Complexity and entropy density analysis of the Korean stock market

J. B. Park*, J. W. Lee*, H.-H. Jo, J.-S. Yang, and H.-T. Moon
Department of Physics, Korea Advanced Institute of Science and Technology
Guseong-dong, Yuseong-gu, Daejeon 305-701, Republic of Korea

Abstract

In this paper, we studied complexity and entropy density of stock market by modeling ε-machine of Korean Composition Stock Price Index (KOSPI) from year 1992 to 2003 using causal-state splitting reconstruction (CSSR) algorithm.

Keywords: Econophysics, computational mechanics, ε-machine, statistical complexity, entropy density.

1. Introduction

Computational mechanics (CM) has been studied in various fields of science. It has been applied to analyze abstract models such as cellular automata [1,2] and Ising spin system [3], as well as natural data in geomagnetism [4]. In this paper, we analyzed financial time series using CM to find the statistical complexity and the entropy density of Korean stock market. Empirical time series in financial market have been investigated by using various methods such as rescaled range (R/S) analysis to test presence of correlations [5] and detrended fluctuation analysis (DFA) to detect long-range correlations embedded in seemingly non-stationary time series [6]. We believe that CM enables the complexities and structures of different sets of data to be quantifiably compared and that it directly discovers intrinsic causal structure within the data [4].

In order to study the statistical complexity and the entropy density in CM, we used causal-state splitting reconstruction (CSSR) algorithm [7] to model ε-machine of Korean Composition Stock Price Index (KOSPI) from year 1992 to 2003. From this, we analyzed the result relating to efficient market hypothesis (EMH).

2. Principles

2.1. Information theory

Claude Shannon first suggested quantity called the entropy $H[X] = -\sum_x \Pr(x)\log_2 \Pr(x)$ of a discrete random variable X with a probability mass function $\Pr(x)$, which is the intuitive notion of measuring information [8]. Let A be a countable set of symbols of time series and let S be a random variable for A, and s is its realization. If a block of string with L consecutive variables is denoted as $S^L = S_1, \ldots, S_L$, then Shannon entropy of length L can be defined as

$$H(L) = -\sum_{s_1 \in A} \cdots \sum_{s_L \in A} \Pr(s_1, \cdots, s_L)\log_2 \Pr(s_1, \cdots, s_L)\,. \qquad (1)$$

As block length L increases, it can be assumed that $H(L)$ increases since there is more information content among the variables. So, entropy density is defined as

$$h_\mu \equiv \lim_{L\to\infty}\left[H(L+1) - H(L)\right]. \qquad (2)$$

For the finite-length L the entropy density is defined as,

$$h_\mu(L) \equiv H(L) - H(L-1), \;\; L = 1, 2, \ldots\,. \qquad (3)$$

The entropy density shows how random the next symbol is [9].

2.2. ε-machine

An infinite string $\overleftrightarrow{S}$ can be divided into two semi-infinite halves, i.e., a future $\vec{S}$ and a history $\overleftarrow{S}$. A causal state is defined as a set of histories that have the same conditional distribution for all the futures. ε is a function that maps each history to the sets of histories, each of which corresponds to a causal state.

$$\begin{aligned}\varepsilon(\overleftarrow{s}) = \{\overleftarrow{s}^{'} \mid P(\vec{S}^L = \vec{s}^L \mid \overleftarrow{S} = \overleftarrow{s}) = P(\vec{S}^L = \vec{s}^L \mid \overleftarrow{S} = \overleftarrow{s}'),\\ \vec{s}^L \in \vec{S}^L, \overleftarrow{s}^{'} \in \overleftarrow{S}, L \in \mathbb{Z}^+\}.\end{aligned} \qquad (4)$$

The transition probability $T_{ij}^{(a)}$ denotes the probability of generating a symbol a when making the transition from state $\mathbb{S}_i$ to state $\mathbb{S}_j$ [9,10].

The combination of the function ε from histories to causal states with the labeled transition probabilities $T_{ij}^{(a)}$ is called the ε-machine [10], which represents a computational model underlying the given time series.

2.3. Statistical complexity

Given the ε-machine, statistical complexity is defined as

$$C_\mu \equiv -\sum_{\{\mathbb{S}_i\}} \Pr(\mathbb{S}_i)\log_2 \Pr(\mathbb{S}_i)\,. \qquad (5)$$

This quantity measures the minimum amount of historical information required to make optimal forecasts of bits [9,11].

The logarithm of the number of causal states called the topological complexity is defined as

*These authors contributed equally.

$C_0 = \log_2 \|\mathbb{S}\|$ [9]. This is the upper bound of the statistical complexity.

3. Analysis

3.1. Complexity analysis

Using the CSSR algorithm we have constructed the ε-machines of the KOSPI price change from 1992 to 2002 for a time window of 1 year. We measured and analyzed the statistical complexity and the entropy density of the data.

To construct the ε-machine, we converted the price change log return series $R = R_1 R_2 \cdots R_t \cdots$ to a new binary time series $F = F_1 F_2 \cdots F_t \cdots$. Each function of R_t and F_t is defined below.

$$R_t \equiv \ln Y_{t+\Delta t} - \ln Y_t \tag{6}$$

$$F_t = \begin{cases} 0 & \text{for } \mathrm{R}(t) < 0 \\ 1 & \text{for } \mathrm{R}(t) > 0 \end{cases} \tag{7}$$

Y_t is the price at time t of series $Y = Y_1 Y_2 \cdots Y_t \cdots$ and Δ t is the time interval which is initially set to 1 minute. We used only intra-day returns to avoid the discontinuous jumps due to the overnight effects.

The ε-machines were constructed for a time window of 1 year shifting each month from April 1992 to July 2002. The block size *L* was set to 7, which gave the most reliable results among different time intervals.

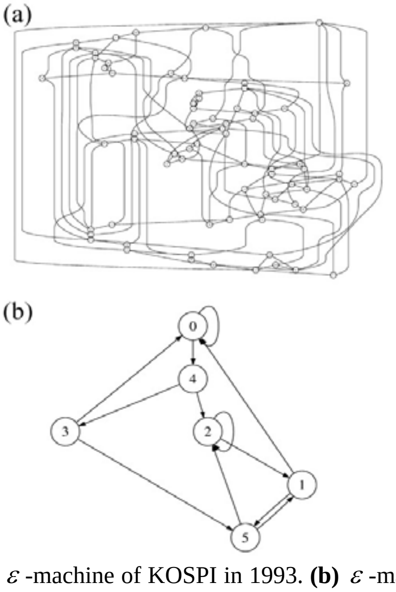

Fig. 1. (a) ε-machine of KOSPI in 1993. **(b)** ε-machine of KOSPI in 2002 (numbers in the circles are node numbers).

Fig. 1 shows two different ε-machines of the KOSPI price change at different periods of time. Each node is a causal state. The number of causal states decreases from 72 states in year 1993 to 6 states in year 2002. By definition, we can see the fluctuation of the number of causal states with the topological complexity C_0 of each ε-machines.

We measured the statistical complexity as well as the topological complexity. We measured only for the 1 minute interval case because as the time interval expends, the total numbers of data points decrease in the factor of $1/\Delta t$. Therefore fewer causal states and low complexity result due to the loss of information.

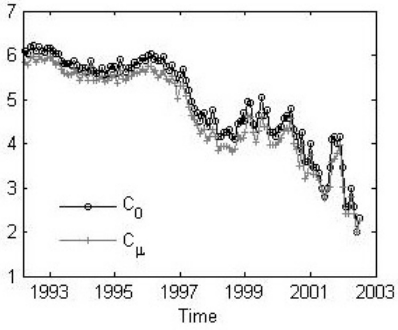

Fig. 2. The topological complexity of 1 minute interval time series (○) and the statistical complexity of 1 minute interval time series (+) from April 1992 to June 2003.

The topological complexity has a tendency of decreasing through time except for the years between 1997 and 1999. This range of time is representing the effect of Asian financial crisis. The (+) graph of Fig. 2 shows the statistical complexity C_μ of ε-machines. The statistical complexity also decreases similar to the topological complexity, because the topological complexity is the upper bound of the statistical complexity by definition. As the gap between these two complexities becomes smaller the causal states distribute more uniformly.

We define a new time series F', which counts the repeated numbers of 0's and 1's in time series *F*.

$$\begin{aligned} F &= 110001110\cdots = 1^{n_1} 0^{n_2} 1^{n_3} 0^{n_4} \cdots \\ F' &= m_1 m_2 \cdots m_i \cdots \\ m_i &= n_i \times \alpha_i \qquad \alpha_i = \begin{cases} 1 & \text{for alphabet 1} \\ -1 & \text{for alphabet 0} \end{cases} \end{aligned} \tag{8}$$

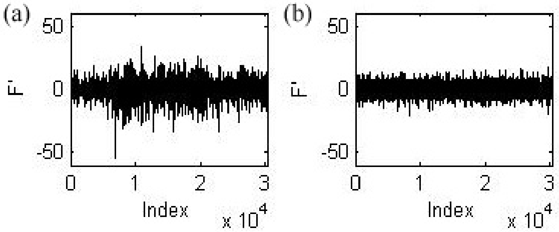

Fig. 3. (a) Time series of F' in 1993. **(b)** Time series of F' in 2002.

In Fig. 3, each peak of the graph denotes duration of consecutive increase or decrease of price. Peaks are more uniformly distributed in year 2002 than 1993 where the standard deviation is 3.03321 and 2.87634, respectively. As peaks in 2002 are more uniformly distributed than in 1993, there are less number of causal states in ε-machine. Thus, both the topological complexity and statistical complexity are decreased.

As the velocity of information flow became faster, the KOSPI log return distribution changes [12]. The velocity of information flow in 2002 became faster than in 1993. So information is rapidly delivered to the agents and immediately applied to the market prices. Immediate price change reduces the correlation between the future and the past, which increases the randomness of the pattern in time series *F*. Therefore, the statistical complexity of the KOSPI price change decreases.

If all the available information is instantly processed as it reaches the market, then the market is said to be efficient. According to the EMH [13], the distribution of the log return becomes Gaussian. This means that the price change pattern is totally random. J.-S. Yang have used tail index of the price change distribution to analyze this property of KOSPI [12]. Also similar phenomenon was found in Japan around 1990 [14,15].

The statistical complexity of the ten-year KOSPI price change shows that Korean market is becoming closer to EMH.

3.2. Entropy density analysis

We now analyze the entropy density $h_\mu(L)$ of the time series *F*. We set *L* as 7, which is the optimal point that $h_\mu(L)$ is closest to h_μ under the limit of the algorithm we used. The time window is set to 1 year and shifting each month, which is in the same manner as the complexity analysis.

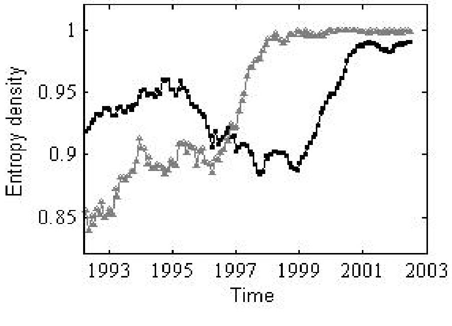

Fig. 4. Entropy density of 1 minute interval time series (□) and Entropy density of 10 minute interval time series (△) from April 1992 to June 2003.

The entropy density was measured for the case of 10 minute time interval series as well as 1 minute case. We measured both 10 minute and 1 minute time interval series because we were able to extract noticeable results by changing the time interval of the series. When determining weather the given series is more random or not, loss of data points becomes less important. This is due to the nature of entropy density which measures the randomness of the series.

In the case of time interval of 1 minute, entropy density increased during year 1992 to 1995 and 1999 to 2003. Between year 1995 and 1999 the entropy decreased. Different from 1 minute case, the entropy density for 10 minute case increases continuously with a little fluctuation during year 1992 to 1998. Moreover, 10 minute case saturates to 1 after year 1998 but 1 minute case saturates after year 2001. The entropy density is closer to 1 when the pattern is more random.

To explain this peculiar phenomenon, we compared entropy density with autocorrelation of the time series *F*. Autocorrelation of *F* is defined by

$$A(\tau) = \frac{\langle F_t F_{t+\tau} \rangle - \langle F_t \rangle^2}{\langle F_t^2 \rangle - \langle F_t \rangle^2}\,, \tag{9}$$

where τ is the time lag.

We picked 4 specific regions of time series F to measure the entropy density and autocorrelation for each region separately. These are January of 1993, January of 1994, December of 1997, and January of 2002.

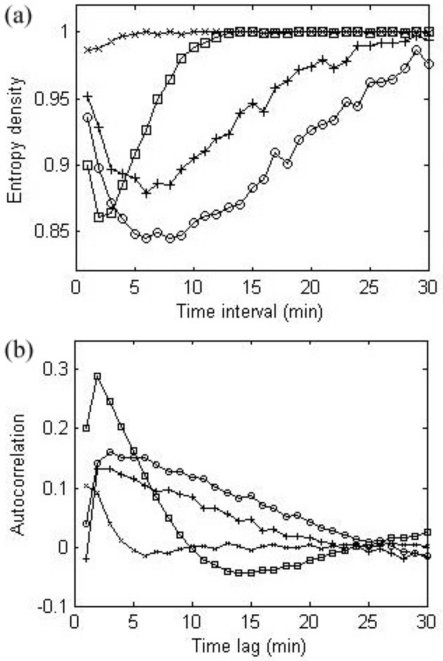

Fig. 5. (a) Entropy density of Jan. 1993 (○), Jan. 1994 (+), Dec. 1997 (□), and Jan. 2002 (×). **(b)** Autocorrelation of Jan. 1993 (○), Jan. 1994 (+), Dec. 1997 (□), and Jan. 2002 (×).

Comparing (a) and (b) of Fig. 5, we found the opposite tendency of entropy density and autocorrelation. When there is high correlation among the series, the randomness of the series decreases. Therefore the entropy density decreases as well. In Fig.

4, the entropy density of 10 minute interval series is lower than 1 minute interval series between years of 1992 and 1997. This indicates more correlations in 10 minute interval series than 1 minute interval series in this period of time. The entropy density of 1 minute interval becomes lower than the entropy density of 10 minute interval after 1997 until 2003. There are more correlations in the 1 minute interval series than 10 minute intervals in this period of time.

The stock market price changes only when the agents in the stock market buy or sell. Since the agents of the stock market make decision based on the information they have, the velocity of information flow becomes the most important factor of changing rate of price in the stock market. We define the time required to deliver information as Δt_i , and the time interval we chose to measure the entropy density as Δt_m . The information delivery time Δt_i is directly proportional to the average price change cycle. If Δt_m is shorter than Δt_i , it is in smaller scales than the changing periods of the series. Therefore this series contain noises. When Δt_m is longer than Δt_i , it becomes hard to find correlations between each interval of Δt_m . As Δt_m expands longer than the average changing period of the series, the correlations between each interval gradually disappear.

Between the years 1992 and 1997 shown in Fig. 4, the information delivery time Δt_i was more close to 10 minute than 1 minute. The 1 minute interval contained more noise than 10 minute, so the series of 1 minute interval appears more random than 10 minute interval series. Hence, the entropy density of 1 minute case is higher than 10 minute case this period of time. As the velocity of information becomes faster, Δt_i becomes closer to 1 minute each year. From 1997 to 2003, Δt_i is more close to 1 minute than 10 minute. The 10 minute interval becomes longer than the information delivery time Δt_i , so the correlations between each interval decrease. The entropy density of 10 minute interval series becomes higher than 1 minute interval series.

By generalizing the above discussion, we can find Δt_m which is closest to Δt_i . This particular time interval Δt_m shows the highest correlation of the series at each time window. This can be derived from the time interval which minimizes the entropy density. The minimum entropy density of each time window is defined as

$$h_{\mu,\,\min} = \min_{\Delta t_m} \left(h_\mu(\Delta t_m)\right) = h_\mu(\Delta t_{\min}) \qquad (10)$$

where $\Delta t_{\min}$ is the particular time interval we described.

In Fig. 5 (a), the entropy density of January 1993 (○) is minimized at 8 minute time interval. The time interval $\Delta t_{\min}$ is 6 minute in January 1994 (+) , 2 minute in December 1997 (□) , and 1 minute in January 2002 (×) .

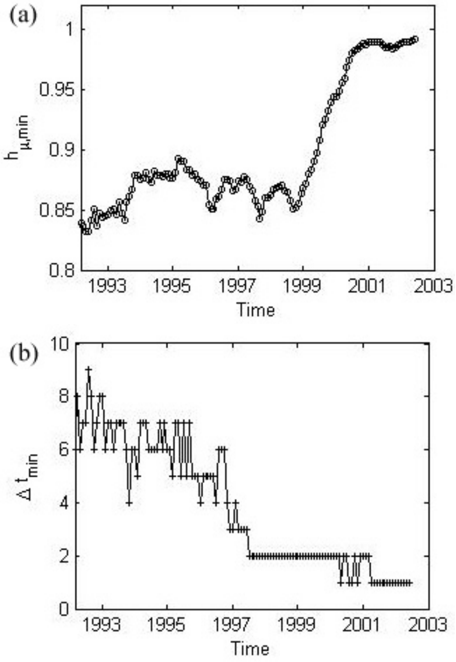

Fig. 6 (a) Minimum entropy density at each month from April 1992 to June 2003, **(b)** Time interval of series which minimizes the entropy density at each month from April 1992 to June 2003.

Fig. 6 (a) shows the minimum entropy density $h_{\mu,\,\min}$ of each month defined by equation (10). From 1992 to 1999, $h_{\mu,\,\min}$ fluctuates inside the range of 0.8~0.9. After year 1999, $h_{\mu,\,\min}$ increases rapidly and saturates to 1 until year 2003. We define this as the optimal line of entropy density.

In Fig. 6 (b), the time interval $\Delta t_{\min}$ has a tendency of decreasing. The local fluctuations are due to the limited resolution of the time interval. The time interval $\Delta t_{\min}$ is the most correlated interval and also the closest interval to Δt_i . The long term decrease of $\Delta t_{\min}$ from 1992 to 2003 represents the decrease of the information delivery time Δt_i in this period of time. After 1997, $\Delta t_{\min}$ decreases more rapidly. This period of time is when the super-high speed internet service was started in Korea. The ratio of online traders increased exponentially, and this influenced the stock market price to change more rapidly.

Since there is no standard for measuring the information delivery time Δt_i , we take $\Delta t_{\min}$ as the standard for measuring this. We call this quantity effective delay of information (EDI). When the market is idealized with EMH [13], EDI becomes 0. EDI of the Korean stock market is 1 minute in year 2003. We can measure the efficiency of the stock market by measuring EDI.

4. Conclusion

The statistical complexity and the entropy density in CM are intuitive and powerful concept to study complicated nonlinear sequences derived from physical systems. We analyzed the statistical complexity and the entropy density of the KOSPI price change from 1992 to 2002 by using ε-machines constructed from the CSSR algorithm.

We found the effect of the information flow velocity on the Korean stock market by measuring the statistical complexity and the entropy density of the ten-year KOSPI price change. The statistical complexity has a tendency of decreasing. This phenomenon indicates more uniform price change distribution due to the increase of information flow velocity. We defined minimum entropy density as an optimal line to measure the randomness. We also defined the time interval of the series as EDI, which gives the minimum entropy density at each time window. By measuring the minimum entropy density, we found that EDI of the ten-year KOSPI price change has decreased. EDI measures the efficiency of the Korean stock market. Quantitative analysis showed that the efficiency of the Korean market dynamics became close to EMH.

5. Reference

[1] J. E. Hanson, and J. P. Crutchfield, “Computational mechanics of cellular automata: An example,” *Physica D*, 103, pp. 169-189, 1997.

[2] C. R. Shalizi, K. L. Shalizi, and R. Haslinger, “Quantifying Self-Organization with Optimal Predictors.” *Phys. Rev. Lett.*, vol. 93, no. 11, 2004.

[3] J. P. Crutchfield, and D. P. Feldman, “Statistical complexity of simple one-dimensional spin systems,” *Phys. Rev. E*, vol. 55, no. 2, 1997.

[4] R. W. Clarke, M. P. Freeman, and N. W. Watkins, “Application of computational mechanics to the analysis of natural data: An example in geomagnetism,” *Phys. Rev. E*, vol. 67, 2003.

[5] E. E. Peters, “Chaos in order in the capital markets,” Wiely, 1991.

[6] P. Norouzzadeh, and G. R. Jafari, “Application of multifractal measures to Tehran price index,” *Physica A*, vol. 356, 2005**.**

[7] C. R. Shalizi, et al., "An Algorithm for Pattern Discovery in Time Series", SFI Working Paper 02-10-060, 2002

[8] N. J. A. Sloane and A. D. Wyner, editors, “Ce. E. Shannon: Collected Papers”, IEEE Press, 1993.

[9] D. Feldman, “A Brief Introduction to: Information Theory, Excess Entropy and Computational Mechanics”, http://hornacek.coa.edu/dave/Tutorial/index.html, April 1998.

[10] C. R. Shalizi and J. P. Crutchfield, “Computational mechanics: Pattern and prediction, structure and simplicity”, *J. Stat. Phys.*, 104, pp. 819-881, 2001.

[11] J. P. Crutchfield, “The calculi of emergence: computation, dynamics and induction”, *Physica D*, 75, 1994.

[12] J.-S. Yang, S. Chae, W.-S. Jung, H.-T. Moon, “Microscopic spin model for the dynamics of the return distribution of the Korean stock market index”, *Physica A*, 363, pp. 377-382, 2006.

[13] R. N. Mantegna and H. E. Stanley, “An Introduction to Econophysics: Correalations and Complexity in Finance”, Cambridge University Press, Cambridge, 1999.

[14] T. Kaizoji, “Inflation and deflation in financial markets,” *Physica A*, vol. 343, no. 47, pp. 662-668, 2004.

[15] T. Kaizoji, “Heterogeneous agents and nonlinear dynamics,” in T. Lux, S. Reitz, and E. Samanidou, (Eds.), Springer, Berlin, pp. 237, 2005.