Filomat 30:15 (2016), 3999–4013
DOI 10.2298/FIL1615999L

Published by Faculty of Sciences and Mathematics,
University of Niš, Serbia
Available at: http://www.pmf.ni.ac.rs/filomat

Correlation Analysis of Industry Sectors in China's Stock Markets Based on Interval Data

Wen Long[a,b,c], Yeran Tang[a,b,c], Dingmu Cao[a,b]

[a] *Research Center on Fictitious Economy & Data Science, Chinese Academy of Sciences, P.R.China, 100190*
[b] *School of Economics and Management, University of Chinese Academy of Sciences, P.R.China, 100190*
[c] *Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, P.R.China, 100190*

Abstract. Comparing with single value data, interval data is an important data type containing more information. This paper focuses on correlation analysis of interval data and proposes a comprehensive weighted correlation coefficient which combines both trend and range information of a sequence. Applying the new approach of interval data correlation coefficient to China's stock market, we obtain the empirical value of the weight taking CSI 300 as the representative. Further, the correlations between industry sectors as well as the sample stocks within a specific industry sector are analyzed respectively using this method. The empirical studies suggest that the correlation coefficient of interval data has improved the results of traditional correlation coefficient and the temporal fluctuation characteristics of the sequence are better reflected based on interval data.

1. Introduction

The global stock markets develop rapidly in recent years. Take China for example, statistics show that the total number of listed companies in China mainland surged from 10 to more than 2700 in the last two decades. As an important part of the national economy, stock markets have become a main channel to allocate resources. With development of stock market, researchers pay much attention to the analysis of correlations among different industry sectors in stock market. It is mainly because that price fluctuations of one industry sector will probably have impacts on others, which increases the risks of investment. For investors, knowing the correlations of stocks can help them manage risks. According to Markowitz portfolio theory, the risk of a portfolio is related to the correlation coefficients of stocks in portfolio. It shows that the study of stock market correlation is of great significance.

With rapid development of China's stock market, there is a growing quantity of enterprises and transaction data. Meanwhile, the internal structure of the data is increasingly complicated, which is caused by mergers and acquisitions, restructuring, name change, termination of trading, trading halt, etc. Besides, the application of high frequency data also shows diversification of the types and the structure of the data in stock market. When facing large and complex stock market data, the traditional data analysis method can

2010 *Mathematics Subject Classification*. Primary 62P20; Secondary 90B50
Keywords. stock market, industry sector, correlation, interval data.
Received: 25 May 2015; Accepted: 18 July 2016
Communicated by Dr. Alex Maritz and Dr. Charles Xie
Research supported by National Natural Science Foundation of China (No.71101146), University of Chinese Academy of Sciences and the Open Project of Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences.
Email address: longwen@ucas.ac.cn (Wen Long)

hardly meet the needs of information processing and knowledge acquisition. Considering symbolic data analysis (SDA) is a good way to deal with high-dimensional data and can be used to reflect the relationship between stock market industry sectors preferably, this paper applies SDA method to analyze the correlations between China stock market industry sectors, thus provide not only a method of study on industry correlation, but also a new way for interval data correlation analysis. We believe our efforts on this will help to expand the application fields of interval data analysis method.

Based on Pearson correlation coefficient, this paper puts forward a new method of stock market correlation analysis, which is to construct a weighted correlation coefficient of temporal interval data combining its trend with its volatility. Specifically, we first show the difference of correlation coefficients between single value and interval data in extreme cases. Then, combining interval data pretreatment techniques, we propose a new weighted method to convert interval data sequence into a single value data sequence. After that, in order to determine the weight, we analyze the gradient ratio of the correlation coefficient, and increase the weight as much as possible until the correlation coefficient tends to be stable. At last, we use proposed weight to calculate the comprehensive weighted correlation coefficient, which is the correlation coefficient between stock markets industry sectors based on interval data.

We apply our interval data correlation coefficient method into stock market in China using CSI 300 as representative samples, and get experience value of the weight to investigate the relationships between industry sectors in China's stock market. The empirical studies include two aspects. One is to compare our results with the correlation coefficient of single value data, and the other is to discuss the correlation of different sample stocks within the same sectors by using the proposed method.

2. Literature Review

Due to the scale and complexity of the stock market data, varieties of new data analysis methods are put forward and get a wide range of applications. Symbolic data analysis (SDA) is one of the methods to discover knowledge from large amounts of data.

2.1. Characteristics of Interval Data

The basic concept of SDA analysis is proposed by Diday at international classification association conference in 1988. After that, it is widely used and continuously developed in both of theory and practice. Its analysis objective includes the numerical data, categorical data, interval data, multi-valued data and distributed data. The basic idea is to formulate symbolic object samples and reduce the space dimension according to the pretreatment of the original sample (Bock, Diday, 2000).

Assume x_1 and x_2 are the data in sample space. If $x_1 \leq x_2$, $[x_1, x_2]$ is a interval data. We can define x_1 and x_2 as upper and lower limit, and define $(x_1 + x_2)/2$ and $(x_2 - x_1)/2$ as midpoint and radius. The method to construct an interval data can be maximum value and minimum value of samples under different condition, or can be a number of sample's maximum value and minimum value under the same condition. There are two main channels to get interval data sample. One is caused by the observation error or the uncertainty of expert opinions. These uncertain results can be expressed by interval data to increase the reliability. The other is to form symbolic sample and reduce dimension of the sample space to discover knowledge from huge amounts of data. Interval data is one of the symbolic data. Stock price can be converted to $X = [x_1, x_2]$ by minimum price and maximum price in a day.

Interval data can reflect both of the central tendency and the discrete tendency, so it is an important type of symbolic data. When analyzing interval data descriptively, assume samples uniform distribution independently, then their statistics such as mean, variance and standard deviation can be calculated according to the joint distribution function (Bock, Diday, 2000; Billard, Diday, 2006). Other methods to analyze interval data are also partly based on assumption of uniform distribution, such as midpoint method of principal component analysis (Cazes, 1997) and regression analysis method combining midpoint with radius (Lima-Neto, Carvalho, 2008). Compared with single value data, interval data using midpoint method can avoid influence of outliers, and combining radius, interval data can even reflect volatility. Therefore, using interval data to analyze correlations of stock markets can receive more information and more comprehensive reflection about original data. Nevertheless, traditional multivariate statistical methods cannot

be applied directly to interval data. Some preprocessing should be taken to convert interval data into single value data first.

2.2. Related Research

The researches of SDA method include descriptive analysis, visual processing, regression analysis, principal component analysis, discriminant analysis, cluster analysis, similarity research and other fields. Our paper mainly applies the first four kinds of analysis methods, which can be summarized as follows:

- Descriptive Analysis
  Bertrand, Goupil (2000), Billard, Diday (2003, 2006), and Gioia, Lauro (2006) have expounded the literature review about descriptive analysis on symbolic data in detail. A descriptive analysis of the interval data are under the assumption of uniform distribution. Based on this assumption, some statistics such as the mean, variance, correlation coefficient and covariance of the interval data can be calculated according to the traditional empirical distribution function and joint distribution function.

- Visual Processing
  The visual processing method frequently used in the SDA is mainly zoom star chart and the factor loading diagram with sample principal plane graph shown in principal component analysis results. Noirhomme-Fraiture (2002) proposes that zoom star chart can be used to describe 2D and 3D symbolic data. Through their special form, the characteristics of different samples and their fluctuation regularity can be compared intuitively.

- Regression Analysis
  Billard, Diday (2000) put forward linear regression model suitable for interval data using SDA. The model regards upper and lower limits of the interval as different samples, and they share the same coefficient. They assume that the upper and lower limits of the interval data are independent, so they can establish regression model respectively without any constraint of their coefficients. Lima-Neto, Carvalho (2008) hold the view that using midpoint and radius of the interval data to build model is more reasonable than using upper and lower limits. Considering the constraint between upper and lower limits, they add some constraints to ensure the interval data radius are positive, and lower limits are no more than upper limits. Maia (2008) compares the theoretical and empirical researches of autoregressive model (AR), moving average regression model (ARIMA) and artificial neural network (ANN), and finds the mixed model of these three models is effective considering the accuracy and applicability. Giordani, Paolo (2011) apply Least Absolute Shrinkage and Selection Operator (Lasso) method (Tibshirani, 1996) into regression analysis of interval data. Under the premise of retaining original data information, they try to make the coefficients of midpoints and radius regression models equal.

- Principal Component Analysis
  Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables. The number of principal components is less than or equal to the number of original variables, so this method can reduce dimension of original data. Cazes (1997) and Chouakria (1998) propose the vertex principal component analysis (VPCA) and the central principal component analysis (CPCA), which apply principal component analysis into interval data. Specifically, VPCA regards each of the interval data sample as a hyperrectangle, and converts them into a high dimensional real matrix for traditional PCA analysis. However, the computation of VPCA will exponentially increase with the increase number of variables. CPCA substitute the midpoint for each interval data to simplified calculation, but it will lose original information about volatility. Lauro and Palumbo (2000) regard midpoint and radius as different variables for their further analysis. Douzal (2011) points

out that VPCA only contains part of information about interval data volatility. Le-Rademacher and Billard (2012) use covariance matrix of symbolic data to extract the principal component, which can contain all of the volatility information and solve the dimension problem caused by VPCA method.

To sum up, SDA method can be used to analyze some complex systems like stock markets, and it still can be further developed on many aspects, including correlation analysis. The existing correlation analysis methods using SDA can be divided into two ways. One is based on descriptive analysis. That is, under the assumption of uniform distribution, after constructing traditional empirical distribution function and joint distribution function, the correlation of interval data can be calculated according to the traditional single value data method. The other is based on regression analysis. The coefficient of linear regression model can be regarded as the correlation of two variables. However, both of these methods have some problems. The correlation coefficients based on descriptive analysis are under the assumption of uniform distribution, which cannot be met in most cases, and the information of fluctuation range is not considered. Meanwhile, the correlation coefficients based on regression analysis need to build a number of regression models when analyzing multivariate correlation, which causes complicated calculation. Besides, the existing regression analysis for interval data almost builds models to the lower and upper limits respectively, which is hard to integrate.

Based on these existing methods and Pearson correlation coefficient, our paper proposes a new method to analyze the correlation of interval data in stock markets. We construct a weighted correlation coefficient of interval data combining its trend with its volatility, and calculate the empirical value of weight in China stock markets. Then we analyze the correlation characteristics of China stock markets industry sectors using our method to prove the effectiveness of the proposed method and expand the application fields of interval data analysis method.

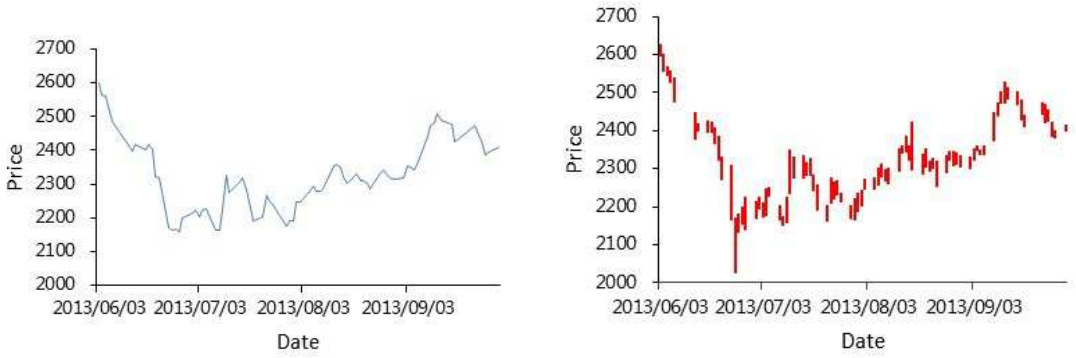
3. Interval Data Correlation Coefficient

3.1. Compared with Single Value Data in Stock Markets

When measuring the correlations in stock markets, Pearson correlation coefficient is most widely used. For two closing price sequence of the industry sectors $\{x_i\}$ and $\{y_i\}$, the formulation of Pearson correlation coefficient is:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

Where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.



(a) Closing Price (b) Max & Min Price

Figure 1: Stock Price of CSI 300 (2013.6-2013.9)

Evidently, Pearson correlation coefficient is for single value data, and cannot reflect volatility of stock price. When calculating correlation coefficient in stock markets, the existing measure methods mainly

adopt closing price, minimum and maximum price a day, but these measurements overlook volatility of stocks. Figure 1 shows the difference between single value data and interval data when describing stock price. The interval data constructed by maximum and minimum price (Figure 1(b)) can not only reflect the trend of closing price (Figure 1(a)), it can also reflect price volatility, which contains more information about stock markets.

In extreme cases that the closing prices are equal, the value of two sequences will be regarded as the same when using single value data. However, the range of these sequences can be largely different. For example, in 2012, the closing prices of energy index are 2611.8 both on Oct 11^{st} and Oct 17^{th}, but their maximum prices are 2632.3 and 2641.0, and their minimum prices are 2580.8 and 2610.5. When analyzing their correlations, single value data will lose volatility information about stock price. The situation like this sample often happens, and interval data can reflect information of original samples more precisely.

We perform some further study to analyze the superiority of interval data compared with single value data. Through two extreme cases, we calculate the Pearson correlation coefficients between CSI 300 stocks and the index based on closing price sequence and range sequence respectively. The time interval of the sample is from Jan 1^{st}, 2013 to Sep 30^{th}, 2013. Figure 2 shows the difference of the correlation coefficients between closing price and range price. The samples between two dotted lines on the right suggest that the difference of correlation coefficients between closing price and range price is about 0-0.5, and the coefficient based on closing price is higher. The samples above the left dotted line suggest that some differences are more than -0.5, which means the coefficients based on range price have deficiency when stock price changes negatively.

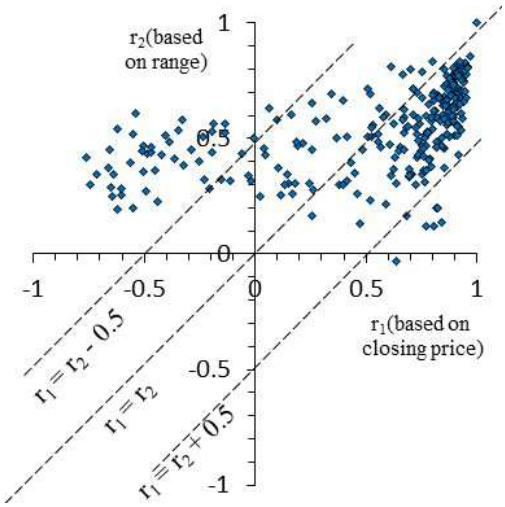
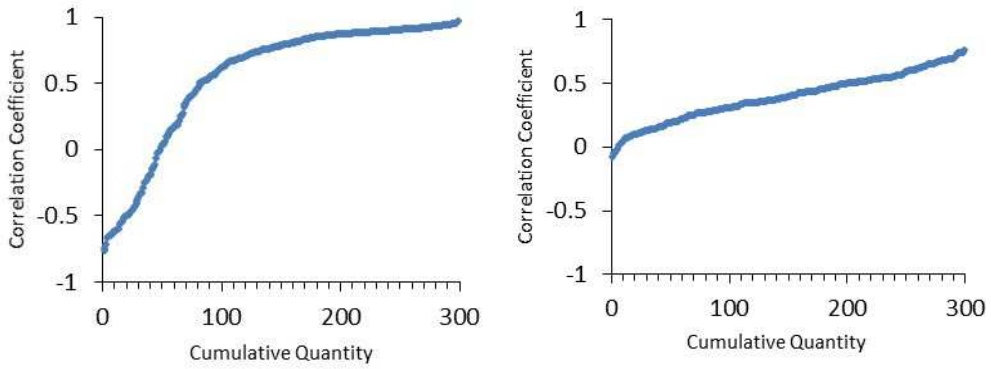


Figure 2: Correlation Coefficient of CSI 300 and Constituent Stocks

After ascending sorting the correlation coefficient demonstrated in figure 2, it can be seen in figure 3(a) that correlation coefficients based on closing price sequence are distributed from -0.8 to 1, and are relatively large. Most of them are more than 0.6, and almost 50% of the coefficients are more than 0.8. The correlation coefficients based on range price sequence are distributed from -0.2 to 1, and the distribution is relative equilibrium. Only 5 coefficients are less than 0, which can be explained by substitutional relationships of stock and index. The price of some stocks get higher, the others may become lower, but the volatilities are similar. Therefore, the coefficients based on range data which can express the range of volatility are positive relations.



(a) Based on Closing Price (b) Based on Range

Figure 3: Correlation Coefficient of CSI 300 and Constituent Stocks

It can be seen that the correlation coefficient based on the closing price can better express the stock price of negative change. However, nearly half of the coefficients between stocks and CSI 300 are above 0.8, which suggests it can hardly distinguish the high relevance stocks. The correlation coefficients based on the range price can hardly express the change direction of the stock price, but have more differentiation in a certain extent. Therefore, it is necessary to combine absolute value with range to analyze correlations. In this paper, we use interval data to replace the traditional single value data, combine absolute value with range, and construct comprehensive weighted correlation coefficient to analyze correlation characteristics in stock markets.

3.2. Interval Data Weighting Method

For an interval data sequence $[x_1, x_2]$, define x_R as its volatility term, which express the changing range, represented as $x_R = x_2 - x_1$. Define x_k as the trend term, which express the absolute value of the data, such as upper limit, lower limit, midpoint, and center of gravity, $x_k \in [x_1, x_2]$. The formulation of comprehensive weighted correlation coefficient method to convert interval data to single value data is:

$$x_i = \alpha \cdot x_k + (1 - \alpha) \cdot x_R \tag{2}$$

Where, α called comprehensive weighted correlation coefficient, which measures the relative influence of two sequences, and $\alpha \in [0, 1]$.

For interval data sequences $\{x_i\}$ and $\{y_i\}$, after using weighting method in formulation (2) to convert interval data to single value data, the correlation coefficient of interval data can be calculated based on Pearson correlation coefficient:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{3}$$

Where, $x_i = \alpha \cdot x_k + (1 - \alpha) \cdot x_R$, $y_i = \alpha \cdot y_k + (1 - \alpha) \cdot y_R$, $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$.

Notice that x_R precisely defined as gap between upper limit and lower limit of the interval, and x_K is reflect the absolute value of the data sequence which have different ways to value. In most cases, the center of gravity is considered, but this method has exceptions in some special cases. In stock markets, the closing price of the stocks is considered as representative data. When study on stocks or stock index X, define X_{ct} as closing price sequence, and let $x_K = X_{ct}$, which is a typical single value data. Define X_{ut} as maximum price and X_{lt} as minimum price in a certain time horizon, then $[X_{lt}, X_{ut}]$ is the interval data sequence of stock X, and $X_{ct} \in [X_{lt}, X_{ut}]$. Define x_R as the range of stock price, and let $X_R = X_{ut} - X_{lt}$. Above these, in interval data sequence of stock markets, the formulation of comprehensive weighting method to convert interval data to single value data is:

$$X_i = \alpha \cdot X_{ct} + (1 - \alpha) \cdot (X_{ut} - X_{lt}) \tag{4}$$

3.3. Value of Weights

To evaluate the value of weights α is to analyze whether there is α in formulation (2) to combine absolute value with range in an interval data. Considering extreme cases, when $\alpha = 1$, sequence is a traditional single value data, which expresses the absolute value. When $\alpha = 0$, sequence represents range of volatility. In theory, different applications have different empirical value of the weights, and $\alpha \in [0, 1]$, which makes the sequence contains the information of extreme cases and can be relatively stable.

According to our analysis, the correlation coefficients based on single value data (when $\alpha = 1$) is stable, but have low differentiation. The correlation coefficients based on range (when $\alpha = 0$) have relatively high differentiation. The main purpose of combining with range index is to make differentiation of traditional single value data more significant. Thus, the principle of weights evaluation is that under the condition of correlation coefficients stabilization, the value of α should be as small as possible.

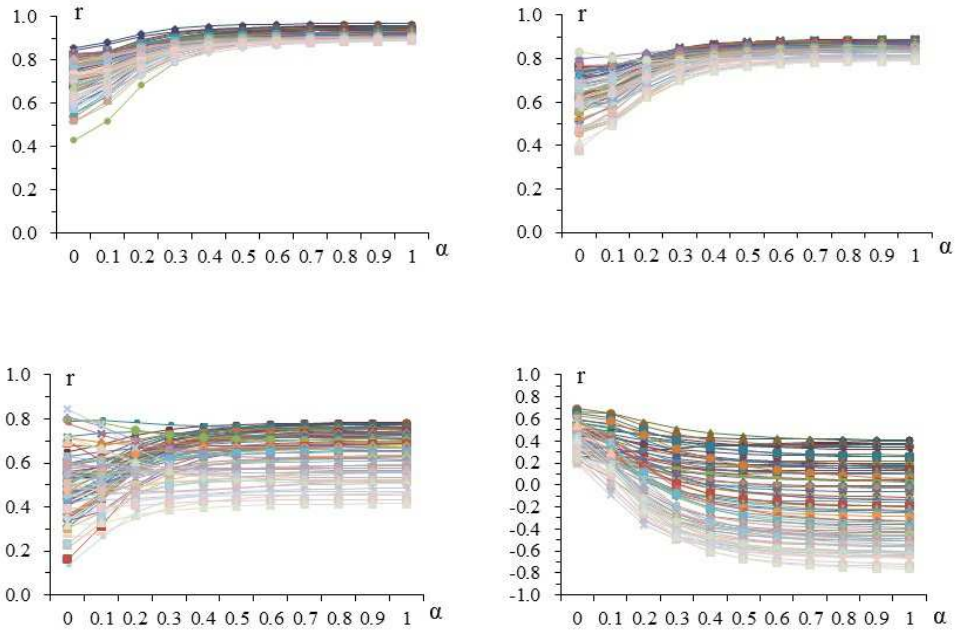


Figure 4: Correlation Coefficients between Constituent Stocks and CSI 300

Note: Due to space limitation, the 300 constituent stocks are divided into 4 charts and only 75 sample stocks are shown in each chart. The abscissa axis represents value of α, and ordinate represents correlation coefficients between sample stocks and CSI 300 calculated by comprehensive weighting method.

We try to get empirical value of α with CSI 300 which is a representative sequence of stock market in China. Starting α from 0, until the correlation coefficient in formulation (3) is stable, the corresponding α is the empirical weights of China stock markets. Define gradient ratio as the conception that the change rate of the correlation coefficient caused by a unit adding of α. Assume step size $i = 0.1$. Starting α from 0, and note $\alpha = 0$ as α^0, $\alpha = 0.1$ as $\alpha^{0.1}$, ..., $\alpha = 1$ as α^1. Define comprehensive weighting correlation coefficient of sequence $\{X\}$ and $\{Y\}$ as $r_{xy,\alpha}$, and define its gradient ratio as $n_{xy,\alpha}$. When $\alpha = 0.1$, $n_{xy,\alpha^{0.1}} = \frac{r_{xy,\alpha^{0.1}} - r_{xy,\alpha^0}}{r_{xy,\alpha^0}}$, ..., when $\alpha = 1$, $n_{xy,\alpha^1} = \frac{r_{xy,\alpha^1} - r_{xy,\alpha^{0.9}}}{r_{xy,\alpha^{0.9}}}$. Assume the correlation coefficient is stable when $n_{xy,\alpha} < \mu$, and k is the number of stable coefficient. After the trend of k mutation and gradually flat, the corresponding α is the weights of comprehensive weighting method. We choose 300 sample stocks to calculate their correlation

coefficients with CSI 300 from $\alpha = 0$ to $\alpha = 1$ and the step size is 0.1. The line chart is shown in Figure 4. The time interval of the sample is from Jan 1^{st}, 2013 to sep 30^{th}, 2013.

Table 1: Gradient ratio of Correlation Coefficients between Constituent Stocks and CSI 300

α	Number of lines within 1% gradient ratio	Number of lines within 2% gradient ratio	Number of lines within 3% gradient ratio	Number of lines within 4% gradient ratio	Number of lines within 5% gradient ratio
0.1	3	8	11	14	19
0.2	1	2	7	9	10
0.3	4	8	20	42	66
0.4	21	96	**154**	**190**	**212**
0.5	**121**	**192**	227	244	254
0.6	178	220	246	260	265
0.7	196	236	260	270	277
0.8	214	250	268	277	283
0.9	227	261	275	282	284
1.0	236	269	279	283	287

Table 1 shows the distribution of the gradient ratio under different standard ($\mu = 1\%, 2\%, ..., 5\%$). For example, when $\alpha = 0.1$, there are 3 lines whose gradient ratio are under 1%, and 8 lines whose gradient ratio are under 2%. When the value of μ is fixed, larger gradient ratio leads to larger number of stable lines. Table 1 suggests that, the increase of α causes more stable lines, which meets our expectation that the correlation coefficient becomes stable with the increase of α. Besides, the increase of α represents high proportion of closing price and low proportion of range in sequence when calculating correlation coefficients. When choosing appropriate value of α, the one that lets the increase of number of lines less would be chosen. It can be seen that, the correlation coefficients become stable when $\alpha = 0.4$ (3-5% standard of stabilization), or $\alpha = 0.5$ (1-2% standard of stabilization). Therefore, $\alpha = 0.4$ (or 0.5) is the empirical value of comprehensive weighting method in stock markets.

In order to further verify the applicability of $\alpha = 0.4$, we test the correlation coefficients between 10 industry index and CSI 300 from 2007 to 2013. With the increase of α, coefficients also become stable. Table 2 shows the results of gradient ratio analysis. It suggests that $\alpha = 0.4$ (3-5% standard of stabilization), or $\alpha = 0.5$ (1-2% standard of stabilization) is the empirical value of comprehensive weighting method with interval data of industry index and CSI 300.

Table 2: Gradient ratio of Correlation Coefficients between Industry Index and CSI 300

α	Number of lines within 1% gradient ratio	Number of lines within 2% gradient ratio	Number of lines within 3% gradient ratio	Number of lines within 4% gradient ratio	Number of lines within 5% gradient ratio
0.1	3	7	11	15	23
0.2	16	27	38	46	52
0.3	40	57	62	68	68
0.4	**62**	**74**	**74**	**74**	**74**
0.5	72	74	76	75	74
0.6	74	75	76	75	74
0.7	73	74	76	75	75
0.8	74	76	76	76	75
0.9	74	77	77	76	75
1.0	74	77	77	76	75

Based on the results of two experiments above, we choose $\alpha = 0.4$ as the empirical value of comprehensive weighting method in stock markets. Therefore, in our paper, the formulation to invert single value data into interval data in stock markets is:

$$X_i = 0.4 \cdot X_{ct} + 0.6 \cdot (X_{ut} - X_{lt}) \tag{5}$$

Where X_{ct} is the sequence of closing price, X_{ut} is the maximum price and X_{lt} is the minimum price of the stock sequence.

4. Empirical Test in China Stock Markets

We have proposed a method to calculate correlation coefficient of interval data, and applied it into stock market to get formulation to transfer interval data into single value data. In this part, we analyze correlation in China's stock market with this method. Firstly, we compare the results of interval data and single value data with the development cycle characteristics of stock market in China, and summarize the advantage of using our interval data method. Then, we analyze the stocks associated characteristics in bank sectors of China's stock market in the year of 2013.

4.1. The Advantage of Interval Data Correlation Coefficient

Considering the division of development stage in China's stock market, we study on CSI 300 stock market in China from Jul 2^{nd}, 2007 to Dec 21^{st}, 2012. We divide 300 sample stocks into 10 industries according to industry classification standard, and get the index from all the stocks in each industry. The industries are energy (EN), raw materials (RM), industrial (IN), optional consumption (OC), main consumption (MC), medicine and health care (MH), finance and real estate (FR), information technology (IT), telecommunications (TE), utility industry (UT). It can be seen in Figure 5 that, all of the index are peak in the end of 2007, and at bottom in 2008. After that, the stock market is moving sideways, and inconsistent among different industries. Therefore, we divide stock market into two phases. One is from Jul 2^{nd}, 2007 to the end of 2008, which is a remarkable period. The other is from 2009 to 2012, which is a storming phase.

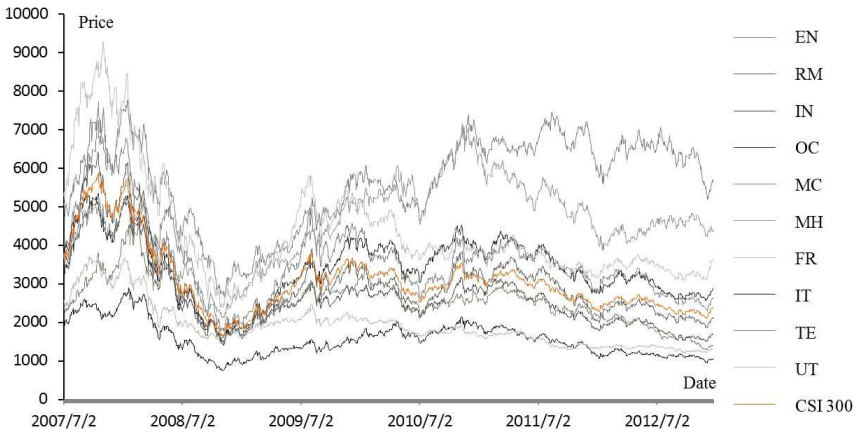


Figure 5: Trends of CSI 300 and Industry Index

(1) The Similarities

The correlation coefficient based on single value data is shown in table 3. It can be seen that, in the first phase (from 2007 to 2008), most of the industries have strong correlation with CSI 300. The value of coefficient is from 0.78 to 0.96. Besides, the correlation between different industries is strong, especially in information technology, utilities, raw materials and industrial. In the second phase (after 2009), the correlation becomes weak, especially the main consumer, medicine and health care and utilities. The analysis

of correlation in different phases suggest that, in remarkable period, the volatilities of each industries are coincident, while in storming phase, each industries will tend to show their own characteristics.

Table 3: Correlation Coefficient Matrix of Index Returns

	CSI300	EN	RM	IN	OC	MC	MH	FR	IT	TE	UT
First Phase:	**2007.7-2008**										
CSI300	1										
EN	0.86	1									
RM	0.94	0.82	1								
IN	0.96	0.81	0.94	1							
OC	0.94	0.77	0.92	0.95	1						
MC	0.86	0.71	0.84	0.87	0.89	1					
MH	0.84	0.69	0.82	0.86	0.88	0.82	1				
FR	0.93	0.75	0.79	0.82	0.79	0.71	0.69	1			
IT	0.89	0.71	0.85	0.9	0.92	0.83	0.84	0.75	1		
TE	0.78	0.71	0.72	0.77	0.75	0.71	0.67	0.68	0.73	1	
UT	0.86	0.72	0.86	0.88	0.86	0.76	0.78	0.7	0.82	0.72	1
Second Phase:	**2009-2012**										
CSI300	1										
EN	0.89	1									
RM	0.92	0.86	1								
IN	0.95	0.83	0.90	1							
OC	0.90	0.77	0.83	0.91	1						
MC	0.73	0.58	0.65	0.73	0.76	1					
MH	0.73	0.58	0.67	0.75	0.76	0.76	1				
FR	0.92	0.78	0.76	0.80	0.75	0.55	0.54	1			
IT	0.79	0.66	0.77	0.84	0.83	0.71	0.77	0.60	1		
TE	0.73	0.64	0.67	0.73	0.70	0.56	0.59	0.63	0.67	1	
UT	0.85	0.74	0.79	0.85	0.80	0.63	0.66	0.74	0.73	0.67	1

To compare single value data with interval data, we calculate the correlation coefficients based on our method proposed above. Table 4 shows the results. It suggest that, compared with two phases, the characteristics of correlations are similar using two different methods. In first phase, the correlations are strong, and each coefficient is more than 0.7, while in second phase, the correlations are decreased.

Table 4: Correlation Coefficient Matrix of Index Returns Based on Interval Data

	CSI300	EN	RM	IN	OC	MC	MH	FR	IT	TE	UT
First Phase:	**2007.7-2008**										
CSI300	1										
EN	0.97	1									
RM	0.99	0.96	1								
IN	1.00	0.96	0.99	1							
OC	0.98	0.93	0.98	0.99	1						
MC	0.95	0.92	0.96	0.96	0.94	1					
MH	0.98	0.94	0.98	0.98	0.98	0.96	1				
FR	0.99	0.96	0.96	0.97	0.95	0.90	0.94	1			
IT	0.96	0.89	0.96	0.97	0.98	0.96	0.97	0.91	1		
TE	0.81	0.81	0.84	0.83	0.80	0.92	0.86	0.74	0.85	1	
UT	0.98	0.94	0.98	0.99	0.98	0.92	0.96	0.96	0.95	0.77	1
Second Phase:	**2009-2012**										
CSI300	1										
EN	0.95	1									
RM	0.94	0.93	1								
IN	0.93	0.86	0.93	1							
OC	0.87	0.82	0.90	0.86	1						
MC	**0.08**	**0.19**	**0.26**	**0.04**	**0.42**	1					
MH	0.54	**0.46**	0.63	0.59	0.82	0.63	1				
FR	0.88	0.81	0.70	0.71	0.58	**-0.25**	**0.17**	1			
IT	0.75	0.64	0.79	0.86	0.89	**0.28**	0.86	**0.43**	1		
TE	0.90	0.85	0.83	0.92	0.76	**-0.08**	**0.41**	0.76	0.74	1	
UT	0.73	0.60	0.56	0.71	**0.39**	-0.59	**0.02**	0.85	**0.42**	0.76	1

Note: The underline represents the strong correlation (coefficient above 0.9), and the bold represents the weak correlation (coefficient from -0.5 to 0.5).

(2) The Differences

The results in table 5 are the results in table 4 minus those in table 3. It can be seen that, in first phase, the correlation coefficients based on our method is higher than traditional method. All of the coefficients of the index and CSI 300 are above 0.81. Since single value sequence only reflects the change of absolute value, while interval sequence can reflect both absolute value and its volatility, these results suggest, combing with volatility, the correlation between sectors are higher than those considering absolute value only, when in the remarkable period. In second phase, Correlation coefficients based on interval data are more likely less than those based on single value data. It suggests that, the correlation of trend is higher than the correlation of volatility between two stock price sequences. Take the main consumption industry and telecommunications for example, correlation coefficient of the former based on interval data is less than based on single value data, while the latter, on the other hand, whose correlation coefficient based on interval data is larger than based on single value data. The correlation coefficients of the second phase in Table 4 show that the main consumption industry is mainly correlated with others weakly, which suggests the correlation only based on trend term. When combing with volatility to adjust correlation coefficients, the correlation of the main consumption with others is weakened, or even become negative. It is quite reasonable, because the volatility of telecommunication industry should be more impressionable of other industries, and the trend of the main consumption industry is more impressionable, their industry characteristics are significant in the second phase. The numerical comparison suggests that, interval data

has a greater degree of differentiation than single value in terms of the correlation coefficients, and it can reflect more industry characteristics.

Table 5: Difference of Correlation Coefficient with Two Methods

	CSI300	EN	RM	IN	OC	MC	MH	FR	IT	TE	UT
First Phase:	**2007.7-2008**										
CSI300	0.00										
EN	0.11	0.00									
RM	0.05	0.15	0.00								
IN	0.03	0.15	0.05	0.00							
OC	0.04	0.16	0.06	0.04	0.00						
MC	0.09	0.21	0.11	0.09	0.05	0.00					
MH	0.13	0.25	0.16	0.12	0.09	0.14	0.00				
FR	0.06	0.21	0.17	0.15	0.16	0.19	0.24	0.00			
IT	0.07	0.18	0.11	0.07	0.06	0.13	0.13	0.15	0.00		
TE	0.03	0.10	0.12	0.06	0.05	0.21	0.19	0.06	0.13	0.00	
UT	0.12	0.23	0.12	0.11	0.12	0.15	0.18	0.25	0.12	0.05	0.00
Second Phase:	**2009-2012**										
CSI300	0.00										
EN	0.06	0.00									
RM	0.02	0.07	0.00								
IN	-0.02	0.03	0.04	0.00							
OC	-0.03	0.06	0.07	-0.04	0.00						
MC	-0.65	-0.40	-0.40	-0.69	-0.33	0.00					
MH	-0.19	-0.13	-0.04	-0.17	0.06	-0.13	0.00				
FR	-0.04	0.03	-0.06	-0.09	-0.17	-0.81	-0.38	0.00			
IT	-0.04	-0.02	0.02	0.02	0.06	-0.43	0.10	-0.17	0.00		
TE	0.16	0.21	0.17	0.19	0.06	-0.64	-0.18	0.14	0.06	0.00	
UT	-0.12	-0.14	-0.22	-0.14	-0.41	-1.22	-0.64	0.11	-0.32	0.10	0.00

Note: The underline represents negative number, which suggests the correlation coefficient based on interval data is less than that based on single value data.

4.2. Correlation of Sample Stocks in Bank Sectors

After summarizing the advantage of interval data correlation coefficient, we use our method to analyze correlation characteristics of the bank sector in China's stock market in the year of 2013. Table 6 shows the correlation coefficients of 16 listed banks with CSI 300 in different weights of our calculation formulation.

When weights $\alpha = 1$, the sequence equals to closing price sequence of the stock, and when weights $\alpha = 0$, the sequence equals to range sequence of the stock price. The results of $\alpha = 1$ suggest banks, especially the large state-owned banks such as bank of China, bank of communications, have high correlation with closing price sequence of stock market, and its stock prices are consistent with the whole market. The results of $\alpha = 0$ suggest the large difference among large state-owned banks when it comes to correlation of banks with range sequence of the stock market. Specifically, Bank of China, Industrial and Commercial Bank of China (ICBC), and China Construction Bank have low correlation with range sequence of the stock market, while Huaxia Bank, China Merchants Bank, and Minsheng Bank have high correlation with range sequence of the stock market.

When weights $\alpha = 0.4$, we use comprehensive weighting method to adjust the correlation coefficients. Figure 6 clearly shows that differentiation of correlation is higher than before, which suggests the method

based on interval data can better reflect characteristics of stocks. Meanwhile, the correlation coefficients in Table 6 when $\alpha = 0.4$ can acquire the adjusted coefficients of each bank. It can be seen that, the correlation coefficient of Huaxia Bank increased based on comprehensive weighting method because its closing price sequence is less correlated with the market and its volatility sequence is more correlated with the market. On the other hand, the correlation coefficient of Bank of China and ICBC decreased based on comprehensive weighting method because its closing price sequence is more correlated with the market and its volatility sequence is less correlated with the market.

Table 6: Correlation Coefficients of Bank-CSI 300 in different α

α	0	0.4	1	α	0	0.4	1
Ping An Bank	0.53	0.89	0.89	Bank of Beijing	0.74	0.93	0.95
Bank of Beijing	0.68	0.94	0.95	Agricutural Bank	0.60	0.89	0.91
SPD Bank	0.59	0.86	0.90	Bank of Comunications	0.65	0.94	0.94
Huaxia Bank	0.76	0.83	0.80	ICBC	0.55	0.80	0.87
Minsheng Bank	0.69	0.74	0.75	Everbright Bank	0.69	0.90	0.91
Merchants Bank	0.74	0.89	0.90	Construction Bank	0.55	0.84	0.87
Bank of Nanjing	0.71	0.93	0.94	Bank of China	0.56	0.92	0.95
Industrial Bank	0.65	0.88	0.87	China Citic Bank	0.61	0.89	0.91

As a result, the correlation coefficients based on comprehensive weighting method combine the characteristics of trend with volatility of the stocks. Using this method to analyze stock market has a great significance to both investors and the managers. For value investors, they pay more attention to the long-term performance of the stock, so they need the correlation of stocks trend with others for diversification. For speculators, they are more focused on the short-term volatility of the stock price, so they need more information about the volatility correlation of the stocks with others for assets allocation. For managers, the correlation coefficients based on comprehensive weighting can reflect characteristics of stocks more accurately, so they can better grasp the regulation of the market from different angles.

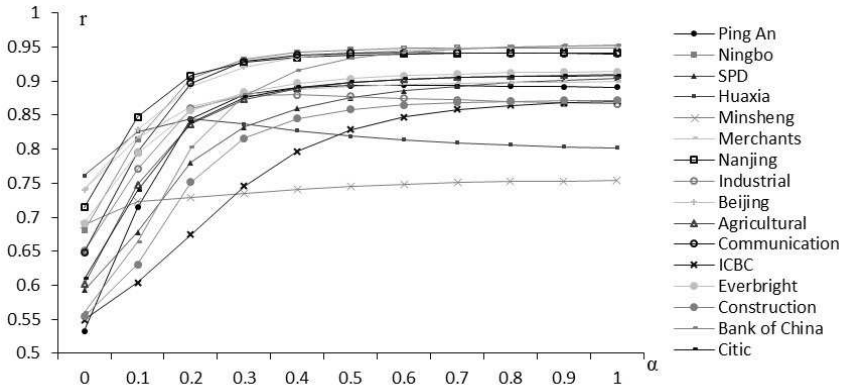


Figure 6: Correlation Coefficients of Bank-CSI 300 in different α

5. Conclusions

The rapid development of stock market provides a great convenience to information disclosure, dissemination and sharing. Meanwhile, the correlation between stocks brings new challenge to the stability of the market. A large number of literatures have studied the correlation between industry sectors involving multiple times and more contrastive analysis of the econometric model, but its data indicators are single, mainly based on the closing price. The analyses based on single value data have shortcomings on range

reflection of the stocks, which makes the results easier be affected by outliers and lose comprehensiveness. Our paper proposes a comprehensive weighting method to analyze the characteristics of stock market industry sector in China based on interval data.

We obtain the empirical value of the weight which is calculated as 0.4 by taking CSI 300 as the representative of China's stock market. When applying our method into stock market in China, we firstly analyze the correlation of industry sectors and then take bank sectors for example to analyze the correlation characteristics of sample stocks within one industry sector. The empirical results suggest our method based on interval data can reflect the correlation characteristics of stock market industry sectors preferably and verify the effectiveness of the proposed interval data correlation coefficients. Our method also expands the application field of interval data analysis. Meanwhile, The study on interval data correlation theory is worthy of further discussion:

Firstly, when we apply our method to stock market, we define the range of the interval data as maximum price minus minimum price in one day, and define the absolute value as closing price. However, in different cases, the variables that represent the range and absolute value of the interval data may be different. Besides what we used above, in the future we may discuss the different variables' influence on the value of comprehensive weight. Considering the characteristics of interval data is to reflect both absolute value and change range, we can also define the volatility term of the interval data as opening price minus closing price, and measure the trend term using some other important data points such as maximum price, minimum price, opening price, midpoint, center of gravity, etc. The influence of multiple measurements on our comprehensive weight can be compared as well.

Secondly, expand the application field of our method. When discussing the empirical weight of our comprehensive weighting method, we consider CSI 300 as the representative of stock market in China. In the future, we hope to expand it to other stock markets even other fields to test universality of our method.

Finally, in this paper, the measurement of the interval data correlation is based on extreme cases and the gradient ratio analysis to determine the empirical value of the correlation coefficients. The future work can further discuss other methods to measure the empirical value of weight.

References

[1] Ao. Liang, Fu. Hui-Min, Integral estimate method for interval censored data, Journal of Aerospace Power 22.2(2007) 175–179.
[2] Bertrand. P, Goupil. F, Descriptive Statistics for Symbolic Data, Analysis of Symbolic Data, Springer Berlin Heidelberg, 2000.
[3] Billard. L, E. Diday, Symbolic data analysis: conceptual statistics and data mining, Chichester: Wiley, 2006.
[4] Billard. L, Diday. E, Regression analysis for interval-valued data, In Data Analysis, Classification and Related Methods, Proceedings of the Seventh Conference of the International Federation of Classification Societies (IFCS'00), Belgium: Springer, (2000) 369–374.
[5] Billard. L, Diday. E, Symbolic regression analysis, In Classification, Clustering and Data Analysis, Proceedings of the Eighenth Conference of the International Federation of Classification Societies (IFCS'02), Poland: Springer, (2002) 281–288.
[6] Billard. L, Diday. E, From the statistics of data to the statistics of knowledge: symbolic data analysis, J. Amer. Statist. Assoc, 98 (2003) 470–487.
[7] Bock. H. H, Diday. E, Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data, Heidelberg: Springer-Verlag, 2000.
[8] Cazes. P, A. Chouakria, E. Diday, Y. Schektman, Extension de l'analyse en composantes principales a des domnees de type intervalle, Revue de Statistique appliquée, 45.3 (1997) 5–24.
[9] Chouakria. A, Extension des methodes d'analyse factorielle a des donnees de type intervalle, Ref (1998).
[10] Chouakria. A, P. Cazes, E. Diday, Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data, Berlin: Springer-Verlag, 2000.
[11] Diday. E, Noirhomme-Fraiture. M, Symbolic Data Analysis and the SODAS Software, Chichester: Wiley, 2008.
[12] Diday. E, The symbolic approach in clustering and related methods of data analysis: the basic choices, In Classification and Related Methods of Data Analysis (eds. H.-H. Bock), Proc. of IFCS, 87 (1988) 673–684.
[13] Douzal-Chouakria. A, L. Billard, E. Diday, Principal component analysis for interval-valued observations, Statistical Analysis and Data Mining, 4.2 (2011) 229-246.
[14] Gioia. F, Lauro. C, Principal Component Analysis on Interval Data, Computational Statistics, 21 (2006) 343–363.
[15] Giordani. P, H. Kiers, Principal component analysis of symmetric fuzzy data. Computational statistics and data analysis, 45.3 (2004) 519–548.
[16] Giordani. P, Linear regression analysis for interval-valued data based on the Lasso technique, Technical Report n. 7, Department of Statistical Sciences, Sapienza University of Rome, 2011.
[17] Guo. Jun-peng, Li. Wen-hua, Gao. Feng, Descriptive statistics and analysis of interval symbolic data with general distribution, System Engineering-Theory and Practice, 12 (2011) 2367–2372.

[18] Guo. Jun-peng, Li. Wen-hua, Analysis of validity of the PCA for interval data, Journal of systems engineering, 02 (2009) 226–230.
[19] Hu. Yan, Wang. Hui-Wen, An Interval Data Factor Analysis Method and Its Application, Application of statistics and management, 04 (2004) 53–58.
[20] Lauro. C, F. Palumbo, Principal component analysis of interval data: a symbolic data analysis approach, Computational statistics, 15.1 (2000) 73-C87.
[21] Le-Rademacher. J, L. Billard, Symbolic covariance principal component analysis and visualization for interval-valued data, Journal of Computational and Graphical Statistics, 21.2 (2012) 413C-432.
[22] Li. Wen-Hua, Guo. Jun-Peng, Methodology and application of regression analysis of interval-type symbolic data, Journal of management science in China, 04 (2010) 38–43.
[23] Lima-Neto. E. A, F. D. A. T. D. Carvalho, Centre and range method for fitting a linear regression model on symbolic interval data, Computational Statistics and Data Analysis, Vol. 52 (2008) 1500–1515.
[24] Lima-Neto. E. A, F. D. A. T. D. Carvalho, Constrained linear regression models for symbolic interval-valued variables, Computational Statistics and Data Analysis, 54 (2010) 333–347.
[25] Long. Wen, Dingmu. Cao, The Style and Structure of Chinese Stock Market in 2005-2010: Based on Symbolic Principal Component Analysis, Business Intelligence and Financial Engineering (BIFE), 2012 Fifth International Conference on. IEEE, (2012).
[26] Long. Wen, Guan. Rong, Wang. Hui-Wen, Data analysis based on the symbol of China's stock market style plate market research, Chinese society for the study of modern management, The third (2008) of management annual meeting proceedings China, Chinese society for the study of management modernization, 10 (2008).
[27] Maia. A. L. S, F. D. A. T. D. Carvalho, T. B. Ludermir, Forecasting models for interval-valued time series, Neurocomputing 71 (2008) 3344–3352.
[28] Noirhomme-Fraiture. M, Visualization of Large Data Sets: The Zoom Star Solution, International Electronic Journal of Symbolic Data Analysis, (2002).
[29] Tibshirani. R, Regression shrinkage and subset selection with the lasso, Journal of the Royal Statistical Society B, 58.1 (1996) 267–288.
[30] Wang, Hui-Wen, Li. Yan, Guan. Rong, A Comparison Study of Two Methods for Principal Component Analysis of Interval Data, Journal of Beijing university of Aeronautics and Astronautics (Social Sciences Edition), 04 (2011) 86–89.
[31] Wang. Li-Yuan, Hu. Yan, Wang. Hui-Wen, The Canonical Correlation Analysis of Interval Data and Its Application in the Stock Market, System engineering-theory and practice, 01 (2005) 128–133.
[32] Wang. Yan, Zhang. Yin, Wang. Hui-Wen, Research on the Present Conditions of Journals from Different Disciplines Based on Interval Data Analysis, Application of statistics and management,01 (2012) 134–141.