IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 19, 2021, accepted January 20, 2021, date of publication February 1, 2021, date of current version February 9, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3055967

Research on Systemic Financial Risk Measurement Based on HMM and Text Mining: A Case of China Financial Market

ZHANFENG LI, YANLI CAI, AND SHULAN HU
School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan 430073, China

Corresponding author: Shulan Hu (hu_shulan@zuel.edu.cn)

This work was supported in part by the Project of the National Social Science Foundation of China: the Research on the Relationship between Major Risk Events and Stock Market Liquidity and Volatility under Grant 17BTJ034, and in part by the Project of the National Social Science Foundation of China: A Study on Unbalanced and Rebalancing Development of Inclusive Finance from a Spatial Perspective under Grant 18BJL077.

ABSTRACT The paper considered the sensitivity of unstructured network data to external shocks in the financial system, based on HMM applied in the traditional financial indicator system to construct the new composite index. We integrated economic statistical structure data and internet information, to capture the internal correlation and external shocks to financial markets. There appeared to be some evidence that the new index was superior at measuring the systemic financial risk. In addition, according to the new composite index we constructed, China's systemic financial risk was at the medium high level. It is an important task to prevent the systemic financial risk and maintain the stability of macro-economy.

INDEX TERMS Systemic financial risk, Baidu index, hidden Markov model, text mining.

I. INTRODUCTION

Since the 20th century, financial institutions in various countries have been expanding rapidly in the direction of globalization and liberalization, which has brought some challenges as well as opportunities to the world economy. The negative consequences caused by one financial institution's crisis would spread out to other financial institutions which had interest cooperation with it. Eventually the collapse of other financial institutions and even the whole financial system happened, that is systemic financial risk.

The report of the 19th National Congress of the Communist Party of China[1] indicated that China faced severe challenges caused by the growing pains of economic transformation in 2018. An interlacing of old and new issues, a combination of cyclical and structural problems brought changes in what was a generally stable economic performance, some of which caused concern. Meanwhile, China's externally-generated risks are on the rise. In order to ensure the sustainable development of financial and macro-economy market, the importance of systemic financial risk prevention has been gradually highlighted. The conference stressed that government would strengthen monitoring, early warnings, mitigation, control of financial risks and ensure that no systemic risks would emerge. For preventing systemic financial risk, much attention had been focused on the measurement and supervision of systemic financial risk.

In the late 1970s, the Bank for International Settlements(BIS, for short) put forward the concept of preventing systemic financial risk. Academics and relevant regulators all recognized it as three characteristic of harm, infectivity and impact on the real economy. Such as, the macroprudential policy tools and frameworks report issued jointly by International Monetary Fund (IMF, for short), BIS and Financial Stability Board (FSB, for short) [3]; Billio *et al.* (2012) [1]; Steven *et al.* (2013) [2].

There was a process of gradual accumulation before the outbreak of systemic financial risk, which made it possible to prevent systemic financial risk. Such kind of prevention had been paid attention after the financial crisis in 2008. Consequently, a certain system had been formed

The associate editor coordinating the review of this manuscript and approving it for publication was Derek Abbott.

[1]The report content was quoted from China daily.

in the research on systemic financial risk measurement. The research methods could be divided into four categories: the index method ([5]–[7]), the network model ([8]–[11]), the value at risk (VaR) series method ([12]–[14]) and the relevant default method ([15], [16]).

We interested in the composite index method. And the IMF (2009) suggested that the financial stability index constructed by the composite index method could be used as the main basis to measure the systemic financial risk in developing countries under the condition of underdeveloped financial markets.

The index method measured risk by constructing index system and synthesizing composite index with statistical method, such as [5], [6]. The authors chose the dependent variables which could reflect the financial risk and the independent variables which had a certain correlation with the dependent variables to make models for predicting the risk (such as [7]).

The index research of systemic financial risk measurement has attracted considerable attention, where multiple studies have offered many index systems, such as Gerard Jr and klingebiel (1996) [5]. In order to measure the risk more sharply and timely, Illing and Liu (2006) [17] tried to bring high-frequency and dynamic market data into the index system. Later, IMF (2003), Asian Development Bank (2004) and European Central Bank (2005) established successively relevant index systems either. Since the global financial crisis in 2008, a lot of relevant researches have been developed. Such as Wang and Hu (2014) [18]; Xu and Chen (2015) [20]; Wu and Hu (2016) [21];Yang and Wang (2019) [22] .

The challenge was that China's financial market statistical data was limited by the historical length, stability and continuity. It was not suitable to use historical data regression modeling for extrapolation prediction or monitoring methods based on market data. The composite index method was flexible, simple and complex. Therefore, the paper would propose a new index system to measure systemic financial risk based on HMM and text mining to measure systemic risk timely.

II. METHODOLOGY

Financial and economic news is continuously monitored by financial market participants [25]. The keyword search volume reflected the activity and prosperity of the economic market. The time series had not only linear changes and fluctuating clusters, but variance structure and regime changing. The section introduced hidden Markov model (HMM) and its related algorithms, Markov switching model, GARCH model and text mining for later analyzing.

A. HIDDEN MARKOV MODEL

Hidden Markov Model consists of two stochastic processes. The state sequence S^M is a Markov chain that is sometimes called the regime characterized by states and transition probabilities. The states of the chain are externally not visible, therefore ''hidden''. The observation sequence produces emissions observable O^N at each moment, depending on a state-dependent probability distribution and the inference. So, a HMM could define as (A, B, π). Here we considered HMM with $M = 2$, which contained two hidden states.

HMM is a discrete-time stochastic process $\left(S^T, O^T\right)$. The homogeneous Markov chain $S^T := \left\{S_t, t \in \mathbb{Z}^+\right\}$ with states $\mathcal{S} = \{0, 1\}$ and the observations $O^T := \left\{O_t, t \in \mathbb{Z}^+\right\} \in \mathbb{R}$.

(i) The unobserved state sequence $\{S_t\}$ is a time-homogeneous Markov chain with transition probability matrix $A = (a(i, j))$, i.e., for any integer $i, j \in \mathscr{S}$. The state transition probability matrix A is

$$A = (a(i, j)) = (P[S_t = j|S_{t-1} = i]), \tag{1}$$

and initial probability distribution is

$$\pi = (\pi(i_1)) = (P[S_1 = i_1]) \tag{2}$$

where $a(i, j)$ is the probability that the state at time t is j, is given when the state at time $t - 1$ is i. The structure of this stochastic matrix defines the connection structure of the model. The state transition probabilities should satisfy the normal stochastic constraints, $i, j \in \{0, 1\}$, $0 \leqslant a(i, j) \leqslant 1$, and $\sum_{j=0}^{1} a(i, j) = 1$.

(ii) The observations $\{O_t\}$ are conditionally independent giving the sequence of states of the Markov chain S_t, and the conditional distribution of O_t only depends on S_t. The observation symbol probability matrix B is

$$B = (b(i, o_t)) = (P(O_t = o_t|S_t = i)), \quad i \in \mathcal{S}, o_t \in \mathbb{R} \tag{3}$$

where $b(i, o_t)$ is the probability that observation symbol $O_t = o_t$ is emitted by state $S_t = i$. The following stochastic constraints must be satisfied: $i \in \{0, 1\}$, $o_t \in \mathbb{R}$, $0 \leqslant b(i, o_t) \leqslant 1$, and $\sum_{o_t \in \mathbb{R}} b(i, o_t) = 1$.

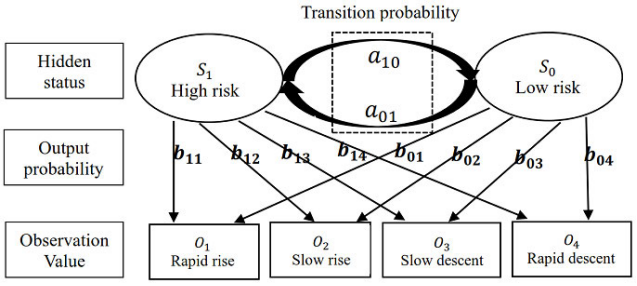

FIGURE 1. Example of HMM on systemic financial risk.

Systemic financial risk changes could be roughly divided into high-risk and low-risk states, such as [18], [19]. The example about systemic financial risk measurement made by 2−state HMM was shown in Figure 1. HMM has two basic properties. The first one is homogeneous hypothesis. HMM assumed that market state S_t in the current period t depended on the state S_{t-1} only, and the transition probability distribution between market states was

$$P(S_t|S_{t-1}, S_{t-2}, \ldots, S_1, O_1, \ldots, O_{t-1}) = P(S_t|S_{t-1}). \tag{4}$$

The second one is observation independence hypothesis. It assumes that the observation O_t is independent of all states and the observations before time t. The hidden state probability of the market risk state S_t mapped to the observation symbol of investor attention O_t was

$$P(O_t|S_t, S_{t-1}, \ldots, S_1, O_1, \ldots, O_{t-1}) = P(O_t|S_t). \quad (5)$$

There were many observation symbols here, such as fluctuations in stock market prices, trading volume, price index fluctuations, and industrial production fluctuations. We adopted the search volume of keywords as the observation value to analyze the potential systemic financial risk.

B. THREE BASIC ALGORITHMS IN HMM

HMM was proposed by Baum and Petrie [26] in 1966 as probability functions of Markov chains. Viterbi ([27]) gave Viterbi algorithm in 1967 which been used for dynamic planning widely. In 1970, Baum *et al.* [38], Baum [39] established the Forward-Backward algorithm in order to estimate the probability of the state where the observed values were. Based on the Forward-Backward algorithm, Dempster *et al.* (1977) [28] applied the Expectation-Maximization (EM) algorithm to establish a more general maximum likelihood estimation method of hidden Markov model parameters. Baum *et al.* [38], Baum [39] gave the local convergence property of the parameter estimation algorithm, named Baum-Welch algorithm in honor of Lloyd Welch [40].

There are three basic problems about HMM: identification problem, decoding problem and learning problem. The Forward-Backward algorithm solves the identification problem, Viterbi algorithm solves the decoding problem, and Baum-Welch algorithm solves the learning problem.

1) FORWARD-BACKWARD ALGORITHM

The Forward-Backward algorithm calculates the probability of a given observation sequence. This algorithm represents the best description of a given observation sequence by recursion to get the probability of any observation sequence.

Algorithm 1 The Forward Algorithm

Input: The observation sequence $O^T = \{O_1, O_2, \ldots, O_T\}$, and hidden state $S^T = \{S_1, S_2, \ldots, S_T\}$
Output: $P(O^T \mid \lambda)$

```
# Set initial value
for each state i ∈ [0, 1].
α(i, 1) = π(i)b(i, o_1),i ∈ [0, 1] do
end for
# Recursion
for each time step t from 1 to T − 1 do
    for each state i ∈ [0, 1] do
        α(i, t + 1) = Σ_{j=0,1} α(j, t)a(j, i)b(i, o_{t+1})
    end for
end for
P(O^T | λ) = α(0, T) + α(1, T)
return P(O^T | λ)
```

Algorithm 2 The Backward Algorithm

Input: The observation sequence $O^T = \{O_1, O_2, \ldots, O_T\}$, and hidden state $S^T = \{S_1, S_2, \ldots, S_T\}$
Output: $P(O^T \mid \lambda)$

```
# Set initial value
for each state i ∈ [0, 1].
β(i, T) = 1,i ∈ [0, 1] do
end for
# Recursion
for each time step t from T − 1 to 1 do
    for each state i ∈ [0, 1] do
        β(i, t) = Σ_{j=0,1} β(j, t + 1)a(i, j)b(j, o_{t+1})
    end for
end for
P(O^T | λ) = π_0 b(0, o_1)β(0, 1) + π_1 b(1, o_1)β(1, 1)
return P(O^T | λ)
```

2) VITERBI ALGORITHM

Viterbi algorithm is widely used for natural language processing, word segmentation and so on. The algorithm can find the most possible state of the observation through dynamic programming.

Algorithm 3 The Viterbi Algorithm

Input: The observation sequence $O^T = \{O_1, O_2, \ldots, O_T\}$, and hidden Markov model $\lambda = (A, B, \pi)$
Output: $S^T = \{S_1, S_2, \ldots, S_T\}$

```
# Set initial value
for each state i ∈ [0, 1]. do
    δ(1, i) = π(i_1)b(i, o_1)
end for
# Update values
for each time step t from 1 to T do
    for each state i ∈ [0, 1] do
        # Get maximum values
        max_δ = −1
        for each state j ∈ [0, 1] do
            tmp = δ(t − 1, j) ∗ a(j, i)
            if tmp> max_δ then
                max_δ = tmp
                pre_index[t][i] = j
            end if
        end for
        # Update values
        δ(t, i) = max_δ ∗ b(i, o_t)
    end for
end for
# Decode, find the maximum result value
decode=[−1 for i in T]
return S^T
```

3) BAUM-WELCH ALGORITHM

Baum-Welch algorithm determines the locally optimal θ by constructing an auxiliary function Q based on EM principle.

The maximum likelihood estimation method is used to estimate the parameters $\lambda = (A, B, \pi)$ by maximizing the probability of observation sequence $O^T = \{O_1, O_2, \ldots, O_T\}$ in $P\left(O^T \mid \lambda\right)$. Baum-Welch recursive algorithm completes the re-estimation of HMM parameters, that is to solve the learning problem. The Baum-Welch algorithm is as follows:

Algorithm 4 The Baum-Welch Algorithm

Input: The observation sequence $O^T = \{O_1, O_2, \ldots, O_T\}$
Output: hidden Markov model $\lambda = (A, B, \pi)$

```
# Set initial value
λ^(0) = (A^(0), B^(0), π^(0))
# Recursion
for time in range(max _iter): do
    Calculate the values of α(i,t),β(i,t),γ(i,t) and ε(i,j,t) respectively, under the current series A, B and π.
    # Update π.
    for each state S ∈ [0, 1] do
        π = γ(i, 1)
    end for
    # Update A.
    tmp1 = np.zeros(T − 1)
    tmp2 = np.zeros(T − 1)
    for each state i ∈ [0, 1] do
        for each state j ∈ [0, 1] do
            for each time step t from 1 to T − 1 do
                tmp1[t] = ε(i, j, t)
                tmp2[t] = γ(i, t)
            end for
            A[i][j] = np.sum(tmp1) / np.sum(tmp2)
        end for
    end for
    # Update B.
    for each state i ∈ [0, 1] do
        for each state k ∈ N_range do
            tmp1 = np.zeros(T)
            tmp2 = np.zeros(T)
            number = 0
            for each time step t from 1 to T do
                if k == O_t then
                    tmp1[t] = γ(i, t)
                    number+
                end if
                tmp2[t] = γ(i, t)
            end for
            if number == 0 then
                B[i][k] = 0
            else
                B[i][k] = sum(tmp1)/sum(tmp2)
            end if
        end for
    end for
end for
return A, B, π
```

C. MARKOV SWITCHING MODEL

Markov Switching Model allows for a given variable to follow a different time series process over different subsample [29]. Linear autoregressive processes with Markov regime are also widely used in several electrical engineering areas including tracking of maneuvering targets [30], failure detection [31] and stochastic adaptive control [32]. In general we can write a model of this kind as:

$$Y_t = \mu_{S_t} + \phi_{S_t} Y_{t-1} + \varepsilon_t, \tag{6}$$

$$\varepsilon_t | \mathscr{F}_{t-1} \sim \text{i.i.d. } N\left(o, \sigma^2\right), \tag{7}$$

$$\mu_{S_t} = \mu_0(1 - S_t) + \mu_1 S_t, \tag{8}$$

$$\phi_{S_t} S_t = \phi_0(1 - S_t) + \phi_1 S_t. \tag{9}$$

The proposal will be to model the regime S_t as the outcome of an unobserved 2-state Markov chain. Y_t is the sequence of samples, and $\mathscr{F}_t$ is the σ-algebra. The mean μ_{S_t}, the coefficient ϕ_{S_t} and the transition probabilities between states $a(i, j)$ are unknown parameters.

The logarithmic likelihood function of Markov transformation model can be obtained as follows:

$$\ln L = \sum_{t=1}^{T} \left[\sum_{s_t=0}^{1} \sum_{s_{t-1}=0}^{1} f\left(y_t | s_t, s_{t-1}, \mathscr{F}_{t-1}\right) f\left(s_t, s_{t-1} | \mathscr{F}_{t-1}\right) \right]. \tag{10}$$

The average duration of regime j is

$$E(D) = \sum_{j=1}^{\infty} jP(D = j) = \frac{1}{1 - a(j, j)}. \tag{11}$$

Keyword search volume fluctuation could be divided into two switching regimes, namely high attention and low attention regime. To establish a dual mechanism conversion of the Markov model on the Baidu index, it could be assumed that $S_t = 0$ as a low attention regime and $S_t = 1$ as a high attention regime, and Markov property was satisfied between state variables. Considering the characteristics of our data, a first-order autoregressive model might be applied.

D. GENERALIZED AUTOREGRESSIVE CONDITIONAL HETEROSKEDASTIC MODEL

GARCH model was proposed by Bollerslev (1986) [33], which was generalized from the seminal work on ARCH model by Engle (1982) [34]. The GARCH(1,1) model made in the paper has the following specification.

$$Y_t = \mu + \sum_{i=1}^{n} \alpha_i Y_{t-i} + \varepsilon_t, \tag{12}$$

$$\varepsilon_t = \eta_t \sqrt{h_t}, \tag{13}$$

$$h_t = \omega + \beta_1 \varepsilon_{t-1}^2 + \gamma_1 h_{t-1}. \tag{14}$$

where Y_t is the interested financial time series, ε_t is the residual series, h_t is its conditional volatility and η_t is an identical and independent innovation sequence. Therefore,

the parameter $\beta_1 + \gamma_1$ measures the volatility persistence, which means how fast the current risk shock to the volatility will die away (Ho *et al.*, 2013) [35]. In order to ensure that h_t is stationary and always positive, Bollerslev (1986) [33] suggested to apply the constraints $\beta_1 + \gamma_1 < 1$ and $\omega > 0$, $\beta_1, \gamma_1 \geqslant 0$.

E. TEXT MINING

The first step to do text mining is word segmentation. English words can be segmentated according to its own space separation, while Chinese has no space. So we need to have a special word segmentation tool. The most of word segmentations are based on statistical methods called language model. For language sequences $W_1, W_2, \ldots, W_n$, language model is to calculate the probability of the sequence, namely $P(W_1, W_2, \ldots, W_n)$. When a sentence is segmented in m ways as follows:

$$\begin{matrix} W_{11} & W_{12} & \ldots & W_{1n_1} \\ W_{21} & W_{22} & \ldots & W_{2n_2} \\ \vdots & \vdots & \ddots & \vdots \\ W_{m1} & W_{m2} & \ldots & W_{mn_m} \end{matrix} \tag{15}$$

The word segmentation is to find the maximum probability, that is

$$\arg\max_i P\left(W_{i1}, W_{i2}, \ldots, W_{in_i}\right). \tag{16}$$

Chinese word segmentation is the basis of other Chinese information processing, and it has many applications, such as machine translation, speech synthesis, automatic classification, automatic proofreading, and so on. Chinese word segmentation may affect some research, but it also brings opportunities for some enterprises. It is necessary to solve the problem of Chinese word segmentation for computer processing technology development.

III. TRADITIONAL COMPOSITE INDEX

Tao and Zhu (2014) [19] thought that the synthetic index method was adopted to build the systemic financial risk monitoring and calibrating system. The synthetic index model was built based on the historical data of the Chinese market. The systemic financial risk index constructed by the index method was almost based on structured statistical data. However, the news, reviews and searches online contained a lot of informations, structured data sometimes missed the part of informations. Liu and Xu (2015) [23] thought that the statistical structured data had different characteristics from the unstructured data online. The statistical structured data had low noise but was often lagged; the unstructured data was updated quickly, but the information was noisy and the data sources and forms were unstable. The timeliness of external shock information, such as economic policy changes, natural disasters and wars, is very important in systemic financial risk measurement. Wu and Chen (2018) [24] extracted information about China's systemic financial risk by text mining and web crawler technology from the articles in the newspaper, and constructed a index to measure China's systemic financial risk level.

The system financial risk measurement system of Wu and Chen [24] contained the text information online only, without data from the financial market and the economic market. Therefore, it failed to capture the endogenous factors in financial markets. According to the research of Tao and Zhu (2014) [19] and others ([18], [20]–[22]), we proposed a new index system. The traditional index system contained seven financial primary indicators including real estate market, stock market, currency market and so on. At the same time, the new index dimension was added as the eighth primary indicator which reflected the volume of explosive information online and the attention to financial markets.

A. TRADITIONAL INDEX FOR FINANCIAL MARKETS

In order to measure systemic financial risk scientifically, the selection of indicators should cover all aspects involved in the financial market as much as possible. It was necessary to include not only internal risk indicators, but also external equilibrium indicators. We analyzed the causes of systemic financial risk from two aspects of internal and external factors.

(i) The internal causes of systemic financial risk in China mainly include:

Firstly, the frangibility of financial system. Financial frangibility is risk accumulation in all financial areas. Due to the imbalance of social financing structure, the excessive proportion of indirect financing in the banking system, and the mismatching of assets and liabilities of some financial institutions represented by the ''shadow banking system'', the vulnerability of the financial system had been increased.

Secondly, the rapid development of financial innovation and comprehensive operation. Under the separate regulatory system, regulatory arbitrage, regulatory vacuum and other issues have emerged. The more prominent is the rapid development of cross industry and cross market financial products, such as some asset management businesses. What's more, some importance financial holding companies in finance system led to the transferring and diffusing of risks by its strong association with many other industries and institutions.

Thirdly, lending activities except with banks increased. These activities have evaded the financial regulatory requirements, such as capital adequacy ratio and deposit loan ratio. It weakened the effect of macro-control and strengthened the complexity, relevance and infectivity of the financial system.

Lastly, the moral hazard of the financial system. The central bank took the risk that should be taken by financial institutions or investors. The financial institutions had the impulse to engage in high-risk business excessively, the public risk awareness was weak, and the local government still intervenes in the financial industry.

(ii) The external factors of systemic financial risk were as follows. China was facing with excess capacity and high debt ratio of enterprises which lead to the increase of

non-performing loans in the banking industry, the appearance of inflation, and the distortion of capital allocation.

This situation increased a hidden risk to the sustainable and healthy development of the financial industry. In short, systemic financial risk was often the result of the interaction and co-evolution between the internal vulnerability and external factors.

There were many measurement methods of systemic financial risk in China financial market by Bisias, Flood, Lo & Valavanis (2012); Allen, Bali & Tang (2012) and the actual research of Tao Ling & Zhu Ying (2016). Through analyzing the correlation between indicators and deleting the highly relevant indicators, we selected seven market dimensions contained 23 indicators to build a composite indicator system of China's systemic financial risk.

The specific indicators of each market dimension had different impacts on systemic financial risk. Based on the research above, indicators were divided into three groups: negative, positive and two-way. Negative group indicated that the change of the variable would cause the reverse change of the market economy; positive group indicated that the change of the variable would cause the same direction change of the market economy, and the two-way group indicated that the risk change with standard deviation of the index. See Table 1 for specific indicators, and the significance of each indicator was shown in the appendix.

B. SYSTEMIC FINANCIAL RISK INDEX SYNTHESIS

The KMO value obtained through the test of principal component analysis was 0.840 and the probability-value of Bartlett's test of sphericity was 0 with the seven dimension indicators, so the samples were suitable for factor analysis. Principal component analysis was carried out on the variables of seven dimensions, 23 indicators. Seven principal components were gained after rotation according to the principle that the eigenvalue was larger than 1. The cumulative contribution rate was 85.938%. Finally, we obtained the synthesis by the simple weighted method. The time series diagram of the traditional composite index was shown in Figure 2:

It was possible to explain the changes in the traditional composite index properly from the empirical perspective of China's economic operation. The composite index had obvious seasonal fluctuations. Due to the Spring Festival in China, the composite index was at the tough point from January to February every year. Meanwhile, the index diagram could roughly showed the changing stages of China's economic market in recent years:

1) From March 2011 to May 2012, financial risks intensified again manifested by excessive loan growth, continuous expansion of debt quotas and severe overcapacity. Fixed asset investment, industrial added value, and GDP growth continued to decline. The composite index rose again.
2) From May 2012 to May 2014, the stock market was in a relative stable stage. There was sufficient flow in the currency market, interest rates gradually decreased, the growth rate of foreign exchange reserves re-entered the rising channel and the macro economy improved. The composite index fluctuated steadily.
3) From May 2014 to May 2015, in the first half of the year the market relied on leveraged funds to rise steadily. The passion of stock speculation and profit-making from the entire population appeared, and financial market funds gathered. As a result, the risk

TABLE 1. Composite indicator of systemic financial risk.

Market name	Variable name	Variable properties
Real estate market	Cumulative year-on-year increase in sales of commercial housing	Negative
	Unit sales price of commercial housing increased year-on-year	Negative
Stock market	Average price-earnings ratio	Two-way
	Year-on-year market value growth	Positive
Foreign exchange market	Foreign exchange accounts	Negative
	Effective RMB exchange rate	Negative
	Year-on-year growth in foreign exchange reserves	Negative
	Export value	Negative
	Year-on-year imports	Negative
Bond Market	6-month ChinaBond corporate bond and central coupon spread	Positive
	Spreads on 5-year Treasury bonds and 3-month Treasury bonds	Positive
	ChinaBond Composite Index (Total Value)/ Wealth Index	Positive
Government department	Industrial added value growth rate	Negative
	CPI year on year	Negative
	Cumulative total investment in urban fixed assets	Negative
Currency market	7-day repurchase fixed profit in the interbank market	Positive
	One-week and one-year SHIBOR interest rate spreads	Positive
	SHIBOR-LIBOR 1 week interest rate difference	Negative
Financial Institutions	M2 YoY growth rate / M1 YoY growth rate	Positive
	Medium and long-term loans / total loans	Positive
	Deposit ratio of financial institutions	Positive
	Growth of loans / Growth of industrial added value	Positive
	Short-term loan growth rate / industrial added value growth rate	Positive

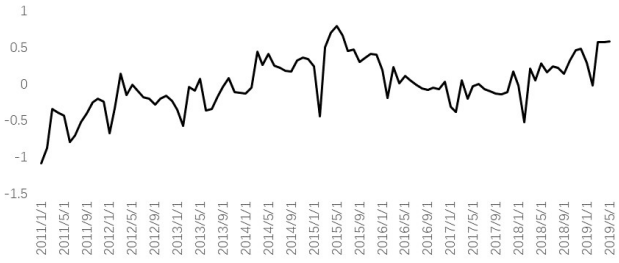

FIGURE 2. The time series of traditional composite index.

of funds accumulated gradually and the the composite index rose gradually.

4) From May 2015 to February 2016, the stock price collapsed, the policy was adjusted continuously and the capital allocation was handled recklessly. Investors suffered heavy losses and financial markets were affected significantly. Economic markets in other dimensions had also suffered a certain impact and the composite index was in a downward channel.
5) From February 2016 to April 2018, the stock market experienced large fluctuations, foreign exchange reserves continued to flow out and the problem of overcapacity continued to be exposed. As a result, the composite index continued to rise. It was in a steady fluctuation but still at a relatively high level.
6) Since April 2018, Sino-US trade friction had erupted, private enterprises had faced with credit risk, infrastructure investment had fallen, the stock market had plummeted, real estate bubbles had increased, and market expectations had been chaotic. In July 2018, the United States imposed tariffs on Chinese products which worth $34 billion. Subsequently, Sino-US trade frictions continued to escalate. China's foreign trade, foreign investment utilization and technological cooperation were affected. Deep-seated problems were beginning to emerge such as weak technological innovation capabilities, fragile industrial and supply chains. These problems overlapped and affected each other, leading to the continuous accumulation of financial risks and increasing economic fluctuations

IV. NEW COMPOSITE INDEX

The composite index based on the traditional economic data was almost in the top state when the financial systemic risk occured [19]. Compared the inflection point change and trend of the composite index with the influence range of major risk events, the measurement of the traditional systemic financial risk index were consistent with the real events, but lack of predictability. In order to improve the predictability, we would develop a new index system which is more complete than the tradiontional index system.

A. FINANCIAL HIGH-FREQUENCY WORD ACQUISITION AND KEYWORD SCREENING

Financial news and related information dissemination had a certain impact on financial risks. The occurrence of systemic risk events could cause high attention of investors. And when the attention of investors to a financial event rose sharply, it would cause some fluctuations in the financial market. Especially, an urgent risk event might cause high attention in a short time.

Considering the time-sensitive, stability and totality of data, we obtained keywords related to systemic financial risk from the East Money website[2] as the main resources. The search data could not be downloaded directly from Baidu, so we get the post text data of all the big financial bars through the web crawler, such as stocks bar, fund bar and futures bar in East Money website from January 2011 to July 2019. Then, the crawled text data was classified and frequency counted as shown in Table 2:

TABLE 2. Frequency of words.

Words	Frequency	Words	Frequency
Company	305732	Retail investor	178921
market	269402	shareholder	178021
stock	264657	Daily limit	177729
announcement	242398	funds	159823
Main force	239299	stock market	143245
Combination	219902	Bank	142098
Shares	218061	Bounce	132903
China	211454	Down	129876
share price	200235	Firm offer	127987
investment	199192	index	119982

The frequency refers to the number of times that the keyword appeared in the text of post bar in sample period. It reflected indirectly the economic prosperity and financial market activity. Here, the top 70 words in the frequency ranking were selected as the keyword lexicon, as shown in Table 3. We crawled the daily search volume data of each keyword in the lexicon and selected the top 20 for analysis.

TABLE 3. High-frequency word screening results.

category	Key words
Candidate Thesaurus (70)	Company; Market; Stock; Announcement; Main; Portfolio; Shares; China; Stock-Price; Investment; Retail; Shareholders; Daily Limits; Funds; stock market; broader-market; rebound; down; firm index; market trend; buy; positive; gold; Oil; suspension; performance; hopes; reduce; bank; technology; opportunities; new-shares; drop-limit; listing; operating; agencies; Securities; Rise; Acquisition; Chip; Restructuring; Shipment; Dividend; Value; Sector; New-Third-Board; Financing; Callback; Crude-Oil; Financing; Risk; Cost; Unraveling; Time; Technology; Trends; Major; Clearance; Holding; Stock Makers; Close; Pulling; Late; Market Rate; Increase Stock; Picking; USD
Selected Key words (20)	Market; Main-force; Bank; Stock; Stock-market; company; share price; Good; Suspension; Performance; Buy; funds; shareholder; Ship; Financing; opportunity; Down; Shares; limit up; limit down

B. A NEW DIMENSION

We obtained 102 samples through the keywords' monthly search volume from January 2011 to July 2019. We made the model on the logarithmic difference of the data.

[2]To demonstrate the effectiveness of proposed model, four different sources of China finance news are collected. The news articles about stocks, foreign exchange, bond and so on from the East Money, text based technical analysis of each stock from Straight flush, user comments from both Weibo and Baidu Finans platforms are gathered.Limited by the length of the paper, we only showed results based on the East Money

The model for analyzing the growth rate of attention was a first-order lagging model with switching regime, according to the sample characteristics, log-likelihood value and information quantity criterion. The estimation results of (6)-(9) were as follows:

TABLE 4. The first-order lagging Markov switching model estimation results.

parameter	μ_0	ϕ_0	μ_1	ϕ_1	σ^2
estimated value	-0.0163	-0.6533***	0.0564	0.6386***	-1.6017***
Standard deviation	0.0316	0.1404	0.0430	0.1490	0.0813
Log-likelihood	$\ln L = 2.396868$				
Information criterion	$\mathrm{AIC} = 0.0912$		$\mathrm{SC} = 0.2724$		

According to Table 4, the two values μ_0 and μ_1 were not significant in 1% significance. The growth rate of keyword searches could be described by a two-state Markov switching model.

The first regime had a negative growth rate, because the parameter μ_0 was -0.0163, ϕ_0 was -0.6533. That was to say the changes of attention for financial markets in the first regime was relatively stable. The attention of investors had small oscillation amplitude changes. It was because investors would be distracted by a large number of information in the absence of risk events.

The second regime had dramatic changes, with investors' interest in financial markets rising rapidly and dramatically. μ_1 was 0.0564, and ϕ_1 was 0.6386. It was clear that investors' attention on the financial market had an explosive growth once into this regime. The changes of investors' attention under the high-risk regime were much greater than under the low-risk regime, which was consistent with the actual situation.

TABLE 5. The average duration of two regimes.

Transition probability	Probability value	Status	Average duration (months)
p_{00}	0.5677	Low attention in the financial market	2.3134
p_{11}	0.3805	High attention in the financial market	1.6142

The transition probabilities between the attention regime switching were shown in Table 5. We could get $p_{00} = 0.5677$, that is, when the market keyword search volume in the current month was in a low-attention state, the probability of staying in the original regime next month was 0.5677. $p_{11} = 0.3805$ was the probability still staying in the high-atttention state. $p_{01} = 1 - p_{00} = 0.4323$ was the probability of switching the low-attention state to the high-attention state. According to the analysis of regime duration in economic phenomenon, the average duration of low attention in the financial market was 2.3134 months, and the average duration of high attention in the financial market was 1.6142 months.

It could be seen that the stopping time in the low-attention state was longer than in the high-attention state. The high-attention state and impact caused by a systemic financial risk event would last about 48 days. It was shown that the financial market was maintained a relatively stable situation, but its fluctuation mechanism could not be ignored.

Combined with the previous analysis, the paper considered adding the eighth dimension to improve systemic financial risk index system. The specific indicators were shown in Table 6. The eighth dimension reflected the concerns and intentions of investors. Compared with the traditional indicators, the new composite index had a higher data quality.

TABLE 6. The index system of eighth dimension.

	Variable name	Economic significance	Variable properties
Investor attention of finance	Monthly search volume of each selected keyword	Reflect the potential of systemic financial risk events	Positive
	Monthly percentage of high attention days	When the proportion is high, the probability of systemic financial risk is high	Positive
	Average duration of high attention	Reflect the impact of potential systemic financial risk events	Positive

Among them, the monthly Baidu search volum of the key words, such as ''market'', ''main force'' and ''stock market'', were selected from the above. Now searching online was the main way for the public to obtain the detailed information of the financial market. According to the analysis above, the stopping time of the low-attention state was 1.43 times as much as of the high-attention state. Systemic financial risk events ferment and their impacts increased as the length of the investors' attention grew. Therefore, the duration of attention variable could be used to measure the impact of risk events.

V. EMPIRICAL RESULTS

One important characteristic of systemic financial risk was its impact on the real economy based on the previous analysis. Therefore, the paper measured the systemic financial risk from the impact of the real economy, and selected the growth rate of real industrial value-added as a representative of macroeconomic variables. The sample period was from January 2011 to July 2019. The growth rate of real industrial value-added was supplemented by interpolation, and seasonal factors were eliminated through seasonal differences.

A. GARCH MODEL

In the paper, we made GARCH model on industrial value-added according to Engle(2018)[36] to measure systemic financial risk. the daily average return was equal to 20.78% with a standard deviation of 0.1859. The returns were positively skewed, while the excess kurtosis suggested leptokurtic behaviour.

The Jarque-Bera test showed that the sample data had skewness and kurtosis matching a normal distribution, while the ARCH test for conditional heteroskedasticity confirmed that there exist ARCH effects in the returns of the industrial value-added. So, the autoregressive model for the conditional mean could not reflect the sequence characteristics fully, an autoregressive conditional heteroskedasticity model should be added to measure conditional variance. In addition, according to the results of both the Augmented Dickey-Fuller (ADF) and Phillips-Perron (PP) unit-root tests stationarity was guaranteed.

TABLE 7. Result of GARCH model.

	μ	α_1	ω	β_1	γ_1
Coefficient	0.009169	0.9687	0.001489	0.527021	0.384532
P-Value	0.1918	0.0000	0.0000	0.0011	0.0000

Table 7 showed the estimation results of the GARCH models. The two information criteria selected the AR(1)-GARCH (1,1) model. All the parameter estimates were statistically significant for the AR(1)-GARCH(1,1) model. While the results of the ARCH and Q tests, which had been used as diagnostic tests, indicated that the selected AR(1)-GARCH(1,1) model was appropriate for the industrial value-added. The prediction accuracy of GARCH model for the systemic financial risk was 64.12%, while the prediction accuracy of composite index was 32.81%. The GARCH model had a higher prediction accuracy, but single GARCH model could not analyze the correlation between different markets or different assets. Multivariate GARCH model was usually used to analyze the correlation between multiple markets, but there were some limitations in parameter estimation and multivariate distribution assumption [37]. Therefore, the paper would continue to study the composite index method to supplement the missing part.

B. COMPOSITE INDEX EVALUATION

Following the method of Giglio, Kelly and Pruitt (2016), we performed the following autoregression on the growth rate of real industrial value-added (Y_t):

$$Y_t = c + \sum_{i=1}^{p} \alpha_i Y_{t-i}. \tag{17}$$

The autoregressive order p was determined by the AIC criterion, and the macroeconomic shock was represented by the residual term of the autoregression. The results of the auto-regression model for the growth rate of real industrial value-added were shown in Table 8:

TABLE 8. Autoregressive model results of the growth rate of real industrial value-added.

	c	y_{t-1}	y_{t-3}	y_{t-12}
Coefficient	-0.112108	0.518077	0.207980	-0.237455
(P value)	(0.0006)	(0.0000)	(0.0227)	(0.0037)

The timing diagrams of the macroeconomic and composite indices were shown in Figure 3.

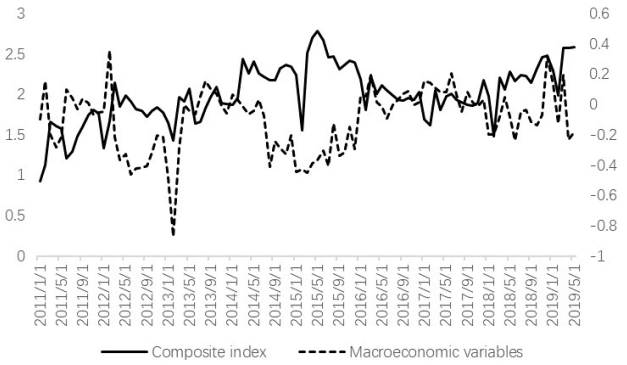

FIGURE 3. Macroeconomic variables and composite index time series.

Figure 3 was a line chart of the systemic financial risk composite index and macroeconomic index over time during the sample period. In order to facilitate the observation of their trends, the paper had standardized these two indices. It could be seen clearly from Figure 3 that the overall trend of these indices were similar, and the convergence of volatility was high relatively. During the domestic economic downturn in 2012 and the second half of 2015, China's stock market fell sharply and risk oscillating rose. While it had fluctuated violently and frequently around 2015, which reflected the repeated intensification of financial risk. Here, we introduced the prediction trend accuracy to evaluate the risk prediction ability of the systemic financial risk composite index. Prediction trend accuracy, that was measured the average of correct times a model predict the trend of risk. The prediction trend accuracy of new composite index is 59.6%, while the accuracy of traditional composite index is only 21.3%.

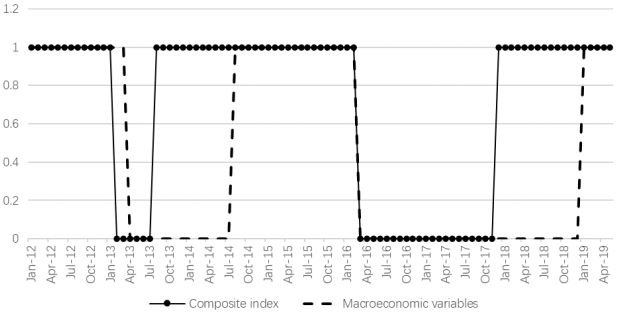

FIGURE 4. Macroeconomic variables and high volatility regional map of composite index.

In Figure 4, it is shown that the new composite index had a certain leading over the volatility of macroeconomic variables. Taking the second half of 2013 and the first half of 2014 as examples, the leveraged funds accessed to the financial market, financial market capital accumulated, so the

TABLE 9. The index of first-order lagging measured the degree of risk in each market.

	Real estate market index est_t	Stock market index sto_t	Foreign exchange market index exc_t	Bond market index bond_t	Government sector index gov_t	Money market index mon_t	Financial Institution Index fin_t	Macro-economic index mac_t
Macroeconomic	0.2523	0.0402	-0.1101	0.0903	-0.0604	0.0470	-0.0441	0.6033
index	0.0024	0.4264	0.0153	0.0211	0.3629	0.4383	0.3068	0.0000
mac_{t-1}	0.1018	0.0072	0.0664	0.0603	0.0096	0.0070	0.0121	0.3692
composite	-0.2060	0.2191	0.2764	0.0222	0.2461	-0.0953	-0.2126	-0.3835
index	0.0481	0.0003	0.0000	0.6500	0.0020	0.2010	0.0000	0.0029
z_{t-1}	0.0447	0.1444	0.2761	0.0024	0.1052	0.0189	0.1855	0.0983

[1] The first line of data in each cell was the regression coefficient, the second line was the t-test of the corresponding coefficient, and the third line was the goodness of fit of the regression model.

systemic financial risk had gradually gathered. Meanwhile, the new composite index had rose sharply based on the increasing volume of attention in financial markets, and captured the potential risks of financial markets in advance. In the third quarter of 2014 and the first quarter of 2016, the stock market was faced with big ups and downs, as the new composite index reacted to the situation. In 2016-2017, it showed a lower level of risk. At this stage, as we could see from Figure 3 that the new composite index was little leading but more cooperative than the fluctuation of the real economy.In 2018, China's foreign trade frictions escalated and the domestic economy would carry out structural reforms. In July 2018, the Sino-US trade war broke out, China's foreign trade and macroeconomics had been greatly affected. According to the central economic work conference in December 2014, China is in the ''new normal'' of economy. Compared with the stock market crash in 2015, the new composite index has fallen and stayed at a stable medium and high level now. The new composite index could capture the high degree of attention on these issues, to measure the accumulation of systemic financial risk. Comparing the new composite index with the macroeconomic index in high-risk areas, it takes several weeks to transmit the finacial risks to the real economy and to cause shocks in the real market. The macroeconomic index was lagged to measure systemic financial risk, but the new composite index was more sensitive to identify the high-risk. For example, in the second half of 2013 and the beginning of 2018, high risks were identified at an early stage. And the high-risk measurement accuracy of the composite index was 69.7%.

According to figures 3 and 4, China's systemic financial risk level was in the later stages of the financial crisis, the systemic financial risk index was falling back gradually and economy was improving. With the economy out of recession, the systemic financial risk index fluctuated slightly at a low level. When the capital market bubble stacked up until the stock market crash, the risk of the financial system soared again and the relevant policies of the stock market came out. After the stage, the financial market entered a stable period. Now the real estate bubble accumulated and the external trade situation was severe. The systemic financial risk index rose gradually after a low level in 2017. At present, China's economy was facing the pain of transformation, while the real economy was struggling ahead and the systemic financial risk index remained at a medium-high level. Therefore, the regulatory authorities should be on guard against systemic financial risk.

Table 9 showed that the new composite index had a significant leading for the indices of financial sectors compared with the macroeconomic index, in addition to the real estate market and the bond market. For example, the new composite index could explain the 14.44% volatility in the stock market in the next period. The out-of-sample forecast statistics of the composite index were more significant than the macroeconomic index, because the new composite index could provide more information than the macroeconomic index only based on its historical data. The new composite index was significant in the regression forecast of the five markets dimensions. This showed that each index of different financial sectors could capture some of the risk factors from different angles. On the other hand, it also showed that China's systemic financial risk affected the real economy from different market paths. All in all, the new composite index could help to get a better risk prediction for the stock market, foreign exchange market, government departments, and financial institutions, with R^2 all around 10%.

Meanwhile, the significance results of the new composite index predicting the first-order lagging and second-order lagging of seven markets index were almost the same in Table 10, except for the foreign exchange market. The forecast significance for first-order lagging foreign exchange market was stronger than second-order lagging. For the macroeconomic index, its autoregressive prediction was superior significant than that of the composite index in first-order lagging. However, the prediction result of second-order lagging was to the opposite, the new composite index was superior to the macroeconomic index with a significant difference. This result indicated that the new composite index could predict the macroeconomic shocks two months in advance. The self-change rules could be predicted one-period in advance through the autoregression and the external shocks could be captured by the new composite index with two months in advance.

TABLE 10. The index of second-order lagging measured the degree of risk in each market.

	Real estate market index est_t	Stock market index sto_t	Foreign exchange market index exc_t	Bond market index bond_t	Government sector index gov_t	Money market index mon_t	Financial Institution Index fin_t	Macro-economic index mac_t
Macroeconomic	0.2568	0.0679	-0.0658	0.1015	-0.0052	0.0388	-0.0989	0.4043
index	0.0019	0.1637	0.1512	0.0096	0.9378	0.5258	0.0175	0.0001
mac_{t-2}	0.1076	0.0227	0.0241	0.0763	0.0001	0.0048	0.0646	0.1758
composite	-0.2480	0.1978	0.2835	0.0017	0.2371	-0.0941	-0.1684	-0.6034
index	0.0165	0.0007	0.0000	0.9718	0.0029	0.2114	0.0009	0.0000
z_{t-2}	0.0658	0.1261	0.2930	0.0000	0.0993	0.0183	0.1227	0.2567

[1] The first line of data in each cell was the regression coefficient, the second line was the t-test of the corresponding coefficient, and the third line was the goodness of fit of the regression model.

VI. CONCLUSION

This paper proposed a new method to measure systemic financial risk. The study began with traditional measurement index system, including the internal and external risk analysis of financial market and index selection. Then a new index dimension based on HMM, drawing on search volume for keywords was given. The new index showed superior performances compared to the traditional one.

The traditional composite index composed by Tao and Zhu [19] and Wu and Chen [24] could reflect the financial risk level of our country in the sample period, and Wu and Chen [24] found the risk coming paths. The traditional index captured high-risk points, such as, the "money shortage" event in 2013, the "stock market crash" event in 2015 and the trade war between China and the United States in 2018 with lagging. The new index alerted the systemic risk about 1-2 months in advance and ensured the risk measurement functioned continuously and effectively through the attention of the financial market. The macroeconomic index had a lagged response after the shock due to the complete dependence on the market information response. Additionally, the new composite index was superior predictable than the macroeconomic index in the high-risk area. When located in the low-risk area, the new composite index was little leading and much cooperative with the fluctuation of the macroeconomic index.

The composite index constructed in the paper showed that China's systemic financial risk was at mid-high level in these years. There were some changes in the statistical caliber of China's financial data during the transferring of economic system. In addition, many problems appeared in China's financial market, such as financial market instruments and trading products were insufficient and the risk hedging mechanism was not complete. The new composite index composed of investors' attention and the financial structure data improved the foresight and accuracy of early warnings to systemic financial risk. But there were some deficiencies in the analysis of the complexity and relevance characteristics of the financial system. In our following research, we would carry out the various possibilities and paths of the risk evolution process. So, market regulators could optimize the market behavior by supervising important parameters, and control the risk level within the acceptable range finally.

APPENDIX

A. SOME SUPPLEMENTS FOR FORWARD-BACKWARD ALGORITHM

We introduced forward probability and backward probability for the convenience of calculation. The forward probability is written as $\alpha\,(i,t)$,

$$\alpha\,(i,t) = P(o_1, o_2, \ldots, o_t, S_t = i). \tag{18}$$

The backward probability is written as $\beta\,(i,t)$,

$$\beta\,(i,t) = P(o_{t+1}, o_{t+2}, \ldots, o_T | S_t = i). \tag{19}$$

$\gamma\,(i,t)$ represents the probability that the state at time t is $S_t = i$ with the observation sequence O^T is given.

$$\begin{aligned}\gamma\,(i,t) &= P\left(S_t = i \mid O^T, \lambda\right) \\ &= \frac{\alpha\,(i,t)\,\beta\,(i,t)}{\sum_i \sum_j \alpha\,(i,t)\,a(i,j)b\,(j,o_{t+1})\,\beta\,(j,t+1)}.\end{aligned} \tag{20}$$

B. SOME SUPPLEMENTS FOR VITERBI ALGORITHM

According to equation (20), we have a definition of $\gamma\,(i,t) = P\left(S_t = i \mid O^T, \lambda\right)$. Equation (8) can find the optimal state at each moment, but the state combination may not be the sequence state with the maximum of the whole. We found the global optimal solution based on Bayes' formula.

$$\begin{aligned}S^T &= \arg\max_{S^T} P\left(S^T \mid O^T\right) \\ &= \arg\max_{S^T} P\left(S^T, O^T\right).\end{aligned} \tag{21}$$

At this point, the problem turned to find the state sequence which maximized probability $P\left(S^T, O^T\right)$. We defined $\delta(T)$ before solving the optimal solution:

$$\delta(T) = \max_{S^{T-1}} P\left(S^T, O^T\right), \tag{22}$$

$$\delta(i,T) = \max_{S^{T-1}} P\left(S^{T-1}, S_T = i, O^T\right). \tag{23}$$

In order to maximize the entire sequence, we should initialize firstly.

$$\delta(1) = P\left(S_1, O_1\right) = P\left(S_1\right) P\left(O_1 \mid S_1\right), \tag{24}$$

$$\delta(i,1) = P\left(S_1 = i, O_1\right) = \pi(i_1) b(i, o_1). \tag{25}$$

We make recursion after initialization to find $S_T = i$ which maximize $\delta(i, T)$.

$$\begin{aligned}\delta(T) &= \max_{S^{T-1}} P\left(S^T, O^T\right) \\ &= P\left(O^T \mid S^T\right) \max_{S^{T-1}} \left[P\left(S_T \mid S_{T-1}\right) \delta\left(T-1\right)\right], \end{aligned} \tag{26}$$

$$\begin{aligned}\delta(i, T) &= \max_{S^{T-1}} P\left(S^{T-1}, S_T = i, O^T\right) \\ &= b(i, o_T) \max_j \left[a\left(i, j\right) \delta(j, T-1)\right]. \end{aligned} \tag{27}$$

We found the last state $S_T = i$ in the maximization sequence. Then we could find the penultimate state from back to front by the dynamic programming. That is to find $S_{T-1} = j$ satisfied $\delta(i, T) = b(i, o_T)a\left(i, j\right) \delta(j, T-1)$. And so on, the maximum possible state of each moment could be found out step by step.

C. SOME SUPPLEMENTS FOR BAUM-WELCH ALGORITHM

Baum-Welch algorithm is usually used to solve the problem of parameter estimation. The algorithm usually constructed an auxiliary function Q, which is completed by EM algorithm. EM algorithm is a method for finding the maximum likelihood estimation of parameter for incomplete data. It was mainly composed of two steps: the first step was to find the expectation, the second step was to maximize the expectation. The two steps were carried out alternately, so that the estimated model parameters gradually approached the real parameters. Until a certain convergence condition was met, the iteration was stopped. EM algorithm solved the problems of traditional maximum likelihood estimation method in solving practical problems, and had important value for speech processing and other fields.

$\varepsilon\left(i, j, t\right)$ represents the probability that the states at t and $t+1$ are i and j respectively, with the observation sequence O^T is given.

$$\begin{aligned}\varepsilon\left(i, j, t\right) &= \frac{P\left(S_t = i, S_{t+1} = j, O^T \mid \lambda\right)}{P\left(O^T \mid \lambda\right)} \\ &= \frac{\alpha\left(i, t\right) a(i, j) b\left(j, o_{t+1}\right) \beta\left(j, t+1\right)}{\sum_i \sum_j \alpha\left(i, t\right) a(i, j) b\left(j, o_{t+1}\right) \beta\left(j, t+1\right)}. \end{aligned} \tag{28}$$

The maximum likelihood estimation of the parameter is:

$$\widehat{\lambda} = \arg\max_{\lambda} P\left(O^T \mid \lambda\right) = \arg\max_{\lambda} L\left(\lambda\right). \tag{29}$$

The log likelihood function is:

$$L\left(\lambda\right) = \ln P\left(O^T \mid \lambda\right). \tag{30}$$

The problem to find the parameter λ makes the probability $P\left(O^n \mid \widehat{\lambda}\right)$ maximum is to solve the functional extremum values. However, the training set is usually limited and the calculation is too complex to realize in many practical problems. We could find the local optimum only, thus Baum-Welch algorithm appears.

$$\begin{aligned}&P\left(O^n \mid \widehat{\lambda}\right) \\ &= \sum_{i=1}^{M} \alpha\left(i, t\right) \beta\left(i, t\right) \\ &= \sum_{i=1}^{M} \sum_{j=1}^{M} \alpha\left(i, t\right) a(i, j) b\left(j, o_{t+1}\right) \beta\left(j, t+1\right). \end{aligned} \tag{31}$$

E steps: We define an auxiliary function Q instead of likelihood function $P\left(O^T \mid \lambda\right)$, and record the expectation of log likelihood function as auxiliary function $\left(\lambda, \widehat{\lambda}\right)$, which can be expressed as formula (30).

$$\begin{aligned}Q &= E\left[\ln P\left(O^T, S^T\right)\right] \\ &= \int dS^T P\left(S^T \mid O^T\right) \ln P\left(O^T, S^T\right) \\ &= \int dS^T P\left(S^T \mid O^T\right) \\ &\quad \times \ln\left[P(S_1) \prod_{t=2}^{T} P\left(S_t \mid S_{t-1}\right) \prod_{t=1}^{T} P\left(O_t \mid S_t\right)\right] \\ &= \sum_{i_1} P\left(S_1 = i_1 \mid O^T\right) \ln P\left(S_1 = i_1\right) \\ &\quad + \sum_{ij} \sum_{t=2}^{T} P\left(S_t = i, S_{t-1} = j \mid O^T\right) \\ &\quad \times \ln P\left(S_t = i \mid S_{t-1} = j\right) \\ &\quad + \sum_{i} \sum_{t=1}^{T} P\left(S_t = i \mid O^T\right) \ln P\left(O_t \mid S_t = i\right). \end{aligned} \tag{32}$$

M steps: From the auxiliary function Q, we can see that the model parameters are independent in each part, and the model parameters can be updated by maximizing each item on the right side of the equation.

$$\lambda^* = \arg\max_{\lambda^{n+1}} Q\left(\lambda^{n+1}, \lambda^n\right)$$

D. SOME SUPPLEMENTS FOR MARKOV SWITCHING MODEL

In order to estimate the parameters, under the condition of past information set $\mathscr{F}_{t-1}$, the joint distribution density of y_t, s_t and s_{t-1} was expressed as:

$$f\left(y_t, s_t, s_{t-1}\right) = f\left(y_t | s_t, s_{t-1}, \mathscr{F}_{t-1}\right) f\left(s_t, s_{t-1} | \mathscr{F}_{t-1}\right). \tag{33}$$

Then, the edge distribution $f\left(y_t | \mathscr{F}_{t-1}\right)$ could be expressed as:

$$\begin{aligned}&f\left(y_t | \mathscr{F}_{t-1}\right) \\ &= \sum_{s_t=0}^{1} \sum_{s_{t-1}=0}^{1} f\left(y_t, s_t, s_{t-1} | \mathscr{F}_{t-1}\right) \\ &= \sum_{s_t=0}^{1} \sum_{s_{t-1}=0}^{1} f\left(y_t | s_t, s_{t-1}, \mathscr{F}_{t-1}\right) f\left(s_t, s_{t-1} | \mathscr{F}_{t-1}\right). \end{aligned} \tag{34}$$

TABLE 11. Composite indicator of systemic financial risk.

Market name	Variable name	Economic significance	Variable properties
real estate market	Cumulative year-on-year increase in sales of commercial housing	Reflect the prosperity of the real estate market	Negative
	Unit sales price of commercial housing increased year-on-year	Represent the price situation in the real estate market. Risk and price moved in the same direction	Negative
Stock market	Average price-earnings ratio	Represent the valuation of the stock market, the risk increased with the deviation	Two-way
	Year-on-year market value growth	Represent the prosperity	Positive
Foreign exchange market	Foreign exchange accounts	The decline in the growth rate of foreign exchange accounts represents a decline in trade volume or a rapid outflow of hot money	Negative
	Effective RMB exchange rate	Devaluation of local currency is one of the expression of the crisis. The higher the effective currency interest rate, the lower the risk	Negative
	Year-on-year growth in foreign exchange reserves	Represents the ability to resist risks, the stronger the ability to resist risks, the smaller the risk	Negative
	Export value	Represents the trade activity	Negative
	Year-on-year imports	Represents the trade activity	Negative
Bond Market	6-month ChinaBond corporate bond and central coupon spread	The difference interest rate between the corporate bond and the low-risk central bank bill change in the same direction as risk.	Positive
	Spreads on 5-year Treasury bonds and 3-month Treasury bonds	The difference interest rate between long-term bond and short-term bond change in the same direction as risk	Positive
	ChinaBond Composite Index (Total Value)/ Wealth Index	The inclining to buy bond assets and sell equity assets change in the same direction as risk	Positive
Government department	Industrial added value growth rate	reflect comprehensive economic strength	Negative
	CPI year on year	Representative inflation level	Negative
	Cumulative total investment in urban fixed assets	Representative investment status	Negative
Currency market	7-day repurchase fixed profit in the interbank market	Represents the short-term funding supply and demand relationship	Positive
	One-week and one-year SHIBOR interest rate spreads	Represents long-term fund borrowing spreads	Positive
	SHIBOR-LIBOR 1 week interest rate difference	Represents the interest rate difference	Negative
Financial Institutions	M2 YoY growth rate / M1 YoY growth rate	Economic efficiency	Positive
	Medium and long-term loans / total loans	Reflect loan liquidity	Positive
	Deposit ratio of financial institutions	Represents the ability of financial institutions to resist risks	Positive
	Growth of loans / Growth of industrial added value	Increased risk as it increases	Positive
	Short-term loan growth rate / industrial added value growth rate	Increased risk as it increases	Positive

The logarithmic likelihood function of Markov transformation model can be obtained as follows:

$$\ln L = \sum_{t=1}^{T} \left[\sum_{s_t=0}^{1} \sum_{s_{t-1}=0}^{1} f\left(y_t | s_t, s_{t-1}, \mathscr{F}_{t-1}\right) f\left(s_t, s_{t-1} | \mathscr{F}_{t-1}\right) \right]. \tag{35}$$

Based on the logarithmic likelihood function described above, for any $i, j \in \{0, 1\}$ we can have (36), as shown at the top of the next page. Among this, $P\left(S_t = j|\mathscr{F}_t\right) = P\left(S_t = j, S_{t-1} = i|\mathscr{F}_t\right) P\left(S_t = j|\mathscr{F}_t\right)$ is called filtering probability. According to Baum forward equation, it can be regarded as a composite function which depends on the observation sequence x_t of the previous period and the filtering probability $P\left(S_{t-1} = i|\mathscr{F}_t\right)$. The filter probabilities and log likelihood values at each time were obtained by iterating the T-Time value of $P\left(S_t = j|\mathscr{F}_t\right)$ in $t = 1, 2, \ldots, T$ into the log likelihood function. According to the filtering probability, $P\left(S_t|\mathscr{F}_t\right)$ could be obtained. If this value

$$P(s_t = j, s_{t-1} = i|\mathscr{F}_{t-1}) = P(S_t = j, S_{t-1} = i) f(S_{t-1} = i|\mathscr{F}_{t-1}) = \frac{f(Y_t|S_t = j, S_{t-1} = i, \mathscr{F}_{t-1}) f(S_t, S_{t-1}|\mathscr{F}_{t-1})}{\sum_{S_t=0}^{1}\sum_{S_{t-1}=0}^{1} f(Y_t|S_t = j, S_{t-1} = i, \mathscr{F}_{t-1}) f(S_t, S_{t-1}|\mathscr{F}_{t-1})}. \quad (36)$$

was substituted into the above two formulas, the smoothing probability of each period could be obtained.

Finally, we could calculate the average duration of a regime based on the transition probability. Suppose D is the duration of regime j, then:

$$\begin{aligned}
&(i)\ D = 1, \text{if} \quad S_t = j \quad \text{and} \quad S_{t+1} \neq j; \text{then},\\
&\qquad P(D = 1) = 1 - a(j, j);\\
&(ii)\ D = 2, \text{if} \quad S_t = S_{t+1} = j \quad \text{and} \quad S_{t+2} \neq j; \text{then},\\
&\qquad P(D = 2) = a(j, j)(1 - a(j, j));\\
&(iii)\ D = 3, \text{if} \quad S_t = S_{t+1} = S_{t+2} = j \text{ and } S_{t+3} \neq j; \text{then},\\
&\qquad P(D = 3) = a(j, j)^2(1 - a(j, j));\\
&\qquad \cdots
\end{aligned} \quad (37)$$

So, the average duration of regime j is

$$E(D) = \sum_{j=1}^{\infty} jP(D = j) = \frac{1}{1 - a(j, j)}. \quad (38)$$

E. SCHEDULE OF TABLE 5

See Table 11.

REFERENCES

[1] M. Billio, M. Getmansky, A. W. Lo, and L. Pelizzon, "Econometric measures of connectedness and systemic risk in the finance and insurance sectors," *J. Financial Econ.*, vol. 104, no. 3, pp. 535–559, Jan. 2012.
[2] S. L. Schwarcz, "Systemic risk," *Georgetown Law J.*, vol. 97, no. 1, pp. 193–249, Mar. 2008.
[3] *Macroprudential Policy Tools and Frameworks: Update to G20 Finance Ministers and Central Bank Governors*, IMF, Washington, DC, USA, FSB, Moscow, Russia, and BIS, Basel, Switzerland, Feb. 2011, p. 2.
[4] C.-L. Huang, O. K. Abass, and C.-P. Yu, "Triclosan: A review on systematic risk assessment and control from the perspective of substance flow analysis," *Sci. Total Environ.*, vols. 566–567, pp. 771–785, Oct. 2016.
[5] C. Gerard, Jr., and D. Klingebiel, *Bank Insolvencies Cross-Country Experience*. Rochester, NY, USA: Social Science Electronic Publishing, 1999.
[6] R. Kalra, "Financial stress: What is it, how can it be measured, and why does it matter?" *CFA Dig.*, vol. 40, no. 1, pp. 5–50, Feb. 2010.
[7] J. F. O. Bilson, "Recent developments in monetary models of exchange rate determination," *Staff Papers*, vol. 26, pp. 201–223, Jun. 1979.
[8] D. Hans and G. Nguyen, *Interbank Exposures: An Empirical Examination of Contagion Risk in the Belgian Banking System*, vol. 3. Rochester, NY, USA: Social Science Electronic Publishing, 2007, pp. 123–171.
[9] M. Drehmann and N. Tarashev, "Measuring the systemic importance of interconnected banks," *J. Financial Intermediation*, vol. 22, no. 4, pp. 586–607, Oct. 2013.
[10] X. L. Gong and J. Bian, "Quantifying sector-level balance sheet contagion—China's macro-financial risk analysis," *Econ. Res. J.*, vol. 45, no. 7, pp. 79–99, 2010.
[11] X. Y. Fan, D. P. Wang, and L. B. Liu, "Scale, relevance and measurement of systemically important banks in China," *J. Financial Res.*, vol. 389, no. 11, pp. 16–30, 2012.
[12] W. J. Baumol, "An expected gain-confidence limit criterion for portfolio selection," *Manage. Sci.*, vol. 10, no. 1, pp. 174–182, Oct. 1963.
[13] P. Artzner, "Application of coherent risk measures to capital requirements in insurance," *North Amer. Actuarial J.*, vol. 3, no. 2, pp. 11–25, 1999.
[14] T. Adrian and K. Brunnermeier, "CoVaR: A method for macroprudential regulation," Federal Reserve Bank, New York, NY, USA, Staff Rep. 348, 2009.
[15] A. Lehar, "Measuring systemic risk: A risk management approach," *J. Banking Finance*, vol. 29, no. 10, pp. 2577–2603, Oct. 2005.
[16] M. S. Basurto and C. Goodhart, "Banking stability measures," *IMF Work. Papers*, vol. 23, no. 9, pp. 202–209, 2009.
[17] M. Illing and Y. Liu, "Measuring financial stress in a developed country: An application to Canada," *J. Financial Stability*, vol. 2, no. 3, pp. 243–265, Oct. 2006.
[18] C. L. Wang and L. Hu, "An empirical research on early–warming of financial risk in China," *J. Financial Res.*, vol. 411, no. 9, pp. 99–114, 2014.
[19] L. Tao and Y. Zhu, "On China's financial systemic risks," *J. Financial Res.*, vol. 432, no. 6, pp. 18–36, 2016.
[20] D. L. Xu and S. L. Chen, "Research on the measurement of systemic financial risk based on financial stress index," *Econ. Perspect.*, vol. 4, pp. 69–78, Apr. 2015.
[21] P. Wu and H. F. Hu, "The construction of financial risk index FRI in China and the test of economic forecast," *Statist. Decis.*, vol. 446, no. 2, pp. 120–123, 2016.
[22] X. X. Yang and Y. F. Wang, "Construction of systemic risk measurement index and analysis of early warning ability in China—Based on dynamic factor model of mixing data," *South China Finance*, vol. 514, no. 6, pp. 3–15, 2019.
[23] T. X. Liu and X. F. Xu, "Can Internet search behavior help to forecast the macro economy?" *Econ. Res. J.*, vol. 50, no. 12, pp. 68–83, 2015.
[24] Y. K. Wu and Q. P. Chen, "Measurement of China's systemic financial risk based on text mining and Web crawler technology," *Jianghuai Tribune*, vol. 5, pp. 70–75, May 2018.
[25] K. Mishev, A. Gjorgjevikj, I. Vodenska, L. T. Chitkushev, and D. Trajanov, "Evaluation of sentiment analysis in finance: From lexicons to transformers," *IEEE Access*, vol. 8, pp. 131662–131682, 2020, doi: 10.1109/ACCESS.2020.3009626.
[26] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *Ann. Math. Statist.*, vol. 37, no. 6, pp. 1554–1563, Dec. 1966.
[27] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 2, pp. 260–269, Apr. 1967.
[28] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood estimation from incomplete data via the *EM* algorithm (with discussion)," *J. Roy. Stat. Soc. B, Methodol.*, vol. 39, no. 1, pp. 1–38, 1977.
[29] J. D. Hamilton, *Time Series Analysis*. Princeton, NJ, USA: Princeton Univ. Press, 1994, pp. 690–699.
[30] Y. Bar-Shalom and X.-R. Li, *Estimation and Tracking-Principles, Techniques, and Software*. Boston, MA, USA: Artech House, 1993.
[31] J. Tugnait, "Adaptive estimation and identification for discrete systems with Markov jump parameters," *IEEE Trans. Autom. Control*, vol. AC-27, no. 5, pp. 1054–1065, Oct. 1982.
[32] A. Doucet, A. Logothetis, and V. Krishnamurthy, "Stochastic sampling algorithms for state estimation of jump Markov linear systems," *IEEE Trans. Autom. Control*, vol. 45, no. 2, pp. 188–202, Feb. 2000.
[33] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity," *J. Econometrics*, vol. 31, no. 3, pp. 307–327, Apr. 1986.
[34] R. F. Engle, "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation," *Econometrica, J. Econ. Soc.*, vol. 50, no. 4, pp. 987–1007. 1982.
[35] K.-Y. Ho, Y. Shi, and Z. Zhang, "How does news sentiment impact asset volatility? Evidence from long memory and regime-switching approaches," *North Amer. J. Econ. Finance*, vol. 26, pp. 436–456, Dec. 2013.
[36] R. Engle, "Systemic risk 10 years later," *Annu. Rev. Financial Econ.*, vol. 10, no. 1, pp. 125–152, Nov. 2018.

[37] Y. H. Wei and S. Y. Zhang, "Multivariate copula-GARCH model and its applications in financial risk analysis," *J. Appl. Statist. Manage.*, vol. 26, no. 3, pp. 432–439, 2007.
[38] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Statist.*, vol. 41, no. 1, pp. 164–171, Feb. 1970.
[39] L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," *Inequalities*, vol. 3, no. 1, 1972, pp. 1–8.
[40] J. Garcia-Frias and J. D. Villasenor, "Turbo decoding of hidden Markov sources with unknown parameters," in *Proc. Data Compress. Conf. (DCC)*, Salt Lake City, UT, USA, Mar. 1998, pp. 159–168.

ZHANFENG LI was born in Jinzhou, Liaoning, China, in 1963. He received the bachelor's degree in engineering from the Beijing Institute of Technology, in 1985, and the master's and Ph.D. degrees from the Zhongnan University of Economics and Law. He is currently a Professor and the Associated Dean of the School of Statistics and Mathematics, Zhongnan University of Economics and Law. He is also the Director of the China Association of Mathematical Economics. His research interests include financial econometrics, econometric analysis and economic forecasting, and decision-making. He is a member of the Hubei Statistical Association.

YANLI CAI was born in Jingmen, Hubei, China, in 1991. She received the bachelor's and master's degrees in statistics from the Zhongnan University of Economics and Law, where she is currently pursuing the Ph.D. degree. Her research interest includes financial statistics.

SHULAN HU was born in Jingde Zhen, Jiangxi, in 1981. She received the B.S. degree in mathematics and the M.S. degree and the Ph.D. degree in probability and mathematical statistics from Wuhan University, Wuhan, China, in 2003 and 2008, respectively. She finished the postdoctoral position from INRIA, France, in 2011. She is currently a Professor with the Zhongnan University of Economics and Law. Her research interests include probabilistic theory of hidden Markov models, stochastic algorithm, econometrics, and relative applications in economy and finance.