Applied Mathematics and Computation 268 (2015) 399–412

Contents lists available at ScienceDirect

Applied Mathematics and Computation

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

Compositional segmentation of time series in the financial markets

Yi Yin, Pengjian Shang*, Jianan Xia

Department of Mathematics, Beijing Jiaotong University, No. 3 of Shangyuan Residence Haidian District, Beijing, 100044, PR China

ARTICLE INFO

Keywords:
Entropic segmentation
Segments
Time irreversibility
Detrended fluctuation analysis (DFA)
Financial markets

ABSTRACT

We introduce an entropic segmentation algorithm and apply it to decompose the financial sequences into compositionally homogeneous domains. To probe more about the nature of the financial time series, we investigate the statistical properties of the segment from the view of segmentation position and segment length first. We reveal some important and interesting conclusions and information hidden in these time series of stock markets. Then, we focus on the study of the intrinsic properties for each segment in the time series from two aspects: time irreversibility and correlation. The fluctuations on the time irreversibility and the scaling exponent all support that the segments present compositional heterogeneity and verify the segmentation. Meanwhile, we conclude that time irreversibility is inherent in the stock time series and verifies that stock markets are nonequilibrium systems essentially even though segmentation. Moreover, the scaling exponents for each segment point out that the traditional detrended fluctuation analysis is not applicable to measure the correlation for the whole original time series of stock market.

1. Introduction

Recently, stock markets have become active areas and attracted much attention. Stock market indices are important measures of financial and economical performance and are normally used to benchmark the performance of stock portfolios. The financial market has been referred to as an example of a complex system consisting of many interacting components. It is important to explore the knowledge containing in the stock market indices. There are nonstationary time series of data with statistical properties that change under time translations in many different fields, while the time series of stock market indices is exactly such kind of nonstationary time series in the financial field. Thus, the series of data is heterogeneous which means that one can obtain different values of the mean, standard deviation or higher moments depending on the time interval where they are calculated. It is more complicated for the analysis of such time series because many statistical methods are valid based on the assumption of stationarity of the analyzed data. The correlation existing in the correlated time series can be easily misidentified as nonstationarities [1]. Hence, sequence structure can be adequately revealed through segmentation algorithms. The article [2] proposed one segmentation algorithm, which is conceptually simple and computationally efficient. This method can decompose a sequence into homogeneous subsequences (patches or domains), while different partitions of the sequence are obtained at different statistical significance levels by different threshold value. The homogeneity vanishes when the method is applied to

* Corresponding author. Tel.:+86 10 51684410; fax: +86 13641228092.
E-mail address: pjshangdu@163.com (P. Shang).

http://dx.doi.org/10.1016/j.amc.2015.06.061

more complex, long-range correlated sequences. By relaxing the threshold value, we segment the homogeneous subsequences under a previous higher threshold value further and find new domains within other domains.

As for the segmentation, we can consider it as a "detrending procedure" in the sense that it can be applied to remove the effects of nonstationarity (e.g. daily periodicities in solar irradiation, sleep–wake differences in heart rate, etc.) and investigate the more subtle fluctuations that may reveal intrinsic correlation properties of the dynamics of the system under study [3–7]. However, the nonstationarity itself also is regarded as an important feature of the complex system. As is known to us, for physiological time series, their nonstationarity properties change from healthy to pathological conditions [8–11] and under different physiological states [12–18], in DNA the regions with higher concentrations or densities of a certain dinucleotide (CpG islands) are related to the presence of genes [19–21]. Meanwhile, the dynamics of human gait [22] and the human electrocardiogram [23] can be analyzed by nonlinear time series analysis methods. Besides, it has been shown that the greater the inhomogeneity in the distribution of a word along a text the higher its relevance to the text [24,25] and the expansion–contraction of the economy relates closely to different volatility periods of stock market records [26], while the distribution of periods with different Internet activities is closely related to the congestion state of the net [27].

It is noticed that traditional methods are adequate for mostly stationary stochastic processes, while they can be problematic for complex nonstationary time series [28] because these methods cannot distinguish sequences of different complexities directly. As a result, there is a strong demand for a new measure which is able to deal with heterogeneity containing in the complex nonstationary time series. Objectiveness, mathematical tractability and consistence with the intuitive notion of what the complexity is about [29] are the requirements for a good definition of complexity. Whereas, the algorithmic complexity [30] (maximum for randomness), other derived measures [31] and those based on mutual information [32] are not completely satisfactory. In fact, internal property accounts for the number of domains and their compositional heterogeneity rather than any reference to environments or ensembles.

Therefore, the identification of the different compositional patches in a sequence is an important issue in current study of complex system, since it may be one of the key steps in understanding the inner structure of complex system. Simultaneously, the segmentation algorithm has been validated for the artificially generated sequences in the previous studies [2,33,34]. Consequently, in this paper, we employ the segmentation algorithm, based on the Jensen–Shannon divergence [35], which can be used to segment these sequences into statistically significant, compositionally homogeneous patches, to the time series of US and Chinese stock markets. The statistical properties of the segment should be investigated to learn more about the nature of the sequence in the complex system. Besides, it is of great interest to detect the change on the time irreversibility, which is a fundamental property of far-from-equilibrium systems related to the unidirectionality of the energy flow across the boundaries of the system [36], and long-range correlations for the compositional patches obtained by segmenting a sequence by means of the asymmetry index (Ai) [37] and detrended fluctuation analysis (DFA) [38].

The article is structured as follows: in Section 2, we introduce the segmentation algorithm. Section 3 shows data description and Section 4 provides the statistical properties of applying the segmentation algorithm based on the Jensen–Shannon divergence to the financial time series and the results and analysis of the time irreversibility and long-range correlation for the compositional patches. Finally, we conclude the paper in Section 5.

2. Entropic segmentation method

Here we define compositional domains as subsequences with a different base composition in comparison to the two adjacent subsequences, at a given level of statistical confidence. The domains obtained by this definition are neighborhood dependent. Thus, for comparing adjacent subsequences and deciding the segmentation positions which divide different domains, it is necessary to use some measure to quantify the difference between compositional domains. Hence, our aim is to divide a sequence into segments in such a way as to maximize the compositional divergence between the resulting domains. For this purpose, we need a segmentation method which is capable of detecting shifts in sequence composition and thus locating the possible borders or edges between adjacent domains.

In spite of the widely-used sliding window along the sequence, it can be found that the detection of diffuse borders of variable extent betrays an undesirable dependence on the sliding-window size used to scan the sequence [39]. As a result, we believe that an overall view of the whole sequence may better serve a decision about segmentation, while performing the segmentation procedure iteratively is able to allow a halt at any point and be computationally simple. The segmentation procedure applied in this paper can be described as follows.

First, we introduce a measure to make a segmenting decision. In this paper, we use the Jensen–Shannon divergence measure [35] to quantify the distance between probability. Let $S = \{a_1, a_2, \ldots, a_N\}$ be a sequence composed of N symbols from the alphabet $A = \{A_1, \ldots, A_k\}$; take a given position n $(1 \le n < N)$ and consider the two resulting subsequences:

$$S^{(1)} = \{a_1, a_2, \ldots, a_n\}, \quad S^{(2)} = \{a_{n+1}, a_{n+2}, \ldots, a_N\}$$

and let

$$\mathscr{F}^{(1)} = \left\{f_1^{(1)}, \ldots, f_k^{(1)}\right\}, \quad \mathscr{F}^{(2)} = \left\{f_1^{(2)}, \ldots, f_k^{(2)}\right\}$$

be the respective vectors of relative symbol frequencies, i.e., $f_i^{(1)}$ is the relative proportion of the symbol A_i in $S^{(1)}$ and $f_i^{(2)}$ is the relative proportion of the symbol A_i in $S^{(2)}$.