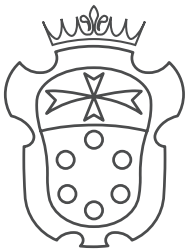


SCUOLA
NORMALE
SUPERIORE

Faculty of Sciences

PhD Thesis in
Computational Methods and Mathematical Models for Sciences and Finance

35 cycle

Shannon entropy and high frequency financial time series

Scientific Disciplinary Sector **SECS-S/06**

Candidate
dr. Andrey Shternshis

Supervisors

Prof. Stefano Marmi
Prof. Piero Mazzarisi

Academic year 2022–2023

Abstract

This thesis considers the problem of evaluating a degree of market efficiency. A market is called efficient if the financial asset prices fully reflect all available information. In the weak form of the Efficient Market Hypothesis (EMH), the information set only includes historical prices. In this form, the EMH implies that it is impossible to predict an asset's future price based on current prices. That is, market efficiency means that no profitable trading strategy can increase the expected profit of the buy-and-hold strategy. In the scientific literature, theoretical arguments support the Efficient Market Hypothesis. At the same time, several empirical results reject the hypothesis for various markets in different countries. In addition, several strategies, including algorithmic trading strategies and statistical and machine learning methods, are designed to increase the expected profit. The performance of such methods is another signal of not efficient markets. In this thesis, I aim to assess the degree of market efficiency based on the measure of Shannon entropy, thus testing the Efficient Market Hypothesis. No predictability of prices implies the maximum uncertainty for any symbolization of the dynamics, which can be captured by an entropy measure attaining its maximum under the EMH. A measure significantly smaller needs to be interpreted as a signal of market inefficiency. An entropy value significantly smaller than the maximum defines an inefficiency of the market in the considered period. A well-known measure of randomness for symbolic sequences is Shannon entropy. The Shannon entropy is widely used to measure randomness in many fields, such as physics, finance, biology, and medicine. The Shannon entropy represents the average amount of uncertainty removed with the transmission of a symbol generated by a random process. In this thesis, I discretize price returns into a finite alphabet to calculate the Shannon entropy. Each symbol of an alphabet represents the direction of a price or the magnitude of price increments.

This thesis examines four research projects on the topics of measuring the predictability of high-frequency price returns time series. In the first research project, I propose a method for filtering out data regularities. Data regularities are empirical properties of price returns that allow the prediction of specific dynamical patterns. For instance, these regularities include intraday volatility, volatility clustering, and microstructure noise. Price predictability associated with data regularities implies a decrease in the entropy value. However, such regularities are not profitable. Therefore, the first preliminary goal is to filter out data regularities before estimating the degree of market efficiency. In this research, I investigate price staleness as one of the sources of regularity. Price staleness is a lack of price adjustments generating spurious 0-returns. First, I show inefficient time intervals have more 0-returns than efficient ones. Then, I propose a method for filtering out spurious 0-returns from the data. In this method, I distinguish two sources of 0-returns: price rounding and price staleness. Using simulations, I show that the proposed method preserves, on average, the 0-returns because of price rounding only. I set all the other 0-returns as missing values under the assumption of price staleness for them. The presence of missing values opens up several methodological challenges. For instance, I modify a method for estimating volatility in the case of incomplete data to make a more accurate estimate. Moreover,

I propose a modification of the Empirical Frequencies Method for estimating entropy in the presence of missing values. Finally, I prove that the entropy estimator is consistent under the assumption of independence for the two data generating processes, for both symbols and missing values.

In this thesis, I investigate Exchange Traded Funds (ETFs) traded at the New York Stock Exchange. I work with a one-minute frequency of price returns time series. I concentrate on the analysis of weekly time intervals. Dividing the whole period of time into non-overlapping intervals allows the observation of the entropy dynamics over time. I define a time interval as inefficient in two steps: (i) estimating the Shannon entropy with the modified Empirical Frequencies Method; (ii) using empirical quantiles obtained from Monte Carlo simulations of entirely random sequences to define a confidence interval for the measure. I test each interval with a significance level of 0.01. Finally, I define the overall degree of market inefficiency as the fraction of inefficient time intervals detected throughout the whole period and for each asset separately. I show that the degree of inefficiency decreases by filtering out the sources of data regularity one by one. In other words, the residuals display larger Shannon entropy values after the filtering process. The degree of inefficiency of the ETF market for weekly time intervals at a one-minute frequency is equal to 1.35%. Therefore, it is possible to conclude that the ETF market is inefficient but only partially and for particular sub-periods. In addition to weekly time intervals, I consider monthly and quarterly time scales. The degree of market inefficiency calculated for monthly intervals is about 11%. The degree of market inefficiency based on quarterly intervals is approximately 10%. For the sake of explanation, I further study the patterns of symbols that are repeating with considerable frequency. I name them periodic patterns. For example, among them, the most frequent pattern is the switching sign of non-zero returns at each trading minute. I show a significant correlation between periodic patterns and a low entropy estimate for some ETFs. Finally, I introduce co-inefficiency for a group of assets. By definition, several assets are co-inefficient if they are inefficient in the same time intervals. I find statistically significant co-inefficiency for all considered lengths of time intervals: weeks, months, and quarters. For example, there are three ETFs that have co-inefficiency in the first quarter of 2009. Moreover, I show that co-inefficiency can be detected for cointegrated prices, assuming that 0-returns due to price staleness are associated with small trading volumes.

In the second research project, I investigate the efficiency of the Moscow stock exchange from 2012 to 2021. I focus on filtering out volatility clustering and price staleness in this study. The estimations of volatility and a degree of price staleness are mutually connected. Price staleness generates spurious 0-returns that the effect of price rounding can not explain. In order to compute the probability of price rounding, a proper estimator of the volatility is needed. On the other hand, spurious 0-returns due to price staleness tend to underestimate the value of volatility. For this reason, I propose a new method for filtering out heteroscedasticity and price staleness while considering the relationship between the two quantities. I estimate the volatility by using the modification of exponentially weighted moving average method applied to price returns. I exclude the impact of 0-returns due to price staleness on the estimation of volatility. The proposed method has two advantages: it is computationally efficient and permits real-time analysis. First, the formula for volatility has the only parameter that can be adjusted. Second, the estimation of volatility and probability of price rounding is computed minute by minute using values obtained at the previous time step.

I present the analysis of market efficiency for the stocks of 18 companies from 5 different industries. To make the test of the Efficient Market Hypothesis more robust, I consider discretizations simultaneously in three- and four-symbols alphabets. I conclude that the degree of inefficiency for the Moscow Stock Exchange is larger than 80%. I investigate the pairs of stocks that exhibit the most significant degree of inefficiency. I show that months where the random-

ness of the stock prices attains its minimum group together. The most inefficient months for the stock pairs are detected in 2014. Based on the frequencies of blocks of symbols, I determine what behavior of prices repeats most often for the inefficient time intervals. Moreover, I show that it is possible to predict the price direction with a probability larger than one-half in a causal way, i.e. using the data from the first half of the month to make predictions on the second half. It is interesting to study also if there exists some commonality in the dynamics of groups of stocks associated with market inefficiency. To this end, I use the Kullback-Leibler distance to devise a clustering analysis of stocks. I show that the market can be clustered into three groups of stocks. In particular, I show that banks and gas companies cluster together. Finally, I propose a measure of closeness between a pair of assets. I introduce a proper discretization to describe co-movements of prices. After estimating the entropy of the obtained symbolic sequences, I point out that market inefficiency displays some dependence on the sector to which companies belong.

The entropy of price returns is time-varying. Real-world systems are generally non-stationary, with an entropy value that is not constant in time. In the third research project, I propose a hypothesis testing procedure to test the null hypothesis of equal entropy values on two different intervals. The alternative hypothesis is a significant variation. To compute the z-score for hypothesis testing, I approximate the variance of the entropy's estimator. Since entropy may change over time, I aim to find a local estimation of entropy and its variance. In other words, the goal is to characterize the distribution of the entropy estimator, in particular the variance, for finite samples. To characterize the estimator's variance, I first obtain the explicit formulas of the central moments of the multinomial distribution describing the estimation of the Shannon entropy. I find the approximation of the variance up to the fourth order of the length of the sequence. With the same rationale, it is possible to extend the approximation further. The variance depends on unknown values of entropy and probabilities of events. For this reason, I derive an unbiased estimation of the variance that can be obtained using empirical probabilities calculated from a sequence. The expected value of this estimation is the variance of the entropy estimator. Then, I find the optimal length of the rolling window used for estimating the time-varying Shannon entropy by optimizing a novel self-consistent criterion: the optimal length maximizes the z-score of the hypothesis testing in-sample. This choice for the criterion comes from considerations about the bias-variance trade-off. The validity of the proposed methodology is supported by numerical experiments.

In the application part, I test the null hypothesis of equal entropy values for adjacent but non-overlapping intervals. I use the novel methodology to test for time-varying regimes of entropy for stock price dynamics. In particular, I consider the case of meme stocks. Starting from January 2021, meme stocks experienced a dramatic increase in prices and volumes driven by the long trades of many individual investors. I empirically show the existence of periods of market inefficiency for meme stocks. A sharp increase in prices and trading volumes is characterized by a statistically significant decrease in the Shannon entropy. In particular, a low level of entropy is identified even before the price spike of the GameStop stock. Moreover, I show that the minimum entropy values detected for the meme stocks are noticeably lower than those attained by IT companies such as Apple and Microsoft.

In the fourth research topic, I explore the efficiency of markets in a new direction. I consider the case of ultra-high frequency of price returns time series (tick-by-tick). More stylized facts, like fat tails, must be appropriately accounted for at higher frequencies. I explore whether stylized facts such as jumps can cause price predictability. Finally, I study whether the entropy value in a previous time interval increases the probability of detecting a significantly low value at the next time. In other words, I check if there exists an effect of clustering together inefficient time intervals at ultra-high frequency. I test if the entropy of a time interval is at the maximum, corresponding to a measure of complete efficiency. To this end, I find the parameters of the

gamma distribution that describes the entropy's estimator as a random variable. The theoretical quantiles of the gamma distribution help quickly test for ultra-high frequency data's randomness. In particular, multiple tests within the same trading day permits to localize the presence of price predictability.

This thesis considers the problem of measuring and testing market efficiency from different perspectives. I investigated many cases associated with different markets in several countries. I explored the sources of data regularities and the connections between them. I considered price time series at different frequencies. Also, I worked with several types of price discretization. I also proposed a new approach for determining significant entropy values based on a statistical testing procedure instead of relying on Monte Carlo simulations.

List of publications

The four research projects discussed in the abstract are the following.

Published papers

- [Shternshis et al., 2022a] A. Shternshis, P. Mazzarisi, and S. Marmi. Measuring market efficiency: The Shannon entropy of high-frequency financial time series. *Chaos, Solitons & Fractals*, 162:112403, 2022. doi: 0.1016/j.chaos.2022.112403

- [Shternshis et al., 2022b] A. Shternshis, P. Mazzarisi, and S. Marmi. Efficiency of the Moscow stock exchange before 2022. *Entropy*, 24(9):1184, 2022. doi: 10.3390/e24091184

Other works

- [Shternshis and Mazzarisi, 2022] A. Shternshis and P. Mazzarisi. Variance of entropy for testing time-varying regimes with an application to meme stocks. arXiv preprint arXiv:2211.05415, 2022. doi: 10.48550/arXiv.2211.05415. *Under review in the Decisions in Economics and Finance.*

- A. Shternshis and S. Marmi. Testing price predictability of ultra-high frequency data. *In preparation*

Acknowledgements

I would like to thank my scientific advisors, Professors Stefano Marmi and Piero Mazzarisi, for the productive and fruitful conversations and efforts invested in joint projects and my growth in the academic environment. I express my gratitude to the supervisors for their time for discussions and advice, which allow me to grow in writing articles and presenting research results. My special thanks to Prof. Marmi for his support during my doctoral studies, which coincide with the difficult period of the pandemic and the war. I would also like to thank Dr. Lucio Calcagnile for the frequent and helpful discussions of the first paper included in this thesis.

I also express my gratitude to my family: parents, brother, and grandmothers for supporting me and my decisions. I am grateful to my wife for being with me during our move to Italy, scientific conferences, and all the pleasant days. Finally, I would like to acknowledge the contribution of my cat Vasilisa: she is always around and does what she thinks is important.

Contents

Chapter 1

Introduction

I start by discussing the Efficient Market Hypothesis. I give a review of theoretical and empirical works on verification, revision, and testing the hypothesis. Then, I explain how market efficiency relates to the Shannon entropy. In this introductory chapter, I give several ways to define entropy of a random stationary process. At the end of this chapter, I discuss the contributions presented in the thesis.

1.1 Market efficiency

The hypothesis about efficient markets states that prices of assets traded in financial markets incorporate all available information. Markets where the Efficient Market Hypothesis (EMH) is satisfied are called efficient. Taking into account transaction costs, [Jensen, 1978] formulated the definition of market efficiency as follows: If it is impossible to make economic profits by trading on the basis of an information set, the market is called efficient with respect to this information set. In this context, an economic profit refers to a risk-adjusted profit net of all costs. [Fama, 1970] reviewed the theoretical and empirical literature on testing the hypothesis about efficient markets. In his seminal work published in 1970, he concluded that the efficient markets model performed well with a few exceptions. In 1991, E. F. Fama did a new review of the literature concerning the testing of the efficient markets hypothesis [Fama, 1991]. In the twenty years between these reviews, the number of studies suggesting positive predictability of returns increased.

Findings supporting the efficient markets hypothesis were also collected in [Yen and Lee, 2008]. The authors also discussed empirical evidence inconsistent with the EMH. One evidence against the hypothesis was presented, for example, in [Bondt and Thaler, 1985]. The authors stated an overreaction hypothesis based on which investors overreact to unexpected positive and negative news. The hypothesis states that a considerable price movement in one direction is followed by a subsequent price movement in the opposite direction. The hypothesis about assets overreaction was later confirmed in [Chopra et al., 1992] and [Fluck et al., 1997]. This behavior causes market inefficiency since it allows one to predict a stock's performance based on previous returns. However, it was shown that the overreaction hypothesis is associated with stocks with high trading costs [Lesmond et al., 2004]. [Shiller, 2003] explained overreaction discovered in prices by a feedback model: When a price rises creating profits for some investors, it can attract attention of other investors and heighten expectations for further price increases. [Malkiel, 2003] summarized known market anomalies and predictable patterns. [Ricciardi and Simon, 2000] made a review of behavioral finance literature where psychology of investors is

taken into account and can explain some market anomalies. [Fama and French, 1996] stated that some of known price anomalies are explained by the three-factor model proposed in [Fama and French, 1993]. However, the authors noted that the three-factor model does not explain the persistence of price returns [Jegadeesh and Titman, 1993] which suggests market inefficiency.

[Lo and MacKinlay, 1999] stated that stock prices do not follow random walks. They rejected the Efficient Market Hypothesis using a simple specification test [Lo and MacKinlay, 1987]. According to the Adaptive Markets Hypothesis proposed by [Lo, 2004] arbitrage opportunities allowing to increase an average profit appear from time to time. However, profit opportunities disappear when they are discovered and exploited. Nevertheless, new opportunities of profit continue to appear in the market. For example, [Mclean and Pontiff, 2016] found that abnormal returns associated with market anomalies decline after academic publications about these anomalies. Existence of profitable strategies that allow to receive an additional profit is an argument against the Efficient Market Hypothesis. [Lo et al., 2000] formalized the methods of technical analysis and showed that they have a predictive ability. As stated in [Hsu et al., 2016], technical analysis has a predictive power to generate excess returns for developed and emerging currencies. [Hudson and Urquhart, 2021] showed that technical trading rules give significant profitability for traders in cryptomarkets even after taking into account transaction costs. A review of papers on profitability of technical analysis was presented in [Park and Irwin, 2007]. [Los, 1998] concluded that the Hypothesis about Market Efficiency should be rejected for Asian stock markets. He found a serial dependence in weekly price changes and suggested that found price trends can be profitably exploited. [Teixeira and De Oliveira, 2010] showed that the method for trading based on the nearest neighbor classification outperforms the buy-and-hold strategy including transaction costs for 12 out of 15 stocks traded at the Brazil Stock Exchange. [Akyildirim et al., 2022] concluded that arbitrage opportunities exist in the Istanbul Stock Exchange. According to the results of the article, profitable trading strategies can be constructed with the usage of forecasting methods such as neural networks, nearest neighbors, or a random forest classifier. [Brasileiro et al., 2017] suggested an automatic trading method that outperforms alternative approaches including the buy-and-hold strategy.

Assuming no commission fees, market efficiency implies full unpredictability of prices. If the information used for predicting future prices is only previous price observations, the form of the Efficient Market Hypothesis is called weak. If a market is efficient, a trader can not exceed the profit made when the trader buys the asset and does not trade it. In the case of market efficiency, the expected value of a future price is the current value of the price. Therefore, prices in an efficient market follow a martingale model. In mathematical finance, the martingale property is one of the assumptions made when modeling prices [LeRoy, 1973, Lucas, 1978, Kyle, 1985, Black and Scholes, 1973, Heston, 1993]. On the other hand, in inefficient markets, the martingale property may not accurately describe price behavior. Besides the weak form of the Efficient Market Hypothesis, there are semi-strong and strong forms. In the semi-strong form, prices adjust to publicly available information, e.g., annual earnings and stock splits. The strong form states that all information, including non-public, is completely incorporated into current asset prices. The set of information for each subsequent form of efficiency (weak, semi-strong, strong) contains the previous one. That is, rejecting the weak form of the Efficient Market Hypothesis, other forms are rejected by definition. For companies issuing stocks, market efficiency implies that the price of the stock of the company already reflects information about decisions of the company and the valuation of these decisions by investors. In a market with complete information at each time, the matching of supply and demand should incorporate all the information in the market price. Thus, a company's market capitalization, that is the product of the number of outstanding stocks and the current market value of one stock, permits one to know the fair price of the company. Therefore, we have discussed the significance of the hypothesis of efficient

markets from three perspectives: company capitalization, trading strategies, and mathematical models.

Sometimes, instead of the term market efficiency, *information efficiency* is used [Williams, 1999, Boehmer and Kelley, 2009, Grossman, 1995, Lee, 2012]. The efficiency of markets is determined by the set of information and how quickly prices begin to reflect it. For instance, [Dann et al., 1977] found that prices incorporate information about a large volume of transactions during 10-15 minutes. [Busse and Clifton Green, 2002] found that positive TV news are incorporated into prices within one minute, while prices take 15 minutes to reflect negative TV news. [Grossman, 1995] stated that markets are not informationally efficient due to dynamic reallocation of investor funds unrelated to the use of new information. [Lee, 2012] showed that information efficiency of stock prices relates to readability of quarterly reports of the companies.

According to [Grossman and Stiglitz, 1980], prices can not fully reflect the available information since information is not free. Accordingly, if the markets are efficient, then those who have spent resources on obtaining information will not receive a profit. [Bouchaud et al., 2009] noted that markets can not be fully efficient, because traders who possess information would have the same profits as other traders. That is, there would be no motivation for informed traders to trade in an efficient market. According to [Bouchaud, 2005], random fluctuations of asset prices are explained by the presence of informed traders and market makers, whose profit comes from the bid-ask spread. Information received by traders can not be used instantly, but it can be reflected in prices after some short period of time. Therefore, in this thesis, I explore the case of prices at a high frequency. Thus, I consider prices provided at the end of each minute of a trading day. I also explore ultra-high frequency data that refer to the recording of every execution [Engle, 2000]. For such microscopic data, predictability of time series is expected [Lillo and Farmer, 2004]: long memory of buy and sell orders is caused by news arrivals, the feedback model [Shiller, 2003], or splitting large order into small pieces [Kyle, 1985, Chan and Lakonishok, 1995]. By aggregating prices to lower time frequencies, information on both the microstructure of price dynamics and the impact of news on prices is lost.

1.2 Entropy

By definition, in an efficient market, all available information is already incorporated into the current price of an asset traded at the market. Therefore, a future price contains all the new information that was not available one time step before. We consider changes in prices as a realization of a random process whose outputs are symbols from some finite alphabet. The averaged amount of information gained with each new symbol of a process is called Shannon entropy. L. Gulko in his paper about the entropic market hypothesis [Gulko, 1999] says "If entropy is an index of the collective market uncertainty about the future price changes, then in informationally efficient markets the entropy will be maximized." As noted by [Billingsley, 1965], the amount of information that is received with a new symbol of a sequence is the same as the amount of uncertainty before obtaining that symbol. Indeed, the greater the uncertainty about which symbol will be received next, the more new information will be obtained after its occurrence. That is, the state of market efficiency should correspond to the maximum value of entropy. We use the Shannon entropy as a measure of the randomness of a sequence. The less the degree of efficiency, the more predictable price returns [Eom et al., 2008].

The Shannon entropy is defined for a stationary process that generates symbols from a finite alphabet A. A process is called stationary if the probabilities of symbols and blocks of symbols are constant over time. The realization of a process is called typical if the empirical distribution of blocks of symbols of any length converges to its theoretical probability. If almost every realization

of a stationary process is typical, the process is called ergodic. That is, characteristics of an ergodic process can be recovered from one typical realization[1]. We denote by x_1^n a realization of the process with length n and by $p(x_1^n)$ the probability of the appearance of a particular sequence x_1^n. The larger the length of the sequence, the less the probability of the appearance of a certain realization of the process. For ergodic processes, the decrease of $p(x_1^n)$ is almost surely (a.s.) exponential in n. A constant exponential rate is called entropy rate, h.

Theorem 1 (The Shannon-McMillan-Breiman Theorem [Breiman, 1957])**.** *Let p be an ergodic measure on the space A^∞, where A is finite, $x_1^n \in A^n$. There is a non-negative number h called entropy rate such that*

$$\lim_{n\to\infty} \frac{1}{n} \log_{|A|} \frac{1}{p(x_1^n)} = h \text{ a.s.}$$

From now on, we use log with the convention that the base is the size of the alphabet, $|A|$. The probability of occurrence of a particular realization decreases faster with a larger value of the entropy rate. If all symbols appear independently with the same probability, then the entropy rate is equal to 1. This is its maximum possible value obtained in the case of maximum uncertainty.

The measure of uncertainty for a random variable, X, is also determined by entropy. Given probability distribution $p(x), x \in A$, of the discrete random variable X, entropy $H(X)$ is defined as the following [Shannon, 1948].

$$H(X) = -\sum_{x\in A} p(x) \log p(x), \tag{1.1}$$

Using the constraint $\sum_{x\in A} p(x) = 1$, it can be shown that entropy attains its maximum when all symbols from alphabet A have the same probability. For the entropy of a process, this formula has the following form.

Definition 1. *Let X be a stationary random process with a finite alphabet A and a measure p. A k-th order entropy of X is*

$$H_k(X) = -\sum_{x_1^k \in A^k} p(x_1^k) \log p(x_1^k) \tag{1.2}$$

with the convention $0 \log 0 = 0$.

Definition 2. *A process entropy of a process X is defined as the ratio of k-th order entropy and the length of blocks, k, when k tends to infinity.*

$$h(X) = \lim_{k\to\infty} \frac{H_k(X)}{k}$$

The proof of this theorem in the case when the source generated symbols is a Markov process can be found in "A mathematical theory of communication" by [Shannon, 1948]. For ergodic processes, process entropy coincides with entropy rate[2]. That is, for estimating the entropy rate, the definition of process entropy can be used. Further, we assume that both quantities are equal and call them entropy. We also assume that all processes under consideration are ergodic, that is, we can calculate entropy using one given realization.

[1] More precisely, the property of ergodic processes is described in the typical-sequence theorem (Theorem 1.4.1 in [Shields, 1996])

[2] For the proof, see Theorem 1.6.9 in [Shields, 1996].

In addition to interpretation of entropy in terms of the probabilities of symbolic blocks, there are other definitions of entropy for a process and its realization. An alternative approach for evaluation of a process entropy is using recurrence times [Willems, 1989]. The recurrence time, R_k, is the time of waiting for the occurrence of the same block of symbols:

$$R_k(x_1^n) = \inf\{m \geq 1 : x_{m+1}^{m+k} = x_1^k\}$$

According to the recurrence-time theorem [Wyner and Ziv, 1989, Kontoyiannis, 1998], exponential growth of the recurrence time with the growth of k is the process entropy. The higher the entropy, the longer it takes before the sequence repeats its beginning.

Theorem 2. *For any ergodic process with entropy h,*

$$\lim_{k\to\infty} \frac{1}{k} \log\left(R_k(x_1^n)\right) = h \ \ a.s.$$

Furthermore, entropy relates to a coding algorithm. Given a sequence of length n, x_1^n, Lempel-Ziv (LZ) algorithm [Lempel and Ziv, 1976] compresses it to a new sequence, ω_n. The new sequence ω_n consists of new words that are blocks of symbols with different lengths. Starting from the word that is the first symbol of the sequence x_1^n, new words occur according to the following rule. The next new word is the shortest word not yet included in the word list, but presented in the given sequence. The more random the given sequence, the more words needed to code the sequence. More precisely, as shown by [Ziv, 1978], the number of words used, $C(n)$, multiplied by $\frac{\log n}{n}$ converges to the process entropy almost surely.

Theorem 3. *For any ergodic process with entropy h,*

$$\lim_{n\to\infty} \frac{\log n}{n} C(n) = h \ \ a.s.$$

where $C(n)$ is the number of words used in the LZ algorithm [Lempel and Ziv, 1976].

1.3 Contribution of the following chapters

The next chapters are devoted to estimating the entropy of financial time series. The analysis carried out examines the changes in price predictability over time and the sources of this predictability. Also, the results allow testing the hypothesis of efficient markets. We show that the level of inefficiency depends on the considered market. Also, the conclusions of the EMH testing depend on both the chosen length of the time interval and the frequency of the data. We distinguish two reasons for the predictability of financial time series. The first reason is the empirical properties of price returns and the second reason is inefficiency of a market that implies the development of profitable trading strategies.

We discuss a method for estimating entropy and statistical properties of the entropy estimation in Chapter 2. In order to obtain an estimate of process entropy from a realization of a finite length, we rely on the definition of k-th order entropy from Definition 1. We use the Empirical Frequencies method [Marton and Shields, 1994]. This method involves estimating the probabilities of blocks of symbols blocks empirically. First, we review this method in Section 2.1. The modification of this method proposed by [Grassberger, 2008] is discussed in Section 2.2. We propose a modification of the method useful in a case when a part of data is missing in Section 2.3. Finally, we derive approximations of the expected value and variance of the entropy estimation in Section 2.4. To obtain the approximations up to the desired accuracy, we obtain a

recursive formula for the central moments of the multinomial distribution. Moreover, we obtain the unbiased estimation of the variance of the entropy estimation.

In Chapter 3, we analyze the structure of financial time series in terms of the contribution of stylized facts of financial markets to price predictability. We propose the method for filtering out four empirical properties of price returns, that we call data regularities, before estimating a degree of market efficiency in Section 3.3. The method includes filtering out intraday volatility pattern, volatility clustering, price staleness, and microstructure noise (Sections 3.4-3.8). The three-steps method for filtering out data regularities was proposed in [Calcagnile et al., 2020], while some steps were also applied in [Hsieh, 1991, Goldman, 2006]. We introduce the step of filtering out price staleness in Section 3.6. We show that price staleness increases a degree of price predictability using simulated and real data. We demonstrate that the method for filtering out 0-returns increases the entropy value and leaves those 0-returns that can be explained by price rounding. In Section 3.5, we take into account that price staleness affects the estimation of volatility and modify and optimize the method for volatility estimation proposed in [Sucarrat and Grønneberg, 2020]. Moreover, we take advantage of the Empirical Frequencies method and consider calculated empirical frequencies corresponding to price changes. More precisely, we show in Section 3.7 that there is a periodic pattern in price directions connected with a low entropy value.

Chapter 4 is devoted to price rounding and probability of getting a 0-return. We start with the probability of rounding introduced in [Bandi et al., 2020]. In Section 4.1, we provide a correction of this formula. With the provided correction, the sum of the probabilities that a price moves by a certain number of ticks is equal to one. We modify this formula by setting a non-zero bid-ask spread Section 4.2. Further, we expand this formula by considering t-distribution for price returns in Section 4.3. This modification takes into account fat tails of price returns.

In Chapter 5, we investigate the efficiency of the ETF market. We consider in more detail the last two steps of filtering out data regularities, namely price staleness and microstructure noise. In Section 5.2, we compare several approaches for filtering out 0-returns. The approaches take into account bid-ask spread and trading volumes. In Section 5.3, we discuss several methods for filtering out microstructure noise and compare them with each other. We discuss the phenomena of co-inefficiency found for the group of assets and its connection with cointegration of prices in Section 5.4. We extend the analysis by considering different lengths of time intervals where entropy is calculated. In Section 5.5, we consider monthly and quarterly time intervals. Finally, we discuss results obtained after filtering out data regularities for different lengths of time intervals in terms of the Efficient Market Hypothesis in Section 5.6. For instance, we show that the degree of inefficiency for the group of ETFs on weekly time intervals is slightly greater than the significance level for testing the EMH, 1%. The degree of inefficiency for the same group of assets on monthly time intervals is about 11%.

Chapter 6 investigates the degree of inefficiency of the Moscow stock exchange in the time period from 2012 to 2021. First, we modify the method for filtering out volatility clustering and price staleness in Section 6.3. We take into consideration a fact that estimation of volatility and a degree of price staleness are connected. The effect of price staleness is 0-returns that imply underestimation of volatility. We propose a modification of the exponentially weighted moving average estimation of volatility that takes into account these 0-returns. That is, we introduce a simple approach for estimating volatility and filtering out the 0-returns. For the estimation and filtering process, we use historical prices to apply the method for a real-time analysis. We conduct comparison analysis for options of the method in Section 6.4. We show that including price staleness into analysis decreases errors in volatility estimation. In Section 6.5, we update the method for detecting inefficient time intervals using two different discretizations simultaneously. Double-checking of the statistical significance of entropy values makes the results more robust.

We apply the novel methods for filtering out data regularities and estimating a degree of market efficiency to the real data in Section 6.6. First, we conclude that the degree of market inefficiency for monthly time intervals is equal to 82%. Then, we investigate in more detail two stocks that have the smallest efficiency rates corresponding to price randomness. We show that prices of the chosen pair of stocks demonstrate the lowest values of entropy in the years 2014 and 2015. Moreover, we continue considering frequencies of blocks of symbols and study what behavior of price is repeated more often during inefficient months. Finally, we introduce an illustrative example of a trading strategy based on frequencies of blocks. We show that this trading strategy allows us to predict a direction of price in an inefficient month. Finally, we study a common behavior for a group of stocks in Section 6.7. First, we use the Kullback-Leibler distance to group stocks into clusters. Using this distance, we demonstrate that the stocks of banks and oil companies traded at the Moscow Stock Exchange cluster together. Then, we introduce entropy of co-movement and co-movement divergence as alternative measures of closeness used for stock market clustering. These two measures suggest that stocks of companies belonging to the same cluster group in one cluster according to their price movements.

In Chapter 7, we introduce a statistical test for detecting changes in the value of entropy. We test the null hypothesis about equal entropies of two time series in Section 7.2. In such a way, we get rid of using time-demanding Monte-Carlo simulations used for constructing confidence intervals. Utilizing a z-score used in the hypothesis testing, we construct the method for determining an optimal bandwidth of time interval used to estimate entropy in Section 7.3. We show that the method allows us to find the length of interval with entropy different from the rest of the sequence in Section 7.4. We apply the statistical test for all adjacent intervals with the optimal bandwidth to meme and IT stocks in Section 7.6. We detect significantly low entropies for both types of stocks. Moreover, we show that entropy of price returns of meme and IT stocks are time-varying during 2020 and 2021. That is, we conclude that the process generated price returns is not stationary even after filtering out all mentioned data regularities. Furthermore, we show that the entropy of some meme stocks became statistically significantly low before the observed boom of their prices in January 2021. We discover the connection between low entropy values and high trading volumes, although information about trading volumes is not used to calculate entropy.

Chapter 8 contributes both to the methodological part of the thesis and to the application to a new dataset with a full record of transactions. First, we describe ultra-high frequency data for the group of assets in Section 8.2. The difference between this dataset and high-frequency data in other chapters is that price aggregation by minutes is not applied. Then, we propose a statistical test for unpredictability of time series in Section 8.3. We employ our estimations of the bias and variance of entropy estimation to test if the entropy of a given sequence is at the maximum. Knowing the distribution of entropy estimation allows us to make multiple tests for unpredictability controlling the percentage of false positive errors in hypothesis testing. We focus on the analysis of the tick-by-tick data of Apple stock in Section 8.4. In particular, we examine properties of assets prices during days with the revealed predictability of these prices. We compare values such as daily price changes, the fraction of 0-returns, and the amount of assets traded, averaged over days with and without detected predictability. Finally, we expand our analysis to assets that differ in average prices, volatility, and the frequency of trading in Section 8.5.

Chapter 2

The computation of the Shannon entropy

In this chapter, we discuss how to estimate the entropy of a sequence of symbols from a finite alphabet. To evaluate the randomness of a sequence, we estimate the entropy using the *Empirical Frequencies* method [Marton and Shields, 1994]. In Section 2.2, we discuss the modification of the method proposed in [Grassberger, 1988, 2022] that reduces the bias of entropy estimation. For applications where a part of the data is missing, we modify the Empirical Frequencies method in Section 2.3. Finally, we investigate the estimation of the entropy of a random process as a random variable in Section 2.4. We present results discussed in Sections 2.3 and 2.4 in articles [Shternshis et al., 2022a] and [Shternshis and Mazzarisi, 2022], respectively.

2.1 Empirical Frequencies method

The Empirical Frequencies (EF) method is used to estimate entropy from a given finite sequence of symbols [Marton and Shields, 1994]. The method includes the calculation of empirical probabilities of blocks of symbols and substituting them in the formula for the entropy in Equation 1.2. The process is assumed to be ergodic[3] with a positive entropy h. We define a shift-invariant Borel probability measure p on the space A^∞ of sequences $x = \{x_n\}$ drawn from a finite alphabet A. Let $k \leq n$ and let p_k be the true distribution of k-blocks (blocks with length of k). Fixing the length of blocks of symbols, we consider empirical frequencies that are the actual amount of times when a block of symbols appears in the given sequence. For each $a_1^k \in A^k$, empirical frequencies are

$$f(a_1^k|x_1^n) = \#\{i \in [1, n-k+1] : x_i^{i+k-1} = a_1^k\} \tag{2.1}$$

where $x_1^n \in A^n$ and $x_i^{i+k-1} = x_i \dots x_{i+k-1}$.

Definition 3. *For each $a_1^k \in A^k$, empirical probabilities are defined as*

$$\hat{p}_k(a_1^k|x_1^n) = \frac{f(a_1^k|x_1^n)}{n-k+1} \tag{2.2}$$

A naive estimation of k-th order entropy from Equation 1.2 is defined as follows.

[3]Statistical features of an ergodic process can be deduced from a single typical realization.

Definition 4. *Empirical k-entropy is defined by*

$$\hat{H}_k(x_1^n) = -\sum_{a_1^k} \hat{p}_k(a_1^k|x_1^n) \log\left(\hat{p}_k(a_1^k|x_1^n)\right) \tag{2.3}$$

According to the following Theorem introduced in [Shields, 1996] (Theorem II.3.5-6), the Empirical Frequencies method gives a consistent estimate of the process entropy.

Theorem 4. *If p is an ergodic measure of entropy $h > 0$, if $k(n) \to \infty$ as $n \to \infty$, and if $k(n) \leq \frac{\log n}{h}$, then*

$$\lim_{n\to\infty} \frac{1}{k(n)} \hat{H}_{k(n)}(x_1^n) = h \ \textit{a.s.}$$

According to Theorem 4, if the length of blocks is not larger than $\log n/h$, then the scaled estimation converges to the process entropy. The restriction on the length of the blocks allows to get a sufficient number of blocks for a careful estimation of the probabilities of blocks. Thus, setting k less than $\lfloor \log(n) \rfloor$, the estimation of the process entropy is obtained by the following formula.

$$\hat{h}_k = \frac{\hat{H}_k}{k}$$

To write an efficient code for the entropy calculation, we notice that we can use notation for symbols from 0 to $|A|-1$ for an alphabet with size $|A|$ so that $A = \{0, 1, \ldots |A|-1\}$. Then, each block of k symbols corresponds to the number from 0 to $|A|^k - 1$ in the number system with base $|A|$. For example, the block of repeating symbol 0 k times is numbered by 0.

2.2 Grassberger's correction

The convergence of the estimate by the Empirical Frequencies method is achieved in Theorem 4 when the length of the sequence tends to infinity. However, when the length of the sequence is finite, the empirical probabilities have estimation errors. These errors cause a downward bias in the estimation of entropy. The estimator introduced by [Grassberger, 2008], $\hat{h}_k^G$, is defined in order to correct for the bias, so that $E(\hat{h}_{k(n)}^G) \approx h$ for samples of length n. More precisely, let $f_i, i = 0, \ldots, M-1$, be the empirical frequencies of all possible k-blocks defined in Eq. 2.1, where $M = |A|^k$. The number of blocks in consideration is given by $n_b = n - k + 1$. Then, the entropy estimate

$$\hat{H}_k = -\sum_{i=0}^{M-1} \frac{f_i}{n_b} \log \frac{f_i}{n_b} = \log(n_b) - \frac{1}{n_b} \sum_{i=0}^{M-1} f_i \log f_i$$

is replaced by

$$\hat{H}_k^G = \log(n_b) - \frac{1}{n_b} \sum_{i=0}^{M-1} f_i \log\left(\exp G(f_i)\right), \tag{2.4}$$

where the sequence $G(i)$ is defined recursively as

$$
\begin{aligned}
G(1) &= -\gamma - \ln(2) \\
G(2) &= 2 - \gamma - \ln(2) \\
G(2n+1) &= G(2n) \\
G(2n+2) &= G(2n) + \frac{2}{2n+1},\ n \geq 1
\end{aligned}
$$

with the Euler–Mascheroni constant $\gamma = \lim_{n\to\infty} \left(\sum_{k=1}^{n} \frac{1}{k} - \ln n\right) \approx 0.577$. We estimate the process entropy as

$$\hat{h}_k^G = \frac{\hat{H}_k^G}{k} \tag{2.5}$$

2.3 Entropy estimation in case of missing values

In some applications of information theory, a certain number of symbols generated by a random process may be not available. We refer to [Kim et al., 2012, Cirugeda-Roldan et al., 2014, Dong et al., 2019] as the examples of such applications. We adopt the method of Empirical Frequencies for the case of the presence of missing values in data. First, we need to choose a suitable value for the length of blocks, k. After choosing a correct value of k, we consider partitions of the sequence in blocks that do not contain missing values. If k is chosen properly so that it can be used to estimate entropy from a given finite realization of a process, the value of k is called admissible.

Definition 5. *A non-decreasing sequence $k(n) \leq n$ is admissible if*

$$\lim_{n\to\infty} |\hat{p}_{k(n)}(\cdot|x_1^n) - p_{k(n)}| = 0 \ \ a.s.$$

where the distance between two measures p and q on A^k is

$$|p - q| = \sum_{a_1^k} |p(a_1^k) - q(a_1^k)|.$$

We will rely on two theorems formulated for the case of complete data. See [Marton and Shields, 1994] for their proofs.

Theorem 5. *If $k(n) \geq \log(n)/(h-\epsilon)$, where h is the entropy of the process, $\epsilon > 0$, then $k(n)$ is not admissible for the process.*

Theorem 6. *If a process is i.i.d., Markov, or ϕ-mixing and $k(n) \leq \log(n)/(h+\epsilon)$, where h is the entropy of the process, $\epsilon > 0$, then $k(n)$ is admissible for the process.*

Now, let's denote the amount of k-blocks as $n_b(k)$. The exact value of $n_b(k)$ is unknown without knowing the location of missing values. The proof of the Theorem 5 is based on the fact that, when k is large, the number of all distinct blocks in the sequence is bounded by the value $|A|^{k(h-\epsilon)}$ that is not enough to observe all "typical" blocks for the process and thus to estimate the probabilities. Replacing $n(k)$ by $n_b(k)$ we can repeat the proof and update the lower bound of not admissible $k(n_b)$ to be equal to $\log(n_b)/(h-\epsilon)$. We set $|A|$ as the base of the logarithm.

We aim to prove the following theorem for the case of having missing values in the output of the random process. Missing values are written over the symbols of the process into consideration.

Theorem 7. *Assume that the processes of generating symbols and missing values are independent. If a process generating symbols is i.i.d., Markov, or ϕ-mixing and $k(n_b) \leq \log(n_b)/(h+\epsilon)$, then $k(n_b)$ is admissible for the process.*

Proof of Theorem 7. Since $k(n_b) \leq \log(n_b)/(h+\epsilon) \leq \log(n)/(h+\epsilon)$, then $k(n_b)$ is admissible if the data is complete. Note that the number of blocks is greater than or equal to $|A|^{(k/(h+\epsilon))} > 0$. Let's fix a_1^k. Without missing values $\hat{p}_k^n = \hat{p}_k(a_1^k|x_1^n) = \frac{f_k^n}{n-k+1}$ and $\hat{p}_k^n \to p_k$ a.s. When passing to the limit, it is assumed that $n \to \infty$. If $N(n)$ values are missing, then empirical frequencies of observed blocks (blocks without missing values), $\bar{p}_k$, are defined as follows.

$$\begin{aligned}\bar{p}_k(a_1^k|x_1^n) &= \frac{f_k^n - c(n)}{n-k+1-d(n)} \\ &= \frac{\hat{p}_k^n(n-k+1) - c(n)}{n-k+1-d(n)}\end{aligned}$$

where $d(n)$ is the total number of blocks eliminated, $N(n) \leq d(n) \leq k(n)N(n)$, and $c(n)$ is the number of blocks a_1^k eliminated, $0 \leq c(n) \leq k(n)N(n)$. $n-k+1-d(n) = n_b > 0$.

There are two possible cases: I. $d(n)/(n-k+1) \to 1$ II. $d(n)/(n-k+1) \to C, 0 \leq C < 1$.

Case I is in contradiction with the fact $n-k+1-d(n) = n_b > 0$ since $1 - \frac{d(n)}{n-k+1} \to 0$.

Case II: Assume that if a symbol x_i is missing, it has a label (subindex) equal to 1, $I(x_i) = 1$, and otherwise $I(x_i) = 0$. Let's introduce $B_k \in A^k$ such that $B_k = \{x_i^{i+k-1}, \exists j \in \{i, \ldots, i+k-1\} : I(x_j) = 1, i \in \{1 \ldots n-k+1\}\}$. B_k is all blocks containing missing values. Applying the Birkhoff's ergodic theorem [Katznelson and Weiss, 1982, Kamae, 1982] with the characteristic function χ_B, we get that $p(B_k|x_1^n) \to p(B_k)$. Taking $D_{a_1^k} = B_k \cap [a_1^k]$ and proceeding in the similar way, we also get that $p(D_{a_1^k}|x_1^n) \to p(D_{a_1^k})$. Here, $[a_1^k] = \{x_i^{i+k-1} : a_j = x_{j+i-1}, j \in \{1, \ldots, k\}, i \in \{1, \ldots, n-k+1\}\}$.

Now, $\frac{d(n)}{n} \to p(B_k) = C$ and by independence $\frac{c(n)}{n} \to p(D_{a_1^k}) = Cp_k(a_1^k)$. Therefore,

$$\frac{\hat{p}_k^n(n-k+1) - c(n)}{n-k+1-d(n)} = \frac{\hat{p}_k^n - \frac{c(n)}{n-k+1}}{1 - \frac{d(n)}{n-k+1}} \to \frac{p_k - Cp_k}{1-C} = p_k$$

Therefore, the values of k such that $k(n_b) \leq \log(n_b)/(h+\epsilon)$ are admissible in the case of missing values. □

In practice, we take $k = max(K : K < log(n_b(K)))$.

2.4 Variance and mean of entropy estimation

In this section, we investigate statistical properties of a plug-in estimator of entropy from Equation 1.1. More precisely, we obtain a formula for the variance of the estimator of the Shannon entropy using central moments of binomial and multinomial distributions. The central moments are calculated by a new recursive approach. This helps us to find the approximation of the variance with an accuracy of order $O(n^{-4})$, where n is the length of the sequence of events. It is possible to further extend such an approximation by using the proposed approach to compute higher orders of the central moments associated with the multinomial distribution.

Let's assume that there are M events which can appear with probabilities $p_0, p_1, \ldots, p_{M-1}$, $\sum_{j=0}^{M-1} p_j = 1$. We assume that all p_j are positive, because zero probabilities do not affect the

entropy value $H = -\sum_{j=0}^{M-1} p_j \ln p_j$, where ln is the natural logarithm. If events appear independently n times in total, the frequencies of events f_0, f_1, f_{M-1} follow a multinomial distribution. Each frequency is distributed as Binomial $B(p_j, n)$. Therefore, the estimation of p_j, $\hat{p}_j = \frac{f_j}{n}$, is distributed as $B(p_j, n)/n$. The aim of this section is to find the variance of a random variable $\hat{H} = -\sum_{j=0}^{M-1} \hat{p}_j \ln \hat{p}_j$. The variable $\hat{H}$ is an empirical entropy from Equation 2.3 with the base e of the logarithm assuming that blocks $a_1^k \in A^k$ are generated independently. If all blocks have a non-zero probability, $M = |A|^k$, where $|A|$ is the size of the alphabet A.

Theorem 8 (Approximation of variance)**.** *Let's assume that f_j, $j = 0 \dots M-1$, are distributed as multinomial variables $f^M(p_0, \dots, p_{M-1}, n)$ and $\hat{H} = -\sum_{j=0}^{M-1} \frac{f_j}{n} \ln \frac{f_j}{n}$. Then,*

$$Var(\hat{H}) = \frac{1}{n}\left[-H^2 + \sum_j p_j \ln^2(p_j)\right] + \frac{1}{n^2}\left[\frac{M}{2} - \frac{1}{2}\right] + \frac{1}{6n^3}\left[(1-H)\sum_j \frac{1}{p_j} - \sum_j \frac{\ln p_j}{p_j} - 1\right] + O(n^{-4}) \tag{2.6}$$

We first give useful propositions for proving Theorem 8. In particular, we present a recursive formula for the central moments of the multinomial distribution of $(\hat{p}_1, \hat{p}_2)$ in Proposition 1. Also, we find the expression for the expectation of $\hat{p}^2 \ln^2(\hat{p})$ that appears in $\hat{H}^2$ using the Taylor expansion in Proposition 4. Using all needed central moments of binomial and multinomial distributions, we calculate $Var(\hat{H}) = E(\hat{H}^2) - E(\hat{H})^2$. Further, we derive unbiased estimation of the variance in Theorem 9.

Proposition 1 (Central moments of multinomial distribution)**.** *Central moments of random variables following multinomial distribution $f^M(p_1, p_2, n)$ divided by n (/n) can be defined recursively using the formulas below.*

$$\begin{aligned} \mu_{1,0} &= 0, \mu_{1,1} = -\frac{p_1 p_2}{n} \\ \mu_{m+1,k} &= \frac{p_1}{n}\left[(1-p_1)\frac{\partial}{\partial p_1}\mu_{m,k} - p_2 \frac{\partial}{\partial p_2}\mu_{m,k} + (1-p_1) m \mu_{m-1,k} - p_2 k \mu_{m,k-1}\right] \end{aligned} \tag{2.7}$$

Proof of Proposition 1.

$$\mu^M_{m,k}(p_1, p_2, n) = \sum_{x_1 \geq 0, x_2 \geq 0, x_1 + x_2 \leq n} (x_1 - np_1)^m (x_2 - np_2)^k \frac{n!}{x_1! x_2! (n - x_1 - x_2)!} p_1^{x_1} p_2^{x_2} q^{n - x_1 - x_2}$$

where $\mu^M_{m,k}$ is the (m,k)-central moment of the multinomial distribution and $q = 1 - p_1 - p_2$. We can show that

$$\begin{aligned} \frac{\partial}{\partial p_1}\mu^M_{m,k} &= -nm\mu^M_{m-1,k} + \frac{1-p_2}{p_1 q}\mu^M_{m+1,k} + \frac{1}{q}\mu^M_{m,k+1} \\ \frac{\partial}{\partial p_2}\mu^M_{m,k} &= -nk\mu^M_{m,k-1} + \frac{1-p_1}{p_2 q}\mu^M_{m,k+1} + \frac{1}{q}\mu^M_{m+1,k} \end{aligned}$$

Solving the system for $\mu^M_{m+1,k}$, we get that

$$\mu^M_{m+1,k} = p_1\left[(1-p_1)\frac{\partial}{\partial p_1}\mu^M_{m,k} - p_2 \frac{\partial}{\partial p_2}\mu^M_{m,k} + (1-p_1) mn \mu^M_{m-1,k} - p_2 kn \mu^M_{m,k-1}\right]$$

Taking into account that $\mu^M_{m,k} = n^{m+k}\mu_{m,k}$, we obtain the result

$$\mu_{m+1,k} = \frac{p_1}{n}\left[(1-p_1)\frac{\partial}{\partial p_1}\mu_{m,k} - p_2 \frac{\partial}{\partial p_2}\mu_{m,k} + (1-p_1) m \mu_{m-1,k} - p_2 k \mu_{m,k-1}\right]$$

and by symmetry

$$\mu_{m,k+1} = \frac{p_2}{n}\left[(1-p_2)\frac{\partial}{\partial p_2}\mu_{m,k} - p_1\frac{\partial}{\partial p_1}\mu_{m,k} + (1-p_2)k\mu_{m,k-1} - p_1 m\mu_{m-1,k}\right]$$

□

Proposition 2 (Central moments of binomial distribution)**.** *Central moments of the binomial distribution $B(p,n)/n$ can be defined recursively using the formulas below.*

$$\begin{aligned}\mu_0 &= 1\\ \mu_1 &= 0\\ \mu_{m+1} &= \frac{p(1-p)}{n}\left[m\mu_{m-1} + \frac{\partial}{\partial p}\mu_m\right]\end{aligned} \tag{2.8}$$

This is a special case of the previous proposition where $p_2 = k = 0$. It is known as the Renovsky formula [Riordan, 1937].

Proposition 3 (Bias of $\hat{p}\ln\hat{p}$)**.**

$$E(\hat{p}\ln\hat{p}) = p\ln p + \sum_{m=2}^{\infty}\frac{(-1)^m}{m(m-1)p^{m-1}}\mu_m \tag{2.9}$$

where $\hat{p} \sim B(p,n)/n$ and μ_m is its central m-moment.

The result of Proposition 3 was obtained by [Basharin, 1959]. It is derived by using the Taylor expansion around p.

$$\hat{p}\ln\hat{p} = p\ln p + (1+\ln p)(\hat{p}-p) + \sum_{m=2}^{\infty}\frac{(-1)^m}{m(m-1)p^{m-1}}(\hat{p}-p)^m$$

Taking the expected value, we obtain formula (2.9). A random variable following the Binomial distribution $B(p,n)$ divided by n has the mean p and the variance $\frac{p(1-p)}{n}$.

Proposition 4 (Second moment of $\hat{p}\ln\hat{p}$)**.** *Let $\hat{p} \sim B(p,n)/n$. Then,*

$$\begin{aligned}&E\left(\hat{p}^2\ln^2(\hat{p})\right) = p^2\ln^2(p) + (\ln^2 p + 3\ln p + 1)\mu_2 +\\ &+4\sum_{m=1}^{\infty}(-1)^{m+1}\left[\ln p - S_{m-1} + \frac{3}{2}\right]\frac{\mu_{m+2}}{m(m+1)(m+2)p^m}\end{aligned} \tag{2.10}$$

where $S_m = \sum_{k=1}^{m}\frac{1}{k}$.

Proof of Proposition 4. We consider the Taylor expansion of $\hat{p}^2\ln^2(\hat{p})$.

$$\begin{aligned}&\hat{p}^2\ln^2(\hat{p}) = p^2\ln^2(p) + 2p\ln p(\ln p + 1)(\hat{p}-p) + (\ln^2 p + 3\ln p + 1)(\hat{p}-p)^2 +\\ &+4\sum_{m=1}^{\infty}(-1)^{m+1}\left[\ln p - S_{m-1} + \frac{3}{2}\right]\frac{(\hat{p}-p)^{m+2}}{m(m+1)(m+2)p^m}\end{aligned}$$

This expression can be obtained by noticing that derivatives of $p^2\ln^2(p)$ starting from the third take the form

$$\frac{a_m\ln p + b_m}{p^m}$$

where $a_{m+1} = -ma_m$; $mb_m + b_{m+1} = a_m$ with $a_1 = 4$; $b_1 = 6$. The solution of the system is $a_m = 4(-1)^{m+1}(m-1)!$ and $b_m = 4(-1)^m(m-1)!(S_{m-1} - \frac{3}{2})$. The solution is unique because of the uniqueness of the Taylor series. Taking the expected value, we get the result. □

Proposition 5 (Covariance of $\hat{p}\ln\hat{p}$)**.** *Let $\hat{p}_1, \hat{p}_2 \sim f^M(p_1, p_2, n)/n$. Then,*

$$\begin{aligned}
&E(\hat{p}_1 \ln\hat{p}_1 \hat{p}_2 \ln\hat{p}_2) = p_1 p_2 \ln p_1 \ln p_2 + (\ln p_1 + 1)(\ln p_2 + 1)\mu_{1,1} + \\
&+ \sum_{m=2}^{\infty} \frac{(-1)^m}{m(m-1)} \left[p_1 \ln p_1 \frac{1}{p_2^{m-1}} \mu_{0,m} + p_2 \ln p_2 \frac{1}{p_1^{m-1}} \mu_{m,0} \right] + \\
&+ \sum_{m=2}^{\infty} \frac{(-1)^m}{m(m-1)} \left[(\ln p_1 + 1) \frac{1}{p_2^{m-1}} \mu_{1,m} + (\ln p_2 + 1) \frac{1}{p_1^{m-1}} \mu_{m,1} \right] + \\
&+ \sum_{m=2}^{\infty} \sum_{k=2}^{\infty} \frac{(-1)^{m+k}}{m(m-1)k(k-1) p_1^{m-1} p_2^{k-1}} \mu_{m,k}
\end{aligned} \tag{2.11}$$

where $\mu_{m,k}$ are (m,k)-central moments of $f^M(p_1, p_2, n)/n$.

Proof of Proposition 5.

$$\begin{aligned}
&\hat{p}_1 \ln\hat{p}_1 \hat{p}_2 \ln\hat{p}_2 = \sum_{m=0}^{\infty} \frac{(\hat{p}_1 - p_1)^m}{m!} \frac{d^m}{dp_1^m}(p_1 \ln p_1) \sum_{k=0}^{\infty} \frac{(\hat{p}_2 - p_2)^k}{k!} \frac{d^k}{dp_2^k}(p_2 \ln p_2) = \\
&= p_1 p_2 \ln p_1 \ln p_2 + p_1 \ln p_1 (\ln p_2 + 1)(\hat{p}_2 - p_2) + p_2 \ln p_2 (\ln p_1 + 1)(\hat{p}_1 - p_1) + \\
&+ p_1 \ln p_1 \sum_{k=2}^{\infty} \frac{(-1)^k}{k(k-1)p_2^{k-1}} (\hat{p}_2 - p_2)^k + p_2 \ln p_2 \sum_{m=2}^{\infty} \frac{(-1)^m}{m(m-1)p_1^{m-1}} (\hat{p}_1 - p_1)^m + \\
&+ (\ln p_1 + 1)(\hat{p}_1 - p_1) \sum_{k=2}^{\infty} \frac{(-1)^k}{k(k-1)p_2^{k-1}} (\hat{p}_2 - p_2)^k + (\ln p_2 + 1)(\hat{p}_2 - p_2) \sum_{m=2}^{\infty} \frac{(-1)^m}{m(m-1)p_1^{m-1}} (\hat{p}_1 - p_1)^m + \\
&+ (\ln p_1 + 1)(\ln p_2 + 1)(\hat{p}_1 - p_1)(\hat{p}_2 - p_2) + \sum_{m=2}^{\infty} \sum_{k=2}^{\infty} \frac{(-1)^{m+k}}{m(m-1)k(k-1)p_1^{m-1}p_2^{k-1}} (\hat{p}_1 - p_1)^m (\hat{p}_2 - p_2)^k
\end{aligned}$$

Therefore,

$$\begin{aligned}
&E(\hat{p}_1 \ln\hat{p}_1 \hat{p}_2 \ln\hat{p}_2) = p_1 p_2 \ln p_1 \ln p_2 + \\
&+ p_1 \ln p_1 \sum_{k=2}^{\infty} \frac{(-1)^k}{k(k-1)p_2^{k-1}} \mu_{0,k} + p_2 \ln p_2 \sum_{m=2}^{\infty} \frac{(-1)^m}{m(m-1)p_1^{m-1}} \mu_{m,0} + \\
&+ (\ln p_1 + 1) \sum_{k=2}^{\infty} \frac{(-1)^k}{k(k-1)p_2^{k-1}} \mu_{1,k} + (\ln p_2 + 1) \sum_{m=2}^{\infty} \frac{(-1)^m}{m(m-1)p_1^{m-1}} \mu_{m,1} + \\
&+ (\ln p_1 + 1)(\ln p_1 + 1)\mu_{1,1} + \sum_{m=2}^{\infty} \sum_{k=2}^{\infty} \frac{(-1)^{m+k}}{m(m-1)k(k-1)p_1^{m-1}p_2^{k-1}} \mu_{m,k}
\end{aligned}$$

□

Now, we derive the formula for the variance of entropy estimation given in Theorem 8.

Proof of Theorem 8.

$$\begin{aligned}
Var(\hat{H}) &= E(\hat{H}^2) - E(\hat{H})^2 \\
&= \sum_{j=0}^{M-1} E(\hat{p}_j^2 \ln^2 \hat{p}_j) + \sum_{j=0}^{M-1} \sum_{i=0, i\neq j}^{M-1} E(\hat{p}_j \ln\hat{p}_j \hat{p}_i \ln\hat{p}_i) - E(\hat{H})^2
\end{aligned}$$

For calculations we need all moments of orders n^{-1}, n^{-2}, n^{-3} obtained using Equations 2.7 and 2.8.

$$\begin{aligned}
\mu_2 &= \frac{p(1-p)}{n} \\
\mu_3 &= \frac{p(1-p)(1-2p)}{n^2} \\
\mu_4 &= \frac{3p^2(1-p)^2}{n^2} + \frac{p(1-p) - 6p^2(1-p)^2}{n^3} \\
\mu_5 &= \frac{10p^2(1-p)^2(1-2p)}{n^3} + O(n^{-4}) \\
\mu_6 &= \frac{15p^3(1-p)^3}{n^3} + O(n^{-4})
\end{aligned} \tag{2.12}$$

$$\begin{aligned}
\mu_{2,1} &= -\frac{p_1p_2(1-2p_1)}{n^2} \\
\mu_{3,1} &= -\frac{3p_1^2(1-p_1)p_2}{n^2} + \frac{6p_1^2(1-p_1)p_2 - p_1p_2}{n^3} \\
\mu_{4,1} &= -\frac{10p_1^2(1-p_1)(1-2p_1)p_2}{n^3} + O(n^{-4}) \\
\mu_{5,1} &= -\frac{15p_1^3(1-p_1)^2p_2}{n^3} + O(n^{-4})
\end{aligned}$$

$$\begin{aligned}
\mu_{2,2} &= \frac{p_1p_2(1-p_1)(1-p_2) + 2p_1^2p_2^2}{n^2} + \frac{p_1p_2 - 2p_1p_2(1-p_1)(1-p_2) - 4p_1^2p_2^2}{n^3} \\
\mu_{3,2} &= \frac{10p_1^2p_2^2(1-2p_1) + p_1p_2(1-p_1-p_2)(1-5p_1)}{n^3} \\
\mu_{3,3} &= -\frac{9p_1^2p_2^2(1-p_1)(1-p_2) + 6p_1^3p_2^3}{n^3} + O(n^{-4}) \\
\mu_{4,2} &= \frac{3p_1^2(1-p_1)p_2((1-p_1)(1-p_2) + 4p_1p_2)}{n^3} + O(n^{-4})
\end{aligned}$$

Moments with $m + k \leq 4$ coincide with results obtained in [Harris, 1975, Ouimet, 2021]. After summing up $E[\hat{p_j} \ln \hat{p_j}]$ in Eq. 2.9 for all j, the expression becomes

$$E(\hat{H}) = -E(\sum_j \hat{p}_j \ln(\hat{p}_j)) = H - \frac{M-1}{2n} + \frac{1}{12n^2}\left(1 - \sum_{j=0}^{M-1}\frac{1}{p_j}\right) + \frac{1}{12n^3}\sum_{j=0}^{M-1}\left(\frac{1}{p_j} - \frac{1}{p_j^2}\right) + O(n^{-4}) \tag{2.13}$$

where $H = -\sum_j p_j \ln(p_j)$, and $\mu_2, \mu_3, \mu_4, \mu_5, \mu_6$ are used. Similar estimates of the bias of entropy estimation were obtained in other works, see, e.g., [Harris, 1975, Schürmann and Grassberger, 1996, Victor, 2000].

$$\begin{aligned}
E(\hat{H})^2 = H^2 &+ \frac{1}{n}\left[-(M-1)H\right] + \frac{1}{n^2}\left[\frac{M^2}{4} - \frac{M}{2} + \frac{1}{4} + \frac{H}{6}\left(1 - \sum_{j=0}^{M-1}\frac{1}{p_j}\right)\right] + \\
&+ \frac{1}{n^3}\left[\frac{H}{6}\sum_{j=0}^{M-1}\left(\frac{1}{p_j} - \frac{1}{p_j^2}\right) - \frac{M-1}{12}\left(1 - \sum_{j=0}^{M-1}\frac{1}{p_j}\right)\right]
\end{aligned}$$

The approximation of the second moment of $\hat{p}\ln(\hat{p})$ from Eq. 2.10 is

$$\begin{aligned}
&E(\hat{p}^2\ln^2(\hat{p})) = p^2\ln^2(p) + \frac{1}{n}\left(\ln^2 p + 3\ln p + 1\right)p(1-p) + \\
&+ \frac{1}{n^2}\left[\left(\frac{5}{6}p^2 - p + \frac{1}{6}\right)\ln p + \frac{7}{4}p^2 - \frac{5}{2}p + \frac{3}{4}\right] + \\
&+ \frac{1}{n^3}\left[\frac{1}{6}\left(p^2 - p\right)\ln p + \frac{p^2}{3} - \frac{p}{2} + \frac{1}{12} + \frac{1}{12p}\right] + O(n^{-4})
\end{aligned}$$

The approximation of the covariances from Eq. 2.11 is

$$\begin{aligned}
&E(\hat{p}_1\ln\hat{p}_1\hat{p}_2\ln\hat{p}_2) = \\
&= p_1p_2\ln p_1\ln p_2 + \frac{1}{n}\left[-(\ln p_1+1)(\ln p_2+1)p_1p_2 + \frac{1}{2}\left(p_1\ln p_1(1-p_2) + p_2\ln p_2(1-p_1)\right)\right] + \\
&+ \frac{1}{n^2}\left[\frac{5}{12}p_1p_2(\ln p_1 + \ln p_2) + \frac{1}{12}\left(\frac{p_1}{p_2}\ln p_1 + \frac{p_2}{p_1}\ln p_2\right) + \frac{1}{4}(1 + 7p_1p_2 - p_1 - p_2)\right] + \\
&+ \frac{1}{n^3}\left[\frac{1}{12}p_1\ln p_1\left(p_2 + \frac{1}{p_2^2}\right) + \frac{1}{12}p_2\ln p_2\left(p_1 + \frac{1}{p_1^2}\right) + \frac{1}{3}p_1p_2 + \frac{1}{24}\left(\frac{p_1}{p_2} + \frac{p_2}{p_1} + \frac{1}{p_1} + \frac{1}{p_2} - p_1 - p_2\right)\right] + \\
&+ O(n^{-4})
\end{aligned}$$

Summing up for all indexes j, i of the second moments and covariances, we get that

$$\begin{aligned}
E(\hat{H}^2) = H^2 &+ \frac{1}{n}\left[-H^2 + \sum_j p_j\ln^2(p_j) - (M-1)H\right] + \\
&+ \frac{1}{n^2}\left[\frac{H}{6}\left(1 - \sum_j\frac{1}{p_j}\right) + \frac{1}{4}M^2 - \frac{1}{4}\right] + \\
&+ \frac{1}{n^3}\left[\frac{M}{12}\sum_j\frac{1}{p_j} + \frac{1}{12}\sum_j\frac{1}{p_j} - \frac{1}{12} - \frac{M}{12} - \frac{1}{6}H\sum_j\frac{1}{p_j^2} - \frac{1}{6}\sum_j\frac{\ln p_j}{p_j}\right] + O(n^{-4})
\end{aligned}$$

Therefore,

$$Var(\hat{H}) = \frac{1}{n}\left[-H^2 + \sum_j p_j\ln^2(p_j)\right] + \frac{1}{n^2}\left[\frac{M}{2} - \frac{1}{2}\right] + \frac{1}{6n^3}\left[(1-H)\sum_j\frac{1}{p_j} - \sum_j\frac{\ln p_j}{p_j} - 1\right] + O(n^{-4})$$

□

To estimate the variance of entropy estimation, $Var(\hat{H})$, from a given random sequence, we find the unbiased estimation of the variance of entropy estimation in Theorem 9.

Theorem 9. *Let's assume that f_j, $j = 0 \ldots M-1$, are distributed as multinomial variables $f^M(p_0, \ldots, p_{M-1}, n)$ and $\hat{H} = -\sum_{j=0}^{M-1}\frac{f_j}{n}\ln\frac{f_j}{n}$. In addition, let's assume that all events appear*

at least once. Then, $Var(\hat{H}) = E(\hat{Var}) + O(n^{-4})$*, where*

$$\hat{Var} = \frac{1}{n}(\sum_j \hat{p}_j \ln^2 \hat{p}_j - \hat{H}^2) + \frac{1}{n^2}\left[\sum_j \hat{p}_j \ln^2 \hat{p}_j - \hat{H}^2 - M\hat{H} - \sum_j \ln \hat{p} - \frac{M}{2} + \frac{1}{2}\right]$$
$$+ \frac{1}{n^3}\left[\sum_j \hat{p}_j \ln^2 \hat{p}_j - \hat{H}^2 - M\hat{H} - \sum_j \ln \hat{p}_j - \frac{\hat{H}}{3}\sum_j \frac{1}{\hat{p}_j} - \frac{1}{3}\sum_j \frac{\ln \hat{p}_j}{\hat{p}_j} - \frac{1}{12}\sum_j \frac{1}{\hat{p}_j} - \frac{M^2}{4} - \frac{M}{2} + \frac{5}{6}\right] \tag{2.14}$$

Proof of Theorem 9. From the proof of Theorem 8 we know that

$$E(\hat{H}^2) = H^2 + \frac{1}{n}\left[-H^2 + \sum_j p_j \ln^2(p_j) - MH + H\right] +$$
$$+ \frac{1}{n^2}\left[\frac{H}{6} - \frac{H}{6}\sum_j \frac{1}{p_j} + \frac{1}{4}M^2 - \frac{1}{4}\right] + O(n^{-3})$$

and

$$E(M\hat{H}) = MH + \frac{M - M^2}{2n} + O(n^{-2}).$$

We can show using Taylor series and moments μ_2, μ_3, μ_4 that

$$E\left(\sum_j \hat{p}_j \ln^2 \hat{p}_j\right) = \sum_j p_j \ln^2 p_j + \frac{1}{n}\left(\sum_j \ln p_j + H + M - 1\right) +$$
$$+ \frac{1}{n^2}\left(\frac{1}{6}\sum_j \frac{\ln p_j}{p_j} + \frac{M}{2} - \frac{1}{4}\sum_j \frac{1}{p_j} + \frac{H}{6} - \frac{1}{4}\right) + O(n^{-3})$$

and

$$E\left(\sum_j \ln \hat{p}_j\right) = \sum_j \ln p_j + \frac{1}{n}\left(\frac{M}{2} - \frac{1}{2}\sum_j \frac{1}{p_j}\right) + O(n^{-2}).$$

We get the result by using the equation

$$E\left(-\frac{\hat{H}}{3}\sum_j \frac{1}{\hat{p}_j} - \frac{1}{3}\sum_j \frac{\ln \hat{p}_j}{\hat{p}_j} - \frac{1}{12}\sum_j \frac{1}{\hat{p}_j}\right) = -\frac{H}{3}\sum_j \frac{1}{p_j} - \frac{1}{3}\sum_j \frac{\ln p_j}{p_j} - \frac{1}{12}\sum_j \frac{1}{p_j} + O(n^{-1})$$

and substituting all equations in the formula for $E(\hat{Var})$.

$$E(\hat{Var}) = \frac{1}{n}\left[-H^2 + \sum_j p_j \ln^2(p_j)\right] + \frac{1}{n^2}\left[\frac{M}{2} - \frac{1}{2}\right] + \frac{1}{6n^3}\left[(1-H)\sum_j \frac{1}{p_j} - \sum_j \frac{\ln p_j}{p_j} - 1\right] + O(n^{-4})$$

□

Chapter 3

Filtering data regularities

The chapter is organized as follows. We briefly discuss the method for determining predictability of price returns using the Shannon entropy in Section 3.2. We develop the data handling process applied before estimating a degree of market efficiency in Section 3.3. Intraday volatility pattern and volatility clustering are discussed in Sections 3.4 and 3.5. In Section 3.6, we present the method for filtering 0-returns and apply it to simulated and real data. In Section 3.7, we present another source of predictability called periodic patterns. Filtering out microstructure noise is discussed in Section 3.8. All results in this chapter previously appeared in our article [Shternshis et al., 2022a] and its supplementary materials.

3.1 Introduction

The fundamental price of a stock is a quantitative way to assess the intrinsic value of a company. In principle, complete information about a company permits us to know its fair price. When a company is quoted on a stock exchange, the market price of the stock is instead the result of a highly complex process of matching between the supply and demand of traders. In a market with complete information at each time, the matching of supply and demand *should* incorporate all information in the market price. Thus, the *best* forecast is the current observation and the price dynamics is a martingale [Lehmann, 1990]. Very short, this is what is known as the *Efficient Market Hypothesis* (EMH) [Fama, 1970]. When this hypothesis is verified, a market is called *efficient*. However, the definition of an information set which is complete, i.e., including any variable having an impact on price, is usually unfeasible, especially for a quantitative approach. For this reason, it is preferred to work with a weak form of the EMH, that is, the information set is assumed to include only the past observations of the price dynamics.

As stated in [LeRoy, 1989], the hypothesis rejection of a martingale model suggests the existence of trading rules increasing the expected return of some actively managed portfolio with respect to a simple buy-and-hold strategy. In other words, forecasting patterns of price dynamics with a given level of certainty allows devising trading strategies with a positive profit on average. If so, the Efficient Market Hypothesis is not verified and the market is said *inefficient*. Within this context, we neglect the role played by trading costs.

A well-known measure of randomness for symbolic dynamics is the Shannon entropy. It represents the average amount of uncertainty removed with the transmission of each symbol. In the case of financial time series, price dynamics can be opportunely discretized and the Shannon entropy can be computed over the resulting sequence of symbols. This approach was considered, for example, in [Molgedey and Ebeling, 2000] to evaluate predictability of financial time series.

Many measures and methods based on the definition of the Shannon entropy were proposed in recent years with the common goal of studying market efficiency. [Risso, 2009] studied the Shannon entropy as a measure of efficiency for twenty markets, comparing emerging markets with the developed ones. A time-varying entropy of crude oil market efficiency was studied in [Mensi et al., 2012]. [Oh et al., 2015] connected the Shannon entropy with the probability of having market crashes and financial crises. Entropy calculated for energy markets was associated with historical events and climatic factors in [Ruiz et al., 2012]. [Ahn et al., 2019] used entropy to state that the degree of market inefficiency in the Chinese stock market has a strong effect on the economic fundamentals. The Shannon entropy was used to measure uncertainty of cryptocurrency portfolio [Rodriguez-Rodriguez and Miramontes, 2022]. The relationship between entropy and risk measures was investigated in [Pele et al., 2017]. The Shannon entropy as a measure of the risk of a portfolio was also considered in [Dionisio et al., 2006, Ormos and Zibriczky, 2015].

A naive computation of the Shannon entropy for opportunely discretized price dynamics is not, however, the end of the story. There are well-known regularity patterns in financial time series, for instance, daily seasonality or volatility clustering. When not filtered out, such patterns tend to decrease any measure of randomness. Nevertheless, no profitable strategies can be built upon the regularity patterns. Thus, there is a need for devising a computational method for the evaluation of the Shannon entropy that takes into account such regularity patterns. The first study in this direction was presented in [Calcagnile et al., 2020], where volatility clustering, intraday seasonality, and microstructure noise were filtered out before the computation of the Shannon entropy as a measure of efficiency for the Exchange Traded Funds (ETF). Volatility clustering and microstructure noise were also filtered out in [Goldman, 2006]. In this chapter, we propose a computational methodology for entropy estimation, by accounting for many regularity patterns in high-frequency financial time series, in particular including price staleness [Bandi et al., 2020, Kolokolov et al., 2020]. The genuine estimation of the Shannon entropy is used to determine the degree of randomness of the time series of price returns.

More specifically, we start from the method introduced in [Brownlees and Gallo, 2006] for the detection of outliers, then removing splits and merges. After that, we remove step by step both daily seasonality and volatility patterns. Then, we study the effect of price staleness on entropy estimation. In particular, the presence of persistent 0-returns in a row because of the lack of price adjustments or very small trading volumes tends to decrease any estimate of entropy. Persistent 0-returns are converted into a persistent sequence of the same symbol with the effect of larger predictability. Nevertheless, no trading strategy is able to exploit such a persistence pattern. In fact, the presence of price staleness is possible because of very low liquidity or any trading order to be executed should go deep in the limit-order book, thus destroying the persistence pattern of 0-returns. We first show empirically that price staleness tends to decrease the estimate of entropy. Then, we build a method for filtering out 0-returns associated with price staleness. The final step discussed is filtering out microstructure noise.

3.2 Shannon entropy

The unpredictability of price returns in an efficient market implies maximum uncertainty, which can be captured by an entropy measure. The entropy attains its own maximum under the Efficient Market Hypothesis. A measure significantly smaller needs to be intended as a signal of market inefficiency. We consider the Shannon entropy computed over sequences of random variables. The discussion on the Shannon entropy, its definition, and interpretations are given in Section 1.2. The Shannon entropy is defined as the average amount of information that the process transmits with each symbol [Shannon, 1948]. The uncertainty of transmission is proportional to

the expected value of the logarithm of the probability of receiving a sequence of symbols. To estimate entropy of a sequence, we apply the Empirical Frequencies method that we discuss in Section 2.1.

3.2.1 Discretization

The Shannon entropy is defined over a finite alphabet. Prices move on a discrete grid and the minimum price variation is bounded by a tick size. However, the huge amount of possible discrete variations combined with the absence of an upper bound for them makes the computation of entropy infeasible in practical applications. Hence, we build a coarse-grained grid in such a way that the patterns of price variations have a more direct interpretation: "the price goes up", "the price is stationary", or "the price goes down". More specifically, we consider 2-symbols and 3-symbols alphabets. The less the size of the alphabet, the larger the length of the blocks of symbols can be considered. Long blocks allow us to track longer dependencies in price returns time series.

We define price returns as $r_t = \ln\left(\frac{P_t}{P_{t-1}}\right)$, where P_t is the price at time t and $\ln()$ is the natural logarithm. For the binary alphabet, we distinguish positive returns from negative returns.

$$s_t^{(2)} = \begin{cases} 0, r_t < 0, \\ 1, r_t > 0 \end{cases} \tag{3.1}$$

The case $r_t = 0$ is not considered and is removed from the sequence of symbols. The subsamples of the sequence splitted by the presence of 0-returns are then concatenated. This type of discretization is invariant to any multiplicative factor. In particular, volatility of returns in a given period is not important for entropy computation.

The ternary alphabet is obtained by labeling returns according to two tertiles of the empirical distribution of returns, namely

$$s_t^{(3)} = \begin{cases} 1, r_t \leq \theta_1, \\ 0, \theta_1 < r_t \leq \theta_2, \\ 2, \theta_2 < r_t, \end{cases} \tag{3.2}$$

where θ_1 and θ_2 denote the two tertiles of the empirical distribution of the time series $r_t = \{r_1, \ldots, r_n\}$. In other words, θ_1 and θ_2 divide the sorted data r_t into three parts, each containing a third of the total number of the returns. This gives almost the same amount of single symbols $\{0, 1, 2\}$ in the sequence $s_t^{(3)}$. We assume that $\theta_1 < 0$ and $\theta_2 > 0$, thus symbol 0 represents the interval of small price variations. This type of discretization is invariant under the addition of a constant term to each value of returns. Thus, the mean of return dynamics does not affect the entropy computation.

3.2.2 Confidence intervals

Given an estimate of the Shannon entropy, its statistical significance is studied by using Monte Carlo simulations.[4] A random walk after the discretization in 2-symbols or 3-symbols, as described in Section 3.2.1, is a Bernoulli sequence with equal probabilities for the occurrence of each symbol. We define a time series of returns as unpredictable if the entropy estimate is consistent

[4]We use simulations here because they take less computation than iterating over all cases. From combinatorics we know that the number of outcomes to distribute n blocks over M values is equal to $\binom{n+M-1}{M-1}$. This value increases rapidly with the increase of n.

with the entropy of the corresponding Bernoulli process. Any violation is interpreted as a signal of inefficiency for that particular time series of returns.

The bound of significance is computed as follows. We consider lengths of sequences that are multiples of 10. For each considered length, we simulate 10^5 Bernoulli sequences with $p = [\frac{1}{2}, \frac{1}{2}]$ for the binary alphabet and $p = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$ for the ternary alphabet. Then, we compute the Shannon entropy for all of them. For each length, we find the 99% confidence interval (CI) associated with the distribution of entropy estimates. Then, to determine CI for lengths that are not multiples of 10, we use a piecewise linear interpolation. We define a time series as *inefficient* in a given time interval if the estimated entropy of the time series in this interval is less than the bound of 99% one-sided CI of the Bernoulli process with the same length. The length of Bernoulli sequences needed to construct CI is taken according to the number of blocks, n_b. More precisely, we take $l = n_b + k - 1$, where k is the length of blocks. If we detect the presence of inefficiency in a particular interval, as described in this section, we refer to such an interval as an *inefficient interval* or an *interval with inefficiency*. Otherwise, we call time interval *efficient* if it is not inefficient and entropy is at the maximum.

3.3 Financial datasets and data handling

We consider two high frequency datasets of return time series. The first one contains the time series of the prices of the 100 most liquid stocks belonging to the Russell 3000 Index from 02.01.1998 to 23.06.2017. The second dataset contains the time series of the prices of Exchange Traded Funds (ETFs) from 02.01.2003 to 01.12.2009. ETFs are designed to track market indexes. We consider 1-minute closing price data during a regular U.S. trading session, from 9:30 to 16:00. The choice of high liquidity is thus motivated by the need to consider stocks which are traded very frequently, in such a way that the price dynamics are observed at a one minute time scale. If there is no trading at some specific minute, the missing value for the price is reconstructed as the last price available. By using this method, each trading day contains 390 data points. We use a proprietary intraday financial time series dataset provided by `kibot.com`. Tables 3.1 and 3.2 show the list of tickers for ETFs and stocks, respectively.

Table 3.1: List of ETFs

ticker	name ETF	index tracked
SPY	SPDR S&P 500	S&P 500 Index
DIA	DIAMONDS Trust Series 1	Dow Jones Industrial Average Index
IWM	iShares Russell 2000 Index	Russell 2000 Index
EWJ	iShares MSCI Japan Index	MSCI Japan Index
XLE	Energy Select Sector SPDR	Energy Select Sector Index
XLF	Financial Select Sector SPDR	Financial Select Sector Index
XLU	Utilities Select Sector SPDR	Utilities Select Sector Index
IVV	iShares S&P 500 Index	S&P 500 Index
XLB	Materials Select Sector SPDR	Materials Select Sector Index
IWO	iShares Russell 2000 Growth Index	Russell 2000 Growth Index

We divide a time period into intervals in order to test the presence of market inefficiency in each particular time interval. We concentrate on weekly non-overlapping intervals consisting of 5 working days. Weekly time interval consists of $5 \cdot 390 = 1950$ data points. A week is chosen as a relatively short interval. We can assume that the data is stationary over short time

Table 3.2: List of Stock tickers

MSFT	MO	AAPL	LLY	AIG	CAT	ADBE	CL	FDX	EA
CSCO	HD	BAC	AMZN	SPLS	SCHW	GLW	PAYX	KR	NKE
INTC	HPQ	TXN	MCD	XLNX	LOW	CA	DUK	BBBY	MXIM
ORCL	DIS	PG	ABT	COST	IP	RIG	EMR	NEM	MAT
GE	MRK	JNJ	SLB	WFC	MMM	LUV	DOW	NTAP	COF
AMAT	WMT	DD	MDT	JPM	ALL	WMB	INTU	SO	SYMC
IBM	KO	BMY	MSI	WBA	GPS	CTXS	ADI	CAG	LMT
PFE	AMGN	MU	AXP	SBUX	BBY	KMB	CVS	LRCX	CCL
C	BA	T	HAL	XRX	BK	BHI	USB	BSX	TER
TWX	PEP	QCOM	AMD	KLAC	AA	UTX	HON	NE	JCP

Data for stocks are provided by Kibot. For the description of Stocks' symbols we refer to `http://www.kibot.com/Historical_Data/Russell_3000_Historical_Intraday_Data.aspx`.

periods. Moreover, short intervals allow more local tracking of entropy changes in each time interval. Also, the shorter the interval, the less information from the past is required to calculate entropy. Dividing a time series into short intervals also helps to measure the degree of market inefficiency. We calculate it as the percentage of inefficient weeks for the considered set of assets in the market. If the percentage is less or equal to 1 percent, the level of significance for testing the EMH, we interpret it as a perfect randomness of prices in the market.

Before applying our methodology for the entropy estimation, we perform a data handling process. We remove outliers, interpreted as values in the dataset with no economic sense, and splits. Then, sources of the regularities in prices are filtered out, e.g., seasonality and volatility patterns [Cont, 2001, Wood et al., 1985, Bulla and Bulla, 2006], in order to focus on the hidden sources of market inefficiency.

The data handling process is in six steps.

Step 1. Removing outlier values from the dataset;

Step 2. Detecting possible splits where the return is greater than 0.2. We delete such returns from the dataset;

Step 3. Filtering out daily seasonalities;

Step 4. Filtering out heteroskedasticity;

Step 5. Filtering out price staleness;

Step 6. Filtering out microstructure noise.

Steps 1 and 2 represent a data cleaning process. We use the [Brownlees and Gallo, 2006] algorithm of an outlier detection . The algorithm identifies price values which are too distant from the mean value with respect to the standard deviation. The algorithm removes a price P_i if

$$|P_i - \bar{P}_i(k)| \geq cS_i(k) + \gamma,$$

where $\bar{P}_i(k)$ and $S_i(k)$ are respectively the δ-trimmed sample mean and standard deviation of the k price records closest to time i. The δ lowest and the δ highest observations are discarded when the mean and the standard deviation are calculated from the sample. Parameters are taken as $k = 20, \delta = 10\%, c = 5, \gamma = 0.05$.

A stock split from Step 2 is a change in the number of company's shares and in the price of the single share such that a market capitalization does not change. We check the condition $|r| > 0.2$ in the price return series to detect unadjusted splits.

Steps from 3 and 6 consist in filtering out the *data regularities* presented in the data. Time series of price returns are characterized by some regularities related to market patterns, which may apparently suggest the possibility of building up trading strategies to make risk-free profits [Pagan, 1996]. For example, microstructure effects result in a non-zero autocorrelation of returns at a high frequency. However, any trading strategy that tries to exploit such an effect has a non-trivial impact on the price dynamics with the result of zero profit on average [Fama and Blume, 1966, Tsutsui et al., 2005]. Similar considerations can be drawn also for intraday patterns and volatility clustering. The impact of such effects on the estimation of entropy for return time series was already considered in [Calcagnile et al., 2020]. Moreover, data regularities contribute non-stationary components in price returns time series. In this context, stationarity means that mean and variance of a random process do not change over time. Intraday volatility implies time-varying volatility during a trading day. Volatility clustering refers to not constant volatility for different trading days. However, the stationarity and also ergodicity of a process is a critical assumption in determining the entropy of the process [Shannon, 1948]. Getting rid of data regularities, we turn the process into a stationary one and, accordingly, we can use methods for calculating the entropy, e.g., [Marton and Shields, 1994, Wyner and Ziv, 1989]. The whitening procedure starts with removing the intraday volatility pattern getting deseasonalized returns and continues with removing the long memory contribution to returns due to volatility getting standardized returns. Interestingly, there is another source of regularity characterizing time series of returns. It leads to apparent inefficiency which can not be however exploited to build profitable strategies at high frequency trading. This source of regularity is the presence of 0-returns in data, which lowers the estimate of entropy. It can be interpreted as a spurious effect and must be removed before consideration of market efficiency. We discuss all mentioned data regularities in detail in subsequent sections.

3.4 Intraday volatility pattern

The volatility of intraday returns has periodic behavior. It is higher near the opening and the closing of the market, showing a U-shaped profile every day. For empirical evidence of the U-shaped intraday pattern of stock returns in NYSE, see [Wood et al., 1985]. Intraday volatility in the Japanese stock market is discussed in [Andersen et al., 2000]. We filter out the intraday volatility pattern from price return series by using the following model with intraday volatility factors. If $\bar{R}_{d,t}$ is the raw return of day d and intraday time t, we define deseasonalized returns as

$$\tilde{R}_{d,t} = \frac{\bar{R}_{d,t}}{\xi_t}, \tag{3.3}$$

where

$$\xi_t = \frac{1}{N_{days}} \sum_d \frac{|\bar{R}_{d,t}|}{s_d},$$

N_{days} is the number of days in the sample and s_d is the standard deviation of the returns of day d. The procedure also normalizes the values of overnight returns that tend to have larger magnitudes than the other 389 returns. Figure 3.1 shows ξ_t, for the ETF SPY where t passes from 9:31 to 15:59. All picks appear every half hour (from the largest at 10:00 to 15:30).

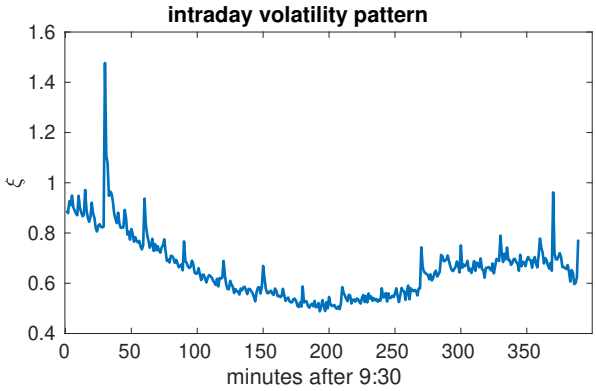


Figure 3.1: Intraday volatility pattern for the ETF SPY

3.5 Volatility clustering

Volatility clustering refers to the fact that large returns tend to be followed by other large returns of either sign, and vice versa for small returns. The deseasonalized returns $\tilde{R}$ defined by Eq. 3.3 are still heteroskedastic since different days can have different levels of volatility. The volatility clustering needs to be filtered out by re-scaling each observation by the estimated value of the volatility at that time. The same way of filtering out heteroskedasticity is used in [Hsieh, 1991], where volatility is estimated by Exponential GARCH model [Bollerslev, 1986]. In order to remove this heteroskedasticity, we estimate the volatility $\hat{\sigma}_t$ and define the standardized returns by

$$r_t = \frac{\tilde{R}_t}{\hat{\sigma}_t}.$$

3.5.1 Volatility estimation

For a reason that will be clear from the next section about price staleness, we choose an algorithm for volatility estimation in the case of missing observations [Sucarrat and Grønneberg, 2020]. It is based on the Expectation-Maximization algorithm (EM) [Dempster et al., 1977], but 0-returns are set as missing values and updated after each step of the numerical maximization of a likelihood function.

The volatility is assumed to follow a GARCH(1,1) model

$$\sigma_t^2 = \mu_0 + \alpha\sigma_{t-1}^2 + \beta\tilde{R}_{t-1}^2$$

and the estimation of parameters $\theta = \{\mu_0, \alpha, \beta\}$ is obtained by using the following 4-steps algorithm.

1. Choose initial values of θ and calculate $\sigma_t^2(\theta)$;
2. Estimate the missing values as $E[\tilde{R}_t^2] = \sigma_t^2$;
3. Using the maximum likelihood estimation, find new values of θ and, hence, new estimation of volatility;
4. Continue steps 2 and 3 until stopping criteria are satisfied.

We make several changes in the method. First, we consider not only 0-returns as missing values but also returns after each 0-return. Second, we calculate the likelihood function using all available data including reconstructed returns. The comparison of performances of different approaches can be found in Section 3.5.2. The section also presents the optimization of parameters. An alternative approach for volatility estimation where the only parameter can be optimized is discussed in Section 6.3.3.

3.5.2 Numerical results for volatility estimation

The goal of this section is to modify method proposed in [Sucarrat and Grønneberg, 2020] in order to achieve a better accuracy in the volatility estimation. We take the following model of prices $\tilde{P}_t$, $t = 1 \dots N$.

$$
\begin{aligned}
P_t &= P_0 + \int_0^t \sigma_s P_s dW_s^1 \\
q_t &= q_0 + \int_0^t \nu dW_s^2 \\
\tilde{P}_t &= P_t(1 - B_t) + \tilde{P}_{t-1} B_t \\
&\begin{cases} B_t & = 1 \text{ with probability } q_t \\ B_t & = 0 \text{ with probability } 1 - q_t \end{cases}
\end{aligned}
$$

where W^1 and W^2 are two independent Brownian motions with $N = 5 \times 10^5$, $P_0 = 50$, $\sigma_t = 10^{-3}$. The probability of spurious 0-returns [Bandi et al., 2020, Kolokolov et al., 2020, Bandi et al., 2017, Zhu and Liu, 2023] is given by q_t with $q_0 = 0.2$ and $\nu = 10^{-3}$.

We estimate volatility $\hat{\sigma}_t$ using 8 different models. For each model, we calculate the error $\sum_t((\hat{\sigma}_t - \sigma_t)^2)$. Methods 1-4 use 0-returns as missing values, methods 5-8 use also values after 0-returns as missing. Odd methods {1 3 5 7} calculate the likelihood function only for not missing values. Even methods {2 4 6 8} calculate the likelihood function using also estimation of missing values. Also, we define different initial values. The methods {1 2 5 6} apply an approach used in MATLAB. The model is transformed to an equivalent ARMA model [Makridakis and Hibon, 1997]. The initial ARMA values are solved using the Yule-Walker equations [Walker, 1931]. The methods {3 4 7 8} use $\{mean(\tilde{R}^2), \epsilon, \epsilon\}$, where ϵ is a predetermined small value. The results are in Table 3.3 below.

Table 3.3: Errors for 8 methods for calculating volatility.

Method	1	2	3	4	5	6	7	8
Error	0.104	**0.034**	0.041	0.052	0.109	**0.031**	0.040	0.047

All errors are averaged over 100 simulations. Minimum values of errors are highlighted in bold.

The smallest squared error is obtained when we define values after 0-returns as missing values, calculate the likelihood function for all values, and use initial values from the MATLAB estimation. This method is chosen as a modification of the approach proposed in [Sucarrat and Grønneberg, 2020] and is used further for volatility estimation until Chapter 6. In Chapter 6, we propose the modification of exponential moving average for volatility estimation.

We have 4 parameters to optimize in the selected method. We use quasi-Newton algorithm [Dennis and Moré, 1977] for maximization of the likelihood function. Gradient estimation is done

by the forward difference formula; the Hessian is estimated using the BFGS algorithm [Shanno, 1970]. Constraints on the parameters are implemented using the interior-point algorithm with the logarithmic barrier function [Potra and Wright, 2000]. The algorithm of volatility estimation is stopped if the gradient of the likelihood function is small enough or step size is less than the predefined value ϵ. δ is used as a minimum step size and the norm of gradient in the stopping criteria. Parameter c is used in the first Wolfe condition [Wolfe, 1969]. Finally, λ is the multiplier of the logarithmic penalty function. We set $\epsilon = 10^{-11}$ as a minimum value that gives two distinct values of gradients for two close values of arguments during all our tests. δ is also set to be equal to 10^{-11} since we take it not smaller than ϵ. Default values of c and λ are 10^{-4} and 10^{-6}, respectively. We fix c and conduct the experiment for different values of λ in Table 3.4.

Table 3.4: Errors for different values of λ.

$\log_{10}\lambda$	-1	-2	-3	-4	-5	-6
Error	0.0358	0.035	0.0325	**0.0323**	0.0331	0.0339

All errors are averaged over 100 simulations. The minimum error is highlighted in bold.

Then, we set $\lambda = 10^{-4}$ and find c that gives a minimum error in Table 3.5.

Table 3.5: Errors for different values of c.

$\log_{10} c$	-1	-2	-3	-4	-5	-6
Error	**0.03197**	0.03238	0.03227	0.03227	0.03227	0.03227

All errors are averaged over 100 simulations. The minimum error is highlighted in bold.

Errors for the different values of c are close to each other. We keep the value $\lambda = 10^{-4}$ and choose $c = 10^{-1}$ for our analysis.

3.6 Zeros as a source of predictability

0-returns in financial time series arise because of many effects including rounding, no trading, and price staleness. The 0-returns occurring because of no trading implies a spurious autocorrelation of time series [Lo and MacKinlay, 1987]. Moreover, except for the non-trading, there is also the effect of *price staleness* in the data shown in [Bandi et al., 2020]. The authors of the article define price staleness as a lack of price adjustments yielding 0-returns. The effect of staleness is one of the features that distinguish real data from prices following a random walk. To explain the phenomenon of price staleness in the data, we refer to the work [Bandi et al., 2017]:

"Classical models of price formation postulate that informed traders react to new information not yet reflected in the transaction price of a security and transact if the trade guarantees a profit net of execution costs (e.g., [Glosten and Milgrom, 1985] and [Kyle, 1985]). Thus, due to lack of trading, a security with higher transaction costs should experience less frequent price updates and a larger number of 'small' returns than a security with a lower cost of transacting. Similarly, uninformed traders may not just buy and sell randomly. They may also react to the size of transaction costs and choose not to trade should these costs be considered too large."

The presence of *spurious zeros*, zeros that appear due to no trading or no price adjustments, affects any estimate of the Shannon entropy. The value of entropy as a measure of randomness is affected by the 0-returns in the data since the large amount of 0-returns makes a time series

predictable. When 0-returns are persistent in time because of no trading or no price adjustments, the price dynamics look predictable because the price is constant in time. However, such an effect can not be seen as market inefficiency since no profitable strategy can be implemented in this case.

In the next sections, we show empirically that the 0-returns are one of the sources of data regularities. We construct a method for filtering out 0-returns due to price staleness in Section 3.6.3. We test the method for filtering out spurious 0-returns first on simulated data and then on the real dataset. The results for the simulated data are in Section 3.6.4. We show that spurious 0-returns generated non-uniformly change the measure of entropy of the return time series. However, the entropy as well as the amount of 0-returns due to rounding goes back to its genuine value by implementing the method. We show that entropy of price returns of ETFs increases after filtering out 0-returns in Section 3.6.5.

3.6.1 Influence of 0-returns on the entropy value

We first investigate the impact of 0-returns on the estimation of entropy for the 2-symbols discretization proposed in Eq. 3.1. The presence of clustering and intraday patterns in volatility does not influence the 2-symbols discretization. In this case, the data whitening process described in Sections 3.4 and 3.5 has no impact on the estimation of entropy, and any signal of the price inefficiency is not linked to intraday volatility or heteroscedasticity.

We focus the analysis on the set of 100 most liquid stocks belonging to the Russell 3000 index. For each stock, we consider the time series of returns for each week of the period from 02.01.1998 to 23.06.2017. For each stock, we average the fraction of 0-returns separately for *inefficient* weeks and for *efficient* weeks. We show the averaged fraction of 0-returns for weeks with inefficiency and for weeks without inefficiency in Fig. 3.2. The result points out an evident correlation between the fraction of 0-returns and a low entropy, a signal of the inefficiency of market dynamics. This supports empirically that the inefficient weeks are characterized by the presence of a large number of 0-returns.[5]

A larger fraction of 0-returns implies a smaller sample size for the time series of symbols. However, this is not the reason for a biased estimation. First, the estimator proposed by Grassberger (Eq. 2.4) is applied to correct for the bias associated with finite sample sizes. Second, we show in the next Section 3.6.2 that if one artificially shortens the sample length by aggregating returns or bootstrapping, the entropy estimate is not affected and has no downward bias.

3.6.2 Dependence between entropy estimation and length of sequence

In this section, we aim to test dependence between the entropy estimation and the length of sequence. For the study we take the price of the Microsoft Corporation stock (MSFT). Then, we discretize returns using the 2-symbols discretization. There are 980 weeks in the considered time period from 02.01.1998 to 23.06.2017. The minimum length of a 2-symbols sequence is obtained in week 352 and is equal to 896. First, we calculate the entropy value for every week. Then, we use sampling of the binary sequences to decrease its length for each week. For every week, we consider 1000 sub-samples of binary symbols choosing only 896 symbols for the sign of returns keeping their ascending order in time. Then, we calculate the average value of entropy of all samples and plot it for each week with the estimated value of entropy of the week in Fig. 3.3.

[5]Despite the fact that 0-returns are removed from the discretized sequence, they may affect the 2-symbols sequence by changing the ratio between positive and negative returns, the sample size of the resulting blocks of symbols after concatenation, and the values after 0-returns.

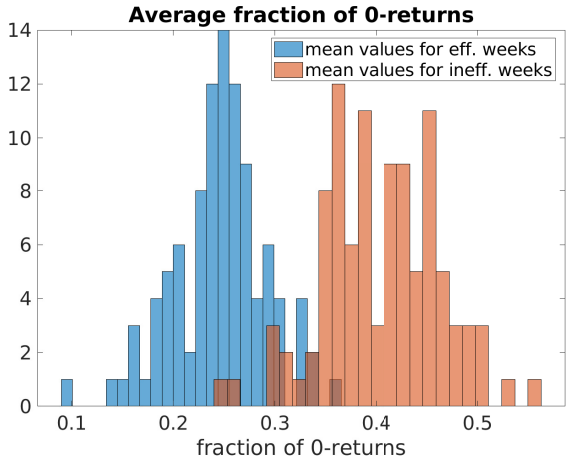


Figure 3.2: The fraction of 0-returns for weeks with and without inefficiency presented in two histograms with 25 bins

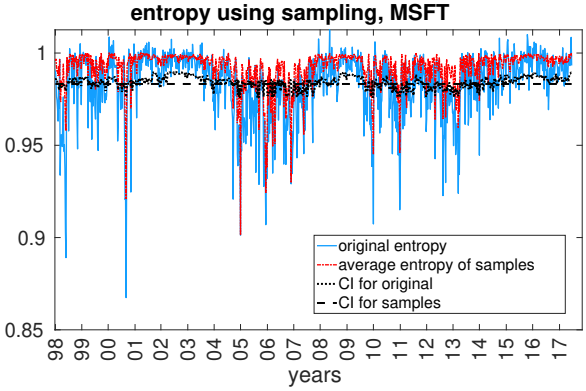


Figure 3.3: The estimated entropy and the averaged entropy of samples for the MSFT with the corresponding 99% Confidence Intervals

The entropy value calculated using sampling does not decrease. On the contrary, the entropy of shortened sequences has a larger value. A probable explanation is the destruction of existing dependencies in the returns since we choose samplings with some gaps between symbols.

The other way to show that entropy does not decrease with the amount of non-zero returns decreasing is by aggregating data to a less frequency. The amount of information does not increase with aggregation. Thus, the entropy value does not decrease with a smaller length. To show it empirically, we aggregate data to 5 minutes, so we reduce the lengths of sequences by about five times. We present the results in Fig. 3.4.

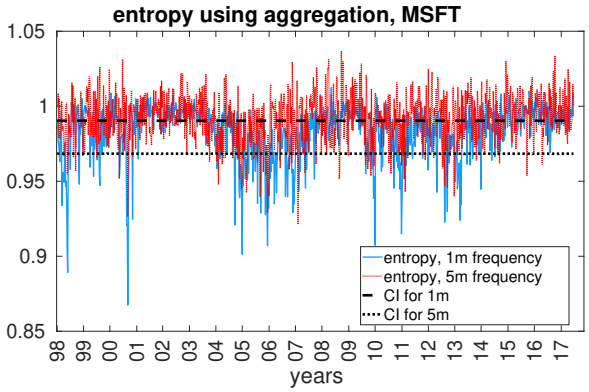


Figure 3.4: The entropies calculated for 1 minute and 5 minutes frequencies for the MSFT with the corresponding 99% Confidence Intervals

When we aggregate data to 5 minutes, the maximum possible length of sequence is 390. In spite of the small length of sequence, the entropy is close to 1. We make the conclusion that a low entropy value relates to a large fraction of 0-returns but not to the length of the sequence.

3.6.3 Filtering out spurious 0-returns

Since we find correlation between the fraction of 0-returns and entropy value, we develop an algorithm to filter out spurious 0-returns. There are two effects resulting in 0-return for the price dynamics as pointed out in [Bandi et al., 2020]: The first is the result of a price discretization due to the tick size, and the second is an economic phenomenon linked to traded volumes. The second channel consists of a high transaction cost and the absorption of bounded volumes causing no price changes and leading to price staleness. In the first case, since the price does not change due to rounding, the return is equal to zero at the minimum resolution available. On the other hand, the 0-returns due to the price staleness effect hide information about the underlying asset. Such 0-returns should be considered as *spurious*. Disentangling these two effects is thus crucial to the end of entropy estimation.

Following [Bandi et al., 2020, Kolokolov et al., 2020], we model price staleness by assuming that observed transaction prices are the result of the coupling of a random walk with a Bernoulli process for the occurrence of the spurious 0-returns.

$$\tilde{P}_t = \bar{P}_t(1 - B_t) + \tilde{P}_{t-1}B_t, \tag{3.4}$$

where $t = 1, \ldots, N$, B_t are Bernoulli variables with values 0 or 1, $\bar{P}_t$ is an efficient price rounded following the Geometric Brownian Motion (GBM), and $\tilde{P}_t$ is an observed price. According to this model, the efficient price is diffusive, even when we do not observe it because of price staleness.

A random price affected by price staleness can be considered as subdiffusive Brownian motion [Magdziarz et al., 2011].

By assuming the process (Eq. 3.4) for the price dynamics of an asset, we aim to infer from real data the probability for each 0-return to be generated because of rounding or price staleness, then filtering out 0-returns which are likely associated with the second effect. To this end, the probability of getting a 0-return because of rounding within some given tick size is obtained in Equation 3.5 under the following approximations. The price follows the Geometric Brownian Motion, so that the returns are normally distributed. The bid-ask spread is set to be 0.

$$p_t = erf(R_t) + \frac{1}{R_t\sqrt{\pi}}(\exp(-R_t^2) - 1) \tag{3.5}$$

where $R_t = \frac{d}{\bar{P}_t\hat{\sigma}_t\sqrt{2\Delta}}$, $erf(x)$ is the Gauss error function, d is the tick size, Δ is a time step, $\bar{P}$ is the rounded price, and $\hat{\sigma}_t$ is an estimation of volatility at time t.[6] We give the proof of this formula and its generalizations in Chapter 4. The result is obtained by considering the probability that the price moves slightly so that it is rounded to the same value as one time step ago. The obtained probability is approximated using the observed price and the volatility estimation.

Given p_t at each time step t, the expected number of 0-returns due to rounding is the sum of all p_t within the considered time period, i.e., $N_{save} = \sum_{t=1}^{N} p_t = \bar{p}N$, where $\bar{p}$ is an average probability. The variance of the amount of 0-returns is equal to $V = \bar{p}(1-\bar{p})N$. If the observed number of 0-returns, N_{real}, is not significantly larger than the expected one according to the model (Eq. 3.5), i.e., $N_{real} \leq N_{save} + 1.96\sqrt{V}$, we do not filter out any 0-return. Otherwise, in order to filter out 0-returns due to price staleness, we replace by missing values the 0-returns which appear not due to the rounding according to the approach below.

A 0-return is considered as spurious according to the following method called *probability-based*. An expected time when a 0-return appears due to rounding is determined when the expected number of 0-returns due to rounding, $Z(t) = \sum_{l=1}^{t} p_l$, jumps to a new integer value, $\lfloor Z(t) \rfloor - \lfloor Z(t-1) \rfloor = 1$. Then, moving from $t = 0$ to the final time, the expected time when a 0-return appears due to rounding is matched with the closest time with a real 0-return in the time series[7]. We save these 0-returns, but set other 0-returns at the amount $N_s = N_{real} - N_{save}$ as missing values. We assume that N_s 0-returns appear due to price staleness. We set not only 0-returns due to price staleness but also the values at the consecutive time step as missing for the reasons we discuss below in Remark 1. The main feature of this approach is that *we use only information on prices for its implementation*: we calculate the probability of getting 0-returns due to rounding using prices, the estimation of volatility, and the tick size. We consider this *probability-based* approach as basic and use it for the analysis. Modifications of this approach including the usage of the information about traded volumes and the estimation of a bid-ask spread are in Section 5.2.1.

3.6.4 Filtering 0-returns on simulated data

To include 0-returns into the analysis, we consider here the three-symbols alphabet, where one of the symbols corresponds to returns between the two tertiles of the empirical distribution (Eq. 3.2). We aim to study here the effect of 0-returns on the estimation of entropy for simulated time series.

We model prices as the Geometric Brownian motion, $P_t = P_0 + \int_0^t P_s\sigma dW_s$, setting an initial price equal to 50 and a constant volatility equal to 10^{-3}. We take a time step equal to 1

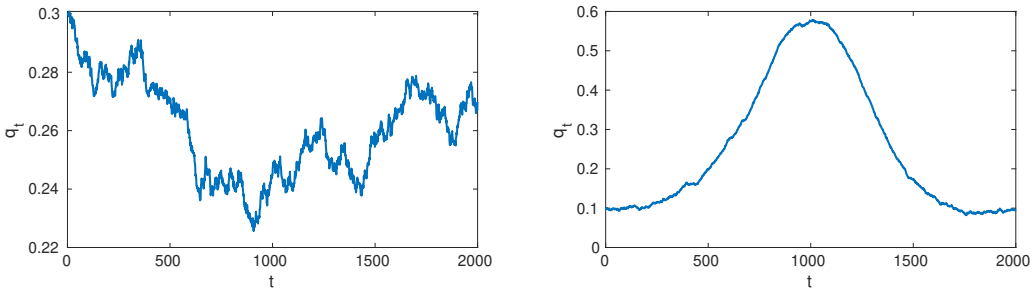
[6] Returns used for the volatility estimation are deseasonalized in a preliminary step.

[7] If we need to save a 0-return from the sequence of 0-returns, we place it at the beginning of this sequence.

minute and simulate 2000 data points. Then, we generate spurious 0-returns with a probability $q_t = q_0 + \int_0^t \mu_s ds + \int_0^t \nu dW_s$ with constant $\nu = 10^{-3}$ and three different values of μ_t and q_0. The first two values present cases where the spurious 0-returns are distributed uniformly but with different extents. The third case simulates a scenario where 0-returns cluster together. The first choice of the probability function is $q_0 = 0.3$, $\mu_t = 0$. The second choice of the probability function is $q_0 = 0.2$, $\mu_t = 0$. The third choice of the probability function is

$$q_0 = 0.1$$
$$\mu_t = -\frac{1}{400^2}(t - 1000)\exp\left(-(t - 1000)^2/400^2\right)$$

The examples of the first and third functions are in Fig. 3.5a and Fig. 3.5b, respectively.



(a) uniform probability

(b) bell-shaped probability

Figure 3.5: Examples of the probabilities of getting spurious 0-returns

We aim to test the method we have developed in the previous section for identifying spurious 0-returns. After detecting the spurious 0-returns and setting them as missing values, the entropy value should increase so that the time series is indistinguishable from a realization of a random walk with some missing values. The filtering method of 0-returns consists of several steps. First, we find those 0-returns that are related to rounding using Equation 3.5. The remaining 0-returns are then considered as 0-returns due to price staleness and are replaced with missing values. Section 2.3 discusses how entropy is calculated when some values of a sequence are missing. We keep missing values in the sequence, but consider the partitions of the time series in blocks that do not contain missing values of symbols. Other common approaches to deal with missing values are concatenating observed sequences and using an interpolation to replace a missing observation with its reconstructed value. The former may create new patterns containing parts of concatenated blocks and the latter incorporates predictable patterns instead of the missing values. The description of the chosen method for the entropy estimation and the proof of the consistency of the entropy estimator are in Section 2.3.

We calculate the entropies only for the last 1950 data points. First, we calculate the entropy of initial return time series, then with the additional 0-returns, and then after setting the spurious 0-returns as missing values. We simulate 1000 time series with no more than 1/3 of 0-returns. The results are shown in Tables 3.6- 3.8 below. The columns of the tables represent the mean entropy for all samples, its standard deviation, the number of samples that are not defined as inefficient, the number of 0-returns averaged, and its standard deviation.

The large amount of spurious 0-returns added uniformly does not sufficiently decrease the entropy value. We take two probability functions with different mean values 0.2 and 0.3. In both cases, on average, sequences with additional 0-returns have the entropy value close to the

Table 3.6: Results of filtering 0-returns with the 1st choice of probability function q_t

time series	mean entropy	std. of entropy	N. of efficient series	N. of 0-returns	std. for 0-returns
GBM	1.0002	0.0025	993	153.38	13.09
After adding 0-returns	1.0002	0.0026	994	619.77	23.77
After setting missing values	0.9978	0.0083	1000	147.96	4.99

Table 3.7: Results of filtering 0-returns with the 2nd choice of probability function q_t

time series	mean entropy	std. of entropy	N. of efficient series	N. of 0-returns	std. for 0-returns
GBM	1.0001	0.0027	987	155.49	12.31
After adding 0-returns	1.0002	0.0026	994	508.49	51.77
After setting missing values	0.9992	0.0061	1000	147.65	4.69

value of the initial time series. Indeed, as proved in [Phillips and Yu, 2007], if the probability of appearing of a spurious 0-return is constant, then price returns keep a martingale property.

Price return series with missing values have entropy lower than the initial sequences. Since the lower bound of confidence interval (CI) also decreases, all sequences after filtering out 0-returns appear to have no inefficiencies. (We compare each estimate with the lower bound of 99% CI for the entropy of Bernoulli sequences, which is about 0.9935 for the length of 1950). Moreover, when we make the distribution sharper and bell-shaped, as in the third case, the estimate of entropy decreases significantly after adding spurious 0-returns. However, when we use the probability-based method described in Section 3.6.3, the entropy becomes closer to its initial value. The other important aspect is that the probability-based method keeps the amount of 0-returns in the sequence quite close to the number of 0-returns appearing by rounding the efficient price in all three cases.

We expect similar results for real data. If spurious 0-returns are distributed uniformly, setting the spurious 0-returns as missing values does not affect the value of entropy. However, if the 0-returns cluster together, for example, in the presence of high transaction costs, then a predictable pattern due to spurious 0-returns needs to be removed from the time series for a genuine estimation of entropy.

Remark 1. *There are two reasons to consider returns after 0-returns as missing values too.*

I. If, according to Eq. 3.4, an observed price is hidden for one or more minutes, the first non-zero return is the sum of all returns that were hidden. Thus, we do not have the value for the return at this minute, but only know the aggregate information.

II. Denoting as missing values returns after 0-returns increases the accuracy of the estimation of volatility. See Fig. 3.6 as an example. The value of volatility is on the x-axis in the range from 5×10^{-4} to 2×10^{-3}. Two mean values of the estimations of volatility in the case of the second choice of the probability function are on the y-axis. The closer the scatter plot to the diagonal, the more precise the estimation.

Table 3.8: Results of filtering 0-returns with the 3rd choice of probability function q_t

time series	mean entropy	std. of entropy	N. of efficient series	N. of 0-returns	std. for 0-returns
GBM	1.0002	0.0027	995	153.78	13.04
After adding 0-returns	0.9905	0.0051	287	613.02	27.64
After setting missing values	0.9984	0.0058	1000	149.27	5.05

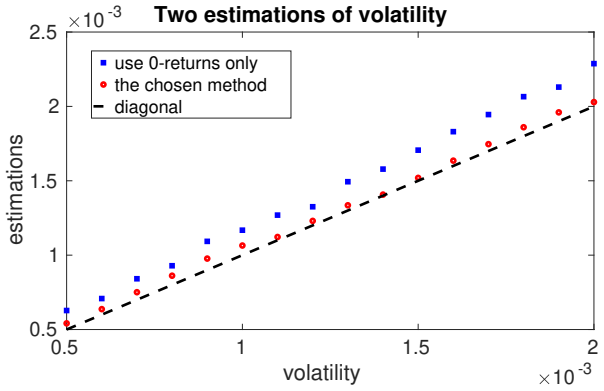


Figure 3.6: The volatility and two estimations. The blue squares are the estimations of volatility using only 0-returns as missing values. The red circles are the estimations of volatility using 0-returns and the values after 0-returns as missing values.

3.6.5 Filtering 0-returns on real data

To test the probability-based approach on real data, we take the SPY ETF, which aims to track the Standard & Poor's 500 Index, and SPDR Dow Jones Industrial Average ETF Trust (DIA). The tick size is $d = 0.01$, and the time step is $\Delta = 1$ minute. We discretize returns after filtering out daily seasonality and heteroskedasticity. We use the probability-based approach described in Section 3.6.3. We apply the approach only to weeks with less than 1/3 of 0-returns and less than 10 consecutive minutes with 0-volume. Fig. 3.7a and Fig. 3.7b show the entropy of these returns including all 0-returns and the entropy calculated after filtering out 0-returns.

For the ETF SPY, we observe 330 weeks. 36 of them are associated with the estimate of entropy lower than the 99% confidence bound after the step of filtering out heteroskedasticity. After applying the method for filtering staleness there are 10 inefficient weeks, but only 4 of them are from the group of previously inefficient weeks. For the ETF DIA, we observe 333 weeks with 42 inefficient weeks after the step of filtering heteroskedasticity. After applying the method for filtering 0-returns there are 12, but only 8 of them are from previously inefficient weeks.

The main conclusions we can make from this section are the following.

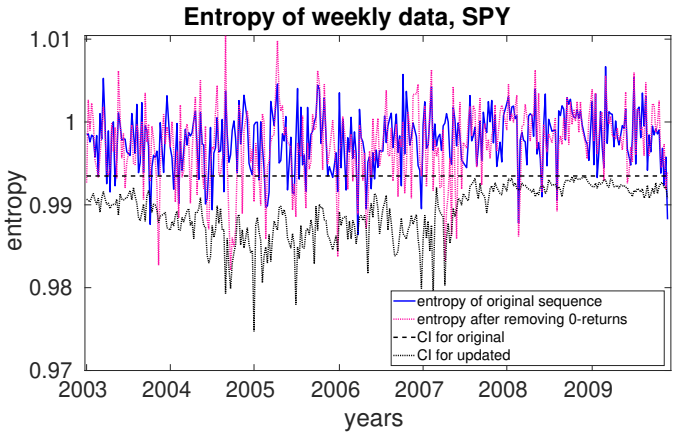
1. On average, the method for filtering out 0-returns increases the entropy value.

2. There are weeks where a low entropy value still can not be explained by intraday volatility pattern, heteroscedasticity, and price staleness.

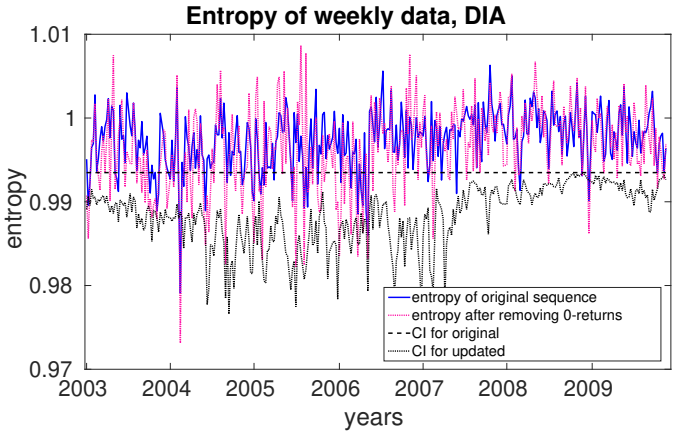
3. After filtering out 0-returns, our method can determine inefficiency for a week which we considered as "efficient" before filtering out 0-returns. Two possible explanations are random fluctuations of the entropy measure as a random variable and detecting low values of entropy that were not detected before due to a high level of confidence. We discuss this issue in more detail in the next Section 3.6.6.

3.6.6 Inefficiency after filtering 0-returns

In this section, we analyze such particular weeks that are detected as inefficient only after filtering out spurious 0-returns. We define a *probability level* as the fraction of entropy values calculated for 10^5 Bernoulli sequences that are greater than the entropy of the time series.



(a) SPY

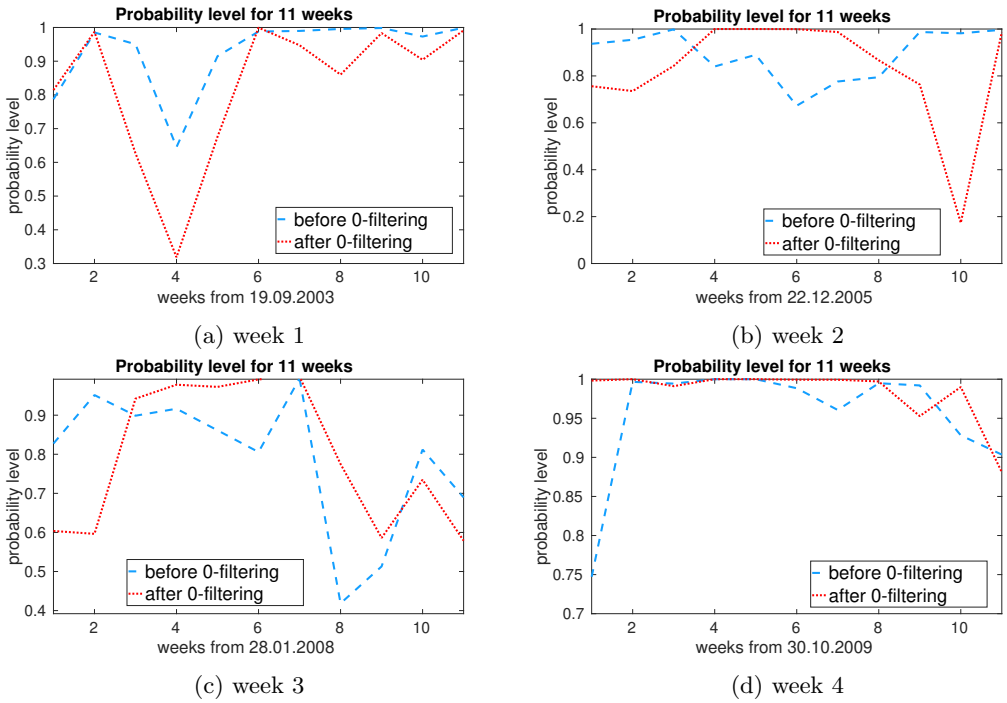


(b) DIA

Figure 3.7: Entropies calculated for the 3-symbols discretization before and after filtering out 0-returns for the ETFs SPY and DIA with the corresponding 99% Confidence Intervals

When we filtered out 0-returns, we determined 12 weeks with inefficiency for the ETF DIA, but 4 of them were not classified in such a way before filtering staleness. In order to investigate the dynamics of entropy, we take a new week with inefficiency, one week before, and one week after. We move a weekly time window day by day, so that the week with the new inefficiency is the 6th interval. For each interval, we find the probability level associated with the entropy value. Fig. 3.8 presents the results for all four new weeks with inefficiency. As before, we define an interval with inefficiency if the probability level is larger than 0.99.

We notice that for the first and the fourth cases in Fig. 3.8 the value of the probability level before filtering staleness is greater than 0.98. That is, the probability level does not change its value significantly after filtering 0-returns, but crosses the confidence level that is set to be 0.99. For the third case, the 7th interval is classified as inefficient before and after filtering 0-returns. It has 80% data in common with the 6th interval, which is classified as inefficient only after filtering staleness. This makes results more coherent for two adjacent intervals. However, for the second case, an interval with detected inefficiency before filtering 0-returns is the third, which



(a) week 1

(b) week 2

(c) week 3

(d) week 4

Figure 3.8: Probability level associated with entropy for 4 cases with the new inefficiencies of the ETF DIA.

has only 40% data in common with the 6th interval. We conclude that one of the reasons for the appearance of new inefficient weeks is a slight increase in the probability level that is lower than the level of confidence equal to 0.99. On the contrary, as in the example of the second case, identifying inefficient weeks may be the case of the extreme realization of the test statistic: the estimation of entropy is in the 1% tail of the entropy distribution associated with a fully random Bernoulli process.

We investigate another possible reason for appearing new inefficiencies and the source of remaining inefficiencies in the next section.

3.7 Periodic patterns

The goal of this section is to investigate a reason for low entropy values by considering the values of empirical frequencies. When estimating entropy, the frequencies of all possible k-blocks are calculated. We are interested in finding some repeating patterns in blocks that appear to be the most frequent, which may cause a decrease in the entropy estimation. For each inefficient week of the ETF DIA after filtering out 0-returns, we write down the most frequent block(s) of symbols of length k which are met while calculating the entropy. The complete table with results is Table 3.9. The table's values represent blocks in 3-symbols. The most frequent blocks in weeks with the new inefficiencies are **000121**, 011210, **020121**, **021002**, 201211, 202111, **210120**. We highlight in bold blocks where the same characters 1 or 2 do not appear in a row. This may mean that the price fluctuates around its average value. We assume that such an effect can be observed, for example, with a bid-ask bounce. That is, such patterns occur when there is no movement in

the efficient price, but transactions occur at both the bid and ask prices. Then, we show when the existence of this pattern is statistically significant.

Table 3.9: The most frequent blocks in the ETF DIA.

12112	**12121**	21121	**000121**	011210	**012001**	**020121**	**021002**	110200	110201
112211	**120012**	**120121**	**121001**	201211	202111	**210120**	**210210**	**212100**	221122

For each inefficient week in the ETF DIA, the most frequent block is recorded after filtering out 0-returns. The blocks highlighted in bold describe a pattern of changing signs with each new non-zero return. Each block is the most frequent only in one inefficient week. The block length is 5 for the first three blocks and 6 for the other blocks.

If we consider only positive and negative returns, we move back for a moment to the 2-symbols discretization. Let choose $k = \lfloor \log_2(n) \rfloor$, where n is the length of a 2-symbols sequence. Let's consider 2 sequences '1010...' and '0101...'. We know that the expected amount of blocks with these sequences for the process with the entropy $h = 1$ is $n_b/2^{(k-1)}$, where $n_b = n - k + 1$. Also, we construct a 99% CI using the formula for the standard deviation from the binomial distribution.

$$\mathbb{p} = \frac{1}{2^{(k-1)}}$$
$$\mathbb{q} = 1 - \mathbb{p}$$
$$\hat{\sigma}^2_{pp} = n_b \mathbb{p}\mathbb{q}$$

Definition 6. *If the actual amount of blocks '1010...' and '0101...' in a sequence, n_{pp}, is greater than the upper bound of CI, that is,*

$$n_{pp} > \mathbb{p} n_b + q_\alpha \hat{\sigma}_{pp},$$

where q_α is a quantile of the normal distribution, we determine the sequence as a sequence with periodic patterns (PP).

These periodic patterns may appear due to a low price value with respect to the tick size and a low volatility (which are also reasons for 0-returns generated by rounding). If a price fluctuates randomly around the mean value, crossing the same tick size twice in the opposite directions is more likely than crossing two tick sizes in the same direction.

We have markers for inefficiency and periodic patterns for each week. Using the hypergeometric distribution[8], we may conclude if there is dependence between detected inefficiency and PP. Thus, we test the following hypotheses.

$\mathcal{H}_0$: the appearance of a week with inefficiency and the appearance of a week with periodic patterns are *independent*.

$\mathcal{H}_a$: the appearance of a week with the inefficiency and the appearance of a week with periodic patterns are *dependent*.

We say that $\mathcal{H}_0$ is rejected if

$$n_{ineff+pp} > m_h + q_\alpha \hat{\sigma}_h \tag{3.6}$$

[8]The hypergeometric distribution describes the probability of k successes in n draws without replacement from the finite population of size N that contains $K \leq N$ success outcomes.

where $n_{ineff+pp}$ is the number of weeks with inefficiency and periodic patterns simultaneously; $m_h = \frac{nK}{N}$; $\hat{\sigma}_h^2 = \frac{nK(N-n)(N-K)}{N^2(N-1)}$; N is the total number of weeks; K is the number of weeks with inefficiency; n is the number of weeks with PP.

The results for SPY are that we have **no** 95% confidence that the random processes "occurrence of a week with inefficiency" and "occurrence of a week with periodic patterns" are dependent. On the contrary, for DIA we do have 95% confidence that there is the dependence for the occurrence of weeks with inefficiency and weeks with PP.

3.8 Microstructure noise

An observed price includes various microstructure effects caused by transaction costs and price rounding. The difference between the efficient price and the observed price with the microstructure effects is called *microstructure noise*. In general, each new observation in a return time series depends on the previous values. A model that can explain the presence of a positive autocorrelation of data is the following. Assume that an observed price $\tilde{P}_t$ differs from the efficient price P_t by some error term u_t, namely

$$\ln \tilde{P}_t = \ln P_t + u_t,$$

and the observed market return $\tilde{r}_t$ is

$$\tilde{r}_t = r_t + u_t - u_{t-1},$$

where r_t is the return of the efficient price. The observed return is affected by the error term associated with the log-price at the previous time step. The microstructure noise tends to be positively autocorrelated in time [Jacod et al., 2017]. If the error term u_t follows an autoregressive $AR(1)$ process, then the observed returns are described by an $ARMA(1,1)$ process. For more detailed analysis of the structure of microstructure noise with respect to different types of traders we refer to [Diebold and Strasser, 2013]. The reciprocal of standard deviation of microstucture noise term is interpreted as a measure of market efficiency in [Boehmer and Kelley, 2009] since microstructure noise alienates observed prices from efficient prices following a martingale model. Instead, we estimate a degree of market efficiency only after filtering out the effect of microstructure noise.

The effect of such a noise term on the estimation of entropy was described in [Calcagnile et al., 2020] by considering both $AR(1)$ and $MA(1)$ models. In particular, the authors found that larger (in absolute value) autoregressive coefficients are associated with lower values of the Shannon entropy. This intuition is exploited by [Ito and Sugiyama, 2009], who used a time-varying autocorrelation of stock returns as a measure of market inefficiency for the U. S. stock market.

Here, we consider a further step of filtering based on the estimation of an ARMA model [Makridakis and Hibon, 1997] on the time series of returns after filtering out 0-returns. We select parameters (P,Q) of an ARMA(P,Q) model describing the data by using the BIC criterion [Schwarz, 1978]. We study the residuals in order to remove any autocorrelation pattern from data. We consider only ARMA(P,Q) models with $P + Q \leq 5$. After filtering out 0-returns and replacing them by missing values, we use the methodology introduced in [Jones, 1980] to deal with the estimation of an ARMA(P,Q) model with missing observations. In particular,we use the Kalman filter[9]. Other approaches for determining the order of an ARMA model including the maximization of entropy and an out-of-sample testing are presented in Section 5.3.1.

[9]We apply the algorithm as in Section 3.5.2 to find the parameters of ARMA model using the likelihood function of residuals. We set $\epsilon = \delta = 10^{-5}$, $c = 10^{-1}$. We set $\lambda = 0$ checking the constraints on the parameters.

Summing up, the Shannon entropy as a measure of randomness is used in the range of articles [Risso, 2008, 2009, Mensi et al., 2012, Oh et al., 2015]. In contrast to the mentioned works, we filtered different sources of data regularities of market dynamics before calculating the degree of market efficiency. The method for filtering out data regularities was first introduced by [Calcagnile et al., 2020]. In this chapter, we introduce price staleness [Bandi et al., 2020, Kolokolov et al., 2020] as a source of apparent inefficiency besides the data regularities caused by daily seasonalities, heteroscedasticity, and microstructure noise. Price staleness creates spurious 0-returns in the data. We construct the method for detecting spurious 0-returns according to the probability of their occurrence. We set spurious 0-returns as missing values. Thus, we build and apply the modification of the Empirical Frequencies method [Marton and Shields, 1994] for calculating entropy in the case of the presence of missing values in the data. We show that 0-returns cause a false detection of inefficiency.

The last step of data whitening is filtering out microstructure noise. We show that for some ETFs there is a clear dependence between a low entropy estimation for discretized returns and the presence of periodic patterns, i.e., the switching sign of non-zero returns for each trading minute. Both periodic patterns and microstructure noise may have their origin in a bid-ask bounce. We use a fitting ARMA model to get rid of these effects.

Chapter 4

Random price movements on a discrete grid

The previous chapter poses the problem of finding the number of 0-returns that result from rounding of a completely random price. In this chapter, we find the probability that, after rounding, the price remains at the same value. Also, we generalize this formula in several different ways. First, we find the probabilities that a price that follows the Geometric Brownian Motion moves by k ticks, where k is different from 0. Then, we include a non-zero bid-ask spread, which extends the formula of the probability of obtaining a 0-return of the rounded price. Finally, we consider the case where price changes are not normally distributed and have fatter tails. The formula for the rounding probability presented in the next section allows us to separate the 0-returns resulting from rounding with 0-returns that have the economic meaning of price staleness as discussed in Section 3.6.3. Some of 0-returns are replaced with missing values to reduce the effect of price staleness on price's degree of predictability.

We have published the results of the next two sections in article [Shternshis et al., 2022a].

4.1 Rounding a price with Gaussian increments

In this section, we calculate an approximate value of the amount of 0-returns generated by rounding an efficient price. We consider a model for an efficient price following the Geometric Brownian Motion.

$$P_t = P_0 + \int_0^t \sigma_s P_s dW_s$$

Assuming that the price is rounded up to tick size d, we can find the probability that the efficient price will not change using the rounded (observed) price $\bar{P}$, the estimation of volatility $\hat{\sigma}$, and the sampling frequency Δ. The probability that $\bar{P}_{t+1} = \bar{P}_t$ given $\bar{P}_t$ is

$$\begin{aligned} p_t = \mathbf{P}[\bar{P}_t - \frac{d}{2} < P_{t+1} < \bar{P}_t + \frac{d}{2}] \\ \bar{P} = P + x \\ x \in U_{-\frac{d}{2},\frac{d}{2}} \end{aligned}$$

where $U_{-\frac{d}{2},\frac{d}{2}}$ stands for the uniform distribution from $-\frac{d}{2}$ to $\frac{d}{2}$. That is, as shown in [Bandi et al., 2020], p_t is equal to

$$\frac{1}{d}\int_{-\frac{d}{2}}^{\frac{d}{2}}\int_{x-\frac{d}{2}}^{x+\frac{d}{2}} f_{N(0,P_t\sigma_t\sqrt{\Delta})}(z)dzdx,$$

where f_N is the normal density function. Indeed, averaging over all values of x, we get that

$$p_t = \frac{1}{d}\int_{-\frac{d}{2}}^{\frac{d}{2}} I_t(x)dx$$

where

$$\begin{aligned} I_t(x) = \mathbf{P}[\bar{P}_t - \frac{d}{2} < P_{t+1} < \bar{P}_t + \frac{d}{2}] &= \mathbf{P}[P_t + x - \frac{d}{2} < P_{t+1} < P_t + x + \frac{d}{2}] \\ &= \mathbf{P}[x - \frac{d}{2} < P_{t+1} - P_t < x + \frac{d}{2}] \\ &= \int_{x-\frac{d}{2}}^{x+\frac{d}{2}} f_{N(0,P_t\sigma_t\sqrt{\Delta})}(z)dz \end{aligned}$$

Then, we estimate $P_t\sigma_t$ by $\bar{P}_t\hat{\sigma}_t$, where the returns that are used for the estimation of volatility are deseasonalized as discussed in Section 3.4.

$$I_t(x) \approx \frac{1}{2}[erf(\frac{x+\frac{d}{2}}{\sqrt{2}s_t}) - erf(\frac{x-\frac{d}{2}}{\sqrt{2}s_t})] \tag{4.1}$$

where $s_t = \bar{P}_t\hat{\sigma}_t\sqrt{\Delta}$ and $erf(x) = \frac{2}{\sqrt{\pi}}\int_0^y \exp(-t^2)dt$ is the Gauss error function. Using integration by parts $\int erf(y)dy = y \cdot erf(y) + \frac{1}{\sqrt{\pi}}\exp(-y^2)$, we obtain the result

$$\begin{aligned} \frac{1}{d}\int_{-\frac{d}{2}}^{\frac{d}{2}} erf(\frac{x+\frac{d}{2}}{\sqrt{2}s})dx &= \frac{\sqrt{2}s}{d}\left(\frac{d}{\sqrt{2}s}erf(\frac{d}{\sqrt{2}s}) + \frac{\exp(-\frac{d^2}{2s^2}) - 1}{\sqrt{\pi}}\right) \\ \frac{1}{d}\int_{-\frac{d}{2}}^{\frac{d}{2}} erf(\frac{x-\frac{d}{2}}{\sqrt{2}s})dx &= \frac{\sqrt{2}s}{d}\left(\frac{d}{\sqrt{2}s}erf(-\frac{d}{\sqrt{2}s}) + \frac{-\exp(-\frac{d^2}{2s^2}) + 1}{\sqrt{\pi}}\right) \end{aligned} \tag{4.2}$$

$$p_t(R_t, 0) = \frac{1}{d}\int_{-\frac{d}{2}}^{\frac{d}{2}} I_t(x)dx = erf(R_t) + \frac{1}{\sqrt{\pi}R_t}(\exp(-R_t^2) - 1) \tag{4.3}$$

where $R_t = \frac{d}{s_t\sqrt{2}} = \frac{d}{\bar{P}_t\hat{\sigma}_t\sqrt{2\Delta}}$ and the second argument 0 stands for the amount of ticks that the price moves. Then, the result is extended for the approximation of probability that the price moves by k ticks, $p_t(R_t, k)$.

Using the similar procedure, we obtain equation for probabilities $p_t(R_t, k)$ where k can be different from 0.

$$\begin{aligned} p_t(R_t, k) = \frac{k+1}{2}erf\left((k+1)R_t\right) + \frac{k-1}{2}erf\left((k-1)R_t\right) - k \cdot erf(kR_t) \\ + \frac{1}{2\sqrt{\pi}R_t}\exp(-(k+1)^2R_t^2) + \frac{1}{2\sqrt{\pi}R_t}\exp(-(k-1)^2R_t^2) - \frac{1}{\sqrt{\pi}R_t}\exp(-k^2R_t^2) \end{aligned} \tag{4.4}$$

The sum of probabilities $p_t(R_t, k)$ over all values of k and any fixed value of $R_t > 0$ is equal to 1. In the next sections, we obtain the modification of Equation 4.3 by including a bid-ask spread and considering t-distribution for price returns.

4.2 Including bid-ask spread

In this section, we include a bid-ask spread, c, to modify the formula (4.3). We assume that c is constant and that the probabilities of the bid price and the ask price to be the close price of a minute are equal and independent. Thus, we aim to find a probability

$$p_t^c = \mathbf{P}[\bar{P}_t - \frac{d}{2} < P_{t+1} + c_1 < \bar{P}_t + \frac{d}{2}]$$
$$\bar{P} = P + x + c_2$$
$$c_2 - c_1 = \begin{cases} 0 \text{ with probability } \frac{1}{2} \\ c \text{ with probability } \frac{1}{4} \\ -c \text{ with probability } \frac{1}{4} \end{cases}$$

With probability $\frac{1}{2}$, there are two bid prices or two ask prices in a row, thus $p_t^c = erf(R_t) + \frac{1}{\sqrt{\pi}R_t}(\exp(-R_t^2) - 1)$.

With probability $\frac{1}{4}$, the ask price follows the bid price and $I_t(x)$ from Eq. 4.1 is replaced by $I_t(x + c)$. Using integration by parts for $erf(y)$ as in Equations 4.2, we get that

$$p_t^c = \frac{c+d}{d} erf\left(\frac{c+d}{\sqrt{2}s_t}\right) - \frac{c}{d} erf\left(\frac{c}{\sqrt{2}s_t}\right) + \frac{\sqrt{2}s_t}{d\sqrt{\pi}}\left[\exp\left(-\frac{(c+d)^2}{2s_t^2}\right) - \exp\left(-\frac{c^2}{2s_t^2}\right)\right] \quad (4.5)$$

Similarly, if the bid price follows the ask price,

$$p_t^c = \frac{c-d}{d} erf\left(\frac{c-d}{\sqrt{2}s_t}\right) - \frac{c}{d} erf\left(\frac{c}{\sqrt{2}s_t}\right) + \frac{\sqrt{2}s_t}{d\sqrt{\pi}}\left[\exp\left(-\frac{(c-d)^2}{2s_t^2}\right) - \exp\left(-\frac{c^2}{2s_t^2}\right)\right] \quad (4.6)$$

We notice that the sum of the last two probabilities from Equations 4.5 and 4.6 is the probability that the price moves by $k = \frac{c}{d}$ tick sizes up or down according to Equation 4.4. Therefore, the required probability that the price does not change its value due to rounding if bid-ask spread is equal to c is

$$p_t^c = \frac{1}{2} erf(R_t) + \frac{k+1}{4} erf((k+1)R_t) + \frac{k-1}{4} erf((k-1)R_t) - \frac{k}{2} erf(kR_t) +$$
$$+ \frac{1}{2\sqrt{\pi}R_t}\left(\exp(-R_t^2) - 1 - \exp\left(-k^2R_t^2\right)\right) + \frac{1}{4\sqrt{\pi}R_t}\left[\exp\left(-(k+1)^2R_t^2\right) + \exp\left(-(k-1)^2R_t^2\right)\right] \quad (4.7)$$

where $k = \frac{c}{d}$.

4.3 Rounding a price with increments having fat tails

Another modification of Equation 4.3 is considering price increments that have t-distribution instead of normally distributed increments. That is, the distribution function has the following form.

$$F_S(z, \nu, s_t) = \frac{1}{2} + C(\nu)\frac{z}{\sqrt{\nu}s_t}\, {}_2F_1\left(\frac{1}{2}, \frac{\nu+1}{2}, \frac{3}{2}, -\frac{z^2}{\nu s_t^2}\right) \quad (4.8)$$

where $C(\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi}\Gamma(\frac{\nu}{2})}$ and ${}_2F_1$ is the hypergeometric function given below [Temme, 2003]. Here, s_t is a parameter standing for deviation of price returns.

$$ {}_2F_1(a,b,c,z) = 1 + \sum_{n=1}^{\infty} \frac{(a)_n (b)_n}{(c)_n} \frac{z^n}{n!} \tag{4.9} $$

where $(x)_n = \prod_{k=1}^{n}(x+k-1)$, the left index of ${}_2F_1$ stands for the number of parameters in the numerator a, b and the right index is the amount of parameters in the denominator, c.

In particular, we get normal distribution when $\nu \to \infty$ [Chidichimo and Thorsley, 2001]:

$$ \lim_{\nu\to\infty} {}_2F_1(a,\nu,c,z/\nu) = {}_1F_1(a,c,z) $$

and $\frac{2z}{\sqrt{\pi}} {}_1F_1(\frac{1}{2},\frac{3}{2},-z^2)$ is the Taylor expansion of the Gauss error function $erf(z)$.

A property that can be obtained from the definition of the hypergeometric function (Eq. 4.9) is the following.

$$ \frac{d}{dz} {}_2F_1(a,b,c,z) = \frac{ab}{c} {}_2F_1(a+1,b+1,c+1,z) $$

That is,

$$ \int {}_2F_1(a,b,c,z)dz = \frac{c-1}{(a-1)(b-1)} {}_2F_1(a-1,b-1,c-1,z) \tag{4.10} $$

Based on the method used in Section 4.1, we calculate the probability of rounding

$$ \begin{aligned} p_t(\nu) &= \frac{1}{d}\int_{-\frac{d}{2}}^{\frac{d}{2}} \int_{x-\frac{d}{2}}^{x+\frac{d}{2}} f_S(z,\nu,s_t)dzdx \\ &= \frac{1}{d}\int_{-\frac{d}{2}}^{\frac{d}{2}} \left[F_S\left(x+\frac{d}{2},\nu,s_t\right) - F_S\left(x-\frac{d}{2},\nu,s_t\right)\right] dx \\ &= \frac{1}{d}(I_t^+ - I_t^-) \\ I_t^+ &= \int_{-\frac{d}{2}}^{\frac{d}{2}} F_S(x+\frac{d}{2},\nu,s_t)dx \\ I_t^- &= \int_{-\frac{d}{2}}^{\frac{d}{2}} F_S(x-\frac{d}{2},\nu,s_t)dx \end{aligned} $$

where $f_S(z,\nu,s_t) = \frac{d}{dz}F_S(z,\nu,s_t)$ is the density function of t-distribution. From Eq. 4.8,

$$ \begin{aligned} I_t^+ &= \frac{d}{2} + C(\nu)\int_{-\frac{d}{2}}^{\frac{d}{2}} \frac{x+\frac{d}{2}}{\sqrt{\nu}s_t} {}_2F_1\left(\frac{1}{2},\frac{\nu+1}{2},\frac{3}{2},-\frac{(x+\frac{d}{2})^2}{\nu s_t^2}\right) dx \\ &= \frac{d}{2} + \frac{C(\nu)\sqrt{\nu}s_t}{2}\int_0^{\frac{d^2}{s_t^2\nu}} {}_2F_1\left(\frac{1}{2},\frac{\nu+1}{2},\frac{3}{2},-y\right) dy \end{aligned} $$

where $y = \frac{\left(x+\frac{d}{2}\right)^2}{\nu s_t^2}$. Using Eq. 4.10, we get that

$$I_t^+ = \frac{d}{2} + \frac{C(\nu)\sqrt{\nu}s_t}{\nu-1}\left[{}_2F_1\left(-\frac{1}{2}, \frac{\nu-1}{2}, \frac{1}{2}, -y\right)\right]_0^{\frac{d^2}{s_t^2\nu}}$$
$$= \frac{d}{2} + \frac{\sqrt{\nu}C(\nu)s_t}{\nu-1}\left({}_2F_1\left(-\frac{1}{2}, \frac{\nu-1}{2}, \frac{1}{2}, -\frac{d^2}{s_t^2\nu}\right) - 1\right)$$

Using substitution $y = \frac{\left(x-\frac{d}{2}\right)^2}{\nu s_t^2}$, we can show that

$$I_t^- = \frac{d}{2} + \frac{\sqrt{\nu}C(\nu)s_t}{\nu-1}\left(-{}_2F_1\left(-\frac{1}{2}, \frac{\nu-1}{2}, \frac{1}{2}, -\frac{d^2}{s_t^2\nu}\right) + 1\right)$$

That is,

$$p_t(\nu) = \frac{1}{d}(I_t^+ - I_t^-) = 2\frac{s_t}{d}\frac{\sqrt{\nu}C(\nu)}{\nu-1}\left({}_2F_1\left(-\frac{1}{2}, \frac{\nu-1}{2}, \frac{1}{2}, -\frac{d^2}{s_t^2\nu}\right) - 1\right)$$

When $\nu \to \infty$, $\frac{2\sqrt{\nu}C(\nu)}{\nu-1} \to \sqrt{\frac{2}{\pi}}$ and we obtain another expression for Equation 4.3 in terms of a confluent hypergeometric function, ${}_1F_1$.

$$p_t(\nu) \to \frac{1}{\sqrt{\pi}}\frac{\sqrt{2}s_t}{d}{}_1F_1\left(-\frac{1}{2}, \frac{1}{2}, -\frac{d^2}{2s_t^2}\right) \text{ when } \nu \to \infty$$

The result for the case $\nu = 1$ can be obtained in a similar way since the probability function is known and it represents the Cauchy distribution. When $\nu = 2$, $b = c$. In this case,

$${}_2F_1(a, b, b, z) = (1-z)^{-a}$$

Therefore, when $\nu = 2$

$$p_t(\nu = 2) = \frac{\sqrt{2}s_t}{d}\left(\sqrt{1 + \frac{d^2}{2s_t^2}} - 1\right)$$

Chapter 5

Measuring market efficiency

In this chapter, we estimate a degree of efficiency of the ETF market using the Shannon entropy. We consider two sources of predictability of financial time series. A part of the predictability of price returns series comes from stylized facts of financial markets. We filter out such data regularities as described in Section 3.3. The last two steps in the filtering process is removing spurious 0-returns due to price staleness and linear terms due to microstructure noise. We explore the degree of predictability left after each of these steps. We consider different methods for filtering out price staleness and microstructure noise and carry out their comparative analysis. The degree of predictability that remains after filtering out all the data regularities is the degree of market inefficiency. A degree of inefficiency smaller than predetermined significance level suggests the confirmation of the Efficient Market Hypothesis.

We have published the results of this chapter in article [Shternshis et al., 2022a] and its supplementary materials.

5.1 Introduction

Different approaches were proposed to test *market inefficiency*, all of them are with the common rationale of measuring how much an empirical price dynamics is far from the assumption of complete randomness (martingale hypothesis). A time-varying autocorrelation of stock returns was proposed as a measure of the degree of market inefficiency for the U.S. stock market [Ito and Sugiyama, 2009]. The R/S statistics and the Hurst exponent were used to rank the efficiency of emerging equity markets [Cajueiro and Tabak, 2004, 2005]. The Hurst exponent was measured on Bitcoin data to compare it with mature markets [Drożdż et al., 2018]. Generalized Hurst exponent approach was applied to measure efficiency in Middle East and North Africa markets [Sensoy, 2013]. An exponentially weighted generalized Hurst exponent was applied to daily stock prices in [Morales et al., 2012]. It was, however, shown that Hurst exponent is not necessarily describe correlations in price returns [McCauley et al., 2007]. That is, different values of the Hurst exponent can be consistent with the Efficient Market Hypothesis [Bassler et al., 2006]. Detrended fluctuation analysis was used to investigate correlations in developed and emerging markets in [Beben and Orłowski, 2001]. Multifractal detrended fluctuation analysis was applied to investigate the efficiency of stock and credit markets [Shahzad et al., 2017]. This parameter shows significant differences between emerging and developed markets [Zunino et al., 2008, Rizvi et al., 2014].

The algorithmic complexity of return time series was applied as a measure of the relative efficiency of financial markets [Giglio et al., 2008]. In [Shmilovici et al., 2003], the algorithmic

complexity was used to check the Efficient Market Hypothesis. The BDS test [Broock et al., 1996] was used to explore chaotic behavior of stock prices [Hsieh, 1991]. Finally, the approximate entropy [Pincus et al., 1991] was proposed as a measure of market efficiency over time for different markets [Alvarez-Ramirez et al., 2012, Pincus and Kalman, 2004, Duan and Stanley, 2010, Oh et al., 2007]. The general idea of these methods is to compare the characteristic of the time series with the value corresponding to a completely random process.

We measure market inefficiency using the Shannon entropy on a weekly basis as the percentage of inefficient weeks among all assets taken into consideration. The main empirical result obtained for the real dataset in Section 5.2 is that the amount of inefficient weeks decreases significantly after filtering out 0-returns due to price staleness. In the last step of the methodology, we study the effect of microstructure noise. After filtering out all mentioned sources of the data regularities, it is possible to conclude that the ETF market is *not* efficient at a high frequency (1-minute) on weekly time interval. However, the signal of market inefficiency for weekly time intervals is weak. Since the procedure of detecting an inefficient week for a specific asset is made with a high level of confidence, the occurrence in the data of 10 ETFs cases of the co-inefficiency is investigated. We investigate whether appearance of inefficiencies is related to prices cointegration. Finally, we consider the evolution of entropy in time and at different time scales. We consider monthly and quarterly time intervals to analyze market inefficiency using rolling windows but with the greater length.

5.2 Price predictability after filtering out price staleness

The goal of this section is to assess the impact of the process of filtering 0-returns on the ETF market and to test the dependence between the remaining weeks with inefficiency and the presence of periodic patterns discussed in Section 3.7. We take the set of 10 ETFs and test them for detecting weeks with inefficiency before and after filtering staleness. The characteristics we are interested in are the total amount of weeks; the amount of weeks with inefficiency before filtering staleness; the amount of weeks with inefficiency after filtering staleness; the amount of new inefficient weeks that appear only after the process of filtering 0-returns; those new inefficient weeks which contain periodic patterns according to Definition 6. Finally, we test the null hypothesis, $\mathcal{H}_0$, about the independence of the events of occurring an inefficient week and a week with periodic patterns according to Equation 3.6. We construct Table 5.1 with the results for all 10 ETFs.

We conclude that for all 10 ETFs the number of inefficient weeks decreases after applying the filtering of 0-returns. The number of detected weeks with inefficiency remains the same only in the case of the ETF EWJ. The algorithm filters out data regularities from all 6 weeks considered for the ETF IVV. That is, we can not detect any statistically significant decrease in entropy for the ETF IVV by using non-overlapping weekly time windows.

Finally, we note that the percentage of the total number of inefficient weeks is 3.46. However, we detect a predictable time series with 99% of confidence. Since 3.46% is significantly greater than 1%, we can conclude that there are other unaccounted sources of inefficiency in the price dynamics. For this reason, we discuss further the role of microstructure noise, another stylized fact of financial markets which is key for the entropy estimation of return time series at a high frequency.

Next sections contain the comparison of results for the different approaches for filtering 0-returns.

Table 5.1: Results of filtering out 0-returns with the probability-based approach.

ETF	N. of weeks	Ineff. before filtering 0-returns	Ineff. after filtering 0-returns	New ineff.	New ineff. with periodic patterns	Reject $\mathcal{H}_0$ with 95%	Reject $\mathcal{H}_0$ with 99%
SPY	330	36	10	6	0	0	0
DIA	333	42	12	4	1	1	0
IWM	294	34	5	3	0	0	0
EWJ	14	3	3	1	1	1	0
XLE	231	12	7	3	1	0	0
XLF	120	23	16	4	1	1	1
XLU	94	10	6	4	0	0	0
IVV	161	6	0	0	0	0	0
XLB	124	4	2	2	0	0	0
IWO	148	13	3	3	0	0	0

For the last two columns, 1 indicates the rejection of $\mathcal{H}_0$ from Section 3.7; 0 indicates a failure to reject $\mathcal{H}_0$.

5.2.1 Approaches for identifying spurious 0-returns

Here, we consider four approaches for filtering out 0-returns.

- We present the *probability-based* in Section 3.6.3. We use the formula for probability that a price following a Geometric Brownian Motion does not change its value due to rounding (Eq. 4.3). Then, we label some 0-returns appearing due-to rounding proportionally to the probability. Then, we set all other 0-returns and the returns after them as missing values.

- In a *volume-based* approach, a 0-return is set as a missing value when associated with zero or small volume of trading. In particular, the amount of 0-returns we keep is the sum of probabilities of getting 0-returns due to rounding, $N_{save} = \sum_{t=1}^{N} p_t$. The remaining 0-returns at the amount of $N_c = N_{real} - N_{save}$ are set as missing values. First, we set N_{0v} 0-returns corresponding to 0-volumes as missing values. Then, if $N_c > N_{0v}$, $N_c - N_{0v}$ 0-returns are set as missing values according to the smallest volumes.

  This approach uses more information than the previous one. It includes not only prices but also trading volumes. The approach fits the logic of the appearance of staleness in the data: traders react to high transaction costs. They trade less and the price updates less frequently.

- As a generalized version of two approaches above, we consider a *mixed* approach. First, we define all returns with 0-volume as missing values. Then, we implement the probability-based approach, but consider only 0-returns that do not relate to 0-volumes. We update the values of probabilities p_t of getting a 0-return by setting it equal to 0, if at time t or one time step ago, $t-1$, there is no trading. We need two consecutive price observations to define a 0-return due to rounding. More precisely, let

$$T_t = \begin{cases} 1, \text{ if } V_t > 0 \text{ and } V_{t-1} > 0 \\ 0, \text{otherwise} \end{cases}$$

be the index of trading activity at both times $t-1$ and t, where V_t is the volume traded at time t. Then, the probability of rounding $\bar{p}_t$ is modified as

$$\bar{p}_t = \begin{cases} p_t \text{ if } T_t = 1 \\ 0, \text{ otherwise} \end{cases}$$

This also affects the values of $\bar{N} = \sum_{t=1}^{N} T_t$, N_{real}, and N_{save}. Then, we set 0-returns as missing values by the probability-based approach. Finally, they are merged with 0-returns denoting cases of no-trading.

This approach combines the rationale of the two previous methods. First, we remove non-trading minutes from consideration. Then, we use the probability of getting a 0-return which is obtained from the past prices. As always, we set as missing not only values due to staleness or no-trading but also the values at the consecutive time step.

- Now, we include also the information on the bid-ask spread. In order to estimate the bid-ask spread, first, we apply the method suggested by [Roll, 1984] for each separate day and get an effective spread. We set an effective spread equal to 0 for days with a positive autocorrelation of price returns. Then, we multiply each effective spread by a midpoint, which is the average of the high and low prices recorded each day. Finally, we average results for each month to obtain the estimation of bid-ask spread, c. In order to take into consideration the bid-ask spread, we use formula (4.7) of the probability of getting a 0-return from Section 4.2. Then, we repeat the methodology of the *probability-based* approach using the updated function and call the obtained approach *b/a-adjusted*.

In the next chapter, 0-returns corresponding to minutes without trading are not taken into account, which is most compatible with the mixed approach.

5.2.2 Comparison of approaches for filtering 0-returns on the ETF market

We apply the four approaches for filtering 0-returns to the group of 10 ETFs. Table 5.2 with the comparison of all four methods is below. We are interested in the percentage of inefficient weeks after filtering out 0-returns; the percentage of weeks that become inefficient after applying the approaches; the number of ETFs that have no inefficient weeks (called efficient ETFs in the Table); and the number of ETFs with dependence between inefficiency and periodic patterns for 0.95 and 0.99 levels of confidence.

The probability-based approach shows the lowest fraction of inefficient weeks and new inefficient weeks. The mixed approach shows a smaller percentage of all inefficient weeks and new inefficient weeks than the volume-based approach.

The b/a-adjusted method leaves slightly more inefficient weeks than the probability-based approach. However, the amount of 0-returns to save in the time series does not differ significantly. See Fig. 5.1 for the values of $\frac{N_{save}-N^c_{save}}{N_{save}}$ for 10 ETFs, where N_{save} and N^c_{save} are the expected amount of 0-returns due to rounding based on the probability-based and b/a-adjusted approaches, respectively. We notice that the volatility estimation, needed to calculate the degree of price staleness in the probability-based approach, already includes bias due to a bid-ask spread. We assume that the b/a-adjusted method can be used if an estimator of volatility is unbiased to a bid-ask spread.

Table 5.2: Comparison of four methods for filtering out 0-returns.

Method	% of ineff. weeks	% of new ineff. weeks	N. of efficient ETFs	N. of ETFs with rejected $\mathcal{H}_0$ at 95%	N. of ETFs with rejected $\mathcal{H}_0$ at 99%
probability-based	**3.46**	**1.62**	1	3	1
volume-based	4.76	2.49	0	1	1
mixed	4.27	2.33	0	2	2
b/a-adjusted	3.84	1.78	0	1	1

$\mathcal{H}_0$ states that the appearance of a week with inefficiency and the appearance of a week with periodic patterns are independent. The smallest values in the first two columns are in bold.

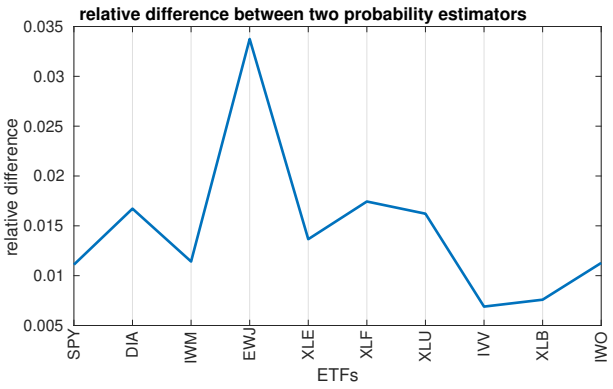


Figure 5.1: The difference between numbers of 0-returns needed to be saved according to the probability-based method and its b/a-adjusted version divided by the value obtained for the former method.

5.3 Price predictability after filtering out microstructure noise

In this section, we investigate the degree of predictability after filtering out microstructure noise. For the analysis of results, we introduce *co-inefficiency* for a group of assets.

Definition 7. *Two or more assets have co-inefficiency if they have the same time interval with inefficiency. The co-inefficiency of n_{co} assets in a time interval is a statistically significant event if*

$$n_{co} > \mathbb{p}\mathbb{n} + q_\alpha \sqrt{(1-\mathbb{p})\mathbb{p}\mathbb{n}},$$

where $\mathbb{n}$ is the number of assets considered in the time interval, $\mathbb{p}$ is the probability of detecting inefficiency in randomly chosen time interval and asset, and q_α is a quantile of the normal distribution.

Independence of appearing inefficient time intervals for several assets is tested by analogy of defining periodic patterns (Definition 6). We reject the hypothesis of the independence of appearing inefficiencies for a given time period with the confidence of $\alpha = 0.95$. Results for 10 ETFs are in Table 5.3. Entropy of weekly time intervals after filtering out microstructure noise is presented in Figure 5.2.

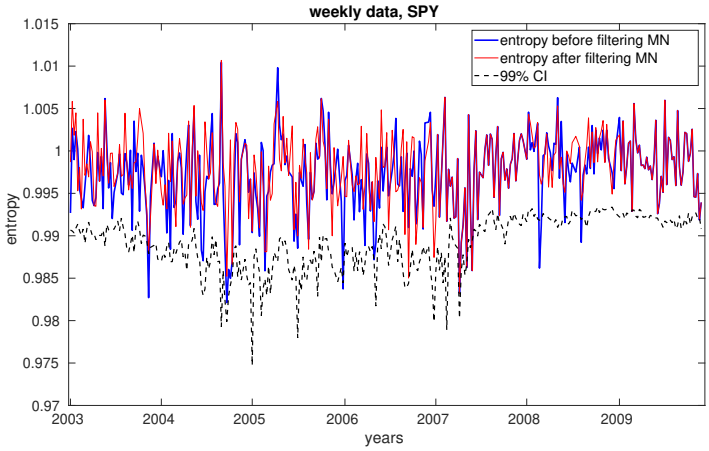


Figure 5.2: Entropy of the ETF SPY before and after filtering out microstructure noise (MN) with corresponding 99% Confidence Interval

Filtering microstructure noise removes the dependence between weeks with periodic patterns and weeks with inefficiency. Moreover, filtering microstructure noise gives a low percentage of new inefficient weeks equal to 0.27%, which makes the results consistent with the previous step of filtering. The number of co-inefficiencies is equal to 2. Both of them are statistically significant and are highlighted in bold in Table 5.3. Two next sections are devoted to the comparison of results for different approaches for filtering microstructure noise.

Filtering microstructure noise gives a low percentage of inefficient weeks equal to 1.35%. A degree of inefficiency calculated as the fraction of inefficient week for each particular asset is given in Figure 5.3.

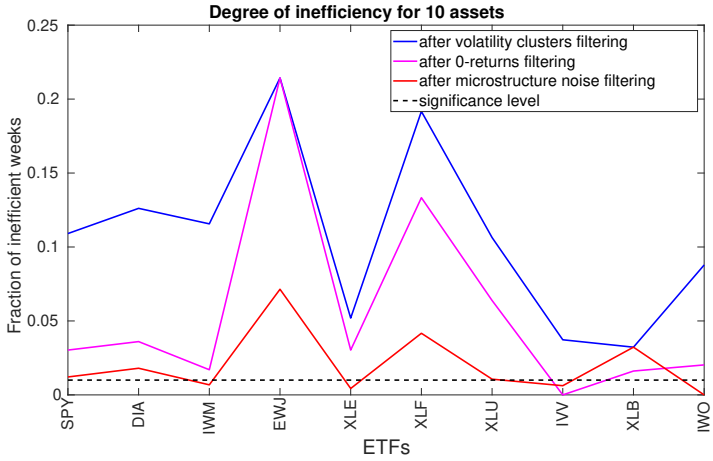


Figure 5.3: A degree of inefficiency for the set of 10 ETFs

This is a clear signal that the Efficient Market Hypothesis is not totally realistic for real-world market dynamics. Indeed, after filtering all known sources of data regularities, the percentage of weeks detected as inefficient is greater than 1%, namely the significance level used in our testing procedure. Thus, the market is not totally efficient at a one minute time scale. Regarding the last step of filtering, namely, considering residuals of an ARMA model, we should also notice that we filter out all possible linear dependencies for returns together with the effect of the microstructure noise. As a consequence, the resulting measure of market inefficiency excludes as well any linear effect of predictability.

5.3.1 Approaches for selecting ARMA model

We test five different approaches for identifying an ARMA model order (P and Q) for filtering out microstructure noise (MN). A step common for all approaches is to find P and Q for a time interval with the size of one year. We derive from an idea of taking such a model that gives the maximum entropy value of residuals. Then, we apply a more standard way to choose a model based on the Bayesian information criterion (BIC) [Schwarz, 1978]. This method is used in the previous section. The next pair of approaches are constructed so that they use only previously obtained returns for each week, so they may be used for the analysis in real time. Finally, we use the Model Confidence Set (MCS [Hansen et al., 2011]) which considers all possible ARMA models together without pairwise testing. We consider only models where $P + Q \leq 5$. The approaches are listed below.

1. *Maximum entropy in-sample.* We fix values P, Q for each year, so that the entropy of residuals reaches its maximum value. For each week, we choose P and Q corresponding to the year at which it starts.

2. *BIC in-sample.* We choose values P, Q for each year, so that BIC attains its minimum value. $BIC = -\ln L - \ln(N)(P + Q)$, where L is the likelihood function of residuals assuming that they are distributed normally and N is the size of available data excluding missing values.

3. *Maximum entropy, out-of-sample.* For each month, we choose P and Q calculated from the data for the previous 12 months. To restrict sharp changes in values P and Q, we follow

the rule below: If the previous values of P and Q give entropy of the current year greater than 1 (using Grassberger's estimator), we keep these values. Otherwise, we change at most one value P or Q by 1 to maximize the entropy value.

4. *BIC, out-of-sample.* We use the BIC criterion for finding P and Q for each month based on the previous yearly data. To forbid large deviations of parameters, we use another penalty function. We use $\ln(N)(|P - P_0| + |Q - Q_0|)$ instead of $\ln(N)(P + Q)$, where P_0 and Q_0 are the previous values of P and Q.

5. *Model Confidence Set.* To determine P and Q for each month, we use the previous year. We divide the year into 12 parts, taking 11/12 of the data for a training dataset, and the last 1/12 for a validation dataset. For each of 21 ARMA(P,Q) models where $0 \leq P+Q \leq 5$, we find ARMA parameters. We find errors of predictions $e_t = Y_t - \hat{Y}_t$, where, Y_t is a return sequence from the validation set and $\hat{Y}_t$ is its prediction. Then, we use the squared error as a loss function, $L_t = e_t^2$. Finally, we construct a Model Confidence Set, which contains a set of models with the smallest losses with 95% confidence interval. We take a model for a month with the smallest number of parameters with the advantage of the smallest Q.

5.3.2 Comparison of approaches for filtering microstructure noise

The five methods for filtering microstructure noise are compared in this section. Table 5.4 is a comparison table for all five methods. The criteria for the comparison of five methods for filtering microstructure noise are almost the same as for the filtering out 0-returns. A column about the number of co-inefficiencies is added.

The methods from second to fourth present close results for the total percentage of inefficient weeks with the minimum obtained by the method using BIC on the past data. The small amount of new inefficient weeks in the case of using MCS is explained by the fact that the white noise model ARMA(0, 0) is usually enough to have small errors of prediction with the comparison of the other ARMA(P,Q) models. The BIC used in-sample also gives a low percentage of new inefficient weeks. Also, this approach is the only one that removes dependence between periodic patterns and inefficiency from the data. The first approach is the only one that increases the amount of ETFs without inefficient weeks to 3.

The number of co-inefficiencies changes from 1 to 2 for different approaches, while the total number of weeks with co-inefficiencies is equal to 5. Since each separate case of detecting inefficiency is a rare event with a probability of about 1-2 percent, we consider all cases of co-inefficiencies for weekly time intervals to be statistically significant.

5.4 Co-inefficiency and cointegration

5.4.1 Cointegration

In the previous sections, we discuss the issue of (in)efficiency of the ETF market for weekly time intervals. Some questions may require more careful study. For instance, reasons for the occurrence of co-inefficiencies are not yet investigated. A research interest is to determine if there is some correlation between assets at a time when both of them have a week with inefficiency. Therefore, in this section, we test assets for cointegration [Engle and Granger, 1987]. Cointegration is a property which means that there is a long-term relationship between individual economic variables, which leads to some mutual change.

Definition 8. *A time series is integrated of order d if* $(1 - L)^d X_t$ *is a stationary process, where* L *is the lag operator and* $1 - L$ *is the first difference.*

$$(1-L)X_t = X_t - X_{t-1}$$

Definition 9. *If $(Y_1, Y_2, \ldots, Y_n)$ is a collection of time series variables each integrated of order d, and there exist coefficients $a_i, i = 1 \ldots n$, such that $\sum_{i=1}^{n} a_i Y_i$ is integrated of order less than d, then $(Y_1, Y_2, \ldots, Y_n)$ are cointegrated.*

A question we are interested in is if two or more ETFs have co-inefficiency, does it imply that they also are cointegrated. An idea is to check a hypothesis that two ETFs have the same inefficient week because of the same structure of returns of this week. We check all weeks except the first and the last one, thus there are 346 weeks. We apply the Engle-Granger cointegration test [Engle and Granger, 1987] on a weekly basis for all time series we find co-inefficiency. We use the augmented Dickey–Fuller test [Dickey and Fuller, 1979] for testing stationarity. The results of the test are in Table 5.5. There is the cointegration only in the last case, which was found using MCS.

5.4.2 Testing the volume-based approach for filtering 0-returns

We aim to test here another approach for filtering out the sources of data regularities. We change the method for identification 0-returns due to staleness to the volume-based approach presented in Section 5.2.1. We leave the "BIC in-sample" approach for filtering out microstructure noise. Detailed results for 10 ETFs are in Table 5.6. The inference about cointegration is in Table 5.7.

Using this set of methods, we conclude that there is cointegration in all 3 cases where co-inefficiency is detected. Thus, we detect co-inefficiency for the group of assets that have the common price dynamics using the volume-based approach. Moreover, it is natural to detect co-inefficiency for the pair of ETFs SPY and IVV, since they track the same index and so have 333 weeks with cointegration out of 346 weeks.

We can conclude that different approaches give us not only different weeks with co-inefficiencies but also different results about cointegration in these weeks. In other words, the analysis of cointegration is sensitive to an approach for filtering out 0-returns. We may interpret detecting co-inefficiency as robust testing of markets' predictability for a specific time period. Appearing a week with co-inefficiency, especially for assets with cointegrated prices, is unlikely to be a coincidence.

5.5 Detecting inefficiency with larger time intervals

We concentrate on weekly time intervals in the previous sections. On a weekly basis, we conclude that we filter out the main sources of data regularities, so that the percentage of intervals with inefficiency is close to the level of significance. However, we restricted ourselves by the short time intervals. Here, we consider monthly and quarterly time intervals to analyze market efficiency using rolling windows but with a greater length. Finally, the dataset with the period of about 2.5 years is considered in Section 5.5.3.

5.5.1 Monthly time intervals

For monthly time intervals, we construct confidence intervals using Monte Carlo simulations for sequences with length increments equal to 100 and then use linear interpolation as described in Section 3.2.2. Here, we omit months with more than 1/3 of 0-returns as usual. We apply the volume-based approach for filtering out 0-returns and then select an ARMA model based on the BIC criteria for each year. The results are in Table 5.8.

Table 5.3: Results for weekly time intervals of filtering out the microstructure noise.

ETF	N. of weeks	Ineff. before ARMA fitting	Ineff. after ARMA fitting	New ineff.	Reject $\mathcal{H}_0$ with 95%	Weeks with ineff.
SPY	330	10	4	0	0	03-May-2007 - 09-May-2007 24-May-2007 - 31-May-2007 23-Feb-2009 - 27-Feb-2009 13-Nov-2009 - 19-Nov-2009
DIA	333	12	6	2	0	14-Apr-2003 - 21-Apr-2003 14-Jan-2004 - 21-Jan-2004 12-Feb-2004 - 19-Feb-2004 04-Feb-2008 - 08-Feb-2008 17-Nov-2008 - 21-Nov-2008 **16-Dec-2008 - 22-Dec-2008**
IWM	294	5	2	0	0	02-Dec-2008 - 08-Dec-2008 **16-Dec-2008 - 22-Dec-2008**
EWJ	14	3	1	0	0	16-Oct-2009 - 22-Oct-2009
XLE	231	7	1	0	0	21-Sep-2005 - 27-Sep-2005
XLF	120	16	5	0	0	22-Sep-2008 - 26-Sep-2008 20-Oct-2008 - 24-Oct-2008 03-Jun-2009 - 09-Jun-2009 **11-Sep-2009 - 17-Sep-2009** 09-Oct-2009 - 15-Oct-2009
XLU	94	6	1	0	0	**11-Sep-2009 - 17-Sep-2009**
IVV	161	0	1	1	0	22-Feb-2006 - 28-Feb-2006
XLB	124	2	4	2	0	28-Feb-2007 - 06-Mar-2007 16-Jul-2007 - 20-Jul-2007 16-Apr-2008 - 22-Apr-2008 21-May-2008 - 28-May-2008
IWO	148	3	0	0	0	

New inefficient weeks are found in the comparison with the results obtained with the probability-based approach. When testing $\mathcal{H}_0$ from Section 3.7, 1 indicates the rejection of the hypothesis. Co-inefficiencies in the last column are in bold.

Table 5.4: Comparison of five methods for filtering out microstructure noise.

Methods	% of ineff. weeks	% of new ineff. weeks	N. of efficient ETFs	N. of ETFs with PP	N. of co-ineff.
1.maxEn in-sample	1.46	0.49	3	1	2
2.BIC in-sample	1.35	**0.27**	1	0	2
3.maxEn out-of-sample	1.38	0.69	1	1	1
4.BIC out-of-sample	**1.32**	0.86	2	1	1
5.MCS	2.53	**0.06**	1	1	2

The smallest values in the first two columns are in bold. New inefficient weeks are found in the comparison with the results obtained with the probability-based approach. The presence of periodic patterns in weeks with inefficiency is defined with 95% of confidence.

Table 5.5: Analysis on co-inefficiency and cointegration with the probability-based approach.

Co-inefficiency	Is there cointegration?	Total number of weeks with cointegration
XLE-XLF: 13-Aug-2009 - 19-Aug-2009	No	73
DIA-IWM: 16-Dec-2008 - 22-Dec-2008	No	130
EWJ-XLF: 16-Oct-2009 - 22-Oct-2009	No	168
XLF-XLU 11-Sep-2009 - 17-Sep-2009	No	152
DIA-EWJ 16-Dec-2008 - 22-Dec-2008	Yes	146

The second column determines if there is cointegration in the week with co-inefficiency. The third column presents the total number of weeks with cointegration for the group of assets.

Table 5.6: Results for weekly time intervals in the case of using the volume-based approach.

ETF	N. of weeks	Ineff. before ARMA	Ineff. after ARMA	New ineff.	Reject $\mathcal{H}_0$ with 95%	Weeks with ineff.
SPY	330	11	4	1	0	18-May-2006 - 24-May-2006 **19-Apr-2007 - 25-Apr-2007** **24-May-2007 - 31-May-2007** 23-Feb-2009 - 27-Feb-2009
DIA	333	17	3	1	0	14-Jan-2004 - 21-Jan-2004 12-Feb-2004 - 19-Feb-2004 26-Jun-2008 - 02-Jul-2008
IWM	294	13	6	0	0	07-Sep-2005 - 13-Sep-2005 22-Feb-2006 - 28-Feb-2006 04-Apr-2007 - 11-Apr-2007 **19-Apr-2007 - 25-Apr-2007** 13-Oct-2008 - 17-Oct-2008 03-Sep-2009 - 10-Sep-2009
EWJ	14	2	0	0	0	
XLE	231	11	6	3	0	21-Sep-2005 - 27-Sep-2005 08-Mar-2006 - 14-Mar-2006 28-Aug-2006 - 01-Sep-2006 **19-Apr-2007 - 25-Apr-2007** 30-Jul-2007 - 03-Aug-2007 06-Nov-2009 - 12-Nov-2009
XLF	120	14	3	0	0	25-Jul-2008 - 31-Jul-2008 08-Sep-2008 - 12-Sep-2008 09-Oct-2009 - 15-Oct-2009
XLU	94	10	0	0	0	
IVV	161	1	2	1	0	**24-May-2007 - 31-May-2007** **13-Feb-2009 - 20-Feb-2009**
XLB	124	6	3	0	0	11-Mar-2008 - 17-Mar-2008 30-Apr-2008 - 06-May-2008 **13-Feb-2009 - 20-Feb-2009**
IWO	148	3	3	0	0	11-Feb-2008 - 15-Feb-2008 06-Oct-2008 - 10-Oct-2008 14-Apr-2009 - 20-Apr-2009

New inefficient weeks are found in the comparison with the results obtained with the volume-based approach. When testing $\mathcal{H}_0$ from Section 3.7, 1 indicates the rejection of the hypothesis. Co-inefficiencies in the last column are in bold.

Table 5.7: Analysis on co-inefficiency and cointegration, the volume-based approach is used.

Co-inefficiency	Is there cointegration?	Total number of weeks with cointegration
SPY - IWM - XLE 19-Apr-2007 - 25-Apr-2007	Yes	195
SPY - IVV 24-May-2007 - 31-May-2007	Yes	333
IVV - XLB 13-Feb-2009 - 20-Feb-2009	Yes	202

The second column determines if there is cointegration in the week with co-inefficiency. The third column presents the total number of weeks with cointegration for the group of assets.

Table 5.8: Results for monthly time intervals in the case of using the volume-based approach.

ETF	N. of months	Ineff. before ARMA	Ineff. after ARMA	New ineff.	Reject $\mathcal{H}_0$ with 95%	Reject $\mathcal{H}_0$ with 99%	Months with ineff.
SPY	83	10	6	1	1	0	**Oct-2003** Aug-2004 **Feb-2005** Jan-2007 **Apr-2007 Sep-2009**
DIA	83	21	3	1	0	0	**Feb-2005 Jun-2006** Jun-2007
IWM	74	21	17	3	0	0	**Oct-2003** Jul-2004 Nov-2004 **Jun-2005 Sep-2005** Oct-2005 Dec-2005 Jan-2006 Feb-2006 **May-2006** Oct-2006 **Apr-2007** Jul-2007 **Dec-2007 Mar-2008** Feb-2009 **Nov-2009**
EWJ	3	1	0	0	0	0	
XLE	58	13	10	2	1	0	**Feb-2005** May-2005 **Jun-2005** Aug-2005 **Sep-2005** Nov-2006 **Mar-2007 Apr-2007** Nov-2007 **Dec-2007**
XLF	30	5	5	0	1	1	Jul-2009 Aug-2009 **Sep-2009 Oct-2009 Nov-2009**
XLU	24	8	3	1	0	0	**Mar-2008** May-2009 **Nov-2009**
IVV	41	1	1	0	0	0	**May-2006**
XLB	31	7	5	0	0	0	**Jun-2006 Mar-2007** May-2008 Aug-2008 Jun-2009
IWO	37	6	3	0	0	0	Oct-2008 **Oct-2009 Nov-2009**

New inefficient months are found in the comparison with the results obtained with the volume-based approach. When testing $\mathcal{H}_0$ from Section 3.7, 1 indicates the rejection of the hypothesis. Co-inefficiencies in the last column are in bold.

The percentage of months with inefficiency after filtering 0-returns is 20.04%. The percentage of months with inefficiency after filtering microstructure noise is 11.42% with the amount of new inefficient months appearing after the last stage of filtering data regularities is 1.72%. For monthly intervals, dependence between inefficiencies and periodic patterns is still in the data: There is the rejection of the null hypothesis of the independence for SPY and XLE for the 95% level of confidence and for XLF for the 99% level of confidence. The total percentage of inefficient weeks is significantly greater than 1%. This means that for monthly data it is insufficient to filter out the four mentioned sources of regularity to conclude that the remaining signal of market inefficiency is weak. Thus, either some sources of regularities that cause the detection of inefficient months are not considered, or there exist trading strategies that can be applied for months but not for the weeks since more data are analyzed. Since periodic patterns still exist in the data even after ARMA fitting, other ways to choose an ARMA model or some additional techniques can be used for filtering out this pattern, which can be the effect of a bid-ask bounce and price discreteness. Finally, there are a lot of intersections of months with inefficiency. We report below all groups with statistically significant co-inefficiency: SPY-IWM: Oct-2003; SPY-DIA-XLE: Feb-2005; IWM-XLE: Jun-2005, Sep-2005, Apr-2007, Dec-2007; IWM-XLF-XLU-IWO: Nov-2009.

5.5.2 Quarterly time intervals

We extend the analysis also for quarters. The results are in Table 5.9. The percentage of quarters with inefficiency after filtering 0-returns is 12.18%. The percentage of quarters with inefficiency after filtering microstructure noise is 10.26% with the amount of new inefficient quarters appearing after the last stage of filtering regularities is 1.28%. There is the only statistically significant co-inefficiency found for the group of ETFs DIA-XLU-XLB for the first quarter of 2009. For 95% level of confidence, there is no dependence between periodic patterns and quarters with inefficiency. As in the case of monthly time intervals, the amount of quarters with inefficiency is statistically significant, that means either unaccounted sources of regularities or existing profitable trading strategies.

Table 5.9: Results for quarterly time intervals in the case of using the volume-based approach.

ETF	N. of quarters	Ineff. before ARMA	Ineff. after ARMA	New ineff.	Reject $\mathcal{H}_0$ with 95%	Quarters with ineff.
SPY	28	3	2	0	0	Q3-2007, Q1-2008
DIA	28	0	1	1	0	**Q1-2009**
IWM	25	3	2	0	0	Q1-2004, **Q4-2007**
EWJ	1	1	0	0	0	
XLE	20	5	5	0	0	Q1-2005, Q1-2006, Q2-2006, Q4-2006, Q1-2007
XLF	10	3	1	0	0	Q2-2009
XLU	8	2	2	0	0	Q3-2008, **Q1-2009**
IVV	13	0	1	1	0	**Q4-2007**
XLB	10	1	1	0	0	**Q1-2009**
IWO	13	1	1	0	0	Q4-2009

Q1-Q4 denotes quarters of each year. New inefficient quarters are found in the comparison with the results obtained with the volume-based approach. When testing $\mathcal{H}_0$ from Section 3.7, 1 indicates the rejection of the hypothesis. Co-inefficiencies in the last column are in bold.

5.5.3 Yearly time interval

In this section, we make a comparison between the results of the article [Calcagnile et al., 2020] and our methodology with the new step of filtering 0-returns. We consider the time interval from the 13th July 2006 to the 1st December 2009. We take for the analysis the same 55 ETFs as in the work of [Calcagnile et al., 2020] but exclude such assets that have the fraction of minutes with 0-volume greater than 1/3. We reproduce the analysis from the work [Calcagnile et al., 2020] in the following way. We do not fill the dataset with minutes without trading and calculate standardized returns using the exponentially weighted moving average (EWMA) for estimating volatility.

$$\hat{\sigma}_t = \mu_1^{-1} \alpha \sum_{i>0} (1-\alpha)^{i-1} |r_{t-i}|,$$

where $\mu_1 = \sqrt{\frac{2}{\pi}}$ and $\alpha = 0.05$.

An ARMA(P,Q) is chosen with the restriction $P + Q \leq 6$ with the smallest BIC obtained by MATLAB function armax.

For the same group of ETFs we make the analysis including the step of filtering out price staleness: reconstruct missing price values, calculate standardized returns using expectation-maximization algorithm (EM, [Sucarrat and Grønneberg, 2020]), filter out 0-returns using the volume-based approach from Section 5.2.1 and consider ARMA residuals using the BIC. We calculate a relative entropy

$$\tilde{h}_k = \frac{\hat{h}_k^G}{\hat{h}_1^G}$$

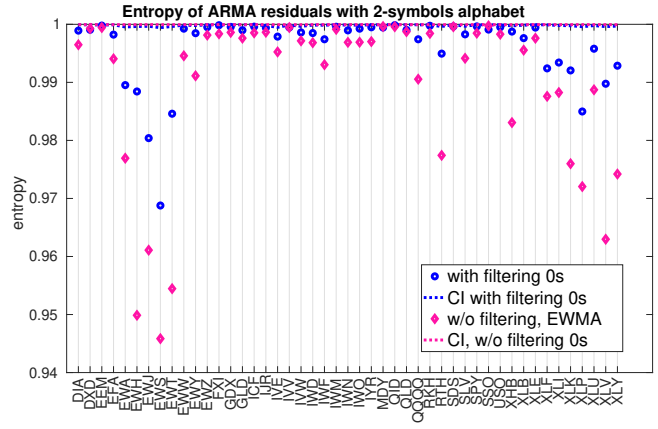
to filter out the difference in the frequencies of single symbols. Here, $\hat{h}^G$ stands for the Grassberger's estimation (Eq. 2.5) We set $k = 10$ for the 2-symbols discretization and $k = 7$ for the 3-symbols discretization and present results in Figures 5.4a and 5.4b.

The conclusion is more straightforward for 2-symbols: adding the step of filtering 0-returns due to staleness increases the entropy of discretized time series. A representative situation for 3-symbols is either a significant increase in entropy values or a slight decrease. Note that confidence intervals for the case of filtering 0-returns is less than in the case of complete data. However, there are no ETFs that can be defined as efficient for both types of discretization and both types of analysis.

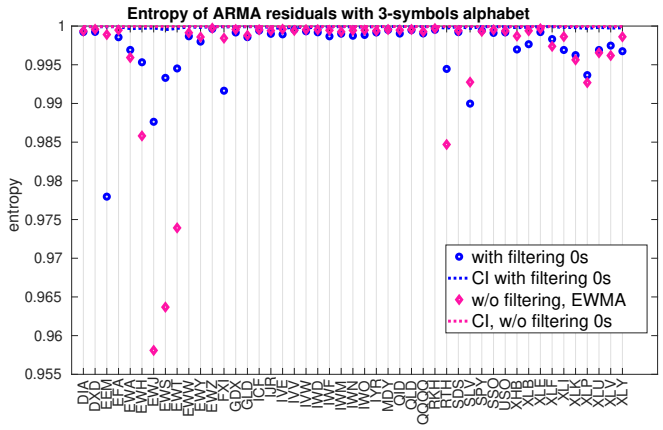
To provide a more complete picture, we also use the mixed-based approach from Section 5.2.1. The results are in Figures 5.5a and 5.5b. The main difference between approaches is the appearance of four assets that are not defined as inefficient even for the considered time period of about 2.5 years for 2-symbols discretization. However, if we move to 3-symbols, there are still no return time series that can be defined as fully unpredictable.

5.6 Discussion on the efficiency of the ETF market

We have studied the efficiency of the ETF market. After the implementation of the multi-step filtering method, which allows to remove daily patterns, heteroscedasticity, 0-returns, and microstructure noise, the fraction of weeks with inefficiency decreases to a value slightly greater than 1 percent. Taking into account the 99% level of confidence for the test for inefficiency, we can conclude that there exists a slight evidence of inefficiency in the ETF market at a high-frequency time scale. [Molgedey and Ebeling, 2000] came to the similar conclusion and stated that the Dow Jones Index is not fully random, but nearly random. We have shown that data



(a) 2-symbols sequence



(b) 3-symbols sequence

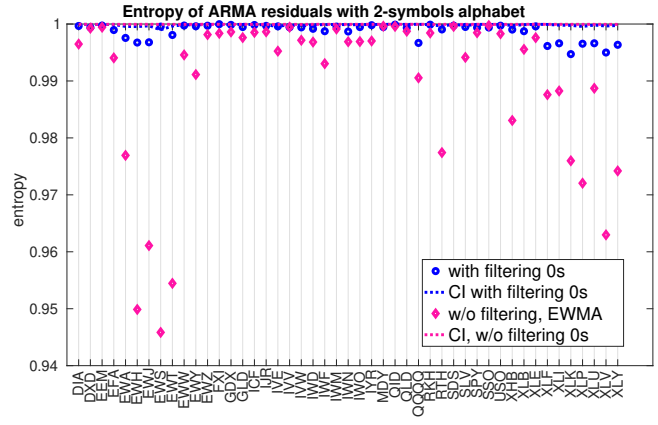
Figure 5.4: Relative entropies calculated with and without filtering out 0-returns for the group of ETFs. Filtering 0-returns is made using the volume-based approach.

regularities contribute to a measure of inefficiency. Such stylized facts are empirical properties of price returns. Accordingly, when modeling prices in efficient or nearly efficient markets, such properties can be taken into account. For example, the agent-based model proposed in [McGroarty et al., 2019] reproduces clustered volatility and autocorrelation of returns.

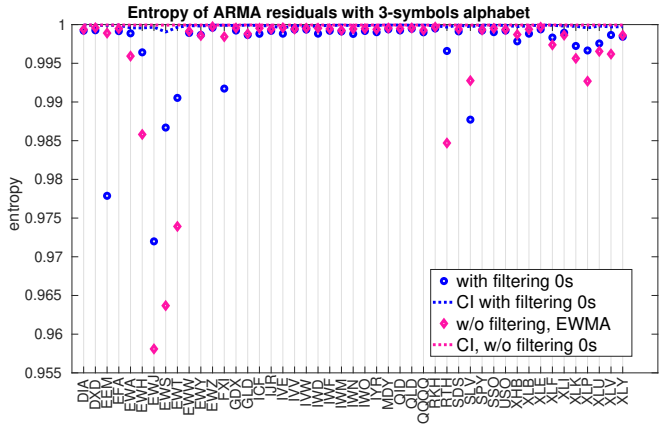
Since some ETFs aim to track indexes having some stocks in common, it is natural to expect that some inefficient weeks appear simultaneously due to exogenous events in the markets. That is, there can be some exogenous cause that leads to more predictability for the several assets at the same time. We indeed have detected co-inefficiency for the ETFs SPY and IVV that are designed to track the S&P 500 stock market index.

The measure of market inefficiency, by analogy with the case of weeks, is equal to about 11 and 10 percent for months and quarters, respectively. A transition from weeks to months significantly increases the measure of inefficiency. Important differences between the two time intervals are the length of blocks taken into consideration and the amount of data in which different predictable patterns may be found. Since the degree of inefficiency varies depending on

the chosen interval length, we consider the choice of the optimal length in Chapter 7.



(a) 2-symbols sequence



(b) 3-symbols sequence

Figure 5.5: Relative entropies calculated with and without filtering out 0-returns for the group of ETFs. Filtering 0-returns is made by the mixed approach.

Chapter 6

Inefficiency of the Russian stock market

The chapter introduces four original contributions. First, we construct a method for filtering out heteroskedasticity and price staleness. This filtering process helps to identify a true degree of market inefficiency. Second, we calculate the degree of market inefficiency for the period of 10 years using monthly intervals. We conclude that the degree of market inefficiency for the Moscow Stock Exchange was greater than 80%. Third, we determine which pair of stocks exhibits the largest amount of inefficiency, as measured by estimating Shannon's entropy on their high-frequency price time series. We show that months where the predictability of stock prices attains its maximum cluster together. We find out the form of stock price behavior that is repeated most often during inefficient time intervals. Finally, we estimate the closeness of price movements using three measures of entropy. Based on these results, we cluster together groups of stocks for which the Efficient Market Hypothesis is rejected, thus pointing out how market inefficiency displays some dependence on the financial sector they belong to.

The chapter is organized as follows. Section 6.2 describes the dataset. In Section 6.3, we present the method for filtering out both heteroskedasticity and price staleness. Section 6.4 presents the results for simulated data. We calculate entropy of price returns after filtering out data regularities in Section 6.6. We introduce and discuss methods for stock prices clustering in Section 6.7. The obtained results for efficiency of the Moscow stock exchange are discussed in Section 6.8. Almost all results in this chapter are published in our article [Shternshis et al., 2022b].

6.1 Introduction

When prices reflect all available information, the market is called efficient [Samuelson, 1965]. One way to claim the efficiency of a market is by testing the Efficient Market Hypothesis (EMH) [Malkiel, 2003]. In its weak form, the EMH considers that the last price incorporates all the past information about market prices. If the weak form of EMH is rejected, previous prices help to predict future prices. For traders, market efficiency means that analyzing the history of previous prices does not help to design a strategy that produces abnormal profits. For a company issuing shares, market efficiency means that the cost of its share already reflects all information about the valuation and decisions of the company. The EMH is of great interest also in research. Mathematical models of an asset price are usually based on the assumption that the price follows

a martingale: the expected value of a future price is the current value of the price. If the EMH is rejected, there should be an estimation of the future price that is better than its current value. In such a case, new models should be created.

A review of studies investigating the EMH was presented by [Fama, 1970, 1991]. The martingale hypothesis was also tested later. It was shown that the efficiency of a market depends on the development of the country [Kim and Shamsuddin, 2008]. Moreover, the martingale hypothesis was confirmed on short time intervals, but it may be violated on longer intervals [Linton and Smetanina, 2016]. In addition, there is a range of strategies designed to increase an expected profit. High-frequency and algorithmic trading strategies were discussed in [Mandes, 2016]. Statistical and machine learning methods for high frequency trading were reviewed in [Huang et al., 2019]. The existence of such profitable strategies contradicts the Efficient Market Hypothesis.

According to [Grossman and Stiglitz, 1980], a degree of market inefficiency determines the effort investors are willing to expend to gather and trade on information. The goal of this chapter is to investigate the degree of stock market efficiency of the Moscow stock Exchange using the Shannon entropy. We quantify the degree of market inefficiency and the degree of price randomness. We aim to distinguish between price predictability due to stylized facts of financial time series [Cont, 2001] and due to market inefficiency. For instance, some regularity patterns were connected to calendar effects [French, 1980, Mills et al., 2000]. In particular, we consider volatility clustering and price staleness as data regularities needed to be filtered out. Based on the behavior of stock prices, we group them into clusters using several measures. Combining stocks into one cluster means a common price behavior that moves prices away from complete randomness. Before estimating the degree of market efficiency, we need to dispose of regularities that make prices more predictable but that do not imply any profitable strategies. A method for filtering regularities was introduced in [Calcagnile et al., 2020]. However, such a process of filtering is not usually applied in other research studies. In fact, deviations of price behavior from perfect randomness may be the result of some known regularity pattern, such as volatility clustering or daily seasonality, but not a signal of market inefficiency.

One of the innovations of this chapter is a new method for filtering data regularities, that allows to estimate volatility and a degree of price staleness minute by minute. We process data by filtering regularities of financial time series including volatility clustering and price staleness. Price staleness [Bandi et al., 2020] is defined as a lack of price adjustments yielding 0-returns. Traders may trade less because of high transaction costs and so the price does not update. Price staleness produces an extra amount of 0-returns called *spurious 0-returns*. The other source of 0-returns in the time series is price rounding. Estimations of volatility and the degree of price staleness are mutually connected: Spurious 0-returns appear due to price staleness tend to underestimate volatility. At the same time, volatility estimation is needed to calculate the expected amount of 0-returns due to rounding.

One method for estimating volatility in the presence of spurious 0-returns was presented in [Sucarrat and Grønneberg, 2020]. It uses expectation-maximization algorithm [Dempster et al., 1977] to estimate returns in the places of all 0-returns and uses the GARCH(1,1) model to estimate volatility [Bollerslev, 1986]. The maximization of the likelihood function appearing at each step of the considered algorithm requires several parameters for numerical optimization. If the estimation of volatility is sensitive to these parameters then they may affect the entropy of returns standardized by volatility and the amount of 0-returns in the time series. In this chapter, we suggest a modification of moving average volatility estimations that require an adjustment of the only parameter that can be defined using out-of-sample testing. The idea is to adopt a simple method for volatility estimation such that price staleness is taken into consideration. Moreover, while estimating volatility, we filter out spurious 0-returns.

The degree of market efficiency was measured for many countries [Patra and Hiremath, 2022].

Stock indices for 20 countries were considered in [Risso, 2008]. The efficiency of 11 emerging markets and the U.S. and Japan markets was measured in [Cajueiro and Tabak, 2004]. U.S. stock markets were considered in a recent paper [Alvarez-Ramirez and Rodriguez, 2021]. A review of articles about Baltic countries was presented in [Degutis and Novickytė, 2014]. A degree of uncertainty of Chinese [Ahn et al., 2019], Tunisian [Mahmoud et al., 2014], Mexican [Coronel-Brizio et al., 2007], and Portuguese [Dionisio et al., 2006, Pascoal and Monteiro, 2014] stock markets was also considered by using entropy measures. However, the efficiency of the Russian stock market has not yet been analyzed. In this chapter, we present an analysis of market efficiency based on the estimation of Shannon entropy for a group of 18 stocks of Russian companies from five industries.

Entropy measures are widely used in mathematical finance. The conditional entropy was used in [London et al., 2001] to measure the randomness in stock and exchange markets at different time scales. A generalization of the Shannon entropy, the Tsallis entropy [Tsallis, 1988], was proposed as a risk measure during financial crises and crashes [Gradojevic and Caric, 2017, Gençay and Gradojevic, 2017]. The permutation transition entropy was introduced in [Zhao et al., 2020] to measure the complexity of financial time series. [Marschinski and Kantz, 2002] and [Kwon and Yang, 2008] studied the transfer entropy introduced in [Schreiber, 2000] to investigate the strength and the direction of the information transfer in the U.S. stock market. The cross sample entropy was applied to quantify the asynchrony of the two exchange rates return series [Liu et al., 2010]. Finally, an entropy measure was used to identify different types of trading behaviors based on historical prices and news in [Liu et al., 2020]. Here, we use the Shannon entropy as a measure of price returns randomness.

Our main goal is to measure a degree of efficiency of the Moscow Stock Exchange. The data taken for the study are reviewed in the next section.

6.2 Moscow Stock Exchange

We study the Moscow Stock Exchange. We consider close prices aggregated at one-minute time scale. In particular, we select only minutes of the main trading session from 10:00 to 18:40. The time interval covers ten years from 2012 to 2021. The time period is divided into monthly time intervals. We take 18 companies, 16 of them are from five sectors: oil industry, metallurgy, banks, telecommunications, and electricity. All stocks are listed in Table 6.1. There are 2520 trading days. Assuming that there are 520 minutes in each trading day, there are 1,310,400 trading minutes in total. We use the [Brownlees and Gallo, 2006]'s algorithm of an outlier detection. All data are provided by Finam Holdings (`https://www.finam.ru/`).

In the next section, we filter data regularities from financial time series. Then, we calculate the degree of efficiency of the market using the Shannon entropy. Finally, we use the resulting time series to group stocks into clusters. In particular, we use hierarchical clustering with the Kullback–Leibler distance discussed in Section 6.7.1.

6.3 Estimation of volatility and a degree of price staleness

6.3.1 Exponentially weighted moving average

To estimate a degree of market efficiency, we first should eliminate the known patterns of predictability, such as a daily seasonality. Financial agents operating in the market tend to trade less in the middle of a day. It is reflected in prices, but this pattern in trading volume should be filtered out to detect genuine patterns of inefficiency. Other known regularities include volatility

Table 6.1: Stocks of Russian companies traded at Moscow Exchange.

Ticker	Company	Sector	Size	Outliers
GAZP	Gazprom	Oil	1,307,427	50
LKOH	Lukoil	Oil	1,287,582	192
ROSN	Rosneft	Oil	1,270,592	130
SNGS	Surgutneftegaz	Oil	1,211,809	11
TATN	Tatneft	Oil	1,191,390	174
SBER	Sberbank	Bank	1,309,402	37
VTBR	VTB Bank	Bank	1,287,330	0
CHMF	Severstal	Metal	1,214,735	157
NLMK	Novolipetsk Steel	Metal	1,194,324	58
GMKN	Nornikel	Metal	1,272,769	197
MTLR	Mechel	Metal	1,084,990	161
MAGN	Magnitogorsk Iron and Steel Works	Metal	1,106,771	13
MTSS	Mobile TeleSystems	Telecommunications	1,153,527	260
RTKM	Rostelecom	Telecommunications	1,140,798	134
HYDR	RusHydro	Electricity	1,252,584	0
RSTI	Rosseti	Electricity	1,094,244	0
AFLT	Aeroflot	Airline	1,083,552	123
MGNT	Magnit	Food retailer	1,184,223	544

For each company, we specify the ticker of stock, its sector, the size of data, and the amount of outliers removed. The size is given in the amount of minutes with trading activity.

clustering, price staleness, and microstructure noise. See Appendix 6.9 for a guide on filtering out data regularities. The contribution of this chapter is that it devises a simple method for filtering volatility clustering and price staleness. One of the methods used to estimate volatility is the exponentially weighted moving average (EWMA).

We define price returns as $r_t = \ln\left(\frac{P_t}{P_{t-1}}\right)$, where P_t is the last price available at time t. In order to estimate volatility σ_t, we apply the exponentially weighted moving average [Hunter, 1986] of values $\mu_1^{-1}|r_i|$, $i < t$, where $\mu_1 = \sqrt{\frac{2}{\pi}}$.

$$\hat{\sigma}_t = Sig_1(\alpha, r_{t-1}, \hat{\sigma}_{t-1}) = \alpha\mu_1^{-1}|r_{t-1}| + (1-\alpha)\hat{\sigma}_{t-1} \tag{6.1}$$

This form of exponential moving average was used in [Calcagnile et al., 2020]. Here, $E[|r_t|] = \mu_1\sigma_t$ is used assuming that returns are normally distributed, $r_t \sim N(0, \sigma_t)$. More weights are provided for the more recent data. An alternative formula based on expectation $E[r_t^2] = \sigma_t^2$ is described as follows.

$$\hat{\sigma}_t^2 = Sig_2(\alpha, r_{t-1}, \hat{\sigma}_{t-1}) = \alpha r_{t-1}^2 + (1-\alpha)\hat{\sigma}_{t-1}^2 \tag{6.2}$$

A large value of return increases the value of volatility. The current value of volatility reflects all available values of returns and changes slowly if the value of α is small. This method for volatility estimation can be considered as the kernel estimator [Fan and Wang, 2008], there the kernel is the exponential density rotated about a vertical axis.

We follow the approach suggested by [Morgan et al., 1996] (p. 97) to find optimal values of α in Equations 6.1 and 6.2. The value of α is selected so that it minimizes error $Er_\sigma(\alpha) =$

$\sum_t (\hat{\sigma}(\alpha)_t^2 - r_t^2)^2$. In order to minimize $Er_\sigma(\alpha)$ as a function of the only parameter $0 < \alpha < 1$, we apply Brent's algorithm[10] [Brent, 1971]. We modify the exponential moving average method in Section 6.3.3 so that it removes a bias due to the effect of price staleness discussed in the next section.

6.3.2 Estimation of price staleness

Let us define an *efficient* price, P^e, as a continuous process following a Geometric Brownian Motion.

$$P_t^e = P_0^e + \int_0^t \sigma_s P_s^e dW_s$$

An observed price moves along a discrete grid. Possible price values are multiples of the tick size, d.

$$P_t = d \cdot \left[\frac{P_t^e}{d} \right]$$

If the efficient price changes insignificantly, the return of the rounded price is equal to 0. Analogically, if the return of rounded price is 0, the return of efficient price has a value close to 0. We use Equation 4.3 to estimate the probability that a return of rounded price has zero value[11]. It is derived in Section 4.1 by considering the probability that a price following a Geometric Brownian Motion moves less than one tick size, assuming that price increments are normally distributed.

Apart from price discreteness, financial assets have a subdiffusive nature of an asset, that is existing of periods of price stagnation during which the price does not change [Magdziarz et al., 2011]. Thus, there is another source for obtaining 0-returns, namely price staleness [Bandi et al., 2020]. Price staleness represents a regularity pattern of the dynamics, namely, the efficient (fundamental) price of an asset is not updated because of a number of reasons, such as no transactions because of high cost, which makes trading unprofitable for agents. A degree of price staleness is time-varying and have not constant intraday pattern [Zhu and Liu, 2023]. Price staleness results in a persistence pattern of "*spurious*" 0-returns. In particular, such a pattern tends to reduce any estimation of the volatility. Therefore, we need to filter out 0-returns due to price staleness while retaining 0-returns due to rounding for a genuine estimation of volatility.

We save 0-returns in the amount of the sum of past values of the probability in Equation 4.3. We set other 0-returns as missing values. We adopt this method to estimate the degree of price staleness together with volatility in the next section.

6.3.3 Modification of exponentially weighted moving average

In this Section, we present a modification of the EWMA that takes into consideration the effect of price staleness. Our modification of the EWMA is based on the suggestion for estimating volatility σ_t as $\hat{\sigma}_{t-1}$ (i.e., by setting $\alpha = 0$), if the value of r_{t-1} is missing because of price staleness. That is, there is no new information from returns to update the value of volatility. We update the estimation of volatility and price staleness minute-by-minute. This method has the clear advantage of making the online inference possible by processing data in real time.

[10] The method is available in Python by using the function scipy.optimize.minimize_scalar. Alternatively, we could use the golden-section search [Kiefer, 1953] that requires the boundary of the search and the only parameter for the stopping criteria.

[11] We estimate the tick size using a two-step procedure for each month. First, we find the amount of significant digits in price. Then, we determine the most frequent increment in ordered prices. The time step between the end and start of the main trading session is set as 1 minute. Moreover, we consider any time gap without trading more than 2 hours as the closure of the market. We set the time step to be equal to 1 minute for these gaps.

The aim of the method is to estimate volatility and filter out spurious 0-returns due to price staleness. Some 0-returns appear due to price rounding. A 0-return is defined as a value due to rounding and is saved in the sequence if the sum of all p_t from Equation 4.3 moves to a new integer value. These 0-returns are saved in the data. First, we set the number of 0-returns "to save" $N_{save} = 0$ and the first value of a cumulative function $Z_1 = 0$. Thus, initially, each appearance of 0-returns does not affect the value of volatility. The cumulative function is updated $Z_t = Z_{t-1} + p_t$, if r_{t-1} is not defined as missing due to staleness. Each time when $\lfloor Z(t) \rfloor - \lfloor Z(t-1) \rfloor = 1$, N_{save} is increased by 1.

We notice that the first non-zero return after a row of 0-returns due to staleness is the sum of all missing returns generated by a hidden efficient price. This return is also set as missing. However, the value of return used for estimating volatility is calculated as its expected value: $\hat{r}_{t-1} = \frac{r_{t-1}}{\sqrt{N_0+1}}$, where N_0 is the amount of missing values strictly before the non-zero return r_{t-1}. The same is also referred to initially missing values, e.g., due to no-trading or errors in collecting the data.

Another assumption is that a 0-return appears due to staleness if the previous return had the 0-value and was defined to appear due to staleness. We include this rule since we assume that it is more likely that two consecutive 0-returns appear due to high transaction costs than due to rounding (that is, simply speaking, two outcomes of generating Gaussian random variables are less than a tick size).

Generally, for the estimation of volatility at time t we should consider three cases: P_{t-1} is missing (or minute $t-1$ is non-trading), $r_{t-1} = 0$, $r_{t-1} \neq 0$. Thus, the algorithm is the following.

6.3.4 Pseudocode

We provide the algorithm for the case of Sig_1 from Eq. 6.1, which is used in the application for real data. All codes used for this chapter are available in [Shternshis et al., 2022b]. We remove all 0-returns that start the sequence.

Step 0: $\hat{\sigma}_1 = |r_1|/\mu_1$; $Z_1 = 0$, $N_{save} = 0$; $N_0 = 0$

For t from 2 to N, where N is the length of time series:

Step 1:

- If r_{t-1} is missing: $\hat{\sigma}_t = \hat{\sigma}_{t-1}$; Increase N_0 by the amount of consecutive missing prices
- Else if $r_{t-1} = 0$:
  - If $N_{save} > 0$ and $N_0 = 0$: $N_{save} = N_{save} - 1$, $\hat{\sigma}_t = Sig_1(\alpha, 0, \hat{\sigma}_{t-1})$
  - Else: $\hat{\sigma}_t = \hat{\sigma}_{t-1}$, $N_0 = N_0 + 1$, $r_{t-1} =$ missing
- Else: $\hat{\sigma}_t = Sig_1(\alpha, \frac{r_{t-1}}{\sqrt{N_0+1}}, \hat{\sigma}_{t-1})$, $N_0 = 0$

Step 2:

- Calculate p_t (Equation 4.3)
- If r_{t-1} is not missing, $Z_t = Z_{t-1} + p_t$
- If $\lfloor Z(t) \rfloor - \lfloor Z(t-1) \rfloor = 1$, $N_{save} = N_{save} + 1$

Finally, we check if the effect of staleness exists in the price time series.

$$\bar{p} = \frac{\sum_{t=1}^{N} p_t}{N}$$
$$V = \bar{p}(1-\bar{p})N$$

If $N_{real} \leq \sum_t p_t + 1.96\sqrt{V}$, we leave the time series without placing any missing values, where N_{real} is the initial amount of 0-returns. The value of α can be selected using a training set. The optimal value of α minimizes the mean of $(\hat{\sigma}_t^2 - r_t^2)^2$.

6.4 Testing methods for filtering out data regularities

The goal of this section is to assess the accuracy of the estimation of volatility and the degree of price staleness. We aim to choose a method that produces the least errors with respect to the estimation for further analysis on real data. We take the following model of an observed price $\tilde{P}_t$ and $t = 1 \dots 2N$:

$$P_t = \int_0^t \sigma_s P_s dW_s^1$$
$$\tilde{P}_t = P_t(1 - B_t) + \tilde{P}_{t-1} B_t \qquad B_t = \begin{cases} 1 \text{ with probability } q_t \\ 0 \text{ with probability } 1 - q_t \end{cases}$$
$$q_t = q_0 + \int_0^t \mu_s ds + \int_0^t \nu dW_s^2$$

where W^1 and W^2 are two independent Brownian motions with the length of $2N$, $N = 10^5$; an initial price is $P_0 = 100$; and $\nu = 10^{-4}$. $B = 1$ stands for the case when price is not updated due to price staleness. Prices are rounded to two digits, thus, the tick size is $d = 0.01$. We consider four choices for q_t and σ_t listed below.

$$q_t^1 = 0$$
$$q_t^2 = 0.1 + \int_0^t \nu dW_s^2$$
$$q_t^3 = 0.2 + \int_0^t \nu dW_s^2$$
$$q_t^4 = 0.2 + \int_0^t \mu_s^4 ds + \int_0^t \nu dW_s^2$$
$$\mu_t^4 = 0.8\pi/N \cos(8t\pi/N)$$
$$\sigma_t^1 = 5 \times 10^{-4}$$
$$\sigma_t^2 \sim ARCH(1.75 \times 10^{-7}, 0.2, 0.1)$$
$$\sigma_t^3 \sim GARCH(1.25 \times 10^{-8}, 0.1, 0.85)$$
$$\sigma_t^4 \sim GARCH(1.25 \times 10^{-8}, 0.15, 0.8)$$

We consider four cases for price staleness: the absence of price staleness; two stochastic probabilities with different constant means; a periodic mean. For all four cases for volatility, the unconditional expected value of σ_t is 5×10^{-4}. The first choice of volatility is a constant. Then, we consider the ARCH model [Engle, 1982] with two lagged values, where 0.2 and 0.1 correspond to the first and the second lags, respectively. Volatility values directly depend only on the previous returns values. The dependency on the previous return should be reflected in the value of smoothing parameter. The third and fourth choices are GARCH(1,1) models [Bollerslev, 1986], where the last parameter (0.85 or 0.8) stands for the coefficients for lagged variances. We consider two sets of parameters for a GARCH model, giving less persistence to the fourth model.

We divide the data into two equal parts with the size N. The first part is a training set for finding optimal values of α from Equations 6.1 and 6.2. The second part is a testing set for calculating errors represented in Tables 6.2 and 6.3. We compare two methods that use Sig_1

and Sig_2 for volatility estimation. For each method, we find the optimal value of α. In addition, we set a fixed value of alpha, $\alpha = 0.05$, as a benchmark for the comparison. We also apply non-modified EWMA estimation from Section 6.3.1 with selected optimal value of α to show the contribution of filtering 0-returns to the accuracy of volatility estimation. We simulate 10^3 prices for each model.

Table 6.2 represents a mean absolute percentage error (MAPE) that is $\frac{1}{N}\sum_t |\frac{\hat{\sigma}_t - \sigma_t}{\sigma_t}|$ for six different approaches. These approaches differ in the choice of a function for volatility, the value of α, and the presence of missing values. Table 6.3 represents three values for each of the two methods using Sig_1 and Sig_2 for volatility estimation. The first value is the optimal value of α. The second is $Er_N = |\frac{N_{round}N_A}{N_0 N} - 1|$, where N_{round} is the amount 0-returns that would appear due to rounding (before adding the effect of staleness in the simulated data), N_A is the amount of remaining non-missing returns, and N_0 is the amount of 0-returns. Er_N represents the absolute error of the proportion of 0-returns that remain in the data and are defined as 0-returns due to rounding. The third value is the proportion of data set as missing values (that is, $1 - \frac{N_A}{N}$).

It can be seen from Table 6.2 that the method that more often produces the lowest value of MAPE is with fixed $\alpha = 0.05$ and Sig_1 used for volatility estimation. Moreover, for almost all cases, filtering 0-returns makes the volatility estimate more accurate. The error of the amount of 0-returns due to rounding is smaller for the function Sig_1 than for the function Sig_2 for all 16 cases.

After the comparison of the two functions of volatility estimation, we decide to use Sig_1, which uses absolute values of returns, in the next sections. We fix the value of smoothing parameter α as 0.05 for the simplicity of further analysis.

6.5 Detection of inefficiency

We perform four steps to determine if the time interval is efficient or not. The steps are filtering out data regularities, discretization of filtered price returns, estimating entropies, and detecting significantly low entropy values. First, we filter out data regularities that are discussed in Chapter 3 and Section 6.3. For the brief discussion we refer to Appendix 6.9. We estimate the entropy of the filtered return time series using Equation 2.4.

We consider all blocks of length k that do not contain missing values. We take the following value of k:

$$k = max(K : K < \lfloor \log(n_b(K)) \rfloor), \tag{6.3}$$

where $n_b(k)$ is the number of blocks of length k. The restriction on a value of k allows having enough blocks to estimate probabilities appearing in k-th order entropy [Marton and Shields, 1994]. The base of the logarithm is the size of alphabet A (3 or 4).

Table 6.2: Results for volatility estimation.

model	MAPE, method v_1	MAPE, method v_2	MAPE with $\alpha = 0.05$, v_1	MAPE with $\alpha = 0.05$, v_2	MAPE w/o filtering 0-returns, v_1	MAPE w/o filtering 0-returns, v_2
σ_1, q^1	0.0193 (0.0007,0.0507)	**0.017** (0.0014,0.0406)	0.0975 (0.0955,0.0995)	0.0897 (0.0878,0.0915)	0.0193 (0.0007,0.0507)	**0.017** (0.0014,0.0406)
σ_1, q^2	**0.0607** (0.0245,0.1017)	0.0629 (0.0293,0.1057)	0.095 (0.093,0.0972)	0.0914 (0.0893,0.0936)	0.0862 (0.0459,0.131)	0.0674 (0.0294,0.1154)
σ_1, q^3	**0.0737** (0.0333,0.1278)	0.0756 (0.033,0.1338)	0.0948 (0.0928,0.0971)	0.0915 (0.0894,0.094)	0.138 (0.0917,0.1863)	0.0888 (0.0368,0.1592)
σ_1, q^4	**0.0716** (0.0323,0.1213)	0.0739 (0.0354,0.1268)	0.0949 (0.0926,0.0973)	0.0913 (0.089,0.0937)	0.1404 (0.1022,0.1875)	0.0873 (0.0405,0.1516)
σ_2, q^1	0.1121 (0.1082,0.121)	0.1183 (0.1146,0.1244)	0.1459 (0.1438,0.1481)	0.1446 (0.1422,0.147)	**0.1118** (0.108,0.1207)	0.1179 (0.1144,0.1243)
σ_2, q^2	0.1359 (0.1163,0.1715)	0.1411 (0.1237,0.1765)	0.1462 (0.1439,0.1487)	0.1489 (0.1457,0.1526)	**0.1341** (0.1043,0.1819)	0.1407 (0.1193,0.1832)
σ_2, q^3	**0.146** (0.1198,0.1958)	0.1519 (0.1266,0.1981)	0.1473 (0.1449,0.1499)	0.1496 (0.1464,0.1534)	0.1649 (0.1196,0.2271)	0.1589 (0.123,0.222)
σ_2, q^4	**0.146** (0.1205,0.1912)	0.15 (0.1256,0.1986)	0.1472 (0.1447,0.1498)	0.1494 (0.1463,0.1532)	0.1696 (0.1274,0.2261)	0.1571 (0.1223,0.2239)
σ_3, q^1	0.1479 (0.1446,0.1513)	**0.1473** (0.1442,0.1505)	0.1495 (0.1467,0.1522)	**0.1473** (0.1442,0.1502)	0.1479 (0.1446,0.1513)	**0.1472** (0.1441,0.1503)
σ_3, q^2	0.1592 (0.1485,0.1891)	0.1613 (0.1508,0.1857)	**0.1529** (0.149,0.1574)	0.1546 (0.1497,0.1598)	0.1622 (0.144,0.2033)	0.1628 (0.1491,0.1978)
σ_3, q^3	0.1681 (0.1528,0.2048)	0.171 (0.1556,0.2178)	**0.1567** (0.1525,0.1616)	0.1584 (0.1536,0.1639)	0.1904 (0.154,0.2477)	0.1815 (0.1546,0.2464)
σ_3, q^4	0.1668 (0.1527,0.1997)	0.1701 (0.1555,0.2178)	**0.1568** (0.1525,0.1613)	0.1583 (0.1537,0.1633)	0.192 (0.1591,0.246)	0.181 (0.1556,0.2455)
σ_4, q^1	0.1897 (0.1856,0.1952)	**0.1873** (0.1838,0.1911)	0.1881 (0.1844,0.1918)	0.1924 (0.1879,0.1968)	0.1897 (0.1856,0.1952)	0.1873 (0.1837,0.1911)
σ_4, q^2	0.2035 (0.1906,0.2454)	0.2057 (0.1921,0.2474)	**0.1954** (0.1891,0.2022)	0.2037 (0.1961,0.2119)	0.2049 (0.1836,0.2617)	0.2079 (0.1902,0.2642)
σ_4, q^3	0.2146 (0.1965,0.2623)	0.2166 (0.1996,0.2757)	**0.2015** (0.1951,0.2077)	0.2101 (0.2026,0.2177)	0.2318 (0.1912,0.307)	0.2294 (0.1988,0.3082)
σ_4, q^4	0.214 (0.1967,0.2591)	0.2155 (0.1986,0.2689)	**0.2013** (0.1951,0.2088)	0.2097 (0.2023,0.2185)	0.2338 (0.1976,0.3064)	0.2286 (0.1988,0.306)

The first column indicates a model. Columns 2 and 3 represent results for two methods described in Section 6.3.3. Columns 4 and 5 are for the same methods but with the fixed value of α. Columns 6 and 7 show the error of the standard EWMA approach with the selected optimal value of α. 95% CI is presented below each averaged statistic. v_1 stands for using Sig_1; v_2 stands for using Sig_2. The minimum value for each row is in bold.

Table 6.3: Results for filtering out 0-returns.

model	α for v_1	α for v_2	Er_N, v_1	Er_N, v_2	Fraction of data deleted, v_1	Fraction of data deleted, v_2
σ_1, q^1	0.0027 (0.0,0.0137)	0.0022 (0.0,0.0103)	0.0006 (0.0,0.0)	0.0015 (0.0,0.0259)	0.0001 (0.0,0.0)	0.0003 (0.0,0.0052)
σ_1, q^2	0.0228 (0.0033,0.0569)	0.0259 (0.0044,0.067)	0.0094 (0.0004,0.026)	0.011 (0.0005,0.0295)	0.2005 (0.0814,0.3244)	0.2008 (0.0818,0.3247)
σ_1, q^3	0.0335 (0.006,0.0902)	0.0379 (0.0058,0.1063)	0.0106 (0.0005,0.0288)	0.0121 (0.0005,0.0336)	0.3661 (0.2474,0.481)	0.3659 (0.246,0.4797)
σ_1, q^4	0.0314 (0.0056,0.0824)	0.036 (0.007,0.0966)	0.0104 (0.0004,0.0283)	0.0122 (0.0007,0.0355)	0.3628 (0.2521,0.4717)	0.3626 (0.2515,0.4713)
σ_2, q^1	0.0039 (0.0,0.0161)	0.0037 (0.0006,0.0146)	0.0149 (0.0,0.0438)	0.035 (0.0,0.0586)	0.0029 (0.0,0.0092)	0.0067 (0.0,0.0134)
σ_2, q^2	0.035 (0.0059,0.0903)	0.0367 (0.0054,0.1021)	0.0209 (0.0012,0.0448)	0.0319 (0.0039,0.0601)	0.2016 (0.0856,0.3249)	0.2032 (0.0878,0.3263)
σ_2, q^3	0.0489 (0.0079,0.1326)	0.0556 (0.0091,0.1476)	0.0217 (0.001,0.0473)	0.0275 (0.0022,0.0603)	0.3706 (0.25,0.4835)	0.371 (0.2518,0.4836)
σ_2, q^4	0.049 (0.0093,0.1268)	0.0525 (0.0082,0.1484)	0.0206 (0.0012,0.0443)	0.0274 (0.0013,0.0571)	0.3645 (0.2495,0.4695)	0.3651 (0.2491,0.4692)
σ_3, q^1	0.0424 (0.0349,0.0527)	0.048 (0.0392,0.0603)	0.0034 (0.0,0.0337)	0.0089 (0.0,0.0402)	0.0007 (0.0,0.0067)	0.0018 (0.0,0.0085)
σ_3, q^2	0.0672 (0.0267,0.1456)	0.0767 (0.0249,0.1592)	0.0155 (0.0006,0.0404)	0.02 (0.0009,0.0551)	0.1995 (0.0826,0.3248)	0.1997 (0.0826,0.3251)
σ_3, q^3	0.0825 (0.0264,0.1734)	0.0985 (0.0301,0.2371)	0.018 (0.0011,0.0477)	0.0243 (0.0008,0.0702)	0.3678 (0.2444,0.4746)	0.3671 (0.2421,0.4751)
σ_3, q^4	0.0788 (0.0266,0.163)	0.0969 (0.0325,0.2362)	0.0178 (0.001,0.0463)	0.0222 (0.0007,0.067)	0.3623 (0.2466,0.476)	0.3615 (0.2486,0.4739)
σ_4, q^1	0.0819 (0.0696,0.1037)	0.0904 (0.0757,0.119)	0.0013 (0.0,0.0248)	0.0047 (0.0,0.0329)	0.0003 (0.0,0.0052)	0.0011 (0.0,0.0075)
σ_4, q^2	0.1132 (0.0534,0.2359)	0.1339 (0.0576,0.2925)	0.0185 (0.0007,0.0564)	0.0265 (0.0008,0.087)	0.1993 (0.0765,0.3287)	0.1982 (0.077,0.3257)
σ_4, q^3	0.1338 (0.0557,0.2678)	0.1596 (0.0597,0.3734)	0.0214 (0.0009,0.0621)	0.0321 (0.001,0.1119)	0.3687 (0.2419,0.4823)	0.3669 (0.2378,0.4817)
σ_4, q^4	0.1317 (0.0571,0.263)	0.1556 (0.0613,0.3541)	0.0211 (0.0008,0.0667)	0.0315 (0.0011,0.108)	0.3641 (0.2599,0.4823)	0.3625 (0.2564,0.4822)

Values of α, errors of the number of 0-returns due to rounding, and the fraction of data set as missing values. The first column indicates a model. 95% CI is presented below each averaged statistic. v_1 stands for using Sig_1; v_2 stands for using Sig_2.

6.5.1 Alphabets with 3 and 4 symbols

The Shannon entropy is computed over a finite alphabet. To measure Shannon's entropy, we need to keep the length of blocks of symbols, k, sufficiently large. A predictable behavior of returns can be seen on blocks of greater length and may not be noticeable on blocks of smaller length. For this reason, we consider 3-symbols and 4-symbols discretizations using empirical quantiles:

$$s_t^{(3)} = \begin{cases} 1, r_t \leq \theta_1, \\ 0, \theta_1 < r_t \leq \theta_2, \\ 2, \theta_2 < r_t, \end{cases} \qquad s_t^{(4)} = \begin{cases} 0, r_t \leq Q_1, \\ 1, Q_1 < r_t \leq Q_2, \\ 2, Q_2 < r_t \leq Q_3, \\ 3, Q_3 < r_t, \end{cases}$$

where θ_1 and θ_2 are tertiles and Q_1, Q_2, and Q_3 are quartiles. The tertiles divide data into three equal parts. The quartiles divide data into four equal parts. Q_2 is also the median of the empirical distribution of returns. For the later analysis, we will need a discretization describing the behavior of a pair of stocks:

$$s_t^{(p)} = \begin{cases} 0, r_t^{(1)} \leq m_1 \text{ and } r_t^{(2)} \leq m_2, \\ 1, r_t^{(1)} \leq m_1 \text{ and } r_t^{(2)} > m_2, \\ 2, r_t^{(1)} > m_1 \text{ and } r_t^{(2)} \leq m_2, \\ 3, r_t^{(1)} > m_1 \text{ and } r_t^{(2)} > m_2, \end{cases} \tag{6.4}$$

where $r_t^{(1)}$ and $r_t^{(2)}$ are two time series of price returns, and m_1 and m_2 are their medians.

6.5.2 Efficiency rate

We determine if the value of entropy is significantly low relative to the case of perfect randomness. We detect inefficiency in the time interval using Monte Carlo simulations. We regard a Brownian motion as absolutely unpredictable. First, we define the length of sequences as $l = n_b(k) + k - 1$. Then, we simulate 10^4 realizations of Brownian motions with Gaussian increments and the length l. For each realization, we calculate entropy using 3- and 4-symbols discretizations. Then, we find the first percentile of the obtained entropies for each discretization. These percentiles are the bounds of 99% of the Confidence Interval (CI) for testing market efficiency. Finally, we define an *efficiency rate* as the ratio of the entropy of the time interval and the bound of CI. If the efficiency rate is less than 1 for at least one type of discretization; *we define the time interval as inefficient*. We provide testing for inefficiency twice using different discretizations because the unique testing may not be robust. An example where 3-symbols discretization is not enough to detect predictability of a sequence is given in Appendix 6.10.

6.6 Entropy of stock prices

We calculate $18 \cdot 120 = 2160$ efficiency rates for each type of discretization, where 18 is the amount of stocks and 120 is the amount of months in 10 years. We define a degree of inefficiency as the fraction of 2160 months that are defined as inefficient according to Section 6.5.2. The degree of inefficiency for the chosen group of stocks traded at the Moscow Exchange is 0.823. In Section 5.5.1, we find that the degree of inefficiency for the U.S. ETF market is about 0.11

Table 6.4: The degree of inefficiency for each stock.

Ticker	Degree of inefficiency	For 3 symbols only	For 4 symbols only
GAZP	0.725	0.392	0.675
LKOH	0.65	0.342	0.542
ROSN	0.742	0.392	0.708
SNGS	0.725	0.4	0.625
TATN	0.617	0.392	0.525
SBER	0.725	0.433	0.658
VTBR	0.842	0.592	0.792
CHMF	0.858	0.55	0.692
NLMK	0.8	0.467	0.692
GMKN	0.733	0.475	0.608
MTLR	0.992	0.783	0.975
MAGN	0.833	0.65	0.758
MTSS	0.967	0.7	0.942
RTKM	0.942	0.683	0.908
HYDR	0.892	0.75	0.8
RSTI	0.917	0.742	0.875
AFLT	0.983	0.775	0.95
MGNT	0.842	0.667	0.742

Fraction of inefficient months using 3-symbols and 4-symbols discretization.

for monthly time intervals and the 3-symbols discretization only. This difference in the degrees of inefficiency can be explained by the hypothesis that developed markets have a high level of efficiency. This hypothesis is confirmed by empirical analysis done in several works including [Risso, 2009, Zunino et al., 2008, Rizvi et al., 2014, Kim and Shamsuddin, 2008]. The degree of inefficiency for each stock and discretization is presented in Table 6.4. We notice that the 4-symbols discretization contributes to a larger amount of inefficient months compared to the 3-symbols discretization. That is, the 4-symbols discretization appears to have a more predictable structure than the 3-symbols discretization.

Figure 6.1 shows the minimum value of efficiency rates among all months for each stock.

There are two most notable deviations from one for MLTR stocks (Mechel, mining and metals company) and RSTI (Rosseti, power company). We investigate them in the next section. For the other 16 stocks, the minimum value of efficiency rate is attained for the AFLT stock, and it is equal to 0.933 (0.964) for three (four) symbols.

6.6.1 Analysis of stocks MLTR and RSTI

We plot the values of efficiency rates for monthly intervals for the MLTR and RSTI stocks in Figure 6.2.

Both types of discretization show coherent results. For MLTR, there are two notable decreases in the efficiency rates at the beginning of 2014 and in the middle of 2016. For both types of discretizations, the eight months with the lowest efficiency rate (in the ascending order of time) are January–February and May–October of 2014. For each month, we write down the most frequent block of symbols in Table 6.5. Note that block 1111 for the 4-symbols discretization appears as the most frequent for 6 months out of 8 for MLTR. The block denotes a slight decrease

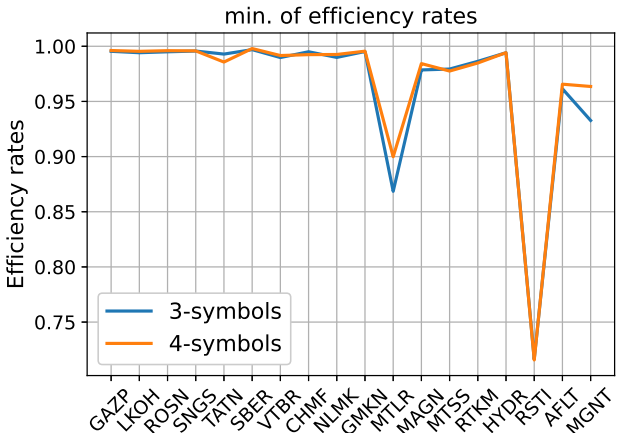


Figure 6.1: Minimum of efficiency rates for 18 stocks using 3- and 4-symbols discretizations.

in price for 4 minutes in a row. The meaning of the last two columns of Table 6.5 is discussed later.

For RSTI, there are two sharp decreases in 2014 and 2015. There are 11 months that have the lowest efficiency rates that are in common for both discretizations. These months are April–September of 2014 and June–October of 2015. Note that these inefficient months cluster together and are not distributed uniformly among the entire time period of 10 years. This is the signal of a condition inside or outside the market that affects the inefficiency of the stocks for more than one month.

6.6.2 Simple trading strategy

We construct a simple trading strategy on discretized returns to test the predictability of future returns. We consider blocks of length 4 obtained by the 4-symbols discretization. For each month, we divide blocks into two halves. The discretization is made using only the first half of a month. We consider the sequences of the first three symbols of each block. If the empirical probability of obtaining 0 or 1 after the sequence of three symbols in the first half is greater than 0.5 , this sequence is from group D (decreasing). If the empirical probability of obtaining 2 or 3 after the sequence of three symbols is greater than 0.5, this sequence is from group I (increasing). Then, for the second half of the month, we determine a success if symbols 0 or 1 follow a sequence from group D or if symbols 2 or 3 follow a sequence from group I. Then, we calculate the fraction of successes. Thus, it is the probability of making a profit: sell after group D or buy after group I. For example, we expect that after 111, the next symbol would be 1 according to Table 6.5. That is, after this block, a trader can sell a stock. In the case of market efficiency, this probability is equal to 0.5. The fourth column of the Table 6.5 shows the probabilities of success for a filtered return time series. The fifth column stands for the original return time series without filtering out data regularities.

For all cases, the probability is greater than 0.5. Obviously, the probabilities for the original return time series are greater than for the filtered return time series. The reason is that predictability for the original return time series follows from the data regularities.

The same analysis is performed for the RSTI stock. Eleven months with the lowest efficiency rates are presented in Table 6.6. For the RSTI stock, the simple trading strategy provides the

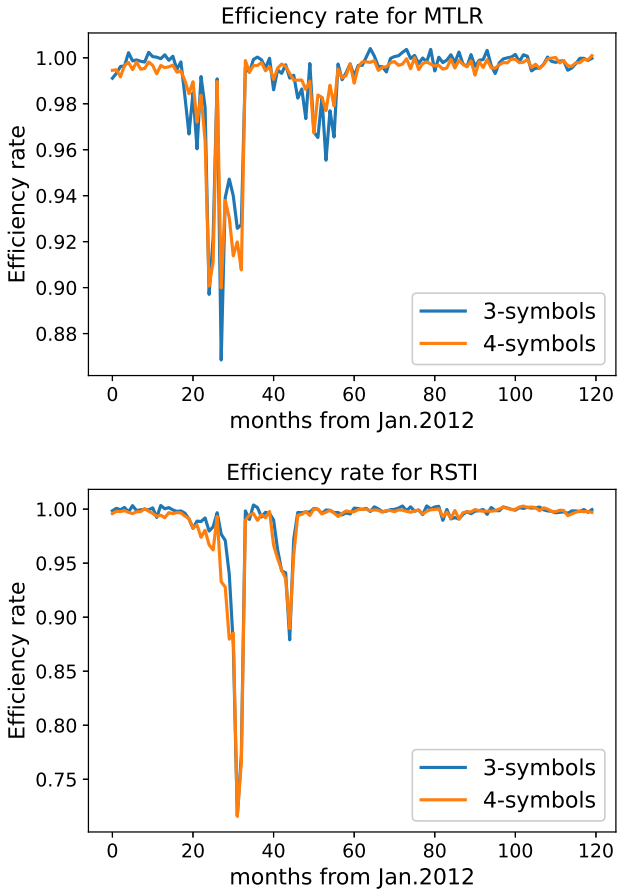


Figure 6.2: Efficiency rate for the MLTR and RSTI stocks using 3- and 4-symbols discretizations.

fraction of successes (of predicting increases and decreases in price) greater than 0.5 for all 11 months. The frequent behavior of the price of RSTI during the chosen months is a slight increase in price for several minutes in a row denoted by symbol 2.

The simple trading strategy is an illustrative example of market inefficiency. In fact, such a strategy could result in no profit when used in practice because it does not take into account the costs of transaction and other trading frictions. Moreover, the filtering of daily seasonality pattern is made by using the entire period of analysis. That is, this method can not be applied in real time. Finally, we consider blocks containing only observed returns by neglecting the missing values from the analysis. Thus, the application of such a strategy in practice should be integrated with the case when a missing value follows a sequence of three symbols.

6.7 Stock Market Clustering

Most of the month-long time intervals are identified as inefficient. However, is there some dependence between two stocks that are inefficient at the same time? Here, we are interested if there is a common behavior for a group of stocks that may cause market inefficiency.

Table 6.5: The most frequent blocks appearing for the Stock MLTR and probabilities of success of the simple trading strategy.

Months of 2014	The most frequent block, 3-s	The most frequent block, 4-s	prob. of success, filtered	prob. of success, original
Jan.	00000	1111	0.64	0.75
Feb.	00000	2222	0.64	0.74
May	00000	1111	0.61	0.73
June	22222	1111	0.60	0.73
July	11111	1111	0.62	0.74
Aug.	00000	1111	0.61	0.76
Sep.	00000	1111	0.63	0.74
Oct.	120120	0303	0.55	0.6

The first column represents months with the lowest efficiency rates. Columns 2 and 3 are the most frequent blocks in 3- and 4-symbols discretization. Columns 4 and 5 are the probability of the success of the simple trading strategy for filtered and original price returns.

6.7.1 Kullback–Leibler distance

In addition to estimating the entropy of one time series, we can also consider the difference between two time series. Kullback–Leibler divergence [Kullback and Leibler, 1951] is used to measure similarity between two distributions for two discrete probability distributions $\mathbb{P}$ and $\mathbb{L}$.

$$KL(\mathbb{P}|\mathbb{L}) = \sum_i \hat{p}_i \log \frac{\hat{p}_i}{\hat{l}_i} \tag{6.5}$$

We use empirical probabilities $\hat{p}_i$ and $\hat{l}_i$ defined in Eq. 2.2. Since the Kullback–Leibler divergence is asymmetric, we consider the distance between two time series proposed in [Benedetto et al., 2002].

$$D(\mathbb{P}, \mathbb{L}) = \frac{KL(\mathbb{P}|\mathbb{L})}{\hat{H}^G(\mathbb{P})} + \frac{KL(\mathbb{L}|\mathbb{P})}{\hat{H}^G(\mathbb{L})} \tag{6.6}$$

where $\hat{H}^G(\mathbb{P})$ and $\hat{H}^G(\mathbb{L})$ are estimations of entropy from Eq. 2.4 associated with empirical probabilities $\mathbb{P}$ and $\mathbb{L}$. The greater the distance of $D(\mathbb{P}, \mathbb{L})$, the more probability distributions $\mathbb{P}$ and $\mathbb{L}$ differ.

6.7.2 Clustering by Kullback–Leibler distance

We measure the similarity of discretized filtered returns by using the Kullback–Leibler (KL) distance (Equation 6.6). We use k, the length of blocks, as the maximum value suitable for both sequences according to Equation 6.3. The 4-symbols discretization is used. Using the Kullback–Leibler distance for all pairs of stocks, we cluster them in three groups using hierarchical clustering with the UPGMA algorithm [Sokal and Michener, 1958]. The result is in Figure 6.3. Combining companies into one cluster means that their stocks have a common behavior that is not related to the value of volatility, the degree of price staleness, and the structure of microstructure noise.

It can be seen that banks and oil companies are clustered together (right cluster in Fig. 6.3). There is a group of four stocks (RTKM, HYDR, AFLT, and MGNT) that have nothing in common at first glance. The remaining group (left part of Fig. 6.3) mainly consists of metallurgy companies. However, there is no visible distinction between the stocks of banks and oil companies.

Table 6.6: The most frequent blocks appearing for the Stock RSTI and probabilities of success of the simple trading strategy.

Months	The most frequent block, 3-s	The most frequent block, 4-s	prob. of success, filtered	prob. of success, original
Apr.2014	212121	0111	0.63	0.77
May.2014	00000	1111	0.61	0.73
June.2014	00000	1111	0.6	0.73
July.2014	00000	2222	0.62	0.74
Aug.2014	00000	2222	0.61	0.76
Sep.2014	000000	22222	0.63	0.74
June.2015	00000	2222	0.54	0.61
July.2015	00000	1111	0.55	0.6
Aug.2015	00000	2222	0.54	0.6
Sep.2015	00000	2222	0.55	0.61
Oct.2015	11111	0111	0.56	0.62

The first column represents months with the lowest efficiency rates. Columns 2 and 3 are the most frequent blocks in 3- and 4-symbols discretization. Columns 4 and 5 are the probability of the success of the simple trading strategy for filtered and original price returns.

According to the clustering tree, two telecommunications companies differ significantly, as well as electricity companies.

Finally, two stocks with the lowest efficiency rates, RSTI and MLTR, are the furthest (in the sense of KL distance) from any other stock. That is, there are no stocks that behave similarly to these two stocks.

6.7.3 Entropy of co-movement

Now, we consider another measure of difference between two stocks: the entropy of co-movement. We calculate the Shannon entropy of the discretized time series describing the movement of a pair of prices presented in Equation 6.4. We consider only minutes that are in common for both stocks. For these minutes, we consider values of residuals obtained after ARMA fitting. The result is in Figure 6.4.

Two companies related to telecommunications are a separate cluster. Three metallurgy companies (MAGN, CHMF, and NLMK) also cluster together. Stocks relating to oil and bank companies form the other cluster. The same cluster, with the exception of the TATN (oil industry), was also formed in the previous section. The "closeness" of stocks GAZP and SBER is detected either in this and in the previous section. The three stocks on the left that join other stock clusters last are the stocks with the lowest values of efficiency rates as can be seen from Fig. 6.1.

Some clusters may form on the basis that companies belong to the same industry. The division of companies into industries is noticeable from the dendrogram in Figure 6.4. However, this criterion does not explain all clusters. For instance, GMKN from metallurgy is in the cluster of oil companies and banks.

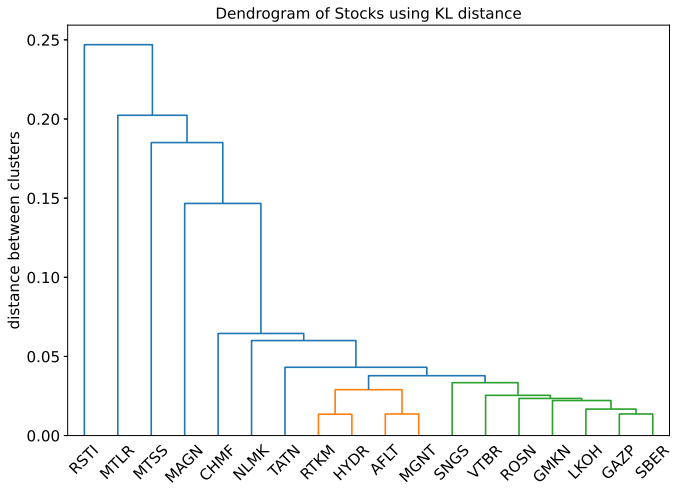


Figure 6.3: Hierarchical clustering tree using KL distance. The threshold for clustering into groups is 0.035.

6.7.4 Co-movement divergence

Using the entropy of co-movement, we conclude that price co-movements can be partially explained by belonging companies to the same industry such as metallurgy or telecommunications. In this section, we present entropy of time series that describes behavior of two prices and avoid a drawback of the entropy of co-movement. Entropy of discretization described in this section satisfies the metric axiom: the entropy calculated for a price with itself is equal to zero. This is achieved with the following discretization.

$$s_t^{(d)} = \begin{cases} 0, r_t^{(1)} \leq m_1 \text{ and } r_t^{(2)} \leq m_2, \\ 0, r_t^{(1)} > m_1 \text{ and } r_t^{(2)} > m_2, \\ 1, r_t^{(1)} \leq m_1 \text{ and } r_t^{(2)} > m_2, \\ 1, r_t^{(1)} > m_1 \text{ and } r_t^{(2)} \leq m_2, \end{cases}$$

We call the Shannon entropy of obtained discretization co-movement divergence, since the entropy is non-negative and symmetric. The results of clustering with the usage of co-movement divergence is given in Figure 6.5.

We observe the same three clusters as in the Section 6.7.3. The minimum distance is again found between the stocks of Sberbank and Gazprom. With the threshold for dividing dendrogram into clusters equal to 0.991, we detect telecommunication cluster (MTSS, RTKM), and metallurgy cluster (MAGN, CHMF, NLMK). Three stocks that are the furthest from discussed clusters are MLTR, RSTI, and AFLT as in the previous section.

6.8 Discussion on efficiency of the Moscow Stock Exchange

We have investigated the predictability of the Moscow Stock Exchange. We are interested in a measure of market inefficiency that is not related to known sources of regularity in financial time series. Usually, these sources are not filtered out, and accordingly, their impact is taken

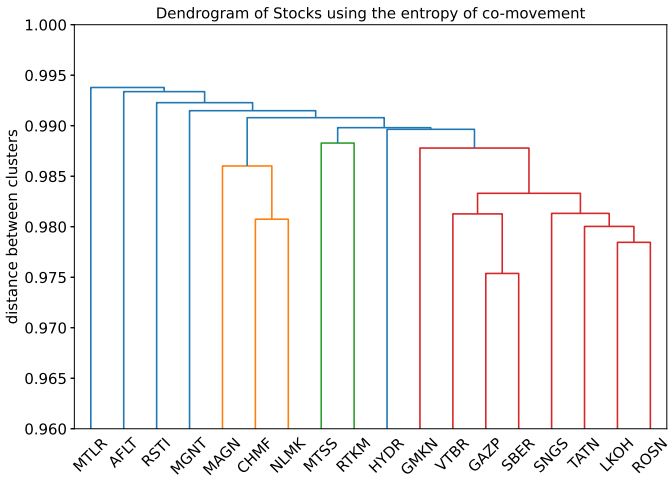


Figure 6.4: Hierarchical clustering tree using the entropy of co-movement. The threshold for clustering into groups is 0.989.

into account in the degree of price predictability (see, e.g., [Molgedey and Ebeling, 2000, Mensi et al., 2012, Risso, 2008]).

We have focused on two sources of regularity, namely volatility clustering and price staleness. The process of filtering volatility clustering was performed in [Calcagnile et al., 2020] by estimating volatility using the exponentially weighted moving average. Filtering out heteroskedasticity by using the Exponential GARCH model was applied in [Hsieh, 1991]. We have developed a modification of the volatility estimation by taking into consideration the effect of price staleness. Price staleness produces spurious 0-returns that affect the estimation of volatility. Another approach of estimating volatility in the case of presence 0-returns was proposed in [Sucarrat and Grønneberg, 2020] where all 0-returns are reevaluated during an expectation-maximization algorithm. In our approach, we separate 0-returns that may have resulted from rounding and from price staleness. Thus, we also filter out data regularity due to price staleness. Our approach combining the estimates of volatility and the degree of staleness can be used for a real-time analysis, since only past observations are used.

One of the clear advantage of the proposed approach relies in its simplicity: There is only one smoothing parameter in the method that can be optimized using historical data. We fix the value of smoothing parameter equaled 0.05. In the literature, the smoothing parameter α is usually taken close to 0. Using the principle of the best one-step forecasting, the smoothing parameter was set to 0.06 for the daily data and to 0.03 for the monthly data [Morgan et al., 1996]. The value of the parameter α was set to be equal 0.12 for in-sample testing and 0.22 for out-of-sample testing in [Bollen, 2015]. [Hunter, 1986] suggested using $\alpha = 0.2 \pm 0.1$.

We use the Shannon entropy as a measure of randomness to infer the degree of inefficiency of the Moscow Stock Exchange. We use two types of the discretization of return time series to test efficiency more reliably for each month. The 4-symbols discretization helps to find more price movements that lead to market inefficiency than the 3-symbols discretization. Approximately 82% of months over the period from 2012 to 2021 are defined as inefficient. According to [Risso, 2009], a higher level of efficiency corresponds to more developed markets. [Zunino et al., 2008] and [Rizvi et al., 2014] came to the same conclusion. Deviation from efficiency is a frequent phenomenon in various markets. For example, the authors of work [Giglio et al., 2008] concluded

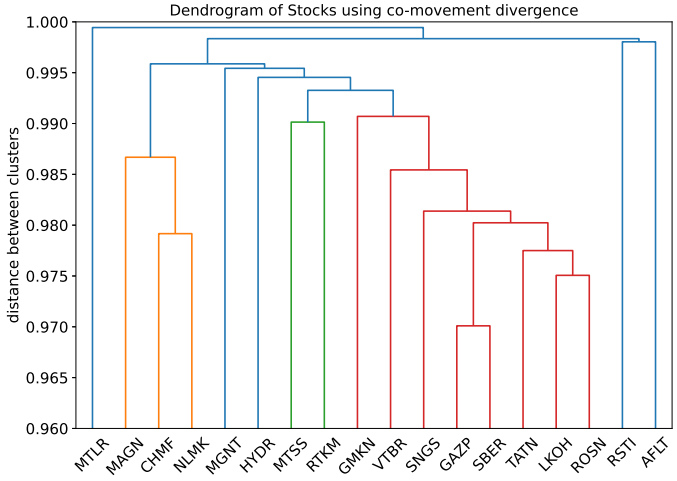


Figure 6.5: Hierarchical clustering tree using the co-movement divergence. The threshold for clustering into groups is 0.991.

that the Colombo Stock Exchange is only 10.5% efficient while the Pakistan Stock Exchange is 23.7% efficient. [Cajueiro and Tabak, 2004] showed that Asian markets are less efficient than Latin American markets. The authors of [Mahmoud et al., 2014] estimated the efficiency of the Tunisian stock market as 97%. There are periods of inefficiency for some stocks traded at the Tel-Aviv stock exchange as stated in [Shmilovici et al., 2003]. Short periods of inefficiency were also detected for U.S. stock markets in [Alvarez-Ramirez and Rodriguez, 2021].

By investigating the discretized values of filtered price returns, we come to the following conclusions:

- Even after filtering out all known sources of regularity, most months contain signals of market inefficiency.
- The most inefficient months are grouped together for two stocks exhibiting the lowest efficiency rates.
- For such months, discretized price returns before and after filtering out data regularities are predictable.
- We have introduced the entropy of co-movement and co-movement divergence for stock clustering. Stock prices display common patterns that have an interpretation in terms of the sector to which stocks belong to.
- The stocks of banks and oil companies cluster together in terms of co-movement of prices in the case of the Moscow stock exchange.

The proposed method for measuring market efficiency using the Shannon entropy can be applied in other markets of different countries. In this study, we use monthly time intervals for entropy calculation. Our work in the next chapter is related to the optimization of the length of return time series. Moreover, in the next chapter we solve the problem of finding a significant decrease in entropy without using Monte Carlo simulations. We also switch to a higher frequency (less than one minute) to analyze the predictability of financial time series in the last chapter.

6.9 Appendix: data cleaning and whitening

This section is devoted to data handling process applied before estimating a degree of market inefficiency. A detailed review of the process of filtered out data regularities is presented in Chapter 3. The procedure below is used for the data presented in this chapter and also is applied in the next chapter.

6.9.1 Outliers and splits

We use the method for outlier detection introduced in [Brownlees and Gallo, 2006]. The algorithm finds price values that are too far from the mean in relation to the standard deviation. The algorithm deletes a price P_i if

$$|P_i - \bar{P}_i(k)| \geq cS_i(k) + \gamma,$$

where $\bar{P}_i(k)$ and $S_i(k)$ are, respectively, a δ-trimmed sample mean and the standard deviation of the k price recorded closest to time i. The $\delta\%$ of the lowest and the $\delta\%$ of the highest observations are discarded when the mean and standard deviation are calculated from the sample. The parameters are $k = 20, \delta = 5, c = 5, \gamma = 0.05$.

Then, we check condition $|r| > 0.2$ in the return series to detect unadjusted splits. A split is a change in the number of company's shares and in the price of the single share such that a market capitalization does not change. There are no unadjusted splits found.

6.9.2 Intraday volatility pattern

The volatility of intraday returns has periodic behavior. The volatility is higher near the opening and the closing of the market. It shows a U-shaped profile every day. The intraday volatility pattern from the return series is filtered by using the following model. We define deseasonalized returns as follows:

$$\tilde{R}_{d,t} = \frac{\bar{R}_{d,t}}{\xi_t},$$

where

$$\xi_t = \frac{1}{N_{days}} \sum_d \frac{|\bar{R}_{d,t}|}{s_d},$$

$\bar{R}_{d,t}$ is the raw return of day d and intraday time t, s_d is the standard deviation of absolute returns of day d, and N_{days} is the number of days in the sample.

6.9.3 Heteroskedasticity and price staleness

Different days have different levels of the deviation of the deseasonalized returns $\tilde{R}$. We define the standardized returns as

$$r_t = \frac{\tilde{R}_t}{\hat{\sigma}_t}.$$

If a transaction cost is high, the price is updated less frequently, even if trading volume is not zero. This effect is called price staleness [Bandi et al., 2020] and is discussed in Section 6.3.2. We identify 0-returns appearing due to rounding (and not due to price staleness) using Equation 4.3. Other 0-returns are set as missing values. In order to remove this heteroskedasticity and price staleness, we estimate the volatility $\hat{\sigma}_t$ and a degree of price staleness in Section 6.3.3.

6.9.4 Microstructure noise

The last step in filtering data regularities is filtering out microstructure noise. The microstructure effects are caused by transaction costs and price rounding. We consider the residuals of an ARMA(P,Q) model of the standardized returns after filtering out 0-returns. We apply the method introduced in [Jones, 1980] to find the residuals of an ARMA(P,Q) model by using the Kalman filter. We select the values of P and Q that minimize the value of BIC [Schwarz, 1978] such that $P + Q \leq 6$. The values of P and Q are chosen for each calendar year and are used for the next year. For the year 2012, we select $P = 0$ and $Q = 1$ corresponding to a MA(1) model.

6.10 Appendix: Predictable time series with entropy at maximum

The goal of this section is to construct a price model where entropy is high because of discretization. This model shows that a high entropy value may be caused by discretization, but not because of the randomness of a return time series.

There are equal probabilities of having symbols 0, 1, and 2. Symbol 1 corresponds to log-returns, r, equal to -0.4, and 2 corresponds to log-returns equal to 0.4. The structure of symbol 0 is more complicated. It covers three other symbols: $3, 4$, and 5. They correspond to log-returns $-0.3, 0.1$, and 0.2, respectively. One of the symbols $3, 4$, or 5 appears with probabilities depending on the previous value of these symbols. The probabilities are presented in the Table 6.7. Having a symbol presented in a column, there are probabilities of obtaining a symbol presented in a row.

Table 6.7: Transition probabilities

first symbol	$\cdot 3$	$\cdot 4$	$\cdot 5$
$3\cdot$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$
$4\cdot$	$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{3}$
$5\cdot$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{6}$

Rows stand for the first symbol of a block, columns stand for the second symbol.

The model implies an average zero return. However, a trading strategy that increases a profit exists. After 3, a trader should buy, and after 4 and 5 the trader should sell. However, the entropy of a 3-symbols series is at its maximum, which should imply an absence of profitable strategies.

Considering the same example with the 4-symbols discretization, we obtain $Q_1 = -0.4$,

$Q_2 = 0.1$, and $Q_3 = 0.4$. Therefore, we have the following discretization of returns.

$$s_t^{(4)} = \begin{cases} 0, r_t = -0.4, \\ 1, r_t = -0.3 \text{ or } r_t = 0.1, \\ 2, r_t = 0.2, \\ 3, r_t = 0.4. \end{cases}$$

Thus, we can distinguish returns $r = 0.2$ from the others using the 4-symbols discretization. Table 6.7 provides the following probabilities for the blocks of two symbols and from the 4-symbols discretization: $p(11) = \frac{7}{162}$, $p(12) = p(21) = \frac{5}{162}$, $p(22) = \frac{1}{162}$. Noting that $p(0) = p(3) = \frac{1}{3}$, $p(1) = \frac{2}{9}$, and $p(2) = \frac{1}{9}$, we calculate that

$$H_1 = -\frac{2}{3}\log\left(\frac{1}{3}\right) - \frac{2}{9}\log\left(\frac{2}{9}\right) - \frac{1}{9}\log\left(\frac{1}{9}\right) \approx 0.946 < 1$$

and

$$H_2 = -\frac{1}{2}(\frac{7}{162}\log\left(\frac{7}{162}\right) + \frac{5}{81}\log\left(\frac{5}{162}\right) + \frac{1}{162}\log\left(\frac{1}{162}\right) + \\ + \frac{4}{9}\log\left(\frac{1}{9}\right) + \frac{8}{27}\log\left(\frac{2}{27}\right) + \frac{4}{27}\log\left(\frac{1}{27}\right)) \approx 0.944 < H_1$$

Chapter 7

Statistical test for changes in entropy value

The goal of this chapter is to propose a methodology to identify significant changes in the value of the Shannon entropy. We move away from using Monte Carlo simulations to determine if the value of entropy is significantly low. Instead, we use a statistical test that allows to detect low entropy values in a faster way. We introduce a rigorous procedure to test the hypothesis that the entropy value associated with two different sequences (defined on the same finite alphabet) is statistically equivalent. To this end, we need to solve two problems. The first problem is to find the variance of the Shannon entropy's estimator obtained by the Empirical Frequencies method. We obtain the approximation of the variance in Theorem 8. The second problem is to find the optimal length of a rolling window used to estimate a time-varying entropy of a time series. We introduce a novel self-consistent criterion to select the optimal bandwidth w. Given a sample of size N, we first define a counting function of the percentage of entropy variations for non-overlapping time series of length w within such sample. Under the assumption of a finite number of *true* entropy variations, the percentage of estimated variations becomes negligible when w is close to the minimum ($w = 1$), while it is zero by definition when w attains its maximum ($w = N$). Then, we show by simulations that such counting function has one maximum corresponding to the optimal bandwidth.

Section 7.2 presents the test of equality of entropies. We propose a method for choosing the length of a sliding window in Section 7.3. We test the hypothesis about equal entropies on simulated data in Section 7.4. For instance, we determine the power and size of the statistical test. Also, we show that we are able to determine the length of interval with a different value of entropy using the novel approach. Our interest regarding application part is to investigate the case of meme stocks, that we discuss in Section 7.5. The hypothesis is tested for the price returns of meme stocks and IT stocks in Section 7.6. We show periods of inefficiency for both types of stocks. Moreover, we notice that in 2020 and 2021 meme stocks were more predictable than IT stocks. We also note that the method for determining optimal bandwidth can be considered as a test for stationarity. More precisely, we show that the price returns of the stock CRM were stationary in 2019 in sense of the constant entropy while other stocks exhibit time-varying behavior of entropy.

The results of this chapter are also available in our research paper [Shternshis and Mazzarisi, 2022].

7.1 Introduction

The Shannon entropy is widely used as a measure of randomness in many fields, such as finance, physics, medicine, and biology. The Shannon entropy of protein sequences was calculated in [Strait and Dewey, 1996, Cristadoro et al., 2023]. Entropy was used in analyzing electroencephalogram signals [Dong et al., 2019, Bezerianos et al., 2003, Liang et al., 2015]. Entropy was used to analyze geoelectrical signals [Telesca et al., 2014]. The Shannon entropy was used for testing homogeneity of galaxy distributions in [Pandey and Sarkar, 2015]. The approximate entropy was introduced for the applications in medical data [Pincus et al., 1991] and later was used for financial time series [Pincus and Kalman, 2004]. For the example of time-varying entropy in financial markets during crises we refer to [Stosic et al., 2016].

One of the main applications of entropy estimation in finance is to measure the randomness of price returns. When the price incorporates all relevant information, the market is called efficient and the price dynamics is a martingale [Fama, 1970, Samuelson, 1965]. As such, the Shannon entropy takes the maximum allowed value and can then be used as a measure of market efficiency. For the review of methods for testing the martingale property and nonlinear dynamics in financial time series, we refer to the article [Barnett and Serletis, 2000]. The drivers of market dynamics are however the result of the complex process of matching the supply and the demand of a large number of investors. It is easy to imagine that the market does not necessarily reflect all relevant information at certain times, because of the complex nature of the price formation mechanism. Moreover, feedback loops, irrational agents, market panic and speculation, or coordination of retail investors driven by non-economic reasons (like for the case of the Gamestop [Mancini et al., 2022]) are just a few examples of mechanisms which can potentially create booms and busts [Agliari et al., 2018, Chan and Santi, 2021], thus driving the price far away from its fundamental value. In such cases, the price dynamics may display some level of predictability and, as such, the market is inefficient. This can be captured by some low value of the Shannon entropy.

In a number of works, the hypothesis of market efficiency is relaxed, by accounting for the possibility of periods of inefficiencies. In order to capture such an effect, Shannon entropy is considered as time-varying and it is computed by using a window rolling over the period of interest, see, e.g., [Molgedey and Ebeling, 2000, Mensi et al., 2012, Risso, 2008, Olbryś and Majewska, 2022]. It is important to notice that many patterns of price dynamics may jeopardize the estimate of entropy. For example, long memory of volatility can be the result of a regime switching behavior [Susmel, 2000, Lobo and Tufte, 1998, Malik et al., 2005] and, as a consequence, the estimate of any other dynamic pattern can be affected. For this reason, it is important to filter out any known pattern of market predictability, such as heteroscedasticity or seasonality, before using the estimate of entropy as measure of market efficiency. Interestingly, even after such filtering, the price dynamics often continues to display some predictability that is captured by low values of entropy estimates, signaling a not complete efficiency of the market [Calcagnile et al., 2020, Shternshis et al., 2022b]. A crucial question is whether a drop in the entropy estimate at some period is statistically significant or it is only a fluctuation consistent with the hypothesis of market efficiency.

The variance of the Shannon entropy is the key quantity to use in order to determine if two entropy estimates associated with two different sequences are statistically equal. In fact, it is possible to define a z-score given the variance of the Shannon entropy, thus determining the corresponding p-value of statistically equal estimates. [Basharin, 1959] obtained the first order approximation of the variance of the entropy estimator $\hat{H}$ calculated using the Empirical

Frequencies method.

$$D_1(\hat{H}) = \frac{1}{n}(\sum_j p_j \ln^2 p_j - H^2) \tag{7.1}$$

where p_j are the set of the probabilities of possible events and n is the length of the sequence of events. The same result was later obtained in [Dávalos et al., 2019]. However, Eq. 7.1 holds in the asymptotic regime, that is when the length of the sequence becomes arbitrarily large, i.e., $n \to \infty$. Here, we aim to estimate the time-varying entropy for finite samples, i.e., using a finite length n of the sequence. Moreover, Eq. 7.1 is not a consistent estimator of the variance in the case of equal probabilities. When all probabilities p_j are equal, $D_1(\hat{H}) = 0$. Thus, a more accurate approximation for the variance is needed. We have obtained a formula (2.6) for the variance of the estimator of the Shannon entropy as a sum of central moments of binomial and multinomial distributions. The central moments are calculated by a new recursive approach. This helps us to find the approximation of the variance with an accuracy of order $O(n^{-4})$. In general, it is possible to further extend such an approximation by using the proposed approach to compute higher orders of the central moments associated with the multinomial distribution. Interestingly, [Ricci et al., 2021] found that the naive estimator of the variance approximation $D_1(\hat{H})$ in Eq. 7.1 has a bias term of order $O(n^{-2})$. On the contrary, we show in Theorem 9 that our proposed estimation (Eq. 2.14) for the variance of entropy is unbiased.

By leveraging on the explicit formulation of the variance of the estimator of entropy, we are able to define a statistical test for entropy variation. In [Matilla-García, 2007], the author suggested a statistical test for independence of symbolic dynamics. The test statistics is related to the permutation entropy. In our research, we are not restricted to the case when the benchmark value of the entropy is its maximum. A rejection of the null hypothesis signals statistically significant variation between any two possible values of entropy.

Another research problem relies on finding the optimal length, i.e., the bandwidth, of the time window when testing for entropy variation between two subsequent time series. This is a problem of bias-variance tradeoff: reducing the length of the window allows to obtain a timely estimate of entropy, i.e., small bias, at the expense of increasing the variance of the estimator, and vice versa.

We use the novel methodology to find significant changes in entropy on simulated and real data. We investigate changes in the efficiency of the New York Stock Exchange with a special focus on meme stocks. The lower the entropy of the price return time series, the higher the price predictability. In particular, we examine the GameStop case, whose price increased significantly in January 2021. We find that significant drops of Shannon entropy can be interpreted as early-warning signals of market turmoil.

Variations in value of entropy identify changes in predictability level of prices. The predictability of prices, in turn, means a violation of the martingale property of prices. If a price follows not a martingale, but a strict local martingale it is called bubble [Cox and Hobson, 2005]. That is, the current price of a bubble exceeds the expected discounted future price under the risk-neutral measure. Several approaches for detecting bubbles were proposed in the literature. For instance, the approaches are based on using options data [Fusari et al., 2020] and volatility estimations [Jarrow et al., 2011]. In the current chapter, we propose the method for determining violation of martingale hypothesis using two entropy estimations of price returns.

7.2 Statistical test for equal entropies

Using the estimation of entropy, we aim to conclude if two entropy values significantly differ. Let's take two sequences with entropies H_1 and H_2. Let $\hat{H}_1$ and $\hat{H}_2$ be estimations of entropies

with variances of estimations Var_1 and Var_2, respectively.

$$\mathcal{H}_0 : H_1 = H_2$$
$$\mathcal{H}_a : H_1 \neq H_2$$

Let assume that $H_1 = H_2$. Then, a z-score

$$z = \frac{\hat{H}_2 - \hat{H}_1}{\sqrt{Var_2 + Var_1}} \tag{7.2}$$

is distributed with a zero mean and the variance equal to 1. We reject $\mathcal{H}_0$ if $|z|$ is larger than a quantile corresponding to 99% of confidence. The quantile is defined empirically in Section 7.4.1.

7.3 Determining optimal bandwidth

We assume that a time series into consideration can have time-varying entropy, thus we need to choose a *bandwidth* w, that is the length of a rolling window where entropy is estimated. We aim to detect significant changes in entropy, thus we can not take w too large so that the window covers several intervals with different entropy values. On the other hand, if the process at some period is stationary, we aim to take w as large as possible to improve the accuracy of the entropy estimation[12]. In terms of the bias-variance trade-off, if the rolling window covers a period where the process is stationary (with constant entropy), the error from a bias is eliminated. However, taking w too small implies a large variance and thus it may be impossible to distinguish a change in entropy from estimation errors. Thus, our goal is to find an optimal parameter w.

We introduce the following approach to determine the optimal bandwidth. We test for all adjacent non-overlapping intervals. Then, we choose a bandwidth that allows us to detect the maximum z-score. More precisely, we aim to maximize[13] an objective function $f(w)$ below.

$$f(w) = \begin{cases} \max(|z(w)|), \text{ if } \%\{|z(w)| > q_{99}\} > 1\% \\ -\frac{1}{w}, \text{ otherwise} \end{cases} \tag{7.3}$$

where q_{99} is a 99% quantile for the empirical distribution of z. The intuition for the objective function is that small values of w give small values of the z-score since the variance, $\hat{Var}$, is $O(w^{-1})$. Large values of w may not be able to catch the largest change in entropy value. Finally, if we can not reject the hypothesis about constant entropy, we aim to choose w as large as possible, thus we maximize $-\frac{1}{w}$. Since we are interested in detecting changes in entropy, $\max(|z|)$ is always non-negative and $-\frac{1}{w}$ is always negative. We give more intuition in Section 7.4.3 by plotting Fig. 7.3.

To make at least one test, we set the upper bound for the bandwidth as $n_{max} = \lfloor \frac{n}{2} \rfloor$, where n is the total length of the sequence. We may define the optimal bandwidth using a training set. To set the lower bound we refer to Theorem 9. In this theorem, we assume that all different events appear in the sequence, thus M is known as the number of all different events. We discuss this assumption in the Remark below.

[12]The Equations (2.6) and (2.13) show that the larger the length of a sequence, the less the variance and the downward bias of the entropy estimation.

[13]To find the maximum, we use the function scipy.optimize.minimize_scalar with method=bounded and xatol= 1 in Python v.3.9.5.

Remark 2. *Number of appearing different events, $\hat{M}$, is defined as $\hat{M} = M - \sum_{j=0}^{M-1} I\{\hat{p}_j = 0\}$, where I is an indicator function. Thus,*

$$E[\hat{M}] = M - \sum_{j=0}^{M-1} (1 - p_j)^n.$$

The error term $\sum_{j=0}^{M-1}(1 - p_j)^n$ attains its minimum when all $p_j = \frac{1}{M}$. It grows as a probability approaches 0; that is, when there is an event that can happen with a tiny but not zero probability, so that we may not observe it in a finite sequence of events. Such an event does not greatly affect the entropy value, since $p \ln p \to 0$ as $p \to 0$. We need to fix the minimum length of a sequence to assume that all events appear in the sequence. To this aim, we introduce the following rule. The length of the sequence is taken such that the minimum error is less than 0.01. That is, one event with a non-zero probability does not appear in 1 case out of 100.

$$\begin{aligned}
\sum_{j=0}^{M-1} (1 - p_j)^n &< 0.01 \\
M(1 - \frac{1}{M})^n &< 0.01 \\
n &> \frac{\ln \frac{0.01}{M}}{\ln \frac{M-1}{M}} \\
n_{min} &= \lceil \frac{\ln \frac{0.01}{M}}{\ln \frac{M-1}{M}} \rceil
\end{aligned}$$

7.4 Simulation study

Before applying the proposed statistical test to real data, we first simulate random symbolic sequences. In Section 7.4.1, we find the quantile using in a testing procedure and check if this quantile is admissible for short sequences and multiple testings. In Section 7.4.2, we investigate the power and size of the statistical test. In Section 7.4.3, we apply our method to a non-stationary process. For the further analysis of simulated and real data, we fix $A = \{0, 1, 2, 3\}$.

7.4.1 Empirical quantile

Determining quantile of entropy's distribution

In this section, we find the quantiles of the distribution of the z-score from Eq. 7.2. Intuitively, the larger M (the number of different events), the more the entropy estimator (as the sum of random variables) tends to be normally distributed. [Basharin, 1959] showed that the distribution of the entropy estimator is asymptotically normal. However, if all probabilities are equal, the distribution of the scaled entropy estimation converges to χ^2-squared distribution [Zubkov, 1974]. That is, quantiles of normal distribution can be used if the entropy is not near the maximum. In case when entropy is close to the possible maximum, the normal distribution for testing equality of two entropies is inappropriate. To show this, we set large values for the length of a sequence and the length of blocks. The difference between two χ^2-distributions we are interested in is no longer a χ^2-distribution. The density function of the difference takes a complex form

and is derived in [Mathai, 1993]. Making a statistical test based on χ^2-distribution of entropy estimation without subtraction is postponed for the last chapter.

We perform 2×10^4 Monte Carlo simulations. We simulate two sequences with length $n = 2 \times 10^5$ of 4 symbols with equal probabilities. We set $k = 7$ so that $M = 4^7 = 16384$. When probabilities are equal, the expression for the variance (Eq. 2.6) becomes

$$Var(\hat{H}_{max}) = \frac{M-1}{2(n-k+1)^2} + \frac{M^2-1}{6(n-k+1)^3}$$

We use this value as given instead of estimating the variance from the sequence. Then, for two sequences we find the value of $\Delta\hat{H} = \hat{H}_2 - \hat{H}_1$. Since $\Delta\hat{H}$ is the difference of two independent variables, its variance is $2Var(\hat{H})$. An empirical 99% (95%) quantile of the empirical distribution of $|z|$ is 3.30722 (2.54542). Thus, even if the length of a sequence and the amount of blocks are quite large, the tails of the empirical distribution are thicker than for the normal distribution. We use the empirical quantiles for further analysis. If an absolute value of z-score is larger than the quantile, the difference between entropy values is statistically significant.

Testing short independent sequences

For the further analysis, we fix $k = 4$. Therefore, the maximum of the k-th order entropy is $k \ln |A| = 4 \ln 4$, $M = 256$. Now, we aim to test the obtained empirical quantiles for shorter sequences. Here, we set $n = 2 \times 10^3$ and make 2×10^4 simulations. Using quantiles from the previous subsection, the false positive rate is 1.025% (4.86%).[14] We consider these results as acceptable for keeping the found values of quantiles for the rest of the chapter.

Testing a fully random sequence

Here, we simulate one sequence with four equiprobable symbols and length $N = 2 \times 10^7$ and divide it by overlapping intervals with the length $n = 2 \times 10^3$. We consider all differences with the gap equal to n, so that there are no common blocks between two intervals. The false positive rate according to the 99% (95%) empirical quantile is 0.9976% (4.4608%), that is quite close to the significance level. Thus, q_{99} from Eq. 7.3 is taken as 3.30722.

7.4.2 Power and size of the test

We consider a process with the 4-symbols alphabet constructed as follows. The probability of repeating a symbol is τ. All other probabilities are equal, that is, the probability of having, for example, 0 after 1 is $\frac{1-\tau}{3}$. Thus, the value of the 4th order entropy is given by

$$H(\tau) = -2\left(\tau \ln \frac{\tau}{4} + (1-\tau) \ln \frac{1-\tau}{12}\right) \tag{7.4}$$

The further τ from $\frac{1}{4}$, the lower the entropy. $\tau = \frac{1}{4}$ corresponds to equiprobable symbols and the maximum of entropy. First, we construct two sequences with $H(\frac{1}{4})$ and length $n = 10000$. We estimate the variance of entropy's estimator using Equation 2.14 and perform hypothesis testing using Equation 7.2. The false positive rate (size) is 0.86% obtained by generating two sequences 2×10^4 times. Now, we simulate one sequence with $H(\frac{1}{4}) \approx 5.54518$ and the other with a different entropy value to measure the power of the test. The results are in Table 7.1. Clearly, the larger the difference between entropies, the larger the power of the test. Even when $\tau = 0.31$, we detect changes in entropies in 100% cases.

[14] Empirical quantiles from this experiment are equal to 3.31684 and 2.52907, respectively.

Table 7.1: The power of the hypothesis testing with different τ.

τ	$H(\tau)$	Power (%)
0.28	5.5405	56.28
0.29	5.53692	94.556
0.3	5.53237	99.915
0.31	5.52688	100

Power is the probability to reject $\mathcal{H}_0$ given that the null hypothesis is false. The results are obtained with 2×10^4 Monte Carlo simulations.

7.4.3 Non-stationary process

We are interested in detecting significant changes in the value of entropy in a non-stationary process. We consider a sequence with 4 symbols with a parameter that changes the value of entropy in the middle of the sequence. The length of a sequence is $N = 30000$. We set $k = 4$, thus $M = 256$. For this sequence, $n_{min} = \lceil \ln \frac{0.01}{M} / \ln \frac{M-1}{M} \rceil = 2594$ and $n_{max} = \lfloor \frac{N-k+1}{2} \rfloor = 14998$. We divide the sequence into three parts as shown in Fig. 7.1. The first and the last parts are simulated as the sequences with the maximum entropy and four equiprobable symbols. The first part has the length of 10000. The part with the length l in the middle has entropy $H(\tau)$ given by Eq. 7.4.

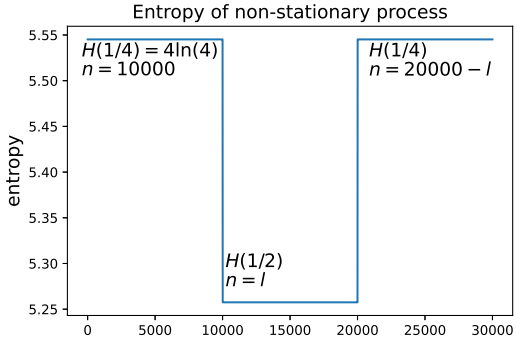


Figure 7.1: Illustration of stepwise entropy.

The variance of the entropy's estimator becomes time-varying, thus we estimate it using Equation 2.14. We apply the method for determining an *optimal bandwidth*, w_{opt}, from Section 7.3. First, we fix $\tau = 0.5$ and take l from n_{min} to 10000 with the increment of 100. For each l, we calculate the optimal bandwidth and plot it in Fig. 7.2. The figure shows that the optimal bandwidth corresponds to the length of the interval in the middle. Identifying properly the length of interval with the different entropy allows us to estimate entropy in this period in the most accurate way.

Figure 7.3 shows the plot of the objective function (Eq. 7.3) for 4 random iterations with different values of l. All plots have one global maximum. The larger l, the larger the argument of the maxima corresponding to w_{opt}. When $l = 10000$ and w is close to n_{max}, the percent of statistically significant changes in entropy is less than 1%, thus $f(w)$ becomes negative. If the process generating the sequence is stationary ($\tau = 0.25$), the range of the objective function may be negative. The objective function of one realization of the stationary process is given in

Fig. 7.4. Since the function is monotonically increasing, the maximum is attained at n_{max}.

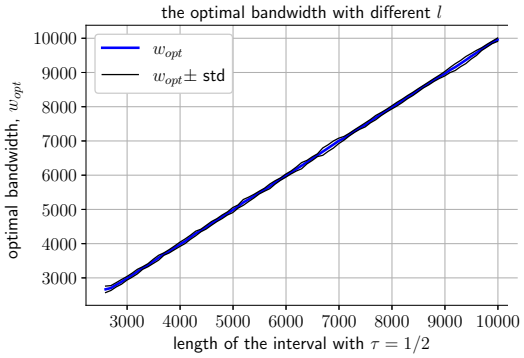


Figure 7.2: Optimal bandwidth for different values of l. The mean and standard deviation (std) are calculated over 100 iterations.

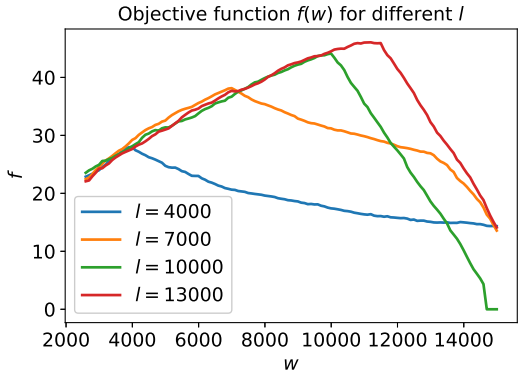


Figure 7.3: Objective function for different values of l.

Finally, we fix $l = 10000$ and take τ from 0.25 to 0.5 with the increment of 0.01. For each τ, we calculate the optimal bandwidth and plot it in Fig. 7.5. $\tau = 0.25$ corresponds to the case of a stationary process and thus w is close to the possible maximum. This corresponds to the intuition of choosing the bandwidth: if the process is stationary, the larger the bandwidth, the more accurate the entropy estimate. The deviation of w_{opt} when $\tau = 0.25$ is high because of the first type error in hypothesis testing. Even if the process is stationary, $\mathcal{H}_0$ may be rejected more than in 1% of cases that leads to a positive value of $f(w)$. Such a large deviation can be reduced using Bonferroni or Šidák corrections [Šidák, 1967] if the distribution of the z-score is known. We note that when the entropy in the middle slightly differs from the maximum ($\tau = 0.26$), the method may not detect the interval with different entropy and set w_{opt} close to the maximum. The larger τ, the closer w_{opt} to the length of the interval in the middle and the less the standard deviation of w_{opt}.

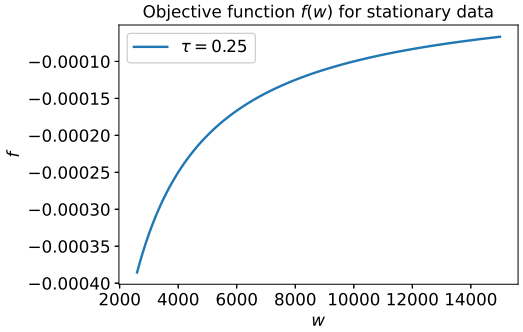


Figure 7.4: Objective function for one realization of stationary process.

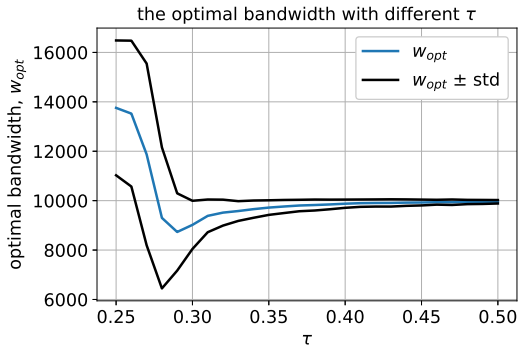


Figure 7.5: Optimal bandwidth for different values of τ. The mean and standard deviation (std) are calculated over 400 iterations.

7.5 Dataset: meme and IT stocks

In early 2021, the eyes of everyone were on the New York Stock Exchange because of one of the most terrific surges in prices in its history. Starting from January 2021, GameStop and multiple other stocks that were heavily shorted experienced a dramatic increase in their share price and volume. Upward movement of prices and high trading activity were driven by the long trades of a huge number of individual investors, who fomented such a coordinated action on the Reddit social platform. When the prices were hitting their all-time highs, the attention of everyone was focused on such shares, which became known as "meme stocks". Then, as the end of January 2021 approached, several retail broker-dealers temporarily limited certain operations in some of these stocks and options. As a consequence, the trend started to revert, even if such a turmoil period has had a permanent impact on GameStop and the other meme stocks. The stocks have maintained a higher price level with respect to the period before and have been characterized by higher volatility. A precise description of the GameStop case can be found in the report by the staff of the U.S. Securities and Exchange Commission[15]. GameStop is a videogame retailer. The performance of the company declined because of the the shift of video game sales to online platforms. In 2020, the share price fell below one dollar. GameStop's price began to increase noticeably on January 13, when the closing price rose to $31.40 from $19.95. By January 27,

[15]See https://www.sec.gov/files/staff-report-equity-options-market-struction-conditions-early-2021.pdf

GameStop stock closed at a high of $347.51 per share. The following day, stock prices jumped further to an intraday high of $483.00.

There are two crucial aspects that have led to the sharp increase in meme stock prices. First, the aggregation of the orders sent to broker-dealers via online trading platforms, such as Robinhood, in the hands of a few off-exchange market makers permitted to negotiate good agreements, which have then resulted in incentives or no fee at all for the end customers. Such an absence of trading frictions driven by FinTech innovations has had a positive feedback effect on retail trading. The second and more important aspect is the coordination in long trades by a huge number of individual investors, which has been possible because of online social networks like Reddit. The coordination has shaped up to be an act of rebellion against short-selling professional investors who had allegedly targeted the meme stocks. The two effects combined have created a first sharp increase in the prices, which has been further amplified by the coverage of the short positions by professional investors. The increasing interest in buying meme stocks and the resulting feedback dynamics have then led the intermediaries to extraordinary operations. Without entering into the discussion about the manipulative nature of such trading behavior coordinated via social networks and the interpretation of the GameStop rally as a revenge against Wall Street for the 2007-2008 crisis, it is evident that the case of meme stocks has represented a period when the market was out of the ordinary. Here, the focus is on market (in)efficiency during such a period. In particular, we find that the clear predictability of the price pattern results in a signal of inefficiency in terms of the Shannon entropy and statistically significant variations of the Shannon entropy have an interpretation as early-warnings.

We consider three meme stocks that became popular in late 2020 and 2021. We take a one-minute frequency from 9:00 to 15:59. We investigate stocks of the companies GameStop (GME), Bed Bath & Beyond (BBBY), and AMC Entertainment Holdings (AMC). In addition, we consider three well-known IT companies, expecting more persistent entropy for their stocks.[16] These companies are Apple (AAPL), Salesforce (CRM), and Microsoft (MSFT). Using return time series, we define a 4-symbols alphabet as follows.

$$s_t^{(4)} = \begin{cases} 0, r_t \leq Q_1, \\ 1, Q_1 < r_t \leq Q_2, \\ 2, Q_2 < r_t \leq Q_3, \\ 3, Q_3 < r_t, \end{cases} \tag{7.5}$$

where Q_1, Q_2, Q_3 are quartiles of the empirical distribution of returns r_t. Here, r_t are price returns time series after filtering out data regularities: intraday volatility pattern, heteroskedasticity, price staleness, and microstructure noise that we discuss in Section 6.9. Price returns before filtering out data regularities are defined as $\bar{R}_t = \ln \frac{P_t}{P_{t-1}}$, where P_t is a price at time t. Data regularities are empirical properties of price returns [Cont, 2001] that make the price returns time series more predictable but do not imply any profitable trading strategy. For instance, intraday volatility pattern refers to the fact that the volatility of intraday returns has periodic behavior. It is higher near the opening and the closing of the market [Wood et al., 1985], because traders tend to trade less in the middle of a day.

7.6 Empirical application: the case of meme stocks

A market in which prices always fully reflect available information is called efficient [Fama, 1970]. In a weak form of the efficient market hypothesis, the information set is historical prices. Thus,

[16] We use a proprietary intraday financial time series dataset provided by `http://www.kibot.com`.

if a market is efficient in the weak form, a future price can not be predicted better than its current value. Full uncertainty about future values of prices implies the maximum degree of randomness of the price returns time series. Thus, the entropy of return time series should attain its maximum if the market is efficient. If a market is efficient and the entropy of the price returns is always at the maximum, the hypothesis $\mathcal{H}_0$ should be failed to reject. The rejection of $\mathcal{H}_0$ for two non-overlapping intervals implies that the entropy is time-varying. Low values of entropy relatively to entropy estimated in the past indicate predictability of price returns.

7.6.1 Meme stocks

Here, we consider stocks GME, BBBY, and AMC. We set the year 2019 as a training set. The period from 01.01.2020 to 20.07.2021 is a testing set. First, we filter out the data regularities. We define an intraday volatility pattern, fit an autoregressive moving average (ARMA) model, and find an optimal bandwidth, w_{opt}, using the training set. Volatility and the degree of price staleness are defined minute by minute. The ARMA model is needed to filter out microstructure noise; volatility estimation is needed to filter out heteroskedasticity. Quartiles Q_1, Q_2, Q_3 used for discretization in Eq. 7.5 are defined using the return time series of the training set after filtering out the data regularities.

We calculate the Shannon entropy of the discretized sequence $s_t^{(4)}$ of the testing set using the rolling window with length w_{opt}. Comparing entropies of two adjacent intervals, we have three possible outputs: entropy decreases, entropy increases, entropy does not change significantly. We present results for each stock in Table 7.2.

We plot the entropy estimation, price, and trading volumes in Figure 7.6 for the stock GME. Figures 7.7 and 7.8 show estimated entropies, prices, and trading volumes for the stocks BBBY and AMC, respectively. We mark on the plots where entropy has significant changes. Red dots correspond to statistically significant decrease in entropy. Green dots stands for statistically significant increase in entropy. The dots at the same time are plotted in figures for the prices and trading volumes. We report below when entropy starts to decrease first time or after an increase. We note that the entropy had been already defined as statistically significant low for the stock AMC when record volumes occurred and the price rose sharply for the first time in January 2021.

For GME, there are two series of decreases. They start from the entropies calculated at the periods [17.07.20 15:46 to 14.09.20 15:34] and **[7.12.20 15:38 to 19.01.21 09:42]**. For BBBY, there are three series of decreases. They start from the entropies calculated at the periods [06.03.20 13:50 to 05.05.20 09:54], **[10.12.20 11:47 to 28.01.21 10:59]**, and [06.05.21 15:16 to 21.06.21 14:45]. For AMC, there are four series of decreases. They start from the entropies calculated at the periods [05.03.20 15:13 to 17.04.20 12:47], [10.08.20 13:08 to 08.09.20 11:27], [08.03.21 14:09 to 30.03.21 10:26], and **[05.05.21 14:46 to 27.05.21 15:10]**. The time intervals highlighted in bold correspond to sharp increases in price values.

7.6.2 IT stocks

We take the stocks of Apple, Salesforce, and Microsoft for the comparison. Calculated entropies are plotted in Fig. 7.9; the optimal bandwidths are recorded in Table 7.2. The entropy of the price returns of these stocks exhibits time-varying behavior. For each stock, statistically significant changes in entropy are detected. However, the entropy does not fall as much as in the case of meme stocks. The price returns time series for the stock CRM in 2019 is defined as stationary since $f(w_{opt})$ is negative.

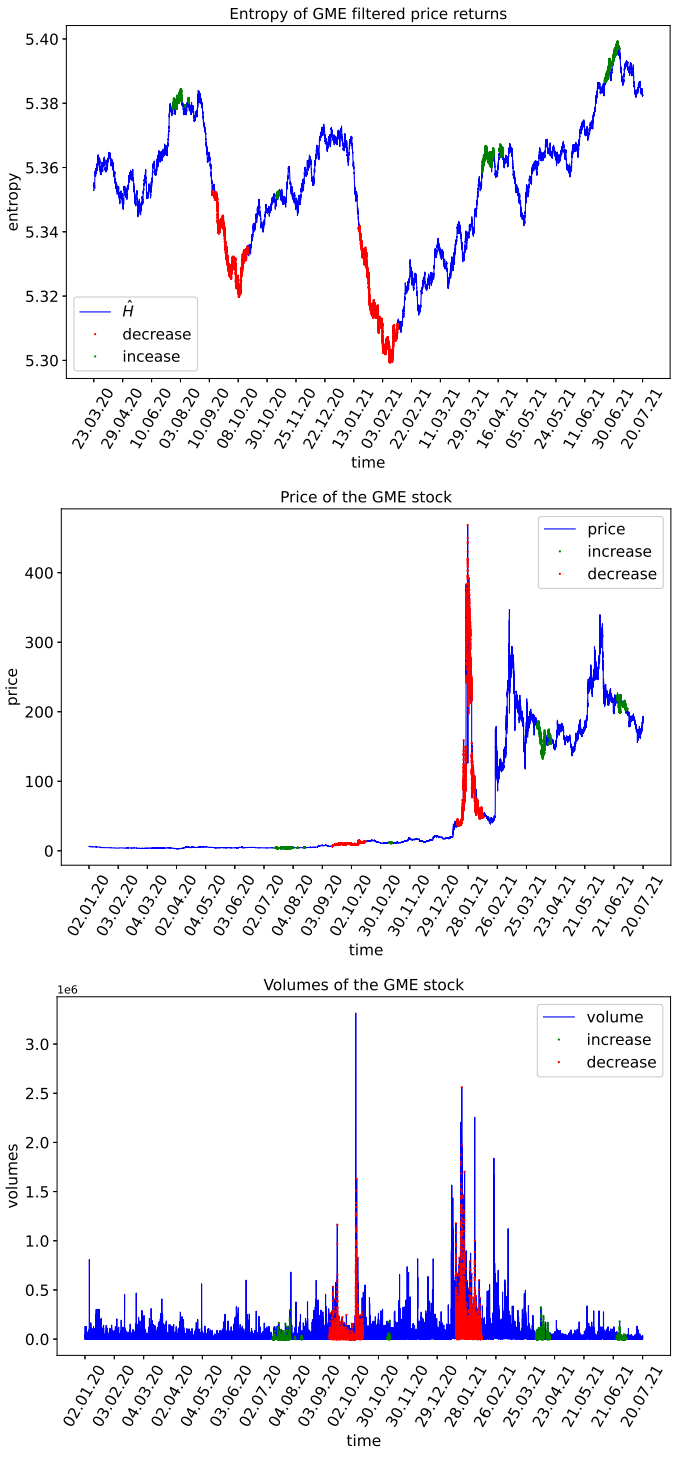


Figure 7.6: Entropy, price, volume of the GME stock. Dots correspond to statistically significant changes in entropy.

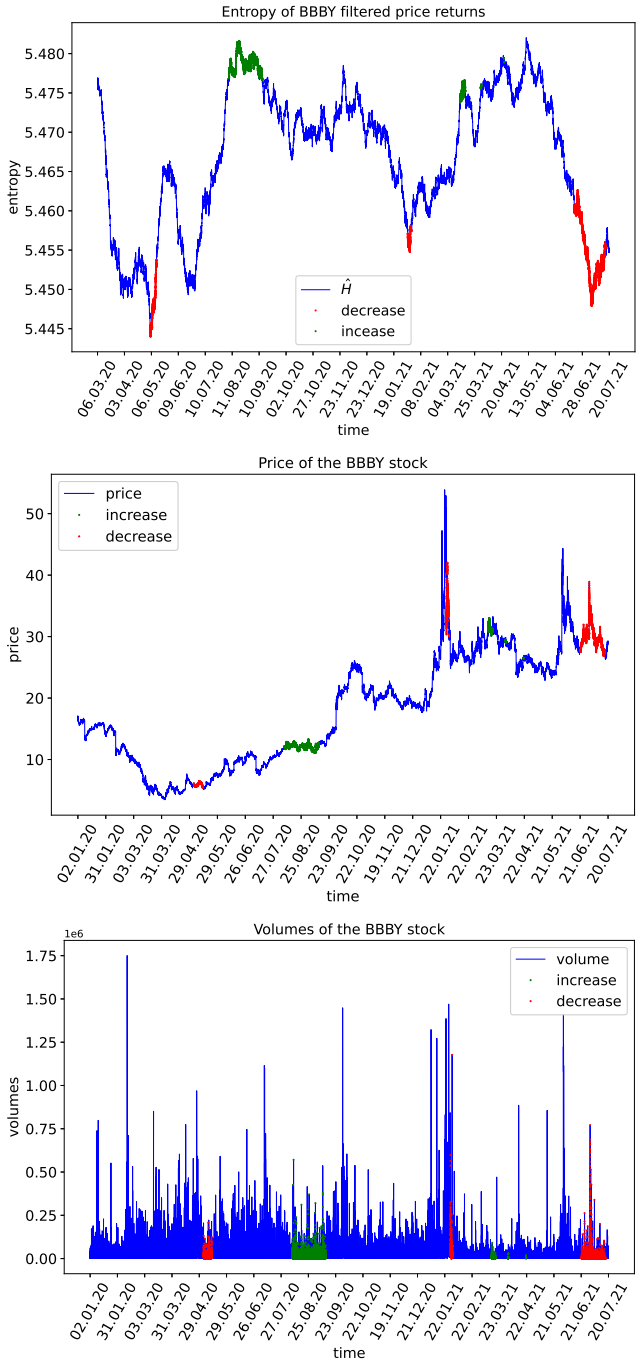


Figure 7.7: Entropy, price, volume of the BBBY stock. Dots correspond to statistically significant changes in entropy.

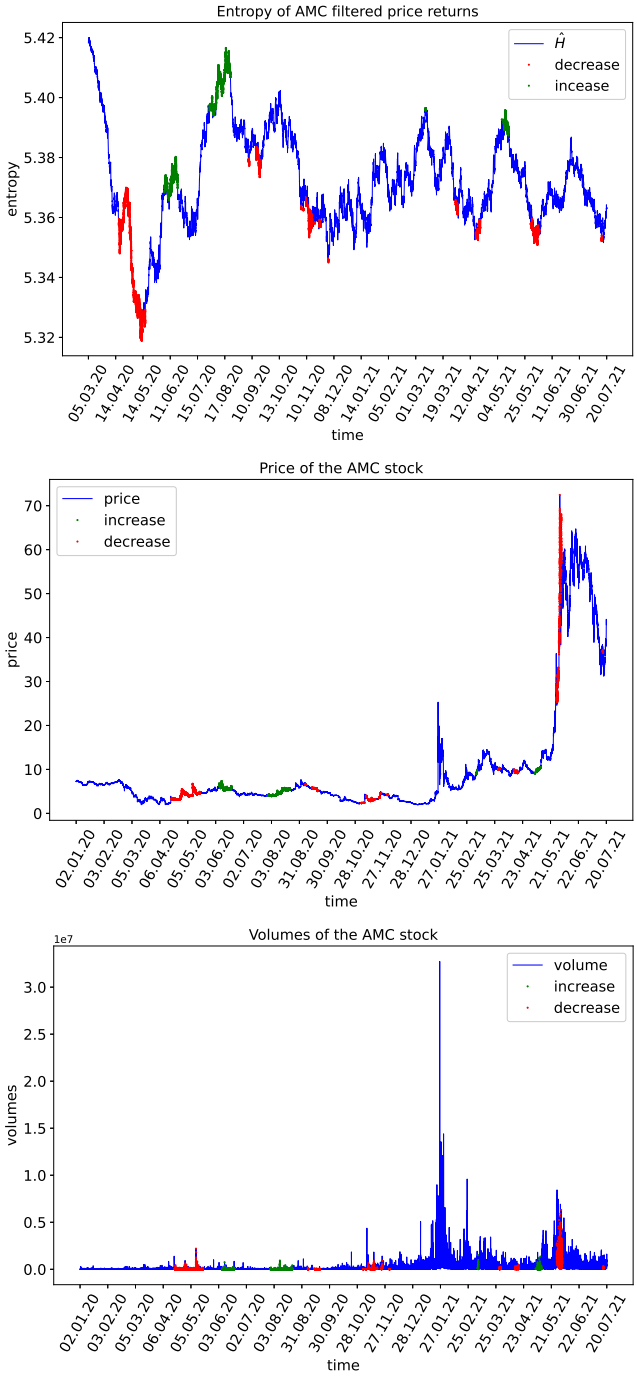
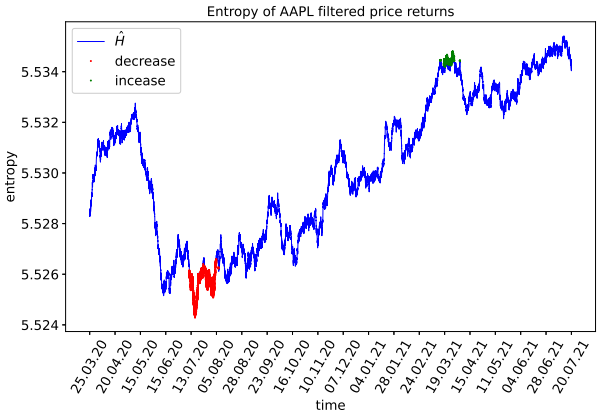
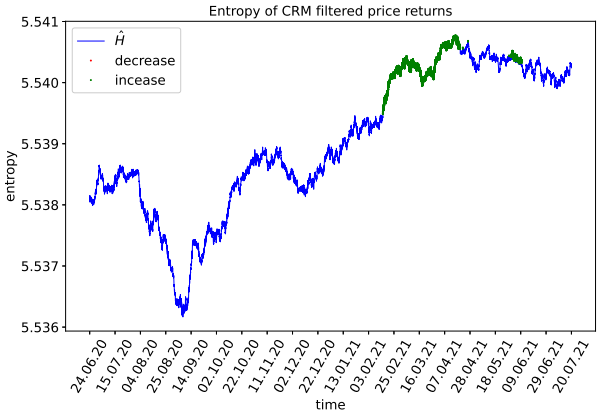


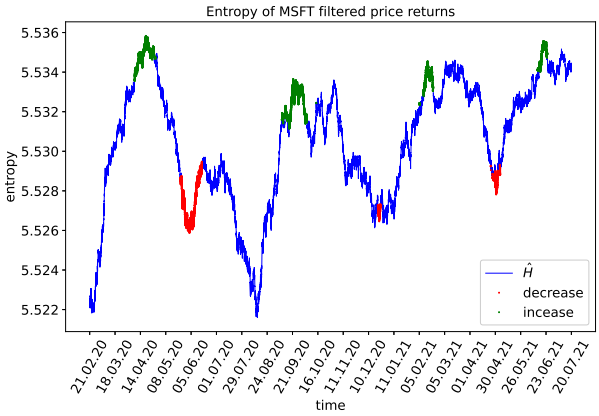
Figure 7.8: Entropy, price, volume of the AMC stock. Dots correspond to statistically significant changes in entropy.



(a) AAPL



(b) CRM



(c) MSFT

Figure 7.9: Entropy of IT Stocks. Dots correspond to statistically significant changes in entropy.

Table 7.2: Optimal bandwidth and hypothesis testing for meme and IT stocks.

Stock	w_{opt}	$f(w_{opt})$	n_{max}	Number of tests	Number of increases	Number of decreases
GME	6707	3.881	14387	89909	5672	12468
BBBY	10218	4.975	22542	89248	7399	7831
AMC	5845	5.052	16374	93639	6372	8107
AAPL	20313	9.323	33865	104531	1495	6137
CRM	44321	$-2.256 \cdot 10^{-5}$	44322	55189	17651	0
MSFT	12539	5.438	41413	121887	15969	7271

$f(w_{opt})$ is calculated on the training set. The number of tests is the amount of adjacent time intervals with lengths w_{opt} in the testing set. The increases and decreases are statistically significant changes in entropy between two adjacent intervals.

7.6.3 Quarterly training sets

A reason for the increase in predictability, when entropy falls, may be a change in the structure of data regularities. For instance, due to the growing popularity of meme stocks, the intraday volatility pattern could change from the year 2019 to the year 2021. If the behavior of traders has changed, the intraday volatility pattern from the training set does not filter out the data regularity in the testing set. In order to filter the data regularities more carefully, we update an estimation of intraday volatility pattern and a fitted ARMA model using quarterly intervals starting from the first quarter of 2019. This quarter is also used to filter out the data regularities in-sample. An optimal bandwidth is defined using the year 2019. The results are shown in Table 7.3 and Fig. 7.10.

Table 7.3: Optimal bandwidth and hypothesis testing for stocks using quarterly training sets

Stock	w_{opt}	$f(w_{opt})$	n_{max}	Number of tests	Number of increases	Number of decreases
GME	7240	15.351	14343	89042	20179	10933
BBBY	20673	25.432	22596	64582	3789	11354
AMC	8060	14.062	16343	89855	6992	12003
AAPL	21515	24.064	33849	104809	24966	7191
CRM	27647	5.027	44287	88858	8204	2058
MSFT	35462	8.602	41378	76696	24200	11

$f(w_{opt})$ is calculated on the training set. The number of tests is the amount of adjacent time intervals with lengths w_{opt} in the testing set. The increases and decreases are statistically significant changes in entropy between two adjacent intervals.

In all six cases, entropy still exhibits time-varying behavior. Compared to the previous setting, where the training set is one year, the number of statistically significant changes in entropy for the stocks GME and AAPL increases. Also, in all cases, the maximum value of the objective function, indicating how much the entropies of two adjacent intervals differ, increases.

There are two sequences of statistically significant decreases in entropy value for the stock GME. They start from the entropies calculated at the periods [03.08.20 12:31 to 22.09.20 13:50]

and [29.12.20 11:23 to 27.01.21 13:15]. For the stock AMC[17], there are three series of low entropy values. They start from the entropies calculated at the periods [25.03.20 12:31 to 13.05.20 14:17], [30.09.20 14:21 to 13.11.20 13:31], and [05.05.21 11:58 to 04.06.21 14:51]. For both stocks, the last time intervals correspond to a sharp increase in the prices. For the stock BBBY, the entropy becomes statistically significantly low starting from the interval [07.08.20 11:10 to 18.11.20 11:16]. Thus, entropy was low at the time of the rapid growth in the price and trading volumes.

We can conclude that a possible change in the structure of data regularities over time is not the cause of the changes in the values of entropy. As in the previous section, the entropies of IT companies do not drop as low as in the case of meme stocks. Therefore, there is more predictability in the prices of meme stocks than in the prices of IT companies.

7.7 Discussion on the meme stocks

We introduce a novel procedure of hypothesis testing to determine if two discrete sequences with the same alphabet have different entropy values. We use it to detect changes in the entropy value of a sequence. Since entropy can change over time, we can not choose the length of the interval for measuring entropy as large as possible. For this reason, we have introduced two contributions. First, we have found the approximation of entropy's variance that is used in hypothesis testing. Our formula is more precise regarding the length of the interval compared to the formulas proposed by [Basharin, 1959] and [Harris, 1975]. Second, we have proposed the method for finding an optimal length of the interval. We have shown that this method is suitable for determining how to split a sample into time series that display statistically different values of Shannon entropy.

We apply the novel method to the return time series of meme stocks. We have found time intervals when entropy is statistically low with respect to other periods, thus signaling a high predictability of the price pattern. In particular, we focus on three meme stocks, then comparing them with three more standard IT companies, namely GME, BBBY, AMC, AAPL, CRM, and MSFT. We have found that entropy changes over time for all six stocks in a period between January 2020 and July 2021. We can also say that the entropy was time-varying in the year 2019 for all stocks except CRM.

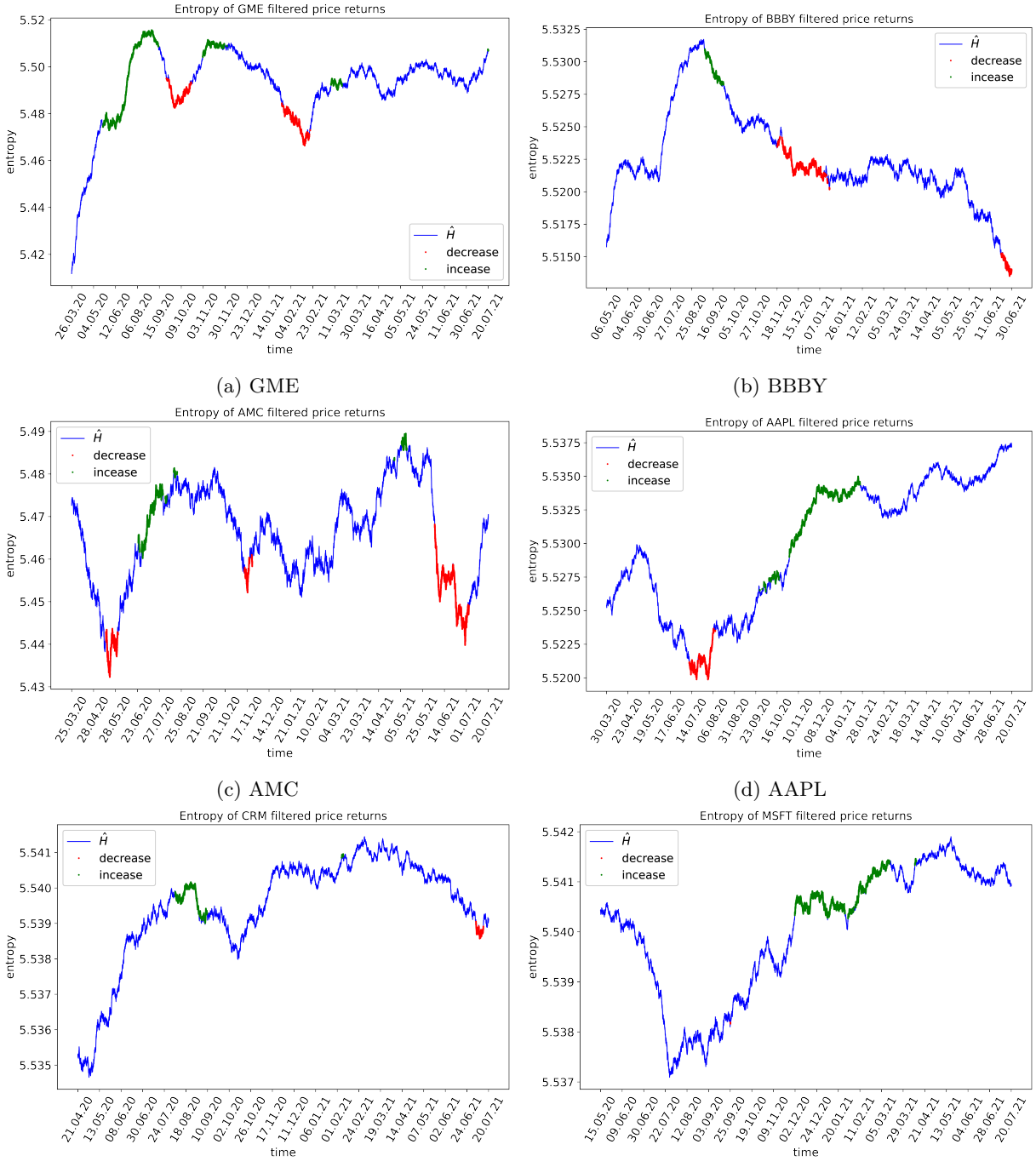
Low entropy values found for each stock are indicative of the predictability of return time series. We filter data regularities in two ways, and therefore, we believe that the entropy values obtained are related to the measure of market inefficiency. The deviation of the U.S. stock market from efficiency is also detected in other articles, e.g., [Alvarez-Ramirez and Rodriguez, 2021, Giglio et al., 2008, Molgedey and Ebeling, 2000].

From the entropy plots, we notice that entropy for meme stocks falls lower than for IT stocks. This indicates the existence of the periods of high predictability for meme stocks compared to IT stocks. For the GME stock, a low level of entropy is identified prior to the price spike. Moreover, a fall in entropy corresponds more to the growth in trading volumes than to the increase in the stock price according to the both series of low entropy values. For the AMC stock, entropy was low when the price went up in January 2021. Entropy fell again before the price attained its maximum in June 2021. For the BBBY stock, the period from December 2020 to the end of January 2021 is also characterized by a low entropy, although it is discovered late enough to be considered a price increase warning. The low entropy values found for the BBBY stock are more likely to correspond first to the price drop in February 2021 and then to another spike in June 2021.

[17]Non-zero returns in the first three quarters of 2019 are not enough to find the intraday volatility pattern for all minutes for the stock AMC. This generates missing values in the filtered return time series of the year 2019.

In the case of GameStop and AMC Entertainment Holdings, growth in prices and trading volumes is characterized by a drop in the entropy value. Interestingly, such a drop occurs before the boom observed in January 2021. That is, some regularity pattern in the price dynamics, that appeared before all the news spread the market, leads to a statistically significant signal of market inefficiency. Given the observed timing, such a signal can also be interpreted as an early-warning of turmoil period for the stock. Another indicator for tracking fragility of banking sector was proposed in [Kibritçioglu, 2003]. A range of indicators for early-warning system for currency crises was discussed in [Kaminsky et al., 1998]. The authors of [Babecký et al., 2014] considered the list of 28 indicators and test their significance for identifying banking and currency crises. Our contribution to this field is to suggest entropy as one of the possible early-warning indicators. We note that the indicator based on a decrease in entropy is obtained only using history of price returns.

Today, financial markets are inherently high-dimensional due to the plethora of instruments composing the portfolios of investors. At the same time, they are highly challenging to monitor, displaying more and more complex cycles, booms and bursts of prices, economic bubbles and so on, all of them representing severe risk factors for the portfolios. In this high-dimensional and complex context, the existence of online early-warnings of market inefficiency is key. In fact, such signals allow to anticipate the periods of turmoil, thus covering or, at least, mitigating the portfolio risk associated with such events, with potential stabilizing effects for the whole market.



(a) GME

(b) BBBY

(c) AMC

(d) AAPL

(e) CRM

(f) MSFT

Figure 7.10: Entropy of stocks with quarterly training sets. Dots correspond to statistically significant changes in entropy.

Chapter 8

Testing price predictability of ultra-high frequency data

In this chapter, we investigate the predictability of tick-by-tick data. To our best knowledge, we present the first research investigating predictability of ultra-high frequency data by estimating the Shannon entropy. We investigate empirical properties of price returns and the difference in the properties of returns for unpredictable and predictable time series. We study empirical facts about returns already discussed in this dissertation, such as non-zero autocorrelation, as well as other facts, such as price jumps.

In Section 8.2, we present a chosen group of assets with a set of characteristics such as mean price, daily trading volumes, and number of transactions. In Section 8.3, we propose a statistical test for investigating the predictability of symbols of a sequence. We compare the entropy of price returns with its possible maximum. Significantly low values of entropy are defined by an asymptotic gamma distribution. In Section 8.4, we start testing the predictability of real data from the stock of Apple Inc. We measure predictability of daily time intervals considering different lengths of blocks, aggregations levels, and months. Then, we consider the problem of optimal time interval used to detect low values of entropy. We perform multiple testing for sub-intervals of days to detect the presence of predictability in a particular time period during a day. Then, we test the reversibility of tick-by-tick data by estimating entropy production. Our research interest is to investigate if signs of price returns are reversible in time. In Section 8.5, we study what separates the unpredictable from the predictable sequences for various stocks and trading days. For inefficient days, we localize the presence of predictability inside inefficient days.

8.1 Introduction

[Engle, 2000] defined ultra-high frequency data as the full record of transactions and their associated characteristics. [Bouchaud et al., 2002] investigated statistical properties of the limit order books of several stocks. In particular, the authors stated that the distribution of changes in limit order prices exhibits a power-law tail. [Engle and Russell, 1998] noted that the longest duration between transactions appeared in the middle of trading days. According to the authors, clustering of transactions appears because of size of bid-ask spread and gathering of informed traders. U-shapes of frequencies of large trades, small trades, and market orders were also discovered in [Biais et al., 1995]. In another paper by Engle, [Engle, 2000], he showed that intraday volatility has the similar pattern and attains its minimum in the middle of a trading day. Moreover,

significant coefficients of ARMA(1,1) model were detected by [Engle, 2000]. Highly dependent microstructure noise was stated by [Robert and Rosenbaum, 2010]. [Lillo and Farmer, 2004] conducted a statistical test and concluded that signs of market orders, executed limit orders, and cancellations of limit orders are long-memory processes. One of the explanations for the long memory given in the article is an arrival of news that move the mid price, that is the middle between the best ask and best bid prices. For more details and discussion on market and limit orders, we refer to works [Biais et al., 1995, Foucault et al., 2013, Gould et al., 2013]. [Lillo and Farmer, 2004] also showed that trading volumes of transactions are a long memory process. Long memory of signs of trades were also detecting in [Bouchaud et al., 2003] by estimating the rate of correlations decay. The authors suggested that these correlations are due to dividing one order into several small parts. According to [Epps and Epps, 1976], changes in stock prices between transactions are associated with trading volumes. Some stylized facts including fat tails of price returns, volatility clustering, and the leverage effect [Bouchaud et al., 2001] were discussed in [Bouchaud, 2005].

According to [Lillo and Farmer, 2004] and [Bouchaud et al., 2003], long memory found for signs of orders does not imply inefficiency of the market. The possible explanations for the long memory consistent with market efficiency are fluctuations in market liquidity [Lillo and Farmer, 2004] and interaction between market orders and limit orders [Bouchaud et al., 2003]. Moreover, the high speed of occurrence of new orders makes it difficult to predict the next price before it appears in such a short period of time. Taking into account the fact that we do not consider transaction costs, we can not assure that the predictability we find on ultra-high frequency indicates the presence of profitable trading strategies. However, we can examine the periods in which we find price predictability. We explore what price characteristics distinguish days with predictability from others. Also, we study the duration in transaction time of periods of predictability.

This chapter presents three main contributions. First, we develop a statistical test for the predictability of a sequence. Applying the test, we get rid of Monte-Carlo simulations and thus determine significantly low values of entropy fast. We compare entropy estimation of a sequence with the possible maximum, but not with entropy of another sequence as we propose in Section 7.2. Difference between entropy estimation and its maximum follows gamma distribution as shown in Zubkov [1974] and Brouty and Garcin [2023]. Using simulations, we show that the entropy of the autoregressive process is statistically significantly low according to the gamma distribution even if the autoregressive parameter slightly differs from 0. Second, we exploit the statistical properties of the Shannon entropy's estimator to explore predictability of ultra-high frequency data. Some of the properties are stylized facts of limit orders such as long memory of signs, volatility clustering, and autocorrelation of returns. Other considered characteristics relate to trading activity measured by the number of transactions and daily volumes. Finally, we apply Šidák [1967]' correction to make several test of price predictability inside one trading day. In such a way, we localize intervals that we call inefficient and find the duration of periods of predictability. Detecting intervals with different degrees of predictability allows us to conclude that prices of executed transactions at ultra-high frequency are not stationary processes in the sense of constant entropy.

To estimate Shannon entropy, we consider the discretization that distinguishes between positive and negative returns. We show that the degree of predictability decreases when we aggregate data by a number of transactions. We find that the probability of repeating the same sign of return is one of the features which describe predictable sequences. For a group of stocks, this probability is significantly high during inefficient days, that is consistent with the long memory of the signs of returns. However, we discover that for other stocks, the probability of repeating symbols can be statistically low for inefficient days.

Another property describing inefficient days is the number of non-zero returns and trading volumes. For some stocks, we also found a high autocorrelation value inherent in inefficient days. Finally, we demonstrate that usually there is one inefficient interval with a short length inside an inefficient day. Usually, two inefficient intervals, as well as two inefficient days, do not cluster together. That is, after an interval with detected predictability, the next interval does not display a notable predictability level. However, considering all transactions of CCL stock (Carnival Corp.), we are able to detect several inefficient time intervals going in a row.

8.2 Tick-by-tick dataset

We explore limit order book data [Gould et al., 2013] downloaded from LOBSTER (`www.lobsterdata.com`). We consider the executions of visible and hidden orders of a group of assets. For the analysis, we take a set of stocks from various industries that differ in average price, volatility, number of transactions, waiting times between transactions, and trading volumes. In addition, we consider the ETF SPY that is designed to track the S&P 500 Index. We take the time period from 01.08.2022 to 21.11.2022, that is 80 trading days in total. Tickers of all chosen assets and their properties are presented in Table 8.1. For each day, the considered daily time period is from 9:30 to 16:00, that is 390 minutes in total. Times of transactions are recorded with the precision of one nanosecond.

Table 8.1: Assets and characteristics of prices

Asset	Ticker	Mean price	Standard deviation of price	Trading volume	N. of transactions	Average time between transactions
Apple Inc.	AAPL	153.47	0.93	12,184,032	136,136	0.165
Microsoft Corporation	MSFT	251.78	1.37	4,529,093	84,342	0.269
Tesla Inc.	TSLA	388.02	3.81	8,686,354	178,704	0.127
Intel Corporation	INTC	30.15	0.20	7,055,642	38,255	0.595
Eli Lilly and Company	LLY	327.33	1.73	370,050	11,404	2.086
Snap Inc.	SNAP	10.67	0.14	4,967,779	18,521	1.358
Ford Motor Company	F	13.93	0.10	4,468,175	12,954	1.815
Carnival Corporation & plc	CCL	9.24	0.12	5,874,376	15,372	1.518
SPDR S&P 500 ETF	SPY	390.52	1.56	9,136,137	95,181	0.246

Average time is given in seconds. Numeric values are averaged over days.

Considering each occurred transaction, we work with data in transaction time [Oomen, 2006]. Discretization is made by distinguishing between positive and negative returns. 0 corresponds to price decreasing, 1 corresponds to price increasing. Thus, alphabet A is $\{0, 1\}$ and a symbolic sequence is obtained according to binary discretization from Eq. 3.1. All 0-returns are removed. Removing 0-returns from discretization, we move from transaction time to tick time [Griffin and Oomen, 2008], where a price is recorded only if the price is changed and moves to another level in a discrete grid.

8.3 Statistical test for the value of entropy

We conduct a statistical test to determine the significance of an entropy value as a degree of unpredictability. We develop an idea from the previous chapter, namely, Section 7.2. Now, we construct quantiles for one entropy estimation instead of the difference between estimations. This allows us to use the properties of familiar distributions, that are gamma of normal distributions as we discuss later in this section.

Using a statistical test, we aim to determine whether a given sequence of symbolic outputs is generated from a completely random process with the maximum value of entropy. A sequence has outputs taken from a finite set of M possible events. Probabilities of appearing of outputs are denoted as $p_0 \dots p_{M-1}$. Empirical frequencies are the amount of times when each output appears in the sequence. Dividing empirical frequencies by the length of the sequence we get empirical probabilities $\hat{p}_0 \dots \hat{p}_{M-1}$. We assume that each empirical frequency has the binomial distribution and set of all frequencies has the multinomial distribution.

In order to estimate the predictability of a sequence x_1^n, we consider non-overlapping blocks of k symbols from a finite alphabet A. For symbols generated independently, considering blocks without overlapping allows us to assume that all blocks appear independently, and thus, empirical frequencies follow the multinomial distribution. Since we use non-overlapping intervals, the number of blocks is $n_b = \lfloor \frac{n}{k} \rfloor$. If all blocks have positive probability of appearing, then $M = |A|^k$, where $|A|$ is the size of alphabet A. In the previous chapters, we use overlapping intervals, so that the amount of blocks is $n - k + 1$. A larger number of blocks allows to estimate the probabilities of the appearance of blocks more accurately and hence to obtain more robust results for testing the value of entropy. The development of the statistical test using entropy estimation with overlapping blocks is a possible improvement of the approach discussed in this chapter.

8.3.1 Bias and variance of entropy estimation

We estimate the entropy of a process as shown in Equation 2.3, $\hat{H} = -\sum_{j=0}^{M-1} \hat{p}_j \ln \hat{p}_j$. If all probabilities are equal, then estimation of entropy follows gamma distribution. Otherwise, entropy estimation has the normal distribution [Zubkov, 1974]. To make a statistical test, we need to know the bias and variance of entropy estimation. All formulas are given up to the third order of n_b in Section 2.4. Under the assumption that all probabilities are equal, the formula for the variance (Eq. 2.6) simplifies as follows.

$$Var^{\Gamma}(\hat{H}) = \frac{M-1}{2n_b^2} + \frac{M^2-1}{6n_b^3} \tag{8.1}$$

This formula is used for the variance of the gamma distribution for testing if the entropy is at the maximum. To estimate the bias of entropy estimation we use Equation 2.13 that gives the expected value of entropy estimation.

$$\Delta(\hat{H}) = H - E(\hat{H}) = \frac{M-1}{2n_b} - \frac{1}{12n_b^2}\left(1 - \sum_{j=0}^{M-1}\frac{1}{p_j}\right) - \frac{1}{12n_b^3}\sum_{j=0}^{M-1}\left(\frac{1}{p_j} - \frac{1}{p_j^2}\right) \tag{8.2}$$

The expression of the bias is obtained by approximating $\hat{H}$ by Taylor sequence around $p_0 \dots p_{M-1}$, taking the expected value, and substituting central moments of the binomial distribution. A similar result was obtained in [Treves and Panzeri, 1995]. For the case when all probabilities are equal, the bias is equal to

$$\Delta^{\Gamma} = \frac{M-1}{2n_b} + \frac{M^2-1}{12n_b^2} + \frac{M^3-M^2}{12n_b^3} \tag{8.3}$$

The shape $\bar{k}$ and scale $\bar{\theta}$ of the gamma distribution $\Gamma(\bar{k}, \bar{\theta})$ are defined from the variance and bias as follows.

$$\bar{\theta} = \frac{Var^{\Gamma}}{\Delta^{\Gamma}}$$
$$\bar{k} = \frac{\Delta^{\Gamma}}{\bar{\theta}}$$

The similar but asymptotic values for the parameters of gamma distribution were recently obtained in Brouty and Garcin [2023] in a different setting of testing.

For the case when not all probabilities are equal, we use normal distribution and approximate the bias and variance using the plug-in estimations. That is, we substitute empirical probabilities into formulas for the bias and variance of entropy estimation given by Equations 2.13 and 2.6, respectively. For instance, the estimation for bias is

$$\Delta^{N} = \frac{M-1}{2n_b} - \frac{1}{12n_b^2}\left(1 - \sum_{j=0}^{M-1} \frac{1}{\hat{p}_j}\right) \tag{8.4}$$

where superscripts Γ and N in Equations 8.3 and 8.4 stand for Gamma and Normal distributions, respectively, used for determining quantiles.

Interestingly, the plug-in estimation is the unbiased estimation of bias from Equation 8.2. Indeed, we obtain Equation 8.2 by rewriting $\frac{1}{\hat{p}_j}$ from Eq. 8.4 using the Taylor expansion and substituting the central moments of binomial distribution from Eq. 2.12. We plug in empirical probabilities into the formula of the variance Eq. 2.6 when probabilities p_j, $j = 0 \dots M-1$ are assumed to be not the same and get the following formula.

$$Var^{N}(\hat{H}) = \frac{1}{n_b}\left[-\hat{H}^2 + \sum_j \hat{p}_j \ln^2(\hat{p}_j)\right] + \frac{1}{n_b^2}\left[\frac{M}{2} - \frac{1}{2}\right] + \frac{1}{6n_b^3}\left[(1-\hat{H})\sum_j \frac{1}{\hat{p}_j} - \sum_j \frac{\ln \hat{p}_j}{\hat{p}_j} - 1\right]$$

Remark 3. *The plug-in estimation of the variance is biased, however, removing the bias of the variance estimation may lead to a negative value of the variance estimation. Indeed, as stated in [Ricci et al., 2021], the plug-in estimation of the first order approximation of the variance,* $Var_1 = \frac{1}{n_b}\left[-H^2 + \sum_j p_j \ln^2(p_j)\right]$*, is normally distributed with the mean* $\frac{Var_1}{n_b} + \frac{\gamma}{n_b^2}$*, where* $\gamma = MH + M - 1 - Var_1 + \sum_j \ln(p_j)$*.*

By setting $p_j = \frac{1}{M}$*,* $j = 0 \dots M-1$*, we note that the first term of the variance is equal to 0 and* γ *becomes equal to* $M-1$*. Meanwhile, the second term of the variance approximation is only* $\frac{M-1}{2n_b^2}$*.*

8.3.2 Test for predictability

After all the preliminaries given in the previous section, we aim to determine if a given process is not fully random. Our null and alternative hypotheses in this chapter are the following.

$\mathcal{H}_0$: Appearance of a new symbol in sequence is *independent* from the previous history of the sequence.

$\mathcal{H}_a$: Appearance of a new sequence *depends* on past observations of the sequence.

First, we determine if entropy is at the maximum for $k = 1$. To this aim, we check if the difference between the maximum of entropy, $\ln(M)$ and the entropy estimation is larger than 99% quantile of gamma distribution q_{99}^{Γ} with mean from Eq. 8.3 and variance from Eq 8.1. If the difference from the maximum is close to 0, that is, less than q_{99}^{Γ}, we conclude that symbols have the same probability of appearing. In this case, we repeat the test for $k > 1$ to test if blocks of symbols have different probabilities. if not stated otherwise, we use $k = 2$.

In the case when symbols have different probabilities, the normal distribution is used to detect predictability of blocks of symbols. If symbols appear independently, then k-th order entropy, that is the entropy estimation for blocks of length k, is the entropy of single symbols multiplied by k [Shannon, 1948]. Thus, we test if the value k-th order entropy is less than the first percentile of the normal distribution with the mean $k \cdot \hat{H}_1^G - \Delta^N$, where $\hat{H}_1^G$ is the unbiased estimation of entropy obtained for $k = 1$ from Eq. 2.4 and Δ^N is the bias from Eq. 8.4. We note that we select a mean for the normal distribution that is a random variable, because it depends on $\hat{H}_1^G$. However, we expect that this estimate, firstly, is not biased, and secondly, it has a low deviation since a large number of blocks is used for estimating empirical frequencies when $k = 1$. The case with a normal distribution may not be considered separately, for example, if all symbols have equal probabilities, which can be achieved by choosing a discretization of price returns.

If entropy is statistically significantly low for $k > 1$ in comparison with case $k = 1$, we say that the interval where entropy is estimated is inefficient. Otherwise, we call the time interval efficient to distinguish between two types of intervals.

8.3.3 Simulations: Bernoulli and autoregressive model

First, we conduct our test using simulations. We simulate two processes with length $n = 10^6$. The first one is Bernoulli sequence with a probability of appearing 0 equal to p. Other process is auto-regressive process with parameter AR. The process is then discretized: negative values are denoted by 0, positive values are denoted by 1. For different values of k, we plot the difference between the maximum value of entropy and the corresponding 99% quantile of the gamma distribution. Figure 8.1 shows the results. For fully random processes, when $p = 0.5$ or $AR = 0$, the entropy is almost at the maximum. For other values, significant deviations from the maximum are found for all values of k.

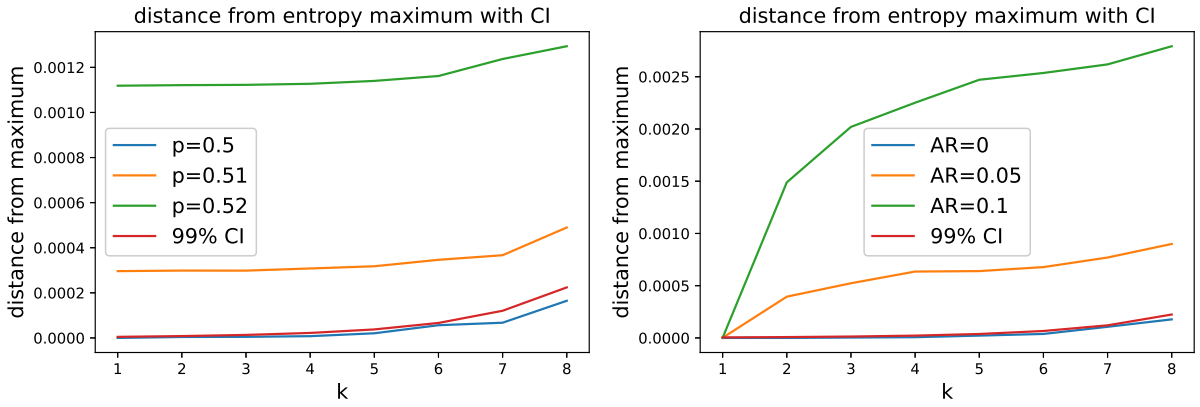
Next sections are about the predictability of executed orders for the group of assets described in Section 8.2.

8.4 Predictability of Apple's limit order book

8.4.1 Probabilities of single symbols and pairs

We move to real data analysis and start from investigating the limit order book of AAPL stock. We first consider the August 1 and August 24 of the year 2022 to illustrate how entropy estimation behaves for different days and lengths of blocks. For these two days, we present durations in seconds between each transaction in Figure 8.2.

The time precision is nanoseconds. With a given precision, some transactions happen at the same time. Such simultaneous events may correspond to the event when the volume of a market order is larger than the volume of the best buy or sell limit order. Another possibility is that market orders from different traders are executed automatically at some specific price. We exclude such events from the analysis by summing volumes and consider the first price available at each nanosecond with trading activity. Considering different values of k, we present results of



(a) Bernoulli

(b) AR model

Figure 8.1: Relative difference from the maximum for Bernoulli and AR models. Results are obtained by simulations of processes with length $n = 10^6$. All values are normalized by the maximum entropy, $\ln M$.

the hypothesis testing in Figure 8.3. To construct confidence intervals, we use a fitted gamma distribution with the mean and variance as discussed in Section 8.3.1.

For the 1st of August 2022, entropy is low even for $k = 1$. That is, the amount of positive and negative returns significantly differs. In such a case, a low entropy for $k > 1$ is expected. However, for $k = 2$ we calculate that $p(00) > p(11) > p(10) > p(01)$. Therefore, there is a pattern of repeating signs of price returns. For the 24th of August 2022, entropy is high for single symbols. However, it decreases when $k = 2$. Again, $p(00) > p(11) > p(01) > p(10)$. Similarly, blocks 00000 and 11111 are more frequent than other blocks with length $k = 5$ for August 1. For August 24, the most frequent blocks with length $k = 5$ are 10000 and 11111. The two days selected for analysis, August 1 and 24, show that the number of positive and negative returns can be statistically equal or significantly different depending on a trading day. More frequent repetitions of pairs of identical symbols is a consequence of a long memory of order signs. We are interested in discovering other causes of predictability, but not only repetition of symbols. For this reason, in the next section we consider data with high frequency, but already aggregated. We use transaction time, that is, we aggregate by a number of transactions.

8.4.2 Detecting predictability in transaction time

In this section, we examine the predictability of Apple's stock over several months and with different aggregation levels. An *aggregation level* is a number of transactions taken as one time step. We plot on Fig. 8.4 values of the cumulative distribution function (cdf) of the gamma distribution corresponding to obtained entropies with different aggregation levels.

The higher the value of cdf, the larger the degree of predictability. If cdf is less than 0.99, we can conclude that price returns do not exhibit significant predictability with the significance level $\alpha = 0.01$. We observe that predictability level decreases with the growth of aggregation level. Moreover, we see that the larger aggregation is needed in August to rid of the predictability of trades. We plot the fraction of days which is determined as inefficient for each aggregation level in Fig. 8.5. August contains a larger amount of days with significant predictability of orders in comparison to the autumn months. With larger aggregation, the fraction of inefficient days

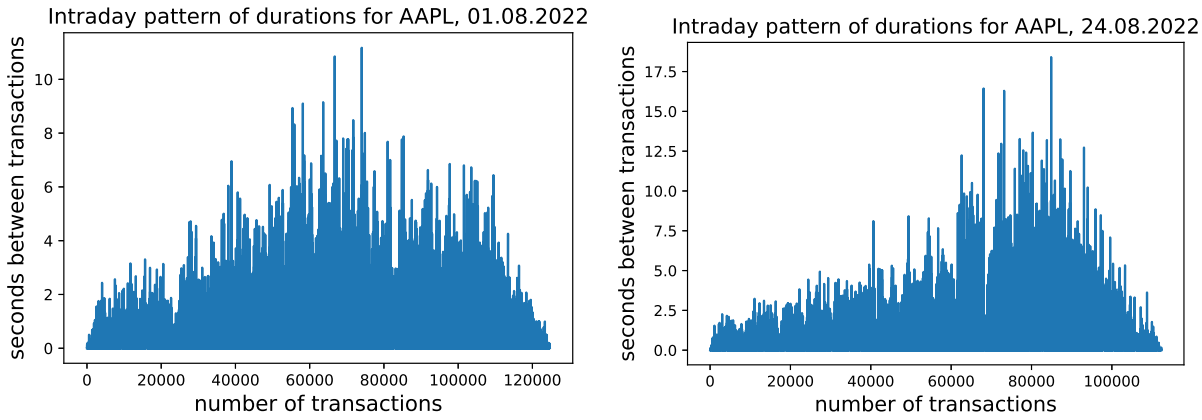


Figure 8.2: Time in seconds between each transaction during two trading days, 01.08.2022 and 24.08.2022

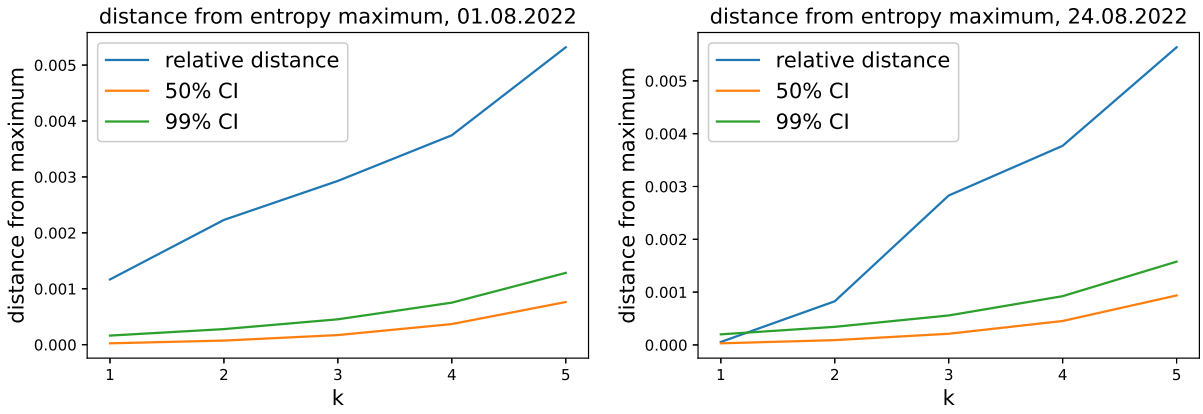


Figure 8.3: difference between entropy estimation and its maximum divided by $\ln(M)$, 01.08.2022 and 24.08.2022

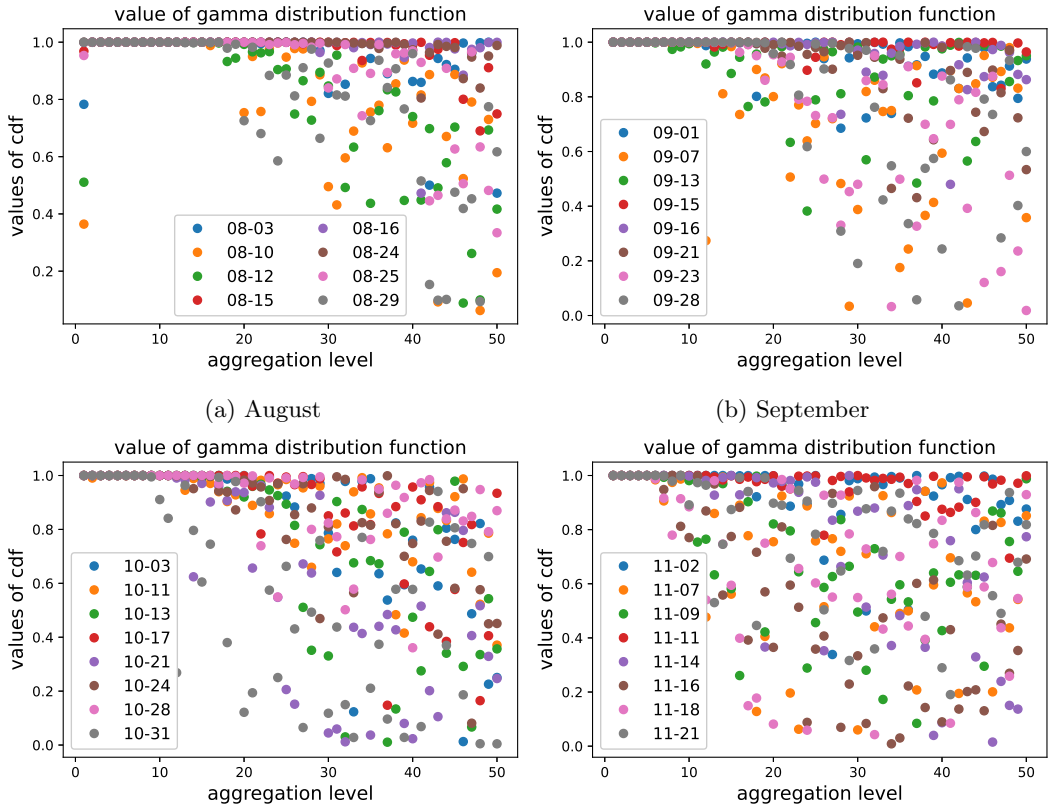
decreases but not monotonically. Also, we observe some outliers in August where the price is unpredictable for small aggregation values.

We check if entropy for $k = 2$ is significantly low than for $k = 1$. If $p(0)$ and $p(1)$ differ significantly, we use the normal distribution to test if entropy is significantly lower than unbiased estimation calculated with $k = 1$. Otherwise, we use the gamma distribution to test if entropy is too far from the possible maximum. Among 80 days, there are 66 inefficient days. 42 of them are tested using the gamma distribution. There are other 6 days that are efficient for both $k = 1, 2$.

8.4.3 Statistics of inefficient time intervals

Now, we consider different parameters of price returns time series and check if there is a dependence between them and predictability. The list of parameters follows.

- First, we compute $p(00) + p(11)$ to determine if predictability for $k = 2$ appears because the price moves in the same direction.



(a) August

(b) September

(c) October

(d) November

Figure 8.4: Values of the gamma cumulative distribution function of the distance of entropy estimation from its maximum for different months and aggregation levels in transaction time

- We calculate $|p(1) - p(0)|$ to check if predictability is caused by the difference in the amount of price increases and decreases during a trading day. We also write down the magnitude of daily changes in a price to determine if the predictability appears when the price significantly changes. For the same reason, we record the mean value of price returns.

- Then, we calculate the fraction of 0-returns and the amount of non-zero returns that is the length of the symbolized sequence.

- We are interested in autocorrelation of non-zero returns as well as in autocorrelation of their magnitude values. Estimating autocorrelation, we consider only non-zero returns.

- Then, we consider the distribution of price returns. We fit empirical price returns distribution[18] and record the degrees of freedom ν, scale, and shift parameters. The smaller value of ν, the fatter tails of price returns.

- Finally, we are interested in if there is a significant difference in trading volumes.

[18]Here, we use scipy.stats.t.fit in Python.

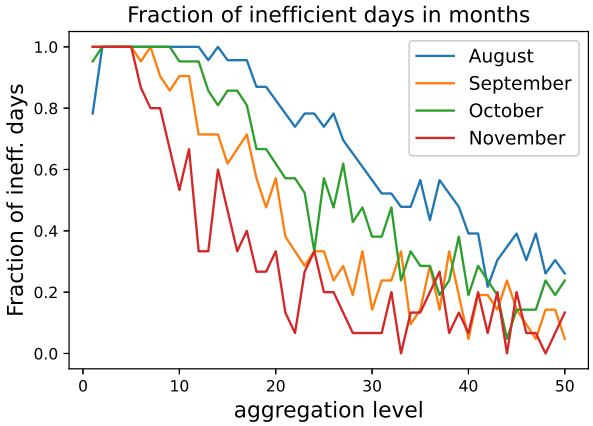


Figure 8.5: Fraction of inefficient days for months from August to November for different aggregation levels in transaction time

Later we include jumps into such analysis of dependence between stylized facts and predictability. We calculate mean values of the described parameters for the AAPL stock in Table 8.2. We test if there is a statistical significance in mean values for efficient and inefficient days. P-value less than 0.05 stands for significant difference with 95% of confidence.

According to the test for the difference in mean value of $p(00) + p(11)$, the price direction is more persistent inefficient days. Inefficient days contain more non-zero price changes and have less fraction of 0-returns. Trading volumes for inefficient days are significantly greater than for efficient days. We suppose that there is more motivation for traders to execute their orders during days labelled as inefficient.

To address the reason for the predictability of the ultra-high frequency time series, which is not related to repeating buy or sell orders, we aggregate data by $a = 30$ transactions. Besides the stylized facts recorded in Table 8.2, we also compare the fraction of jumps detected among all price returns. For the detection of jumps, we use the method described by [Lee and Mykland, 2007]. To employ this test on ultra-high frequency data, we average price returns as suggested by [Christensen et al., 2014].[19] For the aggregated data, we identify only two inefficient days, namely 30.08 and 01.09. All differences in statistics found for data with all executions are vanished, that is, there are no significant difference in average statistics. In particular, the p-value for testing equal fractions of jumps is 0.957.

8.4.4 Localization of inefficient intervals

In this section, we consider the length of the interval used to estimate entropy. In previous sections, we investigate daily time periods. There are two days where the predictability of pairs of symbols is detected when the data is aggregated. The research question is whether there is a smaller time interval inside an inefficient day for which significant predictability can be detected. The motivation for searching for the smaller interval is to localize when price predictability occurs. Moreover, the smaller the time interval, the less information from the past is needed to

[19] We use the square root of the amount of price returns as the window size used in the method [Lee and Mykland, 2007]. The method [Christensen et al., 2014] requires pre-averaging of price returns. We use the same number of transactions for aggregation and pre-averaging, that is $a = 30$ for AAPL. Jumps are defined with significance level of 1%.

Table 8.2: Statistics obtained for efficient and inefficient days of AAPL without aggregation.

parameter	mean for inefficient days	mean for efficient days	p-value
$p(00)+p(11)$	0.545	0.514	1.509×10^{-7}**
$\lvert p(1)-p(0)\rvert$	0.016	0.020	0.408
magnitude of daily log-price increment	0.0143	0.0145	0.963
mean price returns	-4.23×10^{-8}	9.10×10^{-8}	0.124
fraction of 0-returns	0.594	0.636	5.397×10^{-6}**
number of non-zero returns	34,242	26,813	0.003**
magnitude of autocorrelation of non-zero returns	0.036	0.022	0.006**
magnitude of autocorrelation of absolute values	0.173	0.221	0.112
ν of t-distribution	1.683	1.372	0.134
scale of t-distribution	1.438×10^{-5}	5.432×10^{-6}	0.0007**
magnitude of shift of t-distribution	9.99×10^{-8}	8.263×10^{-8}	0.688
daily volume	12,533,484	10,532,829	0.034*

In the last column, * is rejection of equal means with 0.05 significance and ** stands for 0.01 significance.

calculate entropy. Thus, a small rolling window allows to detect price predictability faster.

Since rolling window inside a trading day implies multiple tests, the [Šidák, 1967] correction of the significance level is used. Moreover, to make the conducted tests independent, we consider non-overlapping time intervals inside a trading day. For each trading day, we aim to detect inefficiency for S partitions with significance level $1-0.99^{1/S}$. We record the minimum value of S starting from $\lfloor \frac{n_b}{n_{min}} \rfloor$ where $n_{min} = \lceil \ln\left(\frac{0.01}{M}\right) / \ln\left(1-\frac{1}{M}\right) \rceil$. The amount of partitions and the number of partition (from 1 to S) where inefficiency is detected are in Table 8.3. For each day, there is only one from S intervals where inefficiency is found. Thus, predictability disappears from the time interval to the next subsequent non-overlapping interval. We show that predictability is present in the middle of trading day on August 30. On September 1, predictability is detected at the end of the trading day, namely in the last of 47 sub-intervals.

Table 8.3: Partition of inefficient days for AAPL.

day	S	N. $\in [1, S]$
30.08	3	2
01.09	47	47 (16)

Values in brackets correspond to the number of interval (from 1 to S) where entropy is significantly low with 0.01 level of significance, without the Šidák correction.

8.4.5 Entropy production

Finally, we are interested to test if price directions are revertible in time. We estimate entropy production that is Kullback-Leibler divergence associated with the probability measure of a process and its reversal [Cristadoro et al., 2023, Benoist et al., 2018]. Entropy production is a measure of the irreversibility of the process. In other words, entropy production measures Kullback-Leibler divergence (Eq. 6.5) between probabilities of outputs and probabilities of reversal outputs.

Definition 10. *Let X be a stationary random process with a finite alphabet A and a measure p. Entropy production of X is*

$$ep(X) = -\lim_{k\to\infty} \frac{1}{k} \sum_{x_1^k \in A^k} p(x_1^k) \log \frac{p(x_1^k)}{p(\hat{x}_1^k)}$$

where $\hat{x}_1^k = u(x_k)u(x_{k-1})\dots u(x_1)$ and $u : A \to A$ is an involution.

We try two possible types of involution: $u^{id}(0,1) = (0,1)$, $u^s(0,1) = (1,0)$. Involution u^{id} leaves symbols unchanged, while u^s swaps symbols.

As proved by [Cristadoro et al., 2023], ratio of recurrence times from Theorem 2, $\frac{1}{k}\log\frac{\hat{R}_k}{R_k}$, almost surely converges to entropy production. Here,

$$\begin{aligned} R_k &= \inf\{m \geq 1 : x_{m+k}^{m+2k-1} = x_1^k\} \\ \hat{R}_k &= \inf\{m \geq 1 : x_{m+k}^{m+2k-1} = \hat{x}_1^k\} \end{aligned}$$

We use another way of estimating entropy production also given by [Cristadoro et al., 2023] involving match lengths instead of recurrence times.

$$\begin{aligned} L_m &= \sup_{1\leq r\leq m} \{k \geq 1 : x_{r+k}^{r+2k-1} = x_1^k\} \\ \hat{L}_m &= \sup_{1\leq r\leq m} \{k \geq 1 : x_{r+k}^{r+2k-1} = \hat{x}_1^k\} \\ \hat{ep}_m &= \frac{\log m}{\hat{L}_m} - \frac{\log m}{L_m} \\ ep &= \lim_{m\to\infty} ep_m \end{aligned} \tag{8.5}$$

For each day, we consider discretized data without aggregation and calculate $\hat{ep}_{100}$ for 100 random initial points. That is, the first symbol is chosen randomly to obtain 100 estimations of $\hat{ep}_{100}$. Mean values of entropy productions with standard deviations are presented in Figure 8.6 for two types of involution.

We can conclude that time series of price directions are reversible for the considered time period for the AAPL stock. There is no significant difference in the process generating signs of price returns and the process reversed in time. We also do not detect irreversible time series of daily time intervals with the involution that changes signs of returns.

8.5 Predictability of executed orders

Here, we consider all assets presented in Table 8.1 to collect more results about predictability of data, factors that influence on this predictability, and irreversibility of signs of orders.

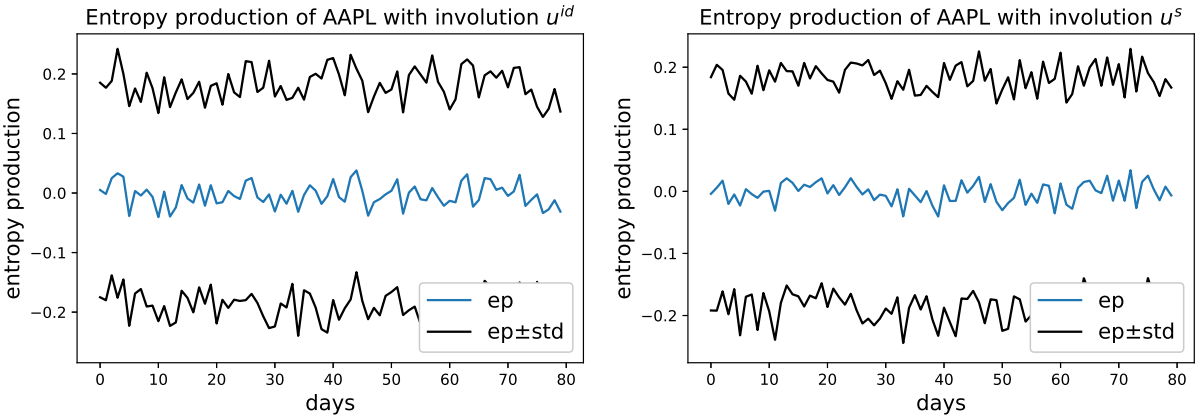


Figure 8.6: Entropy production of AAPL stock with two involutions

First, we conduct analysis of statistics of efficient and inefficient days of the Microsoft stock. Without aggregation by transactions, we find 77 inefficient days. Detailed results are given in Table 8.4. We conclude that days with detected predictability are characterized by larger daily price changes, trading volumes, and the number of price changes during a day.

For the stock TSLA, we detect only one efficient day without aggregation. We notice that the sum of empirical probabilities, $p(00)$ and $p(11)$, are significantly large even with aggregation level $a = 50$. All statistics are given in Table 8.5. With $a = 50$, ten inefficient days are also characterized by larger amount of price changes during a day and smaller degree of freedom of the fitted t-distribution of price returns.

Ten inefficient days for the stock TSLA with aggregation by $a = 50$ transactions are in Table 8.6. For each of these days, we make the test for the predictability of shorter time intervals. We conclude that for 9 days out of 10, daily time intervals can be narrowed so that the predictability is detected only for a part of an inefficient day. Moreover, we specify the approximate duration of inefficient time intervals and its location in terms of the amount of total intervals and the index number of time intervals with inefficiency. For example, we notice that for the days 16.08, 03.10, and 14.10, time intervals with predictability are at the beginning of the trading days.

We find only 25 inefficient days for the stock INTC without aggregation. They differ significantly only by a high probability of repeating price direction. Aggregating by $a = 20$ transactions, we find only 6 inefficient days that have significantly high autocorrelation. We filter out significant autocorrelation by fitting an ARMA model with constant mean and $0 \leq P + Q \leq 3$. After filtering out linear dependencies by ARMA(P,Q) model, we denote values of residuals less or greater than 0 by two different symbols of alphabet A for discretization. Estimating the entropy of discretized residuals, we detect only 5 inefficient days. The five days are listed in Table 8.7.

Table 8.4: Statistics obtained for efficient and inefficient days of MSFT without aggregation.

parameter	mean for inefficient days	mean for efficient days	p-value
$p(00) + p(11)$	0.556	0.514	0.107
$\|p(1) - p(0)\|$	0.022	0.026	0.613
magnitude of daily log-price increment	0.013	0.008	0.013*
mean price returns	-1.397×10^{-8}	1.038×10^{-7}	0.743
fraction of 0-returns	0.502	0.473	0.292
number of non-zero returns	30,258	10,971	0.033*
magnitude of autocorrelation of non-zero returns	0.027	0.068	0.160
magnitude of autocorrelation of absolute values	0.197	0.268	0.281
ν of t-distribution	1.943	1.988	0.158
scale of t-distribution	2.489×10^{-5}	2.925×10^{-5}	0.441
magnitude of shift of t-distribution	4.432×10^{-7}	5.615×10^{-7}	0.485
daily volume	4,650,933	1,394,246	0.012*

In the last column, * is rejection of equal means with 0.05 significance.

Table 8.5: Statistics obtained for efficient and inefficient days of TSLA with aggregation $a = 50$.

parameter	mean for inefficient days	mean for efficient days	p-value
$p(00) + p(11)$	0.545	0.520	4.059×10^{-7}**
$\|p(1) - p(0)\|$	0.025	0.024	0.721
magnitude of daily log-price increment	0.025	0.022	0.746
mean price returns	-7.707×10^{-6}	-6.375×10^{-7}	0.062
fraction of 0-returns	0.032	0.039	0.210
number of non-zero returns	2,758	2,093	0.003**
magnitude of autocorrelation of non-zero returns	0.058	0.041	0.273
magnitude of autocorrelation of absolute values	0.128	0.142	0.404
ν of t-distribution	1.996	4.291	0.024*
scale of t-distribution	4.143×10^{-4}	4.163×10^{-4}	0.921
magnitude of shift of t-distribution	2.018×10^{-5}	1.733×10^{-5}	0.475
daily volume	10,833,692	8,378,487	0.188
fraction of jumps	7.163×10^{-5}	7.942×10^{-5}	0.928

In the last column, * is rejection of equal means with 0.05 significance, and ** stands for 0.01 significance.

Table 8.6: Partition of inefficient days for TSLA.

day	S	N. $\in [1, S]$
01.08	3	9
08.08	29	17
10.08	54	20 (37)
16.08	3	1
03.10	43	4 (27 38 43)
07.10	11	8
12.10	1	1
14.10	47	3 (47)
20.10	6	5
10.11	14	3

Values in brackets correspond to the number of interval (from 1 to S) where entropy is significantly low with 0.01 level of significance, without the Šidák correction.

Table 8.7: Partition of inefficient days for INTC.

day	S	N. $\in [1, S]$
12.09	4	4
12.10	2	1
13.10	17	6
24.10	2	1, 2
28.10	24	14

For all five days, there is a time interval with a smaller length where entropy is statistically significantly low. For all days except 24.10, there is the only interval with a significant degree of predictability. That is, having one interval with a low level of entropy, the next one (and the previous one) is no longer inefficient. Predictability disappears as soon as we consider the interval following the other interval where the low entropy is found. The exception is October 24 when both halves of the day are time intervals with low entropy values.

The stock LLY is traded less often and with less amount of transactions in comparison to the already mentioned stocks. For the stock LLY, results are given in Table 8.8. Without aggregation, we detect only 73 inefficient days. The fraction of 0-returns is significantly smaller for these inefficient days, than for the other 7 days. The difference in this fraction is also discovered for other stocks. Moreover, in the case of LLY, inefficient days are characterized by a low autocorrelation of non-zero returns.

Table 8.8: Statistics obtained for efficient and inefficient days of LLY without aggregation.

parameter	mean for inefficient days	mean for efficient days	p-value
$p(00) + p(11)$	0.593	0.553	0.112
$\|p(1) - p(0)\|$	0.034	0.073	0.173
magnitude of daily log-price increment	0.012	0.017	0.288
mean price returns	6.835×10^{-8}	-2.696×10^{-6}	0.387
fraction of 0-returns	0.396	0.414	0.043*
number of non-zero returns	5,073	3,221	0.110
magnitude of autocorrelation of non-zero returns	0.061	0.105	0.005**
magnitude of autocorrelation of absolute values	0.195	0.145	0.055
ν of t-distribution	1.652	1.053	0.126
scale of t-distribution	4.654×10^{-5}	3.826×10^{-5}	0.674
magnitude of shift of t-distribution	1.194×10^{-6}	2.36×10^{-6}	0.416
daily volume	383,708	227,210	0.072

In the last column, * is rejection of equal means with 0.05 significance, and ** stands for 0.01 significance.

For the stock of Snapchat, we find 22 inefficient days. Results are collected in Table 8.9. These days are characterized by a low probability of repeating symbols, high trading volumes, a large fraction of 0-returns, and high autocorrelation of non-zero returns. unlike other discussed stocks, inefficient days of SNAP exhibit larger fraction of 0-returns in comparison with efficient days.

In the case of the stock SNAP, the probability of repeating the direction of price is less than 0.5 for inefficient days. Therefore, the probability of changing price direction is significantly high for inefficient days. We already discuss such a periodic pattern of signs in Section 3.7 but at one-minute frequency. Already with aggregation each pair of transactions, $a = 2$, all differences between 13 inefficient and 67 efficient days disappear.

For the data of the stock of Ford, we detect 13 inefficient days. The results are in Table 8.10. Here, inefficient days differ in a set of parameters. Firstly, they are more likely to have a pattern

Table 8.9: Statistics obtained for efficient and inefficient days of SNAP without aggregation.

parameter	mean for inefficient days	mean for efficient days	p-value
$p(00)+p(11)$	0.464	0.508	0.003**
$\|p(1)-p(0)\|$	0.014	0.017	0.328
magnitude of daily log-price increment	0.034	0.030	0.557
mean price returns	1.19×10^{-7}	-1.208×10^{-6}	0.275
fraction of 0-returns	0.660	0.638	0.015*
number of non-zero returns	3,663	2,951	0.086
magnitude of autocorrelation of non-zero returns	0.168	0.057	8.169×10^{-5}**
magnitude of autocorrelation of absolute values	0.205	0.173	0.035*
ν of t-distribution	1.511	1.385	0.493
scale of t-distribution	2.526×10^{-5}	5.262×10^{-5}	0.112
magnitude of shift of t-distribution	1.896×10^{-7}	7.572×10^{-7}	0.124
daily volume	6,475,660	4,395,545	0.020*

In the last column, * is rejection of equal means with 0.05 significance, and ** stands for 0.01 significance.

of changing symbols, indicating an increase or decrease in price. The same pattern at ultra-high frequency is detected for the stock SNAP as well. Also, on average, inefficient days contain a greater percentage of 0-returns and a greater number of changes in price. Additionally, we find a significantly higher correlation for returns as well as a greater degree of freedom of t-distribution for inefficient days.

Using aggregation by $a = 5$ trades, we obtain only 7 inefficient days presented in Table 8.11. They differ statistically from 73 efficient days only by significantly low autocorrelation of absolute values of non-zero returns. For 3 days out of 7, we do not find a shorter interval with a statistically low level of entropy. For the remaining four, taking the Šidák correction into account, we can find one inefficient sub-interval of a smaller length. To do this, we divide the day into several equal parts, i.e., into 3, 4, or 23 parts.

Now, we analyze the stock of Carnival Corp., that is tour operator company. There are 17 inefficient days found for the stock CCL. These days differ from efficient days by the larger value of autocorrelation and the scale of the fitted t-distribution of price returns. We remove such significant differences by considering the residuals of an ARMA model. Investigating residuals, we find only 16 inefficient days presented in Table 8.12.

Since we analyze data of the CCL stock without aggregation, we have a complete data. First, for the data without aggregation, we notice that some inefficient intervals cluster together. We highlight such intervals in Table 8.12. Then, we detect days when predictable time intervals occur several times during a day, i.e., on 09.09 and 16.11 as shown in the Table. Here, we do not conduct analysis on news and their influence on price predictability. As in the whole thesis, we examine only prices and trading volumes. However, we notice that Carnival Corp. announced an increase in Internet service price on board on September 9. On November 15, Carnival Corp. announced that it would issue a new debt worth one billion. On November 16, the price

Table 8.10: Statistics obtained for efficient and inefficient days of F without aggregation.

parameter	mean for inefficient days	mean for efficient days	p-value
$p(00) + p(11)$	0.459	0.491	0.002**
$\|p(1) - p(0)\|$	0.023	0.022	0.789
magnitude of daily log-price increment	0.023	0.014	0.12
mean price returns	-6.211×10^{-8}	7.512×10^{-7}	0.352
fraction of 0-returns	0.729	0.687	0.011*
number of non-zero returns	2,763	2,046	0.026*
magnitude of autocorrelation of non-zero returns	0.181	0.124	0.001**
magnitude of autocorrelation of absolute values	0.236	0.225	0.458
ν of t-distribution	1.973	1.724	0.001**
scale of t-distribution	3.695×10^{-15}	1.534×10^{-5}	0.008**
magnitude of shift of t-distribution	9.118×10^{-16}	2.627×10^{-7}	0.062
daily volume	6,048,314	4,161,430	0.064

In the last column, * is rejection of equal means with 0.05 significance, and ** stands for 0.01 significance.

of Carnival stock decreased by 13%. We can only suppose that the predictability of prices at high frequencies is due to the presence of news. This hypothesis is consistent with the empirical results for the set of stocks. When new information is available, traders are more active. The trading activity is reflected in changes in the price and in the volume of stocks traded as shown for the stocks AAPL, MSFT, and TSLA. For further consideration of price reaction to news, we refer to a review by [Corrado, 2011].

Last, we consider the price of the ETF SPY. There are 69 inefficient days with a statistically high probability of moving price in the same direction. The similar results we obtain for stocks AAPL, TSLA, and INTC. Moreover, efficient days are characterized by a significant difference between numbers of price increasing and price decreasing as shown in Table 8.13. Using the aggregation level $a = 40$, we remain with only two inefficient days presented in Table 8.14.

Finally, we investigate reversibility of price returns signs for all 9 assets under consideration. The results are similar to those we get for the Apple stock. There is no evidence that price returns discretized by their signs are irreversible in time. We make such a conclusion by estimating entropy production and match lengths as shown in Equation 8.5.

Table 8.11: Partition of inefficient days for F.

day	S	N. $\in [1, S]$
01.08	1	1
09.08	4	4
25.08	1	1
14.09	1	1
20.09	3	1
21.10	4	3
27.10	23	7 (3 17)

Values in brackets correspond to the number of interval (from 1 to S) where entropy is significantly low with 0.01 level of significance, without the Šidák correction.

Table 8.12: Partition of inefficient days for CCL.

day	S	N. $\in [1, S]$
01.08	52	6 (4 31 44 50 52)
03.08	51	**50** (25 **43 44** 47 **49 51**)
09.08	63	43 (56)
29.08	1	1
09.09	59	**37 38 41 42 53 57 58** (46 48 **51 52** 55)
21.09	84	3 84 (42 59 81)
29.09	9	6
30.09	147	**147** (12 14 16 21 **30 31** 43 103 141 **143 144 145 146**)
05.10	61	**60** (18 28 43 58 **61**)
13.10	34	5
20.10	4	2
25.10	12	**12** (**11**)
03.11	76	**76** (44 51 54 **74 75**)
08.11	11	8
14.11	65	47 (50)
16.11	63	9 **57 59** (**3 4** 36 55 **58** 63)

Values in brackets correspond to the number of interval (from 1 to S) where entropy is significantly low with 0.01 level of significance, without the Šidák correction. Consecutive intervals for one day are in bold.

Table 8.13: Statistics obtained for efficient and inefficient days of SPY without aggregation.

parameter	mean for inefficient days	mean for efficient days	p-value
$p(00) + p(11)$	0.531	0.511	0.0001**
$\|p(1) - p(0)\|$	0.012	0.021	0.009**
magnitude of daily log-price increment	0.009	0.011	0.598
mean price returns	-4.813×10^{-9}	5.429×10^{-8}	0.502
fraction of 0-returns	0.441	0.455	0.203
number of non-zero returns	40,090	29,211	0.09
magnitude of autocorrelation of non-zero returns	0.020	0.024	0.62
magnitude of autocorrelation of absolute values	0.128	0.136	0.882
ν of t-distribution	1.963	1.826	0.422
scale of t-distribution	1.807×10^{-5}	1.557×10^{-5}	0.206
magnitude of shift of t-distribution	2.129×10^{-7}	3.126×10^{-7}	0.115
daily volume	9,416,945	7,372,443	0.212

In the last column, ** is rejection of equal means with 0.01 significance.

Table 8.14: Partition of inefficient days for SPY.

day	S	N. $\in [1, S]$
17.08	1	1
01.09	3	3

8.6 Discussion on the predictability at ultra-high frequency

We study the predictability of stock prices at ultra-high frequency. Considering signs of price changes, we construct binary sequences for all recorded executed transactions. We have proposed a statistical test to determine if blocks of the signs exhibit a significant degree of predictability. Applying the test, we distinguish efficient days from inefficient days. For conducting the test for low entropy values, we take advantage of knowing the distribution of entropy estimation [Zubkov, 1974] and the bias and variance of the estimation obtained in Section 2.4. A further development of the test is deriving distribution of entropy estimation obtained with overlapping blocks of symbols.

It is known that signs of trades have a long memory in such a microscopic view of transaction data [Lillo and Farmer, 2004, Bouchaud et al., 2003]. We have shown that the probability that the price of an asset has two subsequent movements in the same direction is one of the factors affecting the predictability of the prices. According to [Lillo and Farmer, 2004, Bouchaud et al., 2003], such a long memory does not imply inefficiency of the considered market, because such type of predictability is compensated by fluctuations in transaction costs and liquidity and by interaction between market makers and informed traders. To consider predictability related not only to long memory, we use aggregation by the number of transactions. That is, in this chapter, we work with transaction time, but not with calendar time as in previous chapters. As we discuss in Section 3.4, there is a data regularity called intraday volatility pattern observed in calendar time. In transaction time, there is an irregular in seconds time spacing that was empirically shown by [Engle and Russell, 1998, Biais et al., 1995]. We demonstrate such a pattern for the AAPL stock in Figure 8.2. An advantage of using transaction time was discussed by [Oomen, 2006]: In general, sampling in transaction time allows to reduce the errors of estimation of realized variance.

We have shown that the degree of predictability decreases with the increase of aggregation level. Moreover, we have demonstrated that prices recorded in August are more predictable than prices in the autumn months. In order to get rid of predictability in the days of August, we need to aggregate a larger number of transactions than in the autumn.

If transactions appear at extremely high frequencies, e.g., less than one second on average, then the larger fraction of days is defined as inefficient and characterized by repeating signs of price returns. It is also worth noting that for some stocks (SNAP, F), days with inefficient time intervals are described by a significantly low probability that price moves twice in the same direction. The repetition of price direction is explained by the long memory caused by the appearance of news, the reaction to them, and the splitting of one order into parts. Meanwhile, we explain the pattern of changing price direction in Section 3.7 by a bid-ask bounce and a random walk of a price described by a low mean value and a low volatility that are characteristics of stocks SNAP and F. We have detected dependence between such pattern of changing price and inefficiency at a higher (one minute) frequency in Section 5.2. Since the average time between transactions of SNAP and F are larger than for stocks AAPL, MSFT, TSLA, and INTC, we suppose that detection of such periodic patterns depends on the frequency of trading.

For 3 out of 9 stocks under consideration (INTC, SNAP, CCL), non-zero price returns of inefficient days have high autocorrelation. Highly significant coefficients of an AR model for tick-by-tick data were empirically investigated by Engle [2000] and Robert and Rosenbaum [2010]. Some stylized facts of price returns at ultra-high frequency data including fat tails of return distribution and volatility clustering were discussed in [Bouchaud, 2005]. To explore fat tails of price returns, we estimate degrees of freedom of the fitted t-distribution of the price returns. For the stock TSLA, we discover that price returns of inefficient days have fatter tails than returns of efficient days. However, we have the opposite result for the stock of Ford, where efficient days

are described by price returns with fatter tails. We check volatility clustering by measuring the autocorrelation of absolute values of returns. The autocorrelation is significantly greater during inefficient days for the stocks AAPL and SNAP. Again, we obtain the opposite result for the stock F.

We notice that inefficient days of the AAPL stock are characterized by larger trading volumes, the larger amount of non-zero price changes, and the less fraction of 0-returns in comparison with efficient days. We assume that such a difference in characteristics for the two groups of days is associated with the activity of traders in the presence of news on days marked as inefficient. This assumption is also based on the research papers on price reaction to news. [Grinblatt et al., 1984] empirically showed that stock prices react to announcements about stock dividends and splits. [Chan, 2003] stated that public news affect monthly price returns. [Huynh and Smith, 2017] showed that weekly price returns react to the attention to news and their tone.

We apply the correction proposed in [Šidák, 1967] to make multiple tests for predictability for short intervals during inefficient days. In most cases, we can identify one inefficient interval by dividing a day into equal intervals in transaction time. In such a way, we determine both the position of this interval relative to the time of the day and its duration. For the stock CCL, we have found several groups of such inefficient intervals following each other. Finally, we have applied the method proposed by [Cristadoro et al., 2023] to investigate reversibility of signs of price returns. Applying entropy production Benoist et al. [2018], we conclude that process generating price returns signs is reversible in time.

Conclusions

This thesis contributes to the topic of measuring efficiency of financial markets and unpredictability of time series. A crucial consideration when evaluating a degree of market efficiency is that there are two sources of predictability of the prices of financial assets. The first source is the deviation of a market from the state of efficiency when prices follow martingale models. The second source is stylized facts of financial markets that are empirical properties that describe price returns. We have constructed a four-steps method for filtering out data regularities that are such stylized facts that increase the predictability of price returns. The four data regularities considered in the thesis are intraday volatility pattern, volatility clustering, price staleness, and microstructure noise. First, we have shown that price staleness decreases entropy, a measure of randomness. Second, we have discussed a range of methods for filtering out price staleness, volatility clustering, and microstructure noise and have made the comparative analysis of the methods. Then, we have proposed a method that filters out the effects of volatility clustering and price staleness simultaneously. We pay attention that the effect of price staleness tends to decrease an estimation of volatility. A degree of price staleness is determined through the probability of price rounding up to its previous value, and this probability, in turn, is determined using the volatility estimation. The method we have proposed allows to update estimation of volatility and the probability of price rounding minute by minute. We have developed a formula for estimating the probability of rounding. We have extended this formula to include the parameters responsible for the bid-ask spread and the thickness of the tails of price returns distribution.

Having a method for filtering data regularities, we first evaluate the degree of efficiency of ETFs traded at New York Stock Exchange. We calculate the entropy of price returns at one-minute frequency using weekly, monthly, and quarterly time intervals. For weekly time intervals, we conclude that the ETF market is not totally efficient under the assumption of low transaction costs, however, the prices of ETFs are close to be fully random. We have shown that the degree of inefficiency of the group of ETFs increases from about 1.5% to 11% when moving from weekly to monthly intervals. Then, we have investigated the Moscow stock exchange. We infer that a degree of inefficiency for the Russian stock market using monthly time intervals is about 82%. We conclude that the degree of inefficiency depends on the length of intervals used for analysis and on the market under consideration.

Analyzing the Moscow stock exchange, we have performed stock market clustering. Using entropy estimation of sequences denoting co-movement of prices, we conclude that stocks of companies belonging to the same industry cluster together. We have detected a pair of stocks exhibiting the highest degree of inefficiency. We have shown that the most inefficient months cluster together during years 2014, 2015, and 2016. Most of the inefficient months are described by a frequent decreasing or increasing of the price for several minutes in a row. The entropy of returns of these two stocks then increases over time in the period under consideration until 2022. For the Russian stock market, we observe significant deviations from the Efficient Market Hypothesis. A more appropriate hypothesis might be the Adaptive Markets Hypothesis by [Lo,

2004] according to which arbitrage opportunities may exist until they are discovered and become known. Similarly, if gamblers systematically guess the numbers on roulette, the casino upgrades the equipment or tightens control over the game.

We have considered the problem of determining the optimal length of a rolling window used for entropy estimation. We have proposed a novel approach based on the count of significant changes in entropy values. The approach is also suitable for testing the stationarity of the process in the sense of a constant value of entropy during a considered time period. To check if entropy changes significantly, we have introduced a statistical test. To test if two entropies differ significantly, we have found the formula for the variance of entropy estimation depending on the length of the sequence and the set of probabilities. To make an unbiased estimation of the variance, we have introduced a random variable whose expected value is the variance of entropy estimation. The optimal length of rolling window for entropy estimation can be defined using a training set. The optimal length depends on the stock under consideration. For instance, the optimal length in calendar time is about one quarter of a year for the stocks of Apple and GameStop. For the IT stocks, like Apple, and meme stocks, like GameStop, we have captured time-varying behavior of entropy of price returns for the years 2020 and 2021. We use quarterly and yearly time intervals for filtering out data regularities, and thus conclude that detected significantly low values of entropy correspond to periods of a higher degree of inefficiency.

Furthermore, we move from data at one-minute frequency to ultra-high frequency data, where all transactions are recorded. With this microscopic view, we get more information about price dynamics and therefore we get lower values of the entropy of signs of price returns. We have shown that a degree of randomness increases with the increase of aggregation level in transaction time for the stock of Apple. For frequently traded stocks, we have demonstrated that statistically predictable price returns time series display the pattern of moving price in one direction. We also note that inefficient days detected at the ultra-high frequency are usually characterized by high trading activity expressed through trading volumes, daily price changes, and the amount of price changes during a day.

Investigating the time-varying behavior of entropy of meme stocks, we consider entropy as an early-warning indicator of price predictability. For two meme stocks of GameStop and AMC Entertainment Holdings, we have detected low values of entropy before the sharp increases in prices recorded in January 2021. Moreover, we have shown that statistically low entropy values correspond to high trading volumes. It is worth noting that the entropy estimation is based only on previous price values and does not take trading volumes into account. Therefore, we believe that entropy is a useful tool for detecting changes in predictability levels. We have shown on the example of meme stocks that a low entropy value is an indicator for economic bubbles, bursts and booms of prices. The relationship between entropy values and trading volumes suggests that the entropy of returns is an indicator of the possible coordination of traders and manipulations in the markets.

Bibliography

A. Agliari, A. Naimzada, and N. Pecora. Boom-bust dynamics in a stock market participation model with heterogeneous traders. *Journal of Economic Dynamics and Control*, 91:458–468, 2018. doi: 10.1016/j.jedc.2018.04.007. Special Issue in Honour of Prof. Carl Chiarella.

K. Ahn, D. Lee, S. Sohn, and B. Yang. Stock market uncertainty and economic fundamentals: an entropy-based approach. *Quantitative Finance*, 19(7):1151–1163, 2019. doi: 10.1080/14697688.2019.1579922.

E. Akyildirim, A. F. Bariviera, D. K. Nguyen, and A. Sensoy. Forecasting high-frequency stock returns: A comparison of alternative methods. *Annals of Operations Research*, 313(2):639–690, 2022. doi: 10.1007/s10479-021-04464-8.

J. Alvarez-Ramirez and E. Rodriguez. A singular value decomposition entropy approach for testing stock market efficiency. *Physica A*, 583:126337, 2021. doi: 10.1016/j.physa.2021.126337.

J. Alvarez-Ramirez, E. Rodriguez, and J. Alvarez. A multiscale entropy approach for market efficiency. *International Review of Financial Analysis*, 21:64–69, 2012. doi: 10.1016/j.irfa.2011.12.001.

T. G. Andersen, T. Bollerslev, and J. Cai. Intraday and interday volatility in the japanese stock market. *Journal of International Financial Markets, Institutions and Money*, 10(2):107–130, 2000. doi: 10.1016/S1042-4431(99)00029-3.

J. Babecký, T. Havránek, J. Matějů, M. Rusnák, K. Šmídková, and B. Vašíček. Banking, debt, and currency crises in developed countries: Stylized facts and early warning indicators. *Journal of Financial Stability*, 15:1–17, 2014. doi: 10.1016/j.jfs.2014.07.001.

F. M. Bandi, D. Pirino, and R. Renò. Excess idle time. *Econometrica*, 85(6):1793–1846, 2017. doi: 10.2139/ssrn.2199468.

F. M. Bandi, A. Kolokolov, D. Pirino, and R. Renò. Zeros. *Management Science*, 66(8):3466–3479, 2020. doi: 10.1287/mnsc.2019.3527.

W. A. Barnett and A. Serletis. Martingales, nonlinearity, and chaos. *Journal of Economic Dynamics and Control*, 24(5):703–724, 2000. doi: 10.1016/S0165-1889(99)00023-8.

G. P. Basharin. On a statistical estimate for the entropy of a sequence of independent random variables. *Theory of Probability & Its Applications*, 4:333–336, 1959. doi: 10.1137/1104033.

K. E. Bassler, G. H. Gunaratne, and J. L. McCauley. Markov processes, hurst exponents, and nonlinear diffusion equations: With application to finance. *Physica A: Statistical Mechanics and its Applications*, 369(2):343–353, 2006. doi: 10.1016/j.physa.2006.01.081.

M. Beben and A. Orłowski. Correlations in financial time series: established versus emerging markets. *The European Physical Journal B-Condensed Matter and Complex Systems*, 20(4): 527–530, 2001. doi: 10.1007/s100510170233.

D. Benedetto, E. Caglioti, and V. Loreto. Language trees and zipping. *Physical Review Letters*, 88:048702, 2002. doi: 10.1103/PhysRevLett.88.048702.

T. Benoist, V. Jakšić, Y. Pautrat, and C.-A. Pillet. On entropy production of repeated quantum measurements i. general theory. *Communications in Mathematical Physics*, 357:77–123, 2018. doi: 10.1007/s00220-017-2947-1.

A. Bezerianos, S. Tong, and N. Thakor. Time-dependent entropy estimation of eeg rhythm changes following brain ischemia. *Annals of biomedical engineering*, 31:221–32, 2003. doi: 10.1114/1.1541013.

B. Biais, P. Hillion, and C. Spatt. An empirical analysis of the limit order book and the order flow in the paris bourse. *The Journal of Finance*, 50(5):1655–1689, 1995. doi: 10.1111/j.1540-6261.1995.tb05192.x.

P. Billingsley. *Ergodic theory and information*, volume 1. Wiley New York, 1965.

F. Black and M. Scholes. The pricing of options and corporate liabilities. *Journal of political economy*, 81(3):637–654, 1973. doi: 10.1086/260062.

E. Boehmer and E. K. Kelley. Institutional Investors and the Informational Efficiency of Prices. *The Review of Financial Studies*, 22(9):3563–3594, 2009. doi: 10.1093/rfs/hhp028.

B. Bollen. What should the value of lambda be in the exponentially weighted moving average volatility model? *Applied Economics*, 47:853–860, 2015. doi: 10.1080/00036846.2014.982853.

T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Economics*, 31:307–327, 1986. doi: 10.1016/0304-4076(86)90063-1.

W. F. M. D. Bondt and R. Thaler. Does the stock market overreact? *The Journal of Finance*, 40(3):793–805, 1985. doi: 10.2307/2327804.

J.-P. Bouchaud. The subtle nature of financial random walks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 15(2):026104, 2005. doi: 10.1063/1.1889265.

J.-P. Bouchaud, A. Matacz, and M. Potters. Leverage effect in financial markets: The retarded volatility model. *Physical Review Letters*, 87:228701, 2001. doi: 10.1103/PhysRevLett.87.228701.

J.-P. Bouchaud, M. Mézard, and M. Potters. Statistical properties of stock order books: empirical results and models. *Quantitative finance*, 2(4):251, 2002. doi: 10.1088/1469-7688/2/4/301.

J.-P. Bouchaud, Y. Gefen, M. Potters, and M. Wyart. Fluctuations and response in financial markets: the subtle nature of 'random' price changes. *Quantitative Finance*, 4(2):176, 2003. doi: 10.1088/1469-7688/4/2/007.

J.-P. Bouchaud, J. D. Farmer, and F. Lillo. Chapter 2 - how markets slowly digest changes in supply and demand. In T. Hens and K. R. Schenk-Hoppé, editors, *Handbook of Financial Markets: Dynamics and Evolution*, Handbooks in Finance, pages 57–160. North-Holland, San Diego, 2009. doi: 10.1016/B978-012374258-2.50006-3.

R. C. Brasileiro, V. L. Souza, and A. L. Oliveira. Automatic trading method based on piecewise aggregate approximation and multi-swarm of improved self-adaptive particle swarm optimization with validation. *Decision Support Systems*, 104:79–91, 2017. doi: 10.1016/j.dss.2017.10.005.

L. Breiman. The individual ergodic theorem of information theory. *The Annals of Mathematical Statistics*, 28(3):809–811, 1957.

R. P. Brent. An algorithm with guaranteed convergence for finding a zero of a function. *The computer journal*, 14(4):422–425, 1971.

W. A. Broock, J. A. Scheinkman, W. D. Dechert, and B. LeBaron. A test for independence based on the correlation dimension. *Econometric Reviews*, 15(3):197–235, 1996. doi: 10.1080/07474939608800353.

X. Brouty and M. Garcin. A statistical test of market efficiency based on information theory. *Quantitative Finance*, 0(0):1–16, 2023. doi: 10.1080/14697688.2023.2211108.

C. Brownlees and G. Gallo. Financial econometric analysis at ultra-high frequency: Data handling concerns. *Computational Statistics & Data Analysis*, 51(4):2232–2245, 2006. doi: 10.1016/j.csda.2006.09.030.

J. Bulla and I. Bulla. Stylized facts of financial time series and hidden semi-markov models. *Computational Statistics & Data Analysis*, 51(4):2192–2209, 2006. doi: 10.1016/j.csda.2006.07.021.

J. A. Busse and T. Clifton Green. Market efficiency in real time. *Journal of Financial Economics*, 65(3):415–437, 2002. doi: 10.1016/S0304-405X(02)00148-4.

D. Cajueiro and B. Tabak. Ranking efficiency for emerging markets. *Chaos, Solitons and Fractals*, 22:349–352, 2004. doi: 10.1016/j.chaos.2004.02.005.

D. Cajueiro and B. Tabak. Ranking efficiency for emerging markets ii. *Chaos, Solitons and Fractals*, 23:671–675, 2005. doi: 10.1016/j.chaos.2004.05.009.

L. Calcagnile, F. Corsi, and S. Marmi. Entropy and efficiency of the etf market. *Computational Economics*, 55:143–184, 2020. doi: 10.1007/s10614-019-09885-z.

J. C. Chan and C. Santi. Speculative bubbles in present-value models: A bayesian markov-switching state space approach. *Journal of Economic Dynamics and Control*, 127:104101, 2021. doi: 10.1016/j.jedc.2021.104101.

L. K. C. Chan and J. Lakonishok. The behavior of stock prices around institutional trades. *The Journal of Finance*, 50(4):1147–1174, 1995. doi: 10.1111/j.1540-6261.1995.tb04053.x.

W. S. Chan. Stock price reaction to news and no-news: drift and reversal after headlines. *Journal of Financial Economics*, 70(2):223–260, 2003. doi: 10.1016/S0304-405X(03)00146-6.

M. C. Chidichimo and M. D. Thorsley. Asymptotic expansion of 2 f 1 (a, b; c; x) in terms of confluent hypergeometric functions. *Journal of Mathematical Physics*, 42(11):5371–5378, 2001. doi: 10.1063/1.1407835.

N. Chopra, J. Lakonishok, and J. R. Ritter. Measuring abnormal performance: Do stocks overreact? *Journal of Financial Economics*, 31(2):235–268, 1992. doi: 10.1016/0304-405X(92)90005-I.

K. Christensen, R. C. Oomen, and M. Podolskij. Fact or friction: Jumps at ultra high frequency. *Journal of Financial Economics*, 114(3):576–599, 2014. doi: 10.1016/j.jfineco.2014.07.007.

E. Cirugeda-Roldan, D. Cuesta-Frau, P. Miro-Martinez, and S. Oltra-Crespo. Comparative study of entropy sensitivity to missing biosignal data. *Entropy*, 16(11):5901–5918, 2014. doi: 10.3390/e16115901.

R. Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236, 2001. doi: 10.1080/713665670.

H. Coronel-Brizio, A. Hernández-Montoya, R. Huerta-Quintanilla, and M. Rodríguez-Achach. Evidence of increment of efficiency of the mexican stock market through the analysis of its variations. *Physica A*, 380:391–398, 2007. doi: 10.1016/j.physa.2007.02.109.

C. J. Corrado. Event studies: A methodology review. *Accounting & Finance*, 51(1):207–234, 2011. doi: https://doi.org/10.1111/j.1467-629X.2010.00375.x.

A. M. Cox and D. G. Hobson. Local martingales, bubbles and option prices. *Finance and Stochastics*, 9:477–492, 2005. doi: 10.1007/s00780-005-0162-y.

G. Cristadoro, M. Degli Esposti, V. Jakšić, and R. Raquépas. Recurrence times, waiting times and universal entropy production estimators. *Letters in Mathematical Physics*, 113(1):19, 2023. doi: 10.1007/s11005-023-01640-8.

L. Y. Dann, D. Mayers, and R. J. Raab. Trading rules, large blocks and the speed of price adjustment. *Journal of Financial Economics*, 4(1):3–22, 1977. doi: 10.1016/0304-405X(77)90034-4.

A. Degutis and L. Novickytė. The efficient market hypothesis: a critical review of literature and methodology. *Ekonomika*, 93:7–23, 2014. doi: 10.15388/Ekon.2014.2.3549.

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–22, 1977. doi: 10.1111/j.2517-6161.1977.tb01600.x.

J. E. Dennis, Jr. and J. J. Moré. Quasi-newton methods, motivation and theory. *SIAM Review*, 19(1):46–89, 1977. doi: 10.1137/1019005.

D. A. Dickey and W. A. Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366a):427–431, 1979. doi: 10.1080/01621459.1979.10482531.

F. X. Diebold and G. Strasser. On the Correlation Structure of Microstructure Noise: A Financial Economic Approach. *The Review of Economic Studies*, 80(4):1304–1337, 2013. doi: 10.1093/restud/rdt008.

A. Dionisio, R. Menezes, and D. Mendes. An econophysics approach to analyse uncertainty in financial markets: an application to the portuguese stock market. *The European Physical Journal B*, 50:161–164, 2006. doi: 10.1140/epjb/e2006-00113-2.

X. Dong, C. Chen, Q. Geng, Z. Cao, X. Chen, J. Lin, Y. Jin, Z. Zhang, Y. Shi, and X. D. Zhang. An improved method of handling missing values in the analysis of sample entropy for continuous monitoring of physiological signals. *Entropy*, 21(3), 2019. doi: 10.3390/e21030274.

S. Drożdż, R. Gębarowski, L. Minati, P. Oświęcimka, and M. Watorek. Bitcoin market route to maturity? evidence from return fluctuations, temporal correlations and multiscaling effects. *Chaos*, 28(7):071101, 2018. doi: 10.1063/1.5036517.

W.-Q. Duan and H. Stanley. Volatility, irregularity, and predictable degree of accumulative return series. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 81:066116, 2010. doi: 10.1103/PhysRevE.81.066116.

A. Dávalos, M. Jabloun, P. Ravier, and O. Buttelli. On the statistical properties of multiscale permutation entropy: Characterization of the estimator's variance. *Entropy*, 21(5), 2019. doi: 10.3390/e21050450.

R. F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007, 1982. doi: 10.2307/1912773.

R. F. Engle. The econometrics of ultra-high-frequency data. *Econometrica*, 68(1):1–22, 2000. doi: 10.1111/1468-0262.00091.

R. F. Engle and C. W. J. Granger. Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, 55(2):251–276, 1987. doi: 10.2307/1913236.

R. F. Engle and J. R. Russell. Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrica*, 66(5):1127–1162, 1998. doi: 10.2307/2999632.

C. Eom, G. Oh, and W.-S. Jung. Relationship between efficiency and predictability in stock price change. *Physica A: Statistical Mechanics and its Applications*, 387(22):5511–5517, 2008. doi: 10.1016/j.physa.2008.05.059.

T. W. Epps and M. L. Epps. The stochastic dependence of security price changes and transaction volumes: Implications for the mixture-of-distributions hypothesis. *Econometrica*, 44(2):305–321, 1976. doi: 10.2307/1912726.

E. F. Fama. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417, 1970. doi: 10.2307/2325486.

E. F. Fama. Efficient capital markets: ii. *The Journal of Finance*, 46:1575, 1991. doi: 10.2307/2328565.

E. F. Fama and M. E. Blume. Filter rules and stock-market trading. *The Journal of Business*, 39(1):226–241, 1966. doi: 10.1086/294849.

E. F. Fama and K. R. French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56, 1993. doi: 10.1016/0304-405X(93)90023-5.

E. F. Fama and K. R. French. Multifactor explanations of asset pricing anomalies. *The Journal of Finance*, 51(1):55–84, 1996. doi: 10.1111/j.1540-6261.1996.tb05202.x.

J. Fan and Y. Wang. Spot volatility estimation for high-frequency data. *Statistics and its Interface*, 1(2):279–288, 2008. doi: 10.4310/SII.2008.v1.n2.a5.

Z. Fluck, B. G. Malkiel, and R. E. Quandt. The Predictability of Stock Returns: A Cross-Sectional Simulation. *The Review of Economics and Statistics*, 79(2):176–183, 1997. doi: 10.1162/003465397556764.

T. Foucault, M. Pagano, and A. Röell. Limit Order Book Markets. In *Market Liquidity: Theory, Evidence, and Policy*. Oxford University Press, 2013. ISBN 9780199936243. doi: 10.1093/acprof:oso/9780199936243.003.0007.

K. R. French. Stock returns and the weekend effect. *Journal of financial economics*, 8(1):55–69, 1980. doi: 10.1016/0304-405X(80)90021-5.

N. Fusari, R. Jarrow, and S. Lamichhane. Testing for asset price bubbles using options data. *Johns Hopkins Carey Business School Research Paper*, 20(12), 2020.

R. Gençay and N. Gradojevic. The tale of two financial crises: An entropic perspective. *Entropy*, 19(6):244, 2017. doi: 10.3390/e19060244.

R. Giglio, R. Matsushita, A. Figueiredo, I. Gleria, and S. D. Silva. Algorithmic complexity theory and the relative efficiency of financial markets. *EPL (Europhysics Letters)*, 84(4):48005, 2008. doi: 10.1209/0295-5075/84/48005.

L. R. Glosten and P. R. Milgrom. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14(1):71–100, 1985. doi: 10.1016/0304-405X(85)90044-3.

E. Goldman. Testing efficiency of the ruble-sterling foreign-exchange market under the gold standard. *Empirical Economics*, 31(2):449–477, 2006. doi: 10.1007/s00181-005-0025-6.

M. D. Gould, M. A. Porter, S. Williams, M. McDonald, D. J. Fenn, and S. D. Howison. Limit order books. *Quantitative Finance*, 13(11):1709–1742, 2013. doi: 10.1080/14697688.2013.803148.

N. Gradojevic and M. Caric. Predicting Systemic Risk with Entropic Indicators. *Journal of Forecasting*, 36:16–25, 2017. doi: 10.1002/for.2411.

P. Grassberger. Finite sample corrections to entropy and dimension estimates. *Physics Letters A*, 128(6-7):369–373, 1988. doi: 10.1016/0375-9601(88)90193-4.

P. Grassberger. Entropy estimates from insufficient samplings. *arXiv: Data Analysis, Statistics and Probability*, 2008.

P. Grassberger. On generalized schürmann entropy estimators. *Entropy*, 24(5):680, 2022. doi: 10.3390/e24050680.

J. E. Griffin and R. C. Oomen. Sampling returns for realized variance calculations: tick time or transaction time? *Econometric Reviews*, 27(1-3):230–253, 2008. doi: 10.1080/07474930701873341.

M. S. Grinblatt, R. W. Masulis, and S. Titman. The valuation effects of stock splits and stock dividends. *Journal of Financial Economics*, 13(4):461–490, 1984. doi: 10.1016/0304-405X(84)90011-4.

S. J. Grossman. Dynamic asset allocation and the informational efficiency of markets. *The Journal of Finance*, 50(3):773–787, 1995. doi: 10.1111/j.1540-6261.1995.tb04036.x.

S. J. Grossman and J. E. Stiglitz. On the impossibility of informationally efficient markets. *The American Economic Review*, 70(3):393–408, 1980.

L. Gulko. The entropic market hypothesis. *International Journal of Theoretical and Applied Finance*, 02(03):293–329, 1999. doi: 10.1142/S0219024999000170.

P. R. Hansen, A. Lunde, A. Lunde, and J. M. Nason. The Model Confidence Set. *Econometrica*, 79(2):453 – 497, 2011. doi: 10.3982/ECTA5771.

B. Harris. The statistical estimation of entropy in the non-parametric case. Technical report, Wisconsin Univ-Madison Mathematics Research Center, 1975.

S. L. Heston. A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options. *The Review of Financial Studies*, 6(2):327–343, 1993. doi: 10.1093/rfs/6.2.327.

D. A. Hsieh. Chaos and nonlinear dynamics: Application to financial markets. *The Journal of Finance*, 46(5):1839–1877, 1991. doi: 10.2307/2328575.

P.-H. Hsu, M. P. Taylor, and Z. Wang. Technical trading: Is it still beating the foreign exchange market? *Journal of International Economics*, 102:188–208, 2016. doi: 10.1016/j.jinteco.2016.03.012.

B. Huang, Y. Huan, L. D. Xu, L. Zheng, and Z. Zou. Automated trading systems statistical and machine learning methods and hardware implementation: a survey. *Enterprise Information Systems*, 13:132–144, 2019. doi: 10.1080/17517575.2018.1493145.

R. Hudson and A. Urquhart. Technical trading and cryptocurrencies. *Annals of Operations Research*, 297(1):191–220, 2021. doi: 10.1007/s10479-019-03357-1.

J. S. Hunter. The exponentially weighted moving average. *Journal of Quality Technology*, 18:203–210, 1986. doi: 10.1080/00224065.1986.11979014.

T. D. Huynh and D. R. Smith. Stock price reaction to news: The joint effect of tone and attention on momentum. *Journal of Behavioral Finance*, 18(3):304–328, 2017. doi: 10.1080/15427560.2017.1339190.

M. Ito and S. Sugiyama. Measuring the degree of time varying market inefficiency. *Economics Letters*, 103:62–64, 2009. doi: 10.1016/j.econlet.2009.01.028.

J. Jacod, Y. Li, and X. Zheng. Statistical properties of microstructure noise. *Econometrica*, 85(4):1133–1174, 2017. doi: 10.3982/ECTA13085.

R. Jarrow, Y. Kchia, and P. Protter. How to detect an asset bubble. *SIAM Journal on Financial Mathematics*, 2(1):839–865, 2011. doi: 10.1137/10079673X.

N. Jegadeesh and S. Titman. Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, 48(1):65–91, 1993. doi: 10.2307/2328882.

M. C. Jensen. Some anomalous evidence regarding market efficiency. *Journal of Financial Economics*, 6(2):95–101, 1978. doi: 10.1016/0304-405X(78)90025-9.

R. H. Jones. Maximum likelihood fitting of arma models to time series with missing observations. *Technometrics*, 22(3):389–395, 1980. doi: 10.1080/00401706.1980.10486171.

T. Kamae. A simple proof of the ergodic theorem using nonstandard analysis. *Israel Journal of Mathematics*, 42:284–290, 1982. doi: 10.1007/BF02761408.

G. L. Kaminsky, J. S. Lizondo, and C. Reinhart. Leading indicators of currency crises. *IMF Staff Papers*, 1998(004):A001, 1998. doi: 10.5089/9781451974515.024.A001.

Y. Katznelson and B. Weiss. A simple proof of some ergodic theorems. *Israel Journal of Mathematics*, 42:291–296, 1982.

A. Kibritçioglu. Monitoring banking sector fragility. *The Arab Bank Review*, 5(2):51–66, 2003.

J. Kiefer. Sequential minimax search for a maximum. *Proceedings of the American Mathematical Society*, 4:502, 1953. doi: 10.2307/2032161.

J. H. Kim and A. Shamsuddin. Are asian stock markets efficient? evidence from new multiple variance ratio tests. *Journal of Empirical Finance*, 15:518–532, 2008. doi: 10.1016/j.jempfin.2007.07.001.

K. K. Kim, H. J. Baek, Y. G. Lim, and K. S. Park. Effect of missing rr-interval data on nonlinear heart rate variability analysis. *Computer Methods and Programs in Biomedicine*, 106(3):210–218, 2012. doi: 10.1016/j.cmpb.2010.11.011.

A. Kolokolov, G. Livieri, and D. Pirino. Statistical inferences for price staleness. *Journal of Econometrics*, 218(1):32–81, 2020. doi: 10.1016/j.jeconom.2020.01.021.

I. Kontoyiannis. Asymptotic recurrence and waiting times for stationary processes. *Journal of Theoretical Probability*, 11(3):795–811, 1998. doi: 10.1023/A:1022610816550.

S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951. doi: 10.1214/aoms/1177729694.

O. Kwon and J.-S. Yang. Information flow between composite stock index and individual stocks. *Physica A: Statistical Mechanics and its Applications*, 387(12):2851–2856, 2008. doi: 10.1016/j.physa.2008.01.007.

A. S. Kyle. Continuous auctions and insider trading. *Econometrica*, 53(6):1315–1335, 1985. doi: 10.2307/1913210.

S. S. Lee and P. A. Mykland. Jumps in Financial Markets: A New Nonparametric Test and Jump Dynamics. *The Review of Financial Studies*, 21(6):2535–2563, 2007. doi: 10.1093/rfs/hhm056.

Y.-J. Lee. The effect of quarterly report readability on information efficiency of stock prices. *Contemporary Accounting Research*, 29(4):1137–1170, 2012. doi: 10.1111/j.1911-3846.2011.01152.x.

B. N. Lehmann. Fads, martingales, and market efficiency. *The Quarterly Journal of Economics*, 105(1):1–28, 1990. doi: https://doi.org/10.2307/2937816.

A. Lempel and J. Ziv. On the complexity of finite sequences. *IEEE Transactions on Information Theory*, 22(1):75–81, 1976. doi: 10.1109/TIT.1976.1055501.

S. F. LeRoy. Risk aversion and the martingale property of stock prices. *International Economic Review*, 14(2):436–446, 1973. doi: 10.2307/2525932.

S. F. LeRoy. Efficient capital markets and martingales. *Journal of Economic Literature*, 27(4):1583–1621, 1989.

D. A. Lesmond, M. J. Schill, and C. Zhou. The illusory nature of momentum profits. *Journal of Financial Economics*, 71(2):349–380, 2004. doi: 10.1016/S0304-405X(03)00206-X.

Z. Liang, Y. Wang, X. Sun, D. Li, L. J. Voss, J. W. Sleigh, S. Hagihira, and X. Li. Eeg entropy measures in anesthesia. *Frontiers in computational neuroscience*, 9:16, 2015. doi: 10.3389/fncom.2015.00016.

F. Lillo and J. D. Farmer. The long memory of the efficient market. *Studies in Nonlinear Dynamics & Econometrics*, 8(3), 2004. doi: 10.2202/1558-3708.1226.

O. Linton and E. Smetanina. Testing the martingale hypothesis for gross returns. *Journal of Empirical Finance*, 38:664–689, 2016. doi: 10.1016/j.jempfin.2016.02.010.

A. Liu, J. Chen, S. Y. Yang, and A. G. Hawkes. The flow of information in trading: An entropy approach to market regimes. *Entropy*, 22(9):1064, 2020. doi: 10.3390/e22091064.

L.-Z. Liu, X.-Y. Qian, and H.-Y. Lu. Cross-sample entropy of foreign exchange time series. *Physica A: Statistical Mechanics and its Applications*, 389(21):4785–4792, 2010. doi: 10.1016/j.physa.2010.06.013.

A. W. Lo. The adaptive markets hypothesis. *The Journal of Portfolio Management*, 30(5):15–29, 2004. doi: 10.3905/jpm.2004.442611.

A. W. Lo and A. C. MacKinlay. Stock market prices do not follow random walks: Evidence from a simple specification test. Working Paper 2168, National Bureau of Economic Research, 1987.

A. W. Lo and A. C. MacKinlay. *A Non-Random Walk Down Wall Street*. Princeton University Press, Princeton, 1999. ISBN 9781400829095. doi: 10.1515/9781400829095.

A. W. Lo, H. Mamaysky, and J. Wang. Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. *The Journal of Finance*, 55 (4):1705–1765, 2000. doi: 10.1111/0022-1082.00265.

B. J. Lobo and D. Tufte. Exchange rate volatility: Does politics matter? *Journal of Macroeconomics*, 20(2):351–365, 1998. doi: 10.1016/S0164-0704(98)00062-7.

M. London, A. Evans, and M. Turner. Conditional entropy and randomness in financial time series. *Quantitative Finance*, 1(4):414–426, 2001. doi: 10.1088/1469-7688/1/4/302.

C. A. Los. Nonparametric efficiency testing of asian stock markets using weekly data. *Centre for Research in Financial Services Working Paper*, 99(01), 1998.

R. E. Lucas. Asset prices in an exchange economy. *Econometrica*, 46(6):1429–1445, 1978. doi: 10.2307/1913837.

M. Magdziarz, S. Orzeł, and A. Weron. Option pricing in subdiffusive bachelier model. *Journal of Statistical Physics*, 145:187–203, 2011. doi: 10.1007/s10955-011-0310-z.

I. Mahmoud, S. Sebai, K. Naoui, and H. Jemmali. Market informational efficiency of tunisian stock market: the contribution of shannon entropy. *European Journal of Economics, Finance and Administrative Sciences*, pages 6–17, 2014.

S. Makridakis and M. Hibon. Arma models and the box–jenkins methodology. *Journal of Forecasting*, 16(3):147–163, 1997. doi: 10.1002/(SICI)1099-131X(199705)16:3⟨147::AID-FOR652⟩3.0.CO;2-X.

F. Malik, B. T. Ewing, and J. E. Payne. Measuring volatility persistence in the presence of sudden changes in the variance of canadian stock returns. *Canadian Journal of Economics/Revue canadienne d'économique*, 38(3):1037–1056, 2005. doi: 10.1111/j.0008-4085.2005.00315.x.

B. G. Malkiel. The efficient market hypothesis and its critics. *Journal of economic perspectives*, 17(1):59–82, 2003. doi: 10.1257/089533003321164958.

A. Mancini, A. Desiderio, R. Di Clemente, and G. Cimini. Self-induced consensus of reddit users to characterise the gamestop short squeeze. *Scientific reports*, 12(1):13780, 2022. doi: 10.1038/s41598-022-17925-2.

A. Mandes. Algorithmic and High-Frequency Trading Strategies: A Literature Review. MAGKS Papers on Economics 201625, Philipps-Universität Marburg, Faculty of Business Administration and Economics, Department of Economics, 2016.

R. Marschinski and H. Kantz. Analysing the information flow between financial time series. *The European Physical Journal B*, 30:275–281, 2002. doi: 10.1140/epjb/e2002-00379-2.

K. Marton and P. C. Shields. Entropy and the Consistent Estimation of Joint Distributions. *The Annals of Probability*, 22(2):960 – 977, 1994. doi: 10.1214/aop/1176988736.

A. Mathai. On noncentral generalized laplacianness of quadratic forms in normal variables. *Journal of Multivariate Analysis*, 45(2):239–246, 1993. doi: 10.1006/jmva.1993.1036.

M. Matilla-García. A non-parametric test for independence based on symbolic dynamics. *Journal of Economic Dynamics and Control*, 31(12):3889–3903, 2007. doi: 10.1016/j.jedc.2007.01.018.

J. L. McCauley, G. H. Gunaratne, and K. E. Bassler. Hurst exponents, markov processes, and fractional brownian motion. *Physica A: Statistical Mechanics and its Applications*, 379(1):1–9, 2007. doi: 10.1016/j.physa.2006.12.028.

F. McGroarty, A. Booth, E. Gerding, and V. Chinthalapati. High frequency trading strategies, market fragility and price spikes: an agent based model perspective. *Annals of Operations Research*, 282(1):217–244, 2019. doi: 10.1007/s10479-018-3019-4.

R. D. Mclean and J. Pontiff. Does academic research destroy stock return predictability? *The Journal of Finance*, 71(1):5–32, 2016. doi: 10.1111/jofi.12365.

W. Mensi, C. Aloui, M. Hamdi, and K. Nguyen. Crude oil market efficiency: An empirical investigation via the shannon entropy. *Économie internationale*, 129:119–137, 2012. doi: 10.3917/ecoi.129.0119.

T. C. Mills, C. Siriopoulos, R. N. Markellos, and D. Harizanis. Seasonality in the athens stock exchange. *Applied Financial Economics*, 10(2):137–142, 2000. doi: 10.1080/096031000331761.

L. Molgedey and W. Ebeling. Local order, entropy and predictability of financial time series. *European Physical Journal B*, 15:733–737, 2000. doi: 10.1007/s100510051178.

R. Morales, T. Di Matteo, R. Gramatica, and T. Aste. Dynamical generalized hurst exponent as a tool to monitor unstable periods in financial time series. *Physica A: statistical mechanics and its applications*, 391(11):3180–3189, 2012. doi: 10.1016/j.physa.2012.01.004.

J. Morgan, J. Longerstaey, R. Limited, R. Ltd, M. Spencer, and M. G. T. C. of New York. *RiskMetrics: Technical Document*. J. P. Morgan, 1996. URL `https://www.msci.com/documents/10199/5915b101-4206-4ba0-aee2-3449d5c7e95a`.

G. Oh, S. Kim, and C. Eom. Market efficiency in foreign exchange markets. *Physica A: Statistical Mechanics and its Applications*, 382(1):209–212, 2007. doi: 10.1016/j.physa.2007.02.032.

G. Oh, H. yong Kim, S.-W. Ahn, and W. Kwak. Analyzing the financial crisis using the entropy density function. *Physica A: Statistical Mechanics and its Applications*, 419:464–469, 2015. doi: 10.1016/j.physa.2014.10.065.

J. Olbryś and E. Majewska. Regularity in stock market indices within turbulence periods: The sample entropy approach. *Entropy*, 24(7), 2022. doi: 10.3390/e24070921.

R. C. A. Oomen. Properties of realized variance under alternative sampling schemes. *Journal of Business & Economic Statistics*, 24(2):219–237, 2006. doi: 10.1198/073500106000000044.

M. Ormos and D. Zibriczky. Entropy-based financial asset pricing. *PLOS ONE*, 9(12):1–21, 2015. doi: 10.1371/journal.pone.0115742.

F. Ouimet. General formulas for the central and non-central moments of the multinomial distribution. *Stats*, 4:18–27, 2021. doi: 10.3390/stats4010002.

A. Pagan. The econometrics of financial markets. *Journal of Empirical Finance*, 3(1):15–102, 1996. doi: 10.1016/0927-5398(95)00020-8.

B. Pandey and S. Sarkar. Testing homogeneity in the Sloan Digital Sky Survey Data Release Twelve with Shannon entropy. *Monthly Notices of the Royal Astronomical Society*, 454(3): 2647–2656, 2015. doi: 10.1093/mnras/stv2166.

C.-H. Park and S. H. Irwin. What do we know about the profitability of technical analysis? *Journal of Economic surveys*, 21(4):786–826, 2007. doi: 10.1111/j.1467-6419.2007.00519.x.

R. Pascoal and A. M. Monteiro. Market efficiency, roughness and long memory in psi20 index returns: Wavelet and entropy analysis. *Entropy*, 16(5):2768–2788, 2014. doi: 10.3390/e16052768.

S. Patra and G. S. Hiremath. An entropy approach to measure the dynamic stock market efficiency. *Journal of Quantitative Economics*, 20(2):337–377, 2022. doi: 10.1007/s40953-022-00295-x.

D. T. Pele, E. Lazar, and A. Dufour. Information entropy and measures of market risk. *Entropy*, 19:226, 2017. doi: 10.3390/e19050226.

P. C. Phillips and J. Yu. Information loss in volatility measurement with flat price trading. *Cowles Foundation Discussion Paper*, 2007.

S. Pincus and R. Kalman. Irregularity, volatility, risk, and financial market time series. *Proceedings of the National Academy of Sciences of the United States of America*, 101:13709–14, 2004. doi: 10.1073/pnas.0405168101.

S. Pincus, I. Gladstone, and R. Ehrenkranz. A regular statistic for medical data analysis. *Journal of clinical monitoring*, 7:335–345, 1991. doi: 10.1007/BF01619355.

F. A. Potra and S. J. Wright. Interior-point methods. *Journal of computational and applied mathematics*, 124(1-2):281–302, 2000. doi: 10.1016/S0377-0427(00)00433-7.

L. Ricci, A. Perinelli, and M. Castelluzzo. Estimating the variance of shannon entropy. *Physical Review E*, 104:024220, 2021. doi: 10.1103/PhysRevE.104.024220.

V. Ricciardi and H. K. Simon. What is behavioral finance? *Business, Education & Technology Journal*, 2(2):1–9, 2000.

J. Riordan. Moment recurrence relations for binomial, poisson and hypergeometric frequency distributions. *The Annals of Mathematical Statistics*, 8:103–111, 1937.

W. A. Risso. The informational efficiency and the financial crashes. *Research in International Business and Finance*, 22:396–408, 2008. doi: 10.1016/j.ribaf.2008.02.005.

W. A. Risso. The informational efficiency: the emerging markets versus the developed markets. *Applied Economics Letters*, 16(5):485–487, 2009. doi: 10.1080/17446540802216219.

S. A. R. Rizvi, G. Dewandaru, O. I. Bacha, and M. Masih. An analysis of stock market efficiency: Developed vs islamic stock markets using mf-dfa. *Physica A: Statistical Mechanics and its Applications*, 407:86–99, 2014. doi: 10.1016/j.physa.2014.03.091.

C. Y. Robert and M. Rosenbaum. A New Approach for the Dynamics of Ultra-High-Frequency Data: The Model with Uncertainty Zones. *Journal of Financial Econometrics*, 9(2):344–366, 2010. doi: 10.1093/jjfinec/nbq023.

N. Rodriguez-Rodriguez and O. Miramontes. Shannon entropy: An econophysical approach to cryptocurrency portfolios. *Entropy*, 24(11):1583, 2022. doi: 10.3390/e24111583.

R. Roll. A simple implicit measure of the effective bid-ask spread in an efficient market. *The Journal of Finance*, 39(4):1127–1139, 1984. doi: 10.1111/j.1540-6261.1984.tb03897.x.

M. d. C. Ruiz, A. Guillamón, and A. Gabaldón. A new approach to measure volatility in energy markets. *Entropy*, 14(1):74–91, 2012. doi: 10.3390/e14010074.

P. A. Samuelson. Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review*, 6(2):41–49, 1965.

T. Schreiber. Measuring information transfer. *Physical Review Letters*, 85:461–464, 2000. doi: 10.1103/PhysRevLett.85.461.

T. Schürmann and P. Grassberger. Entropy estimation of symbol sequences. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 6(3):414–427, 1996. doi: 10.1063/1.166191.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978. doi: 10.1214/aos/1176344136.

A. Sensoy. Generalized hurst exponent approach to efficiency in mena markets. *Physica A: Statistical Mechanics and its Applications*, 392(20):5019–5026, 2013. doi: 10.1016/j.physa.2013.06.041.

S. J. H. Shahzad, S. M. Nor, W. Mensi, and R. R. Kumar. Examining the efficiency and interdependence of us credit and stock markets through mf-dfa and mf-dxa approaches. *Physica A*, 471:351–363, 2017. doi: 10.1016/j.physa.2016.12.037.

D. F. Shanno. Conditioning of quasi-newton methods for function minimization. *Mathematics of Computation*, 24(111):647–656, 1970. doi: 10.1090/S0025-5718-1970-0274029-X.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

P. C. Shields. *The Ergodic Theory of Discrete Sample Paths*. American Mathematical Society, 1996.

R. J. Shiller. From efficient markets theory to behavioral finance. *Journal of Economic Perspectives*, 17(1):83–104, 2003. doi: 10.1257/089533003321164967.

A. Shmilovici, Y. Alon-Brimer, and S. Hauser. Using a stochastic complexity measure to check the efficient market hypothesis. *Computational Economics*, 22:273–284, 2003. doi: 10.1023/A:1026198216929.

A. Shternshis and P. Mazzarisi. Variance of entropy for testing time-varying regimes with an application to meme stocks. *arXiv preprint arXiv:2211.05415*, 2022. doi: 10.48550/arXiv.2211.05415.

A. Shternshis, P. Mazzarisi, and S. Marmi. Measuring market efficiency: The shannon entropy of high-frequency financial time series. *Chaos, Solitons & Fractals*, 162:112403, 2022a. doi: 0.1016/j.chaos.2022.112403.

A. Shternshis, P. Mazzarisi, and S. Marmi. Efficiency of the moscow stock exchange before 2022. *Entropy*, 24(9):1184, 2022b. doi: 10.3390/e24091184.

R. R. Sokal and C. D. Michener. A statistical method for evaluating systematic relationships. *Univ. Kansas, Sci. Bull.*, 38:1409–1438, 1958.

D. Stosic, D. Stosic, T. Ludermir, W. de Oliveira, and T. Stosic. Foreign exchange rate entropy evolution during financial crises. *Physica A: Statistical Mechanics and its Applications*, 449:233–239, 2016. doi: 10.1016/j.physa.2015.12.124.

B. Strait and T. Dewey. The shannon information entropy of protein sequences. *Biophysical Journal*, 71(1):148–155, 1996. doi: 10.1016/S0006-3495(96)79210-X.

G. Sucarrat and S. Grønneberg. Risk estimation with a time-varying probability of zero returns. *Journal of Financial Econometrics*, pages 1–32, 2020. doi: 10.1093/jjfinec/nbaa014.

R. Susmel. Switching volatility in private international equity markets. *International Journal of Finance & Economics*, 5(4):265–283, 2000. doi: 10.1002/1099-1158(200010)5:4⟨265::AID-IJFE132⟩3.0.CO;2-H.

L. A. Teixeira and A. L. I. De Oliveira. A method for automatic stock trading combining technical analysis and nearest neighbor classification. *Expert systems with applications*, 37(10):6885–6890, 2010. doi: 10.1016/j.eswa.2010.03.033.

L. Telesca, M. Lovallo, G. Romano, K. I. Konstantinou, H.-L. Hsu, and C.-c. Chen. Using the informational fisher–shannon method to investigate the influence of long-term deformation processes on geoelectrical signals: An example from the taiwan orogeny. *Physica A: Statistical Mechanics and its Applications*, 414:340–351, 2014. doi: 10.1016/j.physa.2014.07.060.

N. M. Temme. Large parameter cases of the gauss hypergeometric function. *Journal of Computational and Applied Mathematics*, 153(1):441–462, 2003. doi: 10.1016/S0377-0427(02)00627-1. Proceedings of the 6th International Symposium on Orthogonal Polynomials, Special Functions and their Applications, Rome, Italy, 18-22 June 2001.

A. Treves and S. Panzeri. The upward bias in measures of information derived from limited data samples. *Neural Computation*, 7(2):399–407, 1995. doi: 10.1162/neco.1995.7.2.399.

C. Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of Statistical Physics*, 52:479–487, 1988. doi: 10.1007/BF01016429.

Y. Tsutsui, K. Hirayama, T. Tanaka, and N. Uesugi. Can We Make Money with Fifth-order Autocorrelation in Japanese Stock Prices? ISER Discussion Paper 0639, Institute of Social and Economic Research, Osaka University, 2005.

J. D. Victor. Asymptotic bias in information estimates and the exponential (bell) polynomials. *Neural Computation*, 12(12):2797–2804, 2000. doi: 10.1162/089976600300014728.

G. T. Walker. On periodicity in series of related terms. *Proceedings of the Royal Society A*, 131: 518–532, 1931. doi: 10.1098/rspa.1931.0069.

F. M. Willems. Universal data compression and repetition times. *IEEE Transactions on Information Theory*, 35(1):54–58, 1989. doi: 10.1109/18.42176.

L. V. Williams. Information efficiency in betting markets: a survey. *Bulletin of Economic Research*, 51(1):1–39, 1999. doi: 10.1111/1467-8586.00069.

P. Wolfe. Convergence conditions for ascent methods. *SIAM Review*, 11(2):226–235, 1969. doi: 10.1137/1011036.

R. A. Wood, T. H. McInish, and J. K. Ord. An investigation of transactions data for nyse stocks. *The Journal of Finance*, 40(3):723–739, 1985. doi: 10.2307/2327796.

A. D. Wyner and J. Ziv. Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression. *IEEE Transactions on Information Theory*, 35 (6):1250–1258, 1989. doi: 10.1109/18.45281.

G. Yen and C.-f. Lee. Efficient market hypothesis (emh): Past, present and future. *Review of Pacific Basin Financial Markets and Policies*, 11(02):305–329, 2008. doi: 10.1142/S0219091508001362.

X. Zhao, M. Ji, N. Zhang, and P. Shang. Permutation transition entropy: Measuring the dynamical complexity of financial time series. *Chaos, Solitons & Fractals*, 139:109962, 2020. doi: 10.1016/j.chaos.2020.109962.

H. Zhu and Z. Liu. On bivariate time-varying price staleness. *Journal of Business & Economic Statistics*, 0(0):1–14, 2023. doi: 10.1080/07350015.2023.2174547.

J. Ziv. Coding theorems for individual sequences. *IEEE Transactions on Information Theory*, 24(4):405–412, 1978. doi: 10.1109/TIT.1978.1055911.

A. M. Zubkov. Limit distributions for a statistical estimate of the entropy. *Theory of Probability & Its Applications*, 18(3):611–618, 1974. doi: 10.1137/1118080.

L. Zunino, B. M. Tabak, A. Figliola, D. G. Pérez, M. Garavaglia, and O. A. Rosso. A multifractal approach for stock market inefficiency. *Physica A: Statistical Mechanics and its Applications*, 387(26):6558–6566, 2008. doi: 10.1016/j.physa.2008.08.028.

Z. Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967. doi: 10.1080/01621459.1967.10482935.