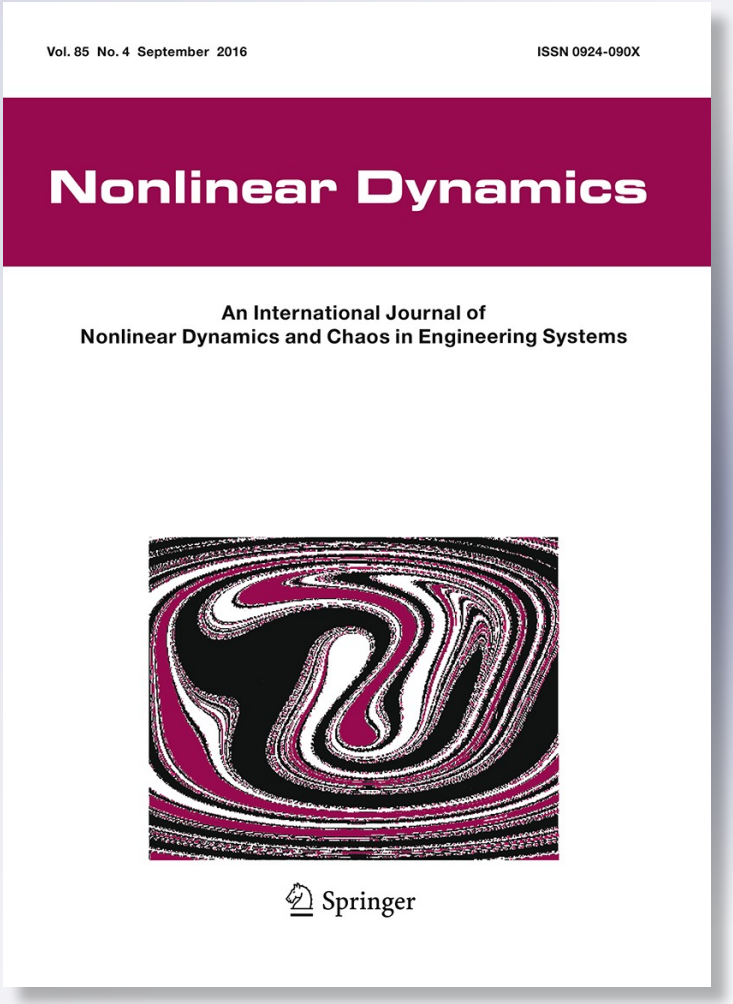
The similarity analysis of financial stocks based on information clustering

Qiang Tian, Pengjian Shang & Guochen Feng

Nonlinear Dynamics
An International Journal of Nonlinear Dynamics and Chaos in Engineering Systems

ISSN 0924-090X
Volume 85
Number 4

Nonlinear Dyn (2016) 85:2635-2652
DOI 10.1007/s11071-016-2851-9



Springer

Nonlinear Dyn (2016) 85:2635–2652
DOI 10.1007/s11071-016-2851-9

ORIGINAL PAPER

The similarity analysis of financial stocks based on information clustering

Qiang Tian · Pengjian Shang · Guochen Feng

Received: 1 September 2015 / Accepted: 13 May 2016 / Published online: 26 May 2016

Abstract Similarity in time series is an important feature of dynamical systems such as financial systems, with potential use for clustering of series in system. Here, we mainly introduce a novel method: the reconstructed phase space information clustering method to analyze the financial markets. The method is used to examine the similarity of different sequences by calculating the distances between them, which the main difference from previous method is the way to map the original time series to symbolic sequences. Here we make use of the state space reconstruction to construct the symbolic sequences and quantify the similarity of different stock markets and exchange rate markets considering the chaotic behavior between the complex time series. And we compare the results of similarity of artificial and real data using the modified method, information categorization method and system clustering method. We conclude that the reconstructed phase space information clustering method is effective to research the close relationship in time series and for short time series especially. Besides, we report the results of similarity of different exchange rate time series in different periods and find the effect of the exchange rate regime in 2008 on the time series. Also we acquire some characteristics of exchange rate time series in China market, especially for the top four trading partners of China.

Keywords Reconstructed phase space · Time delay · Embedding dimension · Similarity index · Phylogenetic tree

1 Introduction

In recent years, there has been a lot of interest in the research of time series especially for financial markets, which have becoming active and attract more and more attention. In general, the major research of financial markets focuses on the stock markets and exchange rate markets. These markets are remarkably well-defined complex systems with a large number of interacting units that conform to the underlying economic trends. Physicists are currently contributing to the modeling of complex systems by using tools and methodologies developed in statistical mechanics and theoretical physics. Econophysics [1–3] is the term used to denote the application of statistical mechanics to economic systems. A range of methods have been introduced to investigate financial time series as a reflection of economic trends. Typical applications on time series deal with tasks such as clustering similarity search, and prediction. And the similarity between financial time series is specific to the application domain and also an important feature of the dynamics of financial markets.

Q. Tian · P. Shang (✉) · G. Feng
Department of Mathematics, School of Science, Beijing Jiaotong University, Beijing 100044, People's Republic of China
e-mail: pjshang@center.njtu.edu.cn

As of now, many various methods have been used to quantify the similarity in time series. Pincus proposed the approximate entropy (ApEn) to quantify the concept of changing complexity [4–6], and it was applied to measure the biologic time series [7,8]. Moreover, Richman et al. [9,10] analyzed the shortcomings of the ApEn method and introduced a new method called sample entropy (SampEn) under a broad range of conditions; it was widely applied in clinical cardiovascular studies. After that the Cross-ApEn and Cross-SampEn were put forward to measure the similarity of two distinct time series [11], and Cross-SampEn then was applied to investigate synchronism and cross-correlation of stock markets [12]. All of methods above were constructed based on quantifying the regularity of time series, and aimed at estimating the system complexity of financial markets initially. Furthermore, Costa et al. [13–15] introduced the multiscale entropy by taking account of the timescales and applied it to measure the complexity of biologic systems. Then, multiscale cross-sample entropy [16] was proposed to analyze the similarity of two series under different timescales based on multiscale entropy and the Cross-SampEn. For further analysis, multiscale time irreversibility [17] was also proposed to classify the financial markets and obtained similar results. In addition, there were lots of other methods to analyze the similarity of different stock markets, such as detrended Cross-Correlation Analysis [18], multiscale detrended fluctuation analysis (MSDFA) and multiscale detrended cross-correlation analysis (MSDCCA) [19], three-phase clustering method [20] and Time-Varying Copula-GARCH Model [21]. Also there were some studies on the property of the exchange rate fluctuations in exchange rate markets. Previous researches mostly addressed on the characteristics of their fluctuations adopting diverse functional forms, such as power laws [22–24], Gaussian function [25], and superimposed Gaussian function [26]. And the detrended fluctuation analysis (DFA) was used to investigate the scaling behavior in the fluctuations of exchange rates [25,27–29]. For some foreign exchange markets, it had been found that the exchange rate variations possess the multifractal nature by applying the multifractal detrended fluctuation analysis (MFDFA) [30], structure functions [31–34], and multifractal model of asset returns (MMAR) [35]. Besides, someone studied the statistical and multifractal properties of the yuan exchange rate index based on the DMA and MFDMA methods [36]. On the other hand, a new method called the multifractal asymmetric detrended cross-correlation analysis method (MF-ADCCA) [37] to investigate the scaling behavior of the cross-correlations among the Chinese stock market, the China exchange market, and the US stock market.

Defining similarity is nontrivial. In 2003, Yang et al. [38,39] proposed the method by the measurement of the similarity between two complex sequences. They developed a novel information-based similarity index to detect and quantify hidden dynamical structures in the human heart rate time series using tools from statistical linguistics. For time series tend to be very long, it was a good method that the time series can be mapped to binary symbolic sequences and the dissimilarity index can be calculated through rank frequency. They also applied the method of constructing phylogenetic trees [40]to arrange different groups of samples on a branching tree to best fit the pairwise distance measurements. For better consequences, Peng et al. proposed another definition for the weighting factor by using Shannon entropy [41], and they gave another definition of the dissimilarity index; then, they applied the novel definition for an information categorization approach [42], biologic signals [43], SARS Coronavirus [44], and the patterns of blood pressure signals [45].

Here in our study, we also propose a novel clustering method to calculate the similarity between the time series based on state space reconstruction. State space reconstruction is the first step in nonlinear time series analysis of data from chaotic systems including estimation of invariants and prediction. It can make full use of the characteristics of the whole time series and not just consider the adjacent points of the time series. The state space reconstruction was proposed by Packard [46], and now the method is widely used in the chaotic time series. As an example of chaotic time series, we can make use of the method to map financial time series to symbolic time series; it should be a good way to detect the underlying features of them.

In previous studies, a number of works focused on the analysis of physiologic signals by using the method above. The measurement of similarity is a linguistic analysis algorithm, and it is a kind of pattern analysis method. Based on this method, we take account of the situation of stock time series and exchange time series by selecting proper values of parameters, and

then, we analyze the similarity of these two markets. However, most of previous methods are applied to the time series directly and the length of the time series should be same. The novel method is based on symbolic sequences, so we do not need to consider the length of sequences, just choosing the proper m-bit words. Furthermore, the financial market dynamics are driven by a number of complex factors: index at the same level, the subindex, the economic data, trading sentiment, trading prices, weight, and other stock information. For this type of intrinsically noisy system, it may be useful to simplify the dynamics via mapping the output to symbolic sequences, where we choose state space reconstruction in comparison with the binary sequences denoted by 1 and 0 [47]. The resulting symbolic sequence retains important features of the dynamics generated by the underlying dynamical system, but it is tractable enough to be analyzed as a symbolic sequence. So we turn the complex stock time series to symbolic sequences, which is the first significant step to apply this method. And the state space reconstruction takes into account the whole characteristics of each letter in word not just the adjacent one. In addition, traditional clustering method with Euclidean distances just mentioned the difference of values at different time, neglecting the dynamics characters of time series. The novelty of the information clustering method is that it incorporates elements of both information-based and word statistics-based categories, because the rank-order difference of each word statistics is weighted by its underlying information using Shannon entropy. Furthermore, the composition of these basic elements captures global information related to usage of respective elements in financial time series. Also, the distance plots and phylogenetic trees based on the dissimilarity index can give us direct information about different markets to analyze the similarity among them.

The reminder of the paper is organized as follows. In the following section, we present the details of the novel reconstructed phase space information clustering method. Section 3 describes the data used in our work, including the Autoregressive Fractionally Integrated Moving Average (ARFIMA) series, stock time series, and exchange rate time series in Chinese market. In Sect. 4, we compared the clustering results obtained using various similarity measures such as information categorization method (ICM), the reconstructed phase space information (RPSI) clustering method, and system clustering with squared Euclidean distances (SCE), and also study various currencies against yuan exchange rate and the effect of exchange rate regime in 2008 on the similarity of the top four exchange rate time series in China exchange rate market. Finally, we summarize the findings of this paper in the last section.

2 Reconstructed phase space information cluster method

In 2007, Peng et al. [43] proposed the information categorization method, which is the modification based on the method proposed by Yang et al. [38,39]. The method provided a new measurement of the similarity between two time series. But this method just considers the correlation of adjacent value for the m-bit word, not accounting for the relationship about every value. For it maps the time series to binary symbolic sequences by Eq. 1.

$$I_n = \begin{cases} 0 & \text{if} \quad x_n \le x_{n-1}, \\ 1 & \text{if} \quad x_n > x_{n-1}. \end{cases} \tag{1}$$

which x_t is the value of time series at time t; then, we can get a new symbol series I_n.

To counterweight these facts, we propose another way to map the original series to symbol time series. Here we build the symbol time series by state space reconstruction [46]. State space reconstruction is the first step in nonlinear time series analysis of data from chaotic systems including estimation of invariants and prediction.

Therefore, we expand the one-dimensional time series into a higher- dimensional reconstructed phase space (RPS) and construct a new method called reconstructed phase space information(RPSI) clustering method. For a given time series $\{x_t\}_{t=1}^{N}$, where x_t is the value in time t. The method can be summarized as follows:

Step 1 Let $X_j^{m,\tau}$ be the set of points x_t from j to $j+(m-1)\tau$, and $X_j^{m,\tau} = \{x_j, x_{j+\tau}, \ldots, x_j + (m-1)\tau\}$ where $j = 1, 2, \ldots, T-(m-1)\tau$, and $m \ge 2$ and $\tau \ge 1$ denote, respectively, the embedding dimension and time delay. Each of the $N = T-(m-1)\tau$ subvectors is a sin-

gle motif out of possible ones (representing all unique orderings of m different real numbers).

Step 2 Let $\omega_k^{m,\tau}$ be the permutation of $X_j^{m,\tau}$ in a nondecreasing order using the alphabet $A = \{0, 1, \ldots, m-1\}$, which is called an m-bit word. For the permutation of $X_j^{m,\tau}$ may be same, here $\omega_k^{m,\tau}$ is unique ordering for them.

Step 3 Count the occurrences of different words, and then sort them in descending order by frequency of occurrences. Then, we can get the value of $p_1(\omega_k^{m,\tau})$and $R_1(\omega_k^{m,\tau})$, which represents the frequency and rank of occurrences of word $\omega_k^{m,\tau}$ in time series S_1, $p_2(\omega_k^{m,\tau})$ and $R_2(\omega_k^{m,\tau})$ for time series S_2 analogously.

Step 4 Calculate the distance $D_m(S_1, S_2)$ between the two time series which incorporate the likelihood of each word. The distance is defined as Eq. (2),

$$D_m(S_1, S_2) = \frac{1}{m! - 1} \sum_{k=1}^{m!} \left| R_1(\omega_k^{m,\tau}) - R_2(\omega_k^{m,\tau}) \right| F(\omega_k^{m,\tau}) \tag{2}$$

where

$$F(\omega_k^{m,\tau}) = \frac{1}{Z}[-p_1(\omega_k^{m,\tau}) \log p_1(\omega_k^{m,\tau}) - p_2(\omega_k^{m,\tau}) \log p_2(\omega_k^{m,\tau})] \tag{3}$$

And the normalization factor Z is given by

$$Z = \sum\nolimits_j [-p_1(\omega_k^{m,\tau}) \log p_1(\omega_k^{m,\tau}) - p_2(\omega_k^{m,\tau}) \log p_2(\omega_k^{m,\tau})] \tag{4}$$

Similarly, the sum is divided by the value $m! - 1$ to keep the value in the same range of [0, 1].

For phase space reconstruction, it is very important to detect the proper value of embedding dimension and time delay. If the embedding dimension was too small, some neighbor points may be close to each other because of the projection from some higher dimension down to their lower dimension. Instead, higher embedding dimension might lead to excessive computation during assessment of parameters. Also, the selection of time delay affects the result of analysis. In our study, we choose the method of false nearest neighbors [48] to determine the minimum embedding dimension. And the first local minimum of the average mutual information function proposed in [49] is a good way to identify the best value for time delay.

3 Data

3.1 Artificial time series

According to the Autoregressive Fractionally Integrated Moving Average(ARFIMA) models [50], we construct several series of length N. As we know, the ARFIMA models can model the cross-correlation between two ARFIMA series for any given strength of coupling between them, according to the equations below:

$$x_t = [WX_t + (1 - W)Y_t] + \varepsilon_t \tag{5}$$

$$y_t = [(1 - W)X_t + WY_t] + \varepsilon_t' \tag{6}$$

$$X_t = \sum_{n=1}^{\infty} a_n(d_1) x_{t-n} \tag{7}$$

$$Y_t = \sum_{n=1}^{\infty} a_n(d_2) y_{t-n} \tag{8}$$

$$a_n(d) = d\Gamma(n - d)/(\Gamma(1 - d)\Gamma(n + 1)) \tag{9}$$

where ε_t and ε_t' are two different i.i.d. Gaussian variables with zero mean and variance = 1. $\Gamma(x)$is the Gamma function, and the scaling parameters d_1 and d_2 vary with the range of (−0.5, 0.5). The parameter W denotes the strength of the coupling and varies from 0.5 to 1, where $W = 0.5$ gives the highest cross-correlation, while $W = 1$ represents the total absence of correlation.

In our simulation, we have a collection of different time series generated by ARFIMA models to analyze the similarity among them and construct phylogenetic trees for these series.

3.2 Stock time series of different areas

The time series obtained from various stock markets consist of the daily records of ten stock indices listed in Table 1 during the period 1991–2013. We obtain data from the website [51].com. And the lengths of ten stock time series are all different from each other because

Table 1 The list of ten stock indices

Area	Index
Asia	1 ShangZheng
	2 ShenCheng
	3 HSI
	4 Nikkei225 (Japan)
America	5 S&P500
	6 NASDAQ
	7 DJIA
Europe	8 DAX (Germany)
	9 CAC40 (France)
	10 FTSE (U.K.)

they are belong to different areas and the number is all above 5000.

3.3 Exchange rate time series of Chinese foreign exchange market

On July 21, 2005, the People's Bank of China (PBC) reformed the exchange rate regime by moving into a managed floating exchange rate regime based on market supply and demand with reference to a basket of currencies. Up to the end of 2011, the top four trading partners of China were the European Union, the US, Japan, and Hong Kong. Along with further reforms and opening-up in China, the US dollar (USD), euro (EUR), British pound (GBP), and Japanese yen (JPY) acted as the invoicing and settlement currencies in the Chinese foreign trade, while the trade contracted between China and nondollar countries become significantly more frequent [36]. And the trading relation between China and more areas become closer than before, it enhanced the exchange trade with yuan exchange rate, such as the Australian dollar (AUD), Canadian dollar (CAD), Hong Kong dollar (HKD), Malaysian ringgit (MYR), and Russian rouble (RUB). Therefore, we select the USD/CNY (Chinese Yuan), HKD/CNY, 100JPY/CNY, EUR/CNY, GBP/CNY, AUD/CNY, CAD/CNY, CNY/MYR, and CNY/RUB exchange rates as the foreign exchange rates in the Chinese foreign exchange market. Then, we can get 9 exchange rate time series as analysis object. Furthermore, the empirical data in this paper choose the daily median price of various exchange rates from November 28, 2011, to October 23, 2014, in our study.

4 Analysis and results

In this part, we performed experiments to analyze the ability of information clustering to distinguish between ARFIMA time series, and also for different financial time series. We compared the clustering results obtained using three similarity measures: information categorization method (ICM), the reconstructed phase space information (RPSI) clustering method, and system clustering with squared Euclidean distances (SCE). Then experiments were conducted on both simulated and real datasets to analyze the properties among them using the information clustering method.

4.1 Comparison of three clustering methods

We perform clustering on a database of ARFIMA time series and analyze the results. We generate four groups which consist of 100 ARFIMA time series, with the parameters $(d,W) = [(0.5,0.5),(0.5,1),(0.1,1),(0.1,0.5)]$ for the time series in each group, respectively. We form ten collections from these time series and run clustering on each of the groups. Collections 1–6 are built by selecting two from these four groups.

Here we use the similarity metric to measure the accuracy and quality of clustering. The procedure is as follows:

Given two clusterings $G = G_1, \ldots, G_k$(say the "ground truth")and $A = A_1, \ldots, A_k$(obtained using clustering method), the cluster similarity metric is defined as $Sim(G, A) = (\sum_i \max_j Sim(G_i, A_j))/k$, where $Sim(G_i, A_j) = 2|G_i \cap A_j|/(|G_i| + |A_j|)$. Then, $Sim(G, A)$ can be used to evaluate the clustering results [52]. Note that this similarity measure will return 0 if the two clusterings are completely dissimilar and 1 if they are the same, and the results range from 0 to 1, when it approaches 1 it means the result better. Therefore, we calculate $Sim(G, A)$ for the six collections with the three different methods in Table 2.

Table 2 shows the cluster similarity metric obtained when each of the collections was clustered using the different similarity measures. The cluster similarity metric using the RPSI is the highest and is always above 0.60. This indicates that the objects are clustered with a high confidence level. And the results of system clustering are the lowest comparing with others. The above

Table 2 The quality of clustering results with three methods (ICM, RPSI, and SCE)

Collection	ICM	RPSI	SCE
1	0.500	0.600	0.467
2	0.600	0.633	0.467
3	0.567	0.633	0.533
4	0.700	0.700	0.533
5	0.600	0.667	0.500
6	0.500	0.633	0.500

results affirm that RPSI clustering method is better than the two other distance measures.

Similarly, we get results from stock markets and exchange rate markets in comparison with the three methods and find the advantage of the information categorization method. Therefore, we use the RPSI clustering method to analyze the characters of artificial time series and financial time series, and we also acquire the results with the other two methods for comparison in the next section.

4.2 Numerical results for artificial time series

In our study, we generate a set of ARFIMA series with $N = 5000$ data using four different values of the parameters $d, W) = [(0.5, 0.5), (0.5, 1), (0.1, 1), (0.1, 0.5)]$. In our study, each couple of ARFIMA series is generated with the same parameter d. And we construct ten series for each four parameters. Pervious research presented that the cross-correlation between sequences is highest when $W = 0.5$, and it would decrease as W approached 1. So we choose four sets of series as parameters above by applying the ICM and RPSI to analyze the properties for different parameters. As we know, if two time series are similar in their rank order of the words, the scattered points will be located near the diagonal line. Therefore, the average deviation of these scattered points away from the diagonal line, which means that greater distance indicates less similarity and vice versa.

For ICM, the parameter m has little difference in the result when the length of series is long enough, so we choose $m = 8$ in our study. For each couple of ARFIMA series, we compute the distances of each two sequences among the ten series and obtained 45 distances of these series for each parameter. Table 3 presents the means, minimum, maximum, standard deviations (SDs), and coefficients of variation (CVs) of the distances for four kinds of parameters. In this case, the d parameter is set to $d = 0.1$, W parameter is set to $W = 1$, and the calculated distances are quite different with others, although the mean value of the distances is small but the SDs and CVs are the highest between them. And the gap between the minimum and maximum is large, the difference of similarity among them is large too. When $d = 0.5$, the distance for $W = 1$ is larger than for $W = 0.5$, as expected that the similarity for $W = 1$ is smaller. It shows the similar result for $d = 0.1$. However, the difference for various parameter d is not obvious when we choose the same parameter W. Like the ICM, we then map the ten series to symbolic time series by using RPSI clustering method. The construction of phase space requires the specification of values for the time delay and the embedding dimension m. In this case, the embedding dimension selected by false nearest neighbors algorithm is 5 and the time delay we choose $\tau = 1$ by computing the mutual information function. Also we calculated the means, minimum, maximum, standard deviations (SDs), and coefficients of variation (CVs) of the distances in Table 4. Similarly, the distances for $W = 1$ are clearly larger than for $W = 0.5$, and the SDs or CVs do not change obviously for the four condition. It reflects that this method is better to analyze the similarity between them. And we can also find that the distances for $d = 0.1$ are larger than for $d = 0.5$, which corresponds that the similarity decreases when the value of parameter d decreases.

Table 3 The list of means, minimum, maximum, SDs, and CVs of distances at different d and W for ICM

d	W	Means	Minimum	Maximum	SDs	CVs
0.5	0.5	0.103425	0.088188	0.118044	0.007002	0.067703
0.5	1	0.112378	0.07853	0.167742	0.025419	0.226189
0.1	1	0.100069	0.053202	0.253018	0.074182	0.741312
0.1	0.5	0.063301	0.052849	0.078084	0.005425	0.085706

Table 4 The list of means, minimum, maximum, SDs, and CVs of distances at different d and W for the RPSI clustering method

d	W	Means	Minimum	Maximum	SDs	CVs
0.5	0.5	0.120415	0.093545	0.150244	0.012246	0.101696
0.5	1	0.130218	0.102717	0.155061	0.011933	0.09164
0.1	1	0.303678	0.247383	0.355265	0.022955	0.075591
0.1	0.5	0.290572	0.231943	0.33936	0.023838	0.082039

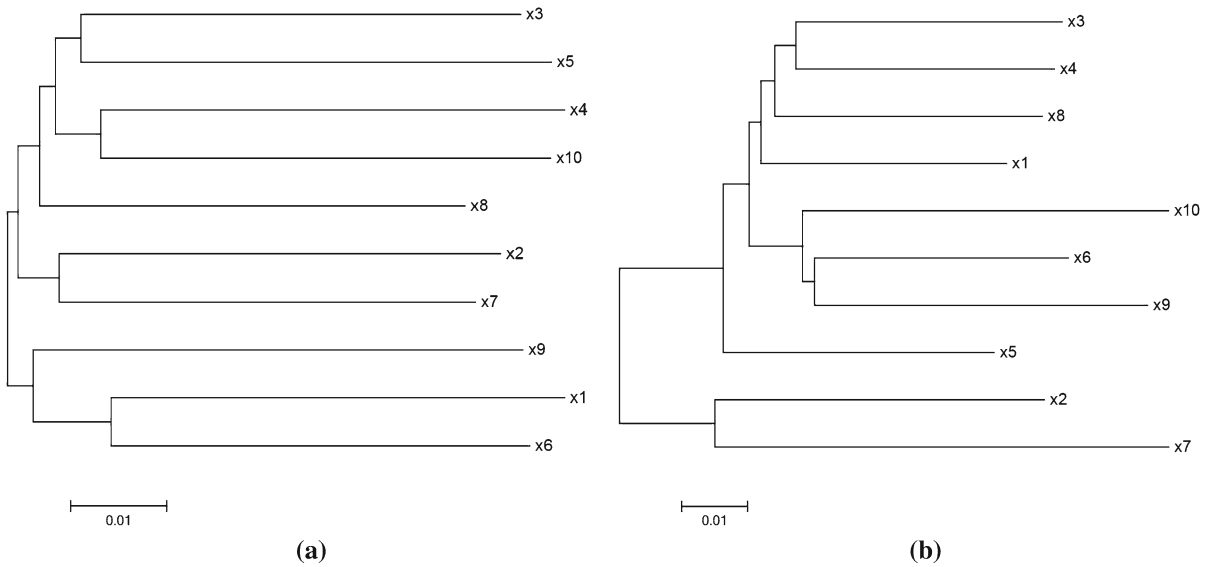


Fig. 1 Phylogenetic trees generated according to the distances between ten ARFIMA series with parameters **a** $d = 0.5$ and $W = 0.5$, **b** $d = 0.5$ $W = 1$ when applying ICM for $m = 8$

But the result of the dependence of parameter d on the distance may be misleading when applying the first method.

A phylogenetic tree is a branching diagram or "tree" showing the inferred relationships among various sequences based upon similarities and differences in their physical characteristics. To present the relationship among the ten simulated series for different parameters, we construct phylogenetic trees based on the distances by using the information categorization method(ICM) and PRSI clustering method. Figures 1 and 3 present four phylogenetic trees of the ten ARFIMA series with four different parameters when applying ICM. And Figs. 2 and 4 show the phylogenetic trees of the same series by using PRSI clustering method for $m = 5$ and $\tau = 1$. In these trees, we can conclude that the distances for $W = 1$ are larger than the distances for $W = 0.5$, which means that the similarity for $W = 1$ is smaller than for $W = 0.5$. And the lengths of branch with the case for $d = 0.5$ and $W = 0.5$ are closely to each other, and also for $d = 0.1$ and $W = 0.5$. It reflects that the similarity of sequences is higher for the case of $W = 0.5$, which corresponds to the definition of AFRIMA models. In order to contrast of results of ICM and RPSI clustering method, we compare the trees by merging Fig. 1 with Fig. 2, and Fig. 3 with Fig. 4. We can conclude that the results of the two methods are approximate when choosing the same value of $d = 0.5$, but the results of ICM are more obvious than RPSI clustering method when choosing $d = 0.1$. The ARFIMA series are randomly generated and consist of high complexity, so when we analyze the similarity of them, the phylogenetic trees is just not enough and we should consider the means, minimum, maximum, SDs, and CVs of distances calculated by using the two methods.

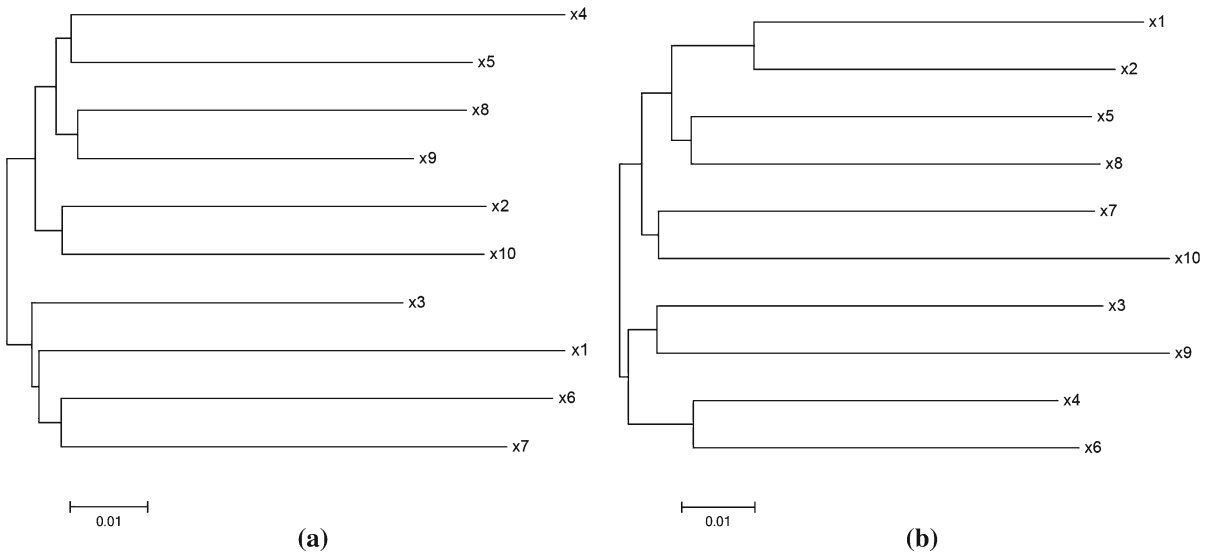


Fig. 2 Phylogenetic trees generated according to the distances between ten ARFIMA series with parameters **a** $d = 0.5$ and $W = 0.5$, **b** $d = 0.5$ and $W = 1$ when applying RPSI clustering method for $m = 5$ and $\tau = 1$

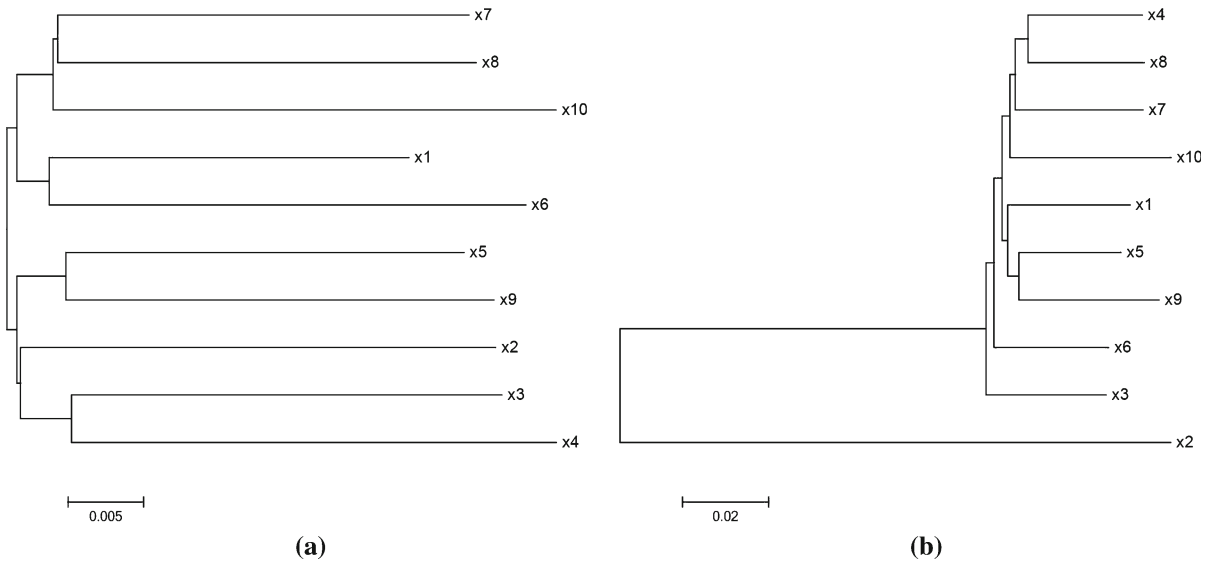


Fig. 3 Phylogenetic trees generated according to the distances between ten ARFIMA series with parameters **a** $d = 0.1$ and $W = 0.5$, **b** $d = 0.1$ and $W = 1$ when applying ICM for $m = 8$

4.3 Dependence of distances on different parameters

To better clarify the dependence of distances to both the strength of coupling W and the scaling parameter τ, we show how distance varies for different values of these two parameters. And we also calculate the distances by using the two methods above with different parameters to analyze how these parameters impact on the results.

Fig. 5 represents the curves of distances obtained by using the two methods with fixed scaling parameter $d = 0.4$ and changing the coupling values W from 0.5 to 1, with step 0.1. And Fig. 6 presented the distances curves with fixed coupling values $W = 0.9$ and changing the scaling parameter d from 0.1 to 0.5, with step 0.1.

The curves in Fig. 5 reflect the distances are getting higher as the coupling decreases, or, in other words,

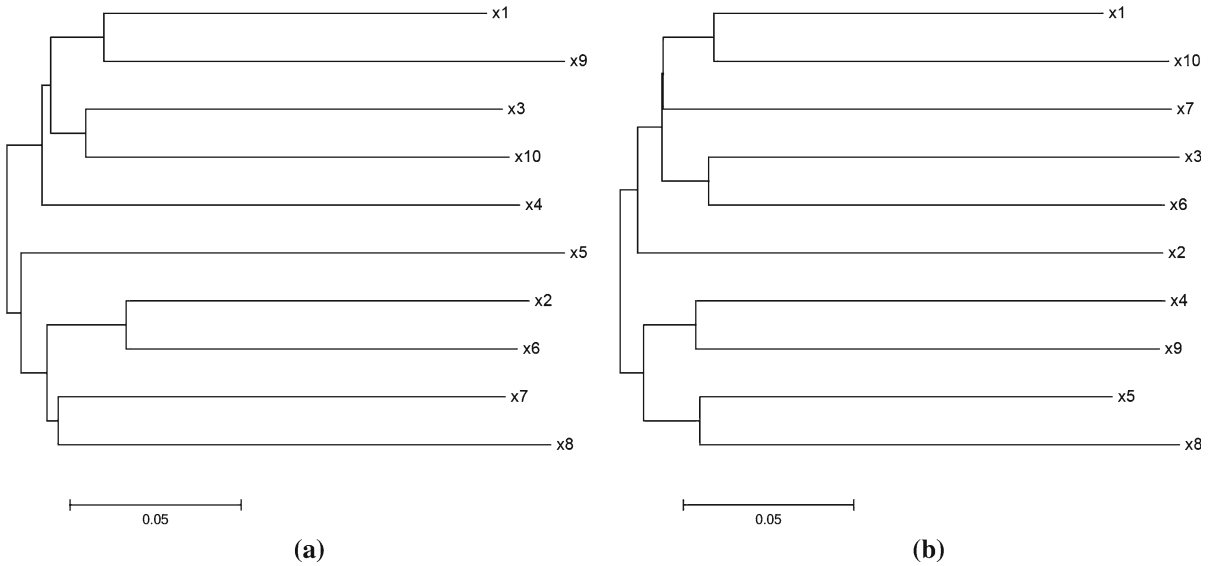


Fig. 4 Phylogenetic trees generated according to the distances between ten ARFIMA series with parameters **a** $d = 0.1$ and $W = 0.5$, **b** $d = 0.1$ and $W = 1$ when applying RPSI clustering method for $m = 5$ and $\tau = 1$

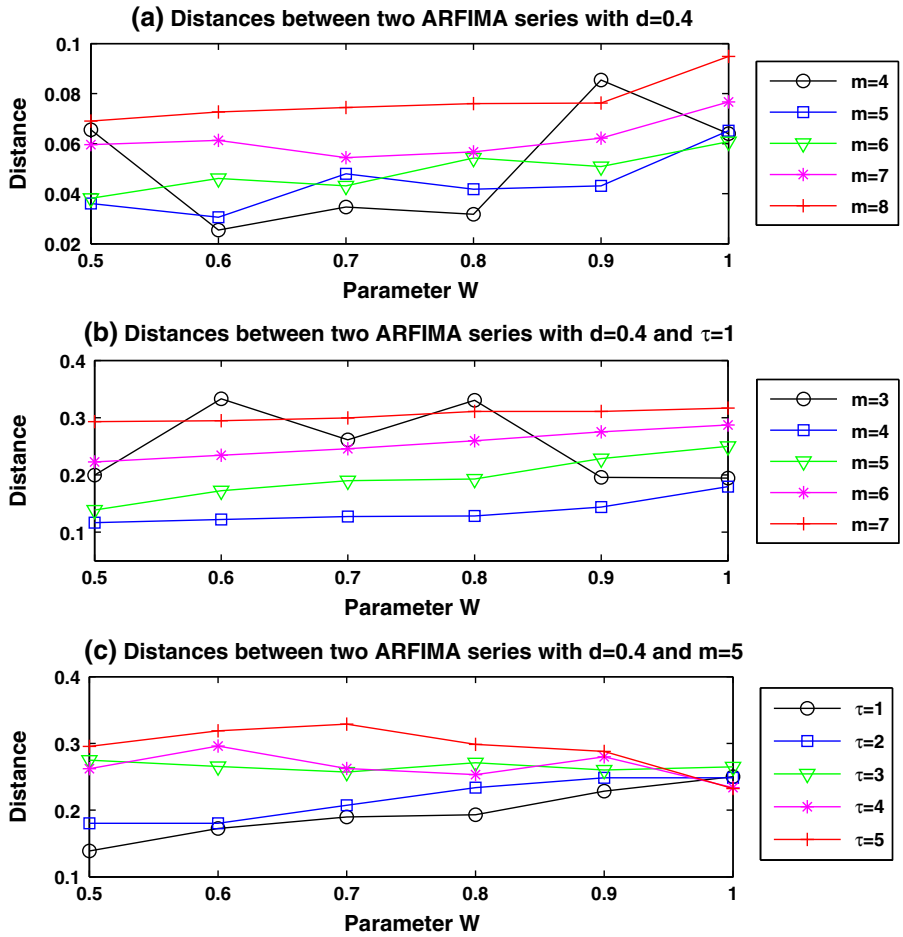


Fig. 5 The distances between two ARFIMA series with fixed scaling parameter $d = 0.4$ and changing the coupling values W from 0.5 to 1 **a** for ICM method with changing parameters m from 4 to 8, **b** for RPSI clustering method with fixed time delay $\tau = 1$ and changing embedding dimension m from 3 to 7, **c** for RPSI clustering method with fixed embedding dimension $m = 5$ and changing time delay $\tau = 1$ to $\tau = 5$

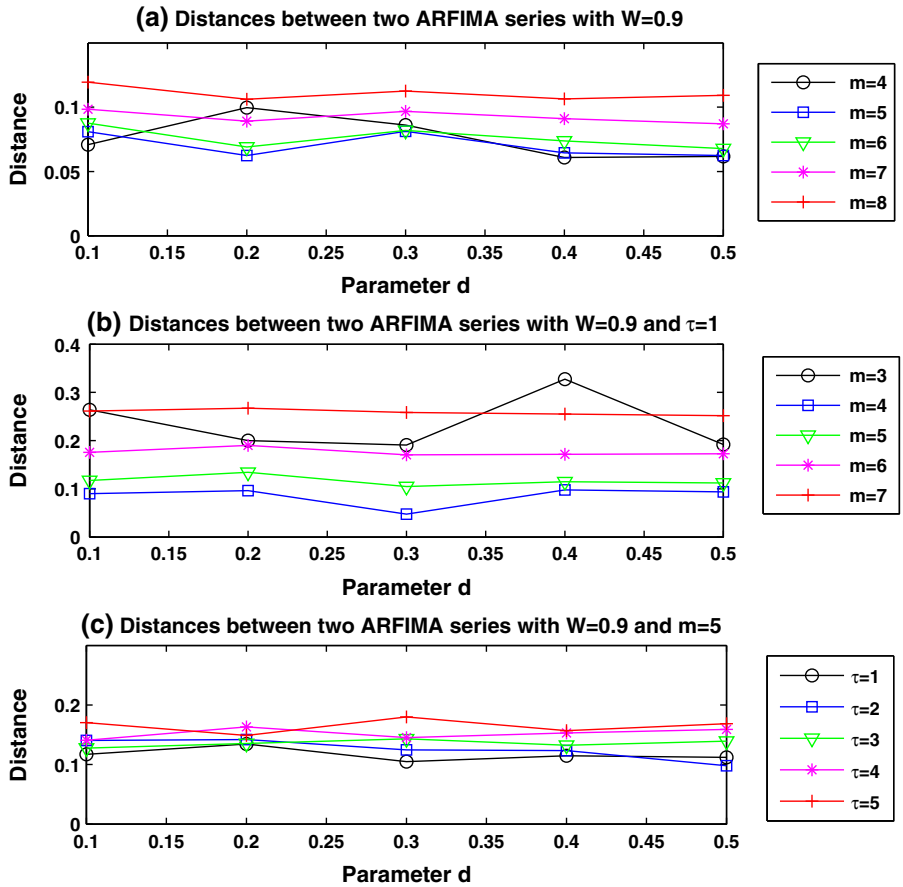


Fig. 6 The distances between two ARFIMA series with fixed coupling values $W = 0.9$ and changing the scaling parameter d from 0.1 to 0.5 **a** for ICM method with changing parameters m from 4 to 8, **b** for RPSI clustering method with fixed time delay $\tau = 1$ and changing embedding dimension W from 3 to 7, **c** for RPSI clustering method with fixed embedding dimension $W = 0.9$ and changing time delay $\tau = 1$ to $\tau = 5$

as W increases, which means the similarity is getting lower as W increases. Besides, we can find the results with the parameter $m = 8$ for ICM method and parameters $m = 5$, $\tau = 1$ for RPSI clustering method are close to ideal values. Fig. 5a presents that the lower values of m may influence the results, which Fig. 5b, c reflects the results may be misleading when we do not choose the optimal embedding dimension and time delay selected by false nearest neighbors algorithm and mutual information function. And the distances of ICM are lower than that of RPSI clustering method, which may be not good for us to construct appropriate phylogenetic trees. The curves in Fig. 6 reflect the distances are approximate for different parameters d, which means the similarity is closer as d changes. Similarly, the change trends with the parameter $m = 8$ for ICM method and parameters $m = 5$, $\tau = 1$ for RPSI clustering method are flatter than others.

Therefore, we can conclude that the two clustering methods are applicable to other the ARFIMA models with different parameters. And we can find the results with optimal embedding dimension $m = 5$ and time delay $\tau = 1$ for RPSI clustering method are more obvious than others, so it is important to calculate the optimal parameters using the false nearest neighbors algorithm and mutual information function when we choose the RPSI clustering method. By contrast, the optimal parameter m for ICM we select just considering the length of the time series may lead inaccurate results to analyze the similarity between series.

4.4 Analysis of stock markets in different areas

We investigate ten stock time series in three different areas from April 3, 1991, to December 31, 2013. Different stock time series in different areas show different characteristics among them, and there may be related to others. First, we use ICM to investigate the underlying relationship between them. Therefore, we map the

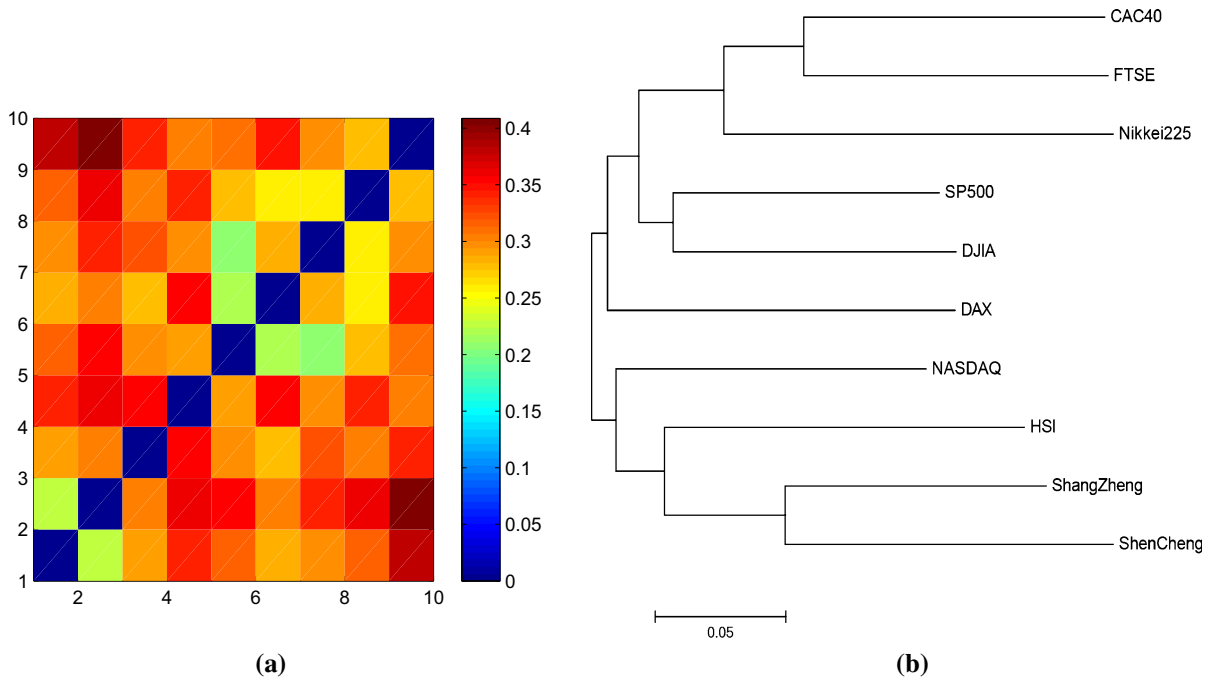


Fig. 7 Distance plots (**a**) and Phylogenetic tree (**b**) generated according to the distances between ten stock indices from 1991 to 2013 with information categorization method for $m = 8$

Table 5 The list of means, minimum, maximum, and SDs of distances between stock time series for two methods

Method	Means	Minimum	Maximum	SDs	CVs
ICM	0.310701	0.210135	0.408706	0.043538	0.140127
RPSI	0.059079	0.045673	0.075843	0.006837	0.115734

ten stock time series to binary symbolic time series. Then, we calculate the distances between the symbolic time series for $m = 8$. Fig. 7 shows the distance plots and phylogenetic tree of the 10 stock indices for $m = 8$. The two figures reflect the high similarity of the time series in America stock markets, also in Chinese stock markets. And the similarity between America and Europe markets is closer than that with Chinese markets. For the Japan stock markets are influenced heavier by America and Europe markets than by Chinese markets, so it locates in the two markets which is away from Chinese markets. However, due to its unique economic development frame, Nikkei225 locates a single branch. Besides, the color of distance plots also reflect the same result as the phylogenetic trees.

To compare the RPSI clustering method with ICM, we select the values $m = 5$ and $\tau = 1$ for RPSI clustering method. Similarly, we calculate the distances among these ten stock series and present their means, minimum, maximum, standard deviations (SDs), and coefficients of variation (CVs) of them in Table 5. In Table 5, we can get the results that the distances among these stock time series with the ICM are obviously higher than the RPSI clustering method, and the consequence in the first method fluctuates significantly. Though the mean and SDs of for the RPSI clustering method are smaller than the first method, the similar value of CVs show that the two method are both useful for the analysis. If we want to reflect the distribution of points against diagonal line clearly, it should select the RPSI clustering method.

Besides, Fig. 8 presents the distance plots and phylogenetic tree of the ten stock indices for optimal $m = 5$ and $\tau = 1$ selected by the methods as in Sect. 4.2 with RPSI clustering method. And the results of other parameters m and τ are not good. These neighboring stocks

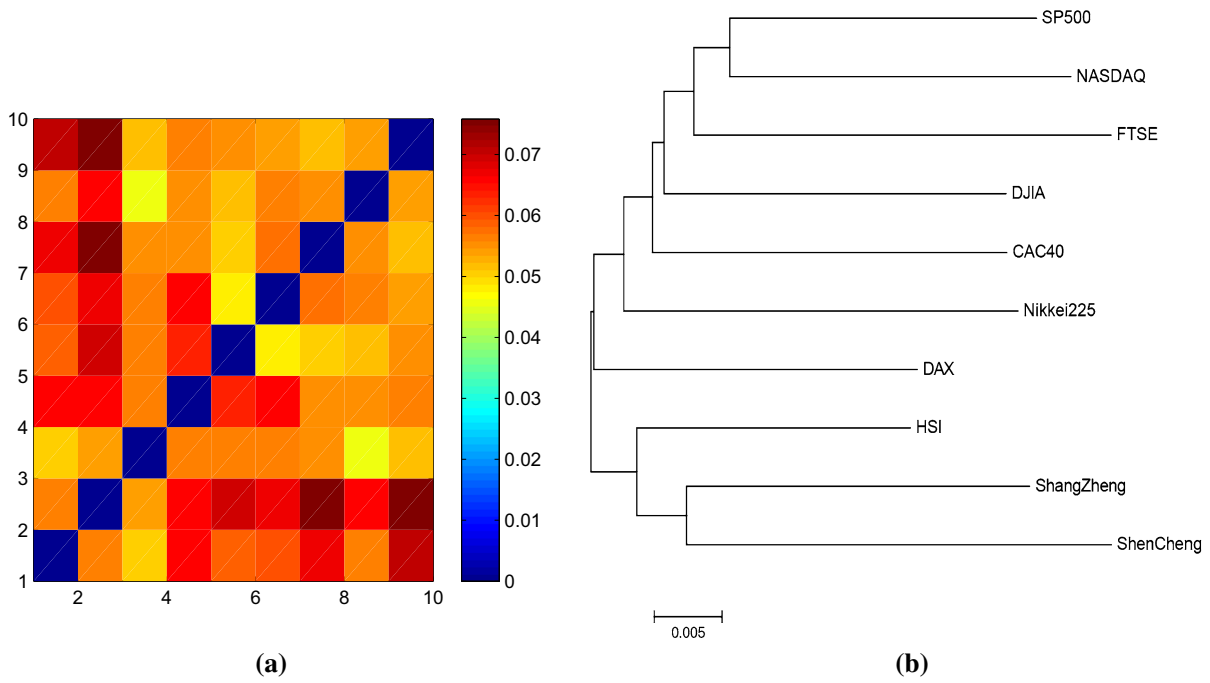


Fig. 8 Distance plots (**a**) and Phylogenetic tree (**b**) generated according to the distances between ten stock indices from 1991 to 2013 with RPSI clustering method for $m = 5$ and $\tau = 1$

time series may share many common features. And the results with RPSI clustering method reflect more obviously.

We can find both Figs. 7b and 8b show that these time series are dived by two clusters. One cluster is from Asian area except for Nikkei225 index, and another cluster is from America and Europe areas. But for Fig. 7b the branch of NASDAQ index may be misleading for the results. Furthermore, the difference of colors for the distance plots can also reflect the similarity between these stock time series obviously. These markets present small difference in the same area, and large difference in different area. In conformity with the previous study using ICM, one can observe that the America and Europe stock markets are belong to a big branch, and ShangZheng, ShenCheng, and HSI indices are belong to one branch but not with the Nikkei225 index.

As we know, the Hong Kong stock market and Japan stock market are two specific markets in Asian. Both of them have some relations with America and Europe markets, but they are still influenced by the mainland of China. However, the economy of Hong Kong market is more close to Chinese market, so it located the branch which belongs to the Chinese market. And the Nikkei225 index locates a single branch for its specificity as same as the conclusion for ICM. In spite of this, the clustering of these stock markets is in accord with actual situation. The distribution of tree corresponds to the actual stock markets similarly.

In addition, we get the tree from the distances calculated by the system clustering in Fig. 9a. And the clustering result for different areas is not obvious as the ICM and RPSI clustering method, and it reflects the method is not suitable to the similarity measurement of these stock time series. The results of cluster similarity metric using the RPSI clustering method also show the modified method is better to analyze the similarity between these stock time series.

4.5 Analysis of Chinese foreign exchange market

Just like the procedure in Sect. 4.2 and 4.4, we apply the RPSI clustering method mentioned in Sect. 2 and ICM to analyze the similarity between exchange rate time series. The length of time series we select is 702; it is far less than those in Sects. 4.2 and 4.4. If we

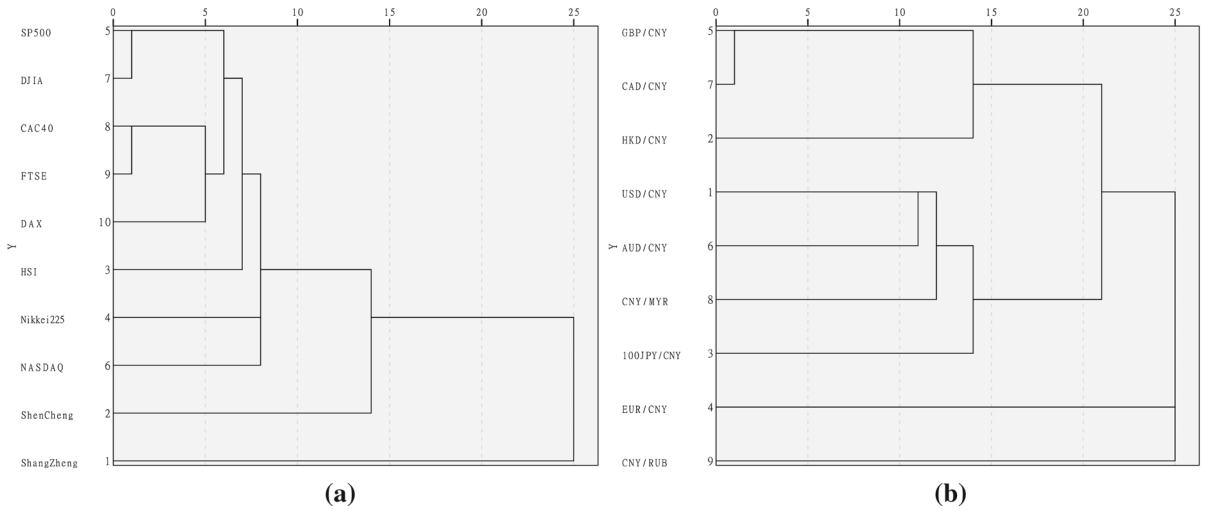


Fig. 9 Trees generated according to the distances between ten stock indices from 1991 to 2013 (**a**) and exchange rates time series (**b**) with system clustering method

still choose the parameter $m = 8$ for ICM, the distance is so small that we cannot show the relationship between them. Therefore, we select the maximum value of m but the distances of them are not too small. Through several experiments, we choose $m = 6$ finally. Figure 10 displays the distance plots and phylogenetic trees for the nine exchange rate time series by the ICM with $m = 6$. And Fig. 11 presents the distance plots and phylogenetic trees for the nine exchange rate time series by RPSI clustering method with $m = 5$ and $\tau = 1$. Table 6 shows the means, minimum, maximum, SDs, and CVs of distances by applying these two methods.

From Fig. 10, we can investigate the similarity between the time series clearly. It reflects that USD/CNY and HKD/CNY exchange rate shares closer relation, then with CNY/RUB and these three belong to one branch. Then, 100JPY/CNY and GBP/CNY, EUR/CNY, CNY/MYR, AUD/CNY and CAD/CNY locate another two branches. For Fig. 11, we can find that the relationship of USD/CNY and HKD/CNY, 100JPY/CNY and GBP/CNY, AUD/CNY, and CAD/CNY are similar to the result of the first method. Only CNY/RUB, EUR/CNY, and CNY/MYR's location change. And Table 6 reflects that the distances of the ICM method are larger than the RPSI clustering method, as the number of words we select for the RPSI clustering method is 120, but 64 for ICM. More kinds of word may cause more diversity of distribution of these words; it will cause the decrease in the distances. However, the values of CVs for the two methods are similar; it shows that these two methods are both available to the similarity analysis of time series. We still can find some phenomenon from the results reflected in the figure by these two methods. As known to all, the US dollar, euro, British pound, and Japanese yen act as the top four currencies in the Chinese foreign trade, it means that there are close relationship between them, we can get the similar result from the figure roughly. Also, all the figures suggest that the similarity between US dollar and Hong Kong dollar is higher than others, because the Hong Kong government considered pegging Hong Kong dollar to US dollar in case of significantly dropping of rate after 1980s. The USD/CNY and HKD/CNY almost have the same tendency from the data for many years. Every country has own exchange rate arrangement, it influenced on the price of exchange rate. The study is based on dataset of People's Bank of China, and we take account of different currencies against yuan exchange, but if one selects another exchange rate from different countries, we may get different clusterings. Above all, if the time series are short like the exchange rate time series, in my opinion, the RPSI clustering method should be better for it can make sure the values of parameters by some methods not by computing the distances many times to determine the value of parameter m. But for long time series, the two methods are both available, we can

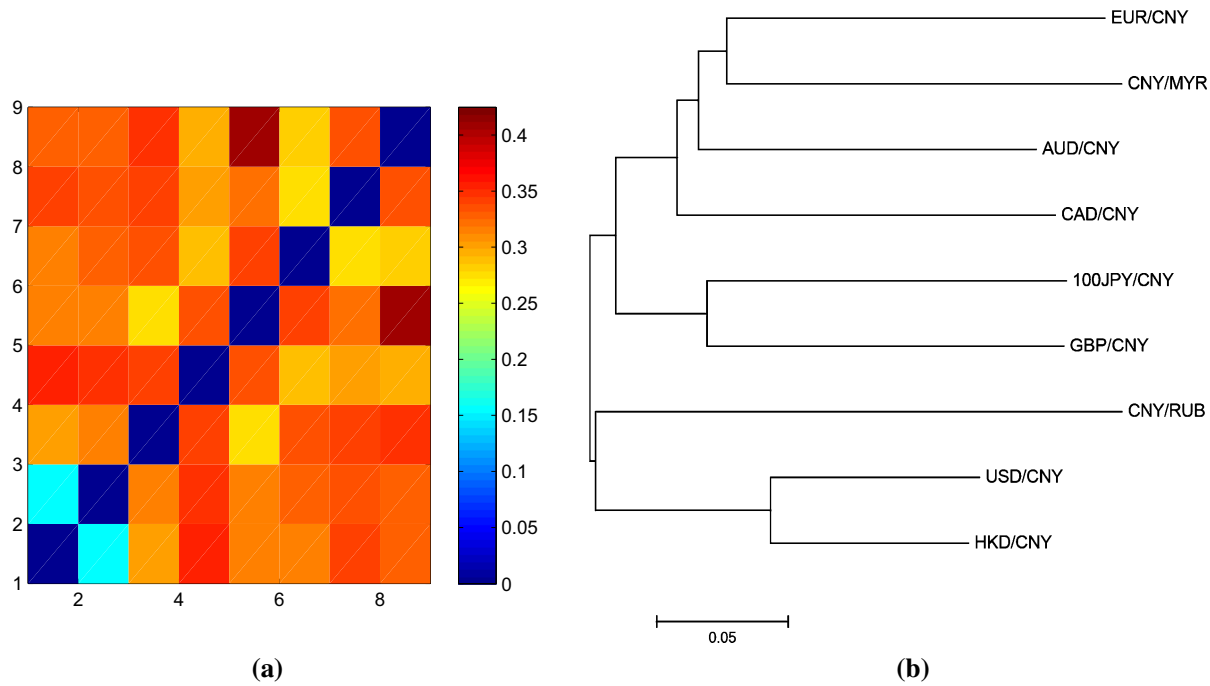


Fig. 10 Distance plots (**a**) and Phylogenetic tree (**b**) generated according to the distances between nine exchange rate time series from 2011 to 2014 with ICM for $m = 6$

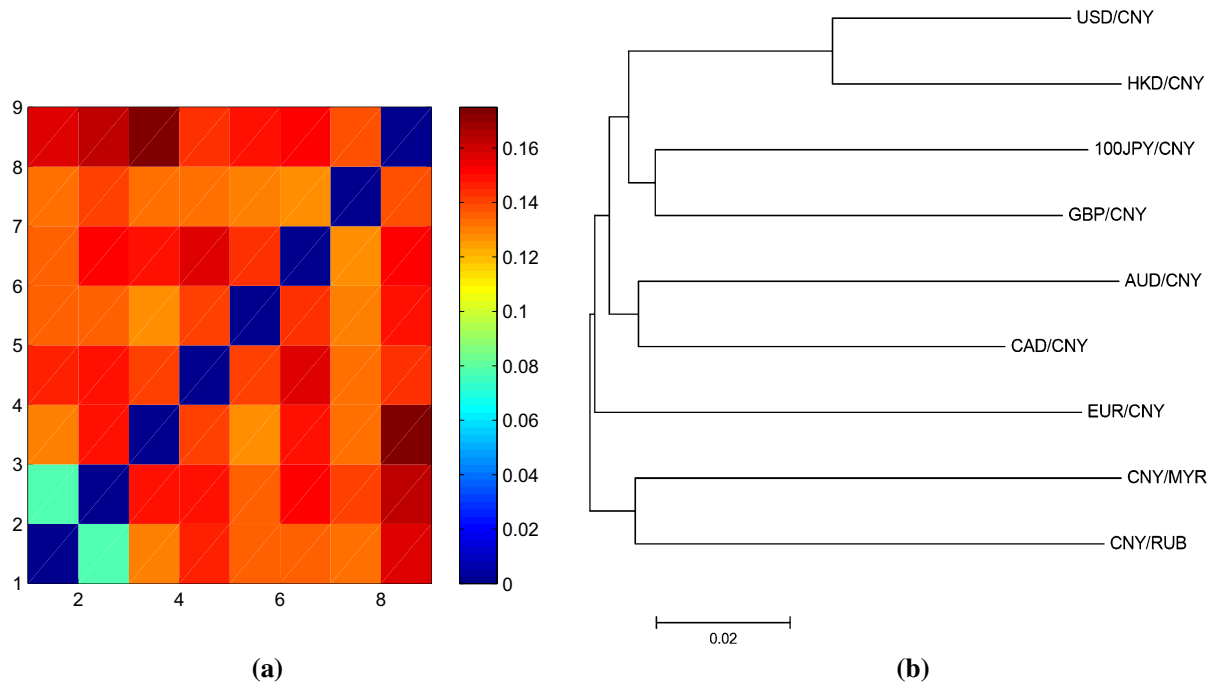


Fig. 11 Distance plots (**a**) and Phylogenetic tree (**b**) to generated according to the distances between nine exchange rate time series from 2011 to 2014 with RPSI clustering method for $m = 5$ and time delay $\tau = 1$

Table 6 The list of means, minimum, maximum, and SDs of distances between stock time series for two methods

Method	Means	Minimum	Maximum	SDs	CVs
ICM	0.330868	0.155119	0.424827	0.046221	0.139696
RPSI	0.142652	0.079008	0.175021	0.01547	0.108449

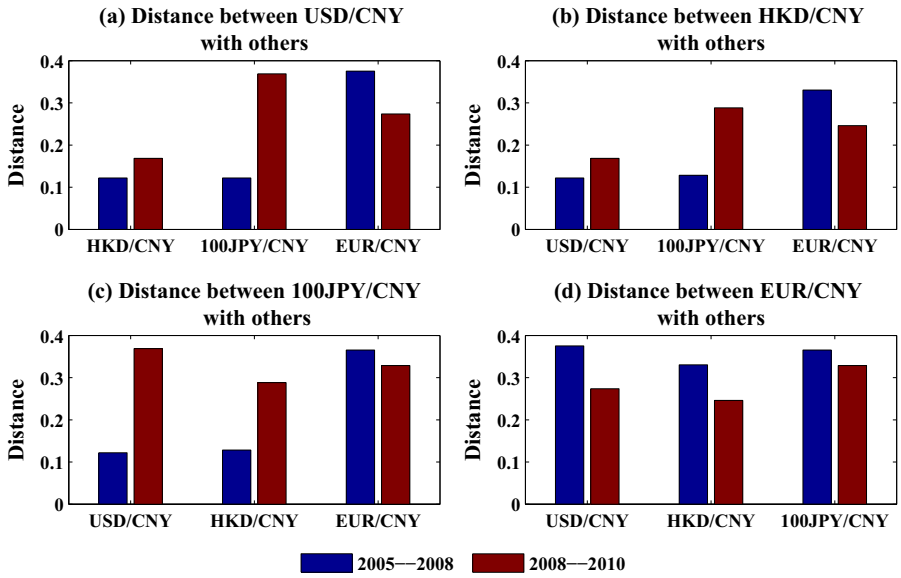


Fig. 12 The distances between **a** USD/CNY and the others, **b** HKD/CNY and the others, **c** 100JPY/CNY and the others, **d** EUR/CNY and the others by the information categorization method with $m = 5$ and time delay $\tau = 1$ in Period A and B, respectively

choose the proper method considering the underground of the actual market. Also we get the trees of exchange rate time series using system clustering. From Fig. 9b, we can find that the system clustering is also not appropriate for these time series for the result is not satisfied for us. Besides, we also calculate the cluster similarity metric and find the RPSI clustering method is more suitable to analyze these short time series as exchange rate time series.

China has experienced several reforms in its exchange rate regime which has become a hot economic issue. On July 21, 2005, the People's Bank of China announced a 2.1% appreciation of the yuan against the dollar and reformed the exchange rate regime by moving into a managed floating exchange rate regime based on market supply and demand with reference to a basket of currencies. Since the modification, the yuan was pegged to a basket of currencies with a possible slow revaluation rather than solely pegged to dollar. In July 2008, China halted the yuan's rise to cope with the global economic crisis. Since then, the yuan has been held at about 6.83 per dollar. On June 19, 2010, the PBC decided to proceed further with reform of the yuan exchange rate regime and to enhance its flexibility. It means that the yuan has been pegged to a basket of currencies again. The exchange rate regime reforms may affect the effective exchange rate of the yuan. We wonder whether the change in China's exchange rate regime in July 2008 has had an effect on the effective exchange rate of the yuan after the exchange rate reform on the July 21, 2005. In order to study the impact of the exchange rate policy modification in July 2008 on the effective exchange rate of the yuan after the reform on the July 21, 2005, we select the USD/CNY, HKD/CNY, 100JPY/CNY, and EUR/CNY exchange rate and divide the exchange rate time series (from July 21, 2005 to June 18, 2010) into two periods. Period A starts from July 21, 2005 to June 30, 2008, and period B is from July 1, 2008, to June 18, 2010. Then, we employ the RPSI clustering method mentioned in Sect. 2 to calculate the distances and analyze the effect of exchange rate regime in 2008 on the change in similarity between them.

Figure 12 displays that the distances between every exchange rate time series and others when we select the information categorization method with $m = 5$ for period A, B. The similarity between USD/CNY (HKD/CNY) and others shows small change for the

two periods except for 100JPY/CNY, and the correlation between EUR/CNY and others become closer after 2008, especially for 100JPY/CNY. Therefore, the exchange rate regime in 2008 influences on the similarity between the 100JPY/CNY and USD/CNY (HKD/CNY) severely, but for 100JPY/CNY, it still stay close relationship with EUR/CNY. In July 2008, China halted the yuan's rise in case of the global economic crisis, and the rate of 100JPY/CNY continued to decrease after 2008. Because of the economic crisis, many countries had taken a series of measures to ensure the downtown of economic. But different countries may choose various methods considering their policy and economic environment. As we have researched above, the US dollar and Hong Kong dollar have the similar tendency. Besides, the US dollar (Hong Kong dollar) against yuan exchange rate had larger relationship with Japanese yen because China put on different regimes on US dollar and Japanese yen for different politic factors. At the beginning of 2008, the rate of inflation in China is higher than in Japan, it may be one of reasons that the 100JPY/CNY changes so sharply for the exchange rate regime in 2008. From the year of 2008, Euro exchange the yen rose unilaterally, so the similarity between 100JPY/CNY and EUR/CNY become larger than others. In conclusion, the comparison of similarity index indicates some phenomenon in exchange rate markets before and after the exchange rate regime in 2008. Also, we can get similar conclusion for the change in the relationship between them before and after 2008 with ICM for $m = 5$.

5 Conclusions

In this paper, to detect the quality of similarity measure of time series, we consider three methods: information categorization method, reconstructed phase space clustering method, and system method with squared Euclidean distances. Both of the two information clustering methods consist of the rank-frequency method and the distance calculation. The rank-frequency method does not require that the length of the time series should be the same, which is convenient for calculation unlike other methods. The new distance measure, defined in Eq. (2), incorporates both a probabilistic weighting factor given by Shannon's entropy and a term related to the number of words. The major difference of the two methods is the way to map the original sequences to symbol sequences. The previous method maps the time series to binary time series, which just consider the adjacent value in the m-bit word. And the values of m have little effect on the result. In general, when we calculate the distances of time series if the series are long enough, we choose $m = 8$; but for the short series, we should make sure the distances are not too small and select the value of m considering the actual sequences. The modified method takes account of the relationship about each value in the m-bit word and we should select appropriate value of dimension embedding and time delay. Especially for time delay, different value may cause various consequences; we need to ensure the proper value of time delay under different background. In pervious study, there are many ways to determine the dimension embedding and time delay; in our study, we choose the false nearest neighbors algorithm and mutual information function.

Here we get the similarity of ARFIMA models, the daily stock markets and China exchange rate market by investigating the distances of these time series with these three methods above. By comparing the similarity metrics, we can find the RPSI clustering method is a highly effective similarity measure for time series. And we get the conclusion that the two information clustering methods are both useful if the difference of the tendency on the time series is large, but if these time series have close relationship at first sight, the RPSI clustering method may be better to analyze the small difference in properties of time series. Also we get similar results for these two methods like in [53]. For China exchange rate markets, we also select the two methods to analyze the exchange rate time series for different time periods. By calculating the distances between the nine currencies against yuan exchange rate time series, we get the phylogenetic trees for two methods and acquire some characteristics of exchange rate time series in China market, especially for the top four trading partners of China: European Union, the USA, Japan, and Hong Kong. Then, we select the top four exchange rate time series and compare the similarity between them in two periods (2005–2008 and 2008–2010); we conclude that the exchange rate regime in 2008 has some effect on the correlation of the four exchange rate time series and we can find that the USD/CNY and HKD/CNY have close relationship, as well the 100JPY/CNY and EUR/CNY. Both of the two methods provide a good choice to classify the different stock time series, but for the change in similarity between exchange rate time series, the

information clustering method based on RPSI should be better.

In summary, we introduce a novel quantitative measurement of similarity among symbolic sequences based on their chaotic properties. This derivation is based on the generic statistical physics assumptions and, therefore, can be applied to a wide range of problems. With the simple measure of similarity, we can categorize different types of symbolic sequences by using standard clustering algorithms. Clustering algorithm is an important tool of knowledge discovering in data mining. This clustering method of symbolic sequences may provide very useful information about the underlying dynamical processes that generate these sequences. Also it is different from the previous clustering method, it provides a novel calculation method of distances between the time series considering its statistic characteristics and complexity. Furthermore, we can change the way mapping the time series to symbolic sequences, like the two method under the study, and their different results can give us good suggestion for choosing proper method to analyze the time series. Maybe the definition of the dissimilarity indices is different for different objects of study. The RPSI clustering method is potentially useful because of its ability to take into account both macroscopic structures and the microscopic details of the dynamics. Now, we make a preliminary research by using the modified method, expecting that the method can be applied to do further study on stock markets, transport system, or foreign exchange rate markets and cross-correlation of stock markets and exchange rate markets.

Acknowledgments Financial support by the Fundamental Research Funds for the Central Universities (2015YJS167) is gratefully acknowledged.

References

1. Mantegna, R.N., Stanley, H.E.: Introduction to Econophysics: Correlations and Complexity in Finance. Cambridge University Press, Cambridge (2000)
2. Mantegna, R.N., Stanley, H.E.: Scaling behaviour in the dynamics of an economic index. Nature **376**, 46–49 (1995)
3. Stanley, H.E.: The fragility of interdependency: coupled networks switching phenomena. Bull. Am. Phys. Soc. **58**, 000583 (2013)
4. Pincus, S.M.: Approximate entropy as a measure of system complexity. Proc. Natl. Acad. Sci. **88**, 2297–2301 (1991)
5. Pincus, S.M.: Approximate entropy (ApEn) as a complexity measure. Chaos Interdiscip. J. Nonlinear Sci. **5**, 110–117 (1995)
6. Pincus, S.M.: Quantifying complexity and regularity of neurobiological systems. Quant. Neuroendocrinol. **28**, 336–363 (1995)
7. Pincus, S.M., Viscarello, R.R.: Approximate entropy: a regularity measure for fetal heart rate analysis. Obstet. Gynecol. **79**, 249–255 (1992)
8. Schuckers, S.A.C.: Use of approximate entropy measurements to classify ventricular tachycardia and fibrillation. J. Electrocardiol. **31**, 101–105 (1998)
9. Richman, J.S., Moorman, J.R.: Physiological time-series analysis using approximate entropy and sample entropy. Am. J. Physiol. Heart Circ. Physiol. **278**, H2039–H2049 (2000)
10. Lake, D.E., Richman, J.S., Griffin, M.P., Moorman, J.R.: Sample entropy analysis of neonatal heart rate variability. Am. J. Physiol. Regul. Integr. Comp. Physiol. **283**, R789–R797 (2002)
11. Liu, L.Z., Qian, X.Y., Lu, H.Y.: Cross-sample entropy of foreign exchange time series. Phys. A Stat. Mech. Appl. **389**, 4785–4792 (2010)
12. Shi, W., Shang, P.: Cross-sample entropy statistic as a measure of synchronism and cross-correlation of stock markets. Nonlinear Dyn. **71**, 539–554 (2013)
13. Costa, M., Goldberger, A.L., Peng, C.K.: Multiscale entropy analysis of complex physiologic time series. Phys. Rev. Lett. **89**, 705–708 (2002)
14. Costa, M., Goldberger, A.L., Peng, C.K.: Multiscale entropy analysis of biological signals. Phys. Rev. E **71**, 021906 (2005)
15. Thuraisingham, R.A., Gottwald, G.A.: On multiscale entropy analysis for physiological data. Phys. A Stat. Mech. Appl. **366**, 323–332 (2006)
16. Xia, J., Shang, P.: Multiscale entropy analysis of financial time series. Fluct. Noise Lett. **11**, 333–342 (2012)
17. Xia, J., Shang, P., Wang, J., Shi, W.: Classifying of financial time series based on multiscale entropy and multiscale time irreversibility. Phys. A Stat. Mech. Appl. **400**, 151–158 (2014)
18. Lin, A., Shang, P., Zhao, X.: The cross-correlations of stock markets based on DCCA and time-delay DCCA. Nonlinear Dyn. **67**, 425–435 (2012)
19. Yin, Y., Shang, P.: Modified DFA and DCCA approach for quantifying the multiscale correlation structure of financial markets. Phys. A Stat. Mech. Appl. **392**, 6442–6457 (2013)
20. Aghabozorgi, S., Teh, Y.W.: Stock market co-movement assessment using a three-phase clustering method. Expert Syst. Appl. **41**, 1301–1314 (2014)
21. Czapkiewicz, A., Majdosz, P.: Grouping stock markets with time-varying Copula-GARCH model. Financ. Uver Czech J. Econ. Financ. **64**, 144–159 (2014)
22. Ausloos, M., Ivanova, K.: Correlations between reconstructed EUR exchange rates versus CHF, DKK, GBP, JPY and USD. Int. J. Mod. Phys. C **12**, 169–195 (2001)
23. Xu, Z., Gencay, R.: Scaling, self-similarity and multifractality in FX markets. Phys. A Stat. Mech. Appl. **323**, 578–590 (2003)

24. Yoon, S.M., Choi, J., Lee, C.C., Yum, M.K., Kim, K.: Dynamical volatilities for yen–dollar exchange rates. Phys. A Stat. Mech. Appl. **359**, 569–575 (2006)
25. Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P.: The distribution of realized exchange rate volatility. J. Am. Stat. Assoc. **96**, 42–45 (2001)
26. Ghashghaie, S., Breymann, W., Peinke, J., Talkner, P., Dodge, Y.: Turbulent cascades in foreign exchange markets. Nature **381**, 767–770 (1996)
27. Ausloos, M.: Statistical physics in foreign exchange currency and stock markets. Phys. A Stat. Mech. Appl. **285**, 48–65 (2000)
28. Ausloos, M., Ivanova, K.: Introducing false EUR and false EUR exchange rates. Phys. A Stat. Mech. Appl. **286**, 353–366 (2000)
29. Muniandy, S., Lim, S., Murugan, R.: Inhomogeneous scaling behaviors in Malaysian foreign currency exchange rates. Phys. A Stat. Mech. Appl. **301**, 407–428 (2001)
30. Norouzzadeh, P., Rahmani, B.: A multifractal detrended fluctuation description of Iranian rial–US dollar exchange rate. Phys. A Stat. Mech. Appl. **367**, 328–336 (2006)
31. Ivanova, K., Ausloos, M.: Low q-moment multifractal analysis of Gold price, Dow Jones industrial average and BGL-USD exchange rate. Eur. Phys. J. B **8**, 665–669 (1999)
32. Schmitt, F., Schertzer, D., Lovejoy, S.: Multifractal analysis of foreign exchange data. Appl. Stoch. Models Data Anal. **15**, 29–53 (1999)
33. Schmitt, F., Ma, L., Angounou, T.: Multifractal analysis of the dollar–yuan and euro–yuan exchange rates before and after the reform of the peg. Quant. Financ. **11**, 505–513 (2010)
34. Baviera, R., Pasquini, M., Serva, M., Vergni, D., Vulpiani, A.: Correlations and multi-affinity in high frequency financial datasets. Phys. A Stat. Mech. Appl. **300**, 551–557 (2001)
35. Fisher, A.J., Calvet, L.E., Mandelbrot, B.B.: Multifractality of Deutschemark/US dollar exchange rates. Cowles Foundation Discussion Paper (1997)
36. Wang, D.H., Yu, X.W., Suo, Y.Y.: Statistical properties of the yuan exchange rate index. Phys. A Stat. Mech. Appl. **391**, 3503–3512 (2012)
37. Cao, G., Cao, J., Xu, L., He, L.Y.: Detrended cross-correlation analysis approach for assessing asymmetric multifractal detrended cross-correlations and their application to the Chinese financial market. Phys. A Stat. Mech. Appl. **393**, 460–469 (2014)
38. Yang, A.C.C., Hseu, S.S., Yien, H.W., Goldberger, A.L., Peng, C.K.: Linguistic analysis of the human heartbeat using frequency and rank order statistics. Phys. Rev. Lett. **90**, 108103(2003)
39. Yang, A.C.C., Hseu, S.S., Yien, H.W., Goldberger, A.L., Peng, C.K., et al.: Yang et al. Reply. Phys. Rev. Lett. **92**, 109802 (2004)
40. Reijmers, T., Wehrens, R., Daeyaert, F., Lewi, P., Buydens, L.M.: Using genetic algorithms for the construction of phylogenetic trees: application to G-protein coupled receptor sequences. Biosystems **49**, 31–43 (1999)
41. Shannon, C.E.: A note on the concept of entropy. Bell Syst. Tech. J **27**, 379–423 (1948)
42. Yang, A.C.C., Peng, C.K., Yien, H.W., Goldberger, A.L.: Information categorization approach to literary authorship disputes. Phys. A Stat. Mech. Appl. **329**, 473–483 (2003)
43. Peng, C.K., Yang, A.C.C., Goldberger, A.L.: Statistical physics approach to categorize biologic signals: from heart rate dynamics to DNA sequences, Chaos: An Interdisciplinary. J. Nonlinear Sci. **17**, 015115 (2007)
44. Goldberger, A.L., Peng, C.K.: Genomic classification using an information-based similarity index: application to the SARS coronavirus. J. Comput. Biol. **12**, 1103–1116 (2005)
45. Yeh, J.R., Lin, T.Y., Shieh, J.S., Chen, Y., Peng, C.K.: A novel blocking index based on similarity measurement applied in distinguishing the patterns of blood pressure signals at dynamically transitional situation. Biomed. Eng. Appl. Basis Commun. **20**, 107–114 (2008)
46. Packard, N., Crutchfield, J., Farmer, J., Shaw, R.: Geometry from a time series. Phys. Rev. Lett. **43**, 712–715 (1980)
47. Ashkenazy, Y., Ivanov, P.C., Havlin, S., Peng, C.K., Goldberger, A.L., Stanley, H.E.: Magnitude and sign correlations in heartbeat fluctuations. Phys. Rev. Lett. **86**, 1900 (2001)
48. Kennel, M.B., R, B., HD, A.: Determining embedding dimension for phase-space reconstruction using a geometrical construction. Phys. Rev. A **45**, 3403–3411 (1992)
49. Fraser, A.M., Swinney, H.L.: Independent coordinates for strange attractors from mutual information. Phys. Rev. A **33**, 1134–1140 (1986)
50. Balocchi, R., Varanini, M., Macerata, A.: Quantifying different degrees of coupling in detrended cross-correlation analysis. Europhys. Lett. **101**, 20011–20016(6) (2013)
51. http://finance.yahoo.com
52. Anguelov, D., Gavrilov, M., Indyk, P., Motwani, R.: Mining the stock market: which measure is best. In: 6th American International Conference on Knowledge Discovery & Data Mining, pp. 487–496 (2000)
53. Tian, Q., Shang, P., Feng, G.: Financial time series analysis based on information categorization method. Phys. A Stat. Mech. Appl. **416**, 183–191 (2014)