Counting the number of different scaling exponents in multivariate scale-free dynamics: Clustering by bootstrap in the wavelet domain

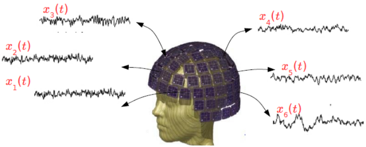
ICASSP 2022, Singapore

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

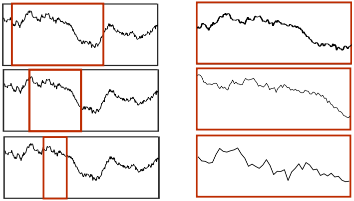
[1]ENSL, CNRS, Laboratoire de physique, Lyon, France.
[2]IRIT, Univ. Toulouse, CNRS, Toulouse, France.
[3]Math. Dept., Tulane University, New Orleans, USA.

Multivariate self-similarity

1) Multivariate setting



2) Univariate self-similarity

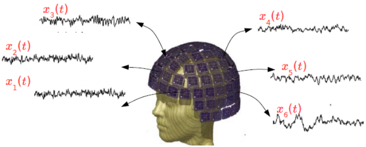


$B_H(t)$ characterized by $0 < H < 1$

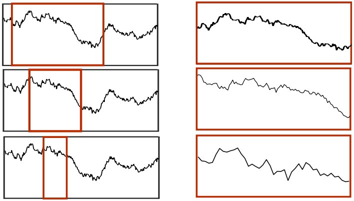
$\Rightarrow$ Multivariate self-similarity: $\underline{H} = (H_1, \dots, H_M)$

Multivariate self-similarity

1) Multivariate setting



2) Univariate self-similarity



$B_H(t)$ characterized by $0 < H < 1$

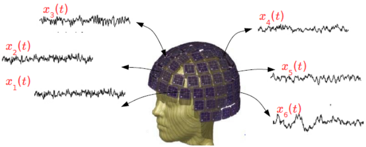
$\Rightarrow$ Multivariate self-similarity: $\underline{H} = (H_1, \ldots, H_M)$

Goals:

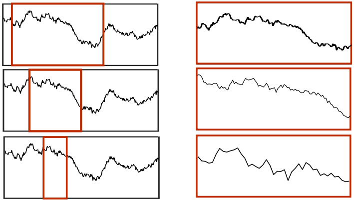
- Count the number of H_m actually different

Multivariate self-similarity

1) Multivariate setting



2) Univariate self-similarity



$B_H(t)$ characterized by $0 < H < 1$

$\Rightarrow$ Multivariate self-similarity: $\underline{H} = (H_1, \ldots, H_M)$

Goals:

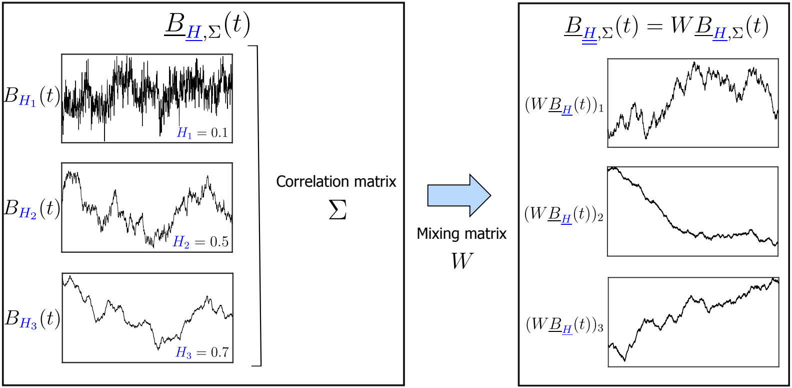
- Count the number of H_m actually different
- Count the number of components with same H_m

Outline

1. Scaling exponent clustering procedure
   - Multivariate self-similarity model
   - Estimation
   - Testing procedure

2. Does it work ?
   - Half-normal statistic
   - Bootstrap and half-normal law
   - Clustering performance

Multivariate self-similarity model [Didier et al., 2011]



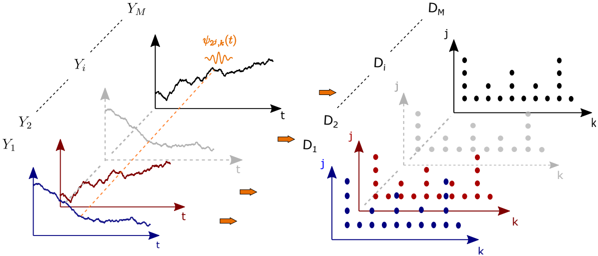
$\underline{B}_{\underline{\underline{H}},\Sigma}(t)$ characterized by the matrix $\underline{\underline{H}} = W\text{diag}(\underline{H})W^{-1}$

By convention: $0 < H_1 \leq \ldots \leq H_M < 1$

Step 1: estimation of $\underline{H} = (H_1, \ldots, H_M)$

Multivariate estimation

- Multivariate wavelet transform of $Y = W\underline{B}_{\underline{H},\Sigma}$:
  - ψ_0: mother wavelet
  - $D_m(2^j, k) = \langle 2^{-j/2}\psi_0(2^{-j}t - k) | Y_m(t)\rangle$
  - $D(2^j, k) = (D_1(2^j, k), \ldots, D_M(2^j, k))$



- Wavelet spectrum ($M \times M$ matrix):

$$S_{m_1,m_2}(2^j) = \frac{1}{N_j}\sum_{k=1}^{N_j} D_{m_1}(2^j, k) D_{m_2}(2^j, k)^*, \; N_j = \frac{N}{2^j}, \; N\text{: sample size}$$

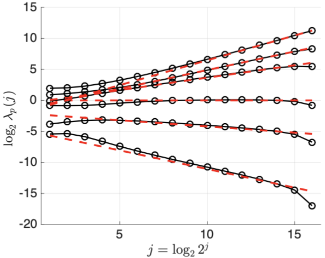
Multivariate estimation [Didier and Abry, 2018]

Eigenvalues of $S(2^j)$:

$$S(2^j) = U(2^j) \begin{bmatrix} \lambda_1(2^j) & 0 & \cdots & \cdots & 0 \\ 0 & \lambda_2(2^j) & \cdots & \cdots & 0 \\ 0 & 0 & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ 0 & 0 & \cdots & 0 & \lambda_M(2^j) \end{bmatrix} U(2^j)^T$$

- $Y = W\underline{B}_{\underline{H},\Sigma}$ self-similar
  $\Rightarrow$ asymptotical power law: $\lambda_m(2^j) \propto 2^{j(2H_m+1)}$
- Linear regression on log-eigenvalues:

$$\hat{H}_m = \frac{1}{2}\sum_{j=j_1}^{j_2} \omega_j \log_2 \lambda_m(2^j) - \frac{1}{2}$$



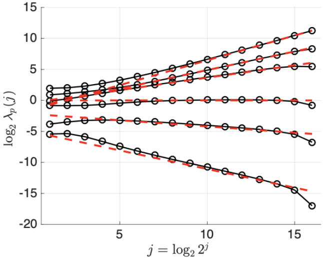
Multivariate estimation [Didier and Abry, 2018]

Eigenvalues of $S(2^j)$:

$$S(2^j) = U(2^j) \begin{bmatrix} \lambda_1(2^j) & 0 & \cdots & \cdots & 0 \\ 0 & \lambda_2(2^j) & \cdots & \cdots & 0 \\ 0 & 0 & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ 0 & 0 & \cdots & 0 & \lambda_M(2^j) \end{bmatrix} U(2^j)^T$$

- $Y = W\underline{B}_{\underline{H},\Sigma}$ self-similar
  $\Rightarrow$ asymptotical power law: $\lambda_m(2^j) \propto 2^{j(2H_m+1)}$
- Linear regression on log-eigenvalues:

$$\hat{H}_m = \frac{1}{2}\sum_{j=j_1}^{j_2} \omega_j \log_2 \lambda_m(2^j) - \frac{1}{2}$$



Issue:

Different numbers of wavelet coefficients to compute $S(2^j)$ between scales 2^j
$\Rightarrow$ $\lambda_m(2^j)$ have different bias across scale 2^j
$\Rightarrow$ bias corrected estimation [Lucas et al., EUSIPCO 2021]

Testing $H_m = H_{m+1}$

By convention: $0 < H_1 \leq \ldots \leq H_M < 1$

- Test formulation:
  - $M-1$ hypotheses:

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

  - Estimates $\hat{H}_m \to$ sorting $\hat{H}_{\tau(1)} < \ldots < \hat{H}_{\tau(M)}$
  - Statistics

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

Testing $H_m = H_{m+1}$

By convention: $0 < H_1 \leq \ldots \leq H_M < 1$

- Test formulation:
  - $M-1$ hypotheses:
  $$\mathcal{H}_0^{(m)} : H_m = H_{m+1}, \quad m = 1, \ldots, M-1$$
  - Estimates $\hat{H}_m \rightarrow$ sorting $\hat{H}_{\tau(1)} < \ldots < \hat{H}_{\tau(M)}$
  - Statistics
  $$\tilde{\delta}_m = \hat{H}_{\tau(m+1)} - \hat{H}_{\tau(m)}$$
- Test statistic: $\tilde{\delta}_m$, approximated by a half-normal distribution, under $\mathcal{H}_0^{(m)}$
$$f(\tilde{\delta}_m | H_m = H_{m+1}) = \frac{1}{\sigma_m} \sqrt{\frac{2}{\pi}} \exp\left(-\frac{\tilde{\delta}_m{}^2}{2\sigma_m{}^2}\right)$$

Testing $H_m = H_{m+1}$

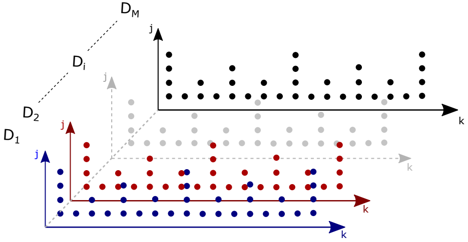
By convention: $0 < H_1 \leq \ldots \leq H_M < 1$

- Test formulation:
  - $M-1$ hypotheses:
    $$\mathcal{H}_0^{(m)} : H_m = H_{m+1}, \quad m = 1, \ldots, M-1$$
  - Estimates $\hat{H}_m \rightarrow$ sorting $\hat{H}_{\tau(1)} < \ldots < \hat{H}_{\tau(M)}$
  - Statistics
    $$\tilde{\delta}_m = \hat{H}_{\tau(m+1)} - \hat{H}_{\tau(m)}$$
- Test statistic: $\tilde{\delta}_m$, approximated by a half-normal distribution, under $\mathcal{H}_0^{(m)}$
  $$f(\tilde{\delta}_m | H_m = H_{m+1}) = \frac{1}{\sigma_m}\sqrt{\frac{2}{\pi}} \exp\left(-\frac{\tilde{\delta}_m{}^2}{2\sigma_m{}^2}\right)$$
- Test decision:
  $$\text{rejects } \mathcal{H}_0^{(m)} \text{ if } \tilde{\delta}_m > \gamma_m(\sigma_m)$$

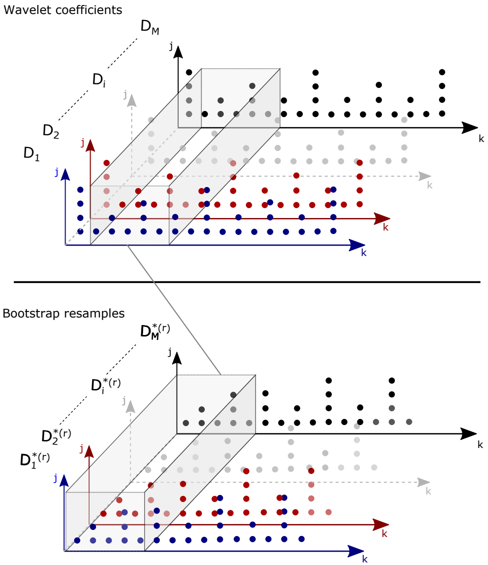
Issue: σ_m unknown $\Rightarrow$ estimation under $\mathcal{H}_0^{(m)}$ from a single observation
$\Rightarrow$ Bootstrap resampling

Multivariate wavelet block-bootstrap resamples

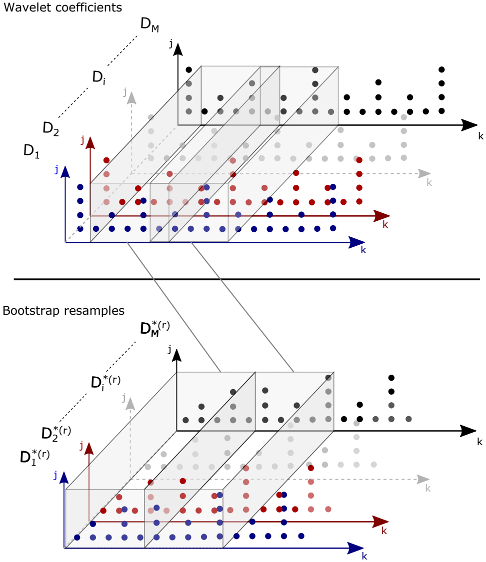
Wavelet coefficients



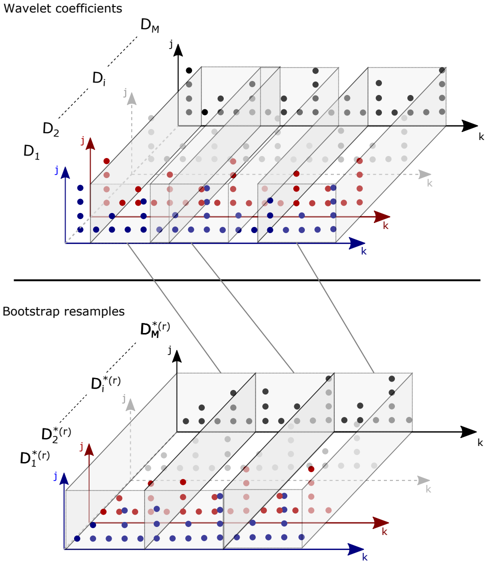
Multivariate wavelet block-bootstrap resamples



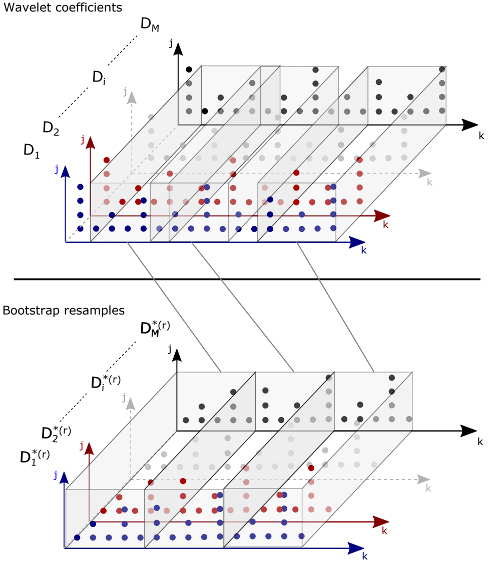
Multivariate wavelet block-bootstrap resamples



Multivariate wavelet block-bootstrap resamples

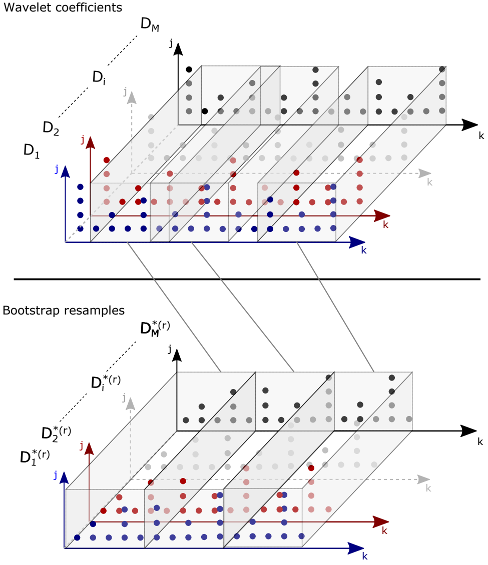


Multivariate wavelet block-bootstrap resamples



$\Rightarrow$ R wavelet coefficient resamples
$$D^{*(r)} = (D_1^{*(r)}, \ldots, D_M^{*(r)})$$

Multivariate wavelet block-bootstrap resamples



$\Rightarrow$ R wavelet coefficient resamples

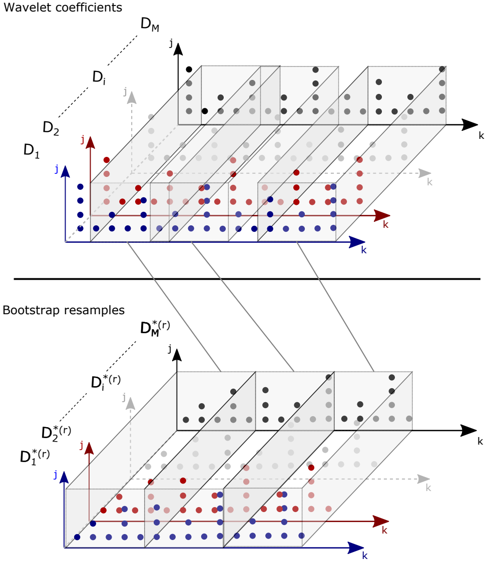
$$D^{*(r)} = (D_1^{*(r)}, \dots, D_M^{*(r)})$$

$$\Downarrow$$

R Bootstrap estimates

$$\underline{\hat{H}}^{*(r)} = (\hat{H}_1^{*(r)}, \dots, \hat{H}_M^{*(r)})$$

Multivariate wavelet block-bootstrap resamples



$\Rightarrow$ R wavelet coefficient resamples

$$D^{*(r)} = (D_1^{*(r)}, \dots, D_M^{*(r)})$$

$$\Downarrow$$

R Bootstrap estimates

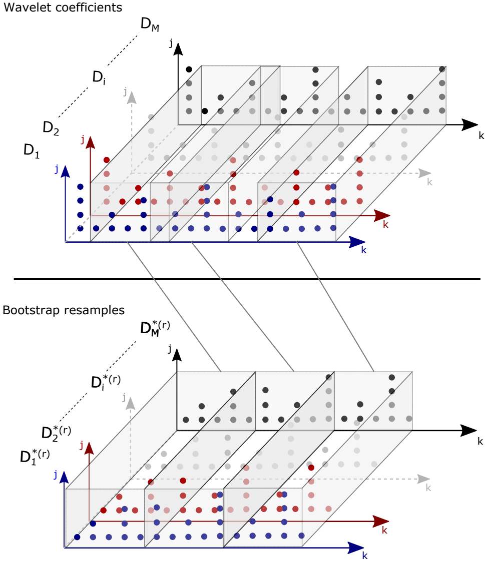
$$\underline{\hat{H}}^{*(r)} = (\hat{H}_1^{*(r)}, \dots, \hat{H}_M^{*(r)})$$

$$\Downarrow$$

Simulate half-normal hypotheses:

$$\bar{H}_m^{*(r)} = \hat{H}_m^{*(r)} - \langle \hat{H}_m^{*} \rangle$$

Multivariate wavelet block-bootstrap resamples



$\Rightarrow$ R wavelet coefficient resamples

$$D^{*(r)} = (D_1^{*(r)}, \ldots, D_M^{*(r)})$$

$$\Downarrow$$

R Bootstrap estimates

$$\underline{\hat{H}}^{*(r)} = (\hat{H}_1^{*(r)}, \ldots, \hat{H}_M^{*(r)})$$

$$\Downarrow$$

Simulate half-normal hypotheses:

$$\bar{H}_m^{*(r)} = \hat{H}_m^{*(r)} - \langle \hat{H}_m^{*} \rangle$$

$$\Downarrow$$

Ordered estimates:

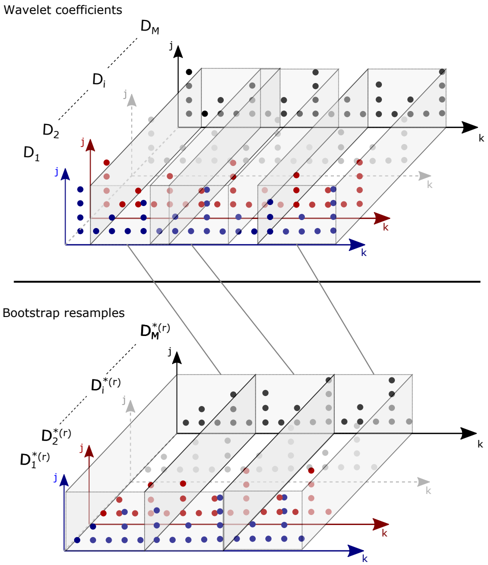
$$\bar{H}_{\tau^*(r,1)}^{*(r)} < \ldots < \bar{H}_{\tau^*(r,M)}^{*(r)}$$

$$\Downarrow$$

Bootstrap statistics

$$\tilde{\delta}_m^{*(r)} = \bar{H}_{\tau^*(r,m+1)}^{*(r)} - \bar{H}_{\tau^*(r,m)}^{*(r)}$$

Multivariate wavelet block-bootstrap resamples



$\Rightarrow$ R wavelet coefficient resamples

$$D^{*(r)} = (D_1^{*(r)}, \ldots, D_M^{*(r)})$$

$$\Downarrow$$

R Bootstrap estimates

$$\underline{\hat{H}}^{*(r)} = (\hat{H}_1^{*(r)}, \ldots, \hat{H}_M^{*(r)})$$

$$\Downarrow$$

Simulate half-normal hypotheses:

$$\bar{H}_m^{*(r)} = \hat{H}_m^{*(r)} - \langle \hat{H}_m^{*} \rangle$$

$$\Downarrow$$

Ordered estimates:

$$\bar{H}_{\tau^*(r,1)}^{*(r)} < \ldots < \bar{H}_{\tau^*(r,M)}^{*(r)}$$

$$\Downarrow$$

Bootstrap statistics

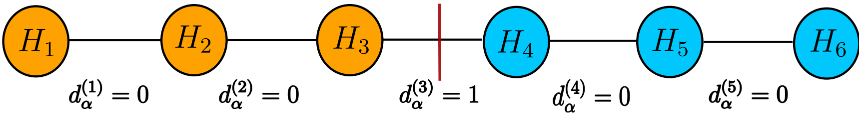
$$\tilde{\delta}_m^{*(r)} = \bar{H}_{\tau^*(r,m+1)}^{*(r)} - \bar{H}_{\tau^*(r,m)}^{*(r)}$$

$$\Downarrow$$

$$\hat{\sigma}_m^{*2} = \mathrm{Var}^*(\tilde{\delta}_m^*)\left(1 - \frac{2}{\pi}\right)$$

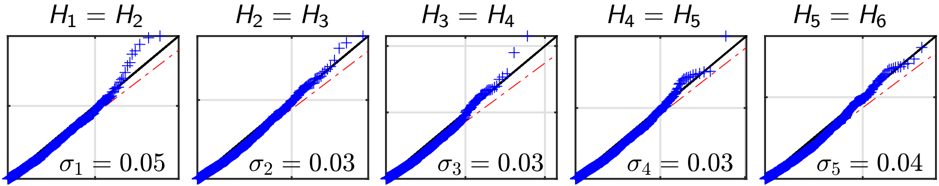
Clustering strategy

- p-values: $p_m^* = 1 - F_{\mathcal{HN}}\left(\tilde{\delta}_m / \hat{\sigma}_m^*\right)$
- Multiple hypothesis test (Benjamini-Hochberg) corrections:
  - α: false discovery rate
  - $p_{\pi(m)}^*$: sorted p-values of the test
  - $d_\alpha^{(m)} = 1$ if $p_{\pi(m)}^* < \frac{\alpha}{M-1} m$
- Clustering



Null distribution of $\tilde{\delta}_m$

- Monte Carlo simulations
  - $N_{MC} = 1000$ realizations
  - $M = 6$ components
  - sample size $N = 2^{16}$

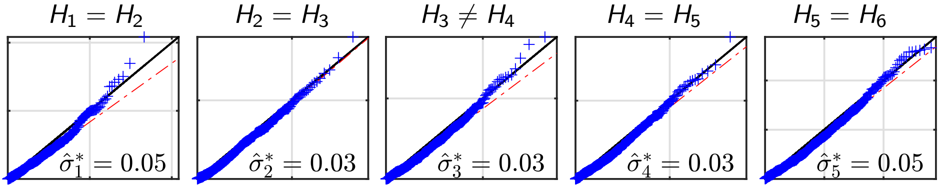


Quantile-quantile plot: Monte Carlo $\tilde{\delta}_m$ against half-normal distribution

$\Rightarrow$ Under $H_m = H_{m+1}$, $\tilde{\delta}_m$ is half-normal

Bootstrap null distribution estimation

- Null hypothesis and alternative hypotheses:



Quantile-quantile: Bootstrap $\tilde{\delta}_m^*$ against half-normal distribution

$\Rightarrow$ Under both null hypothesis and alternative hypothesis, $\tilde{\delta}_m^*$ is half-normal

Bootstrap scale parameter estimation

- σ_m vs. $\hat{\sigma}_m^*$
- Null hypothesis for any pair: $H_1 = \ldots = H_M$
- Monte Carlo average $\pm$ standard deviation

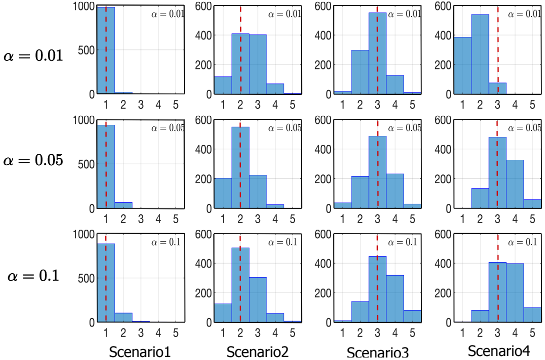
	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
σ_m	4.62	3.01	2.72	2.97	4.29
$\hat{\sigma}_m^*$	4.74 $\pm$0.59	3.08 $\pm$0.25	2.73 $\pm$0.19	2.93 $\pm$0.20	4.12 $\pm$0.31

$\Rightarrow$ Bootstrap estimates $\hat{\sigma}_m^*$ well approximate σ_m

Clustering performance

Histograms of the estimated numbers of clusters for several α

Scenario1 (1 cluster): $\underline{H} = (0.8, 0.8, 0.8, 0.8, 0.8, 0.8)$
Scenario2 (2 clusters): $\underline{H} = (0.6, 0.6, 0.6, 0.8, 0.8, 0.8)$
Scenario3 (3 clusters): $\underline{H} = (0.4, 0.4, 0.6, 0.6, 0.8, 0.8)$
Scenario4 (3 clusters): $\underline{H} = (0.4, 0.6, 0.6, 0.6, 0.8, 0.8)$



Clustering performance

Quantification

NMI: Normalized Mutual Information
(joint entropy of ground truth partition and estimated partition)
ARI: Adjusted Rand Index
(pairs of elements correctly separated or correctly gathered)

Monte Carlo average ± 95% confidence interval

$\alpha = 0.05$	Scenario1	Scenario2	Scenario3	Scenario4
NMI	n/a	0.66 ± 0.02	0.87 ± 0.01	0.79 ± 0.01
ARI	0.94 ± 0.02	0.60 ± 0.03	0.68 ± 0.02	0.59 ± 0.02

Scenario1 (1 cluster): $\underline{H} = (0.8, 0.8, 0.8, 0.8, 0.8, 0.8)$
Scenario2 (2 clusters): $\underline{H} = (0.6, 0.6, 0.6, 0.8, 0.8, 0.8)$
Scenario3 (3 clusters): $\underline{H} = (0.4, 0.4, 0.6, 0.6, 0.8, 0.8)$
Scenario4 (3 clusters): $\underline{H} = (0.4, 0.6, 0.6, 0.6, 0.8, 0.8)$

Conclusion

Achieved:

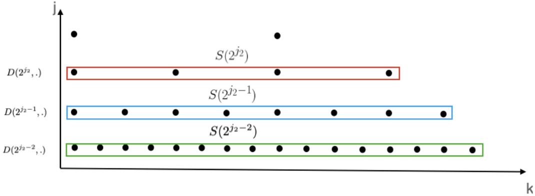
- From a single observation
- Testing procedure for $M - 1$ pairwise hypotheses from ordered estimates
- Clustering of self-similarity exponents

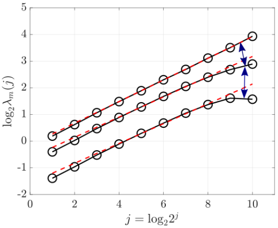
Perspectives:

- Test based on comparing all pairs $(H_m, H_{m'})$
- Large dimension: number of components M $\approx$ sample size N
- Application to real data: drowsiness detection [Lucas et al., EMBC 2022]

Repulsion effect

Gap between eigenvalues larger than expected at each scale





Issue: Few coefficients at large scales

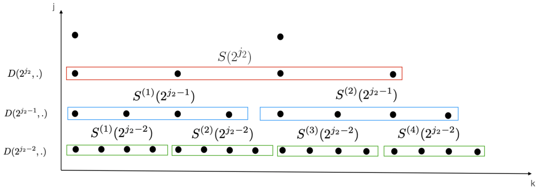
$\Rightarrow$ repulsion effect: important bias when $H_1 = \ldots = H_M$

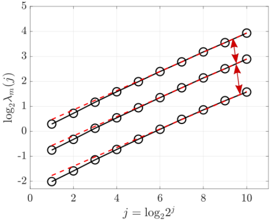
$\Rightarrow$ repulsion effect increases with scale 2^j

Bias corrected estimation [Lucas et al., EUSIPCO 2021]

$$S^{(w)}(2^j) \triangleq \frac{1}{n_{j_2}} \sum_{k=1+(w-1)n_{j_2}}^{wn_{j_2}} D(2^j, k)D(2^j, k)^*, \quad w = 1, \dots, 2^{j-j_2}, \quad n_{j_2} = \frac{N}{2^{j_2}}$$

Wavelet spectra for same numbers of wavelet coefficients





- Eigenvalues of $S^{(w)}(2^j)$: $\{\lambda_1^{(w)}(2^j), \dots, \lambda_M^{(w)}(2^j)\}$
  $\rightarrow$ similar repulsion at all scales $j \in \{j_1, \dots, j_2\}$
- Averaged log-eigenvalues: $\vartheta_m(2^j) \triangleq 2^{j_2-j} \sum_{w=1}^{2^{j-j_2}} \log_2(\lambda_m^{(w)}(2^j))$
- Linear regression on averaged log-eigenvalues $\vartheta_m(2^j)$

Adjusted Rand Index

2 partitions of $\mathcal{V} = \{1, \ldots, M\}$: $U = \{U_1, U_2, \ldots, U_R\}$, $V = \{V_1, V_2, \ldots, V_C\}$.

$$RI = (a+b)/\binom{M}{2}$$

- a: number of pairs of elements of $\mathcal{V}$ in the same subset in V and in the same subset in U
- b: number of pairs of elements of $\mathcal{V}$ in different subsets in V and in different subsets in U

$$ARI = \frac{index - expected\ index}{maximum\ index - expected\ index}$$

where $index = a + b$.

Normalized mutual information

2 partitions of $\mathcal{V} = \{1, \ldots, M\}$: $U = \{U_1, U_2, \ldots, U_R\}$, $V = \{V_1, V_2, \ldots, V_C\}$.

$$NMI = \frac{H(U) + H(V) - H(U, V)}{\sqrt{H(U)H(V)}}$$

where:

$$H(U) = -\sum_{i} q_{i,.} \log_2(q_{i,.})$$

$$H(V) = -\sum_{j} q_{.,j} \log_2(q_{.,j})$$

$$H(U, V) = -\sum_{i,j} q_{i,j} \log_2(q_{i,j})$$

with:

- $q_{i,j} = P(U_i \cap V_j)$ the proportion of elements both in U_i and V_j
- $q_{i,.} = P(U_i)$ the proportion of elements in U_i
- $q_{.,j} = P(V_j)$ the proportion of elements in V_j