Clustering self-similarity exponents of multivariate time series by a bootstrap in the wavelet domain

Journées du GDR AMA 2021, Porquerolles, France

Charles-Gérard Lucas, Patrice Abry, Herwig Wendt, Gustavo Didier

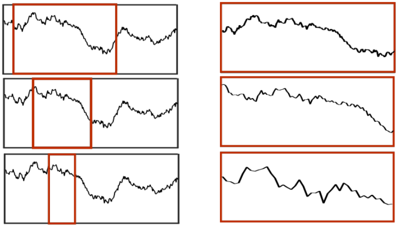
Laboratoire de Physique, ENS de Lyon

October 1, 2021

Gliederung

Univariate self-similarity

Scale-free dynamics



$$\{X(t)\}_{t\in\mathbb{R}} \overset{fdd}{=} \{a^H X(t/a)\}_{t\in\mathbb{R}}, \forall a > 0$$

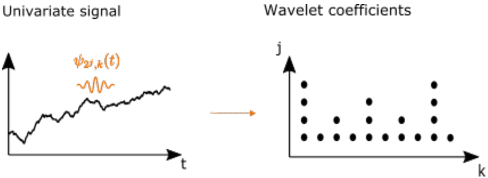
$$0 < H < 1$$

Goal: estimation of H

Univariate estimation of H (Flandrin et al., 1992)

Univariate wavelet transform:

- $D_X(2^j, k) = \langle 2^{-j/2}\psi_0(2^{-j}t - k) | X(t) \rangle$
- ψ_0: mother wavelet



Wavelet spectrum

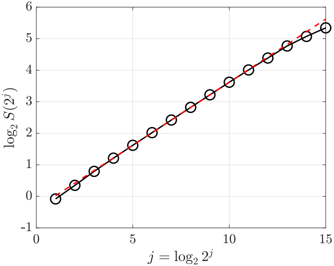
$$S(2^j) = \frac{1}{N_j} \sum_{k=1}^{N_j} D_X(2^j, k)^2 \in \mathbb{R}$$

$$N_j = \frac{N}{2^j},\ N\text{: sample size}$$

X self-similar

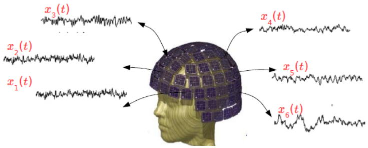
$\Rightarrow$ power law: $S(2^j) \propto 2^{j(2H-1)}$

Linear regression in a log-log diagram



Multivariate self-similarity

Collection of signals



Multivariate setting

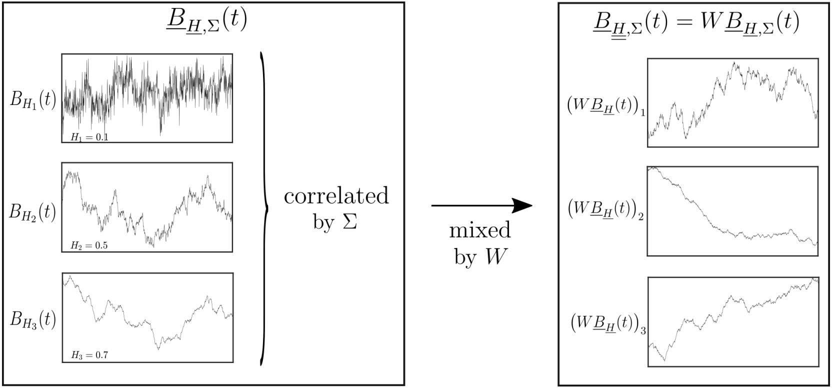
Multivariate self-similarity exponent

$$\underline{H} = (H_1, \ldots, H_M)$$

where $0 < H_1 \leq \ldots \leq H_M < 1$

Goal: estimating the groups of equal self-similarity exponents in $\underline{H}$

Multivariate self-similarity (Didier et al., 2011)



$$\{\underline{B}_{\underline{\underline{H}},\Sigma}(t)\}_{t\in\mathbb{R}} \overset{fdd}{=} \{a^{\underline{\underline{H}}}\underline{B}_{\underline{\underline{H}},\Sigma}(t/a)\}_{t\in\mathbb{R}}, \forall a > 0$$

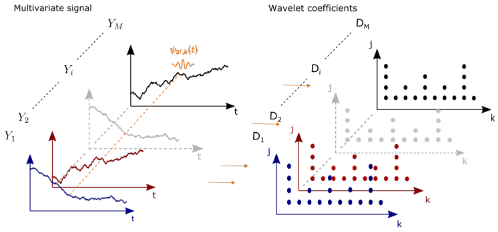
$$\underline{\underline{H}} = W\mathrm{diag}(\underline{H})W^{-1}$$

Goal: estimation of $\underline{H}$

Gliederung

Multivariate estimation

Multivariate wavelet transform of $Y = W\underline{B}_{H,\Sigma}$: $D(2^j,k) = (D_1(2^j,k),\ldots,D_M(2^j,k))$



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

$$S_{m_1,m_2}(2^j) = \frac{1}{N_j}\sum_{k=1}^{N_j} D_{m_1}(2^j,k)D_{m_2}(2^j,k)^*$$

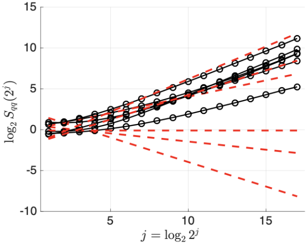
$N_j = \frac{N}{2^j}$, N: sample size

$Y = W\underline{B}_{H,\Sigma}$ self-similar

$\Rightarrow$ *mixture* of M^2 power laws when $W \neq I$:

$$S_{m_1,m_2}(2^j) = \sum_{k=1}^{M}\sum_{n=1}^{M} A_{k,n}^{(m_1,m_2)} 2^{j(H_k+H_n-1)}$$

Linear regression in a log-log diagram



Estimation of H (Didier and Abry, 2018)

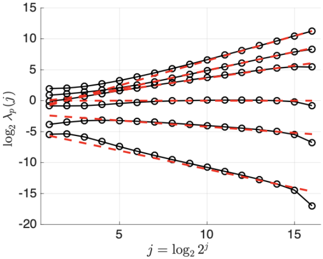
Eigenvalue decomposition:

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

$Y = W\underline{B}_{\underline{H},\Sigma}$ self-similar
$\Rightarrow$ Asymptotical power law:
$\lambda_m(2^j) \propto 2^{j(2H_m - 1)}$

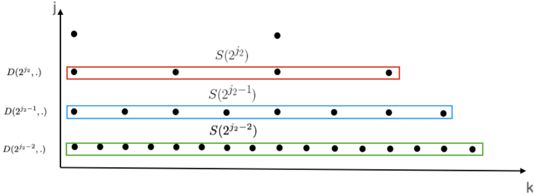
Linear regression on log-eigenvalues:

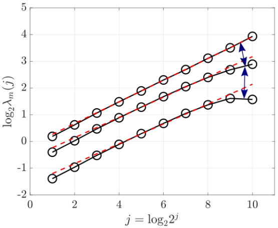
$$\hat{H}_m = \frac{1}{2} \sum_{j=j_1}^{j_2} \omega_j \log_2 \lambda_m(2^j) + \frac{1}{2}$$



Repulsion effect

Gap between eigenvalues larger than expected at each scale





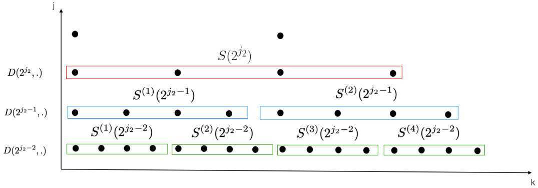
Few coefficients $\Rightarrow$ repulsion effect : important bias when $H_1 = \ldots = H_M$

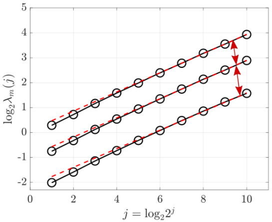
Issue: repulsion effect increases with scale 2^j

Bias corrected estimation

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

Wavelet spectra for same numbers of wavelet coefficients





Eigenvalues of $S^{(w)}(2^j)$: $\{\lambda_1^{(w)}(2^j),\ldots,\lambda_M^{(w)}(2^j)\}$
$\rightarrow$ similar repulsion at all scales $j \in \{j_1,\ldots,j_2\}$

Averaged log-eigenvalues: $\bar{\lambda}_m(2^j) \triangleq 2^{j_2-j} \sum_{w=1}^{2^{j-j_2}} \log_2(\lambda_m^{(w)}(2^j))$

Linear regression on averaged log-eigenvalues $\bar{\lambda}_m(2^j)$

Gliederung

Testing the equalities $H_m = H_{m+1}$

Single observation $\underline{H} = (H_1, \ldots, H_M)$

Fluctuation of the estimator: maybe $H_i = H_j$ despite $\hat{H}_i \neq \hat{H}_j$

$\Downarrow$

Tests for $H_m = H_{m+1} \Rightarrow$ clustering using the inequalities $H_1 < \ldots < H_M$

$$\boxed{\text{Testing } H_m = H_{m+1}}$$

Testing procedure on sorted estimates: $\hat{H}_{\tau(1)} < \ldots < \hat{H}_{\tau(M)}$

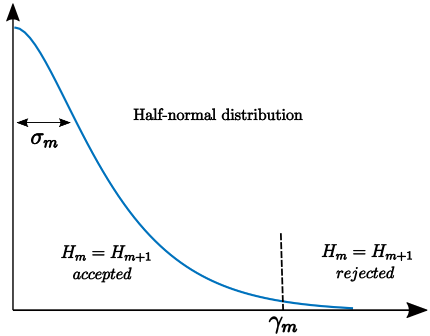
$M-1$ test statistics: $\tilde{\delta}_m = \hat{H}_{\tau(m+1)} - \hat{H}_{\tau(m)}$

Behavior of the test statistics $\tilde{\delta}_m$?

Test statistics $\tilde{\delta}_m$ under $H_m = H_{m+1}$

Half-normal distribution: $f(\tilde{\delta}_m | H_m = H_{m+1}) = \frac{\sqrt{2}}{\sigma_m \sqrt{\pi}} \exp\left(-\frac{\tilde{\delta}_m^2}{2\sigma_m^2}\right)$

σ_m: scale parameter



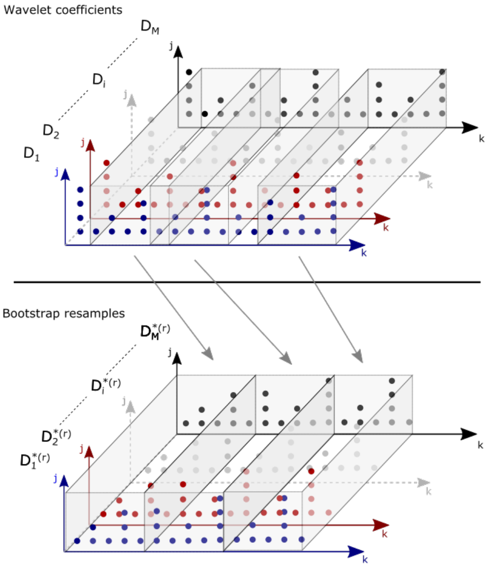
Test decision:

rejects $H_m = H_{m+1}$ if $\tilde{\delta}_m > \gamma_m$

γ_m: critical value to select
$\Rightarrow$ need for σ_m

Issue: single observation $\Rightarrow$ scale parameter σ_m unknown
$\longrightarrow$ estimation of σ_m by Bootstrap resampling

Multivariate wavelet block-bootstrap resamples



$\Rightarrow$ R wavelet coefficient resamples

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

$$\Downarrow$$

R Bootstrap estimates

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

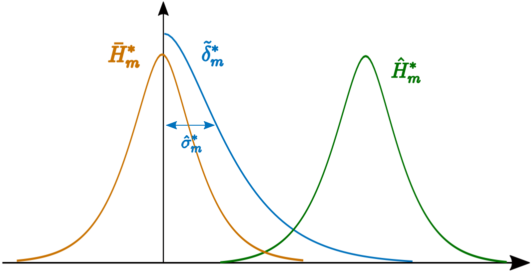
Bootstrap scale paramter estimate

Bootstrap test statistics reproducing null hypotheses

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

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

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



$$f(\tilde{\delta}_m^*) \approx f(\tilde{\delta}_m | H_m = H_{m+1}) \Rightarrow \boxed{\sigma_m^2 \approx \hat{\sigma}_m^{*2} = \mathrm{Var}^*(\tilde{\delta}_m^*) \left(1 - \frac{2}{\pi}\right)}$$

Test decisions

Test p-values:

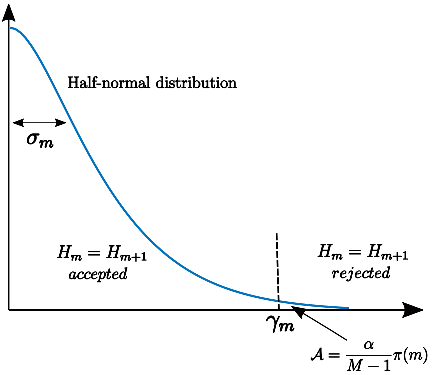
$$P(x > \tilde{\delta}_m | H_m = H_{m+1}) \approx 1 - F\left(\frac{\tilde{\delta}_m}{\hat{\sigma}^*_m}\right) \triangleq p^*_m$$

F: standardized half-normal cumulative distribution function

Benjamini-Hochberg decisions:

$$\text{rejects } H_m = H_{m+1} \text{ if } p^*_m < \frac{\alpha}{M-1}\pi(m)$$

$p^*_{\pi(m)}$: sorted p-values of the test
α: significance level



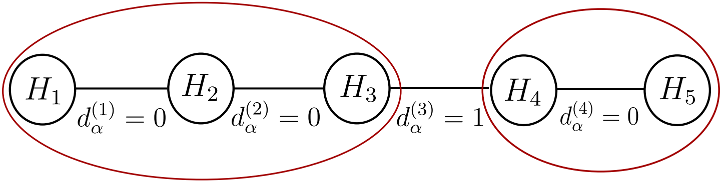
Gliederung

Clustering strategy

$$M - 1 \text{ binary decisions: } d_\alpha^{(m)} = 1 \Leftrightarrow H_m = H_{m+1} \text{ rejected}$$
$$\text{Inequalities } H_1 < \ldots < H_M$$

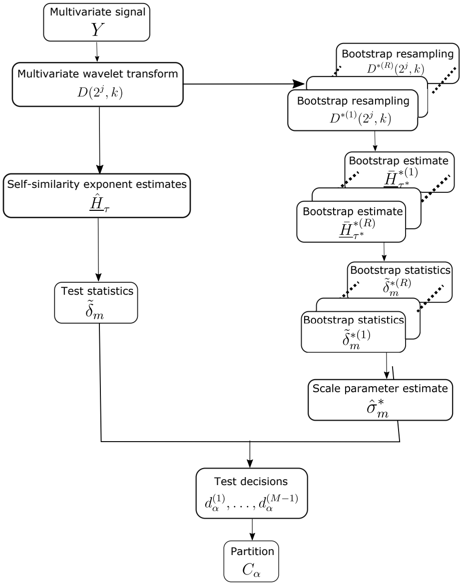
Natural clustering:

$$C_\alpha(m) = \sum_{m'=1}^{m} D_\alpha(m'), \quad D_\alpha = (1, d_\alpha^{(1)}, \ldots, d_\alpha^{(M-1)})$$



$$C_\alpha = [1\ 1\ 1\ 2\ 2]$$

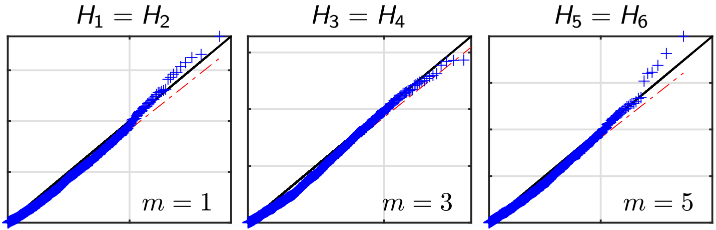
Clustering procedure



Gliederung

Half-normal distribution of $\tilde{\delta}_m$

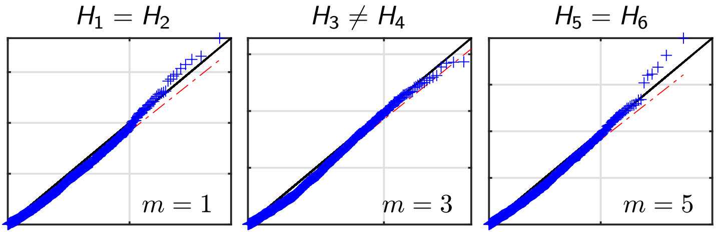
Monte Carlo simulations
$N_{MC} = 1000$ realizations, $M = 6$ components, sample size $N = 2^{16}$



Quantile-quantile plot under $H_1 = \ldots = H_6 = 0.8$
Monte Carlo $\tilde{\delta}_m$ against half-normal distribution

$\Rightarrow$ Confirms the half-normal distribution of $\tilde{\delta}_m$ under $H_m = H_{m+1}$

Bootstrap procedure assessment



Quantile-quantile-plots under $\underline{H} = (0.6, 0.6, 0.6, 0.8, 0.8, 0.8)$
Bootstrap $\tilde{\delta}^*_m$ against half-normal distribution

$\Rightarrow$ Bootstrap statistics $\tilde{\delta}^*_m$ well approximate the null distribution of $\tilde{\delta}_m$:
$f(\tilde{\delta}^*_m) \approx f(\tilde{\delta}_m | H_m = H_{m+1})$ for any hypothesis

Bootstrap scale parameter assessment

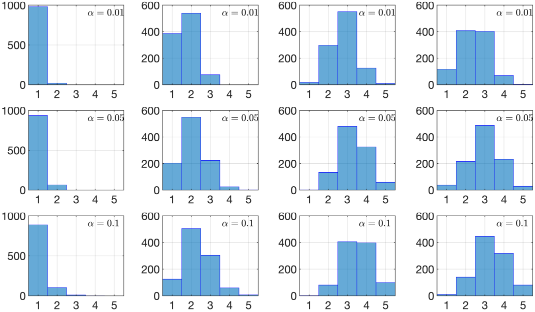
Scale parameters σ_m and bootstrap scale parameter estimates $\hat{\sigma}_m^*$
(Monte Carlo average and standard deviation)
for $H_1 = \ldots = H_M = 0.8$

	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
$\sigma_m \times 10^2$	1.65	1.16	1.01	1.06	1.50
$\hat{\sigma}_m^* \times 10^2$	1.65 ± 0.13	1.10 ± 0.07	0.99 ± 0.06	1.07 ± 0.06	1.51 ± 0.09

$\Rightarrow$ Bootstrap estimates $\hat{\sigma}_m^*$ well approximates the scale parameters σ_m

Clustering performance

$$\underline{H} = (\underbrace{H_1, \ldots, H_1}_{M_1}, \underbrace{H_2, \ldots, H_2}_{M_2}, \underbrace{H_3, \ldots, H_3}_{M_3})$$



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

Histograms of the estimated numbers of clusters for several α

Scenario1: $(M_1, M_2, M_3) = (1, 0, 0)$, $H_1 = 0.8$
Scenario2: $(M_1, M_2, M_3) = (3, 3, 0)$, $(H_1, H_2) = (0.6, 0.8)$
Scenario3: $(M_1, M_2, M_3) = (2, 2, 2)$, $(H_1, H_2, H_3) = (0.4, 0.6, 0.8)$
Scenario4: $(M_1, M_2, M_3) = (1, 3, 2)$, $(H_1, H_2, H_3) = (0.4, 0.6, 0.8)$

Clustering performance

$$\underline{H} = (\underbrace{H_1, \ldots, H_1}_{M_1}, \underbrace{H_2, \ldots, H_2}_{M_2}, \underbrace{H_3, \ldots, H_3}_{M_3})$$

NMI: joint entropy of ground truth partition and estimated partition
ARI: pairs of elements correctly separated or correctly gathered

Clustering performance with 95% confidence interval
for significance level $\alpha = 0.05$.

	Scenario2	Scenario3	Scenario4
NMI	0.66 ± 0.02	0.87 ± 0.01	0.79 ± 0.01
ARI	0.60 ± 0.03	0.68 ± 0.02	0.59 ± 0.02

Scenario2: $(M_1, M_2, M_3) = (3, 3, 0)$, $(H_1, H_2) = (0.6, 0.8)$
Scenario3: $(M_1, M_2, M_3) = (2, 2, 2)$, $(H_1, H_2, H_3) = (0.4, 0.6, 0.8)$
Scenario4: $(M_1, M_2, M_3) = (1, 3, 2)$, $(H_1, H_2, H_3) = (0.4, 0.6, 0.8)$

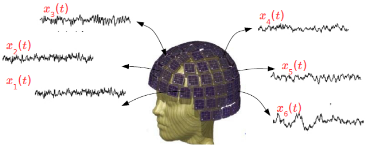
Conclusion

Achieved:

- Bias corrected estimation of multivariate self-similarity exponents
- Multivariate wavelet domain bootstrap procedure
- Clustering of self-similarity exponents from a single observation

Perspectives:

- Can we build a clustering strategy based on $M(M-1)/2$ tests for $H_i = H_j$, $i, j = 1, \ldots, M$?
- Large dimension: number of components M $\approx$ sample size N



Thank you for your attention !

Alternative hypotheses