Eur. Phys. J. B (2014) 87: 168
DOI: 10.1140/epjb/e2014-50108-3

THE EUROPEAN PHYSICAL JOURNAL B

Regular Article

Information-theoretic approach to lead-lag effect on financial markets

Paweł Fiedor[a]

Cracow University of Economics, Rakowicka 27, 31-510 Kraków, Poland

Received 16 February 2014 / Received in final form 28 April 2014
Published online 1 August 2014

Abstract. Recently the interest of researchers has shifted from the analysis of synchronous relationships of financial instruments to the analysis of more meaningful asynchronous relationships. Both types of analysis are concentrated mostly on Pearson's correlation coefficient and consequently intraday lead-lag relationships (where one of the variables in a pair is time-lagged) are also associated with them. Under the Efficient-Market Hypothesis such relationships are not possible as all information is embedded in the prices, but in real markets we find such dependencies. In this paper we analyse lead-lag relationships of financial instruments and extend known methodology by using mutual information instead of Pearson's correlation coefficient. Mutual information is not only a more general measure, sensitive to non-linear dependencies, but also can lead to a simpler procedure of statistical validation of links between financial instruments. We analyse lagged relationships using New York Stock Exchange 100 data not only on an intraday level, but also for daily stock returns, which have usually been ignored.

1 Introduction

Financial markets are becoming increasingly complex adaptive systems. Nonetheless, economists lack a fundamental theory behind their complex behaviour even at times when their structure has been much simpler. This lack of theory has many consequences. First, other scientists, notably physicists, can study those systems without worrying about the intricacies of economic theories. Second, it also leads to an assumption that the time series describing stock returns are unpredictable [1]. Within this paradigm the evolution of stock prices can only be explained by random processes. Additionally, the Efficient-Market Hypothesis [2] proposes that all information is reflected in the prices and therefore that it is not possible to predict future prices based on the past (equivalently the prices walk randomly). Weaker variants of the hypothesis state that only past prices are included in the current ones, thus rendering impossible predictions which are based only on the past prices. Such a hypothesis would then mean that there can be no lead-lag effect (where a change in one price at a given time leads to a similar change in another price at a specific later time) on the financial markets, making the analysis in this paper pointless. But the Efficient-Market Hypothesis has been continually disproved in many ways since the 1980s, and in fact the support for it has dwindled among researchers. Particularly researchers analysing NYSE stock returns [3,4] show that the data can be compressed, thus showing that the stock returns are not random, as then no compression would be possible. We have performed similar tests on New York and Warsaw exchanges in the recent past [5]. If the price changes of stocks are not random there then arises a possibility that the data is structured. Thus researchers are encouraged to explore methods of modelling this structure and analysing real-world markets.

The standard assumption that price formation processes are stochastic leaves researchers with a question of whether these processes are independent for different financial instruments, or whether there exist relationships based on known or unknown common economic factors driving these formation processes. Tools developed to model physical systems [6–8] are often used to analyse the interdependencies between financial instruments. The largest effort has been used in understanding correlations in financial markets for daily [9–14] and intraday time scales [15–17]. In the recent years other measures of similarity have also been introduced, including Granger causality analysis [18], partial correlation analysis [19], both of which try to quantify how one financial instrument provides information about another instrument, and mutual information together with mutual information rate [20], both of which aim at including non-linear relationships in the analysis. All of these methods aim for a single goal, that is the discovery of meaningful information in the increasingly complex adaptive systems of financial markets.

The most common analysis uses synchronous correlations of equity returns. Such analyses have shown that financial markets have a nested structure, in which stock returns are driven by a common factor, and stocks

[a] e-mail: s801dok@wizard.uek.krakow.pl

themselves are organised in groups defined by economic sector. The correlations inside those groups are higher than the average pair correlation. One can also find second order groups. The correlations can, of course, be exchanged for another well-defined similarity measure such as mutual information [20]. This is well corroborated, as the same results have been obtained using substantially different methods, ranging from random matrix theory [21], through principal component analysis [22], through hierarchical clustering [9], to correlation-based networks [9,23,24] and mutual information-based networks [20]. The methods developed to construct dependency networks may be grouped into two categories: threshold-based methods and topological methods. Both categories start with a sample similarity measure (correlation matrix, mutual information matrix, etc.). Then using the threshold method a threshold is set on the similarity measure, and a network is constructed so that only links between nodes whose pairwise similarity measure is larger than the threshold are present. With lowering the threshold value a more complex hierarchy emerges, and groups of stocks progressively merge to form larger groups, until they form the whole market. Such threshold networks are very robust with regards to the uncertainty in the similarity measure, but it is difficult to find a single threshold value which could accurately display the nested structure of the similarity matrix of stock returns. Topological methods, on the other hand, construct dependency networks, such as the minimal spanning tree (MST) [9,20,23,24] or the planar maximally-filtered graph (PMFG) [20,25,26]. These are based on the ranking of empirical similarity measures. The resulting networks are intrinsically hierarchical and therefore easy to present as a graph, but this approach is less stable than threshold methods with respect to the statistical uncertainty in the data. Furthermore such an approach does not necessarily present information about the statistical significance of the similarity measures [27].

On the other hand, very few inquires have been performed looking into networks of lagged correlations [28,29]. The above-described methods of constructing dependency networks cannot be easily extended to the analysis of directed lagged correlations or similarity measures in financial markets. The lagged interdependencies in stock returns are quite small even at short time horizons, therefore an analysis is strongly influenced by the statistical uncertainty of the estimation process. The use of topological methods is difficult, as they only take into consideration the ranking of similarity measures, and not their actual values. Thus many links in such a network may indeed be statistically insignificant if we use lagged dependencies. On the other hand, threshold methods are difficult to apply because it is difficult to find an appropriate threshold level. Additionally, these methods use the same threshold for all stock pairs, which is a problem in the analysis of lagged relationships, as the statistical significance of a lagged similarity measure is likely to vary across stocks (for example due to differences in volatility).

In reference [29] a method for filtering a lagged correlation matrix into a network of statistically-validated directed links that takes into account the heterogeneity of stock return distributions has been introduced. This has been done by associating a p-value with each observed lagged-correlation and then setting a threshold on p-values, i.e. setting a level of statistical significance corrected for multiple hypothesis testing. They have applied this method to analyse the structure of lagged relationships between intraday equity returns on US equity markets.

In this paper we extend this analysis in two ways. First, we extend this methodology to include non-linear relationships. Second, we also analyse daily lagged relationships. It is well known that financial markets, and particularly time series describing returns on financial instruments, involve terms that are not first order. There is now strong evidence of the existence of non-linear dynamics in stock returns [30–34], market index returns [35–39], and currency exchange rate changes [30,40–43]. Meanwhile Pearson's correlation coefficient is strictly not sensitive to any non-linear dependencies. Therefore an analysis using correlation can miss important features of any dynamical system, particularly financial markets. Thus we find the assumptions that only linear dependencies are relevant in financial markets, found in the hierarchical clustering methodologies used in econophysics, unsupportable.

We contrast correlation coefficient with the measure of mutual information (I_S) [44], which is more general. In fact $I_S = 0$ if and only if the two studied random variable are strictly independent. Mutual information is then a natural measure which can be used to extend the similarity measure to make it sensitive to non-linear dependencies, and has been successfully used in some applications [45–47]. Recently we have used it in the creation of dependency networks on financial markets [20]. Mutual information measures how much information two studied stochastic processes share and has been used to enhance the understanding of the brain in neuroscience [48–50], to characterise [51,52] and model various complex and chaotic systems [53–55], and also to quantify the information capacity of a communication system [56]. Additionally mutual information provides a convenient way to identify the most relevant variables with which to describe the behaviour of a complex system [57], which is of paramount importance in modelling those systems, and indeed to the methodology of this paper [20].

Furthermore, we have found in our earlier studies [5] that while intraday stock returns deviate from the Efficient-Market Hypothesis much more strongly than daily returns, the latter themselves are not fully random processes, thus we will also look into lead-lag relationships in the daily stock returns. We believe that lead-lag effect will be much smaller in daily stock returns, but it may not be negligible nonetheless.

The paper is organised as follows. In Section 2 we present the method used to filter and validate statistically significant lagged correlations and introduce a method for statistical validation of significant mutual information between financial instruments. In Section 3 we analyse the structure of NYSE at different frequencies using the

presented methodology. In Section 4 we discuss the results. In Section 5 we conclude the study.

2 Methods

Here we present the methodology of statistically validating lagged correlations for the purpose of network analysis presented in reference [29]. On this basis we will present our extended methodology, which includes non-linear dependencies. For this purpose we will also need to define mutual information, its properties, and estimators.

Curme et al. [29] begin the analysis by calculating the matrix of logarithmic returns over given intraday time-horizons. Let us denote the most recent price for stock n occurring on or before time t during the studied period by $p_n(t)$. The opening price of the stock is defined as the price of its first transaction of the studied period. Additionally, τ is the time horizon. Then for each stock the logarithmic returns are sampled,

$$r_{n,t} \equiv \log\left(p_n(t)\right) - \log\left(p_n(t-\tau)\right), \tag{1}$$

every τ minutes (days, seconds) throughout the studied period. These time series constitute columns in a matrix R. Then R is filtered into two matrices, A and B, in which returns during the last period of λ are excluded from A and returns during the first period of λ are excluded from B (thus λ denotes the lag). From these matrices an empirical lagged correlation matrix C is constructed using the Pearson correlation coefficient of columns of A and B,

$$C_{m,n} = \frac{1}{T-1}\sum_{i=1}^{T}\frac{\left(A_{m,i} - \langle A_m\rangle\right)\left(B_{n,i} - \langle B_n\rangle\right)}{\sigma_m \sigma_n}, \tag{2}$$

where $\langle A_m\rangle$ and σ_m are the mean and sample standard deviation, respectively, of column m of A, and T is the number of rows in A (and B). Curme et al. [29] set the lag λ to be one time horizon τ.

The matrix C can be seen as a weighted adjacency matrix for a fully connected, directed graph. Such matrix needs to be filtered, and to find a threshold of statistical significance Curme et al. [29] apply a shuffling technique [58]. The rows of A are shuffled repeatedly without replacement in order to create a large number of surrogate time series of returns. After each shuffling the lagged correlation matrix is recalculated as $\widetilde{C}$ and compared to the empirical matrix C. For each shuffling there is an independent realisation of $\widetilde{C}$. Then matrices U and D are constructed, where $U_{m,n}$ is the number of realisations for which $\widetilde{C}_{m,n} \geq C_{m,n}$, and $D_{m,n}$ is the number of realisations for which $\widetilde{C}_{m,n} \leq C_{m,n}$.

From matrix U a one-tailed p-value is associated with all positive correlations as the probability of observing a correlation that is equal to or higher than the empirically-measured correlation. Similarly, from D a one-tailed p-value is associated with all negative correlations. Curme et al. [29] set the threshold at the standard $p = 0.01$. The statistical threshold must be adjusted to account for multiple comparisons. Curme et al. [29] use the conservative Bonferroni correction and a less conservative FRD adjustment which both depend on the sample size of N stocks. In particular Bonferroni correction works as follows: p/N^2. For $N = 100$ it gives $0.01/100^2$, thus in such case a construction of 10^6 independently shuffled surrogate time series is required. If $U_{m,n} = 0$ then a statistically-validated positive link from stock m to stock n ($p = 0.01$, Bonferroni correction) can be associated. Likewise, if $D_{m,n} = 0$ a statistically-validated negative link from stock m to stock n is associated. In this way Curme et al. [29] construct the Bonferroni network [59].

Curme et al. [29] also construct a less constrained network based on False Discovery Rate protocol [60]. In practice, in the FDR network the threshold for the matrices U or D is not zero but a positive integer. From this threshold the links in C can be filtered to construct the FDR network [59]. The Bonferroni network is a subgraph of the FDR network. This method makes no assumptions about the return distributions, and also imposes no topological constraints on the Bonferroni or FDR networks [29].

Since this method only analyses strictly linear relationships we use mutual information instead of Pearson's correlation coefficient. To extend such measure to include non-linear dependencies we propose to base the topological arrangement of the nodes in a network on the mutual information. Mutual information is most often defined in the context of Shannon's entropy [61], which is a measure of uncertainty of a random variable X:

$$H(X) = -\sum_i p\left(x_i\right)\log_2 p\left(x_i\right) \tag{3}$$

summed over all possible outcomes $\{x_i\}$ with respective probabilities of $p(x_i)$. Joint (X, Y) and conditional $H(X|Y)$ entropies are also defined for two variables.

We can also define mutual information in Shannon's sense [61]. For two discrete random variables X and Y mutual information between them is defined as:

$$I_S(X,Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}, \tag{4}$$

where $p(x, y)$ is the joint probability distribution function of X and Y and $p(x)$ and $p(y)$ are the marginal probability distributions. For continuous variables the definition is analogous using probability density functions. Equivalently using entropy mutual information is defined as:

$$I_S(X,Y) = H(X) + H(Y) - H(X,Y). \tag{5}$$

Mutual information measures information shared between the two variables, therefore covering both linear and non-linear dependencies, hence using it to describe dependencies in financial markets seems natural. Mutual information is non-negative and $I_S(X, X) = H(X)$. We also note that for easy estimation we need discrete data, while stock returns are not discrete, thus we need to discretise them. For discussion of this step see below and [5,20].

We also need an estimator of entropy for practical purposes. There is a large number of estimators and a presentation of these can be found in [62–66]. In this study we use the Schurmann-Grassberger estimate of the entropy of a Dirichlet probability distribution, which is thought to be the best choice for most applications [67]. The Schurmann-Grassberger estimator is a Bayesian parametric procedure which assumes samples distributed following a Dirichlet distribution:

$$\hat{H}(X) = \frac{1}{m+|\chi|N}\sum_{x\in\chi}(\#(x)+N)(\psi(m+|\chi|N+1) - \psi(\#(x)+N+1)), \quad (6)$$

where $\#(x)$ is the number of data points having value x, $|\chi|$ is the number of bins from the discretisation step, m is the sample size, and $\psi(z) = d\ln\Gamma(z)/\mathrm{d}z$ is the digamma function. The Schurmann-Grassberger estimator assumes $N = 1/|\chi|$ as the prior [68].

Based on this definition we proceed with the method presented in [29] only exchanging correlation coefficient with mutual information. Since mutual information does not distinguish between positive and negative relationships we do not need both U and D and can settle with U. This is not a problem as in this analysis the direction of the relationship is not particularly important and can be easily found anyway. Similar analyses have been performed outside of economic systems [69]. Nonetheless, a less computationally expensive method can be presented, without introducing very strong assumptions. It has been shown that the mutual information between independent random variables (X & Y), when estimated from relative frequencies, follows a very good approximation of a gamma distribution with parameters $a = (|X|-1)(|Y|-1)/2$ and $b = 1/(N\ln 2)$ [70,71]:

$$I_S(X,Y) \sim \Gamma\left(\frac{1}{2}(|X|-1)(|Y|-1), \frac{1}{N\ln 2}\right), \quad (7)$$

where N is the sample size and $|X|$ and $|Y|$ denote the numbers of realizations of the random variables X and Y.

Here we briefly explain why that is the case. Using the natural logarithm in the entropy expression we can expand the expression for mutual information $I_S(X,Y)$ into a Taylor series about expansion point $p_{XY} \equiv p_X p_Y$ and obtain:

$$I_S(X,Y) \approx \frac{1}{2}\sum_x\sum_y \frac{(p(x,y)-p(x)p(y))^2}{p(x)p(y)}. \quad (8)$$

This expression relates to the χ^2 test with the same constant factor of $2N$. The direct proof that the above has a gamma distribution is beyond the scope of this paper and will not be presented. However, the same fact can be easily derived from knowing the χ^2 test variable follows a χ^2 distribution (given the null hypothesis is true). Since $I_S = (\chi^2)/(2N\ln 2)$, we can scale the χ^2 distribution by the factor $2N\ln 2$ and obtain a gamma distribution [70,71].

Therefore, to determine the significance of $I_S(A_m, B_n)$ from a sample study of length N at a significance level p, we check the condition:

$$I_S(A_m,B_n) \geq \Gamma_{1-p}\left(\frac{1}{2}(|A_m|-1)(|B_n|-1), \frac{1}{N\ln 2}\right), \quad (9)$$

where $\Gamma_{1-p}(a,b)$ denotes the $(1-p)$-quantile of the gamma distribution. This is sound as under the null hypothesis A_m and B_n are independent. As in reference [29], we need to adjust p using Bonferroni or any other less-conservative correction. We will use this method instead of shuffling, as the latter has already been analysed in reference [29]. Both methods should give reasonably similar results.

3 Materials and results

To find mutual information-based lagged relationships in practice we have taken log returns for 98 securities out of 100 which constitute the NYSE 100, excluding two with incomplete data. These log returns are intraday (one-minute intervals). The data cover 15 days between the 21st October 2013 and the 8th of November 2013. The choice of data length as much smaller than used in reference [29] is explained in two ways. First, for empirical applications it is often required to see fast dynamics and not dynamics evolving over decades. Second, the choice of data spanning over many years would raise questions about the homogeneousness of the studied sample. Additionally, we note that our dataset has a length of over 3000 intervals, which is sufficient. To analyse daily relationships we also look at the daily price time series of 91 securities traded on the New York Stock Exchange (NYSE100) (the 9 missing stocks were excluded due to missing data). The data has been downloaded from Google Finance database[1] and was up to date as of the 11th of November 2013, going back 10 years. The data is transformed in the standard way for analysing price movements, that is so that the data points are the log ratios between consecutive closing prices for a given period, as defined above, and those data points are, for the purpose of estimating mutual information, discretised into 4 distinct states (binary discretisation would discard volatility). The states represent equal parts, so each state are assigned the same number of data points. This design means that the model has no unnecessary parameters and proved to be very efficient [5,72,73]. The choice of quartiles is largely irrelevant (equivalently one can choose 8 or 16 bins), see the discussion in reference [5].

We have set the p-value to 0.01 and corrected it using conservative Bonferroni correction obtaining the corrected p-value (α) of roughly $\alpha = 10^{-6}$. As Bonferroni correction is thought to be excessive we also show the results for $\alpha = 10^{-5}$, $\alpha = 10^{-4}$ and $\alpha = 10^{-3}$. The first gives very conservative results, while the last choice will be very relaxed in testing the statistical significance. We use the appropriate gamma distribution for the validation of mutual information. Moreover, while Curme et al. [29] set λ to be equal to τ, we will first use the same setup and compare

[1] http://www.google.com/finance/

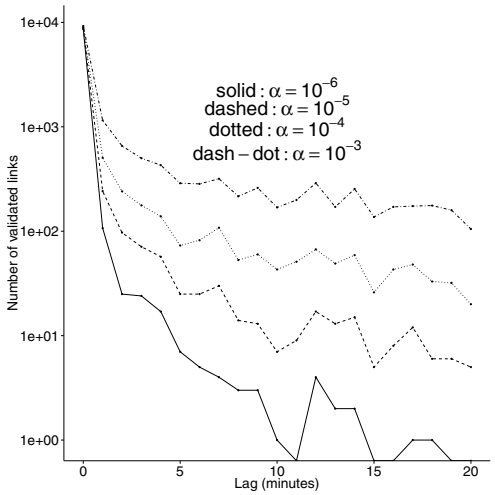

Fig. 1. Number of pairs of stocks with statistically significant correlation-based links for different values of lag λ (equal to price sampling frequency τ). Each line validates the correlation coefficient at the specified adjusted p-value for multiple comparisons.

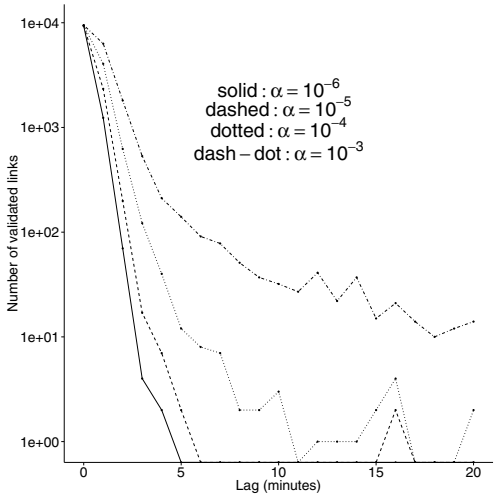

Fig. 2. Number of pairs of stocks with statistically significant mutual information-based links for different values of lag λ (equal to price sampling frequency τ). Each line validates the mutual information at the specified adjusted p-value for multiple comparisons.

both approaches, but then also set τ to be equal to the interval in the data (1 min or 1 day), and only use λ as a variable. We find this setup more informative than the one used in reference [29], since traders are usually interested in short-term changes. As in reference [29], we impose no topological restraints on the networks.

In Figure 1 we present the number of validated (statistically significant) correlation-based links between pairs of stocks for a given time lag λ equal to price sampling frequency τ for intraday stock returns, based on the methodology of reference [29]. In Figure 2 we present the number of validated (statistically significant) mutual information-based links for a given shift of λ equal to price sampling frequency τ for intraday stock returns, based on

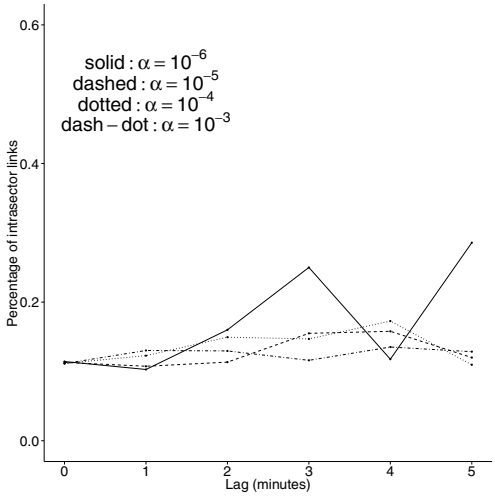

Fig. 3. Percentage of statistically significant correlation-based links between stocks belonging to the same economic sector in all statistically significant links for different values of lag λ (equal to price sampling frequency τ). Each line validates the correlation coefficient at the specified adjusted p-value for multiple comparisons.

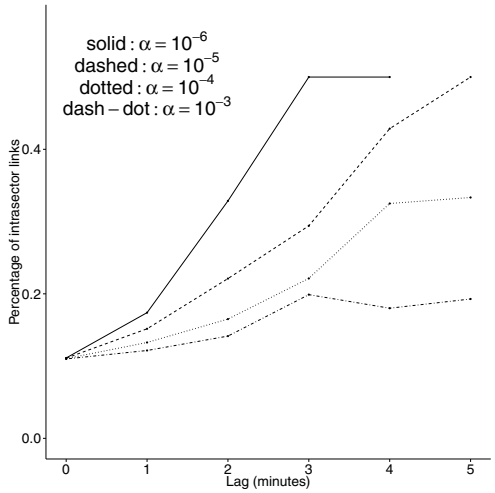

Fig. 4. Percentage of statistically significant mutual information-based links between stocks belonging to the same economic sector in all statistically significant links for different values of lag λ (equal to price sampling frequency τ). Each line validates the mutual information at the specified adjusted p-value for multiple comparisons.

the presented methodology. Note that for $\lambda = 0$ we create a synchronous network, and that while the significance of Pearson's correlation coefficients was determined by a shuffling technique, the statistical significance of mutual information was determined using percentiles of a gamma distribution. In these, solid lines show conservative results which test the statistical significance at an adjusted p-value $\alpha = 10^{-6}$ (Bonferroni network), while dashed, dotted, and dash-dot lines present results at less-conservative adjusted p-values of 10^{-5}, 10^{-4}, and 10^{-3}, respectively; this layout remains the same for Figures 1–6. To further compare these approaches we show what percentage of the

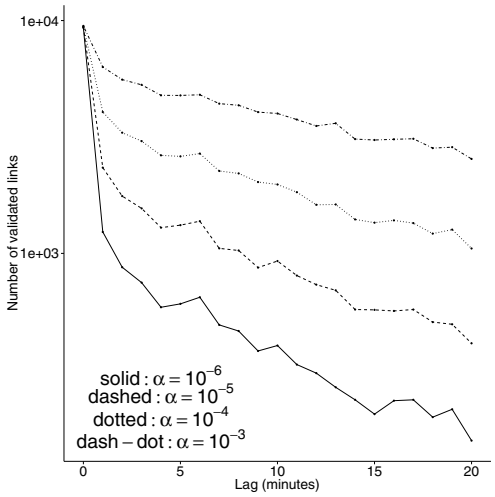

Fig. 5. Number of pairs of stocks with statistically significant mutual information-based links for different values of lag λ (with price sampling frequency τ set to one minute). Each line validates the mutual information at the specified adjusted p-value for multiple comparisons.

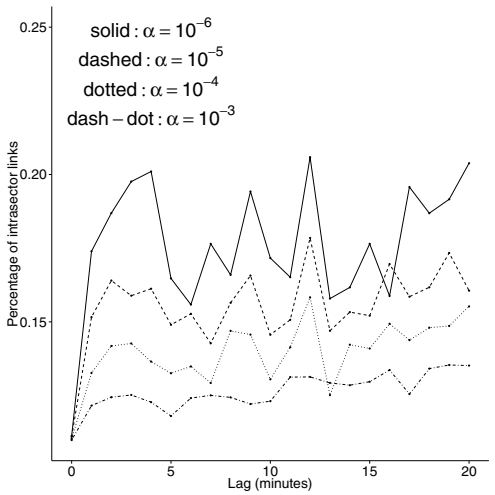

Fig. 6. Percentage of statistically significant mutual information-based links between stocks belonging to the same economic sector in all statistically significant links for different values of lag λ (with price sampling frequency τ set to one minute). Each line validates the mutual information at specified adjusted p-value for multiple comparisons.

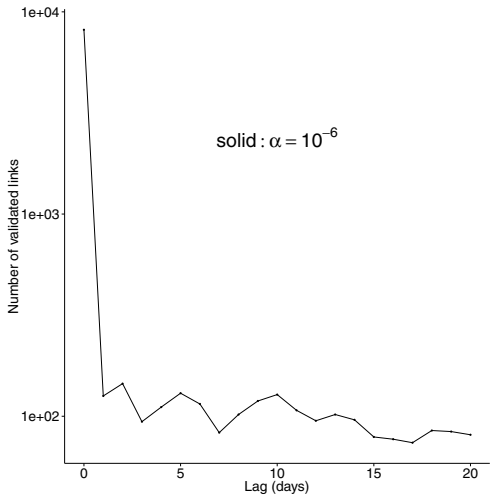

Fig. 7. Number of pairs of stocks with statistically significant mutual information-based links for different values of lag λ (with price sampling frequency τ set to one day), at an adjusted p-value of 10^{-6}.

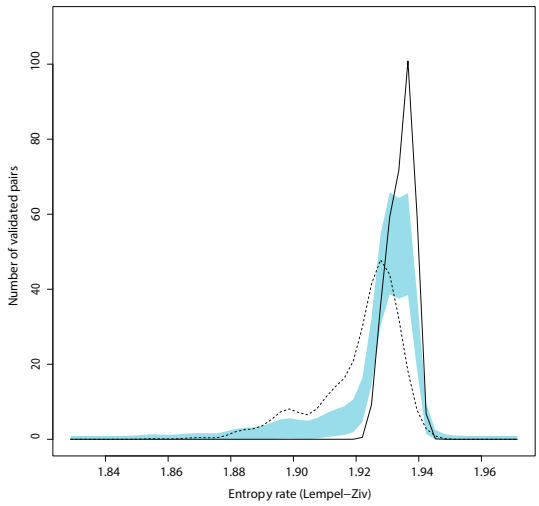

Fig. 8. Distribution of the average of two Shannon's entropy rates within pairs of stocks with statistically significant (dashed line) and statistically insignificant (solid lines) mutual information-based links, with no lag ($\lambda = 0$), at an adjusted p-value of 10^{-6}. The band corresponds to the permutation test of equality of both distributions.

validated links are links within the same economic sector (as defined by NYSE). We present this in Figure 3 for correlation and Figure 4 for mutual information, for varying lag ($\lambda = \tau$). In Figure 5 we present the number of validated mutual information-based links for a given shift λ for intraday stock returns, this time with the price sampling frequency τ set to one minute for all values of λ. The percentage of the validated links within a sector for this setup is presented in Figure 6. The networks themselves are less illustrative and have been ignored.

In Figure 7 we present the number of validated mutual information-based links for a given shift of λ, and price sampling frequency τ of one day, for daily stock returns. Note that for $\lambda = 0$ we create a synchronous network. In Figures 8 and 9 we show how the validation of links is connected with the entropy rate of the underlying time series (average for the two stocks in a link) for varying values of λ (dashed line denotes significant links, solid line denotes other links). In other words, we show the distributions of Shannon's entropy rates within two groups, one consisting of pairs of stocks with statistically significant links, and another without such links. We calculate entropy rate using the same data, with the Lempel-Ziv algorithm; see reference [5] for details. The entropy rate measures the average information content produced by an information

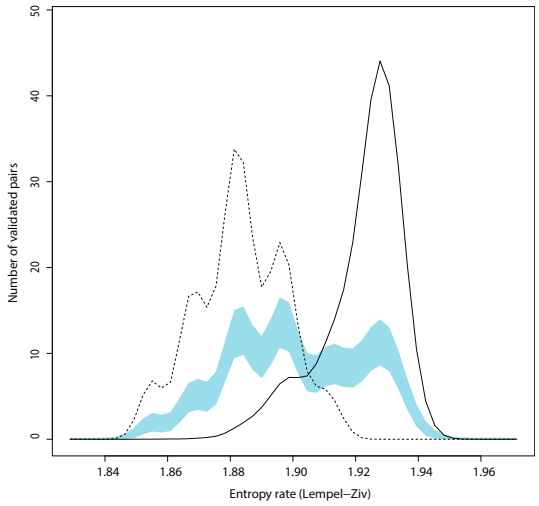

Fig. 9. Distribution of the average of two Shannon's entropy rates within pairs of stocks with statistically significant (dashed line) and statistically insignificant (solid lines) mutual information-based links, with lag of 1 day ($\lambda = 1$), at an adjusted p-value of 10^{-6}. The band corresponds to the permutation test of equality of both distributions.

source per unit of time, thus the larger the value of the entropy rate the more random, or less predictable, the price formation process is, in this particular case 2 being the theoretical maximum for fully random processes and 0 being the minimum for fully predictable ones. We note that the results are virtually identical for λ between 1 and 20, thus only two examples have been presented.

4 Discussion

The discussion is divided into two parts. In the intraday analysis we want to show that the methodology based on mutual information gives good results compared to the methodology based on Pearson's correlation, and in fact we also want to show that they are improved. In the daily analysis we want to show that even at this level there are some inefficiencies worth looking into.

We start with the intraday analysis. At lag $\lambda = 0$, or in the synchronous analysis, there are no arbitrage opportunities, since the price changes for synchronised stocks happen simultaneously. There is then no market pressure to dissipate such synchronisation. But for lag $\lambda > 0$, statistically significant relationships, which we study and call lead-lag effects, constitute arbitrage opportunities, as knowing one price change may allow an analyst to know another price change at a later time, thus creating a possibility for profit. In liquid markets we would therefore expect a pressure for these opportunities to dissipate very quickly, as market participants quickly use them for profit. In other words, we expect the number of statistically significant lead-lag links to decay very fast with increasing lag λ. And when we compare the results obtained with our methodology with the results based on the methodology from reference [29] for $\tau = \lambda$, we see in Figures 1 and 2 that when using the mutual information the number of statistically significant links decays much faster with increasing time lag than it does in results based on correlation. In fact for conservative Bonferroni corrections of p-value (solid lines) the lead-lag effect dissipates for mutual information after 5 min, but for correlations only after 10 min. We also see that if the p-value is only weakly adjusted for the multiple comparisons ($\alpha = 10^{-3}$) the results seem to contain a lot of statistical uncertainty, which we see as small numbers of validated links even for large values of λ. On this basis we find that both methodologies perform well in validating the lead-lag relationships between financial instruments, with the one based on mutual information perhaps being closer to our expectations. We need another way to analyse them to confirm that the difference between the two approaches is meaningful and does not result from a meaningless shift. In other words, we want to show that our methodology is not only different, but also more accurate. It is well corroborated that financial markets are structured according to economic sectors, thus we would expect a good method to find a larger than average portion of intrasector links among the validated links. In Figures 3 and 4 we can see that while correlation-based approach does not lead to the discovery of links which are strongly interconnected based on economic sectors (the percentage is roughly the same for $\lambda = 0$ and $\lambda > 0$), the analysis using mutual information finds relationships which are more often within a sector (the percentage of intrasector links grows with the time lag, and also is clearly dependent on the adjusted p-value α, which is not the case for correlation-based results), which is in better agreement with other studies mentioned in Section 1. This corroborates the usefulness of our methodology, and proves the results are meaningful and not merely different.

Further, we used a different setup with price sampling frequency τ equal to one minute. There are two reasons for this. First, we do not see a point in artificially decreasing the price sampling resolution with increasing time lag. Second, we believe that such an analysis may be more interesting to the market practitioners. For $\tau = 1$ we would expect more statistically significant links, as the frequency is higher, and thus the dynamics faster. And indeed we find that the number of statistically validated links decreases much more slowly for this setup for all values of α, as can be seen on Figure 5. This hints that there is a significant amount of inefficiency in the market at higher price sampling resolutions, which means that there are potentially a lot of arbitrage opportunities, when the market is analysed at this faster pace. We also find that these links are to a significant degree based on economic sectors (as compared to unconstrained networks), and that this degree depends on the adjusted p-value α, as can be seen in Figure 6. We thus conclude that the market is quite far from the Efficient-Market Hypothesis at such small intervals, which is corroborated not only by reference [29], but also studies not using network approaches [5].

Now we turn to the analysis of daily stock returns. It is often ignored, as studies show that daily stock returns are much closer to being random and ruled by

the Efficient-Market Hypothesis than intraday stock returns [5], thus researchers expect to find no significant time lagged relationships. But we see in Figure 7 that while there is an enormous drop of the number of validated links between synchronous ($\lambda = 0$) and asynchronous ($\lambda > 0$) networks, there is still a large number of links (around 100) which are present even at large values of λ. Curious as to whether these result stem from statistical uncertainty and noise, or are meaningful, we compare the predictability of the underlying price formation processes between two groups: the validated pairs (dotted lines) and the non-validated pairs (solid lines), as can be seen in Figures 8 and 9 in the form of kernel densities of Shannon's entropy rates, for the values of λ between 0 and 1 (the results are virtually the same for λ between 1 and 20). In Figure 8 we see that in synchronous networks there is only a small difference in predictability between the two groups. But we also find that the stock returns involved in validated pairs in asynchronous networks are on average significantly more predictable (in terms of Shannon's information entropy) than the ones not involved in validated pairs (the reference band presented is associated with the permutation test for equality), as can be seen in Figure 9. We are therefore inclined to say that these links are not strictly a noise in the data, but present a serious deviation from the Efficient-Market Hypothesis in the daily stock returns for certain stocks. To investigate how such inefficiencies arise, we have also calculated Shannon's entropy rates for a binary alphabet (instead of the alphabet with four letters used previously). Such a setup ignores volatility levels. We have found that the difference between groups seen in Figure 9 is largely reduced when entropy rates are calculated using prices discretised with a binary alphabet, thus we believe that these inefficiencies (lead-lag effects) are due to spatiotemporal synchronisation of volatility on the financial markets. This inquiry is non-trivial however and will require further exhaustive studies to fully grasp the meaning of this violation of the assumption of market effectiveness.

5 Conclusions

In this paper we have presented a methodology for statistically validating lead-lag relationships between financial instruments, which is able to account for non-linear dependencies in the financial markets, and find statistical significance of these dependencies in much simpler and immensely less computationally expensive manner (especially for large sets of stocks) than previous methodologies. We have also applied this methodology to analysing daily and intraday price changes for NYSE 100 stocks, and have found it to perform well. While we cannot directly show the inclusion of the non-linear relationships since we do not know the financial market's structure a priori, we were able to show that the results are different and in some respects better than those obtained using previous methodologies. The results obtained using mutual information appear to be more meaningful than those obtained using Pearson's correlation, since they decay quicker with increasing time lag, and provide significantly larger proportion of intrasector links. While the results for intraday data are not surprising, the results for daily data show that there are statistically validated links, which we would not expect in the daily price changes, and that these persist even for large values of time lag and cannot be easily explained as noise in the data. An inquiry using Shannon's entropy rates leads us to believe that these are due to spatiotemporal synchronisation of volatility, but this problem will require further exhaustive studies. Further studies should also be performed to analyse the usefulness and robustness of this methodology on other markets, both geographically (other world markets) and objectively (currency exchange rates, stock indices, etc.). A more exhaustive study with varying lag parameters (τ and λ) should also be performed to further understand the deviation of the Efficient-Market Hypothesis on different time scales.

References

1. P.A. Samuelson, Ind. Manage. Rev. **6**, 41 (1965)
2. J. Tobin, J. Money Credit Bank. **1**, 15 (1969)
3. A. Lo, A. MacKinlay, Rev. Finance Stud. **1**, 41 (1988)
4. A. Shmilovici, Y. Alon-Brimer, S. Hauser, Comput. Econom. **22**, 273 (2003)
5. P. Fiedor, Frequency Effects on Predictability of Stock Returns, in *Proceedings of the IEEE Computational Intelligence for Financial Engineering & Economics 2014*, edited by A. Serguieva, D. Maringer, V. Palade, R.J. Almeida (IEEE, London, 2014), pp. 247–254
6. B.B. Mandelbrot, J. Business **36**, 394 (1963)
7. L.P. Kadanoff, Simulation **16**, 261 (1971)
8. R.N. Mantegna, Physica A **179**, 232 (1991)
9. R. Mantegna, Eur. Phys. J. B **11**, 193 (1999)
10. P. Cizeau, M. Potters, J. Bouchaud, Quant. Finance **1**, 217 (2001)
11. K. Forbes, R. Rigobon, J. Finance **57**, 2223 (2002)
12. B. Podobnik, H. Stanley, Phys. Rev. Lett. **100**, 084102 (2008)
13. T. Aste, W. Shaw, T.D. Matteo, New J. Phys. **12**, 085009 (2010)
14. D. Kenett, T. Preis, G. Gur-Gershgoren, E. Ben-Jacob, Europhys. Lett. **99**, 38001 (2012)
15. G. Bonanno, F. Lillo, R. Mantegna, Quant. Finance **1**, 96 (2001)
16. M. Tumminello, T.D. Matteo, T. Aste, R. Mantegna, Eur. Phys. J. B **55**, 209 (2007)
17. M. Munnix, R. Schafer, T. Guhr, Physica A **389**, 4828 (2010)
18. M. Billio, M. Getmansky, A. Lo, L. Pelizzon, J. Finance Econ. **104**, 535 (2012)
19. D. Kenett, M. Tumminello, A. Madi, G. Gur-Gershgoren, R. Mantegna, E. Ben-Jacob, PLoS ONE **5**, e15032 (2010)
20. P. Fiedor, Phys. Rev. E **89**, 052801 (2014)
21. L. Laloux, P. Cizeau, M. Potters, J. Bouchaud, Int. J. Theoret. Appl. Finance **3**, 391 (2000)
22. D. Fenn, M. Porter, S. Williams, M. McDonald, N. Johnson, N. Jones, Phys. Rev. E **84**, 026109 (2011)
23. G. Bonanno, G. Caldarelli, F. Lillo, R. Mantegna, Phys. Rev. E **68**, 046130 (2003)
24. J. Onnela, A. Chakraborti, K. Kaski, J. Kertesz, Physica A **324**, 247 (2003)

25. M. Tumminello, T. Aste, T.D. Matteo, R.N. Mantegna, Proc. Natl. Acad. Sci. USA **102**, 10421 (2005)
26. M. Tumminello, T. Aste, T.D. Matteo, R.N. Mantegna, Eur. Phys. J. B **55**, 209 (2007)
27. M. Tumminello, C. Coronnello, F. Lillo, S. Micciche, R. Mantegna, Int. J. Bifurcat. Chaos **17**, 2319 (2007)
28. N. Huth, F. Abergel, `arXiv:1111.7103` (2011)
29. C. Curme, M. Tumminello, R. Mantegna, H. Stanley, D. Kenett, `arXiv:1401.0462` (2014)
30. W.A. Brock, D.A. Hsieh, B. LeBaron, *Nonlinear Dynamics, Chaos, and Instability. Statistical Theory and Economic Evidence* (MIT Press, Cambridge, 1991)
31. M. Qi, J. Bus. Econ. Stat. **17**, 419 (1999)
32. D. McMillan, Int. Rev. Econ. Finance **10**, 353 (2001)
33. D. Sornette, J. Andersen, Int. J. Mod. Phys. C **13**, 171 (2002)
34. K. Oh, K. Kim, Expert Syst. Appl. **22**, 249 (2002)
35. P.H. Franses, D.V. Dijk, J. Forecasting **15**, 229 (1996)
36. A. Abhyankar, L. Copeland, W. Wong, Econ. J. **105**, 864 (1995)
37. P. Chen, Stud. Nonlinear Dyn. Econom. **1** (1996)
38. A. Abhyankar, L. Copeland, W. Wong, J. Bus. Econ. Stat. **15**, 1 (1997)
39. P.A. Ammermann, D.M. Patterson, Pacific-Basin Finance Journal **11**, 175 (2003)
40. D. Hsieh, J. Business **62**, 339 (1989)
41. R. Meese, A. Rose, Rev. Econ. Stud. **58**, 603 (1991)
42. C. Brooks, Appl. Finance Econ. **6**, 307 (1996)
43. M. Qi, Y. Wu, J. Empir. Finance **10**, 623 (2003)
44. T. Cover, J. Thomas, *Elements of Information Theory* (John Wiley & Sons, 1991)
45. F. Zhou, J. He, W. Zhong, Mutual Information based Minimum Spanning Trees Model for Selecting Discriminative Genes, in *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering* (2007), pp. 1051–1055
46. F. Zhou, J. He, W. Zhong, Y. Pan, Int. J. Comput. Biol. Drug Des. **2**, 187 (2009)
47. A.C. Muller, S. Nowozin, C.H. Lampert, in *Pattern Recognition* (Springer, Berlin, 2012), Chap. Information Theoretic Clustering Using Minimum Spanning Trees
48. O. Sporns, D.R. Chialvo, M. Kaiser, C.C. Hilgetag, Trends Cogn. Sci. **8**, 418 (2004)
49. N. Brenner, O. Agam, W. Bialek, R. de Ruyter van Steveninck, Phys. Rev. Lett. **81**, 4000 (1998)
50. N. Brenner, O. Agam, W. Bialek, R. de Ruyter van Steveninck, Phys. Rev. E **66**, 031907 (2002)
51. J. Donges, Y. Zou, N. Marwan, J. Kurths, Eur. Phys. J. Special Topics **174**, 157 (2009)
52. M. Palus, V. Komarek, T. Prochazka, Z. Hrncir, K. Sterbova, IEEE Eng. Med. Biol. **20**, 65 (2001)
53. A.M. Fraser, H.L. Swinney, Phys. Rev. A **33**, 1134 (1986)
54. U. Parlitz, in *Nonlinear Modeling – Advanced Black-Box Techniques* (Kluwer Academic Publishers, Boston, 1998), Chap. Nonlinear Time-Series Analysis
55. H. Kantz, T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge University Press, Cambridge, 2004)
56. S. Haykin, *Communication Systems* (John Wiley & Sons, New York, 2001)
57. F. Rossi, A. Lendasse, D. Francois, V. Wertz, M. Verleysen, Chemometr. Intell. Lab. **2**, 215 (2006)
58. B. Efron, R. Tibshirani, *An introduction to the bootstrap* (CRC press, 1993)
59. M. Tumminello, S. Miccichè, F. Lillo, J. Piilo, R. Mantegna, PLoS ONE **6**, e17994 (2011)
60. Y. Benjamini, Y. Hochberg, J. R. Statist. Soc. B **57**, 289 (1995)
61. C.E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948)
62. J. Beirlant, E. Dudewicz, L. Gyorfi, E. van der Meulen, Int. J. Math. Stat. Sci. **6**, 17 (1997)
63. G. Darbellay, I. Vajda, IEEE T. Inform. Theory **45**, 1315 (1999)
64. L. Paninski, Neural Comput. **15**, 1191 (2003)
65. C. Daub, R. Steuer, J. Selbig, S. Kloska, BCM Bioinformatics **5**, 118 (2004)
66. W. Nemenman, W. Bialek, R. de Ruyter van Steveninck, Phys. Rev. E **69**, 056111 (2004)
67. J. Bonachela, H. Hinrichsen, M. Munoz, J. Phys. A **41**, 202001 (2008)
68. T. Schurmann, P. Grassberger, Chaos **6**, 414 (1996)
69. D. François, V. Wertz, M. Verleysen, The permutation test for feature selection by mutual information, in *European Symposium on Artificial Neural Networks, 2006*, pp. 239–244
70. B. Goebel, Z. Dawy, J. Hagenauer, J. Mueller, An Approximation to the Distribution of Finite Sample Size Mutual Information Estimate, in *Proc. IEEE Intl. Conf. Comm.* (2005)
71. Z. Dawy, B. Goebel, J. Hagenauer, C. Andreoli, T. Meitinger, J. Mueller, IEEE/ACM Trans. Comput. Biol. Bioinf. **3**, 47 (2006)
72. R. Steuer, L. Molgedey, W. Ebeling, M. Jiménez-Montaño, Eur. Phys. J. B **19**, 265 (2001)
73. N. Navet, S.H. Chen, in *Natural Computing in Computational Finance*, edited by T. Brabazon, M. O'Neill (Springer, 2008), Vol. 100