Identifying high-dimensional self-similarity based on spectral clustering applied to large wavelet random matrices

Oliver Orejola, Gustavo Didier, Herwig Wendt, Patrice Abry

Identifying high-dimensional self-similarity based on spectral clustering applied to large wavelet random matrices

Oliver Orejola, Gustavo Didier
Mathematics Department,
Tulane University, New Orleans, USA
{oorejola,gdidier}@tulane.edu

Herwig Wendt
Université de Toulouse
CNRS, IRIT, Toulouse (FR)
herwig.wendt@irit.fr

Patrice Abry
Université de Lyon, ENS de Lyon
CNRS, Lab. de Physique, Lyon (FR)
patrice.abry@ens-lyon.fr

***Abstract*—In the modern world, systems are routinely monitored by multiple sensors. This generates "Big Data" in the form of large collections of time series. On the other hand, scale-invariant (fractal) systems are of great interest in signal processing since they naturally emerge in several fields of application. In this work, we put forward an algorithm for the statistical identification of high-dimensional self-similarity. In the threefold limit as dimension, sample size and scale go to infinity, the method builds upon spectral clustering applied to large wavelet random matrices to consistently estimate the Hurst modes (exponents) and, hence, characterize high-dimensional self-similarity. Monte Carlo simulations show that the proposed methodology is efficient for realistic sample sizes.**

I. Introduction

Context: scale invariance. *Scale invariance* has been observed in signals obtained from a wide variety of contexts in physics and engineering [1], [2]. Unlike with many classical statistical mechanical systems, a signal X is called scale-invariant, or *fractal*, when its temporal dynamics lack a characteristic scale. In this case, modeling revolves around *scaling exponents*, since they characterize how the behavior of the system is related across a continuum of scales. A cornerstone model of scale invariance is fractional Brownian motion (fBm) [3]. FBm is the only Gaussian, stationary-increment, self-similar process. The latter property means that its finite-dimensional distributions (f.d.d.) are scale-invariant, i.e., $\{B_H(t)\}_{t\in\mathbb{R}} \overset{\text{f.d.d.}}{=} \{a^H B_H(t/a)\}_{t\in\mathbb{R}}$, $a > 0$, where the scaling exponent $0 < H < 1$ is called the *Hurst* parameter. The estimation of H plays a key role in signal processing tasks such as characterization, diagnosis, classification and detection. The so-named *wavelet transform* provides the analytical basis for well-established estimation methodologies for H [4].

Challenge: high-dimensional time series. The modeling of self-similarity in applications has remained so far based on the univariate fBm model. Yet, in the modern world of "Big Data," a plethora of sensors monitor natural and artificial systems, generating large data sets in the form of several joint time series. In neuroscience, for example, the number of macroscopic brain activity time series ranges from hundreds (MEG data) to several tens of thousands (fMRI data) [2]. In climate studies, dealing with large numbers of measured components has become standard [5]. For such high-dimensional data, a multitude of scaling laws – i.e., *Hurst multimodality* – implies distinct large-scale behavior along possibly non-canonical coordinate axes. Ignoring Hurst multimodality in data may also lead to arbitrarily large *estimation biases* [6], [7].

Related work: self-similarity in high dimensions. The recently proposed *operator fractional Brownian motion* (ofBm) [8]–[10] provides the fundamental scale-invariant model in a high-dimensional time series setting [6]. The use of ofBm in applications leads to a critical question: *how many different scaling laws exist in the possibly very large number of time series?* The so-named *wavelet eigenanalysis methodology* [6], [7] is founded in the behavior of the eigenvalues of *wavelet random matrices*. It was shown that the methodology provides efficient and robust estimation of Hurst exponents in both multivariate and high-dimensional settings – i.e., respectively, for a fixed number p of time series and as $p(n) \to \infty$ as the sample size n grows [7], [11], [12].

In the statistical signal processing literature, the problem of identifying the number and properties of sources in multivariate or high-dimensional noisy signals has been studied for decades [13]–[16]. Examples of the proposed techniques include principal component analysis, factor analysis and sparse graphical Gaussian models [17]. Nevertheless, there has been a paucity of estimation methodologies for both *high-dimensional* and *scale-invariant* signals. Efforts in this direction include [18], [19]. The study of high-dimensional limits is a key issue both in theory and in applications involving scale invariance. In fact, in multiscale analysis, the ratio between the sample size and the scale of analysis is inherently *not* large compared to the number of components *at coarse scales*. A key related difficulty is the study of random matrices under *temporal dependence* [20]–[22].

Spectral clustering and wavelet random matrices. Part of

G.D.'s long term visits to ENS de Lyon were supported by the school, the CNRS and the Simons Foundation collaboration grant #714014. H.W. supported by ANR-18-CE45-0007 MUTATION, France.

the motivation for this article is *finite-sample performance*. It is well-known that the graph Laplacian (matrix) bridges the gap between useful low-dimensional representations, such as graphs, and complex objects, such as manifolds [23]. In the context of wavelet random matrices, even though the logarithmic empirical spectral distribution (log-e.s.d.) ultimately converges to a measure $\pi(dH)$ [24], in modeling such convergence can be a delicate issue both in statistical and numerical senses. Nevertheless, the graph Laplacian-based *spectral clustering* method, from the Machine Learning literature [25], [26], holds promise as a way of tackling the issue. This is so because it can be used in converting the estimation of the measure $\pi(dH)$ into the quantifiable detection of a finite number of (Hurst) modes.

Goal, contributions and outline. The goal of this work is to put forth and study the properties of a spectral clustering-based algorithm for identifying high-dimensional fractality as characterized by the distribution of Hurst exponents. This is done in the context of the three-way limit as the *sample size*, the *number of components* and the *scale* go to infinity at a fixed rate. To this end, the definitions and properties of ofBm and the high-dimensional model are summarized in Sections II-A and II-B. Wavelet eigenanalysis-based estimation is briefly sketched in Section II-C. The proposed three-step algorithm is described in Section II-D. Section III contains the core contribution of this work. There we explain, first, how the incorporation of multiscale behavior calls for a modification of the high-dimensional limits used in random matrix theory. Second, we describe the high-dimensional behavior of the log-e.s.d. of wavelet random matrices. Third, we establish the consistency of the proposed algorithm.

Furthermore, these theoretical results are tested in practice by means of extensive Monte Carlo simulations based on synthetic ofBm sample paths (Section IV-A). First, we use simulations to visualize high-dimensional fractality and observe the clustering problem at hand (Section IV-B). Second, we test the proposed algorithm's practical performance (Section IV-C). The reported results demonstrate that the proposed algorithm has satisfactory performance over samples of the sizes typically encountered in neuroscience (cf. [2]) and can be readily applied to real-world high-dimensional data.

II. Self-similarity analysis and modeling

A. Operator fractional Brownian motion

Operator fractional Brownian motion (ofBm) is a canonical model for multidimensional scale-invariant structures in real-world data. We briefly recall its definition and some properties (see [9] for the general definition and properties of ofBm).

Let $\underline{B}_{\underline{H},\mathbf{\Sigma}}(t) = \big(B_{H_1}(t),\ldots,B_{H_p}(t)\big)_{t\in\mathbb{R}}$ denote a collection of p possibly correlated fBm components defined by their individual self-similarity exponents $\underline{H} = (H_1,\ldots,H_p)$, $0 < H_1 \leq \ldots \leq H_p < 1$. Let $\mathbf{\Sigma}$ be a pointwise covariance matrix with entries $(\mathbf{\Sigma})_{\ell,\ell'} = \sigma_\ell\sigma_{\ell'}\rho_{\ell,\ell'}$, where σ_ℓ^2 are the variances of the components and $\rho_{\ell,\ell'}$ their (pairwise) correlation coefficients. We define ofBm as the stochastic process $\underline{B}_{\mathbf{P},\underline{H},\mathbf{\Sigma}}(t) := \mathbf{P}\underline{B}_{\underline{H},\mathbf{\Sigma}}(t)$, where $\mathbf{P}$ is a real-valued, $p\times p$ invertible matrix that mixes the components (changes the scaling coordinates) of $\underline{B}_{\underline{H},\mathbf{\Sigma}}(t)$. OfBm consists of a multivariate Gaussian self-similar process with stationary increments. Moreover, it satisfies the (operator) self-similarity relation

$$\{\underline{B}_{\mathbf{P},\underline{H},\mathbf{\Sigma}}(t)\}_{t\in\mathbb{R}} \overset{\text{f.d.d.}}{=} \{a^{\mathbf{H}}\underline{B}_{\mathbf{P},\underline{H},\mathbf{\Sigma}}(t/a))\}_{t\in\mathbb{R}}, \qquad (1)$$

$\forall a > 0$. In (1), the matrix (Hurst) exponent is given by $\mathbf{H} = \mathbf{P}\text{diag}(\underline{H})\mathbf{P}^{-1}$, and $a^{\mathbf{H}} := \sum_{k=0}^{+\infty}\log^k(a)\mathbf{H}^k/k!$, where $\overset{\text{f.d.d.}}{=}$ stands for the equality of finite-dimensional distributions.

B. High-dimensional model

To model the complexity of a high-dimensional fractal system, and also to reflect the modellers' ignorance, we assume Hurst exponents trickle into the system randomly, based on some unknown distribution. So, let $\pi(dH)$ be a discrete distribution of Hurst exponents with ordered support $\{H_1,\ldots,H_r\}$, $r\in\mathbb{N}$. Given a vector $\underline{H}\in(0,1)^p$ of i.i.d. samples from $\pi(dH)$, the process

$$Y(t) := \underline{B}_{\mathbf{P},\underline{H},I}(t) \qquad (2)$$

as defined in Section II-A is, *conditionally on* $\underline{H}$, a p-variate ofBm with Hurst matrix $\mathbf{H}$. Given a time series $\{Y(t)\}_{t=1,\ldots,n}$, we further assume $p = p(n) \to \infty$ as $n\to\infty$, under which (2) is a *high-dimensional model* and has $p = p(n)\to\infty$ Hurst exponents (on models of the general form (2) under weak dependence, see, for instance, [27]).

C. Scaling in the wavelet domain

Multivariate wavelet transform. Let ψ be a mother wavelet, i.e., a real-valued function such that $\int_{\mathbb{R}}\psi^2(t)dt = 1$. For all $k,j\in\mathbb{Z}$, the multivariate DWT of $\{Y(t)\}_{t\in\mathbb{R}}$ is defined as $D_Y(2^j,k) := (D_{Y_1}(2^j,k),\ldots,D_{Y_p}(2^j,k))$, where $D_{Y_\ell}(2^j,k) := \langle 2^{-j/2}\psi(2^{-j}t-k)|Y_\ell(t)\rangle \in \mathbb{R}$ for $\ell\in\{1,\ldots,p\}$. For a detailed introduction to wavelet transforms, see [28]. It can be shown that the wavelet coefficients $\{D_Y(2^j,k)\}_{k\in\mathbb{Z}}$ of p-variate ofBm $Y = \underline{B}_{\mathbf{P},\underline{H},\mathbf{\Sigma}}$ satisfy, for every fixed octave j, the operator self-similarity relation $\{D_Y(2^j,k)\}_{k\in\mathbb{Z}} \overset{\text{f.d.d.}}{=} \{2^{j(\mathbf{H}+\frac{1}{2}I)}D_Y(1,k)\}_{k\in\mathbb{Z}}$ [6], [7].

Wavelet random matrices and high-dimensional eigenanalysis. Given *any* p-variate process $\{Y(t)\}_{t\in\mathbb{R}}$ (in particular, model (2)), the sample wavelet spectrum (variance) at octave $j = j_1,\ldots,j_2$ is given by the $p\times p$ wavelet random matrices

$$\mathbf{S}_Y(2^j) = \frac{1}{n_j}\sum_{k=1}^{n_j} D_Y(2^j,k)D_Y(2^j,k)^* \qquad (3)$$

where n is the time series (sample) size and $n_j \simeq n/2^j$ is the number of wavelet coefficients available at scale 2^j. It was shown in [6], [7] that, in general, estimation based on the entrywise multiscale behavior of $\mathbf{S}_Y(2^j)$ is arbitrarily biased and effectively meaningless. So, let $\lambda_1(2^j),\ldots,\lambda_p(2^j)$ be the *eigenvalues* of the random matrix $\mathbf{S}_Y(2^j)$ as in (3). Notably, it was further shown in [6], [7] that the vector of Hurst exponents $\underline{H}$ can be efficiently estimated by means of the *weighted* wavelet log-eigenvalues

$$\widetilde{H}_\ell = \Big(\sum_{j=j_1}^{j_2} w_j\log_2\lambda_\ell(2^j)\Big)\Big/2 - \frac{1}{2}, \quad \ell = 1,\ldots,p, \qquad (4)$$

where $\sum_j w_j = 0$ and $\sum_j j w_j = 1$. In fact, (4) has good statistical performance in both Gaussian and non-Gaussian frameworks [6], [7], [11], [12].

D. Algorithm for the estimation of Hurst modes

In this work, we propose an algorithm for Hurst distribution estimation that comprises **three steps**. So, let $\widetilde{\mathbf{H}} = \{\widetilde{H}_1, \ldots, \widetilde{H}_p\}$ (see (4)). For a fixed threshold $\varepsilon > 0$, we define $G_\varepsilon(\widetilde{\mathbf{H}}) = (V, E)$ to be the *graph induced by a ε–threshold* where $V = \widetilde{\mathbf{H}}$ and $E = \{e_{\widetilde{H}_i, \widetilde{H}_j} | \{|\widetilde{H}_i - \widetilde{H}_j|) < \varepsilon,\ i \neq j\}$. That is, $G_\varepsilon(\widetilde{\mathbf{H}})$ is the graph obtained by connecting points $\widetilde{H}_i$ and $\widetilde{H}_j$ that lie within a distance ε of one another. In particular, $G_\varepsilon(\widetilde{\mathbf{H}})$ has *graph Laplacian*

$$L_\varepsilon(\widetilde{\mathbf{H}}) := D_\varepsilon(\widetilde{\mathbf{H}}) - A_\varepsilon(\widetilde{\mathbf{H}}). \tag{5}$$

In (5), $A_\varepsilon(\widetilde{\mathbf{H}}) := \left[\mathbb{1}_{\{|\widetilde{H}_i - \widetilde{H}_j| < \varepsilon,\ i \neq j\}}\right]_{1 \leq i \leq j \leq p}$ is the *adjacency matrix*. Also, the *degree matrix* $D_\varepsilon(\widetilde{\mathbf{H}})$ is given by a diagonal matrix with entries $D_\varepsilon(\widetilde{\mathbf{H}})_{ii} := \sum_{j=1}^{p} A_\varepsilon(\widetilde{\mathbf{H}})_{ij}$, $i = 1, \ldots, p$. **Step 1** of the algorithm, depicted next as pseudocode, uses (5) to construct clusters of estimated Hurst exponents $\widetilde{H}_1, \ldots, \widetilde{H}_p$.

Step 1: spectral clustering
Input: $\varepsilon > 0$ and $\widetilde{\mathbf{H}} = \{\widetilde{H}_1, ..., \widetilde{H}_p\}$.
• Compute the eigenvalues, $\{\theta_\ell\}_{1 \leq \ell \leq p}$, and corresponding eigenvectors $\{\mathbf{u}_\ell\}_{1 \leq \ell \leq p}$ of $L_\varepsilon(\widetilde{\mathbf{H}})$ (5)
• Let $\widehat{r} = |\{\theta_\ell = 0 | 1 \leq \ell \leq p\}|$
• Let $\mathbf{U} \in \mathbb{R}^{p \times \widehat{r}}$ be a matrix with columns $\mathbf{u}_1, ..., \mathbf{u}_{\widehat{r}}$
• For $i = 1, ..., p$, let $\mathbf{y}_i \in \mathbb{R}^{\widehat{r}}$ be the vector corresponding to the $i-$th row of $\mathbf{U}$
• Use the $k-$means algorithm (e.g., [29]) to organize the points $\{\mathbf{y}_i\}_{1 \leq i \leq p}$ into clusters $C'_1, ..., C'_{\widehat{r}}$
Output: clusters $C_1, ..., C_{\widehat{r}}$ with $C_j = \{\widetilde{H}_i | \mathbf{y}_i \in C'_j\}$

Step 2 of the algorithm, shown next in the form of pseudocode, constructs an estimate of the Hurst distribution by taking averages over clusters.

Step 2: Hurst distribution estimation
Input: $\varepsilon > 0$, clusters $C_1, ..., C_{\widehat{r}}$ from **Step 1**
• For each C_i let $\widehat{H}_i = \text{mean}(C_i)$ and $\pi(\widehat{H}_i) = |C_i|/p$
Output: estimates $\widehat{H}_1, ..., \widehat{H}_{\widehat{r}}$ and $\pi(\widehat{H}_1), ..., \pi(\widehat{H}_{\widehat{r}})$

Note that **Steps 1** and **2** depend on the initial choice of threshold $\varepsilon > 0$. In **Step 3**, this choice is converted into a *model selection* procedure. Given the clusters $\mathcal{C}_\varepsilon = \{C_1, ..., C_{\widehat{r}}\}$ obtained in **Step 2**, the criterion for model selection is given by the so-called *intra-cluster standard deviation* $\text{ICSD}(\mathcal{C}_\varepsilon)$. More precisely, for $\min_i |C_i| > 1$,

$$\text{ICSD}(\mathcal{C}_\varepsilon) = \sum_{i=1}^{\widehat{r}} \Big(\frac{1}{|C_i| - 1} \sum_{\widetilde{H}_\ell \in C_i} (\widetilde{H}_\ell - \widehat{H}_i)^2 \Big)^{1/2}. \tag{6}$$

Step 3 is described next in the form of pseudocode.

Step 3: model selection
Input: $M > 0$ and $m \in \mathbb{N}$.
• For $k = 1, ..., m$ and $\varepsilon_k := kM/m$, set $\varepsilon = \varepsilon_k$ and follow **Steps 1** and **2** to obtain clusters $\mathcal{C}_{\varepsilon_k} = \{C_{1,k}, ..., C_{r_k,k}\}$ and $\text{ICSD}(\mathcal{C}_{\varepsilon_k})$
Output: $\varepsilon_{k_*(n)}$ with the smallest corresponding $\text{ICSD}(\mathcal{C}_{\varepsilon_k})$

III. High-dimensional limits

A. High dimensions: wavelet log-e.s.d. in the three-way limit

This work involves high-dimensional limits. To be more precise, recall that in a classical statistical setting the sample size goes to infinity whereas the number of components remains fixed, i.e., $n \to \infty$ and $p = p_0$. By contrast, in high-dimensional settings one often considers the *two-way* limit $\lim_{n \to \infty} p(n)/n = c \in [0, \infty)$, e.g., for sample covariance matrices [30].

On the other hand, modeling scale-invariance by means of (4) further calls for the scaling limit $(j_1, j_2) \to \infty$. Hence, overall the analysis of high-dimensional fractal behavior by means of large wavelet random matrices involves considering the nonstandard, *three-way* limit

$$n, p, j \to \infty,\ \frac{p}{n/2^j} \to c \in [0, \infty) \tag{7}$$

(see [7], [24], [31]). Providing results in the framework of (7) is a non-trivial contribution of this paper, both theoretically and numerically.

B. High-dimensional fractality: the Hurst exponents distribution

Mathematical results in [7], [24], [31] imply that, in the three-way limit (7) and under (2), the eigenvalues of the sample wavelet spectrum behave as

$$\frac{\log_2 \lambda_\ell(2^j)}{j} = 2H_\ell + 1 + o_{\mathbb{P}}(1), \tag{8}$$

$\ell = 1, \ldots, p$, where $o_{\mathbb{P}}(1)$ vanishes in probability. In fact, eigenvalue scaling relations of the type (8) are shown to hold for several high-dimensional instances, such as in the case of the wavelet eigenvalue bulk behavior for (2) (see [24]) as well as for non-Gaussian signal-plus-noise models (see [7], [31]).

Recall the fundamental properties of the graph Laplacian L, namely: $(\mathbf{GL}_1)$ if the underlying graph G is a disjoint union of simple graphs $G_1, \ldots, G_\kappa$, then $\dim \text{Ker}(G) = \kappa$; $(\mathbf{GL}_2)$ the number of non-zero entries of each eigenvector $\mathbf{u}_\ell$ in $\text{Ker}(L)$ is equal to the number of vertices in G_ℓ.

So, assume measurements are given by the high-dimensional model (2). Fix $\varepsilon > 0$. Because of (8) as well as of $(\mathbf{GL}_1)$ and $(\mathbf{GL}_2)$, **Steps 1** and **2** of the algorithm generate clusters of wavelet log-eigenvalues that, with probability going to 1, eventually decouple and coalesce around each Hurst mode $H_1, \ldots, H_r$. Moreover, based on **Step 3**, choosing $\varepsilon > 0$ by means of minimizing $\text{ICSD}(\mathcal{C}_\varepsilon)$ enforces near-optimal clustering over finite n. This way, we obtain the key result of this paper [32].

Fundamental property of the Hurst distribution estimation algorithm: *The dimension $\widehat{r}$ of the kernel of the graph Laplacian converges in probability to the number r of Hurst modes. Moreover, for each eigenvector $\mathbf{u}_1, \ldots, \mathbf{u}_{\widehat{r}}$ in the kernel of $L_{\varepsilon_{k_*(n)}}(\widetilde{\mathbf{H}})$, the proportion of non-zero entries of*

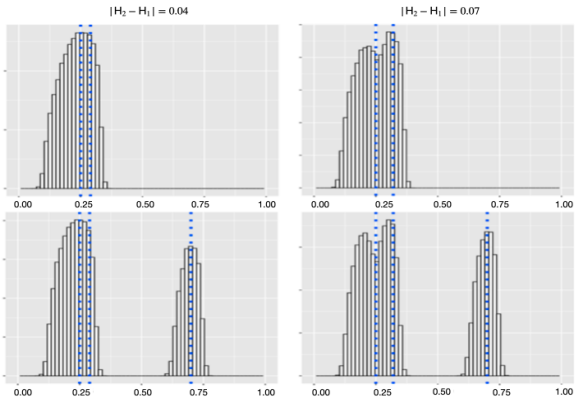

Fig. 1. **Distribution of weighted wavelet log-eigenvalues $\tilde{H}_q$ in the three-way limit $\lim_{n\to\infty} p\ 2^j/n$.** The measurements are given by a $p-$variate ofBm *conditionally on* the Hurst matrix $\mathbf{H} = \mathbf{P}\mathrm{diag}(H_1, \ldots, H_p)\mathbf{P}^{-1} \in \mathbb{R}^{p^2}$. **(Top row, bimodal)** Left and right plots, respectively: $H_1, \ldots, H_p$ are sampled from $\pi(dH)$ supported on $\{0.25, 0.29\}$ and $\{0.25, 0.32\}$. For a small difference in Hurst modes, $\Delta = 0.04$, although bimodal, visually the distribution appears unimodal (left). However for a sufficiently large difference in Hurst modes, $\Delta = 0.07$, a bimodal distribution appears (right). **(Bottom row, trimodal)** Left and right plots, respectively: $H_1, \ldots, H_p$ are sampled from $\pi(dH)$ supported on $\{0.25, 0.29, 0.7\}$ and $\{0.25, 0.32, 0.7\}$. For small minimum difference in Hurst modes, $H_2 - H_1 = 0.03$, although trimodal, visually the distribution appears bimodal (left). However for a sufficiently large difference in Hurst modes, $H_2 - H_1 = 0.07$, a trimodal distribution appears (right).

$\mathbf{u}_i$, $i = 1, \ldots, r$, *converges in probability to* $\pi(H_i)$, *and also* $\widehat{H}_i \xrightarrow{\mathbb{P}} H_i$.

This key theoretical result is further inspected by means of simulations in Section IV-C and is illustrated in Fig. 2.

IV. Finite-sample performance assessment

A. Monte Carlo simulation setting

To assess the practical relevance of the theoretical results stated in Section III, we make use of Monte Carlo simulations based on 1000 independent realizations of p-variate measurements $\underline{B}_{\mathbf{P},\underline{H},I} = \mathbf{P}\underline{B}_{I,\underline{H},I}$, where $\mathbf{P}$ is a randomly chosen orthogonal matrix and $\{\underline{B}_{I,\underline{H},I}(t)\}_{t\in\mathbb{R}}$ made up of independent univariate fBms with $n = 2^{14}$ (such sample sizes are realistic, for example, in the context of the analysis of infraslow brain activity [2]). The univariate fBms are generated using the **R** package `somebm` (see [33]). The p-dimensional vector $\underline{H}$ is obtained by drawing p i.i.d. samples from $\mathrm{Unif}(H_1, H_2)$, for each realization independently. Here, we limit ourselves to the consideration of $H_1 = 0.25$, and $H_2 = 0.25 + \Delta$, with $\Delta \in [0.0, 0.10]$. The wavelet transformation is generated by means of Mallat's algorithm based on a Daubechies filter with $N_\psi = 2$. In our simulations, we set the dimension to $p = 2^6$. The range of regression scales is chosen to be $j_1 = 2$ to $j_2 = 5$. Hence, $p/n_j < 1$ for $j = j_1, \ldots, j_2$, implying that all the wavelet random matrices used in the regression procedure have full rank. In **Step 3**, $m = 10$ and M is heuristically picked to be $M := \log_2\left(\frac{\lambda_p(2^1)}{\lambda_1(2^1)}\right)\frac{1}{j_2}$. From [24] it follows that such an M is a reasonable upper bound for the *width* of the wavelet log-e.s.d.

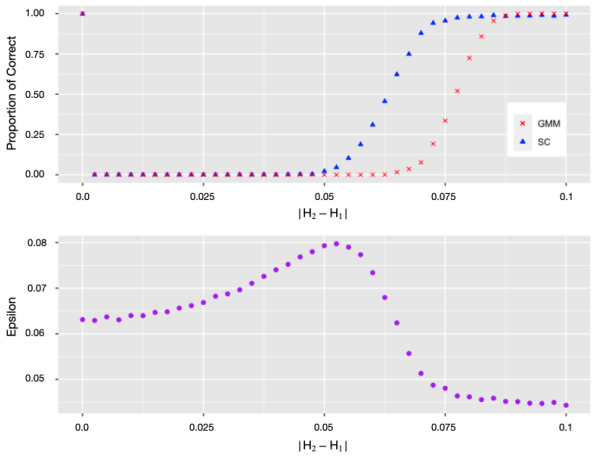

Fig. 2. **Algorithm performance analysis (top plot).** Proportion of correct identification of the number of Hurst modes over 1000 Monte Carlo runs. (Blue) The proposed method (Red) Gaussian mixture model-based clustering. **Optimal $\varepsilon > 0$ chosen via model selection (bottom plot).** The average ε computed over 1000 Monte Carlo simulations chosen by means of model selection that provides the smallest ICSD.

B. High-dimensional fractality in practice

To illustrate the practical relevance of the property (8) and the clustering problem at hand, Fig. 1 provides an example of the high-dimensional behavior of the wavelet log-e.s.d. In both plots, top row, the true distribution of Hurst exponents is bimodal. However, for small Δ ($\Delta = 0.04$), the wavelet log-eigenvalue distribution visually appears unimodal. This contrasts with large Δ ($\Delta = 0.07$). There, a strikingly bimodal distribution emerges in the wavelet eigenspectrum (right). An analogous phenomenon is observed for trimodal instances (Fig. 1, bottom row).

C. Estimation performance

To evaluate the performance of the proposed algorithm, we test it for a variety of bimodal Hurst distributions $\pi(dH)$. For each distribution, we compute the accuracy of spectral clustering by means of verifying if $\widetilde{r} = r := |\mathrm{supp}\ \pi(dH)|$. We compare the proposed spectral clustering-based method with the so-named Gaussian mixture mode-based clustering (GMM) [34]. The **R** package `mclust` [35] is used to implement GMM-based clustering. GMM is based on parameterized finite univariate Gaussian mixture models. Models are estimated by EM algorithm (Expectation-Maximization) initialized by hierarchical model-based agglomerative clustering. The optimal model is then selected according to BIC.

For $\pi(dH) = \mathrm{Unif}(H_1, H_2)$, the results are reported in Fig. 2. They show that the proposed method begins to correctly identify two modes near $\Delta = 0.05$, and consistently identifies two modes when $\Delta \geq 0.075$. This stands in contrast to GMM's capabilites. That is, GMM begins to correctly identify two modes near $\Delta = 0.07$, and consistently identifies two modes when $\Delta \geq 0.085$.

The choice of threshold $\varepsilon > 0$, following **Step 3**, is reported in Fig. 2. The results are consistent with Fig. 1. In fact, for $\Delta < 0.05$ the proposed algorithm consistently fails to correctly identify r. Accordingly, from Fig. 1 (top left) the wavelet log-e.s.d. is visually *uni*modal (i.e., unimodality is the "best" model). Hence, a larger ε is necessary to capture this unimodality. By contrast, for $\Delta > 0.7$, the proposed method identifies r correctly as the wavelet log-e.s.d. appears appropriately bimodal (Fig. 1, right). Accordingly, the model selection procedure (**Step 3**) picks a smaller ε.

Overall, these results confirm that the proposed algorithm is operational, has satisfactory statistical performance and can be readily applied in the study of high-dimensional fractality.

V. Conclusions and perspectives

In this paper, we build upon wavelet random matrices and the spectral clustering method to construct a three-step algorithm for the identification of high-dimensional fractal systems. In the threefold limit as dimension, sample size and scale go to infinity, the method consistently estimates the Hurst modes. In addition, Monte Carlo simulations for realistic sample sizes demonstrate that the proposed methodology displays efficient and robust finite-sample performance in the estimation of the Hurst distribution.

Real data modeling calls for the investigation of the more general case where the components of $Y(t)$ in (2) are not conditionally Gaussian, or are perturbations thereof. This requires further mathematical efforts beyond the scope of this article. Future work also includes applications in the analysis of neuroscientific data, following up on studies such as [36].

References

[1] D. Sornette, *Critical Phenomena in Natural Sciences: Chaos, Fractals, Selforganization and Disorder: Concepts and Tools*, Springer Science & Business Media, 2006.

[2] D. La Rocca, H. Wendt, V. van Wassenhove, P. Ciuciu, and P. Abry, "Revisiting functional connectivity for infraslow scale-free brain dynamics using complex wavelets," *Frontiers in Physiology*, vol. 11, pp. 1651, 2021.

[3] P. Flandrin, "Wavelet analysis and synthesis of fractional Brownian motion," *IEEE Trans. Info. Theory*, vol. 38, pp. 910 – 917, March 1992.

[4] D. Veitch and P. Abry, "A wavelet-based joint estimator of the parameters of long-range dependence," *IEEE Trans. Info. Theory*, vol. 45, no. 3, pp. 878–897, 1999.

[5] F. A. Isotta, C. Frei, V. Weilguni, M. Perčec Tadić, P. Lassegues, B. Rudolf, V. Pavan, C. Cacciamani, G. Antolini, S. M. Ratto, and M. Munari, "The climate of daily precipitation in the Alps: development and analysis of a high-resolution grid dataset from pan-Alpine rain-gauge data," *Int. J. Climatol.*, vol. 34, no. 5, pp. 1657–1675, 2014.

[6] P. Abry and G. Didier, "Wavelet estimation for operator fractional Brownian motion," *Bernoulli*, vol. 24, no. 2, pp. 895–928, 2018.

[7] P. Abry and G. Didier, "Wavelet eigenvalue regression for n-variate operator fractional Brownian motion," *J. Multivar. Anal.*, vol. 168, pp. 75–104, November 2018.

[8] J. D. Mason and Y. Xiao, "Sample path properties of operator-self-similar Gaussian random fields," *Theory Probab. Appl.*, vol. 46, no. 1, pp. 58–78, 2002.

[9] G. Didier and V. Pipiras, "Integral representations and properties of operator fractional Brownian motions," *Bernoulli*, vol. 17, no. 1, pp. 1–33, 2011.

[10] G. Didier and V. Pipiras, "Exponents, symmetry groups and classification of operator fractional Brownian motions," *J. Theor. Probab.*, vol. 25, pp. 353–395, 2012.

[11] P. Abry, H. Wendt, and G. Didier, "Detecting and estimating multivariate self-similar sources in high-dimensional mixtures," in *IEEE Stat. Signal Process. Workshop*, 2019, pp. 1–5.

[12] B.C. Boniece, H. Wendt, G. Didier, and P. Abry, "On multivariate non-Gaussian scale invariance: fractional Lévy processes and wavelet estimation," in *Proc. Eur. Signal. Process. Conf. (EUSIPCO)*, 2019, pp. 1–5.

[13] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, Academic Press, 2010.

[14] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, 2004.

[15] A. P. Liavas and P. A. Regalia, "On the behavior of information theoretic criteria for model order selection," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1689–1695, 2001.

[16] J. P. C. L. Da Costa, A. Thakre, F. Roemer, and M. Haardt, "Comparison of model order selection techniques for high-resolution parameter estimation algorithms," in *Proc. 54th Intern. Scient. Colloq. (IWK'09)*, 2009.

[17] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, vol. 4, Prentice-Hall, New Jersey, 2014.

[18] R. Zhang, P. Robinson, and Q. Yao, "Identifying cointegration by eigenanalysis," *J. Am. Stat. Assoc.*, vol. 114, no. 526, pp. 916–927, 2018.

[19] O. Orejola, G. Didier, P. Abry, and H. Wendt, "Hurst multimodality detection based on large wavelet random matrices," in *Proc. Eur. Signal. Process. Conf. (EUSIPCO)*, 2022, pp. 2131–2135.

[20] H. Liu, A. Aue, and D. Paul, "On the Marčenko–Pastur law for linear time series," *Ann. Statist.*, vol. 43, no. 2, pp. 675–712, 2015.

[21] L. Erdős, T. Krüger, and D. Schröder, "Random matrices with slow correlation decay," in *Forum of Math. Sigma*. Cambridge Univ. Press, 2019, vol. 7.

[22] F. Merlevède and M. Peligrad, "On the empirical spectral distribution for matrices with long memory and independent rows," *Stochastic Process. Appl.*, vol. 126, no. 9, pp. 2734–2760, 2016.

[23] Y.-H. He and S.-T. Yau, "Graph Laplacians, Riemannian manifolds, and their machine-learning," *Math. Comp. and Geom. of Data*, vol. 2, no. 1, pp. 1–48, 2022.

[24] P. Abry, G. Didier, O. Orejola, and H. Wendt, "On the empirical spectral distribution of large wavelet random matrices based on mixed-Gaussian fractional measurements in moderately high dimensions," `https://arxiv.org/pdf/2401.02815.pdf`, pp. 1–52, 2024.

[25] R. Kannan, S. Vempala, and A. Vetta, "On clusterings: good, bad and spectral," *Journal of the ACM*, vol. 51, no. 3, pp. 497–515, 2004.

[26] M. Filippone, F. Camastra, F. Masulli, and S. Rovetta, "A survey of kernel and spectral methods for clustering," *Pattern Recogn.*, vol. 41, no. 1, pp. 176–190, 2008.

[27] J. Chang, B. Guo, and Q. Yao, "Principal component analysis for second-order stationary vector time series," *Ann. Statist.*, vol. 46, no. 5, pp. 2094–2124, 2018.

[28] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, San Diego, CA, 1998.

[29] I. S. Dhillon, Y. Guan, and B. Kulis, "Kernel k-means: spectral clustering and normalized cuts," in *Proc. 10th ACM SIGKDD Intern. Conf. on Knowledge Disc. and Data Mining*, 2004, pp. 551–556.

[30] Z. Bai and J. Silverstein, *Spectral Analysis of Large Dimensional Random Matrices*, Springer, 2010.

[31] P. Abry, B. C. Boniece, G. Didier, and H. Wendt, "On high-dimensional wavelet eigenanalysis," *arXiv preprint arXiv:2102.05761*, 2021.

[32] Oliver Orejola, *Essays on random matrix theory and applications*, Ph.D. thesis, New Orleans, LA, May 2024.

[33] J. Huang, *somebm: some Brownian motions simulation functions*, 2013, R package version 0.1.

[34] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *J. Amer. Statist. Assoc.*, vol. 97, no. 458, pp. 611–631, 2002.

[35] L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery, *mclust 5: clustering, classification and density estimation using Gaussian finite mixture models*, 2016.

[36] C.-G. Lucas, P. Abry, H. Wendt, and G. Didier, "Epileptic seizure prediction from eigen-wavelet multivariate self-similarity analysis of multi-channel EEG signals," in *Proc. Eur. Signal. Process. Conf. (EUSIPCO)*, 2023, pp. 970–974.