

Article

Multiscale Model Selection for High-Frequency Financial Data of a Large Tick Stock by Means of the Jensen–Shannon Metric

Gianbiagio Curato ¹ and Fabrizio Lillo ^{2,3,*}

¹ Scuola Normale Superiore, Piazza dei Cavalieri 7, Pisa 56126, Italy;

E-Mail: gianbiagio.curato@sns.it or altair2020@libero.it

² Scuola Normale Superiore, Piazza dei Cavalieri 7, Pisa 56126, Italy

³ Dipartimento di Fisica e Chimica, Viale delle Scienze, Palermo 90128, Italy

* Author to whom correspondence should be addressed; E-Mail: fabrizio.lillo@sns.it;

Tel.:+39 050509159.

Received: 20 October 2013; in revised form: 18 November 2013 / Accepted: 11 December 2013 /

Published: 16 January 2014

Abstract: Modeling financial time series at different time scales is still an open challenge. The choice of a suitable indicator quantifying the distance between the model and the data is therefore of fundamental importance for selecting models. In this paper, we propose a multiscale model selection method based on the Jensen–Shannon distance in order to select the model that is able to better reproduce the distribution of price changes at different time scales. Specifically, we consider the problem of modeling the ultra high frequency dynamics of an asset with a large tick-to-price ratio. We study the price process at different time scales and compute the Jensen–Shannon distance between the original dataset and different models, showing that the coupling between spread and returns is important to model return distribution at different time scales of observation, ranging from the scale of single transactions to the daily time scale.

Keywords: Jensen–Shannon divergence; multiscale analysis; model selection; high frequency financial data; Markov-switching modeling

1. Introduction

The complexity of market behavior has fascinated physicists and mathematicians for many years [1]. One of the main sources of interest comes from the difficulty of modeling the rich dynamics of asset

prices. In fact, since the beginning of the last century, a large set of statistical regularities of price dynamics has been identified, including the asymptotically power-law distribution of returns, their lack of linear correlations, but the presence of very persistent higher order correlations, the slow convergence to the Gaussian distribution, scaling properties, multifractality, *etc.* [2–4]. The modeling activity has been correspondingly very intense, considering models both in discrete and in continuous time, and including random walks, Levy processes, stochastic volatility models, multifractal models, *etc.* [1,5–8]. However, up to now, there is no consensus on a model that is able to reproduce all the statistical regularities, and therefore, there is a growing interest toward methods allowing one to discriminate among different models those more suited to describe financial data.

A specific challenge is the modeling of how the return distribution changes at different time scales [3]. Due to the presence of fat-tailed distributions, also at very short time scales, and non-linear time correlations, the dynamics of the price-change distribution is far from trivial and not well described by any model. The problem becomes even more dramatic when one wants to describe the price-change distribution also at the shortest time scales, *i.e.*, when the discrete nature of trading appears. Trading and, correspondingly, price changes occur at discrete time. Moreover, an asset price cannot assume arbitrary values, but it is constrained to live in a grid of values fixed by the exchange. The tick size is the smallest interval between two prices, *i.e.*, the grid step. Since tick size can be a sizable fraction of the asset price, when seen at small time scales, price movement appears as a (non-trivial) random walk on a grid, with jumps occurring at random times, while at large time scales, one can probably forget the microstructural issues and describe the dynamics with a more traditional stochastic differential equation or time series approach. One of the main methodological problem is, therefore, to have a method to compare data and model predictions *at different time scales*.

In this work, we propose to perform multiscale model selection for financial time series by using the Jensen–Shannon distance [9–11], and we specifically consider the case of models describing the high frequency dynamics of large tick assets, *i.e.*, assets where the ratio between tick and price is relatively large [12,13]. We perform the model selection at different scales m , representing the level of aggregation of the time series. In other words, given the return time series, $x(t)$, we study the properties of the probability distribution of its sums $y_m = \sum_{t=1}^m x(t)$. It is important to clarify that we do not perform a goodness-of-fit test at different scales m defining a p -value relative to a specific statistic, *e.g.*, Kolmogorov–Smirnov statistic, *etc.* [14]. Our analysis consists, instead, in the comparison between the probability distribution computed from empirical data and those computed from synthetic data generated by specific statistical models. The discrepancy is measured by the Jensen–Shannon distance. In particular, by considering a class of models recently proposed [15], we show that models containing the coupling between price and spread, as well as the time correlation of spread outperform other models without these characteristics in describing the change of the shape of the return distribution across scales.

The paper is organized as follows. In Section 2, we illustrate the definitions of the Jensen–Shannon divergence and distance, and we characterize the unavoidable bias, due to the finiteness of the data sample. In Section 3, we illustrate the statistical models of mid-price and spreads dynamics developed in [15]. Moreover, we apply the Jensen–Shannon distance criteria to select among three competing models of the dynamics of the price of a large tick asset, namely Microsoft. Finally in Section 4, conclusions and perspectives are discussed.

2. Jensen–Shannon Distance

Distance or divergence measures are of key importance in a number of theoretical and applied statistical inference and data processing problems, such as estimation, detection, compression and model selection [16]. Among the proposed measures, one of the best known is the Kullback–Leibler (KL) divergence between two distributions, $D(\mathbf{p}||\mathbf{q})$ [17], also called *relative entropy*. It is a measure of the inefficiency of assuming that the distribution is \mathbf{q} when the true distribution is \mathbf{p} . It is used in many different applications, such as econometrics [18], clustering analysis [19], multivariate analysis [20,21], neuroscience [22] and discrete systems [23]. We will limit the following discussion to discrete probability distributions, but the results can be generalized to probability density functions.

Let X be a discrete random variable with support of definition \mathcal{X} and probability mass function $p(x)$, $x \in \mathcal{X}$. If $q(x)$ is another probability mass function defined on the same support, \mathcal{X} , the KL-divergence is defined as:

$$D_{KL}(\mathbf{p}||\mathbf{q}) = \sum_{x \in \mathcal{X}} p(x) \log \left(\frac{p(x)}{q(x)} \right) \quad (1)$$

where the base of the logarithm is two. We use the convention that $0 \log(0/0) = 0$ and the convention, based on continuity arguments, that $0 \log(0/q) = 0$. If there is any symbol, $x \in \mathcal{X}$, such that $p(x) > 0$ and $q(x) = 0$, then $D_{KL}(\mathbf{p}||\mathbf{q})$ is undefined. This means that distribution \mathbf{p} has to be absolutely continuous with respect to \mathbf{q} for KL-divergence to be defined [24]. It is well known that $D_{KL}(\mathbf{p}||\mathbf{q})$ is non-negative and additive, but not symmetric [24]. In order to overcome this problems, Lin [11] defined a new symmetric divergence, called L divergence:

$$D_L(\mathbf{p}, \mathbf{q}) = D_{KL}(\mathbf{p}||\mathbf{m}) + D_{KL}(\mathbf{q}||\mathbf{m}) \quad (2)$$

where $\mathbf{m} = (\mathbf{p} + \mathbf{q})/2$ is the “mean” probability mass function. $D_L(\mathbf{p}, \mathbf{q})$ vanishes if and only if $\mathbf{p} = \mathbf{q}$. The L divergence is symmetric and bounded by $D_L(\mathbf{p}, \mathbf{q}) \leq 2$. It is worth noticing that the L divergence can be expressed in terms of the Shannon entropy as:

$$D_L(\mathbf{p}, \mathbf{q}) = 2H\left(\frac{\mathbf{p} + \mathbf{q}}{2}\right) - H(\mathbf{p}) - H(\mathbf{q}) \quad (3)$$

i.e., it is the difference of entropy between the mean distribution, \mathbf{m} , and the sum of the entropies of \mathbf{p} and \mathbf{q} . The generalization of the L divergence is the Jensen–Shannon divergence [11], defined as:

$$Div_{JS}(\mathbf{p}, \mathbf{q}) = H(\pi_1\mathbf{p} + \pi_2\mathbf{q}) - \pi_1H(\mathbf{p}) - \pi_2H(\mathbf{q}) \quad (4)$$

where $\pi_1, \pi_2 \geq 0$, $\pi_1 + \pi_2 = 1$ are the weights of the probability distributions, \mathbf{p} and \mathbf{q} , respectively. According to this new definition, $D_L(\mathbf{p}, \mathbf{q}) = 2Div_{JS}(\mathbf{p}, \mathbf{q})$, for $\pi_1 = \pi_2 = 1/2$. Endres *et al.* [9] found that the square root of D_L is a metric, *i.e.*, it fulfills the triangle inequality. They named this new information metric the Jensen–Shannon distance, \mathcal{D}_{JS} :

$$\mathcal{D}_{JS}(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{x \in \mathcal{X}} \left(p(x) \log \frac{2p(x)}{p(x) + q(x)} + q(x) \log \frac{2q(x)}{p(x) + q(x)} \right)} \quad (5)$$

The bounds of this distance are: $0 \leq \mathcal{D}_{JS} \leq \sqrt{2}$. The Jensen–Shannon divergence is used also in statistical mechanics [25], quantum mechanics [26], thermodynamics [27], networks [28], particle physics [29], biology [30] and cosmology [31].

In this paper, we are interested in using the Jensen–Shannon distance as a method for selecting among a set of models the one that best describes a given dataset. We are concerned with the case when our data is represented by a discrete time series of length N . When considering different competing models, we search for the best model describing the probability distribution of the aggregation (*i.e.*, sum) of the time series at different time scales m . Moreover, the use of Jensen–Shannon distance allows us to compare two empirical distributions.

To be more specific, consider the random variable, x , taking values from the set $\mathbf{x} = (x_1, \dots, x_k)$ with probabilities $\mathbf{p} = (p_1, \dots, p_k)$. Given N observations of the time series, $x(t)$, $t = 1, \dots, N$, one builds a histogram $\mathbf{n} = (n_1, \dots, n_k)$, where n_i is the number of times the outcome was x_i . The frequency vector $\mathbf{f} = (f_1, \dots, f_k) = (n_1/N, \dots, n_k/N)$ is an estimator of the probability distribution, \mathbf{p} . We want to perform a statistical analysis at different scales of aggregation, *i.e.*, we study the probability distribution, \mathbf{p}_m , and frequency distribution, \mathbf{f}_m , of the sum $\sum_{t=1}^m x(t)$, where the value, m , defines the scale. The probability distribution of the elementary process $x(t)$, corresponding to $m = 1$, is denoted by $\mathbf{p}_{m=1} = \mathbf{p}$. If the initial dataset had N values, the scale, m , is limited by $1 \leq m \leq N$. The number of experimental data available at each aggregation scale m reduces to $N_m \equiv \lfloor N/m \rfloor$, because we sum the experimental data, which belong to the N_m non-overlapping windows of length m .

In order to select the best model that describes the data at all aggregation scales, we compute the Jensen–Shannon distances for various values of m , *i.e.*, $\mathcal{D}_{JS}(\mathbf{p}_m, \mathbf{f}_m)$. We estimate \mathbf{p}_m according to different statistical models, and we select the one that minimizes $\mathcal{D}_{JS}(\mathbf{p}_m, \mathbf{f}_m)$ for the different values of m . As will be clear below, we will also need to compute the distance between two frequency distributions in order to compare the two different datasets, $\mathcal{D}_{JS}(\mathbf{f}_{1,m}, \mathbf{f}_{2,m})$. In this case, we assume that the length of the two datasets is the same $N_1 = N_2 = N$.

It is important to stress that even if we knew the *true* distribution, \mathbf{p}_m , the distance, $\mathcal{D}_{JS}(\mathbf{p}_m, \mathbf{f}_m)$, inferred from a finite sample of data, would be larger than zero. The fluctuations of \mathbf{f}_m from dataset to dataset may not only result in fluctuations of the numerical values of \mathcal{D}_{JS} , but also in a systematic shift, *i.e.*, bias, of the numerical values of \mathcal{D}_{JS} . This bias is identified with the expectation value, $E[\mathcal{D}_{JS}(\mathbf{p}_m, \mathbf{f}_m)] \neq 0$, for the various values of the scale, m . The bias is also present if we compute the distance, $\mathcal{D}_{JS}(\mathbf{f}_{1,m}, \mathbf{f}_{2,m})$, between two frequency vectors that are computed from datasets representing the same stochastic process.

The concept of a systematic bias of the numerical values of Jensen–Shannon divergence, Div_{JS} , is well known in the literature, and it is connected to the systematic bias in the estimation of entropy. It follows directly from Jensen inequality [17] that the expected value, $E[H(\mathbf{f})]$, of the entropy computed from an ensemble of finite-length sequences cannot be greater than the theoretical value, $H(\mathbf{p})$, of the entropy computed from the (unobservable) probabilities:

$$E[H(\mathbf{f})] \leq H(\mathbf{p}) \quad (6)$$

where the expectation is defined over the ensemble of finite-length i.i.d. sequences generated by the probability distribution, \mathbf{p} . It can be shown that the expected value of the observed entropy is systematically biased downwards from the true entropy:

$$E[H(\mathbf{f})] = H(\mathbf{p}) - \frac{k-1}{2N \ln 2} + O(N^{-2}) \quad (7)$$

where k is the number of components of the probability and frequency vectors, \mathbf{p} and \mathbf{f} , and N is the ensemble size. This result was obtained by Basharin [32] and Herzel [33], who pointed out that to the first order, $O(1/N)$, the bias is independent of the actual distribution, \mathbf{p} . The term of order $O(1/N^2)$ involves the unknown probabilities $\mathbf{p} = (p_1, \dots, p_k)$ and cannot be estimated in general [34–36].

Grosse *et al.* [37] derived an analytical approximation of the expected value of $Div_{JS}(\mathbf{f}_1, \mathbf{f}_2)$ between two i.i.d. sequences of length N coming from the same probability distribution, which is:

$$E [Div_{JS}(\mathbf{f}_1, \mathbf{f}_2)] = \frac{k - 1}{4N \ln 2} + O(N^{-2}) \tag{8}$$

Clearly, also in the case of the Jensen–Shannon distance, \mathcal{D}_{JS} , there is a systematic positive bias.

2.1. A Simple Binomial Model

In this section, we present a toy example of the use of Jensen–Shannon distance for model selection. The purpose of the section is mostly didactical and serves to show the multiscale procedure and the issues related to the finiteness of the sample that will be present also in the real financial case described in the next section.

Let us consider a process, which at scale $m = 1$ is a binomial i.i.d. process, *i.e.*, $\mathbf{p}_{m=1}$ is described by $B(n, p_B)$, where p_B describes the probability of success. The sum of m i.i.d. binomial variables is still described by a binomial distribution [38], *i.e.*, $\mathbf{p}_m = (p_{m,1}, \dots, p_{m,k})$ is described by $B(nm, p_B)$, and its support is a set composed by $k = nm + 1$ elements. Given a time series of length N , at each aggregation scale, m , we have $N_m \equiv \lfloor N/m \rfloor$ observations from non-overlapping windows, and we measure the frequency vector $\mathbf{f}_m = (n_{m,1}/N_m, \dots, n_{m,k}/N_m)$, where $n_{m,i}$ is the number of occurrences of the event, i , at scale m .

The probability distribution of empirical frequencies is given by the multinomial distribution:

$$p(\mathbf{f}_m) = \lfloor N/m \rfloor! \prod_{i=1}^{nm+1} \frac{p_{m,i}^{n_{m,i}}}{n_{m,i}!} \tag{9}$$

In principle, one can compute exactly the moments of the distances, $\mathcal{D}_{JS}(\mathbf{p}_m, \mathbf{f}_m)$ and $\mathcal{D}_{JS}(\mathbf{f}_m^1, \mathbf{f}_m^2)$, which are:

$$E[\mathcal{D}_{JS}^b(\mathbf{p}_m, \mathbf{f}_m)] = \sum_{n_{m,1}, \dots, n_{m,k}=0}^{\lfloor N/m \rfloor} \lfloor N/m \rfloor! \prod_{i=1}^k \frac{p_{m,i}^{n_{m,i}}}{n_{m,i}!} \left(\sum_{i=1}^k \left(f_{m,i} \log \left(\frac{2f_{m,i}}{f_{m,i} + p_{m,i}} \right) + p_{m,i} \log \left(\frac{2p_{m,i}}{f_{m,i} + p_{m,i}} \right) \right) \right)^{b/2} \tag{10}$$

and

$$E[\mathcal{D}_{JS}^b(\mathbf{f}_m^1, \mathbf{f}_m^2)] = \sum_{n_{m,1}^1, \dots, n_{m,k}^1=0}^{\lfloor N/m \rfloor} \sum_{n_{m,1}^2, \dots, n_{m,k}^2=0}^{\lfloor N/m \rfloor} (\lfloor N/m \rfloor!)^2 \prod_{j=1}^{\lfloor N/m \rfloor} \prod_{l=1}^{\lfloor N/m \rfloor} \frac{(p_{m,j}^1)^{n_{m,j}^1}}{n_{m,j}^1!} \frac{(p_{m,l}^2)^{n_{m,l}^2}}{n_{m,l}^2!} \left(\sum_{i=1}^k \left(f_{m,i}^1 \log \left(\frac{2f_{m,i}^1}{f_{m,i}^1 + f_{m,i}^2} \right) + f_{m,i}^2 \log \left(\frac{2f_{m,i}^2}{f_{m,i}^1 + f_{m,i}^2} \right) \right) \right)^{b/2} \tag{11}$$

These expressions can be used to compute the mean and variance of the Jensen–Shannon distance, as well as of the Jensen–Shannon divergence.

The computational problem with these expectations are the values of $k = nm + 1$ and of N , because the number of categories of the multinomial distribution grows dramatically with the scale, m . The support of the multinomial distribution for the scale, m , has a number of elements:

$$n.e. = \binom{\lfloor N/m \rfloor + nm}{nm} = \frac{\prod_{i=1}^{nm} (\lfloor N/m \rfloor + i)}{(nm)!} \tag{12}$$

For example if $N = 1000$ and $n = 2, m = 1$, we have that the number of elements is $n.e. \approx 5 \times 10^5$.

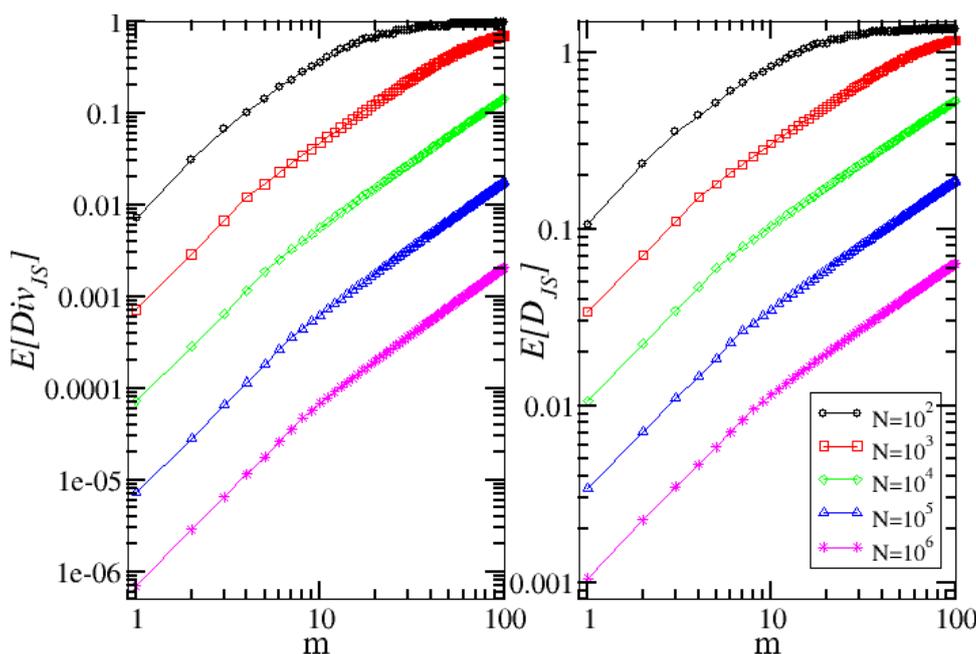
To handle this problem, we compute these expectations by means of Monte Carlo simulations, and we replace ensemble averages with sample averages, *i.e.*, for example:

$$E[\mathcal{D}_{JS}(\mathbf{p}_m, \mathbf{f}_m)] \approx \langle \mathcal{D}_{JS}(\mathbf{p}_m, \mathbf{f}_m) \rangle \equiv \frac{1}{N_r} \sum_{j=1}^{N_r} \mathcal{D}_{JS}(\mathbf{p}_m, \mathbf{f}_m(\mathbf{n}_{m,j})) \tag{13}$$

where $\langle \dots \rangle$ represents the ensemble average and j represents the j -th simulation of the set of N_r .

We first consider the problem of the finite sample bias in the computation of the Jensen–Shannon divergence and distance. Specifically, we compute $E[\mathcal{D}_{JS}(\mathbf{f}_m^1, \mathbf{f}_m^2; N)]$ and $E[Div_{JS}(\mathbf{f}_m^1, \mathbf{f}_m^2; N)]$ as a function of the time series length, N , when the two frequency vectors are taken by two independent realizations of the same binomial model. We study the two information functionals in the range $m = [1, \dots, 100]$ for the values $N = 10^2, 10^3, 10^4, 10^5, 10^6$, as reported in Figure 1.

Figure 1. $E[Div_{JS}(\mathbf{f}_m^1, \mathbf{f}_m^2; N)]$ (left) and $E[\mathcal{D}_{JS}(\mathbf{f}_m^1, \mathbf{f}_m^2; N)]$ (right) for the binomial model as a function of the aggregation scale, m , and for different values of time series length, N . Results are obtained from numerical simulations, and the plots are in log-log scale.



As expected, the bias decreases with N and increases with m . By using the result in Equation (8) for sequences of i.i.d observations, we are able to compute analytically the shape of the initial part of the curve corresponding to the Jensen–Shannon divergence. In fact, in our framework, we should perform the following substitutions in Equation (8), $N \rightarrow N/m$ and $k \rightarrow nm + 1$, and we thus obtain that the scaling of the Jensen–Shannon divergence as a function of N and m for the binomial model is:

$$E[Div_{JS}(\mathbf{f}_m^1, \mathbf{f}_m^2; N)] = \frac{nm^2}{4N \ln(2)} + O\left(\frac{m^2}{N^2}\right) \tag{14}$$

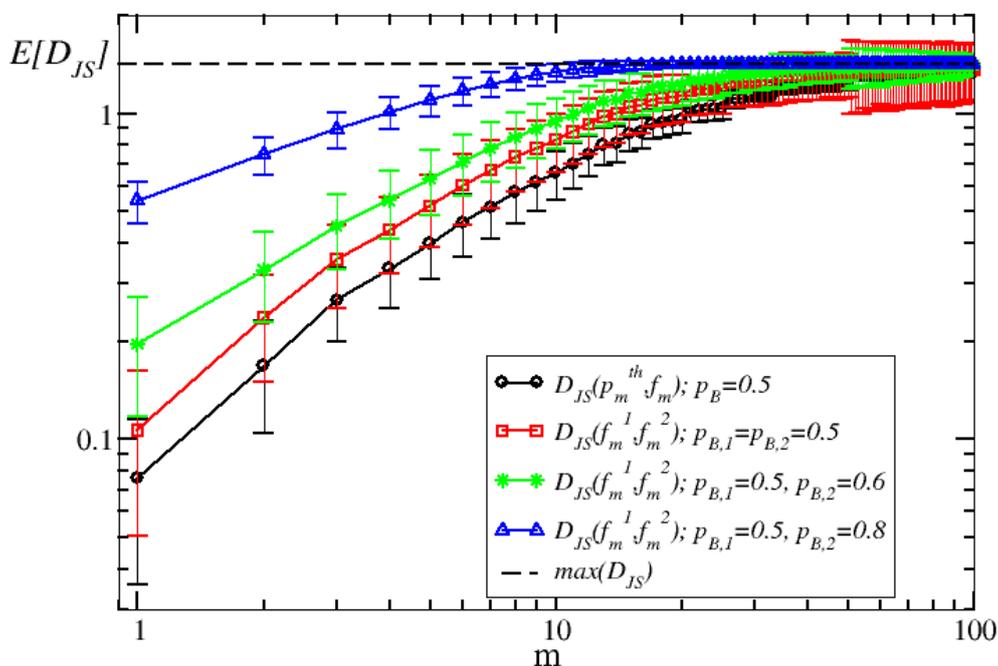
This approximation is more and more valid when N increases, as we can observe in Figure 1. A power-law fit $f(x) = cx^e$ on the initial part of the curves in the case $N = 10^6$ gives

$c = (7.1 \pm 0.1) \times 10^{-7}, e = (2.0 \pm 0.1)$. This is in agreement with the power law of exponent 2 of Equation (14) and with the coefficient $n / (4N \ln(2)) \approx 7.2 \times 10^{-7}$.

In the case of the Jensen–Shannon distance, we do not have any analytic result and limit ourselves to a power-law fit of the initial part of the curve. For the case $N = 10^6$, the fit gives $c = (1.1 \pm 0.1) \times 10^{-3}, e = (1.0 \pm 0.1)$. The initial part of the curve appears to scale linearly with scale m , i.e., $E [Div_{JS}(\mathbf{f}_m^1, \mathbf{f}_m^2; N)] \propto m$.

In order to illustrate how to perform model selection with the Jensen–Shannon distance, we consider the case of an (artificial) sample generated from the binomial model with $p_B = 0.5$. We then compare the Jensen–Shannon distance between this sample and another realization of the model with the same parameter and of a realization of the model with different parameter $p_B \neq 0.5$. As expected, Figure 2 shows that the expected value of the Jensen–Shannon distance between two samples generated by the model with the same parameter is always smaller than the distance between two samples with a different parameter. Moreover, the distance between a sample and the true probability distribution is smaller than the distance between two samples of the same model. This simple observation suggests to us a procedure for selecting models by using the Jensen–Shannon distance.

Figure 2. Expectations and standard deviations of the Jensen–Shannon distance between two samples of the binomial model with the same parameter $p_{B,1} = p_{B,2} = 0.5$ (red squares) and with different parameters (green diamonds and blue triangles). The black circles are an estimation of the Jensen–Shannon distance between a sample and the true model.



Specifically, suppose that $\mathbf{f}_{m=1}^{sam}$ represents the frequency vector computed from the sample of length N , but we do not know the true model that generates it. Suppose we have a statistical model, from which we are able to simulate an output of the same length. In this case, we can compute a frequency, $\mathbf{f}_{m=1}^{mod}$, from our reference model. To compare the two processes at different scales m , we compute the frequencies, \mathbf{f}_m^{sam} and \mathbf{f}_m^{mod} , from the sums of the initial sample over $\lfloor N/m \rfloor$ non-overlapping windows.

If we have different competing models M_1, M_2, \dots , we generate synthetic samples of length N and compute the distances $\mathcal{D}_{JS}(\mathbf{f}_m^{sam}, \mathbf{f}_m^{mod}; l)$, where the index, l , runs on the possible different models. The model that minimizes the Jensen–Shannon distance at different scales m is the model that reproduces the data better. It is clear that even if we had the true model, the minimum distance at different scales will be different from zero. This is because, as we have seen before, $E[D_{JS}(\mathbf{f}_m^1, \mathbf{f}_m^2; N)]$ is larger than zero, even when the two samples come from the real model. As we will see in the financial case in the next section, one can split the real sample into two subsamples of length $N/2$ and compute their Jensen–Shannon distance, to be used as a reference line with respect to the Jensen–Shannon distance between the data and the models.

3. Application to High Frequency Financial Data

In this section, we use the above multiscale procedure, based on the Jensen–Shannon distance, in order to select the best statistical model in the particular case of models describing the high frequency price dynamics of a large tick asset. The models used here were introduced by Curato and Lillo [15], and data refer to NASDAQ (National Association of Securities Dealers Automated) stocks at the time scale of single transactions, traded during July and August, 2009 (see [15] for more details).

3.1. Bid-Ask Spread and Price Dynamics

In financial markets, there are two important prices at each time t : the ask price, $p_{ASK}(t)$, and the bid price, $p_{BID}(t)$. A customer that wants to buy (sell) a certain volume of the stock submits a buy (sell) market order, which is executed at the ask (bid) price, p_{ASK} (p_{BID}). From these two prices, we define the mid-price $p_{mean}(t) = (p_{ASK}(t) + p_{BID}(t))/2$. Our models are defined in transaction time, which is an integer counter of events defined by the execution of a market order, *i.e.*, $t \in \mathbb{N}$. Note that if a market order is executed against several limit orders, our clock advances only by one unit. The price of the order cannot assume arbitrary values, but it can be placed on a grid of fixed values determined by the exchange. The grid step is defined by the tick size, and it is measured in the currency of the asset. The presence of a finite tick size implies that the bid and ask prices can be represented by integer numbers, *i.e.*, $p_{ASK}, p_{BID} \in \mathbb{N}$, for which the unit is represented by the tick size. Our models are defined by the dynamics of two stochastic variables, *i.e.*, mid-price changes $x(t, m)$ between m consecutive transactions and the bid-ask spread, $s(t)$:

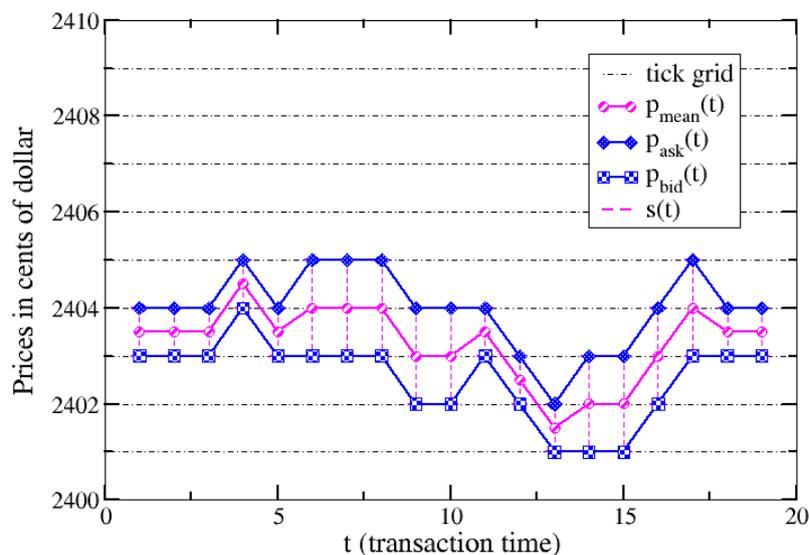
$$\begin{aligned} x(t, m) &= p_{mean}(t + m) - p_{mean}(t) \\ s(t) &= p_{ASK}(t) - p_{BID}(t) \end{aligned} \quad (15)$$

We measure the mid-price changes $x(t)$ in units of half tick size, *i.e.*, $x(t) \in \mathbb{Z}$, and the bid-ask spread, $s(t)$, in units of one tick size, *i.e.*, $s(t) \in \mathbb{N}$. The value of the integer, m , describes the time scale of observation of the price process. Here, we are interested in large tick size assets. Their principal property is that the possible values of spreads and mid-price changes belong to a small set of integer numbers. For example, in the investigated stock, it is $s(t) \in \{1, 2\}$ and $x(t, m = 1) \in \{-2, -1, 0, 1, 2\}$. An example of mid-price and spread dynamics is given in Figure 3.

3.2. Markov Dynamics

We compare three different models for price dynamics in transaction time proposed in [15]. The first model, called $M0$ model, is defined by price changes $x(t)$ that are independent from the spread process, $s(t)$. Instead, in the other two models, *i.e.*, the MS and MS_B models, there is a coupling between the process of price changes, $x(t)$, and the spread process, $s(t)$. We now define the price-change processes relative to the time scale $m = 1$, setting $x(t) = x(t, m = 1)$.

Figure 3. Dynamics of the mid-price, $p_{mean}(t)$, and bid-ask spread, $s(t)$, on the price grid determined by the finite tick size of \$0.01.



M0 model. The model is defined by an i.i.d process for $x(t)$, where the unconditional distribution, $p(x(t))$, reproduces the empirical distribution of price changes. In this case, $p(x(t) | s(t)) = p(x(t))$, *i.e.*, we have independence between the two variables, x and s .

MS model. This model is defined by a particular coupling between the price changes and spread dynamics. We start from the description of the spread process, $s(t)$, because this process will be independent from the process, $x(t)$, whereas $x(t)$ will be the dependent variable.

It is well known that the spread process, $s(t)$, is autocorrelated in time [39,40]. In our models, the spread process, $s(t)$, is represented by a stationary Markov(1) process:

$$P(s(t) = j | s(t - 1) = i, s(t - 2) = k, \dots) = P(s(t) = j | s(t - 1) = i) = p_{ij} \tag{16}$$

where $i, j \in \mathbb{N}$ are spread values. The spread process is described by the two-state transition matrix, $B \in M_{2,2}(\mathbb{R})$:

$$B = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}$$

where the normalization is given by $\sum_{j=1}^2 p_{ij} = 1$. We find [15] that the coupling is not directly defined by the spread process, $s(t)$, but by the kind of transition between spreads. Starting from the

$s(t)$ process, we can define a new stationary Markov(1) process, $z(t)$, that describes the stochastic dynamics of transitions between states $s(t)$ and $s(t + 1)$ as:

$$\begin{aligned}
 z(t) = 1 & \quad \text{if} \quad s(t + 1) = 1, s(t) = 1 \\
 z(t) = 2 & \quad \text{if} \quad s(t + 1) = 2, s(t) = 1 \\
 z(t) = 3 & \quad \text{if} \quad s(t + 1) = 1, s(t) = 2 \\
 z(t) = 4 & \quad \text{if} \quad s(t + 1) = 2, s(t) = 2
 \end{aligned} \tag{17}$$

This Markov(1) process is defined on four possible states and is characterized by the four-state transition matrix, $M \in M_{4,4}(\mathbb{R})$:

$$M = \begin{pmatrix} p_{11} & p_{12} & 0 & 0 \\ 0 & 0 & p_{21} & p_{22} \\ p_{11} & p_{12} & 0 & 0 \\ 0 & 0 & p_{21} & p_{22} \end{pmatrix}$$

where we have several forbidden transitions. For example, the transition $z(t) = 2 \rightarrow z(t + 1) = 1$ is impossible, because it corresponds to the following transitions for spreads: $[s(t) = 1, s(t + 1) = 2] \rightarrow [s(t + 1) = 1, s(t + 2) = 1]$; while, clearly, there is only one value for $s(t + 1)$. We can now define a Markov-Switching model, or Hidden Markov model, for the price changes, $x(t)$:

$$\begin{aligned}
 P(x(t) = \pm 2 | z(t) = 1; \theta) &= \theta_1 \\
 P(x(t) = 0 | z(t) = 1; \theta) &= 1 - 2\theta_1 \\
 P(x(t) = \pm 1 | z(t) = 2; \theta) &= 1/2 \\
 P(x(t) = \pm 1 | z(t) = 3; \theta) &= 1/2 \\
 P(x(t) = \pm 2 | z(t) = 4; \theta) &= \theta_4 \\
 P(x(t) = 0 | z(t) = 4; \theta) &= 1 - 2\theta_4
 \end{aligned} \tag{18}$$

These conditioning rules are imposed by the discreteness of the price grid (see Figure 3 and [15]). In this model, we impose perfect symmetry between positive and negative values of price changes $x(t)$. The model is defined by four parameters, $p_{11}, p_{21}, \theta_1, \theta_4$, that can be estimated from the data.

MS_B model. This model is a limit case of the MS model. In this case, the spread process is an i.i.d Bernoulli process defined by $P(s(t) = 1) = p_B$. Though $s(t)$ is an i.i.d process, $z_B(t)$ is a Markov(1) process defined by:

$$M_B = \begin{pmatrix} p_B & (1 - p_B) & 0 & 0 \\ 0 & 0 & p_B & (1 - p_B) \\ p_B & (1 - p_B) & 0 & 0 \\ 0 & 0 & p_B & (1 - p_B) \end{pmatrix}$$

The conditioning rules for price changes are the same as those of Equation (18). The model now is defined only by three parameters: p_B, θ_1, θ_4 .

For the next section, it is useful to quantify the number of possible states of the variables, $x(t, m)$, for different scales m . For our models, the number of states grows linearly with m , i.e., $k = 1 + 4m$. We stress that we do not have analytic expressions of probability distributions for the process, $x(t, m > 1)$.

This is due to the fact that the model is relatively complicated, also because it is correlated in time and, therefore, not i.i.d.. For scales $m > 1$, we study our models only by Monte Carlo simulations, *i.e.*, we generate a sample of N observations of the processes defined at scales $m = 1$, and then, we study the properties of $y_m = \sum_{j=1}^m x(j)$ on non-overlapping windows of length m .

3.3. Multiscale Model Selection

Our problem is now how to select the model that reproduces the data better. We focus our selection problem on the ability of the models to reproduce the price-change process at different time scales. The selection problem does not involve the spread process, $s(t)$. In this case, we study a sample of $N = 348,253$ price-change observations from the MSFT (Microsoft) stock.

We study this selection problem by means of the concepts developed in Section 2.1. First, we compute the Jensen–Shannon distance between two realizations of the real process. To this end, we divide the sample into two non-overlapping samples, each of length $N/2$, and we compute the two frequency vectors, \mathbf{f}_m^1 and \mathbf{f}_m^2 , for each value of m . We then compute the Jensen–Shannon distance, $\mathcal{D}_{JS}(\mathbf{f}_m^1, \mathbf{f}_m^2; N/2)$. It is clear that this is only one of the possible values of the random variable, \mathcal{D}_{JS} , and we expect that it will be affected by some kind of fluctuations. Then, we generate $N_r = 25$ synthetic samples of length $N/2$ of processes corresponding to our three models. In this way, we compute N_r different frequencies \mathbf{f}_m^{model} that allow us to compute the sample averages:

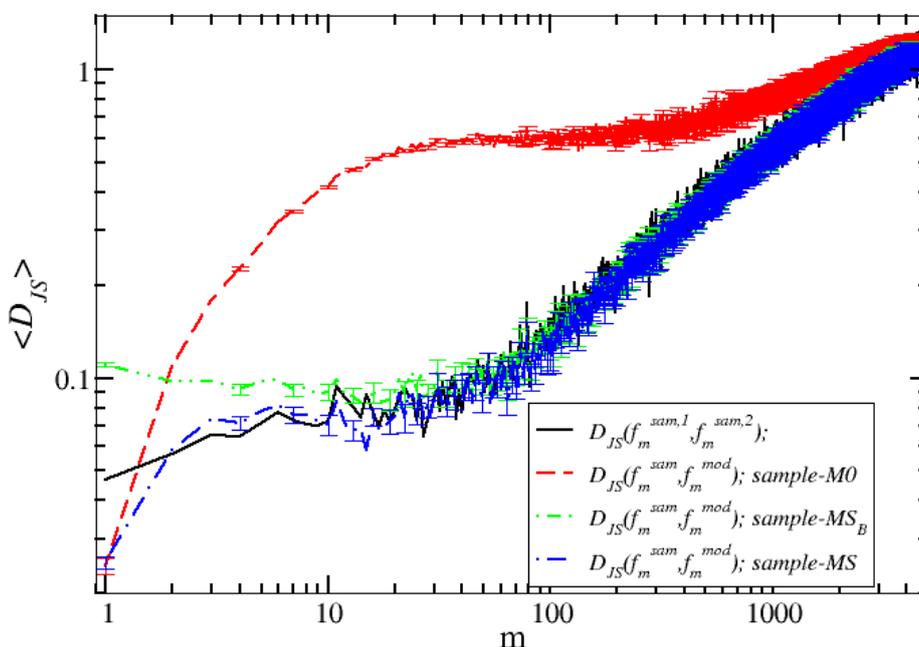
$$\langle \mathcal{D}_{JS}^q(\mathbf{f}_m^{sample}, \mathbf{f}_m^{model}; N/2) \rangle = \frac{1}{N_r} \sum_{i=1}^{N_r} \mathcal{D}_{JS}^q(\mathbf{f}_m^{sample}, \mathbf{f}_m^{model,i}; N/2) \quad (19)$$

from which we can compute a mean and a standard deviation value for the Jensen–Shannon distance for each value of m . The results for the different models are reported in Figure 4. The model that reproduces the empirical data better, *i.e.*, which is closer to $\mathcal{D}_{JS}(\mathbf{f}_m^1, \mathbf{f}_m^2; N/2)$, is the MS model. It is important to notice that also the MS_B model reproduces the empirical data for values of the scale $m > 10$. In fact, for $m > 10$, the MS and MS_B models have the same ability to reproduce the empirical data. The conditioning rules of Equation (18) are critical in order to reproduce the data for values of $m > 10$. The model $M0$, instead, appears to reproduce the data better for the scale of a single transaction, *i.e.*, $m = 1$. This is only the consequence of the fact that the probability distribution of $M0$ models is exactly the same as the empirical distribution of price changes for single transactions, *i.e.*, it reproduces the small asymmetry of the real distribution between positive and negative values of price changes. Instead, MS and MS_B have symmetric distributions for price changes for single transactions. We can observe that the three models reproduce the data for $m > 3,000$ equally well. This corresponds to a real time of the order of one hour, *i.e.*, the daily time scale. This time scale can be interpreted as the one after which the microscopic details of price formation and market microstructure are not relevant anymore in describing the dynamics of the shape of the return distribution. In other words, the coupling of the price-change process and of the bid-ask spread process appears to be the key to understand the dynamics of prices for a large tick stock from the time scale of single transaction to the daily time scale.

Our analysis has been performed by using the Jensen–Shannon distance. However, other distances between probability distributions exist, such as the Kolmogorov–Smirnov distance [14], the Euclidean distance and the Hellinger distance [41]. We have repeated the analysis with these distances, and

we found that its ability to select between competing models is smaller than that of Jensen–Shannon distance. In particular, Kolmogorov–Smirnov distance is able to discriminate models only until the scale $m \approx 100$, which means that for this measure of discrepancy after $m = 100$, the models reproduce the sample distribution in the same way. In our framework, the distance with major discriminant power should be able to discriminate models for high values of the aggregation scale, m . The Hellinger and Euclidean distances, instead, have a discriminant power that is similar to that of Jensen–Shannon distance.

Figure 4. Mean and standard deviation of \mathcal{D}_{JS} between Microsoft data and three models, namely $M0$, MS and MS_B (see the text). The black line is the distance, \mathcal{D}_{JS} , between the two subsamples of the real data obtained by splitting the sample in two. We do not display the error bars for each value of m , but only for 25% of them.



4. Conclusions

One important issue for the study of price dynamics is the selection and validation of a statistical model against the empirical data. Usually, financial time-series models that work well at a fixed time scale do not work comparably well at different time scales. The Jensen–Shannon distance analysis that we have performed enables us to perform an accurate test of goodness of our statistical models and to select among a pool of competing models. We have performed the same model selection procedure with different statistical distances. We find that their power to discriminate between different competing models is not larger than that of Jensen–Shannon distance. Moreover for the Jensen–Shannon distance, we have a good control of the finite sample properties.

Our analysis demonstrates that, for large tick assets, the coupling between mid-price dynamics and spread dynamics is important to account for the mid-price dynamics from the time scale of a single transaction to the time scale of one trading day.

We believe that the described method, based on the Jensen–Shannon distance, could be used also in contexts different from the financial one investigated in this work. This method could be useful each time we want to perform a multiscale test for a model against the empirical data samples.

Acknowledgments

The authors acknowledge partial support by the grant, SNS11LILLB “Price formation, agents heterogeneity, and market efficiency”.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Bouchaud, J.-P.; Potters, M. *Theory of Financial Risks: From Statistical Physics to Risk Management*; Cambridge University Press: New York, NY, USA, 2003.
2. Cont, R. Empirical properties of asset returns: Stylized facts and statistical issues. *Quantit. Financ.* **2001**, *1*, 223–236.
3. Gopikrishnan, P.; Plerou, V.; Amaral, L.A.N.; Meyer, M.; Stanley, H.E. Scaling of the distribution of fluctuations of financial market indices. *Phys. Rev. E* **1999**, *60*, 5305–5316.
4. Mandelbrot, B.B. *Fractals and Scaling in Finance*; Springer: New York, NY, USA, 1997.
5. Cont, R.; Tankov, P. *Financial Modelling with Jump Processes*; Chapman & Hall/CRC Press: Boca-Raton, FL, USA, 2004.
6. Bacry, E.; Delour, J.; Muzy, J.F. Modelling financial time series using multifractal random walks. *Physica A* **2001**, *299*, 84–92.
7. Ding, Z.; Granger, W.J.; Engle, R.F. A long memory property of stock market returns and a new model. *J. Empir. Financ.* **1995**, *1*, 83–93.
8. Micciche, S.; Bonanno, G.; Lillo, F.; Mantegna, R.N. Volatility in financial markets: Stochastic models and empirical results. *Physica A* **2002**, *314*, 756–761.
9. Endres, D.M.; Schindelin J.E. A new metric for probability distributions. *IEEE Trans. Inform. Theor.* **2003**, *49*, 1858–1860.
10. Connor, R.; Cardillo, F.A.; Moss, R.; Rabitti, F. Evaluation of Jensen–Shannon Distance over Sparse Data. In *Similarity Search and Applications*; Brisaboa, N., Ed.; Springer-Verlag Berlin: Heidelberg, Germany, 2013; pp. 163–168.
11. Lin, J. Divergence measures based on the shannon entropy. *IEEE Trans. Inform. Theor.* **1991**, *37*, 145–151.
12. Robert, C.Y.; Rosenbaum, M. A new approach for the dynamics of ultra-high-frequency data: The model with uncertainty zones. *J. Financ. Econometr.* **2011**, *9*, 344–366.
13. Garèche, A.; Disdier, G.; Kockelkoren, J.; Bouchaud, J.P. Fokker-planck description for the queue dynamics of large tick stocks. *Quantit. Financ.* **2012**, *12*, 1395–1419.
14. Conover, W.J. A Kolmogorov goodness-of-fit test for discontinuous distributions. *J. Am. Stat. Assoc.* **1972**, *67*, 591–596.

15. Curato, G.; Lillo, F. Modeling the coupled return-spread high frequency dynamics of large tick assets. **2013**, arXiv: 1310.4539.
16. Basseville, M. Divergence measures for statistical data processing—An annotated bibliography. *Signal Process.* **2013**, *93*, 621–633.
17. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2006.
18. Smith, A.; Naik, P.A.; Tsai, C. Markov-switching model selection using Kullback-Leibler divergence. *J. Econometr.* **2006**, *134*, 553–577.
19. De Domenico, M.; Insolia, A. Entropic approach to multiscale clustering analysis. *Entropy* **2012**, *14*, 865–879.
20. Contreras-Reyes, J.E; Arellano-Valle, R.B. Kullback-Leibler Divergence measure for multivariate skew-normal distributions. *Entropy* **2012**, *14*, 1606–1626.
21. Tumminello, M.; Lillo F.; Mantegna R.N. Kullback-Leibler distance as a measure of the information filtered from multivariate data. *Phys. Rev. E* **2007**, *76*, 031123:1–031123:12.
22. Quiroga, R.Q.; Arnhold, J.; Lehnertz, K.; Grassberger, P. Kulback-Leibler and renormalized entropies: Applications to electroencephalograms of epilepsy patients. *Phys. Rev. E* **2000**, *62*, 8380–8386.
23. Roldán, É.; Parrondo, J.M.R. Entropy production and Kullback-Leibler divergence between stationary trajectories of discrete systems. *Phys. Rev. E* **2012**, *85*, 031129:1–031129:12.
24. Kullback, S. *Information Theory and Statistics*; Dover Publications: New York, NY, USA, 1968.
25. Crooks, G.E.; Sivak, D.A. Measures of trajectory ensemble disparity in nonequilibrium statistical dynamics. *J. Stat. Mech.* **2011**, P06003.
26. Majtey, A.P.; Lamberti, P.W.; Prato, D.P. Jensen–Shannon divergence as a measure of distinguishability between mixed quantum states. *Phys. Rev. A* **2005**, *72*, 052310:1–052310:6.
27. Crooks, G.E.; Measuring thermodynamic length. *Phys. Rev. Lett.* **2007**, *99*, 100602:1–100602:4.
28. Carpi, L.C.; Rosso, O.A.; Saco, P.M.; Ravetti, M.C. Analyzing complex networks evolution trough information theory quantifiers. *Phys. Let. A* **2011**, *375*, 801–804.
29. Chekmarev, S.F. Information entropy as a measure of nonexponentiality of waiting-time distributions. *Phys. Rev. E* **2008**, *78*, 066113:1–066113:7.
30. Felizzi, F.; Comoglio, F. Network-of-queues approach to B-cell-receptor affinity discrimination. *Phys. Rev. E* **2012**, *85*, 061926:1–061926:18.
31. Hosoya, A.; Buchert, T.; Morita, M. Information entropy in cosmology. *Phys. Rev. Lett.* **2004**, *92*, 141302:1–141302:4.
32. Basharin, G.P. On a statistical estimate for the entropy of a sequence of independent random variables. *Theor. Prob. Appl.* **1959**, *4*, 333–338.
33. Herzel, H.; Schmitt, A.O.; Ebeling, W. Finite sample effects in sequence analysis. *Chaos Soliton. Fract.* **1994**, *4*, 97–113.
34. Schürmann, T.; Grassberger, P. Entropy estimation of symbol sequences. *Chaos* **1996**, *6*.
35. Roulston, M.S. Estimating the errors on measured entropy and mutual information. *Physica D* **1999**, *125*, 285–294.

36. Grassberger, P. Finite sample corrections to entropy and dimension estimates. *Phys. Lett. A* **1988**, *128*, 369–373.
37. Grosse, I.; Bernaola-Galván, P.; Carpena, P.; Román-Roldán, R.; Oliver J.; Stanley, H.E. Analysis of symbolic sequences using the Jensen–Shannon divergence. *Phys. Rev. E* **2002**, *65*, 041905:1–041905:16.
38. Johnson, N.L.; Kemp, W.A.; Kotz, S. *Univariate Discrete Distributions*, 3rd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2005.
39. Plerou, V.; Gopikrishnan, P.; Stanley, H.E. Quantifying fluctuations in market liquidity: Analysis of the bid-ask spread. *Phys. Rev. E* **2005**, *71*, 046131:1–046131:8.
40. Ponzi, A.; Lillo, F.; Mantegna, R.N. Market reaction to a bid-ask spread change: A power-law relaxation dynamics. *Phys. Rev. E* **2009**, *80*, 016112:1–016112:12.
41. Guha, S.; McGregor, A.; Venkatasubramanian, S. Streaming and sublinear approximation of entropy and information distances. **2005**, arXiv:cs/0508122.

© 2014 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).