Contents lists available at ScienceDirect

Physica A

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

Compositional segmentation and complexity measurement in stock indices

Haifeng Wang*, Pengjian Shang, Jianan Xia

Department of Mathematics, School of Science, Beijing Jiaotong University, Beijing 100044, PR China

HIGHLIGHTS

- We introduce a complexity measure based on the entropic segmentation called sequence compositional complexity (*SCC*) into the analysis of financial time series.
- We find that the values of *SCC* of some mature stock indices are likely to be lower than the *SCC* values of some immature index data.
- If we classify the indices with the method of *SCC*, the financial market of Hong Kong has more similarities with mature foreign markets than Chinese ones.

ARTICLE INFO

Article history:
Received 2 April 2015
Received in revised form 13 June 2015
Available online 8 September 2015

Keywords:
Sequence compositional complexity
Entropic segmentation
Financial systems
Long-range correlation

ABSTRACT

In this paper, we introduce a complexity measure based on the entropic segmentation called sequence compositional complexity (*SCC*) into the analysis of financial time series. *SCC* was first used to deal directly with the complex heterogeneity in nonstationary DNA sequences. We already know that *SCC* was found to be higher in sequences with long-range correlation than those with low long-range correlation, especially in the DNA sequences. Now, we introduce this method into financial index data, subsequently, we find that the values of *SCC* of some mature stock indices, such as *S*&*P*500 (simplified with *S*&*P* in the following) and *HSI*, are likely to be lower than the *SCC* value of Chinese index data (such as *SSE*). What is more, we find that, if we classify the indices with the method of *SCC*, the financial market of Hong Kong has more similarities with mature foreign markets than Chinese ones. So we believe that a good correspondence is found between the *SCC* of the index sequence and the complexity of the market involved.

1. Introduction

There is a common agreement that economical and financial systems can be considered as complex systems of multiple interacting subsystems. So concepts and methodologies that solve problems in complex systems from other fields such as physics, biology, etc., are transferred naturally to analyze the economical and financial data. Among them, the methods of complexity measure from physical sciences are the ones that are most frequently transferred to financial area. For instance, Kyoung-jae Kim proposed the *GA* (Genetic Algorithm) approach for instance selection in artificial neural networks for financial forecasting in 2000 [1]; Huaning Hao studied the short-term forecasting of stock price based on genetic-neural network [2]; Hong and Stein [3] studied interaction among market participants from the perspective of behavioral finance [3]; Ying Shangjun, Wei Yiming depicted the complexity of the stock market based on cellular automata and fractal

[*] Corresponding author.
E-mail address: 14121602@bjtu.edu.cn (H. Wang).

http://dx.doi.org/10.1016/j.physa.2015.08.057
0378-4371/© 2015 Elsevier B.V. All rights reserved.

structure features and got a conclusion that there is a obvious correlation between the complexity of the stock market and the psychology of investors [4]. In recent studies, the Becks model is employed to elucidate price fluctuations in markets and is related to superstatistics [5]. What is more, methods on the study of correlations in time series are also taken into attention [6–13].

These researches on the complexity of stock data are still at a groping stage, a common solution or a well-developed method to quantify the complexity of financial sequence is still far from agreed as the results of these methods have a strong correlation with the sample data, and once we put our research into data of samples with various characteristics, even data of one sample during different periods, there would be dramatic changes among the results. In order to deal with this problem, a new method which can surmount the timing difference and variety difference should be recommended.

A good definition of complexity must be objective and mathematically tractable, yet consistent with the intuitive notion of what the complexity is about. In this paper, we introduce the *SCC* (Sequence Compositional Complexity) to deal directly with the complex heterogeneity in nonstationary financial sequences. The plot of *SCC* as a function of significance level provides a profile of sequence structure at different length scales [14]. It is a good measure to deal with the complex heterogeneity of nonstationary financial sequences directly. And by means of it, we cannot only acquire the value of quantification of the complexity but also adequately reveal the sequence through segmentation algorithms addicted to *SCC*.

The structure of the paper is organized as follows. In the rest part of this paper, we will first introduce the *SCC* in detail. Then we symbolize the data from different financial markets and analyze them through SCC method. Next, we compare the profiles of these sequences structure to get some different features in the index data between financial markets in China and developed countries of the west. Finally, we draw conclusions from the discussions above and present them in the last part.

2. *SCC* method

Our aim is to measure the complexity of a sequence by dividing it into segments in such a way so as to maximize the compositional divergence between the compositional domains, so we define compositional domains at a given level of statistical confidence with a different base composition compared to the two subsequences which are divided adjacently and neighborhood dependent [15–20]. Some measures of the difference between compositions need to be found to compare adjacent subsequences and decide whether they are different domains. In order to check out the difference and whether the adjacent subsequences are different domains, we use the *Jensen–Shannon* divergence measure to meet this demand [14].

For two subsequences S_1 and S_2,

$$JS_2(S_1 - S_2) = H[S] - \left(\frac{l_1}{L}H[S_1] + \frac{l_2}{L}H[S_2]\right) \geq 0 \tag{1}$$

where l_1, l_2 are the lengths of S_1 and S_2, $L = l_1 \bigoplus l_2$ (concatenation), and $H[\cdot] = -\sum p \log_2 p$ is the Shannon entropy of the probability distribution p obtained from base frequencies in the corresponding subsequence. The value of n corresponding to the maximum JS_2 along a given sequence segment is assumed to separate subsegments with the maximum compositional differentiation between them. Therefore, the compositional homogeneity is higher within each of the two resulting subsegments than in the parental segment. By doing this, we obtain a value that can represent the complexity difference between the new subsequences and their parental series.

For each index, we first separate it into two part as S_1 and S_2 at the position of $j = 2$, then let $j = j + 1$ and repeat this process until $j = N - 1$, N is the length of the whole sequence, for each position, we obtain a value of JS_2, and we define the max of them as the $JS_2(1)$ of this whole sequence. We note the position of the max value as k_1, and divided by it, we obtain two subsequences. Next, we find the JS_2 of each sequence, and define the max of them as $JS_2(2)$, and note that position as k_2. Repeat this process for n times, we can separate the original sequence more and more stable and gain bigger value of JS_n which is defined as

$$JS_n(s) = \sum_{k=1}^{n-1} JS_2(k). \tag{2}$$

In order to obtain the compositional complexity of the sequence, we should separate the original sequence as much as possible. But JS_n increases with separating times which means that we can get the max value of JS_n if we separate the sequence for $N - 1$ times into N parts. But this makes no sense. So we should set a significant level s, and rely on s, we can decide whether the segment should be taken or not. Statistical confidence is established by calculating the probability that the given divergence value (or lower) appears in a random sequence (with the same length and base composition), once it has been randomly split. For short subsequences, the probability is exactly computed from the hypergeometric distribution; for large ones, Ref. [15] gives us a method to obtain the statistic corresponding value of different significances.

Profiles for pure random sequences begin to depart appreciably from zero for low s values ($s < 80\%$), meaning that the profiles are being contaminated by spurious complexity, due to statistical fluctuations. Therefore, we choose the range $80\% < s < 100\%$ for all sequence comparisons [14].

The JS_n measure fits the requirements for representing the compositional complexity of the sequence. What is more, through the segmentation, we obtain homogeneous subsequences from the original non-stationary one. For the application of this trait into financial time series, we will discuss it in the further study.