Physica A 570 (2021) 125777

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Information measure for long-range correlated time series: Quantifying horizon dependence in financial markets

Linda Ponta [a], Pietro Murialdo [b], Anna Carbone [b,*]

[a] Universitá Cattaneo LIUC, Castellanza, Italy
[b] Politecnico di Torino, corso Duca degli Abruzzi 24, 10129 Torino, Italy

ARTICLE INFO

Article history:
Received 27 June 2020
Received in revised form 30 October 2020
Available online 4 February 2021

Dataset link: www.bloomberg.com/professional, https://project.inria.fr/fraclab/

Keywords:
Complex systems
Information Measures
Long-range dependence
Financial Markets

ABSTRACT

Market dynamics is quantified via the cluster entropy $S(\tau, n) = \sum_j P_j(\tau, n) \log P_j(\tau, n)$, an information measure with $P_j(\tau, n)$ the probability for the clusters, defined by the intersection between the price series and its moving average with window n, to occur with duration τ. The cluster entropy $S(\tau, n)$ is estimated over a broad range of temporal horizons M, for raw and sampled highest-frequency data of US markets. A systematic dependence of $S(\tau, n)$ on M emerges in agreement with price dynamics and correlation involving short and long range horizon dependence over multiple temporal scales. A comparison with the price dynamics based on Kullback–Leibler entropy simulations with different representative agent models is also reported.

1. Introduction

Entropy-based methodologies have demonstrated the ability to quantify *heterogeneity* and *dynamics* of complex systems [1–3], thus, have found several relevant applications in interdisciplinary contexts as biology, economics and finance. In particular, the entropy ability to quantify *heterogeneity* and *dynamics* has been exploited for portfolio selection, as an alternative to traditional methods based on Markowitz covariance and Sharpe single-index models [4–11] and for market evolution models in terms of stochastic functions [12–19].

Equilibrium prices p_t of traded securities can be represented as the conditional expectation of the discounted future payoff z_t:

$$p_t = E\left[\frac{m_{t+1}}{m_t} z_{t+1}\right] \quad , \tag{1}$$

where m_t is the pricing kernel and m_{t+1}/m_t is the *stochastic discount factor*. The pricing kernel m_t is factorizable into a function of the consumption growth μ_{t+1} times a model specific term ψ_t:

$$m_t = \mu_{t+1}\psi_t \quad . \tag{2}$$

The simple consumption-based asset pricing model identifies the kernel as a parametric function of the consumption growth C_t. In the framework of time-separable power utility representative agent models, the function μ_{t+1} is simply proportional to $\Delta C_t = \log(C_t/C_{t-1})$. More sophisticated agent behaviours have been suggested to explain puzzling phenomena such as amplitude and cross-sectional dispersion of returns among different categories of financial assets,

* Corresponding author.
E-mail address: anna.carbone@polito.it (A. Carbone).

https://doi.org/10.1016/j.physa.2021.125777
0378-4371/

equity premia and risk-free rates. Pricing kernel dispersion and dynamics with different representative agents have been modelled by using the Kullback–Leibler entropy in [16], thus extending the work [13] aimed at quantifying standard deviation and volatility to yield the pricing kernels bounds. A relative entropy minimization approach is put forward in [17] to extract the model dependent term μ_{t+1} and quantify the minimum amount of extra information to be embedded in the standard pricing kernel models for reproducing asset returns correctly. The Kullback–Leibler divergence between the probability distribution functions of the components μ_{t+1} and ψ_t has been used as criterion to estimate the deviation of m_{t+1} with respect to the simple consumption flow growth model [17].

An information theoretical measure, the *moving average cluster entropy*, has been proposed in [20,21] and applied to quantify heterogeneity of human chromosomes [22] and of financial markets [23,24]. The physical interpretation of the cluster entropy is related to the theory of irreversible processes and the concept of local equilibrium. By assuming that the system exists in a state of equilibrium in every elemental volume surrounding an arbitrary point, the state functionals (e.g. entropy) are described by state variables depending on space and time. Additionally, space and time can be generalized and described by fractional rather than integer variables [25–28]. The entropy change dS in a time interval dt is given by:

$$dS = dS_{int} + dS_{ext} \tag{3}$$

with dS_{int} and dS_{ext} the entropy change produced respectively by endogenous processes and by exchanges of energy and matter with system's exterior.

Ref. [23] shows that the cluster entropy of the volatility series, estimated over a constant temporal horizon, takes values depending on each market, as opposed to the cluster entropy of the price series, approximately invariant across markets over the same horizon. These findings led to develop the *Market Heterogeneity Index*, a tool able to estimate the portfolio weights over a constant time horizon. The *Market Heterogeneity Index*, defined as the integral of the cluster entropy function, provides a cumulative figure allowing a straightforward comparison with the portfolio weights obtained by the Sharpe ratio approach. A key advantage of the cluster entropy approach is to not require a symmetric Gaussian distribution of returns, which is quite elusive in real-world financial assets and thus hinders, in principle, the application of Markowitz-based portfolio models. Then, the study [24] was addressed to analyse the behaviour of the cluster entropy with artificial models of financial markets such as Generalized Autoregressive Conditional Heteroskedastic (GARCH), Autoregressive Fractionally Integrated Moving Average (ARFIMA), Fractional Brownian Motion (FBM) etc.

The present study builds upon the results of [23,24] and extends the cluster entropy method to quantify market price heterogeneity and dynamics over different temporal horizons in real world assets. The approach is implemented on tick-by-tick prices of NASDAQ, DJIA and S&P500 from Jan 1 to Dec 31 2018 with length $N = 6982017$, $N = 5749145$ and $N = 6142443$ respectively, downloaded from www.bloomberg.com/professional. Further details are provided in Table 1. The three assets have been selected for this investigation based on homogeneity criteria, being traded in the same country, with same currency and comparable number of transactions over time. The asset similarity rules out that differences in the price evolution might be due to exogenous causes. The extent of the investigated horizons is another feature ensuring that the observed behaviour is genuinely related to the intrinsic price dynamics rather than exogenous factors. Hence, the maximum range of the investigated horizons is taken equal to one year and the cluster entropy analysis has been performed on multiple sets of raw and sampled data from one to twelve monthly horizons M. A systematic dependence of the cluster entropy of the asset prices on varying temporal horizons has been observed. Such a dependence could be related to the intrinsic dynamics rather than cross market variations.

To further substantiate the results, a direct comparison is made between the horizon dependence obtained respectively by the moving average cluster entropy and the Kullback–Leibler relative entropy with representative agent models of price evolution.

The manuscript is organized as follows. The main relationships relevant to the implementation of the *Cluster Entropy* approach are recalled in Section 2. The analysed data sets, financial assets and artificially generated series, are described in Section 3. *Cluster Entropy* and *Market Dynamic Index* of the price series as a function of the temporal horizon M are reported in Section 4. Artificially generated Fractional Brownian paths are used as reference to validate the deviations observed in real-world markets. Statistical significance via standard T-paired test is reported. A comparison against the Kullback–Leibler entropy obtained by simulating the pricing kernel with different representative agent models and other concluding remarks are reported in Section 5.

2. Method

The main definitions underlying the cluster entropy approach [20–22] are briefly recalled henceforth. The method builds upon the idea of Claude Shannon to quantify the 'expected' information contained in a message extracted from a sequence $\{x_t\}$ with a probability distribution $P(x_t)$ by using the entropy functional [29]:

$$S[P(x_t)] = \int_X P(x_t) \log P(x_t) dx_t \quad , \tag{4}$$

that for discrete sets writes:

$$S[P(x_t)] = \sum_X P(x_t) \log P(x_t) \quad . \tag{5}$$

Consider the time series $\{x_t\}$ of length N and the moving average $\{\widetilde{x}_{t,n}\}$ of length $N-n$ with n the moving average window. The time sequence $x(t)$ is partitioned in *clusters* by the intersection with its moving average $\tilde{x}_n(t)$. The simple moving average is defined at each t as the average of the n past observations from t to $t-n+1$,

$$\tilde{x}_n(t) = \frac{1}{n}\sum_{i=1}^{n} x(t-i). \tag{6}$$

Note that while the original series is defined from 1 to N, the moving average series is defined from 1 to $N-n$ because n samples are necessary to initialize the series. Consecutive intersections between the time series and the moving average series yield a partition into a series of *clusters*. Each cluster is defined as the portion of the time series $x(t)$ between two consecutive intersection of $x(t)$ itself and its moving average $\tilde{x}_n(t)$.

The Eq. (6) is a smoothing function of the time series, which can be interpreted as a linear regression: $x(t) = \widetilde{x}_n(t) + \epsilon_n$ where ϵ_n is the error in the estimate with expected value equal to 0 [30,31]. The function $\widetilde{x}_n(t)$ is obtained by locally averaging the $x(t)$ somehow close to t as:

$$\widetilde{x}_n(t) = \sum_{i=1}^{n} W_i(t) x_i(t) \tag{7}$$

where $W_i(t)$ are called *weights* and decrease as t_i is far from t. Such estimators require a finite or countably infinite partition $\mathcal{C}_{n,j} = \left\{C_{n,1}, C_{n,2}, \ldots\right\}$ over the time series domain to estimate $\widetilde{x}_n(t)$ by averaging $x(t)$'s with the corresponding t's in $\mathcal{C}_{n,j}$, i.e.:

$$\widetilde{x}_n(t) = \sum_{i=1}^{n} \frac{I_{\{t_i \in \mathcal{C}_{n,j}\}}}{\sum_{i=1}^{n} I_{\{t_i \in \mathcal{C}_{n,j}\}}} x_i(t) \tag{8}$$

with $t_i \in \mathcal{C}_{n,j}$ and I a relevant function defined over the set $\mathcal{C}_{n,j}$, thus:

$$W_i(t) = \frac{I_{\{t_i \in \mathcal{C}_{n,j}\}}}{\sum_{l=1}^{n} I_{\{t_l \in \mathcal{C}_{n,j}\}}} \quad . \tag{9}$$

The simple moving average is an example of a broad class of linear estimators such as the *naive kernel* or *window kernel* approach with $I\{\|t\| \leq h\}$, i.e. $\widetilde{x}_n(t)$ is obtained by averaging $x(t)$'s such that the distance between t_i and t is not greater than h and the *k-nearest neighbour approach*, with the weight $W_{n,i}(t)$ equals $1/k$ if t_i is among the k nearest neighbours of t, and equals 0 otherwise. In general, one uses a weighted average of the $x(t)$ where the weights (i.e., the influence of $x(t)$ on the value of the estimate at t) depend on the distance between t_i and t. An extensive investigation and further generalizations to higher order moving average polynomials and trend estimators can be found in [30,31].

The function $\{\widetilde{x}_{t,n}\}$ generates a partition $\{\mathcal{C}_{n,j}\}$ of non-overlapping clusters between two consecutive intersections of $\{x_t\}$ and $\{\widetilde{x}_{t,n}\}$ for each n. Each cluster j has duration:

$$\tau_j \equiv \|t_j - t_{j-1}\| \tag{10}$$

where the instances t_{j-1} and t_j refer to two consecutive intersections.

The probability distribution function $P(\tau, n)$ can be obtained by ranking the number of clusters $\mathcal{N}(\tau_1, n), \mathcal{N}(\tau_2, n), \ldots, \mathcal{N}(\tau_j, n)$ according to their length $\tau_1, \tau_2, \ldots, \tau_j$ for each n. A stationary sequence of clusters $\mathcal{C}$ is generated with probability distribution function varying as [22]:

$$P(\tau, n) \sim \tau^{-\alpha} \mathcal{F}(\tau, n) \quad . \tag{11}$$

The factor $\mathcal{F}(\tau, n)$, taking the form $\exp(-\tau/n)$, accounts for the finite size effects when $\tau \gg n$, resulting in the drop-off of the power-law and the onset of the exponential decay. The cluster entropy writes (the details of the derivation can be found in [20,22]):

$$S[P(\tau_j, n)] = \sum_j P(\tau_j, n) \log P(\tau_j, n) \quad , \tag{12}$$

that, by using Eq. (11), simplifies to:

$$S(\tau, n) = S_0 + \log \tau^{\alpha} + \frac{\tau}{n} \quad , \tag{13}$$

where S_0 is a constant, $\log \tau^{\alpha}$ and τ/n are related respectively to the terms $\tau^{-\alpha}$ and $\mathcal{F}(\tau, n)$.

To clarify the meaning of the terms appearing in Eq. (13), it is worthy of remarking that, for isolated systems, the entropy increase dS is related to the irreversible processes spontaneously occurring within the system (as mentioned in the Introduction and Eq. (3)). The entropy tends to a constant value as a stationary state is asymptotically reached ($dS \geq 0$). For open systems interacting with the environment, the increase is given by a term dS_{int} due to the irreversible processes spontaneously occurring within the system, and a term dS_{ext} due to the irreversible processes arising from external

interactions. The logarithmic term in Eq. (13) is related to the intrinsic entropy change dS_{int}. It is indeed independent of n, i.e. of the method used for partitioning the sequence, which plays here the role of the external interaction. The logarithmic term is of the form of a Boltzmann entropy $S = \log \Omega$, where Ω is the maximum volume occupied by the isolated system. The quantity τ^D is proportional to the volume occupied by the random walker. Whenever τ could reach the maximum size N of the sequence, the second term on the right side would write $\log N^D$. The term τ/n represents the excess entropy S_{ext} (excess noise) added by the partition process. It comes into play when the sequence is partitioned in clusters, thus it depends on n.

To gain further insight in the meaning of the entropy $S(\tau, n)$, the source entropy rate s_∞ is calculated for Eq. (13). The source entropy rate is a measure of the excess randomness and increases as the cluster coding process becomes noisier. By using the definition and Eq. (5), the source entropy rate writes:

$$s_\infty \equiv \lim_{\tau\to\infty} \frac{S(\tau, n)}{\tau} = \frac{1}{n} \quad . \tag{14}$$

The excess randomness of the clusters is found to be inversely proportional to n and, thus, becomes negligible in the limit of $n \to \infty$. Such a behaviour is clearly related to disorder increase with increasing cluster lengths at constant n.

The minimum value of the entropy $S(\tau, n) = 0$ is obtained for the fully ordered (deterministic) set of clusters with duration $\tau = 1$ from Eq. (13) in the limit $n \sim \tau \to 1$ with $S_0 \to -1$. Conversely, the maximum value of the entropy $S(\tau, n) = \log N^\alpha$ is obtained when $n \sim \tau \to N$ (with N the maximum length of the sequence). This condition corresponds to the maximum randomness (minimum information) carried by the sequence, when the longest cluster, coinciding with the whole series, is obtained.

For Fractional Brownian Motions, the exponent α is equal to the fractal dimension $D = 2-H$ with H the Hurst exponent of the time series. The term $\log \tau^\alpha$ can be thus interpreted as a generalized form of the Boltzmann entropy $S = \log \Omega$, where $\Omega = \tau^D$ corresponds to the fractional volume occupied by the fractional random walker. The term τ/n represents an excess entropy (excess noise) added to the intrinsic entropy term $\log \tau^D$ by the partition process. It depends on n and is related to the finite size effect discussed above.

Moreover, we stress the difference between the time series partitions obtained either by using equal size boxes or moving average clusters. For equal size boxes, the excess noise term τ/n becomes a constant and can be included in the constant term, thus the entropy reduces to the logarithmic term, which corresponds to the intrinsic block entropy of an ideal fractional random walk [3]. When a moving average partition is used, an excess entropy term τ/n emerges from the spreading introduced in the probability distribution function by the random partitioning process operated by the moving average intersections.

To summarize entropy properties in a single figure, a cumulative information measure has been defined as follows:

$$I(n) = \int_1^{\tau_{max}} S(\tau, n) d\tau \quad , \tag{15}$$

which, for discrete sets, reduces to:

$$I(n) = \sum_{\tau=1}^{\tau_{max}} S(\tau, n) \quad . \tag{16}$$

In Section 4, Eq. (16) will be estimated over real-world and artificial time series at varying time horizons M. The measure will be indicated by $I(M, n)$ and referred to as *Market Dynamic Index*. Furthermore, the *Horizon Dependence* $H(M, n)$ can be defined as the variation of $I(M, n)$:

$$H(M, n) = I(M, n) - I(1, n) \quad , \tag{17}$$

referred to the first horizon ($M = 1$).

Values of $H(M, n)$ obtained by using Eq. (17) based on the cluster entropy will be compared with the values obtained on different representative agent models of the pricing kernel dynamics by a measure of relative entropy. The pricing kernel accounts for the stochastic dynamic evolution of asset returns, which in their turn contain information about the pricing kernel. The analysis is based on the *Kullback–Leibler* entropy of the actual probability distribution of the prices with respect to the risk-adjusted probability. It has been argued that a realistic asset pricing model should have substantial one-period entropy and modest horizon dependence to justify equity mean excess returns and bond yields at once [16]. The *Kullback–Leibler* entropy of the continuous probability measure $P(x_t)$ with respect to some probability measure $P^*(x_t)$, writes:

$$K(P\|P^*) = \int_X P(x_t) \log \left\{ \frac{P(x_t)}{P^*(x_t)} \right\} dx_t \quad . \tag{18}$$

Eq. (18) can be interpreted as the expectation of the function $\log\{P(x_t)/P^*(x_t)\}$ with respect to the probability $P(x_t)$:

$$K(P\|P^*) = E\left[\log \left\{ \frac{P(x_t)}{P^*(x_t)} \right\} \right] \quad . \tag{19}$$

The relative entropy given by Eq. (18) reduces to Eq. (4) for constant probability $P^*(x_t)$.

Table 1
Asset description. Details of NASDAQ, S&P500 and DJIA data sets as downloaded from the Bloomberg terminal. Tick duration (time interval between individual transactions) is about one second for the three markets.

Ticker	Name	Country	Currency	Members	Length
NASDAQ	Nasdaq Composite	US	USD	2570	6982017
S&P500	Standard & Poor 500	US	USD	505	6142443
DJIA	Dow Jones Industrial Average	US	USD	30	5749145

Table 2
Exploratory data analysis. Mean (μ), standard deviation (σ), skewness (s), kurtosis (k) and Hurst exponent $\hat{H}$ estimated over linear return series of the three financial indexes. The Hurst Exponent is obtained by mean of the Detrending Moving Average algorithm [30,31].

Ticker	$\mu \cdot 10^{-3}$	σ	s	$k \cdot 10^{3}$	$\hat{H}$
NASDAQ	−0.04	0.37	−10.16	58.23	0.53
S&P500	−0.03	0.12	−7.72	17.92	0.52
DJIA	−0.26	1.23	−4.66	8.83	0.51

It is worth mentioning the relation between the cluster entropy approach adopted in this work, the multiscale entropy (MSE) and its variants [32–34]. The multiscale entropy provides insights into the complexity of fluctuations over a range of time scales and thus extends the standard one-sample entropy. The computational implementation of the multiscale entropy implies a coarse graining of the time series at increasing time resolutions. Coarse graining the data basically means averaging different numbers of consecutive points to create different scales or resolutions of the signal. In the cluster entropy approach, the coarse graining of the signal is performed through the moving average, an average at increasing time resolutions. The multiscale entropy analysis aims at quantifying the interdependence between entropy and scale, achieved by evaluating sample entropy of univariate time series coarse grained at multiple temporal scales, thus enabling the assessment of the dynamical complexity of the system.

3. Data

Prices p_t and returns r_t of market indices traded in the US, namely NASDAQ, S&P500 and DJIA are investigated in this manuscript. In this section we provide a few general definitions and properties of relevance to the data analysis performed in Section 4. Linear returns can be defined by:

$$r_t = p_t - p_{t-h} \quad , \tag{20}$$

or by:

$$r_t = \frac{p_t}{p_{t-h}} \quad , \tag{21}$$

where p_t is the price at time t, with $0 < h < t < N$ and N the maximum length of the time series. Alternatively, the log-returns can be defined as:

$$r_t = \log p_t - \log p_{t-h} \quad . \tag{22}$$

Eqs. (20)–(22) are definitions, thus if linear returns are defined according to Eq. (21) they are related to the logarithmic returns by the following equation $r_t^{log} = \log(r_t^{lin})$ [35].

Data sets have been downloaded from the terminal www.bloomberg.com/professional. For each index, data include tick-by-tick prices p_t from January to December 2018. Details (Ticker; Extended name; Country; Currency; Members; Length) as provided by Bloomberg for the three assets are reported in Table 1. The length of each index refers to the year 2018 (last column).

As a general overview of the asset properties, we report the first four moments of the linear returns' distributions, respectively the mean $\mu = E(x_t)$, variance $\sigma^2 = E(x_t - \mu)^2$, skewness $s = E(x_t - \mu)^3/\sigma^3$ and kurtosis $k = E(x_t - \mu)^4/\sigma^4$, in Table 2. Estimates of the Hurst exponent $\hat{H}$ are also reported (last column in Table 2). All results are obtained over the entire financial series data ranging from January 2018 to December 2018. Estimates of the Hurst Exponent are obtained by means of the Detrending Moving Average algorithm [30,31]. Given a moving average window n, for each observation in the series, a backward m.a. $\tilde{x}_{n,backward}(t) = \frac{1}{n}\sum_{k=0}^{n-1} x(t-k)$, computes the mean of n past observations, a centred m.a. $\tilde{x}_{n,centred}(t) = \frac{1}{n}\sum_{k=-(n-1)/2}^{(n-1)/2} x(t-k)$, computes the mean of $n/2$ past observations and $n/2$ future observations and a forward m.a. $\tilde{x}_{n,forward}(t) = \frac{1}{n}\sum_{k=-(n-1)}^{0} x(t-k)$ computes the mean of n future observations. Fig. 1 reports linear returns and norm plots of linear returns for the U.S. financial indexes. Results show the departure of the linear return distributions from normality.

In this study, different temporal horizons have been considered as monthly integer multiples of one-month period ranging from $M = 1$ up to $M = 12$. To perform the cluster entropy analysis over equally spaced sequences with constant

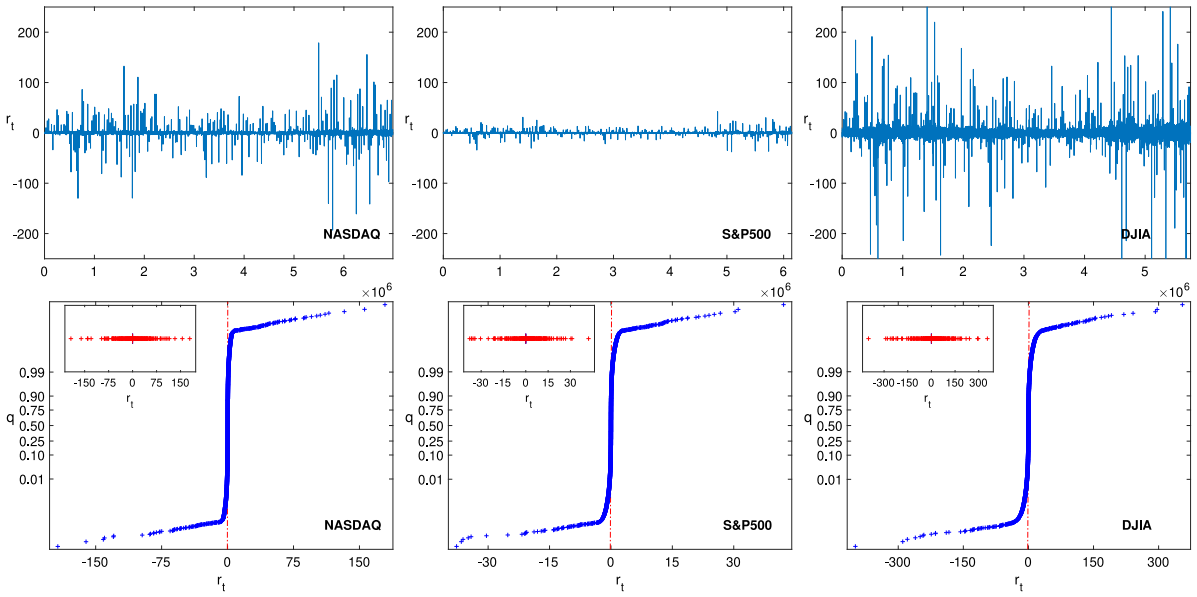

Fig. 1. Plot of linear returns for NASDAQ, S&P500 and DJIA (first row). Quantile–return plot for NASDAQ, S&P500 and DJIA (second row). These plots compare the distribution of the three indexes' linear returns with a normal distribution: on the *x*-axis, values of returns are reported, with mean close to 0 as found in Table 1, while on the *y*-axis quantile values are reported (the scale is not linear but a function of the interquantile distance). The distribution underlying indexes' linear returns cannot be assumed normal due to the heavy fat tails at the extreme of the distribution. The insets show box plots. Boxes for every indexes are very narrow because of the high variability of linear returns at the extreme of the distribution, that confirms the fat tails distribution mentioned above.

Table 3
Data length. 1st column reports the temporal horizon M (number of periods in month units). The lengths N of the price series for each temporal horizon M for the three assets are reported in 2nd, 3rd and 4th columns.

M	NASDAQ	S&P500	DJIA
1	586866	516635	516644
2	1117840	984046	984101
3	1704706	1500662	1500764
4	2291572	2017282	1623779
5	2906384	2558504	2165044
6	3493250	3075125	2681708
7	4069315	3580946	3187571
8	4712062	4146769	3753440
9	5243029	4614186	4220774
10	5885781	5180006	4786624
11	6461845	5685826	5292487
12	6982017	6142443	5749145

lengths, raw data have been sampled, thus yielding equally spaced data with equal length over different horizons. The sampling frequency is defined by dividing the length of the series corresponding to the longest horizon by the length of the shortest one and rounding the ratio to the nearest integer. The individual lengths of the subsequences referred to the twelve time periods are reported for each index in Table 3.

Consider for example the S&P500 market (2nd column in Table 3). Minimum and maximum values of the horizon are respectively $M = 1$ (January with $N = 516635$) and $M = 12$ (twelve months from January to December with $N = 6142443$). N is the number of tick-by-tick data in the considered horizon. Looking at Table 3, the minimum length of the time series is for the horizon $M = 1$, which is different for the three markets as the number of transactions in each financial market is different. In the following, the length of S&P500 is taken as the minimum reference value to perform the entropy analysis.

For the sake of validation, computational tests have been performed on artificial series. The artificial series have been generated by means of the FRACLAB tool (https://project.inria.fr/fraclab/) with lengths N corresponding to those of the financial markets under investigation (Table 3). Further details are reported in Sections 4 and 5.

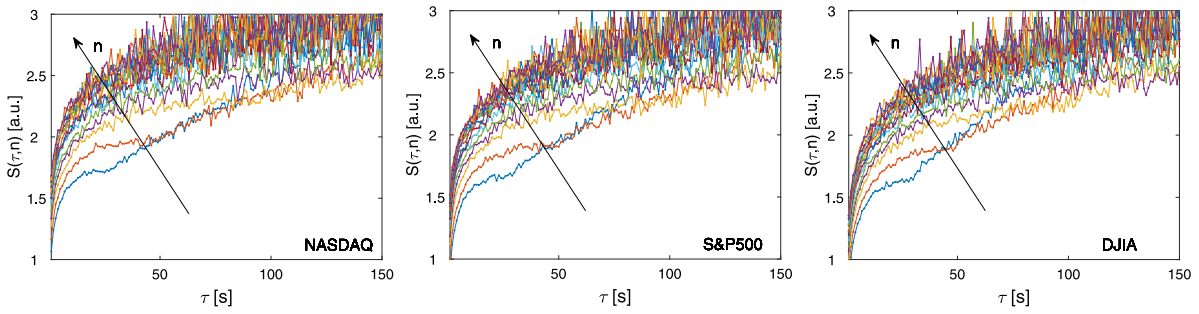

Fig. 2. Cluster entropy $S(\tau, n)$ vs cluster duration τ for the time series of the prices (raw data) respectively of the market indices NASDAQ, S&P500 and DJIA described in Table 1. The series lengths are $N = 586866$, $N = 516635$ and $N = 516644$ respectively for NASDAQ, S&P500 and DJIA as given in Table 3. The curves refer to one period, i.e. the first month of tick-by-tick data ($M = 1$). Different curves in each figure refer to different values of the moving average window n as indicated by the arrow.

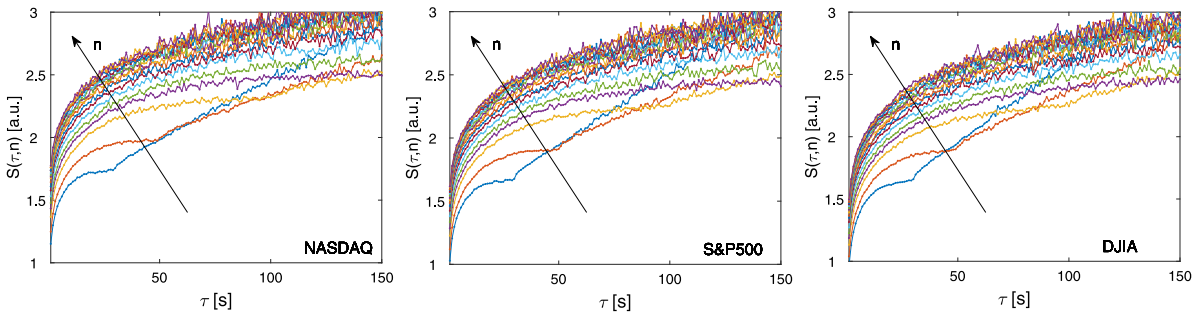

Fig. 3. Cluster entropy $S(\tau, n)$ vs cluster duration τ for the time series of the prices (raw data) respectively of the market indices NASDAQ, S&P500 and DJIA described in Table 1. The series lengths are $N = 6982017$, $N = 6142443$ and $N = 5749145$ respectively for NASDAQ, S&P500 and DJIA as given in Table 3. The curves refer to twelve periods, i.e. the whole year 2018 of tick-by-tick data ($M = 12$). Different curves in each figure refer to different values of the moving average window n as indicated by the arrow.

4. Results

Probability distribution $P(\tau, n)$ and cluster entropy $S(\tau, n)$ have been estimated on a large set of price series by means of the procedure summarized in Section 2. The series of NASDAQ, S&P500 and DJIA indexes described in Section 3 have been used for the investigation.

The cluster entropy $S(\tau, n)$ is obtained by taking the intersection of the asset prices p_t and its moving average $\tilde{p}_{t,n}$ for different moving average window n [20–22]. For each window n, the *clusters* are defined as the subsets $\{p_t : t = s, \ldots, s - n\}$ between two consecutive intersections. It is worth mentioning that the *clusters* are defined as the portions of the series between death/golden crosses according to the technical trading rules. Therefore, the information content of the cluster has a straightforward connection with the technical trading perspective on the price and volatility series. Following [22], the probability distribution function $P(\tau, n)$ is obtained by ranking the clusters according to their characteristic size (the duration τ).

Fig. 2 shows the cluster entropy $S(\tau, n)$ estimated on raw data prices. Plots refer to one month of data ($M = 1$). Lengths are $N = 586866$, $N = 516635$ and $N = 516644$ respectively for NASDAQ, S&P500 and DJIA (first row of Table 3).

Fig. 3 shows the cluster entropy $S(\tau, n)$ estimated on raw data prices for horizon of twelve months ($M = 12$). Lengths are $N = 6982017$, $N = 6142443$ and $N = 5749145$ respectively for NASDAQ, S&P500 and DJIA (last row of Table 3).

Fig. 4 shows the cluster entropy $S(\tau, n)$ estimated on sampled data of the price series. Plots refer to the first month of data ($M = 1$). All the series have same length $N = 492035$.

Fig. 5 shows the cluster entropy $S(\tau, n)$ estimated on sampled data of the price series. Plots refer to twelve months ($M = 12$). All the series have same length $N = 492035$. Different curves in each figure correspond to moving average values varying from $n = 30$ s, $n = 50$ s, $n = 100$ s, $n = 150$ s, $n = 200$ s ... up to $n = 1500$ s (with step 100 s).

The entropy curves shown in Figs. 2–5 exhibit a behaviour consistent with Eq. (13). At small values of the cluster duration ($\tau < n$), entropy behaves as a logarithmic function. At large values of the cluster duration ($\tau > n$), the entropy curves increase linearly with the τ/n term dominating. $S(\tau, n)$ is n-invariant for small values of τ, while its slope decreases as $1/n$ at larger τ, as expected according to Eq. (13), meaning that clusters with duration $\tau > n$ are not power-law correlated, due to the finite-size effects introduced by the partition with window n. Hence, they are characterized by a value of the entropy exceeding the curve $\log \tau^D$, which corresponds to power-law correlated clusters. Same cluster

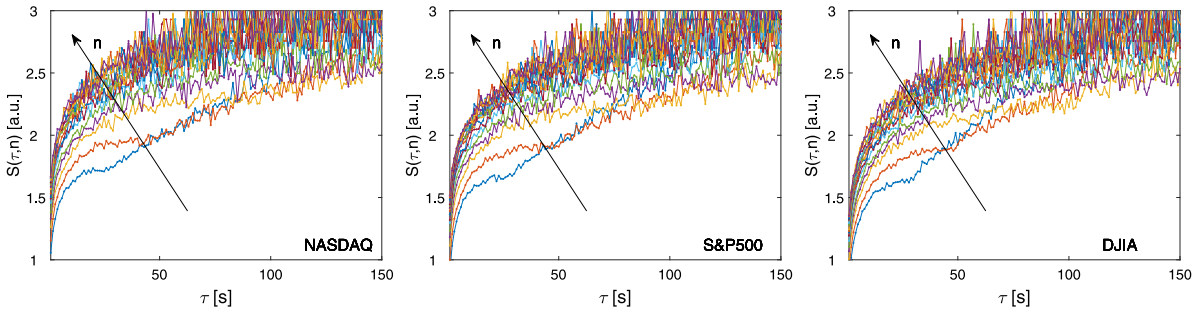

Fig. 4. Cluster entropy $S(\tau, n)$ vs cluster duration τ for the time series of the prices (sampled data) respectively of the market indices NASDAQ, S&P500 and DJIA described in Table 1. Figures refer to the first month of data ($M = 1$). All time series have same length $N = 492035$ by the sampling procedure described in Section 3. Different curves refer to different values of the moving average window n as indicated by the arrow.

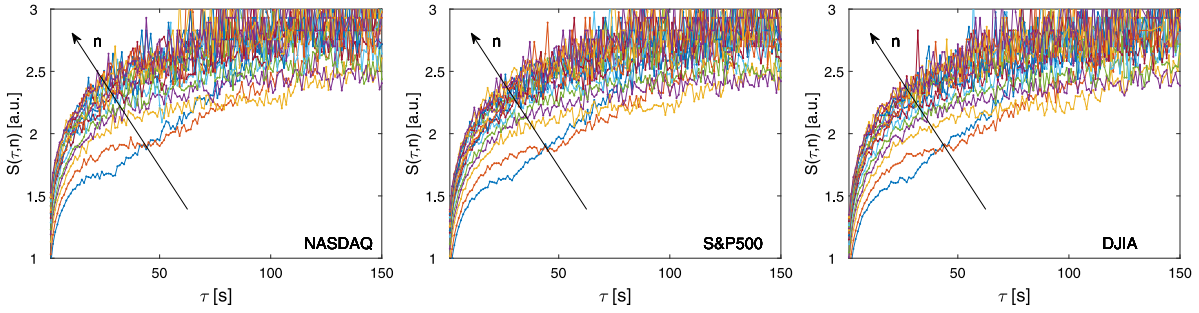

Fig. 5. Cluster entropy $S(\tau, n)$ vs cluster duration τ for the time series of the prices (sampled data) respectively of the market indices NASDAQ, S&P500 and DJIA described in Table 1. Figures refer to twelve months of data ($M = 12$). All time series have same length $N = 492035$ by the sampling procedure described in Section 3. Different curves refer to different values of the moving average window n as indicated by the arrow.

duration τ can be generated by different values of the moving average window n. At given τ value, larger entropy values are obtained as n increases. The entropy curves $S(\tau, n)$ of the NASDAQ, S&P500 and DJIA prices shown in Figs. 2–5 are representative of a quite general behaviour observed in several markets.

How to quantify the horizon dependence of the asset prices by using the cluster entropy $S(\tau, n)$ is discussed in the following. The *Market Dynamic Index* $I(M, n)$ is estimated by using Eq. (16) with the values of the entropy $S(\tau, n)$ of the asset prices p_t over several periods M on raw and sampled data. The first period ($M = 1$) of the price sequences corresponds to January 2018 for all the assets. Multiple periods have been built by considering $M = 2$ (January and February 2018) and, so on, up to $M = 12$ (one year from January to December 2018). Details concerning lengths of the series corresponding to the temporal horizons M are reported in Table 3. $I(M, n)$ is plotted in Fig. 6 for the NASDAQ, S&P500 and DJIA prices. A dependence of the function $I(M, n)$ on the horizon M can be observed. At small scales (i.e. small n and τ) $I(M, n)$ is the same for all M implying that the horizon dependence $H(M, n)$ is negligible. Conversely, at large n values, i.e. with a broad range of cluster lengths τ spanning more than one decades of values in the power law distribution, a horizon dependence $H(M, n)$ varying with M is found. By comparing the behaviour of the different assets a dependence of the function $I(M, n)$ on the horizon M is observed, especially at large scales. The *Market Dynamic Index* $I(M, n)$ varies more significantly for NASDAQ than for S&P500 and DJIA prices.

5. Discussion and conclusions

In this section, the results of the cluster entropy analysis obtained on real-world assets are discussed and compared with: (i) artificially generated series of Fractional Brownian Motion; (ii) artificially generated series of different representative agent models.

The *Market Dynamic Index* $I(M, n)$ and the *Horizon Dependence* $H(M, n)$ defined by Eqs. (16), (17) have been estimated on NASDAQ, S&P500 and DIJA assets. To understand the behaviour of real-world markets, simulations have been performed on Fractional Brownian Motion series with assigned Hurst exponent H generated via FRACLAB. In traditional financial theory, price increments are usually considered to be approximately *i.i.d.* and the Hurst exponent of financial processes is generally assumed to be $H \sim 0.5$. Fig. 7 shows cluster entropy curves for FBM series with $H = 0.5$. Each plot refers respectively to FBM segments ranging over one ($M = 1$), six ($M = 6$) and twelve ($M = 12$) monthly horizons. For the sake of comparison, FBM series have been generated with total length equal to that of the NASDAQ index in the whole 2018 ($N = 6982017$); then, the artificial series has been divided into twelve segments according to the monthly

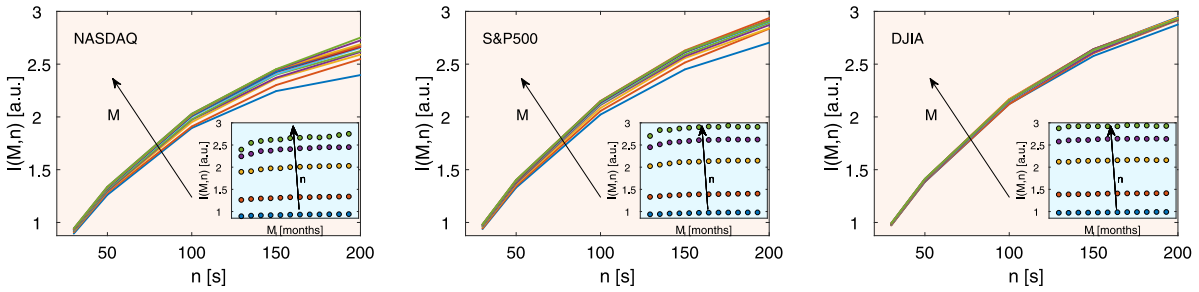

Fig. 6. *Market Dynamic Index* $I(M, n)$ as a function of the moving average window n, calculated according to Eq. (16) for the prices respectively of the NASDAQ, S&P500 and DJIA indexes as described in Table 1. Different curves in each figure refer to horizon varying from one ($M = 1$) to twelve months ($M = 12$). In particular, this set of curves corresponds to time series length $N = 492035$ by the sampling procedure described in Section 3. $I(M, n)$ has been evaluated as the integral of the entropy curves $S(\tau, n)$ similar to those shown in Fig. 4. The insets show the *Market Dynamic Index* $I(M, n)$ as a function of the horizon M. Symbols with different colours refer to different values of the moving average window n as indicated by the arrow (namely $n = 30$ s, $n = 50$ s, $n = 100$ s, $n = 150$ s and $n = 200$ s).

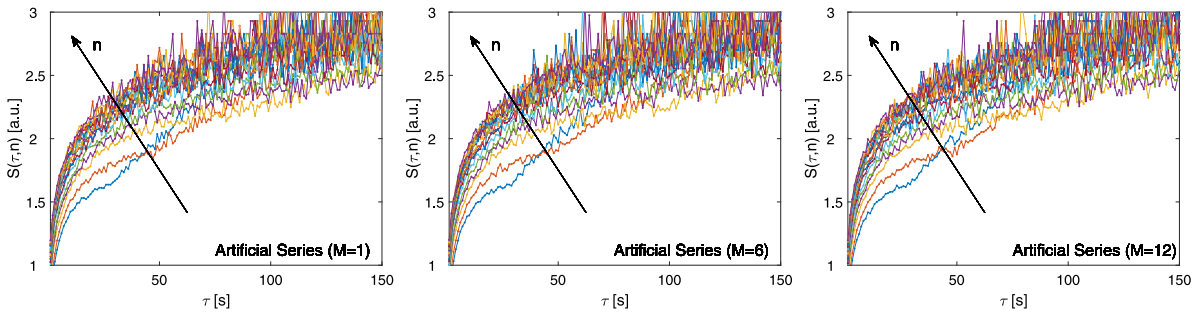

Fig. 7. Cluster entropy $S(\tau, n)$ vs cluster duration τ for the time series of the Fractional Brownian Motion with $H = 0.5$. Figures refer to one ($M = 1$), six ($M = 6$) and twelve ($M = 12$) time horizons of data. All time series have same length $N = 492035$ obtained by the sampling procedure described in Section 3 applied to the FBMs. Different curves refer to different values of the moving average window n as indicated by the arrow.

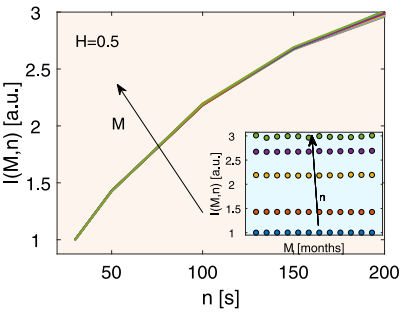

Fig. 8. *Market Dynamic Index* $I(M, n)$ as a function of the moving average window n, calculated according to Eq. (16) for the FBM with $H = 0.5$. Different curves in each figure refer to horizon varying from one ($M = 1$) to twelve months ($M = 12$). In particular, this set of curves corresponds to time series length $N = 492035$ with sampling frequency calculated as described in Section 3. $I(M, n)$ has been evaluated as the integral of the entropy curves $S(\tau, n)$ in Fig. 7. The insets show the *Market Dynamic Index* $I(M, n)$ as a function of the horizon M. Symbols with different colours refer to different values of the moving average window n as indicated by the arrow ($n = 30$ s, $n = 50$ s, $n = 100$ s, $n = 150$ s and $n = 200$ s).

structure of the NASDAQ series in 2018 (reported in Table 1) and sampled to obtain twelve series of the same lengths. Details of the sampling procedure are described in [24]. As it can be observed in Fig. 8, for Fractional Brownian Motion, $I(M, n)$ has quite a constant value at different horizons M and moving average windows n, thus exhibiting a different behaviour compared to the real world assets shown in Fig. 6.

To quantify the departure of real market series' from the artificially generated series behaviour, statistical significance is estimated and results are reported in Table 8. The *T-paired* statistics compares the moments (means, variances and higher order moments) that are obtained on two independent datasets according to either the same statistical measure or the same experiment. The comparison is performed between the real world time series (e.g. the S&P 500 assets) and artificial data series (the simple Fractional Brownian Motion with $H = 0.5$) by using the same information measure (the moving average cluster entropy). The *T-paired test* validates the null hypothesis that the cluster entropy values obtained

Table 4
Market Dynamic Index $I(M,n)$ and *Market Horizon Dependence* $H(M,n)$. The indexes $I(M,n)$ and $H(M,n)$ are calculated by using the relationships Eqs. (16), (17) with the entropy data plotted in Figs. 4 and 5 for NASDAQ. The spanned horizons M range from one to twelve months (from $M=1$ to $M=12$). The values of the moving average window n are reported in the 1st column. The reference values (entropy first period $I(1)$) in the third, fourth and fifth column have been taken equal to those of the consumption growth models respectively with Power Utility ($I(1)=0.0049$), Recursive Utility ($I(1)=0.0214$) and Difference Habit ($I(1)=0.0197$) [16].

Nasdaq composite index (NASDAQ)				
n	Entropy Indexes	Power Utility	Recursive Utility	Difference Habit
30	$I(12)$	0.0052	0.0226	0.0208
	$H(12)$	0.0003	0.0012	0.0011
50	$I(12)$	0.0052	0.0227	0.0209
	$H(12)$	0.0003	0.0013	0.0012
100	$I(12)$	0.0052	0.0229	0.0211
	$H(12)$	0.0003	0.0015	0.0014
150	$I(12)$	0.0054	0.0234	0.0215
	$H(12)$	0.0005	0.0020	0.0018
200	$I(12)$	0.0056	0.0246	0.0226
	$H(12)$	0.0007	0.0032	0.0029

Table 5
Market Dynamic Index $I(M,n)$ and *Market Horizon Dependence* $H(M,n)$ for the S&P500 index. Other details as in Table 4.

S&P 500 index (S&P500)				
n	Entropy Indexes	Power Utility	Recursive Utility	Difference Habit
30	$I(12)$	0.0051	0.0224	0.0206
	$H(12)$	0.0002	0.0010	0.0009
50	$I(12)$	0.0052	0.0226	0.0208
	$H(12)$	0.0003	0.0012	0.0011
100	$I(12)$	0.0052	0.0227	0.0209
	$H(12)$	0.0003	0.0013	0.0012
150	$I(12)$	0.0052	0.0229	0.0211
	$H(12)$	0.0003	0.0015	0.0014
200	$I(12)$	0.0053	0.0230	0.0212
	$H(12)$	0.0004	0.0016	0.0015

on real-world financial markets and those obtained on the artificial series (FBMs with $H=0.5$) come from distributions with equal mean and variance. In the case of the NASDAQ index, the *p-value* ranges between $0.5154 \leq p \leq 0.7584$. This confirms that the NASDAQ index exhibits a behaviour quite different from an independent elementary stochastic process of price variations, as the FBM with $H=0.5$ would imply. In the case of the S&P500 index, the *p-value* ranges between $0.7399 \leq p \leq 0.9248$. Therefore, the S&P500 exhibits an intermediate tendency to behave as ideal market. In the case of the DJIA index, the *p-value* ranges in the interval $0.8892 \leq p \leq 0.9434$, suggesting a behaviour closer to an independent stochastic process.

A comparison with results of asset price dispersion and dynamics [16], based on the Kullback–Leibler divergence in terms of the ratio between the true and risk-adjusted distribution of the pricing kernels m_t is offered. In [16], the *Horizon Dependence* is defined as the difference $H(M)=I(M)-I(1)$ with $I(M)$ given by:

$$I(M)=\frac{EL_t(m_{t,t+M})}{M}, \tag{23}$$

where $EL_t(m_{t,t+M})$ is the average of the relative entropy of the pricing kernel, and $I(1)$ is calculated at the horizon $M=1$. The authors argued that a proper representative agent model should have substantial entropy, to account for the mean excess returns, and modest horizon dependence to account for the small premiums on long bonds. A summary of the horizon dependence $H(M)$ obtained by different representative agent models according to [16] is reported in Table 7.

Values of the *Market Dynamic Index* $I(M,n)$ and *Horizon Dependence* $H(M,n)$, defined respectively by Eq. (16) and Eq. (17), are reported in Tables 4–6 for the NASDAQ, S&P500 and DJIA. The one-period cluster entropy value $I(1,n)$ has been taken equal to the one-period entropy (lower bound entropy) $I(1)=0.0049$, $I(1)=0.0214$ and $I(1)=0.0197$ respectively for power utility, recursive utility and difference habit representative agent models of the consumption growth [16]. The value $I(12,n)$ has been obtained from the curves in Fig. 6 for NASDAQ, S&P500 and DJIA. $H(12,n)$ is obtained as the difference between $I(12,n)$ and $I(1,n)$ on account of Eq. (17).

The *Market Dynamic Indexes* $I(M,n)$ (data in Tables 4–6) have been plotted in Fig. 9 (circles) for the NASDAQ, S&P 500 and DJIA. The values of the *Market Dynamic Indexes* $I(M,n)$ are referred to the lower bound value $I(1)$. As a guide

Table 6
Market Dynamic Index $I(M,n)$ and *Market Horizon Dependence* $H(M,n)$ for the DJIA index. Other details as in Table 4.

Dow Jones Industrial Average Index (DJIA)				
n	Entropy Indexes	Power Utility	Recursive Utility	Difference Habit
30	$I(12)$	0.0050	0.0218	0.0201
	$H(12)$	0.0001	0.0004	0.0004
50	$I(12)$	0.0050	0.0219	0.0201
	$H(12)$	0.0001	0.0005	0.0004
100	$I(12)$	0.0050	0.0217	0.0200
	$H(12)$	0.0001	0.0003	0.0003
150	$I(12)$	0.0050	0.0219	0.0201
	$H(12)$	0.0001	0.0005	0.0004
200	$I(12)$	0.0050	0.0218	0.0201
	$H(12)$	0.0001	0.0004	0.0004

Table 7
Entropy Index $I(1) = EL_t\left(m_{t,t+1}\right)$, $I(\infty)$, *Horizon Dependence* $H(120) = I(120) - I(1)$ *and* $H(\infty) = I(\infty) - I(1)$ as defined in Ref. [16]. Kullback–Leibler entropy approach is used for obtaining the entropy indexes $I(1)$ and $I(\infty)$ and horizon dependences $H(120)$ and $H(\infty)$ for representative agent models with constant variance (top), stochastic variance (middle) and jumps (bottom) .

Constant variance				
	Power Utility	Recursive Utility	Ratio Habit	Difference Habit
$I(1)$	0.0049	0.0214	0.0049	0.0197
$I(\infty)$	0.0258	0.0232	0.0003	0.0258
$H(120)$	0.0119	0.0011	−0.0042	0.0001
$H(\infty)$	0.0208	0.0018	−0.0047	0.0061
Stochastic variance				
	Recursive Utility 1	Recursive Utility 2	Campbell Cochrane	–
$I(1)$	0.0218	0.0249	0.0230	–
$I(\infty)$	0.0238	0.0293	0.0230	–
$H(120)$	0.0012	0.0014	0	–
$H(\infty)$	0.0020	0.0044	0	–
with Jumps				
	IID w/Jumps	Stochastic Intensity	Constant Intensity 1	Constant Intensity 2
$I(1)$	0.0485	0.0512	1.2299	0.0193
$I(\infty)$	0.0485	0.0542	15.730	0.0200
$H(120)$	0	0.0025	9.0900	0.0005
$H(\infty)$	0	0.0030	14.5000	0.0007

Table 8
T-paired test. 1st column reports the horizon M. 2nd, 3rd and 4th column report the probability p to reject the null hypothesis that the cluster entropy values have same mean and variance for NASDAQ, S&P500, DJIA and for the Fractional Brownian Motion with $H = 0.5$ at varying horizons M.

Probability p			
M	NASDAQ	S&P500	DJIA
1	0.5154	0.7399	0.8892
2	0.6026	0.8335	0.9257
3	0.6470	0.8588	0.9332
4	0.6631	0.8814	0.9283
5	0.6823	0.9018	0.9417
6	0.7124	0.9246	0.9534
7	0.7162	0.9224	0.9461
8	0.7288	0.9309	0.9618
9	0.7370	0.9479	0.9645
10	0.7409	0.9336	0.9570
11	0.7542	0.9321	0.9519
12	0.7584	0.9248	0.9434

for the eyes, lines between the symbols are drawn. The variation of the *Market Dynamic Index* $I(M,n)$ can be directly compared to the values $I(120)$ and $I(\infty)$ estimated in [16] for the power utility, recursive utility and difference habit

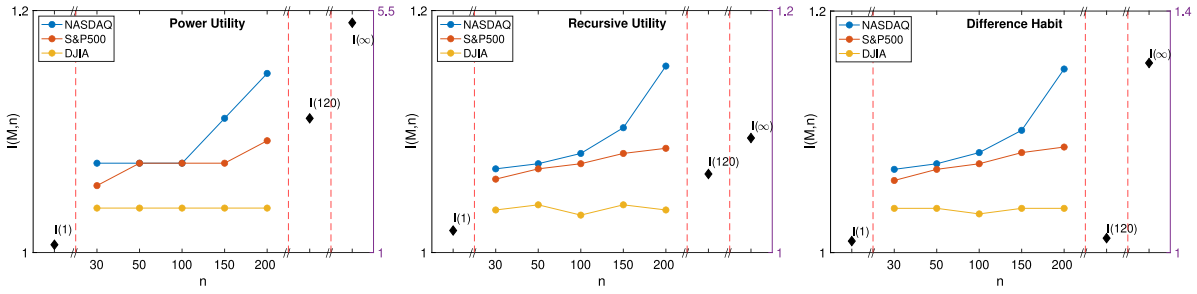

Fig. 9. *Market Dynamic Index* $I(M, n)$ as a function of the moving average window n (left axis) for NASDAQ, S&P500 and DIJA (data reported in Tables 4, 5, 6). Cluster entropy data for the three markets are plotted with reference to the lower bound entropy values $I(1) = 0.0049$ for *Power Utility* (left panel), $I(1) = 0.0214$ for *Recursive Utility* (middle panel) and $I(1) = 0.0197$ for *Difference Habit* (right panel) agent models respectively. The value of the entropy index $I(120)$ and the upper bound $I(\infty)$ are also plotted (values on the right axis).

representative agent models. Bearing in mind that the cluster entropy increase is referred to twelve months (namely M varies from 1 to 12 over the year 2018), Fig. 9 zooms the *one-year horizon dependence* out of the *ten-years horizon dependence* investigated in [16]. Overall, the comparison with the Kullback–Leibler entropy based on the pricing kernel shows that the two approaches yield consistent results. Furthermore Fig. 9 shows that the increasing trend exhibited by the entropy index with recursive utility more faithfully reproduces those of the real world data compared to other models. Actually, the difference habit and power utility representative agent models respectively under-estimate and over-estimate the increasing trends and the asymptotic values of the entropy production rate.

In order to complete our understanding of the behaviour of the cluster entropy $S(\tau, n)$ and the market dynamic index $I(M, n)$, it is convenient to recall and further elaborate on the meaning of the entropy change dS in a time interval dt (Eq. (3)) rewritten here:

$$dS = dS_{int} + dS_{ext}$$

with dS_{int} and dS_{ext} related to the endogeneous and exogeneous irreversible processes respectively. The assumption of local equilibrium implies the existence of multiple characteristic times (multiscale process), namely, the time required to reach the equilibrium for the whole system and the time required to reach the equilibrium locally (i.e. in a limited volume, compared to the size of the system). The variation of entropy with time t in a local volume can be written in the form:

$$\frac{dS}{dt} = \sigma - \mathrm{div}\,(\boldsymbol{j}_s) \tag{24}$$

where $\sigma = \sum_i X_i J_i$ is the endogenous entropy production (positive or equal to zero according to the second law of thermodynamics); j_s is the entropy flux, which depends on the thermodynamic flux densities J_i; X_i are thermodynamic forces. The maximum entropy production principle states that if irreversible forces X_i act, the actual flux J_i maximizes the entropy production. In isolated systems, the maximum entropy production Eq. (24) reduces to:

$$\frac{dS}{dt} = \sigma \quad , \tag{25}$$

implying that an isolated system tends to the state with maximum entropy along the shortest possible path.

The entropy change dS/dt, with the entropy production rates σ and $\mathrm{div}\,(\boldsymbol{j}_s)$ can be related to the derivative of Eq. (13), by taking the cluster duration τ as the time unit, i.e.:

$$\frac{dS}{d\tau} = \frac{\alpha}{\tau} + \frac{1}{n} \quad , \tag{26}$$

with the first term referring to the entropy produced by the intrinsic processes and the second term referring to the entropy produced by the interactions of the system with its environment.

To interpret the behaviour of the cluster entropy results for the three assets, we recall that the exponent α is related to the Hurst exponent by the relation $\alpha = 2 - H$ (with $0 < H < 1$). In the range of cluster duration $\tau < n$ when the entropy behaviour is dominated by the logarithmic term, the excess entropy τ/n can be neglected. Hence, the entropy production rate is given by:

$$\sigma = \frac{2 - H}{\tau} \quad , \tag{27}$$

with the dependence on τ consistent with the entropy production rate of the simple diffusion process [28]. This equation shows that the entropy production rate is positive for any $\tau > 0$ and thus the process can be classified as irreversible. The entropy production decreases with τ and tends to zero for $\tau \to +\infty$.

The relationship Eq. (27) suggests that the entropy production rate depends on the correlation exponent H. This dependence is reflected in the behaviour of the three assets as will be discussed here below. In order to compare the entropy production and cope with fluctuations and discreteness of the data, the time-averaged entropy over a time interval has been considered (Eq. (16) with $\tau_{max} \approx n$). The *Market Dynamic Index* $I(M, n)$ provides a cumulative estimate of the endogenous entropy production in the system. It depends on the Hurst exponent H ($\alpha = 2 - H$). By comparing integrals estimated over same range of cluster duration τ, like those in Fig. 6, one can note that the *Market Dynamic Index* $I(M, n)$ exhibits a slower increase for NASDAQ ($H = 0.53$) compared to S&P500 ($H = 0.52$) and DIJA ($H = 0.51$). This entropy production behaviour is consistent with the smaller value of $\alpha = 2 - H$ for the NASDAQ compared to S&P500 and DIJA.

However, the difference in the Hurst exponent H is not enough to account for the dispersion of the entropy index $I(M, n)$ observed at large n for NASDAQ in particular. The discrete time version of fractional Brownian motion with Hurst exponent H is the fractionally differenced white noise or ARFIMA $(0, d, 0)$ where d is the fractional differencing parameter related to the Hurst coefficient as $d = H - 1/2$. A more general class of processes are the ARFIMA (p, d, q) models with the parameter p and q related to the autoregressive and moving average terms. The ARFIMA (p, d, q) model might account for the dispersion of the entropy curve. This effect might be related to a local variability and short term correlation as one can deduce from [24] where cluster entropy simulations with ARFIMA (p, d, q) processes with parameter $p, q \neq 0$ are reported.

An interesting perspective could be offered by arguing on the emergence of long memory as a consequence of the aggregation of short memory processes [36]. Formally, it was shown that the autocorrelation function of the aggregated process,

$$x_t = \sum_{i=1}^{\mathcal{N}_\mathcal{A}} x_{i,t}, \quad x_{i,t} = \phi_i x_{i,t-1} + \varepsilon_{i,t}$$

where $\left(\varepsilon_{i,t}\right)_{t\in\mathbb{N}}$ are i.i.d. white noise for each i, converges to that of a long memory process as $\mathcal{N}_\mathcal{A} \to \infty$ if $(\phi_i)_{i\in\mathbb{N}}$ are i.i.d. random variables. Real world applications have shown that $\mathcal{N}_\mathcal{A}$ does not need to be large and even a few superimposed entities allow x_t to exhibit long memory. The effect of aggregation can be relevant to the cluster entropy behaviour. The three assets are obtained by aggregating 2570, 505 and 30 members (see Table 1). They appear to be driven by a large number of superimposed dynamical entities characterized by short-term fluctuations.

As a concluding remark, the cluster entropy method demonstrates its ability to quantify dispersion, intrinsic dynamics and horizon dependence of price returns. The comparison with the Kullback–Leibler entropy results has evidenced that the cluster entropy approach might provide complementary criteria to select among different representative agent models of financial markets.

Future directions of this study have been devised aimed at deepening the insights in the endogenous sources of market dynamics and at extending the definition of portfolio weights (proposed in Ref. [23]), to include horizon dependence in the optimal portfolio choice.

CRediT authorship contribution statement

Linda Ponta: Methodology, Resources. **Pietro Murialdo:** Data curation, Investigation. **Anna Carbone:** Conceptualization, Methodology, Writing-review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability statement

NASDAQ, DIJA, SP500 indexes have been downloaded from the Bloomberg terminal: www.bloomberg.com/professional, which is accessible to several organizations worldwide or by activating a subscription via the registration page.

Furthermore we provide a set of series of different lengths (simulating the different horizons) that have been generated artificially by means of the FRACLAB tools freely accessed at: https://project.inria.fr/fraclab/ and a MATLAB code to reproduce the results shown in the Figures.

Acknowledgement

This work has been supported by FuturICT 2.0 a FLAG-ERA Initiative within the Joint Transnational Calls 2016, Italy, Grant Number: JTC-2016-004

Table 9
File and workflow description.

Name	Description
TimeSeriesGenerator.m	→ generate time series by using FracLab.
SampledSeries.m	→ create data structures and sample the data.[a]
Prices.m	→ create and store price vectors.
Entropy.m	→ evaluate the cluster entropy.[b]
MarketDynamicIndex.m	→ evaluate the Market Dynamic Index.[c]
DMIfigure.m	→ plot the Market Dynamic Index.[d]
DMA.m	function
DMAbackward.m	function
DMAcentered.m	function
DMAforward.m	function
ComputeClusterProbability.m	function

[a]It has the variable "Sampled" as input. If "Sampled = 1" returns sampled data; If "Sampled = 0" returns unsampled (raw) data.
[b]It Yields Figs. 2–5 and 7.
[c]According to Eq. (16).
[d]Yields Figs. 6 and 8.

Table 10
Data input and output.

File name	Input (data & functions)	Output (data & figures)
TimeSeriesGenerator.m	None	Data1.mat, . . . , Data12.mat
SampledSeries.m	Data1.mat - Data12.mat	DataSampled0.mat; DataSampled1.mat
Prices.m	DataSampled0.mat, DataSampled1.mat	PricesData1.mat, . . . , PricesData12.mat.
Entropy.m	PricesData1.mat, . . . , PricesData12.mat; DMA.m; ComputeClusterProbability.m	Simulations_Complete_Data1.mat, . . . , Simulations_Complete_Data12.mat
MarketDynamicIndex.m	Simulations_Complete_Data1.mat, Simulations_Complete_Data12.mat	MDI_n_Prices_Data.mat
DMI_figure.m	MDI_n_Prices_Data.mat	None
DMA.m	DMA_backward.m; DMA_centred.m; DMA_forward.m	None

Appendix. Supporting information

This section describes the files available in the Dropbox Shared Folder: https://www.dropbox.com/sh/9pfeltf2ks0ewjl/AACjuScK_gZxmyQ_mDFmGHoya?dl=0, in particular the following items can be found:
DATA contains raw and sampled data .zip folders. These data can be generated and analysed by using the codes described here below. Horizon.zip contains the horizon (.MAT) data estimated for market series and artificial series shown in Figs. 6 and 8. These data have been used to estimate the horizon dependence reported in Table 7. These data have been used to perform the test whose results are reported in Table 8.
CODES contains all the MATLAB codes used for the analysis (Codes.zip).
Main steps and input/output resources are summarized respectively in Table 9 and in Table 10:

1. Generate the time series by using the *"TimeSeriesGenerator.m"* file (alternatively use any other time series in the specified format).
2. Create a structured data with option sampled/unsampled the series with the *"SampledSeries.m"* file.
3. Create a vector of prices by means of *"Prices.m"*
4. Evaluate the Market Dynamic Index by means of *"MarketDynamicIndex.m"*
5. Plot the MarketDynamic Index by means of *"MDI_figure.m"* (Figs. 6 and 8).
6. Evaluate the entropy by means of *"Entropy.m"* (to produce Figs. 2 and 3 for unsampled data series; to produce Figs. 4 and 5 for sampled data series).

References

[1] J.P. Crutchfield, Between order and chaos, Nat. Phys. 8 (1) (2012) 17–24.
[2] C. Bandt, B. Pompe, Permutation entropy: A natural complexity measure for time series, Phys. Rev. Lett. 88 (17) (2002) 174102.
[3] P. Grassberger, I. Procaccia, Characterization of strange attractors, Phys. Rev. Lett. 50 (5) (1983) 346.
[4] G.C. Philippatos, C.J. Wilson, Entropy, market risk, and the selection of efficient portfolios, Appl. Econ. 4 (3) (1972) 209–220.
[5] J. Ou, Theory of portfolio and risk based on incremental entropy, J. Risk Financ. 6 (1) (2005) 31–39.
[6] A.K. Bera, S.Y. Park, Optimal portfolio diversification using the maximum entropy principle, Econometric Rev. 27 (4–6) (2008) 484–512.
[7] V. DeMiguel, L. Garlappi, R. Uppal, Optimal versus naive diversification: How inefficient is the $1/n$ portfolio strategy? Rev. Financ. Stud. 22 (5) (2009) 1915–1953.
[8] M. Ormos, D. Zibriczky, Entropy-based financial asset pricing, PLoS One 9 (12) (2014) e115742.

[9] P. Fiedor, A. Hołda, Information-theoretic approach to quantifying currency risk, J. Risk Financ. (2016).
[10] L. Chen, R. Gao, Y. Bian, H. Di, Elliptic entropy of uncertain random variables with application to portfolio selection, Soft Comput. (2020) 1–15.
[11] S. Lahmiri, S. Bekiros, Randomness, informational entropy, and volatility interdependencies among the major world markets: The role of the COVID-19 pandemic, Entropy 22 (8) (2020) 833.
[12] L. Molgedey, W. Ebeling, Local order, entropy and predictability of financial time series, Eur. Phys. J. B 15 (4) (2000) 733–737.
[13] L.P. Hansen, R. Jagannathan, Implications of security market data for models of dynamic economies, J. Political Econ. 99 (2) (1991) 225–262, URL http://www.jstor.org/stable/2937680.
[14] L.P. Hansen, Nobel lecture: Uncertainty outside and inside economic models, J. Political Econ. 122 (5) (2014) 945–987, http://dx.doi.org/10.1086/678456, arXiv:https://doi.org/10.1086/678456.
[15] L.P. Hansen, T.J. Sargent, Macroeconomic Uncertainty Prices, Working Paper Series 25781, National Bureau of Economic Research, 2019, http://dx.doi.org/10.3386/w25781,
[16] D. Backus, M. Chernov, S. Zin, Sources of entropy in representative agent models, J. Finance 69 (1) (2014) 51–99.
[17] A. Ghosh, C. Julliard, A.P. Taylor, What is the consumption-CAPM missing? An information-theoretic framework for the analysis of asset pricing models, Rev. Financial Stud. 30 (2) (2017) 442–504, http://dx.doi.org/10.1093/rfs/hhw075.
[18] V. Dimitrova, M. Fernández-Martínez, M. Sánchez-Granero, J. Trinidad Segovia, Some comments on bitcoin market (in) efficiency, PLoS One 14 (7) (2019) e0219243.
[19] A.M. Puertas, M.A. Sánchez-Granero, J. Clara-Rahola, J.E. Trinidad-Segovia, F.J. de las Nieves, Stock markets: A view from soft matter, Phys. Rev. E 101 (3) (2020) 032307.
[20] A. Carbone, G. Castelli, H.E. Stanley, Analysis of clusters formed by the moving average of a long-range correlated time series, Phys. Rev. E 69 (2004) 026105.
[21] A. Carbone, H.E. Stanley, Scaling properties and entropy of long-range correlated time series, Physica A 384 (1) (2007) 21–24.
[22] A. Carbone, Information measure for long-range correlated sequences: The case of the 24 human chromosomes, Sci. Rep. 3 (2013) 2721.
[23] L. Ponta, A. Carbone, Information measure for financial time series: Quantifying short-term market heterogeneity, Physica A 510 (2018) 132–144, URL http://www.sciencedirect.com/science/article/pii/S0378437118308100.
[24] P. Murialdo, L. Ponta, A. Carbone, Long-range dependence in financial markets: A moving average cluster entropy approach, Entropy 22 (6) (2020) 634, Publisher: Multidisciplinary Digital Publishing Institute.
[25] P. Glansdorff, I. Prigogine, Thermodynamic Theory of Structure, Stability and Fluctuations, Wiley-Interscience, 1971.
[26] F. Schlögl, On dynamics of small fluctuations from a steady state, Phys. Lett. A 36 (3) (1971) 193–194.
[27] G. Nicolis, Y. De Decker, Stochastic approach to irreversible thermodynamics, Chaos 27 (10) (2017) 104615.
[28] Y. Luchko, Entropy production rates of the multi-dimensional fractional diffusion processes, Entropy 21 (10) (2019) 973.
[29] C.E. Shannon, A mathematical theory of communication, Part I, Part II, Bell Syst. Tech. J. 27 (1948) 623–656.
[30] S. Arianos, A. Carbone, C. Türk, Self-similarity of higher-order moving averages, Phys. Rev. E 84 (4) (2011) 046113.
[31] A. Carbone, K. Kiyono, Detrending moving average algorithm: Frequency response and scaling performances, Phys. Rev. E 93 (6) (2016) 063309.
[32] M. Costa, A.L. Goldberger, C.-K. Peng, Multiscale entropy analysis of complex physiologic time series, Phys. Rev. Lett. 89 (6) (2002) 068102.
[33] H. Niu, J. Wang, Quantifying complexity of financial short-term time series by composite multiscale entropy measure, Commun. Nonlinear Sci. Numer. Simul. 22 (1–3) (2015) 375–382.
[34] A. Humeau-Heurtier, The multiscale entropy algorithm and its variants: A review, Entropy 17 (5) (2015) 3110–3123.
[35] R. Cont, Empirical Properties of Asset Returns: Stylized Facts and Statistical Issues, Taylor & Francis, 2001.
[36] C.W. Granger, Long memory relationships and the aggregation of dynamic models, J. Econometrics 14 (2) (1980) 227–238.