entropy

MDPI

Article

Exploring the Relationship among Predictability, Prediction Accuracy and Data Frequency of Financial Time Series

Shuqi Li and Aijing Lin *

School of Science, Beijing Jiaotong University, Beijing 100044, China; 18121635@bjtu.edu.cn
* Correspondence: ajlin@bjtu.edu.cn

Received: 17 August 2020; Accepted: 1 December 2020; Published: 6 December 2020

Abstract: In this paper, we aim to reveal the connection between the predictability and prediction accuracy of stock closing price changes with different data frequencies. To find out whether data frequency will affect its predictability, a new information-theoretic estimator P_{lz}, which is derived from the Lempel–Ziv entropy, is proposed here to quantify the predictability of five-minute and daily price changes of the SSE 50 index from the Chinese stock market. Furthermore, the prediction method EEMD-FFH we proposed previously was applied to evaluate whether financial data with higher sampling frequency leads to higher prediction accuracy. It turns out that intraday five-minute data are more predictable and also have higher prediction accuracy than daily data, suggesting that the data frequency of stock returns affects its predictability and prediction accuracy, and that higher frequency data have higher predictability and higher prediction accuracy. We also perform linear regression for the two frequency data sets; the results show that predictability and prediction accuracy are positive related.

Keywords: entropy rate; predictability; entropy difference; financial time series

1. Introduction

As the most essential task of financial market analysis, price analysis has been paid more and more attention, even though the support for the strong version of the efficient market hypothesis (EMH) [1–3] has decreased since the 1980s [4,5]. If the EMH is of some relevance to reality, then a market would be very unpredictable due to the possibility for investors to digest any new information instantly [6,7]. However, new evidence challenges the EMH with many empirical facts from observations, e.g., the leptokurtosis and fat tail of the non-Gaussian distribution, especially the fractal market hypothesis (FMH) [8,9]. In addition, Beben and Orlowski [10] and Di Matteo et al. [11–13] found that emerging markets were likely to have a stronger degree of memory than developed markets, suggesting that the emerging markets had a larger possibility of being predicted.

Traditionally, econometricians and econophysicists are more interested in predictability of price changes in principle and in practice. The notion of predictability of the time series can be explained by the memory effects of the past values. Using entropy to measure the degree of randomness and the predictability of a series has been a topic for a long time; it goes back almost to the very beginning of the development of communication and information theory.

In this paper, we propose a new information-theoretic predictability estimator P_{lz} for financial time series, which is derived from the Lempel–Ziv estimator [14–16]. The P_{lz} quantifies the contributions of the past values by reducing the uncertainty of the forthcoming values in time series. Then we use the prediction method EEMD-FFH [17] to find some connections between the predictability and prediction accuracy of financial time series.

The paper is organized as follows. In the Methodology section, we introduce two predictability estimators which measure the magnitude of predictability of time series, D_{norm} [18] and P_{lz} proposed in this paper, and a prediction algorithm EEMD-FFH. In the Numerical Simulation section, we apply these two estimators to the artificial simulation numerical analysis, the logistic map. The Financial Time Series Analysis and Prediction section is the most important part in this paper, which draws two important results. Finally, the Conclusion section is devoted to the summing-up and future studies.

2. Methodology

2.1. Entropy Rate

Entropy rate can be used to estimate the predictability of time series. It is a term derivative of the entropy, which measures the uncertainty of a random process in theory of information. Let $x = \{x_t\}$, $t = 1, 2, \cdots, T$ be a time series realization. The Shannon entropy [19] of a random variable at time t is defined as:

$$H[x_t] = -\sum_{x_t \in \Theta} p(x_t) \log_2 p(x_t)$$

where $p(x_t)$ is the probability distribution of x_t and Θ is sample space. Shannon also introduced the entropy rate, which generalizes the notion of entropy. For a stochastic process $X = \{X_t\}$, $t = 1, 2, \cdots, T$, the entropy rate is given by

$$H(X) = \lim_{T \to \infty} \frac{1}{T} H(X_1, X_2, \cdots, X_T)$$

The right side can be interpreted as the average entropy of each random variable in the stochastic process. If the process satisfies the stationarity condition, the entropy rate can also be expressed as a conditional entropy rate

$$H(X) = \lim_{T \to \infty} H(X_T | X_1, X_2, \cdots, X_{T-1})$$

It denotes the uncertainty in a quantity at time T having observed the complete history up to that point.

2.2. Entropy Rate Estimation

Entropy rate estimation has been paid more and more attention over the last 10 years due to the fact that the real entropy is known in very few isolated applications, one of the main reasons being the crucial practical importance of information-theoretic techniques in neurosciences. Entropy rate estimators can be classified into two categories [20]:

i. The "plug in" (also called maximum-likelihood) technique and its modifications. The main principle of these methods is to compute the empirical frequencies of different patterns in the data, and then calculate the entropy of the empirical distribution. Due to the cost of calculation and limits on the data size, the "plug in" method cannot reveal the signal with long term time dependency.

ii. Estimators based on data compression methods, such as Lempel–Ziv (LZ) [14–16] and context-tree weighting (CTW) [21,22]. This kind of approach is used to speed up the convergence and improve the performance in capturing long term time dependency.

In this study, we use two estimators which fall into the above two categories, entropy difference [18] belonging to the "plug in" class, and the new estimator we propose P_{lz} belonging to the Lempel–Ziv class).

2.2.1. D_{norm} Predictability Estimator

Consider a time series $X = \{x_t\}, t = 1, 2, \cdots, T$. The entropy rate at time t for a stationarity process is defined as $H[x_t | x_1, x_2, \cdots, x_{t-1}]$. We assume that the underlying system can be approximated by

a p-order Markov process. Then the value of the current moment is only related to the previous p moments. Hence, we can simplify the entropy rate:

$$H\left[x_t \left| x_1, x_2, \cdots, x_{t-1}\right.\right] = H\left[x_t \left| x_{t-p}, x_{t-p+1}, \cdots, x_{t-1}\right.\right]$$

$$\equiv H\left[x_t \left| x_{t-1}^{(p)}\right.\right] = \sum_{x_t, x_{t-1}^{(p)} \in \Theta} p\left(x_t, x_{t-1}^{(p)}\right) \log_2 \frac{p\left(x_t, x_{t-1}^{(p)}\right)}{p\left(x_{t-1}^{(\pi)}\right)}$$

After we consider the past values, the uncertainty of the time series will not increase; therefore, $H\left[x_t \left| x_{t-1}^{(p)}\right.\right] \leq H\left[x_t\right]$. Now we can define the entropy difference (ED) as D, which is the difference between the entropy and entropy rate and is non-negative.

$$D = H\left[x_t\right] - H\left[x_t \left| x_{t-1}^{(p)}\right.\right]$$

The right side can be interpreted as the contributions of the past values to reduce the uncertainty at time t. If the underlying process is a random walk, then $D = 0$. That is to say, the past values provide no information for current time. $0 < D \leq H\left[x_t\right]$ indicates that the process has time autocorrelation; thus, the past values can help to improve the predictability at time t. Due to the lower bound and upper bound being certain of D: $0 < D \leq H\left[x_t\right]$, we can normalize it to interval $[0, 1]$.

$$D_{norm} = \frac{H\left[x_t\right] - H\left[x_t \left| x_{t-1}^{(p)}\right.\right]}{H\left[x_t\right]} = 1 - \frac{H\left[x_t \left| x_{t-1}^{(p)}\right.\right]}{H\left[x_t\right]}, \quad 0 \leq D_{norm} \leq 1$$

D_{norm} measures the predictability of time series. When D_{norm} tends to 0, the time series is unpredictable. If D_{norm} is approximately 1, $H\left[x_t \left| x_{t-1}^{(p)}\right.\right] \approx 1$, the time series can be predicted completely at time t.

We proceeded with three numerical simulations to apply D_{norm} to different time series respectively. The data size is 10,000 points. The results are like those in Table 1. The first row is the entropy and D_{norm} of a deterministic time series $1, 1, \ldots, 1$. The simulation results are consistent with our intuitive understanding, namely, the uncertainty is 0 or the predictability is 1. The results of repeated pattern indicates that for a repeat pattern time series, the entropy is 1 up to the upper bound, and the underlying time series can be predicted totally when past values were considered. For the pure random series, the predictability equals to 0. That is, we cannot predict a pure random time series even we consider its past values. The result is dependent on three factors: sample size, the efficiency of the estimator and the quality of the random generator. Hence, it is easy to understand why the entropy is 0.9997 not equal to 1.

Table 1. The entropy and D_{norm} of three time series.

Time Series (N = 10,000)	Entropy	D_{norm}
All 1: 1, 1, …, 1	0	1
Repeated pattern (0, 1): 0, 1, 0, 1,…, 0, 1	1	1
Generate random integer from 1, 2	0.9997	0

2.2.2. P_{lz} Predictability Estimator

After Kolmogorov defined the complexity as the size of the minimum binary code that produces this time series in 1965 [23], complexity has been widely used to estimate entropy rate. Jacob Ziv and Abraham Lemple in 1977 designed a practical algorithm called Lempel–Ziv [16] to measure the complexity in the Kolmogorov sense, which also can identify the randomness of a time series. On this basis, there may entropy rate estimators were derived. One of them was created by Kontoyiannis in

1998 [24] (it will denoted as H_{lz} in the better). H_{lz} was widely used, and proved to have good statistical properties and better practical performance than other Lempel–Ziv estimators [24].

Consider a time series $X = \{x_t\}, t = 1, 2, \cdots, T$. The H_{lz} estimator is defined as:

$$H_{lz} = \left(\frac{1}{T} \sum_{i=1}^{T} L_i \right)^{-1} \log_2 T$$

where T is the size of the underlying time series and L_i is the length of the shortest substring starting from time i that does not appear as a contiguous substring in the prior values.

It has been proved that H_{lz} converges to the entropy rate with probability one as T approaches infinity for a stationary ergodic Markov process.

We calculate the estimator H_{lz} values of some simulated time series, which were generated at random from the alphabet $\{1, 2, 3, 4, 5, 6, 7, 8\}$ with different data sizes T. The theoretical entropy is equal to $-\sum_{i=1}^{8} (1/8) \log_2 (1/8) = 3$. As shown in the left part of Figure 1, H_{lz} converges to this theoretical value as data size T tends to infinity.

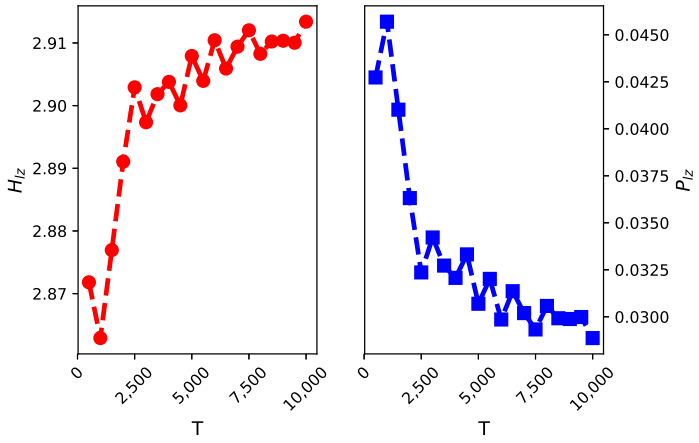

Figure 1. We calculate the estimator values (H_{lz} on the left side and the right for P_{lz}) for different data volumes $T = \{500, 1000, \ldots, 10{,}000\}$. The results indicate that H_{lz} converges to the theoretical entropy 3 and P_{lz} converges to 0 as T grows.

In our study, we propose a new predictability estimator P_{lz}, which is derived from the estimator H_{lz}. P_{lz} is defined as:

$$P_{lz} = 1 - H_{lz} / \hat{S}$$

where $\hat{S} = \log_2 S$ and S is the number of distinct states of the symbolized data. The estimator H_{lz} is normalized into interval $[0, 1]$. We will introduce the detailed discretized method and its necessity in Section 3.

2.3. EEMD-FFH Prediction Algorithm

In this paper we will use a particular method EEMD-FFH [17] to find out whether the predictability of time series is related to the prediction accuracy of a particular algorithm.

The EEMD method is used to decompose a time series into a series of intrinsic mode functions (IMFs) and one residue. It has been widely used in various industries [25,26], and was proposed by Huang et al. [27]. Based on the EEMD model, there is a hybrid prediction model called EEMD-FFH [17,28] that integrates MKNN (for predicting high frequency IMFs), ARIMA (for predicting low frequency IMFs) and quadratic regression (for residue wave) models.

The operation steps of EEMD-FFH are as follows:

Step 1. Decompose the time series $X(t)$ via EEMD

$$X(t) = \sum_{i=1}^{n} c_i + r_n.$$

Step 2. Use different models to predict IMFs of different frequencies

$$IMF_i \Longrightarrow Result_i$$

$$r_n \Longrightarrow Result_r$$

Step 3. Sum up the results to get the prediction value

$$\hat{X} = \sum_{i=1}^{n} Result_i + Result_r$$

In the experimental section of this article, we use this algorithm to predict daily financial data and five-minute high-frequency financial data.

3. Numerical Simulation

In this section, we consider a nonlinear system, the logistic map, to test the two predictability estimators as mentioned above. Chaos in dynamical systems has been investigated over a long period of time. With the advent of fast computers, the numerical investigations on chaos have increased considerably over the last two decades, and by now, a lot is known about chaotic systems. One of the simplest and most transparent system exhibiting order to chaos transition is the logistic map [29]. The logistic map is a discrete dynamical system defined by

$$x_{t+1} = px_t(1 - x_t)$$

with $0 \leq x_t \leq 1$. Thus, given an initial value (seed) x_0, the series x is generated. Here the subscript t plays the role of discrete time. The behavior of the series as a function of the parameter p is interesting. A thorough investigation of logistic map has already been done [29]. Here, without going into detailed discussion, we simply note that

- The logistic map has $x = 0$ and $x = (p-1)/p$ as fixed points. That is, if $x_i = 0$ or $x_t = (p-1)/p$, then $x_{t+1} = x_t$.
- For $p < 1$, $x = 0$ is an attractive (stable) fixed point. That is, for any value of the seed x_0 between 0 and 1, x_t approaches 0 exponentially.
- For $1 \leq p \leq 3$, $x = (p-1)/p$ is an attractive fixed point.
- For $3 < p < 3.56995$, the logistic map shows interesting behavior such as repeated period doubling, appearance of odd periods, etc.
- Most values of p beyond 3.56995 exhibit chaotic behavior.

Here, we set $p = 3.7$ and let the data length $N = 10^5$. The initial value of x_0 is set to 0.5.

As only one equation is described in the logistic map, x_t changes no information with other variables. We added Gaussian white noises to the original time series x_t with different strengths to obtain a composite time series, $y_t = x_t + \lambda\epsilon_t$. ϵ_t is the Gaussian white noise (with zero mean and unit variance). $\lambda \geq 0$ is a parameter that tunes the strength of noise. x_t is the real signal corrupted by the external noise ϵ_t, and λ determines the signal–noise ratio. The larger the λ, the smaller the signal–noise ratio.

We used k-means clustering to discretize the data with added Gaussian white noises into b distinct clusters. b is a pre-defined parameter that determines the number of clusters. In Figure 2, we show the values of D_{norm} and P_{lz} on different bins $b = 5, 6, \cdots, 14$, with the noise strength parameter λ from

0.01 to 0.1 with a step of 0.01. The result indicates that the predictability of the time series decreased with increasing λ, as the signal–noise ratio became lower. D_{norm} and P_{lz} reach values close to 0.1 when $\lambda = 0.1$, so it is hard to predict the composite time series when we add more Gaussian white noise into it. Moreover, the predictability of a logistic map has no obvious relationship with the number of bins. This result is consistent with [30], which found the choice of bins is largely irrelevant to the estimation results. Here, the parameter b for the k-means clustering was 10. This experimental setup has been proven to be very efficient at revealing the randomness of the original data [31].

From the above numerical simulations, we were able to conclude that the two estimators have good performances in estimating the randomness or predictability of the system (the predictability of the time series decreases with increasing λ), so we carried out the following real financial data experiments.

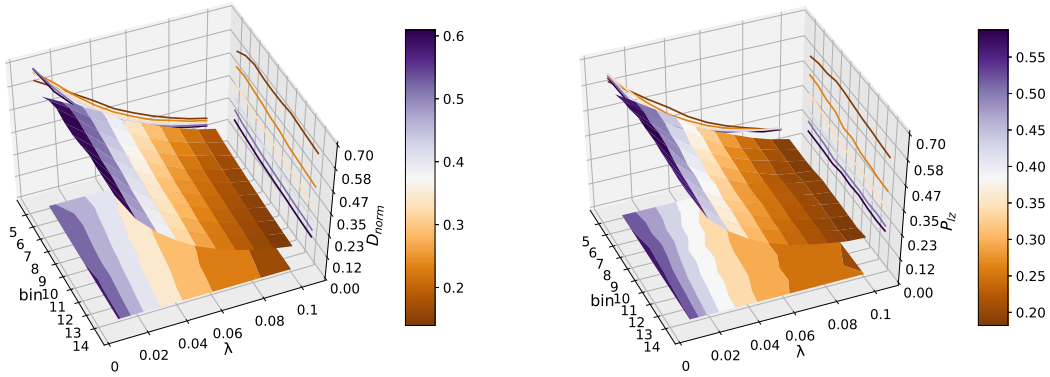

Figure 2. This figure shows that the change of predictability of the logistic map with increasing λ and *bin*. D_{norm} and P_{lz} reach values close to 0.1 when $\lambda = 0.1$, so it is hard to predict the composite time series when we add more Gaussian white noises into it. Moreover, the predictability of logistic map has no obvious relationship with the number of bins, which we can see in the projection of the 3D image. There is not much difference between the two estimators, eliminating the effect of different statistics on the same systems.

4. Financial Time Series Analysis and Prediction

4.1. Data and Stock Selection

In order to assess whether the five-minute high-frequency financial data or daily financial data has a larger possibility of being predicted, we estimated the entropy rate of close price of the stocks that make up the SSE (Shanghai Securities Exchange) 50 index. This index is based on scientific and objective methods to select the most representative 50 stocks with the large scale and good liquidity in the Shanghai stock market, able to form sample stocks, so as to comprehensively reflect the overall situation of a group of leading enterprises with the most market influence in Shanghai's stock market. The data have been found at URL http://www.10jqka.com.cn/ and were up to date as of the 13 January 2019, going 10 years back. Only complete records, i.e., five-minute and daily data with valid values for both 50 stocks, were admitted; invalid values were filtered out. In reality, non adjacent data may become adjacent data because of this procedure, but the relatively small number of invalid values compared to the valid values prevents a statistically significant impact [32]. The original close data of SSE 50 stocks cannot reasonably be assumed as stationary, a property for a time series yet essential for the validity of the forthcoming analysis. A classical solution to solve this problem is to

define some new variables which can be considered stationary or at least asymptotically stationary [33]. The usual transforms for raw time series $X = \{x_t\}, t = 1, 2, \cdots, T$ are as follows:

$$\begin{aligned} Increment \quad & \delta x_\tau(t) := x(t+\tau) - x(t) \\ Return \quad & r_\tau(t) := \delta x_\tau(t)/x(t) \\ Log-Return \quad & s_\tau(t) := \ln[x(t+\tau)] - \ln[x(t)] \end{aligned}$$

The choice of the variable does not affect the outcome of the present work; in fact, in the high-frequency regime they are approximately identical or proportional to each other [33]. We will use the log-returns in the forthcoming analysis. The usual quantity employed to characterize the fluctuation in financial data is the so called volatility, here defined as

$$\mathrm{vol}_\Delta(t) := \frac{1}{\Delta} \sum_{i=1}^{\Delta} |s_\tau(t+i)|$$

where the parameter Δ refers to the chosen length of the time-window and τ (in our cases always $\tau = 1$ day&5 min) denotes the basic time scale. The average values of the whole 50 stocks log-returns are $\hat{s}_{1\mathrm{day}}(t) \simeq \pm 1 \times 10^{-4}$ and $\hat{s}_{5\mathrm{min}}(t) \simeq \pm 1 \times 10^{-5}$, while the absolute log-returns, also interpretable as estimates of the 5min and daily volatility, have mean values of $\hat{vol}_{1\mathrm{day}}(t) \simeq \hat{vol}_{5\mathrm{min}}(t) \simeq 6 \times 10^{-4}$. However, as is widely known, the strength of fluctuations in financial data is subject to long-term correlated oscillations. Still, in concordance with other authors [33], we assume a sufficiently long financial time series to be asymptotically stationary, i.e., leading to relevant results for the long-term statistical properties of the analyzed data. The distributions of the Shannon entropy for daily data and five-minute high-frequency are shown in Figure 3. μ and σ represent mean and standard deviation respectively.

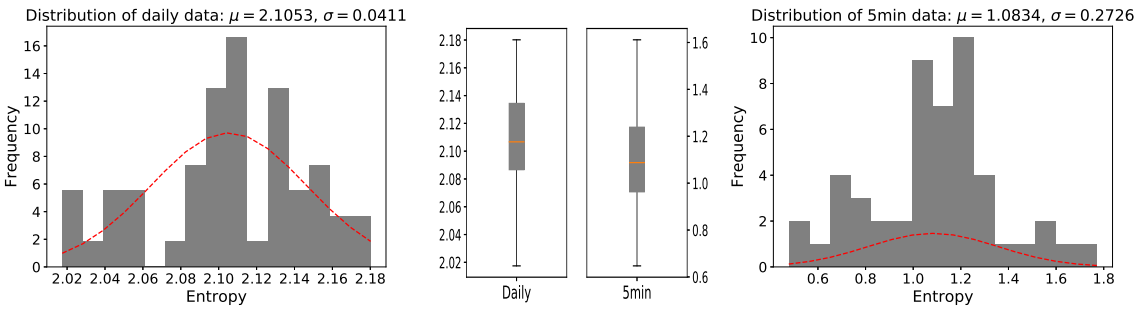

Figure 3. Distribution of the entropy of the stocks that make up the SSE 50 index (the left one shows log-returns of consecutive daily closing prices, while the right one shows log-returns of consecutive 5min closing prices). In every histogram, a normal distribution with the same mean and standard deviation is plotted. Discrete cases of two distributions in the two box-plots are shown.

The results show that the entropy of 5 min closing prices is lower than that of daily closing prices. This is not very surprising, since high entropy has been observed even for larger time scales [34]. We considered 20 stocks, which included the five highest entropy stocks and the five lowest entropy stocks of the daily data and the same choice for the five-minute high-frequency financial data. In order to eliminate the influence of multi-scale on entropy calculation and stock selection, we used a coarse-graining algorithm by amplification in different proportions (range from 0 to 20). Then we calculated the average value and the median value for the coarse-graining dataset to choose high entropy stocks and low entropy stocks. The stocks selected by average value and median value were totally identical. After removing overlapping stocks, we obtained 20 stocks, and the detailed calculation results are shown in Table 2. These 20 stocks are considered in the experiments in the next sections.

Table 2. The selected stocks by calculating the average value and the median value for the coarse-graining dataset.

Data Type		Daily				5 min			
Method	Average		Median		Average		Median	
High entropy	**GYFL601138**	2.416	**GYFL601138**	2.507	**HRYY600276**	1.592	**HRYY600276**	1.772
	ZXJT601066	2.385	**ZXJT601066**	2.346	**HTZQ601236**	1.454	**HTZQ601236**	1.611
	ZGTB601601	2.193	**ZGTB601601**	2.261	**ZSYH600036**	1.452	**ZSYH600036**	1.598
	XHBX601336	2.119	**ZGGH601111**	2.152	**XYYH601166**	1.417	**GZMT600519**	1.536
	ZGGH601111	2.118	**XHBX601336**	2.148	**GZMT600519**	1.372	**XYYH601166**	1.436
Low entropy	**HTZQ600837**	1.310	**ZGPA601318**	1.380	**ZGZC601766**	0.661	**ZGZT601390**	0.666
	ZGGL601888	1.299	**HTZQ600837**	1.343	**ZGZT601390**	0.660	**WHHX600309**	0.658
	ZGPA601318	1.276	**ZGGL601888**	1.295	**WHHX600309**	0.605	**ZGZC601766**	0.647
	HEZJ600690	1.260	**HEZJ600690**	1.215	**SDHJ600547**	0.579	**SDHJ600547**	0.542
	SAGD600703	0.930	**SAGD600703**	0.979	**ZGJZ601668**	0.524	**ZGJZ601668**	0.478

To give the evidence that the raw time series are not random, we compared the entropy of the raw time series with the entropy of randomly shuffled variants of the original data, which is also called surrogate testing. With such a preprocessing, all potential correlations in the original time series were destroyed; 100 shuffled time series for each raw time series (before the homogeneous partitioning) were generated and their average entropy was measured. The distributions of shuffled data are different to those in the original time series, and the average entropy is much larger, as can be seen in Figure 4. This provides evidence that there are temporal dependencies in the data we analyzed, and it makes sense for us to calculate their degrees of predictability.

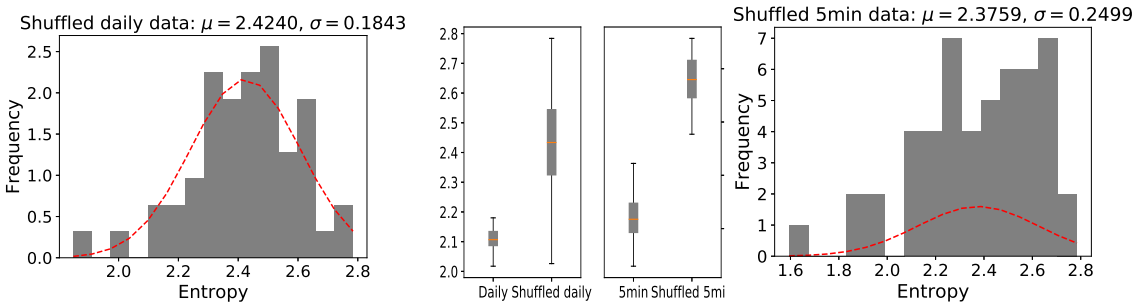

Figure 4. Distribution of the shuffled data entropy of the stocks that make up the SSE 50 index (the left one shows log-returns of shuffled daily closing prices while the right one shows log-returns of shuffled 5min closing prices). In every histogram, a normal distribution with the same mean and standard deviation is plotted. The entropy of surrogate time series is much larger than that of the raw data in the middle box-plots.

4.2. Estimating the Predictability of Different Frequency Time Series Based on D_{norm} and P_{lz}

In this section, we use the 20 stocks already obtained in Table 2 to calculate the predictability of daily data and five-minute high-frequency data respectively, based on D_{norm} and P_{lz}. The main question asked in this paper is whether daily price changes are more or less predictable than intraday (five-minute high-frequency) price changes. The reason why we use these two predictability estimators is to make the experiment more credible, and the two estimators are not compared in this paper.

We divide those 20 stocks into four groups, every group including five stocks, as shown in Figure 5, which we obtained in Section 4.1. Then for every part we calculate the predictability of daily data and five-minute high-frequency data respectively, based on D_{norm} and P_{lz}. In Figure 6, the left panels show the predictability of every group based on D_{norm}, and the right panels show the predictability of every part based on P_{lz}.

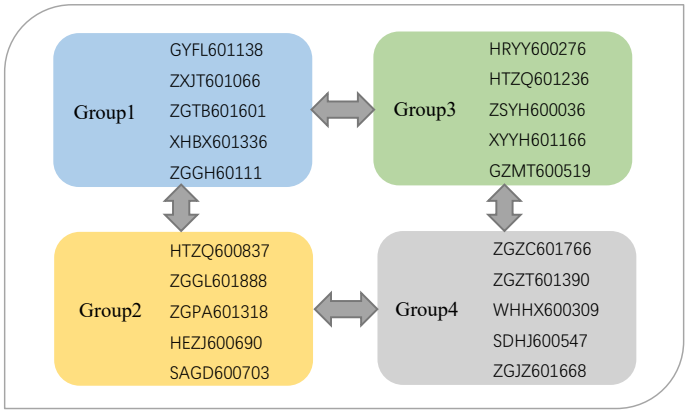

Figure 5. The 20 selected stocks we obtained in Section 4.1 are assigned to four groups to test whether daily price changes are more or less predictable than intraday (five-minute high-frequency) price changes.

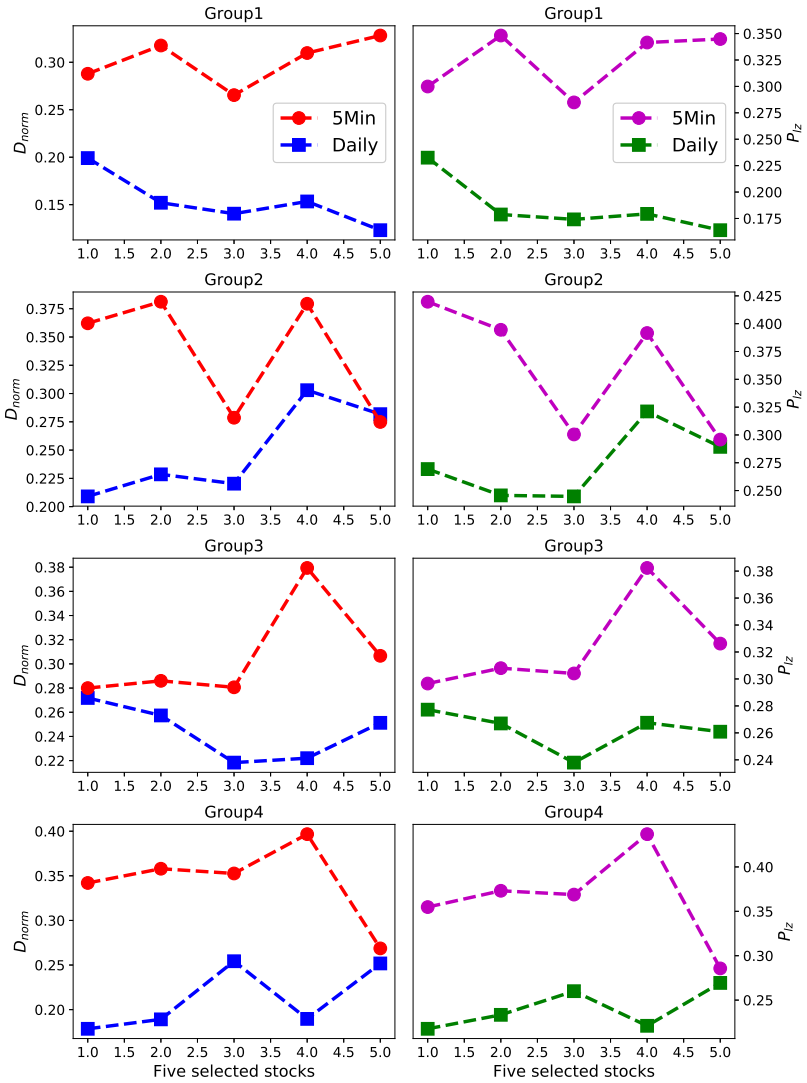

Figure 6. The predictability of daily data and five-minute high-frequency data respectively based on predictability estimators D_{norm} and P_{lz}. Left panels show the results D_{norm} for group1–4, respectively. Right panels show the results P_{lz} for group1–4, respectively.

Surprisingly, for every stock of every group the predictability value of five-minute high-frequency data is obviously much larger than daily data, which means that five-minute high-frequency price changes are more predictable than daily price changes. The experimental results strongly suggest that the predictability of time series is related to the frequency of the data itself. From this conclusion we raise another question: whether the predictability of time series is related to the prediction accuracy of a particular algorithm. In the next section we will focus on this question.

4.3. Comparing the Prediction Accuracies of Different Frequency Time Series Based on EEMD-FFH

In last section, the experimental results strongly suggest that five-minute high-frequency price changes are more predictable than daily price changes. Then, is it possible that high frequency time series, which are more predictable, have higher prediction accuracy? In this section, we detail another experiment based on EEMD-FFH algorithm to explore this.

In order to assess the performance of EEMD-FFH for data at different frequencies, we use an indicator, root mean squared error (RMSE).

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(x_t - \hat{x}_t)^2}$$

where x_t represents the raw data; $\hat{x}_t$ represents the prediction value; n is the number of prediction points; smaller RMSE means higher accuracy.

Table 3 tabulates the average RMSE of five stocks in every group. The last 200 points for every time series have been predicted and the set containing these 200 points was our testing set. For every one of them we can see that the five-minute high-frequency financial data have higher prediction accuracy than daily data. We also show RMSE of every group in Figure 7, where the obvious difference is more intuitive.

Table 3. RMSE of the two different data frequencies based on the EEMD-MKNN algorithm.

Group	Daily Data	5 min Data
1	0.0530	0.0063
2	0.0609	0.0073
3	0.0491	0.0047
4	0.0625	0.0056

To show the relationship between the predictability and prediction accuracy, we conducted correlation analysis. We calculated the Pearson correlation coefficient and Spearman correlation coefficient between the predictability and RMSE in different frequency, as shown in the following Table 4. These results reveal that the connection between these two concepts exists, and they are negatively correlated.

Table 4. Correlation coefficient between the predictability and RMSE.

	Daily data		5 min data	
	Pearson	Spearman	Pearson	Spearman
P_{lz}	−0.6081	−0.7018	−0.7022	−0.7649
D_{norm}	−0.6223	−0.6837	−0.7129	−0.7589

To explore the relationship further, we also performed linear regression for the two frequency data sets (every set includes 20 stocks of 4 groups). In this linear regression model, the value of predictability is an independent variable and the value of RMSE is the dependent variable. The scatter plot and fitted lines are shown in Figure 8. Every regression line shows that the predictability and RMSE are negatively correlated. RMSE is root mean squared error; high RMSE denotes low prediction accuracy—that is to say the predictability and prediction accuracy are positively related.

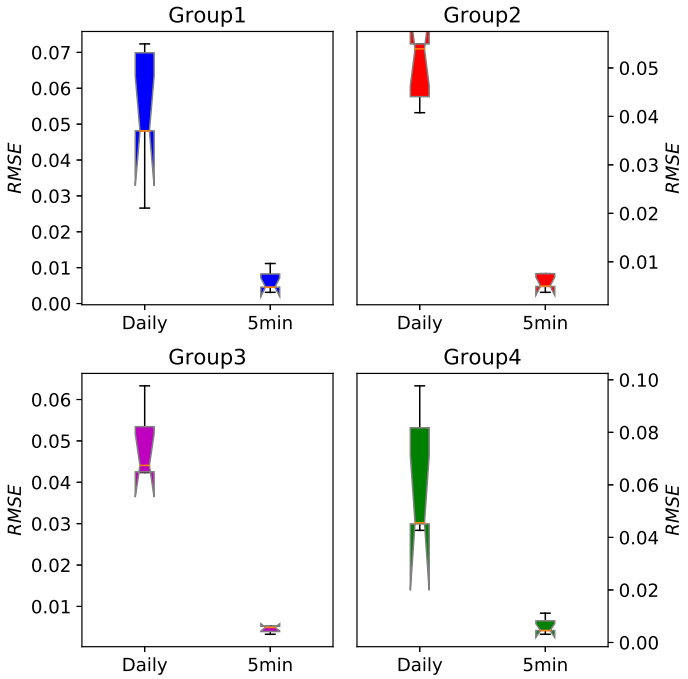

Figure 7. The boxplot of prediction error RMSE for the daily data and five-minute data based on the EEMD-FFH algorithm. For every group we can see that the five-minute high-frequency financial data have significantly higher prediction accuracy.

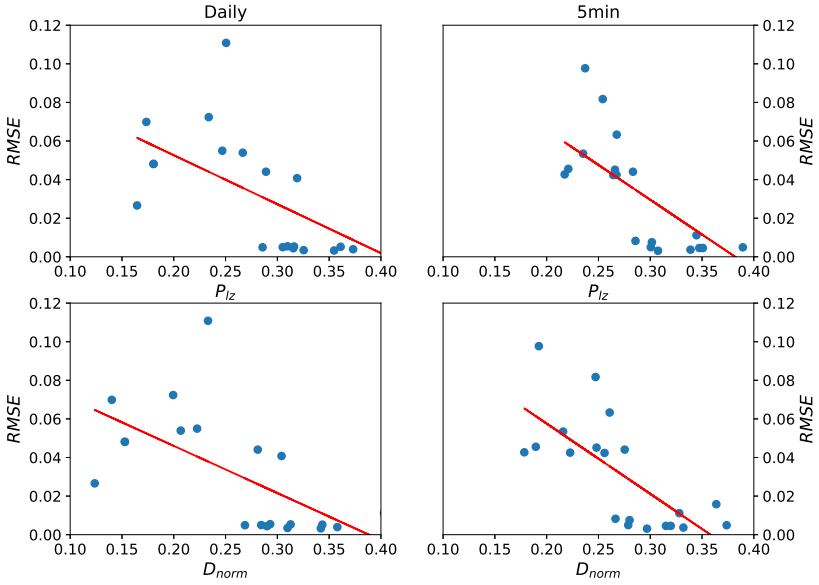

Figure 8. The scatter plot and regression fitted lines of daily data (**left panels**) and five-minute high-frequency data (**right panels**) respectively, based on predictability estimators D_{norm} and P_{lz}. Every regression line shows that the predictability and RMSE are negatively correlated—that is to say the predictability and prediction accuracy are positive related.

In statistics, the predictability of time series belongs to the category of time series analysis, which is different from the prediction accuracy based on a forecasting method. In this experiment, our goal was to see if the two were related. Surprisingly, the analysis results indicate that predictability fits prediction accuracy perfectly, and we found that the five-minute high-frequency financial data have higher predictability and prediction accuracy than daily data.

5. Conclusions

In this paper, we introduced a new information-theoretic predictability estimator P_{lz} for financial time series, which was derived from the Lempel–Ziv estimator. The P_{lz} quantifies the contributions of the past values by reducing the uncertainty of the forthcoming values in the time series. We limited ourselves to the stocks constituting SSE 50 index because they are primary components of the Chinese market, to do an experiment to explore whether data's frequency would effect its predictability. The results strongly suggest that five-minute high-frequency price changes are more predictable than daily price changes. Additionally, we used the prediction method EEMD-FFH to find some connections between the predictability and prediction accuracy. Here, the empirical evidence suggests that there is a strong positive relationship between these two concepts—this is, higher frequency data have higher predictability and higher prediction accuracy.

Further studies should be performed to confirm whether these results are robust and valid for other stock markets as well. Another important study is to find whether different prediction methods will change the result that a strong positive relationship exists between predictability and prediction accuracy for different frequency financial data.

Author Contributions: Conceptualization, S.L.; Methodology, S.L.; Software, S.L.; Validation, S.L.; Formal Analysis, S.L.; Investigation, S.L.; Resources, A.L.; Data Curation, S.L.; Writing—Original Draft Preparation, S.L.; Writing—Review & Editing, A.L.; Visualization, S.L.; Supervision, A.L.; Project Administration, A.L.; Funding Acquisition, A.L. Both authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (grant number 61673005).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lim, T.; Lim Xiu Yun, J.; Zhai, R. History of the efficient market hypothesis. *Int. J. Manag. Sci. Bus. Res.* **2012**, *1*, 11.
2. Malkiel, B.G. The efficient market hypothesis and its critics. *J. Econ. Perspect.* **2003**, *17*, 59–82. [CrossRef]
3. Malkiel, B.G. Efficient market hypothesis. In *Finance*; Springer: Berlin/Heidelberg, Germany, 1989; pp. 127–134.
4. Lee, C.C.; Lee, J.D. Energy prices, multiple structural breaks, and efficient market hypothesis. *Appl. Energy* **2009**, *86*, 466–479. [CrossRef]
5. Yen, G.; Lee, C.F. Efficient market hypothesis (EMH): Past, present and future. *Rev. Pac. Basin Financ. Mark. Policies* **2008**, *11*, 305–329. [CrossRef]
6. Malkiel, B.G.; Fama, E.F. Efficient capital markets: A review of theory and empirical work. *J. Financ.* **1970**, *25*, 383–417. [CrossRef]
7. Lin, A.; Liu, K.K.; Bartsch, R.P.; Ivanov, P.C. Dynamic network interactions among distinct brain rhythms as a hallmark of physiologic state and function. *Commun. Biol.* **2020**, *3*, 1–11.
8. Peters, E.E. *Fractal Market Analysis: Applying Chaos Theory to Investment and Economics*; John Wiley & Sons: Hoboken, NJ, USA, 1994.
9. Weron, A.; Weron, R. Fractal market hypothesis and two power-laws. *Chaos Solitons Fractals* **2000**, *11*, 289–296. [CrossRef]
10. Beben, M.; Orłowski, A. Correlations in financial time series: Established versus emerging markets. *Eur. Phys. J. -Condens. Matter Complex Syst.* **2001**, *20*, 527–530. [CrossRef]
11. Di Matteo, T.; Aste, T.; Dacorogna, M.M. Scaling behaviors in differently developed markets. *Phys. Stat. Mech. Appl.* **2003**, *324*, 183–188. [CrossRef]
12. Di Matteo, T.; Aste, T.; Dacorogna, M.M. Long-term memories of developed and emerging markets: Using the scaling analysis to characterize their stage of development. *J. Bank. Financ.* **2005**, *29*, 827–851. [CrossRef]
13. Lin, A.; Liu, K.K.; Bartsch, R.P.; Ivanov, P.C. Delay-correlation landscape reveals characteristic time delays of brain rhythms and heart interactions. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* **2016**, *374*, 20150182. [CrossRef] [PubMed]

14. Farach, M.; Noordewier, M.; Savari, S.; Shepp, L.; Wyner, A.; Ziv, J. On the entropy of DNA: Algorithms and measurements based on memory and rapid convergence. In Proceedings of the Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, San Francisco, CA, USA, 1 January 1995; pp. 48–57.
15. Kontoyiannis, I.; Algoet, P.H.; Suhov, Y.M.; Wyner, A.J. Nonparametric entropy estimation for stationary processes and random fields, with applications to English text. *IEEE Trans. Inf. Theory* **1998**, *44*, 1319–1327. [CrossRef]
16. Ziv, J.; Lempel, A. A universal algorithm for sequential data compression. *IEEE Trans. Inf. Theory* **1977**, *23*, 337–343. [CrossRef]
17. Li, S.; Liu, X.; Lin, A. Fractional frequency hybrid model based on EEMD for financial time series forecasting. *Commun. Nonlinear Sci. Numer. Simul.* **2020**, 89, 105281. [CrossRef]
18. Zhao, X.; Liang, C.; Zhang, N.; Shang, P. Quantifying the Multiscale Predictability of Financial Time Series by an Information-Theoretic Approach. *Entropy* **2019**, *21*, 684. [CrossRef]
19. Shannon, C.E. A note on the concept of entropy. *Bell. Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
20. Gao, Y.; Kontoyiannis, I.; Bienenstock, E. From the entropy to the statistical structure of spike trains. In Proceedings of the 2006 IEEE International Symposium on Information Theory, Seattle, WA, USA, 9–14 July 2006; pp. 645–649.
21. Willems, F.M.; Shtarkov, Y.M.; Tjalkens, T.J. The context-tree weighting method: Basic properties. *IEEE Trans. Inf. Theory* **1995**, *41*, 653–664. [CrossRef]
22. Kennel, M.B.; S ens, J.; Abarbanel, H.D.; Chichilnisky, E. Estimating entropy rates with Bayesian confidence intervals. *Neural Comput.* **2005**, *17*, 1531–1576. [CrossRef]
23. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
24. Kontoyiannis, I. Asymptotically optimal lossy Lempel-Ziv coding. In Proceedings of the 1998 IEEE International Symposium on Information Theory, Cambridge, MA, USA , 16–21 August 1998; p. 273.
25. Flandrin, P.; Rilling, G.; Goncalves, P. Empirical mode decomposition as a filter bank. *IEEE Signal Process. Lett.* **2004**, *11*, 112–114. [CrossRef]
26. Huang, N.E. A Study of the Characteristics of White Noise Using the Empirical Mode Decomposition Method. *Proc. Math. Phys. Eng. Sci.* **2004**, *460*, 1597–1611.
27. Huang, N.E.; Shen, Z.; Long, S.R.; Wu, M.C.; Shih, H.H.; Zheng, Q.; Yen, N.C.; Tung, C.C.; Liu, H.H. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. Lond. Ser. Math. Phys. Eng. Sci.* **1998**, *454*, 903–995. [CrossRef]
28. Zhang, N.; Lin, A.; Shang, P. Multidimensional k-nearest neighbor model based on EEMD for financial time series forecasting. *Phys. Stat. Mech. Appl.* **2017**, *477*, 161–173. [CrossRef]
29. Feigenbaum, M.J. Quantitative universality for a class of nonlinear transformations. *J. Stat. Phys.* **1978**, *19*, 25–52. [CrossRef]
30. Fiedor, P. Frequency effects on predictability of stock returns. In Proceedings of the IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr), London, UK, 27–28 March 2014; pp. 247–254.
31. Steuer, R.; Molgedey, L.; Ebeling, W.; Jiménez-Montano, M. Entropy and optimal partition for data analysis. *Eur. Phys. J.* **2001**, *19*, 265–269. [CrossRef]
32. Ghashghaie, S.; Breymann, W.; Peinke, J.; Talkner, P.; Dodge, Y. Turbulent Cascades in Foreign Exchange Markets. *Nature* **1996**, *381*, 767–770. [CrossRef]
33. Mantegna, R.N.; Stanley, H.E.; Chriss, N.A. An Introduction to Econophysics: Correlations and Complexity in Finance. *Phys. Today* **1999**, *53*, 70. [CrossRef]
34. Park, J.B.; Lee, J.W.; Jo, H.H.; Yang, J.S.; Moon, H.T. Complexity and entropy density analysis of the Korean stock market. In Proceedings of the 9th Joint Conference on Information Sciences, Kaohsiung, Taiwan, 8–11 October 2006; pp. 305–309.