A NOVEL INTEGRATED APPROACH FOR ANALYZING THE FINANCIAL TIME SERIES AND ITS APPLICATION ON THE STOCK PRICE ANALYSIS

Le Hu, Jimin Ye, Jianwei E

School of Mathematics and Statistics, Xidian University
Xi'an, Shaanxi, 710071, China

(Received February 1, 2021; accepted June 28, 2021)

With the advent of the big data era, the ever-increasing accumulated financial data plays a significant role in people's daily life. The ensemble methods are the most efficient and accurate methods for analyzing the essential phenomenon of the financial data. In this paper, a novel approach, which incorporates the complementary ensemble empirical mode decomposition (CEEMD) and semi-parametric linear non-Gaussian analysis (Spline-LCA), *i.e.*, CEEMD-Spline-LCA, is proposed to analyze the nonlinear and non-stationary financial time series. The proposed CEEMD-Spline-LCA method consists of three steps: In the first place, the CEEMD is applied to obtain Intrinsic Mode Functions (IMFs) of the analyzed data. Next, according to the contribution coefficients between the IMFs and the financial time series, IMFs are reorganized to get a new collection of the IMFs (NIMFs) for the subsequent explanation of influence factors of financial time series. Furthermore, Spline-LCA is utilized to separate the NIMFs into independent components (ICs), reflecting the different inner driving factors. Concentrating the established model on the stock price (the Dow Jones Industrial Average, the oldest stock index generally used), by comparing with real economic indicators, we find that the obtained ICs are close approximation of the exchange rate (U.S. Dollar Index), interest rate (Fed funds rate), GDP growth rate, CPI and major events, respectively.

DOI:10.5506/APhysPolB.52.1163

1. Introduction

Scholars now have the luxury of studying with the financial time series. Some examples are as follows: Intercontinental Exchange operates global markets, which have a tremendous amount of the financial data. The Dow Jones developer platform has at least fifty years of stock price data and has broad experience and deep knowledge of the factors driving market change. A pragmatic approach for analyzing the hidden information in financial data is empirical mode decomposition (EMD). For instance, Zhu *et al.* [1] explained the inner market formation mechanism of the carbon price via the EMD model. Although EMD performs well for non-stationary data, IMFs can only give a relatively broad analysis. Usually, the different driving mechanisms are mutually independent. Hence, an integrated approach based on the EMD and independent components analysis (ICA) is a feasible alternative for finding the hidden information (or ICs), wherein the ICA is utilized to separate ICs from the IMFs.

The Dow Jones Industrial Stock Price, as economy "Barometer", is regarded as the most authoritative and representative stock price index, which plays a prominent role in analyzing the fluctuation behavior of the stock market. Recently, many scholars studied stock price fluctuations. In detail, for the Dow Jones Industrial Average, much of the growth has come from the U.S. capital market environment of low-base rate and low-earnings price ratios by Stein [2]. Schwert [3] pointed out that the historical volatility of industry has no significant effect on the current volatility of the stock market, but the volatility of the stock market has a strong predictive power on the volatility of industrial production. Further, Robert [4] argued that the growth of the Dow Jones Industrial Average needs to be analyzed in the context of the different stages of development of the U.S. industrial economy. Typically, economic fluctuations will affect stock price movements [5, 6]. David [7] found that uncertainty about inflation and earnings are the main influence factors of stock volatility. Also, according to Zeisberger [8], the growth of the Dow Jones Industrial Average is closely related to the development and change of thr American industrial economy. In recent years, scholars have paid more attention to the prediction of stock price and ignored analysis of the reasons that affect stock price fluctuations [9–14].

In order to further study the specific reasons that affect stock volatility, some researchers applied statistical methods to find the reasons more intuitively. For instance, Abdalla [15] thought that a change in the exchange rate would lead to a change in the stock price by establishing a causal relationship between the foreign exchange market and the stock market. Subsequently, Aggarwal [16] used the classic regression method to study the exchange rate and the stock market, and obtained the positive correlation between the exchange rate and the stock market. Engle [17] integrated GDP, short-term interest rate, exchange rate, and inflation index into the spline-GARCH model to study the long-term fluctuation of stock market price index in macroeconomy. The author found that the GDP, inflation index, and interest rate fluctuations are positively correlated with stock market fluctuations. Furthermore, Charles [18] employed a new semi-parametric test based on conditional heteroscedasticity models to study large shocks in the volatility of the U.S. stock market. However, these methods seek the causal relationship between the stock price and several other economic variables in a certain way, *i.e.*, some possible reasons are given firstly, then, the statistical methods are used to select the right reasons.

ICA, as a digital signal processing technique, which aims at separating out ICs from the observed signal under an unknown mixing system, has been applied to the fields, including brain imaging, finance, and image feature extraction [19–22]. In recent years, the combination of mode decomposition method and ICA method have been employed in various nonlinear and non-stationary time series. Li [23] integrated EMD and ICA (EMD-ICA) to separate the source signals from the observed signal with white noise. Further, Li applied it to analyze the mixed signal including EEG and EOG. However, EMD has the disadvantages of mode mixing and boundary effect [24]. Hence, Jurić [25] replaced the EMD with ensemble empirical mode decomposition (EEMD) to measure electromyographic signals in laryngeal muscles. Pramudita *et al.* [26] removed ocular artifacts in EEG signals by using combination of complete EEMD (CEEMD) and ICA. This combination method has been widely applied to analyze the nonlinear and non-stationary data, for example, Xian *et al.* [27] obtained 7 ICs of the original series (gold price), which represent the economic factors of CPI, gold supply and demand, *etc.*; E *et al.* [28] obtained 4 ICs of the original series (crude oil price), which denote the economic factors of the U.S. import crude oil price, GDP, *etc.* However, this method is rarely used to the stock price.

In financial data analysis, the CEEMD method is used to eliminate residual noise in the IMFs by Yeh [29]. The Spline-LCA model is closer to the practice by taking the noise component into consideration, and the idea of dimensionality reduction can reduce the computational constrains [30]. Thereby, integrating the CEEMD and spline-LCA, we propose a novel method (CEEMD-Spline-LCA) to analyze financial data, which possesses some advantages: For one thing, the main influencing factors are found from the data straightforwardly. For another, the change processes of each influencing factor are analyzed directly. Last but not least, this paper mainly uses this method to analyze the stock price and find its internal formation mechanism.

The paper is structured as follows: Section 2 takes a brief look at the CEEMD and Spline-LCA. The CEEMD-Spline-LCA approach is presented and described in Section 3. In Section 4, the Dow Jones Industrial Average is decomposed into several ICs by the proposed method and their ICs are compared with the real metrics. Section 5 is the conclusion.

2. Research methodology

2.1. CEEMD

EMD introduced by Huang *et al.* [31], as a classical signal decomposition technique, has been successfully applied to many fields, such as crude oil price, geophysical studies and rotor startup signals [32–34]. The essence of the EMD is to identify the oscillatory mode according to its characteristic time scale, and to decompose the data accordingly.

EMD can decompose the observed signal into IMFs and a residual. The IMF needs to meet two conditions: (1) The number of extrema and zero crossings must be either equal or differ at most by one; (2) At any point, the mean of the envelopes defined by the local maxima and local minima is zero. Assuming that $x(t)$ is the observed signal, the EMD algorithm consists of the following five steps:

Step 1. Calculate the local extrema points (including maxima and minima points) of the observed signal $x(t)$.

Step 2. A cubic spline line is used to connect all the local maxima as the upper envelope line $u_0(t)$. The procedure is repeated for all the local minima to generate the lower envelope line $v_0(t)$.

Step 3. Find out the mean curve of the upper and lower envelopes, *i.e.*

$$m_0(t) = \frac{u_0(t) + v_0(t)}{2}\,. \tag{1}$$

Then, subtracting the mean $m_0(t)$ from the observed signal $x(t)$, one can yield that

$$h_0(t) = x(t) - m_0(t)\,. \tag{2}$$

Step 4. If $h_0(t)$ satisfies the above two conditions of IMF, then let $c_1(t) = h_0(t)$ be the first IMF and goes to the next step; otherwise, let $x(t) = h_0(t)$ and goes back to Step 1, until getting the first IMF $c_1(t)$.

Step 5. Separate $c_1(t)$ from the rest of $x(t)$ by

$$r_1(t) = x(t) - c_1(t)\,. \tag{3}$$

Repeating Step 1.–Step 4. on $r_1(t)$, we can obtain the second IMF $c_2(t)$. Subsequently, one can yield that

$$\begin{gathered} r_2(t) = r_1(t) - c_2(t)\,, \\ \vdots \\ r_N(t) = r_{N-1}(t) - c_N(t)\,. \end{gathered} \tag{4}$$

This iterative procedure can be stopped via the criterion that the rest of $x(t)$, say, $r_N(t)$, is a monotonic function from which no more IMF is extracted. Particularly, we call the final rest of $x(t)$ as residual function. Hence, the observed signal $x(t)$ can be rewritten as

$$x(t) = \sum_{i=1}^{N} c_i(t) + r_N(t)\,, \tag{5}$$

where N is the number of IMFs.

To overcome some disadvantages of the EMD such as mode mixing and boundary effect, Wu and Huang [35] put forward the EEMD by adding white noise to the observed signal. Although the larger the ensemble, the smaller the relative magnitude of the residual noise, it is too time-consuming in practical applications. Hence, Yeh [29] proposed the CEEMD, which adds white noise to the observed signal (*i.e.* one positive and one negative) in pairs.

Considering the observed signal $x(t)$, the CEEMD consists of the following steps:

Step 1. Add a pair of whitening noise series $\{n_i(t)\}$ and $\{-n_i(t)\}(i = 1, \cdots, Ne)$ to the observed signal $x(t)$

$$a_i^1(t) = x(t) + n_i(t)\,, \tag{6}$$
$$a_i^2(t) = x(t) - n_i(t)\,. \tag{7}$$

Step 2. Use the EMD algorithm to decompose the new series $\{a_i^1\}$ and $\{a_i^2\}$. Then, two IMFs series, $\{b_{ij}^1(t)\}$ and $\{b_{ij}^2(t)\}$ are derived. Note that $\{b_{ij}^1(t)\}$ ($\{b_{ij}^2(t)\}$) means the j^{th} IMF came from the i^{th} new data with positive (negative) noise.

Step 3. Derive the IMFs

$$c_j(t) = \frac{1}{2Ne} \sum_{i=1}^{Ne} \left[b_{ij}^1(t) + b_{ij}^2(t) \right]\,, \tag{8}$$

where Ne denotes the number of white noise pairs added to $x(t)$. It is easy to see that $x(t) \approx \sum_{j=1}^{N+1} c_j(t)$, where N is the number of IMFs determined by some given beforehand condition, and $c_{N+1}(t)$ is recorded as residual.

2.2. Spline-LCA

As an improvement of the classical FastICA algorithm proposed by Hyvärinen and Oja [36], semi-parametric linear non-Gaussian analysis (Spline-LCA) was proposed by Risk *et al.* [30], which achieves dimension reduction and latent variable estimation simultaneously. In the Spline-LCA model, the observed signal vector $\boldsymbol{x} \in \mathbb{R}^T$ is generated according to the model

$$\boldsymbol{x} = \boldsymbol{M}_{\mathrm{s}}\boldsymbol{s} + \boldsymbol{M}_{\mathrm{n}}\boldsymbol{n}\,, \tag{9}$$

where $\boldsymbol{s} = [s_1, \cdots, s_Q]^T$ is the source signal vector of mutually independent non-Gaussian random variables with zero mean and unit variance $(T \geq Q)$; $\boldsymbol{n}$ is $(T-Q)-$ variate normal with $E\{\boldsymbol{n}\} = \boldsymbol{0}$ and $E\{\boldsymbol{n}\boldsymbol{n}^T\}$ nonsingular; $\boldsymbol{M} = [\boldsymbol{M}_{\mathrm{s}}, \boldsymbol{M}_{\mathrm{n}}]$ ($\boldsymbol{M}_{\mathrm{s}} \in \mathbb{R}^{T\times Q}$ and $\boldsymbol{M}_{\mathrm{n}} \in \mathbb{R}^{T\times(T-Q)}$) is the mixing matrix with full rank. It is proved that given an i.i.d. sample of $\boldsymbol{x}$, $\boldsymbol{M}_{\mathrm{s}}$ and $\boldsymbol{s}$ can be estimated uniquely up to scales and permutation [30].

It is well known that the pre-whitening can make the ICA easy to be implemented. Denote the eigenvalue decomposition (EVD) of the covariance matrix of $\boldsymbol{x}$ by $\boldsymbol{\Sigma}_{\mathrm{x}} = \boldsymbol{U}\Lambda\boldsymbol{U}^T$, then, $\boldsymbol{L} = \boldsymbol{U}\Lambda^{-1/2}\boldsymbol{U}^T$ is the symmetric whitening matrix, thus the pre-whitened observed signal vector is $\boldsymbol{z} = \boldsymbol{L}\boldsymbol{x}$. Based on the pre-whitened observed vector $\boldsymbol{z}$, the ICs and the Gaussian noise can be separated using a $T \times T$ orthogonal matrix $\boldsymbol{W}$

$$\left(\hat{\boldsymbol{s}}^T, \hat{\boldsymbol{n}}^T\right)^T = \boldsymbol{W}\boldsymbol{z}\,. \tag{10}$$

Due to the inherent ambiguity in ICA, let $\boldsymbol{W}_{\mathrm{s}}$ denote the sub-matrix of $\boldsymbol{W}$ consisting of the rows of $\boldsymbol{W}$, which correspond to the separated source signals. Then, the way to estimate the ICs is to update the $Q \times T$ dimension matrix $\boldsymbol{W}_{\mathrm{s}}$ such that

$$\hat{\boldsymbol{s}} = \boldsymbol{W}_{\mathrm{s}}\boldsymbol{z} \tag{11}$$

is the estimation of the source signal vector $\boldsymbol{s}$.

Assume that $\{g_q\}$, $q = 1, \ldots, Q$ are the true densities of the ICs. The estimator of $\boldsymbol{W}_{\mathrm{s}}$ is defined as

$$\hat{\boldsymbol{W}}_{\mathrm{s}} = \underset{\boldsymbol{W}_{\mathrm{s}}}{\mathrm{argmax}} \sum_{i=1}^{n}\sum_{q=1}^{Q} \log g_q\left(\boldsymbol{w}_q^T\boldsymbol{z}_i\right)\,, \tag{12}$$

where $\{\boldsymbol{z}_i\}$, $i = 1, \ldots, n$ is an i.i.d. sample of the pre-whitened observed signal vector $\boldsymbol{z}$, and $\boldsymbol{w}_q^T$ is the q^{th} row of $\boldsymbol{W}_{\mathrm{s}}$.

This is a semi-parametric model, an improved fixed-point algorithm (which will be given in following Algorithm 1) can be exploited to estimate the interesting parameter $\boldsymbol{W}_{\mathrm{s}}$ for a given non-parameter part $\{g_q\}$, $q = 1, \ldots, Q$. It is pointed out in Ref. [21] that the score functions (*i.e.*, $r_q \triangleq$

$\frac{\mathrm{d}}{\mathrm{d}s}(\log g_q(s)))$ can be considered as the optional nonlinearities in the ICA algorithm if the probability density functions (pdfs) of the source signals: the non-parameter part $\{g_q\},\ q = 1, \ldots, Q$ are known. Since the ProDenICA (product density estimation for ICA) generally outperformed parametric and kernel ICA methods and its algorithmic complexity is $O(n)$, it is necessary to estimate pdfs of the source signals properly [37, 38].

Assume that the ICs have tilted Gaussian distributions of the form of $\phi(v)\mathrm{e}^{f(v)}$, where $\phi(v)$ represents the standard normal density and $f(v)$ denotes a twice-differentiable function. Then, the estimation of $\{g_q\},\ q = 1, \ldots, Q$ is transformed to the estimation of corresponding $\{f_q\},\ q = 1, \ldots, Q$. This can be obtained by applying the iteratively reweighted penalized least squares algorithm [39].

The algorithm of Spline-LCA is concluded as Algorithm 1. r_q and $r_q^{'}$ are the first and second derivatives of $\log(g_q(v))$, respectively. The sign- and permutation-invariant mean-squared error, $\mathrm{PMSE}(\boldsymbol{W}_1, \boldsymbol{W}_2) = \frac{1}{TQ}$ argmin $||\boldsymbol{W}_1 - \boldsymbol{W}_2 P_{\pm}||_{\mathrm{F}}^2$ ($\boldsymbol{W}_1 \in \mathbb{R}^{T\times Q}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{T\times R}$ with $Q \leq R$), where $||\cdot||_{\mathrm{F}}$ represents the Frobenius norm and $P_{\pm}$ denotes the $R \times Q$ dimension signed permutation matrix, is a novel measure of dissimilarity.

Algorithm 1: The Spline-LCA algorithm

Set the pre-whitened observed data matrix $\boldsymbol{Z} \in \mathbb{R}^{n\times T}$; initial $\boldsymbol{W}_{\mathrm{s}}^0$; tolerance ε; and $m = 0$.

(1) Let $\hat{\boldsymbol{S}}^0 = \boldsymbol{Z}\boldsymbol{W}_{\mathrm{s}}^{0^T}$.

(2) Estimate $\{g_q^{(m+1)}, q = 1, \ldots, Q\}$ via estimation of $\{f_q^{(m+1)}, q = 1, \ldots, Q\}$ using the iteratively reweighted penalized least squares algorithm.

(3) For each $\boldsymbol{w}_q, q = 1, \ldots, Q$ of $\boldsymbol{W}_{\mathrm{s}}$, calculate

$\boldsymbol{w}_q^* = \frac{1}{n}\sum_{i=1}^{n}\{r_q(\boldsymbol{w}_q^{(m)^T}\boldsymbol{z}_i)\boldsymbol{z}_i - r_q^{'}(\boldsymbol{w}_q^{(m)^T}\boldsymbol{z}_i)\boldsymbol{w}_q^{(m)}\}$.

(4) $\boldsymbol{W}_{\mathrm{s}}^* = \boldsymbol{U}^*\boldsymbol{D}^*\boldsymbol{V}^{*T}$ (thin SVD).

(5) $\boldsymbol{W}^{(m+1)} = \boldsymbol{U}^*\boldsymbol{V}^{*T}$.

(6) $\hat{\boldsymbol{S}}^{(m+1)} = \boldsymbol{X}\boldsymbol{W}_{\mathrm{s}}^{(m+1)^T}$.

(7) If $\mathrm{PMSE}(\boldsymbol{W}_{\mathrm{s}}^{(m+1)^T}, \boldsymbol{W}_{\mathrm{s}}^{(m)^T}) \geq \varepsilon$, $m = m + 1$ and repeat (2)–(6), else stop.

---

3. CEEMD-Spline-LCA

Thge financial time series, defined at discrete points on the time axis, is a discrete random process [40], for example, a daily rate of return on stocks. Financial data is subject to collisions from various factors which can be assumed to be $s_1, s_2, \ldots, s_Q$ [21]. The financial time series is a mixture of several ICs in some way, which can be regarded as the multivariate influencing factors with the joint distribution $G(s_1, s_2, \ldots, s_Q)$. Then, in terms of independence,

$$g(s_1, s_2, \ldots, s_Q) = g_1(s_1) \ldots g_Q(s_Q) \,. \tag{13}$$

Note that $g(s_1, \ldots, s_Q)$ is the joint density function of the financial time series and $g_1(s_1), \ldots, g_Q(s_Q)$ are the probability density function (pdf) of the $s_1, \ldots, s_Q$, respectively.

It is very important to seek $s_1, s_2, \ldots, s_Q$ in some way. Since most of the financial data are single-channel data, the mode decomposition method CEEMD is applied to obtain IMFs, $c_1(t), \ldots, c_{N+1}(t)$. However, we can only roughly judge the impact of the observed signal based on the frequency and amplitude of IMFs. Put it another way, the internal information reflected on the components as independent as possible is not abstracted via the CEEMD model. Thus, we adopt the ICA model which belongs to the blind source separation technique with the advantage of unsupervised machine

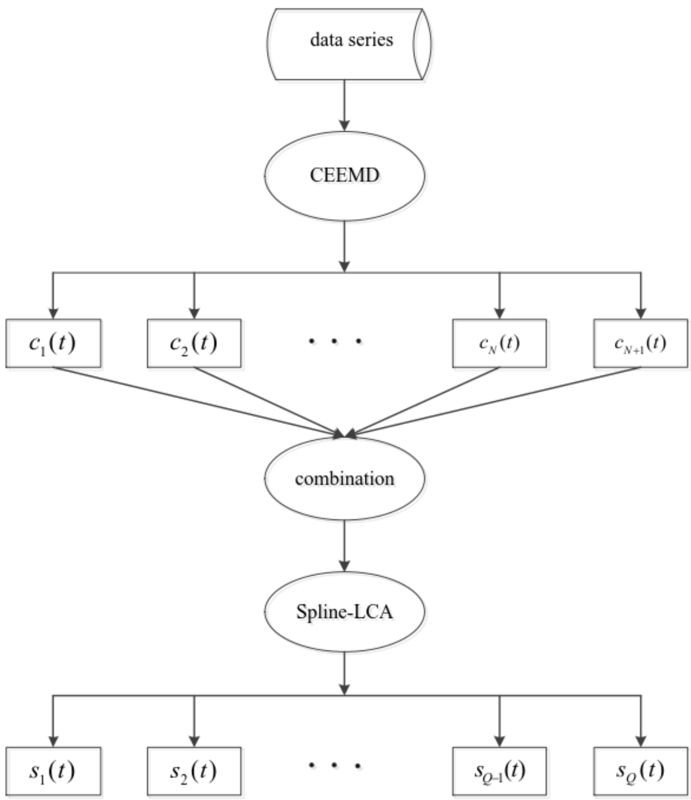

Fig. 1. The structure diagram of CEEMD-Spline-LCA.

learning to analyze the inner formation mechanism of the financial data. Particularly, IMFs, $c_1(t), \ldots, c_{N+1}(t)$, can be expressed linearly by a group of basis consisting of $s_1(t), \ldots, s_Q(t)$

$$\begin{cases} c_1(t) = a_{11}s_1(t) + \cdots + a_{1Q}s_Q(t) \\ c_2(t) = a_{21}s_1(t) + \cdots + a_{2Q}s_Q(t) \\ \cdots \\ c_{N+1}(t) = a_{N+1,1}s_1(t) + \cdots + a_{N+1,Q}s_Q(t)\,, \end{cases} \tag{14}$$

where the coefficient $a_{ij} = [A]_{ij}$ represents the contribution of j^{th} source onto i^{th} mixture.

In this section, we proposed an analytical method, CEEMD-Spline-LCA, whose purpose is to track internal formation mechanism hidden in the financial data. As shown in Fig. 1, the method is described as the following three steps: Firstly, CEEMD is utilized to obtain the IMFs from the financial data. Secondly, according to the selection of the threshold (the IMFs whose contribution coefficients for the financial data are less than threshold will be summed as a new mode function, and the rest will remain the same), the IMFs are reorganized to obtain a collection of new IMFs (NIMFs). Finally, by applying Spline-LCA to the NIMFs, a series of ICs are achieved.

4. Empirical analysis

4.1. Data resource

The Dow Jones Industrial Average (DJIA) is one of the important indicators of the stock market. Throughout the trend of the stock price, it has been going up from 2002 to 2019. However, the stock price plummeted in 2009 and 2019. The DJIA weekly data from April 2002 to August 2020 is selected as the empirical data in the Dow Jones Company, which is plotted in Fig. 2.

4.2. CEEMD-Spline-LCA results

As shown in Fig. 3, the DJIA data can be decomposed into 8 IMFs and one residual via the CEEMD.

Based on amplitude and frequency, we can roughly classify the components into four categories. The first category describes high frequencies, including IMF1, IMF2, and IMF3. These components have high frequencies and low amplitudes, which may be affected by some unexpected events, such as bad weather, good or bad mood, and behavior movement. The second category shows a moderate trend in frequency and amplitude, including IMF4 and IMF5. This section may have a good explanation for some important events, such as the subprime lending crisis of 2007. The third category

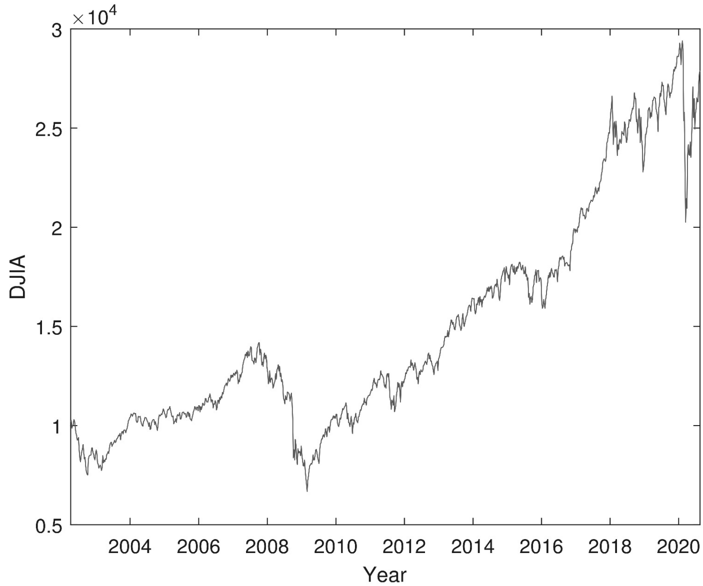

Fig. 2. The weekly DJIA data from April 2002 to August 2020.

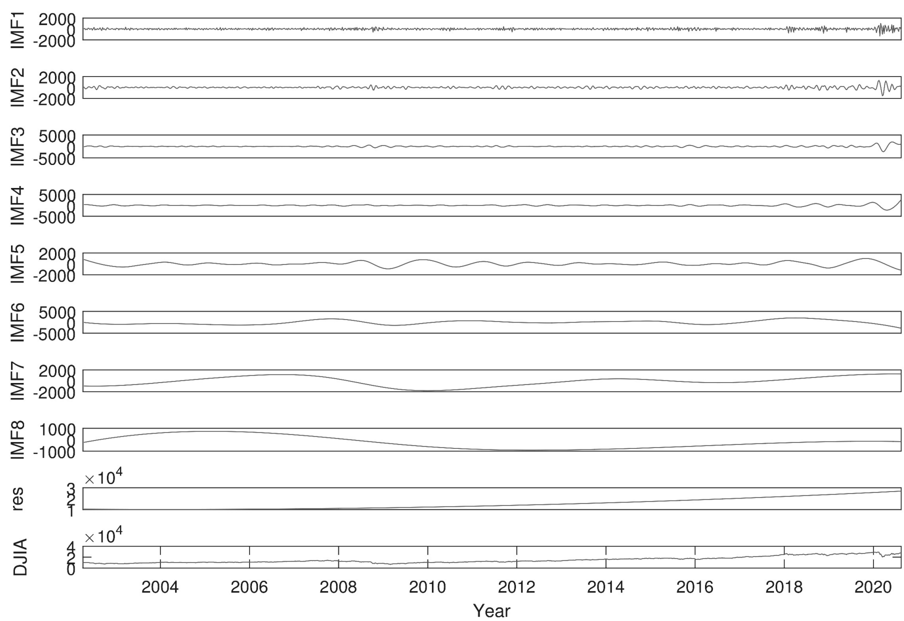

Fig. 3. The IMFs and residual for the DJIA.

(including IMF6, IMF7, and IMF8) describes low frequencies with no sharp fluctuations or periodic changes in this area. Exchange rates, interest rates and CPI may be factors that affect the DJIA. Finally, the residual series is the last category, which reflects the basic trend of the DJIA.

In order to analyze the influence factors of the stock price, we need to implement the multichannel signal construction. Firstly, referring to the excellent paper by Xian *et al.* [27], the contribution coefficients between the aforementioned IMFs and stock price are calculated in Table I. Then, the IMFs are recombined into several new IMFs according to setting the recombination threshold $\lambda = 0.3$, and the recombined IMFs are plotted in Fig. 4. Finally, five ICs, which are going to be analyzed comparatively in the forthcoming subsection, are decomposed by applying the Spline-LCA. The separated results are shown in Fig. 5.

TABLE I

The contribution coefficients of the IMFs and residual for the DJIA.

IMF1	IMF2	IMF3	IMF4	IMF5	IMF6	IMF7	IMF8	Residual
0.1712	0.1806	0.2223	0.2296	0.2568	0.4123	0.3914	0.3664	0.5658

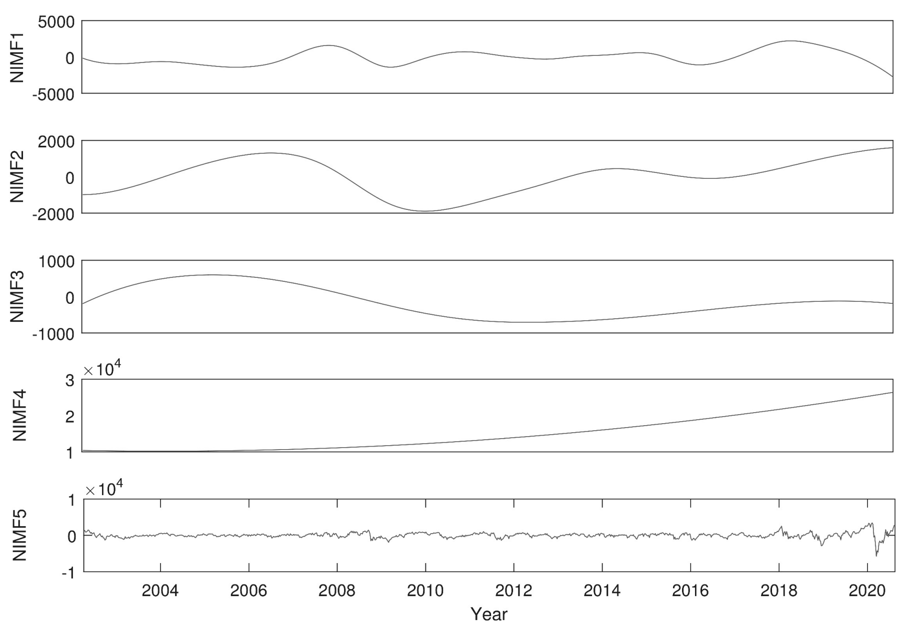

Fig. 4. The recombined new series.

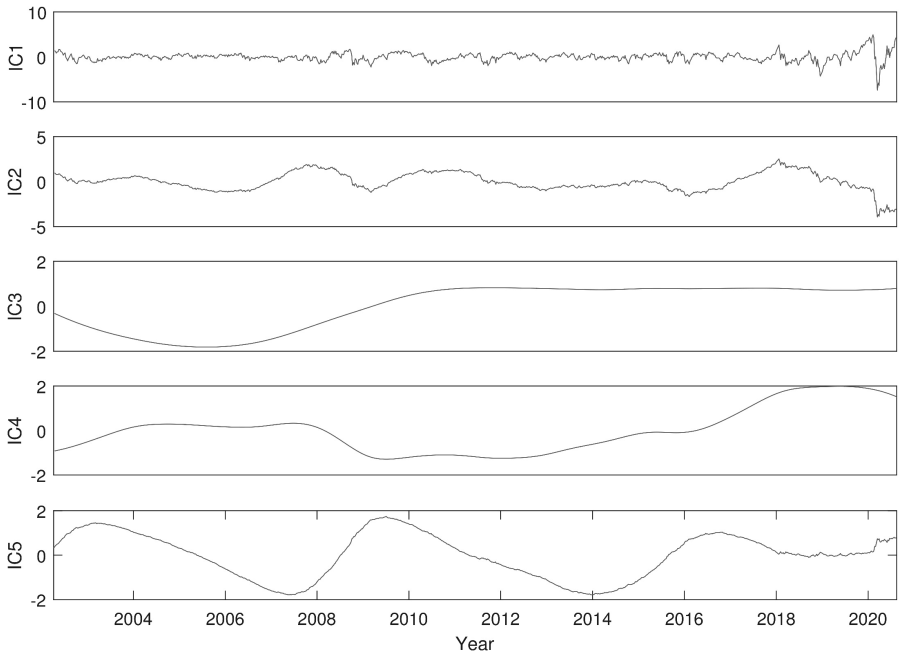

Fig. 5. The ICs for the DJIA from April 2002 to August 2020 by the CEEMD-Spline-LCA.

4.3. Descriptive statistics of ICs

This study trades on Spline-LCA to gain the five ICs and $\boldsymbol{W}_{\mathrm{s}}$

$$\boldsymbol{W}_{\mathrm{s}} = \begin{pmatrix} -0.3689 & 0.2477 & -0.4191 & -0.3343 & 0.7178 \\ -0.6135 & 0.0968 & -0.515 & 0.2888 & -0.5151 \\ -0.0636 & 0.3726 & 0.2428 & 0.8171 & 0.3610 \\ 0.6801 & 0.0676 & -0.6960 & 0.2192 & 0.0219 \\ -0.1449 & -0.8865 & -0.1244 & 0.2984 & 0.2978 \end{pmatrix}. \quad (15)$$

The reconstruction of DJIA is declared via Eq. (16)

$$\begin{aligned} \hat{x}(t) &= \sum_{i=1}^{5} m_i s_i(t) \\ &= 0.3488 s_1 - 1.1286 s_2 + 0.2975 s_3 + 0.7020 s_4 + 0.1528 s_5 \,. \end{aligned} \quad (16)$$

As shown in Table II, the ICs are analyzed in so far as mean, skewness, kurtosis, correlation, J–B, hurst and ADF.

TABLE II

Descriptive statistics of the ICs.

	Mean	Kurt.	Skew.	Cor.	J–B	m_i	Hurst	ADF
IC1	0.6665	11.9764	0.3526	−0.1341	$3239.6000^{(1)}$	0.3488	0.8539	$-7.9919^{(1)}$
IC2	0.7747	4.2381	−0.4124	−0.0451	$88.4253^{(1)}$	−1.1286	1.0000	$-0.8348^{(0)}$
IC3	0.9096	1.8333	−0.7466	0.5689	$143.4749^{(1)}$	0.2975	1.0000	$-3.8600^{(1)}$
IC4	0.7889	2.4503	0.6203	0.7981	$73.5840^{(1)}$	0.7020	1.0000	$-5.8499^{(1)}$
IC5	0.8337	1.9601	0.2531	0.1384	$53.4439^{(1)}$	0.1528	1.0000	$0.0949^{(0)}$

Note: Kurt., Skew., and Cor. denote the kurtosis, skewness, correlation, respectively. J–B is the Jarque–Bera Statistic; m_i is the combination coefficient of the ICs; Hurst represents the Hurst exponent; ADF is the Augmented Dickey Fuller Statistic.

J–B tests whether the time series is subject to a normal distribution. The null hypothesis is that the sample obeys a normal distribution. Index 1 means rejecting the null hypothesis, which demonstrates that the sample does not follow a normal distribution. It has been calculated that all ICs are non-Gaussian. Further, combining kurtosis and skewness, it can be seen that all ICs are super-Gaussian distributions, which have sharp peaks and long tails. This is also consistent with the ICA hypothesis that the components are non-Gaussian.

The Pearson correlation is applied to test the correlation between the ICs and the observed signal. The closer the absolute value of the indicator is to 1, the higher the correlation coefficient. Since the correlation coefficient between IC4 and the observed signal is maximum, this ingredient has the greatest effect on the DJIA.

ADF is used to test whether time series is stationary property. The null hypothesis is that the sample has a unit root, which means the series is stationary. Index 1 represents the rejection of the null hypothesis. Similarly, index 0 means accepting the null hypothesis. Table II shows that IC3 is stationary and the rest are non-stationary.

Hurst tests the memory of the ICs. If the indicator is close to 1, the time series can be described by random walk. When the indicator is between 0.5 and 1, it means that the time series has long-term memory. It can be seen from Table II that all ICs have a long memory.

4.4. ICs trend compare with significant influencing factors

The stock market, as an indicator of overall economic development, is a response and an early warning to regional markets. Applying the CEEMD-Spline-LCA model to the DJIA, the ICs reflecting the internal formation mechanism of the stock price are obtained, which can heuristically be considered as the influence factors of the stock price. For example, changes in the exchange rate will cause changes in the stock price [15]. The growth of stock price requires low-interest rate and low-price earnings ratios [2]. Further, the GDP, short-term risk-free market interest rate and inflation affect stock price fluctuations [17].

4.4.1. IC1

The comparison chart of CPI and IC1 is shown in Fig. 6. These data come from the Bureau of Labor Statistics. In fact, the rate of change in CPI largely reflects the degree of inflation or deflation. From 2008 to 2009, inflation caused the stock market to fall sharply. Under normal circumstances, inflation will have a certain impact on the stock market. However on a broader trend, the DJIA has continued to rise. We can understand that moderate and stable inflation has little effect on the stock price. Within a tolerable range to inflation, when the economy is in a prosperous stage, the stock price will continue to rise. With hyperinflation, the economy will be severely distorted, which causes the stock price to fall. To sum up, it not only stimulates but also suppresses the stock market.

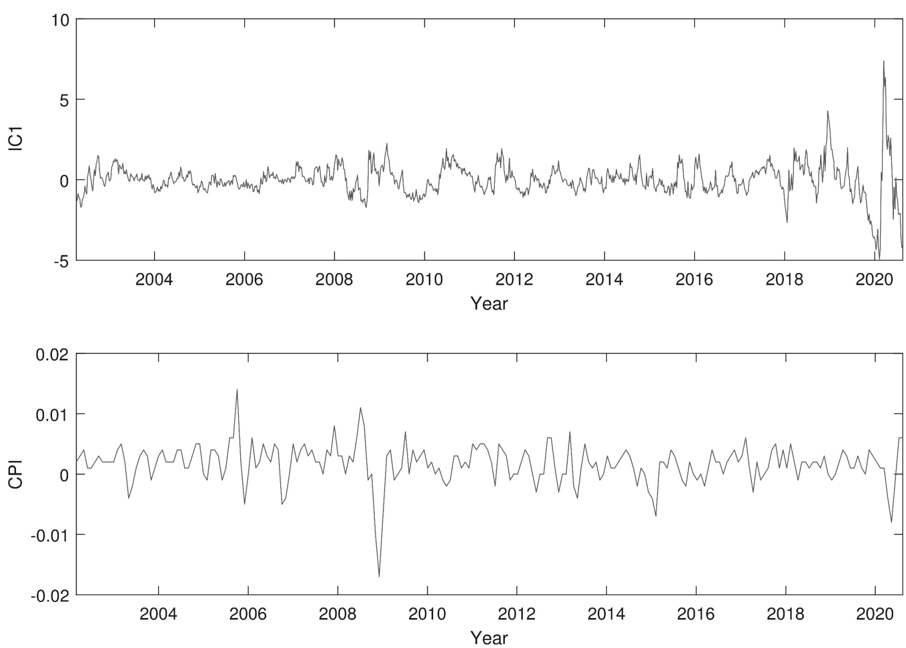

Fig. 6. The IC1 and CPI.

4.4.2. IC2

The USDX is a composite indicator of the exchange rate of the United States dollar in the international foreign exchange market. The data come from Intercontinental Exchange. The comparison of IC2 and USDX is plotted in Fig. 7. The relationship between USDX and DJIA is relatively complex. It has different relevance to the stock market in different periods. In the long run, there is a negative correlation between the two from 2002 to 2010. The decline in USDX will lead to an increase in the price of commodities denominated in the U.S. dollars, triggering an increase in the profits of resource and export companies, which will cause the stock price to rise. From 2014 to 2016, the two are roughly positively correlated. The positive U.S. economy and increased profits of listed companies may be the main reasons. It can be seen that the change of USDX will affect the degree of change of stock price to a certain extent.

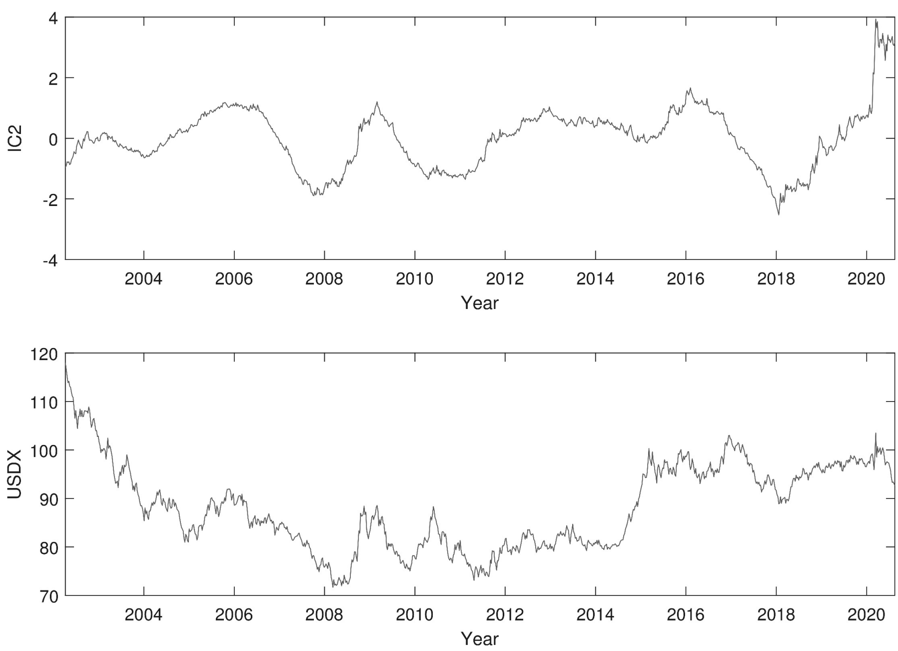

Fig. 7. The IC2 and USDX.

4.4.3. IC3

It can be seen from Table II that IC3 has a strong correlation with the DJIA and long-memory property in terms of the Hurst exponent. As shown in Fig. 8, IC3 is growing steadily from 2010 to 2020. For a more detailed

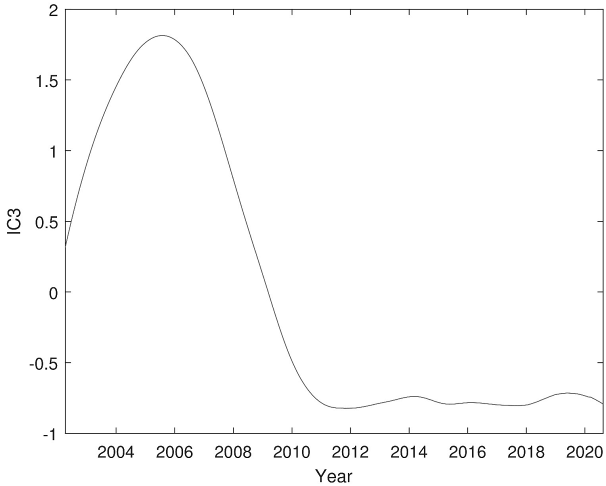

Fig. 8. The IC3.

description, we can choose the GDP growth rate from 2002 to 2019, which is shown in Fig. 9. The data are derived from the United Nations Conference on Trade and Development. The GDP, which indicates the sound development of the national economy, is the basic manifestation of a country's economic achievements. We can see that the two are basically the same. The GDP and DJAI continued to grow from 2002 to 2007, but until 2008, economic growth was negative, and the stock plummeted at the same time. The global

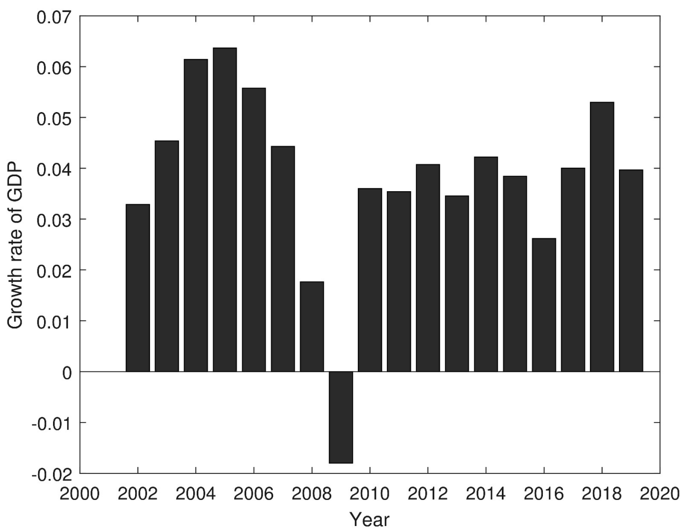

Fig. 9. The annual percentage growth rate of GDP from 2002 to 2019.

economy is mainly affected by the loss of the function of the financial system, which leads to the credit crunch, the rupture of enterprise capital chain, and the contraction of residents' income and wealth. Because the unreasonable economic structure promoted the rapid growth of GDP, it eventually caused the stock price to fall. In the long term, there is a strong correlation between the American macroeconomy and stock market, that is a wave in either direction causes another one to move in the same direction. Hence, good GDP growth and stock price growth are positively correlated.

4.4.4. IC4

The correlation coefficient and combination coefficient of IC4 are 0.7020 and 0.7981, respectively, which implies that IC4 has a long-term impact on the DJIA. With a few exceptions, IC4 has moved in line with the Fed funds rate in Fig. 10. We can get the data from the Federal Reserve Board. Stock market valuations are highly correlated with risk-free returns. Since 2004, the benchmark interest rate has been raised 17 consecutive times in two years, which finally punctured the real estate bubble, and the stock price began to slow down. Since the 2008 financial crisis, the Federal Reserve has begun to cut interest rates for the first time. The low-interest rate from 2009 to 2016 led to the DJIA growth. In 2018, the federal interest rate began to

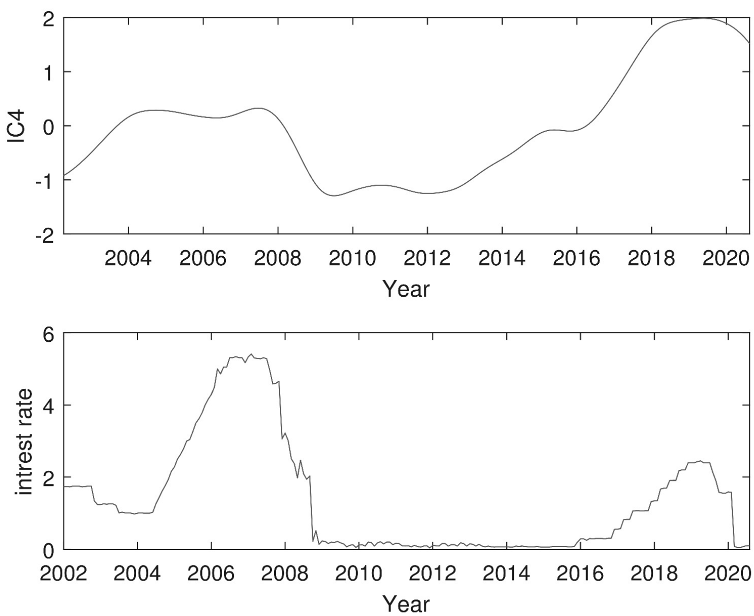

Fig. 10. The IC4 and Fed Fund Rate.

rise three times. This had a certain impact on the stock market, and the stock price began to fall in the fourth quarter. The higher the risk-free yield, the less attractive the stock market offers. When the interest rate rises, part of the funds will be transferred from investing in the stock market to bank deposits and buying bonds, which will reduce the demand for stock in the market and cause the stock price to fall. Conversely, when the interest rate falls, the profitability of savings will decrease, and some of the funds may return to the stock market, thereby expanding the demand for stock and increasing the stock price.

4.4.5. IC5

The trend of IC5 is characterized by randomness. An important influence factor that affects the DJIA volatility is international events (such as wars, financial crises, *etc.*). In the subprime lending crisis of 2007, Russia invades Crimea of 2014 and so on, detailed international events are described in Table III, all show roughly the same fluctuations as IC5. From 2007 to 2009, the basic trend of the stock price is that the decline rate is slow first and then rapid. The DJIA peaked at 14164.53 in 2007. Since then, it has been falling all the way until it bottomed out at 6547.05 in 2009. The main reason is that the subprime crisis has evolved into a global financial crisis. The trade dispute between China and the United States also caused the stock price to fall in 2019. From 2009 to 2018, there were also some important events, such as the oil spill in Mexico and Russia invading Crimea, which caused some fluctuations in the stock price. This also shows that international events will have some impact on the stock price. Although the occurrence of major international events will cause the stock price to surge or plummet, the mechanism of the stock price will slowly return to a normal state.

TABLE III

International events from 2003 to 2019.

Year	International events
2003	the war in Iraq
2007	subprime lending crisis
2010	oil spill in Mexico
2014	Russia invades Crimea
2019	trade disputes between China and the United States

5. Conclusions

It is important to find out the hidden main reasons that cause the financial data fluctuation. In this paper, a novel hybrid approach based on CEEMD and Spline-LCA (called CEEMD-Spline-LCA) is developed for seeking the most essential reasons that affect the fluctuation of the financial data. The financial time series is decomposed into several IMFs by CEEMD. Then, by analyzing the correlation between the financial time series and the IMFs, the new IMFs are obtained. In light of that the different reasons that affect the financial time series are usually independent, the Spline-LCA is used to abstract the ICs from the new IMFs. Simulation results on the weekly DJIA from April 2002 to August 2020 show the abstracted hidden ICs, which reflect the internal formation mechanism and are just the main factors which caused the volatility of the DJIA to some extent. What is more, the ICs are mutually independent, instead of being estimated or predicted jointly. They can be estimated one by one, hence, it lays the foundation for future prediction or classification of the financial time series precisely.

This work is supported in part by the National Natural Science Foundation of China (grant No. 61573014) and in part by the Fundamental Research Funds for the Central Universities and the Innovation Fund of Xidian University (grant No. YJS2107).

REFERENCES

[1] B. Zhu, P. Wang, J. Chevallier, Y. Wei, «Carbon Price Analysis Using Empirical Mode Decomposition», *Comput. Econ.* **45**, 195 (2015).

[2] H. Stein, «Presidential Economics: The Making of Economic Policy from Roosevelt to Clinton», *American Enterprise Institute for Public Policy Research*, 1988.

[3] G.W. Schwert, «Why Does Stock Market Volatility Change Over Time?», *J. Finance* **44**, 1115 (1989).

[4] B. Robert, «What Is Good for Goldman Sachs Is Good for America the Origins of the Present Crisis», Tech. Rep., Institute for Social Science Research, UCLA, 2009.

[5] P.S. Spiro, «The impact of interest rate changes on stock price volatility», *J. Portfolio Manage.* **16**, 63 (1990).

[6] J.H. Cochrane, «Production-Based Asset Pricing and the Link Between Stock Returns and Economic Fluctuations», *J. Finance* **46**, 209 (1991).

[7] D. Alexander, V. Pietro, «Inflation and earnings uncertainty and volatility forecasts: A structural form approach», Tech. Rep., Chicago GSB Research Paper, 2008.

[8] S. Zeisberger, D. Vrecko, T. Langer, «Measuring the time stability of Prospect Theory preferences», *Theor. Decis.* **72**, 359 (2012).

[9] Y. Ruan, L. Alfantoukh, A. Durresi, «Exploring Stock Market Using Twitter Trust Network» in: «IEEE 29$^{\rm th}$ International Conference on Advanced Information Networking and Applications», *IEEE Computer Society*, Los Alamitos, CA, USA 2015, pp. 428–433.

[10] A. Skabar, I. Cloete, «Investigation of the Effect of Training and Prediction Window Sizes on Neural Financial Prediction Models» in: Proceedings of the International Conference on Machine Learning and Cybernetics, 2003, pp. 2133–2137.

[11] L. Yu, H. Chen, S. Wang, K.K. Lai, «Evolving Least Squares Support Vector Machines for Stock Market Trend Mining», *IEEE Trans. Evolut. Comput.* **13**, 87 (2009).

[12] I.E. Livieris *et al.* «Performance Evaluation of an SSL Algorithm for Forecasting the Dow Jones Index Stocks», «2018 9$^{\rm th}$ International Conference on Information, Intelligence, Systems and Applications (IISA)», *IEEE Computer Society*, Los Alamitos, CA, USA 2018.

[13] H. Mehri-Dehnavi, H. Agahi, R. Mesiar, «A New Nonlinear Choquet-Like Integral With Applications in Normal Distributions Based on Monotone Measures», *IEEE Trans. Fuzzy Syst.* **28**, 288 (2020).

[14] S. Ravikumar, P. Saraf, «Prediction of Stock Prices using Machine Learning (Regression, Classification) Algorithms» in: «2020 International Conference for Emerging Technology (INCET)», 2020, pp. 1–5.

[15] I.S.A. Abdalla, V. Murinde, «Exchange rate and stock price interactions in emerging financial markets: evidence on India, Korea, Pakistan and the Philippines», *Appl. Financ. Econ.* **7**, 25 (1997).

[16] R. Aggarwal, «Exchange rates and stock prices: a study of the US capital markets under floating exchange rates», *Akron Bus. Econ. Rev.* **12**, 7 (1981).

[17] R.F. Engle, J.G. Rangel, «The Spline-GARCH Model for Low-Frequency Volatility and Its Global Macroeconomic Causes», *Rev. Financ. Stud.* **21**, 1187 (2008).

[18] A. Charles, O. Darné, «Large shocks in the volatility of the Dow Jones Industrial Average index: 1928–2013», *J. Bank. Finance* **43**, 188 (2014).

[19] A.D. Back, A.S. Weigend, «A First Application of Independent Component Analysis to Extracting Structure from Stock Returns», *Int. J. Neural Syst.* **08**, 473 (1997).

[20] T.-W. Lee, M. Girolami, T.J. Sejnowski, «Independent Component Analysis Using an Extended Infomax Algorithm for Mixed Subgaussian and Supergaussian Sources», *Neural Comput.* **11**, 417 (1999).

[21] A. Hyvärinen, J. Karhunen, E. Oja, «Independent Component Analysis», *John Wiley & Sons, Inc.*, New York, USA 2001.

[22] J. Dieter, J. Ranu, «Encyclopedia of Computational Neuroscience», *Springer-Verlag*, New York 2015.

[23] H. Li, H. Li, «Denoising of electric power system signals by ICA based on EMD virtual channel» in: 2009 Asia-Pacific Power and Energy Engineering Conference, 2009, pp. 1–4.

[24] P. Li, Z. Chen, Y. Hu, «A method for automatic removal of EOG artifacts from EEG based on ICA-EMD» in: 2017 Chinese Automation Congress (CAC), 2017, pp. 1860–1863.

[25] T. Jurić, M. Bonković, M. Rogić, «Application of EEMD-ICA algorithm to EMG signals measured in laryngeal muscles» in: «21st International Conference on Software, Telecommunications and Computer Networks — SoftCOM», 2013, pp. 1–4.

[26] B.A. Pramudita *et al.*, «Removing ocular artefacts in EEG signals by using combination of complete EEMD (CEEMD) — Independent Component Analysis (ICA) based outlier data» in: 2017 International Conference on Robotics, Automation and Sciences (ICORAS), 2017, pp. 1–5.

[27] L. Xian, K. He, K.K. Lai, «Gold price analysis based on ensemble empirical model decomposition and independent component analysis», *Physica A* **454**, 11 (2016).

[28] J. E, Y. Bao, J. Ye, «Crude oil price analysis and forecasting based on variational mode decomposition and independent component analysis», *Physica A* **484**, 412 (2017).

[29] J.R. Yeh, J.S. Shieh, N.E. Huang, «Complementary Ensemble Empirical Mode Decomposition: A Novel Noise Enhanced Data Analysis Method», *Adv. Adap. Data Anal.* **02**, 135 (2010).

[30] B.B. Risk, D.S. Matteson, D. Ruppert, «Linear Non-Gaussian Component Analysis Via Maximum Likelihood», *J. Am. Stat. Assoc.* **114**, 332 (2019).

[31] N.E. Huang *et al.*, «The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis», *Proc. Roy. Soc. A.* **454**, 903 (1998).

[32] X. Zhang, K. Lai, S.Y. Wang, «A new approach for crude oil price analysis based on Empirical Mode Decomposition», *Energ. Econ.* **30**, 905 (2008).

[33] N.E. Huang, Z. Wu, «A review on Hilbert–Huang transform: Method and its applications to geophysical studies», *Rev. Geophys.* **46**, 1 (2008).

[34] G. Gai, «The processing of rotor startup signals based on empirical mode decomposition», *Mech. Syst. Signal Proc.* **20**, 222 (2006).

[35] Z. Wu, N.E. Huang, «Ensemble Empirical Mode Decomposition: A Noise-Assisted Data Analysis Method», *Adv. Adap. Data Anal.* **01**, 1 (2009).

[36] A. Hyvärinen, E. Oja, «Independent component analysis: algorithms and applications», *Neural Networks* **13**, 411 (2000).

[37] T. Hastie, R. Tibshirani, J.H. Friedman, «The Elements of Statistical Learning», *Springer New York*, New York, NY 2009.

[38] B.B. Risk *et al.*, «An evaluation of independent component analyses with an application to resting-state fMRI», *Biometrics* **70**, 224 (2014).

[39] T. Hastie, R. Tibshirani, «Independent Components Analysis Through Product Density Estimation», *Adv. Neural Inform. Proc. Syst.* **15**, 665 (2002).

[40] R.S. Tsay, «Analysis of Financial Time Series», *John Wiley & Sons, Inc.*, New York, USA 2002.