COUNTING DISTINCT MULTIVARIATE SELF-SIMILARITY PARAMETERS USING A BOOTSTRAP-DRIVEN GRAPH CLUSTERING APPROACH

Charles-Gérard Lucas[1]*, Herwig Wendt*[2]*, Patrice Abry*[3]*, Gustavo Didier*[4]

[1]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
[2]CNRS, IRIT, University of Toulouse, Toulouse, France
[3]CNRS, ENS de Lyon, Laboratoire de Physique, Lyon, France
[4]Math. Dept., Tulane University, New Orleans, LA, USA

charles-gerard.lucas@inria.fr, hwendt@n7.fr, patrice.abry@ens-lyon.fr, gdidier@tulane.edu

ABSTRACT

In various modern fields, the multiplicity of sensors in applications may result in potentially numerous scale-free time series that jointly characterize one same system. Multivariate self-similarity analysis tackles the challenge of studying these systems by providing as many self-similarity parameter estimates as available time series. The possibly large amount of self-similarity parameters raises the major issue of identifying the number of actually distinct self-similarity parameters. The present work attains this goal by designing an adapted graph to perform a spectral clustering-type procedure. The proposed graph is weighted using pairwise equality test p-values estimated by a multivariate time-scale block-bootstrap scheme combined with wavelet random matrix eigenanalysis for self-similarity parameter estimation. Numerical experiments on synthetic multivariate data show a very satisfactory performance of the clustering strategy.

***Index Terms*—** Multivariate self-similarity, scaling exponents, multivariate wavelet transform, bootstrap, spectral clustering

1. INTRODUCTION

Context. Scale-free time series do not possess a characteristic scale. Rather, they are characterized by mechanisms relating a large continuum of scales. These mechanisms are controlled by the scaling exponents, and their estimation is the core objective of scale-free analysis. Scale-free temporal dynamics are typical in a wide variety of signals, including Internet traffic, finance, geography, physiological signals, or turbulence, see e.g., [1–6]. Self-similarity is a specific scale-free model in which a single scaling exponent, the Hurst exponent H, controls the temporal dynamics of the time series, and has been extensively used for univariate data modeling. Recently, an extension for multivariate data was proposed, i.e., for M time series collected simultaneously by several sensors. Multivariate self-similarity models scale-free data by using as many Hurst exponents as data components dynamics [7, 8]. However, the number of *distinct* Hurst exponents (sources) may be smaller than the number M of time series. This raises the issues of estimating this number, and also of identifying the exponents that are identical (clustering).

Related work. Operator fractional Brownian motion (ofBm) is a robust and widely used model of multivariate self-similarity. It is a multivariate Gaussian self-similar process with stationary increments and is characterized by the self-similarity parameter vector $\underline{H}$ [7–15]. The wavelet characterization of ofBm leads to an asymptotically jointly Gaussian estimator for $\underline{H}$ [7,8,14,16]. Jointly Gaussian statistics make it straightforward to test the equality of pairs self-similarity parameters H_m, or to test whether all self-similarity parameters are equal, i.e., $H_1 = \ldots = H_M$. In [15, 17], a wavelet-domain block-bootstrap procedure was developed to approximate the unknown covariance of $\underline{H}$ for these test formulations from a single finite M-variate time series. Yet, this approach does not allow for grouping equal self-similarity parameters because the $M(M-1)/2$ pairwise tests possibly yield contradictory decisions. An attempt to address this issue was proposed in [18,19]. It consisted in testing the equality of the $M-1$ pairs of consecutive self-similarity parameters $H_m \leq H_{m+1}, m = 1, \ldots, M-1$. However, already for $M \geq 3$ the ordering operation induces significantly more complex test statistic distributions. This results in poor control of the confidence level and in low test power.

Goals and contributions. To address these issues, the present work proposes a procedure for clustering equal values in $\underline{H}$ starting from a single finite-size multivariate time series. To that end, Section 2 recalls the definition and properties of the ofBm model and details eigen-wavelet-analysis-based self-similarity parameter estimation. Building on this, an original graph clustering procedure is constructed in Section 3 that leverages the multivariate wavelet-domain bootstrap approximations of pairwise test statistics. The key idea is to let the nodes of the graph represent the self-similarity parameters, and also to include the available statistical information in the graph vertices. First, a statistical test for the detection of parameters H_m that differ from all others is proposed, which enables us to detect the isolated nodes of the graph. Second, the remaining edges of the graph are weighted according to the p-values of the pairwise tests for equality of the respective self-similarity parameters. The actual clustering can then be performed by standard spectral clustering. The relevance and performance of this proposed approach is assessed in Section 4 with Monte Carlo experiments conducted on synthetic ofBm with different numbers of clusters in $\underline{H}$, sample sizes and numbers of components M. The results demonstrate that our bootstrap informed graph-based clustering strategy yields satisfactory performance both in terms of counting the number of actually distinct self-similarity parameters and also of counting the number of components that are controlled by each of them.

2. MULTIVARIATE SELF-SIMILARITY

Model. *Operator fractional Brownian motion* (ofBm) is a versatile multivariate extension of fractional Brownian motion (fBm),

G.D. is partially supported by the award NSF DMS 2515732. G.D.s long-term visits to ENS de Lyon were supported by the school, the CNRS and the Simons Foundation collaboration grant #714014.

the unique self-similar centered Gaussian process with stationary increments [11, 20]. The present work focuses on a special case of ofBm, the multivariate fBm (mfBm). It is particularly well adapted to real-world data modeling and is defined by M fBm $X_m(t), m = 1, \ldots, M$, with possibly distinct Hurst exponents $\underline{H} = (H_1, \ldots, H_M)$ and point correlation matrix $\mathbf{\Sigma}$, that are mixed by a $M \times M$ real-valued p.d. matrix $\mathbf{W}$ [16],

$$Y_{\mathbf{W},\underline{H},\mathbf{\Sigma}}(t) \triangleq \mathbf{W}\ \{X_1(t), \ldots, X_M(t)\}_{t\in\mathbb{R}} = \mathbf{W}X(t). \quad (1)$$

Moreover, it satisfies the (operator) self-similarity relation

$$\forall a > 0 : \{Y_{\mathbf{W},\underline{H},\mathbf{\Sigma}}(t)\}_{t\in\mathbb{R}} \overset{\text{f.d.d.}}{=} \{a^{\mathbf{H}} Y_{\mathbf{W},\underline{H},\mathbf{\Sigma}}(t/a))\}_{t\in\mathbb{R}}, \quad (2)$$

where the matrix Hurst exponent is given by $\mathbf{H} = \mathbf{W}\text{diag}(\underline{H})\mathbf{W}^{-1}$, and $a^{\mathbf{H}} = \sum_{k=0}^{\infty} \log^k(a)\mathbf{H}^k/k!$.

Self-similarity parameter estimation. The parameter vector $\underline{H} = (H_1, \ldots, H_M)$ controls the temporal dynamics of $Y_{\mathbf{W},H,\mathbf{\Sigma}}(t)$ and is of central interest in applications. The estimation usually relies on the multivariate discrete wavelet transform (DWT) of $Y_{\mathbf{W},H,\mathbf{\Sigma}}(t)$, which is given by the vectors $D_Y(2^j, k) \triangleq (D_{Y_1}(2^j, k), \ldots, D_{Y_M}(2^j, k))$ for $k \in \mathbb{Z}, j \in \{j_1, \ldots, j_2\}$, where the entries are defined by the component-wise DWT coefficients $D_{Y_m}(2^j, k) \triangleq \int_{\mathbb{R}} 2^{-j/2}\psi_0(2^{-j}t - k)Y_m(t)\mathrm{d}t$, where ψ_0 is a mother wavelet [21]. In particular, the *eigenvalues* $\lambda_m(2^j)$ of the *wavelet spectrum*

$$\mathbf{S}(2^j) \triangleq \frac{1}{n_j}\sum_{k=1}^{n_j} D_Y(2^j, k)D_Y(2^j, k)^T, n_j = \text{card}(D_Y(2^j, \cdot)) \quad (3)$$

asymptotically ($j \to +\infty$) behave as power laws $\lambda_m(2^j) \sim 2^{j(2H_m+1)}$ [7, 8]. This suggests estimation of $\underline{H}$ by regressions

$$\hat{H}_m = \frac{1}{2}\sum_{j=j_1}^{j_2} v_j \log_2 \lambda_m(2^j) - \frac{1}{2}, \quad \forall m = 1, \ldots, M, \quad (4)$$

with linear regression weights v_j as defined in, e.g., [22]. In practice, to reduce eigen-repulsion induced bias in the estimation, a snapshot-based estimator for the wavelet spectrum $\mathbf{S}(2^j)$ is used instead of (3), see [16] for details. The estimator $\hat{\underline{H}} = (\hat{H}_1, \ldots, \hat{H}_M)$ is asymptotically unbiased and jointly Gaussian in the limit of large sample sizes and large scales [7, 8].

3. SELF-SIMILARITY PARAMETER CLUSTERING

In this section, we propose a procedure for detecting groups of equal self-similarity parameters from a single finite-size observation of mfBm that makes use of the asymptotic normality of the estimator $(\hat{H}_1, \ldots, \hat{H}_M)$. In a first step, we design a statistical test for detecting single distinct values for $H_m \in \underline{H}$ (which cannot be detected by spectral clustering); in a second step, we propose a spectral clustering approach to group the remaining self-similarity parameters using a probabilistically weighted similarity graph. In both steps, a bootstrap estimator is used for the unknown parameters.

3.1. Pairwise dissimilarity measure, statistical model and test

It is natural to consider the statistic $\hat{\delta}_{m,m'} = \hat{H}_{m'} - \hat{H}_m$ as a measure for how close two self-similarity parameters H_m and $H_{m'}$, $1 \le m, m' \le M$, are. Indeed, this statistic asymptotically has a Gaussian distribution [7, 8]

$$\hat{\delta}_{m,m'} = \hat{H}_{m'} - \hat{H}_m \sim \mathcal{N}(H_{m'} - H_m, \sigma^2_{m,m'}). \quad (5)$$

This can be used to construct a test with significance α for the equality of a single pair of parameters H_m, $H_{m'}$, i.e., for the hypothesis

$$\mathcal{H}_0^{(m,m')} : H_m = H_{m'}, m' \neq m. \quad (6)$$

The test can be defined by

$$d_\alpha^{(m,m')} = \begin{cases} 1 & \text{if } p_{m,m'} < \alpha \quad (\mathcal{H}_0^{(m,m')} \text{ rejected}) \\ 0 & \text{otherwise} \qquad\quad (\mathcal{H}_0^{(m,m')} \text{ not rejected}) \end{cases} \quad (7)$$

where

$$p_{m,m'} = 2\left(1 - F_{\mathcal{N}(0,\sigma^2_{m,m'})}\left(|\hat{\delta}_{m,m'}|\right)\right) \quad (8)$$

is the probability to observe the value $\hat{\delta}_{m,m'}$ under the hypothesis $\mathcal{H}_0^{(m,m')}$, and $F_{\mathcal{N}(0,\sigma^2_{m,m'})}$ is the centered Gaussian cumulative distribution function with variance $\sigma^2_{m,m'}$.

3.2. Detection of single distinct self-similarity parameters

A self-similarity parameter H_m is called *distinct* if it is different from all other parameters $H_{m'}, m' \neq m$. To detect this situation, we formulate the hypothesis

$$\mathcal{H}_0^{(m)} : \quad \exists m' \neq m : \ H_m = H_{m'}. \quad (9)$$

It is rejected (and, thus, H_m is distinct) if the pairwise hypotheses (6) are rejected for *all* $m' \neq m$,

$$d_\alpha^{(m)} = \prod_{m'\neq m} d_\alpha^{(m,m')} = \begin{cases} 1: & \mathcal{H}_0^{(m)} \text{ rejected} \\ 0: & \mathcal{H}_0^{(m)} \text{ not rejected.} \end{cases} \quad (10)$$

To control its false positive rate, since the test is based on multiple independent tests, the size of the latter must be corrected. Here, we make use of the Benjamini-Hochberg correction [23]

$$d_\alpha^{(m,m')} = \begin{cases} 1 & \text{if } \pi^{-1}(m') \le \underset{j\in\{1,\ldots,M-1\}}{\arg\max}\ j\mathbb{1}_{\hat{p}^*_{m,\pi(j)} < \frac{\alpha}{M-1}j}, \\ 0 & \text{otherwise,} \end{cases} \quad (11)$$

where $\mathbb{1}_{j\in A} = 1$ if $j \in A$ and $\mathbb{1}_{j\in A} = 0$ otherwise, for any set A and π is the permutation that orders the p-values, $\hat{p}^*_{m,\pi(1)} < \ldots < \hat{p}^*_{m,\pi(M-1)}$. Alternatively, one could use the more conservative Bonferroni correction, i.e., a significance level $\alpha' = \alpha/(M-1)$ in (7).

3.3. Probabilistic graph spectral clustering

Similarity graph construction. To detect non-singleton groups of equal values in $\underline{H}$, we make use of spectral clustering, whose principles are briefly recalled below. The originality of our procedure resides in the graph construction, built with nodes representing the parameters H_m, and edge weights $\mathcal{S}_{m,m'}$ that are informed by the statistical model (5) under the null hypothesis $\mathcal{H}_0^{(m,m')}$, and by the test decisions (10) for identifying single distinct self-similarity parameters. Specifically, the weighted graph is defined as the triplet $\mathcal{G} = (\mathcal{V}, \epsilon, \mathcal{S})$ where $\mathcal{V} = \{1, \ldots, M\}$ is the set of vertices related to self-similarity parameters $(H_1, \ldots, H_M)$, $\epsilon = \{(m, m') : m, m' = 1, \ldots, M\}$ is the set of edges and $\mathcal{S}$ is the $M \times M$ symmetric weight matrix, for $m, m' = 1, \ldots, M$:

$$\mathcal{S}_{m,m'} = \begin{cases} p_{m,m'}(1 - d_\alpha^{(m)})(1 - d_\alpha^{(m')}) & m' \neq m, \\ 0 & m' = m. \end{cases} \quad (12)$$

Spectral clustering. We briefly recall the principles of spectral clustering. In this work, we consider the random-walk normalized Laplacian $\mathcal{L}_{rw}$, defined as follows [24–26]:

$$(\mathcal{L}_{rw})_{m,m'} = \begin{cases} 1 & \text{if } m = m' \text{ and } \mathcal{D}_{m,m} \neq 0, \\ -\frac{\mathcal{S}_{m,m'}}{\mathcal{D}_{m,m}} & \text{if } m \neq m' \text{ and } \mathcal{D}_{m,m} \neq 0, \\ 0 & \text{otherwise,} \end{cases} \quad (13)$$

for every $m, m' \in \{1, \ldots, M\}$, where the degree matrix $\mathcal{D}$ is the $M \times M$ diagonal matrix with entries $\mathcal{D}_{m,m} = \sum_{k=1}^{M} \mathcal{S}_{m,k}, m = 1, \ldots, M$. Now let denote $\varphi_1 \leq \ldots \leq \varphi_M$ the eigenvalues of $\mathcal{L}_{rw}$, and $v_1 \leq \ldots \leq v_M$ the corresponding eigenvectors. If the graph $\mathcal{G}$ is composed of N_{C} connected components, the Laplacian $\mathcal{L}_{rw}$ has exactly N_{C} eigenvalues equal to zero, i.e., $\varphi_1 = \ldots = \varphi_{N_{\mathrm{C}}} = 0$. In practice, the number N_{C} of clusters in $\mathcal{G}$ can be estimated from the number of eigenvalues φ_k close to zero, for example by means of the *maximum eigengap* [27],

$$\hat{N}_{\mathrm{C}} = \underset{k \in \{1,\ldots,M-1\}}{\arg\max} \; \varphi_{k+1} - \varphi_k. \quad (14)$$

Then, the clustering of the vertices $\mathcal{V} = \{1, \ldots, M\}$ into $\hat{N}_{\mathrm{C}}$ clusters is obtained by k-means clustering of the rows of the matrix $(v_1, \ldots, v_{\hat{N}_{\mathrm{C}}})$ comprised of the eigenvectors of the graph Laplacian $\mathcal{L}_{rw}$ associated with the $\hat{N}_{\mathrm{C}}$ smallest eigenvalues $\varphi_1, \ldots, \varphi_{\hat{N}_{\mathrm{C}}}$. Finally, because of the Laplacian normalization, spectral clustering cannot identify a node as single if it is linked to other nodes, leading to the need for a prior identification of these nodes using (10) [28].

3.4. Bootstrap p-values

The above self-similarity parameter clustering procedure requires knowledge of the variances $\sigma^2_{m,m'}$ in (5). Since they are unknown, we propose to estimate them using a bootstrap procedure [29, 30]. Specifically, we make use of a block-bootstrap resampling scheme for the multivariate DWT coefficients $D(2^j, k)$, $k = 1, \ldots, n_j$, which preserves their time-scale multivariate dependence structure [15, 17]: At each scale 2^j, R bootstrap resamples

$$D_j^{*(r)} = \left(D^{*(r)}(2^j, 1), \ldots, D^{*(r)}(2^j, n_j) \right), r = 1, \ldots, R, \quad (15)$$

are obtained by drawing with replacement $\lceil n_j / L_B \rceil$ elements from the collection $\{(D(2^j, k), \ldots, D(2^j, k + L_B - 1))\}_{k=1,\ldots,n_j}$ of overlapping blocks of circularized multivariate DWT coefficients of size L_B. For each resample, bootstrap estimates $\mathbf{S}^{*(r)}(2^j)$, $\lambda_m^{*(r)}(2^j)$ and $\hat{H}_m^{*(r)}$ are successively computed using Eqs. (3-4). The bootstrap estimation of the test statistics (5) under the null hypothesis $\mathcal{H}_0^{(m,m')}$ is given by

$$\hat{\delta}^{*(r)}_{m,m'} = \hat{H}^{*(r)}_{m'} - \hat{H}^{*(r)}_{m} - (\hat{H}_{m'} - \hat{H}_m), \; r = 1, \ldots, R, \quad (16)$$

and its empirical distribution is used as an approximation of the distribution of $\hat{\delta}_{m,m'}$ under $\mathcal{H}_0^{m,m'}$. The sample variance of the bootstrap resamples $\hat{\delta}^*_{m,m'}$ is used as an estimate for the variances $\sigma^2_{m,m'}$, and plugged into the expression (8) for the p-values.

4. PERFORMANCE ASSESSMENT

4.1. Monte Carlo simulation

Experimental setup. The proposed procedure is applied to $N_{\mathrm{MC}} = 1000$ ($M = 6$) and $N_{\mathrm{MC}} = 100$ ($M = 20$) independent realizations of M-fBm ($N \in \{2^{16}, 2^{18}\}$) with self-similarity parameters $\underline{H} = (H_1, \ldots, H_M)$ specified below. The covariance matrix $\mathbf{\Sigma}$ has unit diagonal entries and off-diagonal entries $r = 0.5$ ($M = 6$) and $r = 0.4$ ($M = 20$). The $M \times M$ mixing matrix $\mathbf{W}$ was drawn at random and kept fixed for all experiments. Wavelet analysis is performed with the least asymmetric Daubechies 3 mother wavelet and the linear regressions are performed across scales $(j_1, j_2) = (6, 8)$ ($N = 2^{16}$) and $(j_1, j_2) = (8, 10)$ ($N = 2^{18}$). For the bootstrap procedure, $R = 500$ and $L_B = 6$. FDR is set to $\alpha = 0.05$.

Clustering performance assessment. Performance is quantified in terms of Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) [31]: ARI measures the number of pairs of elements correctly gathered in the same cluster or separated in different clusters, and NMI measures the joint entropy of the estimated and correct cluster distributions. Both are defined in $[0, 1]$. Our proposed approach, denoted graph, is compared to that in [18, 19], denoted sort. The following scenarios are studied:

– Scenario1 (1 cluster): $H_1 = \ldots = H_M = 0.8$ ($M = 6, 20$).
– Scenario2 (2 clusters): $H_m \in (0.6, 0.8)$;
  cluster size $(3, 3)$ ($M = 6$) and $(10, 10)$ ($M = 20$).
– Scenario3 (3 clusters): $H_m \in (0.4, 0.6, 0.8)$;
  cluster size $(2, 2, 2)$ ($M = 6$) and $(7, 7, 6)$ ($M = 20$).
– Scenario4 (3 clusters): $H_m \in (0.4, 0.6, 0.8)$;
  cluster size $(1, 3, 2)$ ($M = 6$) and $(13, 6, 1)$ ($M = 20$).

Table 1: **Detection of isolated nodes.** Average detection rate for each individual H_m, $m = 1, \ldots, 6$, using (10) and (11) under Scenario4. The red boxes indicate the isolated node H_1.

N	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$	$m=6$
2^{16}	0.992	0.030	0.032	0.001	0.067	0.064
2^{18}	0.999	0.015	0.017	0.010	0.055	0.055

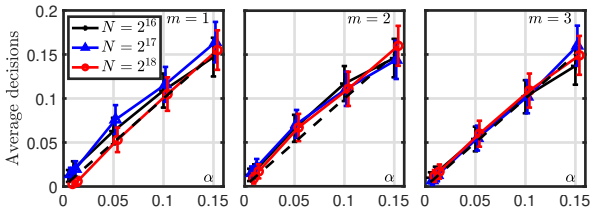

Fig. 1: **False positive rate for isolated nodes.** Average rate (with 95% error bars) of rejecting at least one null hypothesis $\mathcal{H}_0^{(m,m')}$, $m' \neq m$, for $m = 1, 2, 3$ for Scenario1 ($M = 6$) versus FDR α.

4.2. Clustering performance

Single-node detection. The detection performance for isolated nodes (clusters of size 1) is assessed for Scenario4 ($M = 6$). Table 1 reports the average (over Monte Carlo realizations) detection rate for all 6 nodes when using (10) only. Clearly, the isolated node H_1 is detected in almost all cases for both sample sizes, with a false negative rate below 1%. Along the same lines, the false positive rate (for the nodes $H_2, \ldots, H_6$, which belong to larger clusters) is well controlled and of the order of $3 - 4\%$ on average. This is further quantified in Fig. 1, which plots the average rate of rejecting at least one of the null hypothesis $\mathcal{H}_0^{(m,m')}$, $m' \neq m$, for $m = 1, 2, 3$ under Scenario1 ($M = 6$), showing that the average rejection rate closely follows the prescribed FDR level. All in all, this leads us to conclude that the proposed strategy for detecting isolated nodes is effective.

Cluster detection performance. Fig. 2 shows the histograms of the estimated numbers of clusters $\hat{N}_{\mathrm{C}}$ given by (14) for all scenar-

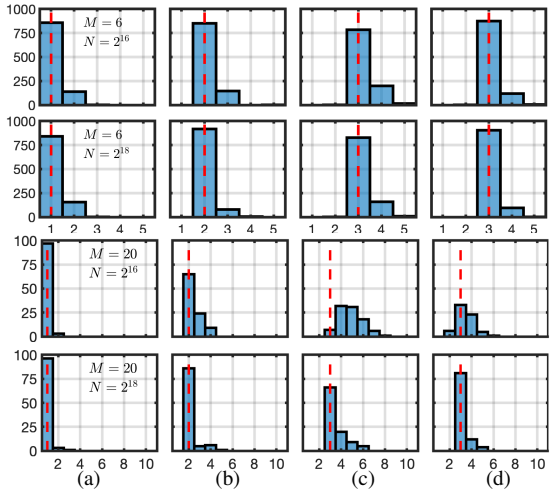

Fig. 2: **Estimation of the number of clusters.** Histograms of $\hat{N}_{\mathrm{C}}$ obtained by spectral clustering for Scenario1-4 and different numbers of components M and sample sizes N. The red dashed lines indicate the exact number of clusters.

ios. For $M = 6$, the proposed procedure detects the correct number of clusters N_{C} with high probability for both sample sizes. For the more difficult case with $M = 20$ components, detection performance is again satisfactory for the larger sample size $N = 2^{18}$, while the proposed procedure overestimates the number of clusters for smaller sample sizes, in particular for Scenario3-4 (3 clusters).

Table 2 (top) further quantifies these observations and reports NMI and ARI values for the proposed clustering strategy (graph). For Scenario1 (single cluster), performance is excellent (with ARI values ranging from 0.84 to 0.97) for both sample sizes and numbers of components. For Scenario2-4, performance are excellent for $M = 6$ components (with NMI and ARI values in excess of 0.9 for both sample sizes). They are less satisfactory for $M = 20$ components for the smaller sample size ($N = 2^{16}$) but also approach NMI and ARI values of 0.9 for the larger sample size ($N = 2^{18}$). Table 2 (bottom) reports the respective results obtained as in [18, 19] (sort), showing that while it is on par with the proposed method for $M = 6$, it fails for large component number ($M = 20$), suggesting that the bootstrap approximation of the test statistic underlying [18, 19] gets increasingly poor for large M.

Self-similarity parameter estimation performance. We finally study if clustering can improve the estimation of the self-similarity parameters $\underline{H}$. To that end, a post-clustering estimate $\hat{\underline{H}}_{\mathrm{av}}$ is defined for which each entry m is given by the average over the estimates $\hat{H}_{m'}$ from (4) of the cluster it belongs to. Table 3 reports the root mean squared error (RMSE) $\sqrt{\widehat{\mathbb{E}}||\hat{\underline{H}} - \underline{H}||^2}$, where $\widehat{\mathbb{E}}$ stands for averaging over Monte Carlo realizations, for $\hat{\underline{H}}$, $\hat{\underline{H}}_{\mathrm{av}}^{\mathrm{sort}}$ and $\hat{\underline{H}}_{\mathrm{av}}^{\mathrm{graph}}$. The results show that the proposed clustering globally yields more accurate estimates for $\underline{H}$, in particular when there are large clusters (Scenario1 and 4) and when sample size is large. Moreover, it significantly outperforms the one proposed in [18, 19] for large number of components M.

Overall, these results show that the proposed bootstrap-driven clustering strategy for estimating the number of distinct self-similarity parameters is operational and yields satisfactory performance for reasonable sample size and number of components.

Table 2: **Clustering performance.** NMI and ARI (Monte Carlo average $\pm$ 95% confidence interval) for Scenario1-4 and different numbers of components M and sample sizes N (best results in bold).

M	N	**graph**	Scenario1	Scenario2	Scenario3	Scenario4
6	2^{16}	NMI	n/a	$\mathbf{0.95 \pm 0.01}$	$\mathbf{0.96 \pm 0.00}$	$\mathbf{0.98 \pm 0.00}$
		ARI	$\mathbf{0.86 \pm 0.02}$	$\mathbf{0.92 \pm 0.01}$	$\mathbf{0.90 \pm 0.01}$	0.94 ± 0.01
	2^{18}	NMI	n/a	$\mathbf{0.98 \pm 0.01}$	$\mathbf{0.98 \pm 0.00}$	$\mathbf{0.99 \pm 0.00}$
		ARI	0.84 ± 0.02	$\mathbf{0.97 \pm 0.01}$	$\mathbf{0.95 \pm 0.08}$	0.97 ± 0.01
20	2^{16}	NMI	n/a	$\mathbf{0.53 \pm 0.03}$	$\mathbf{0.71 \pm 0.02}$	$\mathbf{0.58 \pm 0.03}$
		ARI	$\mathbf{0.97 \pm 0.03}$	$\mathbf{0.49 \pm 0.04}$	$\mathbf{0.54 \pm 0.04}$	$\mathbf{0.41 \pm 0.07}$
	2^{18}	NMI	n/a	$\mathbf{0.69 \pm 0.02}$	$\mathbf{0.90 \pm 0.02}$	$\mathbf{0.90 \pm 0.02}$
		ARI	0.96 ± 0.04	$\mathbf{0.69 \pm 0.03}$	$\mathbf{0.86 \pm 0.03}$	$\mathbf{0.88 \pm 0.04}$
M	N	**sort**	Scenario1	Scenario2	Scenario3	Scenario4
6	2^{16}	NMI	n/a	0.77 ± 0.02	0.95 ± 0.01	$\mathbf{0.98 \pm 0.00}$
		ARI	$\mathbf{0.86 \pm 0.02}$	$\mathbf{0.92 \pm 0.01}$	$\mathbf{0.90 \pm 0.01}$	$\mathbf{0.95 \pm 0.01}$
	2^{18}	NMI	n/a	0.97 ± 0.01	$\mathbf{0.98 \pm 0.00}$	$\mathbf{0.99 \pm 0.00}$
		ARI	$\mathbf{0.99 \pm 0.00}$	0.96 ± 0.01	$\mathbf{0.95 \pm 0.01}$	$\mathbf{0.98 \pm 0.01}$
20	2^{16}	NMI	n/a	0.00 ± 0.01	0.06 ± 0.03	0.12 ± 0.04
		ARI	0.96 ± 0.04	0.00 ± 0.00	0.03 ± 0.02	0.11 ± 0.05
	2^{18}	NMI	n/a	0.17 ± 0.06	0.75 ± 0.06	0.75 ± 0.05
		ARI	$\mathbf{0.97 \pm 0.03}$	0.17 ± 0.06	0.65 ± 0.06	0.62 ± 0.08

Table 3: **Estimation for H.** Global RMSE$\times 10^2$ of the estimates $\hat{\underline{H}}$ and the estimates averaged over detected clusters $\hat{\underline{H}}_{\mathrm{av}}$, for different scenarios and sample sizes N (best results in bold).

M	N		Scenario1	Scenario2	Scenario3	Scenario4
6	2^{16}	$\hat{\underline{H}}$	2.52	3.99	**3.78**	3.26
		$\hat{\underline{H}}_{\mathrm{av}}^{\mathrm{sort}}$	**1.05**	5.38	4.22	3.03
		$\hat{\underline{H}}_{\mathrm{av}}^{\mathrm{graph}}$	1.50	**3.62**	3.79	**2.84**
	2^{18}	$\hat{\underline{H}}$	2.59	2.93	3.06	2.70
		$\hat{\underline{H}}_{\mathrm{av}}^{\mathrm{sort}}$	**1.10**	2.62	2.87	**2.11**
		$\hat{\underline{H}}_{\mathrm{av}}^{\mathrm{graph}}$	1.58	**2.22**	**2.70**	**2.11**
20	2^{16}	$\hat{\underline{H}}$	2.13	**6.53**	**6.47**	**5.81**
		$\hat{\underline{H}}_{\mathrm{av}}^{\mathrm{sort}}$	**0.63**	10.01	15.65	10.89
		$\hat{\underline{H}}_{\mathrm{av}}^{\mathrm{graph}}$	0.74	6.82	6.88	6.15
	2^{18}	$\hat{\underline{H}}$	2.20	**5.29**	4.10	3.46
		$\hat{\underline{H}}_{\mathrm{av}}^{\mathrm{sort}}$	**0.61**	9.12	8.36	6.49
		$\hat{\underline{H}}_{\mathrm{av}}^{\mathrm{graph}}$	0.79	5.32	**3.83**	**3.28**

5. CONCLUSIONS AND PERSPECTIVES

This work develops a clustering strategy to count the number of distinct scaling exponents and group identical ones based on a single multivariate time series. The procedure constructs a weighted graph statistically informed from pairwise self-similarity parameter estimates and an original wavelet-domain block-bootstrap resampling scheme operating on two levels: for identifying isolated nodes, and for quantifying the proximity of pairwise self-similarity parameter estimates. Clustering is then performed using standard spectral clustering with a random-walk normalized Laplacian. Numerical experiments for operator fractional Brownian motion demonstrate that the proposed clustering strategy yields satisfactory performance for various scenarios involving different numbers of components and clusters and can moreover improve self-similarity parameter estimation. Future research will explore alternative formulations for statistically informed graph weights and applications to high-dimensional multivariate data, such as the study of brain dynamics. Code is available at https://github.com/charlesglucas/ofbm_tools.

6. REFERENCES

[1] B. Mandelbrot, "A multifractal walk down wall street," *Scientific American*, vol. 280, no. 2, pp. 70–73, Feb. 1999.

[2] B. B. Mandelbrot, "Intermittent turbulence in self-similar cascades: divergence of high moments and dimension of the carrier," *J. Fluid Mech.*, vol. 62, pp. 331–358, 1974.

[3] T. Nakamura, K. Kiyono, H. Wendt, P. Abry, and Y. Yamamoto, "Multiscale analysis of intensive longitudinal biomedical signals and its clinical applications," *Proc. IEEE*, vol. 104, no. 2, pp. 242–261, 2016. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7386807

[4] R. Fontugne, P. Abry, K. Fukuda, D. Veitch, K. Cho, P. Borgnat, and H. Wendt, "Scaling in Internet traffic: a 14 year and 3 day longitudinal study, with multiscale analyses and random projections," *IEEE/ACM Transactions on Networking*, vol. 25, no. 4, pp. 1–14, 2017.

[5] D. La Rocca, N. Zilber, P. Abry, V. van Wassenhove, and P. Ciuciu, "Self-similarity and multifractality in human brain activity: A wavelet-based analysis of scale-free brain dynamics," *Journal of Neuroscience Methods*, vol. 309, pp. 175–187, 2018.

[6] P. Abry, H. Wendt, S. Jaffard, and G. Didier, "Multivariate scale-free temporal dynamics: From spectral (fourier) to fractal (wavelet) analysis," *Comptes Rendus Physique*, vol. 20, no. 5, pp. 489–501, 2019.

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

[8] ——, "Wavelet eigenvalue regression for n-variate operator fractional Brownian motion," *Journal of Multivariate Analysis*, vol. 168, pp. 75–104, November 2018.

[9] M. Maejima and J. D. Mason, "Operator-self-similar stable processes," *Stochastic Processes and their Applications*, vol. 54, no. 1, pp. 139–163, 1994.

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

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

[12] ——, "Exponents, symmetry groups and classification of operator fractional Brownian motions," *Journal of Theoretical Probability*, vol. 25, pp. 353–395, 2012.

[13] P.-O. Amblard and J.-F. Coeurjolly, "Identification of the multivariate fractional Brownian motion," *IEEE Trans. Signal Process.*, vol. 59, no. 11, pp. 5152–5168, 2011.

[14] H. Wendt, P. Abry, and G. Didier, "Bootstrap-based bias reduction for the estimation of the self-similarity exponents of multivariate time series," in *IEEE Int. Conf. Acoust., Speech, and Signal Proces. (ICASSP)*, Brighton, UK, May 2019.

[15] C.-G. Lucas, P. Abry, H. Wendt, and G. Didier, "Bootstrap for testing the equality of selfsimilarity exponents across multivariate time series," in *Proc. European Signal Processing Conference (EUSIPCO)*, Dublin, Ireland, August 2021.

[16] C.-G. Lucas, G. Didier, H. Wendt, and P. Abry, "Multivariate selfsimilarity: Multiscale eigenstructures for selfsimilarity parameter estimation," *IEEE Transactions on Signal Processing*, 2024.

[17] H. Wendt, P. Abry, and G. Didier, "Wavelet domain bootstrap for testing the equality of bivariate self-similarity exponents," in *Proc. IEEE Workshop Statistical Signal Process. (SSP)*, Freiburg, Germany, June 2018.

[18] C.-G. Lucas, P. Abry, H. Wendt, and G. Didier, "Counting the number of different scaling exponents in multivariate scale-free dynamics: Clustering by bootstap in the wavelet domain," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Singapore, May 2022.

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

[20] V. Pipiras and M. S. Taqqu, *Long-Range Dependence and Self-Similarity*. Cambridge University Press, 2017, vol. 45.

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

[22] 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.

[23] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

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

[25] P. Sarkar and P. J. Bickel, "Role of normalization in spectral clustering for stochastic blockmodels," *The Annals of Statistics*, vol. 43, no. 3, pp. 962–990, 2015.

[26] U. Von Luxburg, M. Belkin, and O. Bousquet, "Consistency of spectral clustering," *The Annals of Statistics*, pp. 555–586, 2008.

[27] A. Azran and Z. Ghahramani, "Spectral methods for automatic multiscale data clustering," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 1. IEEE, 2006, pp. 190–197.

[28] L. Page, "The pagerank citation ranking: Bringing order to the web," Technical Report, Tech. Rep., 1999.

[29] S. N. Lahiri, *Resampling Methods for Dependent Data*. New York: Springer, 2003.

[30] A. M. Zoubir and D. R. Iskander, *Bootstrap techniques for signal processing*. Cambridge University Press, 2004.

[31] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *The Journal of Machine Learning Research*, vol. 11, pp. 2837–2854, 2010.