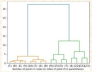
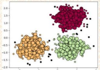
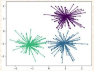
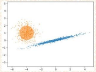
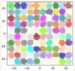
Kullback-Leibler cluster entropy: An inferential tool for long-range correlated data

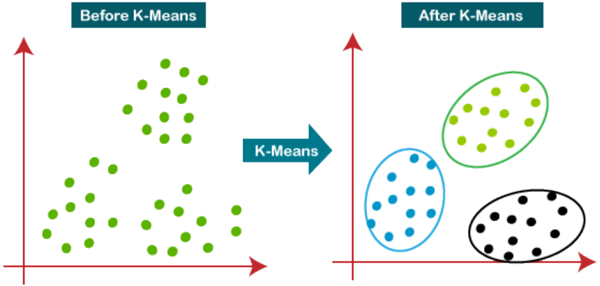
Anna Carbone
Politecnico di Torino
https://www.polito.it/en/staff?p=anna.carbone

21-25 October 2024
TED4LAT Workshop and PhD School
Institut des Systèmes Complexes
Île-de-France
113 rue Nationale Paris

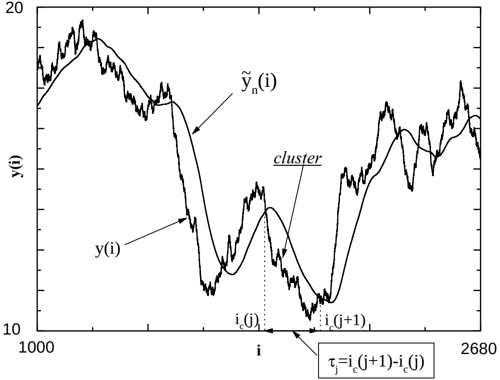
Unsupervised Learning Artificial Intelligence: Clustering

Type of Clustering Algorithm	Visual Overview	Description	Algorithm(s)
Centroid-based		Cluster points based on proximity to centroid	KMeans KMeans++ KMedoids
Connectivity-based		Cluster points based on proximity between clusters	Hierarchical Clustering (Agglomerative and Divisive)
Density-based		Cluster points based on their density instead of proximity	DBSCAN OPTICS HDBSCAN
Graph-based		Cluster points based on graph distance	Affinitiy Propagation Spectral Clustering
Distribution-based		Cluster points based on their likelihood of belonging to the same distribution.	Gaussian Mixture Models
Compression-based		Transform data to a lower dimensional space and then perform clustering	BIRCH

K-mean clustering



Power-law distribution of cluster features



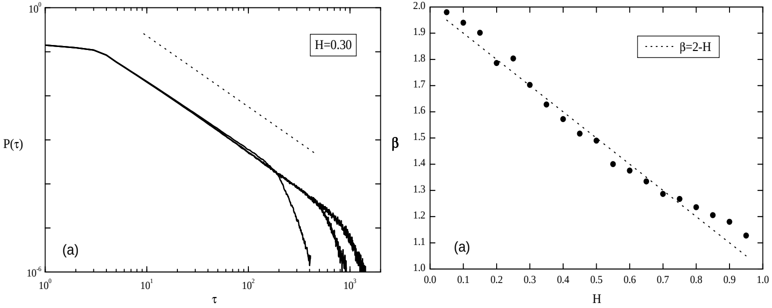
By ranking the number of clusters $\mathcal{N}(\tau_1, n), \ldots \mathcal{N}(\tau_j, n)$ according to the duration $\tau_1, \tau_2, \ldots, \tau_j$ for each n, one has:

$$P(\tau_j, n) = \frac{\mathcal{N}(\tau_j, n)}{\mathcal{N}_C(n)}$$

with $\mathcal{N}_C(n) = \sum_{j=1}^{k(n)} \mathcal{N}(\tau_j, n)$ the total number of clusters

$$\sum_{n=1}^{N} \sum_{j=1}^{\mathcal{N}_C(n)} P(\tau_j, n) = 1 \ .$$

Power-law distribution of cluster features

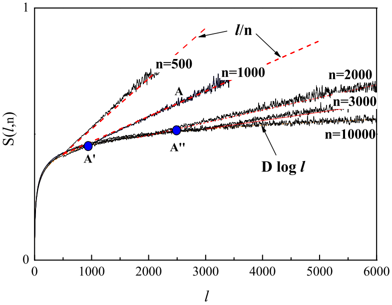


$$P(\tau, n) \sim \tau^{-\alpha} \mathcal{F}(\tau, n)$$

$$\alpha = D = 2 - H$$

$$\mathcal{F}(\tau, n) = e^{-\tau/\tau^*}.$$

The Shannon entropy of the long-range correlated sequence is estimated by counting the clusters at different n corresponding to not-overlapping partitions of the sequences.



$$S(\ell, n) \equiv - \sum_{\mu(\ell,n)} P(\ell, n) \log P(\ell, n).$$

$$S(\ell, n) \sim S_0 + D \log \ell + \frac{\ell}{n} \ .$$

The entropy is the sum of two terms corresponding to power-law (*ordered*) and exponentially (*disordered*) distributed clusters.

[1] Sci. Rep. (2013), Physica A (2018), EPL (2021), SciPost (2022)

Information Theoretical Measures: Shannon Entropy

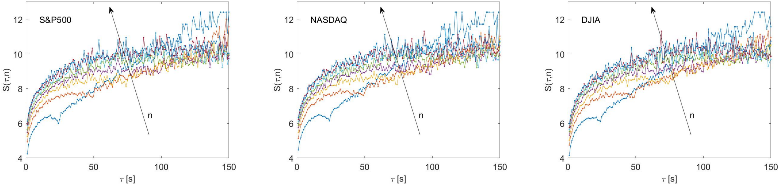
The *source entropy rate* s is defined for the entropy $S(\ell, n)$:

$$s \equiv \lim_{\ell \to \infty} \frac{S(\ell, n)}{\ell} = \frac{1}{n} \quad . \tag{1}$$

The source entropy rate is a measure of the *excess randomness* and increases as the block coding process becomes noisier. The excess randomness of the clusters is found to be inversely proportional to n and, thus, becomes negligible in the limit of $n \to \infty$. One can note that higher entropy rates correspond to steeper slopes of the linear term ℓ/n (smaller n values).

The interest and meaning of the ℓ/n terms will be illustrated for genomic series.

Information Theoretical Measures: Market Dynamics [2]



Cluster entropy $S(\tau, n)$ for the probability distribution function of the volatility series of the linear return of tick-by-tick data of the S&P500, NASDAQ, DJIA. The results refer to the horizon $\mathcal{M} = 1$, i.e twelve monthly periods sampled out of the year 2018. The different plots refer to different values of the moving average window n (here n ranges from $25s$ to $200s$ with step $25s$).

[2] *Physica A (2018) , Entropy (2020)*

Information Theoretical Measures: Multiperiod Portfolio

The cluster entropy index $I_i(n)$ is estimated as:

$$I_i(n) = \sum_{\tau=1}^{m} S_i(\tau, n) + \sum_{\tau=m}^{N} S_i(\tau, n) \quad .$$

The average index I_i is calculated over the set of n values :

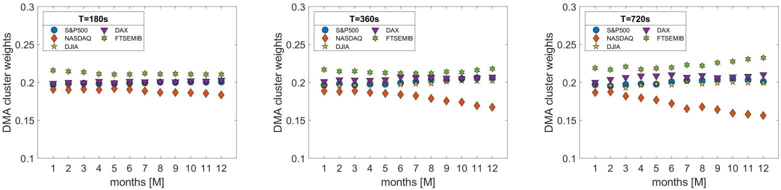
$$I_i = \sum_{n} I_i(n) \quad .$$

The portfolio weights are defined as follows:

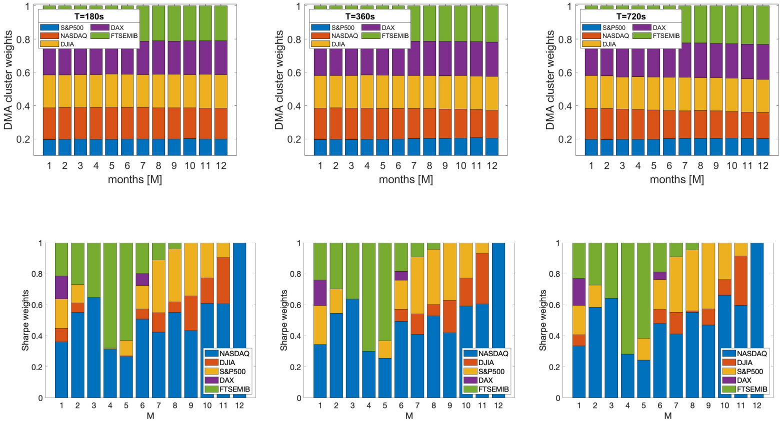
$$w_{i,\mathcal{C}} = \frac{I_i}{\sum_{i=1}^{\mathcal{N}_{\mathcal{A}}} I_i} \quad ,$$

with the condition $\sum_{i=1}^{\mathcal{N}_{\mathcal{A}}} w_{i,\mathcal{C}} = 1$.

Information Theoretical Measures: Multiperiod Portfolio



Information Theoretical Measures: Multiperiod Portfolio



Information Theoretical Measures: Model Inference [3]

The *relative cluster entropy* or *cluster divergence* $\mathcal{D}_{\mathcal{C}}[P\|Q]$ can be defined to compare two probability distributions P and Q, with $\mathcal{D}_{\mathcal{C}}[P\|Q] = 0$ for $P = Q$ with the condition $\mathrm{supp}(P) \subseteq \mathrm{supp}(Q)$:

$$\mathcal{D}_{\mathcal{C}}[P||Q] = \sum_{n=1}^{N} \sum_{j=1}^{\mathcal{N}_C(n)} P(\tau_j, n) \log \frac{P(\tau_j, n)}{Q(\tau_j, n)} \quad .$$

where the quantity $\mathcal{D}_{j,n}[P||Q]$ is estimated in terms of the τ_j and n as follows:

$$\mathcal{D}_{j,n}[P||Q] = P(\tau_j, n) \log \frac{P(\tau_j, n)}{Q(\tau_j, n)} \quad ,$$

where the index j refers to the set of clusters with duration τ_j generated by the partition for a given n and $P(\tau_j, n)$ and $Q(\tau_j, n)$ are their frequencies.

[3] *SciPost (2022)*

By using contiunuous variables and power-law probability distribution functions in the form

$$P(\tau) \sim \tau^{-\alpha_1} \qquad Q(\tau) \sim \tau^{-\alpha_2}$$

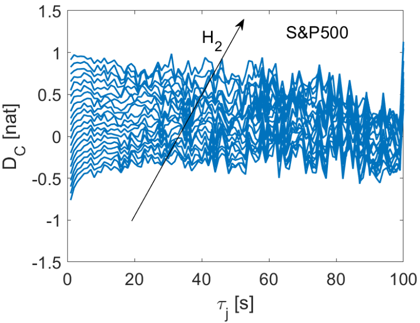
where α_1 and α_2 are the correlation exponents. By using $\alpha_1 = 2 - H_1$ and $\alpha_2 = 2 - H_2$, the relative cluster entropy can be calculated as a definite integral over the interval $[1, \infty]$:

$$D_C[P||Q] = \log \frac{1 - H_1}{1 - H_2} + \frac{H_1 - H_2}{1 - H_1} \ . \qquad (2)$$

$D_C[P||Q] \geq 0$ over the whole range of H_1 and H_2 values thus satisfying the property of the relative entropy to be positive defined. $D_C[P||Q] = 0$ for $H_1 = H_2$ Interestingly, $D_C[P||Q]$ turns out to be a function only of H_1 and H_2

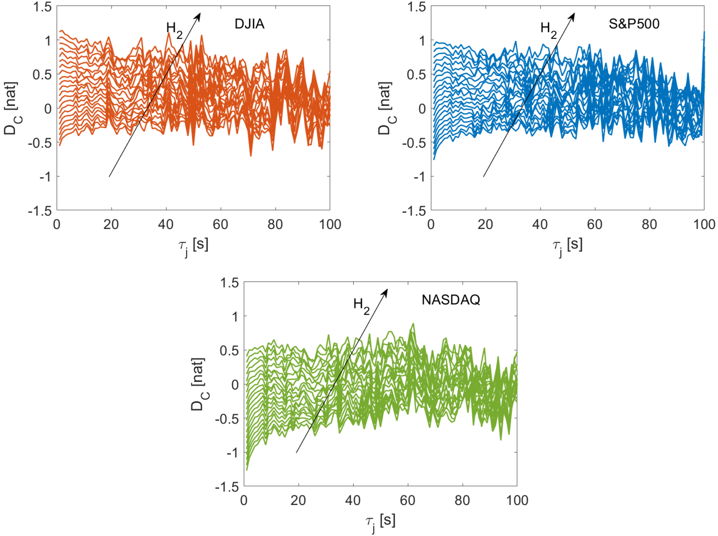
$$D_C[P||Q] = D_C[H_1||H_2]$$

Information Theoretical Measures: Model Inference

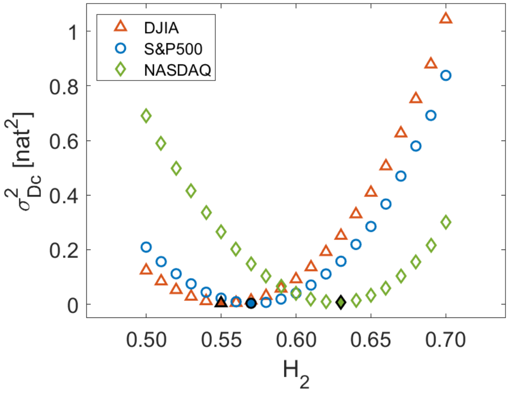


$\mathcal{D}_{j,n}[P||Q]$, summed over the parameter n for $P(\tau_j, n)$ estimated on the S&P500 market price p_t and the model probability $Q(\tau_j, n)$ on *fBms* with Hurst exponent H_2 ranging from 0.50 to 0.70 with step 0.1.

Information Theoretical Measures: Model Inference



Information Theoretical Measures: Model Inference



Thanks for your attention.