ± 0.01	0.98 ± 0.01
Benjamini-Hochberg	2^{16}	NMI	n/a	0.17 ± 0.02	0.51 ± 0.02	0.42 ± 0.02
		ARI	0.99 ± 0.01	0.16 ± 0.02	0.35 ± 0.02	0.30 ± 0.02
	2^{17}	NMI	n/a	0.62 ± 0.03	0.87 ± 0.01	0.87 ± 0.01
		ARI	0.99 ± 0.00	0.62 ± 0.03	0.74 ± 0.02	0.79 ± 0.02
	2^{18}	NMI	n/a	0.97 ± 0.01	0.98 ± 0.00	0.99 ± 0.00
		ARI	0.99 ± 0.00	0.97 ± 0.01	0.95 ± 0.01	0.97 ± 0.01

Cependant, la procédure conduit globalement à des performances satisfaisantes en pratique pour une taille d'échantillon N assez grande, avec des valeurs de NMI et ARI allant respectivement jusqu'à 0.99 et 0.98 pour une taille d'échantillon $N = 2^{18}$. Par ailleurs, la procédure atteint des performances légèrement meilleures avec la correction de Benjamini-Hochberg qu'avec la correction de Bonferroni à faible taille d'échantillon, mais celles-ci deviennent similaires à grande taille d'échantillon N.

Enfin, on peut remarquer que, sous le scénario 1, constitué d'une seule partition, l'ARI est toujours proche de 1. Ceci est lié à la nature conservative du test de rejet de l'hypothèse nulle $\mathcal{H}_0$ associé, comme observé sur la figure 3.25. En effet, le niveau de confiance réel $\hat{\alpha}$, donné par la relation $ARI = 1 - \hat{\alpha}$, surestime le niveau de confiance ciblé α .

Conclusions

La stratégie de partitionnement à partir des p-valeurs $\bar{p}^*_m$ mène à des résultats satisfaisants mais surestime le nombre de partitions lorsqu'il y en a plus d'une. Quant à la stratégie de partitionnement à partir des p-valeurs $\tilde{p}^*_m$, elle nécessite une taille d'échantillon suffisamment grande pour atteindre des performances satisfaisantes en pratique, et en particulier pour surpasser les performances de la stratégie fondée sur les p-valeurs $\bar{p}^*_m$.

3.4.8 Conclusions

La présente section constitue une première tentative de dénombrement des exposants d'autosimilarité effectivement différents, à partir d'une seule observation de taille finie de données multivariées. Elle repose sur la combinaison de la procédure d'estimation présentée dans le chapitre 2 et d'une procédure de $M - 1$ tests par paires construit en ordonnant les exposants d'autosimilarité $H_1 \leq \ldots \leq H_M$.

Les paramètres des tests par paires sont estimés à partir de la procédure de ré-échantillonnage

bootstrap par blocs introduite dans la section 3.2. La complexité de cet estimation, dû au tri des estimées $\hat{H}_1, \ldots, \hat{H}_M$ des $H_1, \ldots, H_M$, a mené à la construction de deux différents types de statistiques bootstrap, les premières permettant de reproduire l'hypothèse nulle globale $H_1 = \ldots = H_M$ tandis que les secondes reproduisent les hypothèses observées ($H_m = H_{m+1}$ vraie ou non, pour chaque $m = 1, \ldots, M-1$). Des simulations de Monte Carlo, réalisées sur des M-mBf synthétiques avec un faible nombre de composantes ($M = 6$), comparent les deux approches et montrent dans quels cas les statistiques bootstrap reproduisent de manière satisfaisante les distributions des statistiques des tests par paires : la première, seulement lorsque tous les exposants d'autosimilarité H_m sont égaux ; la seconde, à la fois lorsque tous les exposants d'autosimilarité H_m sont égaux et lorsqu'il existe des groupes de tailles possiblement différentes avec des exposants d'autosimilarité H_m de valeurs différentes, résultat non trivial. En outre, la procédure globale donne des résultats très satisfaisants pour l'estimation du nombre exact d'exposants d'autosimilarité distincts et pour le regroupement des exposants d'autosimilarité identiques, dans plusieurs scénarios avec un, deux ou trois groupes de valeurs, et pour des groupes de tailles éventuellement déséquilibrées.

Cependant, les distributions des statistiques sous les $M-1$ hypothèses nulles par paires $H_m = H_{m+1}$ ne sont connues que de manière approximative et l'approximation faite est seulement valable pour un petit nombre de composantes M et de petits groupes d'exposants d'autosimilarité égaux. La stratégie est évaluée sur un plus grand nombre de composantes dans la section 3.6, qui en confirme les limites. Une perspective consiste à examiner si l'utilisation de $M(M-1)/2$ hypothèses non triées améliore encore les performances par rapport à l'utilisation de $M-1$ hypothèses.

3.5 Tests d'égalité par paires d'exposants

Cette approche est l'objet d'un article en cours d'écriture intitulé « Multivariate self-similarity parameter counting Spectral clustering using wavelet-domain bootstrap », par C.-G. Lucas, P. Abry, H. Wendt et G. Didier, prévu pour être soumis à une revue internationale.

Cette présente section vise à tester l'égalité entre toutes les paires d'exposants d'autosimilarité $(H_m, H_{m'})$, avec $1 \leq m < m' \leq M$. La mise en jeu de $M(M-1)/2$ hypothèses pour regrouper les M exposants d'autosimilarité H_m par groupe d'exposants d'autosimilarité égaux impose le recours à une stratégie de partitionnement plus élaborée que dans la section précédente. Par exemple, si le test rejette l'égalité $H_1 = H_3$ mais pas les égalités $H_1 = H_2$ et $H_2 = H_3$, le choix des partitions dans lesquelles affecter H_1, H_2 et H_3 n'est pas évident. Une méthode de partitionnement spectral d'un graphe pondéré à l'aide des p-valeurs et décisions de tests par paires est mise en place pour répondre à ce problème.

3.5.1 Formulation des tests

Les hypothèses nulles pour les $M(M-1)/2$ tests par paires sont définies comme suit :

$$\forall\, 1 \leq m < m' \leq M, \quad \mathcal{H}_0^{(m,m')} : \quad H_m = H_{m'}. \tag{3.37}$$

Les hypothèses alternatives sont notées, pour tous $m, m' \in \{1, \ldots, M\}$,

$$\mathcal{H}_1^{(m,m')} : \quad \mathcal{H}_0^{(m,m')} \text{ n'est pas vraie.} \tag{3.38}$$

Comme dans la section précédente, chaque hypothèse est testée à partir d'une observation unique de données M-variées de taille finie et donc du vecteur des M estimées $\underline{\hat{H}} = (\hat{H}_1, \ldots, \hat{H}_M)$, obtenues selon la procédure décrite dans la section 2.2. Étant donnée la distribution asymptotiquement gaussienne du vecteur $\underline{\hat{H}}$, des statistiques naturelles pour tester ces hypothèses sont, pour tout couple d'entiers (m, m') tel que $1 \leq m < m' \leq M$,

$$\hat{\delta}_{m,m'} := \hat{H}_{m'} - \hat{H}_m. \tag{3.39}$$

Quelle que soit l'hypothèse observée ($\mathcal{H}_0^{(m,m')}$ vraie ou non), les statistiques de test $\hat{\delta}_{m,m'}$ suivent asymptotiquement une loi gaussienne de paramètres $\hat{\mu}_{m,m'}$ et $\hat{\sigma}_{m,m'}$. En particulier, sous l'hypothèse nulle $\mathcal{H}_0^{(m,m')}$, la statistique $\hat{\delta}_{m,m'}$ suit asymptotiquement une loi normale centrée, i.e. $\hat{\mu}_{m,m'} = 0$, d'écart-type $\hat{\sigma}_{m,m'}$ inconnu. Le paramètre $\hat{\sigma}_{m,m'}$ est donc approximé à l'aide de la procédure bootstrap introduite dans la section 3.2, ce qui permet l'approximation des p-valeurs $\hat{p}_{m,m'}$ des tests, comme décrit dans la section suivante. Les différentes notations importantes sont illustrées par la figure 3.28.

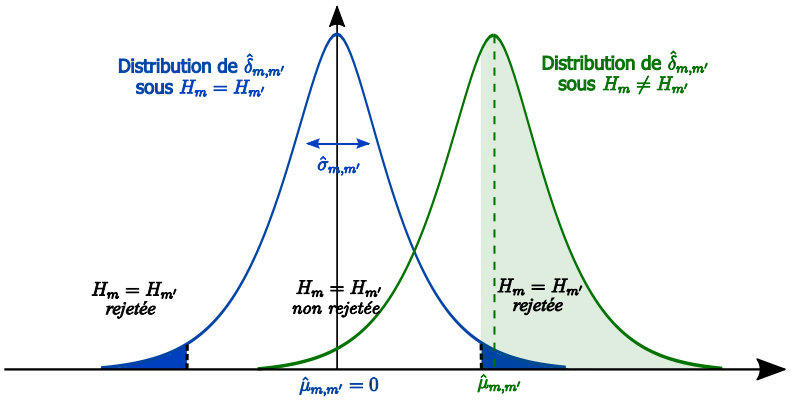


FIGURE 3.28 – **Illustration du test gaussien.** La statistique $\hat{\delta}_{m,m'}$ suit asymptotiquement une loi gaussienne d'écart-type $\hat{\sigma}_{m,m'}$. L'hypothèse nulle $H_m = H_{m'}$ est rejeté lorsque la statistique observée est au-delà des seuils de part et d'autre de l'origine (lignes noires pointillées), car la statistique observée appartient alors aux valeurs les moins probables pour la distribution normale centrée (en bleu). Plus la distribution normale non centrée (en vert), sous une hypothèse alternative donnée, s'en écarte, plus la probabilité de rejeter l'hypothèse nulle $H_m = H_{m'}$ sera élevée.

3.5.2 Estimation des p-valeurs par bootstrap

Estimation bootstrap paramétrique

Pour estimer le paramètre $\hat{\sigma}_{m,m'}$ du test pour l'hypothèse nulle $\mathcal{H}_0^{(m,m')}$, la distribution de la statistique $\hat{\delta}_{m,m'}$ est reproduite par bootstrap sous $\mathcal{H}_0^{(m,m')}$ à partir de R estimées bootstrap d'exposants d'autosimilarité $\underline{\hat{H}}^{*(r)} = (\hat{H}_1^{*(r)}, \ldots, \hat{H}_M^{*(r)})$, avec $r \in \{1, \ldots, R\}$. Les statistiques bootstrap sont ainsi définies, pour tous $1 \leq m < m' \leq M$ et $r \in \{1, \ldots, R\}$, par

$$\hat{\delta}_{m,m'}^{*(r)} := \hat{H}_{m'}^{*(r)} - \hat{H}_m^{*(r)} - \langle \hat{H}_{m'}^* - \hat{H}_m^* \rangle, \tag{3.40}$$

où $\langle \hat{H}_{m'}^* - \hat{H}_m^* \rangle$ correspond à la moyenne sur les réalisations bootstrap,

$$\langle \hat{H}_{m'}^* - \hat{H}_m^* \rangle := \frac{1}{R} \sum_{r=1}^{R} \left(\hat{H}_{m'}^{*(r)} - \hat{H}_m^{*(r)} \right). \tag{3.41}$$

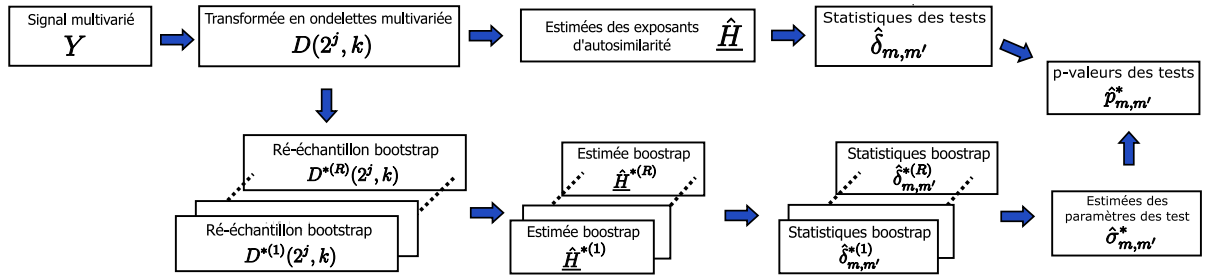
Les paramètres des tests $\hat{\sigma}_{m,m'}$ sont ainsi estimés à l'aide de l'estimateur bootstrap de la variance Var*, pour tous $1 \leq m < m' \leq M$, comme suit :

$$\hat{\sigma}^{*2}_{m,m'} := \text{ Var }^*(\hat{\delta}^*_{m,m'}). \tag{3.42}$$

Ensuite, les p-valeurs $\hat{p}_{m,m'}$ des tests pour $\mathcal{H}_0^{(m,m')}$ peuvent être approximées par bootstrap, pour tous $1 \leq m < m' \leq M$,

$$\hat{p}^*_{m,m'} := 2\left(1 - F_{\mathcal{N}(0,\hat{\sigma}^{*2}_{m,m'})}(|\hat{\delta}_{m,m'}|)\right), \tag{3.43}$$

où $\mathrm{F}_{\mathcal{N}(0,\hat{\sigma}^{*2}_{m,m'})}$ est la fonction de répartition de la loi normale centrée d'écart-type $\hat{\sigma}^*_{m,m'}$. La procédure d'estimation des p-valeurs par bootstrap est résumée par la figure 3.29.



,

FIGURE 3.29 – **Organigramme de la procédure de test bootstrap à $M(M-1)/2$ hypothèses.**

Avec cette approximation des p-valeurs, les puissances des tests pour $\mathcal{H}_0^{(m,m')}$, avec un niveau de confiance $\alpha_{m,m'}$, pour tous $1 \leq m < m' \leq M$,

$$\mathbb{P}\left(\hat{p}^*_{m,m'} < \alpha_{m,m'} \,|\, \mathcal{H}_1^{(m,m')}\right) = \mathbb{P}\left(|\hat{\delta}_{m,m'}| > F^{-1}_{\mathcal{N}(0,\hat{\sigma}^{*2}_{m,m'})}\left(1 - \alpha_{m,m'}/2\right) | \mathcal{H}_1^{(m,m')}\right) \tag{3.44}$$

peuvent également être approximées, par

$$\hat{\pi}(\hat{\delta}_{m,m'}, \hat{\sigma}^*_{m,m'}) := 1 - F_{\mathcal{HN}(\hat{\delta}_{m,m'},\hat{\sigma}^{*2}_{m,m'})}\left(F^{-1}_{\mathcal{N}(0,\hat{\sigma}^{*2}_{m,m'})}\left(1 - \alpha_{m,m'}/2\right)\right), \tag{3.45}$$

où $\mathrm{F}_{\mathcal{HN}(\hat{\mu}_{m,m'},\hat{\sigma}^{*2}_{m,m'})}$ est la fonction de répartition de la loi demi-normale de paramètre de position $\hat{\mu}_{m,m'}$ et de paramètre d'échelle $\hat{\sigma}^*_{m,m'}$. Ici, le paramètre de position $\hat{\mu}_{m,m'}$ est estimé par $\hat{\delta}_{m,m'}$ et non par la moyenne bootstrap $\hat{\mu}^*_{m,m'}$ des $\hat{\delta}^*_{m,m'}$ car la procédure bootstrap reproduit le biais d'estimation.

Estimation bootstrap non paramétrique

Une estimation bootstrap non paramétrique des p-valeurs des tests pour $\mathcal{H}_0^{(m,m')}$ est également possible, pour tous $1 \leq m < m' \leq M$, par

$$\hat{p}^{*(\mathrm{NP})}_{m,m'} := \frac{1}{R}\sum_{r=1}^{R} \mathbb{1}_{|\hat{\delta}_{m,m'}| \leq \left|\hat{\delta}^{*(r)}_{m,m'}\right|}, \tag{3.46}$$

où $\mathbb{1}_{\{x \in A\}} = 1$ si $x \in A$ et $\mathbb{1}_{\{x \in A\}} = 1$ sinon pour tout ensemble A. Ces p-valeurs non paramétriques ne reposent sur aucun a priori sur la distribution de $\hat{\delta}_{m,m'}$ hormis la symétrie, en particulier l'approximation de la distribution des estimées $\hat{H}_m$ par une distribution gaussienne

n'est pas nécessaire. Cependant, ces estimées nécessitent un nombre R d'échantillons bootstrap suffisamment grand, comme étudié numériquement dans la section 3.5.8. De plus, contrairement à l'approche paramétrique, l'approche non paramétrique ne permet pas d'estimer les puissances des tests par paires par bootstrap.

3.5.3 Décisions des tests bootstrap

Les tests par paires construits précédemment aboutissent à la construction de $M(M-1)/2$ décisions, ne permettant pas simplement de partitionner le vecteur des M exposants d'autosimilarité $\underline{H} = (H_1, \ldots, H_M)$. Ces décisions sont tout de même utiles pour vérifier le bon comportement des tests. Elles permettront également de détecter des exposants d'autosimilarité H_m différents de tout autre exposant $H_{m'}$, avec $m' \neq m$, dans la stratégie de partitionnement décrite dans la section suivante.

On construit alors les décisions des tests par paires à partir des p-valeurs paramétriques $\hat{p}^*_{m,m'}$ suivant les procédures de correction introduites dans la section 3.4.4 pour un test à $M(M-1)/2$ hypothèses dans le cas présent. Étant donnée la double indexation des hypothèses $\mathcal{H}_0^{(m,m')}$, les définitions des décisions doivent être adaptées. Ainsi, les décisions issues de la correction de Bonferroni sont définies, pour tous $1 \leq m < m' \leq M$, par

$$\begin{cases} d_\alpha^{(m,m')} = 1 & (\mathcal{H}_0^{(m,m')} \text{ rejetée}) \qquad \text{si } \hat{p}^*_{m,m'} < \frac{\alpha}{M(M-1)/2}, \\ d_\alpha^{(m,m')} = 0 & (\mathcal{H}_0^{(m,m')} \text{ non rejetée}) \qquad \text{sinon,} \end{cases} \tag{3.47}$$

où α est le niveau de confiance du test, et les décisions corrigées par les procédures de Benjamini-Hochberg et Benjamini-Yekutieli sont calculées comme suit :

$$\begin{cases} d_\alpha^{b(\pi(k))} = 1 & (\mathcal{H}_0^{b(\pi(k))} \text{ rejetée}) \qquad \text{si } k \leq \underset{j \in \{1,\ldots,M(M-1)/2\}}{\arg\max} j \mathbb{1}_{\hat{p}^*_{b(\tau(j))} < \frac{\alpha}{M(M-1)/2} c(j)}, \\ d_\alpha^{b(\pi(k))} = 1 & (\mathcal{H}_0^{b(\pi(k))} \text{ non rejetée}) \qquad \text{sinon,} \end{cases} \tag{3.48}$$

pour tout $k \in \{1, \ldots, M(M-1)/2\}$, où τ est la permutation qui ordonne les p-valeurs, $\hat{p}^*_{b(\tau(1))} < \ldots < \hat{p}^*_{b(\tau(M(M-1)/2))}$ avec b une bijection qui associe un couple d'entiers (m, m') tel que $1 \leq m < m' \leq M$ à toute valeur $k \in \{1, \ldots, M(M-1)/2\}$, et la fonction c, définie selon l'équation (3.35), dépend de la procédure de correction choisie (Benjamini-Hochberg ou Benjamini-Yekutieli).

Enfin, l'hypothèse nulle $\mathcal{H}_0$ ($H_1 = \ldots = H_M$) est rejetée si au moins une hypothèse nulle $\mathcal{H}_0^{(m,m')}$ est rejetée pour une paire (m, m'). Ainsi, la décision du test pour $\mathcal{H}_0$ avec un niveau de confiance α est définie par

$$\begin{cases} d_\alpha = 1 & (\mathcal{H}_0 \text{ rejetée}) \qquad \text{si } \exists\, 1 \leq m < m' \leq M |\ d_\alpha^{(m,m')} = 1, \\ d_\alpha = 0 & (\mathcal{H}_0 \text{ non rejetée}) \qquad \text{si } \forall\, 1 \leq m < m' \leq M,\ d_\alpha^{(m,m')} = 0. \end{cases} \tag{3.49}$$

Pour les tests par paires non paramétriques, les décisions des tests pour $\mathcal{H}_0^{(m,m')}$, avec $1 \leq m < m' \leq M$, sont construites suivant les mêmes procédés à partir des p-valeurs non paramétriques $\hat{p}^{*(\mathrm{NP})}_{m,m'}$ et notées $d_\alpha^{(\mathrm{NP},(\mathrm{m},\mathrm{m}'))}$. La décision pour $\mathcal{H}_0$ qui s'ensuit est notée $d_\alpha^{(\mathrm{NP})}$.

3.5.4 Définition d'un graphe des exposants

Pour regrouper les exposants d'autosimilarité $H_1, \ldots, H_M$ de même valeur ensemble, on construit un graphe dont les nœuds représentent les exposants H_m et on cherche à pondérer les arrêtes de sorte que des exposants H_m et $H_{m'}$, avec $m \neq m'$, sont d'autant plus proches que la probabilité d'égalité $H_m = H_{m'}$ est forte.

À cette fin, est défini un graphe pondéré par le triplet $\mathcal{G} = (\mathcal{V}, \epsilon, S)$ où $\mathcal{V} = \{1, \ldots, M\}$ est l'ensemble des *nœuds* associés aux exposants d'autosimilarité $(H_1, \ldots, H_M)$, $\epsilon = \{(m, m') | m, m' = 1, \ldots, M\}$ est l'ensemble des *arrêtes* et les entrées $S_{m,m'}$ de la matrice S de taille $M \times M$ sont les poids des arrêtes. La matrice S est dénommée *matrice de similarité*. La *matrice des degrés* D est alors définie par

$$\forall m, m' \in \{1, \ldots, M\}, \quad \begin{cases} D_{m,m} = \sum_{k=1}^{M} S_{m,k}, \\ D_{m,m'} = 0 \quad \text{si } m \neq m'. \end{cases} \tag{3.50}$$

Plus une p-valeur $\hat{p}^*_{m,m'}$ est grande, plus l'égalité $H_m = H_{m'}$ est probable. En conséquence, les p-valeurs $\hat{p}^*_{m,m'}$ apparaissent comme une pondération adaptée du graphe des exposants d'autosimilarité. Cependant, si un exposant d'autosimilarité H_m est différent de tous les autres exposants $H_{m'}$, la normalisation du laplacien dans le méthode de partitionnement spectral exploitée mène à l'impossibilité d'identifier le nœud m comme une partition d'un seul élément. On propose donc la procédure de détection d'un nœud isolé suivante : un nœud m est isolé si l'égalité entre l'exposant H_m et tout autre exposant $H_{m'}$ est rejetée, c'est-à-dire $d_\alpha^{(m,m')} = 1$ pour tout $m' = 1, \ldots, M$ tel que $m' \neq m$. Cette stratégie est réalisée en définissant une matrice binaire T, rapportant la détection d'exposants d'autosimilarité uniques, qui s'écrit

$$\begin{cases} T_{m,m'} = 0 & \text{si } \prod_{k=m+1}^{M} d_\alpha^{(m,k)} \prod_{k=1}^{m-1} d_\alpha^{(k,m)} = 1, \\ T_{m,m'} = 1 & \text{sinon,} \end{cases} \tag{3.51}$$

pour tous $m, m' \in \{1, \ldots, M\}$. La matrice T peut être vue comme la matrice d'adjacence du graphe non pondéré où tous les exposants d'autosimilarité H_m sont connectés excepté les exposants uniques. Le niveau de confiance prédéfini est fixé à $\alpha = 0.05$ pour cette stratégie. La procédure de correction pour obtenir les décisions $d_\alpha^{(m,m')}$, $1 \leq m < m' \leq M$, utilisée ici est celle de Benjamini-Hochberg, en raison de résultats numériques comparant les différentes procédures détaillés dans la section 3.5.8.3.

Finalement, la matrice de similarité S est définie par, pour tous $m, m' \in \{1, \ldots, M\}$,

$$\begin{cases} S_{m,m'} = S_{m',m} = \hat{p}^*_{m,m'} T_{m,m'} & \text{si } m < m', \\ S_{m,m} = 0. \end{cases} \tag{3.52}$$

Avec cette définition, l'arrête entre des exposants H_m et H'_m, avec $m' > m$, est pondérée par 0 si l'un des deux exposants est considéré unique et par la p-valeur $\hat{p}^*_{m,m'}$ sinon.

3.5.5 Partitionnement spectral

Le partitionnement spectral repose sur la définition du laplacien du graphe dont les propriétés sont intéressantes (voir Filippone et collab. (2008)). Dans ce travail, on considère le

laplacien de la marche aléatoire $\mathcal{L}_{rw}$ associé à une matrice de similarité S, défini par

$$\begin{cases} (\mathcal{L}_{rw})_{m,m'} = 1 & \text{si } m = m' \text{ et } D_{m,m} \neq 0, \\ (\mathcal{L}_{rw})_{m,m'} = -\frac{S_{m,m'}}{D_{m,m}} & \text{si } m \neq m' \text{ et } D_{m,m} \neq 0, \\ (\mathcal{L}_{rw})_{m,m'} = 0 & \text{sinon,} \end{cases} \tag{3.53}$$

pour tous $m, m' \in \{1, \ldots, M\}$, avec D la matrice des degrés associée à S.

Notons $\varphi_1 \leq \ldots \leq \varphi_M$ les valeurs propres ordonnées de $\mathcal{L}_{rw}$. Si un graphe est constitué de N_C composantes connexes alors le laplacien $\mathcal{L}_{rw}$ est une matrice ayant exactement N_C valeurs propres nulles, i.e. $\varphi_1 = \ldots = \varphi_{N_C} = 0$. En pratique, l'estimation du nombre de partitions N_C repose alors sur l'estimation du nombre de valeurs propres proches de 0. Cette estimation est communément réalisée par seuillage des plus petites valeurs propres au moyen du *maximum eigengap* (Azran et Ghahramani, 2006), c'est-à-dire la plus grande différence entre valeurs propres successives,

$$\forall k \in \{1, \ldots, M-1\}, \quad e_k := \varphi_{k+1} - \varphi_k. \tag{3.54}$$

Ainsi, le nombre de partitions N_C est estimé comme le nombre de valeurs propres proches de 0 au sens des différences entre valeurs propres successives,

$$\hat{N}_C = \underset{k \in \{1,\ldots,M-1\}}{\arg\max} \; e_k. \tag{3.55}$$

Finalement, le partitionnement des nœuds $m \in \{1, \ldots, M\}$ en $\hat{N}_C$ partitions est effectué à travers le partitionnement des lignes de la matrice $V = (\mathbf{v}_1, \ldots, \mathbf{v}_{\hat{N}_C})$ des vecteurs propres du laplacien $\mathcal{L}_{rw}$ associés aux plus petites valeurs propres $\varphi_1, \ldots, \varphi_{\hat{N}_C}$. En pratique, le partitionnement des lignes de V est effectué par l'algorithme des k-moyennes (Steinhaus, 1957). La stratégie de partitionnement spectral est illustrée par la figure 3.30.

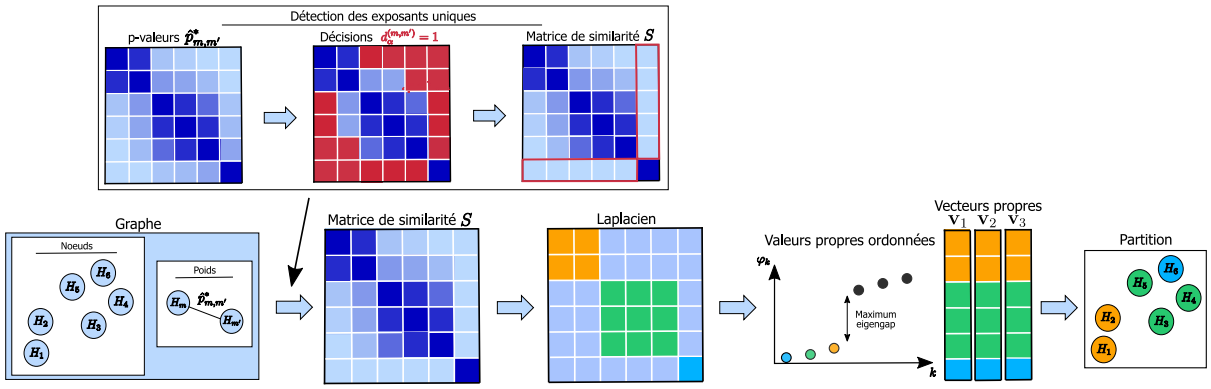


Figure 3.30 – **Illustration de la méthode de partitionnement du graphe.** Le graphe est originellement pondéré par les p-valeurs des tests $\hat{p}^*_{m,m'}$, puis les exposants d'autosimilarité uniques sont détectés à l'aide des décisions $d_\alpha^{(m,m')}$, aboutissant à la construction de la matrice de similarité S. Une procédure de partitionnement spectral est ensuite appliquée à S. À noter que, contrairement à ce qui est illustré pour faciliter la compréhension, les coefficients diagonaux de la matrice de similarité S valent 0.

3.5.6 Matrice de similarité PageRank

La matrice de similarité S précédemment introduite ne peut-être contrôlée par un paramètre. Suivant des travaux initiés par Esteban Bautista Ruiz durant sa thèse (Bautista Ruiz, 2019),

on propose de construire une nouvelle matrice de similarité S_η, paramétrée par une valeur réelle η, à l'aide du vecteur PageRank (Brin et Page, 1998). Pour un vecteur $\mathbf{y}$ de taille M, le vecteur Page Rank $\hat{\mathbf{f}}_\mathbf{y} \in \mathbb{R}^M$ du graphe $\mathcal{G}(\mathcal{V}, \epsilon, S)$ est défini comme la solution de

$$\arg\min_{\mathbf{f}} \quad \mathbf{f}^T D^{-1} \mathcal{L}_{rw} D^{-1} \mathbf{f} + \eta(\mathbf{f}-\mathbf{y})^T D^{-1}(\mathbf{f}-\mathbf{y}), \tag{3.56}$$

avec $\eta > 0$ un paramètre à valeur réelle. Le premier terme, dit de *régularisation*, correspond à une distance quadratique entre les entrées de $\mathbf{f}$ pondérée par les poids du graphe. Le second terme, dit d'*attache aux données*, correspond à une distance quadratique entre les vecteurs $\mathbf{f}$ et $\mathbf{y}$. Le paramètre η est ainsi un paramètre de régularisation spatiale : plus η est grand, moins les entrées du vecteur $\mathbf{f}$ dépendent de la structure du graphe. Cette solution s'écrit explicitement,

$$\hat{\mathbf{f}}_\mathbf{y} = \eta(\mathcal{L}_{rw}^T + \eta \mathbb{I})^{-1}\mathbf{y}. \tag{3.57}$$

Pour construire la matrice de similarité paramétrée S_η, appelée *matrice PageRank*, on associe à chaque nœud $m \in \{1, \dots, M\}$ un vecteur PageRank comme suit :

$$\hat{\mathbf{f}}_\eta^{(m)} := \eta(\mathcal{L}_{rw}^T + \eta \mathbb{I})^{-1}\mathbf{e}^{(m)}, \tag{3.58}$$

où $\mathbf{e}^{(m)}$ est un vecteur tel que $\mathbf{e}_k^{(m)} = 1$ si $m = k$ et $\mathbf{e}_k^{(m)} = 0$ sinon, pour $k = 1, \dots, M$. Autrement dit, le vecteur $\hat{\mathbf{f}}_\eta^{(m)}$ est la solution du problème (3.56) pour le vecteur $\mathbf{y} = \mathbf{e}^{(m)}$, correspondant à l'identifiant du nœud m. Ainsi, l'entrée $\hat{\mathbf{f}}_\eta^{(m)}$ quantifie la proximité du nœud m aux autres nœuds du graphe. Lorsque η est très grand, le vecteur $\hat{\mathbf{f}}_\eta^{(m)} \approx \mathbf{e}^{(m)}$ indique que le nœud m est très éloigné de tous les autres, tandis que pour $\eta \approx 0$, toutes les entrées de $\hat{\mathbf{f}}_\eta^{(m)}$ sont quasiment nulles.

Finalement, on définit la matrice PageRank paramétrée par η comme suit :

$$S_\eta = \mathcal{D}^{-\frac{1}{2}} \, \mathcal{F}_\eta \, \mathcal{D}^{-\frac{1}{2}} - \mathbb{I}, \tag{3.59}$$

où les entrées de $\mathcal{F}_\eta$ sont définies, pour tout $m, m' \in \{1, \dots, M\}$, par

$$(\mathcal{F}_\eta)_{m,m'} = \min((\hat{\mathbf{f}}_\eta^{(m)})_{m'}, (\hat{\mathbf{f}}_\eta^{(m')})_m), \tag{3.60}$$

et

$$\mathcal{D} = \begin{pmatrix} (\mathcal{F}_\eta)_{1,1} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & (\mathcal{F}_\eta)_{M,M} \end{pmatrix}. \tag{3.61}$$

La procédure de partitionnement spectral décrite dans la section 3.5.5 est alors appliquée au graphe $(\mathcal{V}, \epsilon, S_\eta)$, c'est-à-dire le graphe dont les nœuds sont les exposants d'autosimilarité H_m et les arrêtes sont pondérées par les entrées de S_η. En pratique, pour éviter les erreurs numériques sur les degrés des nœuds isolés liées à l'inversion matricielle dans le vecteur PageRank, le partitionnement est effectué sur le graphe $\mathcal{G}_\eta = (\mathcal{V}, \epsilon, S_\eta \odot T)$, où $\odot$ dénote le produit de Hadamard (point par point).

3.5.7 Synthèse des notations et formules

Formulaire récapitulatif sur les tests gaussiens

Les différentes quantités relatives aux tests des hypothèses $\mathcal{H}_0^{(m,m')}$ $(H_m = H_{m'})$ pour $1 \leq m < m' \leq M$ sont rappelées dans le tableau suivant.

Définition	Estimateur bootstrap	
	paramétrique	non paramétrique
Statistiques des tests par paires		
$\hat{\delta}_{m,m'}$	$\hat{\delta}^{*(r)}_{m,m'},\ r = 1, \ldots, R$	$\hat{\delta}^{*(r)}_{m,m'},\ r = 1, \ldots, R$
Paramètres des statistiques		
$(\hat{\mu}_{m,m'}, \hat{\sigma}_{m,m'})$	Aucun	$(\hat{\mu}^*_{m,m'}, \hat{\sigma}^*_{m,m'})$
P-valeurs des tests par paires		
$\tilde{p}_{m,m'}$	$\hat{p}^{*(\mathrm{NP})}_{m,m'} = \frac{1}{R}\sum_{r=1}^{R} \mathbb{1}_{\lvert\hat{\delta}_{m,m'}\rvert \leq \lvert\hat{\delta}^{*(r)}_{m,m'}\rvert}$	$\hat{p}^*_{m,m'} = 2\left(1 - F_{\mathcal{N}(0,\hat{\sigma}^{*2}_{m,m'})}(\lvert\hat{\delta}_{m,m'}\rvert)\right)$
Puissance des tests par paires		
$\hat{\pi}(\tilde{\mu}_m, \tilde{\sigma}_m)$	Aucun	$\hat{\pi}(\hat{\delta}_{m,m'}, \hat{\sigma}^*_{m,m'})$ (Eq. (3.45))
Décision de rejet des tests par paires		
Inconnue	$d_\alpha^{(\mathrm{NP},(m,m'))}$	$d_\alpha^{(m,m')}$
Décision de rejet du test pour $\mathcal{H}_0$		
Inconnue	$d_\alpha^{(\mathrm{NP})}$	d_α

Formulaire récapitulatif sur le partitionnement spectral

Les différentes quantités relatives au partitionnement spectral sont rappelées dans le tableau suivant.

Matrice de similarité	S (Eq. (3.52)) ou S_η (Eq. (3.59))
Degré du nœud m	$D_{m,m} = \sum_{k=1}^{M} S_{m,k}$
Laplacien du graphe	$\mathcal{L}_{\mathrm{rw}}$ (Eq. (3.53))
Valeurs propres du laplacien	$\varphi_1, \ldots, \varphi_M$
Eigengap	$e_k = \varphi_{k+1} - \varphi_k$
Nombre de partitions estimé	$\hat{N}_{\mathrm{C}} = \underset{k \in \{1,\ldots,M-1\}}{\arg\max}\ e_k$

3.5.8 Performances des estimateurs, du test et du partitionnement

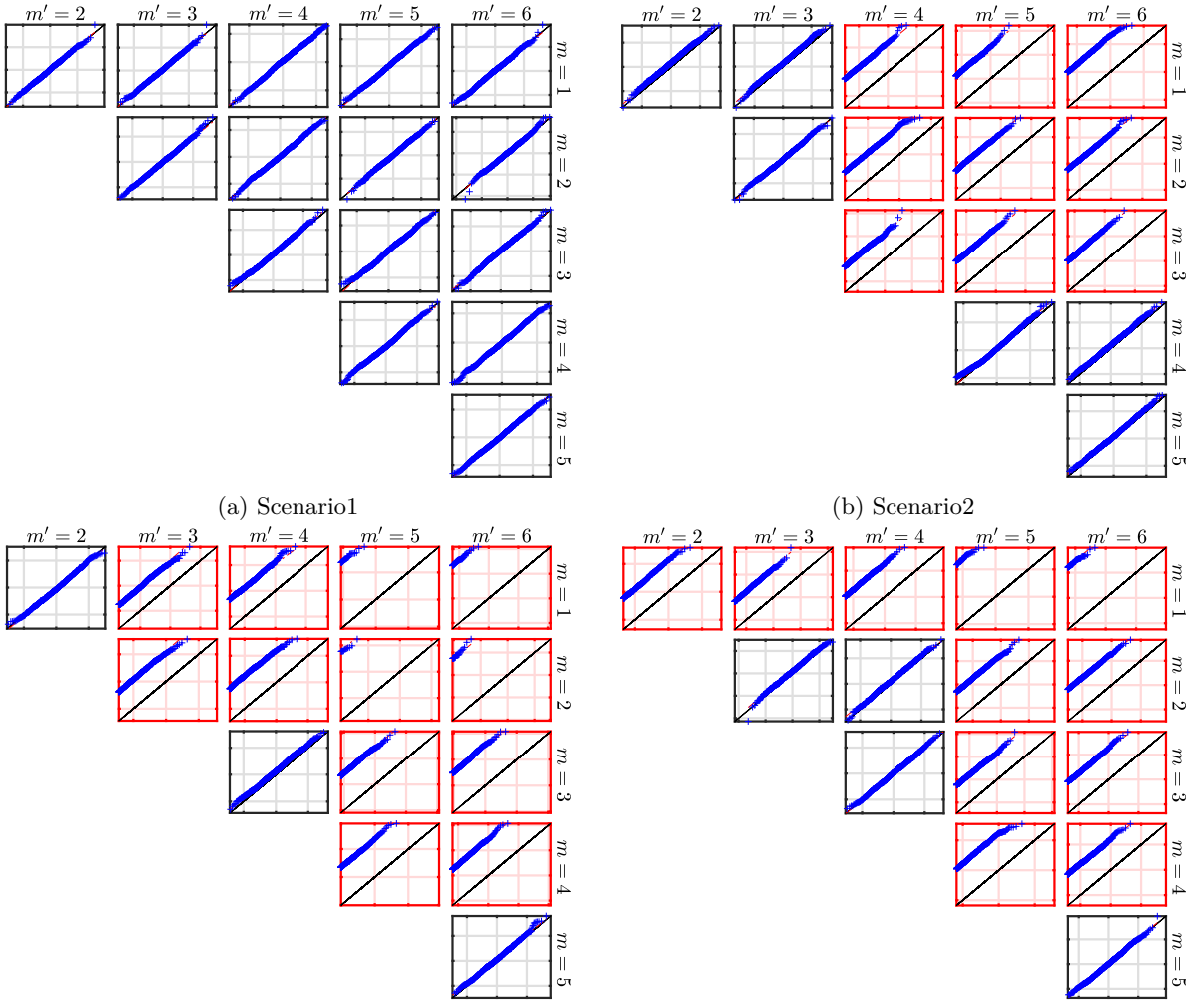
Les performances de la stratégie de partitionnement présentée ci-dessus sont évaluées sur des M-mBf synthétiques au travers de simulations de Monte Carlo dont la configuration est présentée dans la section 3.4.7.

3.5.8.1 Propriétés des statistiques

Les tests par paires reposent sur la normalité asymptotique des statistiques $\hat{\delta}_{m,m'}$ et l'écart de la moyenne $\hat{\mu}_{m,m'}$ de $\hat{\delta}_{m,m'}$ à 0 sous une hypothèse alternative $\mathcal{H}_1^{(m,m')}$, pour $1 \leq m < m' \leq M$. Pour vérifier ce comportement, la figure 3.31 montre les diagrammes quantile-quantile des distributions empiriques des statistiques de test $\hat{\delta}_{m,m'}$ contre des distributions gaussiennes centrées d'écart-types $\hat{\sigma}_{m,m'}$ estimés par Monte Carlo pour $1 \leq m < m' \leq M$, les différents scénarios et une taille d'échantillon $N = 2^{16}$. Comme prévu, sous les hypothèses nulles $\mathcal{H}_0^{(m,m')}$, les distributions des statistiques $\hat{\delta}_{m,m'}$ sont très bien approximées par des lois normales centrées dans les différents scénarios, ce qui valide l'a priori $\hat{\mu}_{m,m'} \approx 0$. De plus, sous des hypothèses alternatives $\mathcal{H}_1^{(m,m')}$, quel que soit le scénario, les distributions des statistiques $\hat{\delta}_{m,m'}$ sont bien approximées par des lois normales non centrées, et leurs moyennes s'écartent d'autant plus de 0 que les exposants d'autosimilarité H_m diffèrent (i.e. $|\hat{\mu}_{m,m'}| \gg 0$).

On peut cependant observer que, dans le scénario 3, l'écart de $\hat{\mu}_{m,m'}$ à 0 n'est pas le même pour différentes hypothèses alternatives $\mathcal{H}_1^{(m,m')}$ correspondant à un même écart $|H_{m'} - H_m|$. En effet, cet écart est plus important pour les couples (m, m') tels que $m \in \{3, 4\}$ et $m' \in \{5, 6\}$ que pour les couples (m, m') tels que $m \in \{1, 2\}$ et $m' \in \{3, 4\}$. Pourtant, pour ces différents couples, l'écart entre exposants d'autosimilarité est le même : $|H_{m'} - H_m| = 0.2$. Ceci est lié au fait que les estimées $\hat{H}_1, \ldots, \hat{H}_M$ ont des biais légèrement différents, fait illustré dans la section 2.5 par la figure 2.11.

Ces observations confirment la pertinence de tests construits à partir de ces statistiques. En effet, les a priori permettant le calcul des p-valeurs $\hat{p}^*_{m,m'}$ des tests par paires, à savoir l'égalité $\hat{\mu}_{m,m'} = 0$, et la gaussianité de $\hat{\delta}_{m,m'}$, sont approximativement valides. De plus, le comportement de $\hat{\delta}_{m,m'}$ diffère significativement entre l'hypothèse nulle $\mathcal{H}_0^{(m,m')}$ et une hypothèse alternative $\mathcal{H}_1^{(m,m')}$.



(a) Scenario1

(b) Scenario2

(c) Scenario3

(d) Scenario4

FIGURE 3.31 – **Distributions des statistiques $\hat{\delta}_{m,m'}$.** Diagrammes quantile-quantile de $\hat{\delta}_{m,m'}$ au travers des réalisations de Monte Carlo contre une loi normale centrée de paramètre $\hat{\sigma}_{m,m'}$ estimé par Monte Carlo pour $1 \leq m < m' \leq M$, les différents scénarios et une taille d'échantillon $N = 2^{16}$. Les cases rouges indiquent les écarts à $\mathcal{H}_0^{(m,m')}$. *Les distributions des statistiques $\hat{\delta}_{m,m'}$ sont bien approximées par des distributions normales, dont les moyennes s'écartent de* 0 *sous des hypothèses alternatives $\mathcal{H}_1^{(m,m')}$.*

3.5.8.2 Comportement des tests par paires

Tests paramétriques

Pour vérifier le comportement des tests par paires paramétriques, il faut s'assurer que les statistiques bootstrap $\hat{\delta}^*_{m,m'}$ suivent bien des lois gaussiennes et que leurs écart-types $\hat{\sigma}^*_{m,m'}$ approximent correctement les paramètres des tests $\hat{\sigma}_{m,m'}$, pour $1 \leq m < m' \leq M$.

En premier lieu, la figure 3.32 montre les diagrammes quantile-quantile des distributions empiriques des statistiques bootstrap $\hat{\delta}^*_{m,m'}$ pour une réalisation de Monte Carlo choisie arbitrairement contre des distributions gaussiennes centrées d'écart-types $\hat{\sigma}^*_{m,m'}$ pour $1 \leq m < m' \leq M$, les différents scénarios et une taille d'échantillon $N = 2^{16}$. Quelle que soit l'hypothèse ($\mathcal{H}_0^{(m,m')}$ vraie et $\mathcal{H}_0^{(m,m')}$ non vraie), les statistiques $\hat{\delta}^*_{m,m'}$ sont très bien approximées par des lois normales centrées (i.e. $\hat{\mu}^*_{m,m'} \approx 0$), comme attendu puisque la procédure bootstrap vise à reproduire le comportement de $\hat{\delta}_{m,m'}$ sous l'hypothèse nulle $\mathcal{H}_0^{(m,m')}$.

Ces résultats assurent la bonne construction de l'estimateur bootstrap $\hat{\sigma}^*_{m,m'}$ des paramètres de test $\hat{\sigma}_{m,m'}$. Pour quantifier la qualité de l'estimation, le tableau 3.7 rapporte les valeurs des paramètres de test $\hat{\sigma}_{m,m'}$ estimés par Monte Carlo et de leurs estimées bootstrap $\hat{\sigma}^*_{m,m'}$ moyennées sur les réalisations de Monte Carlo pour $1 \leq m < m' \leq M$ et les différents scénarios. On remarque tout d'abord que les paramètres $\hat{\sigma}_{m,m'}$ diffèrent peu entre les scénarios, ce qui signifie que seuls les moyennes $\hat{\mu}_{m,m'}$ diffèrent entre l'hypothèse nulle $\mathcal{H}_0^{(m,m')}$ et une hypothèse alternative $\mathcal{H}_1^{(m,m')}$ De plus, dans tous les cas ($\mathcal{H}_0^{(m,m')}$ vraie et $\mathcal{H}_0^{(m,m')}$ non vraie), les des paramètres bootstrap $\hat{\sigma}^*_{m,m'}$ sont de très bonnes estimées des $\hat{\sigma}_{m,m'}$.

En complément, la figure 3.33 montre les diagrammes quantile-quantile des p-valeurs $\hat{p}^*_{m,m'}$ des tests par paires estimées par Monte Carlo contre une loi uniforme pour les différents scénarios. Comme attendu théoriquement, sous les hypothèses nulles $\mathcal{H}_0^{(m,m')}$, les distributions des p-valeurs $\hat{p}^*_{m,m'}$ sont bien approximées par une loi uniforme, ce qui suggère le comportement adéquat de l'estimateur bootstrap $\hat{\sigma}^*_{m,m'}$ (censé suivre une loi de Student). Sous des hypothèses alternatives $\mathcal{H}_1^{(m,m')}$, les distributions des p-valeurs $\hat{p}^*_{m,m'}$ sont très éloignées de distributions uniformes, d'autant plus pour des grands écarts entre exposants d'autosimilarité $|H_{m'} - H_m|$.

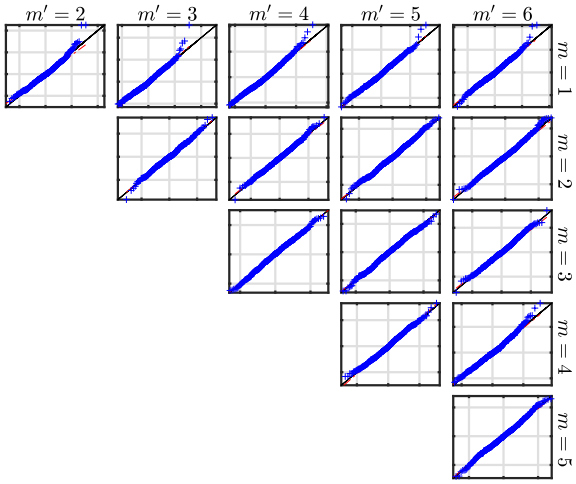
Pour quantifier les résultats précédents, le tableau 3.8 rapporte les puissances empiriques des tests par paires obtenues par moyenne des décisions non corrigées de rejet $\hat{p}^*_{m,m'} < \alpha_{m,m'}$ sur les réalisations de Monte Carlo et leurs estimées bootstrap $\hat{\pi}(\hat{\delta}_{m,m'}, \hat{\sigma}^*_{m,m'})$ (cf. Eq. (3.45)) pour un niveau de confiance prédéfini $\alpha_{m,m'} = 0.05$. Ces résultats montrent que :

(*i*) sous l'hypothèse nulle $\mathcal{H}_0^{(m,m')}$, comme prévu, le taux de fausse détection est bien reconstruit, et ce dans tous les scénarios ;

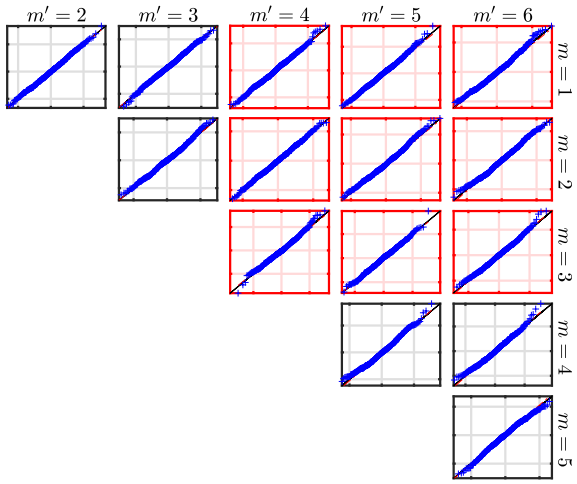
(*ii*) sous les hypothèses alternatives $\mathcal{H}_1^{(m,m')}$, les puissances de tests sont satisfaisantes, surtout pour des exposants H_m et $H_{m'}$ de valeurs éloignées ;

(*iii*) et les estimées bootstrap $\hat{\pi}(\hat{\delta}_{m,m'}, \hat{\sigma}^*_{m,m'})$ sont de bonnes approximations des puissances empiriques des tests sous les hypothèses alternatives $\mathcal{H}_1^{(m,m')}$, mais les surestiment sous les hypothèses nulles $\mathcal{H}_0^{(m,m')}$.

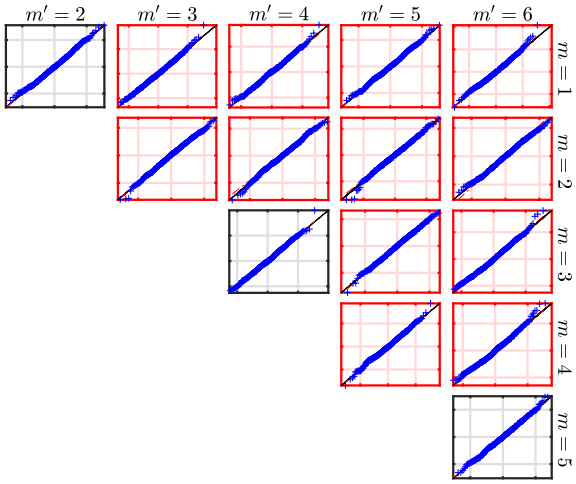
On observe également que, dans le scénario 1, la puissance empirique est plus importante pour les couples (m, m') tels que $m \in \{3, 4\}$ et $m' \in \{5, 6\}$ que pour les couples (m, m') tels que $m \in \{1, 2\}$ et $m' \in \{3, 4\}$, ce qui est en accord avec l'écart de $\hat{\mu}_{m,m'}$ à 0 observée dans la figure 3.31, figure qui relate le comportement de la distribution de $\hat{\delta}_{m,m'}$.



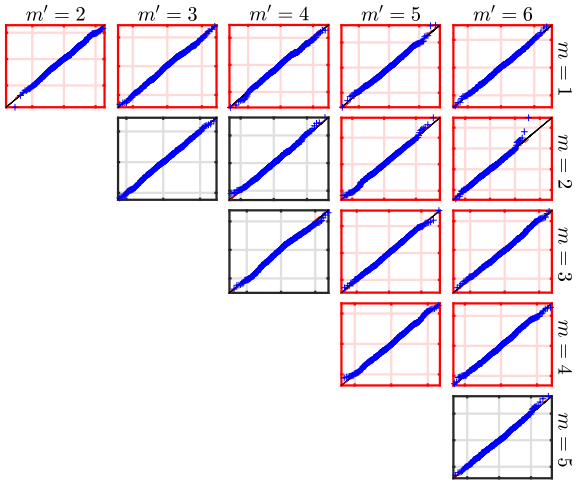
(a) Scenario1



(b) Scenario2



(c) Scenario3



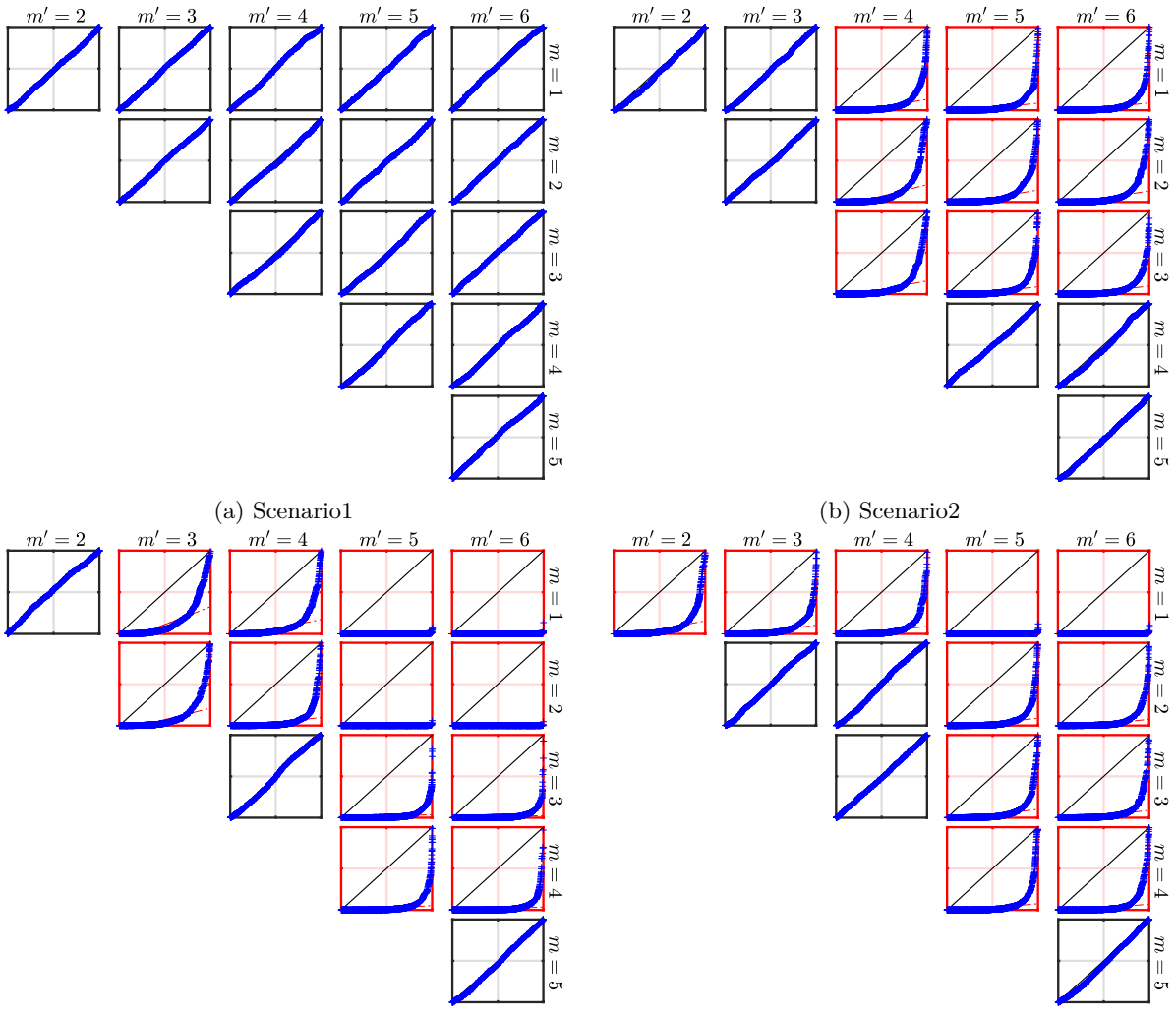
(d) Scenario4

FIGURE 3.32 – **Distributions des $\hat{\delta}^*_{m,m'}$.** Diagrammes quantile-quantile de $\hat{\delta}^*_{m,m'}$ pour une réalisation de Monte Carlo contre une loi normale centrée d'écart-type $\hat{\sigma}^*_{m,m'}$ pour $1 \leq m < m' \leq M$, les différents scénarios et une taille d'échantillon $N = 2^{16}$. Les cases rouges indiquent les écarts à $\mathcal{H}_0^{(m,m')}$. *Les distributions des statistiques bootstrap $\hat{\delta}^*_{m,m'}$ sont bien approximées par des distributions normales centrées.*

TABLEAU 3.7 – **Estimation des paramètres des tests $\hat{\sigma}_{m,m'}$.** Estimées bootstrap $\hat{\sigma}^*_{m,m'}$ des paramètres $\hat{\sigma}_{m,m'}$ (moyennes de Monte Carlo ± intervalles de confiance à 95 %) pour $1 \leq m < m' \leq M$, les différents scénarios et une taille d'échantillon $N = 2^{16}$. Les cases rouges correspondent aux écarts par rapport à $\mathcal{H}_0^{(m,m')}$.

$\times 10^2$		2	3	4	5	6	m'/m		2	3	4	5	6	m'/m
$\hat{\sigma}_{m,m'}$	Scenario1	7.81	7.63	7.59	7.35	7.48	1	Scenario2	7.86	8.01	7.66	7.05	7.74	1
$\hat{\sigma}^*_{m,m'}$		7.59 ±0.74	7.51 ±0.76	7.42 ±0.77	7.50 ±0.77	7.65 ±0.80			7.77 ±0.79	7.76 ±0.86	7.56 ±0.83	7.33 ±0.81	7.69 ±0.84	
$\hat{\sigma}_{m,m'}$			6.83	6.84	6.86	7.00	2			7.51	7.44	7.02	7.53	2
$\hat{\sigma}^*_{m,m'}$			6.54 ±0.57	6.70 ±0.62	6.82 ±0.65	6.98 ±0.71				7.17 ±0.72	7.13 ±0.72	6.84 ±0.73	7.18 ±0.73	
$\hat{\sigma}_{m,m'}$				5.82	6.36	6.43	3				7.10	6.70	7.06	3
$\hat{\sigma}^*_{m,m'}$				5.63 ±0.53	6.19 ±0.55	6.38 ±0.62					6.86 ±0.71	6.49 ±0.70	6.87 ±0.75	
$\hat{\sigma}_{m,m'}$					5.73	6.18	4					5.27	7.13	4
$\hat{\sigma}^*_{m,m'}$					5.47 ±0.49	5.89 ±0.58						5.51 ±0.60	6.61 ±0.69	
$\hat{\sigma}_{m,m'}$						5.45	5						6.26	5
$\hat{\sigma}^*_{m,m'}$						5.39 ±0.56							5.87 ±0.60	

$\times 10^2$		2	3	4	5	6	m'/m		2	3	4	5	6	m'/m
$\hat{\sigma}_{m,m'}$	Scenario3	6.32	7.95	7.56	7.39	7.27	1	Scenario4	8.22	7.82	7.78	7.99	7.65	1
$\hat{\sigma}^*_{m,m'}$		6.83 ±0.77	7.82 ±0.86	7.73 ±0.83	7.65 ±0.82	7.69 ±0.82			7.87 ±0.87	7.80 ±0.86	7.61 ±0.87	7.70 ±0.89	7.70 ±0.88	
$\hat{\sigma}_{m,m'}$			7.36	7.20	7.08	6.87	2			7.66	7.88	7.64	7.44	2
$\hat{\sigma}^*_{m,m'}$			7.11 ±0.79	6.85 ±0.73	6.79 ±0.77	6.65 ±0.74				7.42 ±0.80	7.26 ±0.78	7.33 ±0.80	7.32 ±0.81	
$\hat{\sigma}_{m,m'}$				7.12	7.38	7.23	3				6.94	7.37	7.15	3
$\hat{\sigma}^*_{m,m'}$				6.99 ±0.75	7.18 ±0.84	7.11 ±0.83					6.68 ±0.70	7.06 ±0.76	7.07 ±0.77	
$\hat{\sigma}_{m,m'}$					7.20	7.03	4					7.23	7.31	4
$\hat{\sigma}^*_{m,m'}$					6.78 ±0.75	6.67 ±0.79						6.60 ±0.73	6.62 ±0.76	
$\hat{\sigma}_{m,m'}$						6.94	5						7.00	5
$\hat{\sigma}^*_{m,m'}$						6.39 ±0.70							6.43 ±0.70	



(a) Scenario1

(b) Scenario2

(c) Scenario3

(d) Scenario4

FIGURE 3.33 – **Distributions des p-valeurs paramétriques $\hat{p}^*_{m,m'}$.** Diagrammes quantile-quantile de $\hat{p}^*_{m,m'}$ au travers des réalisations de Monte Carlo contre une loi uniforme pour $1 \leq m < m' \leq M$, les différents scénarios et une taille d'échantillon $N = 2^{16}$. Les cases rouges indiquent les écarts à $\mathcal{H}_0^{(m,m')}$. *Les distributions des p-valeurs bootstrap paramétriques $\hat{p}^*_{m,m'}$ sont bien approximées par des distributions uniformes sous les hypothèses nulles $\mathcal{H}_0^{(m,m')}$, et s'en écartent sous des hypothèses alternatives $\mathcal{H}_1^{(m,m')}$.*

TABLEAU 3.8 – **Estimation des puissances des tests paramétriques.** Estimées bootstrap $\hat{\pi}(\hat{\delta}_{m,m'}, \hat{\sigma}^*_{m,m'})$ (moyenne de Monte Carlo et intervalle interquartile) des puissances empiriques, obtenues comme la proportion de rejet $\hat{p}^*_{m,m'} < \alpha_{m,m'}$ sur les réalisations de Monte Carlo pour $1 \leq m < m' \leq M$, les différents scénarios, un niveau de confiance $\alpha_{m,m'} = 0.05$ et une taille d'échantillon $N = 2^{16}$. Les cases rouges correspondent aux écarts à $\mathcal{H}_0^{(m,m')}$.

		2	3	4	5	6	m'/m		2	3	4	5	6	m'/m
$\langle \mathbb{1}_{\hat{p}^*_{m,m'}<0.05} \rangle$	Scenario1	0.06	0.06	0.06	0.04	0.05	1	Scenario2	0.07	0.07	0.66	0.73	0.72	1
$\hat{\pi}(\hat{\delta}_{m,m'}, \hat{\sigma}^*_{m,m'})$		0.18 0.07 − 0.22	0.17 0.06 − 0.22	0.18 0.06 − 0.23	0.17 0.06 − 0.21	0.16 0.06 − 0.21			0.18 0.06 − 0.24	0.18 0.06 − 0.24	0.62 0.39 − 0.87	0.67 0.48 − 0.91	0.67 0.47 − 0.91	
$\langle \mathbb{1}_{\hat{p}^*_{m,m'}<0.05} \rangle$			0.06	0.06	0.06	0.05	2			0.06	0.59	0.66	0.67	2
$\hat{\pi}(\hat{\delta}_{m,m'}, \hat{\sigma}^*_{m,m'})$			0.18 0.06 − 0.22	0.17 0.06 − 0.22	0.17 0.06 − 0.21	0.17 0.06 − 0.21				0.18 0.06 − 0.23	0.57 0.32 − 0.84	0.62 0.38 − 0.88	0.62 0.40 − 0.88	
$\langle \mathbb{1}_{\hat{p}^*_{m,m'}<0.05} \rangle$				0.06	0.06	0.06	3				0.63	0.72	0.71	3
$\hat{\pi}(\hat{\delta}_{m,m'}, \hat{\sigma}^*_{m,m'})$				0.18 0.06 − 0.23	0.17 0.06 − 0.22	0.17 0.06 − 0.22					0.60 0.36 − 0.87	0.66 0.47 − 0.91	0.66 0.45 − 0.91	
$\langle \mathbb{1}_{\hat{p}^*_{m,m'}<0.05} \rangle$					0.06	0.07	4					0.04	0.08	4
$\hat{\pi}(\hat{\delta}_{m,m'}, \hat{\sigma}^*_{m,m'})$					0.18 0.06 − 0.21	0.18 0.06 − 0.25						0.16 0.01 − 0.18	0.19 0.02 − 0.26	
$\langle \mathbb{1}_{\hat{p}^*_{m,m'}<0.05} \rangle$						0.06	5						0.09	5
$\hat{\pi}(\hat{\delta}_{m,m'}, \hat{\sigma}^*_{m,m'})$						0.17 0.06 − 0.21							0.19 0.06 − 0.23	

		2	3	4	5	6	m'/m		2	3	4	5	6	m'/m
$\langle \mathbb{1}_{\hat{p}^*_{m,m'}<0.05} \rangle$	Scenario3	0.04	0.48	0.60	1.00	1.00	1	Scenario4	0.63	0.70	0.72	1.00	1.00	1
$\hat{\pi}(\hat{\delta}_{m,m'}, \hat{\sigma}^*_{m,m'})$		0.16 0.06 − 0.19	0.50 0.23 − 0.77	0.57 0.35 − 0.83	0.97 0.98 − 1.00	0.98 0.98 − 1.00			0.60 0.36 − 0.86	0.65 0.44 − 0.91	0.68 0.47 − 0.92	0.98 0.99 − 1.00	0.98 0.99 − 1.00	
$\langle \mathbb{1}_{\hat{p}^*_{m,m'}<0.05} \rangle$			0.59	0.71	1.00	1.00	2			0.07	0.08	0.72	0.76	2
$\hat{\pi}(\hat{\delta}_{m,m'}, \hat{\sigma}^*_{m,m'})$			0.57 0.32 − 0.84	0.67 0.45 − 0.92	0.99 1.00 − 1.00	0.99 1.00 − 1.00				0.18 0.06 − 0.22	0.19 0.06 − 0.26	0.68 0.47 − 0.92	0.69 0.51 − 0.92	
$\langle \mathbb{1}_{\hat{p}^*_{m,m'}<0.05} \rangle$				0.07	0.82	0.86	3				0.07	0.71	0.72	3
$\hat{\pi}(\hat{\delta}_{m,m'}, \hat{\sigma}^*_{m,m'})$				0.18 0.06 − 0.25	0.75 0.61 − 0.97	0.79 0.67 − 0.97					0.18 0.06 − 0.22	0.66 0.44 − 0.91	0.66 0.46 − 0.91	
$\langle \mathbb{1}_{\hat{p}^*_{m,m'}<0.05} \rangle$					0.79	0.84	4					0.74	0.76	4
$\hat{\pi}(\hat{\delta}_{m,m'}, \hat{\sigma}^*_{m,m'})$					0.73 0.56 − 0.95	0.77 0.62 − 0.97						0.69 0.50 − 0.93	0.69 0.51 − 0.94	
$\langle \mathbb{1}_{\hat{p}^*_{m,m'}<0.05} \rangle$						0.08	5						0.07	5
$\hat{\pi}(\hat{\delta}_{m,m'}, \hat{\sigma}^*_{m,m'})$						0.19 0.01 − 0.22							0.19 0.06 − 0.24	

Tests non paramétriques

Les tests par paires non paramétriques reposent sur le fait que la distributions de la statistique bootstrap $\hat{\delta}^*_{m,m'}$ approxime bien la distribution de la statistique $\hat{\delta}_{m,m'}$ sous l'hypothèse nulle $\mathcal{H}_0^{(m,m')}$ et s'en écarte sous une hypothèse alternative $\mathcal{H}_1^{(m,m')}$, pour $1 \leq m < m' \leq M$.

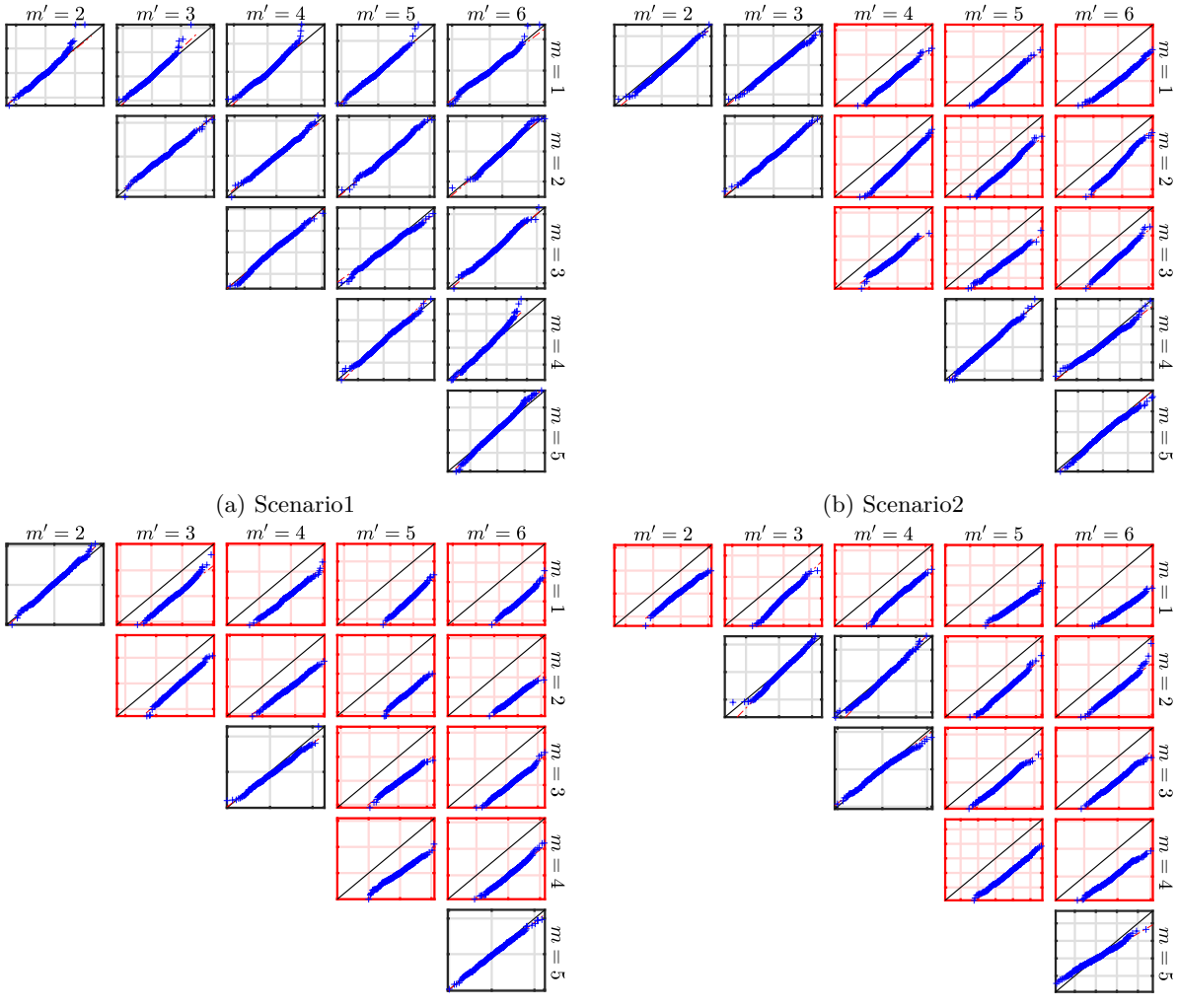
Tout d'abord, la figure 3.34 montre les diagrammes quantile-quantile des distributions empiriques des statistiques bootstrap $\hat{\delta}^*_{m,m'}$ pour une réalisation de Monte Carlo contre les distributions des $\hat{\delta}_{m,m'}$ estimées par Monte Carlo pour $1 \leq m < m' \leq M$, les différents scénarios et une taille d'échantillon $N = 2^{16}$. Comme attendu, les statistiques $\hat{\delta}_{m,m'}$ et les statistiques bootstrap $\hat{\delta}^*_{m,m'}$ ont des distributions très proches sous les hypothèses nulles $\mathcal{H}_0^{(m,m')}$ et très différentes sous des hypothèses alternatives $\mathcal{H}_1^{(m,m')}$. Plus précisément, sous $\mathcal{H}_1^{(m,m')}$, les statistiques $\hat{\delta}_{m,m'}$ et $\hat{\delta}^*_{m,m'}$ suivent approximativement des lois identiques de même variance mais de moyennes différentes, et ces moyennes diffèrent d'autant plus que l'écart entre les exposants d'autosimilarité associés $|H_{m'} - H_m|$ est grand.

En supplément, la figure 3.35 montre les diagrammes quantile-quantile des p-valeurs non paramétriques $\hat{p}^{*(\mathrm{NP})}_{m,m'}$ des tests par paires estimées par Monte Carlo contre une loi uniforme pour $1 \leq m < m' \leq M$ et les différents scénarios. Comme prévu, sous les hypothèses nulles $\mathcal{H}_0^{(m,m')}$, l'approximation des distributions des p-valeurs $\hat{p}^{*(\mathrm{NP})}_{m,m'}$ par une loi uniforme est satisfaisante. Sous des hypothèses alternatives $\mathcal{H}_1^{(m,m')}$, les distributions des p-valeurs $\hat{p}^{*(\mathrm{NP})}_{m,m'}$ ne suivent plus une loi uniforme, d'autant moins que les exposants d'autosimilarité associés H_m et $H_{m'}$ sont différents. Ce comportement assure la reproduction des hypothèses nulles $\mathcal{H}_0^{(m,m')}$ et une puissance de test suffisante.

Pour quantifier les résultats précédents, le tableau 3.9 rapporte les puissances empiriques des tests par paires non paramétriques obtenues comme moyennes des décisions non corrigées de rejet $\hat{p}^{*(\mathrm{NP})}_{m,m'} < \alpha_{m,m'}$ au travers des réalisations de Monte Carlo pour les différents scénarios et un niveau de confiance prédéfini $\alpha_{m,m'} = 0.05$. Le niveau de confiance ciblé $\alpha_{m,m'}$ est bien reconstruit sous $\mathcal{H}_0^{(m,m')}$, et la puissance est grande sous $\mathcal{H}_1^{(m,m')}$ et augmente avec l'écart $|H_{m'} - H_m|$. Ces puissances des test sont similaires aux puissances des tests par paires paramétriques rapportées dans le tableau 3.8.

Conclusions

Les approches paramétriques et non paramétriques permettent toutes deux une bonne reconstruction des hypothèses nulles par paires $\mathcal{H}_0^{(m,m')}$ et atteignent des puissances similaires. La procédure paramétrique permet en plus d'obtenir de bonnes estimations bootstrap des puissances des tests.



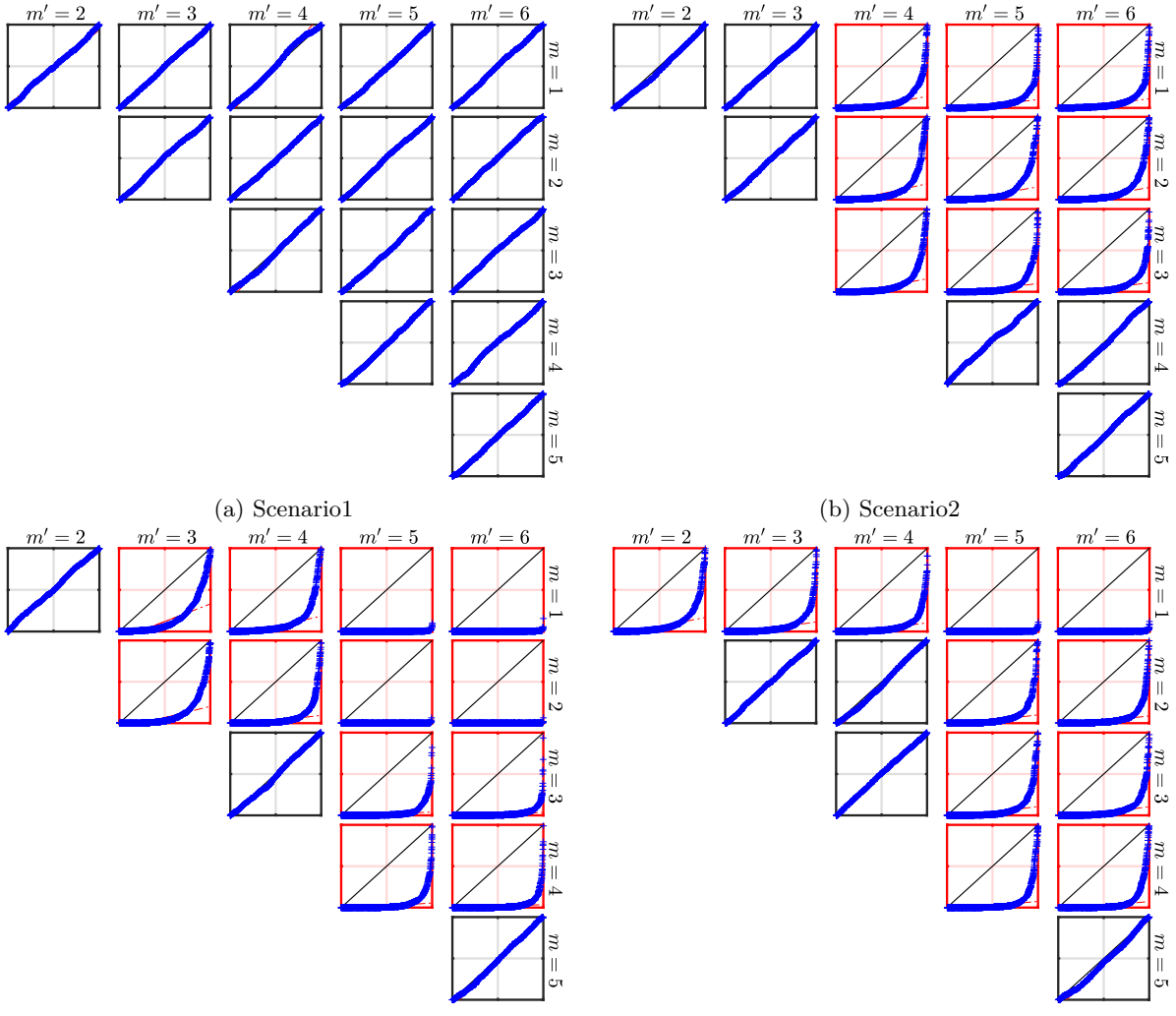
(a) Scenario1

(b) Scenario2

(c) Scenario3

(d) Scenario4

FIGURE 3.34 – **Reproduction des distributions des $\hat{\delta}_{m,m'}$ par $\hat{\delta}^*_{m,m'}$.** Diagrammes quantile-quantile de $\hat{\delta}^*_{m,m'}$ pour une réalisation de Monte Carlo contre $\hat{\delta}_{m,m'}$ au travers des réalisations de Mont Carlo pour $1 \leq m < m' \leq M$, les différents scénarios et une taille d'échantillon $N = 2^{16}$. Les cases rouges indiquent les écarts à $\mathcal{H}_0^{(m,m')}$. *Les distributions des statistiques bootstrap $\hat{\delta}^*_{m,m'}$ et des statistiques $\hat{\delta}_{m,m'}$ sont similaires, de même variance, et de moyennes similaires sous les hypothèses nulles $\mathcal{H}_0^{(m,m')}$ et différentes sous des hypothèses alternatives $\mathcal{H}_1^{(m,m')}$.*



(a) Scenario1

(b) Scenario2

(c) Scenario3

(d) Scenario4

FIGURE 3.35 – **Distributions des p-valeurs non paramétriques $\hat{p}^{*(\mathrm{NP})}_{m,m'}$.** Diagrammes quantile-quantile de $\hat{p}^{*(\mathrm{NP})}_{m,m'}$ au travers des réalisations de Monte Carlo contre une loi uniforme pour $1 \leq m < m' \leq M$, les différents scénarios et une taille d'échantillon $N = 2^{16}$. Les cases rouges indiquent les écarts à $\mathcal{H}_0^{(m,m')}$. *Les distributions des p-valeurs bootstrap non paramétriques $\hat{p}^{*(\mathrm{NP})}_{m,m'}$ sont bien approximées par des distributions uniformes sous les hypothèses nulles $\mathcal{H}_0^{(m,m')}$, et s'en écartent sous des hypothèses alternatives $\mathcal{H}_1^{(m,m')}$.*

TABLEAU 3.9 – **Puissance des tests non paramétriques.** Puissances empiriques obtenues comme moyennes des décisions non corrigées des tests $\hat{p}^{*(\mathrm{NP})}_{m,m'} < \alpha_{m,m'}$ sur les réalisations de Monte Carlo pour $1 \leq m < m' \leq M$, les différents scénarios, un niveau de confiance $\alpha_{m,m'} = 0.05$ et une taille d'échantillon $N = 2^{16}$. Les cases rouges correspondent aux écarts par rapport à $\mathcal{H}_0^{(m,m')}$.

Scenario1	2	3	4	5	6	m'/m
	0.07	0.06	0.06	0.05	0.06	1
		0.07	0.07	0.05	0.06	2
			0.06	0.06	0.06	3
				0.07	0.07	4
					0.06	5

Scenario2	2	3	4	5	6	m'/m
	0.07	0.07	0.65	0.69	0.68	1
		0.07	0.58	0.63	0.65	2
			0.63	0.70	0.71	3
				0.04	0.08	4
					0.09	5

Scenario3	2	3	4	5	6	m'/m
	0.05	0.47	0.56	0.99	1.00	1
		0.61	0.71	1.00	1.00	2
			0.07	0.81	0.85	3
				0.79	0.84	4
					0.08	5

Scenario4	2	3	4	5	6	m'/m
	0.62	0.69	0.70	1.00	1.00	1
		0.07	0.08	0.71	0.72	2
			0.07	0.70	0.69	3
				0.75	0.76	4
					0.07	5

3.5.8.3 Détection des exposants d'autosimilarité uniques

Pour détecter les exposants d'autosimilarité H_m différents de tous les autres exposants $H_{m'}$, avec $m \neq m'$, la matrice de similarité du graphe fait appel aux décisions corrigées des tests par paires, ce qui implique le recours à une des procédures de correction introduites dans la section 3.5.3.

Le comportement des différentes procédures de correction est évalué à travers les décisions paramétriques d_α et non paramétriques $d_\alpha^{(\mathrm{NP})}$ pour l'hypothèse nulle $\mathcal{H}_0$ ($H_1 = \ldots = H_M$). La figure 3.36 rapporte les décisions paramétriques d_α et non paramétriques $d_\alpha^{(\mathrm{NP})}$ moyennées sur les réalisations de Monte Carlo sous l'hypothèse nulle $\mathcal{H}_0$ (Scenario1) pour les différentes procédures de correction en fonction du niveau de confiance prédéfini α. Le test paramétrique garantit la bonne reconstruction du niveau de confiance prédéfini α avec les procédures de Bonferroni et Benjamini-Hochberg, mais est trop conservatif avec la procédure de Benjamini-Yekutieli. En revanche, le test non paramétrique ne reconstruit correctement le niveau de confiance prédéfini α que s'il est suffisamment grand en raison de la nature discrète de la distribution de la statistique bootstrap $\hat{\delta}^*_{m,m'}$. Par ailleurs, le test paramétrique permet une meilleure reconstruction de α que le test non paramétrique. Dans la suite, on ne considère donc que les tests par paires paramétriques.

Ensuite, pour évaluer la méthode de détection d'exposants d'autosimilarité uniques, le tableau 3.10 rapporte le nombre de réalisations de Monte Carlo pour lesquelles un exposant H_m a été détecté comme unique selon la définition de la matrice T (cf. Eq. (3.51)) pour différentes procédures de correction et différentes tailles d'échantillon N sous le scénario 4 (où seul H_1 est unique). L'exposant unique est bien mieux détecté avec la procédure de Benjamini-Hochberg que celle de Bonferroni, notamment pour les tailles d'échantillon $N = 2^{16}$ et $N = 2^{17}$. De plus, l'exposant unique est d'autant mieux détecté que la taille d'échantillon N est grande, ce qui est conforme à l'évolution des puissances de test avec N rapportée précédemment par le tableau 3.8. En revanche, le nombre d'exposants détectés à tort comme uniques augmente aussi avec N, notamment pour la procédure de Benjamini-Hochberg.

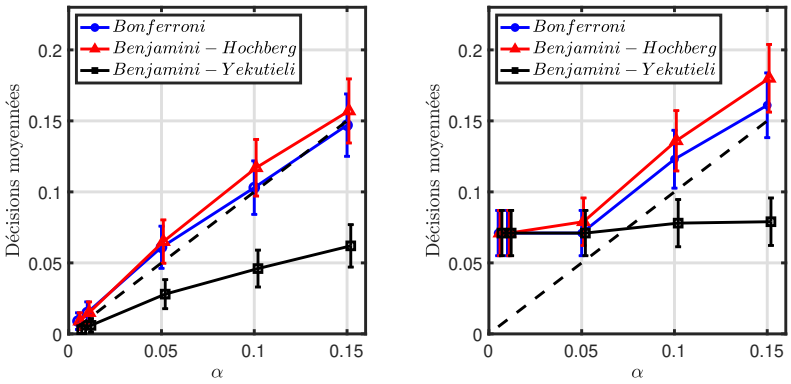


FIGURE 3.36 – **Reproduction de $\mathcal{H}_0$.** Décisions des tests (à gauche) paramétriques d_α et (à droite) non paramétriques $d_\alpha^{(\mathrm{NP})}$ (voir Section 3.5.2) moyennées sur les réalisations de Monte Carlo en fonction du niveau de confiance α pour différentes procédures de correction sous $\mathcal{H}_0$ (Scenario1) avec une taille d'échantillon $N = 2^{16}$. Les lignes noires pointillées indiquent les valeurs théoriques attendues, c'est-à-dire la première bissectrice. *Les procédures de Bonferroni et Benjamini-Hochberg assurent une bonne reconstruction du niveau de confiance ciblé α. L'approche paramétrique permet une reconstruction des petites valeurs de α contrairement à l'approche non paramétrique.*

TABLEAU 3.10 – **Détection des exposants d'autosimilarité uniques.** Nombre de réalisations pour lesquelles un exposant H_m est considéré comme isolé (i.e. le nœud m est de degré nul, $D_{m,m} = 0$) parmi les $N_{\mathrm{MC}} = 1000$ réalisations de Monte Carlo sous le scénario 4 en suivant la stratégie de partitionnement définie en section 3.5.5 à partir des p-valeurs $\hat{p}^*_{m,m'}$ pour différentes procédures de correction avec $\alpha = 0.05$ et différentes tailles d'échantillon N. Les cases rouges indiquent les exposants d'autosimilarité H_m uniques.

N		$m=1$	$m=2$	$m=3$	$m=4$	$m=5$	$m=6$
2^{16}	Bonferroni	128	0	0	0	2	3
	Benjamini-Hochberg	367	1	0	0	13	20
2^{17}	Bonferroni	583	0	0	0	5	2
	Benjamini-Hochberg	869	9	4	7	42	35
2^{18}	Bonferroni	973	0	0	0	5	2
	Benjamini-Hochberg	999	11	14	15	51	49

La procédure de détection des exposants d'autosimilarité uniques présente un comportement adéquat au partitionnement, ce qui assure la bonne construction de la matrice de similarité S du graphe. Étant donnés les résultats obtenus, c'est la procédure de Benjamini-Hochberg à partir des p-valeurs paramétriques $\hat{p}^*_{m,m'}$ qui sera utilisée pour obtenir les décisions $d_\alpha^{(m,m')}$, avec un niveau de confiance $\alpha = 0.05$, dans le reste de ce chapitre.

3.5.8.4 Performances du partitionnement

La stratégie de partitionnement de $\underline{H}$ consiste à exploiter les caractéristiques du laplacien de la matrice de similarité S construite à partir des p-valeurs paramétriques $\hat{p}^*_{m,m'}$ et des décisions $d_\alpha^{(m,m')}$ des tests associés pour $1 \leq m < m' \leq M$. Les résultats précédents assurent la bonne construction des $\hat{p}^*_{m,m'}$ et $d_\alpha^{(m,m')}$, et donc de la matrice S. On s'intéresse désormais à la pertinence de la matrice de similarité S pour effectuer un partitionnement spectral, stratégie introduite dans la section 3.5.5.

En guise d'illustration de la stratégie de partitionnement, la figure 3.37 donne un exemple de matrice de similarité S ainsi que les valeurs propres ordonnées $\varphi_1, \dots \varphi_M$ du laplacien $\mathcal{L}_{\text{rw}}$ du graphe associé sous le scénario 3, pour une réalisation de Monte Carlo et une taille d'échantillon $N = 2^{18}$. La matrice S permet de distinguer facilement les différentes partitions et le maximum eigengap $e_3 = \varphi_4 - \varphi_3$ apparaît ici comme une mesure pertinente de la proximité des valeurs propres $\varphi_1, \varphi_2, \varphi_3$ à 0.

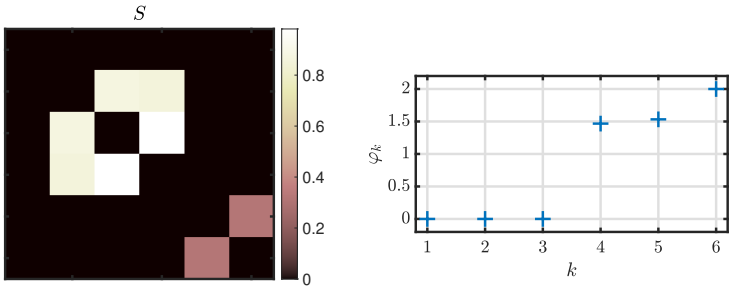
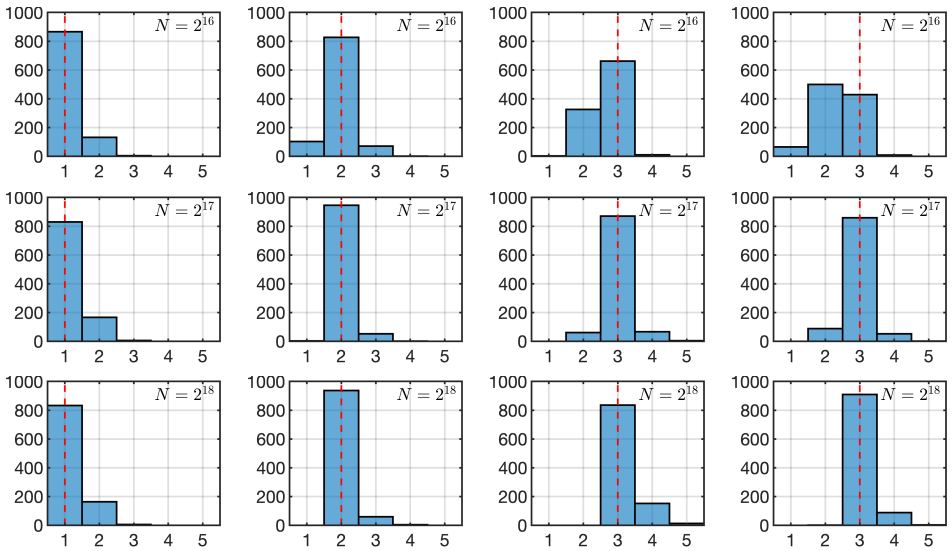


FIGURE 3.37 – **Exemple de matrice de similarité S.** Matrice de similarité S (à gauche) et valeurs propres $\varphi_1, \dots, \varphi_M$ du laplacien du graphe associé (à droite) pour une réalisation de Monte Carlo et une taille d'échantillon $N = 2^{18}$ sous le scénario 3 (composé de 3 partitions à 2 éléments).

Dans cette stratégie, le nombre de partitions est estimé en identifiant le maximum eigengap (Eq. (3.55)), mesure dont la pertinence peut dépendre de la matrice de similarité S choisie. La figure 3.38 présente les histogrammes du nombre estimé de partitions $\hat{N}_{\text{C}}$ pour les différents scénarios et différentes tailles d'échantillons $N = 2^{16}, 2^{17}, 2^{18}$. Pour les différents scénarios et les différentes tailles d'échantillon N, la procédure proposée détecte dans la majorité des cas le nombre correct de partitions, excepté pour le scénario 4 avec $N = 2^{16}$ en raison de la mauvaise détection de l'exposant unique H_1 à faible taille d'échantillon. Ceci démontre que la stratégie de partitionnement conduit à des résultats satisfaisants concernant le nombre de groupes détectés de même valeur H_m, et ce d'autant que la taille d'échantillon N est grande.



(a) Scenario1 (b) Scenario2 (c) Scenario3 (d) Scenario4

FIGURE 3.38 – **Estimation du nombre de partitions.** Histogrammes des nombres de partitions estimés $\hat{N}_{\text{C}}$ suivant la stratégie de partitionnement définie en section 3.5.5 pour (a) 1 partition, (b) 2 partitions et (c-d) 3 partitions pour différentes tailles d'échantillon N. Les lignes rouges pointillées indiquent le nombre exact de partitions.

Ensuite, les partitions sont définies à partir du partitionnement des $\hat{N}_{\mathrm{C}}$ premiers vecteurs propres (ordonnés selon les valeurs propres $\varphi_1 \leq \ldots \leq \varphi_M$) du laplacien $\mathcal{L}_{rw}$ de la matrice de similarité S par l'algorithme des k-moyennes. Le tableau 3.11 rapporte les performances de cette stratégie de partitionnement en termes d'ARI et NMI (cf. Section 3.4.7.6 et Annexe C) pour les différents scénarios et différentes tailles d'échantillon N. La stratégie a des performances satisfaisantes même pour une petite taille d'échantillon $N = 2^{16}$, avec des NMI et ARI allant jusqu'à 86% et 87%, respectivement. Par ailleurs, pour les scénarios 2, 3 et 4 constitués de plusieurs partitions, les performances augmentent avec la taille d'échantillon N. En revanche, pour le scénario 1 constitué d'une seule partition, les performances varient peu avec la taille d'échantillon N.

TABLEAU 3.11 – **Performances de la stratégie de partitionnement.** ARI et NMI (moyenne de Monte Carlo ± intervalle de confiance à 95%) de la stratégie de partitionnement définie dans la section 3.5.5 pour les différents scénarios et différentes tailles d'échantillon N.

N		Scenario1	Scenario2	Scenario3	Scenario4
2^{16}	NMI	n/a	0.75 ± 0.02	0.86 ± 0.01	0.78 ± 0.01
	ARI	0.87 ± 0.02	0.69 ± 0.02	0.71 ± 0.02	0.60 ± 0.02
2^{17}	NMI	n/a	0.95 ± 0.01	0.95 ± 0.01	0.95 ± 0.01
	ARI	0.83 ± 0.02	0.93 ± 0.01	0.89 ± 0.01	0.89 ± 0.01
2^{18}	NMI	n/a	0.99 ± 0.00	0.98 ± 0.00	0.98 ± 0.00
	ARI	0.83 ± 0.02	0.98 ± 0.01	0.95 ± 0.01	0.95 ± 0.00

De façon surprenante, pour une taille d'échantillon $N = 2^{16}$, les performances de partitionnement sont équivalentes entre les scénarios 2, 3 et 4, et ce bien que la détection du nombre de partitions atteigne les meilleures performances dans le scénario 2 et les moins bonnes dans le scénario 4, d'après la figure 3.38. Ceci est dû à une mauvaise affectation des nœuds dans différentes partitions en raison de faibles puissances de tests rapportées dans le tableau 3.8.

Ces résultats valident la stratégie de partitionnement spectral détaillée en section 3.5.5.

3.5.8.5 Performances du partitionnement PageRank

La stratégie de partitionnement du vecteur $\underline{H}$ peut être contrôlée par un paramètre en effectuant le partitionnement spectral sur le graphe $\mathcal{G}_\eta = (\mathcal{V}, \epsilon, S_\eta \odot T)$, pondéré par la matrice de similarité PageRank S_η, donnée par l'équation (3.59), comme décrit dans la section 3.5.6. Les résultats précédents assurent la bonne construction de la matrice de similarité S à partir de laquelle est construite la matrice S_η.

Pour illustrer la stratégie de partitionnement PageRank, la figure 3.39 donne, pour une même réalisation de Monte Carlo, des exemples de matrice de similarité PageRank S_η ainsi que les valeurs propres ordonnées $\varphi_1 \leq \ldots \leq \varphi_M$ du laplacien $\mathcal{L}_{\mathrm{rw}}$ du graphe associé pour plusieurs paramètres η sous le scénario 3 (i.e. 3 partitions) et une taille d'échantillon $N = 2^{18}$. On observe que pour certaines valeurs du paramètre η (grandes en l'occurrence), les trois premières valeurs propres $\varphi_1, \varphi_2, \varphi_3$ du laplacien s'approchent davantage de 0, mais aussi que le maximum eigengap $e_3 = \varphi_4 - \varphi_3$ est plus grand. Ce comportement se répercute sur la forme de la matrice de similarité S_η : les poids $(S_\eta)_{m,m'}$ associés à l'inégalité $H_m \neq H_{m'}$ sont beaucoup plus faibles relativement aux poids $(S_\eta)_{m,m'}$ associés à l'égalité $H_m = H_{m'}$. Ainsi, les paramètre η contrôle la parcimonie de la matrice de similarité PageRank S_η

Tout d'abord, on s'intéresse à l'estimation $\hat{N}_{\mathrm{C}}$ du nombre de partitions N_{C} à l'aide du maximum eigengap (cf. Eq. (3.55)) sur le graphe $\mathcal{G}_\eta$. La figure 3.40 rapporte les nombres de partitions estimés $\hat{N}_{\mathrm{C}}$ (moyennées sur les réalisations de Monte Carlo) dans le cadre du partitionnement PageRank (en rouge) en fonction du paramètre η pour différentes tailles d'échantillon N et les différents scénarios. Pour comparaison, les nombres de partitions estimés à partir de S sont donnés en bleu. On constate que le nombre de partitions estimé $\hat{N}_{\mathrm{C}}$ augmente avec le paramètre η. Dans le cas d'une seule partition (Scenario1), le nombre de partitions est correctement estimé pour des valeurs de η proches de 0 et surestimé pour des valeurs de η plus grandes. Lorsque le vecteur des exposants d'autosimilarité $\underline{H}$ est constitué de plusieurs partitions (en l'occurrence les scénarios 2, 3 et 4), le partitionnement sous-estime fortement le nombre de partitions N_{C} lorsque η est proche de 0 et une taille d'échantillon faible ($N = 2^{16}$ en l'occurrence). De surcroît, des valeurs de η suffisamment grandes permettent au partitionnement d'atteindre des performances d'estimation du nombre de partitions N_{C} similaires, voire supérieures, à celles du partitionnement à partir de la matrice de similarité S.

Enfin, le partitionnement du graphe $\mathcal{G}_\eta$ en $\hat{N}_{\mathrm{C}}$ partitions se fait à travers le partitionnement en k-moyennes des vecteurs propres associés aux $\hat{N}_{\mathrm{C}}$ plus petites valeurs propres $\varphi_1, \ldots, \varphi_{\hat{N}_{\mathrm{C}}}$, méthode détaillée dans la section 3.5.6. Les performances de cette stratégie sont quantifiées dans la figure 3.41 pour le partitionnement PageRank (en rouge) en termes d'ARI et NMI en fonction du paramètre η pour les différents scénarios et différentes tailles d'échantillon N. Pour comparaison, les performances du partitionnement à partir de S sont données en bleu. Pour une partition unique (scénario 1), les performances décroissent lorsque le paramètre η croît, ce qui est en adéquation avec la surestimation du nombre de paritions observée précédemment. En conséquence, le paramètre η apparaît comme un paramètre de contrôle du taux de fausse alarmes associé à l'hypothèse nulle $\mathcal{H}_0$ (présence d'une unique partition). Pour les scénarios 2, 3 et 4, où le vecteur $\underline{H}$ est divisé en plusieurs partitions, aux tailles d'échantillon $N = 2^{16}, 2^{17}$, les performances sont très faibles pour de petites valeurs de η petites mais atteignent des performances satisfaisantes pour des valeurs de η plus grandes, pouvant même surpasser celles du partitionnement à partie de la matrice de similarité S. À la taille d'échantillon $N = 2^{18}$, les performances varient peu avec η et sont similaires à celles du partitionnement à partir de S.

En conclusion, la stratégie de partitionnement PageRank permet d'ajuster le taux de fausses alarmes associé à l'hypothèse nulle $\mathcal{H}_0$ comme le permet la méthode de partitionnement à partir de tests par paires des exposants d'autosimilarité ordonnés présentée dans la section 3.4. Qui plus est, le paramètre η peut être ajusté manuellement en fonction de la forme de la matrice de similarité PageRank S_η et du comportement des valeurs propres du laplacien associé. Cette méthode ne montre cependant pas de meilleures performances que la stratégie de partitionnement à partir de la matrice de similarité S dans des scénarios à plusieurs partitions pour un faible nombre de composantes $M = 6$.

3.5.9 Conclusions

La présente section conçoit une stratégie de dénombrement des exposants d'autosimilarité H_m à partir de $M(M-1)/2$ tests d'égalité par paires d'exposants exploités à l'aide d'une structure de graphe. Les exposants d'autosimilarité H_m, associés aux nœuds du graphe, sont estimés à partir d'une seule observation de taille finie de données multivariées suivant la procédure d'estimation présentée dans le chapitre 2. Les propriétés de l'estimateur permettent une connaissance satisfaisante des distributions des statistiques de test choisies, et les paramètres de ces statistiques sont estimés à partir de la procédure de ré-échantillonnage bootstrap par blocs présentée dans la section 3.2. Les arrêtes du graphe sont alors pondérés à l'aide des tests par paires ainsi

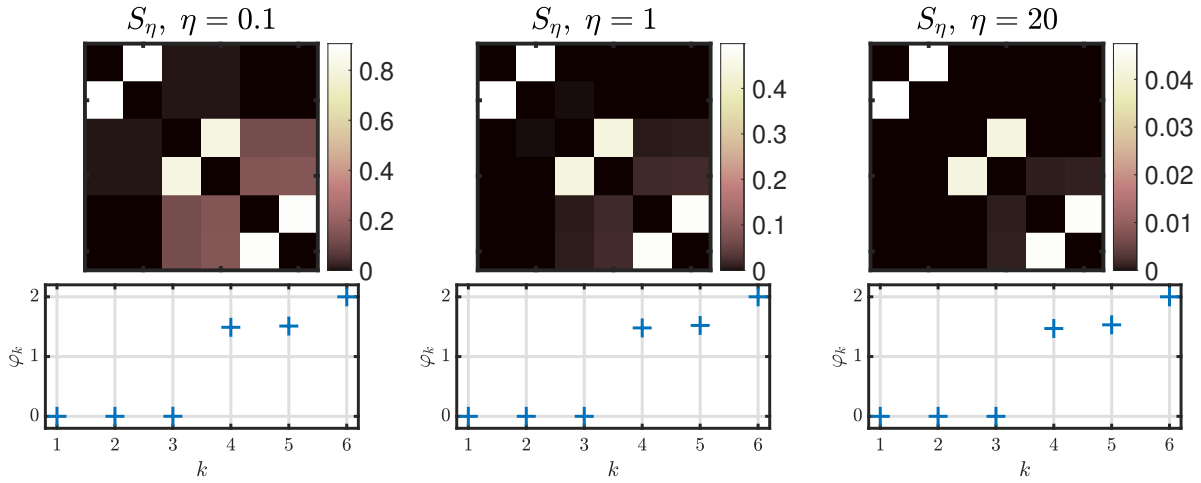
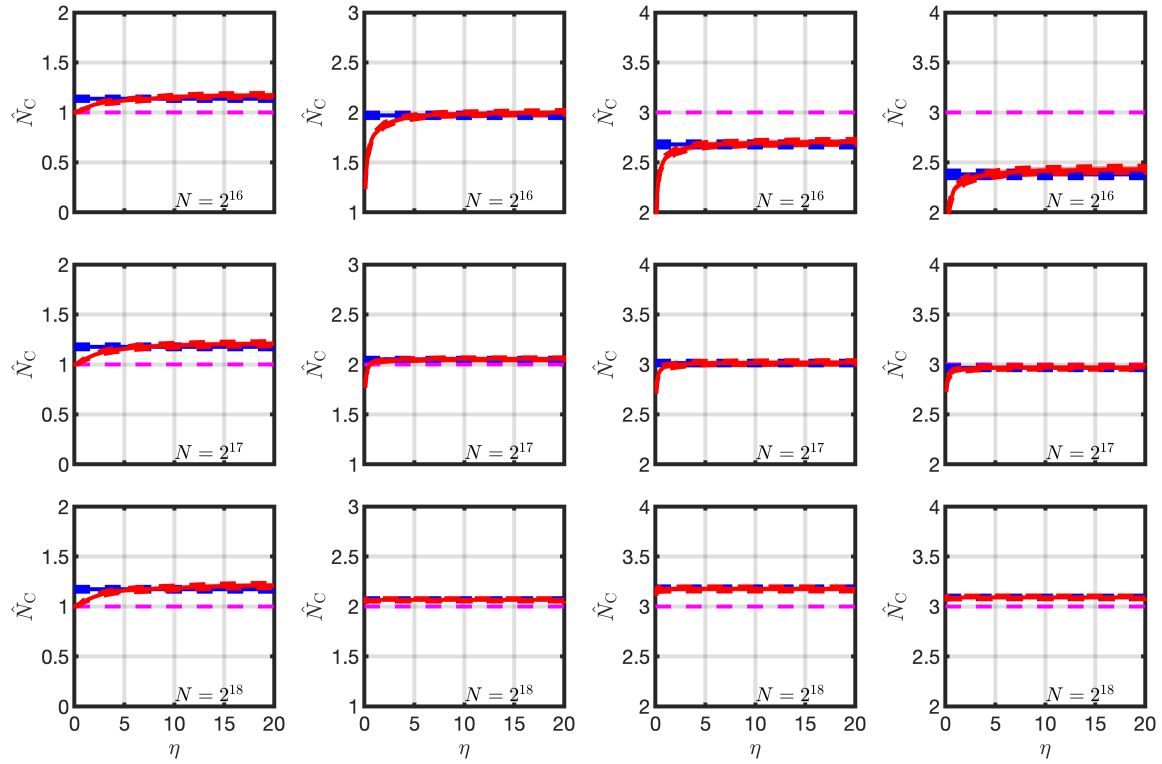
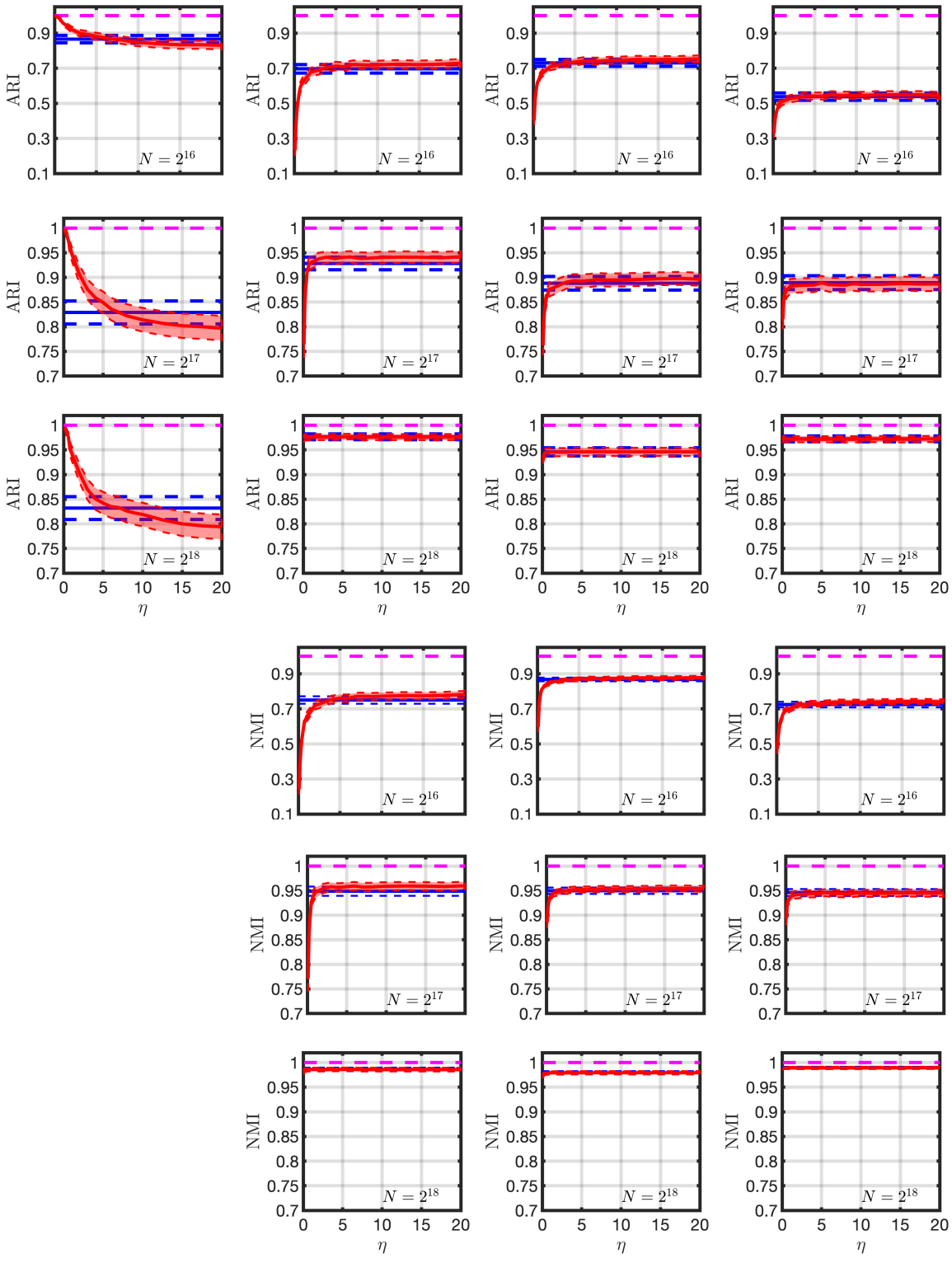


FIGURE 3.39 – **Exemples de matrices de similarité PageRank S_η.** Matrices de similarité PageRank S_η (en haut) et valeurs propres $\varphi_1, \ldots, \varphi_M$, des laplaciens associés (en bas) pour une même réalisation de Monte Carlo, différents paramètres η et une taille d'échantillon $N = 2^{18}$ sous le scénario 3 ($\underline{H} = (0.4, 0.4, 0.6, 0.6, 0.8, 0.8)$).



(a) Scenario1 (b) Scenario2 (c) Scenario3 (d) Scenario4

FIGURE 3.40 – **Estimation du nombre de partitions.** Nombres de partitions estimés $\hat{N}_C$ (moyennes Monte Carlo avec intervalle de confiance à 95%) par la stratégie de partitionnement effectué à partir de (en bleu) la matrice de similarité S et (en rouge) la matrice de similarité PageRank S_η en fonction du paramètre η pour les différents scénarios et (de haut en bas) différentes tailles d'échantillon N. Les lignes magenta pointillées indiquent le nombre exact de partitions.



(a) Scenario1 (b) Scenario2 (c) Scenario3 (d) Scenario4

FIGURE 3.41 – **Performances de la stratégie de partitionnement PageRank.** Performances en termes d'ARI et NMI (moyennes de Monte Carlo ± intervalle de confiance à 95%) de la stratégie de partitionnement définie dans la section 3.5.5 en utilisant (en bleu) la matrice de similarité S définie par l'équation (3.52) et (en rouge) la matrice de similarité PageRank S_η définie par l'équation (3.59) pour les différents scénarios et (de haut en bas) différentes tailles d'échantillon N. Les lignes magenta pointillées indiquent les valeurs idéales de performances, à savoir 1.

construits, pondération traduite par une matrice de similarité.

Deux matrices de similarités sont proposées, l'une paramétrée contrairement à l'autre. La première matrice est construite à partir de la combinaison des p-valeurs des tests et d'une procédure de correction pour un test à hypothèses multiples. La seconde matrice est construite à partir de cette première matrice et du vecteur PageRank, vecteur qui fait appel à un paramètre. Une procédure de regroupement spectral est appliquée à ces deux matrices de similarité et évaluée numériquement sur des simulations de Monte Carlo, réalisées sur des $M = 6$-mBf synthétiques, pour différents nombres de partitions et différentes tailles d'échantillon. Les résultats montrent que la procédure de tests bootstrap est correctement construite. De plus, les deux matrices de similarité permettent d'atteindre de bonnes performances de partitionnement. La stratégie de partitionnement PageRank a l'avantage de permettre de contrôler la détection d'une unique partition. Une étude et une comparaison des deux stratégies de partitionnement sur davantage de composantes, $M = 20$, est faite dans la section 3.6.

3.6 Comparaison des méthodes

Cette section vise à confronter les cinq différentes procédures présentées dans les trois sections précédentes, à savoir la méthode de détection de l'hypothèse nulle $\mathcal{H}_0$ ($H_1 = \ldots = H_M$) par un test du χ^2 (cf. Section 3.3), la stratégie de partitionnement du vecteur $\underline{H} = (H_1, \ldots, H_M)$ à partir de $M - 1$ tests par paires d'exposants ordonnés $H_m \leq H_{m+1}$ avec deux différentes méthodes d'estimation des paramètres des tests par bootstrap (cf. Section 3.4) et la stratégie de partitionnement du graphe d'exposants d'autosimilarité H_m pondéré de façon paramétrée ou non paramétrée (matrice de similarité PageRank) à partir des p-valeurs de $M(M-1)/2$ tests par paires (cf. Section 3.5).

3.6.1 Simulations de Monte Carlo

Les différentes procédures sont comparées grâce à des simulations de Monte Carlo effectuées à partir de $N_{\mathrm{MC}} = 1000$ réalisations indépendantes de M-mBf synthétiques (cf. Section 2.4) avec $M = 6$ et $M = 20$ composantes.

La description de la synthèse des $M = 6$-mBf et de la configuration des procédures d'estimation et de ré-échantillonnage bootstrap associés est donnée dans la section 3.4.7.

Les $M = 20$-mBf synthétiques sont de différentes tailles d'échantillon $N = 2^{17}, 2^{18}, 2^{19}$. Quatre scénarios sont considérés pour le vecteur des exposants d'autosimilarité $\underline{H}$:

(i) le Scenario1 correspond à $H_1 = \ldots = H_M = 0.8$ (1 partition) ;

(ii) le Scenario2 consiste en 2 partitions de taille 10 avec des valeurs égales à 0.6 et 0.8, telles que $\mathcal{H}_0^{(10,11)}$ n'est pas vraie ;

(iii) le Scenario3 consiste en 3 partitions de tailles différentes 7, 7 et 6 avec des valeurs égales à 0.5, 0.6 et 0.7, telles que $\mathcal{H}_0^{(7,8)}$ et $\mathcal{H}_0^{(14,15)}$ ne sont pas vraies ;

(iv) le Scenario4 consiste en 4 partitions de tailles différentes 7, 6, 6 et 1 avec des valeurs égales à 0.4, 0.5, 0.6 et 0.7, telles que $\mathcal{H}_0^{(7,8)}$, $\mathcal{H}_0^{(13,14)}$ et $\mathcal{H}_0^{(19,20)}$ ne sont pas vraies (i.e. H_{20} est unique).

La matrice de covariance Σ de taille $M \times M$ des M-mBf est choisie avec des entrées diagonales fixées à $\Sigma_{m,m} = 2^m$, pour $m = 1, \ldots, M$, et des entrées non diagonales fixées à $\Sigma_{m,m'} = 0.5\sqrt{\Sigma_{m,m}\Sigma_{m',m'}}$, pour $m \neq m' = 1, \ldots, M$. La matrice de mélange inversible W de taille $M \times M$ des M-mBf est choisie au hasard et identique pour toutes les expériences.

L'analyse en ondelettes est effectuée avec l'ondelette de Daubechies la moins asymétrique à $N_\psi = 3$ moments nuls (Daubechies, 1992) sur les échelles d'analyse $2^{j_1} = 2^7$ à $2^{j_2} = 2^{10}$.

3.6.2 Rejets de l'hypothèse nulle

La décision d_α de rejeter l'hypothèse nulle $\mathcal{H}_0$ ($H_1 = \ldots = H_M$) avec un niveau de confiance α peut être construite à partir d'une des procédures de test suivantes :

(*i*) le test du χ^2 de la section 3.3 ;

(*ii*) la procédure de Benjamini-Hochberg sur les $M-1$ tests par paires d'exposants d'autosimilarité ordonnés, comme détaillé dans la section 3.4, où les paramètres des tests $\tilde{\sigma}_m$ peuvent être estimés de deux façons :
- soit par $\bar{\sigma}^*_m$ (cf. Eq. (3.25)) et le test associé est appelé *test demi-normal*,
- soit par $\tilde{\sigma}^*_m$ (cf. Eq. (3.29)) et le test associé est appelé *test normal replié* ;

(*iii*) la procédure de Benjamini-Hochberg sur les $M(M-1)/2$ tests par paires d'exposants d'autosimilarité, comme détaillé dans la section 3.5, à laquelle on fera référence par *test gaussien*.

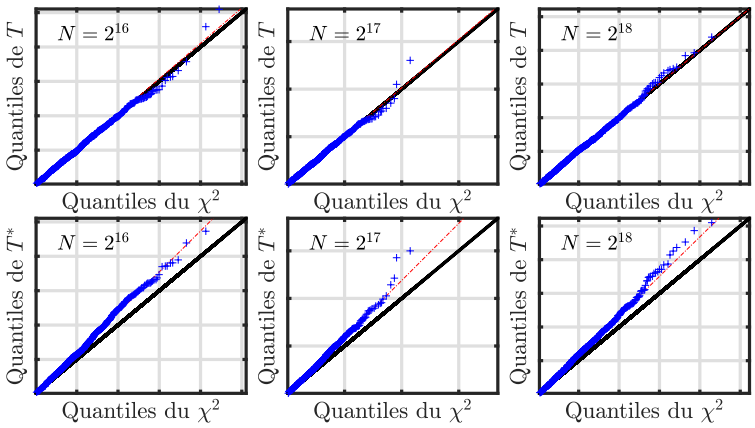
Formulaire récapitulatif sur les procédures de tests

Les différentes quantité relatives aux procédures de tests étudiées sont rappelées dans le tableau suivant.

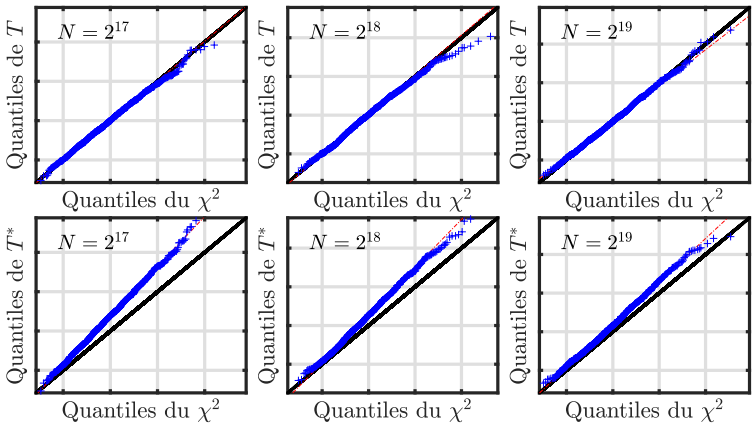
Test	χ^2	Demi-normal	Normal repliée	Gaussien
Nombre de statistiques	1	$M-1$	$M-1$	$M(M-1)/2$
Statistiques	T	$\tilde{\delta}_m$	$\tilde{\delta}_m$	$\hat{\delta}_{m,m'}$
Statistiques bootstrap	T^*	$\bar{\delta}^*_m$	$\tilde{\delta}^*_m$	$\hat{\delta}^*_{m,m'}$
Paramètres de test	$\hat{\Sigma}^*_{\underline{\hat{H}}}$	$\bar{\sigma}^*_m$	$\tilde{\sigma}^*_m$	$\hat{\sigma}^*_{m,m'}$

Limites du test du χ^2

Pour des échelles d'analyse $2^{j_1}, \ldots, 2^{j_2}$ grandes dans la procédure d'estimation de $\underline{H}$, la distribution du vecteur des estimées bootstrap $\underline{\hat{H}}^*$ ne converge pas assez rapidement vers une loi gaussienne multivariée, et finalement la distribution de la statistique bootstrap T^* (cf. Eq. (3.8)) est mal approximée par une loi du χ^2, fait mis en évidence par la figure 3.42. En effet, la figure 3.42 montre les diagrammes quantile-quantile des distributions empiriques de la statistique T (cf. Eq. (3.5)) et de la statistique bootstrap T^* du test du χ^2 en fonction de la distribution théorique du χ^2 sous l'hypothèse nulle $\mathcal{H}_0$ (Scenario1) pour différents nombres de composantes M et différentes tailles d'échantillon N. Quelques soient le nombre de composantes M et la taille d'échantillon N, la statistique T suit approximativement une loi du χ^2, comme prévu par la théorie, mais la distribution de la statistique bootstrap T^* s'en éloigne. Ces résultats contrastent avec les résultats de la section 3.3.5 (en particulier la figure 3.6), où les échelles d'analyse choisies permettaient à la statistique T^* de suivre un comportement adéquat.



(a) $M = 6$



(b) $M = 20$

FIGURE 3.42 – **Distributions de T et T^* sous $\mathcal{H}_0$.** Diagrammes quantile-quantile de (en haut) la statistique T (cf. Eq. (3.5)) et (en bas) la statistique bootstrap T^* (cf. Eq. (3.5)) du test du χ^2 au travers des réalisations de Monte Carlo en fonction de la distribution théorique du χ^2 à $M-1$ degrés de liberté sous l'hypothèse nulle $\mathcal{H}_0$ (Scenario1) à M composantes pour (de gauche à droite) différentes tailles d'échantillon N. *Les distributions des statistiques bootstrap T^* sont mal approximées par des distributions du χ^2, contrairement aux distributions des statistiques T.*

Limites des tests demi-normal et normal replié

Les tests demi-normal et normal replié reposent sur l'a priori de demi-normalité des $M-1$ statistiques $\tilde{\delta}_m$ sous les hypothèses nulles par paires $H_m = H_{m+1}$ (cf. Eq. (3.17)), pour $m = 1, \ldots, M-1$. La figure 3.43 rapporte les diagrammes quantile-quantile des distributions empiriques des statistiques $\tilde{\delta}_m$ des tests par paires d'exposants ordonnés contre une distribution demi-normale de paramètre $\bar{\sigma}_m = \sqrt{\mathrm{Var}(\tilde{\delta}_m)/(1-2/\pi)}$ estimé par Monte Carlo, sous le scénario 1 où $H_m = H_{m+1}$ pour tout $m = 1, \ldots, M-1$, pour $M = 20$ composantes et une taille d'échantillon $N = 2^{18}$. Ces diagrammes montrent que, pour un grand nombre de composantes M, l'approximation des distributions des statistiques $\tilde{\delta}_m$ par des distributions demi-normales est mise en défaut pour toutes les paires (H_m, H_{m+1}). L'a priori sur lequel sont construits les tests n'est donc plus valide.

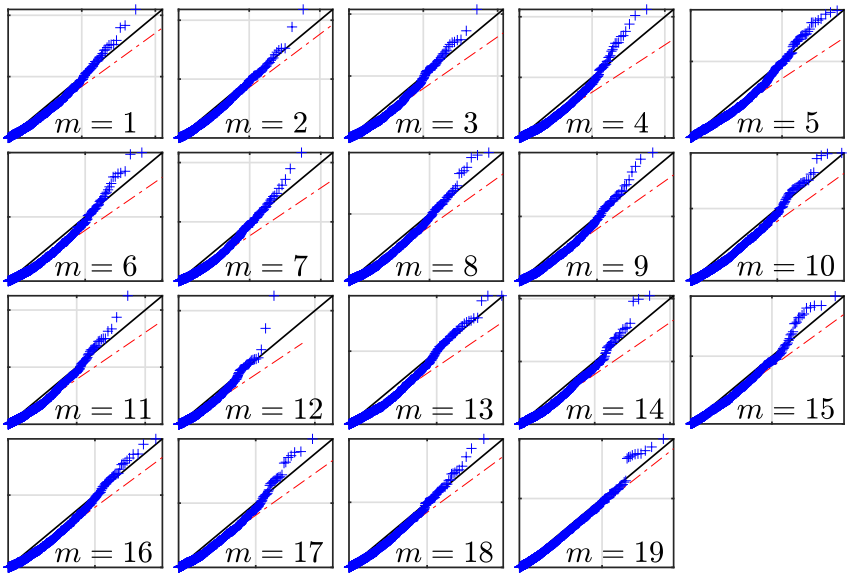


FIGURE 3.43 – **Distributions des $\tilde{\delta}_m$ sous $\mathcal{H}_0$.** Diagrammes quantile-quantile de $\tilde{\delta}_m$ (cf. Eq (3.17)) par rapport à une distribution demi-normale $\mathcal{FN}_{0,\bar{\sigma}_m}$ de paramètre d'échelle $\bar{\sigma}_m = \sqrt{\mathrm{Var}(\tilde{\delta}_m)/(1-2/\pi)}$ estimé par Monte Carlo, pour $m = 1, \ldots, M-1$, sous l'hypothèse nulle $\mathcal{H}_0$ (Scenario1) avec $M = 20$ composantes et pour une taille d'échantillon $N = 2^{18}$. *Les distributions des statistiques $\tilde{\delta}_m$ sont mal approximées par des distributions demi-normales.*

Comparaison pour la reproduction du niveau de confiance

Les proportions de rejet de l'hypothèse nulle $\mathcal{H}_0$ ($H_1 = \ldots = H_M$) des différents tests sont à présent comparées. La figure 3.44 rapporte les décisions de rejet d_α des différentes procédures de test moyennées sur les réalisations de Monte Carlo sous l'hypothèse nulle $\mathcal{H}_0$ (Scenario1) en fonction du niveau de confiance ciblé α pour différents nombres de composantes M et différentes tailles d'échantillon. Le test du χ^2 reconstruit mal le niveau de confiance ciblé α en raison du comportement inadapté de l'estimateur bootstrap de la covariance en jeu $\hat{\Sigma}^*_{\underline{\hat{H}}}$, menant ainsi à un test trop conservatif. En revanche, le test gaussien reconstruit bien α quels que soient le nombre de composantes M et la taille d'échantillon N. Quant aux tests demi-normal et normal replié, ils peinent à reproduire correctement α car les statistiques de test correspondantes (qui sont les mêmes pour les deux tests) sont mal approximées par des distributions demi-normales sous l'hypothèse nulle $\mathcal{H}_0$ (comportement supposé des statistiques pour la construction de ces tests).

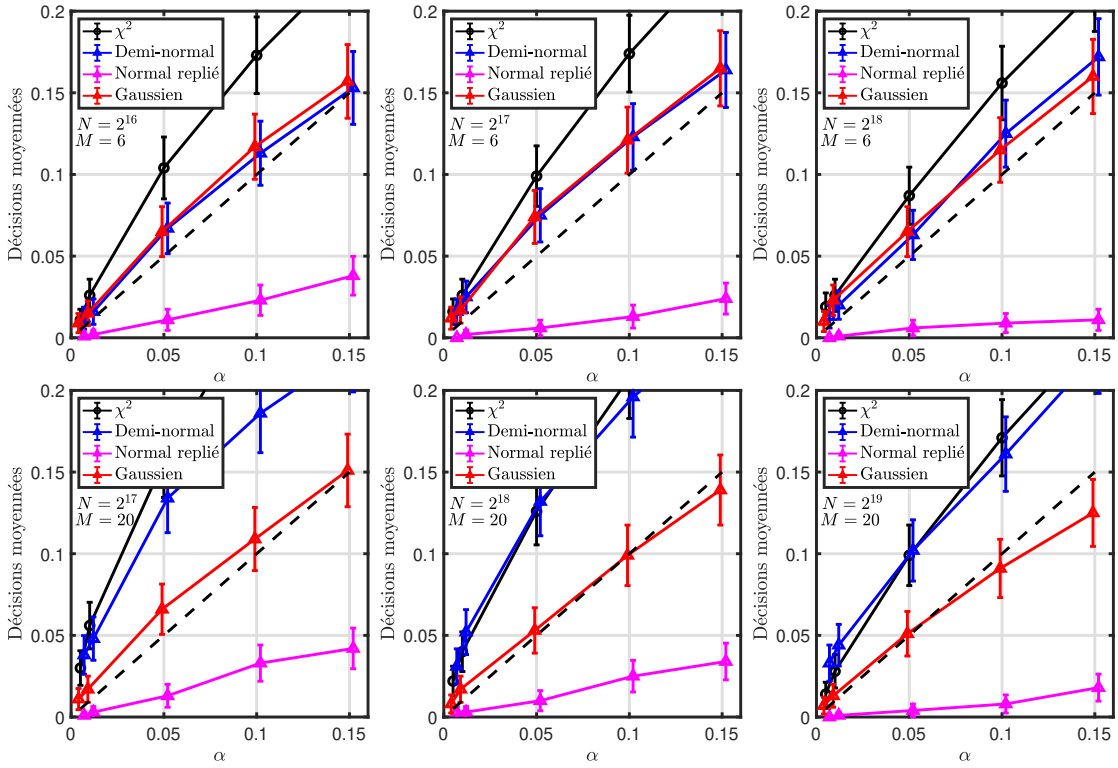


FIGURE 3.44 – **Niveaux de confiance des différents tests.** Décisions des tests moyennées sur les réalisations de Monte Carlo en fonction du niveau de confiance ciblé α pour les différentes procédures de tests étudiées sous $\mathcal{H}_0$ (Scenario1) avec différents nombres de composantes M (de haut en bas) et différentes tailles d'échantillon N (de gauche à droite). Les lignes noires pointillées indiquent les valeurs théoriques attendues, c'est-à-dire la première bissectrice. *Seul le test gaussien garantit la bonne reconstruction du niveau de confiance ciblé α dans tous les cas de figure.*

On remarque par ailleurs que, pour les différentes nombres de composantes M, la taille d'échantillon n'a pas d'impact sur la reconstruction de α pour les différents tests, excepté le test normal replié. En effet, le test normal replié est d'autant plus conservatif que la taille d'échantillon N est grande.

En conclusion, ces résultats montrent que le test gaussien est le plus adapté pour reconstruire l'hypothèse nulle, quels que soient le nombre de composantes et la taille d'échantillon.

3.6.3 Stratégies de partitionnement

On souhaite à présent comparer les performances de partitionnement du vecteur des exposants d'autosimilarité $\underline{H} = (H_1, \ldots, H_M)$ des trois méthodes présentées dans les sections précédentes :

(i) le partitionnement à partir des $M-1$ décisions issues du test demi-normal, appelé *partitionnement demi-normal*;

(ii) le partitionnement à partir des $M-1$ décisions du test normal replié, appelé *partitionnement normal replié*;

(iii) le *partitionnement du graphe* des M exposants d'autosimilarité pondéré par la matrice de similarité S issue du test gaussien (cf. Section 3.5.5).

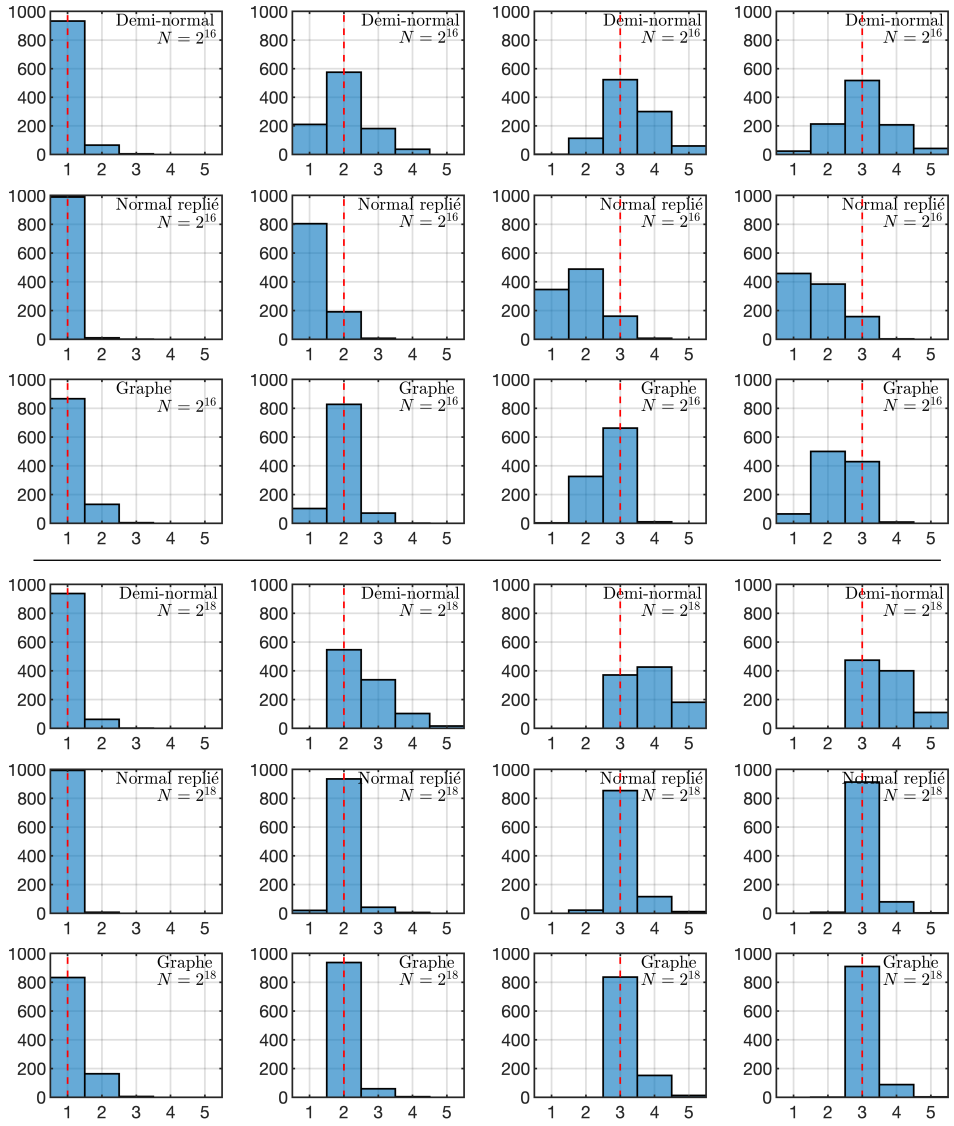
Pour les différentes procédures de tests par paires en jeu dans chacune des méthodes, le niveau de confiance est fixé à $\alpha = 0.05$.

Comparaison pour $M = 6$ composantes

En premier lieu, la figure 3.45 rapporte les histogrammes des nombres estimés de partitions $\hat{N}_{\mathrm{C}}$ pour les différentes méthodes de partitionnement sous les différents scénarios avec $M = 6$ composantes et différentes tailles d'échantillon $N = 2^{16}, 2^{18}$. Lorsque tous les exposants d'auto-similarité sont égaux (Scenario1), le taux de détection d'une partition unique (i.e. proportion de $\hat{N}_{\mathrm{C}} = 1$) obtenu par le partitionnement demi-normal reconstruit le taux de fausses découvertes $\alpha = 0.05$ tandis que celui obtenu par le partitionnement normal replié surestime α, en accord avec les observations de la figure 3.44. Pour ce qui est du partitionnement du graphe, ce taux de détection n'est pas contrôlé par le taux de fausses découvertes α, qui n'intervient que pour la détection des nœuds isolés. Dans les cas des scénarios 2, 3 et 4, où le vecteur $\underline{H}$ est formé de plusieurs partitions, le partitionnement demi-normal estime mal le nombre de partitions N_{C} contrairement au partitionnement du graphe, pour les deux tailles d'échantillon. En effet, le partitionnement demi-normal a tendance à surestimer le nombre de partitions N_{C} en raison de la mauvaise approximation des distributions des $M-1$ statistiques associées aux hypothèses nulles par paires $H_m = H_{m+1}$ par une distribution demi-normale sous une hypothèse alternative $\mathcal{H}_1$. Quant au partitionnement normal replié, il ne donne une estimation du nombre de partitions N_{C} aussi bonne que celle du partitionnement du graphe qu'à partir d'une taille d'échantillon N suffisamment grande (voir la section 3.4.7 pour plus de détails).

Enfin, le tableau 3.12 quantifie les performances des différentes méthodes de partitionnement en termes d'ARI et NMI pour les différents scénarios et différentes tailles d'échantillon N. Le trois méthodes voient leurs performances augmenter avec la taille d'échantillon N. Le partitionnement demi-normal atteint des performances raisonnables, pouvant atteindre celles du partitionnement du graphe à faible taille d'échantillon $N = 2^{16}$, mais qui sont limitées par la surestimation du nombre de partitions N_{C} sous les scénarios à plusieurs partitions (scénarios 2, 3 et 4). Le partitionnement normal replié atteint les performances du partitionnement du graphe à partir de la taille d'échantillon $N = 2^{18}$, car dans ce cas, les statistiques des tests normaux repliés $\tilde{\delta}_m = \hat{H}_{\tau(m+1)} - \hat{H}_{\tau(m)}$ sous une hypothèse alternative $H_{m+1} > H_m$ sont mieux approximées par une loi normale repliée, pour $m = 1, \ldots, M-1$.

Le partitionnement du graphe atteint de bonnes performances, mais ces performances sont difficilement comparables à celles des autres méthodes ici, car les performances d'une unique partition (Scenario1) ne sont pas similaires entre les différentes méthodes. Ceci est dû au fait que les performances du partitionnement du graphe n'est pas contrôlé par un niveau de confiance α. Une comparaison approfondie des performances est cependant possible dans la configuration à $M = 20$ composantes étudiée ci-dessous.



(a) Scenario1 (b) Scenario2 (c) Scenario3 (d) Scenario4

FIGURE 3.45 – **Estimations du nombre de partitions pour $M = 6$.** Histogrammes des nombres estimés de partitions $\hat{N}_C$ selon différentes méthodes de partitionnement sous les différents scénarios avec $M = 6$ composantes et différentes tailles d'échantillon $N = 2^{16}$ (pour les trois premières lignes) et $N = 2^{18}$ (pour les trois dernières lignes). Les lignes rouges pointillées indique le nombre exact de partitions.

TABLEAU 3.12 – **Performances de la stratégie de partitionnement pour $M = 6$.** Performances en termes d'ARI et NMI (moyenne de Monte Carlo ± intervalle de confiance à 95%) obtenues par les différentes stratégies de partitionnement pour les différents scénarios avec $M = 6$ composantes et différentes tailles d'échantillon N.

			Scenario1	Scenario2	Scenario3	Scenario4
$N = 2^{16}$	NMI	Demi-normal	n/a	0.67 ± 0.02	$\mathbf{0.87 \pm 0.01}$	$\mathbf{0.79 \pm 0.01}$
	ARI		0.93 ± 0.02	0.60 ± 0.03	0.69 ± 0.02	0.58 ± 0.02
	NMI	Normal replié	n/a	0.17 ± 0.02	0.51 ± 0.02	0.42 ± 0.02
	ARI		$\mathbf{0.99 \pm 0.01}$	0.16 ± 0.02	0.35 ± 0.02	0.30 ± 0.02
	NMI	Graphe	n/a	$\mathbf{0.75 \pm 0.02}$	0.86 ± 0.01	0.78 ± 0.01
	ARI		0.87 ± 0.02	$\mathbf{0.69 \pm 0.02}$	$\mathbf{0.71 \pm 0.02}$	$\mathbf{0.60 \pm 0.02}$
$N = 2^{17}$	NMI	Demi-normal	n/a	$\mathbf{0.90 \pm 0.01}$	$\mathbf{0.92 \pm 0.01}$	$\mathbf{0.92 \pm 0.01}$
	ARI		0.93 ± 0.02	$\mathbf{0.84 \pm 0.01}$	$\mathbf{0.77 \pm 0.02}$	$\mathbf{0.75 \pm 0.02}$
	NMI	Normal replié	n/a	0.62 ± 0.03	0.87 ± 0.01	0.87 ± 0.01
	ARI		$\mathbf{0.99 \pm 0.00}$	0.62 ± 0.03	0.74 ± 0.02	0.79 ± 0.02
	NMI	Graphe	n/a	0.75 ± 0.02	0.86 ± 0.01	0.78 ± 0.01
	ARI		0.87 ± 0.02	0.69 ± 0.02	0.71 ± 0.02	0.60 ± 0.02
$N = 2^{18}$	NMI	Demi-normal	n/a	0.91 ± 0.01	0.93 ± 0.00	0.93 ± 0.00
	ARI		0.94 ± 0.02	0.83 ± 0.01	0.77 ± 0.01	0.77 ± 0.02
	NMI	Normal replié	n/a	0.97 ± 0.01	$\mathbf{0.98 \pm 0.00}$	$\mathbf{0.99 \pm 0.00}$
	ARI		$\mathbf{0.99 \pm 0.00}$	0.97 ± 0.01	$\mathbf{0.95 \pm 0.01}$	$\mathbf{0.97 \pm 0.01}$
	NMI	Graphe	n/a	$\mathbf{0.99 \pm 0.00}$	$\mathbf{0.98 \pm 0.00}$	0.98 ± 0.00
	ARI		0.83 ± 0.02	$\mathbf{0.98 \pm 0.01}$	$\mathbf{0.95 \pm 0.01}$	0.95 ± 0.00

Comparaison pour $M = 20$ composantes

Pour mieux observer l'avantage de la méthode de partitionnement du graphe, les différentes stratégies sont à présent comparées sur un plus grand nombre de composantes $M = 20$. Pour visualiser le comportement du partitionnement du graphe $\mathcal{G} = (\mathcal{V}, \epsilon, S)$ pour $M = 20$ composantes, la figure 3.46 donne une matrice de similarité S et les valeurs propres ordonnées $\varphi_1 \leq \ldots \leq \varphi_M$ du laplacien associé pour une réalisation de Monte Carlo sous le scénario 3 avec une taille d'échantillon $N = 2^{18}$. On distingue bien les différentes partitions sur la matrice S, et les trois premières valeurs propres φ_1, φ_2, φ_3 sont très proches de 0 relativement aux autres valeurs propres $\varphi_4, \ldots, \varphi_{20}$, comme en témoigne le grand maximum eigengap $e_3 = \varphi_4 - \varphi_3$.

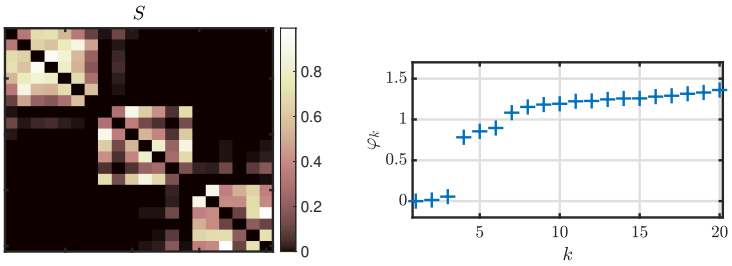
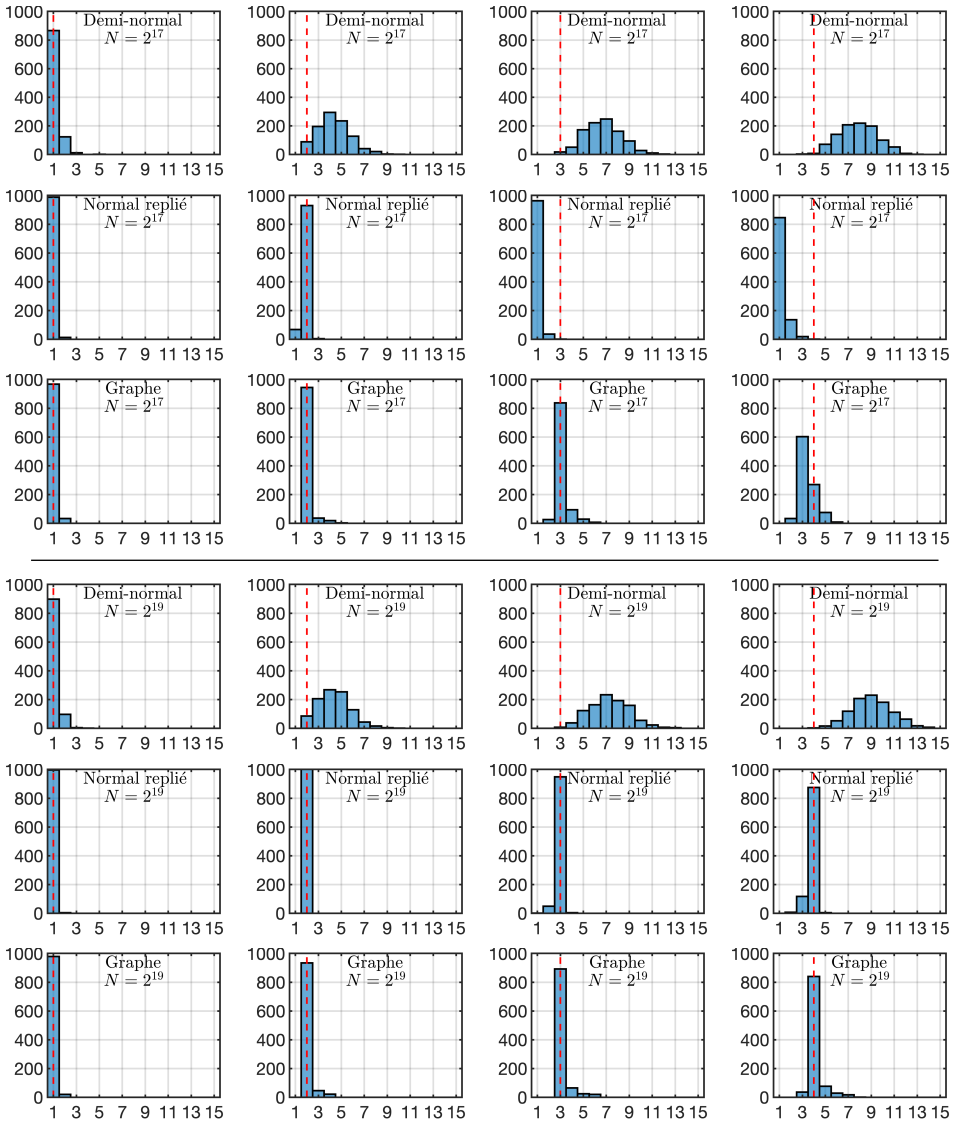


FIGURE 3.46 – **Exemple de matrice de similarité S.** Matrice de similarité S (à gauche) et valeurs propres du laplacien de graph associé $\varphi_1, \ldots, \varphi_M$, (à droite) pour une réalisation de Monte Carlo, sous le scénario 3 (composé de 3 partitions) avec $M = 20$ composantes et une taille d'échantillon $N = 2^{18}$.

La figure 3.47 rapporte les histogrammes des nombres estimés de partitions $\hat{N}_{\mathrm{C}}$ pour les différentes méthodes de partitionnement sous les différents scénarios avec $M = 20$ composantes et différentes tailles d'échantillon $N = 2^{16}$, 2^{18}. Quelle que soit la taille d'échantillon N, le partitionnement demi-normal estime très mal le nombre de partitions N_{C}. Pour ce qui est du partitionnement normal replié, il atteint de bonnes performances d'estimation de N_{C} pour une taille d'échantillon N suffisamment grande et un écart $\Delta H = H_{m+1} - H_m$ suffisamment grand entre deux partitions de valeurs H_m et H_{m+1} (comme c'est le cas dans le scénario 2 où $\Delta H = 0.2$). Dans les autres cas, ses performances sont extrêmement faibles à cause de la mauvaise approximation des lois des statistiques des tests par paires $\tilde{\delta}_m$ par des lois normales repliées, pour $m = 1, \ldots, M-1$. En revanche, le partitionnement du graphe estime le nombre de partitions N_{C} de façon très satisfaisante pour toutes les tailles d'échantillon N et tous les scénarios, en dehors du scénario 4 où cette estimation est affaiblie par la présence d'un exposant d'autosimilarité unique (en l'occurrence H_{20}).

Les performances de partitionnement des différentes approches sont quantifiées dans le tableau 3.13 par l'ARI et la NMI. Comme les performances d'estimation du nombre de partitions N_{C} le laissaient présager, le partitionnement demi-normal donne des résultats bien moins bons que le partitionnement du graphe, et similaires pour toutes les tailles d'échantillon N, avec des ARI et NMI atteignant respectivement 0.80 et 0.68 sous les scénarios à plusieurs partitions. Comme observé sur les simulations à $M = 6$ composantes, les performances du partitionnement normal replié ne sont corrects qu'à partir d'une taille d'échantillon N suffisamment grande. En particulier, ses performances sont extrêmement mauvaises sous les scénarios 3 et 4 (c'est-à-dire pour un écart $H_{m+1} - H_m = 0.2$ entre des partitions de valeurs H_m et H_{m+1}) pour une taille d'échantillon $N = 2^{16}$, et restent très peu satisfaisantes pour une taille d'échantillon $N = 2^{17}$. En revanche, le partitionnement du graphe atteint de très bonnes performances quels que soient le scénario et la taille d'échantillon N, avec des NMI et ARI n'allant pas en deçà de 0.83 et 0.75, respectivement.

En définitive, la stratégie de partitionnement du graphe apparaît comme la plus robuste au



(a) Scenario1 (b) Scenario2 (c) Scenario3 (d) Scenario4

FIGURE 3.47 – **Estimation du nombre de partitions pour $M = 20$.** Histogrammes des nombres estimés de partitions $\hat{N}_C$ pour les différents scénarios avec un nombre de composantes $M = 20$ et différentes tailles d'échantillon $N = 2^{17}$ (pour les trois premières lignes) et $N = 2^{19}$ (pour les trois dernières lignes). Les lignes rouges pointillées indiquent le nombre exact de partitions.

TABLEAU 3.13 – **Performances de la stratégie de partitionnement.** Performances en termes d'ARI et NMI (moyenne de Monte Carlo ± intervalle de confiance à 95%) obtenues par les différentes stratégies de partitionnement pour les différents scénarios à $M = 20$ composantes et tailles d'échantillon N.

			Scenario1	Scenario2	Scenario3	Scenario4
$N = 2^{17}$	NMI	Demi-normal	n/a	0.78 ± 0.01	0.78 ± 0.00	0.78 ± 0.00
	ARI		0.87 ± 0.02	0.68 ± 0.01	0.60 ± 0.01	0.57 ± 0.01
	NMI	Normal replié	n/a	0.93 ± 0.02	0.03 ± 0.01	0.09 ± 0.01
	ARI		$\mathbf{0.99 \pm 0.01}$	0.93 ± 0.02	0.02 ± 0.01	0.05 ± 0.01
	NMI	Graphe	n/a	$\mathbf{0.99 \pm 0.00}$	$\mathbf{0.85 \pm 0.01}$	$\mathbf{0.83 \pm 0.01}$
	ARI		0.97 ± 0.01	$\mathbf{0.98 \pm 0.01}$	$\mathbf{0.79 \pm 0.01}$	$\mathbf{0.75 \pm 0.01}$
$N = 2^{18}$	NMI	Demi-normal	n/a	0.78 ± 0.01	0.79 ± 0.00	0.80 ± 0.00
	ARI		0.87 ± 0.02	0.68 ± 0.01	0.60 ± 0.01	0.57 ± 0.01
	NMI	Normal replié	n/a	$\mathbf{1.00 \pm 0.00}$	0.59 ± 0.03	0.68 ± 0.02
	ARI		$\mathbf{0.99 \pm 0.01}$	$\mathbf{1.00 \pm 0.00}$	0.49 ± 0.02	0.56 ± 0.02
	NMI	Graphe	n/a	0.99 ± 0.00	$\mathbf{0.97 \pm 0.00}$	$\mathbf{0.95 \pm 0.00}$
	ARI		0.96 ± 0.01	0.98 ± 0.01	$\mathbf{0.96 \pm 0.01}$	$\mathbf{0.92 \pm 0.01}$
$N = 2^{19}$	NMI	Demi-normal	n/a	0.78 ± 0.01	0.80 ± 0.00	0.80 ± 0.00
	ARI		0.90 ± 0.02	0.68 ± 0.01	0.61 ± 0.01	0.56 ± 0.01
	NMI	Normal replié	n/a	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.99 \pm 0.00}$	$\mathbf{0.99 \pm 0.00}$
	ARI		$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.98 \pm 0.01}$	$\mathbf{0.98 \pm 0.00}$
	NMI	Graphe	n/a	0.99 ± 0.00	$\mathbf{0.99 \pm 0.00}$	0.98 ± 0.00
	ARI		0.98 ± 0.01	0.98 ± 0.01	0.97 ± 0.01	0.97 ± 0.01

nombre de composantes M, à la taille d'échantillon N, au faible écart entre exposants d'autosimilarité H_m différents et au choix des échelles d'analyse $2^{j_1}, \ldots, 2^{j_2}$.

3.6.4 Matrices de similarité du graphe des exposants

Cette section a pour but d'évaluer la stratégie de partitionnement PageRank, reposant sur la matrice de similarité PageRank S_η, définie dans la section 3.5.6, pour un grand nombre de composantes $M = 20$, et de la confronter à la stratégie de partitionnement du graphe pondéré par la matrice de similarité S (cf. Section 3.5.5).

Pour illustrer le comportement du partionnement PageRank dans un cas à $M = 20$ composantes, la figure 3.48 donne des exemples de matrice de similarité PageRank S_η et les valeurs propres ordonnées $\varphi_1, \leq \ldots \leq \varphi_M$ des laplaciens associés pour plusieurs paramètres η sous le scénario 3 (i.e. 3 partitions) pour une même réalisation de Monte Carlo et une taille d'échantillon $N = 2^{18}$. Le paramètre η permet d'ajuster la parcimonie de la matrice de similarité S_η et donc le comportement des valeurs propres φ_k. Pour certaines valeurs de η, les valeurs propres $\varphi_1, \varphi_2, \varphi_3$, proches de 0, se détachent davantage des autres valeurs propres $\varphi_4, \ldots, \varphi_{20}$, faisant ainsi du maximum eigengap $e_3 = \varphi_4 - \varphi_3$ une mesure plus pertinente pour l'estimation du nombre de partitions.

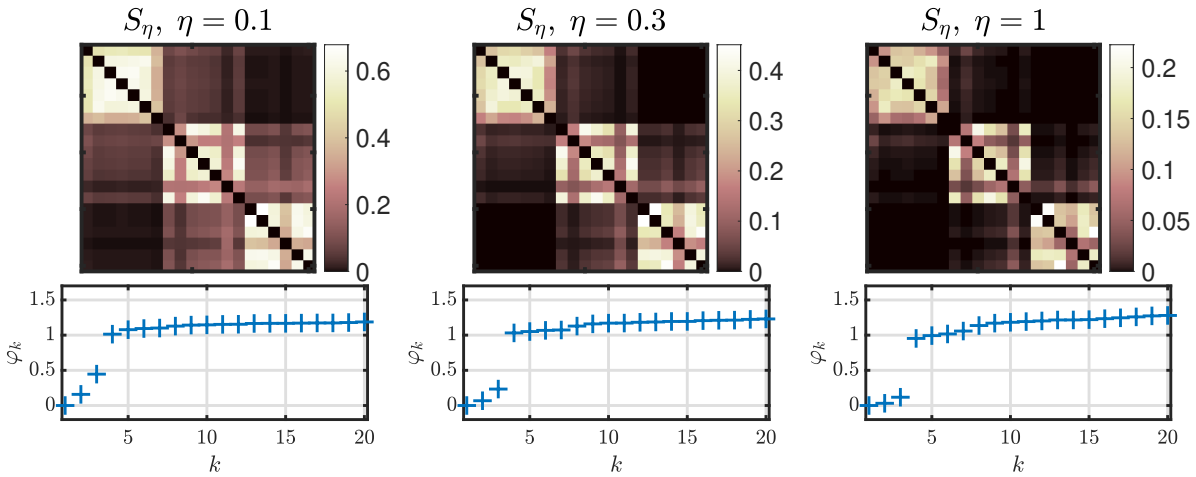
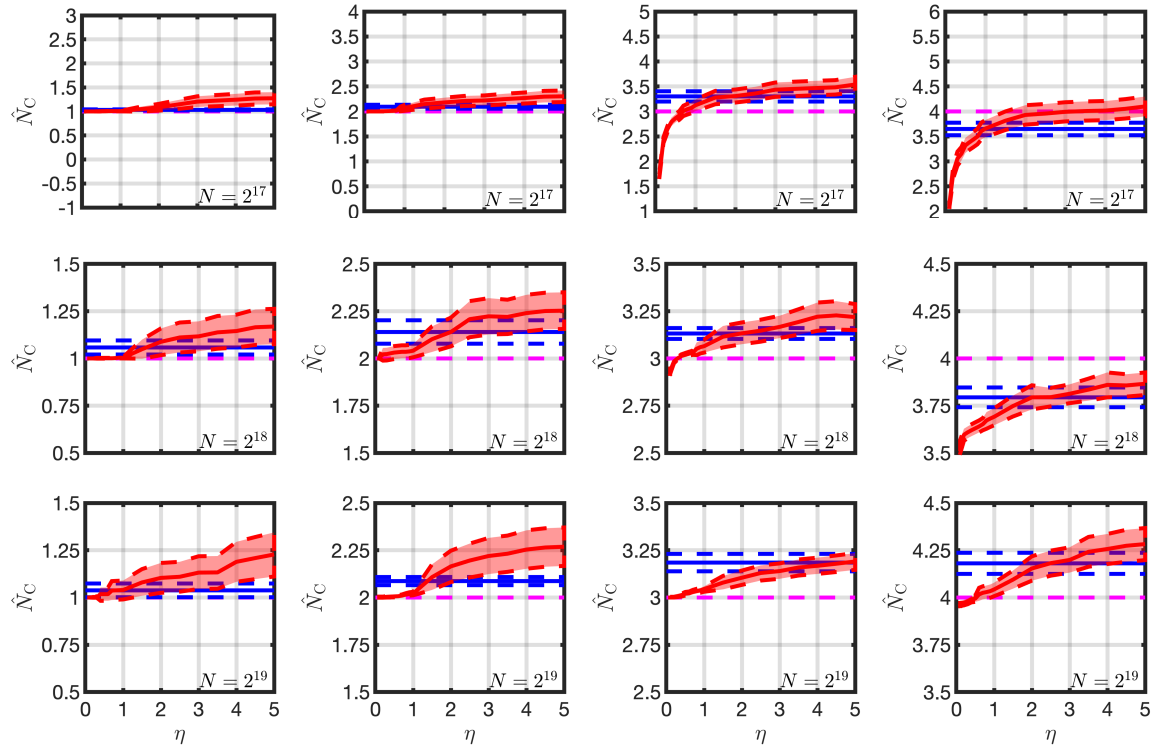


FIGURE 3.48 – **Exemples de matrices de similarité PageRank S_η.** Matrices de similarité PageRank S_η et valeurs propres $\varphi_1, \ldots, \varphi_M$, des laplaciens associés pour une réalisation de Monte Carlo, différents paramètres η et une taille d'échantillon $N = 2^{18}$ sous le scénario 3 avec $M = 20$ composantes.

Pour évaluer l'estimation du nombre de partitions N_C, la figure 3.49 rapporte les nombres de partitions estimés $\hat{N}_C$ en utilisant les matrices de similarité S (en bleu) et S_η (en rouge) moyennés sur les réalisations de Monte Carlo (avec intervalle de confiance à 95%) en fonction du paramètre η pour les différents scénarios à $M = 20$ composantes et différentes tailles d'échantillon N. Comme pour le cas à $M = 6$ composantes (cf. Figure 3.40), le nombre de partitions estimé $\hat{N}_C$ augmente avec le paramètre η. Cela aboutit à une surestimation dans le cas d'une seule partition (scénario 1). Pour les cas à plusieurs partitions (c'est-à-dire les scénarios 2, 3 et 4), le partitionnement sous-estime fortement N_C pour des valeurs de η proches de 0, un écart $\Delta H = |H_{m+1} - H_m|$ entre les valeurs H_m et H_{m+1} de deux partitions faible et une taille d'échantillon N faible, comme c'est le cas pour les scénarios 3 et 4 avec $N = 2^{17}$ et $\Delta H = 0.1$.

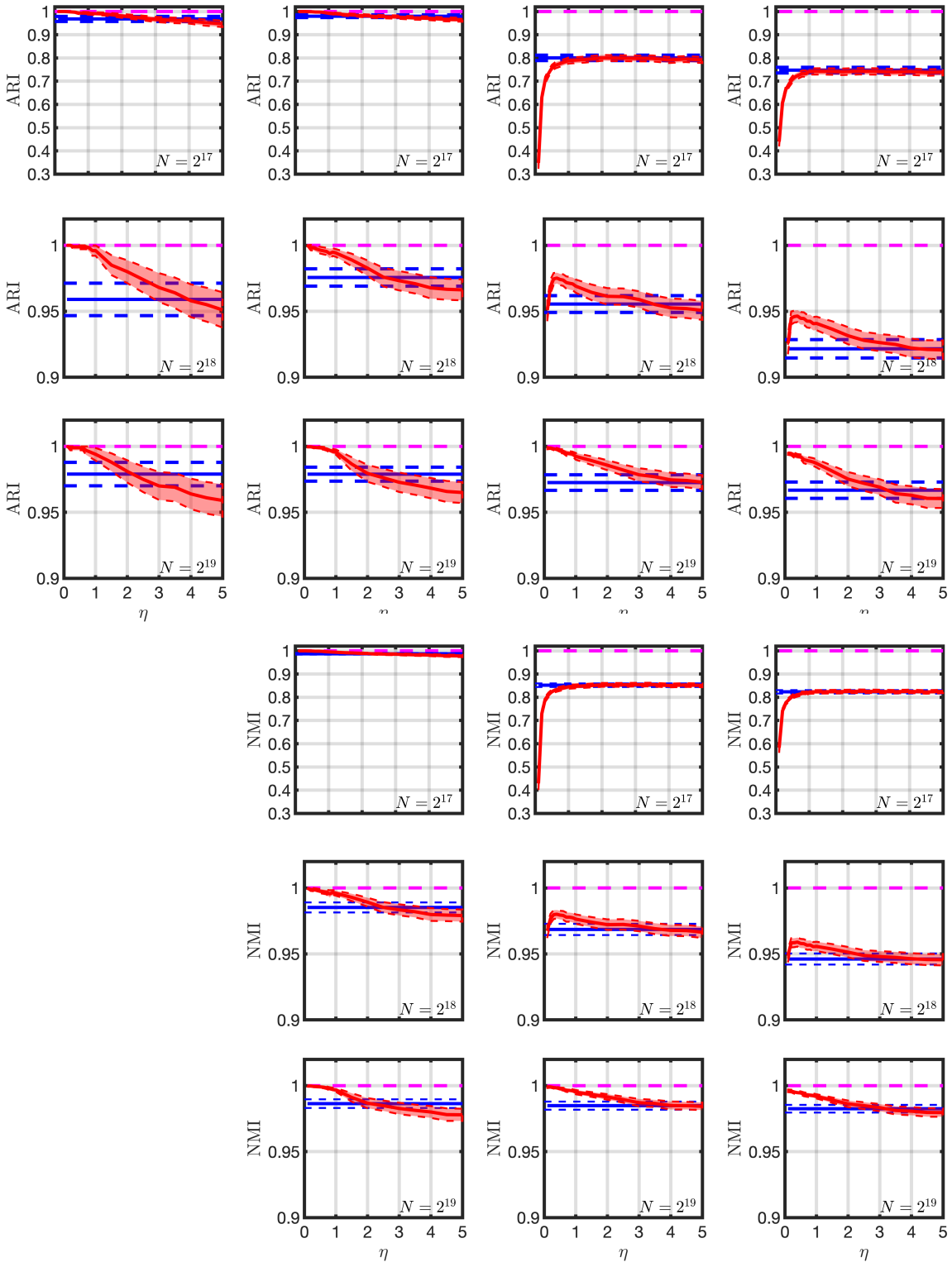


(a) Scenario1 (b) Scenario2 (c) Scenario3 (d) Scenario4

FIGURE 3.49 – **Estimation du nombre de partitions du graphe PageRank.** Nombres estimés $\hat{N}_C$ de partitions (moyennes de Monte Carlo ± intervalle de confiance à 95%) du graphe pondéré par (en bleu) la matrice de similarité S et (en rouge) la matrice de similarité PageRank S_η en fonction du paramètre η pour les différents scénarios à $M = 20$ composantes et (de haut en bas) différentes tailles d'échantillon N. Les lignes magenta pointillées indiquent le nombre exact de partitions.

Enfin, la figure 3.50 montre les performances des partitionnements du graphe pondéré par S et du graphe pondéré par S_η en termes d'ARI et NMI (moyennées sur les réalisations de Monte Carlo) pour les différents scénarios à $M = 20$ composantes et tailles d'échantillon N. À faible taille d'échantillon $N = 2^{16}$, dans les scénarios 3 et 4 où l'écart $|H_{m+1} - H_m|$ entre les valeurs H_m et H_{m+1} de deux partitions est de 0.1, les performances du partitionnement PageRank sont très faibles pour de petites valeurs de η, et croissent avec η jusqu'à atteindre les performance du partitionnement à partir de S. En revanche, dans le scénario 1 constitué d'une seule partition et dans le scénario 2 où $|H_{m+1} - H_m| = 0.2$ (lorsque $H_{m+1} \neq H_m$), les performances du partitionnement PageRank décroissent avec η jusqu'à atteindre les performances du partitionnement à partir de S. Aux tailles d'échantillon $N = 2^{17}, 2^{18}$, pour une certaine plage de paramètres η, les performances du partitionnement PageRank sont meilleures que la stratégie reposant sur la matrice de similarité non paramétrée S.

En conclusion, le paramètre η permet de contrôler la détection d'une partition unique et peut être choisi de façon à optimiser les performances de partitionnement. De petites valeurs de η peuvent conduire à regrouper à tort des exposants d'autosimilarité H_m distincts tandis que de grandes valeurs de η peuvent mener à séparer à tort des exposants d'autosimilarité H_m égaux dans différentes partitions. Le choix du paramètre η peut être fait en fonction de la forme de la matrice de similarité S_η et du comportement des valeurs propres $\varphi_1, \ldots, \varphi_M$ du laplacien associé. En outre, les différents graphiques présentés permettent également de déterminer les paramètres optimaux η pour effectuer le comptage et le regroupement des exposants d'autosimilarité H_m.



(c) Scenario1 (d) Scenario2 (e) Scenario3 (f) Scenario4

FIGURE 3.50 – **Performances de la statrégie de partitionnement PageRank.** Performances en termes d'ARI et NMI (moyennes de Monte Carlo ± intervalle de confiance à 95%) du partitionnement du graphe pondéré par (en bleu) la matrice de similarité S et (en rouge) la matrice de similarité PageRank S_η en fonction du paramètre η pour les différents scénarios à $M = 20$ composantes et (de haut en bas) différentes tailles d'échantillon N. Les lignes magenta pointillées indiquent les valeurs idéales de performances, à savoir 1.

3.6.5 Conclusions

Cette section confronte les différentes approches proposées dans les sections précédentes pour dénombrer et regrouper les exposants d'autosimilarité $H_1, \ldots, H_M$ à partir de leurs estimées $\hat{H}_1, \ldots, \hat{H}_M$. Pour ce, leurs performances sont évaluées sur des simulations de Monte Carlo réalisées sur des M-mBf synthétiques pour des différents nombres de composantes $M = 6$ et $M = 20$.

En plus de ne permettre que la détection d'une partition unique, la procédure de test du χ^2 montre des limites dues à l'estimation bootstrap de la covariance des $\hat{H}_1, \ldots, \hat{H}_M$, qui nécessite des échelles d'analyse suffisamment petites. Quant aux méthodes de partitionnement reposant sur $M-1$ tests de paires d'exposants ordonnés successifs $H_m = H_{m+1}$, pour $m = 1, \ldots, M-1$, les limites ne sont pas dues à la procédure bootstrap, mais à la construction du test dont un a priori repose sur une approximation des distributions des statistiques des tests qui est valide pour un faible nombre de composantes $M = 6$ mais pas pour un plus grand nombre $M = 20$. Ainsi, les tests sur l'ensemble des $M(M-1)/2$ paires d'exposants $H_m = H_{m'}$, avec $1 \leq m < m' \leq M$, montrent des avantages considérables par rapport à ces méthodes.

Le partitionnement spectral du graphe des exposants H_m conçu à partir des $M(M-1)/2$ tests par paires gaussiens montre des performances très satisfaisantes, même pour une faible taille d'échantillon N et un grand nombre de composantes M. De plus, son comportement peut être finement ajusté par un paramètre de contrôle.

3.7 Conclusion

Ce chapitre propose différentes stratégies pour dénombrer et regrouper les valeurs distinctes dans le vecteur des exposants d'autosimilarité $\underline{H} = (H_1, \ldots, H_M)$ à partir de l'estimateur multivarié corrigé $\underline{\hat{H}} = (\hat{H}_1, \ldots, \hat{H}_M)$ introduit dans le chapitre 2. Ainsi, différents tests paramétriques sont construits en estimant leurs paramètres à l'aide d'une procédure de ré-échantillonnage bootstrap par blocs de coefficients d'ondelettes multivariés, permettant alors le partitionnement des exposants d'autosimilarité $\underline{H} = (H_1, \ldots, H_M)$ en partitions associées à des valeurs différentes. En outre, la procédure de ré-échantillonnage bootstrap permet également d'estimer les puissances de test. Les stratégies présentées sont toutes évaluées sur des réalisations de Monte Carlo de M-mBf synthétiques avec divers nombres de composantes M et tailles d'échantillon N.

En premier lieu, un test du χ^2 est conçu pour déterminer si tous les exposants d'autosimilarité sont égaux ou non. Ce test reproduit correctement l'hypothèse nulle dans diverses configurations et est puissant, mais peut être mis en défaut par des échelles d'analyse choisies grandes. Pour dénombrer et regrouper les valeurs des exposants lorsqu'elles sont multiples, une première stratégie consistant à effectuer $M-1$ tests d'égalité par paires $H_m = H_{m+1}$ d'exposants ordonnés successifs $H_1 \leq \ldots \leq H_M$ donne des résultats satisfaisants pour un faible nombre de composantes M, mais atteint des performances limitées lorsque le nombre de composantes M augmente. Cela est dû à la mauvaise approximation de la distribution des statistiques des tests sous les hypothèses nulles par paires $H_m = H_{m+1}$ par des lois demi-normales, a priori sur lequel sont construits les tests.

Un test sur l'ensemble des paires d'exposants $H_m = H_{m'}$, où $1 \leq m < m' \leq M$, est donc proposé pour remédier aux limites des différentes stratégies. Les distributions des statistiques des tests sont très bien approximées par des distributions gaussiennes, conséquence du comportement de l'estimateur multivarié corrigé $\underline{\hat{H}}$, aboutissant à $M(M-1)/2$ tests reproduisant fidèlement les hypothèses nulles $H_m = H_{m'}$ et puissants sous des hypothèses alternatives. Pour traiter les

$M(M-1)/2$ tests ainsi construits, une procédure de partitionnement spectral est effectuée sur un graphe des exposants H_m pondéré par une matrice de similarité construite à partir de ces tests. La difficulté de cette stratégie repose dans la définition d'une matrice de similarité du graphe. Une première matrice de similarité est construite à partir des p-valeurs des tests. Cette dernière procédure est efficace et s'avère être plus robuste que les précédentes à la diminution de la taille d'échantillon N et l'augmentation du nombre de composantes M. Pour paramétrer le graphe des exposants H_m et ainsi permettre d'adapter le comportement du partitionnement spectral du graphe, une construction originale de matrice de similarité est proposée, faisant suite à des travaux initiés par Bautista Ruiz (2019). Cette approche montre des performances supérieures à la précédente pour certaines plages de paramètres.

Chapitre 4

Grande dimension

Sommaire

4.1 Introduction

Jusqu'ici, les performances asymptotiques de l'estimateur multivarié corrigé $\hat{\underline{H}}^{(\mathrm{M,bc})}$ du vecteur des exposants d'autosimilarité $\underline{H} = (H_1, \ldots, H_M)$ proposé dans le chapitre 2 ont été étudiées théoriquement et empiriquement dans la limite asymptotique des grandes tailles d'échantillon N pour un nombre de composantes M fixe. Cependant, dans certaines applications pratiques, telles que la magnétoencéphalographie et l'imagerie par résonance magnétique fonctionnelle (Ciuciu et collab., 2012), le nombre de composantes M est grand comparé à la taille d'échantillon N. Dans ce cas, le nombre de composantes M est alors également grand comparé au nombre $n_j \approx N/2^j$ de coefficients d'ondelettes disponibles à l'échelle 2^j, en particulier aux grandes échelles, auxquelles doit être réalisée l'estimation par ondelettes, comme expliqué dans la section 1.4.3.3. Il est donc important d'étudier l'autosimilarité multivariée dans un autre régime asymptotique où le nombre de composantes M n'est plus fixe mais croît avec la taille d'échantillon N.

Pour cette raison, ce chapitre étudie à présent l'estimation lorsque le nombre de composantes M, la taille de l'échantillon N et les échelles d'analyse $2^{j_1}, \ldots, 2^{j_2}$ tendent conjointement vers l'infini de sorte que le nombre d'échelles d'analyse $j_2 - j_1 + 1$ est fixe et le rapport $M/(N/2^{j_2})$ est asymptotiquement constant. Ce régime asymptotique, dit de *grande dimension*, se résume ainsi à une *triple limite* (Abry et collab., 2022) :

$$\frac{M}{N/2^{j_2}} \underset{M,N,j_2 \to +\infty}{\longrightarrow} c \in [0, +\infty). \tag{4.1}$$

Une description plus formelle du régime asymptotique de la grande dimension est donnée dans l'annexe D.1. Lorsque $c = 0$, la taille d'échantillon N tend infiniment plus vite vers l'infini que M, rappelant le régime asymptotique où M est fixe et N tend vers l'infini, régime alors dit de *faible dimension* par opposition.

Ainsi, une étude empirique du comportement asymptotique de l'estimateur multivarié corrigé $\hat{\underline{H}}^{(\mathrm{M,bc})}$ en grande dimension est réalisée dans la section 4.2 à travers des simulations de Monte Carlo menées sur des M-mBf synthétiques de taille finie. Cette étude numérique rapporte d'abord la limite des outils de faible dimension du chapitre 3, incitant à revisiter les procédures de dénombrement de valeurs distinctes présentes dans $\underline{H}$ et l'estimation de la proportion de chacune de ces valeurs dans $\underline{H}$. Cette étude montre ensuite que le nombre de valeurs distinctes dans $\underline{H}$ est donné asymptotiquement par le nombre de modes dans la distribution des estimées multivariées corrigées $\hat{H}_m^{(\mathrm{M,bc})}$, comportement alors exploité pour développer de nouvelles procédures.

En premier lieu, la section 4.3 porte sur l'élaboration d'une procédure bootstrap pour tester la présence d'une unique valeur dans $\underline{H}$ à partir d'une seule observation de données multivariées. Celle-ci s'inspire d'un test d'unimodalité proposé par Orejola et collab. (2022), où le seuil de rejet est fixé à partir de M-mBf synthétiques. La conception d'une procédure bootstrap permet de pallier le désavantage de l'usage de M-mBf synthétiques, assurant l'utilisation du test sur cette seule observation sans étalonnage préalable. En second lieu, la section 4.4 développe une procédure pour compter les valeurs différentes dans $\underline{H}$ à partir de tests de multimodalité et propose une procédure d'estimation de chacune de ces valeurs et de leur proportion dans $\underline{H}$. La pertinence de ces procédures est évaluée numériquement sur des M-mBf synthétiques pour différentes tailles d'échantillon N et différents nombres de composantes M.

4.2 Étude empirique de l'estimation

4.2.1 Simulations de Monte Carlo

Pour étudier numériquement le comportement de l'estimateur multivarié corrigé $\hat{\underline{H}}^{(\mathrm{M,bc})}$, obtenu selon l'équation (2.3), dans des configurations de grande dimension, des expériences numériques sont menées sur $N_{\mathrm{MC}} = 100$ réalisations de M-mBf synthétiques (cf. Section 2.4).

Dans le contexte de la triple limite (4.1), nous considérons plusieurs nombres de composantes $M \in \{2^4, 2^5, 2^6\}$ et les régressions linéaires pour l'estimation sont effectuées sur des échelles $2^{j_1} \leq 2^j \leq 2^{j_2}$, avec $j_2 - j_1 = 2$, pour différentes tailles d'échantillon N de sorte que $c := M/(N/2^{j_2}) = M/n_{j_2}$ est fixé. Plus la valeur de c est proche de 1, plus le nombre de coefficients d'ondelettes n_{j_2} utilisés pour calculer les $\log_2 \bar{\lambda}_m(2^j)$ est faible comparé au nombre de composantes M. Puisque les outils de faible dimension ne sont censés être adaptées qu'à des valeurs de c proche de 0, le comportement de l'estimateur multivarié corrigé $\hat{\underline{H}}^{(\mathrm{M,bc})}$ est étudié pour deux différentes valeurs $c \in \{1/8, 1/4\}$. De plus, à c fixé, diverses relations entre M, N, j_1 et j_2 sont explorées. Plus précisément, deux gammes d'échelles d'analyse sont examinées à un rapport c fixé : les régressions linéaires sont effectuées soit de $2^{j_1} = M/4$ à $2^{j_2} = M$, soit de $2^{j_1} = M/8$ à $2^{j_2} = M/2$.

Les M entrées du vecteur des exposants d'autosimilarité $\underline{H}$ sont tirés selon une loi uniforme discrète sur le support $\{H_1, \ldots, H_L\}$ avec $L \in \{2, 3\}$, $H_L = 0.8$ et $H_l = H_{l+1} - \Delta H$ où $\Delta H \in (0, 0.2)$. Les matrices de covariance Σ des M-mBf sont les matrices identité $\mathbb{I}$ de taille $M \times M$ et leurs matrices de mélange W sont choisies au hasard parmi l'ensemble des matrices orthogonales de taille $M \times M$ et maintenues fixes pour chaque ensemble (M, N).

Pour la procédure d'estimation, la transformée en ondelettes multivariée discrète est calculée à l'aide de l'ondelette mère de Daubechies à $N_\psi = 2$ moments nuls (Daubechies, 1992).

4.2.2 Limite des outils de faible dimension

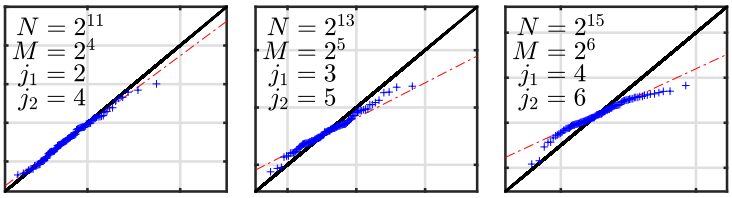
4.2.2.1 Normalité multivariée asymptotique

En faible dimension, i.e. à M fixé et N tendant vers l'infini, le vecteur $\hat{\underline{H}}^{(\mathrm{M,bc})}$ est asymptotiquement gaussien, sous des hypothèses faibles, comme énoncé par le théorème 2.3. En pratique, d'après les résultats rapportés dans la section 2.5, l'approximation de la distribution de $\hat{\underline{H}}^{(\mathrm{M,bc})}$ par une distribution normale multivariée est valide même à faible taille d'échantillon $N = 2^{13}$ pour $M = 6$ composantes. Ces résultats ont ainsi permis de construire les différentes procédures pour tester l'égalité entre les entrées de $\underline{H}$ dans le chapitre 3.

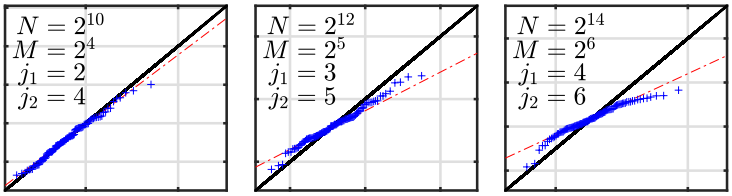
En pratique, le nombre de composantes M est parfois grand comparé à la taille d'échantillon N. On se propose donc d'évaluer l'approximation de la distribution de $\hat{\underline{H}}^{(\mathrm{M,bc})}$ par une loi normale multivariée dans diverses configurations à grands nombres de composantes M et faibles tailles d'échantillon N. Cela revient à vérifier si la distribution du carré de la distance de Mahalanobis de $\hat{\underline{H}}^{(\mathrm{M,bc})}$,

$$T := \left(\hat{\underline{H}}^{(\mathrm{M,bc})} - \mathbb{E}\left[\hat{\underline{H}}^{(\mathrm{M,bc})}\right]\right) \mathrm{Var}\left(\hat{\underline{H}}^{(\mathrm{M,bc})}\right)^{-1} \left(\hat{\underline{H}}^{(\mathrm{M,bc})} - \mathbb{E}\left[\hat{\underline{H}}^{(\mathrm{M,bc})}\right]\right)^T, \tag{4.2}$$

est bien approximée par une distribution du χ^2 à M degrés de liberté. La figure 4.1 rapporte les diagrammes quantile-quantile de la distribution empirique de T contre une loi du χ^2 à M degrés de liberté pour différents nombres de composantes M, tailles d'échantillon N et octaves d'analyse



(a) $c = 1/8$



(b) $c = 1/4$

FIGURE 4.1 – **Approximation par une loi normale.** Diagramme quantile-quantile de la distribution du carré de la distance de Mahalanobis de $\hat{\underline{H}}^{(\mathrm{M,bc})}$ au travers des réalisations de Monte Carlo contre une distribution du χ^2 à M degrés de liberté pour différents nombres de composantes M, tailles d'échantillon N et échelles d'analyses $2^{j_1}, \ldots, 2^{j_2}$ telles que $c = 1/8$ et $c = 1/4$ pour une unique valeur $H_1 = H_2$ présente dans $\underline{H}$.

(j_1, j_2) tels que $c \in \{1/8, 1/4\}$ et des valeurs toutes égales dans $\underline{H}$. La distribution de $\hat{\underline{H}}^{(\mathrm{M,bc})}$ est bien approximée par une loi normale multivariée pour $M = 2^4 = 16$ composantes avec des tailles d'échantillon $N \in \{2^{10}, 2^{11}\}$ mais par pour les plus grands nombres de composantes M.

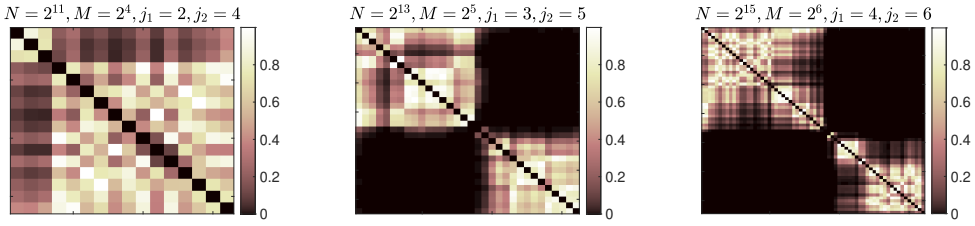
L'approximation de la distribution de $\hat{\underline{H}}^{(\mathrm{M,bc})}$ par une distribution normale multivariée n'est donc pas valable dans des configurations de grande dimension variées.

4.2.2.2 Partitionnement spectral

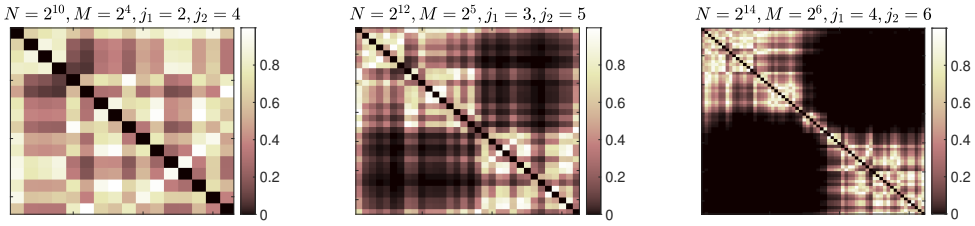
Pour illustrer plus clairement la limite des outils de faible dimension, on se propose d'appliquer la méthode de partitionnement spectral introduite dans la section 3.5 pour compter les exposants distinct dans $\underline{H}$ et regrouper les exposants égaux dans des configurations de grande dimension.

Pour rappel, le partitionnement spectral est réalisé sur un graphe des exposants d'autosimilarité $H_1, \ldots, H_M$ pondéré à partir des p-valeurs des tests par paires pour les hypothèses nulles $H_m = H_{m'}$, pour tous $m \neq m' \in \{1, \ldots, M\}$. La matrice de similarité S du graphe donnée par l'équation (3.52) représente cette pondération. La figure 4.2 donne des exemples de matrices de similarité S pour différents nombres de composantes M, tailles d'échantillon N et octaves d'analyse (j_1, j_2) tels que $c \in \{1/8, 1/4\}$ pour deux valeurs distinctes $H_1 \neq H_2$ présentes dans $\underline{H}$. Idéalement, la matrice de similarité S devrait être constituée de 2 blocs bien distincts.

Pour $M = 2^4$ et $N = 2^{11}$ avec $c = 1/8$ et $M = 2^4$ et $N = 2^{10}$ avec $c = 1/4$, l'approximation gaussienne de l'estimateur $\hat{\underline{H}}^{(\mathrm{M,bc})}$ est valide d'après la figure 4.1 mais la taille d'échantillon N est trop petite comparée à M pour obtenir de bonnes estimées bootstrap des écart-types $\hat{\sigma}_{m,m'}$ des statistiques $\hat{H}^{(\mathrm{M,bc})}_{m'} - \hat{H}^{(\mathrm{M,bc})}_{m}$ des tests, altérant ainsi l'estimation des p-valeurs. Pour des tailles d'échantillon N plus grandes, en plus de la mauvaise estimation des $\hat{\sigma}_{m,m'}$ par bootstrap, la distribution de l'estimateur $\hat{\underline{H}}^{(\mathrm{M,bc})}$ s'écarte d'une distribution normale multivariée d'après la figure 4.1, si bien que des poids du graphe associés à des exposants H_m et H'_m égaux sont souvent très faibles.



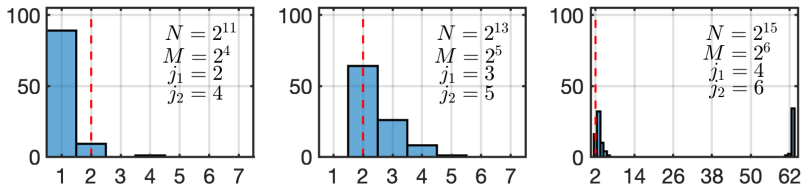
(a) $c = 1/8$



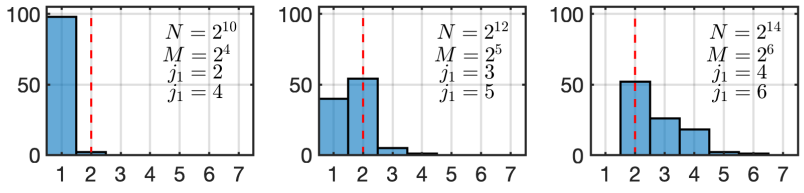
(b) $c = 1/4$

FIGURE 4.2 – **Exemples de matrice de similarité.** Matrices de similarité S pour une réalisation de Monte Carlo et différents nombres de composantes M, tailles d'échantillon N et échelles d'analyses $2^{j_1}, \ldots, 2^{j_2}$ tels que $c = 1/8$ et $c = 1/4$ pour deux valeurs $H_1 = 0.6$ et $H_2 = 0.8$ dans $\underline{H}$.

En complément, la figure 4.3 rapporte les histogrammes des nombres de partitions estimés pour différents nombres de composantes M, tailles d'échantillon N et octaves d'analyse (j_1, j_2) tels que $c = 1/8$ et $c = 1/4$ pour deux valeurs distinctes $H_1 \neq H_2$ présentes dans $\underline{H}$. Comme attendu au vu de la figure 4.2, le nombre de partitions est sous-estimé à faible taille d'échantillon ($N = 2^{11}$ et $N = 2^{10}$) et la surestimation du nombre de partitions augmente avec N, notamment pour $c = 1/8$. On observe en particulier que le nombre de partitions estimé peut être particulièrement grand : la plupart des entrées de $\underline{H}$ sont considérées comme des nœuds isolés par la procédure. Ceci est lié à un taux de rejet important des hypothèses $H_m = H'_m$ en conséquence des faibles p-valeurs obtenues, conséquence de la non-gaussianité de $\hat{\underline{H}}^{(\mathrm{M,bc})}$ et la mauvaise estimation des $\hat{\sigma}_{m,m'}$ par bootstrap.



(a) $c = 1/8$



(b) $c = 1/4$

FIGURE 4.3 – **Estimation du nombre de partitions par partitionnement spectral.** Histogrammes des nombre de partitions estimés par partitonnement spectral de la matrice de similarité S pour différents nombres de composantes M, tailles d'échantillon N et échelles d'analyses $2^{j_1}, \ldots, 2^{j_2}$ tels que $c = 1/8$ et $c = 1/4$ pour pour deux valeurs $H_1 = 0.6$ et $H_2 = 0.8$ dans $\underline{H}$. Les lignes pointillées rouges indiquent le nombre exact de partitions.

Cette procédure est donc inadaptée aux ordres de grandeurs de M et N considérés ici, pourtant réalistes en pratique. Ces résultats motivent la conception d'une nouvelle procédure de test, adaptée à la grande dimension.

4.2.3 Comportement asymptotique en grande dimension de l'estimation

Le cadre de la grande dimension fait l'objet d'un travail en cours de Gustavo Didier, mathématicien à l'université Tulane, et de son doctorant Oliver Orejola, détaillé en Annexe D.2. Ce travail a en particulier mené Orejola et collab. (2022) à conjecturer que le nombre de modes de la distribution empirique des M entrées de l'estimateur multivarié $\hat{H}_1^{(\mathrm{M})}, \ldots, \hat{H}_M^{(\mathrm{M})}$, obtenues à partir de l'équation (1.50), tend asymptotiquement vers le nombre de valeurs distinctes L dans $\underline{H}$ sous la triple limite définie par l'équation (4.1). Plus précisément, la probabilité de chaque mode tend vers la proportion de chacune des valeurs distinctes $H_1, \ldots, H_L$ dans $\underline{H}$. Ce résultat suggère de compter les valeurs présentes dans $\underline{H}$ lorsqu'elles sont multiples en testant la multimodalité de la distribution des estimées multivariées $\hat{H}_1^{(\mathrm{M})}, \ldots, \hat{H}_M^{(\mathrm{M})}$. C'est ainsi la stratégie adoptée par Orejola et collab. (2022) pour tester la présence d'une unique valeur dans $\underline{H}$. Il est naturel de conjecturer que le résultat précédent est également valable pour l'estimateur multivarié corrigé $\hat{\underline{H}}^{(\mathrm{M,bc})}$.

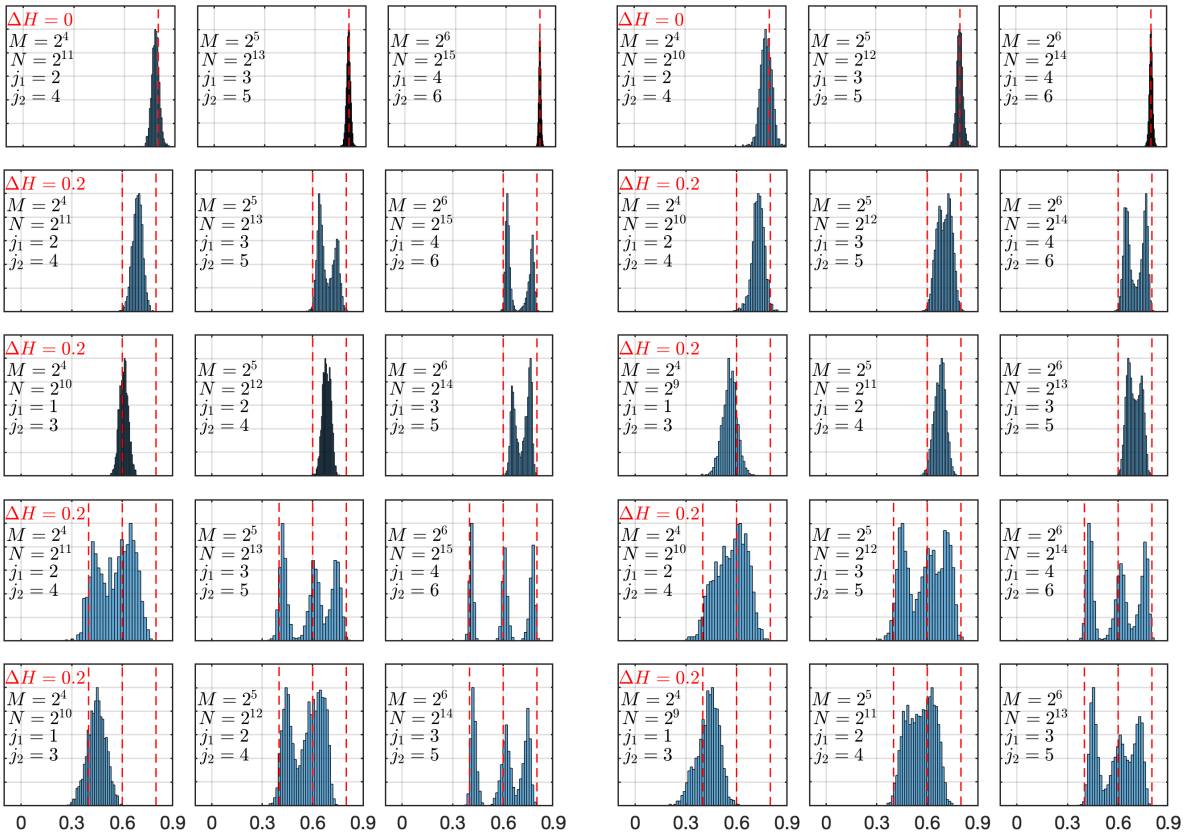
La validité de ce résultat est examinée numériquement pour différentes configurations de grande dimension. La figure 4.4 présente les histogrammes des entrées $\hat{H}_m^{(\mathrm{M,bc})}$ en fonction des entrées $m = 1, \ldots, M$ et des réalisations Monte Carlo pour différents écarts $\Delta H = |H_2 - H_1|$ entre les valeurs H_1 et H_2 présentes dans $\underline{H}$ et les deux valeurs de c avec différentes dimensions M, tailles d'échantillon N et octaves d'analyse (j_1, j_2).

Ces résultats montrent tout d'abord que, lorsque toutes les entrées de $\underline{H}$ sont égales, la distribution des $\hat{H}_m^{(\mathrm{M,bc})}$ est unimodale comme attendu, et que la variance de cette distribution décroît lorsque N, M, j_1 et j_2 augmentent à c fixé, si bien que cette distribution tend vers une distribution discrète de mode unique H_1. Ces résultats montrent également que, lorsqu'il y a $L > 1$ valeurs distinctes présentes dans $\underline{H}$, L modes apparaissent dans la distribution des $\hat{H}_m^{(\mathrm{M,bc})}$ lorsque N, M, j_1 et j_2 augmentent à c fixé, et les valeurs de ces deux modes tendent bien vers les valeurs H_1 et H_2. Ce comportement corrobore le comportement asymptotique en grande dimension des $\hat{H}_m^{(\mathrm{M,bc})}$ conjecturé ici.

En outre, lorsqu'il y a $L > 1$ valeurs distinctes présentes dans $\underline{H}$, on observe que

(i) la convergence de la distribution des $\hat{H}_m^{(\mathrm{M,bc})}$ vers une distribution discrète à L modes est plus rapide à $c = 1/8$ qu'à $c = 1/4$ pour des octaves d'analyse (j_1, j_2) comparables, c'est-à-dire pour $(j_1, j_2) = (2, 4)$ sur les seconde et quatrième lignes, et pour $(j_1, j_2) = (1, 3)$ sur les troisième et cinquième lignes ;

(ii) et cette convergence est similaire entre les deux valeurs de c à même taille d'échantillon $N = 2^{10}$, c'est-à-dire entre la troisième ligne à gauche et la deuxième ligne à droite, et entre la cinquième ligne à gauche et la quatrième ligne à droite.



(a) $c = 1/8$ (b) $c = 1/4$

FIGURE 4.4 – **Distributions des $\hat{\mathbf{H}}_{\mathbf{m}}^{(\mathbf{M,bc})}$ sous la triple limite.** Histogrammes des $\hat{H}_m^{(\mathrm{M,bc})}$ au travers des réalisations de Monte Carlo et des composantes $m = 1, \ldots, M$ pour différents nombres de composantes M, tailles d'échantillon N et échelles d'analyses $2^{j_1}, \ldots, 2^{j_2}$ tels que $c = 1/8$ et $c = 1/4$, et pour différents écarts $\Delta H = |H_{l+1} - H_l|$ entre (lignes rouges pointillées) les valeurs $H_1, \ldots, H_L$ présentes dans $\underline{H}$. *La distribution empirique des entrées de $\underline{\hat{H}}^{(\mathrm{M,bc})}$ tend asymptotiquement vers une distribution discrète de modes les différentes valeurs présentes dans $\underline{H}$.*

4.3 Test d'égalité entre les exposants d'autosimilarité

Cette section est une version étendue de l'article suivant : C.-G. LUCAS, P. ABRY, H. WENDT, G. DIDIER et O. OREJOLA, « Bootstrap based test for the unimodality of estimated Hurst exponents. Performance assessment in a high-dimensional analysis setting », *XXVIVème Colloque Francophone de Traitement du Signal et des Images (GRETSI 2023).*

L'objectif de cette section est de tester si tous les exposants d'autosimilarité sont égaux, c'est-à-dire détecter si $\underline{H}$ contient une seule ou bien plusieurs valeurs.

4.3.1 Test d'unimodalité

Puisque l'unimodalité de la distribution des entrées de $\underline{\hat{H}}^{(\mathrm{M,bc})}$ indique la présence d'une unique valeur dans $\underline{H}$ d'après la section 4.2.3, nous concevons une procédure de test d'unimodalité à partir de la distribution des M entrées $\hat{H}_m^{(\mathrm{M,bc})}$ pour tester l'hypothèse nulle

$$\mathcal{H}_0 : \quad \text{toutes les valeurs de } \underline{H} \text{ sont égales.} \tag{4.3}$$

Plusieurs procédures de test d'unimodalité existent. Suivant Orejola et collab. (2022), la procédure proposée ici repose sur le test d'unimodalité de Hartigan (*dip test*, Hartigan et Hartigan (1985)). Celle-ci consiste à étudier la forme de la fonction de répartition empirique $\hat{F}$ des entrées $\hat{H}_m^{(\mathrm{M,bc})}$ de $\hat{\underline{H}}^{(\mathrm{M,bc})}$, donnée par

$$\forall x \in \mathbb{R}, \quad \hat{F}(x) = \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}_{\left\{\hat{H}_m^{(\mathrm{M,bc})} \leq x\right\}}, \tag{4.4}$$

où $\mathbb{1}_{\{x \in A\}} = 1$ si $x \in A$ et $\mathbb{1}_{\{x \in A\}} = 0$ sinon, pour tout ensemble A.

Soit $\mathcal{U}$ l'ensemble de toutes les fonctions de répartition unimodales : une fonction de $\mathcal{U}$, de mode m, est convexe sur $(-\infty, m]$ et concave $[m, +\infty)$. La statistique de Hartigan, définie par

$$\hat{d} = \inf_{G \in \mathcal{U}} \sup_{x \in \mathbb{R}} |\hat{F}(x) - G(x)|, \tag{4.5}$$

mesure la déviation de la fonction de répartition empirique à une fonction de répartition unimodale. Cette statistique peut être calculée numériquement par l'algorithme AS 217 de Hartigan (1985) reposant sur le théorème 6 de Hartigan et Hartigan (1985), algorithme détaillé dans l'annexe B.3.

Pour un niveau de confiance $\alpha \in (0, 1)$, le test s'écrit

$$\text{rejeter } \mathcal{H}_0 \quad \text{si } \hat{d} > d_\alpha, \tag{4.6}$$

où d_α est un seuil de rejet défini par

$$\alpha = \sup_{H_0 \in (0,1)} \mathbb{P}\left(\hat{d} > d_\alpha \middle| \operatorname{supp}(\pi) = H_0\right). \tag{4.7}$$

Dans Orejola et collab. (2022), d_α est estimé en exploitant la distribution de $\hat{d}$ sous $\mathcal{H}_0$ à partir de M-mBf synthétiques. Dans le présent chapitre, on propose d'estimer d_α à partir d'une seule observation de données multivariées et donc une seule observation de la statistique $\hat{d}$, sans étalonnage préalable. On propose alors d'approximer le seuil d_α en utilisant la procédure boostrap par blocs d'ondelettes décrite dans le chapitre 3. Cette procédure boostrap permet de répliquer le comportement des données observées, évitant un étalonnage à partir de M-mBf synthétiques indépendants des données.

4.3.2 Procédure de test bootstrap

Suivant la procédure de la section 3.2, R échantillons bootstrap $\hat{H}_1^{*(\mathrm{M,bc},r)}, \ldots, \hat{H}_M^{*(\mathrm{M,bc},r)}$ sont obtenus à partir d'un échantillon de taille N et notés $\hat{H}_1^{*(r)}, \ldots, \hat{H}_M^{*(r)}$ pour simplifier la lecture. Ces échantillons sont centrés pour reproduire l'hypothèse nulle $\mathcal{H}_0$, comme suit :

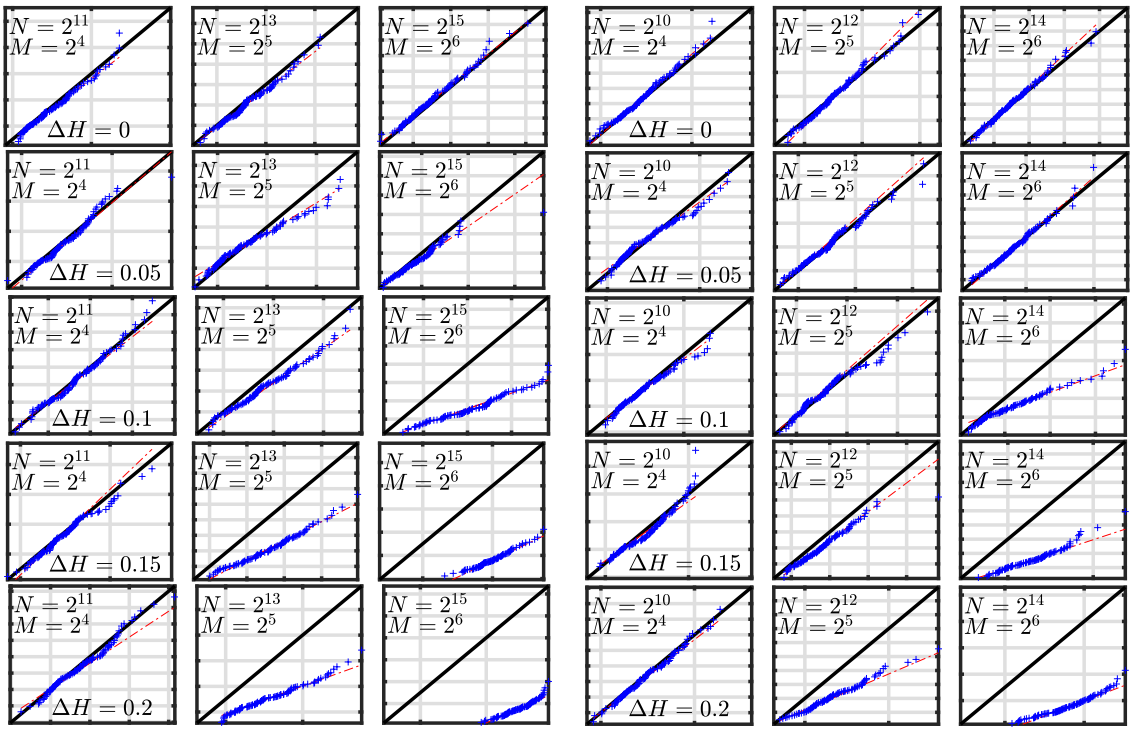
$$\forall r \in \{1, \ldots, R\},\ \forall m \in \{1, \ldots, M\}, \quad \bar{H}_m^{*(r)} = \hat{H}_m^{*(r)} - \langle \hat{H}_m^* \rangle, \tag{4.8}$$

où $\langle \hat{H}_m^* \rangle$ est la moyenne au travers des échantillons bootstrap,

$$\langle \hat{H}_m^* \rangle = \frac{1}{R} \sum_{r=1}^{R} \hat{H}_m^{*(r)}. \tag{4.9}$$

Centrer ainsi les échantillons bootstrap $\hat{H}_m^*$ permet de forcer les moyennes des distributions des $\bar{H}_m^*$ à être égales. Étant donné le comportement asymptotique de l'estimateur multivarié corrigé $\underline{\hat{H}}^{(\mathrm{M,bc})}$ observé dans la section 4.2.3 (donné plus formellement par l'équation (D.3)), on s'attend à ce que la distribution des entrées $\bar{H}_m^{*(r)}$ de chaque échantillons bootstrap $r \in \{1, \ldots, R\}$ soit unimodale. On calcule alors la fonction de répartition empirique $\hat{F}^{*(r)}$ des échantillons $\bar{H}_1^{*(r)}, \ldots, \bar{H}_M^{*(r)}$ pour chaque échantillon bootstrap $r \in \{1, \ldots, R\}$ selon l'équation (4.4), puis la statistique de test $\hat{d}^{*(r)}$ qui en résulte selon l'équation (4.5).

La statistique bootstrap $\hat{d}^*$ doit alors reproduire le comportement de la statistique de test $\hat{d}$ sous $\mathcal{H}_0$, ce qui est évalué numériquement. La figure 4.5 rapporte les diagrammes quantile-quantile de la distribution de la statistique $\hat{d}^*$ pour une réalisation de Monte Carlo contre la distribution de la statistique $\hat{d}$ au travers des réalisations de Monte Carlo pour différents nombres de composantes M, tailles d'échantillon N et échelles d'analyse $2^{j_1}, \ldots, 2^{j_2}$ de sorte que le rapport $c \in \{1/8, 1/4\}$ est fixé, et pour différents écarts $\Delta H = |H_2 - H_1|$ entre les valeurs H_1 et H_2 présentes dans $\underline{H}$. Ces tracés semblent confirmer que, asymptotiquement, la distribution de la statistique $\hat{d}^*$ reproduit bien la distribution de $\hat{d}$ sous $\mathcal{H}_0$ ($\Delta H = 0$) et s'en écarte sous des hypothèses alternatives ($\Delta H > 0$), et ce d'autant plus que c est petit.



(a) $c = 1/8$ (b) $c = 1/4$

FIGURE 4.5 – **Distributions des statistiques bootstrap $\hat{\mathbf{d}}^*$.** Diagrammes quantile-quantile de la distribution de la statistique bootstrap $\hat{d}^*$ pour une réalisation de Monte Carlo contre la distribution de la statistique $\hat{d}$ au travers des réalisations de Monte Carlo pour différents nombres de composantes M, tailles d'échantillon N et échelles d'analyse allant de $2^{j_1} = M/4$ à $2^{j_2} = M$ tels que $c = 1/8$ et $c = 1/4$, et pour (de haut en bas) différents écarts $\Delta H = |H_2 - H_1|$ entre les valeurs H_1 et H_2 présentes dans $\underline{H}$. *La distribution de la statistique bootstrap $\hat{d}^*$ approxime la distribution de la statistique $\hat{d}$ sous hypothèse nulle ($\Delta H = 0$) et s'en écarte sous des hypothèses alternatives.*

Finalement, le seuil d_α du test d'unimodalité (4.6) est estimé par $\hat{d}^*_\alpha = \hat{d}^{*(l)}$, avec $l = 1, \ldots, R$ tel que

$$\text{Card}\left(\left\{r \in \{1, \ldots, R\} |\, \hat{d}^{*(r)} \geq \hat{d}^{*(l)}\right\}\right) = \lfloor (1-\alpha)R \rfloor, \tag{4.10}$$

c'est-à-dire tel que $\hat{d}^{*(l)}$ est plus petit qu'un pourcentage α de statistiques bootstrap $\hat{d}^{*(1)}, \ldots, \hat{d}^{*(R)}$. Autrement dit, la p-valeur du test pour $\mathcal{H}_0$ est approximée par

$$\hat{p}^* := \frac{1}{R} \sum_{r=1}^{R} \mathbb{1}_{|\hat{d}| \leq |\hat{d}^{*(r)}|}, \tag{4.11}$$

où $\mathbb{1}_{\{x \in A\}} = 1$ si $x \in A$ et $\mathbb{1}_{\{x \in A\}} = 0$ sinon pour tout ensemble A, et le test s'écrit

$$\text{rejeter } \mathcal{H}_0 \quad \text{si } \hat{p}^* < \alpha. \tag{4.12}$$

La procédure de test d'unimodalité est résumée par la figure 4.6.

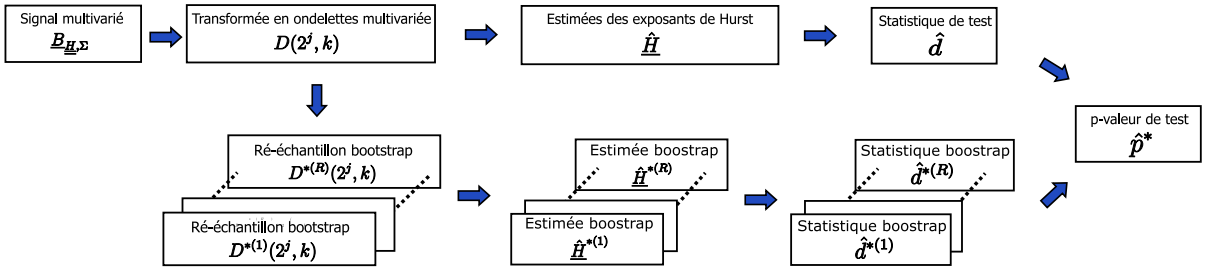


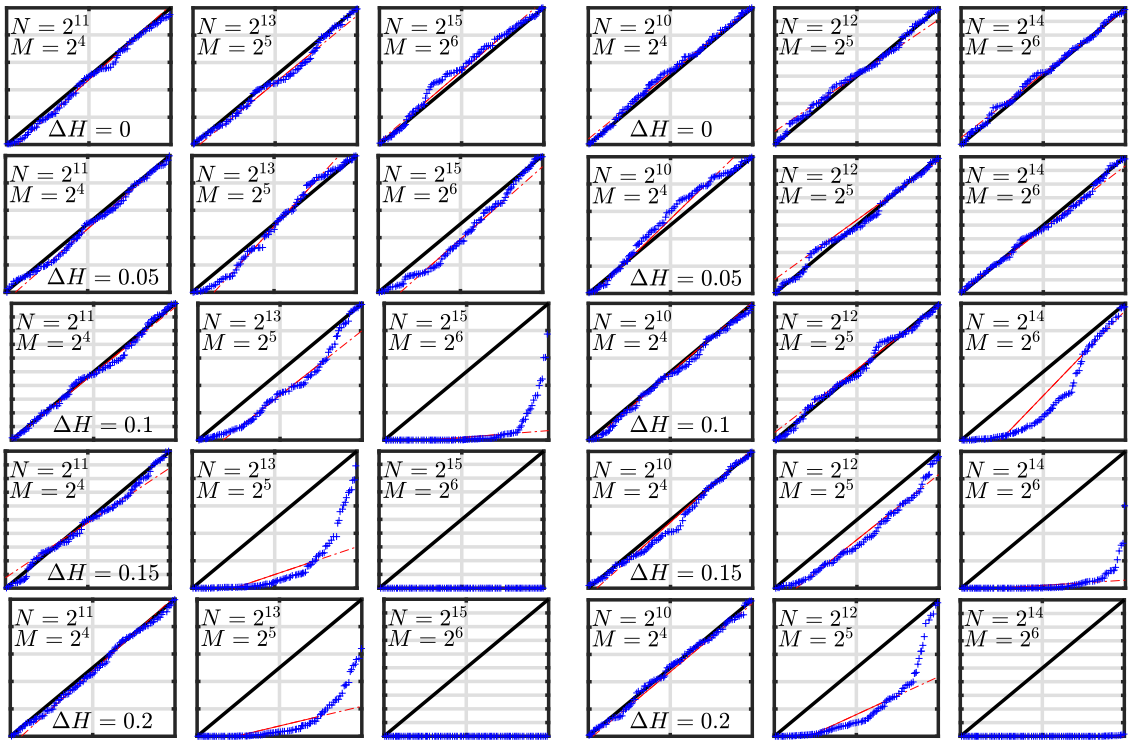
FIGURE 4.6 – **Organigramme de la procédure de test d'unimodalité.**

Pour s'assurer de la bonne construction des p-valeurs bootstrap $\hat{p}^*$, la figure 4.7 rapporte les diagrammes quantile-quantile de la distribution des p-valeurs $\hat{p}^*$ au travers des réalisations de Monte Carlo contre une distribution uniforme pour différents nombres de composantes M, tailles d'échantillon N et échelles d'analyse $2^{j_1}, \ldots, 2^{j_2}$ de sorte que le rapport $c \in \{1/8, 1/4\}$ est fixé et différents écarts $\Delta H = |H_2 - H_1|$ entre les valeurs H_1 et H_2 présentes dans $\underline{H}$. Ces tracés montrent que les p-valeurs bootstrap $\hat{p}^*$ sont bien approximées par des distributions uniformes sous $\mathcal{H}_0$ comme attendu théoriquement, et s'éloignent de distributions uniformes sous des hypothèses alternatives quand N et ΔH augmentent. De plus, cet éloignement est plus rapide pour $c = 1/8$ que pour $c = 1/4$. Ces résultats confirment que les p-valeurs bootstrap $\hat{p}^*$ ont bien le comportement voulu.

4.3.3 Performances du test

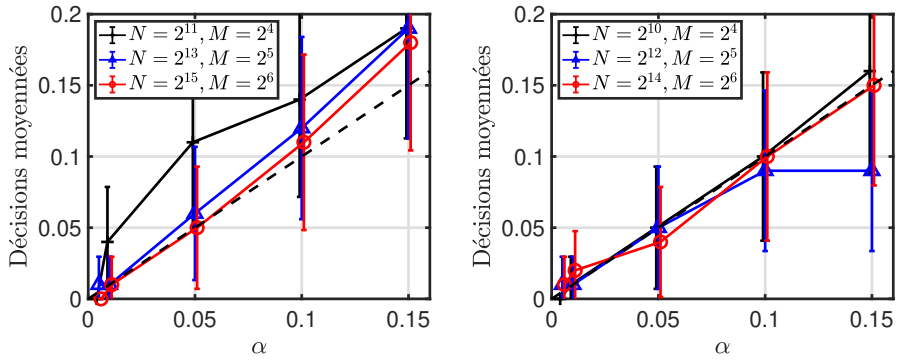
Les performances du test sont évaluées sous la triple limite définie par l'équation (4.1) à partir des simulations de Monte Carlo décrites dans la section 4.2.1. Pour la procédure bootstrap, $R = 500$ blocs de taille $L_B = 4$ (taille du support de l'ondelette) sont ré-échantillonnés à partir des coefficients d'ondelettes multivariés.

Pour vérifier la constuction adaptée de la procédure de test bootstrap, nous évaluons d'abord le comportement de la procédure de test sous l'hypothèse nulle $\mathcal{H}_0$. La figure 4.8 rapporte les décisions $\hat{d} > \hat{d}^*_\alpha$ du test bootstrap de rejet de l'hypothèse nulle $\mathcal{H}_0$ moyennées sur les réalisations de Monte Carlo (avec un intervalle de confiance à 95%) en fonction du niveau de confiance prédéfini α pour différentes triples limites c liées à différentes tailles d'échantillon N, nombres de composantes M et échelles d'analyse $2^{j_1}, \ldots, 2^{j_2}$. Pour les deux valeurs de c, le niveau de confiance ciblé α est bien reconstruit par la procédure bootstrap.



(a) $c = 1/8$ (b) $c = 1/4$

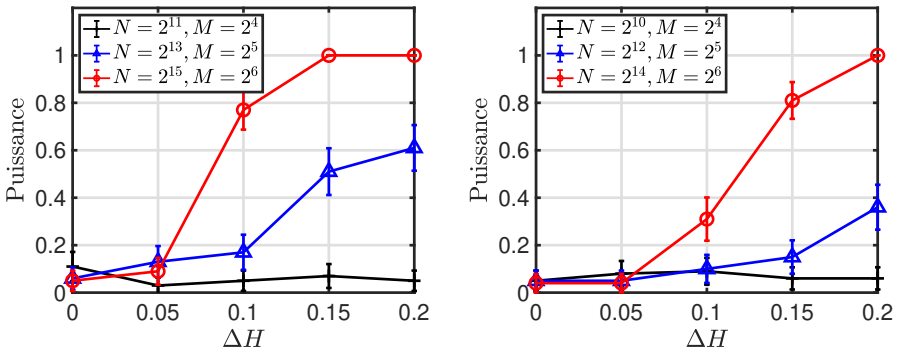
FIGURE 4.7 – **Distributions des p-valeurs bootstrap $\hat{\mathbf{p}}^*$.** Diagrammes quantile-quantile des distributions des $\hat{p}^*$ au travers des réalisations de Monte Carlo contre des distributions uniformes pour différents nombres de composantes M, tailles d'échantillon N et échelles d'analyse allant de $2^{j_1} = M/4$ à $2^{j_2} = M$ tels que $c = 1/8$ et $c = 1/4$, et pour (de haut en bas) différents écarts $\Delta H = |H_2 - H_1|$ entre les valeurs H_1 et H_2 présentes dans $\underline{H}$. *La distribution des p-valeurs bootstrap $\hat{p}^*$ est bien approximée par une distribution uniforme sous hypothèse nulle ($\Delta H = 0$) et s'en écarte sous des hypothèses alternatives.*



(a) $c = 1/8$ (b) $c = 1/4$

FIGURE 4.8 – **Niveaux de confiance.** Décisions de rejet du test moyennées sur les réalisations de Monte Carlo (avec un intervalle de confiance à 95%) par rapport au niveau de confiance prédéfini α sous l'hypothèse nulle $\mathcal{H}_0$ pour nombres de composantes M, tailles d'échantillon N et échelles d'analyse $2^{j_1}, \ldots, 2^{j_2}$ tels que $c = 1/8$ et $c = 1/4$.

Enfin, on quantifie la puissance du test. La figure 4.9 présente la puissance empirique du test (moyenne de Monte Carlo avec un intervalle de confiance à 95%) en fonction de l'écart $\Delta H = |H_2 - H_1|$ entre les valeurs H_1 et H_2 présentes dans $\underline{H}$ pour un niveau de confiance prédéfini $\alpha = 0.05$ et différentes triples limites c maintenues fixes pour différentes tailles d'échantillon N, nombres de composantes M et échelles d'analyse $2^{j_1}, \ldots, 2^{j_2}$. Ces résultats montrent tout d'abord que, à un rapport fixé c, la puissance augmente avec la taille de l'échantillon N, ce qui est en accord avec le comportement multimodal observé en pratique dans la section 4.2.3. En outre, pour un rapport M/N fixe (comparaison entre la ligne noire avec '+' de gauche et la ligne bleue avec 'Δ' de droite, et entre la ligne bleue avec 'Δ' de gauche et la ligne rouge avec 'o' de droite), la puissance augmente avec M, c'est-à-dire pour un plus grand nombre d'échantillons dans le calcul de la fonction de répartition empirique (4.4) impliquée dans le test d'unimodalité.



(a) $c = 1/8$ (b) $c = 1/4$

FIGURE 4.9 – **Puissance du test.** Proportions des décisions de rejet (moyenne de Monte Carlo avec un intervalle de confiance à 95%) par rapport à l'écart $\Delta H = |H_2 - H_1|$ entre les valeurs H_1 et H_2 présentes dans $\underline{H}$ pour un niveau de confiance prédéfini $\alpha = 0.05$, et différents nombres de composantes M, tailles d'échantillon N et échelles d'analyse allant de $2^{j_1} = M/4$ à $2^{j_2} = M$ tels que $c = 1/8$ et $c = 1/4$.

4.4 Dénombrement d'exposants d'autosimilarité

Lorsque plusieurs valeurs sont présentes dans le vecteur des exposants d'autosimilarité $\underline{H}$, il devient nécessaire de compter ces valeurs. D'après la section 4.2.3, le nombre de modes de la distribution des entrées de l'estimateur multivarié corrigé $\underline{\hat{H}}^{(\mathrm{M,bc})}$ correspond au nombre de valeurs distinctes dans $\underline{H}$. Une procédure de test de multimodalité pour compter les exposants distincts est donc décrite dans cette section, puis exploitée pour estimer leurs valeurs et leurs proportions dans $\underline{H}$.

La stratégie présentée dans cette section résulte également d'une collaboration avec Gustavo DIDIER et Oliver OREJOLA.

4.4.1 Tests de multimodalité

4.4.1.1 Formulation des tests

Pour compter le nombre L d'exposants distincts dans $\underline{H}$, la procédure proposée ici repose sur le test de multimodalité de Silverman (SILVERMAN, 1981) appliqué au vecteur $\underline{\hat{H}}^{(\mathrm{M,bc})}$ pour tester l'hypothèse nulle

$$\mathcal{H}_0^{(k)} : \quad \underline{H} \text{ contient au plus } k \text{ valeurs distinctes}, \tag{4.13}$$

pour tout entier $k \geq 2$. L'hypothèse nulle $\mathcal{H}_0^{(k)}$ est donc vraie pour tout $k \geq L$. D'après le comportement étudié dans la section 4.2.3, tester l'hypothèse $\mathcal{H}_0^{(k)}$ peut se ramener à tester l'hypothèse

$$g \text{ a au plus } k \text{ modes}, \tag{4.14}$$

où g est la densité de probabilité des échantillons $\hat{H}_1^{(\mathrm{M,bc})}, \ldots, \hat{H}_M^{(\mathrm{M,bc})}$. Une méthode naturelle pour compter les exposants distincts dans $\underline{H}$ consiste finalement à identifier le plus petit entier $k \geq 2$ pour lequel l'hypothèse $\mathcal{H}_0^{(k)}$ n'est pas rejetée.

4.4.1.2 Statistiques des tests

Fixons $k \geq 2$. La procédure de test de Silverman exploite l'estimateur par noyau gaussien de la densité g à partir de M observations $\underline{\hat{H}} = (\hat{H}_1, \ldots, \hat{H}_M)$, défini par

$$\forall h \in \mathbb{R}, \quad \hat{f}_{\underline{\hat{H}},s}(h) = \frac{1}{M\,s} \sum_{m=1}^{M} \frac{1}{\sqrt{2\pi}} e^{-\frac{(h-\hat{H}_m)^2}{2s^2}}, \tag{4.15}$$

où $s > 0$. La statistique de Silverman est définie comme le paramètre critique $s_{\mathrm{crit}}^{(k)}$ pour lequel l'estimateur par noyau de la densité a au plus k modes,

$$s_{\mathrm{crit}}^{(k)} = \inf \left\{ s \,\middle|\, \hat{f}_{\underline{\hat{H}}^{(\mathrm{M,bc})},s} \text{ a au plus } k \text{ modes} \right\}. \tag{4.16}$$

La statistique de Silverman peut être calculée par dichotomie sur le nombre de modes de $\hat{f}_{\underline{\hat{H}}^{(\mathrm{M,bc})},s}$, en vertu de l'équivalence suivante, démontrée dans Silverman (1981) :

$$s < s_{\mathrm{crit}}^{(k)} \Leftrightarrow \hat{f}_{\underline{\hat{H}}^{(\mathrm{M,bc})},s} \text{ a plus de } k \text{ modes}. \tag{4.17}$$

Ainsi, plus s est grand, plus le noyau gaussien de l'estimateur de la densité (4.15) est étiré et moins la densité estimée $\hat{f}_{\underline{\hat{H}},s}$ a alors de maxima locaux. La statistique de Silverman $s_{\mathrm{crit}}^{(k)}$ mesure ainsi la contraction du noyau gaussien nécessaire pour faire apparaître plus de k modes dans la densité estimée, et l'hypothèse nulle $\mathcal{H}_0^{(k)}$ est rejetée pour de grandes valeurs de $s_{\mathrm{crit}}^{(k)}$.

4.4.1.3 P-valeurs des tests

On fixe $k \geq 2$ et on se propose de recourir la procédure de Silverman (1981) pour estimer la p-valeur du test de Silverman pour l'hypothèse nulle $\mathcal{H}_0^{(k)}$ à partir d'une unique observation de données multivariées.

Écriture des p-valeurs

Tout d'abord, l'équivalence (4.17) a permis à Silverman de proposer une ré-écriture partique de la p-valeur du test de Silverman. On considère une observation $s_{\mathrm{crit}}^{(\mathrm{obs},k)}$ de la statistique de test issue de l'estimateur $\underline{\hat{H}}^{(\mathrm{M,bc})}$. En notant $s_{\mathrm{crit}}^{(k)}$ la statistique de Silverman associée à un vecteur $\underline{\hat{H}} = (\hat{H}_1, \ldots, \hat{H}_M)$ dont les entrées sont tirées selon la distribution g des $\hat{H}_m^{(\mathrm{M,bc})}$, la p-valeur du test de Silverman pour k modes est définie par

$$p_k := \mathbb{P}\left(s_{\mathrm{crit}}^{(\mathrm{obs},k)} < s_{\mathrm{crit}}^{(k)} \,\middle|\, \mathcal{H}_0^{(k)} \right). \tag{4.18}$$

D'après l'équation (4.17), cette p-valeur peut être réécrite comme suit :

$$p_k := \mathbb{P}\left(\hat{f}_{\underline{\hat{H}}, s_{\text{crit}}^{(\text{obs},k)}} \text{ a plus de } k \text{ modes} \,\middle|\, \underline{\hat{H}} = (\hat{H}_1, \ldots, \hat{H}_M) \text{ tirés selon } g \text{ sous } \mathcal{H}_0^{(k)} \right). \quad (4.19)$$

Cette écriture de la p-valeur évite d'avoir à calculer la statistique de Silverman par dichotomie pour chaque échantillon $\underline{\hat{H}} = (\hat{H}_1, \ldots, \hat{H}_M)$.

Estimation bootstrap des p-valeurs

La p-valeur (4.19) du test de Silverman peut être approximée à partir de R ré-échantillons bootstrap $\underline{\bar{H}}^{*(r)} = (\bar{H}_1^{*(r)}, \ldots, \bar{H}_M^{*(r)})$ tirés selon la procédure de Silverman (1981) comme suit :

$$\forall r \in \{1, \ldots, R\}, \forall m \in \{1, \ldots, M\}, \quad \bar{H}_m^{*(r)} = \frac{\hat{\sigma}_{\underline{\hat{H}}^{(\text{M,bc})}}}{\sqrt{\hat{\sigma}^2_{\underline{\hat{H}}^{(\text{M,bc})}} + \left(s_{\text{crit}}^{(\text{obs},k)}\right)^2}} \left(\hat{H}_{I^{(r)}(m)}^{(\text{M,bc})} + s_{\text{crit}}^{(\text{obs},k)} \, \varepsilon_m^{(r)} \right), \quad (4.20)$$

où les $I^{(r)}(m)$ sont tirés uniformément avec remise dans $\{1, \ldots, M\}$, le vecteur $(\varepsilon_1^{(r)}, \ldots, \varepsilon_M^{(r)})$ est un bruit blanc gaussien standard et $\hat{\sigma}_{\hat{H}^{(\text{M,bc})}}$ est l'écart-type (empirique) des entrées de $\underline{\hat{H}}^{(\text{M,bc})}$. Autrement dit, les $\bar{H}_m^{*(r)}$ sont des ré-échantillons tirés uniformément parmi $\hat{H}_1^{(\text{M,bc})}, \ldots, \hat{H}_M^{(\text{M,bc})}$ avec un bruit blanc gaussien additif d'écart-type la statistique de Silverman $s_{\text{crit}}^{(\text{obs},k)}$. Le coefficient multiplicateur en amont dans l'équation (4.20) est un facteur de remise à l'échelle, permettant l'égalité $\text{Var}^*\left(\bar{H}_m^*\right) = \text{Var}\left(\hat{H}_m^{(\text{M,bc})}\right)$ pour tout $m \in \{1, \ldots, M\}$. Ainsi, la distribution des échantillons $\bar{H}_1^{*(r)}, \ldots, \bar{H}_M^{*(r)}$ approxime la distribution $\hat{f}_{\underline{\hat{H}}^{(\text{M,bc})}, s_{\text{crit}}^{(\text{obs},k)}}$.

L'estimée bootstrap de la p-valeur p_k donnée par l'équation (4.19) est alors définie par

$$\hat{p}_k^* := \frac{\text{Card}\left(\left\{ r \in \{1, \ldots, R\} \,\middle|\, \hat{f}_{\underline{\bar{H}}^{*(r)}, s_{\text{crit}}^{(\text{obs},k)}} \text{ a plus de } k \text{ modes} \right\} \right)}{R}. \quad (4.21)$$

Puisque la distribution des échantillons $\bar{H}_1^{*(r)}, \ldots, \bar{H}_M^{*(r)}$ n'approxime pas la distribution g des $\hat{H}_1^{(\text{M,bc})}, \ldots, \hat{H}_M^{(\text{M,bc})}$, la distribution des p-valeurs bootstrap (4.21) peut s'écarter d'une distribution uniforme. En effet, Hall et York (2001) ont mis en évidence le caractère conservatif du test de Silverman dans un cadre assez général, l'estimateur bootstrap de la p-valeur p_k n'étant pas consistant. La procédure de test de Silverman est résumée par la figure 4.10.

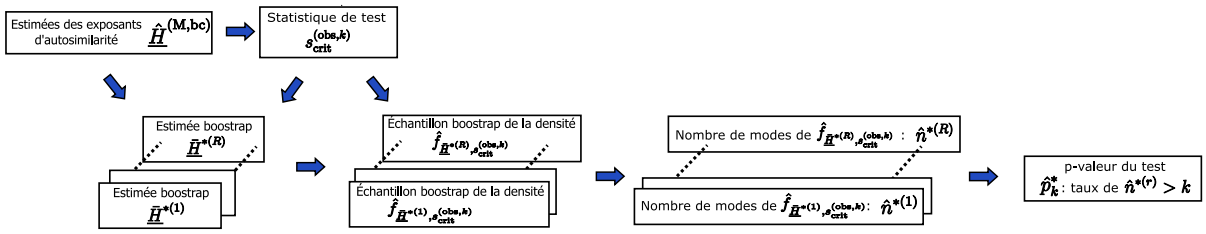


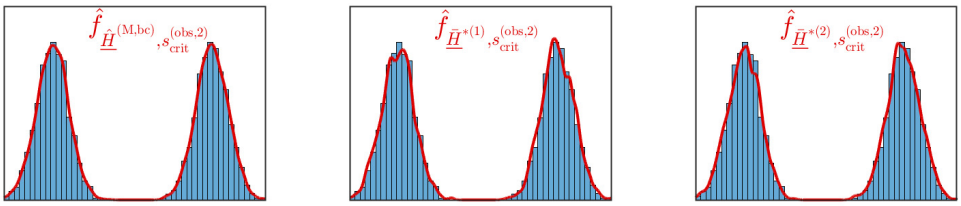
FIGURE 4.10 – **Organigramme de la procédure de test de Silverman à $k \geq 2$ fixé.**

La stratégie d'estimation bootstrap des p-valeurs de Silverman est illustrée par la figure 4.11 pour l'hypothèse nulle $\mathcal{H}_0^{(2)}$ (au plus $k = 2$ modes) dans deux cas différents :

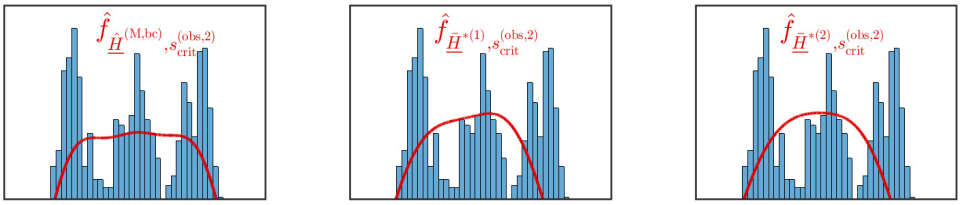
(*i*) lorsqu'il y a $L = 2$ modes dans la distribution des $\hat{H}_m^{(\text{M,bc})}$, situation où $\mathcal{H}_0^{(2)}$ est vraie (a) ;

(*ii*) lorsqu'il y a $L = 3$ modes dans la distribution des $\hat{H}_m^{(\mathrm{M,bc})}$, situation où $\mathcal{H}_0^{(2)}$ n'est pas vraie (b).

La figure 4.11 montre ainsi la densité estimée par noyau gaussien avec le paramètre critique $s_{\mathrm{crit}}^{(\mathrm{obs},2)}$, i.e. la statistique de Silverman pour tester $\mathcal{H}_0^{(2)}$, sur la première colonne et des densités estimées par noyau gaussien à partir des deux ré-échantillons bootstrap $\underline{\bar{H}}^{*(1)}$ et $\underline{\bar{H}}^{*(2)}$ sur la seconde colonne. La p-valeur bootstrap $\hat{p}_2^*$, qui correspond au taux de densités bootstrap ayant plus de 2 modes, doit être élevée dans le cas (*i*) et faible dans le cas (*ii*). Pour $L = 2$, la densité estimée approxime bien la distribution des entrées $\hat{H}_m^{(\mathrm{M,bc})}$, et les deux densités bootstrap associées représentées sont assez proches de cette dernière mais ont plus de 2 modes, suggérant une p-valeur $\hat{p}_2^*$ élevée. Pour $L = 3$, la densité estimée approxime très mal la distribution des entrées $\hat{H}_m^{(\mathrm{M,bc})}$, et les densités bootstrap associées sont également de mauvaises approximations de cette distribution et n'ont qu'un seul mode. Le nombre de modes des densités bootstrap ainsi obtenues peut difficilement excéder 2, entraînant une p-valeur $\hat{p}_2^*$ faible.



(a) $L = 2$ modes dans la distribution des $\hat{H}_m^{(\mathrm{M,bc})}$



(b) $L = 3$ modes dans la distribution des $\hat{H}_m^{(\mathrm{M,bc})}$

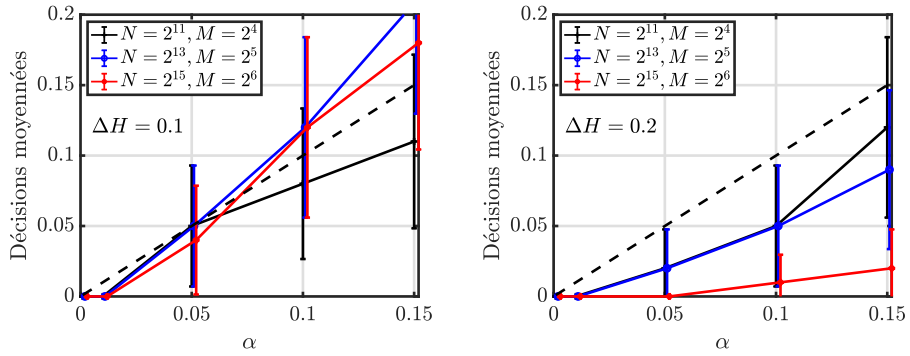
FIGURE 4.11 – **Illustration de la procédure bootstrap de Silverman pour $\mathcal{H}_0^{(2)}$ (au plus $k = 2$ modes).** Pour différents nombres L de modes dans la distribution des $\hat{H}_m^{(\mathrm{M,bc})}$ (de haut en bas), la densité estimée par noyau gaussien à partir de l'estimateur $\underline{\hat{H}}^{(\mathrm{M,bc})}$ avec un paramètre $s_{\mathrm{crit}}^{(\mathrm{obs},2)}$ correspondant à la statistique de Silverman pour $k = 2$ modes (à gauche) est ré-échantillonnée par bootstrap à partir de deux ré-échantillons bootstrap $\underline{\bar{H}}^{*(1)}$ et $\underline{\bar{H}}^{*(2)}$ de l'estimateur $\underline{\hat{H}}^{(\mathrm{M,bc})}$ (au milieu et à droite).

4.4.1.4 Reproduction de l'hypothèse nulle pour 2 modes

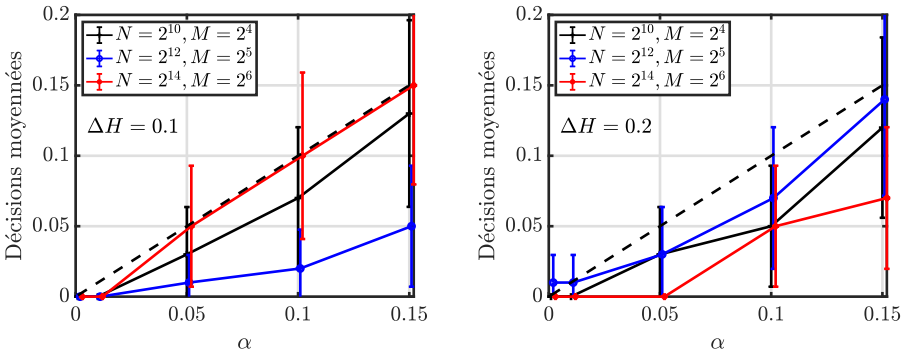
On se propose d'étudier ici le test de Silverman du rejet de l'hypothèse nulle $\mathcal{H}_0^{(2)}$ (au plus 2 modes). Lorsque le vecteur des exposants d'autosimilarité $\underline{H}$ contient exactement $L = 2$ valeurs distinctes, toutes les hypothèses nulles $\mathcal{H}_0^{(k)}$, avec $k \geq L = 2$, sont vraies. En particulier, $\mathcal{H}_0^{(2)}$ est vraie et devrait être rejetée à tort à un taux égal au niveau de confiance α.

La figure 4.12 rapporte les proportions de décisions de rejet $\hat{p}_k^* < \alpha$ de l'hypothèse nulle $\mathcal{H}_0^{(2)}$ (obtenues par moyennes sur les réalisations de Monte Carlo) pour différents écarts $\Delta H = |H_2 - H_1|$ entre les valeurs H_1 et H_2 présentes dans $\underline{H}$ et différentes triples limites c liées à différents nombres de composantes M, tailles d'échantillon N et échelles d'analyse $2^{j_1}, \ldots, 2^{j_2}$. Pour $\Delta H = 0.1$, le taux de fausse alarmes réel reconstruit correctement le taux de fausses alarmes ciblé α lorsque $c = 1/8$, même pour de faibles nombre de composantes M et taille d'échantillon N, en l'occurrence $N = 2^{11}$ et $M = 2^4$. En revanche, lorsque $c = 1/4$, cette reconstruction

n'est bonne que pour de plus grands nombre de composantes M et taille d'échantillon N. Pour $\Delta H = 0.2$, le taux de fausses alarmes ciblé α est mal reconstruit par le taux de fausse alarmes réel et ceci empire à mesure que le nombre de composantes M et la taille d'échantillon N augmentent. Le test de Silverman peut donc être très conservatif, le taux de fausse alarmes réel étant parfois très inférieur au niveau de confiance α, conséquence de la distribution non uniforme des p-valeurs.



(a) $c = 1/8$



(b) $c = 1/4$

FIGURE 4.12 – **Niveaux de confiance pour 2 modes.** Proportions des décisions de rejet (moyenne de Monte Carlo avec un intervalle de confiance à 95%) de l'hypothèse nulle $\mathcal{H}_0^{(2)}$ (au plus 2 modes) en fonction du niveau de confiance α pour deux différents écarts $\Delta H = |H_2 - H_1|$ entre les deux valeurs H_1 et H_2 présentes dans $\underline{H}$, et différents nombres de composantes M, tailles d'échantillon N et échelles d'analyse allant de $2^{j_1} = M/4$ à $2^{j_2} = M$ tels que $c = 1/8$ et $c = 1/4$.

Ces résultats signifient que lorsque $\underline{H}$ contient exactement $L = 2$ valeurs distinctes, le test de Silverman pour $k = 2$ modes va permettre de les détecter correctement avec un taux supérieur à $1 - \alpha$. Des cas où $L > 2$ sont étudiés dans la section 4.4.3.2.

4.4.2 Stratégie d'estimation

Le nombre de valeurs distinctes dans le vecteur des exposants d'autosimilarité $\underline{H}$ correspond au plus petit entier $k \geq 2$ pour lequel l'hypothèse nulle $\mathcal{H}_0^{(k)}$ (au plus k valeurs distinctes dans $\underline{H}$) est vraie. Ce nombre peut donc être estimé par

$$\widehat{L} = \min \{k \geq 2 \,|\, \hat{p}_k^* \geq \alpha\}, \tag{4.22}$$

où les $\hat{p}_k^*$ sont les p-valeurs bootstrap données par l'équation (4.21) pour tout $k \geq 2$ et le niveau de confiance est fixé à $\alpha = 0.05$. Ainsi, pour obtenir $\widehat{L}$, il suffit de parcourir les entiers $k \geq 2$ dans l'ordre croissant jusqu'à ce que l'hypothèse nulle $\mathcal{H}_0^{(k)}$ ne soit pas rejetée, comme détaillé par l'algorithme 1.

Algorithme 1: Estimation du nombre L de valeurs dans $\underline{H}$

Entrées : estimées $\hat{\underline{H}}^{(\mathrm{M,bc})} = (\hat{H}_1^{(\mathrm{M,bc})}, \ldots, \hat{H}_M^{(\mathrm{M,bc})})$, niveau de confiance α ;
Initialisation : $\hat{p}_1^* = 0,\ k = 1$;
Tant que $\hat{p}_k^* < \alpha$ **faire**
    $k = k + 1$;
    Calculer la statistique $s_{\mathrm{crit}}^{(\mathrm{obs},k)}$ associée à $\hat{\underline{H}}^{(\mathrm{M,bc})}$ selon l'équation (4.16) ;
    Tirer des échantillons bootstrap $\bar{\underline{H}}^{*(1)}, \ldots, \bar{\underline{H}}^{*(R)}$ selon l'équation (4.20) ;
    Calculer la p-valeur bootstrap $\hat{p}_k^*$ selon l'équation (4.21) ;
Fin
Sortie : nombre de modes estimé $\hat{L} = k$

Les résultats donnés dans la section 4.2.3 sur le comportement des entrées de $\hat{\underline{H}}^{(\mathrm{M,bc})}$ suggèrent la stratégie suivante pour estimer les proportions respectives $\pi(H_1), \ldots, \pi(H_L)$ des valeurs distinctes $H_1, \ldots, H_L$ dans $\underline{H}$. Le vecteur des estimées multivariées corrigées $\hat{\underline{H}}^{(\mathrm{M,bc})}$ est partitionné par l'algorithme des k-moyennes (Steinhaus, 1957), où le nombre de classes k est donné par le nombre de modes estimé $\widehat{L}$. La moyenne des éléments de chaque partition donne des estimées $\hat{H}_1, \ldots, \hat{H}_{\widehat{L}}$ des valeurs distinctes présentes dans $\underline{H}$ et la proportion d'éléments dans chaque partition donne des estimées de leurs proportions $\hat{\pi}(\hat{H}_1), \ldots, \hat{\pi}(\hat{H}_{\widehat{L}})$.

4.4.3 Performances empiriques

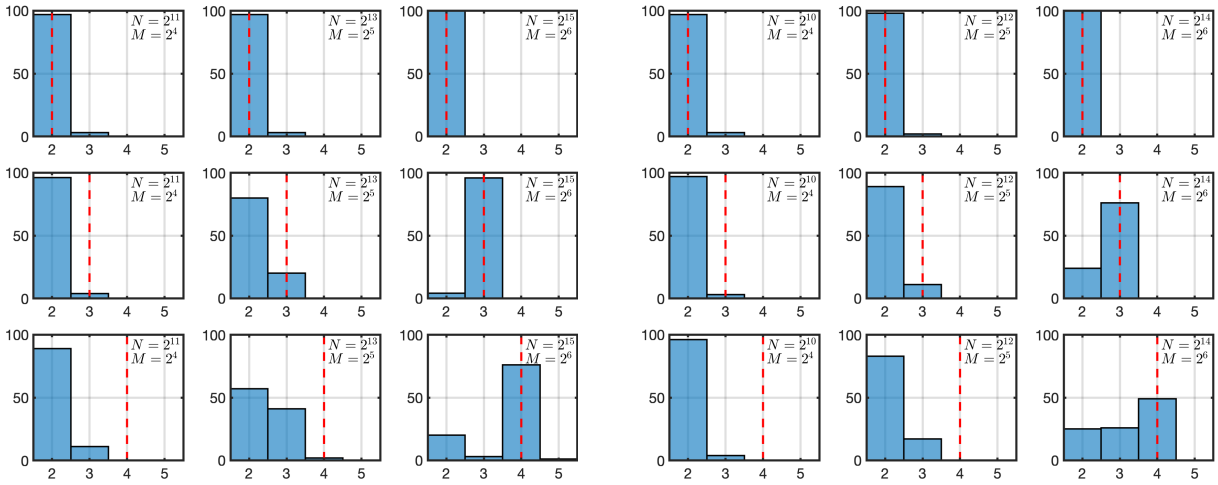
4.4.3.1 Simulations de Monte Carlo

Pour l'étude empirique des performances d'estimation, des simulations de Monte Carlo sont réalisées sur des M-mBf synthétiques selon la configuration présentée dans la section 4.2.1. Différents vecteurs d'exposants d'autosimilarité $\underline{H}$ sont étudiés : les M entrées de $\underline{H}$ sont tirées uniformément dans le support $\{H_1, \ldots, H_L\}$ avec $L \in \{2, 3, 4\}$, $H_L = 0.8$ et $H_l = H_{l+1} - 0.2$ pour tout $l \in \{1, \ldots, L-1\}$. Autrement dit, le vecteur $\underline{H}$ contient L valeurs distinctes présentes en proportion $1/L$ et l'écart entre ces valeurs est de 0.2. Pour la procédure bootstrap, $R = 500$ ré-échantillons des estimées multivariées corrigées sont tirées selon l'équation (4.20).

4.4.3.2 Nombre d'exposants distincts

Pour évaluer l'estimation du nombre d'exposants d'autosimilarité distincts, la figure 4.13 rapporte les histogrammes des estimées $\hat{L}$ des nombres de valeurs distinctes L dans $\underline{H}$, obtenues selon l'équation (4.22), pour différentes valeurs de L et différentes triples limites c liées à différents nombres de composantes M, tailles d'échantillon N et échelles d'analyse $2^{j_1}, \ldots, 2^{j_2}$. Quel que soit le nombre L de valeurs différentes dans $\underline{H}$, les performances d'estimation de L s'améliorent lorsque la dimension M et la taille d'échantillon N augmentent, et ce d'autant plus pour la plus petite valeur de triple limite $c = 1/8$.

Lorsqu'il n'y a que $L = 2$ valeurs différentes dans $\underline{H}$, il y a parfois une surestimation de L mais il y en a moins à mesure que le nombre de composantes M et la taille d'échantillon N croissent. En effet, le nombre de modes est bien détecté dans ce cas avec un taux $1 - \hat{\alpha}$ qui augmente avec M et N, où $\hat{\alpha}$ est le taux de fausses alarmes réel observé sur la figure 4.12. Ainsi, le comportement conservatif du test de Silverman (pour $k = 2$) réduit la surestimation de L.



(a) $c = 1/8$ (b) $c = 1/4$

FIGURE 4.13 – **Estimation du nombre de valeurs distinctes dans $\underline{H}$.** Histogrammes des estimés $\widehat{L}$ du nombresde valeurs distinctes L dans $\underline{H}$ au travers des réalisations de Monte Carlo pour différentes valeurs de L (de haut en bas), un niveau de confiance prédéfini $\alpha = 0.05$, et différents nombres de composantes M, tailles d'échantillon N et échelles d'analyse allant de $2^{j_1} = M/4$ à $2^{j_2} = M$ tels que $c = 1/8$ et $c = 1/4$. Le nombre exact de valeurs différentes L dans $\underline{H}$ est indiqué par une ligne droite pointillée.

Lorsqu'il y a (strictement) plus que 2 valeurs différentes dans $\underline{H}$, la valeur de L est bien estimée dans la majorité des cas seulement pour un nombre de composantes M et une taille d'échantillon N suffisamment grandes, en l'occurrence pour $M = 2^6$ et $N = 2^{15}$ lorsque $c = 1/8$ et pour $M = 2^6$ et $N = 2^{15}$ lorsque $c = 1/8$. La valeur de L a tendance à être sous-estimée pour de plus faibles nombres de composantes M et tailles d'échantillon N. Ce comportement suggère que la puissance des tests de Silverman augmente avec M et N, puisqu'une sous-estimation de L signifie que l'hypothèse nulle $\mathcal{H}_0^{(k)}$ est peu rejetée pour $2 \leq k < L$. De plus, une surestimation de L ne s'observe que pour $c = 1/8$, $N = 2^{15}$ et $M = 2^6$, et cette surestimation est de 1%, suggérant que le test de Silverman pour $k = L$ est également conservatif pour $L > 2$: l'hypothèse nulle $\mathcal{H}_0^{(L)}$ est rejetée avec un taux inférieur au niveau de confiance α.

4.4.3.3 Valeurs et proportions des exposants distincts

Les performances de l'estimation des valeurs $H_1, \ldots, H_L$ et de leurs proportions $\pi(H_1)$, …, $\pi(H_L)$ (cf. Section 4.4.2) sont mesurées par la distance de Wasserstein entre la distribution π et sa distribution estimée $\hat{\pi}$, toutes deux à support inclus dans $[0, 1]$, définie par

$$\mathcal{W}(\pi, \hat{\pi}) = \int_0^1 |F_\pi(h) - F_{\hat{\pi}}(h)| \, \mathrm{d}h, \tag{4.23}$$

où $\hat{F}_\pi$ et $\hat{F}_{\hat{\pi}}$ sont les fonctions de répartition respectivement associées aux distributions π et $\hat{\pi}$. La distance de Wasserstein prend, dans ce cas-ci, ses valeurs dans $[0, 1]$.

L'estimation des valeurs d'exposants d'autosimilarité distincts et de leurs proportions dans $\underline{H}$, i.e. l'estimation de la distribution π, est évaluée par les distances de Wassertein (4.23) rapportées dans le tableau 4.1 pour différents nombres de valeurs distinctes L dans $\underline{H}$ et différentes triples limites c liées à différents nombres de composantes M, tailles d'échantillon N et échelles d'analyse $2^{j_1}, \ldots, 2^{j_2}$. Pour les différentes valeurs de L, les performances d'estimation de π s'améliorent lorsque le nombre de composantes M et la taille d'échantillon N augmentent.

TABLEAU 4.1 – **Estimation des valeurs et proportions d'exposants d'autosimilarité distincts.** Distance de Wasserstein (moyenne de Monte Carlo avec intervalle de confiance à 95%) entre la distribution empirique des valeurs d'exposants d'autosimilarité et la distribution estimée pour différents nombres de valeurs L dans $\underline{H}$ et différents nombres de composantes M, tailles d'échantillon N et échelles d'analyse allant de $2^{j_1} = M/4$ à $2^{j_2} = M$ tels que $c = 1/8$ et $c = 1/4$.

	$c = 1/8$			$c = 1/4$		
	$N = 2^{11}$ $M = 2^4$	$N = 2^{13}$ $M = 2^5$	$N = 2^{15}$ $M = 2^6$	$N = 2^{10}$ $M = 2^4$	$N = 2^{12}$ $M = 2^5$	$N = 2^{14}$ $M = 2^6$
$L = 2$	0.07 $0.04 - 0.09$	0.06 $0.04 - 0.07$	0.03 $0.03 - 0.04$	0.08 $0.05 - 0.10$	0.07 $0.05 - 0.08$	0.05 $0.04 - 0.06$
$L = 3$	0.09 $0.05 - 0.11$	0.06 $0.02 - 0.09$	0.02 $0.01 - 0.07$	0.10 $0.07 - 0.13$	0.09 $0.05 - 0.10$	0.05 $0.02 - 0.09$
$L = 4$	0.11 $0.07 - 0.14$	0.08 $0.04 - 0.11$	0.03 $0.01 - 0.10$	0.1 $0.07 - 0.15$	0.09 $0.04 - 0.12$	0.05 $0.01 - 0.10$

La croissance des performances de l'estimation de π avec N et M est en particulier valable pour $L = 2$ et n'est donc pas seulement lié à l'amélioration de l'estimation de L observée sur les histogrammes 4.13. Cependant, la bonne estimation de $L = 2$ mène à de meilleures performances d'estimation de π dans ce cas-ci par rapport aux configurations où il y a plus que 2 valeurs distinctes dans $\underline{H}$, i.e. $L = 3$ et $L = 4$. En revanche, les performances sont similaires entre $L = 3$ et $L = 4$. Ces résultats suggèrent que cette stratégie est robuste au nombre d'exposants distincts L dans $\underline{H}$ lorsque $L > 2$.

Enfin, on observe que, à même nombre de composantes M, les performances d'estimation sont légèrement meilleures pour la petite valeur de triple limite $c = 1/8$ comparée à $c = 1/4$, c'est-à-dire pour de plus grands nombres $n_{j_1}, \ldots, n_{j_2}$ de coefficients d'ondelettes impliqués dans l'analyse.

4.5 Conclusion

Le présent chapitre a abordé l'étude de l'autosimilarité multivariée en grande dimension, régime asymptotique où le nombre de composantes M, la taille d'échantillon N et les échelles d'analyse tendent vers l'infini. Ce régime asymptotique est contrôlé par un paramètre asymptotiquement constant $c = M/n_{j_2}$, rapport du nombre de composantes M sur le nombre de coefficients d'ondelettes disponible n_{j_2} à la plus grande octave d'analyse j_2. A ainsi été proposée dans ce chapitre une stratégie complète pour compter le nombre de valeurs distinctes dans le vecteur des exposants d'autosimilarité $\underline{H}$, les estimer et estimer la proportion de chacune de ces valeurs.

En premier lieu, le comportement asymptotique de l'estimateur multivarié corrigé $\hat{\underline{H}}^{(\mathrm{M,bc})}$ construit dans le chapitre 2 a été étudié numériquement, à partir de simulations de Monte Carlo de M-mBf synthétiques de taille finie. Les résultats montrent d'abord que le nombre de modes de la distribution des entrées de $\hat{\underline{H}}^{(\mathrm{M,bc})}$ tend asymptotiquement vers le nombre d'exposants distincts dans $\underline{H}$ en grande dimension. S'appuyant sur ce résultat, une procédure pour tester la présence d'une valeur unique dans $\underline{H}$ a d'abord été développée à partir du vecteur $\hat{\underline{H}}^{(\mathrm{M,bc})}$ et de la procédure de ré-échantillonnage bootstrap du chapitre 3. L'étude numérique menée montre que la procédure bootstrap, qui repose sur la statistique de Hartigan, reproduit bien l'hypothèse

nulle et est puissante asymptotiquement, c'est-à-dire pour de grands nombres de composantes M et des tailles d'échantillon réalistes N pouvant être du même ordre que M. Dans le cas où cette dernière procédure rejette la présence d'une unique valeur dans $\underline{H}$, une procédure de multimodalité, reposant sur le test de Silverman, construite également à partir de l'estimateur $\underline{\hat{H}}^{(\mathrm{M,bc})}$ et de la procédure de ré-échantillonnage bootstrap du chapitre 3, a été développée et évaluée numériquement. Cette procédure montre des performances asymptotiques satisfaisantes pour différents nombres $L \geq 2$ d'exposants distincts dans $\underline{H}$, à la fois en termes d'estimation de L et d'estimation des valeurs des exposants et de leurs proportions dans $\underline{H}$.

Finalement, alors que les outils de faible dimension sont adaptés à des cas où le rapport $c = M/n_{j_2}$ est faible, et ce autant pour des grands nombres de composantes M que des petits, les outils de grande dimension sont adaptés à des grands nombres de composantes M et sont d'autant plus efficaces que le rapport c est petit, mais sont aussi valables pour de grands rapports $c = M/n_{j_2}$.

Chapitre 5

Applications biomédicales

Sommaire

5.1 Enjeux

Les rythmes temporels physiologiques et corporels sont bien décrits par des dynamiques arythmiques ou invariantes d'échelle. C'est notamment le cas de l'activité cérébrale macroscopique lente (de fréquence inférieure à 0.1Hz) (Ciuciu et collab., 2014; He, 2011; La Rocca et collab., 2018), de la variabilité du rythme cardiaque (Amaral et collab., 1999; Doret et collab., 2015; Ivanov et collab., 1999; Nakamura et collab., 2016; Yamamoto et Hughson, 1991), de la variabilité de la démarche humaine (BenAbdelkader et collab., 2004; Decker et collab., 2010) ou de la dynamique de la phase de sommeil (Leon et collab., 2022). Cependant, l'analyse de l'invariance d'échelle de ces signaux est généralement réalisée dans un contexte univarié, c'est-à-dire en estimant les exposants d'autosimilarité indépendamment pour chaque série temporelle observée. Pourtant, les données physiologiques sont très souvent constituées de plusieurs séries temporelles enregistrées conjointement pour étudier un même mécanisme biologique ou une pathologie.

Dans ce contexte, les modèles d'autosimilarité multivariée tels que le mouvement brownien opérateur-fractionnaire (mBof) pourraient s'avérer plus efficaces pour rendre compte de la dynamique temporelle conjointe des séries mesurées. Cela n'a jusqu'à présent jamais été tenté. Le présent chapitre vise ainsi à quantifier la pertinence et les avantages de l'autosimilarité multivariée dans les analyses du rythme physiologique et corporel à travers les exemples *(i)* de la détection de la somnolence à partir de données polysomnographiques et *(ii)* de la prédiction des crises d'épilepsie à partir de données d'électroencéphalogrammes (EEG) du cuir chevelu. Des stratégies de classification sont ainsi développées à partir des estimateurs des exposants d'autosimilarité, rappelées dans la section suivante.

5.2 Méthodologie

Pour les différents jeux de données, des séries temporelles M-variées $\{Y(t)\}_{t\in\{1,\dots,N\}}$ sont extraites et analysées par ondelettes selon les différentes procédures décrites dans les chapitres 1 et 2 pour obtenir des estimées des exposants d'autosimilarité $\underline{H}$:

- M *estimées univariées* (cf. Eq. (1.42)) :

$$\forall m \in \{1,\dots,M\}, \quad \hat{H}_m^{(\mathrm{U})} = \frac{1}{2}\left(\sum_{j=j_1}^{j_2} w_j \log_2 S_{m,m}(2^j) - 1\right); \tag{5.1}$$

- $M(M+1)/2$ *estimées multivariées classiques*, pour tenir compte des dépendances temporelles (cf. Eq. (1.48)) :

$$\forall 1 \le m \le m' \le M, \quad \hat{H}_{m,m'}^{(\mathrm{U})} = \frac{1}{2}\left(\sum_{j=j_1}^{j_2} w_j \log_2 |S_{m,m'}(2^j)| - 1\right); \tag{5.2}$$

- M *estimées multivariées* (cf. Eq. (2.3)) :

$$\forall m \in \{1,\dots,M\}, \quad \hat{H}_m^{(\mathrm{M,bc})} = \frac{1}{2}\left(\sum_{j=j_1}^{j_2} w_j \log_2 \bar{\lambda}_m(2^j) - 1\right); \tag{5.3}$$

où $S(2^j)$ désigne le spectre d'ondelettes multivarié empirique. Les estimées corrigées $\log_2 \bar{\lambda}_m(2^j)$ des logarithmes des valeurs propres $\lambda_m(2^j)$ du spectre d'ondelettes multivarié $\mathbb{E}[S(2^j)]$ sont données par l'équation (2.2).

5.3 Détection de la somnolence

Cette section est tirée de l'article suivant : C.-G. Lucas, P. Abry, H. Wendt et G. Didier, « Drowsiness detection from polysomnographic data using multivariate selfsimilarity and eigen-wavelet analysis », dans *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, pp. 2949–2952.

La somnolence est généralement définie comme un état intermédiaire entre l'éveil et le sommeil (Yu et collab., 2018). Il est établi que la somnolence joue un rôle majeur dans les accidents de la route. Par conséquent, la détection de la somnolence à partir de mesures biomédicales mobiles non invasives constitue un enjeu sociétal important. Elle a souvent été réalisée à partir de données d'électromyogramme (EMG), d'électrocardiogramme (ECG) ou d'EEG, en utilisant des outils statistiques non linéaires de traitement du signal tels que l'entropie d'échantillon, l'autosimilarité ou la multifractalité (Ahn et collab., 2016; Cahyadi et collab., 2019; Wang et collab., 2018; Yu et collab., 2018), mais en laissant essentiellement inexplorée l'utilisation de la dynamique invariante d'échelle multivariée (voir a contrario Leon et collab. (2022)).

5.3.1 Jeu de données

5.3.1.1 Description données

Les données utilisées ici sont celles de la base de données polysomnographiques du MIT-BIH disponible à https://physionet.org/content/slpdb/1.0.0/ et documentée dans Goldberger et collab. (2000); Ichimaru et Moody (1999). Les données consistent en une collection de 4 à 7 mesures physiologiques (activités cardiovasculaires, respiratoires ou cérébrales macroscopiques, volume d'éjection, saturation en oxygène, mouvements oculaires ou réponses des muscles du menton) enregistrées pour 16 sujets masculins au Boston's Beth Israel Hospital Sleep Laboratory dans le contexte du syndrome d'apnée obstructive chronique du sommeil ; les enregistrements des deux premiers sujets sont divisés en deux parties consécutives, ce qui donne un total de 18 séries temporelles multivariées (échantillonnées à 250Hz et durant 77 à 390 minutes). Des annotations d'experts sur les stades du sommeil sont disponibles sur toutes les fenêtres, sans chevauchement, de 30 secondes.

5.3.1.2 Pré-traitement des données

Étant donné que ce travail se concentre sur la détection de la somnolence, définie comme des transitions entre le stade *éveillé* et le *stade 1 du sommeil*, l'objectif est d'effectuer une classification de ces deux stades. On utilise donc uniquement les fenêtres temporelles correspondant à ces annotations. De plus, pour étudier les avantages d'une analyse conjointe des signaux par l'autosimilarité multivariée, il est nécessaire d'analyser les différentes modalités des données polysomnographiques dans la même gamme de fréquences. Comme les rythmes respiratoires et cardiaques contiennent des informations pertinentes dans des échelles de temps allant de 1 seconde à 1 minute, seule l'activité cérébrale lente (c'est-à-dire de fréquence inférieure à 0.1Hz) est prise en compte, et les données sont filtrées et ré-échantillonnées à 4Hz. Enfin, pour garantir un nombre cohérent d'attributs pour la détection et la classification, seules les 4 modalités des données polysomnographiques disponibles pour tous les sujets sont utilisées : fréquence cardiaque (FC), pression sanguine (PS), électroencéphalogramme (EEG) et respiration (RESP).

5.3.2 Configuration de l'analyse et de la classification

5.3.2.1 Analyse de séries temporelles physiologiques

L'analyse est effectuée sur des fenêtres glissantes de 2 min, avec 75% de chevauchement entre les fenêtres successives, chacune contenant donc $N = 480 = 4 \times 2 \times 60$ échantillons. La classification sera limitée aux séquences partageant les mêmes annotations pour au moins 4 fenêtres consécutives de 30 secondes. Au total, 1753 et 561 fenêtres de ce type sont disponibles pour l'état éveillé et le stade 1, respectivement.

Les coefficients d'ondelettes multivariés discrets (cf. Section 1.4.1.1) des $M = 4$ séries temporelles de taille $N = 480$, associées aux différentes modalités FC, PS, EEG et RESP, sont calculés en utilisant l'ondelette mère ψ_0 la moins asymétrique de Daubechies à $N_\psi = 3$ moments nuls (Daubechies, 1992). Les régressions linéaires pour l'estimation des exposants d'autosimilarité sont effectuées sur des échelles allant de $2^{j_1} = 2^1$ à $2^{j_2} = 2^4$ correspondant à des fréquences de 1/8 à 2Hz, ou de manière équivalente à des échelles de temps de 1/2 à 8 secondes.

À titre d'exemple, sur les figures 5.1 et 5.2, sont représentées les fonctions de structure univariées $\log_2 |S_{m,m'}(2^j)|$, $1 \leq m \leq m' \leq 4$ et les fonctions de structure multivariées $\log_2 \bar{\lambda}_1(2^j), \ldots,$ $\log_2 \bar{\lambda}_4(2^j)$ pour une fenêtre de 2 minutes, indiquant des lois de puissance et donc une autosimilarité à travers les échelles d'analyse choisies.

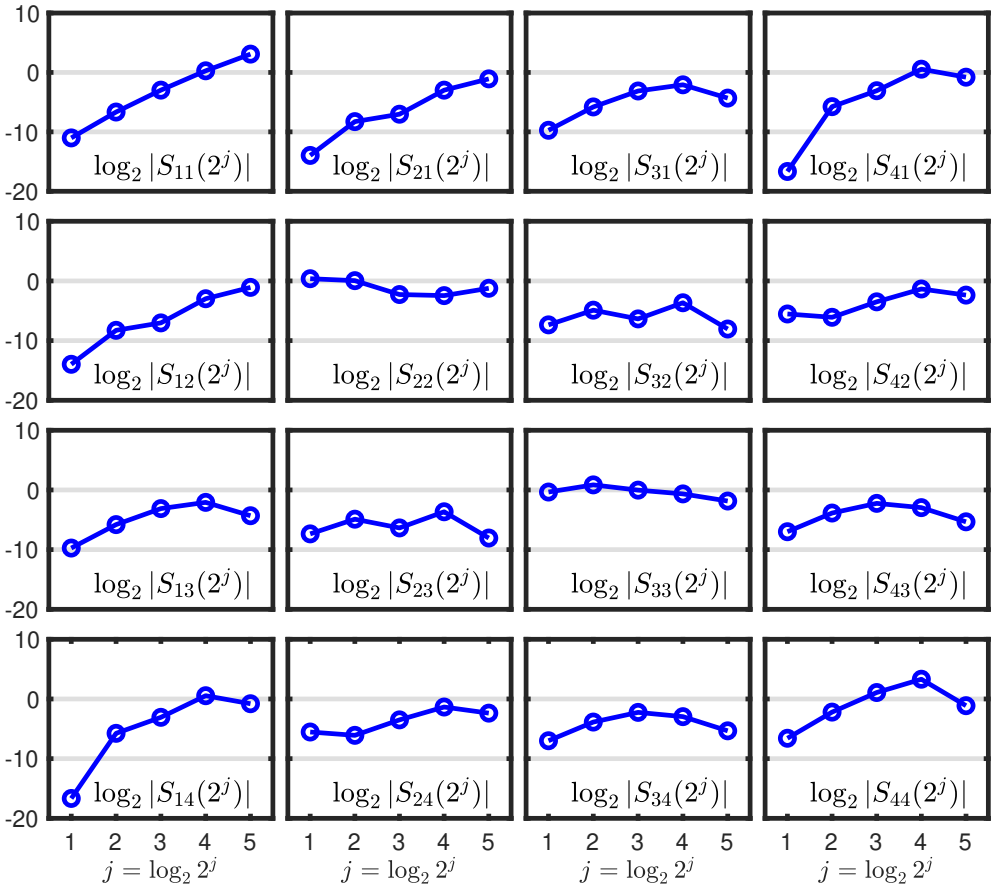


FIGURE 5.1 – **Analyse multivariée classique.** Fonctions de structure multivariée classique $\log_2 |S_{m,m'}(2^j)|$ pour $m, m' \in \{1, \ldots, 4\}$ issues de l'analyse d'une fenêtre de 2 minutes d'un sujet. *Les coefficients du spectre d'ondelettes multivarié empirique $S(2^j)$ de la fenêtre analysée ont des comportements en loi de puissance aux petites échelles, et les coefficients non diagonaux sont non négligeables.*

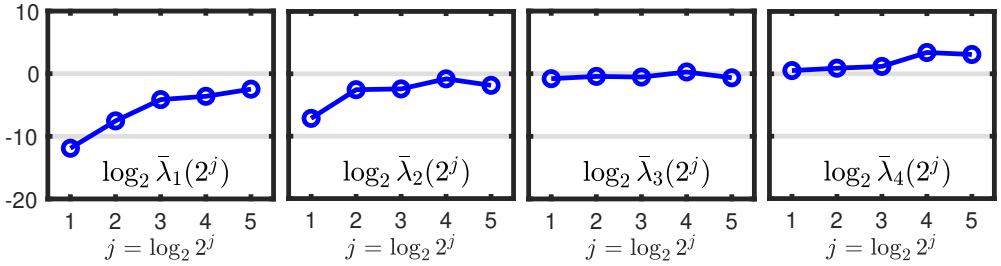


FIGURE 5.2 – **Analyse multivariée.** Fonctions de structure multivariée $\log_2 \bar{\lambda}_m(2^j)$ pour $m \in \{1, \ldots, 4\}$ issues de l'analyse d'une fenêtre de 2 minutes d'un sujet. *Les fonctions de structure multivariée $\bar{\lambda}_m$ de la fenêtre analysée ont un comportement en loi de puissance aux petites échelles.*

5.3.2.2 Méthode et attributs de classification

Nous utilisons comme classificateur une forêt d'arbres décisionnels (BREIMAN, 2001), méthode d'apprentissage consistant à former un grand nombre d'arbres de décision à partir de données ré-échantillonnées avec remise. Chaque arbre est formé à partir d'un sous-ensemble d'attributs choisis aléatoirement parmi les N_f attributs disponibles, afin de réduire la corrélation entre les arbres. La taille des sous-ensembles choisie ici est $\sqrt{N_f}$, comme dans BREIMAN (2001). La décision est prise par vote majoritaire sur l'ensemble des décisions des arbres.

Les forêts d'arbres décisionnels sont réalisés avec $N_{\text{arbres}} \in \{10, 25, 50\}$ arbres. Une matrice de coût diagonale est définie avec les coefficients $w_1 = W_v N_a / N_W$ et $w_2 = N_s / N_W$, où N_a est le nombre de fenêtres liées à l'état « éveillé » (classe 0), N_s est le nombre de fenêtres liées à l'état « stade 1 » (classe 1), $N_W = N_a + N_s$ est le nombre total de fenêtres et $W_v \in [0.001, 6]$ règle le taux de fausses alarmes de la classification. Les performances sont évaluées par validation croisée sur $N_{MC} = 100$ répétitions de la méthode de classification, avec 80% des fenêtres disponibles sélectionnées aléatoirement et indépendamment pour les ensembles d'entraînement.

Les courbes ROC (Receiver Operational Characteristic) sont calculées en faisant varier W_v pendant l'entraînement. Les valeurs d'AUC (Area Under the Curve), correspondant aux aires sous les courbes ROC, sont utilisées comme score de performance.

5.3.3 Classification à un seul attribut

En guise de référence, la classification est d'abord effectuée pour chacune des quatre modalités $m \in \{1, 2, 3, 4\}$ indépendamment, en utilisant comme attribut unique le paramètre d'auto-similarité univarié $\hat{H}_m^{(\mathrm{U})}$. La classification consiste donc simplement à comparer l'attribut $\hat{H}_m^{(\mathrm{U})}$ par rapport à un seuil, et les courbes ROC sont calculées en faisant varier le seuil de classification et représentées sur la figure 5.3.

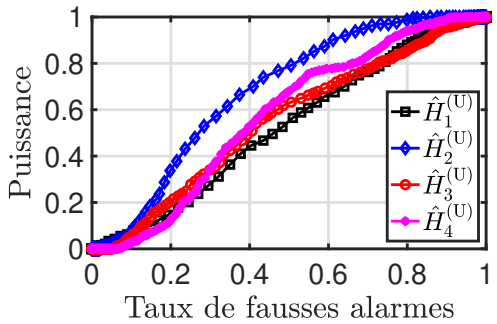


FIGURE 5.3 – **Courbes ROC pour la classification univariée.** *Les meilleures performances sont atteinte par la pression sanguine mais restent modérées.*

Les AUC, rapportés dans le tableau 5.1, montrent que la pression sanguine donne les meilleures performances de classification. Cependant, la performance de la classification à attribut unique reste assez faible.

TABLEAU 5.1 – **AUC pour la classification univariée.**

Attribut	$\hat{H}_1^{(\mathrm{U})}$	$\hat{H}_2^{(\mathrm{U})}$	$\hat{H}_3^{(\mathrm{U})}$	$\hat{H}_4^{(\mathrm{U})}$
AUC	47.29	67.59	55.15	57.44

5.3.4 Classification à plusieurs attributs

L'objectif, à présent, est de quantifier les avantages de l'utilisation conjointe de plusieurs modalités polysomnographiques dans la classification des stades de sommeil liés à la somnolence. Trois stratégies différentes pour combiner les informations provenant des différentes attributs sont testées.

5.3.4.1 Attributs

La méthode la plus simple consiste à concaténer les quatre estimées univariées $\hat{H}_m^{(\mathrm{U})}$ en un vecteur d'attributs de dimension 4,

$$C1 = (\hat{H}_m^{(\mathrm{U})})_{1\leq m\leq 4}. \tag{5.4}$$

Une telle classification conjointe des modalités ne tient pas compte des informations liées à la dépendance temporelle entre les modalités, car elle n'utilise que des attributs calculés indépendamment sur chaque modalité. Cependant, comme l'illustre la figure 5.1, il existe des dépendances temporelles non négligeables à toutes les échelles entre les modalités, ce qui incite à utiliser ces informations pour améliorer la classification.

Pour tenir compte de la dépendance temporelle entre les modalités, une deuxième classification sera effectuée en ajoutant aux 4 attributs univariées $\hat{H}_m^{(\mathrm{U})}$, les $6 = 4\times 3/2 = M(M-1)/2$ estimées multivariées classiques $\hat{H}_{m,m'}^{(\mathrm{U})}$, $m < m'$, obtenues à partir des entrées non diagonales du spectre d'ondelettes multivarié empirique $S(2^j)$, ce qui donne un vecteur d'attributs de dimension 10,

$$C2 = \left((\hat{H}_m^{(\mathrm{U})})_{1\leq m\leq 4}, (\hat{H}_{m,m'}^{(\mathrm{U})})_{1\leq m<m'\leq 4}\right). \tag{5.5}$$

En outre, une contribution originale de ce travail est de promouvoir l'utilisation de l'analyse d'autosimilarité multivariée comme une nouvelle façon de quantifier la dynamique temporelle conjointe. Par conséquent, une troisième classification sera effectuée en ajoutant aux 4 attributs univariées $\hat{H}_m^{(\mathrm{U})}$, les 4 estimées multivariées $\hat{H}_m^{(\mathrm{M,bc})}$, constituant ainsi un vecteur d'attributs de dimension 8,

$$C3 = \left((\hat{H}_m^{(U)})_{1\leq m\leq 4}, (\hat{H}_m^{(\mathrm{M,bc})})_{1\leq m\leq 4}\right). \tag{5.6}$$

5.3.4.2 Performances

Afin de comparer les performances de ces trois stratégies de classification multimodale, un test de Wilcoxon (WILCOXON, 1945) est d'abord appliqué à chaque entrée des vecteurs d'attributs $C1$, $C2$ et $C3$ indépendamment afin de tester l'hypothèse nulle d'une médiane égale entre les deux classes, avec un niveau de confiance fixé à $\alpha = 0.05$. La procédure de correction de

Benjamini-Hochberg contrôlant le taux de fausses découvertes pour les tests à hypothèses multiples (Benjamini et Hochberg, 1995) est appliquée pour obtenir une décision pour chaque entrée du vecteur d'attributs. Le tableau 5.2 indique, pour chacune des trois stratégies, le pourcentage de décisions de rejet de l'hypothèse nulle. Cela montre que, pour les attributs univariés réunis ($C1$), 3 modalités sur 4 donnent des décisions de rejet. Lorsqu'on essaie d'ajouter les dépendances temporelles, les attributs multivariés classiques aboutissent à un pourcentage plus faible de décisions de rejet, tandis que l'approche multivariée proposée pour quantifier la dépendance temporelle augmente le pourcentage de décisions de rejet.

Tableau 5.2 – **Test de Wilcoxon.** Pourcentage de décisions de rejet du test de la somme des rangs de Wilcoxon corrigées par la procédure de Benjamini-Hochberg entre les états « éveillé » et « stade 1 » pour les différents vecteurs d'attributs.

Attributs	$C1$	$C2$	$C3$
Taux de rejet	0.750	0.700	0.875

Pour renforcer cette analyse préliminaire, la figure 5.4 affiche les courbes ROC pour les différents vecteurs d'attributs et le tableau 5.3 rapporte les AUC correspondants. Ces résultats permettent de tirer les conclusions suivantes. Les classifications à plusieurs attributs sont nettement plus performantes que les classifications à attribut unique. Pour les classifications à plusieurs attributs, les performances sont robustes au choix du nombre d'arbres dans la procédure de forêt d'arbres décisionnels. Parmi les classifications à plusieurs attributs, celle utilisant les mesures classiques de la dépendance temporelle ($C2$) n'améliore pas les performances de classification par rapport à la concaténation la plus simple des attributs univariés ($C1$). Au contraire, l'approche multivariée à partir des fonctions de structure $\log_2 \bar{\lambda}(2^j)$ pour sonder la dépendance temporelle, combinée aux attributs univariés ($C3$), améliore considérablement les performances de classification des stades de sommeil liés à la somnolence.

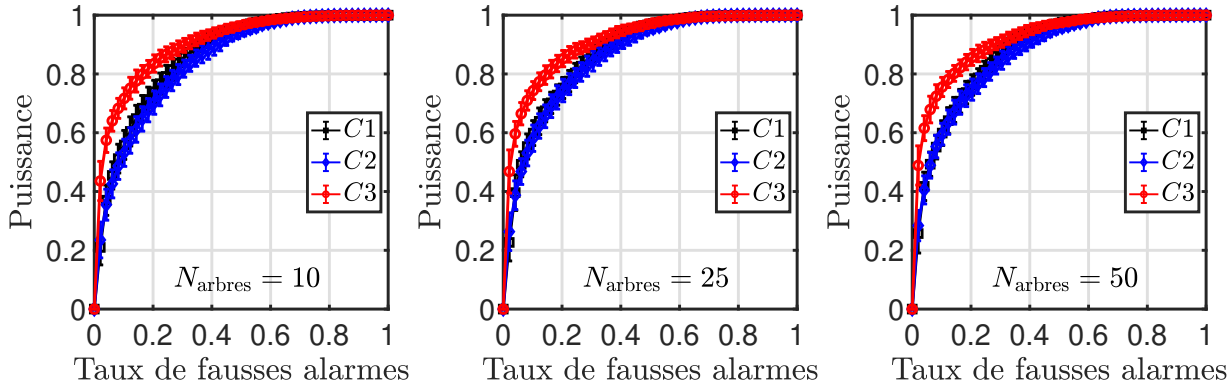


Figure 5.4 – **Courbes ROC pour la classification à plusieurs attributs.** Courbes ROC (moyenne ± écart-type) pour les différentes stratégies de classification avec trois différents nombres d'arbres N_{arbres} (de gauche à droite).

Tableau 5.3 – **AUC (moyenne ± écart-type) pour la classification multivariée.**

AUC	$C1$	$C2$	$C3$
$N_{\text{arbres}} = 10$	85.34 ± 0.70	83.71 ± 0.71	89.31 ± 0.54
$N_{\text{arbres}} = 25$	86.25 ± 0.64	85.29 ± 0.85	90.11 ± 0.54
$N_{\text{arbres}} = 50$	86.68 ± 0.66	85.68 ± 0.68	90.46 ± 0.54

5.3.5 Conclusions

La présente section a d'abord quantifié les avantages de la surveillance multimodale du sommeil dans la classification des phases du sommeil. Ensuite, il a été montré que, par rapport à la simple combinaison par concaténation d'attributs univariés, l'approche multivariée renforce la classification des phases du sommeil. Ce travail quantifie clairement que la dépendance temporelle entre les modalités est porteuse d'informations pertinentes pour l'évaluation de l'état de sommeil. Cela s'explique par le fait que les plus petites valeurs propres du spectre d'ondelettes multivarié sont capables de détecter des dépendances de faible intensité, mais significatives, entre les composantes, que les corrélations croisées classiques par paire peuvent manquer.

5.4 Prédiction de crises d'épilepsie

Cette section est tirée de l'article suivant : C.-G. Lucas, P. Abry, H. Wendt et G. Didier, « Epileptic seizure prediction from eigen-wavelet multivariate selfsimilarity analysis of multi-channel EEG signals », *2023 European Signal Processing Conference (EUSIPCO)*, IEEE.

L'épilepsie, une maladie chronique, consiste en un trouble du système nerveux central, entraînant des crises au cours desquelles le cerveau du patient peut être gravement endommagé. La mise au point de procédures automatisées de prédiction des crises d'épilepsie constitue donc un enjeu crucial et permanent, notamment lorsqu'elles peuvent être mises en œuvre à partir de dispositifs non invasifs et portables d'EEG du cuir chevelu. La prédiction des crises d'épilepsie constitue un sujet de recherche important, souvent étudié à l'aide d'outils tels que la synchronisation et la connectivité fonctionnelle (MohanBabu et collab., 2021), la cohérence de phase (Mormann et collab., 2000), la densité spectrale de puissance (Park et collab., 2011; Zhang, 2015), la puissance des coefficients d'ondelettes dans des bandes de fréquence standard (Saab et Gotman, 2005), les modèles autorégressifs (Chisci et collab., 2010), ou plus récemment les structures d'apprentissage profond (Daoud et Bayoumi, 2019). La dynamique invariante d'échelle a également été impliquée dans la prédiction des crises d'épilepsie, par exemple à partir d'EEG intracrâniens Gadhoumi et collab. (2015) ou d'un unique EEG du cuir chevelu (Domingues et collab., 2019).

5.4.1 Jeu de données

5.4.1.1 Description des données

Les données utilisées dans ce travail consistent en des enregistrements d'EEG multicanaux du cuir chevelu provenant de la base de données CHB-MIT Scalp EEG disponible à https://physionet.org/content/chbmit/1.0.0/ et documenté dans Goldberger et collab. (2000); Shoeb (2009). Ces enregistrements ont été recueillis au Boston Children's Hospital auprès de sujets pédiatriques souffrant de crises épileptiques réfractaires et échantillonnés à 256Hz.

Les enregistrements d'EEG ont été divisés en 23 cas collectés auprès de 22 sujets, composés de 5 hommes et 17 femmes, et annotés avec les débuts et fins de crises épileptiques. Pour chaque cas, entre 22 et 26 signaux d'EEG ont été enregistrés pendant plusieurs heures selon le système international 10-20 de position et de nomenclature des électrodes d'EEG. Les enregistrements durent au moins une heure et seule une partie d'entre eux contient des périodes de crise d'épilepsie.

5.4.1.2 Pré-traitement des données

Dans le présent travail, nous utilisons les 22 premiers canaux d'EEG, de manière à utiliser les mêmes canaux pour tous les sujets. Le travail étant axé sur la prédiction de crises d'épilepsie, l'objectif est d'effectuer une détection des états préictaux, qui sont des périodes se produisant quelques minutes avant le début d'une crise d'épilepsie. Ainsi, les fenêtres correspondant aux états préictaux sont sélectionnées dans les enregistrements contenant des crises tandis que les fenêtres correspondant aux états interictaux (éloignés dans le temps de toute crise épileptique) sont sélectionnées dans les enregistrements sans crise. En pratique, des fenêtres de 2 minutes, soit $N = 30720$ échantillons, sont utilisées.

Pour évaluer quantitativement la performance de l'analyse de l'autosimilarité multivariée par valeurs propres d'ondelettes proposée pour détecter les états préictaux, seuls les sujets dont les données sont constituées d'au moins 110 fenêtres interictales et 10 de fenêtres préictales sont considérés. Ainsi, seuls 8 sujets sont étudiés dans ce travail.

5.4.2 Détection d'états préictaux

5.4.2.1 Configuration de l'analyse

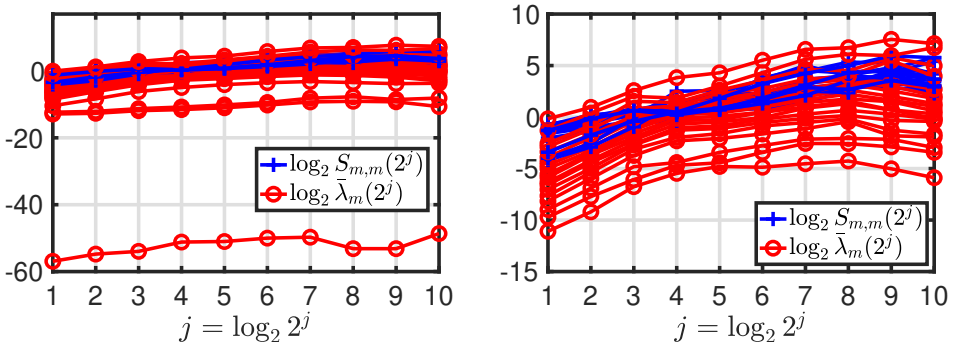
Des transformées en ondelettes (cf. Section 1.4.1.1) de fenêtres de 2 minutes des signaux $M = 22$-variés d'EEG sont réalisées à l'aide de l'ondelette mère ψ_0 de Daubechies à $N_\psi = 2$ moments nuls (Daubechies, 1992). Pour l'estimation des exposants d'autosimilarité, les régressions linéaires sont effectuées sur les échelles allant de $2^{j_1} = 2^1$ à $2^{j_2} = 2^4$, correspondant à des fréquences équivalentes allant de 10Hz à 85Hz, pour lesquelles les signaux d'EEG intracrâniens sont documentés comme ayant une dynamique invariante d'échelle (Gadhoumi et collab., 2015).

5.4.2.2 Analyse multivariée d'une fenêtre

La figure 5.5 (a) compare, pour une fenêtre préictale d'un sujet donné, les $M = 22$ fonctions de structure univariée $\log_2 S_{m,m}(2^j)$ (lignes noires avec le symbole '+') contre les $M = 22$ fonctions de structure multivariée $\log_2 \bar{\lambda}_m(2^j)$ (lignes rouges avec le symbole o'). Alors que les $M = 22$ fonctions d'analyse univariée $\log_2 S_{m,m}(2^j)$ sont très proches, l'analyse multivariée montre clairement que 3 des fonctions de structure multivariée $\log_2 \bar{\lambda}_m(2^j)$ prennent des valeurs bien plus petites que les 19 autres et proches de 0 comparées à celles-ci. Ceci est davantage mis en lumière pour l'échelle 2^4 par la figure 5.6 (a).

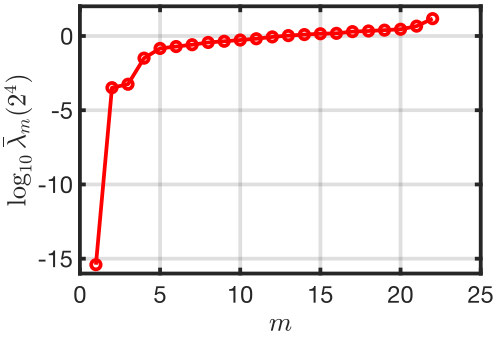
Cela trahit des dépendances linéaires entre les 22 enregistrements d'EEG. En effet, les séries temporelles résultent de la soustraction des mesures des électrodes, certaines étant utilisées plusieurs fois, de sorte qu'une série temporelle est en fait l'addition de plusieurs autres. Une inspection minutieuse des données, illustrée par la figure 5.6 (b), révèle que les paires d'électrodes P3-O1 et Fp2-F4 sont redondantes car les électrodes P3 et O1 sont déjà reliées par le chemin P3-C3-F3-Fp1-F7-T7-P7-O1, et les électrodes Fp2 et F4 sont déjà reliées par le chemin Fp2-F8-T8-P8-O2-P4-C4-F4. Quant aux signaux issus des paires T7-P7 et P7-T7, ils sont parfaitement anti-corrélés et ainsi responsables du très faible ordre de grandeur de la valeur propre $\lambda_1(2^j)$ dans l'analyse des 22 composantes. Cela conduit à supprimer trois enregistrements redondants (T7-P7, P3-O1, FP2-F4) avant d'effectuer l'analyse multivariée, conduisant ainsi à une analyse de l'autosimilarité multivariée de $M = 19$ composantes. Les paires d'électrodes retenues pour l'analyse sont présentées dans la figure 5.6 (b) et l'analyse de l'autosimilarité multivariée des $M = 19$ composantes est présentée dans la figure 5.5 (b).

La figure 5.5 confirme en outre les comportements linéaires à la fois des $\log_2 S_{m,m}(2^j)$ et des

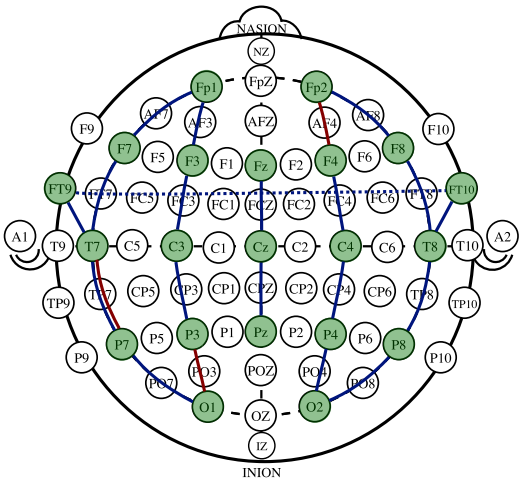


(a) $M = 22$ signaux d'EEG (b) $M = 19$ signaux d'EEG

FIGURE 5.5 – **Analyse d'invariance d'échelle multivariée.** Fonctions de structure univariée $\log_2 S_{m,m}(2^j)$ (en bleu avec le symbole '+') et fonctions de structure multivariée $\log_2 \bar{\lambda}_m(2^j)$ (en rouge avec le symbole 'o') pour une fenêtre préictale associée au sujet 5 avant (à gauche) et après (à droite) suppression des 3 canaux redondants.



(a) **Dépendances linéaires.** Fonction de structure multivariée $\log_{10} \bar{\lambda}_m(2^4)$ en fonction des entrées $m = 1, \ldots, M$, pour une fenêtre préictale associée au sujet 5 pour les $M = 22$ signaux EEG.



(b) **Électrodes de mesure.** 22 électrodes utilisées pour les mesures (en vert), 3 paires d'électrodes supprimées avant analyse (en rouge) et 19 paires d'électrodes retenues pour l'analyse (en bleu).

FIGURE 5.6 – **Suppression de canaux d'EEG redondants.** Parmi les 22 paires d'électrodes de mesure, 3 sont supprimées pour éviter une dépendance linéaire entre les signaux analysés.

$\log_2 \bar{\lambda}_m(2^j)$ à des échelles fines pour $m \in \{1, \ldots, M\}$, et donc la dynamique invariante d'échelle de l'ensemble des signaux analysés.

5.4.2.3 Distributions des estimées des exposants d'autosimilarité

Pour illustrer la capacité de l'analyse d'autosimilarité multivariée à détecter la dynamique temporelle des états préictaux par rapport à celle des états interictaux, la figure 5.7 compare (au moyen de boîtes à moustaches) les distributions des estimées univariées $\hat{H}_m^{(\mathrm{U})}$ (première et troisième colonnes) et multivariées $\hat{H}_m^{(\mathrm{M,bc})}$ (seconde et quatrième colonnes), calculées pour les 8 différents sujets, au travers des fenêtres interictales et préictales disponibles (dont le nombre est donné en légende sur les seconde et quatrième colonnes). La figure 5.7 montre des distinctions plus nettes entre les distributions des estimées $\hat{H}_m^{(\mathrm{M,bc})}$ associées aux états préictaux et les

distributions des estimées $\hat{H}_m^{(\mathrm{M,bc})}$ associées aux états interictaux estimées, pour les différentes entrées m, par rapport aux distributions des estimées $\hat{H}_m^{(\mathrm{U})}$, avec un étalement plus faible des distributions et un chevauchement plus petit entre les distributions.

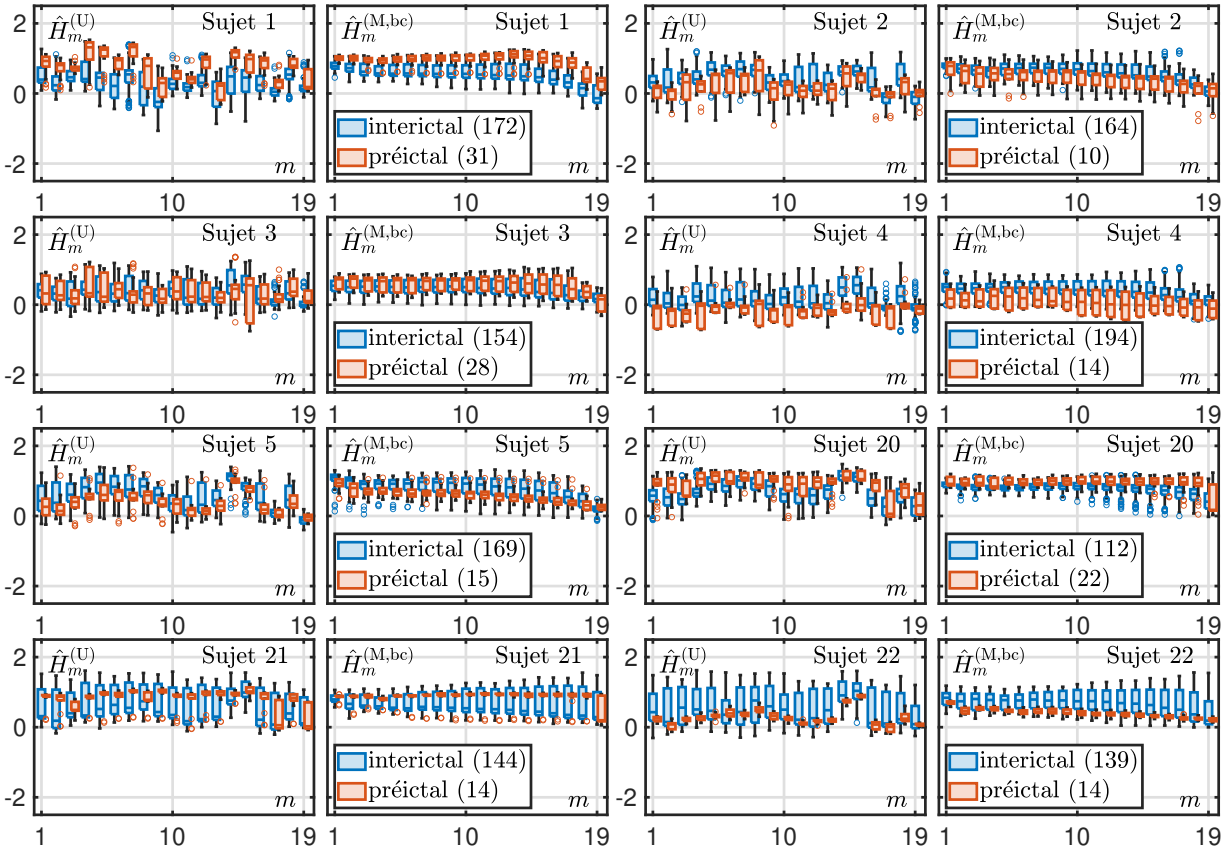


FIGURE 5.7 – **Distributions des estimées exposants d'autosimilarité.** Boîtes à moustaches des estimées univariées $\hat{H}_m^{(\mathrm{U})}$ (première et troisième colonnes) et multivariées $\hat{H}_m^{(\mathrm{M,bc})}$ (deuxième et quatrième colonnes) associées aux états préictaux (en rouge) et interictaux (en bleu) pour les 8 différents sujets.

La figure 5.7 montre également, par des comparaisons entre sujets, que la détection entre les états interictaux et préictaux doit être effectuée pour chaque sujet indépendamment. En effet, les distributions des estimées $\hat{H}_m^{(\mathrm{M,bc})}$ pour les états interictaux diffèrent d'un sujet à l'autre, donc la moyenne entre les sujets estomperait les différences entre les statistiques préictales et interictales. La figure 5.7 indique clairement que

(i) les états préictaux ont une dynamique temporelle qui diffère de celle des états interictaux d'un même sujet,

(ii) et que les états interictaux de différents sujets ont des dynamiques temporelles différentes.

Ce sont les premiers résultats importants de ce travail.

Afin de quantifier les différences entre les distributions des exposants d'autosimilarité estimés entre les états préictaux et interictaux, les p-valeurs $p_m^{(W)}$ du test de la somme des rangs de Wilcoxon, qui teste l'hypothèse nulle d'une médiane égale entre des estimées d'exposants d'autosimilarité ($\hat{H}_m^{(\mathrm{U})}$ ou $\hat{H}_m^{(\mathrm{M,bc})}$) associés aux deux états pour chaque entrée $m \in \{1, \ldots, 19\}$ sont calculées. Ces p-valeurs $p_m^{(W)}$ sont comparées aux seuils de Benjamini-Hochberg (pour la correction d'hypothèses multiples) $d_\alpha^{(W,m)}$, à un taux de fausses découvertes fixé à $\alpha = 0.05$ (BENJAMINI et HOCHBERG, 1995).

Pour quantifier davantage les différences entre la classification à partir des estimées multivariées $\hat{H}_m^{(\mathrm{M,bc})}$ et la classification à partir des estimées univariées $\hat{H}_m^{(\mathrm{U})}$, un score de performance globale à travers les composantes est défini comme la distance signée normalisée entre les p-valeurs ordonnées $p_{\tau(m)}^{(W)}$ et les seuils de Benjamini-Hochberg $d_\alpha^{(W,m)}$,

$$\text{score} = \frac{1}{M}\sum_{m=1}^{M}(d_\alpha^{(W,m)} - p_{\tau(m)}^{(W)}). \tag{5.7}$$

La figure 5.8 compare, pour chacun des 8 sujets, les p-valeurs $p_{\tau(m)}^{(W)}$ pour $\hat{H}_m^{(\mathrm{M,bc})}$ (en rouge avec 'o') et $\hat{H}_m^{(\mathrm{U})}$ (en bleu avec +') aux seuils de Benjamini-Hochberg $d_\alpha^{(W,m)}$ (en lignes pointillées noires) et rapporte les scores correspondants. La figure 5.8 montre que les estimées multivariées $\hat{H}_m^{(\mathrm{M,bc})}$ mènent systématiquement à de plus petites p-valeurs et ainsi des scores globaux plus élevés, et ce, de façon importante pour certains sujets.

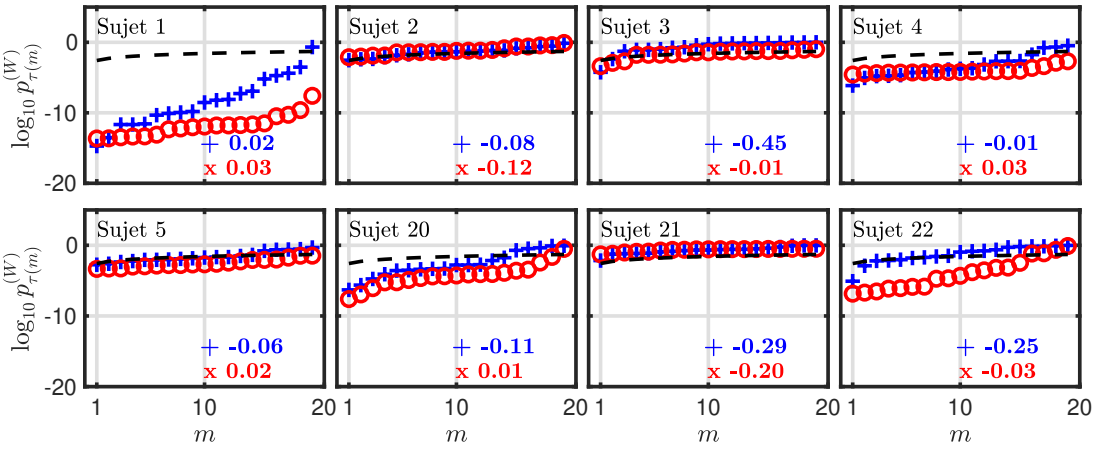


FIGURE 5.8 – **Comparaisons des distributions des estimées des exposants d'autosimilarité d'états préictaux et intericteaux.** Logarithmes des p-valeurs ordonnées $\log_2 p_{\tau(m)}^{(W)}$ du test de la somme des rangs de Wilcoxon entre les distributions des estimées univariées $\hat{H}_m^{(\mathrm{U})}$ (en bleu avec le symbole '+') et multivariées $\hat{H}_m^{(\mathrm{M,bc})}$ (en rouge avec le symbole 'o') des exposants d'autosimilarité associées à des états préictaux et celles associées à des états intericteaux pour les 8 différents sujets, avec le logarithme du seuil de Benjamini-Hochberg (superposé en lignes pointillées noires) pour un taux de fausses découvertes $\alpha = 0.05$. Les scores associés (cf. Eq. (5.7)) sont rapportés pour chaque sujet en bas à droite avec la couleur et le symbole correspondants.

Ces résultats confirment

(i) les différences statistiquement significatives entre la dynamique temporelle des états préictaux et intericteaux individu par individu,

(ii) et la capacité améliorée de l'analyse multivariée $(\hat{H}_1^{(\mathrm{M,bc})}, \ldots, \hat{H}_M^{(\mathrm{M,bc})})$ à évaluer ces différences.

5.4.2.4 Performances de la détection d'états préictaux par sujet

Pour mieux quantifier les avantages de l'analyse d'autosimilarité multivariée pour détecter les états préictaux individu par individu, des courbes ROC sont calculées comme suit, pour chaque sujet indépendamment. Premièrement, 100 fenêtres intericteales sont sélectionnées aléatoirement, à partir desquelles les exposants d'autosimilarité sont estimés et utilisés pour définir les distributions empiriques des estimées des exposants d'autosimilarité H_m sous l'hypothèse

nulle (état interictal). Deuxièmement, pour les N_w fenêtres préictales disponibles et pour N_w fenêtres interictales choisies au hasard (et n'appartenant pas à l'ensemble des 100 fenêtres utilisées pour créer les distributions sous l'hypothèse nulle), les exposants d'autosimilarité sont estimés. Troisièmement, à partir de ces estimations, des p-valeurs sont calculées par comparaison avec les distributions des estimées des H_m sous l'hypothèse nulle puis comparées aux seuils de correction de Benjamini-Hochberg (Benjamini et Hochberg, 1995) pour des comparaisons multiples, avec un taux de fausses découvertes prédéfini α. Une décision de rejet (de l'état interictal) est prise dès que l'une des $M = 19$ p-valeurs est inférieure à ce seuil. Quatrièmement, la moyenne de ces décisions sur les N_w fenêtres préictales et interictales permet de calculer les probabilités de détection correcte et de fausses alarmes pour chaque taux de fausses découvertes prédéfini α. Ces probabilités empiriques sont tracées les unes par rapport aux autres pour obtenir des courbes ROC.

Cette procédure est effectuée indépendamment pour les $M = 19$ estimées univariées $\hat{H}_m^{(\mathrm{U})}$ et multivariées $\hat{H}_m^{(\mathrm{M,bc})}$ ainsi que les $M(M+1)/2 = 190$ estimées multivariées classiques $\hat{H}_{m,m'}^{(\mathrm{U})}$. La figure 5.9 compare, pour chacun des 8 sujets indépendamment, les courbes ROC résultantes et l'aire sous la courbe (AUC) correspondante pour les $\hat{H}_m^{(\mathrm{U})}$ (lignes bleues avec '+'), les $\hat{H}_m^{(\mathrm{M,bc})}$ (lignes rouges avec 'o') et les $\hat{H}_{m,m'}^{(\mathrm{U})}$ (lignes noires avec Δ'). La figure 5.9 montre que l'approche multivariée atteint les performances les plus satisfaisantes. Premièrement, elle est toujours plus performante que la stratégie univariée. Deuxièmement, bien qu'elle soit surpassée par la stratégie multivariée classique pour deux sujet (les sujets 2 et 3), elle fait essentiellement aussi bien et parfois significativement mieux (les sujets 21 et 22) que la stratégie multivariée classique, qui peut montrer une faible sensibilité, en effectuant seulement $M = 19$ tests au lieu de $M(M+1)/2 = 190$.

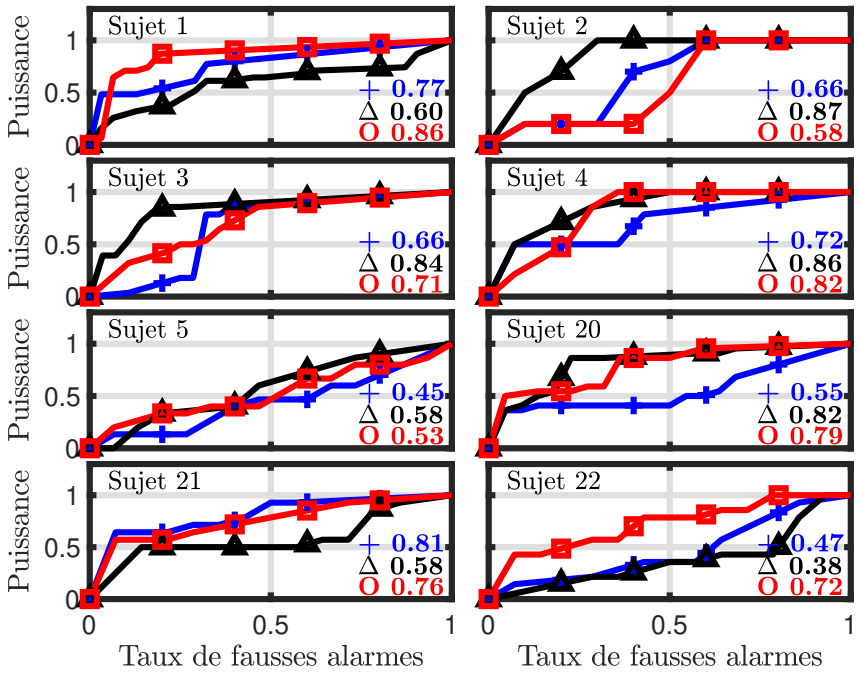


Figure 5.9 – **Prédiction de crises d'épilepsie à partir des estimées des exposants d'autosimilarité.** Courbes ROC et valeurs d'AUC associées des décisions des tests de rejet des états préictaux à partir des distributions des estimées univariées $\hat{H}_m^{(\mathrm{U})}$ (en bleu avec le symbole '+'), multivariées classiques $\hat{H}_{m,m'}^{(\mathrm{U})}$ (en noir avec le symbole 'Δ') et multivariées $\hat{H}_m^{(\mathrm{M,bc})}$ (en rouge avec le symbole 'o') pour les 8 différents sujets.

5.4.3 Conclusions

La présente section a montré la pertinence de la comparaison, par sujet, de la dynamique temporelle invariante d'échelle des données d'EEG multicanaux du cuir chevelu dans les états interictaux et préictaux pour la prédiction des crises épileptiques. Il a également été démontré que la dynamique temporelle multivariée invariante d'échelle évaluée par une analyse d'autosimilarité multivariée à partir des valeurs propres estimées du spectre d'ondelettes multivarié, développée dans le chapitre 2, surpasse les analyses univariée ou multivariée classique.

Conclusion

Bilan

Ce manuscrit s'est intéressé à l'estimation des exposants d'échelle régissant des séries temporelles autosimilaires multivariées. Plus précisément, le travail présenté s'est focalisé sur le modèle du mouvement brownien opérateur-fractionnaire (mBof), caractérisé par le vecteur des exposants d'autosimilarité $\underline{H} = (H_1, \ldots, H_M)$. Après une description des outils d'étude de l'autosimilarité multivariée, constituant l'état de l'art avant cette thèse relatée dans le chapitre 1, un nouvel estimateur, dit multivarié corrigé, de $\underline{H}$ a été proposé dans le chapitre 2. Au même titre qu'un estimateur pré-existant, ce nouvel estimateur est construit à partir de régressions linéaires sur les valeurs propres estimées du spectre d'ondelettes multivarié du mBof au travers des échelles. Dans cette nouvelle procédure, les valeurs propres sont estimées de sorte que leurs biais soient similaires à travers les échelles, évitant un biais de taille finie, induit par le fait que le nombre de coefficients d'ondelettes disponibles à chaque échelle dépend de l'échelle, dont souffrait la procédure d'estimation pré-existante.

Les performances asymptotiques théoriques de l'estimateur multivarié corrigé dans la limite de grandes tailles d'échantillon N sont identiques à celles de l'estimateur pré-existant : il est cohérent et asymptotiquement normal, de variance et corrélation décroissantes avec N, sous les mêmes hypothèses peu restrictives. Afin de comparer les performances empiriques des deux estimateurs, un modèle simplifié issu du mBof, nommé mouvement brownien fractionnaire multivarié (M-mBf), adapté au cas pratique et pouvant facilement être synthétisé numériquement a été proposé et étudié. Grâce à des M-mBf synthétiques, il a pu être montré que les performances pratiques de l'estimateur multivarié corrigé sont robustes et efficaces, mais surtout qu'elles surpassent celles de l'estimateur pré-existant, surmontant le biais de taille finie. Cette étude confirme également la rapide convergence des propriétés asymptotiques théoriques de l'estimateur proposé et montre la faible influence des paramètres du M-mBf sur celles-ci. L'estimateur multivarié corrigé a également été confronté à un estimateur univarié ne prenant pas en compte les dépendances entre les composantes du M-mBf, et s'avère plus adaptée que ce dernier dès lors que le M-mBf ne se résume pas à une collection de mBf. Les outils d'estimation proposés sont ainsi opérationnels pour rendre compte de l'autosimilarité multivariée de signaux du monde réel.

L'estimateur multivarié corrigé a en particulier permis de mettre en lumière l'intérêt du modèle d'autosimilarité multivariée à travers des applications biomédicales dans le chapitre 5. Alors que l'état de l'art avant cette thèse était principalement restreint à une analyse univariée de l'invariance d'échelle des signaux physiologiques, les travaux menés montrent que la prise en compte

des dépendances temporelles dans l'étude de l'invariance d'échelle est une source d'information importante pour réaliser certaines tâches de classification. En effet, l'usage de l'estimateur multivarié corrigé a mené à des performances plus satisfaisantes que l'utilisation de l'estimateur univarié dans les deux applications considérées. En particulier, bien que les performances atteintes pour la prédiction de crises d'épilepsie demeurent limitées, les résultats montrent bien l'avantage de l'approche multivariée par rapport à une approche univariée, encourageant d'explorer davantage l'analyse d'autosimilarité multivariée sur des données d'électroencéphalogramme. Par ailleurs, les outils d'estimation multivariée proposés permettent l'élaboration de stratégies de classification à partir d'autant d'attributs que les outils d'estimation univariée : l'ajout d'informations sur les dépendances temporelles se fait de façon parcimonieuse.

Pour tenir compte des fluctuations de l'estimation, des procédures de dénombrement et regroupement des exposants d'autosimilarité $H_1, \ldots, H_M$ à partir d'une unique observation de données ont été proposées dans le chapitre 3. Les différentes procédures reposent sur l'exploitation d'une méthode de ré-échantillonnage bootstrap par blocs dans le domaine des ondelettes, permettant de conserver la structure de dépendance en temps et en composantes des coefficients d'ondelettes utilisés pour l'estimation. Parmi les trois procédures proposées, une procédure de partitionnement spectral du graphe des exposants d'autosimilarité $H_1, \ldots, H_M$ montre des performances asymptotiques satisfaisantes dans la limite des grandes tailles d'échantillon N, et ce même pour de grands nombres de composantes M. Le graphe est construit à l'aide de $M(M-1)/2$ tests d'égalité entre les exposants $H_1, \ldots, H_M$ valables lorsque l'estimateur multivarié corrigé suit approximativement une loi normale multivariée, comportement ayant été observé dans un cadre assez général dans le chapitre 1 et pouvant être vérifié en pratique sur les échantillons bootstrap. En complément, la procédure de partitionnement spectral ne fournissant pas d'intervalle de confiance, une approche paramétrée de la pondération du graphe a été proposée afin de permettre d'adapter l'estimation du nombre de partitions en pratique. La procédure ainsi construite ne nécessite aucun réglage préalable ni entraînement sur des données et est libre de tout paramètre, excepté le paramétrage possible de la pondération du graphe. Cette procédure est ainsi prête à être exploitée sur des données du monde réel dont la taille d'échantillon N est grande comparée au nombre de composantes M. Bien que cette procédure n'ait pas encore été déployée dans des applications, le nombre d'exposants d'autosimilarité distincts et la taille des groupes d'exposants égaux pourraient être des attributs utiles dans certaines applications.

Enfin, le chapitre 4 a abordé la grande dimension, régime asymptotique où le nombre de composantes M et la taille d'échantillon N tendent conjointement vers l'infini. Il a été mis en évidence empiriquement que le nombre de valeurs distinctes dans $\underline{H}$ est asymptotiquement donné par le nombre de modes de la distribution empirique des estimées multivariées corrigées. En a résulté l'élaboration de nouvelles procédures pour estimer le nombre d'exposants effectivement distincts dans $\underline{H}$, ainsi que pour estimer leurs valeurs et la proportion de chacune de ces valeurs dans $\underline{H}$, à partir d'une unique observation de données. Bien que reposant sur des conjectures, les procédures proposées montrent des performances empiriques satisfaisantes pour de grands nombres de composantes M et des tailles d'échantillon N réalistes proches de M, et sont ainsi prêtes pour une application sur des données du monde réel concordantes. Ainsi, des outils d'estimation, de dénombrement et de partitionnement de $\underline{H}$ sont disponibles autant pour des configurations de faible dimension, c'est-à-dire pour des tailles d'échantillon N grandes comparées au nombre de composantes M, que de grande dimension, c'est-à-dire pour de grands nombres de composantes M pouvant être du même ordre que la taille d'échantillon N. L'analyse d'autosimilarité multivariée est ainsi rendue possible dans une large variété d'applications.

Perspectives

Applications

Bien que le modèle d'autosimilarité multivariée ait été exploité dans différentes applications où son avantage par rapport à une approche univariée a été démontrée, les performances obtenues dans ces applications s'avèrent limitées. Puisque l'approche multivariée semble particulièrement adaptée à la prédiction de crises d'épilepsies, il serait intéressant de réaliser une classification à partir d'un plus grand nombre de fenêtres associées à des états préictaux (survenant avant une crise) pour mettre en place une stratégie de classification adaptée. Par exemple, le recours à des stratégies d'apprentissage automatique, telles que les machines à vecteurs de support, pourrait être envisagé. La réalisation d'autres tâches à partir de données d'électroencéphalographie, telles que le diagnostique de la maladie d'Alzheimer pour lequel la caractérisation de l'invariance d'échelle a déjà été exploitée dans un contexte univarié (Ando et collab., 2021), est également une perspective intéressante. L'utilisation des méthodes de dénombrement et regroupement des exposants d'autosimilarité (en faible dimension) pour de telles tâches de classification mériterait en particulier d'être étudiée. Quant aux outils de grande dimension, ils pourraient être exploités sur des données issues de la magnétoencéphalographie et de l'imagerie par résonance magnétique fonctionnelle, où le nombre de séries temporelles est grand comparé à leur taille (Ciuciu et collab., 2012).

Irréversibilité en temps

Dans certaines applications, les signaux multivariés à analyser ne vérifient pas la propriété de réversibilité en temps. Par exemple, différents signaux acquis par électroencéphalographie peuvent être en retard les uns par rapport aux autres en raison de la conduction volumique : plusieurs sources contribuent aux signaux obtenus par chaque électrode. La distance entre une électrode et une source variant selon la source, les signaux émis par les sources sont mélangés avec des déphasages. Il est donc important d'étendre l'étude menée dans ce manuscrit à un modèle d'autosimilarité multivariée incorporant l'irréversible en temps. Or, les différents théorèmes sur les performances asymptotiques des estimateurs présentés dans les chapitres 1 et 2 n'ont été démontrés que sous l'hypothèse de réversibilité en temps (voir 1.3.2 pour plus de détails sur les hypothèses de travail). Les performances asymptotiques des estimateurs de $\underline{H}$ ont ainsi été étudiées empiriquement pour un M-mBf, processus réversible en temps, dans le chapitre 2. Les outils d'estimation mériteraient d'être également étudiés sur une extension irréversible en temps du M-mBf.

L'irréversibilité en temps du M-mBf pourrait par exemple être contrôlée par l'ajout d'un paramètre (matriciel) η dans la structure de covariance donnée par l'équation (2.22), comme proposé par Amblard et Coeurjolly (2011) en l'absence de mélange ($W = \mathbb{I}$). La fonction de covariance d'une collection de M mBf $\mathcal{B}_1, \ldots, \mathcal{B}_M$ est donnée dans ce dernier cas par

$$\mathbb{E}\left[\mathcal{B}_m(t)\mathcal{B}_{m'}(s)^T\right] = \frac{\sigma_m \sigma_{m'}}{2}\left(w_{m,m'}(t) + w_{m,m'}(-s) - w_{m,m'}(t-s)\right), \tag{5.8}$$

pour tous $m, m' \in \{1, \ldots, M\}$ et pour tous $t, s \in \mathbb{R}$, où

$$w_{m,m'}(t) = \begin{cases} (\rho_{m,m'} - \eta_{m,m'}\,\mathrm{signe}(t))\,|t|^{H_m + H_{m'}} & \text{si } H_m + H_{m'} \neq 1 \\ \rho_{m,m'}\,|t| + \eta_{m,m'}\,t\ln(|t|) & \text{si } H_m + H_{m'} = 1, \end{cases} \tag{5.9}$$

avec σ_m l'écart-type de la composante m et $\rho_{m,m'}$ la corrélation entre les composantes m et m'. Ce sont les paramètres $\eta_{m,m'}$ qui contrôlent ainsi la réversibilité en temps : le processus

ainsi défini est réversible en temps si et seulement si $\eta_{m,m'} = 0$ pour tous $m, m' = 1, \ldots, M$. La fonction de covariance des M mBf mélangés linéairement par une matrice W, constituant ainsi un M-mBf irréversible en temps $\mathcal{B}_{\eta,\Sigma,W,\underline{H}}$, est alors donnée par, pour tous $t, s \in \mathbb{R}$,

$$\mathbb{E}\left[\mathcal{B}_{\eta,\Sigma,W,\underline{H}}(t)\mathcal{B}_{\eta,\Sigma,W,\underline{H}}(s)^T\right] = W\,\mathbb{E}\left[\mathcal{B}(t)\mathcal{B}(s)^T\right]\,W^T. \tag{5.10}$$

Cette relation entre les fonctions de covariance des mBf mélangés et non mélangés est la même que dans l'équation (2.22). Le paramètre d'irréversibilité en temps η intervient donc exclusivement dans la structure de covariance de la collection des M mBf et est par conséquent indépendant du mélange W.

Une étude empirique du comportement de la procédure d'estimation sur des M-mBf irréversibles en temps synthétiques permettrait d'évaluer la nécessité d'adapter la procédure d'estimation. Une perspective pour adapter la procédure d'estimation le cas échéant serait de mesurer la phase du M-mBf par le biais de la transformée en ondelettes complexe, comme cela a été réalisé en l'absence de mélange ($W = \mathbb{I}$) par Coeurjolly et collab. (2013). L'analyse en ondelettes réalisée dans la section 2.4.5 doit alors être reprise pour un M-mBf irréversible en temps pour ajuster les outils d'estimation. Les performances asymptotiques, en particulier la consistance et la normalité multivariée, de ces outils d'estimation devraient ensuite être soumises à une étude substantielle.

Images et anisotropie

Le présent travail se concentre sur l'étude de signaux, mais l'invariance d'échelle se manifeste également dans les images, qui peuvent être multivariées comme en témoigne l'imagerie hyperspectrale. L'extension des outils d'analyse, présentés dans ce manuscrit, à des images représente donc un enjeu important.

En première approche, la relation d'autosimilarité multivariée unidimensionnelle (1.21) peut aisément être étendue pour un champ aléatoire Y, comme suit :

$$\forall a > 0, \quad \{Y(x)\}_{x\in\mathbb{R}^2} \overset{fdd}{=} \left\{a^{\underline{\underline{H}}} Y(x)\right\}_{x\in\mathbb{R}^2}, \tag{5.11}$$

où $\underline{\underline{H}}$ est la matrice de Hurst de taille $M \times M$, contrôlée par le vecteur des exposants d'autosimilarité $\underline{H}$. L'analyse en ondelettes du champ aléatoire Y est alors similaire à l'analyse unidimensionnelle réalisée dans la section 1.4.1.3. L'estimation du vecteur des exposants d'autosimilarité H_m pour un mBof à deux dimensions pose tout de même de nouvelles questions. Les outils d'analyse multi-échelle de l'autosimilarité pour les images (Atto et collab., 2013; Clausel et Vedel, 2013) permettraient d'étendre naturellement la procédure d'estimation multivariée corrigée, proposée dans ce travail, à ce cadre. Toutefois, l'estimateur multivarié corrigé repose sur le calcul, donné par l'équation (2.1), de plusieurs spectres d'ondelettes multivariés empiriques $S^{(w)}(2^j)$ à partir de fenêtres (temporelle) F_w de coefficients d'ondelettes multivariés $\{D_Y(2^j,k)\}_{k\in F_w}$ de même taille $\mathrm{Card}(F_w) = n_{j_2}$ à différentes échelles $2^{j_1} \leq 2^j \leq 2^{j_2}$, pour tout $w \in \{1, \ldots, 2^{j_2-j}\}$. Pour adapter la procédure d'estimation à un champ aléatoire Y, il est nécessaire de définir des fenêtres F_w de coefficients d'ondelettes $D_Y(2^j,\cdot)$ à deux dimensions. Les performances asymptotiques de la procédure résultante exigent une étude théorique et empirique spécifique.

Cependant, dans une image, le caractère invariant d'échelle peut être anisotrope. Dans ce cas, l'autosimilarité multivariée d'un champ aléatoire Y modélisant une telle image peut s'écrire

comme suit (Clausel et Vedel, 2011; Didier et collab., 2018) :

$$\forall a > 0, \quad \left\{ Y\left(a^E x\right) \right\}_{x \in \mathbb{R}^2} \overset{fdd}{=} \left\{ a^{\underline{H}} Y\left(x\right) \right\}_{x \in \mathbb{R}^2}, \tag{5.12}$$

où E est une matrice de taille 2×2 qui caractérise l'anisotropie et, pour toute matrice B, a^B est la matrice définie selon l'équation (1.20). Cette relation signifie que, après un changement d'échelle $x \to a^E x$ qui dépend de la direction, pour tout facteur de dilatation $a > 0$, les statistiques de Y sont covariantes par changement d'amplitude matriciel $Y \to a^{\underline{H}} Y$. L'anisotropie de la dilatation spatiale nécessite d'être prise en compte pour une estimation robuste de $\underline{H}$ conjointement à une estimation de E. Le changement d'échelle matriciel $x \to a^E x$ dans cette relation d'autosimilarité multivariée se répercute nécessairement sur l'analyse en ondelettes du champ aléatoire Y et remet en question les comportements en loi de puissance des valeurs propres du spectre d'ondelettes de Y. L'anisotropie demande donc une étude analytique approfondie et suggère l'élaboration d'une nouvelle procédure d'estimation adaptée. Pour les images univariées anisotropes, une procédure d'estimation de l'exposant d'autosimilarité H reposant sur la transformée en ondelettes hyperboliques a par exemple été proposée (Roux et collab., 2013). L'extension d'un tel outil à des images multivariées pourrait être envisagée. Il faudrait alors étudier la pertinence d'une estimation à partir des valeurs propres du spectre d'ondelettes hyperboliques multivarié.

Annexe A

Démonstrations

A.1 Théorème 2.1 (Lois de puissance asymptotiques)

Le démonstration de ce théorème est une adaptation de celle du théorème 3.1 de Abry et collab. (2022).

Démonstration. Tout d'abord, il faut montrer que, lorsque N tend vers $+\infty$,

$$\forall w \in \{1, \dots, 2^{j_2^0 - j}\}, \quad \forall m \in \{1, \dots, M\}, \quad \frac{\hat{\lambda}_m^{(w)}(a(N)2^j)}{a(N)^{2H_m+1}} \xrightarrow{\mathbb{P}} \xi_m(2^j) > 0, \tag{A.1}$$

où les *valeurs propres asymptotiques ré-échantillonnées* $\xi_m(2^j)$ sont des fonctions déterministes satisfaisant la relation d'échelle

$$\forall m \in \{1, \dots, M\}, \quad \xi_m(2^j) = (2^j)^{2H_m+1} \xi_m(2^0). \tag{A.2}$$

On fixe $m \in \{1, \dots, M\}$, $w \in \{1, \dots, 2^{j_2^0 - j}\}$ et $j \in \{j_1^0, \dots, j_2^0\}$. Supposons pour le moment que la convergence (A.1) est vraie. En écrivant $\widetilde{a}(N) = a(N)2^j$, on a :

$$\xi_m(2^j) = \underset{N\to\infty}{p\text{-lim}} \frac{\hat{\lambda}_m^{(w)}(a(N)2^j)}{a(N)^{2H_m+1}} = \underset{N\to\infty}{p\text{-lim}} \frac{\hat{\lambda}_m^{(w)}(\widetilde{a}(N))}{(\widetilde{a}(N)2^{-j})^{2H_m+1}} = 2^{j(2H_m+1)} \xi_m(2^0), \tag{A.3}$$

où $p\text{-lim}_{N\to\infty}$ désigne la limite en probabilité. Ceci établit la relation entre échelles (A.2).

À présent, il faut montrer l'équation (A.1). L'argument découle d'une simple adaptation de la démonstration établissant le théorème 3.1 de Abry et collab. (2021), correspondant au théorème 1.4 relatant les lois de puissances asymptotiques des $\hat{\lambda}_m(2^j)$.

Plus précisément, dans ce théorème, les mesures sont données par un processus stochastique signal plus bruit de grande dimension de la forme $Y(t) = P(N)X(t) + Z(t)$, où, pour tout $p \geq M$, $Y(t), Z(t) \in \mathbb{R}^p$, $P(N) \in \mathcal{M}(p, M, \mathbb{R})$ et $X(t) \in \mathbb{R}^M$. $X(t)$ est le processus stochastique fractionnaire latent (le "signal"), $Z(t)$ est la composante "bruit", et $P(N)$ est une matrice de coordonnées de rang plein. Supposons que la dimension p est fixe et fixée à M et que $X(t)$ est un mBof vérifiant (OFBM1−3) (cf. Eq. (1.25)–(1.27)). Ensuite, définissons la matrice $P(N)$ comme la matrice identité de taille $M \times M$. De plus, supposons que le terme de bruit $Z(t)$

est identiquement nul. Afin d'établir (A.1), il suffit alors de suivre le reste de l'argument de la preuve du théorème 3.1 dans Abry et collab. (2022). Ceci établit le théorème. □

A.2 Théorème 2.2 (Consistance)

Ce théorème est une conséquence immédiate du théorème A.1.

Démonstration. Fixons $m \in \{1, \dots, M\}$. On peut réécrire

$$\begin{aligned} \hat{H}_m^{(\mathrm{M,bc})} &= \frac{1}{2} \sum_{j=j_1^0}^{j_2^0} w_j \Big(\frac{1}{2^{j_2^0 - j}} \sum_{w=1}^{2^{j_2 - j}} \log_2 \lambda_m^{(w)}(a(N)2^j) \Big) - \frac{1}{2} \\ &= \frac{1}{2} \sum_{j=j_1^0}^{j_2^0} \frac{w_j}{2^{j_2^0 - j}} \sum_{w=1}^{2^{j_2^0 - j}} \left(\log_2 \frac{\lambda_m^{(w)}(a(N)2^j)}{a(N)^{2H_m+1}} - \log_2 a(N)^{2H_m+1} \right) - \frac{1}{2}. \end{aligned} \tag{A.4}$$

D'où, par l'équation (A.1), lorsque N tend vers $+\infty$,

$$\hat{H}_m^{(\mathrm{M,bc})} \xrightarrow{\mathbb{P}} \frac{1}{2}(2H_m + 1) - \frac{1}{2} = H_m. \tag{A.5}$$

Ceci établit la convergence (2.7). □

A.3 Théorème 2.3 (Normalité asymptotique)

Le démonstration de ce théorème est une adaptation de celle du théorème 3.2 de Abry et collab. (2022).

Démonstration. On écrit $a = a(N)$ pour simplifier les notations.

La convergence des valeurs propres $\lambda_m^{(w)}(a2^j))$ du spectre d'ondelettes $\mathbb{E}[S^{(w)}(a2^j))]$ peut être établie par un argument analogue à la démonstration A.1. On a ainsi :

$$\frac{\hat{\lambda}_m^{(w)}(a(N)2^j)}{(a(N)2^j)^{2H_m+1}} \overset{\mathbb{P}}{\sim} \xi_m(2^0), \tag{A.6}$$

lorsque N tend vers $+\infty$, pour tous $m \in \{1, \dots, M\}$, $w \in \{1, \dots, 2^{j_2^0 - j}\}$ et $j \in \{j_1^0, \dots, j_2^0\}$, où ξ_m est défini dans le théorème 1.3.

On établit d'abord que, lorsque N tend vers $+\infty$,

$$\Big\{ \sqrt{n_{a,j_2^0}} \big(\log_2 \hat{\lambda}_m^{(w)}(a2^j) - \log_2 \lambda_m^{(w)}(a2^j) \big) \Big\}_{m \in \{1,\dots,M\},\, w \in \{1,\dots,2^{j_2^0 - j}\}}^{j \in \{j_1^0, \dots, j_2^0\}} \xrightarrow{d} \mathcal{N}(0, \Sigma_\lambda) \tag{A.7}$$

pour une certaine matrice symétrique semi-définie positive Σ_λ. En supposons que l'équation (A.7) est vraie, la convergence (2.8) est une conséquence immédiate de (A.7).

La démonstration de l'équation (A.7) se fait par adaptation de la démonstration du théorème 3 de Abry et Didier (2018a), qui permet de démontrer le théorème 1.7 sur la normalité asymptotique de l'estimateur multivarié $\underline{\hat{H}}^{(\mathrm{M})}$. Fixons une octave $j \in \mathbb{N}$. On définit les matrices

aléatoires d'ondelettes auxiliaires, pour tout $w \in \{1, \dots, 2^{j_2^0 - j}\}$,

$$\widehat{B}_a^{(w)}(2^j) := a^{-(\underline{\underline{H}} + \frac{1}{2}\mathbb{I})} S^{(w)}(a2^j) a^{-(\underline{\underline{H}}^T + \frac{1}{2}\mathbb{I})}, \tag{A.8}$$

Sous les conditions (OFBM1-3) (cf. Eq. (1.25)–(1.27)), en conséquence du théorème 3.1 et du lemme C.2 (étendu à la dimension M) de Abry et Didier (2018b), il existe une suite de matrices déterministes

$$B_a(2^j) = a^{-(\underline{\underline{H}} + \frac{1}{2}\mathbb{I})}\, \mathbb{E}[S^{(w)}(a2^j)]\, a^{-(\underline{\underline{H}}^T + \frac{1}{2}\mathbb{I})} \in \mathcal{S}_{>0}(M, \mathbb{R}), \tag{A.9}$$

ne dépendant pas de w, telles que

$$\left\{ \sqrt{n_{a,j_2^0}} \left(\widehat{B}_a^{(w)}(2^j) - B_a(2^j) \right) \right\}_{w \in \{1, \dots, 2^{j_2^0 - j}\}} \xrightarrow{d} \mathcal{N}(0, \Sigma_{\mathbf{B}}), \tag{A.10}$$

lorsque N tend vers $+\infty$, pour une certaine matrice $\Sigma_{\mathbf{B}} \in \mathcal{S}_{\geq 0}(2^{j_2^0 - j}, \mathbb{R})$. En particulier, pour tout $w \in \{1, \dots, 2^{j_2^0 - j}\}$,

$$\widehat{B}_a^{(w)}(2^j) - B_a(2^j) \xrightarrow{\mathbb{P}} 0. \tag{A.11}$$

Fixons $m \in \{1, \dots, M\}$. Notons que, sous la condition (C0), le théorème 2.1 implique que les valeurs propres $\lambda_m^{(w)}(a2^j)$ de $\mathbb{E}[S^{(w)}(a2^j)]$ sont simples pour un N assez grand. Définissons alors, sur $\mathcal{S}_{\geq 0}(M, \mathbb{R})$, la fonction

$$B \mapsto f_{N,m}(B) := \log_2 \lambda_m \left(\frac{W \operatorname{diag}(a^{H_1}, \dots, a^{H_M})\, B \operatorname{diag}(a^{H_1}, \dots, a^{H_M})\, W^T}{a^{2H_m}} \right). \tag{A.12}$$

Pour n'importe quel $\varepsilon > 0$ fixé, soit $\mathcal{O}_{\varepsilon,a} = \{B \in \mathcal{S}_{\geq 0}(M, \mathbb{R}) |\ \|B - B_a(2^j)\| < \varepsilon\}$. Alors, pour un assez petit $\varepsilon_0 > 0$, comme les valeurs propres $\lambda_m^{(w)}(a2^j)$ de $\mathbb{E}[S^{(w)}(a2^j)]$ sont simples, lorsque N tend vers $+\infty$, la dérivée de la fonction $f_{N,m}$ existe dans l'ensemble ouvert et connexe $\mathcal{O}_{\varepsilon_0,a} \subseteq \mathcal{S}_{>0}(M, \mathbb{R})$. Alors, pour $\mathbf{B} \in \mathcal{O}_{\varepsilon_0,a}$, une application de la proposition 3 de Abry et Didier (2018b) donne une expansion de Taylor,

$$f_{N,m}(B) - f_{N,m}(B_a(2^j)) = \sum_{i_1=1}^{M} \sum_{i_2=1}^{M} \frac{\partial}{\partial b_{i_1,i_2}} f_{N,m}(\breve{B}) \cdot \pi_{i_1,i_2}(B - B_a(2^j)), \tag{A.13}$$

pour une certaine matrice $\breve{B} \in \mathcal{S}_{>0}(M, \mathbb{R})$ appartenant au segment connectant B et $B_a(2^j)$ dans $\mathcal{S}_{>0}(M, \mathbb{R})$. On définit l'évènement $A_N = \{\omega : \widehat{B}_a^{(w)}(2^j) \in \mathcal{O}_{\varepsilon_0,a},\ w \in \{1, \dots, 2^{j_2^0 - j}\}\}$. Par l'équation (A.11), $\mathbb{P}(A_N) \underset{N \to \infty}{\longrightarrow} 1$. D'où, par l'équation (A.13), pour un N assez grand et dans l'ensemble A_N, pour tous $w \in \{1, \dots, 2^{j_2^0 - j}\}$ et $m \in \{1, \dots, M\}$, l'expansion

$$f_{N,m}(\widehat{B}_a^{(w)}(2^j)) - f_{N,m}(B_a(2^j)) = \sum_{i_1=1}^{M} \sum_{i_2=1}^{M} \frac{\partial}{\partial b_{i_1,i_2}} f_{N,m}(\breve{B}_a^{(w)}(2^j)) \cdot \pi_{i_1,i_2}(\widehat{B}_a^{(w)}(2^j) - B_a(2^j)) \tag{A.14}$$

est valable pour toute matrice $\breve{B}_a^{(w)}(2^j) \in \mathcal{S}_{>0}(M, \mathbb{R})$ dans le segment reliant $\widehat{B}_a^{(w)}(2^j)$ à $B_a(2^j)$ dans $\mathcal{S}_{>0}(M, \mathbb{R})$. De plus, on peut écrire

$$\left\{ \frac{\partial}{\partial b_{i_1,i_2}} f_{N,m}(\breve{B}_a^{(w)}(2^j)) \right\}_{i_1,i_2 \in \{1,\dots,M\}} = \left\{ \frac{a^{2H_m+1}}{\lambda_m\left(\breve{S}^{(w)}(a2^j)\right)} \frac{\partial}{\partial b_{i_1,i_2}} \lambda_m \left(\frac{\breve{S}(a2^j)}{a^{2H_m+1}} \right) \right\}_{i_1,i_2 \in \{1,\dots,M\}}, \tag{A.15}$$

où $\breve{S}^{(w)}(a2^j) := W \operatorname{diag}(a^{H_1}, \dots, a^{H_M})\, \breve{B}_a^{(w)}(2^j) \operatorname{diag}(a^{H_1}, \dots, a^{H_M})\, W^T$.

À présent, en procédant comme dans la preuve du théorème 3.2 de Abry et Didier (2018b), on conclut que l'équation (A.7) est valide sous l'hypothèse que les valeurs propres $\lambda_m(a2^j)$ sont simples (condition (C0)), la normalité asymptotique, pour tous $j \in \{j_1^0, \dots, j_2^0\}$, $w \in \{1, \dots, 2^{j_2^0 - j}\}$ et $m \in \{1, \dots, M\}$, étant une conséquence de l'équation (A.10). Ceci démontre l'équation (2.8).

□

A.4 Approximations de la covariance

Les approximations proposées découlent du résultat préliminaire suivant, dont la démonstration repose sur la proposition 3 de Abry et Didier (2018a).

Théorème A.1. *Supposons que les conditions (OFBM1-3) et la condition (C0) sont vérifiées. Supposons également que les coefficients d'ondelettes sont décorrélés à toutes les échelles d'analyse $2^{j_1}, \dots, 2^{j_2}$. Alors, pour tous $m \in \{1, \dots, M\}$, $w \in \{1, \dots, 2^{j_2 - j}\}$ et $j \in \{j_1, \dots, j_2\}$,*

$$\left\{ \frac{\sqrt{N}}{\log_2 e} \left(\log_2 \hat{\lambda}_m(2^j) - \log_2 \lambda_m(2^j) \right) \right\}_{m \in \{1,\dots,M\}} \xrightarrow{d} \mathcal{N}(0, 2\mathbb{I}), \tag{A.16}$$

$$\left\{ \frac{\sqrt{N}}{\log_2 e} \left(\log_2 \hat{\lambda}_m^{(w)}(2^j) - \log_2 \lambda_m^{(w)}(2^j) \right) \right\}_{m \in \{1,\dots,M\}} \xrightarrow{d} \mathcal{N}(0, 2\mathbb{I}). \tag{A.17}$$

Démonstration. Soit une octave $j \in \{j_1, \dots, j_2\}$ fixée. Puisque les coefficients d'ondelettes sont décorrélés, la distribution du spectre d'ondelettes multivarié empirique est une distribution de Wishart,

$$N S(2^j) \overset{d}{=} \text{Wishart}(N, \mathbb{E}[S(2^j)]). \tag{A.18}$$

À présent, par une expansion de Taylor (cf. Proposition 3 de Abry et Didier (2018a)), pour tout $m \in \{1, \dots, M\}$,

$$\begin{aligned} \log_2 \hat{\lambda}_m(2^j) - \log_2 \lambda_m(2^j) &= \sum_{i_1=1}^{M} \sum_{i_2=1}^{M} \frac{\partial}{\partial S_{i_1,i_2}} \log_2 \lambda_m(\breve{S}(2^j)) \cdot \pi_{i_1,i_2}(S(2^j) - \mathbb{E}[S(2^j)]) \\ &= \frac{\log_2 e}{\lambda_m(\breve{S}(2^j))} \sum_{i_1=1}^{M} \sum_{i_2=1}^{M} \frac{\partial}{\partial S_{i_1,i_2}} \lambda_m(\breve{S}(2^j)) \cdot \pi_{i_1,i_2}(S(2^j) - \mathbb{E}[S(2^j)]). \end{aligned} \tag{A.19}$$

pour une certaine matrice $\breve{S}(2^j) \in \mathcal{S}_{>0}(M, \mathbb{R})$ appartenant au segment connectant $S(2^j)$ et $\mathbb{E}[S(2^j)]$ dans $\mathcal{S}_{>0}(M, \mathbb{R})$, où les $\hat{\lambda}_m(2^j)$ sont les valeurs propres de $S(2^j)$, les $\lambda_m(2^j)$ sont les valeurs propres de $\mathbb{E}[S(2^j)]$, les $\lambda_m(\breve{S}(2^j))$ sont les valeurs propres de $\breve{S}(2^j)$ et les π_{i_1,i_2} sont les projections sur la base canonique des matrices de taille $M \times M$.

Cependant, sous la condition (C0), les valeurs propres $\lambda_m(2^j)$ de $\mathbb{E}[S(2^j)]$ sont simples. Ainsi, les vecteurs propres $\mathbf{u}_m(N)$ de $S(2^j)$ convergent en probabilité vers un certain vecteur propre limite $\mathbf{u}_m$, i.e. $\mathbf{u}_m(N) \xrightarrow{\mathbb{P}} \mathbf{u}_m$. D'où, pour tout $m \in \{1, \dots, M\}$,

$$\frac{\partial}{\partial S_{i_1,i_2}} \lambda_m(\breve{S}(2^j)) = \mathbf{u}_m^T(N) \frac{\partial}{\partial S_{i_1,i_2}} \breve{S}(2^j)\, \mathbf{u}_m(N) = \mathbf{u}_m^T(N) \mathbf{1}_{i_1,i_2} \mathbf{u}_m(N) \xrightarrow{\mathbb{P}} \mathbf{u}_m^* \mathbf{1}_{i_1,i_2} \mathbf{u}_m. \tag{A.20}$$

D'où, pour tout $m \in \{1, \dots, M\}$, lorsque N tend vers $+\infty$,

$$\log_2 \hat{\lambda}_m(2^j) - \log_2 \lambda_m(2^j) \overset{\mathbb{P}}{\sim} \frac{\log_2 e}{\lambda_m(2^j)} \sum_{i_1=1}^{M} \sum_{i_2=1}^{M} \mathbf{u}_m^T \mathbf{1}_{i_1,i_2} \mathbf{u}_m \cdot \pi_{i_1,i_2}(S(2^j) - \mathbb{E}[S(2^j)]). \tag{A.21}$$

Puisque les valeurs propres de $\mathbb{E}[S(2^j)]$ sont simples, considérons la décomposition spectrale $\mathbb{E}[S(2^j)] = O\ \mathrm{diag}(\lambda_1(2^j), \dots, \lambda_M(2^j))\ O^T$, où, par souci de simplicité, les valeurs propres sont ordonnées, $\lambda_1(2^j) < \dots < \lambda_M(2^j)$. En multipliant devant chaque observation de vecteur gaussien par O^T, sans perte de généralité, on peut supposer que $O = \mathbb{I}$. Par conséquent, on peut également supposer que $\mathbf{u}_m = \mathbf{e}_m$ dans l'équation (A.21), où les $\mathbf{e}_m$ forment la base canonique. D'où, lorsque N tend vers $+\infty$,

$$\begin{aligned} \left\{ \frac{\sqrt{N}}{\log_2 e} \left(\log_2 \hat{\lambda}_m(2^j) - \log_2 \lambda_m(2^j) \right) \right\}_{m \in \{1,\dots,M\}} &\overset{\mathbb{P}}{\sim} \left\{ \frac{\sqrt{N}}{\lambda_m(2^j)} \pi_{m,m} \left(S(2^j) - \mathbb{E}[S(2^j)] \right) \right\}_{m \in \{1,\dots,M\}} \\ &\overset{d}{\to} \mathcal{N}(0, 2\mathbb{I}). \end{aligned} \tag{A.22}$$

Dans l'équation (A.22), la convergence vers une loi normale provient du fait que $S(2^j)$ suit une loi de Wishart (A.18).

Enfin, en réadaptant la démonstration pour les $S_m^{(w)}(2^j)$, pour tous $m \in \{1, \dots, M\}$, $w \in \{1, \dots, 2^{j_2 - j}\}$ et $j \in \{j_1, \dots, j_2\}$, on obtient également

$$\left\{ \frac{\sqrt{N}}{\log_2 e} \left(\log_2 \hat{\lambda}_m^{(w)}(2^j) - \log_2 \lambda_m^{(w)}(2^j) \right) \right\}_{m \in \{1,\dots,M\}} \overset{d}{\to} \mathcal{N}(0, 2\mathbb{I}). \tag{A.23}$$

□

On propose alors des approximations pour les échelles d'analyse $a(N)2^{j_1^0} \leq a(N)2^j \leq a(N)2^{j_2^0}$. Fixons $m \in \{1, \dots, M\}$. En supposant l'indépendance entre les valeurs propres $\lambda_m(a(N)2^j)$ au travers des échelles $a(N)2^j$, on a :

$$\mathrm{Var}(2\hat{H}_m^{(\mathrm{M})}) = \mathrm{Var}\left(\sum_{j=j_1^0}^{j_2^0} w_j \log_2 \hat{\lambda}_m(a(N)2^j) \right) = \sum_{j=j_1^0}^{j_2^0} w_j^2 \,\mathrm{Var}(\log_2 \hat{\lambda}_m(a(N)2^j)). \tag{A.24}$$

Étant donné le théorème A.1, on propose l'approximation suivante :

$$\mathrm{Var}(2\hat{H}_m^{(\mathrm{M})}) \approx 2(\log_2 e)^2 \sum_{j=j_1^0}^{j_2^0} \frac{w_j^2}{n_{a,j}}. \tag{A.25}$$

De façon similaire, en supposant l'indépendance entre les valeurs propres $\lambda_m^{(w)}(a(N)2^j)$ au travers des échelles $a(N)2^j$ et des fenêtres w, il vient :

$$\mathrm{Var}(2\hat{H}_m^{(\mathrm{M,bc})}) = \mathrm{Var}\left(\sum_{j=j_1^0}^{j_2^0} w_j \log_2 \bar{\lambda}_m(a(N)2^j)\right) \tag{A.26}$$

$$= \sum_{j=j_1^0}^{j_2^0} w_j^2 \,\mathrm{Var}\left(\frac{1}{2^{j_2^0-j}} \sum_{w=1}^{2^{j_2^0-j}} \log_2 \hat{\lambda}_m^{(w)}(a(N)2^j)\right) \tag{A.27}$$

$$= \sum_{j=j_1^0}^{j_2^0} \frac{w_j^2}{2^{j_2^0-j}} \mathrm{Var}\left(\log_2 \hat{\lambda}_m^{(1)}(a(N)2^j)\right) \tag{A.28}$$

$$\approx 2(\log_2 e)^2 \sum_{j=j_1^0}^{j_2^0} 2^{j-j_2^0} \frac{w_j^2}{n_{a,j_2^0}} \tag{A.29}$$

Puis, étant donnée l'égalité $n_{a,j} = n_{a,j_2^0}/2^{j-j_2^0}$, on obtient la même approximation que pour les estimées multivariées $\hat{H}_m^{(\mathrm{M})}$.

Les résultats sur les corrélations découlent quant à eux des décorrélations des logarithmes des valeurs propres estimées $\log_2 \hat{\lambda}_1(a(N)2^j)$, … , $\log_2 \hat{\lambda}_M(a(N)2^j)$ et $\log_2 \hat{\lambda}_1^{(w)}(a(N)2^j)$, … , $\log_2 \hat{\lambda}_M^{(w)}(a(N)2^j)$, pour tous $w \in \{1, \ldots, 2^{j_2^0-j}\}$ et $j \in \{j_1^0, \ldots, j_2^0\}$, suggérées par le théorème précédent.

A.5 Théorème 2.4 (M-mBf)

Démonstration. Puisque les processus $\mathcal{B}_{\underline{\underline{H}},A}$ et $B_{\Sigma,W,\underline{H}}$ sont gaussiens M-variés centrés, il suffit de montrer que leurs fonctions de covariance sont égales.

D'une part, en remplaçant $\underline{\underline{H}}$ par $W\mathrm{diag}(\underline{H})W^{-1}$ dans la covariance du mBof donnée par l'équation (1.22), puis en utilisant l'équation (1.20) et le fait que $\underline{\underline{H}}^k = W\mathrm{diag}(\underline{H})^k W^{-1}$, pour tout $k > 0$, on obtient, pour tout $t \in \mathbb{R}$:

$$\begin{aligned}
\mathbb{E}\left[\mathcal{B}_{\underline{\underline{H}},A}(t)\mathcal{B}_{\underline{\underline{H}},A}(t)^*\right] &= \int_{\mathbb{R}} \left|\frac{e^{itf}-1}{if}\right|^2 \Big(W f_+^{-(\mathrm{diag}(\underline{H})-\frac{1}{2}\mathbb{I})} W^{-1} A A^* (W^T)^{-1} f_+^{-(\mathrm{diag}(\underline{H})^T-\frac{1}{2}\mathbb{I})} W^T \\
&\qquad + W f_-^{-(\mathrm{diag}(\underline{H})-\frac{1}{2}\mathbb{I})} W^{-1} \overline{AA^*} (W^T)^{-1} f_-^{-(\mathrm{diag}(\underline{H})^T-\frac{1}{2}\mathbb{I})} W^T \Big)\, \mathrm{d}f \\
&= W \left(\int_{\mathbb{R}} \left|\frac{e^{itf}-1}{if}\right|^2 \Big(f_+^{-(\mathrm{diag}(\underline{H})-\frac{1}{2}\mathbb{I})} (G \odot \Sigma) f_+^{-(\mathrm{diag}(\underline{H})^T-\frac{1}{2}\mathbb{I})} \right. \\
&\qquad \left. + f_-^{-(\mathrm{diag}(\underline{H})-\frac{1}{2}\mathbb{I})} (G \odot \Sigma) f_-^{-(\mathrm{diag}(\underline{H})^T-\frac{1}{2}\mathbb{I})} \Big)\, \mathrm{d}f \right) W^T \\
&= W \left(\int_{\mathbb{R}} \left|\frac{e^{itf}-1}{if}\right|^2 R(f)\, \mathrm{d}f \right) W^T,
\end{aligned} \tag{A.30}$$

avec $R(f)$ la matrice de taille $M \times M$ dont les entrées sont définies par, pour tous $m, m' \in \{1, \dots, M\}$,

$$R_{m,m'}(f) := G_{m,m'}\, \Sigma_{m,m'}\, |f|^{-(H_m + H_{m'} - 1)}. \tag{A.31}$$

Or, pour tous $m, m' \in \{1, \dots, M\}$,

$$\int_{\mathbb{R}} \left| \frac{e^{itf} - 1}{if} \right|^2 |f|^{-(H_m + H_{m'} - 1)}\, \mathrm{d}f = |t|^{H_m + H_{m'}} \frac{2\pi}{\Gamma(H_m + H_{m'} + 1) \sin\left((H_m + H_{m'})\frac{\pi}{2}\right)}, \tag{A.32}$$

D'où, par définition de G, pour tous $m, m' \in \{1, \dots, M\}$,

$$\int_{\mathbb{R}} \left| \frac{e^{itf} - 1}{if} \right|^2 R_{m,m'}(f)\, \mathrm{d}f = |t|^{H_m + H_{m'}}\, \Sigma_{m,m'}, \tag{A.33}$$

et vient alors la variance du mBof à l'instant $t \in \mathbb{R}$,

$$\mathbb{E}\left[\mathcal{B}_{\underline{\underline{H}},A}(t)\mathcal{B}_{\underline{\underline{H}},A}(t)^T\right] = W\, |t|^{\mathrm{diag}(\underline{H})}\, \Sigma\, |t|^{\mathrm{diag}(\underline{H})}\, W^T. \tag{A.34}$$

D'autre part, à partir de la représentation harmonisable du mBf donnée par l'équation (1.3) et de l'équation (A.32), on obtient, pour tous $m, m' \in \{1, \dots, M\}$ et $t \in \mathbb{R}$,

$$\mathbb{E}\left[\mathcal{B}_m(t)\mathcal{B}_{m'}(t)^T\right] = |t|^{H_m + H_{m'}} \mathbb{E}\left[\mathcal{B}_m(1)\mathcal{B}_{m'}(1)^T\right] = |t|^{H_m + H_{m'}} \Sigma_{m,m'}, \tag{A.35}$$

et donc la variance du M-mBf à l'instant $t \in \mathbb{R}$ s'écrit

$$\mathbb{E}\left[B_{\Sigma,W,\underline{H}}(t) B_{\Sigma,W,\underline{H}}(t)^*\right] = W\, |t|^{\mathrm{diag}(\underline{H})}\, \Sigma\, |t|^{\mathrm{diag}(\underline{H})}\, W^T. \tag{A.36}$$

Finalement, les équations (A.34) et (A.36) donnent l'égalité suivante :

$$\mathbb{E}\left[\mathcal{B}_{\underline{\underline{H}},A}(t)\mathcal{B}_{\underline{\underline{H}},A}(t)^T\right] = \mathbb{E}\left[B_{\Sigma,W,\underline{H}}(t) B_{\Sigma,W,\underline{H}}(t)^T\right]. \tag{A.37}$$

Puis, on utilise le fait que, pour un processus stochastique $\{X(t)\}_{t\in\mathbb{R}}$ nul en zéro à accroissements stationnaires et centrés, on a, pour tous $t, s \in \mathbb{R}$,

$$\mathbb{E}\left[(X(t) - X(s))(X(t) - X(s))^T\right] = \mathbb{E}\left[X(t-s)X(t-s)^T\right], \tag{A.38}$$

et donc, en développant,

$$\mathbb{E}\left[X(t)X(s)^T\right] = \frac{1}{2}\left(\mathbb{E}\left[X(t)X(t)^T\right] + \mathbb{E}\left[X(s)X(s)^T\right] - \mathbb{E}\left[X(t-s)X(t-s)^T\right]\right). \tag{A.39}$$

Ainsi, l'équation (A.37) permet de déduire l'égalité entre les fonctions de covariance

$$\mathbb{E}\left[\mathcal{B}_{\underline{\underline{H}},A}(t)\mathcal{B}_{\underline{\underline{H}},A}(s)^T\right] = \mathbb{E}\left[B_{\Sigma,W,\underline{H}}(t) B_{\Sigma,W,\underline{H}}(s)^T\right]. \tag{A.40}$$

□

Annexe B

Outils pour les tests d'hypothèse

B.1 Paramètre de non-centralité du χ^2

On considère un vecteur $\underline{H}$ dont les M_1 premières entrées valent H_1 et les M_2 autres entrées valent H_2. Supposons le cas idéal où l'estimateur $\underline{\hat{H}}$ n'a ni biais, ni corrélation et des composantes de même variance. Sans perte de généralité, on peut écrire $\Sigma_{\underline{\hat{H}}} = \mathbb{I}$. Le paramètre de non-centralité s'écrit dans ce cas

$$\theta = \|\mu\|^2, \tag{B.1}$$

où $\mu := \mathbb{E}[\underline{\hat{H}} - \langle\underline{\hat{H}}\rangle]$ est constitué de M_1 et M_2 entrées μ_1 et μ_2, respectivement, définies par

$$\begin{aligned}
\mu_1 &:= H_1 - \frac{M_1 H_1 + M_2 H_2}{M} = \frac{M_1}{M}(H_2 - H_1), \\
\mu_2 &:= H_2 - \frac{M_1 H_1 + M_2 H_2}{M} = \frac{M_2}{M}(H_1 - H_2).
\end{aligned}$$

Le paramètre de non-centralité s'écrit alors de façon explicite :

$$\begin{aligned}
\theta &= M_1\mu_1^2 + M_2\mu_2^2 \\
&= M_1\left(\frac{M_2}{M}(H_1 - H_2)\right)^2 + M_2\left(\frac{M_1}{M}(H_2 - H_1)\right)^2 \\
&= M_1 M_2 \frac{M_1 + M_2}{M^2}(H_2 - H_1)^2 \\
&= \frac{M_1 M_2}{M_1 + M_2}(H_2 - H_1)^2.
\end{aligned}$$

B.2 Loi normale repliée

Si une variable aléatoire δ_m suit une loi normale de moyenne $\tilde{\mu}_m$ et d'écart-type $\tilde{\sigma}_m$, alors $\tilde{\delta}_m = |\delta_m|$ suit une loi normale repliée, de densité de probabilité

$$f_{\mathcal{FN}}(\tilde{\delta}_m) := \frac{1}{\sqrt{2\pi\tilde{\sigma}_m^2}}\left(\exp^{-\frac{(\tilde{\delta}_m + \tilde{\mu}_m)^2}{2\tilde{\sigma}_m^2}} + \exp^{-\frac{(\tilde{\delta}_m - \tilde{\mu}_m)^2}{2\tilde{\sigma}_m^2}}\right) \tag{B.2}$$

Le couple $(\tilde{\mu}_m, \tilde{\sigma}_m)$, avec $\tilde{\mu}_m > 0$ et $\tilde{\sigma}_m > 0$, peut être estimé par maximisation de la log-vraisemblance (Tsagris et collab., 2014) à partir d'échantillons indépendants $\{x_1, \ldots, x_n\}$, en résolvant le système d'équations, d'inconnu le couple $(\tilde{\mu}, \tilde{\sigma})$, suivant :

$$\sum_{i=1}^{n} \frac{1 - e^{\frac{2\tilde{\mu}x_i}{\tilde{\sigma}^2}}}{1 + e^{\frac{2\tilde{\mu}x_i}{\tilde{\sigma}^2}}} x_i + n\frac{\tilde{\mu}}{2} = 0, \quad \tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \tilde{\mu}^2. \tag{B.3}$$

Lorsque $\tilde{\mu}_m = 0$, la loi normale repliée est appelée *loi demi-normale* et sa densité de probabilité s'écrit

$$f_{\mathcal{HN}}(\tilde{\delta}_m) := f_{\mathcal{FN}}(\tilde{\delta}_m \,|\, \tilde{\mu}_m = 0) = \sqrt{\frac{2}{\pi\tilde{\sigma}_m^2}} \exp^{-\frac{\tilde{\delta}_m^2}{2\tilde{\sigma}_m^2}} \tag{B.4}$$

En particulier, le paramètre d'échelle d'une loi normale repliée, vérifie la relation :

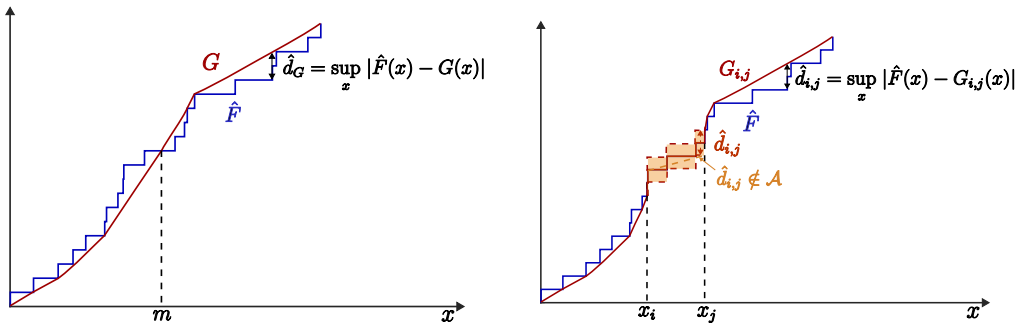
$$\tilde{\sigma}_m^2 = \frac{\mathrm{Var}(\tilde{\delta}_m)}{1 - \frac{2}{\pi}}. \tag{B.5}$$

B.3 Statistique de Hartigan

Pour calculer la statistique de Hartigan,

$$\hat{d} = \inf_{G \in \mathcal{U}} \sup_{x \in \mathbb{R}} |\hat{F}(x) - G(x)|, \tag{B.6}$$

où $\mathcal{U}$ est l'ensemble des fonctions de répartition unimodales, il faut trouver la fonction $G \in \mathcal{U}$ convexe sur $(-\infty, m]$ et concave $[m, +\infty)$ la plus proche de $\hat{F}$. Connaissant le mode m, il est facile de calculer G et donc la statistique de Hartigan, comme illustré par la figure B.1 (a).



(a) Définition de la statistique de Hartigan

(b) Itération de l'algorithme

Figure B.1 – **Illustration du principe de l'algorithme de calcul de la statistique de Hartigan.**

L'algorithme AS 217 de Hartigan (1985) permet de calculer la statistique de Hartigan sans parcourir tous les modes m possibles. Cet algorithme repose sur le théorème 6 de Hartigan et Hartigan (1985), qui s'énonce comme suit :

Théorème B.1 (Hartigan et Hartigan, 1985, Théorème 6)**.** *Soit F une fonction de répartition. Il existe une fonction non-décroissante G vérifiant, pour tout $x_L \leq x_U$,*

(i) G est le plus grand minorant convexe de $F + \hat{d}$ sur $(-\infty, x_L)$,

(ii) G a pente maximale constante sur (x_L, x_U),

(iii) G est le plus petit majorant concave de $F - \hat{d}$ sur $(x_U, +\infty)$,

(iv) $\hat{d} = \sup_{x \notin (x_L, x_U)} |\hat{F}(x) - G(x)| \geq \sup_{x \in (x_L, x_U)} |\hat{F}(x) - G(x)|$.

L'algorithme AS 217, illustré par la figure B.1 (b), procède alors comme suit. Pour l'ensemble des $x_i \leq x_j$ disponibles, l'algorithme

- (i) calcule la fonction $G_{i,j}$ telle que $G_{i,j}$ est le plus grand minorant convexe de $\hat{F}$ sur $(-\infty, x_i]$, $\hat{F} = G_{i,j}$ sur (x_i, x_j) et $G_{i,j}$ est le plus petit majorant concave de $\hat{F}$ sur $[x_j, +\infty)$;
- (ii) calcule la distance entre $\hat{F}$ et $G_{i,j}$ définie par $\hat{d}_{i,j} = \sup_x |\hat{F}(x) - G_{i,j}(x)|$;
- (iii) vérifie si le segment liant $[x_i, \hat{F}(x_i) + \hat{d}_{i,j}/2]$ à $[x_j, \hat{F}(x_j) - \hat{d}_{i,j}/2]$ appartient à $\{(x, y) \mid \hat{F}(x) - \hat{d}_{i,j}/2 \leq y \leq \hat{F}(x) + \hat{d}_{i,j}/2\}$;

La statistique de Hartigan $\hat{d}$ est alors estimée comme étant le minimum des $\hat{d}_{i,j}/2$ sur l'ensemble $\mathcal{A}$ des couples (x_i, x_j) vérifiant (iii). L'ensemble $\mathcal{A}$ correspond à l'ensemble des intervalles (x_i, x_j) où $\hat{F}$ varie suffisamment doucement, ne passant pas abruptement d'une fonction convexe à une fonction concave. La condition d'appartenance d'un intervalle (x_i, x_j) à $\mathcal{A}$ est illustrée dans la figure B.1 (b) : le segment en pointillés oranges doit appartenir à la zone orange.

Annexe C

Mesures de la qualité d'un partitionnement

Soient $U = \{U_1, U_2, \ldots, U_R\}$ et $V = \{V_1, V_2, \ldots, V_C\}$ deux partitions de $\mathcal{V} = \{1, \ldots, M\}$. La similarité entre les deux partitions U et V peut être mesurée par l'indice de Rand ajusté (ARI) et l'information mutuelle normalisée (NMI) décrits ici (Vinh et collab., 2010).

C.1 Indice de Rand ajusté (ARI)

L'indice de Rand est défini par

$$RI := \frac{a+b}{\binom{M}{2}} \tag{C.1}$$

où

- a est le nombre de pairs d'éléments de S qui sont dans le même sous-ensemble de V et le même sous-ensemble de U;
- b est le nombre de pairs d'éléments de S qui sont dans différents sous-ensembles de V et dans différents sous-ensembles de U.

L'indice de Rand ajusté s'écrit alors:

$$ARI := \frac{RI - \mathbb{E}[RI]}{\max(RI) - \mathbb{E}[RI]}. \tag{C.2}$$

En notant $M_{i,j}$ le nombre d'éléments qui sont à la fois dans les partitions U_i et V_j, comme détaillé par le tableau C.1, on a : $a + b = \sum_{i,j} \binom{M_{i,j}}{2}$. Ainsi, il vient :

$$ARI = \frac{\sum_{i,j}\binom{M_{i,j}}{2} - \frac{\sum_i \binom{M_{i,.}}{2}\sum_j \binom{M_{.,j}}{2}}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_i \binom{M_{i,.}}{2} + \sum_j \binom{M_{.,j}}{2}\right] - \frac{\sum_i \binom{M_{i,.}}{2}\sum_j \binom{M_{.,j}}{2}}{\binom{n}{2}}}, \tag{C.3}$$

où $M_{i,.} = \sum_{j=1}^{C} M_{i,j}$ et $M_{.,j} = \sum_{i=1}^{R} M_{i,j}$. Les notations sont synthétisées dans le tableau C.1.

Partition	V_1	V_2	$\cdots$	V_C	Somme
U_1	M_{11}	M_{12}	$\cdots$	M_{1C}	$M_{1.}$
U_2	M_{21}	M_{22}	$\cdots$	M_{2C}	$M_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
U_R	M_{R1}	M_{R2}	$\cdots$	M_{RC}	$M_{R.}$
Somme	$M_{.1}$	$M_{.2}$	$\cdots$	$M_{.C}$	M

TABLEAU C.1 – **Table de contingence.** Définition des nombres M_{ij} d'éléments qui sont à la fois dans les partitions U_i et V_j

L'ARI, qui prend ses valeurs entre -1 et 1, tient compte des regroupements aléatoires et permet de comparer deux partitions de nombres de classes différentes.

C.2 Information mutuelle normalisée (NMI)

L'information mutuelle normalisée est définie par :

$$NMI := \frac{I(U,V)}{\sqrt{H(U)H(V)}} := \frac{H(U)+H(V)-H(U,V)}{\sqrt{H(U)H(V)}}, \tag{C.4}$$

où

$$H(U) := -\sum_i q_{i,.} log2(q_{i,.}), \tag{C.5}$$

$$H(V) := -\sum_j q_{.,j} log2(q_{.,j}), \tag{C.6}$$

$$H(U,V) := -\sum_{i,j} q_{i,j} log2(q_{i,j}), \tag{C.7}$$

avec $q_{i,j} = \mathbb{P}(U_i \cap V_j)$ la proportion d'éléments à la fois dans U_i et V_j, $q_{i,.} = \mathbb{P}(U_i)$ la proportion d'éléments dans U_i et $q_{.,j} = \mathbb{P}(V_j)$ la proportion d'éléments dans V_j.

La NMI prend ses valeurs entre 0, valeur atteinte lorsque U et V sont des variables aléatoires indépendantes, et 1, valeur atteinte lorsque $U = V$.

Annexe D

Formalisme en grande dimension

D.1 Régime asymptotique

Dans le cadre de la grande dimension, le nombre de composantes M n'est pas fixe mais tend vers l'infini lorsque la taille d'échantillon N tend vers l'infini,

$$M(N) \underset{N\to+\infty}{\longrightarrow} +\infty. \tag{D.1}$$

Dans le contexte de l'analyse par ondelettes, l'estimateur multivarié corrigé $\underline{\hat{H}}^{(\mathrm{M,bc})} := (\hat{H}_1^{(\mathrm{M,bc})}, \ldots, \hat{H}_M^{(\mathrm{M,bc})})$ est obtenu, selon l'équation (2.3), par des régressions linéaires sur les logarithmes des valeurs propres estimées $\log_2 \bar{\lambda}_1(2^j)$, …, $\log_2 \bar{\lambda}_{M(N)}(2^j)$ au travers d'échelles d'analyse $2^{j_1(N)} \leq 2^j \leq 2^{j_2(N)}$ allant également vers l'infini lorsque N tend vers $+\infty$. Plus précisément, à l'instar du régime asymptotique présenté dans la section 1.4.3.3, la gamme d'échelles d'analyse est donnée par $\{2^{j_1(N)}, \ldots, 2^{j_2(N)}\} := \{a(N)2^{j_1^0}, \ldots, a(N)2^{j_2^0}\}$, où $a(N)$ est une suite vérifiant l'équation (1.52), et j_1^0 et j_2^0 correspondent respectivement aux plus petite et grande octaves d'analyse à une taille d'échantillon N_0 telle que $a(N_0) = 1$. Ainsi, le nombre d'échelles d'analyse $2^{j_1(N)} \leq 2^j \leq 2^{j_2(N)}$ est constant et vaut $j_2(N) - j_1(N) + 1 = j_2^0 - j_1^0 + 1$.

Le comportement en grande dimension implique alors une *triple limite* (Abry et collab., 2022) :

$$\frac{M(N)}{N/a(N)} \underset{N\to+\infty}{\longrightarrow} c \in [0, +\infty). \tag{D.2}$$

En notant $n_{a,j} = N/(a(N)2^j)$ le nombre de coefficients d'ondelettes disponibles à l'échelle $a(N)2^j$, le rapport $M(N)/n_{a,j}$ est donc asymptotiquement constant pour chaque $j \in \{j_1^0, \ldots, j_2^0\}$ fixé. En particulier, les quantités $\log_2 \bar{\lambda}_m(2^j)$ sont calculées à partir de $a(N)2^{j_2^0}/a(N)2^j = 2^{j_2^0 - j}$ fenêtres de coefficients d'ondelettes de taille $n_{j_2(N)} = n_{a,j_2^0}$ telle que $M(N)/n_{j_2(N)}$ est asymptotiquement constant. Le détail du calcul des $\log_2 \bar{\lambda}_m(2^j)$ est donné dans la section 2.2.

Pour alléger les notations, dans l'ensemble du chapitre 4, le nombre de composantes $M(N)$ est noté M, et les plus petite et grande octaves d'analyses $j_1(N)$ et $j_2(N)$ sont notées j_1 et j_2.

D.2 Comportement asymptotique de l'estimateur multivarié

Formellement, en grande dimension, on considère un mBof qui vérifie les hypothèses (OFBM1-3) (cf. Eq. (1.25)–(1.27)) et dont le vecteur des exposants d'autosimilarité $\underline{H}$ n'est pas déterministe mais est un vecteur de M échantillons indépendants issus d'une même distribution discrète π de support $\{H_1, \ldots, H_L\}$, avec $0 < H_1 \leq \ldots \leq H_L < 1$ et $L \in \mathbb{N}$. Ce modèle permet de fixer le nombre L de valeurs distinctes dans $\underline{H}$, les différentes valeurs $H_1, \ldots, H_L$ prises par les entrées de $\underline{H}$ et les proportions d'entrées $\pi(H_1)$, ..., $\pi(H_L)$ de $\underline{H}$ prenant chacune de ces valeurs, de façon indépendante du nombre de composantes M. L'analyse de l'autosimilarité multivariée en grande dimension revient alors à estimer π, ce qui signifie simplement estimer les valeurs $H_1, ..., H_L$ et leurs proportions $\pi(H_1)$, ..., $\pi(H_L)$.

Dans ce cadre, un travail en cours de Gustavo Didier et Oliver Orejola, de l'université Tulane, montre que la distribution empirique des $M(N)$ valeurs propres $\hat{\lambda}_1(a(N)2^j)$, ..., $\hat{\lambda}_{M(N)}(a(N)2^j)$ du spectre d'ondelettes multivarié empirique $S(a(N)2^j)$ tend asymptotiquement vers la distribution π des entrées de $\underline{H}$ sous la triple limite (D.2). Plus précisément, pour tous $\epsilon > 0$ et $q \in \{1, \ldots, L\}$, lorsque N tend vers $+\infty$,

$$\frac{\mathrm{Card}(\{m \in \{1, \ldots, M(N)\} \mid \log_2 \hat{\lambda}_m(a(N)2^j) \in (\alpha_q - \epsilon, \alpha_q + \epsilon)\})}{M(N)} \xrightarrow{\mathbb{P}} \pi(H_q), \tag{D.3}$$

où $\alpha_q = 2H_q + 1$ et $\{H_1, \ldots, H_L\}$ est le support de π.

Ce résultat a mené Orejola et collab. (2022) à conjecturer que la distribution empirique des $M(N)$ estimées multivariées $\hat{H}_1^{(\mathrm{M})}, \ldots, \hat{H}_{M(N)}^{(\mathrm{M})}$ tend également asymptotiquement vers π. Cela signifie en particulier que le nombre de modes de la distribution empirique des $\hat{H}_1^{(\mathrm{M})}, \ldots, \hat{H}_{M(N)}^{(\mathrm{M})}$ tend asymptotiquement vers le nombre de modes de π.

Références

Abry, P., R. Baraniuk, P. Flandrin, R. Riedi et D. Veitch. 2002, «Multiscale nature of network traffic», *IEEE Signal Processing Magazine*, vol. 19, n° 3, p. 28–46. 11

Abry, P., B. C. Boniece, G. Didier et H. Wendt. 2021, «On high-dimensional wavelet eigenanalysis», *arXiv preprint arXiv:2102.05761*. 187

Abry, P., B. C. Boniece, G. Didier et H. Wendt. 2022, «Wavelet eigenvalue regression in high dimensions», *Statistical Inference for Stochastic Processes*, p. 1–32. 31, 32, 46, 146, 187, 188, 201

Abry, P. et G. Didier. 2018a, «Wavelet eigenvalue regression for n-variate operator fractional Brownian motion», *Journal of Multivariate Analysis*, vol. 168, p. 75–104. 13, 14, 29, 30, 31, 188, 190

Abry, P. et G. Didier. 2018b, «Wavelet estimation for operator fractional Brownian motion», *Bernoulli*, vol. 24, n° 2, p. 895–928. 13, 14, 25, 30, 42, 189, 190

Abry, P., S. G. Roux, H. Wendt, P. Messier, A. G. Klein, N. Tremblay, P. Borgnat, S. Jaffard, B. Vedel, J. Coddington et L. A. Daffner. 2015, «Multiscale anisotropic texture analysis and classification of photographic prints: Art scholarship meets image processing algorithms», *IEEE Signal Processing Magazine*, vol. 32, n° 4, p. 18–27. 11

Abry, P. et D. Veitch. 1998, «Wavelet analysis of long-range dependent traffic.», *IEEE Transactions on Information Theory*, vol. 44, n° 1, p. 2–15. 20

Abry, P., H. Wendt et S. Jaffard. 2013, «When van gogh meets mandelbrot: Multifractal classification of painting's texture», *Signal Processing*, vol. 93, n° 3, p. 554–572. 11

Abry, P., H. Wendt, S. Jaffard et G. Didier. 2019, «Multivariate scale-free temporal dynamics: From spectral (fourier) to fractal (wavelet) analysis», *Comptes Rendus Physique*, vol. 20, n° 5, p. 489–501. 11, 25

Ahn, S., T. Nguyen, H. Jang, J. G. Kim et S. C. Jun. 2016, «Exploring neuro-physiological correlates of drivers' mental fatigue caused by sleep deprivation using simultaneous eeg, ecg, and fnirs data», *Frontiers in human neuroscience*, vol. 10, p. 219. 15, 169

Amaral, L. A. N., A. L. Goldberger, P. C. Ivanov et H. E. Stanley. 1999, «Modeling heart rate variability by stochastic feedback», *Computer physics communications*, vol. 121, p. 126–128. 168

Amblard, P.-O. et J.-F. Coeurjolly. 2011, «Identification of the multivariate fractional brownian motion», *IEEE Transactions on Signal Processing*, vol. 59, n° 11, p. 5152–5168. 12, 44, 183

Amblard, P.-O., J.-F. Coeurjolly, F. Lavancier et A. Philippe. 2012, «Basic properties of the multivariate fractional brownian motion», *Bulletin de la Société Mathématique de France, Séminaires et Congrés*, vol. 28, p. 65–87. 12, 14

Anderson, G. W., A. Guionnet et O. Zeitouni. 2010, *An Introduction to Random Matrices*, *Cambridge Studies in Advanced Mathematics*, vol. 118, Cambridge University Press, Cambridge. 13, 36

Ando, M., S. Nobukawa, M. Kikuchi et T. Takahashi. 2021, «Identification of electroencephalogram signals in alzheimer's disease by multifractal and multiscale entropy analysis», *Frontiers in Neuroscience*, vol. 15, p. 667 614. 183

Atto, A. M. et Y. Berthoumieu. 2011, «Wavelet packets of nonstationary random processes: Contributing factors for stationarity and decorrelation», *IEEE Transactions on Information Theory*, vol. 58, n° 1, p. 317–330. 12

Atto, A. M., Y. Berthoumieu et P. Bolon. 2013, «2-d wavelet packet spectrum for texture analysis», *IEEE Transactions on Image Processing*, vol. 22, n° 6, p. 2495–2500. 12, 184

Atto, A. M., D. Pastor et G. Mercier. 2010, «Wavelet packets of fractional brownian motion: Asymptotic analysis and spectrum estimation», *IEEE Transactions on Information Theory*, vol. 56, n° 9, p. 4741–4753. 12

Azran, A. et Z. Ghahramani. 2006, «Spectral methods for automatic multiscale data clustering», dans *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 1, IEEE, p. 190–197. 108

Bai, Z. et J. W. Silverstein. 2010, *Spectral Analysis of Large Dimensional Random Matrices*, vol. 20, 2nd éd., Springer, New York. 14

Bardet, J.-M., G. Lang, G. Oppenheim, A. Philippe, S. Stoev et M. S. Taqqu. 2003, «Semi-parametric estimation of the long-range dependence parameter: A survey», dans *Theory and applications of Long-range dependence*, édité par P. Doukhan, G. Oppenheim et M. S. Taqqu, Birkhäuser, Boston, p. 557–577. 12, 20

Batty, M., P. Longley et S. Fotheringham. 1989, «Urban growth and form: scaling, fractal geometry, and diffusion-limited aggregation», *Environment and planning A*, vol. 21, n° 11, p. 1447–1472. 11

Bautista Ruiz, E. 2019, *Laplacian Powers for Graph-Based Semi-Supervised Learning*, thèse de doctorat, Université de Lyon. 108, 143

BenAbdelkader, C., R. G. Cutler et L. S. Davis. 2004, «Gait recognition using image self-similarity», *EURASIP Journal on Advances in Signal Processing*, vol. 2004, n° 4, p. 1–14. 168

Benjamini, Y. et Y. Hochberg. 1995, «Controlling the false discovery rate: a practical and powerful approach to multiple testing», *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, n° 1, p. 289–300. 85, 173, 177, 179

Beran, R. 1986, «Simulated power functions», *The Annals of Statistics*, p. 151–173. 73

Bonferroni, C. 1936, «Teoria statistica delle classi e calcolo delle probabilita», *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, vol. 8, p. 3–62. 84

Boniece, B. C., G. Didier, H. Wendt et P. Abry. 2019, «On multivariate non-gaussian scale invariance: fractional lévy processes and wavelet estimation», dans *2019 27th European Signal Processing Conference (EUSIPCO)*, IEEE, p. 1–5. 12

Breiman, L. 2001, «Random forests», *Machine learning*, vol. 45, p. 5–32. 171

Brin, S. et L. Page. 1998, «The anatomy of a large-scale hypertextual web search engine», *Computer networks and ISDN systems*, vol. 30, n° 1-7, p. 107–117. 109

Buzsaki, G. 2006, *Rhythms of the Brain*, Oxford university press. 11

Cahyadi, B., W. Khairunizam, I. Zunaidi, L. H. Ling, A. Shahriman, M. Zuradzman, W. Mustafa et N. Noriman. 2019, «Muscle fatigue detections during arm movement using emg signal», dans *IOP conference series: materials science and engineering*, vol. 557, IOP Publishing, p. 012004. 15, 169

Chisci, L., A. Mavino, G. Perferi, M. Sciandrone, C. Anile, G. Colicchio et F. Fuggetta. 2010, «Real-time epileptic seizure prediction using ar models and support vector machines», *IEEE Transactions on Biomedical Engineering*, vol. 57, n° 5, p. 1124–1132. 174

Ciuciu, P., P. Abry et B. J. He. 2014, «Interplay between functional connectivity and scale-free dynamics in intrinsic fmri networks», *Neuroimage*, vol. 95, p. 248–263. 11, 168

Ciuciu, P., G. Varoquaux, P. Abry, S. Sadaghiani et A. Kleinschmidt. 2012, «Scale-free and multifractal dynamic properties of fmri signals during rest and task», *Frontiers in Physiology*, vol. 3, n° 186. 13, 146, 183

Clausel, M., F. Roueff, M. S. Taqqu et C. Tudor. 2014, «Wavelet estimation of the long memory parameter for hermite polynomial of gaussian processes», *ESAIM: Probability and Statistics*, vol. 18, p. 42–76. 12

Clausel, M., F. Roueff, M. S. Taqqu et C. A. Tudor. 2012, «High order chaotic limits of wavelet scalograms under long-range dependence», *ALEA, Latin American Journal of Probability and Mathematical Statistics*, vol. 10, n° 2, p. 979–1011. 12

Clausel, M. et B. Vedel. 2011, «Explicit construction of operator scaling gaussian random fields», *Fractals*, vol. 19, n° 01, p. 101–111. 12, 185

Clausel, M. et B. Vedel. 2013, «Two optimality results about sample path properties of operator scaling gaussian random fields», *Annals of the University of Bucharest (mathematical series), Proceedings of the "XIème Colloque Franco-Roumain de Mathématiques Appliquées" 4 (LXII)*, p. 375–409. 12, 184

Cochran, W. G. 1952, «The $\chi 2$ test of goodness of fit», *The Annals of mathematical statistics*, p. 315–345. 71

Coeurjolly, J.-F., P.-O. Amblard et S. Achard. 2013, «Wavelet analysis of the multivariate fractional Brownian motion», *ESAIM: Probability and Statistics*, vol. 17, p. 592–604. 13, 184

Daoud, H. et M. A. Bayoumi. 2019, «Efficient epileptic seizure prediction based on deep learning», *IEEE Transactions on Biomedical Circuits and Systems*, vol. 13, n° 5, p. 804–813. 174

Daubechies, I. 1992, *Ten Lectures on Wavelets*, vol. 61, SIAM. 47, 74, 87, 129, 147, 170, 175

Davis, R. L., H. M. Hodges, G. F. Smoot, P. J. Steinhardt et M. S. Turner. 1992, «Cosmic microwave background probes models of inflation», *Physical Review Letters*, vol. 69, n° 13, p. 1856. 11

Decker, L. M., F. Cignetti et N. Stergiou. 2010, «Complexity and human gait», *Revista Andaluza de Medicina del Deporte*, vol. 3, n° 1, p. 2–12. 168

Didier, G., M. M. Meerschaert et V. Pipiras. 2018, «Domain and range symmetries of operator fractional brownian fields», *Stochastic Processes and their Applications*, vol. 128, n° 1, p. 39–78. 12, 185

Didier, G. et V. Pipiras. 2011, «Integral representations and properties of operator fractional brownian motions», *Bernoulli*, vol. 17, n° 1, p. 1–33. 12, 23, 24

Didier, G. et V. Pipiras. 2012, «Exponents, symmetry groups and classification of operator fractional Brownian motions», *Journal of Theoretical Probability*, vol. 25, p. 353–395. 12, 23

Domingues, O. D., P. Ciuciu, D. La Rocca, P. Abry et H. Wendt. 2019, «Multifractal analysis for cumulant-based epileptic seizure detection in eeg time series», dans *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, IEEE, p. 143–146. 15, 174

Doret, M., J. Spilka, V. Chudáček, P. Gonçalves et P. Abry. 2015, «Fractal analysis and hurst parameter for intrapartum fetal heart rate variability analysis: a versatile alternative to frequency bands and lf/hf ratio», *PLoS One*, vol. 10, n° 8, p. e0136 661. 168

Filippone, M., F. Camastra, F. Masulli et S. Rovetta. 2008, «A survey of kernel and spectral methods for clustering», *Pattern recognition*, vol. 41, n° 1, p. 176–190. 107

Flandrin, P. 1992, «Wavelet analysis and synthesis of fractional Brownian motion», *IEEE Transactions on information theory*, vol. 38, n° 2, p. 910–917. 12, 20, 22

Flandrin, P. et P. Abry. 1999, «Wavelets for scaling processes», dans *Fractals: theory and applications in engineering*, Springer, p. 47–64. 22

Frisch, U. 1995, *Turbulence, the Legacy of A.N. Kolmogorov*, Cambridge University Press. 11

Gadhoumi, K., J. Gotman et J.-M. Lina. 2015, «Scale invariance properties of intracerebral eeg improve seizure prediction in mesial temporal lobe epilepsy», *PLoS one*, vol. 10, n° 4, p. e0121 182. 15, 174, 175

Goldberger, A., L. Amaral, L. Glass, J. Hausdorff, P. Ivanov, R. Mark, J. Mietus, G. Moody, C.-K. Peng et H. Stanley. 2000, «Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals», *Circulation*, vol. 101, n° 23, p. e215–e220. 169, 174

Hall, P. et M. York. 2001, «On the calibration of silverman's test for multimodality», *Statistica Sinica*, p. 515–536. 159

Hartigan, J. A. et P. M. Hartigan. 1985, «The dip test of unimodality», *The annals of Statistics*, p. 70–84. 152, 196

Hartigan, P. 1985, «Algorithm as 217: Computation of the dip statistic to test for unimodality», *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 34, n° 3, p. 320–325. 152, 196

He, B. J. 2011, «Scale-free properties of the functional magnetic resonance imaging signal during rest and task», *Journal of Neuroscience*, vol. 31, n° 39, p. 13 786–13 795. 168

He, B. J. 2014, «Scale-free brain activity: past, present, and future», *Trends in cognitive sciences*, vol. 18, n° 9, p. 480–487. 11

Helgason, H., V. Pipiras et P. Abry. 2011, «Fast and exact synthesis of stationary multivariate gaussian time series using circulant embedding», *Signal Processing*, vol. 91, n° 5, p. 1123 – 1133. 45

Ichimaru, Y. et G. Moody. 1999, «Development of the polysomnographic database on cd-rom», *Psychiatry and clinical neurosciences*, vol. 53, n° 2, p. 175–177. 169

Ivanov, P. 2007, «Scale-invariant aspects of cardiac dynamics across sleep stages and circadian phases», *IEEE Engineering in Medicine and Biology Magazine*, vol. 26, n° 6, p. 33–37. 11

Ivanov, P. C., L. A. N. Amaral, A. L. Goldberger, S. Havlin, M. Rosenblum, Z. R. Struzik et H. E. Stanley. 1999, «Multifractality in human heartbeat dynamics», *Nature*, vol. 399, n° 6735, p. 461–465. 168

Kiyono, K., Z. R. Struzik, N. Aoyagi et Y. Yamamoto. 2006, «Multiscale probability density function analysis: non-Gaussian and scale-invariant fluctuations of healthy human heart rate», *IEEE Transactions on Biomedical Engineering*, vol. 53, n° 1, p. 95–102. 11

La Rocca, D., N. Zilber, P. Abry, V. van Wassenhove et P. Ciuciu. 2018, «Self-similarity and multifractality in human brain activity: A wavelet-based analysis of scale-free brain dynamics», *Journal of neuroscience methods*, vol. 309, p. 175–187. 168

Lahiri, S. N. 2003, *Resampling Methods for Dependent Data*, Springer, New York. 70

Lehmann, E. L., J. P. Romano et G. Casella. 2005, *Testing statistical hypotheses*, vol. 3, Springer. 71

Leon, L., H. Wendt, J.-Y. Tourneret et P. Abry. 2022, «A bayesian framework for multivariate multifractal analysis», *IEEE Transactions on Signal Processing*, vol. 70, p. 3663–3675. 168, 169

Mallat, S. 2008, *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*, 3e éd., Academic Press, ISBN 0123743702, 9780123743701. 21

Mandelbrot, B. B. 1975, «On the geometry of homogeneous turbulence, with stress on the fractal dimension of the iso-surfaces of scalars», *Journal of Fluid Mechanics*, vol. 72, n° 3, p. 401–416. 11

Mandelbrot, B. B. 1999, «A multifractal walk down wall street», *Scientific American*, vol. 280, n° 2, p. 70–73. 11

Mandelbrot, B. B. et J. W. V. Ness. 1968, «Fractional brownian motions, fractional noises and applications», *SIAM Review*, vol. 10, n° 4, p. pp. 422–437, ISSN 00361445. 12, 18

McIntosh, A. R. et B. Mišić. 2013, «Multivariate statistical analyses for neuroimaging data», *Annual review of psychology*, vol. 64, p. 499–525. 12

MohanBabu, G., S. Anupallavi et S. R. Ashokkumar. 2021, «An optimized deep learning network model for eeg based seizure classification using synchronization and functional connectivity measures», *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, n° 7, p. 7139–7151. 174

Mormann, F., K. Lehnertz, P. David et C. Elger. 2000, «Mean phase coherence as a measure for phase synchronization and its application to the eeg of epilepsy patients», *Physica D: Nonlinear Phenomena*, vol. 144, n° 3-4, p. 358–369. 174

Nakamura, T., K. Kiyono, H. Wendt, P. Abry et Y. Yamamoto. 2016, «Multiscale analysis of intensive longitudinal biomedical signals and its clinical applications», *Proceedings of the IEEE*, vol. 104, n° 2, p. 242–261. 168

Olsson, J., J. Niemczynowicz et R. Berndtsson. 1993, «Fractal analysis of high-resolution rainfall time series», *Journal of Geophysical Research: Atmospheres*, vol. 98, n° D12, p. 23 265–23 274. 11

Orejola, O., G. Didier, P. Abry et H. Wendt. 2022, «Hurst multimodality detection based on large wavelet random matrices», dans *2022 30th European Signal Processing Conference (EUSIPCO)*, IEEE, p. 2131–2135. 13, 146, 150, 152, 202

Palmer, M. W. 1988, «Fractal geometry: a tool for describing spatial patterns of plant communities», *Vegetatio*, vol. 75, p. 91–102. 11

Park, Y., L. Luo, K. K. Parhi et T. Netoff. 2011, «Seizure prediction with spectral power of eeg using cost-sensitive support vector machines», *Epilepsia*, vol. 52, n° 10, p. 1761–1770. 174

Peebles, P. J. 1989, «The fractal galaxy distribution», *Physica D: Nonlinear Phenomena*, vol. 38, n° 1-3, p. 273–278. 11

Pipiras, V. et M. S. Taqqu. 2017, *Long-Range Dependence and Self-Similarity*, *Cambridge Series on Statistical and Probabilistic Mathematics*, vol. 45, Cambridge University Press, Cambridge, United Kingdom. 12, 20

Roux, S. G., M. Clausel, B. Vedel, S. Jaffard et P. Abry. 2013, «Self-Similar Anisotropic Texture Analysis: The Hyperbolic Wavelet Transform Contribution», *IEEE Trans. on Image Proces.*, vol. 22, n° 11, p. 4353–4363. 12, 185

Saab, M. E. et J. Gotman. 2005, «A system to detect the onset of epileptic seizures in scalp eeg», *Clinical Neurophysiology*, vol. 116, n° 2, p. 427–442. 174

Samorodnitsky, G., M. S. Taqqu et R. Linde. 1996, «Stable non-gaussian random processes: stochastic models with infinite variance», *Bulletin of the London Mathematical Society*, vol. 28, n° 134, p. 554–555. 12, 18

Shoeb, A. H. 2009, *Application of machine learning to epileptic seizure onset detection and treatment*, thèse de doctorat, Massachusetts Institute of Technology. 174

Silverman, B. W. 1981, «Using kernel density estimates to investigate multimodality», *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 43, n° 1, p. 97–99. 157, 158, 159

Steinhaus, H. 1957, «On chords of convex curves», *Bulletin of the Polish Academy of Sciences III*, vol. 5, p. 595–597. 108, 162

Takahashi, M. 1989, «A fractal model of chromosomes and chromosomal dna replication», *Journal of theoretical biology*, vol. 141, n° 1, p. 117–136. 11

Tao, T. 2012, *Topics in Random Matrix Theory*, vol. 132, American Mathematical Society. 13, 37

Tao, T. et V. Vu. 2012, «Random covariance matrices: Universality of local statistics of eigenvalues», *Annals of Probability*, p. 1285–1315. 14

Taqqu, M. S. 2003, «Fractional brownian motion and long-range dependence», *Theory and applications of long-range dependence*, p. 5–38. 19

Taqqu, M. S. et V. Teverovsky. 1998, «On estimating the intensity of long-range dependence in finite and infinite variance time series», *A practical guide to heavy tails: statistical techniques and applications*, vol. 177, p. 218. 12

Tsagris, M., C. Beneki et H. Hassani. 2014, «On the folded normal distribution», *Mathematics*, vol. 2, n° 1, p. 12–28. 84, 196

Turcotte, D. 1990, «Implications of chaos, scale-invariance, and fractal statistics in geology», *Global and Planetary Change*, vol. 3, n° 3, p. 301–308. 11

Veitch, D. et P. Abry. 1999, «A wavelet-based joint estimator of the parameters of long-range dependence», *IEEE Transactions on Information Theory*, vol. 45, n° 3, p. 878–897. 12, 20, 23, 27, 29, 30, 38

Vinh, N. X., J. Epps et J. Bailey. 2010, «Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance», *The Journal of Machine Learning Research*, vol. 11, p. 2837–2854. 98, 199

Wang, F., H. Wang et R. Fu. 2018, «Real-time ecg-based detection of fatigue driving using sample entropy», *Entropy*, vol. 20, n° 3, p. 196. 169

Wendt, H., P. Abry et G. Didier. 2018, «Wavelet domain bootstrap for testing the equality of bivariate self-similarity exponents», dans *2018 IEEE Statistical Signal Processing Workshop (SSP)*, IEEE, p. 563–567. 13, 14, 15, 69

Wendt, H., P. Abry et G. Didier. 2019, «Bootstrap-based bias reduction for the estimation of the self-similarity exponents of multivariate time series», dans *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, p. 4988–4992. 13, 69

Wendt, H., P. Abry et S. Jaffard. 2007, «Bootstrap for empirical multifractal analysis», *IEEE Signal Processing Magazine*, vol. 24, n° 4, p. 38–48. 69, 70

Wendt, H., G. Didier, S. Combrexelle et P. Abry. 2017, «Multivariate hadamard self-similarity: testing fractal connectivity», *Physica D: Nonlinear Phenomena*, vol. 356, p. 1–36. 13, 14, 27, 28

Wilcoxon, F. 1945, «Individual comparisons by ranking methods», *Biometrics*, vol. 1, n° 6, p. 80–83. 172

Yamamoto, Y. et R. L. Hughson. 1991, «Coarse-graining spectral analysis: new method for studying heart rate variability», *Journal of Applied Physiology*, vol. 71, n° 3, p. 1143–1150. 168

Yao, J., S. Zheng et Z. Bai. 2015, *Sample Covariance Matrices and High-Dimensional Data Analysis*, Cambridge University Press. 13, 37

Yekutieli, D. et Y. Benjamini. 1999, «Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics», *Journal of Statistical Planning and Inference*, vol. 82, n° 1-2, p. 171–196. 85

Yu, J., S. Park, S. Lee et M. Jeon. 2018, «Driver drowsiness detection using condition-adaptive representation learning framework», *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, n° 11, p. 4206–4218. 169

Zhang, K. K., Z.and Parhi. 2015, «Low-complexity seizure prediction from ieeg/seeg using spectral power and ratios of spectral power», *IEEE Transactions on Biomedical Circuits and Systems*, vol. 10, n° 3, p. 693–706. 174

Zoubir, A. M. et D. R. Iskander. 2004, *Bootstrap Techniques for Signal Processing*, Cambridge University Press. 69

Résumé

L'invariance d'échelle constitue un paradigme polyvalent de traitement du signal et de l'image, apparaissant dans des champs d'applications nombreux et variés du monde réel et pouvant être formalisé par l'autosimilarité. Cependant, la plupart des études pratiques sont restées jusqu'à présent univariées, se limitant à analyser les différentes composantes d'un même jeu de données indépendamment. Pourtant, les applications les plus récentes impliquent souvent le recours à de nombreux capteurs pour superviser un même système, dont une étude pertinente nécessite une analyse multivariée des séries temporelles résultantes. Des modèles pour l'autosimilarité multivariée ont été récemment proposés, et une procédure d'estimation robuste pour le vecteur des exposants d'autosimilarité caractérisant un tel modèle à partir de représentations multi-échelles a été développée. Cette procédure souffre néanmoins d'un biais important que le présent travail vise, en premier lieu, à réduire en proposant une modification de celle-ci. Les performances d'estimation sont étudiées théoriquement dans les limites asymptotiques des échantillons de grande taille et empiriquement pour des échantillons de taille finie. Ces outils sont appliqués sur des données physiologiques pour réaliser différentes tâches : la détection de la somnolence et la prédiction de crises d'épilepsie. En second lieu, est traitée la question clé de compter le nombre de valeurs réellement distinctes parmi les exposants d'autosimilarité et le nombre d'exposants d'autosimilarité prenant chacune de ces valeurs. Sont ainsi proposées différentes procédures de tests d'égalité entre exposants d'autosimilarité à partir d'un schéma de ré-échantillonage bootstrap par blocs temps-échelle multivariés récemment développé. Enfin, pour tenir compte des ordres de grandeur des données du monde réel, le cadre de la grande dimension, où le nombre de séries temporelles croît avec leur taille, est également abordé.

Abstract

Scale invariance is a versatile signal processing paradigm that appears in many varied real-world applications and can be formalised by self-similarity. However, most practical studies have so far remained univariate, limiting themselves to analyse the different components of the same data set independently. Yet, the most recent applications often involve the use of many sensors to monitor the same system, for which a relevant study requires a multivariate analysis of the resulting time series. Multivariate self-similarity models have recently been proposed, and a robust estimation procedure for the self-similarity exponent vector characterising such a model from multi-scale representations has been developed. However, this procedure suffers from a significant bias which the present work aims, firstly, to reduce by proposing a modification of this procedure. The estimation performance is studied theoretically in the asymptotic limits of large sample sizes and empirically for finite sample sizes. These tools are applied to physiological data to perform different tasks: drowsiness detection and seizure prediction. Secondly, the key issue of counting the number of truly distinct values among the self-similarity exponents and the number of self-similarity exponents taking each of these values is addressed. Different equality testing procedures are thus proposed based on a recently developed multivariate time-scale block-bootstrap resampling scheme. Finally, to account for the orders of magnitude of real-world data, the high-dimensional asymptotic framework where the number of time series increases with their size is also addressed.